

# Study on analyzing questionnaire survey by Monte Carlo Simulation

Jiang Shijie<sup>1</sup>

The Hong Kong  
Polytechnic University  
HongKong, China  
bscqaj@polyu.edu.hk

Peng Yi<sup>2</sup>

The Hong Kong  
Polytechnic University  
HongKong, China  
bsyipeng@polyu.edu.hk

Lv Weisheng<sup>3</sup>

The Hong Kong  
University  
HongKong, China  
wilsonlu@hku.hk

Shen Liyin<sup>4</sup>

The Hong Kong  
Polytechnic University  
HongKong, China  
bsshenn@inet.polyu.edu.hk

**Abstract**—Traditional questionnaire usually takes likert scale and treats the data as discrete and controllable. It overlooks the different opinions chosen in the same scale, moreover does not consider the possibility distribution trend. Monte Carlo Simulation (MCS) is used to explore into the possible trend based on the distribution from the questionnaire. The difference of the results based on original questionnaire and MCS show the trying is beneficial and interesting. However, for this method, the questionnaire survey should be scientifically random. Moreover, foundational research should be conducted to test the validity.

**Keywords**—questionnaire survey; Monte Carlo Simulation (MCS); ranking by mean; fuzzy set theory.

## I. INTRODUCTION

Traditional questionnaire analysis treats the data as discrete and controllable. Though the computation such as average and standard deviation considers the discrete frequency distribution, it stays on the initial data got from survey and overlooks the possibility nature in the process of survey, and hence it does not consider the possibility distribution trend. Monte Carlo Simulation is a useful tool to study the possible probability distribution if trying more times. Hence, we take it to explore into the information

**Project sources:** The Research Project of Postdoctoral Fellow of The Hong Kong Polytechnic University.

concealed behind the survey. The data used is from a questionnaire aimed to find the key assessment indicators for measuring the benefits of investment in rural infrastructure.

## II. MONTE CARLO SIMULATION

Monte Carlo Simulation (MCS) is a numerical method to solve the problems in mathematics, physics, engineering and production management, etc. by randomly sampling relevant stochastic variable or process<sup>[1]</sup>. The basic idea is to construct a stochastic variable or process, the parameter of which is related to the problems, determine the probability distributions of the variables concerned, then sample from these distributions by means of random numbers to obtain data and finally attain the results by the transformation methods defined before based on the statistics<sup>[2]</sup>. Since introduced by Metropolis and Ulam<sup>[3]</sup>, Monte Carlo Simulation has been applied in many disciplines. It can not only solve the deterministic mathematics problems<sup>[4]</sup>, but also the stochastic ones<sup>[5]</sup>. Furthermore, it is used in the disciplines, where data acquisition is hardly possible or consuming much labor and cost, such as engineering, telecommunications and finance<sup>[6]</sup> or project risk assessment<sup>[7]</sup>.

Monte Carlo Simulation is a useful tool applied in a situation where there is uncertain and uncontrollable input information whose probability distribution is known and can be

handled analytically<sup>[2]</sup>, therefore it is used in this paper to explore the possibility distribution trend of the questionnaire. There are various commercial packages available for conducting Monte Carlo simulations analyses. In this

Table 1, based on the probability distribution attained from questionnaire, the Monte Carlo Simulation indicates that with the increase of the survey, the opinion towards specific items shall be concentrated, though it is different from, or even less than the original computation value.

demonstration exercise, the package Crystal Ball was used.

### III. THE DIFFERENCE OF RANKING BY MEAN

As shown in

That's to say, if we conduct more surveys, the result shall be inclined to the simulation results, which is more scientific and useful, especially when we conduct the following analysis, such as ranking by the average value.

**Table 1 The descending rank by the average value**

Indicators	Code	From original questionnaire			From simulation		
		Mean	SD	Ranking	Mean	SD	Ranking
IRR (Internal Rate of Return)	X <sub>11</sub>	6.22	1.67	15	5.81	1.57	15
NPV (Net Present Value)	X <sub>12</sub>	6.08	1.63	17	5.66	1.53	16*
Payback (dynamic)	X <sub>13</sub>	6.37	1.89	14	6	1.73	13*
Loan repayment period	X <sub>14</sub>	5.92	1.58	20	5.46	1.55	19*
EIRR (Economic Internal Rate of Return)	X <sub>15</sub>	5.56	1.60	22	5.14	1.53	22
ENPV (Economic Net Present Value)	X <sub>16</sub>	5.54	1.57	23	5.01	1.67	23
(Direct and indirect) benefit-cost ratio of project	X <sub>17</sub>	5.86	1.53	21	5.35	1.55	21
Employment status	X <sub>21</sub>	6.80	1.50	8	6.3	1.53	7*
Living standard and quality	X <sub>22</sub>	6.82	1.61	7	6.22	1.78	9*
Capability to provide associated facilities	X <sub>23</sub>	7.44	1.25	1	6.88	1.39	1
Culture and education level, hygiene and health level	X <sub>24</sub>	6.63	1.69	11	6.06	1.78	12*
Safety benefit	X <sub>25</sub>	6.82	1.76	6	6.32	1.78	6
Amount of benefit compensation of project stake holders and underprivileged groups	X <sub>26</sub>	6.58	1.70	12	6.09	1.71	11*
Mutual adaptability indicator	X <sub>27</sub>	6.14	1.55	16	5.64	1.56	17*
Social risk level	X <sub>28</sub>	5.98	1.81	19	5.45	1.89	20*
Air pollution index (degree)	X <sub>31</sub>	7.15	1.57	3	6.65	1.58	3
Surface water pollution degree	X <sub>32</sub>	7.22	1.61	2	6.72	1.62	2
Solid waste pollution degree	X <sub>33</sub>	7.07	1.62	4	6.48	1.78	4
Noise pollution index	X <sub>34</sub>	6.706	1.76	9	6.23	1.77	8*
Water and soil loss impact	X <sub>35</sub>	6.86	1.66	5	6.35	1.68	5
Cultural relic and heritage preservation percentage (value)	X <sub>36</sub>	5.99	1.96	18	5.53	1.93	18
Energy saving percentage	X <sub>37</sub>	6.696	1.70	10	6.2	1.72	10
Recycled use percentage of wastes	X <sub>38</sub>	6.40	1.93	13	5.88	2.02	14*

\* indicates that the ranking position is different from that directly by original questionnaire

As illustrated in

Table 1, there is some difference between the ranking from questionnaire and simulation. For example, in the result of that from questionnaire, x22 is ranked 7<sup>th</sup>, x21 8<sup>th</sup> and x34 9<sup>th</sup>. However, in the result of that from simulation, x21 is ranked 7<sup>th</sup>, x34 8<sup>th</sup> and x22 9<sup>th</sup>. In other words, the ranking among the three is quite different if more surveys are conducted, which on the other hand, shows that it is not feasible to rank the indicators by the average value if the concentration trend is not considered. Therefore, the Monte Carlo Simulation is critical for us to find more information based on the

$$\tilde{A} = \mu_{\tilde{A}}(x_{11})/x_{11} + \mu_{\tilde{A}}(x_{12})/x_{12} + \dots = \sum_{i=1}^n \sum_{j=1}^m \mu_{\tilde{A}}(x_{ij})/x_{ij} \quad (\text{Equation 1})$$

Where  $x_{ij}$  is an indicator listed in

Table 1.  $n$  denotes the number of categories, which is 3, and  $m$  is the number of indicators under respect category, which is 8 for the maximum.  $\mu_{\tilde{A}}(x_{ij})$  denotes the degree of membership of  $x_{ij}$  in the fuzzy set  $\tilde{A}$ , and  $\mu_{\tilde{A}}(x_{ij}) \in [0, 1]$ . As reference to the symbols of fuzzy set (Zimmermann, 2001), ‘/’ in  $\mu_{\tilde{A}}(x_{ij})/x_{ij}$  indicates that the degree of membership of  $x_{ij}$  is  $\mu_{\tilde{A}}(x_{ij})$ , and ‘+’ might be seen as and.

As designed by the questionnaire, the significance of a specific indicator can be scored between 1 and 9, and therefore the score of 5 is seen as a neutral level for distinguish importance and unimportance. Thereafter, it is feasible to consider that the probability of over 5 is the degree membership of specific indicator to the group of importance. Based on fuzzy set theory, the possibility for a variable to belong to a group is the degree of membership of the variable in

limited questionnaire.

#### IV. THE DIFFERENCE OF FURTHER PROCESS

The difference exists not only in the ranking by mean, but also in the further process. In our research, we use fuzzy set theory to find out the key assessment indicators (KAIs).

According to the Fuzzy set theory, the symbol  $\tilde{A}$  is used to represent a set of key assessment indicators, noted as KAI set. It is designed as a fuzzy set:

the fuzzy set (Zimmermann, 2001). Hereby, the degree of membership  $\mu_{\tilde{A}}(x_{ij})$  can be described as follows:

$$\mu_{\tilde{A}}(x_{ij}) = \int_5^{\infty} P_{x_{ij}} dx = 1 - P_f \quad (\text{Equation 2})$$

Where  $P_{x_{ij}}$  represents the probability or frequency that occurs in the simulation result of particular indicator, and  $P_f$  indicates that the possibility that the indicator does not belong to the group of key assessment indicators.

As a result, the degree of membership  $\mu_{\tilde{A}}(x_{ij})$  can be calculated using Equation 2. In order to identify whether or not an indicator is a KAI, a benchmark value should be preset. In other words, the indicator shall be considered as a key assessment indicator, if it meets a certain given value ( $\lambda$ ). In this study,  $\lambda=0.85$  is adopted as the criterion to select KAIs. The selected out KAIs based on original

questionnaire and MCS is a little different as shown in

Table 2.

**Table 2 The KAIs selected out by Fuzzy Set theory based on original questionnaire and MCS**

Original KAIs	X <sub>21</sub>	X <sub>22</sub>	X <sub>23</sub>	X <sub>24</sub>	X <sub>25</sub>	X <sub>26</sub>	X <sub>31</sub>	X <sub>32</sub>	X <sub>33</sub>	X <sub>34</sub>	X <sub>35</sub>	X <sub>37</sub>
membership	0.940	0.943	1.000	0.888	0.942	0.913	0.966	0.968	0.951	0.873	0.918	0.885
Simulation KAIs	X <sub>21</sub>	X <sub>22</sub>	X <sub>23</sub>	X <sub>25</sub>	X <sub>31</sub>	X <sub>32</sub>	X <sub>33</sub>	X <sub>35</sub>				
membership	1.000	0.859	1.000	0.864	1.000	0.923	0.856	0.884				

## V. DISCUSSION AND CONCLUSION

Traditional questionnaire analysis treats the data as discrete and controllable. Though the computation such as average and standard deviation considers the discrete frequency distribution, it stays on the initial data got from survey and overlooks the possibility nature in the process of survey, and hence it does not consider the possibility distribution trend. Monte Carlo

Simulation is proposed to explore into the information concealed behind the questionnaire. The difference between the results based on original questionnaire and MCS shows that the trying is beneficial and interesting. However, the assumption is that the questionnaire survey must be scientific and hence the sample is useful to reflect the real distribution. For a new application, more foundational research should be conducted to test the validity.

## REFERENCE

- [1] Z. J. Xu, "The method of Monte Carlo". 1<sup>st</sup> ed., Shanghai: The Science & Technology Press in Shanghai, 1985.
- [2] L. Y. Shen, "Simulation in construction estimation". Working paper No. 8. The Hong Kong Polytechnic University, 1993, pp. 15.
- [3] N. Metropolis, S. Ulam, "The Monte Carlo Method". Journal of the American Statistical Association, Vol. 44, No. 247, 1949, pp. 335-341.
- [4] G. E. Forsythe, R. A. Leibler, "Matrix inversion by a Monte Carlo method". Mathematical Tables and Other Aids to Computation, Vol. 4, No.31, 1950, pp. 127-129.
- [5] H. H. Lean, W. W. Wong, X. B. Zhang, "The sizes and powers of some stochastic dominance tests: a Monte Carlo study for correlated and heteroskedastic distribution". Mathematics and Computers in Simulation, Vol. 79, No. 1, 2008, pp. 30-48.
- [6] A. Z. Cabeza, A. M. Ridao, E. F. Camacho, "Using a risk-based approach to project scheduling: a case illustration from semiconductor manufacturing". European Journal of Operational Research, Vol. 190, No. 3, 2008, pp. 708-723.
- [7] L. Y. Shen, Y. Z. Wu, "Risk concession model for

build/operate/transfer contract projects". Journal of Construction Engineering and Management, Vol. 131, No. 2, 2005, pp. 211-220.