Portfolio Optimization with Percentage Error-Based Fuzzy Random Data for Industrial Production

Mohammad Haris Haikal Othman[1(✉)], Nureize Arbaiy[1], Muhammad Shukri Che Lah[1], and Pei-Chun Lin[2]

[1] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia
harishaikal940322@gmail.com

[2] Department of Information Engineering and Computer Science, Feng Chia University, No. 100, Wenhwa Road, Taichung, Taiwan

Abstract. Data-driven decision-making processes are pervasive in various domains, yet the inherent uncertainties within observational and measurement data can lead to misleading outcomes, particularly in portfolio selection where randomness may seem ambiguous. While existing methodologies recognize the significance of data preprocessing in managing uncertainties such as fuzziness and randomness, a systematic framework to effectively address these challenges is currently lacking. This study aims to bridge this gap by presenting a comprehensive framework tailored to efficiently handle uncertainty during the preprocessing stage. The proposed framework not only acknowledges the importance of data preprocessing but also offers a systematic approach to processing fuzzy random data, thus providing a robust foundation for portfolio selection algorithms. Leveraging fuzzy integers to manage fuzziness and probability distributions to address randomness, our methodology ensures the construction of reliable portfolio selection strategies. The main objective is to optimize selection based on industrial production, effectively managing uncertainty in traditional portfolio selection models. In this proposed approach, fuzziness is handled using fuzzy numbers, and randomness is addressed through probability distributions. The efficacy of this approach is demonstrated in agricultural planning, evaluating five distinct industrial production types: Agriculture, Mining, Manufacturing, Electricity, and Water. The findings underscore the effectiveness of the proposed methodology in managing uncertainties, reducing errors in model development stages, and providing a robust framework for optimal portfolio selection tailored to industrial production contexts, thereby enhancing decision-making processes in uncertain environments.

Keywords: Fuzzy Random Variable · Fuzzy Random Data · Data Pre-processing · Mean-Variance

R. Ghazali et al. (Eds.): SCDM 2024, LNNS 1078, pp. 400–410, 2024.
https://doi.org/10.1007/978-3-031-66965-1_39

1 Introduction

Real-world data is rarely flawless, and uncertainties might come in at any point, reducing modelling and predicting accuracy. Uncertainties in collected data include measurement errors, data quality, representativeness, and bias. Uncertainties inherent in real-world scenarios apply a significant influence on forecasting and decision-making processes, complicating decision [1] parameters across various industries. Data contains fuzziness and randomness because of the ambiguities and vagueness that exist in the real world. Randomness refers to physical uncertainty, while fuzziness is caused by human cognition. Fuzzy random data can be used to describe all the information associated with a measurement result, including systematic and random components to the overall uncertainty.

Portfolio selection models often face challenges in uncertain environments, where security returns are difficult to predict based on historical data alone. To address this, various models have been proposed, such as uncertain portfolio selection models with background risk [2], those based on VaR minimization [3], and multi-period portfolio models considering expert evaluations. These models aim to enhance decision-making by incorporating uncertainty into the portfolio selection process.

The evaluation of the usefulness of portfolio selection models across diverse industrial sectors serves multiple important purposes, including validating the methodology's applicability, assessing its performance comprehensively, and demonstrating its practical relevance. By examining sectors such as Agriculture, Mining, Manufacturing, Electricity, and Water, the study ensures that the proposed methodology addresses sector-specific challenges and is effective in various real-world scenarios. This approach helps validate the robustness and generalizability of the model, making it applicable beyond specific industries.

Data containing fuzzy random uncertainties reduces the accuracy of portfolio selection models by introducing uncertainty and ambiguity into the decision-making process. In modeling uncertain phenomena [4] across different production sectors, the concepts of random variables and fuzzy theory play pivotal roles. While probability theory addresses random events, fuzzy theory offers solutions for handling fuzzy data. However, existing models often treat these uncertainties separately, overlooking their simultaneous occurrence real data.

2 Literature Review

This section provides a background on portfolio selection through the mean-variance model and the utilization of fuzzy numbers.

2.1 Mean-Variance Model

Portfolio selection entails the process of identifying an optimal portfolio. Markowitz introduced a method aimed at determining the optimal portfolio that maximizes returns while minimizing risks, framing such scenarios as portfolio selection problems. Portfolio

selection finds wide application in real-world contexts and has been expanded to address uncertainties [5, 6].

Equation (1) defines the portfolio selection model [7].

$$\begin{aligned} &\min\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i\sigma_{ij}xj\right\} \\ &s.t.\sum_{i=1}^{n} E(r_i)x_i = \rho \\ &\sum_{i=1}^{n} x_i = 1 \\ &x_i \geq 0, i = 1, \ldots, n \end{aligned} \tag{1}$$

In this context, r_i represents the random variable denoting the return, $E(r_i)$ signifies the expected value associated with r_i, σ_{ii} stands for the variance of r_i, σ_{ij} represents the covariance between r_i and r_j, x_i denotes the proportion of capital allocated to security i and ρ signifies the parameter related to the expected return.

2.2 Fuzzy Set

The concept of fuzzy sets is as follows:

Definition 1: Within the context of this analysis, let U represent the universal set of discourse. We define a fuzzy set A on U using a membership function m_A. This function assigns a real value between 0 and 1 to each element x in U, indicating the degree to which x satisfies the characteristic property of A. Formally, we can express a fuzzy set A in U as a collection of ordered pairs $A = \{(x, m_A(x)) : x \in U\}$, where $m_A : U \to [0, 1]$ represents the membership degree of x in A. Formally, a fuzzy number x is described as:

$$m_A(X) = \sum_{i=1}^{n} m_i(X)I_{A_i}(X) \tag{2}$$

where $I_{A_i}(X) = 1$ when $x \in A_i$ and $I_{A_i}(X) = 0$ if $x \notin A_i$.

Definition 2: A fuzzy set A on the universal set U, defined by the membership function $y = m(x)$, is regarded normal if there is at least one element x in U with $m(x) = 1$.

2.3 Fuzzy Random Variables

Data hybridization, which combines fuzzy data with randomness, appears as a major advance in the extension of fuzzy sets [11]. The concept of fuzzy random variables [12] was improved by [13] with a new notion of exclusion, allowing the generalization of integrals for set-valued functions.

Theorem 1. Central Limit Theorem.

Let $X_1, \ldots, X_n$ be the random variables with mean m and variance σ^2. Let

$$Z_n = \frac{n^{\frac{1}{2}}(\overline{X}_n - m)}{\sigma} = \frac{W_n - nm}{n^{\frac{1}{2}}\sigma} \tag{3}$$

$X_1, X_2, \ldots X_n$ are independent, identically distributed random variables with mean μ and standard deviation σ; Z_n is the standardised sample mean. The sample mean, or X_n, is calculated by dividing the sum of the random variables X_i by the sample size, n is the sample size, σ is the standard deviation of each individual random variable, and m is the population mean, or expected value. Z_n converges in distribution to Z as $n \to \infty$. $Zn \to Z \sim N(0, 1)$ is denoted as $n \to \infty$ where Z distribute according to a standard normal distribution function *N (0,1)*.

3 Portfolio Selection Model Using Fuzzy Random Data Based on Percentage Error on Industrial Production Index

3.1 Fuzzy Random Data Pre-processing

This section describes the methods for pre-processing data with fuzzy random variables. To account for real-world uncertainty, we use an approach based on measurement error ranges, as established in previous publications [14–16].

In this analysis, we depict the minimum potential value (A_i) and the highest potential value (B_i). Equation (4) describes the exact procedure used to generate fuzzy data from the initial single-value data points, where p is the percentage of error and ip is index production.

$$\begin{aligned} A_i &= ip - (ip * p\%) \\ B_i &= ip + (ip * p\%) \end{aligned} \tag{4}$$

The original data is transformed into fuzzy intervals $F_i = [A_i, B_i]$ to account for real-world uncertainties. The intervals represent the potential range of each data point. We also define each fuzzy interval by its center point (c_i) and width (w_i). The central point ($c_i = (A_i + B_i)/2$) represents the most likely outcome, while the width ($w_i = B_i - A_i$) indicates the level of uncertainty. Specific formulas for determining these numbers in the Eq. (5).

$$\begin{aligned} c_i &= \frac{A_i + B_i}{2} \forall i = 1, 2, \ldots n \\ w_i &= \frac{B_i - A_i}{2} \forall i = 1, 2, \ldots n \end{aligned} \tag{5}$$

This study addresses randomness in the data using four probability distributions: Normal, Weibull, Gamma, and Logistic. Each distribution generates a p-value, which indicates how well it fits the data. The distribution with the highest p-value, indicating the best fit, is selected for further analysis. After pre-processing with fuzzy random variables, the input is translated into interval form $Y_i = [C_i, W_i]$, representing potential changes around a central point. Within each interval, C_i represents the central point and W_i indicates the width of the interval.

3.2 Fuzzy Random Based Portfolio Selection Model for Industrial Production Index

Let $Y_i = (C_i, W_i)$ represent the set of interval fuzzy numbers with $\forall i = 1,2,\ldots,n$. The expected value and variance of the fuzzy interval data are determined as follows:

Expected value:

$$E(A_i) = (E(c_i), E(w_i)) \tag{6}$$

Variance:

$$\sigma^2(A_i) = (\sigma^2(c_i), \sigma^2(w_i)) \tag{7}$$

E is the expected value of interval fuzzy number and σ^2 is the variance. Portfolio formulation is given as follows:

$$\left.\begin{array}{l} \max \sum_{i=1}^{n} E(c_i)x_i \\ \min \sum_{i=1}^{n} E(w_i)x_i \\ s.t. \sum_{i=1}^{n} \sigma(c_i)x_i \\ \sum_{i=1}^{n} \sigma(w_i)x_i \\ \sum_{i=1}^{n} x_i \leq 1 \\ \sum_{i=1}^{n} \sigma(w_i)x_i \\ x_i \geq 0 \, \forall \, i = 1, 2, \ldots, n \end{array}\right\} \tag{8}$$

max $\sum_{i=1}^{n} E(c_i)x_i$ *and min* $\sum_{i=1}^{n} E(w_i)x_i$ in (8) uses expected value in center and width. Based on the findings in Sects. 3.1, 3.2, we provide an optimized approach to this stage within the method:

1. **Data Collection:** Gather relevant data.
2. **Determine Measurement Error:** Before using the data for fuzzification, it is critical to identify any potential collection inaccuracies. Quantifying this inaccuracy, such as a 5% margin, directs the fuzzification process by establishing acceptable limits for portraying data variability.
3. **Fuzzification Process:**
   - For each data point x_i, calculate the maximum and minimum potential values based on the measurement error. Let's denote these as A_i and B_i respectively.
   - Formulate the fuzzy data interval $F_i = A_i, B_i$.
4. **Calculate Central Point and Width:**
   - Determine the central point c_i of the fuzzy interval F_i.
   - Calculate the width w_i of the fuzzy interval F_i.

5. **Repeat for Each Data Point:** Perform steps 3 and 4 for each data point in the dataset.
6. **Compile Results:** Present the fuzzified data in the form of interval parameters F_i, where each F_i represents a fuzzy data interval with its respective central point and width.
7. **Calculate Probability Distribution Function (PDF):** The central point and width provide clues into distribution types, but further investigation is required for identification. While a distribution's center point and width provide indications about its basic shape (symmetric, skewed), determining the underlying PDF requires more information or analysis, such as higher-order moments or goodness-of-fit tests.
8. **Compute Portfolio Selection Model:** Use the PDF result to identify the risk level to get the most optimize result using Eq. (8).

This strategy provides an efficient method for decision-makers to create selection models while handling data uncertainty.

4 Numerical Experiment

The analysis incorporates data from five distinct production sectors: Agriculture, Mining, Manufacturing, Electricity, and Water. This data is obtained from Malaysia's official data catalog (https://data.gov.my/ms-MY/data-catalogue/ipi) and spans a three-year period, from 2015 to 2023. The primary goal of this study is to identify an investment portfolio that chooses one of these five production sectors.

Table 1. Single form data

Data/Production	Agriculture	Mining	Manufacturing	Electricity	Water
1/1/2010	92.7	99.4	98.8	99.8	99.2
1/2/2010	93.3	97.8	99	99.4	99.1
1/3/2010	96.1	108.6	99.2	100.5	99.2
1/4/2010	95.3	107.7	99.3	99.7	99.9
1/5/2010	94.1	91.4	99.8	100.3	101.5
...	...	...	...	...	...
1/12/2023	123.2	97.4	120	117	118.3

Fuzzy random data pre-processing starts here. The raw data obtained in Table 1 is then fuzzified by using 5% percentage measurement error based on Eq. (4). This resulted in fuzzy interval data in a form of $[a, b]$ where a is minimum data and b is the maximum data. Table 2 shows the percentage error data.

Table 2. Percentage error 5% data $[a_i, b_i]$

Data/Production	Agriculture	Mining	Manufacturing	Electricity	Water
1/1/2010	[88.07, 92.47]	[94.43, 99.15]	[93.86, 98.55]	[94.81, 99.55]	[94.24, 98.95]
1/2/2010	[88.64, 93.07]	[92.91, 97.56]	[94.05, 98.75]	[94.43, 99.15]	[94.15, 98.85]
1/3/2010	[91.30, 95.86]	[103.17, 108.33]	[94.24, 98.95]	[95.47, 100.25]	[94.24, 98.95]
1/4/2010	[90.54, 95.06]	[102.32, 107.43]	[94.34, 99.05]	[94.72, 99.45]	[94.91, 99.65]
...	...	...	...	...	...
1/12/2023	[117.04, 122.89]	[92.53, 97.16]	[114.00,119.70]	[111.15,116.71]	[112.39,118.00]

After the data been processed using the percentage error, the data should be fuzzy data to o and l. The data should be processed such as Table 3.

Table 3. Fuzzy data, center point, and width $[o_i, l_i]$

Date/Production	Agriculture	Mining	Manufacturing	Electricity	Water
1/1/2010	[−2.43, 2.20]	[−2.61, 2.36]	[−2.59, 2.35]	[−2.62, 2.37]	[−2.60, 2.36]
1/2/2010	[−1.85, 2.22]	[−4.17, 2.32]	[−2.40, 2.35]	[−3.01, 2.36]	[−2.70, 2.35]
1/3/2010	[0.88, 2.28]	[6.35, 2.58]	[−2.20, 2.36]	[−1.94, 2.39]	[−2.60, 2.36]
1/4/2010	[0.10, 2.26]	[5.47, 2.56]	[−2.11, 2.36]	[−2.72, 2.37]	[−1.92, 2.37]
...	...	...	...	...	...
1/12/2023	[27.27, 2.93]	[−4.56, 2.31]	[18.05, 2.85]	[14.13, 2.78]	[15.99, 2.81]

After collecting the data, the center point and width of the fuzzy data is identified as in Eq. (8). Table 4 shows the center point and width of the fuzzy data. The probability distribution function is then performed to treat the randomness. Each of the production data will provide 4 types of data which each of them will generate Normal, Log, Gamma, and Weibull distribution. Each of the distributions will provide the p-value. In this paper, the biggest p-value indicates as the best result to treat the randomness.

Table 4. Interval number of the probability distribution function

	center, C	width, W
Agriculture	N(14.7736, 18.5711)	γ (33.4912, 0.0783)
Mining	N(3.0240, 24.6628)	γ (17.2472, 0.1448)
Manufacturing	γ (1.5832, 4.4442)	W(358.2068, 0.0072)
Electricity	W(2.8662, 13.0576)	W(24.6696, 2.7416)
Water	W(1.7513, 10.2081)	W(28.1013, 2.6715)

Table 5 shows 5 types of production are considered and represents in the form of $x_n = (x_1, x_5, \ldots, x_5)$ respectively. Table 4 shows the probability distribution function that has been selected based on the highest p-value. Note that the LOG denotes logistic distribution, was the Weibull Distribution and γ as gamma distribution. The moment estimator is utilized to approximate the expected value and variance each of the vegetable.

Table 5. Expected value and variance

Production		center, C		width, W	
		expected value	variance	expected value	variance
Agriculture	x_1	14.77360	344.88390	2.62131	0.20517
Mining	x_2	3.02401	608.25370	2.49815	0.36184
Manufacturing	x_3	7.03619	488.69613	2.58137	0.01860
Electricity	x_4	2.75506	0.06620	21.95005	74.78710
Water	x_5	1.66756	0.03871	24.98108	101.46823

Table 6 shows the result. Finally, the expected value and variance for each of the production indexes is computed. The data has now completed the pre-processing phase. This pre-processed data is then presented to the portfolio selection model to identify the best portfolio.

The portfolio selection model in Eq. (8) is used to build the Model (9).

$$
\begin{aligned}
&max\ \mathbf{14.77360x_1 + 3.02401_{x2} + 7.03619_{x_3} + 2.75506x_4 + 1.66756_{x5}}\\
&min\ \mathbf{2.62131}x_1 + \mathbf{2.49815}x_2 + \mathbf{2.58137}x_3 + \mathbf{21.95005}x_4 + \mathbf{24.98108}x_5\\
&s.t.\ \ \sqrt{\mathbf{344.88390}x_1} + \sqrt{\mathbf{608.25370}x_2} + \sqrt{\mathbf{488.69613}x_3} + \sqrt{\mathbf{0.06620}x_4} + \sqrt{\mathbf{0.03871}x_5} = k\\
&\quad \sqrt{\mathbf{0.20517}x_2} + \sqrt{\mathbf{0.36184}x_2} + \sqrt{\mathbf{0.01860}x_3} + \sqrt{\mathbf{74.78710}x_4} + \sqrt{\mathbf{101.46823}x_5} \le k\\
&\quad x_1 + x_2 + x_3 + x_4 + x_5 \le 1\\
&\quad x_i \ge 0\ \forall\, i = 1, 2, \ldots, n
\end{aligned}
\tag{9}
$$

Equation (9) is solved using a linear programming approach. Here, we assume that k represents the risk level, and the optimal solution is achieved with $x_n = (1,0,0,0,0)'$, $x_n = (0,1,0,0,0)'$, $x_n = (0,0,1,0,0)'$, $x_n = (0,0,1,0,0)'$, $x_n = (0,0,0,1,0)'$, or $x_n = (0,0,0,0,1)'$. The model's computation halts upon reaching the optimal solution.

5 Result and Discussion

Table 6 shows the optimal solution results. From the results table, the risk of $k = 5.591$ indicates the optimal solution where the expected return is (7.03, 2.58) for a 5% percentage error. The optimum result based on the five industrial production data is manufacturing production.

Table 6. The result with optimal solution

Risk, k	3	4	5	5.591	6
x_n^*	[0,0.6,0,0,0]	[0,0.81,0,0,0]	[0,0.95,0.05,0,0]	[0,0,1,0,0] *	[0, − 0.65, 1.65, 0, 0]
Expected Value	(1.82, 1.34)	(2.4, 1.8)	(3.24, 2.3)	(7.03, 2.58)	[9.66, 2.77]

Risk level k is critical in helping management make portfolio selection decisions. This parameter, shown by the optimal value x^* 5.591 in Table 6, denotes the amount of risk associated with a specific portfolio. In this situation, it indicates that manufacturing output can create the largest return compared to other production sectors, but at a greater risk level.

The fundamental goal of this research is to find industrial production solutions with high potential for returns while controlling associated risks. While risk levels $k = 3, 4,$ and 5 often produce positive expected returns, $k = 5.591$ stands out as having the greatest potential return based on our findings in Table 6. As a result, this model can assist management in prioritizing industries with optimal risk-return profiles to maximize prospective profitability.

A total allocation of 1 is required while aiming for a particular risk level, k = 6, based on the portfolio outcome that has been provided, which is the allocation of resources represented by $x1$ to $x5$. To achieve a risk level of 1.6541, it is specifically necessary to decrease the allocation of $x2$ by roughly 0.651 units and raise the allocation of $x3$. Within the context of this structure, the modifications seek to reallocate resources across the portfolio to satisfy the target degree of risk while abiding by the limitations set forth by the optimization issue. By using this reallocation approach, resources are distributed fairly and in accordance with the designated risk tolerance level.

6 Conclusions

In conclusion, this study introduces a portfolio selection approach designed to effectively address the inherent uncertainties present in real-world data, particularly in industrial production planning. Employing a two-stage strategy, our approach demonstrates resilience and adaptability in navigating uncertainty. Firstly, through the innovative integration of fuzzy random variables, our methodology rigorously cleans and prepares data,

capturing both unpredictability and ambiguity inherent in real-world datasets. Second, using this pre-processed data in a portfolio selection model based on the standard mean-variance paradigm, we optimise portfolio strategies across multiple production sectors while accounting for uncertainty. Particularly, our approach enables more resilient and adaptive portfolio selection strategies, which are critical in the face of industrial production planning uncertainties. Specifically, the first stage utilizes a measurement error method to transform crisp data into fuzzy sets, enriching the dataset and capturing potential variances. Subsequently, key parameters such as fuzzy center and width are derived from these fuzzy sets. To address randomness, our method incorporates probability distributions alongside the preprocessed fuzzy data, providing a comprehensive approach to managing uncertainty. Application of this technique across five distinct industrial sectors—agriculture, mining, manufacturing, electricity, and water—yields promising results, showcasing its ability to identify optimal production yields for each sector. These findings offer valuable insights for strategic planning and decision-making processes across industries, underscoring the resilience and adaptability of our portfolio selection tactics in uncertain environments.

Acknowledgments. This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through Tier 1 (Vot Q507) and Ministry of Higher Education (MOHE) through Fundamental Research Grant Scheme (FRGS) (FRGS/1/2019/ICT02/UTHM/02/7).

References

1. Dorsey, A.H.: Active Alpha: A Portfolio Approach to Selecting and Managing Alternative Investments, vol. 356. Wiley, Hoboken (2011)
2. Wang, S., Xia, Y.: Portfolio Selection and Asset Pricing, vol. 514. Springer, Heidelberg (2012)
3. Tarasi, C.O., Bolton, R.N., Hutt, M.D., Walker, B.A.: Balancing risk and return in a customer portfolio. J. Mark. **75**(3), 1–17 (2011)
4. Markowitz, H.M.: Portfolio selection. J. Financ. **7**(60), 77–91 (1952)
5. Shapiro, A., Dentcheva, D., Ruszczynski, A.: Lectures on Stochastic Programming: Modeling and Theory. SIAM-Society for Industrial and Applied Mathematics (2009)
6. Vakarchuk, R.N., Mäntyniemi, P., Tatevossian, R.E.: On the effect of synthetic and real data properties on seismic intensity prediction equations. Pure Appl. Geophys. **176**, 4261–4275 (2019)
7. Re, C., Suciu, D.: Management of data with uncertainties. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management, pp. 3–8 (2007)
8. Krause, P., Clark, D.: Representing Uncertain Knowledge: An Artificial Intelligence Approach. Springer, Heidelberg (2012)
9. Hasuike, T., Katagiri, H., Ishii, H.: Portfolio selection problems with random fuzzy variable returns. Fuzzy Sets Syst. **160**(18), 2579–2596 (2009)
10. Rubinstein, M.: Markowitz's portfolio selection: a fifty-year retrospective. J. Financ. **57**(3), 1041–1045 (2002)
11. Zhang, Y., Li, X., Guo, S.: Portfolio selection problems with Markowitz's mean–variance framework: a review of literature. Fuzzy Optim. Decis. Making **17**, 125–158 (2018)
12. Amiri, A., Tavana, M., Arman, H.: An integrated fuzzy analytic network process and fuzzy regression method for bitcoin price prediction. Internet Things **25**, 101027 (2024)

13. Uusipaikka, E.: Confidence Intervals in Generalized Regression Models. CRC Press, Boca Raton (2008)
14. Li, B., Teo, K.L.: Portfolio optimization in real financial markets with both uncertainty and randomness. Appl. Math. Model. **100**, 125–137 (2021)
15. Qin, Z.: Mean-variance model for portfolio optimization problem in the simultaneous presence of random and uncertain returns. Eur. J. Oper. Res. **245**(2), 480–488 (2015)
16. Markowitz, H.M.: Foundations of portfolio theory. J. Financ. **46**(2), 469–477 (1991)