

水文頻率分析—核方法

Hydrological Frequency Analysis — Kernel Method

淡江大學水資源及環境工程研究所教授

虞 國 興
Gwo-Hsing Yu

淡江大學水資源及環境工程研究所研究助理

何 志 軒
Chih-Hsuan Ho

摘 要

本研究主要探討 Parzen(1962) 所提核方法之適用性，並與傳統頻率分析中常用之分布 (NOR、LN2、LN3、EV1、PT3 及 LPT3) 做一比較。研究中亦探討五種不同核函數之優劣，並應用核方法分析台灣地區年最大一日暴雨量。

研究結果顯示，核方法所得結果與以傳統方法且正確選定資料所屬之機率分布所得之精確度為差，此乃核方法無須推估資料所屬機率分布所須付之代價，但因兩者所得之精確度差異不大，顯示核方法之適用性；五種不同核函數之整體表現亦無顯著差異，應用時可任意選用；應用於台灣地區年一日最大暴雨量之結因與傳統方法中之最佳分布 LPT3 等差不多。然而由於核方法，不須涉及資料所屬機率分布之判定、資料轉換及偏態係數修正等，應用上較傳統方法為簡便，更具彈性，故核方法值得加以推廣應用於水文頻率分析上。

關鍵詞：頻率分析，無母數，核方法。

ABSTRACT

The major objective of this present study is to investigate the aptness of Kernel method proposed by Parzen (1962). The results of Kernel method and those of obtained by traditional frequency analysis by using Normal, 2-parameter Log Normal, 3-parameter Log Normal, Extreme Value Type I Pearson Type III and Log Pearson Type III distributions are compared. The performance of five different Kernel functions are also investigated in this present study. Besides, the annual maximum 1-day rainfall data in Taiwan area are employed to study the aptness of Kernel method for real data.

The results indicate that the accuracy of Kernel method is worse than those of obtained by the traditional frequency analysis when the probability density function of data obeyed is accurately detected. This is the price Kernel method must pay because it does not need to know the probability density function. However, the differences between these two methods are not significant. This result strongly suggests that Kernel method be quite apt for frequency

analysis. The performance of five different Kernel functions used in the present study are the same. Thus, the choice of Kernel function can be arbitrary. Meanwhile, for the annual maximum 1-day rainfall data, the results of Kernel method are not significantly different with those of obtained by the most appropriate distribution --LPT3. Because Kernel method does not need to identify the probability density function, data transformation or modify the estimated skewness coefficient, it is much easier and flexible than traditional frequency analysis in practice. Therefore, Kernel method is worth of applying in the hydrological frequency analysis.

Keywords : Frequency analysis, Non-parameter, Kernel method.

一、前言

頻率分析係利用過去水文資料，藉由所模擬之機率分布，推估未來水文事件發生之頻率，其於水資源規劃與設計上扮演極重要之角色。一般在推估機率分布上大致可分為二種方法。

(1)參數密度函數估計 (Parametric density function estimation)

(2)無母數密度函數估計 (Nonparametric density function estimation)

傳統頻率分析屬參數密度函數估計，水文上常用之機率分布，計有常態分布 (NOR)、二參數對數常態分布 (LN2)、三參數對數常態分布 (LN3)、極端值 I 型分布 (EV1)、皮爾遜Ⅲ型分布 (PT3) 及對數皮爾遜Ⅲ型分布 (LPT3) 等。由於傳統頻率分析方法涉及機率分布之研判、參數推估之方法選定（動差法及最大概似法）、點繪法公式之選定及資料之繁瑣處理過程（如資料是否需要轉換及偏態係數是否需要修正等）等。因此，當一暴雨事件，欲判斷其發生之頻率時，往往因分析方法之不同而導致不同結果。同時，若分布使用不當，導致估計水文事件之偏差，造成決策上之困擾。

晚近在推求無母數密度函數估計之方法有多種不同方法，如：核方法 (Kernel method)、正交級數法 (orthogonal series method)、最大概似法 (maximum-likelihood method)、直方圖概似法 (histogram-like method)、k 階鄰近法 (k-nearest neighbor method) 等。Parzen (1962) 提出一無母數核密度函數估計 (Non parametric Kernel density function estimation) 的方法，係利用鄰近資料點加權平均估計原始機率密度函數。Abramson(1982) 進一步提出可變核函數密度

估計方法，將不同的樣本觀察點給予不同之參數值。Adamowski(1985) 首次將此觀念應用在頻率分析上，並以點繪法公式推求平滑參數 h ，說明此種方法能適用在洪水頻率分析。核方法應用於頻率分析時僅需推求單一參數，進而推估機率分布，其在分析資料上不須做任何轉換及修正，更不必判斷機率分布，且大多數分布均能以核方法表示，因此較傳統方法方便，更具彈性。

由於目前國內在頻率分析上，尚無利用核方法進行分析研究，因此本研究之主要目的在於了解核方法應用於台灣地區資料頻率分析之適用性，並與頻率分析中常用之機率分布作一比較與探討，研究其中之差異。

本研究主題如下：

(1)核方法於之適用性，並與傳統方法比較。

(2)核方法應用於台灣地區年最大一日暴雨量之適用性。

本文大綱如下：第一節為前言，第二節為理論基礎，第三節為本研究所使用之資料，第四節為結果與討論，最後，第五章為結論。

二、理論基礎

2-1 核方法：

假設有一獨立隨機樣本 $X_1, X_2, X_3, \dots, X_n$ ，樣本數為 n ，具機率密度函數 $f(X)$ ，與分配函數 (Distribution function) $F(X)$ ，Rosenblatt(1956) 提出一權重函數推估方法。

定義：

$$F_n(X) = \frac{X_1, X_2, \dots, X_n \text{ 小於等於 } X \text{ 之個數}}{n}$$

則推估函數為：

表 1 本研究所採用五種核函數

核 函 數		k (·)	
K(1)	Uniform	0.5 0	t ≤1 其他
K(2)	Normal	$\frac{1}{\sqrt{2\pi}} \cdot \exp(-\frac{1}{2}t^2)$	
K(3)	Triangular	1- t 0	t ≤1 其他
K(4)	Biweight	$\frac{15}{16} (1-t^2)^2$ 0	t ≤1 其他
K(5)	Epanechnikov	$\frac{3}{4\sqrt{5}} (1-\frac{1}{5}t^2)$ 0	t ≤√5 其他

2-2 參數之選取：

假設有一理論分布 f 及其推估函數 $\tilde{f}$

則：

$$MISE(\tilde{f}) = E \int [\tilde{f}(x) - f(x)]^2 dx \dots\dots\dots (4)$$

由式(4)得

$$MISE(\tilde{f}) = \int [bias_h(x)]^2 dx + \int var(f(x)) dx \dots\dots\dots (5)$$

式中：

$$\begin{aligned} bias_h(x) &= E[\tilde{f}(x)] - f(x) \\ &= \int \frac{1}{h} K\left\{\frac{x-y}{h}\right\} f(y) dy - f(x) \\ &= \int K(t) f(x-ht) dt - f(x) \end{aligned}$$

以 $y = x - ht$ 代入上式

$$\begin{aligned} &= \int K(t) f(x-ht) dt - f(x) \\ &= \int K(t) \{f(x-ht) - f(x)\} dt \end{aligned}$$

利用泰勒級數 (Taylor series) 展開

$$f(x-ht) = f(x) - h \cdot t \cdot f'(x) + \frac{1}{2} \cdot h^2 \cdot t^2 \cdot f''(x) + \dots$$

則

$$\begin{aligned} bias_h(x) &= -hf'(x) \int tK(t)dt + \frac{1}{2}h^2f''(x) \int t^2K(t)dt + \dots \\ &= \frac{1}{2}h^2f''(x)k_2 + O(h^2) \end{aligned}$$

$$\text{因此 } \int bias_h(x)^2 dx \cong \frac{1}{4}h^4k_2^2 \int f''(x)^2 dx \dots\dots\dots (6)$$

又

$$\begin{aligned} var(f) &= \frac{1}{n} \cdot \int \frac{1}{h^2} \cdot K\left(\frac{x-y}{h}\right)^2 f(y) dy - \frac{1}{n} \left[\int K\left(\frac{x-y}{h}\right) f(y) dy \right]^2 \\ &= \frac{1}{nh} f(x) \cdot \int K(t)^2 dt \end{aligned}$$

$$\text{則 } \int var(\tilde{f}) \cong \frac{1}{nh} \int K(t)^2 dt \dots\dots\dots (7)$$

將式(6)與式(7)代回式(5)得

$$\begin{aligned} \tilde{f}_n(X) &= \frac{F_n(x+h) - F_n(x-h)}{2h} \dots\dots\dots (1) \\ &= \frac{NO. of \{x_1, x_2, \dots, x_n\} \leq x+h - NO. of \{x_1, x_2, \dots, x_n\} \leq x-h}{2nh} \\ &= \frac{1}{2nh} [NO. of \{x_1, x_2, \dots, x_n\} \leq x+h - NO. of \{x_1, x_2, \dots, x_n\} \leq x-h] \\ &= \frac{1}{2nh} \sum [S(x+h) - S(x-h)] \\ &= \frac{1}{2nh} \sum u(x-h, x+h) \end{aligned}$$

式中：S (·) 為階梯函數 (step function)

u (·) 為均勻函數 (uniform function)

其以一臨時性假設權重函數

$$W(t) = \begin{cases} \frac{1}{2} & -1 \leq t \leq 1 \\ 0 & \text{其他} \end{cases}$$

則式(1)可改寫為：

$$\tilde{f}_n(x) = \frac{1}{nh} \sum_{j=1}^n W\left(\frac{x-x_j}{h}\right) = \frac{1}{n} \sum_{j=1}^n W_n(x-x_j) \dots\dots\dots (2)$$

式(2)中

$$W_n(\cdot) = \frac{1}{h} W\left(\frac{\cdot}{h}\right)$$

Parzen(1962) 依據 Rosenblatt 定義核密度函數 (kernel density function) 為：

$$\tilde{f}_n(x) = \frac{1}{n h_n} \sum_{i=1}^n K\left[\frac{x-x_i}{h_n}\right] \dots\dots\dots (3)$$

式中， h_n ：平滑參數 (Smoothing parameter)

K (·)：核函數 (Kernel function)

n：樣本數

平滑參數 h_n 滿足不列條件：

- (1) 對所有 n 而言 h_n 為恆正。
- (2) $\lim_{n \rightarrow \infty} h_n = 0$
- (3) $\lim_{n \rightarrow \infty} n \cdot h_n = \infty$

核函數 K (·) 滿足下列條件：

- (1) K (·) 為一機率密度函數 $\int K(t) dt = 1$
- (2) K (·) 為一對稱性函數 $\int t \cdot K(t) dt = 0$
- (3) $\int t^2 \cdot K(t) dt = k_2 \neq 0$

核方法推估密度函數，其平滑參數 h_n 之選擇可為單一常數，亦可依資料所在位置不同而有不同的值。本研究係採用單一常數，即 $h_n = h$ 。研究中所探討之五種核函數，如表 1 所示。

$$MISE = \frac{1}{4}h^4 \cdot k_2^2 \int f''(x)^2 dx + \frac{1}{nh} K(t)^2 dt + o(n^{-1}) \quad (8)$$

由式(8)使其前二項最小而定義一最佳平滑參數值 h_{opt} (Scott et al, 1981) :

$$h_{opt} = \alpha(K) \cdot \beta(f) \cdot n^{-\frac{1}{5}} \dots\dots\dots (9)$$

式中：

$$\alpha(K) = \left[\int K(y)^2 dy \right]^{-\frac{1}{5}} \cdot \left[\int K(y) \cdot y^2 dy \right]^{-\frac{2}{5}}$$

$$\beta(f) = \left[\int f''(x)^2 dx \right]^{-\frac{1}{5}}$$

由式(9)可知平滑參數依未知理論機率分布 f 、核函數 $K(\cdot)$ 與樣本數 n 而定，然實際上理論分布無法得知，因此在推求式(9)中 $\beta(f)$ 必須先行以一起始猜值 h_0 推估 $\beta(f)$ ，定義：

$$\tilde{\beta}(h_0) = \left(\int \tilde{f}_0''(x)^2 dx \right)^{-\frac{1}{5}} = \tilde{\beta}(\tilde{f}_0)$$

$$\text{式中：} \tilde{f}_0(x) = \frac{1}{n h_0} \sum_{i=1}^n K\left(\frac{x-x_i}{h_0}\right)$$

$\tilde{f}_0$ 為起始猜值 h_0 與樣本資料所構建之推估機率密度函數，由上述可以利用疊代法 (Iteration Method) 求出平滑參數值。

$$h_i = \alpha(K) \cdot \beta(h_{i-1}) \cdot n^{-\frac{1}{5}} \dots\dots\dots (10)$$

式(10)若收斂則可求出一最佳平滑參數值。

選取平滑參數上，Habbema et al(1974) 各 Duin (1976) 以最大概似交互驗證法 (Maximum Likelihood cross validation) 選取最佳平滑參數。

定義函數 $\tilde{f}_{-i}(X)$ 為樣本資料 x_i 以外各點之密度推估函數。

$$\tilde{f}_{-i}(x) = \frac{1}{n-1} \cdot \frac{1}{h} \cdot \sum_{j \neq i} K\left\{\frac{x-x_j}{h}\right\} \dots\dots\dots (11)$$

定義概似函數：

$$L(x_1, x_2, \dots, x_n | h) = \prod_{i=1}^n \tilde{f}_{-i}(h | x_i) \dots\dots\dots (12)$$

當式(12)為最大時，有最佳參數值。

Bowman(1980) 提出利用損失函數 (loss function) 概念。

$$I(p, q) = \int p(x) \log \left\{ \frac{p(x)}{q(x)} \right\} dx$$

其說明當式(12)最大時有最小之損失函數。

假設理論分布 $f(x)$ 滿足下式：

$$\left| \int f(x) \log f(x) dx \right| < \infty$$

由損失函數概念知 $\int f(x) \log \{f(x) / \tilde{f}_{-i}(x)\} dx \rightarrow 0$ 時，

$$\int |\tilde{f}_{-i}(x) - f_n(x)| dx \rightarrow 0 \quad (\text{Geman, 1981}) \quad .$$

定義：

$$\begin{aligned} \frac{1}{n} \log L_n &= \frac{1}{n} \sum_{i=1}^n \log \tilde{f}_{-i}(X_i) \\ &= \frac{1}{n} \sum_{i=1}^n \log \tilde{f}_{-i}(x_i) \approx \int \tilde{f}(x) \log f_n(x) dx \quad (13) \end{aligned}$$

Chow et al(1983) 證明當式(13)最大時，所選取之參數，滿足損失函數 $\rightarrow 0$ 。

則定義一 $CV(h)$ 函數如下 (Silverman, 1986)：

$$CV(h) = \frac{1}{n} \sum_{i=1}^n \log \frac{\tilde{f}_{-i}(x_i)}{\tilde{f}_{-i}(x_i)}$$

式中 $\tilde{f}_{-i}(X_i)$ 如式(11)所示。

假設一樣本點 x_1 與其他各點間距離之最小距離為 d ，當 $h < d$ 時，對有邊界性之核函數 (如：Uniform kernel)，其核函數值為 0，則使式(11) $\tilde{f}_{-i}(X_i) = 0$ ，則式(14) $CV(h)$ 為 $-\infty$ 。至於對無邊界性之核函數 (如：Normal kernel) 其函數值極小， $CV(h)$ 值亦為很小。然當式(14)最大時則可求得一最佳平滑參數值 h_{opt} 。

圖 1 為 7 個隨機樣本資料，圖中虛線為單一核密度函數，而推估之機率密度函數則由數個單一核密度函數相加而成。平滑參數值太小時如圖 1(a)，其所推估之機率密度函數為一非平滑曲線，而當平滑參數增大時，曲線逐漸平滑如圖 1(b) 及 (c) 所示，因此推估之機率密度函數形狀受平滑參數影響甚大。在選取最佳平滑參數時，利用最大概似交互驗證法可避免在 h 值太小時所造成之非平滑性問題。

本研究採用最大概似交互驗證法以選取最佳參數，並推估機率分布，其步驟如下：

- (1) 將資料以大小順序排列，並決定其資料之全距。
- (2) 由資料之全距決定其平滑參數值之範圍 $[h_1, h_2]$ 。
- (3) 利用格點法切割，代入式(14)求出最大 $CV(h)$ ，進而決定最適平滑參數值。
- (4) 以核方法推估其模擬分布。

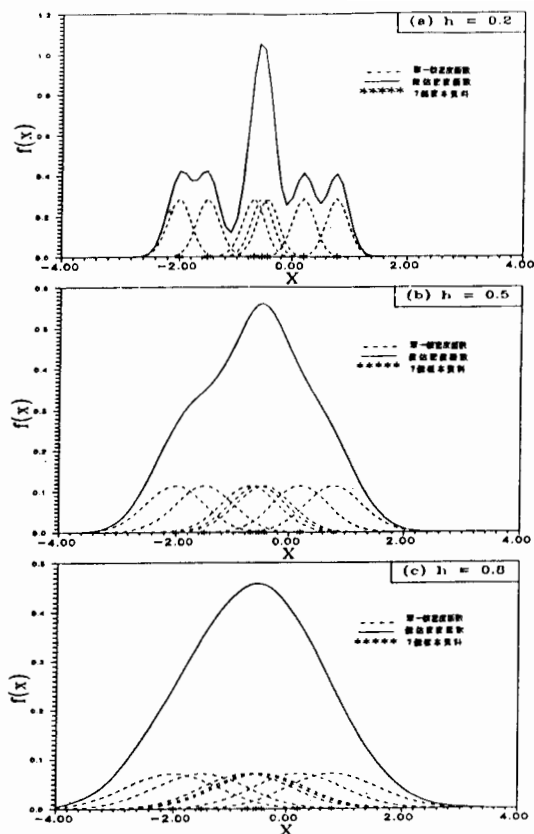


圖 1 不同平滑參數所造成不同平滑特性
[(a)h = 0.2,(b)h = 0.5,(c)h = 0.8]

核方法之計算流程圖如圖 2 所示。

2-3 判斷準則：

2-3-1 合成資料所引用之判斷準則：

本研究利用六種傳統機率分布分別產生合成資料，因其所屬之機率分布為已知，故其各迴歸週期之理論值為已知，為探討核方法於推估機率密度函數之適用性，本研究分別以核方法以及該合成資料所屬之機率密度函數推求其參數，並以 10、20 及 100 年之迴歸週期推估值與理論之差異，探討核方法之適用性。所採用判斷準則計有偏差性(Biasedness)、均方誤差(MSE)、及平均絕對誤差百分比(MAPE, Mean Absolute Percentage Error)。其定義分別如下：

$$s(\tilde{x}_T) = \sqrt{\frac{1}{n-1} \sum (\tilde{x}_T - \bar{\tilde{x}}_T)^2}$$

$$Biasedness = |x_T - E(\tilde{x}_T)|$$

$$MSE = s(\tilde{x}_T)^2 + [x_T - E(\tilde{x}_T)]^2$$

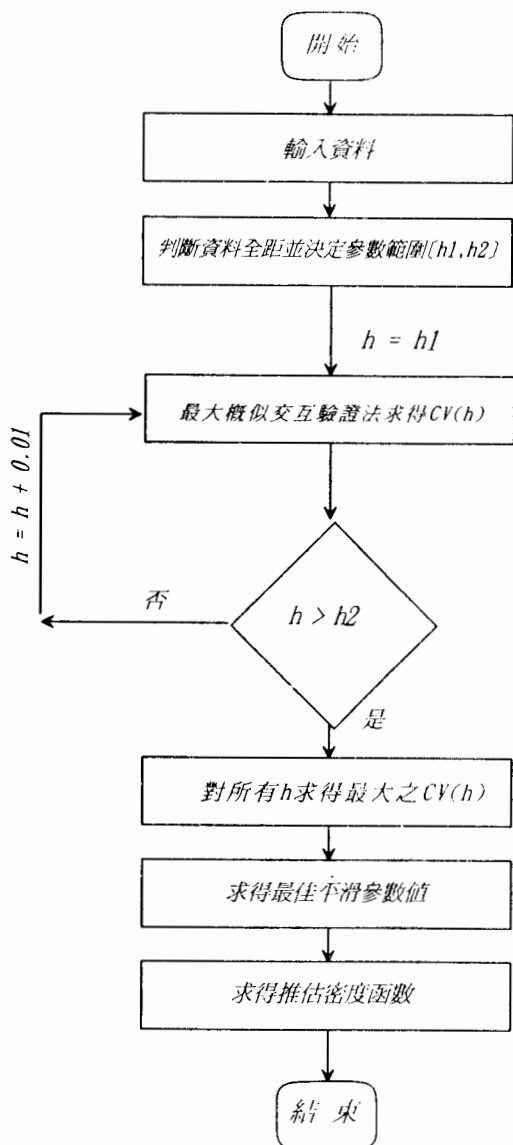


圖 2 核方法計算流程圖

$$MAPE = \frac{1}{n} \sum \frac{|x_T - \tilde{x}_T|}{x_T}$$

其中，

x_T ：為 T 年頻率之理論值

$\tilde{x}_T$ ：T 年頻率之推估值

$\bar{\tilde{x}}_T$ ：T 年頻率推估值之平均值

$E(\tilde{x}_T)$ ：T 年頻率推估值之期望值

$s(\tilde{x}_T)$ ：T 年頻率推估值之標準偏差

2-3-2 實測資料所引用之判斷準則：

由於實測資料之機率分布為未知，無法利用上述合成資料之判斷準則判定何種機率分布較合適於實測資料。本研究採用虞和黃(1992)提出無關機率分布點繪法公式(Distribution Free Plotting Position)：

$$P_j = \frac{j-0.326}{n+0.348} \dots\dots\dots (15)$$

推求各實測資料 x_i 之對應累積機率，再以六種常用分布及以最大概似交互驗證法所得之機率分布推求其推估值 $\tilde{x}_i$ ，並以下列判斷準則判定核方法與傳統方法之優劣。其定義如下：

均方誤差(MSE)。

$$MSE = \frac{1}{n} \sum (x_i - \tilde{x}_i)^2$$

其中，

x_i ：實測資料值

$\tilde{x}_i$ ：推估值

n ：資料個數

三、本研究所使用之資料

3-1 合成資料：

本研究主要目的之一為探討探討核方法之適用性，研究中採用合成資料，因合成資料之理論分布已知。合成資料之基本統計特性如表2所示，其共計產生六種分布，每種分布產生樣本數為30、50、100、200及500五種。每種組合產生100組，故本研究共探討 $6 \times 5 \times 100 = 3000$ 組資料。

表 2 合成資料基本統計特性

分布	平均值	標準偏差	偏態係數	樣本數
NOR	100	25	-----	30,50,100,200,500
LN2	100	25	-----	30,50,100,200,500
LN3	100	25	1.5	30,50,100,200,500
EV1	100	25	-----	30,50,100,200,500
PT3	100	25	1.5	30,50,100,200,500
LPT3	100	25	1.5	30,50,100,200,500

3-2 實測資料：

為探討核方法於實測資料之適用性，本研究分析本省年一日之最大暴雨量，其中北區88站、中區187站、南區208站及東區45站，總計共全省528站。其基本統計特性詳見[1]。

四、結果與討論

由於篇幅所限，本文僅列出部份重要成果，詳細研究成果請參閱(3)

4-1 合成資料：

4-1-1 最小均方誤差法

為探討核方法之適用性，本研究利用合成資料分布已知之特性，以推估分布與理論分布間之最小均方誤差推估平滑參數 h 值。表7、8、9及10所示分別為當樣本數為30、50、100及200時，由六種常用機率分布產生之合成資料以五種核函數推估 h 值之30組平均值及其所對應之均方誤差MSE值。其結果顯示於小樣本時MSE小於 6×10^{-6} 及大樣本時小於 1×10^{-6} 。由此可知以核方法在推求機率分布上並不差，具有相當不錯之精確度。

4-1-2 最大概似交互驗證法與最小均方誤差法之比較

由於實測資料之理論分布為未知，因此最小均方誤差法無法用以推估平滑參數。本研究採用最大概似交互驗證法推求平滑參數，為探討最大概似交互驗證法推求參數之精確度，研究中利用合成資料，比較上述二種參數推估方法間之差異。

表 3 最大概似交互驗證法與最小均方誤差法之結果比較(30組六種合成資料，=200)

核函數	NOR		LN2		LN3		EV1		PT3		LPT3	
	h	MSE	h	MSE	h	MSE	h	MSE	h	MSE	h	MSE
K(1)	5.9	0	5.9	0	4	0	7.5	0	4.8	1	4.8	1
K(2)	2.3	0	2	0	2.1	0	3.7	0	1.9	1	1.7	0
K(3)	7.4	0	6.7	0	7.7	1	9.2	0	6	1	6.9	1
K(4)	7.8	0	7.5	0	7.9	1	9.6	0	6.3	1	7.1	1
K(5)	3.1	0	3.1	0	2.7	1	4.5	0	2.8	1	2.8	1

註：* 單位為 10^{-6}

研究中分別探討當樣本數為 30、50、100 及 200 時，由六種常用機率分布產生之合成資料以兩種參數推估方法於五種核函數下，推估 h 值之 30 組平均偏差值及其所對應之均方誤差 MSE 之平均偏差值 (M.D., Mean deviation)。其結果顯示最小均方誤差法與最大概似交互驗證法所推估之參數所對應之 MSE 間之平均偏差值非常小，約略在 10^{-5} 至 10^{-6} 之間，且當樣本數增大時二者之差異逐漸減小，近乎為 0，(部份結果如表 3 所示)。顯示最大概似交互驗證法推求參數之精確度頗高。

圖 3 所示分別為當樣本數為 30、100 及 500 時，合成資料由 EVI 所產生，並以常態核函數 K [2] 分析，採用最大概似交互驗證法及最小均方誤差法所推估之參數差異比較，其結果顯示當小樣本 ($n = 30$)，最大概似法所推估之參數值較最小均方誤差法大，然於大樣本 ($n = 500$) 時，兩種方法所推估之參數值非常接近。且當樣本數增加時，參數值變小，其滿足理論 h_n 之條件(2)。

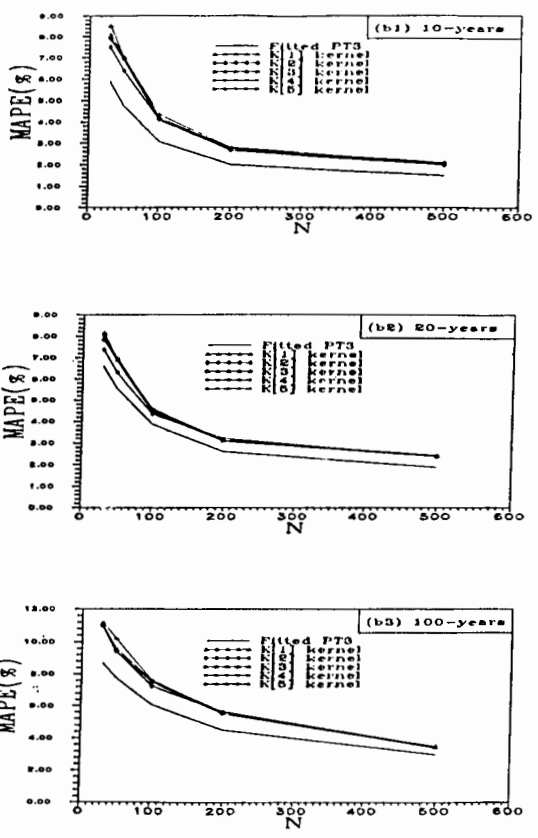


圖 4 五種核函數 MAPE 結果比較 (合成資料：皮爾遜 III 型分布)(b1) 表頻率 10 年 MAPE(b2) 頻率 20 年 MAPE(b3) 頻率 100 年 MAPE

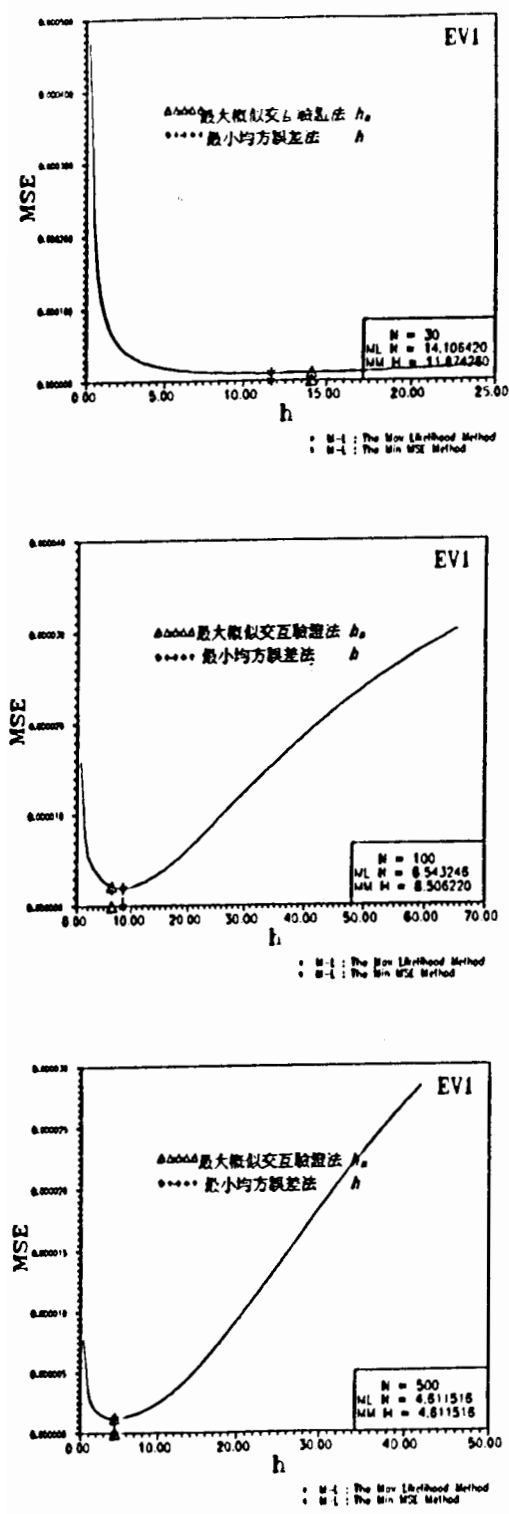


圖 3 最大概似交互驗證法與最小均方誤差法結果比較 (EVI, Normal kernel)

表 4 各種核函數比較(100組，對數皮爾遜Ⅲ型合成資料，n = 500)

核方法			頻率											
核函數	h推估值		10				20				100			
	平均值	標準差	平均值	BIAS	MSE	MAPE	平均值	BIAS	MSE	MAPE	平均值	BIAS	MSE	MAPE
K[1]	11.2	4.4	133.6	1.6	12.3	2.1	148.3	1.6	22.0	2.4	181.7	0.6	91.7	3.9
K[2]	5.9	1.4	133.3	1.3	10.5	1.8	148.0	1.3	21.1	2.4	181.6	0.8	93.9	4.0
K[3]	15.4	5.3	133.5	1.5	12.0	2.0	148.2	1.4	22.3	2.4	181.7	0.7	94.4	4.0
K[4]	16.4	5.5	133.4	1.5	11.6	1.9	148.2	1.4	21.8	2.4	181.7	0.7	94.9	4.0
K[5]	6.5	2.3	133.5	1.6	12.0	2.0	148.3	1.5	21.7	2.4	181.8	0.6	91.6	4.0
傳統方法-LPT3			132.3	0.3	7.5	1.5	146.8	0.1	14.3	2.0	181.4	1.0	57.9	3.3
最佳核函數		理論值	K[2]	K[2]	K[2]	理論值	K[2]	K[2]	K[2]	理論值	K[5]	K[5]	K[1]	
		132.0				146.8				182.4				

4-1-3 五種核函數比較：

合成資料分別由此六種常用分布產生，每種參數組合各產生 100 組，再以合成資料所屬分布與核方法分別推估各迴歸週期之值，以探討五種核函數之優劣，研究中採用偏差性、均方誤差及平均絕對誤差百分比為判斷準則。表 4 所示為當樣本數 $n = 500$ 時，合成資料由 LPT3 分布時之結果。圖 4 所示為合成資料為 PT3 分布時以平均絕對誤差百分比為判斷準則之結果，分析其結果說明如下：

(1) 偏差性：當原始分布為 NOR、LN2 及 EV1，

五種核函數於小樣本 ($N = 30, 50$) 推求 20 年頻率時，偏差性較傳統方法為佳。若原始分布為 LN3、PT3/ 及 LPT3 時，核方法較傳統方法為差。然當大樣本 ($n = 500$) 時，核方法所推估 100 年頻率之偏差性與傳統方法之結果十分接近。當樣本數逐漸增大，偏差性逐漸減少。整體而言，圖 6 至圖 11 顯示中可知五種核函數所得結果差異並不大。

(2) 均方誤差：於小樣本數 ($n = 30, 50$) 時，核方法推估 20 年頻率之 MSE 與傳統方法之結果十分接近，然當 100 年頻率時其差異較大。若原始分布為 LN3、PT3 及 LPT3 時，核方法與傳統方法差異較大，但當樣本數增大時其差異逐漸減少。同時，整體而言，五種核函數所得之結果差異並不大。

(3) 平均絕對誤差百分比：無論在樣本數之大

小，核方法與傳統方法之差異在 2% 之內。整體而言，五種核函數所得結果差異並不大。

4-1-4 點繪法 (Plotting Position) 選取參數之適用性：

Adamowski (1985) 首先應用核方法於水文頻率分析，其以最小均方誤差值推求平滑參數，MSE 之定義如下：

$$MSE = \sum_{i=1}^n [P_j - P_n(x_j)]^2 \quad \text{..... (16)}$$

式中：

P_j ：為超越機率由 Adamowsik(1981) 提出點繪

$$\text{法公式 } P_j = \frac{j-0.25}{n+0.5}$$

$$P_n(x) = \int_x^\infty \tilde{f}(x) dx \quad \text{..... (17)}$$

式中 $P_n(x)$ ：為超越機率 (exceedence probability)

$\tilde{f}(x)$ ：為核函數推估密度分布

將式(3)代入得

$$\begin{aligned} P_n(x) &= \frac{1}{nh} \sum_{i=1}^n \int_x^\infty K\left(\frac{x-x_i}{h}\right) dx \\ &= \frac{1}{nh} \sum_{i=1}^n C_i(x) \quad \text{..... (18)} \end{aligned}$$

式中：

$$C_i(x) = \int_x^\infty K\left(\frac{x-x_i}{h}\right) \quad \text{..... (19)}$$

將式(16)對 h 取一次微分為 0，可得最小 MSE，如下：

$$\sum_{j=1}^n [P_j - P_n(X_j)] \frac{dP_n(X_j)}{dh} = 0 \quad \dots\dots\dots (20)$$

式中：

$$\frac{dP_n(x_j)}{dh} = \frac{1}{nh} \sum_{i=1}^n \left[\frac{dC_i(X_j)}{dh} - \frac{1}{h} C_i(X_j) \right]$$

Adamowski 曾導出 K[1](Uniform Kernel) 核函數所對應之 $C_i(x)$ 函數，本研究亦導出 K[2](Normal Kernel)、K[3](Triangular kernel)、K[4](Biweight Kernel) 及 K[5](Epanechnikov Kernel) 等核函數所對應之 $C_i(X)$ 函數。

然，研究中發現其並無法推求最佳平滑參數 h 值，即為 h 愈小時，其所對應之 MSE 亦愈小。因此，本研究由理論著手並證明當在 h 值趨近零時，式(18)為一點繪法公式（附錄三），因此利用式(16)為判斷準則及式(20)之方法來推求最佳平滑參數之方法並不正確。

表 5 全省 528 站年一日最大暴雨量五種核函數與傳統方法 MSE 結果比較

區域	核函數	NOR	LN2	LN3	EV1	PT3	LPT3
北區	K[1]	52	12	-6	6	-6	-12
	K[2]	78	18	-2	28	-10	-10
	K[3]	70	16	-4	18	-2	-8
	K[4]	70	12	-14	6	-14	-16
	K[5]	62	6	-12	2	-12	-18
中區	K[1]	133	37	27	53	17	13
	K[2]	169	57	47	93	35	23
	K[3]	147	55	47	65	41	31
	K[4]	141	49	37	59	37	25
	K[5]	133	47	29	55	13	9
南區	K[1]	126	28	2	24	0	-8
	K[2]	178	54	12	54	18	-10
	K[3]	162	44	6	40	4	-8
	K[4]	162	38	8	38	2	-16
	K[5]	156	30	12	40	0	-14
東區	K[1]	13	9	-7	7	-7	-5
	K[2]	23	15	-13	3	-11	-19
	K[3]	17	5	-17	3	-17	-19
	K[4]	19	3	-13	-1	-13	-17
	K[5]	23	11	-5	5	-7	-5
全省	K[1]	324	86	16	90	4	-12
	K[2]	448	144	44	178	32	-16
	K[3]	396	120	32	126	26	-4
	K[4]	392	102	18	102	12	-24
	K[5]	374	94	24	102	-6	-28

註：- 表示傳統方法較核方法為佳多出之站數

4-2 實測資料：

表 5 所示係由本省年一日最大暴雨量，以特定核函數與每種分布以傳統方法分析之 MSE 大小相比較，若核方法之 MSE 較小則記 + 1，反之記

- 1。結果顯示於中區任何一種核函數皆較傳統方法六種分布為佳；於北區及東區則核方法較傳統方法中 LN3、PT3 及 LPT3 等三種分布稍差；南區及全省不分區則核方法較傳統方法中 LPT3 稍差。然，就不同核函數之比較，結果顯示五種核函數間之差異不大，此結果與合成資料中所得之結果一致（由於結果表格眾多，僅列出最後分析結果。）

表 6 全省年一日最大暴雨量 K[1] 核函數與傳統方法 MSE 比較總表

判斷準則		北區	中區	南區	東區	全省
K[1]	最佳分布站數	37 (42%)	96 (51%)	100 (48%)	17 (38%)	250 (47%)
	次佳分布站數	4 (5%)	5 (3%)	3 (1%)	2 (4%)	14 (3%)
NOR	最佳分布站數	0	0	0	1	1
	次佳分布站數	0	0	0	1	1
LN2	最佳分布站數	10	17	14	5	46
	次佳分布站數	12	28	27	1	68
LN3	最佳分布站數	5	13	12	7	37
	次佳分布站數	7	29	31	9	76
EV1	最佳分布站數	6	8	10	3	27
	次佳分布站數	8	5	19	2	34
PT3	最佳分布站數	10	25	29	2	66
	次佳分布站數	25	57	68	16	166
LPT3	最佳分布站數	20	28	43	10	101
	次佳分布站數	32	63	60	14	169

註：1 最佳分布站數及次佳分布站數係由核函數與六種傳統方法（共計七種）彼此比較。

2 () 表示與傳統方法比較所佔百分比。

表 6 所示為 K[1] 核函數與傳統方法六種分布應用於本省年最大一日暴雨量時，核方法與六種分布為最佳及次佳分布之站數，研究中採用 MSE 為判斷準則。結果顯示，傳統方法中以 LPT3 分布為最適當，此結果與虞和劉 (1991) 研究發現 LPT3 分布較合適於全省年一日最大暴雨量之結果一致。就全省不分區結果顯示，無論使用何種核函數，選定核方法為最佳分布之站數（次佳分布之站數），約佔全省 528 站中 46 % (5 %)，選定 NOR、LN3、EV1、PT3 及 LPT3 為最佳分布之站數（次佳分布之站數）分別約佔全省 528 站中 0 % (0 %)、8 % (13 %)、8 % (15 %)、6 % (5 %)、11 % (31 %) 及 21 % (31 %)。僅考慮最佳分布之站數時，選定核方法、LN2、LN3、EV1、PT3

及 LPT3 為最佳分布之站數分別為 250, 1, 46, 37, 27, 66 及 101 站 (以 K[1] 為例)。就整體最佳分布之站數排名至最差分布之站數排名共同考慮時, 雖然核方法最佳分布之站數較傳統方法六種分布為多, 但其他排名核方法則較傳統方法中 LPT3 為差。整體而言, 核方法較傳統方法中 LPT3 分布稍差。

五、結 論

- (1) 就合成資料而言, 核方法較以傳統方法且正確選定資料所屬之機率分布所得之精確度為差, 此乃核方法無須推估資料所屬機率分布所須付之代價, 但因兩者所得之精確度差異不大, 顯示核方法之適用性。
- (2) 核方法應用於本省年一日最大暴雨時, 結果顯示於中區時核方法較傳統方法為佳; 北區及東區則傳統方法中 LN3、PT3、及 LPT3 等三種分布較核方法稍佳; 南區及全省不分區而言, 核方法僅較傳統方法中 LPT3 稍差。然, 核方法較傳統方法為簡便, 仍值得推廣使用。
- (3) 五種核函數整體表現彼此差異並不顯著, 應用時可任意選用。
- (4) 最大概似交互驗證法用以推求平滑參數可獲得相當精確之結果。

謝 誌

本研就承蒙行政院國科會 NEC-79-0410-E-032-10 研究計畫之經費補助, 特此致謝。同時, 論文審查者所提供之寶貴意見, 作者由衷感謝。

參考文獻

1. 虞國興、劉振忠, 「台灣地區年一日、二日、三日最大暴雨量及年最大日流量之頻率分析研究」, 農業工程學報, 第 36 卷, 第二期, pp.27-37, 民國 80 年 9 月。
2. 虞國興、黃志強, 「無關機率分布之點繪法公式」, 台灣水利, 第 40 卷, 第三期, pp.22-23, 民國 81 年 9 月。
3. 何志軒, 「水文頻率分析—核方法」, 淡江大學水資源及環境工程研究所碩士論文, pp.78,

民國 81 年 9 月。

4. Abramson, I. S. (1985), "On bandwidth variation in kernel estimation-- a square root law," Ann. Statist. 10, pp.1217-1223.
5. Adamowski, K. (1981), "Plotting formula for flood frequency," Water Resour. Bull., 17(2), pp.197-202.
6. Adamowski, K. (1985), "Nonparametric kernel estimation of flood frequencies," Water Resour. Res., Vol. 21, No. 11 pp.1585-1590.
7. Bowman, A. W. (1984), "An alternative method of cross-validation for the smoothing of density estimation," Biometrika 71 pp.353-360.
8. Chow, Y. S., Geman, S. and Wu, L. D. (1983), "Consistent crossvalidated density estimation," Ann. Statist., 11, pp. 25-38.
9. Duin, R. P. W. (1976), "On the choice of smoothing parameters for Parzen estimators of probability density function," IEEE Trans. Comput., C-25, pp.1175-1179.
10. Geman, S. (1981), "Sieves for nonparametric estimation of density and regression," Report in Pattern Analysis, No. 99, Div of Appl. Math., Brown University.
11. Habbema, J. D. F., Hermans, J. and van der Broek, K. (1974), "A stepwise discrimination program using density estimation," In Buuckman, G. (ed), Compstat Vienna: Physica Verlag, pp. 100-110.
12. Parzen, E. (1962), "On estimation of a probability density Function and mode," Ann. Math. Statist., 33, pp.1065-107613.
13. Rosenblatt, M. (1956), "Remark on some nonparametric estimates of a density function," Ann. Meth. Statist., pp.823-837.
14. Rosenblatt, M. (1971), "Curve estimates," Ann. Meth. Statist., 42, pp. 1815-1842.
15. Scott, D. W., Factor, L. E. (1981), "Monte Carlo study of three databased nonparametric probability density estimators," J. Am. Statist. Assoc., 76, pp.9-15.
16. Singh, R. S. (1980) "Nonparametric kernel estimates of a density function alongwith its derivatives," Colloquia Math. Societatis Janos Bolyai, 32, Nonparametric Statist. Inference, pp.687-711.
17. Silver, B. W. (1986). Density Estimation for Statistics and Data Annlysis. Chapman and Hall, New York.

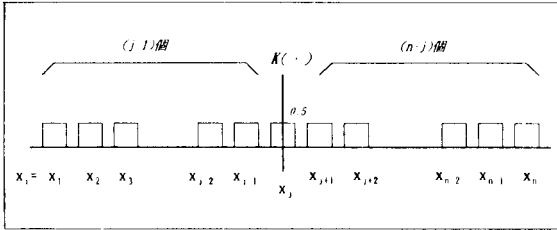
附 錄

證明 $P_n(X_j)$ 爲一點繪法公式

當平滑參數 h 極小，小到使的在 $[X_i - h, X_i + h]$ 範圍內至多只有一個點在其範圍內則，由式(10)之非超越累積機率得：

以 Uniform kernel 爲例，假設：

$$h < |x_i - x_j| \quad | \quad i \neq j \text{ and } K\left(\frac{x - x_i}{h}\right) = 0, \quad \forall \{x \mid x > x_i + h \text{ or } x < x_i - h\}$$



證明示意圖

$$P_n(x_j) = \frac{1}{nh} \cdot \sum_{i=1}^n C_i(x_j)$$

$$= \frac{1}{nh} [C_1(x_j) + C_2(x_j) + C_3(x_j) + \cdots + C_j(x_j) + \cdots + C_n(x_j)]$$

如上圖所示得：

$$= \frac{1}{nh} \cdot \left[\underbrace{h + h + \cdots + 0.5h}_{(j-1)\text{個}} + \underbrace{0 + 0 + \cdots + 0}_{(n-j)\text{個}} \right]$$

$$= \frac{1}{nh} \cdot [(j-1) \cdot h + 0.5h]$$

$$= \frac{1}{n} \cdot \left(j - \frac{1}{2}\right)$$

$$= \frac{(j-0.5)}{n}$$

由上可知當平滑參數小到使之在範圍內無二點以上則該式爲 Hazen 之點繪法公式。

收稿日期：民國 84 年 8 月 20 日

修正日期：民國 84 年 12 月 27 日

接受日期：民國 85 年 1 月 9 日

(上接第96頁)

- fractions in sludge-soil mixtures : 2. Comparative structural chemistry of fulvic acid. Soil Sci. Soc. Am. J. 40:691-697.
15. Syers, J.K., M.G. Browman, G.W. Smillie, and R.B. Corey. 1973. Phosphate sorption by soils evaluated by the Langmuir adsorption equation. Soil Sci. Soc. Am. J. 37:358-363.
16. Van Genuchten, M.Th., and P.J. Wierenga. 1976. Mass transfer studies in sorbing porous media. I. Analytical solutions. Soil Sci. Soc. Am. J. 40:473-480.
17. Voice, T.C., C.P. Rice, W.J. Weber, Jr. 1983. Effect of solids concentration on the sorptive partitioning of hydrophobic pollutants in aquatic systems. Environ. Sci. Technol. 19:789-796.
18. Zasoski, R.J. And R.G. Burau. 1978. A technique for studying the kinetics of adsorption in suspensions. Soil Sci. Soc. Am. J. 42:372-374.

收稿日期：民國 84 年 11 月 5 日

修正日期：民國 85 年 1 月 23 日

接受日期：民國 85 年 2 月 9 日