

# Uncertainty analysis method based on fuzzy random variables and time window selection

**Molin Sun and Zhongyi Zheng**

Navigation College, Dalian Maritime University, 1 Linghai Road, Dalian, China

[molinhabin@sina.com](mailto:molinhabin@sina.com)

**Abstract.** The uncertainty analysis has been conceived as a necessary step in the engineering research in various fields. The purpose of this article is to introduce the uncertainty analysis technique into the process of statistical modelling. Based on the theory of fuzzy random variables, aleatory and epistemic uncertainties in the statistical modelling process can be modeled by fuzzy random variables, which are interpreted as random variables with fuzzy statistical properties. An important aspect of modelling uncertainties lies in the appropriate selection of time window, which is used for the inclusion of data. In order to avoid the arbitrary choice of time window, a simple method for time window selection is proposed by analyzing the uncertainty and the stability of statistic data. Finally, a case study is carried out on the accident statistics. The results show that the proposed methods are effective and can provide more information for decision-making.

## 1. Introduction

The uncertainty analysis investigates the uncertainty of variables that are used in decision-making problems. In other words, the uncertainty analysis can investigate the robustness of an engineering research when the research includes some form of statistical modelling through the quantification of uncertainties in the relevant variables. In general, uncertainty is considered of two different types: aleatory and epistemic uncertainties. The aleatory uncertainty arises from randomness due to inherent variability, and the epistemic uncertainty refers to imprecision due to lack of knowledge or information <sup>[1]</sup>.

There are several methods that can be used to model uncertainties. When there is sufficient information for statistical analysis, probability distributions are often assigned to model aleatory uncertainties <sup>[2]</sup>. In terms of modelling epistemic uncertainties, a number of representation frameworks have been proposed. These include the fuzzy set theory, the possibility theory, the interval analysis and the evidence theory <sup>[3]</sup>. In addition, several approaches have been proposed to model aleatory and epistemic uncertainties simultaneously, such as the theory of fuzzy random variables based approach <sup>[4]</sup>, the possibilistic-Monte Carlo approach <sup>[5]</sup>, the possibilistic-scenario based approach <sup>[6]</sup> and the evidence theory based hybrid approach <sup>[7]</sup>. Among them, the theory of fuzzy random variables based approach has received growing attention because of its representation power and its relative mathematical simplicity. Since aleatory and epistemic uncertainties are very common in the process of statistical modelling, the theory of fuzzy random variables is selected to model uncertainties in this article.

An important aspect of modelling uncertainty lies in the appropriate selection of time window, which is used for the inclusion of data. Traditional empirical approach can lead to either too precautionary or non-conservative estimates of the magnitude of uncertainties based on the arbitrary



choice of the length of time window [8]. Although the time window is an important factor for uncertainties modelling, it has received little empirical study. To reasonably characterize the uncertainty in the engineering research, it is considered necessary to propose a method to select the optimal time window used for the inclusion of data.

## 2. Uncertainties modelling

The theory of fuzzy random variables was first proposed by Kwakernaak in 1978 [9] and can be used to address aleatory and epistemic uncertainties simultaneously. Fuzzy random variables can thus be interpreted as random variables with fuzzy statistical properties, leading to multiple uncertainties. In other words, when using probability distributions to model aleatory uncertainties, parameters of the selected probability distributions are represented by fuzzy numbers instead of crisp numbers.

The selection of probability distributions is an important aspect of modelling aleatory uncertainties. In the process of statistical modelling, the discrete data obtained from counting the numbers of occurrences of events during specified periods of time are often encountered. Since Poisson distribution is a suitable model for modelling the above type of data [10], it is selected as the probability distribution for modelling aleatory uncertainties in this study. According to the Poisson distribution properties, the confidence interval of the number of times an event occurs can be calculated by [10]:

$$\begin{aligned}\lambda_U &= \frac{\chi_{1-\omega/2}^2(2n+2)}{2} \\ \lambda_L &= \frac{\chi_{\omega/2}^2(2n)}{2}\end{aligned}\quad (1)$$

where  $\lambda_U$  and  $\lambda_L$  are the upper and lower boundaries for the confidence interval of the mean value of a Poisson distribution;  $n$  is the number of occurrences of events during a specified time window;  $\omega$  is the significance level of the statistics;  $\chi_{1-\omega/2}^2(2n+2)$  is the  $(1-\omega/2)$ th quantile of the chi-squared distribution with  $(2n+2)$  degrees of freedom;  $\chi_{\omega/2}^2(2n)$  is the  $(\omega/2)$ th quantile of the chi-squared distribution with  $(2n)$  degrees of freedom;  $\chi_{1-\omega/2}^2(2n+2)$  and  $\chi_{\omega/2}^2(2n)$  can be found in the table of chi-squared distribution.

For parameters of the Poisson distribution with epistemic uncertainties, fuzzy numbers are used to model the epistemic uncertainty. According to the interpretation of the uncertainty in these parameters, the range of values of parameters can be estimated roughly by expert judgments. In order to introduce epistemic uncertainties as small as possible, we use directly value ranges rather than select fuzzy distributions to represent epistemic uncertainties. When there are more interpretations of parameters with epistemic uncertainties, the usage of fuzzy distributions is possible.

## 3. Method for time window selection

Time window can be taken as the time interval for which historical data are collected. It starts from the research date and goes backwards [11]. However, traditional empirical approach can lead to either too precautionary or non-conservative estimates of the magnitude of uncertainties based on the arbitrary choice of the length of time window. To reduce the subjectivity, it is considered necessary to propose a simple method for time window selection. It should be noted that the longer the time window is, the more informative data for statistical analysis, and the more accurate uncertainty modelling is. It means that time window used for the inclusion of data should be as long as possible. There are also indications that more recent statistics represent a more conclusive database than old statistics reflecting recent technical developments or new requirements. To coordinate the contradiction described above, the uncertainty degree and the stability degree of statistic data are used as the two indexes to determine the optimal length of the time window, denoted by  $S$  and  $V$ . We, thus, have [12]:

$$S = \lambda_U - \lambda_L \quad (2)$$

$$V = \left( \frac{1}{y} \sum_{i=1}^y (n_i - \bar{n})^2 \right)^{1/2} \quad (3)$$

Where  $S$  can be taken as the spread between the uncertainty boundaries  $[\lambda_L, \lambda_U]$  from the Poisson distribution under a certain level of confidence;  $V$  is defined as the standard deviation;  $n_i$  is the  $i$ -th observed value,  $\bar{n}$  is the mean value of the observed values and  $y$  is the number of observed values; Consistent with the qualitative analysis of the time window selection described above, both  $S$  and  $V$  should be as low as possible. When  $S$  and  $V$  may not reach the minimum at the same time, a comprehensive index  $T$  which is set as the product of  $S$  and  $V$  is proposed. Therefore, the length of the time window could be optimal when we get minimum  $T$ .

#### 4. Case study

Approaches for uncertainties modelling and time window selection illustrated in Section 2 and 3 have been applied to the accident statistics in the engineering report<sup>[13]</sup>. Since our goal here is to apply the proposed methods and to verify their effectiveness, the particular selection of the applied engineering field can be changed if needed. In other words, the proposed methods focus on the type of data which is discrete and obtained from counting the numbers of occurrences of events during specified periods of time. According to the engineering report, the accident statistic data for the period 2000 to 2010 are listed in table 1.

Table 1. The accident statistic data between 2000 and 2010.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
The number of accidents	10	19	21	30	25	36	25	42	47	43	40

##### 4.1. The process of uncertainties modelling

When the values in table 1 are put into equation (1), the confidence interval of the number of accidents for each year 2000–2010 can be calculated under a certain level of confidence. Taking the data in 2010 as an example, the number of accidents is 40 and the corresponding confidence interval can be calculated as<sup>[30, 52]</sup> for the confidence value 0.9. According to the research<sup>[14]</sup>, a certain percentage of accidents are missing from the records and in order to complete the uncertainty modelling of high quality, it was necessary to complement the available data with additional data. We assume the rate of under-reporting is no more than 50% based on the priori knowledge<sup>[15]</sup>. Thus the number of accidents in 2010 can be estimated as<sup>[40, 60]</sup>. When the number of accidents is 60, the confidence interval can be calculated as<sup>[48, 74]</sup> for the same confidence value. Finally, the confidence interval of the number of accidents is estimated as<sup>[30, 74]</sup> considering aleatory and epistemic uncertainties simultaneously. The data in other years can repeat the above computation process to obtain the corresponding confidence interval. The accident statistic data and the corresponding confidence interval are represented by short dashes and line segments in figure 1. For representing the change of the confidence interval more vividly, upper and lower boundaries of the confidence interval are connected with the line, respectively.

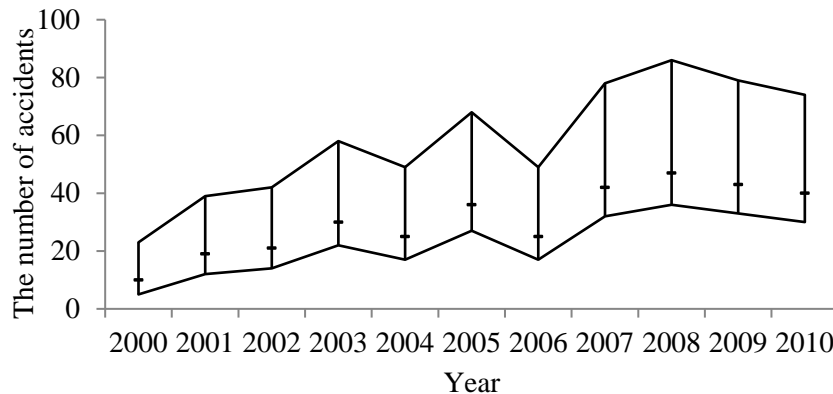


Figure 1. The accident statistic data and the corresponding confidence interval.

#### 4.2. Time window selection

In the process of selecting the length of the time window, the sliding window method is used to calculate  $S$  and  $V$  indexes. Nine time windows will be estimated, from the recent three years (2008-2010), to the recent eleven years (2000-2010). The  $S$  index can be calculated according to the uncertainty model process in Section 4.1 and equation (2). When the values in table 1 are put into equation (3), the  $V$  index can also be obtained. Table 2 shows the calculation results of  $S$  and  $V$  indexes.

Table 2. The calculation results of  $S$  and  $V$  indexes.

Time window	3	4	5	6	7	8	9	10	11
$S$	36	34	30	29	27	26	24	23	22
$V$	3.5	2.9	8.4	7.7	8.7	8.5	9.4	10.1	11.8

As can be seen in table 2,  $S$  and  $V$  may not reach the minimum at the same time. The comprehensive index  $T$ , which is the product of  $S$  and  $V$ , reaches the minimum when the time window is 4. Accident data in recent four years will be used to estimate the magnitude of uncertainties in the process of statistical modelling. Thus the confidence interval of the average number of accidents annually can be calculated as <sup>[38, 72]</sup> for the confidence value 0.9. In order to verify the effectiveness of the proposed methods, the confidence interval calculated above is compared with another two statistical results, which are estimated by the data from recent one year and recent eleven years. According to Section 4.1, the confidence interval of the number of accidents in 2010 is estimated as <sup>[30, 74]</sup>. It is obvious that the confidence interval calculated by the proposed methods is smaller because there are more informative data for statistical analysis and the uncertainty modelling is more accurate. When it comes to the data from recent eleven years, the confidence interval is calculated as <sup>[28, 50]</sup>. The statistical result does not take into account that the number of accidents presents rising trend on the whole, which can be seen in figure 1.

## 5. Conclusions

Uncertainty analysis has been conceived as a necessary step in the process of statistical modelling. In this article, fuzzy random variables are introduced to address aleatory and epistemic uncertainties simultaneously. In addition, a simple method for time window selection is put up to estimate the magnitude of uncertainties, which provides the theoretic foundation and reduces the subjectivity for determining the length of time window in the uncertainty modelling process. Finally, a case study is carried out, which shows that the proposed methods are effective and can provide more information for decision-making.

## References

- [1] Pedroni N, Zio E, Ferrario E, Pisanisi A and Couplet M 2013 *Comput. Struct.* **126** 199–213

- [2] Vierow K, Hogan K, Metzroth K and Aldemir T 2014 *Prog. Nucl. Energ.* **77** 320–328
- [3] Baraldi P and Zio E 2008 *Risk Anal.* **28** 1309–25
- [4] Wang S, Huang G H, Baetz B W and Huang W 2015 *J. Hydrol.* **530** 716–733
- [5] Babuška I and Silva R S 2014 *Comput. Method. Appl. M.* **270** 57–75
- [6] Soroudi A 2012 *IEEE T. Power Syst.* **27** 1283–93
- [7] Clavreul J, Guyonnet D, Tonini D and Christensen T H 2013 *Int. J. Life Cycle Ass.* **18** 1393–1403
- [8] Ballings M and Poel D V D 2012 *Expert Syst. Appl.* **39** 13517–22
- [9] Kwakernaak H 1978 *Inform. Sciences* **15** 1–29
- [10] Johnson N L, Kemp A W and Kotz S 2005 *Univariate Discrete distributions* (Hoboken: Wiley–Interscience) pp 176–177
- [11] Vakili S, Sobell L C, Sobell M B, Simco E R and Agrawal S 2008 *Addict. Behav.* **33** 1123–30
- [12] Zhao J and Lv J 2016 *Transport. Plan. Techn.* **39** 1–13
- [13] Konovessis D *et al* 2015 *Risk Acceptance Criteria and Risk Based Damage Stability Final Report* vol 2 (Norway: European Maritime Safety Agency) pp 28–70
- [14] Psarros G, Skjong R and Eide M S 2010 *Accident Anal. Prev.* **42** 619–625
- [15] Hassel M, Asbjørnslett B E and Hole L P 2011 *Accident Anal. Prev.* **43** 2053–63