Large-Scale Public R&D Portfolio Selection by Maximizing a Biobjective Impact Measure
Igor S. Litvinchev, Fernando López, Ada Alvarez, Member, IEEE, and Eduardo Fernández

Abstract—This paper addresses R&D portfolio selection in social institutions, state-owned enterprises, and other nonprofit organizations which periodically launch a call for proposals and distribute funds among accepted projects. A nonlinear discontinuous bicriterion optimization model is developed in order to find a compromise between a portfolio quality measure and the number of projects selected for funding. This model is then transformed into a linear mixed-integer formulation to present the Pareto front. Numerical experiments with up to 25 000 projects competing for funding demonstrate a high computational efficiency of the proposed approach. The acceptance/rejection rules are obtained for a portfolio using the rough set methodology.

Index Terms—Linear multiobjective optimization, mixed-integer linear model, portfolio optimization, public organization R&D projects.

I. INTRODUCTION

PORTFOLIO optimization problems are very well known and intensively studied in capital investment, in stock market, and in R&D project selection in a private sector. However, in the field of public organizations, the management of R&D resources has not received a similar level of attention.

There are two subproblems strongly related with the portfolio optimization [1]: (a) evaluating individual projects and (b) selecting a portfolio of projects that maximize impact. The literature on subproblem (a) is quite extensive. Some of the related papers are, for example, the works of Henriksen and Traynor [2], Lee et al. [3], and Hsu [4].

In this paper, we focus on an R&D portfolio selection problem in public organizations [subproblem (b)], assuming that the projects have been previously evaluated. The following assumptions apply.

1) A budget that is available to fund the projects is known, and a “call for proposals” is announced to solicit proposals in various areas (e.g., areas of knowledge). The support for a particular area is limited from above and below.

2) The decision on funding is taken once a period, i.e., the so-called static project selection problem is considered [5].

3) Typically, the proposals originate from mutually unrelated institutions/individuals, and thus, public R&D projects can be considered as statistically independent with very small or zero correlation.

4) Intangible assets play a dominant role in the quality measure of a portfolio.

5) There are many candidate projects competing for funding.

A subdivision of the proposals by areas was considered earlier, e.g., in [6] and [7]. However, budget constraints for different areas were not taken into account in previous works. Monetary constraints for specific areas, as mentioned in 1), dictate that it is not possible to apply the solutions postulated by previous works that do not take them into account.

Public organizations are usually concerned with a large number of competing projects, which becomes more critical at a national level. For example, the National Science Foundation (NSF), the most important public R&D management organization in the U.S., received thousands of projects in its areas of knowledge in 2007 and 2008, varying from 1000 up to 9000. Specifically, in the Engineering Area alone, about 10 000 proposals were received in 2007 (see http://dellweb.bfa.nsf.gov/awdfr3/default.asp for more details). The Russian Foundation for Basic Research, a public governmental organization that is similar to NSF, considered in 2007 about 65 000 proposals in ten different calls, i.e., an average of 6500 in each call (http://www.rfbr.ru/pics/22172ref/file.pdf). The Mexican National Council for Science and Technology (CONACYT) received about 1800 proposals in 2006 in its Basic Research Call for Proposals (http://www.siicyt.gob.mx/siicyt/docs/contenido/IGECYT_2007.pdf).

Such a large number of projects require the following:

a) arranging and processing information [8];

b) constructing efficient portfolio optimization techniques [6];

c) justifying/explaining the proposed solutions [8].

Point a) has been approached by an organizational decision support system that is capable of managing and coordinating information of thousands of projects. In our work, we concentrate on points b) and c), constructing suitable techniques for large-scale problems.
To the best of our knowledge, in organizations like NSF, CONACYT, etc., the number of projects competing for funds has increased in each call for proposal in the past ten years. As this growth has, in general, negatively impacted on the quality of the portfolio, NSF, for example, has created working groups to recommend policies to improve the process (http://www.nsf.gov/pubs/2007/nsf0745/nsf0745.pdf).

The aims of this study are aligned with some of the final recommendations made by these working groups, which are also valid for any public or social organization funding R&D projects in a call-for-proposal basis.

Some weak heuristics are popular approaches to address point b). For instance, the approach used by CONACYT, NSF, and other organizations consists mainly of the following steps [9], [20]:

A) distributing projects according to knowledge areas;
B) verifying if the project meets the formal submission criteria;
C) evaluating project attributes by peers using a numerical scale and summing up the grades to get the overall evaluation of the project;
D) assigning funds (by supervisors) according to the ranking generated by the evaluation of each project.

The main point of this approach is to obtain the ranking of projects according to their quality and then to assign funds according to this ranking.

However, using this simple heuristic, portfolio impact is not optimized. It is also important to note that portfolio selection is a decision problem over the set of portfolios but not over the set of projects. Thus, it is necessary to compare the portfolios but not the individual projects. To illustrate this argument, consider the situation in which the peers assign a score of 82 points to project A and 80 points equally to projects B and C. Let the funding required by A be sufficient to support both projects B and C. Then, a higher impact may be obtained by funding both B and C, despite the fact that project A has a stronger individual ranking. Therefore, a reasonable objective could be to maximize the portfolio’s global impact as determined by the number of supported projects and their overall quality.

This problem has been approached in several papers [9], [20]. To some extent, the model used in these works represents the portfolio quality by merging implicitly in a single measure the number of supported projects and their specific evaluations. However, the previous approaches consider the model of portfolio quality as a unique criterion, assuming implicitly reliable and perfect preference information available to determine the model’s parameters. Together with some rough approximations used in the derivation of the model, this is the main reason to consider more explicit and easily measurable criteria to compare portfolios. For example, the number of projects supported in the portfolio may be considered as an explicit objective.

In this work, the problem of portfolio selection is addressed by considering a biobjective formulation. In the solution process of a multiobjective optimization model, many runs of a single-objective optimization algorithm are required to find the best compromise solution. Therefore, the selected optimization algorithm should be highly efficient for instances with several thousands of projects that arise when the concerned governmental organizations distribute funds among competing projects. Moreover, the response times should be very short to keep the attention of the decision maker and enable him or her to explore a larger part of the feasible portfolio’s space. This is an important drawback in [9] and [20], where a time-consuming single-objective evolutionary algorithm was used. In contrast, we present an approach based on a very efficient mixed-integer linear model that makes it possible to compute the Pareto optimal front in seconds, even for instances with several thousands of projects.

In many cases, the portfolio selection is associated implicitly with certain acceptance/rejection rules for projects. From a practical point of view, it is very important to discover and state these rules explicitly. To the best of our knowledge, this issue has not been studied for portfolio problems in public organizations. In our work, this point is addressed by the rough set (RS) methodology [10], [11]. Once the portfolio has been selected, we construct and analyze associated acceptance/rejection rules.

In summary, the contributions of this paper are as follows: 1) to improve the preference model concerning portfolio quality by considering a more flexible biobjective formulation; 2) to propose an efficient solution approach that is capable of computing the nondominated front of large-scale instances; and 3) to propose a way for deriving and explaining the acceptance/rejection rules of the selected portfolio.

The remainder of this paper is organized as follows. In Section II, a brief discussion of the earlier results is presented. A nonlinear discontinuous biobjective model is explained in Section III, while in Section IV, a mixed-integer linear formulation of this model is considered. Section V focuses on calculating the best compromise solutions by the mixed-integer model. Simple decision rules are derived and analyzed in Section VI. Numerical results are reported in Section VII, while the conclusions can be found in Section VIII.

II. SOME BACKGROUND

Multiobjective models to select “the best” R&D project portfolio have been studied, e.g., in [6] and [12]–[14]. For a detailed discussion of previous works, see [5] and [15]. However, all these models focus on portfolios with a relatively few number of projects and are not applicable to large instances with thousands of competing projects. To the best of our knowledge, R&D project portfolio selection for very large instances was first considered by Tian et al. [8] and Klapa and Pinos [6]. While the former focused more on decision support from an organizational point of view, the latter solved portfolio problems with up to 250 projects. It was noted that increasing the number of projects over 250 results in a considerable increase of computation time, but solutions to larger instances were not presented.

In this work, we use basic constructions that are similar to those by Fernández et al. [9], [20]. In the following, we present a short overview of the main results.

First of all, it is necessary to describe the subject of this particular decision problem. When multiple criteria are involved, it is not possible to solve a decision problem without
considering the subjective information from the decision maker. This information is needed when handling the tradeoffs between conflicting goals. Thus, one should accept the relevance of an entity (a person or a group) that represents the interest of the funding organization. This entity is called the supra decision maker (SDM). The preferences and beliefs of this SDM should be modeled to evaluate and select projects.

In [9] and [20], the authors established the following propositions to explain why the risk attitude of the SDM moves away from aversion.

A. The projects are assumed to be statistically independent, and there is no correlation among them. Diversification, an important issue in the portfolio investment problem [16], is presented in a public R&D portfolio problem in a natural way. Although the benefits of diversification with negative correlation cannot be obtained, statistical independence makes it unlikely to obtain weak global results.

B. Since the projects compete for public funding, the group of stakeholders that constitute or represent the SDM is not the owner of the funds that are distributed among the projects (after all, it is public money).

C. The group of stakeholders manages a budget for supporting research projects. This budget will never be considered as a loss but always as an investment. Whenever an acceptable project appears, the SDM is ready to support this project. A possible failure of the project is considered as a lost opportunity rather than a monetary loss.

In the zone of average prizes, where the mass of probability is concentrated, it is natural to suppose that the SDM behaves neutrally toward the risk. Then, a linear utility model seems to be suitable for representing the SDM’s risk attitude in that zone [17]. Some deviations from the linear form may occur in the zones of very high prizes. However, since the probability of these outputs is very low, the expected value of the linear utility model appears to be the SDM’s risk attitude in that zone [17].

The most important parameters of the portfolio quality model proposed in [9] are the following:

\[ N \] total number of projects;
\[ K \] total number of areas to which projects belong;
\[ M_j \] amount of money requested in the project \( j \) proposal;
\[ m_j \] minimal amount of money, suggested by a peer or a decision maker, to guarantee success of project \( j \);
\[ P_G \] total amount of funds available to be distributed among the projects;
\[ P_i^+ \] maximal amount of funds to be distributed among the projects in area \( i \);
\[ P_i^- \] minimal amount of funds to be distributed among the projects in area \( i \);
\[ w_j \] importance assessed by the SDM to project \( j \).

Decision variables \( d_j \) are funds assigned to project \( j \).

In (1), \( x_i = 1 \) if the \( i \)th project is funded and \( x_i = 0 \) otherwise. By definition, \( C' \) is the sum of the certainty equivalents of the projects in the portfolio.

Let \( I = x_1I_1 + x_2I_2 + \cdots + x_NI_N \) be the impact of the entire portfolio. From (1), it follows that

\[ C' = E(I). \]

Let \( C'' \) denote the portfolio’s certainty equivalent. In the linear case, \( C'' = E(I) \), and \( C'' = C' \). Under the assumption of linear utility, the certainty equivalent of the portfolio is equal to the sum of the certainty equivalents of the projects that comprise the portfolio. Taking into account statements A, B, and C, we may use for \( C'' \) the same expression as that in (1).

For \( m_j \leq d_j < M_j \), the SDM is not sure if the project is adequately funded. Then, the proposition that the \( j \)th project is adequately funded can be considered as a fuzzy statement with a degree of truth \( \mu_j(d_j) \). Therefore, (1) can be modified to reflect the fuzziness of the monetary resources handled by each project. The degree of truth \( \mu_j(d_j) \) is presented by a piecewise linear function, increasing on \([m_j, M_j]\), such that \( \mu_j(M_j) = 1 \), \( \mu_j(m_j) = \alpha_j < 1 \), and \( \mu_j(d_j) = 0 \) for \( d_j < m_j \) (see Fig. 1).

Here, \( \alpha_j \) represents the minimal level of membership to the fuzzy predicate that “project \( j \) is adequately funded.”

In (1), \( x_j \) may be considered as the indicator function of the set of supported projects. As fuzzy sets could offer more flexibility than sets, Fernández et al. [20] propose describing \( C'' \) as

\[ C'' \approx C' \approx \sum_{j=1}^{N} c_j \mu_j(d_j) \]

where \( \mu_j(d_j) \) is a membership function which expresses the grade of membership of project \( j \) to the portfolio.
Consequently, the set of supported projects has been transformed into a fuzzy set.

Since the certainty equivalents are given in a ratio scale, \( c_j \) can be replaced by a proportional parameter \( w_j \), and the function

\[
Q = \sum_{j=1}^{N} w_j \mu_j(d_j)
\]

(4)

is a cardinal function on the set of portfolios, i.e., a measure of the portfolio’s quality.

It was assumed that equally evaluated projects from a given area of knowledge have equal certainty equivalents. The set of projects can be partitioned into classes; projects in the same area of knowledge have equal certainty equivalents. The set of the portfolio’s quality.

\[
j \text{ solves certain indifference equations between portfolios. The ratios } w_i/w_j \text{ correspond to a comparison of the respective certainty equivalents. The minimum “weight” } w_L = 1 \text{ corresponds to the lowest class of acceptable projects. Thus, } Q \text{ becomes}
\]

\[
Q' = \sum_{j=1}^{N} w_j' \mu_j(d_j)
\]

(5)

where \( w_j' \) is the common estimated weight for the class to which the project \( j \) belongs to. To estimate \( w_j' \), the decision maker is asked to express strict preference or indifference between two elementary portfolios: one portfolio containing \( n_L \) lowest class projects and the other portfolio containing \( n_j \) projects from the same class that project \( j \) belongs to (\( n_j < n_L \)). By varying \( n_j \) and \( n_L \), an indifference condition should be reached for some values \( n_j^* \) and \( n_L^* \). Then, \( w_j^* = n_j^*/n_L^* \).

The most important sources of imprecision in \( Q' \) come from the following assumptions.

i) The SDM is risk neutral.
ii) Projects in the same class have equal certainty equivalents.
iii) The certainty equivalents of projects are reduced by a factor \( \mu \), as can be seen in (3).
iv) For each class in the project’s space, the SDM is able to express strict preference or indifference between “elementary” portfolios.

The “best” portfolio is obtained by solving the following problem:

\[
\begin{cases}
\max & Q' = \sum_{j=1}^{N} w_j \mu_j(d_j) \\
\text{s.t.} & d \in R_F
\end{cases}
\]

(6)

where \( d = (d_1, \ldots, d_N) \). The feasible region \( R_F \) is determined by the limits for the overall funds, as well as for funds in the specific areas.

III. MODEL OF IMPRECISION BASED ON A BIOBJECTIVE OPTIMIZATION APPROACH

The numerical approach presented in [9] and [20] is difficult to apply to problem instances with thousands of projects. Meanwhile, this scale is typical for many public or social organizations funding R&D activities at a nationwide level. Evolutionary metaheuristics converge very slowly and are not effective for large instances with many projects. Moreover, the authors did not suggest a way to handle the imprecision of the portfolio’s quality model presented earlier.

The main problem is to handle the contradiction between premises ii) and iv). The process of solving indifference equations is difficult and time consuming for the SDM, and its difficulty increases with the number of classes in the project space. On the other hand, the degree of credibility of assumption ii) decreases with the number of classes. Considering also the assumption of linear utility and certain inherent imprecision coming from the process of project valuation, the SDM may hesitate about the real value of small improvements of \( Q' \).

Following [19], we may say that imprecision makes \( Q' \) a pseudocriterion.

To illustrate an imprecision, consider two feasible portfolios \( S_1 \) and \( S_2 \). Denote \( Q'_1 = Q'(S_1) \) and let \( N_1 \) be the cardinality of \( S_1 \) and, similarly, \( Q'_2, N_2 \) for \( S_2 \). Due to the imprecision, there exists a threshold \( \varepsilon \) such that, for \( |Q'_2 - Q'_1| \leq \varepsilon \), the SDM cannot express a strict preference between \( S_1 \) and \( S_2 \), no matter if \( Q'_2 - Q'_1 \) is positive or negative. We can distinguish the following situations.

a) If \( |Q'_2 - Q'_1| \leq \varepsilon \) and \( N_2 > N_1 \), the SDM has a reasonable element to prefer \( S_2 \).

b) If \( |Q'_2 - Q'_1| \leq \varepsilon \) and \( N_1 > N_2 \), the SDM has a reasonable element to prefer \( S_1 \).

c) If \( Q'_2 - Q'_1 > \varepsilon \) and \( N_2 > N_1 \), the SDM can state a clear preference favoring \( S_2 \).

d) If \( Q'_1 - Q'_2 > \varepsilon \) and \( N_1 > N_2 \), the SDM can state a clear preference favoring \( S_1 \).

e) If \( Q'_1 - Q'_2 > \varepsilon \) and \( N_2 > N_1 \), the SDM may hesitate between preference and indifference.

f) If \( Q'_2 - Q'_1 > \varepsilon \) and \( N_1 > N_2 \), the SDM may hesitate between preference and indifference.

For cases e) and f), the final decision depends on the following: the magnitude of \( |Q'_2 - Q'_1| \), the difference \( |N_2 - N_1| \), and the confidence of the SDM on the model \( Q' \). Therefore, portfolio \( S_2 \) is preferred to \( S_1 \) if the first one is a better compromise solution to the bicriterion optimization problem

\[
\max_{S \in R_F} Q'(S), N(S)
\]

(7)

where \( N(S) \) denotes the cardinality of portfolio \( S \).

The nonlinear biobjective mathematical model is then as follows. First, the feasible cardinality of portfolio \( S \).

The nonlinear biobjective mathematical model is then as follows. First, the feasible region is defined by the following constraints:

\[
\sum_{j=1}^{N} \delta_j \geq \Omega
\]

(8)
The complexity of the problem \([15]\) and \([16]\) results basically from the discontinuity of the functions \(\mu_j\) used to characterize the portfolio’s quality. Thus, it is difficult to apply directly heuristic evolutionary algorithms or exact methods. However, the problem structure suggests that transforming the objective function and the feasible region may lead to more tractable formulations.

**IV. LINEAR MIXED-INTEGER MODEL**

Note that, in the problem \([15]\) and \([16]\) subject to constraints \([8]\)–\([14]\), all the components are linear except for the functions \(\mu_j\).

Following the proposal in [20], we present these piecewise functions in a linear way using auxiliary binary variables. The objective is to transform the original nonlinear model into a linear mixed-integer formulation.

Let \(x \) be the amount of funds assigned to the project \((d \) in the previous model), where the project index is omitted for simplicity. Define parameters \(\theta \) and \(\gamma \) such that \(\theta + \gamma x = \alpha \) for \(x = m \) and \(\theta + \gamma x = 1\) for \(x = M\). It is not difficult to verify that the corresponding values are

\[
\theta = \alpha - m(1 - \alpha) \quad \gamma = \frac{1 - \alpha}{M - m}.
\]

Let \(y = 1\) if the project is in the portfolio and \(y = 0\) otherwise. Consider the function \(\mu(y, x) = \theta y + \gamma x\) for \(my \leq x \leq M y \) and \(y \in \{0, 1\}\). For \(y = 1\), \(\mu(1, x)\) coincides with the function \(\mu\) in Fig. 1 for \(x \in [m, M]\). For \(y = 0\), we have \(\mu(0, x) = 0\), or better to say, only \(x = 0\) fulfills restriction \(my \leq x \leq M y\), i.e., if we assume that the project will never be funded higher than is requested in the proposal \((M\) and the project not included in the portfolio is not funded at all, our function \(\mu(y, x)\) coincides with the function presented in Fig. 1.

Then, the problem of finding compromise solutions corresponding to the biobjective model \([15]\) and \([16]\) subject to constraints \([8]\)–\([14]\) can be stated as follows:

\[
\max \pi \sum_{j=1}^{N} w_j(\theta_j y_j + \gamma_j x_j) + (1 - \pi) \sum_{j=1}^{N} y_j
\]

subject to

\[
m_j y_j \leq x_j \leq M_j y_j, \quad j \in \{1, \ldots, N\}
\]

\[
\sum_{j=1}^{N} x_j \leq P_G
\]

\[
P_i^- \leq \sum_{j \in J_i} x_j \leq P_i^+, \quad i \in \{1, \ldots, K\}
\]

\[
y_j \in \{0, 1\}.
\]

The decision variables are \(x_j\) and \(y_j\). Note that, by \([19]\), the project is either supported sufficiently \(x_j \in [m_j, M_j]\) or not supported at all \(x_j = 0\). In the latter case, the project is not included in the portfolio, i.e., constraints \([19]\) and \([22]\) are equivalent to \([13]\) and \([14]\).

The term \(\sum_{j} w_j(\theta_j y_j + \gamma_j x_j)\) in the objective function \([18]\) represents the measure of the portfolio’s quality, \(\sum_j y_j\) gives the total number of (sufficiently funded) projects in the portfolio, and \(\pi \in [0, 1]\) is a weighting factor that allows us to generate efficient Pareto solutions. This model makes restrictions \([8]\) and \([9]\) unnecessary, given that the effect can be achieved by changing the value of \(\pi\).

**V. HOW TO OBTAIN THE BEST COMPROMISE SOLUTION**

The model \([18]\)–\([22]\) is used to get compromise solutions between both objectives (the portfolio’s quality and the number of supported projects). First, a Pareto front is generated, moving \(\pi\) from zero to one with a certain step size. The optimal objective \([18]\) for \(\pi = 1\) corresponds to the best value of the portfolio’s quality, while \(\pi = 0\) provides the largest number of supported projects.
The generation of a Pareto front is a relatively fast process, a fact that has been observed by the authors in all conducted experiments, including the instance of 25 000 projects.

A linear relaxation of the problem (18)–(22) was also solved, considering variables $y_i$ as real numbers from the interval $[0, 1]$. In all the tested instances, the linear relaxation provided a very tight bound for the optimal value of (18). This partially explains the high efficiency of the commercial software package for large instances of the mixed-integer linear model (18)–(22). To get a reasonable presentation of the Pareto front for all instances, it was enough to solve the problem (18)–(22) for 30–40 values of $\pi$.

In all the tests carried out, it was found that the solutions of the Pareto front formed a concave curve, which makes it easier for the decision maker to identify a region of interest and to find out the best compromise solution in that region.

The exploration of the zone of interest can be done using the weighting method that is commonly used for the generation of the Pareto front or, instead, using a reference-point-based method. Then, a “zoom” is done looking for new solutions providing better compromises than the current efficient points.

During the exploration process, the decision maker evaluates or compares solutions, and this task becomes easier if the reference elements (ideal point and nadir) are given. The decision maker can also take as a reference point the solution obtained by the ranking heuristics. The latter are commonly used in large organizations that manage R&D.

The aforementioned procedure allows a fast convergence to the best compromise. The decision maker may take as a reference point the solution obtained for $\pi = 1$ and then move further if he/she does not trust the quality model. This distrust can be expressed by means of a percentage of deviation that he/she is willing to tolerate to incorporate a greater number of projects in the portfolio. Once a satisfactory solution has been found, it can be compared to the heuristic one, realizing how good this solution is in relation to this last one. The process ends when the decision maker chooses what he/she considers a good compromise solution.

We have not found models that provide a simple justification for a portfolio selection. Explaining and justifying a portfolio selection to stakeholders is an important asset to a model and is considered in the following section.

VI. DECISION RULES REPRESENTING A FUNDING POLICY

To build up a portfolio of projects is a complex decision, which is of strategic importance in the organization that is directly responsible for the portfolio’s impact and consequences. Therefore, it is usual that the organization or government, which is in charge of deciding how the portfolio should be built up, expresses its global preferences and beliefs through organization policies. These policies are often a set of if...then...rules (or can be translated into if...then...rules). If a portfolio is obtained by an optimization process, it is desirable to know how close or far it is from the organization policies. This acquires more relevance in the case where some different portfolios are to be presented as recommendations to the SDM or stakeholders. That is why it is important to explain in simple terms why projects are accepted or rejected from a portfolio.

Simple decision rules make it easier to understand the reasons for the decision maker’s judgment, an important issue when she/he has to present her/his conclusions to the SDM or to a stakeholder. Frequently, for any suggested portfolio, decision rules are derived to describe in a synthetic way the funding policy supposedly used to form the portfolio. These rules are then presented to the SDM/stakeholders, together with the original portfolio.

According to [21] and [22], methods for knowledge management and knowledge discovering have been employed in portfolio selection or portfolio optimization to integrate soft and quantitative data in order to make predictions in future market assets. However, there is no evidence that such methods were employed to support the proposal of a solution to the decision makers as it is proposed in this paper.

A. Decision Table

A decision table is an attractive way for a decision maker to express her/his preferences. Slowinski [11] and Greco et al. [23] proved that the logical rules obtained from a decision table are as powerful, for modeling preferences, as other decision supporting methods. The advantage is imposing neither axiomatic requirements about the decision maker’s behavior nor the essence of the decision problem.

In a decision table, there are a set $C$ of condition attributes (those characterizing the objects) and a set $D$ of decision attributes characterizing the preferences of decision makers, where $C \cap D = \emptyset$. Each attribute is represented as a column in the decision table, and the rows of the table correspond to objects. In our case, the objects are projects.

In this paper, they were chosen as condition attributes, the project data included in the optimization model: $m$ is the minimal funding required, $M$ is the maximal funding, $A$ is the knowledge area, and $w$ is the project evaluation. As decision attributes, we chose the funds assigned to a project in the portfolio. Thus, a row of the table represents a project which is described in terms of condition and decision attributes.

Moreover, each row of the decision table can be viewed as a decision rule (where the values in the decision columns are a consequence of values in the condition columns). However, such a set of decision rules can be huge and indeed of little value to extract knowledge about the decision situation represented by the decision table. There are several methods for reducing a set of decision rules.

In this paper, the methodology of RS [24] is used to derive decision rules from a decision table. It was reported in [25] that the RS technique provides better results than a neural network approach and the well-known algorithms C45 and C50.

B. General Description of RS Theory

In this section, we present the main aspects of a classical RS theory.
RS theory rests on the idea of classification, where classification is understood as the ability of an agent (human, robot, etc.) or a group of agents to discern certain phenomena, process, objects, etc., from a concrete reality. Then, knowledge in this context is understood as a “family of classifications” by different patterns in one domain of interest. This family of classifications represents explicit facts of a concrete reality, along with the capacity of deriving or discovering implicit facts from explicit knowledge [26].

This theory is easy to understand and has the following important advantages [27].

1) It provides efficient algorithms for finding hidden patterns in data.
2) It finds reduced sets of data (data reduction).
3) It evaluates the significance of data.
4) It generates minimal sets of decision rules from data.
5) It offers straightforward interpretation of results.
6) It can be used in both qualitative and quantitative data analyses.
7) It identifies relationships that would not be found using statistical methods.

Data analyzed with RS theory are structured in a set of objects (observations, judgments, states, etc.) described by a set of multivalent attributes. A direct representation of the set of objects in a decision table is straightforward.

Let $S = (U, R)$ be a decision table, where $U$ is a set of objects and $R = C \cup D$ is a set of attributes. $C$ and $D$ are the condition and decision attribute sets as defined previously. From a mathematical point of view, a classification in $S$ is nothing more than an equivalence relationship defined in $U$, so in what follows, the term “equivalence relationship” will be used instead of “classification.” The minimal sets of equivalence relationships that ensure the same quality in the approximation instead of “classification.” The intersection for the whole set is referred to as reducts; the intersection of all of them is called a core. Neglecting the elements that form the core leads irremediably to the loss of quality in the classification.

The previously described elements constitute the base of the algorithms which are used for reducing the rows and columns in a decision table without affecting the classification power of the system. They allow obtaining decision algorithms associated to the table.

### C. Applying RS

Before applying the methodology of RS, the condition attribute space is made discrete using the grades (scales) too expensive, expensive, standard, inexpensive, and very inexpensive for attributes $m$ and $M$ and the grades excellent, very good, good, standard, acceptable, more or less, low, and worst for attribute $w$. There are many reasons to conclude that the SDM and peers prefer to use scales with stages described in a natural language [28].

Once the decision table was formed, the rules have to be obtained, calculating the reducts (subsets of condition attributes that generate the same knowledge as the whole set of row descriptors) by columns and rows of the table. In our experience, the best procedure to obtain the reducts was the one based on Learning from Examples Module, version 2 (LEM2) [24]. Finally, the rules are obtained by combining the reducts by rows and columns, to achieve the minimal number of rules with the minimal number of descriptors, while keeping a certain level of preestablished covering and prediction capacity.

Thus, we arrive at a minimal set of decision rules representing the decision policy (relationship between the condition and decision attributes) for the original portfolio. Using this tool, a decision agent can explain, for example, how the very expensive projects are dealt with, whether the excellent projects are properly supported or not, and why the area of the project is important in the assignment. This is an important part of the portfolio optimization process which is necessary in endorsing the portfolio to a higher level decision maker.

A global coverage of 0.952 and a global accuracy of 0.753 for the 25 000-project instance were obtained using RS. For this particular instance, 61 rules were generated.

### VII. Computational Results

Five instances with 40, 400, 1200, 10 000, and 25 000 projects were tested, while $\pi$ was moved from zero to one with a step size of 0.03. In the worst case, the run time did not exceed 20 s on a Sunfire computer with four processors.

In the case of 40 and 400 projects, the instances are the same as that in [29], while the other instances were obtained by using an instance generator developed in [25]. The Pareto front for all the instances has a very similar behavior, so only the Pareto front for the instance with 25 000 projects is shown in Fig. 2.

The solutions corresponding to the Pareto front for 25 000 projects are shown in Table IV.

The commercial software CPLEX v9.0 of ILOG was used for optimization. Some results are shown in Table I.

In Table I, the rows corresponding to $\pi = 1$ present the solutions maximizing the portfolio’s quality subject to constraints (19)–(22). The fourth column gives the degradation of this quality value with respect to its best performance (maximum value). The fifth column shows the increment of the number of supported projects with respect to the solution obtained for $\pi = 1$. For each instance, the first row (marked with $r$) shows the indicators for the solution obtained by ranking heuristics.
We can compare our results with those previously reported for the case of quality maximization for 40 and 400 projects considered in [9]. We were unable to find references for larger instances. Fixing \( \pi = 1 \) in (18)–(22) for the 40-project instance, we obtained 28 projects supported, with the overall portfolio quality 156.574, which, in fact, coincides with the results in [9]. For the problem with 400 projects, our results are also very similar to that in [9] (see Table I).

The use of our biobjective proposal allows identifying other interesting compromise solutions. The decision maker may find a better compromise by balancing the trust in the quality model, with the importance to the increase in the number of supported projects. In Table I, the rows for \( \pi = 1 \) correspond to the total trust in the model of the portfolio’s quality, as opposed to rows with \( \pi = 0.01 \) (\( \pi \approx 0 \)), corresponding to the total distrust.

The solutions corresponding to \( \pi \in [0.3, 0.5] \) present small degradations in the quality model of the portfolio. In these cases, the decision maker basically trusts the quality model but still acknowledges that it has some imprecision. Consider, e.g., the problem with 400 projects. From Table I, we see that, for \( \pi = 0.5 \), the portfolio quality is 1273.93 with 304 projects accepted, i.e., if the decision maker is ready to accept degradation near 1% in quality, he can get a 10.94% increase in the number of accepted projects. This seems to be very reasonable since errors below 5% can be considered inside the range of imprecision described in Sections II and III.

Performing a similar analysis, better compromise solutions for the instances with 1200, 10,000, and 25,000 projects can be identified. These solutions are shown in Table I. To illustrate the analysis, consider the instance with 25,000 projects. The 10.26% increase in the number of supported projects means (in absolute numbers) more than 900 additional research proposals funded. Note that we compared the number of projects obtained by optimizing the quality measure only, where the increase was virtually achieved without degrading the optimal quality value.

For a real-world decision maker with moderate trust in the portfolio’s quality, it seems reasonable to select the portfolio generated for \( \pi \in [0.3, 0.4] \) if he/she accepts a deviation of about 2%–4% in quality, to obtain an increase of 11%–20% in the number of accepted projects. If, for the same instance of 25,000 projects, the decision maker has less trust in the quality model, then he/she can accept a higher deviation (e.g., up to 10%).

Note that, for all problem instances, the solutions obtained by ranking heuristics (row \( r \)) are clearly dominated by the Pareto solutions to the proposed bicriterion problem. This takes place for a wide range of \( \pi \). For example, for a 25,000-project instance with \( \pi = 0.2 \), our solution is similar in quality to the ranking heuristic solution. Meanwhile, our solution has 56% more accepted projects than the ranking solution.

Once the best compromise portfolio has been found, the next step could be to expose implicit policies used in the project selection. The objective of this step is to explain the decision to higher level authorities or even to the public. To do this, the procedure described in Section V is applied. Since the first step is to build the decision table, the decision maker first has to make the condition attributes discrete. Our approach is presented in Table II. In the decision attribute space, only three

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m ) (thousand of dollars)</td>
<td>(0.120) – Inexpensive (0.230, 0.350) – Average ( (0.350, \infty) ) – Very Expensive</td>
</tr>
<tr>
<td>( M ) (thousand of dollars)</td>
<td>(0.120) – Inexpensive (0.230, 0.350) – Average ( (0.350, \infty) ) – Very Expensive</td>
</tr>
<tr>
<td>( w ) (in real numbers)</td>
<td>(0.4, 5.7) - Low (5.7, 8.2) - Acceptable (8.2, 9) - Good ( (9, \infty) ) - excellent</td>
</tr>
</tbody>
</table>

We can compare our results with those previously reported for the case of quality maximization for 40 and 400 projects considered in [9]. We were unable to find references for larger instances. Fixing \( \pi = 1 \) in (18)–(22) for the 40-project instance, we obtained 28 projects supported, with the overall portfolio quality 156.574, which, in fact, coincides with the results in [9]. For the problem with 400 projects, our results are also very similar to that in [9] (see Table I).

The use of our biobjective proposal allows identifying other interesting compromise solutions. The decision maker may find a better compromise by balancing the trust in the quality model, with the importance to the increase in the number of supported projects. In Table I, the rows for \( \pi = 1 \) correspond to the total trust in the model of the portfolio’s quality, as opposed to rows with \( \pi = 0.01 \) (\( \pi \approx 0 \)), corresponding to the total distrust.

The solutions corresponding to \( \pi \in [0.3, 0.5] \) present small degradations in the quality model of the portfolio. In these cases, the decision maker basically trusts the quality model but still acknowledges that it has some imprecision. Consider, e.g., the problem with 400 projects. From Table I, we see that, for \( \pi = 0.5 \), the portfolio quality is 1273.93 with 304 projects accepted, i.e., if the decision maker is ready to accept degradation near 1% in quality, he can get a 10.94% increase in the number of accepted projects. This seems to be very reasonable since errors below 5% can be considered inside the range of imprecision described in Sections II and III.

Performing a similar analysis, better compromise solutions for the instances with 1200, 10,000, and 25,000 projects can be identified. These solutions are shown in Table I. To illustrate the analysis, consider the instance with 25,000 projects. The 10.26% increase in the number of supported projects means (in absolute numbers) more than 900 additional research proposals funded. Note that we compared the number of projects obtained by optimizing the quality measure only, where the increase was virtually achieved without degrading the optimal quality value.

For a real-world decision maker with moderate trust in the portfolio’s quality, it seems reasonable to select the portfolio generated for \( \pi \in [0.3, 0.4] \) if he/she accepts a deviation of about 2%–4% in quality, to obtain an increase of 11%–20% in the number of accepted projects. If, for the same instance of 25,000 projects, the decision maker has less trust in the quality model, then he/she can accept a higher deviation (e.g., up to 10%).

Note that, for all problem instances, the solutions obtained by ranking heuristics (row \( r \)) are clearly dominated by the Pareto solutions to the proposed bicriterion problem. This takes place for a wide range of \( \pi \). For example, for a 25,000-project instance with \( \pi = 0.2 \), our solution is similar in quality to the ranking heuristic solution. Meanwhile, our solution has 56% more accepted projects than the ranking solution.

Once the best compromise portfolio has been found, the next step could be to expose implicit policies used in the project selection. The objective of this step is to explain the decision to higher level authorities or even to the public. To do this, the procedure described in Section V is applied. Since the first step is to build the decision table, the decision maker first has to make the condition attributes discrete. Our approach is presented in Table II. In the decision attribute space, only three
grades are used: 1 represents the complete funding of a project, 0.5 corresponds to the minimal funding, and 0 means that the project is not supported at all. Then, the software RSES v2.0 was used to obtain the rules related to the decision table, while the algorithm LEM2 was used to obtain the reducts. Table III presents the indicators characterizing the quality of the obtained rules for the 25 000-project instance.

For the problem instance with 25 000 projects, 61 rules were obtained initially and then reduced to 19 rules by further analysis. Note that the rules will not be used to create knowledge; they will only be used to summarize the portfolio’s information for the decision maker. The rules are given in Table V. Based on the rules obtained, we may conclude the following.

1) There is a notable trend to support higher impact projects, even if those are considered very expensive.
2) The order of established preference is maintained upon project evaluation and project areas.
3) Inexpensive projects are supported without considering their impact level.
4) If projects are expensive or very expensive and have a low impact level, they are rejected.

By means of these simple statements, it is possible to explain and summarize the decision policy that provides the specific portfolio chosen as the best compromise.

VIII. CONCLUSION

The biobjective model proposed here allows us to reflect the degree of confidence of the decision maker to the model of a portfolio’s quality. This model is used to find a reasonable compromise between the quality measure and the number of supported projects. The solution approach is based on its transformation into a linear mixed-integer optimization model which results in a significant reduction of computational time, particularly compared with previous heuristic approaches.

Decision rules were derived to support a simple explanation of the funding assignment policy. These rules help the decision maker in justifying her/his recommendations for the stakeholders and provide new possibilities for further analysis of a solution.

An interesting area for future research is to include additional resource constraints at the portfolio and/or project level. Moreover, synergies between projects could be considered, although this is more typical for social projects than for R&D projects in public organizations. Further study of the decision rules can
be helpful in complicated decision scenarios, particularly when more than two objectives are considered.

**APPENDIX**

See Tables IV and V.

**ACKNOWLEDGMENT**

The authors would like to thank the three anonymous reviewers for the constructive comments, which helped improve the quality of this contribution.

**REFERENCES**


Ada Alvarez (M’08) received the B.Sc. degree in mathematics from the University of Havana, Havana, Cuba, in 1982 and the Ph.D. degree in mathematics from the Universidad Central “Marta Abreu” de Las Villas, Santa Clara, Cuba, in 1993. She is currently an Associate Professor with the Systems Engineering Postgraduate Program, Autonomous University of Nuevo León, San Nicolás de los Garza, Mexico. Her research interests include combinatorial optimization and heuristics for hard optimization problems, with applications to real-world problems. Dr. Alvarez is a member of the Institute for Operations Research and the Management Sciences and the Operational Research Society.

Eduardo Fernández was born in Cuba in 1951. He received the B.Sc. degree in physics from the University of Havana, Havana, Cuba, in 1973 and the Ph.D. degree in computer-aided design of electronic circuits from the Poznań University of Technology, Poznań, Poland, in 1987. He is currently a Senior Professor with the Faculty of Engineering, Autonomous University of Sinaloa, Culiacán, Mexico. His main areas of interest are mathematical decision models and intelligent decision support systems. He has published 35 refereed papers in those fields. Prof. Fernández is a member of the Mexican National System of Researchers, the International Society on Multiple Criteria Decision Making, the Multicriteria Evaluation and Decision Spanish–American Network, and the EUREKA Network for Knowledge Discovery and Decision Making. He has been nominated twice for the “OR in Development” Prize.