

The Relation Research of Data Quality to Decision Quality and Results in Emergency Decision-making

Shengli Pan, Jun Tian

Information Management and E-Business Department, Xi'an Jiaotong University, Xi'an, China.
1317365392@qq.com, tianjun@mail.xjtu.edu.cn.

Abstract - The data and information usually are inaccurate in emergency decision, which influence decision results and quality. The paper aims explore how to determine relation of data quality to decision results and quality in emergency decision-making. The paper methods apply Bayes decision, involve prior and posterior probability. The paper context includes emergency-making generalization model, match function, maximum expected value function, data quality value, sensitivity analysis of data result and quality, inspect a case. The paper contribution and innovation lie in that supplements a new angle and method to explore relation of data quality to decision result and quality

Index Terms - Emergency decision, Bayes decision, decision result, decision quality, data quality

I. Introduction

It is difficult that make accuracy decision adopting these inaccuracy and incomplete data in limited time when an emergency occurs. Additional, rescues resource supplies and disaster demands are limited and expected value maximum principle is not suit obviously. So, the paper expands research based on above features.

The paper aims explore how to determine relation of data quality to decision results and quality in emergency decision-making. The paper methods apply Bayes decision, involve prior and posterior probability. The paper context includes emergency-making generalization model, match function, maximum expected value function, data quality value, sensitivity analysis of data result and quality, inspect a case.

The paper contribution and innovation lie in that supplements a new angle and method to explore relation of data quality to decision result and quality.

The paper is structured as follows. Section 2 discusses prior research on this field. Section 3.1 structures generalization decision model. Section 3.2 gives match function. Section 3.3 gives data quality matrix. Section 3.4 gives data value. Section 3.5 gives sensitivity analysis of decision results and quality, data quality. Section 4 inspects method through a case. Section 5 summaries the paper and pointes the next research.

II. Relative literatures

According to time and emphasis, some typical relative literatures are being reviewed from different angles.

Cooper (1983) points information decision quality [3]. Eisenhard (1992) considers poor quality data is key reason of poor decision quality [5].

Srinivasan (1999) researches impact of information quality and decision-maker quality on decision quality [13], while other articles previously pay more attention on influence factors of decision-maker to decision quality. Fisher (2003) researches that the impact of experience and time on the use of data quality information in decision making [4]. Wang (1996) proposes that data quality to consumers is "fit to use" [14].

Wonjin Jung (2004) proposes to provide a basis for predicting whether data quality, accessibility data quality, influences decision performance [15]. Frank (2007) uses a fine method that a comparison between the loads on a system S compared with the resistance of the system R as design to measure decision quality in engineering design [6]. Al-Abdullah (2011) refines a framework to enhance decision outcomes from data quality perspective [8].

Chengalur-Smith and Ballou (1999) adopt three data quality factors-complacency, consensus, and consistency - to explore relation between decision quality and information quality [2]. Shankaranarayanan (2006) uses completeness -a data quality dimension-in decision-making [11]. Parsian (2006) research aggregated information used by managers in their decision making processes could suffer from data errors which make an impact on decision quality [9]. Shankaranarayanan (2012) research deeply the impact of data quality metadata to decision making [12].

Robin (1992) gives a method that use data quality measures in decision-making [10], which explicitly proposes using Bayes to judge numerical probability. Albert (1997) point Bayes rules that is a data oriented approach [1]. Zebda (2006) adopted Bayesian decision in quality control decision [17]. Davern (2008) uses decision tree to diagnosing decision quality [7]. Zhang Qing-Qing (2012) gives research about risk-based water quality decision-making under small data using Bayesian network [16], that maybe a future hotspot to general and emergency decision quality.

Different above literature's algorithm level, the paper focuses on how to influence data quality to decision quality and results in emergency decision-making from process control on decision level, which not only give plan and forecast results but also analysis influence mechanism exist in activities.

III. Emergency decision-making and data quality

There are two points need to be mentioned previously. First, data are inaccurate under emergency conditions and a feasible way that adopts Bayes formula to correct data's

inaccurate. Second, disaster demands and rescues resources supplies are limited, common revenue function, loss function and decision rules of maximum expected value are not suit. So, emergency decision should be combed in model, methods, value function and decision rules.

A . Emergency decision-making model

In order to comprehend relative basic concepts, emergency scenario needs to be abstracted out previously: Nature disaster happens in some area. According history records, the probability of serious disaster is 0.8, the probability of common disaster is 0.2. Both are research objectives. The accuracy degree of serious disaster is 0.90, the accuracy degree of common disaster is 0.80. According to population and builds data, approximate 800 may lost in serious disaster, approximate 300 may lost in serious disaster. The capacity of medium size medical corps is 600, the capacity of small size medical corps is 400. Send the medium size medical corps or the small one ?

Generalization emergency model and variables can be abstracted out from scenario as follow.

State set $\Theta = \{\theta\}$, θ represent a state, which may be discrete variable or continuous variable, such as $\Theta = \{\theta_1, \theta_2\}$ or $\Theta = \{100 \leq \theta \leq 1200\}$.

Probability distribution of state variable $P(\theta)$.

Action sets $A = \{\alpha(\theta)\}$, α represent a action that decision maker adopted in state θ .

Disaster demands $d(\theta)$. **Resources supplies** $r(\theta)$.

Case: θ_1 means serous disaster, θ_2 means common disaster, $P(\theta_j) = (0.8 \ 0.2)$, α_1 means sending medium size medical corps, α_2 means sending small one.

B . Value function and decision rules

Emergency decision-making involves several important factors: value function, decision rules, and data quality factor which is researched in 3.3. Value function includes revenue function, loss function, cost function... etc, but better function—match function will be deduced.

Revenue function. $Q_j = Q(\alpha_i, \theta_j)$ Represents revenue that decision maker adopted action α_i in state θ_j , revenue function can be shown by matrix. Case's revenue function:

$$Q = (q)_{ij} = \begin{matrix} & \theta_1 & \theta_2 \\ \alpha_1 & \begin{bmatrix} 600 & 300 \end{bmatrix} \\ \alpha_2 & \begin{bmatrix} 400 & 300 \end{bmatrix} \end{matrix}$$

Calculation details: Expected value of disaster demand is $0.8 \times 800 + 0.2 \times 300 = 700$. For medium size medical corps, expected value of rescue revenue is $0.8 \times 600 + 0.2 \times 300 = 540$. For small size medical corps, expected value of rescue revenue is $0.8 \times 400 + 0.2 \times 300 = 380$.

Obviously, the optimal decision is sending medium medical corps. But the decision is not good. To provide surplus resource is always optimal solution to any disaster under adopting common revenue function and maximum

expectation value. The reason lies in that revenue notion implies that disaster demands and resources supplies are unrestricted, the maximum expectation value is not suit here.

Match function. In order to eliminate defects of common revenue function, the paper structures match function $M(d, r)$ to measure decision quality.

$d(\theta) - r(\theta)$ shows difference between disaster demand and resource supply at state θ , $d(\theta) - r(\theta) > 0$ means resources are surplus, $d(\theta) - r(\theta) = 0$ means resources are just right, $d(\theta) - r(\theta) < 0$ means resource is shortage.

Match function:

$$m_j = |d_i(\theta) - r_j(\theta)| > 0 \quad (1)$$

Case's match function:

$$M = (m)_{ij} = \begin{matrix} & \theta_1 & \theta_2 \\ \alpha_1 & \begin{bmatrix} -200 & 300 \end{bmatrix} \\ \alpha_2 & \begin{bmatrix} -400 & 100 \end{bmatrix} \end{matrix} = \begin{matrix} & \theta_1 & \theta_2 \\ \alpha_1 & \begin{bmatrix} 200 & 300 \end{bmatrix} \\ \alpha_2 & \begin{bmatrix} 400 & 100 \end{bmatrix} \end{matrix}$$

Supplement: $||$ converts match function to monotone function. Giving up $||$, shortage and surplus may be expressed precisely, but which make decision principle complicated. See figure 1.

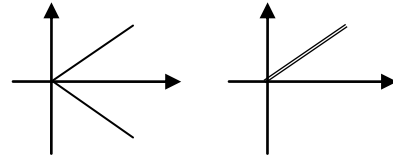


Figure 1 Principle of match function

Decision rules. Minimum expected value of match function is regarded as decision principle, which indicates decision-making quality.

$$m = \min \{E(\alpha_i)\} \quad (2)$$

Calculation details: Expected match value of medium size medical corps is $0.8 \times 200 + 0.2 \times 300 = 220$. Expected match value of small size medical corps is $0.8 \times 400 + 0.2 \times 100 = 340$, $\min\{220, 340\} = 220$.

The optimal decision is sending medium size corps. Obviously, match function overall and precisely evaluates decision quality than revenue function.

Further, expected value can be improved in efficiency view:

$$\beta = \min \{E(\alpha_i) / r\} \quad (3)$$

Obviously, $|220/600| = 0.37 < |340/400| = 0.85$.

C . Data quality matrix

Random event occurred H (discrete variable) or random variable assigned τ (continues variable) shows data value has inspected.

Prior probability $P(H|\theta)$ means H 's reliability after state θ happening. Posterior probability $P(\theta|H)$ means θ 's reliability after H happening.

$P(\theta|H)$ is derived from $P(H|\theta)$ by Bayes function, both represent data quality in decision-making viewpoint.

(1) Data quality prior probability

$$P(H|\theta) = \begin{matrix} H_1 \\ H_2 \\ \dots \\ H_m \end{matrix} \begin{bmatrix} \theta_1 & \theta_2 & \dots & \theta_n \\ P(H_1|\theta_1) & P(H_1|\theta_2) & \dots & P(H_1|\theta_n) \\ P(H_2|\theta_1) & P(H_2|\theta_2) & \dots & P(H_2|\theta_n) \\ \dots & \dots & \dots & \dots \\ P(H_m|\theta_1) & P(H_m|\theta_2) & \dots & P(H_m|\theta_n) \end{bmatrix} \quad (4)$$

$$\text{Case: } \begin{matrix} H_1 \\ H_2 \end{matrix} \begin{bmatrix} P(H_1|\theta_1) & P(H_1|\theta_2) \\ 0.90 & 0.20 \\ 0.10 & 0.80 \end{bmatrix}$$

(2) Data quality posterior probability

Discrete variable form: $\{\theta_j\}$ ($j=1,2,\dots,n$), $\theta_i \cap \theta_j = \Phi$, ($i, j=1,2,\dots,n; i \neq j$), $\sum_{j=1}^n \theta_j = \Omega$, random variable value H .

$$P(\theta_i|H) = \frac{P(H|\theta_i)P(\theta_i)}{\sum_{j=1}^n P(H|\theta_j)P(\theta_j)}, P(\theta_j) > 0, i, j=1,2,\dots,n \quad (5)$$

$$\text{Case: } \begin{matrix} H_1 \\ H_2 \end{matrix} \begin{bmatrix} P(\theta_1|H_1) & P(\theta_2|H_1) \\ 0.9474 & 0.0526 \\ 0.3333 & 0.6667 \end{bmatrix}$$

Continuous variable form: Random variable θ , probability density $p(\theta)$, random variable value τ . $h(\tau)$, density function, $k(\theta|\tau)$ and $\pi(\tau|\theta)$, condition density function.

$$k(\theta|\tau) = \frac{\pi(\tau|\theta)P(\theta)}{\int_{-\infty}^{+\infty} \pi(\tau|\theta)P(\theta)d(\theta)}, h(\tau) > 0 \quad (6)$$

The equation (1) (2) (4) (5) are adopted in section 4.1.

D. Expected value of data quality (EVDQ)

Admittedly, the EVDQ notation is derived from expected value additional information, which measure data quality value and whether to improve data quality. $EVDQ > 0$ means data quality is benefit to decision results and need a action to improve data quality, $EVDQ < 0$ means data quality has nothing to do with decision results and need not a action to improve data quality.

Discrete variable:

$$EVDQ = \sum_{j=1}^n \left[\sum_{i=1}^m [M(\alpha_{opt}, \theta_i) - M(\alpha_i, \theta_i)] P(\theta_i | \alpha_i) \right] P(\alpha_i) - E(\alpha_{opt}) - \sum_{i=1}^m [M(\alpha_i, \theta_i) P(\theta_i | \alpha_i)] P(\alpha_i) \quad (7)$$

Continues variable:

$$EVDQ = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} [M(\alpha(\tau), \theta) - M(\alpha_{opt}, \theta)] k(\theta|\tau) d(\theta) \right] h(\tau) d\tau \quad (8)$$

E. Sensitivity analysis

Usually, decision quality can be judged through comparing different $E(\alpha_i | H_i)$ to obtain function:

$$F(P(\theta_j), P(H_i | \theta_i), P(\theta_i | H_i), m_j) \quad (9)$$

Different decision results and corresponding data quality can be integrated to function in H_i :

$$E(\alpha_{opt} | H_i) < E(\alpha_i | H_i) \quad (10)$$

The equation (9) (10) are adopted in section 4.2.

Discrete variable: when data's reliability is H_i ,

$$\sum_j^n P(\theta_j | H_{opt,j}) \cdot m_{i,j} < \sum_j^n P(\theta_j | H_{i,j}) \cdot m_{i,j} \quad (11)$$

Continues variable:

$$\frac{\pi(\tau|\theta)P(\theta)}{\int_{\alpha_{opt}}^{b_{opt}} \pi(\tau|\theta)P(\theta)d(\theta)} \cdot m_{i,j} < \frac{\pi(\tau|\theta)P(\theta)}{\int_a^{b'} \pi(\tau|\theta)P(\theta)d(\theta)} \cdot m_{i,j} \quad (12)$$

IV. Emergency Decision Cases

A. Calculation

(1) Not involve data quality

$$E(\alpha_1) = \sum_{j=1}^2 m_{1j} P(\theta_j) = 220, E(\alpha_2) = \sum_{j=1}^2 m_{2j} P(\theta_j) = 340, \min\{220, 340\},$$

$$\alpha_{opt} = \alpha_1.$$

(2) Involve data quality

$$\text{When correct state is } H_1, E(\alpha_1 | H_1) = \sum_{j=1}^2 P(\theta_j | H_1) \cdot m_{1j} = 205,$$

$$E(\alpha_2 | H_1) = \sum_{j=1}^2 P(\theta_j | H_1) \cdot m_{2j} = 384, \min\{205, 384\}, \alpha(H_1) = \alpha_1$$

$$\text{When correct state is } H_2, E(\alpha_1 | H_2) = \sum_{j=1}^2 P(\theta_j | H_2) \cdot m_{1j} = 267$$

$$E(\alpha_2 | H_2) = \sum_{j=1}^2 P(\theta_j | H_2) \cdot m_{2j} = 199, \min\{267, 199\}, \alpha(H_2) = \alpha_2$$

(3) Value of data quality

In case, $EVDQ = (220) - (205 \times 0.74 + 199 \times 0.24) = 16$ indicates that the range of involving data quality decision result less 16 persons than not involving data quality. This proves correcting data inaccuracy can improve decision quality.

B. Sensitivity analysis

Influence of data quality to decision quality and result, the value of data quality, the relation of data and data quality will be analysed as follows.

(1) Relation of data quality to decision quality and result

The influence from data quality to decision is displayed in Table 1 and table 2, that show sensitivity analysis of decision quality and data quality under disaster demands and supplies having being given. These data are taken from case.

Table 1 Data quality of partial data

	$P(\theta_j)$		$P(H_1 \theta_j)$		$P(H_2 \theta_j)$	
	$P(\theta_1)$	$P(\theta_2)$	$P(H_1 \theta_1)$	$P(H_1 \theta_2)$	$P(H_2 \theta_1)$	$P(H_2 \theta_2)$
1	0.1	0.9	0.0	1.0	1.0	0.0
2	0.2	0.8	0.1	0.9	0.9	0.1
3	0.3	0.7	0.2	0.8	0.8	0.2
4	0.4	0.6	0.3	0.7	0.7	0.3
5	0.5	0.5	0.4	0.6	0.6	0.4
6	0.6	0.4	0.5	0.5	0.5	0.5
7	0.7	0.3	0.6	0.4	0.4	0.6
8	0.8	0.2	0.7	0.3	0.3	0.7
9	0.9	0.1	0.8	0.2	0.2	0.8
10	1.0	0	0.9	0.1	0.1	0.9

Table 2 Decision results and quality of partial data quality

	$E(\alpha_i)$			$E(\alpha_i H_1)$			$E(\alpha_i H_2)$		
	$E(\alpha_1)$	$E(\alpha_2)$	Opt1	$E(\alpha_1 H_1)$	$E(\alpha_2 H_1)$	Opt2	$E(\alpha_1 H_2)$	$E(\alpha_2 H_2)$	Opt3
1	290	130	130	300	100	100	200	400	200
2	280	160	160	298	109	109	231	308	231
3	270	190	190	291	130	130	237	290	237
4	260	220	220	278	167	167	240	283	240
5	250	250	250	260	220	220	240	280	240
6	240	280	240	240	280	240	240	280	240
7	230	310	230	223	334	223	240	283	240
8	220	340	220	210	371	210	237	290	237
9	210	370	210	203	392	203	231	308	231
10	200	400	200	200	400	200	200	400	200

From table 1 and table 2, $P(H_1 | \theta_1)$ and $P(H_2 | \theta_2)$ are monotone decreasing function, but every optimal decision result is non-monotone function, exist peak value. Figure 1 reflects results of table 1 and table 2. Every opt-decision result is peak shape and two sides match well, which mean decision-making quality better, but a shocked fact is –low quality data can derive high quality decision-making results, such as row 1 and 2 in table 1 and table 2. Conclusions can be summarized as follows:

Conclusion 1: decision-making result and quality are related to data quality.

Conclusion 2: given data, data quality is monotone while decision-making results and duality is peak.

Conclusion 3: high data deduce out high quality decision-making result, but high quality decision-making result may not directly root in high quality data.

(2) The value of data quality

Table 3 The value of data quality

	$E(\alpha_i)$			EVDQ
	Opt1	Opt2	Opt3	
1	130	100	200	60
2	160	109	231	140
3	190	130	237	200
4	220	167	240	241
5	250	220	240	260
6	240	240	240	240
7	230	223	240	220
8	220	210	237	180
9	210	203	231	120
10	200	200	200	40

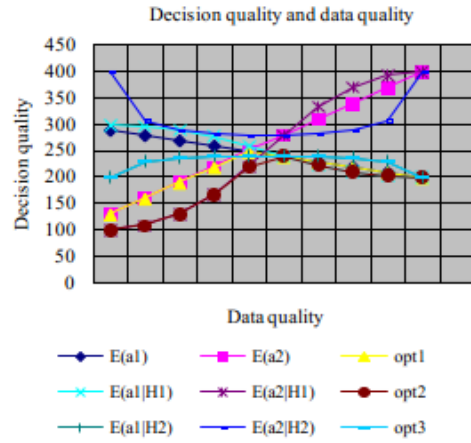


Figure 2 The curve of decision quality and data quality

(3) The relation of data and data quality

Given M and H_1 , the relation of $P(H_2 | \theta_2)$, $P(H_1 | \theta_1)$ and $P(\theta_1)$ is deduced out and displayed as follow in figure 3:

$$\frac{P(H_1 | \theta_1)P(\theta_1)}{[1 - P(H_2 | \theta_2)][1 - P(\theta_1)]} > \frac{m_{11} - m_{21}}{m_{22} - m_{12}} = 1$$

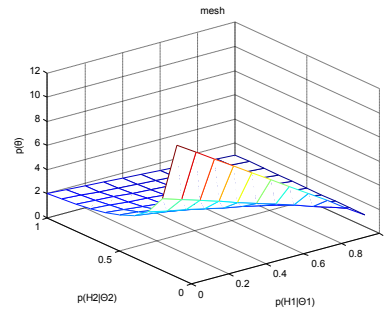


Figure 3 The relation of $P(H_2 | \theta_2)$, $P(H_1 | \theta_1)$ and $P(\theta_1)$

Given M and H_2 , the relation of $P(H_2 | \theta_2)$, $P(H_1 | \theta_1)$ and $P(\theta_1)$ is deduced out and displayed as follow in figure 4:

$$\frac{P(H_2 | \theta_2)[1 - P(\theta_1)]}{[1 - P(H_1 | \theta_1)]P(\theta_1)} > \frac{m_{11} - m_{21}}{m_{22} - m_{12}} = 1$$

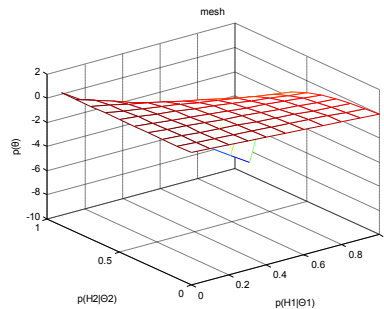


Figure 4 The relation of $P(H_2 | \theta_2)$, $P(H_1 | \theta_1)$ and $P(\theta_1)$

From figure 2 and figure 3, α_i stabilize in limited scope ($0 \leq P(\theta_1), P(H_1 | \theta_1), P(H_2 | \theta_2) \leq 1$), which means the data quality small changes would arouse decision-making results and quality great changes.

V. Summary

The paper proposes generalization emergency decision model and method, reveals the relation of data quality to decision result and quality by introducing Bayes theory for general and emergency decision. The method can be adopted by decision support systems.

However, there still exist some defects. The explore does not involved time that improves data quality. If disaster demand is time function, what is law of data quality to decision results and quality accompany with time? Moreover, sequential and network decisions approximate reality, what is relation of data quality to decision results and data quality? These are worthy to be researched in next.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (NSFC) under Grant #70971108 and 71171157.

References

- [1] Albert, J., Teaching Bayes' rule: A data-oriented approach. *The American Statistician*, Aug 1997, (51)3, 247~253.
- [2] Chengalur-Smith, I., Ballou, D., and Pazer, H., The impact of data quality information on decision making: an exploratory analysis, *IEEE Transactions on Knowledge and Data Engineering*, 1999, 11, 6, 853~864.
- [3] Cooper, R.B., Decision production: a step toward a theory of decision-maker information requirements. *International Conference on Information System*, 1983(45), 251~268.
- [4] Craig, W., Fisher, Indu Shobha, Chengalur-Smith, Donald, P., Ballou, The Impact of Experience and Time on the Use of Data Quality Information in Decision Making. *Information System Research*, Vol. 14, No. 2, 2003, 170~188.
- [5] Eisenhard, K.M. and Zbrackim, M.J. Strategic decision-making, *Strategic Management Journal*, 1992(13), 1~13.
- [6] Frank, A.U., Assessing the quality of data with decision model [P]. *5 the International Symposium on Spatial Data Quality*, 2007,
- [7] Michale J. Davern, Ravi Mantena, Edward A. Stohr. Diagnosing decision quality. *Decision Support System*, 2008, 123~139.
- [8] Muhammand Al-Abdullah, Heinz Roland Weistroffer. A framework to enhance decision outcomes: data quality perspective. *Proceedings of The Southern Association Systems Conference*, Atlanta, GA, USA march 25th-26th, 2011, 1~5.
- [9] Parssian, A., Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions. *Decision Support System*, 2006(42), 1494~1502.
- [10] Robin, A.D., Using data quality measures in decision-making algorithms. *IEEE Expert*, 1992, 63~72.
- [11] Shankaranarayanan, G., Yu Cai. Supporting data quality management in decision-making. *Decision Support System*, 2006(42), 302~317.
- [12] Shankaranarayanan, G., Bin, Zhun., Data quality metadata and decision making. *2012 45th Hawaii international conference on system sciences*, 2012, 1434~1443.
- [13] Srinivasan, R., Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis. *Decision Support System*, 1999(26), 275~286.
- [14] Wang, Y.R., and Strong, D.M., Beyond accuracy: what data quality means to data consumers, *Journal of Management Information Systems*, 1996(12), 5~34.
- [15] Wonjin Jung. A review of research: An investigation of the impact of data quality on decision performance. *Proceedings of the 2004 International Symposium on Information and Communication Technologies*, 2004, 161~171.
- [16] Zhang Qing-Qing, XU Yue, Tian Ye, Zhang Xu-jie. Risk-based water quality decision-making under small data using Bayesian network. *J. Cent. South Univ.* 2012(19):3215~3224.
- [17] Zebda, A., The use of Bayesian decision analysis and strategy matrix for quality control decision. *The Business Review, Cambridge*, 2006(5), 1:287~293.