

探討隱藏馬可夫模式序率模擬月流量系列之 影響因子

Exploring factors of stochastic simulation of monthly streamflow series using the hidden Markov model

國立成功大學水利及海洋工程學系
碩士班研究生
高嘉梅
Chia-Mei Kao

教授
蕭政宗
Jenq-Tzong Shiau

摘 要

台灣地區年平均降雨量約為世界平均值的 2.6 倍，但由於地狹人稠以及降雨空間與時間分布不均，且因全球氣候變遷影響而導致供水的可靠度下降，因此水資源管理顯得更加重要。透過具歷史資料統計特性之流量隨機模擬模式，可繁衍多組水文時間序列，而將這些序列輸入到水資源系統模擬模式中，得以評估系統設計、水庫營運決策的性能和可靠性。因此本研究使用隱藏馬可夫模式(hidden Markov model)進行月流量序率模擬，探討不同影響因子包括資料轉換、狀態數、繁衍組數等，對序率模擬結果準確性的影響。本文選用位於臺灣北部區域蘭陽溪流域的蘭陽大橋測站 1950 到 2021 年的觀測月流量資料為分析案例，分別將原始月流量資料與以對數轉換過之月流量資料建立隱藏馬可夫模式，改變不同狀態數及繁衍組數以繁衍與實際流量等長之月流量序列。繁衍流量以百分誤差(%difference)和相對平均絕對誤差(relative mean absolute difference, RMAD)分析與實際流量之差異程度。研究結果顯示，當繁衍組數越多時，所得到的誤差值越小，而其中又以經對數轉換後四個狀態之隱藏馬可夫模式為最佳，未轉換之原始數據四個狀態之隱藏馬可夫模式次之，而未轉換且為兩個狀態之隱藏馬可夫模式結果為最差。

關鍵詞：隱藏馬可夫模式、序率模擬、月流量

Abstract

The annual average rainfall in Taiwan is approximate 2.6 times that of the world average. However, due to the dense population, spatio-temporal uneven distributed rainfall, and the reduced reliability of water supply caused by global climate change, water resource management is even more important than before. Stochastic simulation model can be used to generate long streamflow series, which possesses the statistical properties of historical data. The reproduced streamflow series provides alternatives instead of historical data to evaluate

performance and reliability of water-resources systems. This study employs the hidden Markov model (HMM) to stochastic simulate the monthly streamflow series and explore the factors influencing the simulated results. The observed monthly streamflow data from 1950 to 2021 at the Lan-Yang Bridge station in the Lan-Yang River basin are selected as the case study in this study. The original and logarithmically transformed monthly streamflow data are used to establish HMMs, respectively. Different state numbers and generation numbers are used to simulate the monthly streamflow series with the same length of the historical data. The simulated results are evaluated by the percentage difference and the relative mean absolute difference (RMAD). The results show that declined errors are observed with increasing generation number. Among HMMs, the performance of the HMM with log-transformed data and 4 states is the best. The HMM with original data and 4 states ranks second, while the HMM with original data and 2 states has the worst performance.

Keywords: hidden Markov model, stochastic simulation, monthly streamflow

一、前言

水資源是民生經濟及工業發展的重要資源，台灣地區年平均降雨量約為世界平均值的 2.6 倍，可說是相當的豐沛，但因地狹人稠、降雨季節及空間分布不均、陡峭的地形使得河川流量快速入海，再加上近年來全球氣候的異常，使得實際可用的水資源相當有限。除了要減少河川水質汙染及水庫淤積的問題外，往往更需要以水工設施進行水流的調節與儲蓄，並配合水資源的運用規劃進行管理。因此藉由分析歷史流量資料，進而了解流量的相關特性，甚至是建立流量之序率模擬模式，都將有利於水資源的運用及管理，使水資源的永續利用得以落實。

流量時間序列的估計對於河流工程和管理項目的設計和實施至關重要，早期在建立水文逕流模式多是藉由蒐集集水區地形及地表相關因子，但其物理因子蒐集不易且費時。而近十年有越來越多研究理論使用序率模式模擬水文資料，因所需資料容易取得且有計算時間較短的優勢，其中以河川流量隨機模擬中最常見的方法之一是使用自迴歸移動平均模式(autoregressive moving average model, ARMA)，ARMA 模式是一種使用自迴歸和移動平均多項式分析平穩線性的時間序列數據方法，線性相加的模式簡單且易於應用，但其隨機模擬時間序列模式無法掌握狀態轉變的特徵，因此可能會對乾旱風險評估和水資源管理技術產生影響。本研究使用的替代序率模式為隱藏馬可夫模式(hidden Markov model, HMM)，HMM 含有隱藏狀態，具有辨識型態的能力，對應至水文領域，以隱藏狀態代表乾與濕季，使得繁衍水文系列具有更穩固的物理意義，可改善 ARMA 無法明確定義乾濕的缺陷。HMM 最初應用在語音辨識(Rabiner, 1989)、生物資訊學領域等應用並且獲得顯著的成果。而隱藏馬可夫模式首次的水文應用是 Zucchini and Guttorp (1991) 在降雨數據的模擬，近年有 Akintug and Rasmussen (2005)、Gelati, et al. (2010)等應用於水文流量模擬。

Pender, et al. (2016)分別運用 HMM 與 ARMA 模式模擬日流量序列，結果顯示 HMM 明顯優於 ARMA 模式。因此本文選用臺灣地區流量站資料建立隱藏馬可夫模式進行月流量序率模擬，探討隱藏馬可夫模式在不同影響因子包括資料轉換、狀態數、繁衍組數等，對序率模擬結果準確性的影響為研究目標，進而選擇最好的影響因子配置運用在月流量序率模擬，透過此研究檢視此模式之適用性與發展性。

二、研究區域與使用資料

本文選取臺灣北部區域蘭陽河流域中的蘭陽大橋測站作為研究對象，以該站 1950 年到 2021 年的觀測月流量資料進行歷史流量資料分析，以下為該流域簡介以及特性說明。

蘭陽溪原名宜蘭濁水溪，發源於南湖大山北麓，與周圍馬當溪及米磨登溪匯集後成為蘭陽溪，全長 73 公里，流域面積約為 978 平方公里，坡陡流急，主流平均坡度為 1/21。其主要支流為宜蘭河、羅東溪、大礁溪、小礁溪及大湖溪等，蜿蜒於雪山山脈與南湖大山山脈之間，由西向東流至三星鄉破布烏始入平地，溪流成網狀流流至蘭陽大橋附近形成幹流，再向東北流經東港附近，最後匯集後注入太平洋。流域內平均年雨量約為 2433 毫米，每年夏秋受到颱風帶來降雨，冬季首當其衝的受到東北季風挾帶的豐沛水氣影響，使得全年無明顯乾季。而根據各流量測站資料得知蘭陽河流域之河川流量以 10 月為最高，4 月最低，其中 2 到 5 月為蘭陽流域之枯水期，然而其蘊藏大量地下水及伏流水之水資源利用價值，使其在 2020-2021 年乾旱得以安然度過。蘭陽溪的污染程度雖不高，但應多加維護以供應宜蘭地區之畜牧、家庭及工業所需之足夠水資源。

蘭陽河流域選用蘭陽大橋測站，資料長度 71 年，其基本資料如表 2-1 所示。而測站月流量統計特性，最大月平均流量測得為 $543.92 \text{ m}^3/\text{s}$ 、最小值 $0.79 \text{ m}^3/\text{s}$ 及平均流量為 $62.23 \text{ m}^3/\text{s}$ ，如表 2-2 所示。

表 2-1 本研究所選流量站基本資料表

站名	站號	流域	河流	資料長度(年)	西元
蘭陽大橋	2560H006	蘭陽溪	蘭陽溪	71	1950-2000, 2002-2021

表 2-2 蘭陽大橋流量站各月份流量統計特性

	單位	一月	二月	三月	四月	五月	六月	七月	八月	九月	十月	十一月	十二月
平均值	(m^3/s)	37.77	33.59	27.74	23.34	32.67	53.41	45.97	68.52	118.2	151.37	93.23	61.43
標準差	(m^3/s)	21.85	22.65	16.7	12.98	23.75	37.77	36.79	57.66	79.97	107.24	59.09	40.12
偏度係數		1.8	1.76	1.61	1.61	2.54	1.6	1.82	1.5	1.16	1.33	1.27	1.82
最大值	(m^3/s)	131.67	119.67	91.7	80.37	159.41	193.6	181.03	260.02	360.17	543.92	281.65	208.04
最小值	(m^3/s)	10.73	8.26	1.21	0.79	2.82	10.23	7.61	10.03	15.81	18.69	12.78	8.36

蘭陽大橋流量站月平均流量分布圖如圖 2-1 所示。

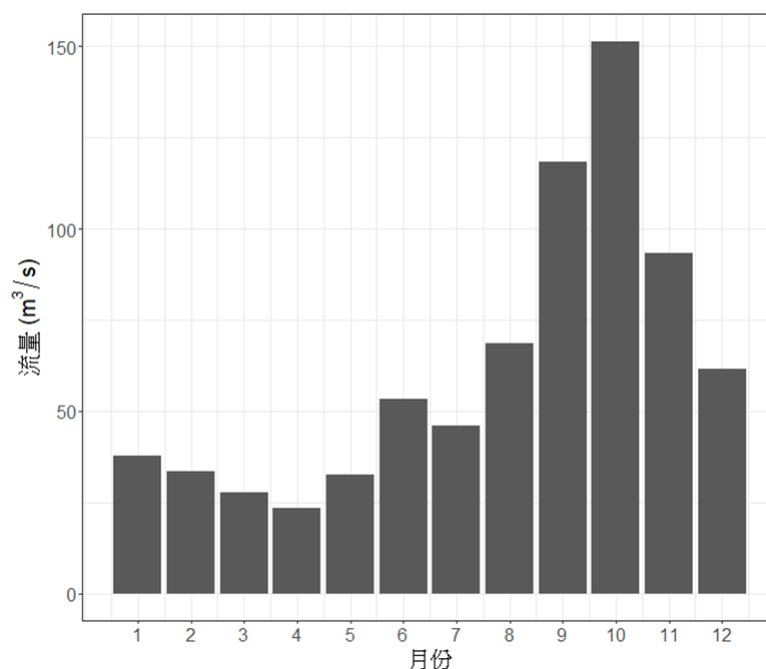


圖 2-1 蘭陽大橋流量站各月份平均流量分布圖

三、研究方法

3.1 隱藏馬可夫模式

Baum and Petrie (1966)首次介紹隱藏馬可夫模式，在建模中除了觀察到狀態隨時間的演化外，也允許藉由潛在的隱藏狀態來繁衍序列。隱藏馬可夫模式包含兩種序率過程，第一個序率過程為隱藏狀態系列，另一個為觀測值系列。於一階馬可夫鏈時， t 時刻的隱藏狀態只與 $t-1$ 時刻有關，滿足馬可夫鏈性質；每個隱藏狀態對應一個連續分布，從中可得到觀測到的時間序列值， t 時刻觀測值只與 t 時刻狀態有關，HMM 示意圖如圖 3-1 所示。

$$\Pr(S_t|S_{t-1})=\Pr(S_t|S_{t-1}) \quad (1)$$

$$\Pr(X_t|X_{t-1}, S_t)=\Pr(X|S_t) \quad (2)$$

上式中 S_t 表示 t 時刻的隱藏狀態， X_t 表示 t 時刻所對應的觀測值。隱藏馬可夫模式由三個參數組成，包括初始機率(Initial probability) π 、 $M \times M$ 之轉移機率矩陣(Transition probability matrix) A 以及發射機率矩陣(Emission probability matrix) B ，可將三個參數寫為：

$$\lambda=[\pi, A, B] \quad (3)$$

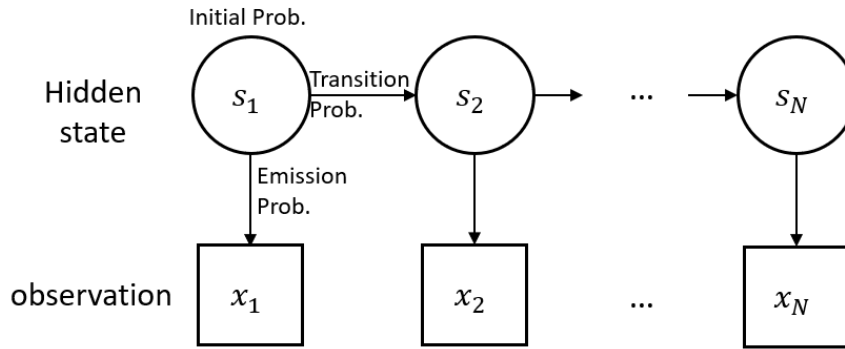


圖 3-1 隱藏馬可夫模式示意圖

3.1.1 參數估計

估計隱藏馬可夫模式參數的方法有很多種，其中常見的有 Thyer and Kuczera (2003) 中使用的貝氏估計(Bayesian estimation)，及 Baum, et al. (1970)所提出的最大期望演算法 (Expectation Maximization algorithm, EM)，又稱 Baum-Welch 演算法等。本文使用最大期望演算法(EM)求得 HMM 參數 λ ，其包含初始機率、轉移機率矩陣以及發射機率矩陣。

3.1.2 序率模擬

在得到隱藏馬可夫模式參數後，需要透過序率模擬方式繁衍與觀測流量資料等長之模擬流量，並且分析各統計特性以檢驗隱藏馬可夫模式對原始流量之還原程度。本研究繁衍模擬流量序列之步驟如下：

1. 利用最大期望演算法獲得隱藏馬可夫模式參數 $\lambda=[\pi, A, B]$
2. 初始機率配合轉移機率矩陣轉換產生狀態序列
3. 將每個狀態所對應的分布參數隨機抽樣，以得到一次與原始等長之模擬流量序列
4. 不斷重複上述 2 及 3 步驟，直到繁衍所需次數之流量序列

3.2 影響因子

不同影響因子的設定可能會帶來不同的結果，因此本研究探討在不同影響因子的組合下，對序率模擬結果準確性的影響。其中影響因子包括資料處理、狀態數及繁衍組數這三種做分析與討論。

3.2.1 資料處理

資料處理的目的多是為了使資料更接近所使用的機率分佈，因此在本文中將資料處理比較分為先將資料做轉換與不轉換兩種方法的模擬結果比較，其中轉換使用的是常見的對數轉換方法，將資料以自然對數進行轉換，如下：

$$y^{(l)} = \ln y \quad (4)$$

上式中 $y^{(l)}$ 為對數轉換後資料； y 為原始資料。而不轉換即直接使用 y 值資料進行隱藏馬可夫模式之序率模擬流程。

3.2.2 狀態數

隱藏馬可夫模式之狀態數決定是很重要的，Thyer and Kuczera (2000) 以研究地區受乾濕季氣候的影響，選擇兩狀態進行研究；Pender, et al. (2016) 則先將流量分 11 組，且每組對應特定狀態，針對高流量用極端值分佈去做改善，但狀態數過多時，可能會使過度擬合的問題發生。本研究為探討狀態數的多寡是否能改善隱藏馬可夫模式序率模擬結果的優劣，分別指定隱藏馬可夫之狀態數為 2 及狀態數為 4，並分析兩者的結果差異。

3.2.3 繁衍組數

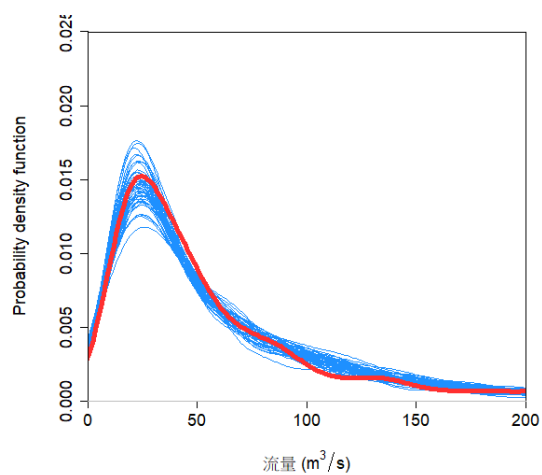
為了測試模擬模式在真實流量序列的適用性，將合成序列的統計數據與記錄值的統計數據進行比較。Pender, et al. (2016) 繁衍了 100 組與歷史流量等長度的流量序列，使其可以對真實數據和模擬數據之間的相關性進行統計分析，以評估可信度；Bracken, et al. (2014) 則繁衍了 1200 組與歷史流量等長度的流量序列。本文以繁衍 100 組為基準，選擇了較少的 50 組繁衍組數以及較多的 200 組組數做比較，以判定影響因子中繁衍組數多寡的差異是否會在評估可信度及統計特性時造成嚴重的影響。

四、結果與討論

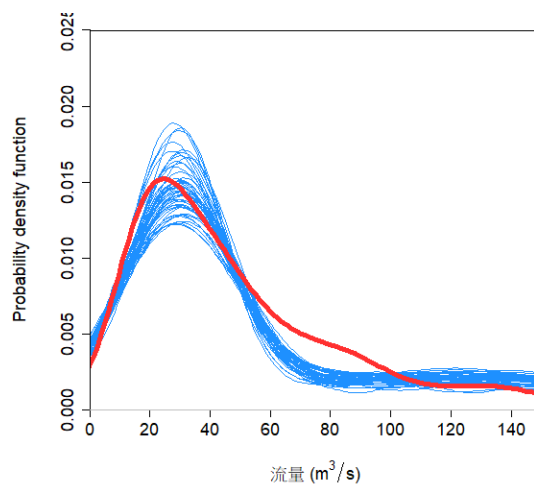
本研究以隱藏馬可夫模式，建立臺灣北部地區之蘭陽溪流域蘭陽大橋流量站之月流量繁衍模式，根據推估所得參數，設定不同影響因子，指定 2 及 4 兩種隱藏馬可夫模式之狀態數，將經由對數轉換及無轉換之原始流量數據分別依照模擬流程各繁衍 50 組和 200 組流量系列，並且將繁衍結果與實際觀測資料進行比較，最後根據研究結果，決定較優影響因子組合之模式。

4.1 模擬流量系列之機率密度圖

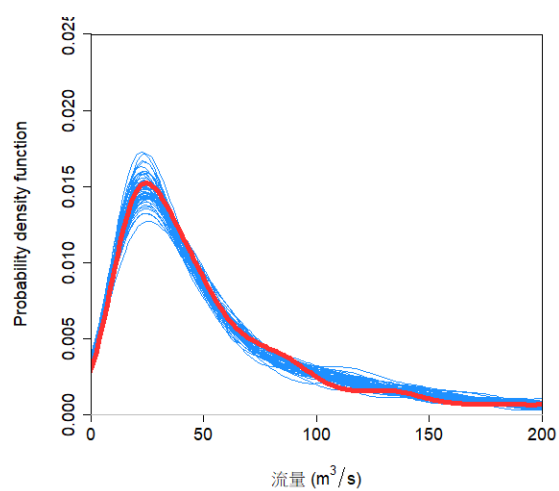
圖 4-1 與圖 4-2 分別為隱藏馬可夫模式所繁衍 50 組和 200 組模擬流量系列個別的 2 和 4 狀態及對數轉換和無轉換的機率密度函數圖。對數轉換不論是繁衍 50 組還是 200 組所繁衍流量機率密度函數(藍色)皆相當接近於實測資料之機率密度函數(紅色)，而以 2 個狀態且資料無轉換之繁衍流量系列最不接近於實測流量資料。



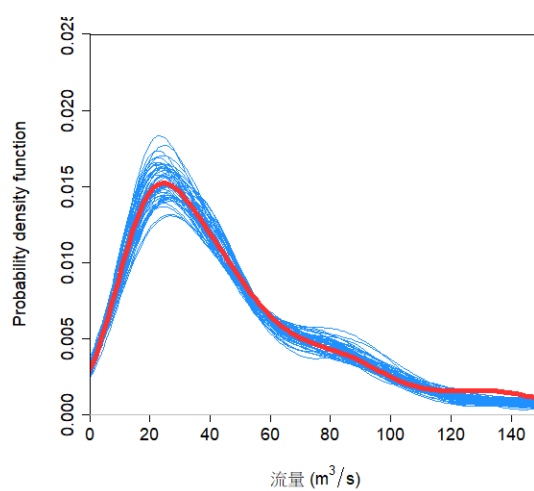
(a)



(b)

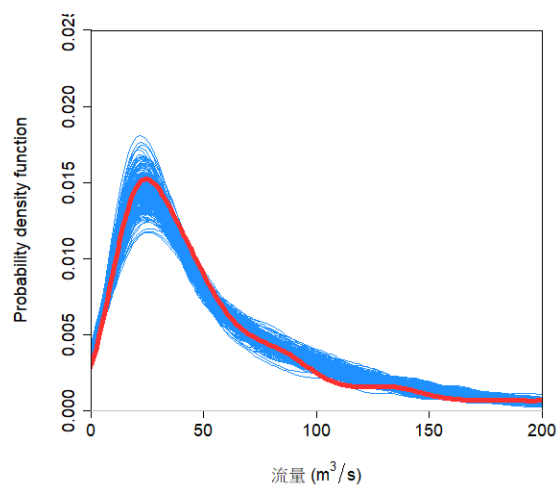


(c)

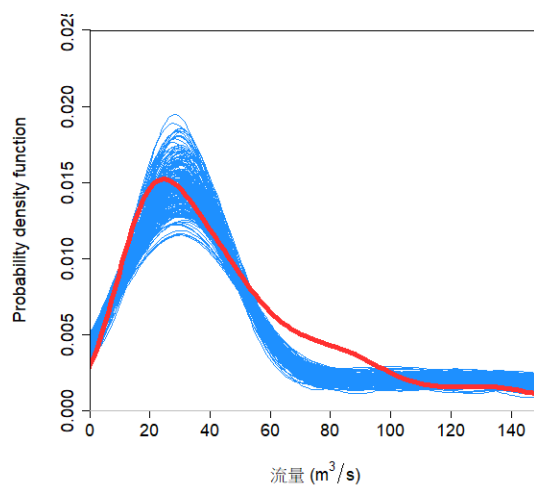


(d)

圖 4-1 HMM 繁衍 50 組流量與實際流量之機率密度函數圖，(a)對數轉換資料具 2 狀態，(b)無轉換資料具 2 狀態，(c)對數轉換資料具 4 狀態，(d)無轉換資料具 4 狀態



(a)



(b)

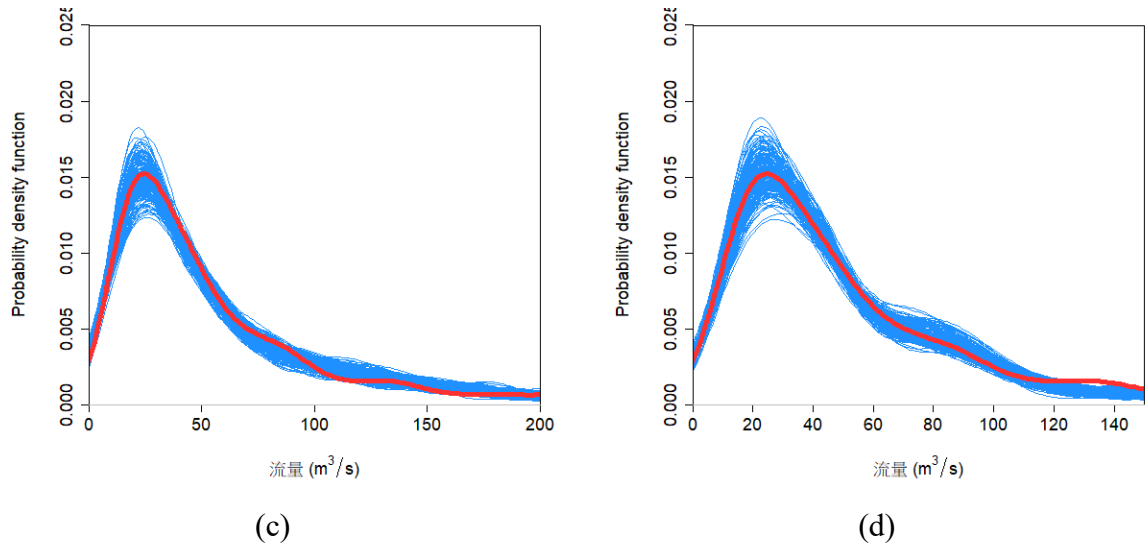
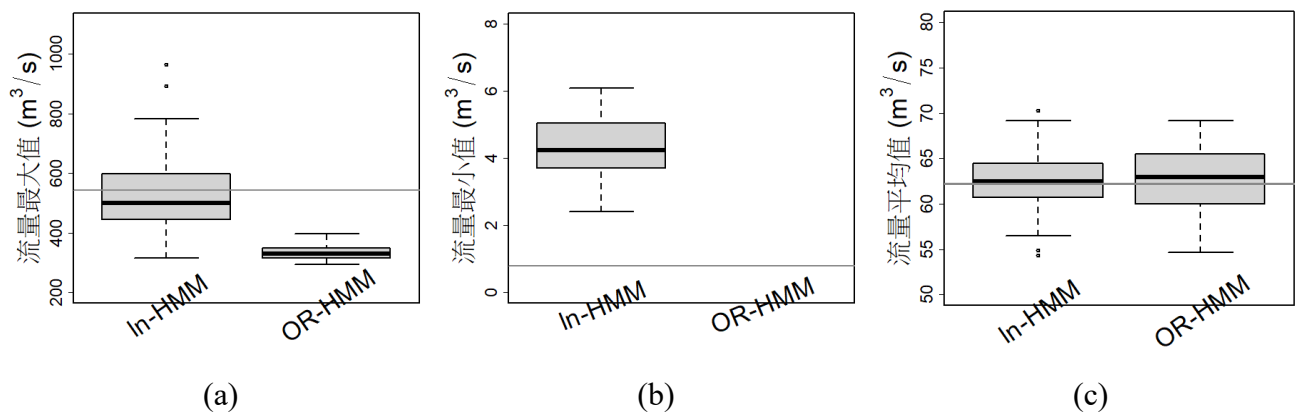


圖 4-2 HMM 繁衍 200 組流量與實際流量之機率密度函數圖，(a)對數轉換資料具 2 狀態，(b)無轉換資料具 2 狀態，(c)對數轉換資料具 4 狀態，(d)無轉換資料具 4 狀態

4.2 模擬流量系列之基本統計量圖

將序率模擬之流量與歷史觀測流量做比較，本研究以最大值、最小值、平均值、標準差及偏態系數五種基本統計量做檢驗，分別呈現 50 組(圖 4-3 及圖 4-4)和 200 組(圖 4-5 及圖 4-6)中兩種狀態的盒狀圖。其中模擬的平均值大多都接近歷史平均值；而最大值與最小值在無轉換之 2 狀態 HMM 結果皆低估很多；標準差可表示資料離散程度，從以下結果顯示模擬之標準差與歷史標準差皆相當接近；偏態系數則是用來評估資料的不對稱性，偏態系數為負代表機率密度函數左側的尾部比右側的長，絕大多數的值位於平均值的右側，偏態系數為正則反之，而偏態系數在無資料轉換的序率模擬結果皆低估。



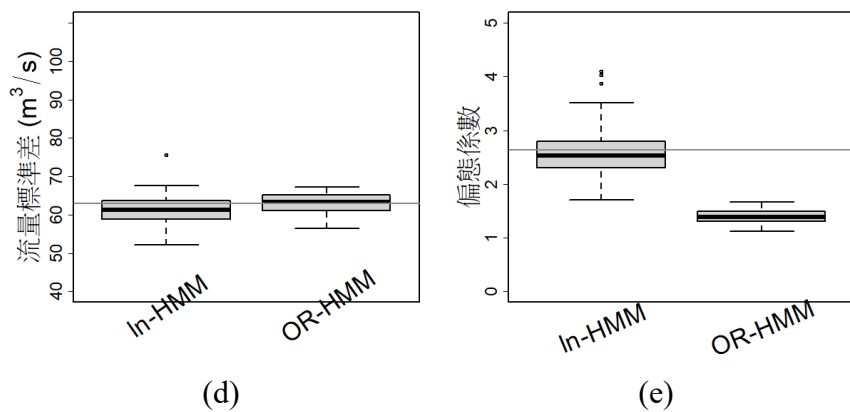


圖 4-3 2 狀態 HMM 繁衍 50 組流量盒狀圖，(a)流量最大值，(b)流量最小值，(c)流量平均值，(d)流量標準差，(e)流量偏態係數

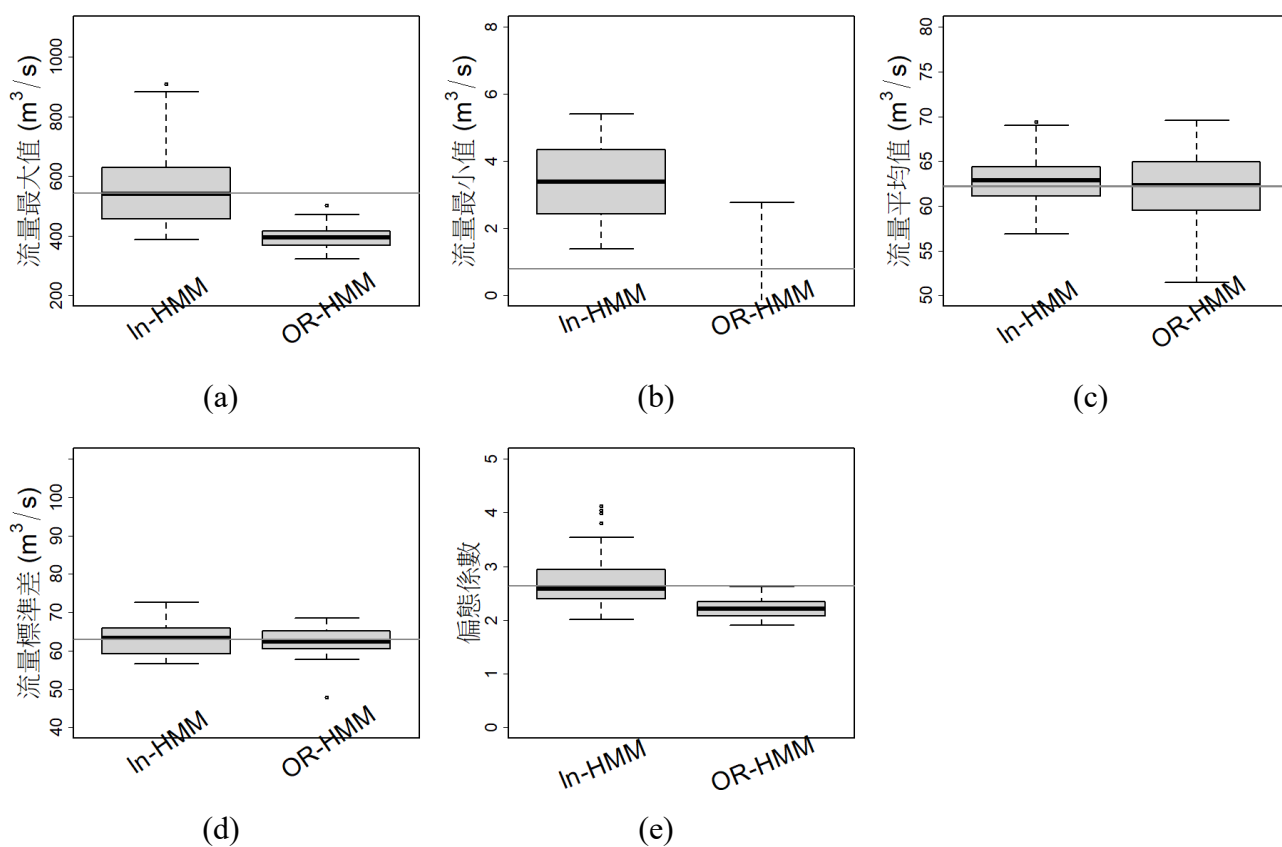


圖 4-4 4 狀態 HMM 繁衍 50 組流量盒狀圖，(a)流量最大值，(b)流量最小值，(c)流量平均值，(d)流量標準差，(e)流量偏態係數

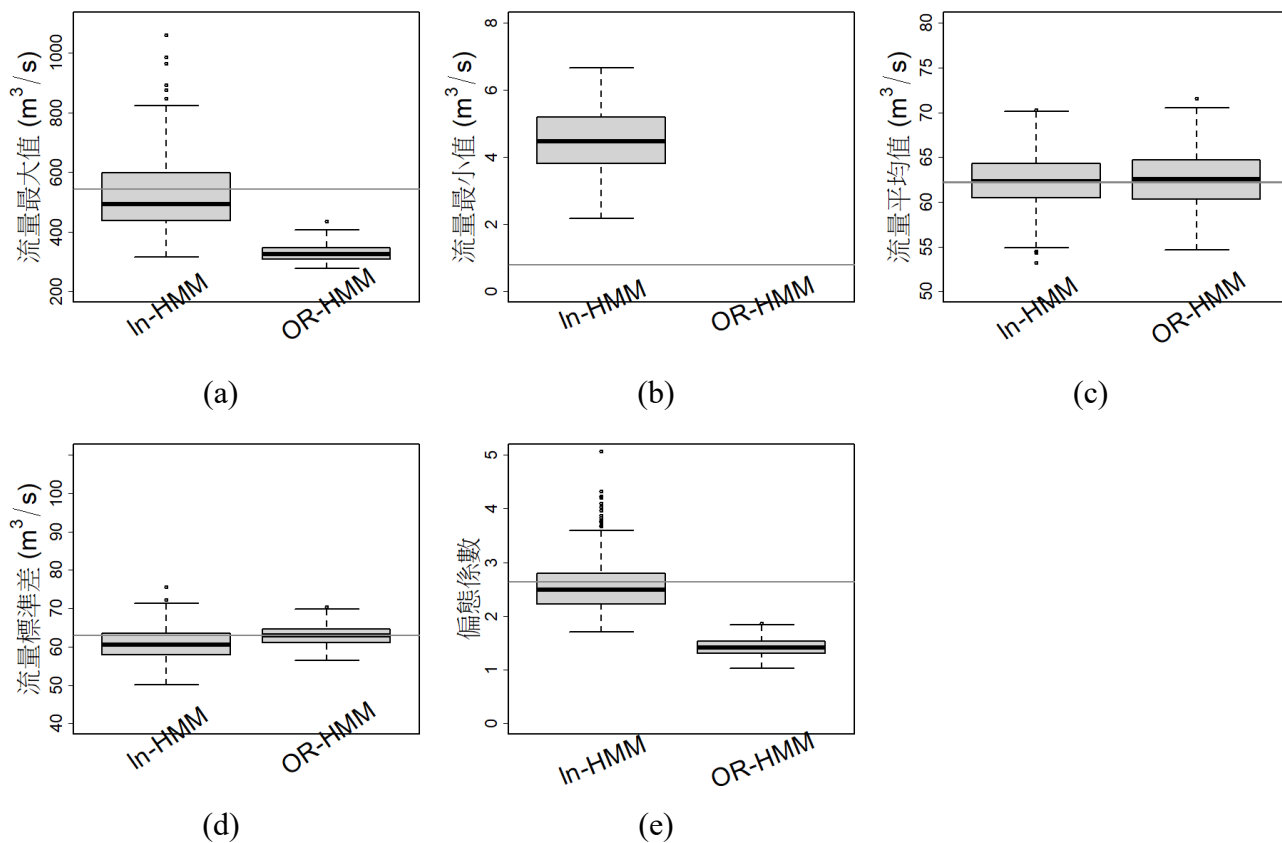
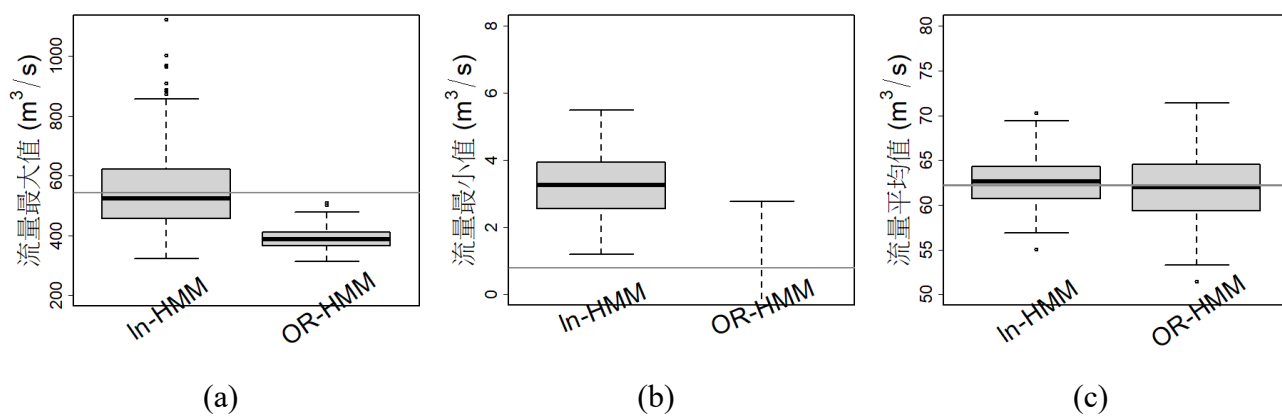


圖 4-5 2 狀態 HMM 繁衍 200 組流量盒狀圖，(a)流量最大值，(b)流量最小值，(c)流量平均值，(d)流量標準差，(e)流量偏態係數



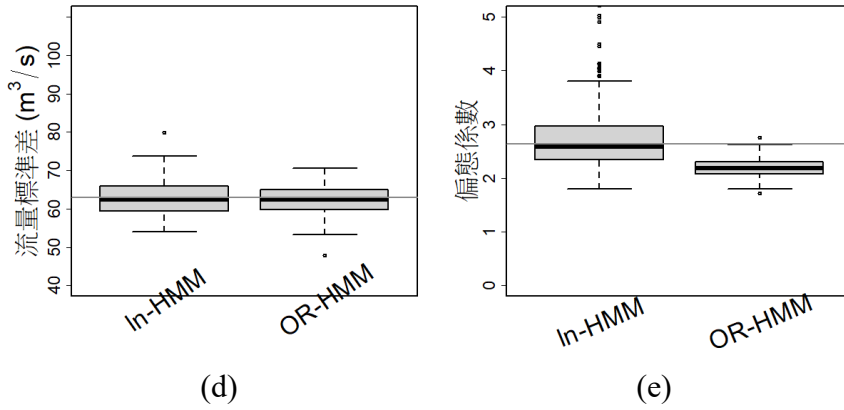


圖 4-6 4 狀態 HMM 繁衍 200 組流量盒狀圖，(a)流量最大值，(b)流量最小值，(c)流量平均值，(d)流量標準差，(e)流量偏態係數

4.3 誤差評估指標

本研究以蘭陽大橋測站之 71 年月流量資料，使用隱藏馬可夫模式並透過改變三種影響因子方式分析序率模擬結果之優劣，因此需要評估指標作為判定準則。而本研究使用的是百分誤差(%difference)及相對平均絕對誤差(relative mean absolute difference, RMAD)做為比較標準，其定義如下：

$$\%difference = \frac{mean(Q_{p,synthetic}) - Q_{p,recorded}}{Q_{p,recorded}} \quad (5)$$

$$RMAD = \sum_{p=1}^n \left| \frac{mean(Q_{p,synthetic}) - Q_{p,recorded}}{Q_{p,recorded}} \right| \times \frac{100\%}{n} \quad (6)$$

上式中 $Q_{p,synthetic}$ 為模擬流量序列第 p 分位數之流量； $mean(Q_{p,synthetic})$ 為模擬產生的 50 個或 200 個流量序列第 p 分位數流量平均值； $Q_{p,recorded}$ 為實際流量序列第 p 分位數之流量數值； n 則為分位數之取點各數，本研究 $n=100$ 。

圖 4-7 取繁衍數量 50 組為例，分別比較 2 狀態及 4 狀態下有無資料轉化的誤差程度，而由下圖可看出流量資料以對數轉換之隱藏馬可夫模式不論是 2 狀態或 4 狀態，在第 10 分位至第 90 分位之間與實際觀測流量誤差範圍皆小於正負 10%；流量資料無轉換之隱藏馬可夫模式在狀態 2 時低流量比實際流量低估很多，且在第 80 分位附近高估達到 30%，但將狀態數增加為 4 個時，誤差有明顯的改善，在第 10 分位至第 90 分位之間與實際觀測流量誤差範圍也皆小於正負 10%。

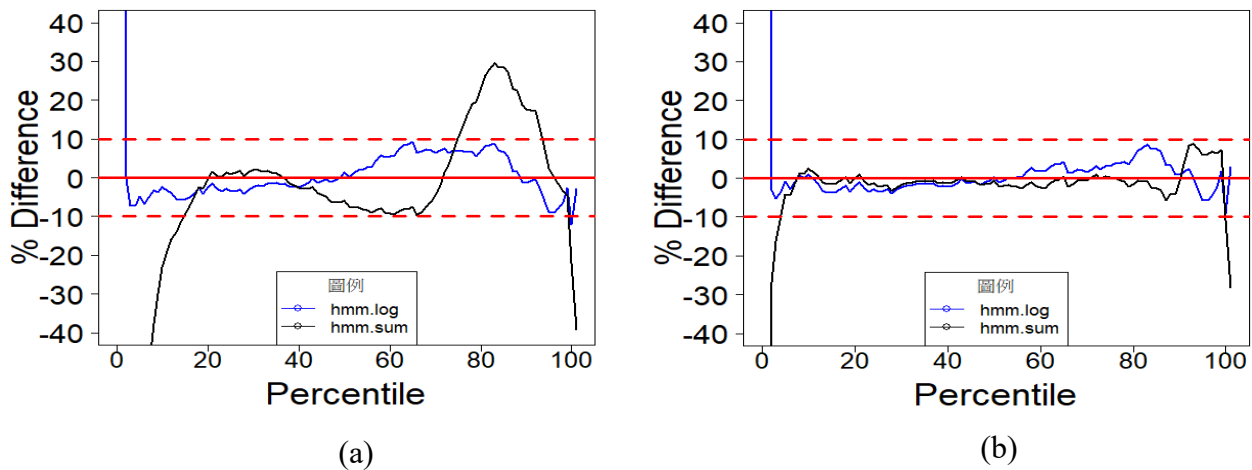


圖 4-7 繁衍 50 組流量之百分誤差(藍色為對數轉換後 HMM；黑色為無轉換後 HMM，
(a)2 狀態，(b)4 狀態)

4.4 模式比較

本研究分析不同影響因子設定下對隱藏馬可夫模式序率模擬的能力，其中包括狀態數的多寡、資料有無處理以及繁衍數量，將不同組合與實際測量資料間的誤差以 RMAD 表示，進而比較選出最佳影響因子配置。

表 4-1 及表 4-2 分別為繁衍 50 組及 200 組流量序列對應不同狀態數和資料處理之 RMAD 誤差值，結果顯示所有組合下皆是以對數轉換方式優於無轉換；其中當狀態數提升時，誤差也會隨之下降，且在無轉換時差距最為顯著；而隨著繁衍組數提升，整體誤差也會有些許改善。比較優劣：繁衍 200 組之 4 狀態對數轉換>繁衍 50 組之 4 狀態對數轉換>繁衍 200 組之 4 狀態無轉換>繁衍 50 組之 4 狀態無轉換，繁衍 50 組之 2 狀態無轉換最差。

表 4-1 蘭陽大橋測站繁衍 50 組流量與實際流量之 RMAD 平均值

	2 狀態	4 狀態
對數轉換	4.266	2.739
無轉換	19.253	2.789

表 4-2 蘭陽大橋測站繁衍 200 組流量與實際流量之 RMAD 平均值

	2 狀態	4 狀態
對數轉換	4.245	2.641
無轉換	18.980	2.785

五、結論

本研究使用隱藏馬可夫模式對臺灣北部蘭陽大橋流量站進行月流量序率模擬，目的探討不同影響因子包括資料轉換、狀態數及繁衍組數，對 HMM 序率模擬結果準確性的影響，而經由計算分析得到以下幾點結論。

1. 當狀態數過少時，若流量資料無先經過處理(如對數轉換)，會導致過大的誤差；而隨著狀態數的增加，資料有無處理進行序率模擬的誤差結果相距不大。
2. 狀態數的增加有助於減少誤差。
3. 繁衍組數的提升能減少誤差但效果有限。
4. 由百分誤差趨勢圖顯示流量在高低極端值的模擬結果較差，常有高估或低估的現象。
5. 以相對平均絕對誤差(RMAD)指標結果顯示，影響因子在繁衍組數設定為 200 組、狀態數為 4 個並且資料先經過對數轉換的資料處理，所得到的誤差最小。

謝誌

本研究承科技部經費補助(MOST111-2221-E-006-042-MY2)，謹致謝忱。

參考文獻

1. Akintug, B., and Rasmussen, P. "A Markov switching model for annual hydrologic time series." Water Resources Research **41**(9), (2005).
2. Baum, L. E., and Petrie, T. "Statistical inference for probabilistic functions of finite state Markov chains." The Annals of Mathematical Statistics **37**(6): 1554-1563, (1966).
3. Baum, L. E., Petrie, T., Soules, G. and Weiss, N. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains." The Annals of Mathematical Statistics **41**(1): 164-171, (1970).
4. Bracken, C., Rajagopalan, B. and Zagana, E. "A hidden Markov model combined with climate indices for multidecadal streamflow simulation." Water Resources Research **50**(10): 7836-7846, (2014).
5. Gelati, E., Madsen, H. and Rosbjerg, D. "Markov-switching model for nonstationary runoff conditioned on El Niño information." Water Resources Research **46**(2), (2010).

6. Pender, D., Patidar, S., Pender, G. and Haynes, H. "Stochastic simulation of daily streamflow sequences using a hidden Markov model." Hydrology Research **47**(1): 75-88, (2016).
7. Rabiner, L. R.. "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE **77**(2): 257-286, (1989).
8. Thyer, M. and Kuczera, G. "Modeling long-term persistence in hydroclimatic time series using a hidden state Markov Model." Water Resources Research **36**(11): 3301-3310, (2000).
9. Thyer, M. and Kuczera, G. "A hidden Markov model for modelling long-term persistence in multi-site rainfall time series 1. Model calibration using a Bayesian approach." Journal of Hydrology **275**(1-2): 12-26,. (2003)
10. Zucchini, W. and Guttorp, P. "A hidden Markov model for space-time precipitation." Water Resources Research **27**(8): 1917-1923, (1991).