International Journal of Intelligent Technologies and Applied Statistics
Vol.4, No.2 (2011) pp.265–283

Model Construction and Residues Analysis with Fuzzy Data

Wen-Xing Li[1*] and Chun-Yi Yu[2]

[1]*Institute for Quantitative Economics, Huaqiao University, Fujian, China*
[2]*Department of Mathematical Sciences, National Chengchi University, Taipei, Taiwan*

ABSTRACT

The application of data classifications in time series analysis and forecasting is rather important. The fuzzy data classification has received much attention recently. It can be applied on various fields such as finance, sociology, biomedicine, electrical engineering and so on. This study is to use the fuzzy data classification to perform an intensive research on the change periods detection and model construction of the interval time series. We use average of the sum of fuzzy entropies to find out interval of the structural changes. Focusing on the time series of intervals, we build a model and make prediction about it. At the end, based on the case study on the population of singles versus, we thoroughly discuss this topic. The result shows that the unemployment rate does significantly correlate with the population of singles, but the "widow's year" does not.

Keywords: Fuzzy data classification; Average of the sum of fuzzy entropies; Change periods; Unemployment rate; Population of singles

1. Introduction

As the trend of getting married late or not getting married emerges, Taiwan faces the problem with low birth rate and aging of population. The issues of the welfare for the aged and age structure of population receive increasing attention too. The heightening unemployment rate, on the other hand, has caused more people to marry late or not marry at all, and this increases the number of people who remain single. Therefore, a study like this that investigates the future trend of singlehood rate becomes more and more important. The fuzzy data analysis and framework for study are recent hot topics. Conventionally, studies recorded in related literature emphasize on performing anti-fuzziness on data, and then classify them. This process is called "fuzzy classification." So-called anti-fuzziness of data, is to describe the

* Corresponding author: liwenxing1981@163.com

nature of fuzziness of subject matters, or to find out the fuzzy relationship between every subject matter. "Classification," on the other hand, is the grouping of data that share the same nature of fuzziness or relationship. However, literature discussing the structural change of data was rare. That was why Wu and Chen [10] combined these two kinds of knowledge to construct a series of procedures that is effective in finding out the structural change within data using the fuzzy classification.

Kim et al. [5] , Pakhira et al. [8], Wang and Zhang [9] and Wu and Yang [11] further constructed different methods of calculation, to investigate the effective way of using the Fuzzy C-Means to determine the cluster in a large number of sets. However, no further study was done on the interval time series. Zeng and Li [14] researched on the fuzzy entropy between intervals, but did not define clearly the interval distance. On the other hand, it is mentioned in the second section of this thesis, that the measurement of interval distance, the fuzzy entropy of intervals, and the fuzzy membership degree of intervals can be used to set up a standard for measurement, to perform fuzzy classification on an interval time series.

A clear definition for the term "change point" is non-existent in semantics. Many phenomena cannot be established with the binary logic of "true" or "false." The fuzzy logic will be useful in this regard. This is the reason why the use of fuzzy theory to identify and classify time sequences is considered. In general, the change points in a time series should be represented as the difference in degree, e.g., a 95% change point. In fact, if the concept of "change point" is based on the fuzzy theory, it would appear the term "change interval" fits into the realistic situation better than "change point." In a linear regression model, Chow [3] examined the significance of change point of a single point of time whose structural change is known.

Liu et al. [7] proposed using the Schwarz Information Criteria to estimate the least square method and the number of change points in segmented multivariate regression. Bai and Perron [2] proposed the use of Wald Test to perform estimation on linear models with multiple structural changes. Andrews and Ploberger [1] applied a number of techniques, namely the Wald, Lagrange Multiplier and Likelihood Ratio tests, to extensively analyze the problems in structural change analysis. Kumar and Wu [6] discovered that the concept of fuzzy logic can be used to effectively determine structural changes in non-linear time sequence. Zhou [15] introduced a new observation in structural change -- the Integrating Bayesian Structural Break Model and the Change Point Detection Methods. Although there have been numerous researchers performing studies on structural change, and countless of methods used, the mathematical inference process is complex, and there still has not been a clear standard in defining change point. Therefore another purpose of this thesis is to find out the structural change in the unemployment rate and effectively analyze it, and then transform its structural change part with the singlehood rate, to discover a better forecasting skill.

As the number of universities in Taiwan increases, it becomes increasingly common for males before 25 to be still pursuing their study. The employment rate for male below 25 is also inconsequential. For these reasons, this thesis sets males between the ages of 25-34 as the object of study, to investigate the relationship between singlehood rate and unemployment rate, and its relationship with the Widow's Year. This will lead to the study of the potential change in the singlehood population in the future and the prediction of its trend, as a reference for the social census research on the marriage rate of the population that is currently single.

Out of Taiwan's population of 23 million, 6.7 million people are single. The Taiwanese males make up 54% of the single population (*Statistical Yearbook of the Republic of China 2009* of the Directorate General of Budget, Accounting and Statistics [DGBAS]) [4]. According to the DGBAS, the Executive Yuan of Taiwan, in the year 2009, the number of Taiwanese aged above 15 who were single exceeded 6 million people -- out of which, 54% were males. The statistical report of DGBAS [4] also reveals that, the single males of 25-34 years old form 13% of the total number of unmarried males above 15 years old. In a family, the male is the breadwinner. In the current situation in Taiwan, of high unemployment rate and high house price, the pressure coming from shouldering the household expenditure is heightening, and it has become increasingly difficult to get married. This is why we see a high singlehood rate in male.

Traditionally, marrying in the Widow's Years is a taboo. Because of the folklore that couples marrying in the Widow's Years will encounter obstacles in their marriage, elders will advise against marrying in the Widow's Years. This will give rise to the rate of unmarried population in the mentioned years. In turn, birth rate will drop, and the rate of population aging will increase. In the future, Taiwan will face the issues of declining birth rate, aging population structure, welfare of the elderly, decreasing overall productivity, etc. Hence, understanding the future development of the singlehood rate of the male population is a big topic.

2. Model construction with fuzzy data

In this study, the fuzzy set concept is used to perform classification and investigation on related time series, and investigate the detection of change interval and fuzzy model construction.

2.1 The role of fuzzy entropy

There are many analysis techniques for time series. The oft-used ones include exponential smoothing, autoregressive integrated moving average (ARIMA), log linear trend, linear trend with seasonal terms, etc. Usually the nature of the time series and the suitability of the model are considered, in order to decide the most appropriate technique to evaluate and forecast the time series. The steps to construct a time series model are: (1) identify; (2) evaluate and diagnose; (3) forecast. In the beginning, determine whether or not the unemployment rate and singlehood rate are relevant. Use regression analysis to clearly point out that both the unemployment rate to the singlehood rate, and the singlehood rate to the unemployment rate are inversely proportional. After that, calculate the ACF and PACF for both these rates, and the CCF between them, in order to obtain the conversion model.

Transfer function model is the extension of the construction method of a univariate time series model to the analysis method of a multivariate time series. Because the analysis method of a single-factor model is affected by time, its ability to predict the future might not be high. In many cases, it is possible to have a set of data whose current observed value is affected by the past observed value, and one or more other set of time series are correlated to the mentioned set of data. This implies that there will be impact transferred to the output series, when there are any changes in the input series. Therefore, due to the possibility of the singlehood rate being affected by unemployment rate of previous periods and the past singlehood rate itself, it is more accurate to use the transfer function mode to perform the conversion of the singlehood rate.

There are a lot of uncertainties in the human's thoughts. If the classical binary logic is used to forcefully classify the thoughts, there will be errors. For example, under the judgment of the binary logic, the interviewee will only be given the option to answer "yes" or "no," but cannot include the uncertainties of the weather in her answer. So the concept of fuzzy set is first brought up by Zadeh [13], as a more complete solution for the multivariate fuzzy phenomena.

The fuzzy theory makes use of the membership function to represent the fuzzy nature of matters, in place of the concept and calculation method of the classical binary logic. Meaning that, in the classical set theory, an element either belongs to a particular set, or it does not. Whereas in the fuzzy theory, the membership degree is used to describe an element -- t he grade of membership of the element to a particular set. For example, the word "happiness" has definitions that differ from a person to another, and hence it has the property of uncertainty. But with the fuzzy theory, it can be shown that Mr. A is 80% happy at this moment. In the fuzzy theory, the full range of membership degree is usually set as 0 to 1. This will enable "happiness" to be quantifiable, and become a distribution of emotion.

When the fuzzy theory is used to examine whether there is any change point in a time series, first we cluster the time series, find out the cluster center, and then use the fuzzy membership degree, fuzzy entropy and other relevant concepts to perform classification. The definition of fuzzy membership degree is as follows:

Definition 2.1 *Fuzzy Membership Degree*

Let a time series be $\{x_t,\ t = 1, 2, \ldots, N\}$, with C_1 and C_2 being two cluster centers of the time series, μ_{it}, $i = 1, 2$. to represent the membership degree of an element x_t in the time series X_t to C_1 and C_2, the membership degree is thus defined as: $\mu_{it} = 1 - \frac{|x_t - C_i|}{\sum_{i=1}^{2} |x_t - C_i|}$

Definition 2.2 *Fuzzy Entropy*

Let a time series be $\{x_t$, t = 1, 2,…,N$\}$, with μ_{it} being the membership degree of x_t to the cluster centers Ci (i = 1, 2,…, k), the fuzzy entropy is thus defined as:

$$\delta(x_t) = -\left(\frac{1}{\mathrm{k}}\right)\sum_{i=1}^{\mathrm{k}}\left[\mu_{it}\ln\left(\mu_{it}\right) + \left(1-\mu_{it}\right)\ln\left(1-\mu_{it}\right)\right]$$

Entropy is a concept in the Thermodynamics study, whose original meaning is the degree at which work can be transformed. The Statistical Physics provides another definition to it: the measure to describe the random motion. In addition, the Probability Theory and Information Theory give it a more common definition: measure of the unboundedness of a random variable, or the measure of the amount of missing information. So fuzzy entropy is used to measure the uncertainty of fuzzy sets, and is an important tool for the processing of fuzzy data, while the membership degree is used to characterize elements that do not clearly belong to some particular sets.

2.2 Distance with interval data

An interval-valued fuzzy set can be viewed as a continuous fuzzy set, which further represents uncertain matters. Take "test results" as an example. In foreign countries, A, B, C and D are used to evaluate a student's result, whereby A represents 100-80 marks, B represents 79-70 marks, C represents 69-60 marks, D represents 59-50 marks, in place of numerical scores. In the past we think that obtaining a high score means learning well. However, does a student who gets 85 marks have better learning ability than another who scores 80 marks? Not necessarily so. That's why the interval-valued fuzzy set resolves the phenomena of uncertainty.

When a sample of interval-valued fuzziness is available, we have to consider the calculation for intervals. Refer to Wu and Yang [11] for interval calculations. However, there is still no complete definition for the measure of interval distance (see [12]). This section will define a well-defined interval distance, and use this definition to calculate the cluster center of interval and the fuzzy membership degree of interval.

How to define a well-defined interval distance? First we represent the interval with (c_i; r_i) with c being the center, r being radius. This way, the interval distance can be considered as the difference of the center plus the difference of the radius. The difference of the center can be seen as the difference in location, and the difference of the radius can be seen as the difference in scale. However, in order to lower the impact of the scale difference on the location difference, we take the *ln* value of the scale difference, and then plus 1 to avoid the *ln* value becoming negative.

Definition 2.3 *Distance between Samples of Interval-Valued Fuzziness*

Assuming U is the universe of discourse. Let $\{s_i = [a_i, b_i], i = 1, 2\}$ be two samples of interval-valued fuzziness from the mentioned universe of discourse U, and $c_i = \frac{a_i + b_i}{2}$, $r_i = \frac{a_i - b_i}{2}$. The distance between the two samples of interval-valued fuzziness s_1 and s_2 is defined as:

$$d(s_1, s_2) = |c_1 - c_2| + \ln(1 + |r_1 - r_2|)$$

Definition 2.4 *Interval Means Square Error (IMSE) of Prediction for a Sample of Interval-Valued Fuzziness*

Let $\{s_i = [a_i, b_i], i = 1, \ldots, N\}$ be an interval time series, with prediction interval being $\hat{s}_i = [\hat{a}_i, \hat{b}_i]$ and $\varepsilon_i = d(s_i, \hat{s}_i)$ beging the error between the prediction interval and the actual interval, thus:

$$IMSE = \frac{1}{l}\sum_{i=N+1}^{N+l} {\varepsilon_i}^2$$

where l is the forecasted expectancy value.

Example 2.1: Let $S = \{[3, 6], [4, 5], [2, 6], [5, 8], [3, 8]\}$ be a set of fuzzy sample from a surveyof expected Salary for the graduated students. $\hat{S} = \{[3, 4], [2, 6], [3, 5], [5, 7], [4, 5]\}$. Then the distance from expected salary and actual salary are computed as Table 1, as follows:

Table 1. Distance for the interval data.

Sample	Estimated Salary	Actual Salary	Distance
1	[3, 6]	[3, 4]	$d(s_1, \hat{s}_1) = \| 3.5 - 4.5 \| + \ln(1 + \| 0.5 - 1.5 \|) = 1.69$
2	[4, 5]	[2, 6]	$d(s_2, \hat{s}_2) = \| 4 - 4.5 \| + \ln(1 + \| 1 - 0.5 \|) = 1.41$
3	[2, 6]	[3, 5]	$d(s_3, \hat{s}_3) = \| 4 - 4 \| + \ln(1 + \| 1 - 2 \|) = 0.69$
4	[5, 8]	[5, 7]	$d(s_4, \hat{s}_4) = \| 6 - 6.5 \| + \ln(1 + \| 1 - 1 \|) = 0.91$
5	[3, 8]	[4, 5]	$d(s_5, \hat{s}_5) = \| 4.5 - 5.5 \| + \ln(1 + \| 0.5 - 2 \|) = 0.09$
$IMSE = \frac{1}{5} \times (1.69^2 + 1.41^2 + 0.69^2 + 0.91^2 + 0.09^2) = 2.12$			

Definition 2.5 *Cluster of Interval-Valued Time Series*

Let $\Psi = \{s_t,\ t = 1, 2,\ldots,N\}$be an interval time series, $k \in N$ be the number of clusters. If there exists a set $J = \{I_i \in \text{interval};\ i = 1, 2,\ldots,k\}$, where by the distance square sum of an element s_t from Ψ and an element I_i from J is the least, then:

$$\text{Min} \sum_{t=1}^{N} \sum_{i=1}^{k} d(e_t, I_i)^2$$

Set $J = \{I_i \in \text{interval};\ i = 1, 2,\ldots,k\}$, is then known as the set of cluster interval for interval time series Ψ.

Example 2.2: The distribution of 27 sets of interval data of unemployment rate (unit: percent) is shown as Figure 1, as follows:

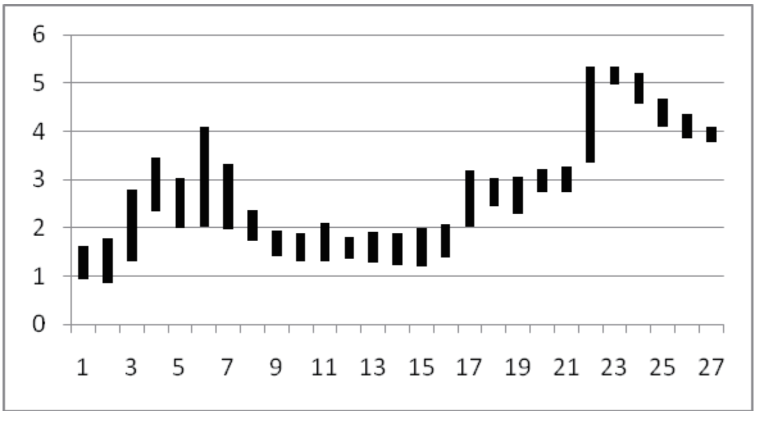

Figure 1. Chart of interval time series.

If we wish to divide the data into two groups, using Definition 2.5, we can obtain two interval clusters $I_1 = (1.83, 2.46)$ and $I_2 = (3.71, 5.23)$ The result of clustering is as Figure 2, as follows:

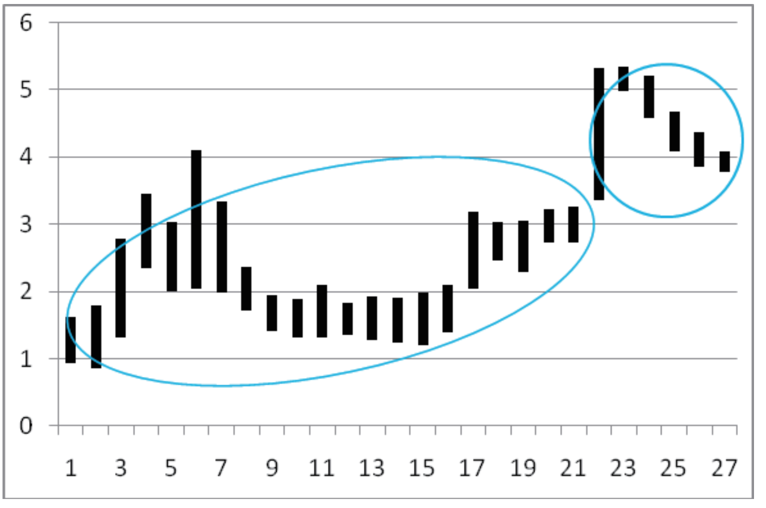

Figure 2. Result of interval clustering.

Definition 2.6 *Fuzzy Membership Degree of Interval*

Let $\psi = \{s_t, t = 1, 2, \ldots, N\}$ be an interval time series, and I_i be the cluster interval; let μ_{it} represent the membership degree of an element x_t from the interval time series Ψ to I_i, where $i = 1, \ldots, k$. The membership degree is thus defined as:

$$\mu_{it} = 1 - \frac{d(x_t, I_i)}{\sum_{i=1}^{k} d(x_t, I_i)}$$

2.3 Change periods

Conventionally, the detection of the structural change in a system takes the change point as the main factor of consideration. However, structural change should be mainly based on variable and not the time value, and it should be a gradually-emerging change interval and not a change point whereby the change happens abruptly at a certain point of time. So the change interval that studies variables, compared with the classical method of investigation for time series, has better descriptive power. Wu and Chen [10] suggested that, the use of fuzzy entropy is effective in identifying whether a structural change happens in a time series. Besides, it can also be used together with the mean cumulated fuzzy entropy of t times, to observe the change in message of fuzzy entropy, based upon which a standard for the classification of change model can be established.

Definition 2.7 *Mean Cumulated Fuzzy Entropy*

Let a time series be $\{x_t\}$, t = 1, 2,…,N , with $\delta(x_t)$ being its fuzzy entropy. The mean cumulated fuzzy entropy is thus defined as:

$$\mathrm{MS}\delta(x_t) = \frac{1}{t}\sum_{i=1}^{t} \delta(x_i)$$

There is usually a threshold level λ set up for fuzzy classification, because no matter it is for nature or humanities, the determination for classification is very subjective and often non-unanimous. Hence an objective measure is needed. According to empirical experience, λ cannot take a value too huge or too small, otherwise the classification cannot be done or too many classes will be created. So a value for λ between 0.001 and 0.1 will be ideal.

Here we wish to find two cluster centers. This is determined based on common experience of empirical analysis and trend of time series. The procedures are as follows:

Step 1: Use the k-means method to find out two cluster centers C_1 and C_2 in time series $\{x_t\}$, and determine the membership degree μ_{it}, $i = 1, 2$ of $\{x_t\}$ to the two cluster centers.

Step 2: Compute the fuzzy entropy $\delta(x_t)$, mean cumulated fuzzy entropy $\mathrm{MS}\delta(x_t) = \frac{1}{t}\sum_{i=1}^{t}\delta(x_i)$, and Median $(\mathrm{MS}\delta(x_t))$ of this series, that correspond to x_t.

Step 3: Take a suitable threshold value λ, classify the mean cumulated fuzzy entropy $\mathrm{MS}\delta(x_t)$ series that correspond with x_t. If the mean cumulated fuzzy entropy $\mathrm{MS}\delta(x_t)$ falls into the interval $[0, \text{Median}(\mathrm{MS}\delta(x_t)) - \lambda)$, we will use 1 to represent Group 1; if $\mathrm{MS}\delta(x_t)$ falls into the interval $[\text{Median}(\mathrm{MS}\delta(x_t)) - \lambda, \text{Median}(\mathrm{MS}\delta(x_t)) + \lambda)$, we use 2 to represent Group 2;and if $\mathrm{MS}\delta(x_t)$ falls into the interval $[\text{Median}(\mathrm{MS}\delta(x_t)) + \lambda, 1]$, 3 will be used to represent Group 3.

Step 4: If the result of classification is inconsistent, we then make adjustment to the result. If it is consistent, go to Step 5.

Step 5: Select an appropriate determination level α. If the number of consecutive samples is greater than $[\alpha N]$, then these consecutive samples belong to the same group. During classification, if more than one group is found, we know that structural change happens in this time series. Thereafter, find the change interval.

3. Empirical studies

In this section, we wish to use the populations of singles and unemployment rate to perform the fuzzy statistical analysis, to find out the change interval of the structural change for unemployment rate, as well as its impact to the singlehood rate and the forecast.

3.1 Source of information

The main source of information is the statistical data from the Directorate General of Budget, Accounting and Statistics (DGBAS) of the Ministry of the Interior. The singlehood rate is calculated by dividing the population of single males in years 1980 to 2006, of ages 25-34 years, with the total population of male of the same age group. In addition, the unemployment rate was also obtained from the DGBAS of the Ministry of the Interior. There were eight Widow's Years within 1980 to 2006, namely 1982, 1985, 1987, 1990, 1993, 1995, 1998, 2001, 2004 and 2006 (see Table 2).

3.2 Transfer function modelling

In this section, we use the transfer function model of time series to construct a model, by first determining CCF of the rate of single men of ages 25-24 year old in years 1980 to 2006, and the unemployment rate of the same years. For such a comparison, the relationship we can obtain is as follows:

$$(1 - B)Y_t = 0.00634 + (-0.148 - 0.083B + 0.492B^2)X_t, \ 1980 \le t \le 2006$$

in which, Y_t is the marriage rate at a point of time t, X_t is the employment rate at the point of time t, and the model of input series being:

$$(1 - 0.325B)(1 - B)X_t = \varepsilon_t, \ 1980 \le t \le 2006$$

And then we investigate the multi-intervention of the Widow's Years. Let I_{t_1} be

$$\begin{cases} 1, t_1 = 1982、1985、1987、1990、1993、1995、1998、2001、2004、2006 \\ 0, \text{otherwise} \end{cases}$$

therefore

$$(1 - B)Y_t = 0.00634 + (-0.148 - 0.083B + 0.492B^2)X_t + CI_{t_1}, \ 1980 \le t \le 2006$$

Table 2. Fuzzy interval of unemployment rate.

Year	Singlehood rate of males of ages 25-34 years.	Unemployment rate	Fuzzy interval of unemployment rate
1980	0.2942	0.0123	(0.0093, 0.0162)
1981	0.2930	0.0136	(0.0086, 0.0178)
1982	0.2959	0.0214	(0.0132, 0.0279)
1983	0.3004	0.0271	(0.0234, 0.0345)
1984	0.3136	0.0245	(0.0201, 0.0303)
1985	0.3255	0.0291	(0.0203, 0.041)
1986	0.3373	0.0266	(0.0198, 0.0333)
1987	0.3592	0.0197	(0.0173, 0.0237)
1988	0.3763	0.0169	(0.0141, 0.0194)
1989	0.3849	0.0157	(0.0131, 0.0188)
1990	0.3994	0.0167	(0.0131, 0.021)
1991	0.4136	0.0151	(0.0135, 0.0182)
1992	0.4209	0.0151	(0.0127, 0.0192)
1993	0.4324	0.0145	(0.0123, 0.019)
1994	0.4556	0.0156	(0.012, 0.0199)
1995	0.4689	0.0179	(0.0138, 0.0209)
1996	0.4774	0.026	(0.0203, 0.0319)
1997	0.4849	0.0272	(0.0245, 0.0303)
1998	0.5019	0.0269	(0.0229, 0.0305)
1999	0.5102	0.0292	(0.0273, 0.0322)
2000	0.5182	0.0299	(0.0273, 0.0327)
2001	0.5380	0.0457	(0.0335, 0.0533)
2002	0.5513	0.0517	(0.0498, 0.0535)
2003	0.5702	0.0499	(0.0458, 0.0521)
2004	0.5982	0.0444	(0.0409, 0.0467)
2005	0.6184	0.0413	(0.0386, 0.0436)
2006	0.6369	0.0391	(0.0378, 0.0409)

Using the Mann-Whitney method to determine whether the marriage rate in the Widow's Years and the marriage rate in the non-Widow's Years have significant impact. Let X be the singlehood rate of males aged 25-34 years old in the non-Widow's Years, Y be the singlehood rate of males aged 25-34 years old in the Widow's Years.

Thereby

H_0: There is no significant relationship between the singlehood rates of males aged 25-34 years old in the non-Widow's Years and the Widow's Years.

H_1: There is significant relationship between the singlehood rates of males aged 25-34 years old in the non-Widow's Years and the Widow's Years.

We found out that:

$$n_1 = 17,\ n_2 = 10,\ W_x = 229,\ W_y = 149$$

The result shows that H_0 cannot be rejected, and concluded that there is no significant relationship between the singlehood rates of males aged 25-34 years old in the non-Widow's Years and the Widow's Years.

3.3 Classifying using the fuzzy classification on threshold transfer function model

In this section, we use the fuzzy classification method to perform classification, in order to construction a more comprehensive model.

Step 1: First, use the k-means method to find out two cluster centers {cluster centers of the unemployment rate $C_1 = 0.0210$, $C_2 = 0.0454$} and {cluster centers of the singlehood rate $C_1 = 0.5331$, $C_2 = 0.3533$} of time series $\{x_t\}$, and determine the membership degree μ_{it}, $i = 1, 2$ of $\{x_t\}$ to these two cluster centers.

Step 2: Compute the fuzzy entropy $\delta(x_t)$, mean cumulated fuzzy entropy $MS\delta(x_t) = \frac{1}{t}\sum_{i=1}^{t}\delta(x_i)$, and the median of this series, that correspond to x_t. For the unemployment rate, Median $(MS\delta(x_t)) = 0.42910$; for the singlehood rate, Median $(MS\delta(x_t)) = 0.471$. Figures 3 and 5 are the charts for mean cumulated fuzzy entropy.

Step 3: Take a suitable threshold value λ, classify the mean cumulated fuzzy entropy $MS\delta(x_t)$ series corresponding to x_t. If the mean cumulated fuzzy entropy $MS\delta(x_t)$ falls into the interval $[0, \text{Median}(MS\delta(x_t)) - \lambda)$, we will use 1 to represent Group 1; if $MS\delta(x_t)$ falls into the interval $[\text{Median}(MS\delta(x_t)) - \lambda, \text{Median}(MS\delta(x_t)) + \lambda)$, we use 2 to represent Group 2; and if $MS\delta(x_t)$ falls into the interval $[\text{Median}(MS\delta(x_t)) + \lambda, 1]$, 3 will be used to represent Group 3. Following the classification method according to the theory, draw classification diagrams like Figures 4 and 6.

Step 4: Select an appropriate significance level α. Here we take $\alpha = 0.2$. If the number of consecutive samples is greater than $[27\alpha] = 6$, we will see the classification as successful; otherwise we will perform inductive classification based on the form of changes. During classification, if more than one group is formed, structural change happens in this time series. We will then find out its change interval.

Take the threshold value $\lambda = 0.01$, and set the change interval of the unemployment rate as years 1985 to 1988, we then consider this a transition period, and construct a new threshold model. But because the classification oscillation before year 1985 is too frequent, it is difficult to place the unemployment rate into any class. We will therefore deem it non-stationary, and only consider the unemployment rate after year 1988 to build the ARIMA model, as follows:

$$\begin{cases} \text{nonstationary}, & 1980 \le t \le 1987 \\ (1-0.0036B)(1-B)X_t = 0.0054 + \varepsilon_t, & 1988 \le t \le 2006 \end{cases}$$

and then construct the conversion model of the unemployment rate after the change interval to the singlehood rate, as follows:

$$(1-0.94B)Y_t = 0.0285 + (0.217 - 0.419B + 0.78B^2)X_t,\ 1988 \le t \le 2006$$

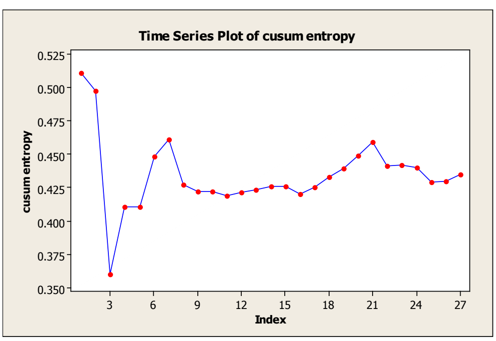

Figure 3. Chart of mean cumulated fuzzy entropy of the unemployment rate.

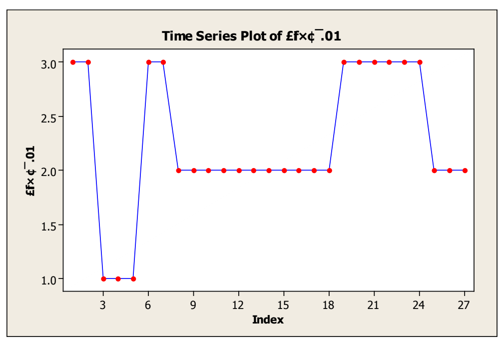

Figure 4. The classification diagram for the figure on the left, obtained by setting $\lambda = 0.01$.

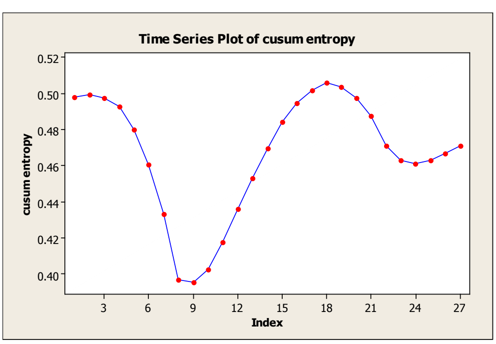

Figure 5. Chart of mean cumulated fuzzy entropy of the singlehood rate.

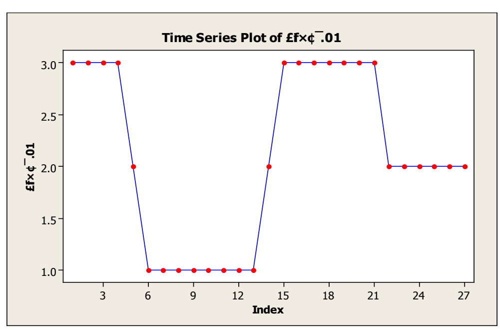

Figure 6. The classification diagram for the figure on the left, obtained by setting $\lambda = 0.01$.

When we take $\lambda = 0.01$, change intervals are found in years 1983 to 1985 as well as years 1992 to 1994. This means structural change happens. So we need to consider the singlehood rate after year 1985. However, the sample size is less than 6, so a model cannot be built. Nonetheless, observing the change intervals, we know that the structural change in the unemployment rate happened earlier than that of the singlehood rate, and in the change period of the unemployment rate, structural change had also occurred on the singlehood rate. Therefore the unemployment rate indeed has significant impact on the singlehood rate.

3.4 Classification of fuzzy time series

We will now perform the classification of fuzzy intervals using the fuzzy classification method, on the unemployment rate intervals. We will then obtain the cluster centers of the fuzzy intervals of the unemployment rate: $I_1 = (0.0183, 0.0245)$, $I_2 = (0.0371, 0.0523)$.

Also, with *Median* $(MS\delta(x_t)) = 0.2460$, from Figure 7 we know that the transition interval for the unemployment rate interval falls between years 1987 and 1988. We consider this a transition period, and construct a new threshold model. But because the classification oscillation before year 1987 is too frequent, it is difficult for us to place the unemployment rate in this period into any class. We will therefore deem it non-stationary, and take only the unemployment rate interval after year 1988 into account to build an ARIMA model, as follows:

$$\text{High model of intervals: } \begin{cases} \text{non} - \text{stationary}, & 1980 \le t \le 1987 \\ (1 - 0.0005B)(1 - B)X_t = 0.002 + \varepsilon_t, & 1988 \le t \le 2006 \end{cases}$$

$$\text{Low model of intervals: } \begin{cases} \text{non} - \text{stationary}, & 1980 \le t \le 1987 \\ (1 - 0.147B)(1 - B)X_t = 0.001 + \varepsilon_t, & 1988 \le t \le 2006 \end{cases}$$

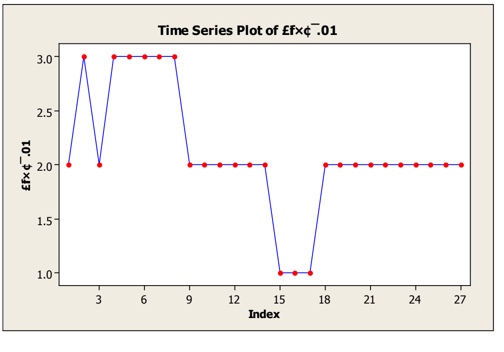

Figure 7. The classification diagram obtained by setting $\lambda = 0.01$.

And then determine the conversion model to the singlehood rate for each of them.

The highest unemployment rate-to-singlehood rate conversion model:

$$(1 - 1.02B)Y_t = 0.0005 - (0.016 - 0.152B^2)X_t,\ 1988 \le t \le 2006$$

The lowest unemployment rate-to-singlehood rate conversion model:

$$(1 - 0.945B)Y_t = 0.0292 + (-0.145 + 0.134B + 0.571B^2)X_t,\ 1988 \le t \le 2006$$

3.5 Comparison with forecast results

After constructing the model, we look at the core interest of this study--the forecasting ability. Table 3 is the comparison results for the forecast of unemployment rate between the best ARIMA model and fuzzy classification ARIMA. We can see that the forecast result of the unemployment rate using the best ARIMA model is not greatly different from that of with the fuzzy classification ARIMA. This could be due to the small sample size taken. Table 4 shows the forecast results obtained by performing fuzzy interval classification on the unemployment intervals, and then construct the threshold ARIMA model. It reveals that the MSE is very small, and this indicates that the result is considerably good.

Table 3. Comparison between the forecasts of the unemployment rate.

Retention phase	Actual value	ARIMA(1, 1, 0)	Fuzzy classification ARIMA(1, 1, 0)
2007	0.0391	0.0390	0.0389
2008	0.0414	0.0397	0.0393
2009	0.0585	0.0406	0.04
MSE		0.000108	0.000116

Table 4. Forecast intervals of the unemployment rate intervals.

Retention phase	Actual value	Fuzzy classification ARIMA(1, 1, 0)
2007	(0.0378, 0.0409)	(0.0388, 0.0421)
2008	(0.038, 0.0503)	(0.04, 0.0433)
2009	(0.0531, 0.0613)	(0.0413, 0.0445)
MSE		0.000111

For the forecast of the singlehood rate, we considered correlating the unemployment rate with the singlehood rate using the transformation. By taking the unemployment rate forecast in Table 3, and forecasting with the unemployment rate-to-singlehood rate transformation, the forecasted singlehood rate is obtained. Table 5 shows the forecast result. In the process of correlation, the fuzzy classification is also an important procedure, because this reduces the number of samples required. But because the overall data are not aplenty, the effect of the fuzzy classification method is not better than the transformation.

Table 5. Comparison of the unemployment rate forecasts.

Retention phase	Actual value	Transformation	Transformation of the unemployment rate
2007	0.6549	0.6545	0.6515
2008	0.6619	0.6710	0.6636
2009	0.6819	0.6873	0.6748
MSE		0.00004	0.00002

Table 6 shows the forecast of singlehood rate obtained by applying the threshold transformation on the intervals of the unemployment rate forecast from Table 4. We know from Tables 5 and 6 that, no matter it is performing the fuzzy classification on the unemployment rate first, and then the transform the unemployment rate to the singlehood rate; or it is performing the fuzzy classification on the intervals of unemployment rate before transforming the result of classification to the singlehood rate to get the forecast or intervals of forecast of the singlehood rate, the differences in MSE is not great. This is because the sample size is not large. However, this provided an approach to effectively obtain the forecast intervals of the singlehood rate, and has demonstrated the uncertainty in singlehood rate.

Table 6. Forecast intervals of the transformation of the singlehood rate.

Retention phase	Actual value	Forecast intervals of the transformation rate
2007	0.6549	(0.653, 0.656)
2008	0.6619	(0.667, 0.675)
2009	0.6819	(0.681, 0.695)
IMSE		0.000129

Overall, the forecasting we made, whether on the unemployment rate or the singlehood rate, is reasonably accurate. The forecasting will improve if classification transformation is used -- we only need to construct a model of the singlehood rate using the end result of classification of the unemployment rate. This way, the required number of data periods is lower (we only need the unemployment rate data after year 1998). In addition, the forecast interval of singlehood rate obtained from forecast of interval will provide a better forecasting ability for the uncertainty in forecasting singlehood rate.

At the end we also find that, for a time series, if the change period of its structural change can be determined, better results on the model construction and forecasting ability can be produced.

4. Conclusion

In the past, researchers emphasized on discrete data when they perform studies on structural change in time sequences. First they examine whether the data are stationary, and then they further analyze and conduct forecasts on variables. Each researcher has his own research direction, and the research approaches vary too, for instance on the choice of variables, whether to find out change point or a few structural change intervals, and so on. However, no intensive study was done on the determination of change period of interval time series. Because of this, this thesis introduces an effective method to resolve this issue. But the common objective that other researchers and us have, is that the products of our research are effective and have high forecasting ability, and can fit into real-life situations. We also hope that with the experience in analyzing the structural change in time series, we can produce effective procedures to detect change interval, and this can be used to do forecasting, in order to fulfill human beings' thirst to grasp the future changes.

In recent years, the unemployment rate in Taiwan is on the rise. The challenges that the Taiwan society faces are not only limited to the disorderly situation caused by the political reconstruction after the first shift of political power in the millennium year, as well as the "Democratic Labor's Pain." As the international economic environment and the cross-Strait economic and trade environment change, the investment and employment markets in Taiwan becomes increasingly challenging, and this created an impact to the supply-and-demand balance in the local employment market. Therefore, it is only after we find out the change interval of the structural change in the unemployment rate, that can we better forecast the future trend of the unemployment rate.

The increase of the singlehood rate causes the decrease in birth rate, and change in age structure. The unemployment rate has influence on a wide range of issues. Among them, this thesis discusses its impact on the singlehood rate. After determining the change interval of the unemployment rate, we can effectively find out the impact and forecast of the unemployment rate that went through structural change, on the singlehood rate. Nowadays there are many literatures that discuss the singlehood rate or the unemployment rate, but one that simultaneously investigates the relationship between the two, and how the Widow's Years affect them, is rare. Also, in this study, after we find out the change interval of the unemployment rate, we apply the conversion model and the intervention model to construct a dynamic model of the singlehood rate, in order to explain the relationship between the singlehood rate, the unemployment rate and the Widow's Years. Subsequently, the non-parametric method was used to detect and analyze the multi-interface interference time points. The result shows that the Widow's Years do not have significant influence on singlehood. At last, a reasonable forecast was made on the unemployment rate versus the single population.

Several topics that are worth researching were discovered during this study. Here is a list of them, for future study:

1. Is there a better way to determine the cluster centers for intervals, so that the relationship between intervals can be described more clearly?
2. This study mainly discusses the analysis of interval data with regards to Z-type, Λ -type, pi-type and S-type fuzzy data. This seems to be worth further research.
3. How does a change interval differ when the threshold value λ and significance level α differ? Meaning that, how to define a threshold value λ and significance level α that are more effective, and find out a more meaningful structural change time.

References

[1] D. W. K. Andrews and W. Ploberger (1994). Optimal tests when a nuisance parameter is present only under the alternative, *Econometrica*, 62, 1383-1414.

[2] J. Bai and P. Perron (1998). Estimating and testing linear models with multiple structural changes, *Econometrica*, 66, 47-78.

[3] G. C. Chow (1960). Testing for equality between sets of coefficients in two linear regression, *Econometrica*, 28, 291-260.

[4] Directorate General of Budget, Accounting and Statistics. (2010). *Statistical Yearbook of the Republic of China 2009*, Directorate General of Budget, Accounting and Statistics, Executive Yuan, Taipei, Taiwan.

[5] D.-W. Kim, K. H. Lee and D. Lee (2004). On cluster validity index for estimation of the optimal number of fuzzy clusters, *Pattern Recognition*, 37, 2009-2025.

[6] K. Kumar and B. Wu (2001). Detection of change points in time series analysis with fuzzy statistics, *International Journal of Systems Science*, 32, 1185-1192.

[7] J. Liu, S. Wu and J. V. Zidek (1997). On segmented multivariate regression, *Statistica Sinica*, 7, 497-525.

[8] M. K. Pakhira, S. Bandyopadhyay and U. Maulik (2005). A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification, *Fuzzy Sets and Systems*, 155, 191-214.

[9] W. Wang and Y. Zhang (2007). On fuzzy cluster validity indices, *Fuzzy Sets and Systems*, 158, 2095-2117.

[10] B. Wu and M.-H Chen (1999). Use of fuzzy statistical technique in change periods detection of nonlinear time series, *Applied Mathematics and Computation*, 99, 241-254.

[11] K.-L. Wu and M.-S. Yang (2005). A cluster validity index for fuzzy clustering, *Pattern Recognition Letters*, 26, 1275-1291.

[12] J.-C. Xie, B. Wu and S. Scriboonchita (2010). Fuzzy estimation methods and their application in the real estate evaluation, *International Journal of Intelligent Technologies and Applied Statistics*, 3, 187-202.

[13] L. A. Zadeh (1965). Fuzzy sets, *Information and Control*, 8, 338-353.

[14] W. Zeng and H. Li (2006). Relationship between similarity measure and entropy of interval valued fuzzy sets, *Fuzzy Sets and Systems*, 157, 1477-1484.

[15] H. D. Zhou (2005). Nonlinearity or structural break? -- Data mining in evolving financial data sets from a Bayesian model combination perspective, *Proceedings of the 38th Hawaii International Conference on System Sciences*, 6, 475-493.