Proceedings of the 1999 Winter Simulation Conference
P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds.

INPUT MODELS FOR SYNTHETIC OPTIMIZATION PROBLEMS

Charles H. Reilly

Department of Industrial Engineering and Management Systems
University of Central Florida
P.O. Box 162450
Orlando, FL 32816, U.S.A.

ABSTRACT

In this paper, we describe and discuss alternative input models for the coefficients in synthetic optimization problems. Synthetic, or randomly generated, problems are often used in computational studies to establish the efficacy of solution methods or to facilitate comparative evaluations of solution methods. The selection of an input model for the coefficients in synthetic optimization problems is important because such a selection may affect the outcome of a computational study. Understanding how an assumed input model affects the characteristics of test problems can assist researchers in their efforts to accurately quantify and interpret the performance of solution methods.

1 INTRODUCTION

When conducting a computational experiment with solution methods for optimization problems, a researcher has to decide on which test problems will be used. Some options are to use real-world test problems, perturbations of real-world test problems, or synthetic test problems. In a simulation context, we can think of these options as corresponding to the use of empirical distributions, fitted-empirical distributions, and theoretical distributions, respectively. Another option is to use standard libraries of test problems. The experimenter's decision on test problem selection can affect the inferences that may be drawn from the results of the computational experiment.

Each option for test problem selection has its pros and cons. For instance, the use of real-world test problems has the advantage of providing results consistent with those for, at least some, example problems encountered in practice. The principal disadvantage is that there may not be a sizable set of such problems to constitute a satisfactory experiment.

Perturbations of real-world examples provide a larger sample space from which to draw test problems. However, the variability across test problems may not adequately represent the variability among practical problem instances.

Libraries of standard test problems can facilitate comparisons across computational studies. However, the inference space is limited to the problems contained in the library.

Synthetic test problem generators offer a virtually infinite supply of test problems. Any number of test problems with specified sizes and properties may be generated. A shortcoming of these generators is that the problems that they generate may bear little resemblance to problems encountered in practice. The selection of an appropriate input model can alleviate this concern, at least to a certain extent.

This paper focuses on input models for synthetic optimization problems that are featured in research papers. Broadly speaking, these problem-generation methods may be classified as independent sampling, implicit correlation induction, and explicit correlation induction.

The paper is organized as follows. For convenience, we define the 0-1 Knapsack Problem, which we use throughout the paper as an example class of optimization problems, in §2. We describe and discuss each of three classes of problem-generations methods, or input models, in §3, 4, and 5, respectively. In §6, we discuss implications of input model selection on the conduct and results of computational experiments. Finally, we list open research questions related to input modeling for synthetic optimization problems in §7.

2 0-1 KNAPSACK PROBLEM

We define the 0-1 Knapsack Problem (KP01) as follows:

$$\text{Maximize} \quad \sum_{j=1}^{n} c_j x_j$$

$$\text{Subject to} \quad \sum_{j=1}^{n} a_j x_j \le b$$

$$x_j \in \{0,1\}, \forall j,$$

where all $c_j > 0$, all $a_j > 0$, $\sum_j c_j > b$, and $\max_j \{a_j\} \le b$. We assume that the c_js are i.i.d. realizations of some random variable C and that the a_js are i.i.d. realizations of some random variable A. Some rule for setting the value of *b* would have to be specified before any instance could be completely generated. For our purpose, such a rule need not be specified.

KP01 is a classical optimization problem for which many solution procedures have been devised. We use it as the basis for our discussion of input models for synthetic optimization problems because of its relatively simple form. Certainly, other optimization problems besides KP01 could be used for this purpose.

3 INDEPENDENT SAMPLING

Almost every computational study on synthetic optimization problems includes some test problems that are generated under independent sampling. Typically, a discrete uniform distribution is assumed for the values of each type of coefficients. Then, coefficient values are generated independently for each coefficient type until all of the needed coefficients are generated.

Let α and β be positive integers. We now present a procedure for generating KP01 coefficients under independent sampling.

Procedure GENER8

1. $a_j \leftarrow A \sim U\{1,2,\ldots,\alpha\}$.
2. $c_j \leftarrow C \sim U\{1,2,\ldots,\beta\}$.

Independent sampling is certainly easy to implement. Under this implementation of independent sampling, every possible KP01 test problem, or combination of coefficient values, is equally likely. The expected correlation between coefficient types (in this case, objective and constraint coefficients) is zero. But due to sampling error, the sample correlation between the coefficient types is not likely to be zero. The number of decision variables in the test problems will affect the distribution of sample correlation values. The larger the test problems, the less dispersed the sample correlation values are likely to be. As the size of the test problems is increased, it will become increasingly rare to find a test problem with even modest correlation between the coefficient types (Reilly 1993).

We would not expect that coefficient types in practical instances of KP01 or many other optimization problems would be uncorrelated, let alone independent. For example, consider a set covering problem in which warehouse sites are to be selected so that some collection of markets may be served at minimum cost. In such a case, we would expect that the cost to build and/or maintain a warehouse at a particular site would be directly related to the number or markets that could be served from that site.

Furthermore, as the size of the test problems is increased, the coefficients in test problems generated under independent sampling will be increasingly likely to pass tests of independence. Often, researchers will include large test problems in their experiments, and doing so affects the variety of the test problems that may be generated, at least in terms of the correlation between coefficient types. Even though there is a tendency to try to solve larger and larger test problems, we see that doing so with problems generated under independent sampling yields test problems that become more and more alike and unrealistic. When considered collectively, these larger test problems represent a smaller portion of the set of all possible test problems.

Conducting a computational study only on test problems generated under independent sampling is likely to produce results consistent with the median performance of the solution method(s) being considered. So that we may get a sense of the range of performance by a solution method, test problems should not be limited to those whose coefficients are generated independently.

4 IMPLICIT CORRELATION INDUCTION

It has been suggested that correlation among coefficient types in synthetic optimization problems can affect solution procedure performance in computational experiments, and consequently, correlation ought to be an experimental factor is such experiments. The conventional wisdom is that an extreme level of correlation between key types of coefficients can produce very challenging test problems. For some problems (such as KP01, Set Covering, Multidimensional Knapsack), strong positive correlation is associated with difficult problem instances. For other problems (such as Generalized Assignment), strong negative correlation is thought to make a problem instance more challenging.

Many researchers include test problems in which correlation is induced between certain types of coefficients in an effort to create more challenging test problems and/or to produce test problems that are more like instances that might be encountered in practice.

When test problems are generated under implicit correlation induction (ICI), parameters for a problem-generation method are specified. The values specified for these parameters imply some population correlation structure between coefficient types.

Unfortunately, the implied correlation levels are not normally quantified in any of the papers in which ICI methods are utilized. Rather, computational results for problems generated under ICI are compared to the results for problems generated under independent sampling. Even when different sets of parameter values are specified for an

ICI method, the test problems generated under ICI are sometimes considered collectively when the results are analyzed.

Calculating correlation levels induced under ICI methods is not difficult. Instead of quantifying the induced correlation levels, qualitative labels are used to distinguish the various of levels of correlation induced. Reilly (1997) provides closed form expressions for the correlation levels induced under ICI methods for some classical optimization problems.

Perhaps the most commonly used form of ICI was introduced by Martello and Toth (1979) for KP01. Additional examples of the use of ICI to generate KP01 instances include Balas and Zemel (1980), Martello and Toth (1988, 1997), Pisinger (1997), and Martello, Pisinger, and Toth (1999).

Similar ICI methods have been used to generate instances of Multidimensional Knapsack Problems (Balas and Martin, 1980; Fréville and Plateau 1994, 1996), Set Covering Problems (Rushmeier and Nemhauser 1993) and Generalized Assignment Problems (Martello and Toth 1981; Amini and Racer 1994).

Let α be a positive integer, and let δ and γ be nonnegative integers. In order to generate coefficients for a KP01 instance with ICI, we might use the following procedure.

Procedure ICI

1. $a_j \leftarrow A \sim U\{1,2,...,\alpha\}$.
2. $t_j \leftarrow T \sim U\{-\delta,-\delta+1,...,\delta\}$.
3. $c_j = a_j + t_j + \gamma$.

With Procedure ICI, the coefficients generated are said to be "weakly correlated" if γ=0, "strongly correlated" if δ=0, and "value independent" if δ=0 and γ=0. When neither δ nor γ is 0, the coefficients are said to be "almost strongly correlated". Reilly (1998) points out that for typical values of these parameters, the induced correlation is over 0.97 for the weakly correlated coefficients and very nearly 1 for the almost strongly correlated coefficients. With value independent problems and strongly correlated coefficients, the induced correlation is perfect.

Reilly (1998a) shows that, under Procedure ICI,

$$\mathrm{Corr}(A,C) = \sqrt{\frac{\alpha^2 - 1}{\alpha^2 + 4\delta(\delta+1) - 1}}.$$

It is unfortunate that correlation levels are apparently not quantified by the researchers who implement ICI problem-generation methods. The computational results reported for different classes of KP01 instances suggest that the performance of solution methods may be significantly affected by relatively minor changes in the correlation level. This observation underscores the importance of selecting an input model for computational experiments and of understanding what properties the resulting test problems will have.

Other types of ICI methods of KP01 include "uncorrelated coefficients with similar weights" (see, e.g., Martello and Toth, 1997) and "inversely strongly correlated coefficients" (see, e.g., Martello, Pisinger, and Toth 1999).

ICI methods are effective because they do indeed induce correlation between selected types of coefficients. It is not clear that the correlation that is induced is indicative of the correlation that would be found among coefficients in real-world instances. ICI methods are not difficult to implement.

A serious drawback with ICI methods is that changes in the coefficients' population correlation structure are confounded with the marginal distributions of coefficient values (Cario *et al*. 1995; Reilly 1997, 1998).

5 EXPLICIT CORRELATION INDUCTION

An alternative to implicit correlation induction is explicit correlation induction. Under explicit correlation induction (ECI), a joint distribution of coefficient values is specified or marginal distributions of values for each type of coefficient and a correlation structure are specified. In either case, the coefficients' population correlation structure is known or may be quantified.

Under ECI, it is easy to control the correlation structure among coefficient types because the correlation structure can be varied without affecting the marginal distributions of coefficient values. As a result, the effect of correlation on solution procedure performance is easier to isolate and measure.

Reilly (1991, 1993) suggests that ECI can be implemented for KP01 by "mixing" coefficient values generated under independent sampling with values generated based on extreme correlation. (Hill and Reilly 1999) extend this idea to multivariate sampling for optimization problems with more than two types of coefficients.)

Let $A \sim U\{1,2,\ldots,\alpha\}$ and $C \sim U\{1,2,\ldots,\beta\}$. Also, let f_A, F_A, and F_A^{-1} be the mass function, cumulative mass function, and inverse cumulative distribution function for *A*, respectively. Then, f_C, F_C, and F_C^{-1} are similarly defined for *C*.

Let ρ^+ be the maximum possible correlation between *A* and *C*. The minimum possible correlation is then $\rho^- = -\rho^+$. If ρ is the desired value of the correlation and if a value for θ is chosen such that $(1-\theta/(\alpha\beta))\rho^- \le \rho \le (1-\theta/(\alpha\beta))\rho^+$ and $0 \le \theta \le 1/(\alpha\beta)$,

then a composite mass function for (*A*,*C*) may be constructed as follows:

$$\lambda^+ g^+(a,c) + \lambda^- g^-(a,c) + \lambda^0 f_A(a) f_C(c),$$

where $g^+(a,c)$ and $g^-(a,c)$ are the maximum- and minimum-correlation distributions for (*A*,*C*), respectively, and $\lambda^+ = (1 - \alpha\beta\theta + \rho / \rho^+)/2$, $\lambda^- = (1 - \alpha\beta\theta - \rho / \rho^+)/2$, and $\lambda^0 = \alpha\beta\theta$.

The following procedure may be used to generate coefficients with explicitly induced correlation:

Procedure ECI

1. Generate $u \sim U(0,1)$.
2. Generate as follows:
   (a) If $u \le \lambda^0$, generate (*a*,*c*) using GENER8.
   (b) If $\lambda^0 < u \le \lambda^0 + \lambda^+$, generate (*a*,*c*) based on $g^+(a,c)$.
   (c) Otherwise, generate (*a*,*c*) based on $g^-(a,c)$.

Procedure ECI is not difficult to implement. The primary advantage of ECI over ICI is that the correlation structure among the coefficient types can be controlled systematically in a computational experiment. With the ECI implementation recommended here, an experimenter also is able to control the entropy of the distribution for (*A*,*C*) by varying the parameter θ (Peterson and Reilly 1995). (The parameter θ represents the smallest joint probability for any possible value of (*A*,*C*). So, by varying θ, an experimenter can effect changes in entropy.)

Of course, the distribution of coefficient values under ECI may not be similar to the distribution of coefficient values observed for practical problem instance. However, the fact that distributional parameters can be systematically varied and controlled without affecting the marginal distributions of coefficient values is a decided advantage of ECI over ICI.

6 DISCUSSION

Computational experiments are usually conducted so that the effectiveness of a solution method can be assessed or so that the performances of alternative solution methods can be compared. There is too little guidance about generating synthetic optimization problems available to researchers who wish to conduct computational experiments. Hall and Posner (1999) provide some general, but helpful, guidelines for experimenters.

It seems that researchers tend to follow the lead of their predecessors when deciding on what types of test problems to use. There is clearly some merit to doing so. However, once a particular problem-generation method is used in one study, it can become the standard approach for generating test problems from a particular class, whether it generates good sets of test problems or not.

We think that it is unfortunate that input models are not examined and understood to the same extent that, say, random number generators are, before they are widely adopted.

We regret that we may leave the reader with the impression that correlation is the most significant distributional factor on solution procedure performance. Correlation is indeed important, but there are other distributional factors that can influence performance of solution methods as well. For example, other factors that matter include distribution family (Loulou and Michaelides 1979) and the range of coefficient values (Yang 1994; Reilly 1998a).

Hooker (1994) advocates the development on an empirical science of algorithms. By recognizing the characteristics of synthetic optimization problems and understanding how those characteristics affect the performance of solution methods, one can better interpret the results of computational experiments and better assess the true capabilities and limitations of solution methods.

Cario *et al.* (1995) conducted a study of the performance of a general-purpose solver (LINDO) on Generalized Assignment Problem instances generated under ICI and ECI. They attempted to facilitate comparisons of results across the problem-generation methods by synchronizing some of the parameters of the distributions of the coefficient values. Based on their work, it appears that ECI instances are more challenging to solve than the comparable ICI instances are. We know of no other study that includes instances generated under both ICI and ECI.

We think that ECI methods offer clear advantages over ICI methods, and even more so over independent sampling. This is not to say that we think that ECI, and in particular the ECI approach we have presented, is the definitive problem-generation method. Much additional research and experimentation with synthetic optimization problems is needed.

7 OPEN QUESTIONS

Interesting open research questions include:

- How should synthetic instances of a given class of optimization problems be generated?
- What distributions of coefficient values should be used when generating instances of a particular class of optimization problem?

- How do the distribution families and the parameters of the assumed distributions of coefficient values affect the performance of solution methods?
- How should relationships between different types of coefficients observed in real-world instances be accounted for in synthetic optimization problems?
- How can the characteristics of practical instances of optimization problems be measured and represented in synthetic problem instances?
- How should the purposes of a computational experiment affect the selection of an input model for the coefficients in synthetic optimization problems?
- How can we "synchronize" instances of a particular optimization problem that are generated with different problem-generation methods?
- How can we generate instances with comparable difficulty for different optimization problems?
- How can the results from computational studies on synthetic optimization problems be used to design more effective solution methods?
- How can we determine in advance the best way to solve an instance of some optimization problem?

REFERENCES

Amini, M. and M. Racer. 1994. Comparison of alternative solution methods for the generalized assignment problem. *Management Science* 40: 868-890.

Balas, E. and C.H. Martin. 1980. Pivot and complement- a heuristic for 0-1 programming. *Management Science* 26: 86-96.

Balas, E. and E. Zemel. 1980. An algorithm for large zero-one knapsack problems. *Operations Research* 28: 1130-1154.

Cario, M., J.J Clifford, R.R. Hill, J. Yang, K. Yang, and C.H. Reilly. 1995. Alternative methods for generating synthetic generalized assignment problems. In *Proceedings of the 4th Industrial Engineering Research Conference*, ed. R. Uszoy and B.W. Schmeiser, 1080-1089. Institute of Industrial Engineers, Atlanta, Georgia.

Fréville, A. and G. Plateau. 1994. An efficient preprocessing procedure for the multidimensional knapsack problem. *Discrete Applied Mathematics* 49: 189-212.

Fréville, A. and G. Plateau. 1996. The 0-1 bidimensional knapsack problem: toward an efficient high-level primitive tool. *Journal of Heuristics* 2:147-167.

Hall, N.G. and M.E. Posner. 1999. Generating experimental data for combinatorial optimization problems. INFORMS National Meeting, May 2, 1999, Cincinnati, Ohio.

Hill, R.R. and C.H. Reilly. 1999. Multivariate composite distributions for coefficients in synthetic optimization problems. *European Journal of Operational Research* (to appear).

Hooker, J.N. 1994. Needed: an empirical science of algorithms. *Operations Research* 42 (2): 201-212.

Loulou, R. and E. Michaelides. 1979. New greedy heuristics for the multidimensional 0-1 knapsack problem. *Operations Research* 27:1101-1114.

Martello, S. and P. Toth. 1979. The 0-1 knapsack problem. In *Combinatorial optimization*, ed. A. Mingozzi and C. Sandi, 237-279. New York: John Wiley and Sons.

Martello, S. and P. Toth. 1981. An algorithm for the generalized assignment problem. In *Proceedings of the 9th IFORS Conference*, ed. J. Brans, 589-603. North-Holland.

Martello, S. and P. Toth. 1988. A new algorithm for the 0-1 knapsack problem. *Management Science* 34 (5): 633-644.

Martello, S. and P. Toth. 1997. Upper bounds and algorithms for 0-1 knapsack problems. *Operations Research* 45 (5): 768-778.

Martello, S., D. Pisinger, and P. Toth. 1999. Dynamic Programming and Strong Bounds for the 0-1 Knapsack Problem. *Management Science* 45 (3): 414-424.

Peterson, J.A. and C.H. Reilly. 1995. Joint probability mass functions for coefficients in synthetic optimization problems. Working Paper 1993-006 (revised). Department of Industrial, Welding, and Systems Engineering, The Ohio State University, Columbus, Ohio.

Pisinger, D. 1997. A minimal algorithm for the 0-1 knapsack problem. *Operations Research* 45 (5): 758-767.

Reilly, C.H. 1991. Optimization problems with uniformly distributed coefficients. In *Proceedings of the 1991 Winter Simulation Conference*, ed. B.L. Nelson, W.D. Kelton, and G.M. Clark, 866-874. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Reilly, C.H. 1993. A comparison of alternative input models for synthetic optimization problems. In *Proceedings of the 1993 Winter Simulation Conference*, ed. G.W. Evans, M. Mollaghasemi, E.C. Russell, and W.E. Biles, 356-364. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Reilly, C.H. 1997. Generating coefficients for optimization problems with implicit correlation induction. In *Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics*, 2438-2443. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Reilly, C.H. 1998a. On the random generation of synthetic 0-1 knapsack problems. Manuscript dated March 1, 1998, Department of Industrial Engineering and Management Systems, University of Central Florida, Orlando, Florida.

Reilly, C.H. 1998b. Properties of synthetic optimization problems. In *Proceedings of the 1998 Winter Simulation Conference*, ed. D.J. Medeiros, E.F. Watson, J.S. Carson, and M.S. Manivannan, 617-621. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Rushmeier, R. and G.L. Nemhauser. 1993. Experiments with parallel branch-and-bound algorithms for the set covering problem. *Operations Research Letters* 13 (5): 277-285.

Yang, J. 1994. A computational study on 0-1 knapsack problems generated under explicit correlation induction. M.S. Thesis. The Ohio State University, Columbus, Ohio.

AUTHOR BIOGRAPHY

CHARLES H. REILLY is Chair and Professor in the Department of Industrial Engineering and Management Systems at the University of Central Florida. He earned a B.A. in mathematics and business administration at St. Norbert College in 1979, and M.S. and Ph.D. degrees in industrial engineering at Purdue University in 1980 and 1983, respectively. His current research interests include the generation of synthetic optimization problems, the evaluation of solution methods for discrete optimization problems, and simulation modeling. Dr. Reilly is a member of IIE, INFORMS, and ASEE.