SIMULATION SUPPORT TO GREY-RELATED ANALYSIS: DATA MINING SIMULATION

David L. Olson[1] and Desheng Wu[2,3]
[1]*Department of Management, University of Nebraska, Lincoln, NE* [2]*Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario* [3]*School of Business, University of Science and Technology of China, Hefei Anhui, China*

Abstract: This chapter addresses the use of Monte Carlo simulation to reflect uncertainty as expressed by fuzzy input. Fuzziness is expressed through grey-related analysis, using interval fuzzy numbers. The method standardizes inputs through norms of interval number vectors. Interval-valued indexes are used to apply multiplicative operations over interval numbers. The method is demonstrated on a practical problem. Simulation offers a more complete understanding of the possible outcomes of alternatives as expressed by fuzzy numbers. The focus is on probability rather than on maximizing expected or extreme values.

Key words: Fuzzy sets, Monte Carlo simulation, grey-related analysis, data mining

1. INTRODUCTION

This chapter addresses the use of Monte Carlo simulation to reflect uncertainty as expressed by fuzzy input. Simulation offers a more complete understanding of the possible outcomes of alternatives as expressed by fuzzy numbers. The focus is on probability rather than on maximizing expected or extreme values. Both weights and alternative performance scores are allowed to be fuzzy. Both interval and trapezoidal fuzzy input can be considered (see Olson and Wu, 2005, 2006).

Fuzzy concepts have long been important in multiple criteria analysis (Dubois, 1980; Gau and Buehrer, 1993; Pawlak, 1982; Pearl, 1988; Pedrycz, 1998). Simulation has been applied to the analytical hierarchy process (AHP) (Levary and Wan, 1998), generating random pair-wise comparison input values. The uncertainty and fuzziness inherent in decision making makes the use of precise numbers problematic in multi-attribute models. Decision makers are usually more comfortable providing intervals for specific model input parameters. Interval input in multi-attribute decision making has been a very active field of research. Methods applying intervals have included (along with many others, see Zhang et al., 2005):

1. Use of interval numbers as the basis for ranking alternatives
   Brans and Vincke, 1985;
   El-Hawary, 1998;
   Chang and Yeh, 2004;
   Kahraman et al., 2004.
2. Error analysis with interval numbers
   Larichev and Moshkovich, 1991.
3. Use of linear programming and object programming with feasible regions bounded by interval numbers
   Roy, 1978;
   Liu et al., 1999;
   Royes et al., 2003.
4. Use of interval number ideal alternatives to rank alternatives by their nearness to the ideal
   Wang et al., 2004.

AHP was presented (Saaty, 1977) as a way to take subjective human inputs in a hierarchy and to convert these to a value function. This method has proven extremely popular. Salo and Hamalainen (1992) published their interval method using linear programming over the constrained space of weights and values as a means to incorporate uncertainty in decision-maker inputs to AHP hierarchies.

The problem of synthesizing ratio judgments in groups was considered very early in AHP (Aczel and Saaty, 1983). Fuzzy AHP was proposed as another way to reflect uncertainty in subjective inputs to AHP in the same group context (Buckley, 1984; 1985a; 1985b). Simulation has been presented as a way to rank order alternatives in the context of AHP values and weights (Levary and Wan, 1998).

Other multiple criteria methods besides AHP have considered fuzzy input parameters. ELECTRE (Roy, 1978) and PROMETHEE (Brans and Vincke, 1985) have always allowed fuzzy input for weights. A multi-attribute method involving fuzzy assessment for selection has been given in the airline safety domain (Chang and Yeh, 2004) and for multiple criteria selection of employees (Royes et al., 2003). Sensitivity in multi-attribute models with fuzzy inputs was considered by Aouam et al. (2003) and in goal programming by Fan et al. (2004). Rough set applications have also been presented (Zaras, 2004). This stream of research has obviously been rich and useful in application. It is extended by grey-related analysis.

2. GREY-RELATED ANALYSIS

Grey system theory was developed by Deng (1982) based on the concept that information is sometimes incomplete or unknown. The intent is the same as with factor analysis, cluster analysis, and discriminant analysis, except that those methods often do not work well when sample size is small and sample distribution is unknown (Wang et al., 2004). With grey-related analysis, interval numbers are standardized through norms, which allow transformation of index values through product operations. The method is simple, practical, and demands less-precise information than other methods. Grey-related analysis and TOPSIS (Hwang and Yoon, 1981; Lai et al., 1994; Yoon and Hwang, 1995) both use the idea of minimizing a distance function. However, grey-related analysis reflects a form of fuzzification of inputs and uses different calculations, to include a different calculation of norms. Feng and Wang (2001) applied grey relation analysis to select representative criteria among a large set of available choices and then used TOPSIS for outranking (Zhang et al., 2005)

Grey-related analysis has been used in a number of applications, In our discussion, we shall use the concept of the norm of an interval number column vector, the distance between intervals, product operations, and number-product operations of interval numbers.

Let $a=[a^-,a^+]=\{x \mid a^- \le x \le a^+, a^- \le a^+, a^-, a^+ \in R\}$.

We call $a=[a^-,a^+]$ an interval number. If $0 \le a^- \le a^+$, we call interval number $a=[a^-,a^+]$ a positive interval number.

Let $X=([a_1^-,a_1^+],[a_2^-,a_2^+],\ldots,[a_n^-,a_n^+])^T$ be an n-dimension interval number column vector.

DEFINITION 1.

If $X = ([a_1^-, a_1^+], [a_2^-, a_2^+], \ldots, [a_n^-, a_n^+])^T$ is an arbitrary interval number column vector, the norm of X is defined here as

$$\| X \| = \max(\max(| a_1^- |, | a_1^+ |), \max(| a_2^- |, | a_2^+ |), \ldots, \max(| a_n^- |, | a_n^+ |)) \quad (1)$$

DEFINITION 2.

If $a = [a^-, a^+]$ and $b = [b^-, b^+]$ are two arbitrary interval numbers, the distance from $a = [a^-, a^+]$ to $b = [b^-, b^+]$, is defined as

$$| a - b | = \max(| a^- - b^- |, | a^+ - b^+ |) \quad (2)$$

DEFINITION 3.

If k is an arbitrary positive real number, and $a = [a^-, a^+]$ is an arbitrary interval number, then $k \cdot [a^-, a^+] = [ka^-, ka^+]$ will be called the number-product between k and $a = [a^-, a^+]$.

DEFINITION 4.

If $a = [a^-, a^+]$ is an arbitrary interval number, and $b = [b^-, b^+]$ are arbitrary interval numbers, we shall define the interval number product $[a^-, a^+] \cdot [b^-, b^+]$ as follows:

$$\text{when } b^+ > 0 \;\; [a^-, a^+] \cdot [b^-, b^+] = [a^- b^-, a^+ b^+] \quad (3)$$

$$\text{when } b^+ < 0 \;\; [a^-, a^+] \cdot [b^-, b^+] = [a^+ b^-, a^- b^+] \quad (4)$$

If $b^+ = 0$, the interval reverts to a point, and thus, we would return to the basic crisp model.

2.1 Steps of Grey-Related Analysis

The principle and steps of the Grey-related analysis method are as follows:

Step 1. Construct decision matrix A with an index number of interval numbers. If the index value of the jth index G_j of feasible plan X_i is an interval number $[a_{ij}^-, a_{ij}^+]$, $i = 1,2,\ldots,m$, $j = 1,2,\ldots,n$, decision matrix A with index number of interval numbers is defined as the follows:

$$A = \begin{bmatrix} [a_{11}^-, a_{11}^+] & [a_{12}^-, a_{12}^+] & \dots & [a_{1n}^-, a_{1n}^+] \\ [a_{21}^-, a_{21}^+] & [a_{22}^-, a_{22}^+] & \dots & [a_{2n}^-, a_{2n}^+] \\ \dots & \dots & \dots & \dots \\ [a_{m1}^-, a_{m1}^+] & [a_{m2}^-, a_{m2}^+] & \dots & [a_{mn}^-, a_{mn}^+] \end{bmatrix} \quad (5)$$

Step 2. Transform the "contrary index" into a positive index .The index is called a positive index if a greater index value is better. The index is called a contrary index if a smaller index value is better. We may transform a contrary index into a positive index if the *j*th index G_j is a contrary index

$$[b_{ij}^-, b_{ij}^+] = [-a_{ij}^+, -a_{ij}^-] \quad i = 1,2,...,m\,. \quad (6)$$

Without loss of generality, in the following discussion, we supposed that all the indexes are "positive indices."

Step 3. Standardize decision matrix A with an index number of interval numbers, obtaining standardizing decision matrix $R = [\, r_{ij}^-, r_{ij}^+ \,]$. If we mark the column vectors of decision matrix A with interval-valued indexes with $A_1, A_2, ..., A_n$, the element of standardizing decision matrix $R = [\, r_{ij}^-, r_{ij}^+ \,]$ is defined as

$$[r_{ij}^-, r_{ij}^+] = \frac{[a_{ij}^-, a_{ij}^+]}{\| A_j \|}, \quad i = 1,2,...,m \qquad j = 1,2,...,n\,. \quad (7)$$

Step 4. Calculate interval number weighted matrix $C = ([\, c_{ij}^-, c_{ij}^+ \,])_{m \times n}$. The formula for the element of interval number weighted matrix C is $C = ([\, c_{ij}^-, c_{ij}^+ \,])_{m \times n}$
where

$$[c_{ij}^-, c_{ij}^+] = [c_j, d_j] \cdot [r_{ij}^-, r_{ij}^+], \quad i = 1,2,...,m \qquad j = 1,2,...,n\,. \quad (8)$$

Step 5. Determine reference number sequence. The element of reference number sequence is composed of the optimal weighted interval number index value for every alternative.

$U_0 = ([u_0^-(1)\ ,\ u_0^+(1)]\ ,\ [u_0^-(2)\ ,\ u_0^+(2)]\ ,\ \ldots\ ,\ [u_0^-(n)\ ,\ u_0^+(n)])$

is a reference number sequence if $u_0^-(j) = \max\limits_{1\le i\le m} c_{ij}^-$, $u_0^+(j) = \max\limits_{1\le i\le m} c_{ij}^+$,

$j = 1, 2, \ldots, n$.

Step 6. Calculate connections between alternatives. First, calculate the connection coefficient $\xi_i(k)$ between the sequence composed of weight interval number standardized index values for every alternative $U_i = ([c_{i1}^-\ ,\ c_{i1}^+], [c_{i2}^-\ ,\ c_{i2}^+],\ \ldots\ ,\ [c_{in}^-\ ,\ c_{in}^+])$ and the reference number sequence $U_0 = ([u_0^-(1),\ u_0^+(1)],\ [u_0^-(2),\ u_0^+(2)],\ \ldots,\ [u_0^-(n),\ u_0^+(n)])$.

The formula for $\xi_i(k)$ is

$$\xi_i(k) = \frac{\min\limits_i \min\limits_k |[u_0^-(k), u_0^+(k)] - [c_{ik}^-, c_{ik}^+]| + \rho \max\limits_i \max\limits_k |[u_0^-(k), u_0^+(k)] - [c_{ik}^-, c_{ik}^+]|}{|[u_0^-(k), u_0^+(k)] - [c_{ik}^-, c_{ik}^+]| + \rho \max\limits_i \max\limits_k |[u_0^-(k), u_0^+(k)] - [c_{ik}^-, c_{ik}^+]|} \tag{9}$$

Here, $\rho \in (0, +\infty)$, and ρ is a resolving coefficient. The smaller ρ is, the greater its resolving power. In general, $\rho \in [0,\ 1]$ .The value of ρ may be changed to reflect the desired degree of resolution.

After calculating $\xi_i(k)$, the connection between the *i*-th plan and the reference number sequence is calculated by the following formula:

$$r_i = \frac{1}{n} \times \sum_{k=1}^{n} \xi_i(k), \qquad i = 1, 2, \ldots m \tag{10}$$

Step 7. Determine optimal plan. The feasible plan X_t is optimal if $r_t = \max\limits_{1\le i\le m} r_i$.

3. MONTE CARLO SIMULATION

Fuzzy inputs can easily be simulated using Monte Carlo simulation models. Interval random numbers over the interval 0−1 can be generated in Monte Carlo simulation directly, and these can be converted to any other uniform range. Simulations can be easier to analyze if they are controlled, using unique seed values to ensure that the difference in simulation output due to random variation was the same for each alternative.

3.1 Trapezoidal Distributed Fuzzy Numbers

The trapezoidal fuzzy input dataset can also be simulated.

X is random number (0 < *rn* < 1).

Definition of trapezoida1 is left 0 in Figure 1; *a*2 is left 1; *a*3 is right 1; and *a*4 is right 0.

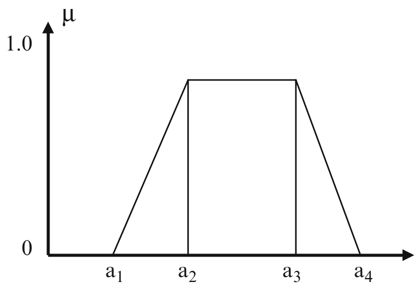


Figure 1. A trapezoidal fuzzy number

J is area of left triangle contingent calculation:

K is area of rectangle

L is area of right triangle

Fuzzy sum = left triangle + rectangle + right triangle = 1

M is the area of the left triangle plus the rectangle (for calculation of *X* value)

X is the random number drawn (which is the area)

If $X \le J$:

$$X = a1 + \sqrt{\frac{X \times (a2 - a1) \times (a4 - a3 + a2 - a1)}{J + L}} \tag{11}$$

If $J \le X \le J+K$:

$$X = a2 + \frac{X - J}{K} \times (a3 - a2) \tag{12}$$

If $J+K \le X$:

$$X = a4 - \sqrt{\frac{(1 - X) \times (a4 - a3) \times (a4 - a3 + a2 - a1)}{J + L}} \tag{13}$$

Our calculation is based on drawing a random number reflecting the area (starting on the left (a1) as 0, ending on the right (a4) as 1), and calculating the distance on the *X*-axis. The simulation software Crystal Ball was used to replicate each model 1000 times for each random number seed. The software enabled counting the number of times each alternative won.

3.2 Grey-Related Decision Tree Models

Grey-related analysis is expected to provide improvement over crisp models by better reflecting the uncertainty inherent in many human analysts' minds. Data mining models based on such data are expected to be less accurate, but hopefully not by very much (Hu et al., 2003). However, grey-related model input would be expected to be more stable under conditions of uncertainty where the degree of change in input data increased.

We applied decision tree analysis to a small set (1000 observations total) of credit card data. Originally, there was one output variable (whether or not the account defaulted, a binary variable with 1 representing default, 0 representing no default) and 65 available explanatory variables. These variables were analyzed, and 26 were selected as representing ideas that might be important to predicting the outcome. The original data set was imbalanced, with 140 default cases and 860 not defaulting. Initial decision tree models were almost all degenerate, classifying all cases as not defaulting. When differential costs were applied, the reverse degenerate model was obtained (all cases predicted to default). Therefore, a new dataset containing all 140 default cases and 160 randomly selected not default cases was generated, where 200 cases were randomly selected as a training set, with the remaining 100 cases used as a test set.

The explanatory variables included five binary variables and one categorical variable, with the remaining 20 being continuous. To reflect fuzzy input, each variable (except for binary variables) was categorized into three categories based on analysis of the data, using natural cutoff points to divide each variable into roughly equal groups.

Decision tree models were generated using the data mining software PolyAnalyst. That software allows setting minimum support level (the number of cases necessary to retain a branch on the decision tree), and a slider setting to optimistically or pessimistically split criteria. Lower support levels allow more branches, as does the optimistic setting. Every time the model was run, a different decision tree was able to be obtained.

But nine settings were applied, yielding many overlapping models. Three unique decision trees were obtained, which are reflected in the output to follow. A total of eight explanatory variables were used in these three decision trees. The same runs were made for the categorical data reflecting grey-related input. Four unique decision trees were obtained, with formulas again given below. A total of seven explanatory variables were used in these four categorical decision trees. All seven models and their fit on test data are given in the Appendix.

These models were then entered into a Monte Carlo simulation (supported by Crystal Ball software). A perturbation of each input variable was generated, set at five different levels of perturbation. The intent was to measure the loss of accuracy for crisp and grey-related models.

The model results are given in the seven model reports in the appendix. Since different variables were included in different models, it is not possible to directly compare relative accuracy as measured by fitting test data. However, the means for the accuracy on test data for each model given in Table 1 show that the crisp models declined in accuracy more than the categorical models. The column headings in Table 1 reflect the degree of perturbation simulated.

Table 1. Mean Model Accuracy

Model	**Crisp**	**0.25**	**0.50**	**1.00**	**2.00**	**3.00**	**4.00**
Cont. 1	0.70	0.70	0.70	0.68	0.67	0.66	0.65
Cont. 2	0.67	0.67	0.67	0.67	0.67	0.66	0.66
Cont. 3	0.71	0.71	0.70	0.69	0.67	0.67	0.66
Cont.	**0.693**	**0.693**	**0.690**	**0.680**	**0.670**	**0.667**	**0.657**
Cat. 1	0.70	0.70	0.68	0.67	0.66	0.66	0.65
Cat. 2	0.70	0.70	0.70	0.69	0.68	0.67	0.67
Cat. 3	0.70	0.70	0.70	0.69	0.69	0.68	0.67
Cat. 4	0.70	0.70	0.70	0.69	0.68	0.67	0.67
Cat.	**0.700**	**0.700**	**0.695**	**0.688**	**0.678**	**0.670**	**0.665**

The fuzzy models were expected to be less accurate, but here they actually average slightly better accuracy. This, however, can simply be attributed to different variables being used in each model. The one exception is that models Continuous 2 and Categorical 3 were based on one variable, V64, the balance-to-payment ratio. The cutoff generated by model Continuous 2 was 6.44 (if V64 was < 6.44, prediction 0), whereas the cutoff for Categorical 3 was 4.836 (if V64 was > 4.835, the category was "high," and the decision tree model was that if V64 = "high," prediction 1, else prediction 0). The fuzzy model here was actually better in fitting the test data (although slightly worse in fitting the training data).

The important point of the numbers in Table 1 is that there clearly was greater degradation in model accuracy for the continuous models than for the categorical (grey-related) models. This point is demonstrated further by the wider dispersion of the graphs in the Appendix.

4. CONCLUSIONS

This chapter has discussed the integration of grey-related analysis and decision making with uncertainty through simulation. Simulation provides a means to better visualize model results and a flexible way to include any level of uncertainty and complexity. Results based on Monte Carlo simulation as a data-mining technique offer more insights to assist our decision making in fuzzy environments by incorporating probability interpretation. Analysis of decision tree models through simulation shows that there does appear to be less degradation in model fit for grey-related (categorical) data than for decision tree models generated from raw continuous data. It must be admitted that this is a preliminary result, based on a relatively small dataset of only one type of data. However, it is intended to demonstrate a point meriting future research. This decision-making approach can be applied to large-scale datasets, expanding our ability to implement data mining and large-scale computing.

The easiest way to apply fuzzy concepts to data mining is to categorize data. This creates the problem of where to set limits between categories. However, reliance on expert judgment can often provide useful limits. If data-mining data are represented through fuzzy concepts, simulation can be applied. Since fuzzy data are probabilistic, simulation seems appropriate. Simulation does involve a lot more work than closed-form (crisp) datasets. However, fuzzy data are often a better representation of real domains.

REFERENCES

Aczel, J., and Saaty, T.L., 1983, Procedures for synthesizing ratio judgments. *Journal of Mathematical Psychology*, **27:** 93−102.

Aouam, T., Chang, S.I., and Lee, E.S., 2003, Fuzzy MADM: An outranking method, *European Journal of Operational Research*, **145**(2)**:** 317–328.

Brans, J.P., and Vincke, Ph., 1985, A preference ranking organization method: The PROMETHEE method. *Management Science*, **31:** 647–656.

Buckley, J.J., 1984, The multiple judge, multiple criteria ranking problem: A fuzzy set approach. *Fuzzy Sets and System*, **13**(1)**:** 25–37.

Buckley, J.J., 1985a, Ranking alternatives using fuzzy members. *Fuzzy Sets and Systems*, **17:** 233−247.

Buckley, J.J., 1985b, Fuzzy hierarchical analysis. *Fuzzy Sets and Systems*, **17:** 233− 247.

Chang, Y.-H., and Yeh, C.-H., 2004, A new airline safety index, *Transportation Research Part B*, **38:** 369−383.

Deng, J.L., 1982, Control problems of grey systems. *Systems and Controls Letters*, **5:** 288−294.

Dubois, D., and Prade, H., 1980, *Fuzzy Sets and Systems: Theory and Applications*, Academic Press, Inc., New York.

El-Hawary, M.E., 1998, *Electric Power Applications of Fuzzy Systems*, The Institute of Electrical and Electronics Engineers Press, Inc., New York.

Fan, Z., Hu, G., and Xiao, S.-H., 2004, A method for multiple attribute decision-making with the fuzzy preference relation on alternatives, *Computers & Industrial Engineering*, **46:** 321−327.

Feng, C.-M., and Wang, R.-T., 2001, Considering the financial ratios on the performance evaluation of highway bus industry, *Transport Reviews*, **21**(4)**:** 449−467.

Gau, W.L., and Buehrer, D.J., 1993, Vague sets, *IEEE Transactions On Systems, Man, And Cybernetics*, **23:** 610−614.

Hu, Y., Chen, R.-S., and Tzeng, G.-H., 2003, Finding fuzzy classification rules using data mining techniques, *Pattern Recognition Letters*, **24**(1−3)**:** 509−519.

Hwang, C.L., and Yoon, K., 1981, *Multiple Attribute Decision Making: Methods and Applications*, Springer-Verlag, New York.

Kahraman, C., Cebeci, U., and Ruan, D., 2004, Multi-attribute comparison of catering service companies using fuzzy AHP: the case of Turkey, *International Journal of Production Economics*, **87:** 171−184.

Lai, Y.-J., Liu, T.-Y., and Hwang, C.-L., 1994, TOPSIS for MODM, *European Journal of Operational Research*, **76**(3)**:** 486−500.

Larichev, O.I., and Moshkovich, H.M., 1991, *ZAPROS: A Method And System For Ordering Multiattribute Alternatives On The Base Of A Decision-Maker's Preferences*, All-Union Research Institute for System Studies, Moscow.

Levary, R.R., and Wan, K., 1998, A simulation approach for handling uncertainty in the analytic hierarchy process, *European Journal of Operational Research*, **106:** 116−122.

Liu, S., Guo, B., and Dang, Y., 1999, *Grey System Theory and Applications*, Scientific Press, Beijing.

Olson, D.L., and Wu, D., 2005, Decision making with uncertainty and data mining, *The 1st International Conference on Advanced Data Mining and Applications (ADMA2005)*, Li, X., Wang, S., and Yang D. Z., eds., Lecture Notes in Computer Science, Springer, Berlin.

Olson, D.L., and Wu, D., 2006, Simulation of fuzzy multiattribute models for grey relationships, *European Journal of Operational Research*, **175**(1)**:** 111−120.

Pawlak, Z., 1982, Rough sets, *International Journal of Information & Computer Sciences*, **11:** 341−356.

Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, CA.

Pedrycz, W., 1998, Fuzzy set technology in knowledge discovery, *Fuzzy Sets and Systems,* **98**(3)**:** 279−290.

Rocco S., and Claudio, M., 2003, A rule induction approach to improve Monte Carlo system reliability assessment, *Reliability Engineering and System Safety*, **82**(1)**:** 85−92.

Roy, B., 1978, ELECTRE III: un algorithme de classement fonde sur une representation floue des preferences en presence de criteres multiple, *Cahiers du Centre Etudes Recherche Operationelle*, **20:** 3−24.

Royes, G.F., Bastos, R.C., and Royes, G.F, 2003, Applicants' selection applying a fuzzy multicriteria CBR methodology, *Journal of Intelligent & Fuzzy Systems*, **14**(4)**:** 167−180.

Saaty, T.L., 1977, A scaling method for priorities in hierarchical structures, *Journal of Mathematical Psychology*, **15:** 234−281.

Salo, A.A., and Hamalainen, R.P., 1992, Preference assessment by imprecise ratio statements, *Operations Research*, **40:** 1053−1061.

Wang, R.-T., Ho, C.-T., Feng, C.-M., and Yang, Y.-K., 2004, A comparative analysis of the operational performance of Taiwan's major airports, *Journal of Air Transport Management*, **10:** 353−360.

Yoon, K., and Hwang, C.L., 1995, *Multiple Attribute Decision Making: An Introduction Sage*, Thousand Oaks, CA.

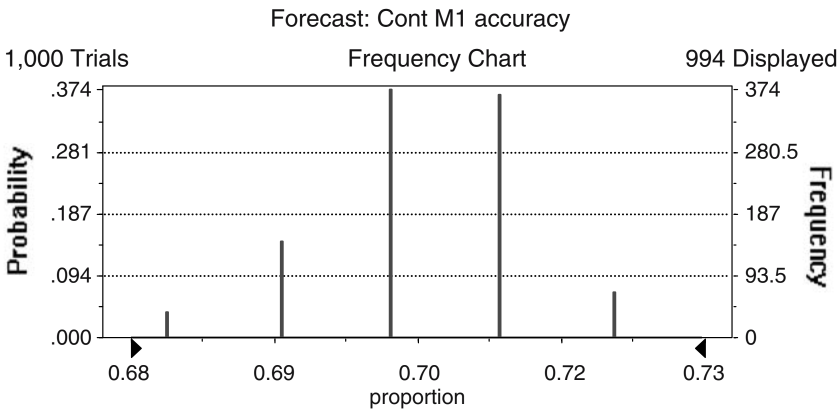
Zaras, K., 2004, Rough approximation of a preference relation by a multi-attribute dominance for deterministic, stochastic and fuzzy decision problems, *European Journal of Operational Research*, **159:** 196–206.

Zhang, J., Wu, D., and Olson, D.L., 2005, The method of grey related analysis to multiple attribute decision making problems with interval numbers, *Mathematical and Computer Modelling*, **42**(9–10)**:** 991–998.

APPENDIX: MODELS AND THEIR RESULTS

Continuous Model 1:

IF(Bal/Pay<6.44,N,IF(Utilization<1.54,Y,IF(AvgPay<3.91,N,Y)))

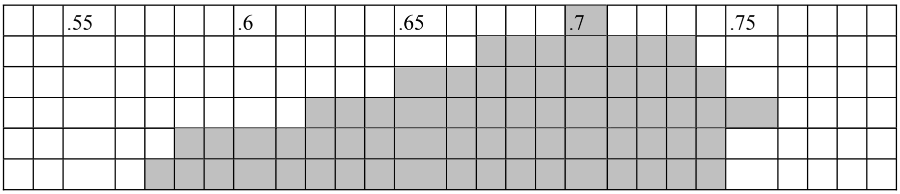


Test matrix:

	Model 0	Model 1	Accuracy
Actual 0	43	16	
Actual 1	14	27	0.70

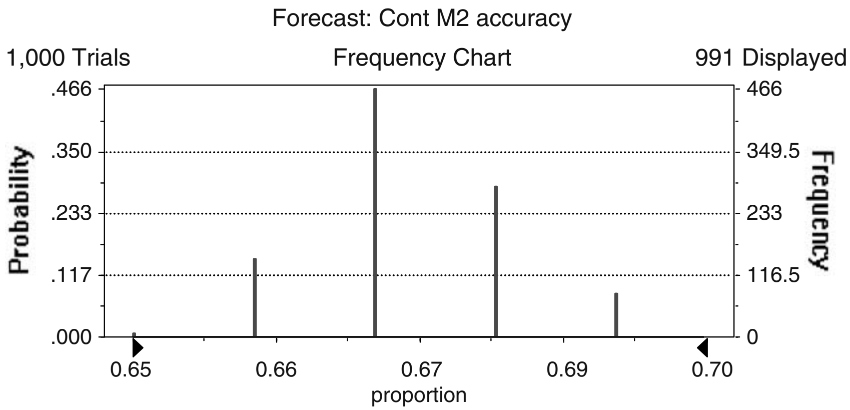
Simulation accuracy of 100 observations, 1000 simulation runs

perturbation	[−0.25,0.25]	0.67−0.73
perturbation	[−0.50,0.50]	0.65−0.74
perturbation	[−1,1]	0.62−0.75
perturbation	[−2,2]	0.58−0.74
perturbation	[−3,3]	0.57−0.74



Continuous Model 2:

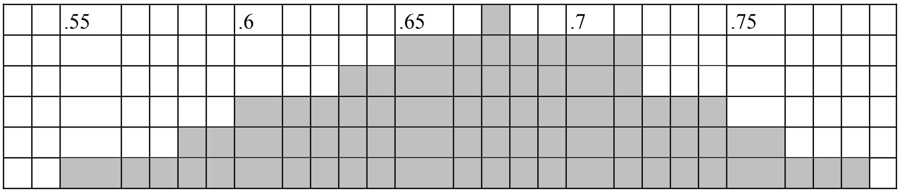
IF (Bal/Pay<6.44,N,Y)



Test matrix:

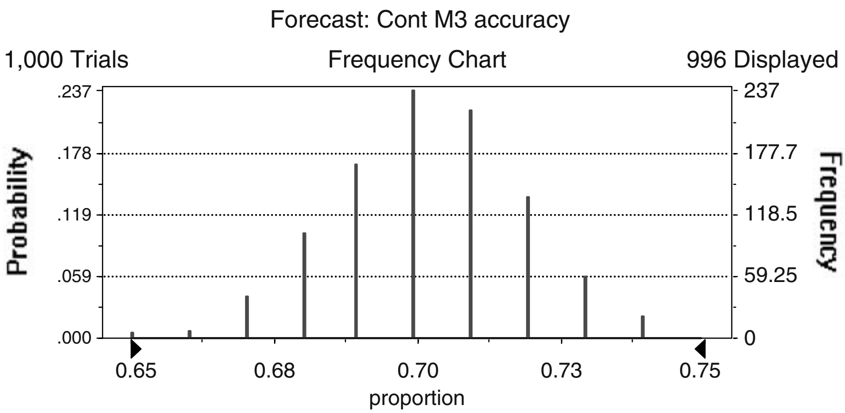
	Model 0	Model 1	Accuracy
Actual 0	40	19	
Actual 1	14	27	0.67

Simulation accuracy of 100 observations, 1000 simulation runs
perturbation [−0.25,0.25] 0.65−0.71
perturbation [−0.50,0.50] 0.63−0.71
perturbation [−1,1] 0.60−0.74
perturbation [−2,2] 0.58−0.75
perturbation [−3,3] 0.55−0.78



Continuous Model 3:

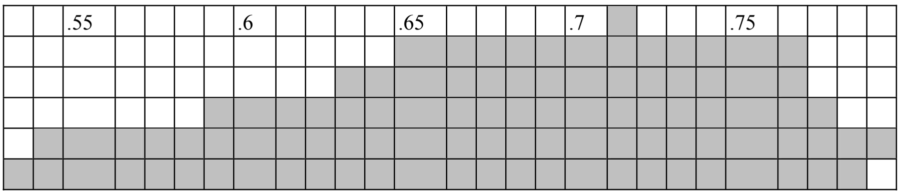
IF(Bal/Pay<6.44,N,IF(Utilization<1.54,Y,IF(AvgRevPay<2.28,Y,N)))



Test matrix:

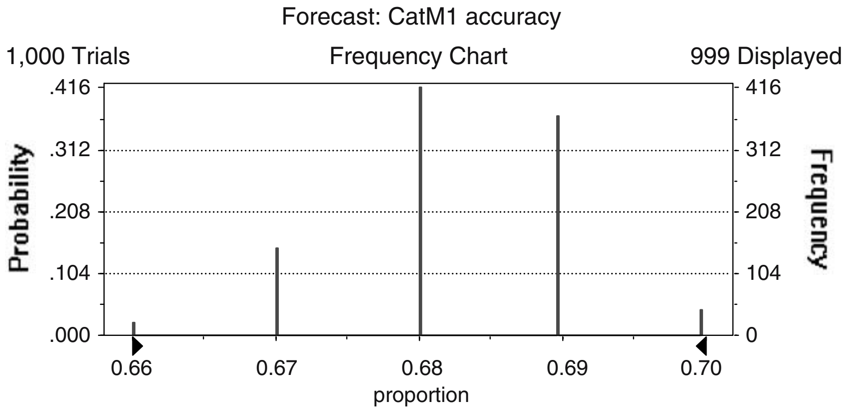
	Model 0	Model 1	Accuracy
Actual 0	44	15	
Actual 1	14	27	0.71

Simulation accuracy of 100 observations, 1000 simulation runs
perturbation [−0.25,0.25]   0.65−0.76
perturbation [−0.50,0.50]   0.63−0.76
perturbation [−1,1]   0.59−0.77
perturbation [−2,2]   0.54−0.79
perturbation [−3,3]   0.53−0.78



Categorical Model 1:

IF(Bal/Pay<6.44,N,IF(Utilization<1.54,Y,IF(AvgRevPay<2.28,Y,N)))

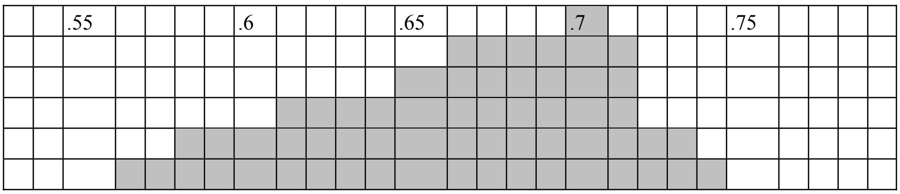


Test matrix:

	Model 0	Model 1	Accuracy
Actual 0	33	26	
Actual 1	5	36	0.70

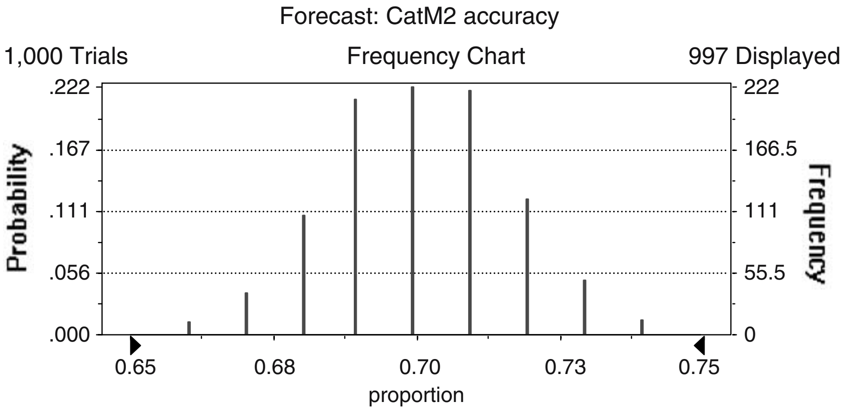
Simulation accuracy of 100 observations, 1000 simulation runs

perturbation [−0.25,0.25]	0.66−0.71
perturbation [−0.50,0.50]	0.64−0.71
perturbation [−1,1]	0.61−0.71
perturbation [−2,2]	0.58−0.73
perturbation [−3,3]	0.56−0.74



Categorical Model 2:

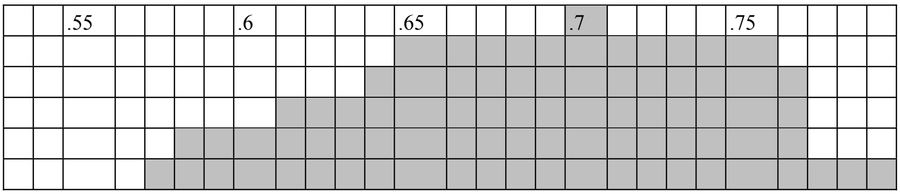
IF(Bal/Pay=`"high",IF(CredLine="low",
IF(CDL="mid",IF(Pur%Bal="low",Y,N),
IF(CDL="low",N,Y))
IF(CredLine="high",IF(CalcIntRate="mid",N,Y),Y),N)



Test matrix:

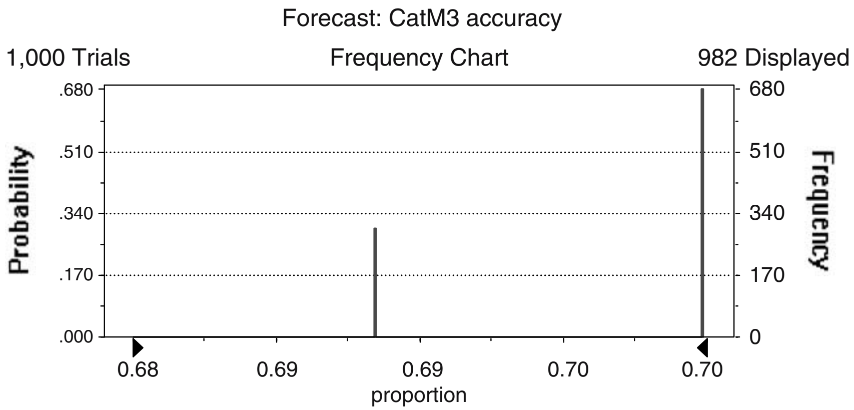
	Model 0	Model 1	Accuracy
Actual 0	42	17	
Actual 1	13	28	0.70

Simulation accuracy of 100 observations, 1000 simulation runs
perturbation [−0.25,0.25] 0.65−0.75
perturbation [−0.50,0.50] 0.64−0.76
perturbation [−1,1] 0.61−0.76
perturbation [−2,2] 0.58−0.76
perturbation [−3,3] 0.57−0.80



Categorical Model 3:

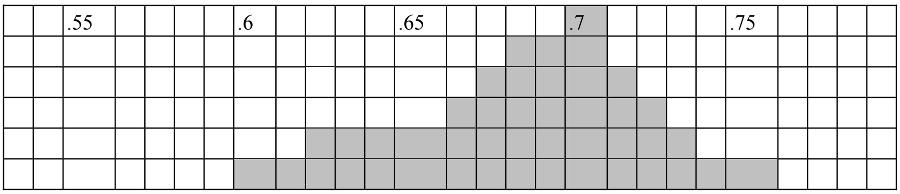
IF(Bal/Pay=”high”,Y,N)



Test matrix:

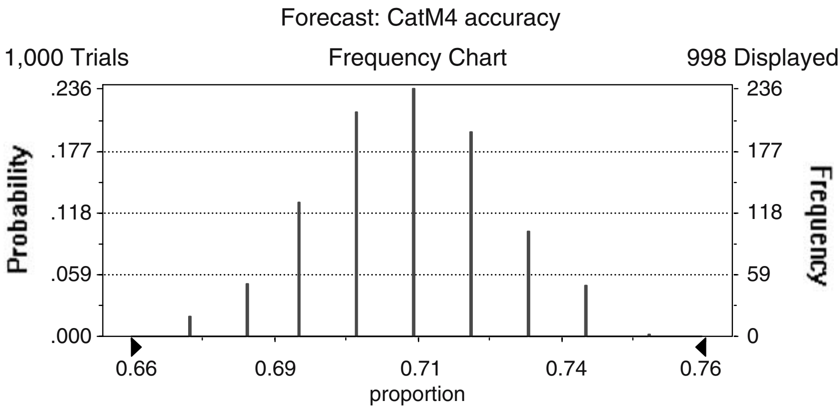
	Model 0	Model 1	Accuracy
Actual 0	33	26	
Actual 1	4	37	0.70

Simulation accuracy of 100 observations, 1000 simulation runs
perturbation [−0.25,0.25] 0.68−0.70
perturbation [−0.50,0.50] 0.67−0.71
perturbation [−1,1] 0.66−0.72
perturbation [−2,2] 0.62−0.73
perturbation [−3,3] 0.59−0.75



Categorical Model 4:

IF(Bal/Pay=”high”,
IF(CredLine=”low”,
IF(CDL=”mid”,IF(Purch%Bal=”low”,Y,N),
IF(CDL=”low”,IF(Residence<.5,Y,N),Y))
IF(CredLine=”high”,IF(CalcIntRate=”mid”,N,Y),Y)



Test matrix:

	Model 0	Model 1	Accuracy
Actual 0	41	18	
Actual 1	12	29	0.70

Simulation accuracy of 100 observations, 1000 simulation runs
perturbation [−0.25,0.25] 0.65−0.76
perturbation [−0.50,0.50] 0.64−0.77
perturbation [−1,1] 0.61−0.77
perturbation [−2,2] 0.58−0.77
perturbation [−3,3] 0.57−0.77