國立政治大學應用數學系

碩士學位論文

模糊資料分類與模式建構探討
-以單身人口數及失業率為例

A study on the fuzzy data classification and model construction - with case study on the population of singles versus unemployment rate

碩士班學生：游鈞毅 撰

指導教授：吳柏林 博士

中華民國九十九年六月二十八日

謝辭

一轉眼時間飛逝，終於從政大畢業，因為本身不是數學系畢業的，所以讀起來有些吃力。但在這遇到了很多人，不厭其煩的讓我請教。尤其是小葉、家盛、B 哥，能夠在政大與你們當同學，真的非常的榮幸。在這兩年中的點點滴滴，會永遠的烙印在我心中。也非常感謝遇到了吳柏林老師，提供給我在寫作論文中了很多不一樣的想法與靈感。更讓我有機會能夠到國外參加研討會，雖然只是去見習，但也受益良多。而在今年五月多的時候，也因為老師的堅持，使我能夠有幸的在研討會中發表論文，這真的是一個很棒的經驗。最後要感謝我的父母親，讓我能夠無憂無慮的讀書，一路走來更支持我的任何決定。希望在往後的人生中，我能夠繼續的做使他們驕傲的兒子。

游鈞毅 謹誌于

國立政治大學應用數學所

中華民國九十九年六月

摘要

資料分類的應用在時間數列的分析與預測過程相當重要。而模糊資料近年來更受到重視，其應用的範圍包含：財金、社會、生醫、電機等各個領域。本研究欲運用模糊資料分類法，對區間時間數列的轉折偵測與模式建構做一個深入探討。主要應用平均累加模糊熵(average of the sum of fuzzy entropies)，找出其結構性改變的區間。並針對區間型時間數列進行模式建構診斷與預測。最後我們以單身人口數與失業率為實列做一個詳細的探討。結果顯示，失業率對單身人口數有顯著的影響而孤鸞年的效應並不顯著。

關鍵字：模糊資料分類 、轉折區間、平均累加模糊熵、失業率、單身人口數

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 .

Keyword: *fuzzy data classification*、*average of the sum of fuzzy entropies*、*change periods*、*unemployment rate* 、*population of singles*

目錄

第一章 前言

近年來晚婚族與未婚族的出現，導致台灣面臨少子化與人口老化問題。緊接著老人安養、人口年齡結構問題越來越受到重視。加上失業率的提升導致更多的人傾向於晚婚或不婚，使得單身人數越來越多。有鑑於此，本研究探討單身族比率的未來走向就變得越來越重要。模糊資料分析與探討架構是目前相當熱門的話題，傳統上有關文獻的研究大都著重於將資料反模糊化，再將資料進行分類，也就是模糊分類。所謂資料反模糊化，乃是指描述事物之模糊性質或找出彼此資料間之模糊關係。而分類是將資料有相同之模糊性資或關係分為同一類。但對於進一步討論資料結構性轉變的文獻並不多見。因此(Wu,1999)將兩者做結合，建構一有效應用模糊分類找出資料間結構性轉變的步驟。

過去的學者皆運用模糊分類法探討研究單時點問題，例：劉浩天(2004)、張弘紋(2010)、陳嘉甄(2009)、林原宏(2005) 。而 Weina Wang (2007)，Malay K(2005)、 Kuo-Lung Wu(2005)，Dae Won Kim(2004)進一步建構不同的演算法，探討在大量集合中如何有效運用 fuzzy k means 找出其 cluster。但對於區間型的時間數列卻無進一步的研究。Wenyi Zeng(2006)探討區間之間的模糊熵，但對於區間距離並無一明確定義。而本篇論文在第二節時提出了區間距離的測量、區間模糊熵、區間模糊隸屬度，可有效的提供一測量標準，來模糊分類一區間時間數列。

「轉折」一詞本身在語意學(semantics)上並無明確的認定，有很多的現象，很難以一般對或錯(true 或 false)的二元邏輯來加以認定，而模糊邏輯正好可以幫忙解決這方面的問題，這是考慮應用模糊理論來辨識與分類時間序列的原因。整題來說，時間數列上的轉折點應該是有程度上的差異，如 95%的轉折點等等。事實上，若以模糊理論的方向來探討轉折二字，轉折「區間」可能要比轉折「點」更能符合現實的狀況。Chow (1960)在線性迴歸模型下，檢定單一個已知結構改變時間點的轉折是否顯著。Liu et al. (1997)認為在多個轉折的線性模型藉由 Schwarz準則 (SIC)作為最小平方法的估計及轉折數的估計。Bai & Perron (1998)認為多個結構位移的線性模型藉由 Wald test 估計。Andrews & Ploberger (1994)採用多種方法廣泛的分析結構改變檢定的問題，以 Wald, Lagrange multiplier, Likelihood ration-like 檢定法。Kumar & Wu (2001)發現在非線性的時間序列藉由模糊邏輯的概念可以有效找出結構轉折。Zhou (2005)提出了創新的結構改變方法 - IntegratingBayesian structural break model和Change point detection methods。雖然眾多學者所做的結構改變分析，其研究結構改變使用的方法不計其數，但在數理

的推論過程都相當的繁瑣且轉折點的定義似乎仍無一明確的標準，因此本論文的目的亦希望找出失業率的結構性改變做有效分析，並對結構改變之部分與未婚率做轉換模式，使其找出更好的預測能力。

黃士滔(2004)、胡愈寧(2004)、許永河(1998)皆提出針對失業率的預測與應用，但未對失業率之本身結構改變做更進一步的研究。吳柏林(1991)- 台灣地區結婚率、 出生率、人口成長率的時間數列模式探討。楊靜利 (2006)-台灣傳統婚配空間的變化與婚姻行為之變遷。都是針對結婚率與時間或教育程度的關係。對於探討單身族未來人口比率的變化文獻並不多見。而近年來台灣大學越來越多的驅勢下，男性25歲前還在就學的情況越來越普遍。25歲以下的男性就業情況也就不顯著。故本篇論文針對25-34歲的男性做為主要研究對像，來探討單身人口比率與失業率、孤鸞年之間關係。藉此探討單身族未來人口比率的變化與趨勢預測。以提供社會人口調查研究的對於單身族婚姻參考。

2300 萬的台灣人口，有 670 萬單身，男性就佔了 54% (主計處 2009 主計年刊)。根據行政院主計處 2009 年台灣 15 歲以上未婚者已突破六百萬大關。其中男性佔了 54%。主計處(2009)的統計資料中也指出，台灣 25-34 歲單身男性佔了此 15 歲以上未婚男性的 13%。男性在婚姻上必須負擔經濟主要供應者的角色，但在目前台灣失業率高漲、房價上升的情況下。負擔家計的壓力愈來愈大，結婚難度也愈來愈高，以至單身男性越來越高。

習俗上，孤鸞年一直是民間所流傳的一個禁忌，民間相信只要在孤鸞年結婚的男女，都會有婚姻上的不幸福，故只要是孤鸞年，家中長輩就會提醒有意結婚的男女，盡量避開在這一年裡結婚。因此連帶也會影響當年的未婚率升高。未婚族導致出生率的降低，使得人口老化比率的提高，未來台灣將面臨少子化、人口結構老化、老人安養、總體生產力衰退的問題。有鑑於此，了解男性單身族比率的未來走向是一大課題。

本文主要分為四節。第一節探討相關方法之文獻，以獲取發展理論架構的依據，並說明分析失業率與未婚率之目的。第二節說明 ARIMA 分析之結構及模糊理論與分類，接著架構新的區間測量方式。第三節為實證分析。最後，本文結論在第四節。

第二章 研究方法

本研究利用模糊集合觀念對相關之時間數列進行分類探討。並對偵測轉折區間與模糊模式之建購做探討。

2.1 ARIMA 轉換模式

所謂的時間數列意指以時間順序的形態所呈現的一連串觀測值，亦即對某種動態系統隨著時間持續記錄所產生有順序觀察值之集合，時間序列分析法的理論早在 1920 年就學者開始提出，Box 和 Jekins(1970)完成了自我迴歸移動平均整合模式(ARIMA model)，從此之後該研究方法即被應用於找出原始數列的變動模式。因此本人運用 ARIMA 模式來找出單身人口的變動模型，與歷年失業率的影響之下有何變動。

有關時間序列模式的選擇，由於時間序列有許多不同的方法，常用的如：指數平滑曲線(Exponential smoothing)、自我迴歸整合性移動平均(Autoregression integrated movingaverage, ARIMA)、對數線性趨勢(Log lineartrend)、線性趨勢含季節性變動(Linear trendwith seasonal terms)等，一般應根據序列本身的型態加上適合度檢定，以決定最合適的方法來進行序列的評估和預測。而時間序列模式建立的步驟為 1.鑑定；2.評估與診斷；3.預測。一開始須先確定失業率與未婚率是否有相關性。運用迴歸分析明顯指出不管是失業率對未婚率。又或者是未婚率對失業率而言都有顯著的負相關性。再分別針對兩者算出各別之 ACF 以及 PACF。接著找出兩者間之 CCF，以便求出其轉換模式。

以下為 1920 年 Box 與 Jenkins 所提出進階的建模技術並且以遞迴的方式對時間數列資料建構模型 ARIMA(p, d, q)：

$$\varphi_p(B)\nabla^d X_t = \theta_q(B)\varepsilon_t$$

其中：B為倒退因子，即 $BX_t = X_{t-1}$

$\varphi_p(B) = (1 - \varphi_1 B^1 - \varphi_2 B^2 - \cdots - \varphi_P B^P)$，φ為自我回歸參數

$\theta_q(B) = \left(1 - \theta_1 B^1 - \theta_2 B^2 - \cdots - \theta_q B^q\right)$，θ為移動平均參數

$\nabla^d = (1 - B)^d$

X_t為一時間數列隨機變數、 ε_t 為white noise、d為差分階數、p 代表自我迴歸級數、q 為移動平均級數。

轉換模式意指將單變量時間數列模式建構法，將其推廣至多元時間數列分析

法。由於單因子模型之對自己受時間影響之分析,對未來之預測能力可能較不正確。而在許多的例子中有可能發生一筆資料其目前的觀測值受到過去的觀測值影響，並且與另一筆(或多筆)時間數列資料具有相關性。亦即當投入數列發生變化時，其將有多少影響傳送到輸出數列之情形。因此未婚率是否受到失業率前幾期影響的可能性或受自己過去所影響，考慮利用轉換函數模式來建立未婚率之轉換模式應會更精確。以下我們將詳細介紹轉換模式之建構

1.轉換函數模式介紹

考慮係由二元隨機過程所產生的時間數列，若將X_t視為投入變數，Y_t視為產出變數而之間的關係可表示為

$$Y_t = U_t + N_t$$

式中U_t：為Y_t之一部分，僅用來解釋X_t部分

N_t：為干擾項(Disturbance Term)與X_t無關。

首先，考慮U_t與X_t之關係，以線性動態關係表示，可記為

$$U_t = \nu_0 X_t + \nu_1 X_{t-1} + \nu_2 X_{t-2} + \cdots = (\nu_0 + \nu_1 B + \nu_2 B^2 + \cdots) X_t = \nu(B) X_t$$

其中，ν_0、ν_1，...為各時期X_t之衝擊反應權數。則

$$Y_t = \nu(B) X_t + N_t$$

其次，考慮干擾項部分，一般干擾項係為非穩定數列，因而符合 ARIMA（p,d,q）模式，即

$$N_t = C + \frac{\theta(B)}{\phi(B)} \alpha_t$$

其中，$\alpha_t \sim N(0, \sigma_a{}^2)$，C 為常數項。因此

$$Y_t = C + \nu(B) X_t + \frac{\theta(B)}{\phi(B)} \alpha_t$$

2.模式鑑定

轉換函數模型的建立過程與單變量時間數列模型構建方法相同,亦是一種遞

迴方式。模型鑑定中需要考慮三個基本問題：

(1) 衝擊反應權重$\nu(B)$的估計。

(2) 干擾項$\frac{\theta(B)}{\phi(B)}\alpha_t$的決定。

(3) 與$\nu(B)$最近似的有理型式$\frac{w(B)}{\delta(B)}$的決定。

以下我們探討CCF方法之模式鑑定：

在轉換函數當中樣本交叉相關係數(Cross Correlation Function；簡稱CCF)作為鑑定模型之主要工具。假設有一組n個觀測值數列相隔k個時差之樣本交叉相關係數定義為

$$r_{xy}(k)=\frac{c_{xy}(k)}{\sqrt{c_{xx}(0)c_{yy}(0)}}，k=0,\pm1,\pm2,\ldots$$

式中

$$c_{xy}(k)=\begin{cases}\frac{1}{n}\sum_{t=1}^{n-k}(\chi_t-\bar{\chi})(y_{t+k}-\bar{y})，k=0,+1,+2,\ldots\\ \frac{1}{n}\sum_{t=1}^{n+k}(y_t-\bar{y})(\chi_{t-k}-\bar{\chi})，k=0,-1,-2,\ldots\end{cases}$$

且$\bar{\chi}$與$\bar{y}$為數列x與y之平均值，又$s_x=\sqrt{c_{xx}(0)}$與$s_y=\sqrt{c_{yy}(0)}$ 故

$$r_{xy}(k)=\frac{c_{xy}(k)}{s_x s_y}，k=0,\pm1,\pm2,\ldots$$

基於上述討論，轉換函數模型之鑑定過程如下：

(1)建構投入數列X_t的ARIMA模型並保留其殘差數列，此步驟一般稱為白噪音畫投入數列，即

$$\alpha_t=\frac{\theta(B)}{\phi(B)}X_t$$

(2)利用(1)投入數列估計後的ARIMA模型將產出數列Y_t轉換，此步驟稱為過濾化產出數列，即

$$\beta_t = \frac{\theta(B)}{\phi(B)} Y_t$$

(3)計算白噪音化投入數列α_t與過濾化產出數列β_t之交叉相關函數CCF來估計衝擊反應權數，即

$$\hat{v}_k = \frac{\hat{\sigma}_\beta}{\hat{\delta}_\alpha} r_{\alpha\beta}(k)$$

(4)利用$\hat{v}_k$的型式與理論v_k圖形匹配，以決定合適的$\frac{\omega(B)}{\delta(B)}$之s與r值以及落後時差b值，即

$$\hat{v} = \frac{\hat{\omega}_s(B)}{\hat{\delta}_r(B)} B^b$$

(5)利用上述決定的轉換函數模型，但假設干擾項為一種白干擾過程，即令$n_t = a_t$，並進行模型參數之估計且並保留其殘差數列值，即可得$\hat{n}_t$為

$$\hat{n}_t = y_t - \hat{v}(B)\chi_t = y_t - \frac{\hat{\omega}_s(B)}{\hat{\delta}_r(B)} B^b \chi_t$$

(6)利用上述第5個步驟之殘差數列，應用單變量模式建構法來認定干擾項的ARIMA模式。

(7)重新認定與估計最終獲取的模式。

2.2 模糊集合與群落隸屬度

人類的思維具有很多的不確定性，如果用傳統二分法強行分類，就會產生錯誤的結論，例如：“天氣好”在傳統的測量尺度下，受訪者只會被問到好或不好兩個選項，並無法把天氣的不確定性包含在其中。因此模糊集合的概念由 Zadeh (1965)首先提出，對於多元複雜的模糊現象，給於較完善的解決方式。

在模糊理論中，乃將傳統二值邏輯的觀念與運算方法，利用隸屬度函數來表示事物的模糊現象。意表在傳統集合論中，一元素只屬於某一集合或不屬於。但在模糊理論中，用隸屬度來表示該元素，隸屬於某一集合的程度。例：“快樂”這一名詞，因個人差異而有所不同，故具有不確定性。但模糊理論卻可以顯示 A

君在此時刻下的 80%的快樂程度。在模糊理論中，隸屬度的全距範圍一般都設定在 0 到 1 之間。這可使“快樂”此一語言變數，成為一感覺性的分布。

應用模糊理論檢定時間數列是否發生轉折時，應先將時間數列分群，找出群落中心。再運用模糊隸屬度與模糊熵等觀念進行分類。其定義如下：

定義 2.1 模糊隸屬度 (Wu and Chen, 1999)
令一時間數列$\{x_t$,t=1,2,…,N$\}$，且 C_1 與 C_2 為時間數列的兩個群落中心令 μ_{it}，$i = 1,2$，表示時間數列X_t中的元素x_t對 C_1、 C_2 的隸屬度，則定義隸屬度為

$$\mu_{it} = 1 - \frac{|x_t - C_i|}{\sum_{i=1}^{2} |x_t - C_i|}$$

定義 2.2 模糊熵（Wu and Chen, 1999）
令一時間數列$\{x_t$, t=1,2,…,N$\}$，μ_{it}表x_t對群落中心C_i(i=1,2,...,k)之隸屬度，則x_t 的模糊熵定義為

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

而熵是熱力學中的一個觀念，它的本意是熱量可以轉變功的程度，統計物理學給予另一解釋：它是描述分子運動無規則的一種度量。而機率論和訊息論又給了它更一般的說明：它是隨即變量無約束度的一種度量，是剩餘資訊量大小的一種度量。所以模糊熵表用來測量模糊集合的不確定性，是處理模糊資料的重要工具。而模糊隸屬度用來描述元素無法明確界定是否屬於給定集合的集合類。

2.3 模糊熵與區間距離

模糊區間集合可視為連續型的模糊集合，能更進一步表示一不確定性的事物，例：“評量成績等第”，在外國學校中常中 A、B、C、D 來評量學生的成績，A 表示為 100-80 分、B 表示 79-70 分、C 表示 69-60 分、D 表示 59-50 分。來代替以往分數的取向，因為以往我們認為分數高就代表的學得好，但 80 分與 85 分就代表著得 85 分的同學學習能力就更好嗎？答案是不一定的。因此模糊區間集合，就解決了此一不確定性現象。

當有了區間模糊樣本，我們必須考慮區間的運算，有關區間運算可參考吳(2005)。但對於其區間距離的測量尚無完備之定義(見，吳 2010)。本節將定義一

較合適(well-defined)區間距離，並應用此定義計算區間聚落中心、區間模糊隸屬度。

如何定義一較合適之區間距離呢？首先我們將區間以$(c_i; r_i)$表示，c 代表中心點，r 代表半徑。因此區間距離可考慮為中心點的差異加上半徑的差異。其中中心點差異可視為位置差異，半徑差異可視為廣度(Scale)差異。但是為了避免廣度差異對位置影響過大，我們將廣度差異化為 *ln*，再加上 1 為避免使 *ln* 為負值。

定義 2.3 區間型模糊樣本之距離

設 U 為一論域，令$\{\chi_i = [a_i, b_i], i = 1,2\}$為自論域 U 中抽出的二個區間模糊樣本，$c_i = \frac{a_i+b_i}{2}$, $r_i = \frac{a_i-b_i}{2}$。定義兩區間模糊樣本χ_1與χ_2之距離為

$$d(\chi_1, \chi_2) = |c_1 - c_2| + \ln(1 + |r_1 - r_2|)$$

例 2.1：令兩筆區間資料$\chi_1 = [2,5]$、$\chi_2 = [3,7]$

則$\chi_1 = (3.5\ ;\ 1.5)$，$\chi_2 = (5\ ;\ 2)$

$$d(\chi_1, \chi_2) = |3.5 - 5| + \ln(1 + |1.5 - 2|) = 1.9$$

例 2.2：令兩筆區間資料$\chi_1 = [3,5]$、$\chi_2 = [4,5]$

則$\chi_1 = (4\ ;\ 1)$，$\chi_2 = (4.5\ ;\ 1)$

$$d(\chi_1, \chi_2) = |4 - 4.5| + \ln(1 + |1 - 1|) = 0.5$$

定義 2.4 區間型模糊樣本之預測均方誤差（mean square error of interval , IMSE)

令$\{\chi_i = [a_i, b_i], i = 1, \dots, N\}$為一區間時間數列，預測區間為$\hat{\chi}_i$=$[\hat{a}_i, \hat{b}_i]$，$\varepsilon_i = d(\chi_i, \hat{\chi}_i)$為預測區間與實際區間的誤差，則

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

其中 l 代表往前預測期數。

例 2.3：大學生預測薪資如下表

往前期數	預測薪資	實際薪資
1	[3,4]	[3,6]
2	[2,6]	[4,5]
3	[3,5]	[2,6]
4	[5,7]	[5,8]
5	[4,5]	[3,8]

則

$\chi_1 = [3,4] = (3.5; 0.5)$、$\hat{\chi}_1$=[3,6]=(4.5;1.5)

$$d(\chi_1, \hat{\chi}_1) = |3.5 - 4.5| + \ln(1 + |0.5 - 1.5|) = 1.69$$

$\chi_2 = [2,6] = (4; 1)$、$\hat{\chi}_2$=[4,5]=(4.5; 0.5)

$$d(\chi_2, \hat{\chi}_2) = |4 - 4.5| + \ln(1 + |1 - 0.5|) = 1.41$$

$\chi_3 = [3,5] = (4; 1)$、$\hat{\chi}_3$=[2,6]=(4;2)

$$d(\chi_3, \hat{\chi}_3) = |4 - 4| + \ln(1 + |1 - 2|) = 0.69$$

$\chi_4 = [5,7] = (6; 1)$、$\hat{\chi}_4$=[5,8]=(6.5;1.5)

$$d(\chi_4, \hat{\chi}_4) = |6 - 6.5| + \ln(1 + |1 - 1|) = 0.91$$

$\chi_5 = [4,5] = (4.5; 0.5)$、$\hat{\chi}_5$=[3,8]=(5.5;2)

$$d(\chi_5, \hat{\chi}_5) = |4.5 - 5.5| + \ln(1 + |0.5 - 2|) = 0.09$$

則根據定義 2.4 ，IMSE= $\frac{1}{5} \times (1.69^2 + 1.41^2 + 0.69^2 + 0.91^2 + 0.09^2) = 2.12$

定義 2.5 區間型時間數列聚落

設$\Psi=\{\chi_t, t=1,2,\ldots\ldots,N\}$為一區間時間數列，$k \in \mathbb{N}$ 為群落個數。若存在一集合$\mathcal{I}=\{I_i \in \text{interval}; i=1,2,\ldots,k\}$，使得 Ψ 中的元素 χ_t與$\mathcal{I}$中的元素I_i的距離平方和為最小，即

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

則稱集合 $\mathcal{I}=\{I_i \in \text{interval}; i=1,2,\ldots,k\}$，為區間時間數列 Ψ 的聚落區間集合。

例 2.4：我們用失業率(單位：百分比)27 筆區間資料，其分佈圖如圖 2.1 所示

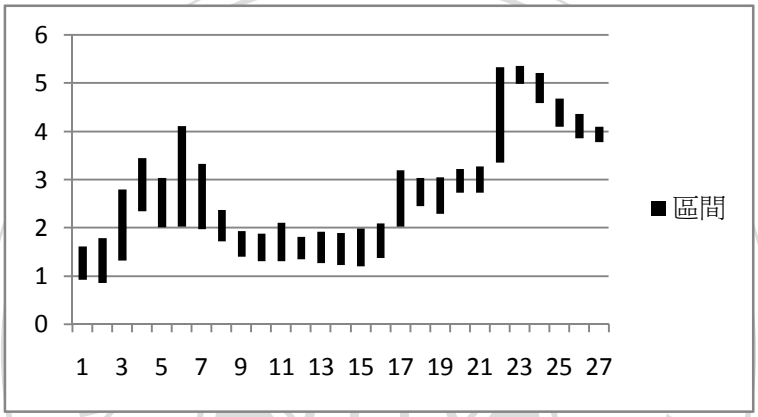


圖 2.1 區間時間數列走勢圖

若我們欲將資料分成兩群，則利用定義 2.5，可得兩個區間聚落$I_1=(1.83,2.46)$，$I_2=(3.71,5.23)$其分群結果如下圖

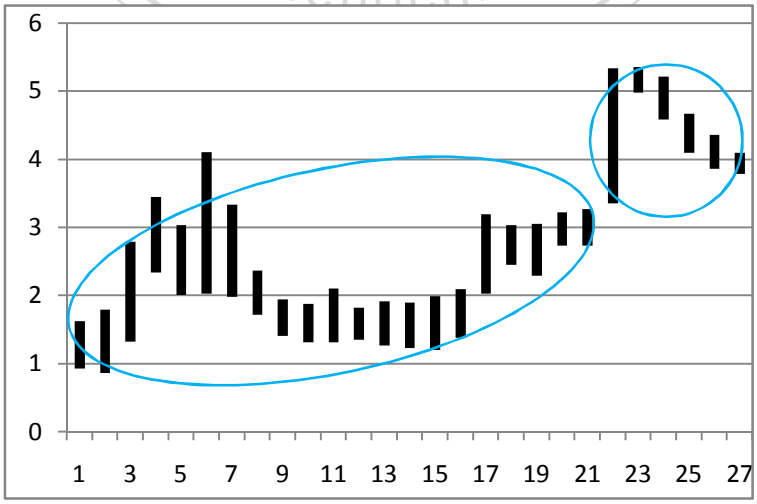


圖 2.2 區間群落結果

定義 2.6 區間模糊隸屬度

令$\psi=\{\chi_t, t=1,2,\cdots\cdots,N\}$為一區間時間數列，且$I_i$為聚落區間，令$\mu_{it}$表示區間時間數列$\Psi$中的元素$x_t$對$I_i$的隸屬度，$i=1,..,k$，則定義隸屬度為

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

2.4 轉折區間

由於傳統上，偵測系統之結構性之改變，大都以考慮轉折點為主，但結構轉折應是以變數值為主而非時間值，應是漸進轉折的轉折區間而非一時點遽變的轉折點，所以探討變數的轉折區間，相較於傳統時間序列研究方法，有更好的解釋能力。吳(1999)提出，運用模糊熵可有效的辨認此時間數列是否有結構性改變的發生。此外，並利用 t 個時間的平均累加模糊熵來觀測模糊熵的訊息變化情況，並據以作為模型轉折分類的標準。

定義 2.7 平均累加模糊熵(Mean Cumulated Fuzzy Entropy) (Wu 1999)

令一時間數列$\{x_t\}$，t =1,2,..,N ，$\delta(x_t)$為其模糊熵，則定義其平均累加模糊熵為：

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

模糊分類通常都會設定一門檻水準(threshold lever, λ)，因為無論是自然或人文科學，判斷分類的認定相當主觀與分歧。故需要一評量尺度。根據實證經驗 λ 不應取太大或太少，否者會有無法分類或分類過多的情況。因此選取 λ 在 0.1 至 0.001 是較好的選擇。

本文我們找兩個聚落中心，此乃根據一般的實證分析經驗及時間數列的走勢研判的(Wu and Chen, 1999)，其分類步驟如下：

步驟1：先利用k-means method找出時間序列$\{x_t\}$的2個群落中心C_1、C_2，並決定$\{x_t\}$對2個群落中心的隸屬度 μ_{it} ，i=1,2。

步驟2：計算出 x_t 對應的模糊熵$\delta(x_t)$、平均累加模糊熵$MS\delta(x_t)=\frac{1}{t}\sum_{i=1}^{t}\delta(x_i)$

此數列的中位數Median($MS\delta(x_t)$)

步驟3：取適當的一門檻值λ,將x_t 對應的平均累加模糊熵$MS\delta(x_t)$數列進行分類。

若平均累加模糊熵MSδ(x_t)落在區間[0,Median(MSδ(x_t))−λ)，我們以1表示第一類組；若MSδ(x_t)落在[Median(MSδ(x_t))−λ,Median(MSδ(x_t))+λ)，以2表示第二類組；若MSδ(x_t)落在區間[Median(MSδ(x_t))+λ,1]，則以3表示第三類組。

步驟4：若分類結果不一致，則對此分類結果做調整。若分類皆相同，則跳過至步驟5。

步驟5：選取適當的判定水準α，若連串的樣本數大於[αN]，則此一連串的樣本歸屬於同一類組。當分類的組數超過一組時，表示此一時間數列發生結構性改變。進而找出其轉折區間。

第三章 實證分析

本節我們想應用單身人口數與失業率進行模糊統計分析，找出失業率之結構改變的轉折區間以及對未婚率之影響與預測。

3.1 資料來源

資料來源為內政部主計處西元 1980 年至 2006 年 25-34 歲以上男性未婚人口數除以 25-34 歲男性人數所得之比率。以及內政部主計處在西元 1980 年至 2006 年的失業率統計數據，由表 3.1 所示。而 1980 年到 2006 年共有八次孤鸞年各為 1982、1985、1987、1990、1993、1995、1998、2001、2004、2006 年。圖 3.1 為表 3.1 之 25-34 歲男性未婚比率的時間數列走勢圖。圖 3.2 為表 3.1 所示之失業率時間數列走勢圖。圖 3.3 為表 3.1 之失業率之模糊區間的時間數列走勢圖。

表 3.1

年份	25-34 歲男性未婚比率	失業率	失業率之模糊區間
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)

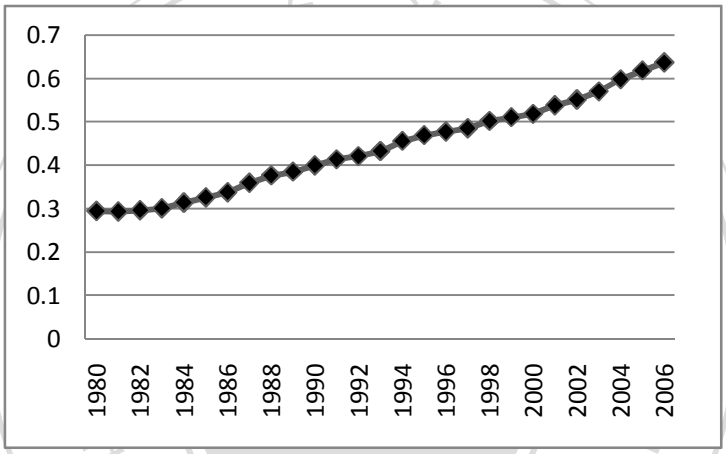


圖 3.1 西元 1980 至 2006 之 25-34 歲男性未婚比率走勢圖

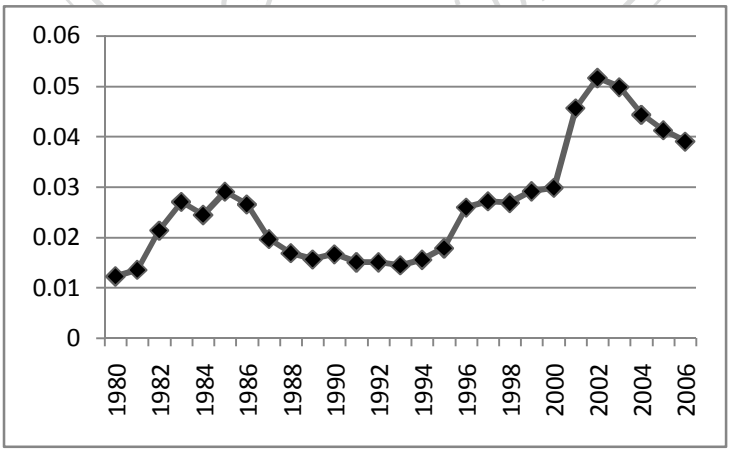


圖 3.2 西元 1980 至 2006 之失業率走勢圖

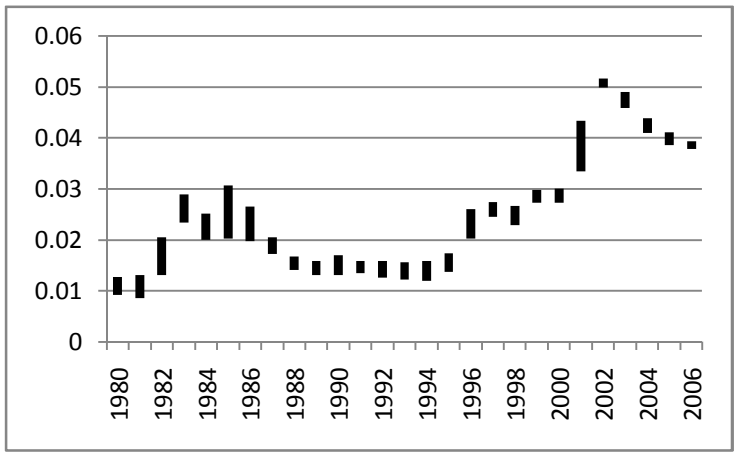


圖 3.3 西元 1980 至 2006 之失業率區間走勢圖

3.2 以轉換模式建構

在本節中利用時間數列之轉換模式建構模型，先將其 25-34 歲男性未婚比率與失業率在 1980 至 2006 年之 CCF 求出，在其比較關係我們可以求的關係是如下：

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

其中　Yt：表時間點 t 的結婚率、Xt：表時間點 t 的失業率

而投入數列之模型為

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

再探討孤鸞年多重介入模式。我們令$\mathrm{I_{t_1}}$為

$$\begin{cases} 1, \mathrm{t_1} = 1982、1985、1987、1990、1993、1995、1998、2001、2004、2006 \\ 0, \mathrm{o.w} \end{cases}$$

則

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

緊接著運用mann-whitney方法判定孤鸞年結婚率與非孤鸞年結婚率是否有顯著影響

令X:非孤鸞年之25-34歲男性未婚比率

Y:孤鸞年之25-34歲男性未婚比率

則

H_0: 非孤鸞年25-34歲男性未婚比率與孤鸞年25-34歲男性未婚比率無顯著關係

H_1: 非孤鸞年25-34歲男性未婚比率與孤鸞年25-34歲男性未婚比率有顯著關係

表3.2

未婚比率	母體	等級
0.294	X	1
0.293	X	2
0.296	Y	3
0.30	X	4
0.314	Y	5
0.325	X	6
0.337	X	7
0.369	Y	8
0.378	X	9
0.385	X	10
0.399	Y	11
0.414	X	12
0.4160	X	13
0.432	Y	14
0.456	X	15
0.469	Y	16
0.477	X	17
0.485	X	18
0.502	Y	19
0.51	X	20
0.518	X	21
0.538	Y	22
0.551	X	23
0.637	Y	24
0.655	X	25
0.622	X	26
0.682	Y	27

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

結果顯示無法拒絕H_0，即認為非孤鸞年之25-34歲男性未婚比率與孤鸞年之25-34歲男性未婚比率無顯著關係。

3.3 以模糊分類法分類並建立門檻轉換模式

在本節中，我們利用模糊分類法來分類，以期能建立更完善的模式。

步驟1：先利用k-means method找出時間序列$\{x_t\}$的2個群落中心{失業率之群落中心C_1=0.0210、C_2=0.0454}、{未婚率之群落中心C_1=0.5331、C_2=0.3533}，並決定$\{x_t\}$對2個群落中心的隸屬度 μ_{it}，i=1,2。

步驟2：計算出 x_t 對應的模糊熵$\delta(x_t)$、平均累加模糊熵$MS\delta(x_t) = \frac{1}{t}\sum_{i=1}^{t}\delta(x_i)$及數列的中位數。則失業率之Median($MS\delta(x_t)$)=0.42910，而未婚率之Median($MS\delta(x_t)$)=0.471。其中平均累加模糊熵的走勢圖如圖3.4和圖3.6。

步驟3：取適當的一門檻值λ,將x_t對應的平均累加模糊熵$MS\delta(x_t)$數列進行分類。若平均累加模糊熵$MS\delta(x_t)$落在區間[0,Median($MS\delta(x_t)$)−λ)，我們以1表示第一類組；若$MS\delta(x_t)$落在[Median($MS\delta(x_t)$)−λ,Median($MS\delta(x_t)$)+λ)，以2表示第二類組；若$MS\delta(x_t)$落在區間[Median($MS\delta(x_t)$)+λ,1]，則以3表示第三類組。在依據理論中分類法，繪成如圖3.5和圖3.7的分類圖。

步驟4：取適當顯著水準α，此時取$\alpha = 0.2$，當連串的樣本數超過[27α]=6時我們才視為分類成功，反之將視其轉折型式歸納分組。當分類的組數超過一組時，表示此一時間數列發生結構性改變。進而找出其轉折區間。

取門檻值λ＝0.01時，失業率轉折區間為西元1985至1988年，我們以此為一轉型期，建立一新的門檻模式，但在1985年之前因分類振盪過於頻繁，很難歸屬於任何一類因此我們視為不穩定狀態，故只需考慮1988年後的失業率做ARIMA模式，如下所示：

$$\begin{cases} \text{不穩定，} 1980 \le t \le 1987 \\ (1 - 0.00357B)(1 - B)X_t = 0.00536 + \varepsilon_t\text{，} 1988 \le t \le 2006 \end{cases}$$

再將轉折區間後的失業率做在對未婚率做轉換模式建構如下：

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

當取λ＝0.01時未婚率在西元1983至1985時及西元1992至1994年發生轉折區間，亦即發生結構性的改變。故知須考慮西元1985年後的未婚率，但因其樣本數過少(size≤ 6)無法建構模式。但卻可以從轉折區間得知，失業率發生結構性的改變較未婚率早，且失業率發生轉折期間未婚率也發生了結構性的變化。因此失業率確實對未婚率有顯著影響。

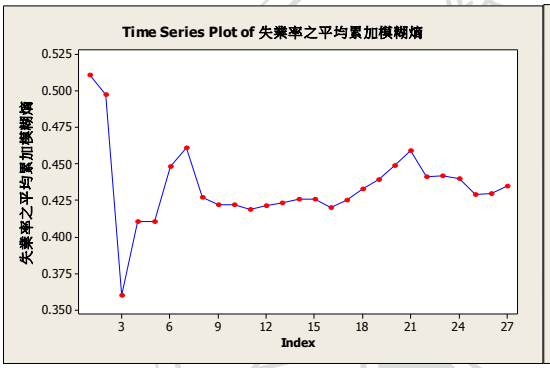


圖3.4 失業率之平均累加模糊熵走勢圖

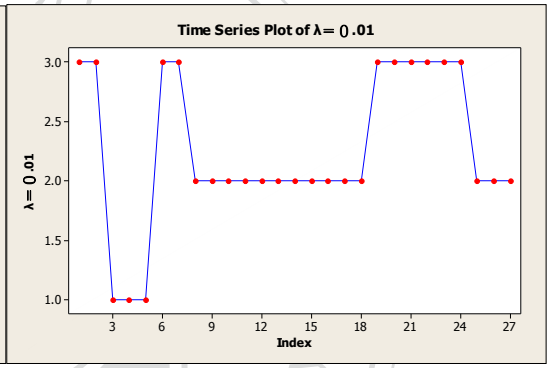


圖3.5 左圖以λ＝0.01所得出的分類圖

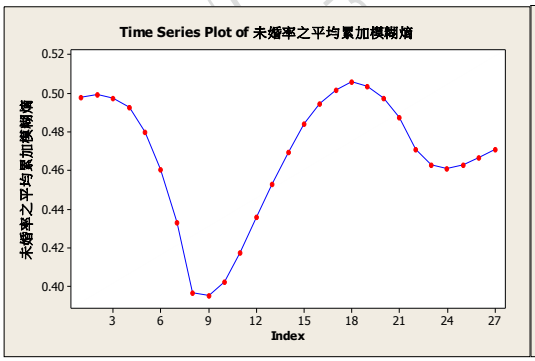


圖3.6 未婚率之平均累加模糊熵走勢圖

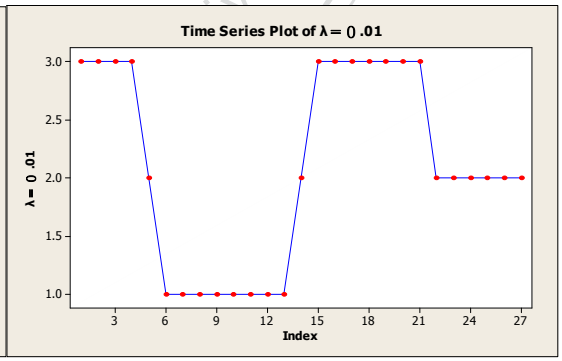


圖3.7 左圖以λ＝0.01所得出的分類圖

3.4 以模糊分類法分類模糊區間

接著我們對失業率區間以模糊分類法分類模糊區間，則我們可得失業率模糊區間聚落中心為$I_1 = (0.0183，0.0245)$、$I_2 = (0.0371，0.0523)$。

且 *Median*(MSδ(x_t))=0.2460，則由圖 3.9 知失業率區間之轉型區間為西元 1987 至 1988 年，我們依此為一轉型期，建立一新的門檻模式。但在 1987 年之前因分類振盪過於頻繁，很難歸屬於任何一類因此我們視為不穩定狀態。故只需考慮 1988 年後的失業率區間做 ARIMA 模式，如下所示：

區間高點模型：$\begin{cases} \text{不穩定，} 1980 \le t \le 1987 \\ (1-0.0005B)(1-B)X_t = 0.002 + \varepsilon_t \text{ ，} 1988 \le t \le 2006 \end{cases}$

區間低點模型：$\begin{cases} \text{不穩定，} 1980 \le t \le 1987 \\ (1-0.147B)(1-B)X_t = 0.001 + \varepsilon_t \text{，} 1988 \le t \le 2006 \end{cases}$

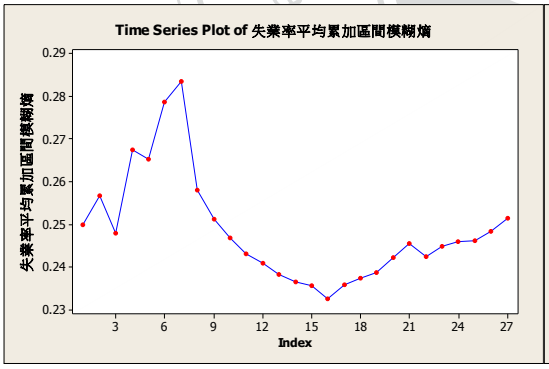
再個別於未婚率求轉換模式

最高失業率對未婚率之轉換模型：

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

最低失業率對未婚率之轉換模型：

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



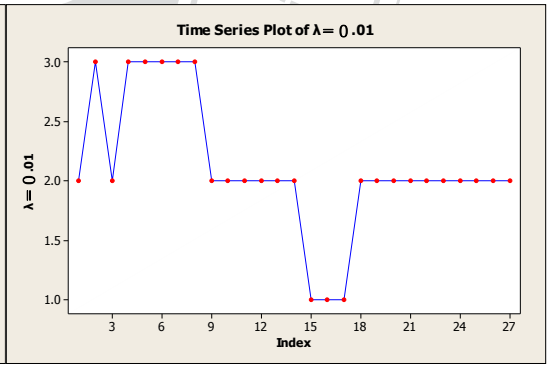


圖3.8失業率之平均累加模糊熵走勢圖　圖3.9 左圖以λ＝0.01所得出的分類圖

3.5 預測結果比較與分析

模式建構之後，最其關心之預測能力。表 3.3 為最佳 ARIMA 模式與模糊分類 ARIMA 之失業率預測結果比較。可以看出最佳 ARIMA 模式與模糊分類 ARIMA 之失業率預測結果差距不大，可能是樣本數不夠多所導致其結果。表 3.4

為將失業率區間利用模糊區間分類後再做門檻 ARIMA 模式之預測結果，發現 MSE 非常的小表示其結果相當不錯。

表3.3 失業率之預測值比較

保留期數	實際值	ARIMA(1,1,0)	模糊分類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

表3.4 區間失業率之預測區間

保留期數	實際區間	模糊區間分類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)
IMSE		0.000111

對於未婚率的預測值我們考慮運用轉換模式來配適。表3.5為利用表3.3之失業率預測值運用失業率與未婚率之轉換模式預測未婚率之預測值。在模式配適過程中模糊分類也是相當重要得一個程序，因為可以減少其需要的樣本數。但因為整體的資料並不多所以模糊分類法並沒有比轉換模式效果來得好。

表3.5 未婚率之預測值比較

保留期數	實際值	轉換模式	失業率模糊分類法轉換模式
2007	0.6549	0.6545	0.6515
2008	0.6619	0.6710	0.6636
2009	0.6819	0.6873	0.6748
MSE		0.000037	0.000022

表3.6為利用表3.4之失業率預測區間再運用門檻轉換模式所得未婚率區間預測值。由表3.5、表3.6可知，不管是在先以失業率做模糊分類，再將分類結果與未婚率做轉換模式。和以失業率區間做模糊分類，再將其分類結果與未婚率做轉換模式，所得之未婚率之預測值或預測區間。所得知MSE或IMSE差異性並不大，

主要的原因是因為樣本數過少所導致。但卻提供了一方法給於未婚率有效的預測區間，更能說明未婚率的不確定性。

表3.6 未婚率轉換模式之預測區間

保留期數	實際值	未婚率轉換模式之區間預測
2007	0.6549	(0.653,0.656)
2008	0.6619	(0.667,0.675)
2009	0.6819	(0.681,0.695)
IMSE		0.000129

整題來看我們做的區間預測值不管是在失業率或是未婚率都有不錯的結果。如果運用分類方法轉換模式結果會更好，只需將失業率做分類出來的後段資料對未婚率做模式建構。此時所需資料期數較少(只需西元 1998 年後的失業率資料)，及在區間預測中較可求得未婚率之預測區間，可對預測未婚率之不確定性給於更好的預測能力。

最後也得到了，一時間數列若能找出其結構轉變的轉折期間，對其模式建構與預測能力能有更好的結果。

第四章 結論

過去學者在時間序列上對於研究結構改變所使用的分析資料多著重於單時點資料。首先檢定資料是否呈現穩定狀態，再進一步對變數進行分析與預測。其中，因學者欲分析方向不同，而各有不同的研究方法，如變數的選擇、欲求出轉折點或者是多個結構改變區間等等。但對於找出區間時間數列之轉折區間並無多加深入研究。為此，本篇論文提出了一有效方法來探討此一問題。但是我們共同的目標都是希望所研究出來的結果是具有效率及預測能力的，能符合現實生活中的現象，並且希望能夠藉由對時間序列在對結構改變分析經驗中，找出一有效的步驟來檢定其轉折區間。進一步用來預測，以達到人類渴望對於未來變化的掌握。

近年來台灣失業率有攀升的現象，台灣社會所面臨的挑戰，除了千禧年首次政黨輪替帶來的政治變遷與「民主陣痛期」的動盪外，也隨著國際經濟環境變化、兩岸經貿環境的改變。使台灣的投資及就業市場更越來越艱困，形成國內就業市場之供需與衝擊。因此，找出失業率結構性改變的轉折區間，才能更有效的預測失業率之未來走向。

未婚率的提高導致出生率的降低，年齡結構的改變。受失業率的影響所涵蓋的範圍甚大，本篇論文單就影響未婚率之走勢來做探討。再找出失業率之轉折區間後，便能有效的找出結構性改變後的失業率對未婚率之影響與預測。目前探討單身人口或失業率的相關文獻很多。但同時探討兩者之間關係與孤鸞年影響的文獻並不多。而本篇論文亦探討，找出失業率的轉折區間後，運用轉換模式與介入模式來建構單身人口比率的動態模式，以敘述單身人口比率與失業率、孤鸞年之間的動態關係。並應用無母數檢定方法對多重介面干擾時點進行偵測分析。結果顯示孤鸞年對單身人口數並無顯著影響。最後並對失業率對單身人口數做一個合理的預測。

本研究找出許多更值得探討的議題，以作為後續的研究：

1. 區間聚落中心是否有更好的找法，以便更能描述區間之間互相的關係。
2. 若將隸屬度改為一函數型，例：Z-type,、Λ -type,、pi-type,和 S-type，本論文之研究方法要如何做調整。
3. 在不同門檻值 λ 及顯著水準α下，其轉折區間有何不同？即要如何定義一更有效的門檻值 λ 及顯著水準α，找出更有意義的結構性改變的時間。

第五章 參考文獻

中文部分

[1]吳柏林 1995 時間數列分析與導論 台北 華泰書局.

[2]吳柏林 2005 模糊統計導論方法與應用 台北 五南書局.

[3]吳柏林 1999 模糊統計分類在臺灣地區失業率分析與預測之應用 中國統計學報 37:1,37-52.

[4]黃士滔 2004 台灣地區失業率預測分析 工程科技與教育刊 ,1:2,257-269.

[5]胡愈寧 2004 整合時間序列資料與總體經濟變數於失業率預測之應用 育達學院學報,139-170.

[6]許永河 1998 台灣地區自然失業率之估計 成功大學學報,33,125-158.

[7]劉浩天 2004 非時變模糊時間數列預測模式之研究 管理研究學報,69-189.

[8]楊靜利 2006 臺灣傳統婚配空間的變化與婚姻行為之變遷 人口學刊. 33 ,1-32.

[9]張弘紋 2010 應用模糊多屬性群體決策方法於研發專案之選擇 專案管理學刊,3:1,74-90.

[10]陳嘉甄 2009 以模糊聚類方法分析數學錯誤概念組型例 教育研究與發展期刊,5:4,159-186.

[11]林原宏 2005 模糊集群 教育研究,138,142-143.

英文部分

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

[2] Y. Yoshinari, W. Pedrycz, K. Hirota.(1993) Construction of fuzzy models through clustering techniques,*Fuzzy Sets and Systems*, 54, 157-165.

[3] A.F.Gómez-Skarmeta, M.Delgado and M.A.Vila,(1999). About the use of fuzzy clustering techniques for fuzzy model identification, *Fuzzy Sets and Systems*,106, 2,179-188 .

[4] Jiulun Fan and Weixin Xie.(1999) Distance measure and induced fuzzy entropy, Fuzzy Sets and Systems, 104:2 , 305-314.

[5] Ioannis K. Vlachos, George D.(2007) Sergiadis Subsethood, entropy, and cardinality for interval-valued fuzzy sets—An algebraic derivation *Fuzzy Sets and Systems*,158,1384-1396.

[6] Michael P. Windham (1981). Cluster validity for fuzzy clustering algorithms *Fuzzy Sets and Systems*, 5:2, 177-185.

[7] Andrews, D. W. K. and Ploberger, W.(1994) Optimal tests when a nuisance parameter is present only under the alternative, *Econometrica*, 62(6), 1383-1414.

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

[9] Kumar, K. and Wu, B. (2001), Detection of change points in time series analysis with fuzzy statistics, *International Journal of Systems Science*, 32(9), 1185-1192.

[10] Zhou, H. D. (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*.

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

[12] San, O.M., Huynh, V., and Nakamori, Y. (2004), An alternative extension of the K-means algorithm for clustering categorical data, Int. J. Appl. Math. *Comput. Sci*,14:2, 241-247.

[13] Weina Wang, Yunjie Zhanga(2007). On fuzzy cluster validity indices, *Fuzzy Sets and Systems* ,58,2095–2117.

[14] Malay K. Pakhira, Sanghamitra Bandyopadhyay, Ujjwal Maulik(2005). A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. *Fuzzy Sets and Systems*, 15:2, 191-214.

[15] Dae-Won Kim, Kwang H. Lee, Doheon Lee(2004). On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition*, 37: 10, 2009-2025.

[16] Kuo-Lung Wu, Miin-Shen Yang(2005). A cluster validity index for fuzzy clustering.*Pattern Recognition Letters,* 26:9, 1275-1291.

[17] Gin-Shuh Liang, Tsung-Yu Chou, Tzeu-Chen Han(2005). Cluster analysis based on fuzzy equivalence relation, *European Journal of Operational Research*, 166: 1,160-171.

[18] Wenyi Zeng, Hongxing Li(2006). Relationship between similarity measure and entropy of interval valued fuzzy sets.*Fuzzy Sets and Systems*, 157: 11, 1477-1484.

[19]Jia-Chun Xie.Berlin Wu.Songsak Sriboonchita(2010). Fuzzy Estimation Methods and their Application in Real Estimation Evaluation.*Internation Journal of Intelligent technique and application statistics.*(will application)