International Journal of
Innovative Management, Information & Production ISME International ⓒ 2014 ISSN 2185-5455
Volume **5**, Number **3**, December **2014** **PP. 65-71**

THE APPLICATION OF FUZZY CORRELATION COEFFICIENT WITH FUZZY INTERVAL DATA

YU-TING CHENG[1], CHIH-CHING YANG[2]
Department of Statistics National Chengchi University
Muzha, Taipei, Taiwan 116
ting@nccu.edu.tw

Department of Statistics National Chengchi University
Muzha, Taipei, Taiwan 116
96354502@nccu.edu.tw

ABSTRACT. *Among the important topics of statistics is to evaluate a proper correlation coefficient with fuzzy data, especially when the data illustrate uncertain, inconsistent and incomplete type. Generally, we use Pearson's Correlation Coefficient to measure the correlation between data with real values. However, when the data are composed of fuzzy interval values, it is not feasible to use such a classical approach to determine the correlation coefficient. This study proposes the computation of fuzzy correlation coefficient with fuzzy interval data. Empirical studies are employed to explain the application for evaluating fuzzy correlation. More related practical phenomena can be explained using the application of fuzzy correlation.*

Keywords: Fuzzy Correlation; Fuzzy Interval Data; Evaluation; Air Pollution; Transportation Engineering.

1. Introduction. In classical statistics, the two-valued logic will be reflected. Investigating the phenomena of nature, socials or economics, fuzzy logic should be applied to account for the full range of possible values. Since Zadeh (1965) developed fuzzy set theory, its applications have been extended to traditional statistical inferences and methods in social or engineering or economics, including medical diagnosis or stock investment systems. For example, a continuing series of studies displayed approximate reasoning methods for econometrics (Lowen, 1990; Ruspini,1991;Dubois & Parde ,1991) and a fuzzy time series model to overcome the bias of stock markets was developed (Wu & Hsu, 2002).

In traditional statistical theory, the observations should be observed under probability distribution. In practice, the observations are sometimes explained by linguistic terms such as "Very important," "Important," "Normal," "Unimportant," "Very unimportant”, or "Maximum value and Minimum value", are only approximately known, rather than equating with randomness. Measuring the correlation coefficient between two variables including fuzziness is a challenge to the classical statistical theory. A lot of studies which investagate the topic of the fuzzy correlation analysis and its application in the social or economic science fields (Bustince and Burillo, 1995; Yu, 1993; Liu and Kao, 2002; Hong, 2006). Such as, Hong and Hwang (1995) and Yu (1993) define a correlation formula to measure the interrelation of intuitionist fuzzy sets. However, the range of their defined correlation is from 0 to 1, which contradicts with the conventional awareness of correlation which should range from -1 to 1. In order to overcome this issue, Chiang and Lin (1999) take random sample from the fuzzy sets and treat the membership grades as the crisp observations. Their derived coefficient is between -1 and 1; however, the sense the fuzziness is gone. Liu and Kao (2002) calculated the fuzzy correlation coefficient based on Zadeh's extension principles. They used a mathematical programming approach to derive fuzzy measures based on the classical definition of the correlation coefficient. Their derivation is very probable; however, in order to use this scheme, the mathematical programming should be required.

In addition, formulas in these studies are quite complicated or required some mathematical programming which really limited the access of some researchers with no strong mathematical background. In this thesis, we propose a simple solution of a fuzzy correlation coefficient without programming. In addition, the provided solutions are based on the classical definition of Pearson correlation which should quite easy and straightforward. The definitions provided in this study can also be used for interval-valued fuzzy data.

The remainder of the paper proceeds as follows. The fuzzy interval correlation is introduced in section 2. Section 3 presents its results of the relationship of the simulation. Section 4 presents its empirical results. Finally, the conclusions are drawn in section 5.

2. Fuzzy Interval Correlation. In general, we need to study the relationship between the variables x and y, the most direct and simple way is to draw a scatter plot, which can approximately illustrate the relationship between these variables such as positive correlation, negative correlation, or non-correlation. Pearson's correlation coefficient is often considered to evaluate that presents a measure of how two random variables□ are linearly related in a sample. The population correlation coefficient, ρ, is defined for two variables x and y by the formula:

$$\rho = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

where (x_i, y_i) is the ith pair observation value, $i = 1,2,3,\ldots,n$. $\bar{x}\,,\bar{y}$ are sample mean for x and y respectively.

In this case, the more positive ρ is, the more positive the association is. This also indicates that when ρ is close to 1, an individual with a high value for one variable will likely have a high value for the other, and an individual with a lower value for one variable will likely to have a low value for the other. On the other hand, the more negative ρ is , the more negative the association is, this also indicate that an individual with a high value for one variable will likely have a low value for the other when ρ is close to -1 and conversely. When ρ is close to 0, this means there is little linear association between two variables. In order to obtain the correlation coefficient, we need to obtain σx2, σy2 and the covariance of x and y. In practice, these parameters for the population are unknown or difficult to obtain. Thus, we usually use rxy, which can be obtained from a sample, to estimate the unknown population parameter. The sample correlation coefficient rxy is expressed as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

Pearson correlation coefficient is a straightforward approach to calculate the relationship between two variables. However, if the variables which considered are not real numbers, but fuzzy data, the formula above is problematic. For example, Mr. Smith who is a new graduate from college expected salary ranges from [45000, 50000] and his expected working hours are [8, 10]. If we collect this kind of data from many new graduates, then the correlation between expected salary and working hours cannot be calculated by us from this data. Suppose IX is the expected salary for each new graduate, IY is the working hours they desired, then the scatter plot for these two sets of fuzzy interval numbers would approximate that shown in Figure 1.

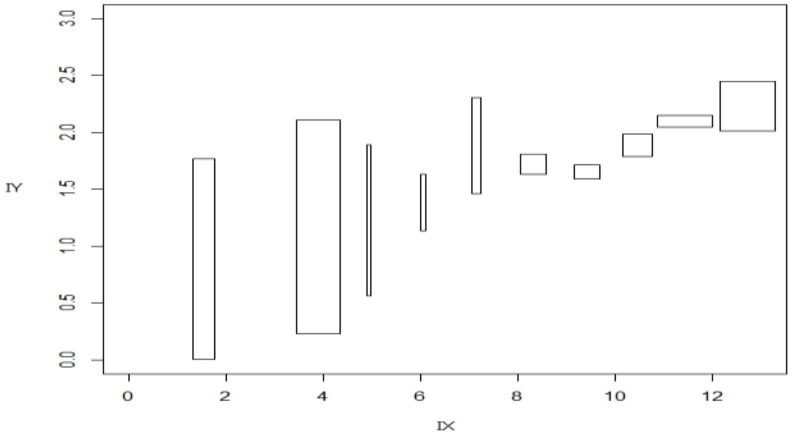

FIGURE 1. Fuzzy correlation with interval data

For the interval valued fuzzy number, we consider to pick out samples from population X and Y. Each fuzzy interval data for the centroids and length of the sample X and sample Y will be considered to calculate the correlation coefficient. In addition, we also employ the maximum value and minimum value of fuzzy interval data to evaluate the correlation coefficient.

In this paper, there are two kinds of fuzzy correlation which are based on the Person's correlation as well as the extension principle Definition 1 and Definition 2, the advantages are that we can compute various samples with fuzzy interval type for the continuous sample.

Definition 1.

Let $(X_i = [a_i, b_i, c_i, d_i], Y_i = [e_i, f_i, g_i, h_i]; i = 1,2,\cdots,n)$ be a sequence of paired trapezoid fuzzy sample on population Ω with its pair of centroid $(cx_i\ , cy_i)$ and pair of area $\|x_i\| = \text{area}(x_i), \|y_i\| = \text{area}(y_i)$.

$$cr_{xy} = \frac{\sum_{i=1}^{n}(cx_i - \bar{cx})(cy_i - \bar{cy})}{\sqrt{\sum_{i=1}^{n}(cx_i - \bar{cx})^2}\sqrt{\sum_{i=1}^{n}(cy_i - \bar{cy})^2}}$$

$$\lambda ar_{xy} = 1 - \frac{\ln(1 + |ar_{xy}|)}{|ar_{xy}|}$$

$$\mathrm{ar}_{xy} = \frac{\sum_{i=1}^{n}(||x_i|| - ||\bar{x}||)(||y_i|| - ||\bar{y}||)}{\sqrt{\sum_{i=1}^{n}(||x_i|| - ||\bar{x}||)^2}\sqrt{\sum_{i=1}^{n}(||y_i|| - ||\bar{y}||)^2}}, \qquad (2)$$

Then fuzzy correlation is defined as:

1. When $\mathrm{cr}_{xy} \geq 0,\ \lambda \mathrm{ar}_{xy} \geq 0$, fuzzy correlation = $(\ \mathrm{cr}_{xy}\ ,\ \min(1, \mathrm{cr}_{xy} + \lambda \mathrm{ar}_{xy}))$
2. When $\mathrm{cr}_{xy} \geq 0,\ \lambda \mathrm{ar}_{xy} < 0$, fuzzy correlation = $(\ \mathrm{cr}_{xy} - \lambda \mathrm{ar}_{xy}\ , \mathrm{cr}_{xy})$
3. When $\mathrm{cr}_{xy} < 0,\ \lambda \mathrm{ar}_{xy} \geq 0$, fuzzy correlation = $(\ \mathrm{cr}_{xy}\ , \mathrm{cr}_{xy} + \lambda \mathrm{ar}_{xy})$
4. When $\mathrm{cr}_{xy} < 0,\ \lambda \mathrm{ar}_{xy} < 0$, fuzzy correlation = $(\ \max\left(-1, \mathrm{cr}_{xy} - \lambda \mathrm{ar}_{xy}\right), \mathrm{cr}_{xy})$

Definition 2.

Let $X_{ji}[a_{1i}\ , a_{2i}]$ and $Y_{ji}[b_{1i}\ , b_{2i}]$ be a sequence of paired fuzzy sample on population Ω. Let

$$r_{jk} = \frac{\sum_{i=1}^{n}\left(a_{ji} - \bar{a_j}\right)\left(b_{ki} - \overline{b_k}\right)}{\sqrt{\sum_{i=1}^{n}\left(a_{ji} - \bar{a_j}\right)^2}\sqrt{\sum_{i=1}^{n}\left(b_{ki} - \overline{b_k}\right)^2}}, j = 1,2, k = 1,2.$$

Then fuzzy correlation is $\left[r_{low}\ , r_{up}\right]$ wit $r_{low} = \bar{r} - s_r h$ and $r_{up} = \bar{r} + s_r$,where

$$\bar{r} = \frac{\sum_{j=1}^{2}\sum_{k=1}^{2} r_{jk}}{4} \text{ and } s_r = \frac{\sum_{j=1}^{2}\sum_{k=1}^{2}\left(r_{jk} - \bar{r}\right)^2}{4}$$

A correlation coefficient is a number between -1 and 1 which measures the degree to which two variables are linearly related. If there is perfect linear relationship with positive slope between the two variables, we have a correlation coefficient of 1; if there is positive correlation, whenever one variable has a high value. Thus, base on the measure of evaluation, the degree of the population correlation coefficient, we will be considered for the correlation of fuzzy interval. As the correlation of fuzzy interval, $\left[r_{low}\ , r_{up}\right]$, is computed then the value of fuzzy correlation can be evaluated that is defined as,

1. When $\left[r_{low}\ , r_{up}\right] \in [\ -0.10, 0.10\ ]$n, the fuzzy correlation is not significant.
2. When $\left[r_{low}\ , r_{up}\right] \in [\ -0.39, -0.11\ ]$ or $[0.11, 0.39\ ]$, the fuzzy correlation is low value.
3. When $\left[r_{low}\ , r_{up}\right] \in [\ -0.69, -0.40\ ]$ or $[0.40, 0.69\ ]$, the fuzzy correlation is middle value.
4. When $\left[r_{low}\ , r_{up}\right] \in [\ -0.99, -0.70\ ]$ or $[0.70, 0.99\ ]$, the fuzzy correlation is high value.

2.1. Simulation studies. In this section, we will employ the Mote Carlo simulation to generate several sequence of fuzzy interval data set and then compare their correlations coefficient with different definition as proposed at the section 2. The distribution for the centroid and area are generated by the normal, uniform, gamma and Cauchy distribution respectively. The procedure to compute correlation coefficient is described below: Table 1 illustrates the result.

1. Step 1. Generate fuzzy set of sequence X with successive 4 points and error term from the underlying distribution.
2. Step 2. Let $Y = aX + e$, calculate the fuzzy data set Y by the fuzzy data set X and error term..

3. Step 3. Find the correlation coefficient from the fuzzy data set by above definitions.

Table 1. The fuzzy interval correlation coefficient for various center and area model with definition 1 and definition 2.

a	Area center	Normal(0,1)	Uniform(0,1)	Gamma(2,2)	Cauchy(0,1)
0.2	Normal	(0.16, 0.20)^1 (0.13, 0.18)^2	(0.05, 0.09)^1 (0.04, 0.08)^2	(0.21, 0.25)^1 (0.17, 0.22)^2	(0.94, 0.98)^1 (0.93, 0.94)^2
	Uniform	(0.18, 0.21)^1 (0.19, 0.22)^2	(0.08, 0.12)^1 (0.07, 0.09)^2	(0.21, 0.25)^1 (0.23, 0.25)^2	(0.94, 0.98)^1 (0.94, 0.95)^2
	Gamma	(0.15, 0.19)^1 (0.11, 0.18)^2	(0.01, 0.04)^1 (0.03, 0.07)^2	(0.19, 0.23)^1 (0.15, 0.21)^2	(0.94, 0.98)^1 (0.93, 0.94)^2
	Cauchy	(-0.02,0.00)^1 (-0.00,0.11)^2	(-0.02,0.00)^1 (0.00, 0.06)^2	(0.00, 0.03)^1 (0.00, 0.14)^2	(0.33, 0.36)^1 (0.30, 0.63)^2
0.5	Normal	(0.33, 0.38)^1 (0.25, 0.39)^2	(0.11, 0.15)^1 (0.08, 0.18)^2	(0.41, 0.45)^1 (0.31, 0.45)^2	(0.95, 0.98)^1 (0.97, 0.98)^2
	Uniform	(0.42, 0.46)^1 (0.40, 0.46)^2	(0.19, 0.23)^1 (0.15, 0.22)^2	(0.49, 0.52)^1 (0.47, 0.53)^2	(0.95, 0.99)^1 (0.98, 0.99)^2
	Gamma	(0.26, 0.30)^1 (0.21, 0.37)^2	(0.06, 0.10)^1 (0.06, 0.17)^2	(0.36, 0.40)^1 (0.26, 0.43)^2	(0.98, 1.00)^1 (0.97, 0.98)^2
	Cauchy	(-0.02,0.00)^2 (0.00, 0.26)^3	(-0.02,0.00)^2 (0.00, 0.14)^3	(-0.02,0.00)^1 (0.00, 0.28)^2	(0.34, 0.37)^1 (0.30, 0.66)^2
0.8	Normal	(0.38, 0.42)^1 (0.30, 0.50)^2	(0.15, 0.19)^1 (0.10, 0.26)^2	(0.50, 0.53)^1 (0.36, 0.56)^2	(0.95, 0.99)^1 (0.98, 0.99)^2
	Uniform	(0.60, 0.64)^1 (0.52, 0.62)^2	(0.29, 0.33)^1 (0.22, 0.32)^2	(0.63, 0.67)^1 (0.59, 0.68)^2	(0.99, 1.00)^1 (0.99, 1.00)^2
	Gamma	(0.37, 0.41)^1 (0.25, 0.48)^2	(0.08, 0.12)^1 (0.07, 0.24)^2	(0.41, 0.44)^1 (0.31, 0.54)^2	(0.95, 0.99)^1 (0.98, 0.99)^2
	Cauchy	(0.00, 0.03)^1 (0.01, 0.35)^2	(-0.03,0.00)^1 (-0.00,0.21)^2	(-0.02,0.00)^1 (0.01, 0.37)^2	(0.34, 0.37)^1 (0.30, 0.66)^2

Note: ^{1} denotes the result by the definition 1; ^{2} denotes the result by the definition 2.

In Table 1, there are some results will be described as follows: (1) when $a = 0.2$, the interval of the correlation coefficient is very close. (2) when $a = 0.5$, the interval of correlation coefficient are close except the distribution of Cauchy. (3) when $a = 0.8$, the estimated interval form definition 3 is bigger than the definition 4 did if the center distributions come from Gamma, Normal, Uniform. While if the distribution comes from Cauchy distribution, we will get a very odd estimation.

2.2. Empirical studies. In general, the transportation engineering will affect the quality of air or climate. Hence, the passenger counts of Taipei MRT system could be considered to investigate the correlation between passenger counts of Taipei MRT system and air pollution, where air pollution include total suspended particles (TSP), air suspended particles (ASP), sulfur dioxide (SO2), ozone(O3) , fallout.

We examined the passenger counts of the Taipei MRT System and air pollution in Taiwan with 170 week samples between January, 1998 and February, 2012. And the form of collected data, maximum and minimum observation, will be showed by fuzzy interval data. The results show the correlation for the passenger counts and air pollution with two approaches of evaluation of correlation coefficient. The results are listed in Table 2.

TABLE 2. Correlations interval based on passenger counts and the air pollution in Taiwan

Fuzzy correlation	TSP	ASP	SO2	O3	Fallout
By definition 1	(-.178, -.142)	(.325, .420)	(.356, .379)	(.370, .425)	(-.181, -.153)
By definition 2	(-.187, -.073)	(.273, .335)	(.166, .552)	(.285, .437)	(-.163, -.150)

In the Table 2, we have the following findings. First, besides the correlation of passenger counts and the TSP and fallout are low significance negative by schemes of definition 1 and definition 2, and this result denotes that the passenger counts of Taipei MRT system increase then that can reduce the value of TSP and fallout. Second, the correlation coefficient is middle level for passenger counts and the ASP, SO2 and O3 by the approach of definition 1, this means the values of ASP, SO2 and O 3 have a lot of effect to the passenger counts. Third, the correlation coefficient is low significance for passenger counts and the ASP, SO2 and O3 by the approach of definition 2, this means the values of ASP, SO2 and O3 have a little effect to the passenger counts, this result show that the passenger counts will affect the air pollution, such as the air pollution of ASP, SO2 and O3 can be affected by the passenger counts of Taipei MRT system.

3. Conclusions. In the progress of the scientific research and analysis, the uncertainty in the statistical numerical data is the important point of the problem where the traditional mathematical computation is hard to be established. If we achieve this artificial accuracy to do causal analysis or measurement, it may lead to the deviation of the causal judgment, the misleading of the decision strategy, or the exaggerated difference between the predicted result and the actual data. As the pattern of data of interval is occurred in transportation engineering or energy environment. Our proposed methods can be applied to make management strategy or decision as the two variables illustrate kind of fuzzy interval data. In other words, this paper employ a simple approach to derive from fuzzy interval measures based on the traditional definition of Pearson correlation coefficient which are easy and straightforward. In the formula we provided, when all observations are real numbers, the developed model becomes the classical Pearson correlation formula.

In practice, many applications are fuzzy in nature. We can absolutely ignore the fuzziness and make the existing methodology for crisp values. However, this will make the researcher over confident with their results. With the methodology developed in this paper, a more realistic correlation is obtained, which provides the decision maker with more knowledge and confident to make better strategies.

REFERENCES

[1] B. Wu, and Y. Hsu (2002), The Use of Kernel Set and Sample Memberships in the Identification of Nonlinear Time Series, *Soft Computing Journal*, vol. 8, no.3, pp.207-216.

[2] C. Yu (1993), Correlation of fuzzy numbers, *Fuzzy Sets and systems*, vol. 55, pp.303-307.

[3] D. A. Chiang and N. P. Lin (1999), Correlation of fuzzy sets, *Fuzzy Sets and Systems*, vol. 102 pp.221-226.

[4] D. Dubois and H. Prade (1991), Fuzzy Sets in Approximate Reasoning, Part 1: Inference with Possibility Distributions, *Fuzzy Sets and Systems*, vol. 40, pp.143-202.

[5] D. Hong (2006), Fuzzy measures for a correlation coefficient of fuzzy numbers under Tw (the weakest t norm)-based fuzzy arithmetic operations, *Fuzzy Sets and Systems*, vol. 176, pp.150-160.

[6] D. Hong and S. Hwang (1995), Correlation of intuitionistic fuzzy sets in probability space, *Fuzzy Sets and Systems*, vol . 75, pp.77-81.

[7] E. Ruspini (1991), Approximate Reasoning: past, present, future, *Information Sciences*, vol. 57, pp.297-317.

[8] H. Bustince and P. Burillo (1995), Correlation of interval-valued intuitionistic fuzzy sets, *Fuzzy Sets and Systems*, vol. 74, pp.237-244.

[9] L. A. Zedah (1965), Fuzzy sets, *Information and Control*, vol. 8, pp.338-353.

[10] R. Lowen (1990), A fuzzy language interpolation theorem, *Fuzzy Sets and Systems*, vol. 34, pp.33-38.

[11] S. Liu and C. Kao (2002), Fuzzy Measures for correlation coefficient of fuzzy numbers, *Fuzzy Sets and Systems*, vol. 128, pp.267-275.