摘要

本研究的目標為發展一套參數導向(parameter-driven)模式以作為分析具有時間序列特徵的計數資料之用。由於該計數的時間序列資料中包括著一些有相關性結構的潛在過程，因此該模型的結果有著非常複雜的可能性，故本研究將導入改良型 EM(Expectation Maximization)演算法進行分析。本改良型 EM 演算法修正原 EM 演算法的步驟 E，在該步驟中將邊際期望值取代了條件期望值以求得目標值。本研究將估計交通流量資料作為實際案例進行分析。流量資料是藉由架設於路邊的偵測器所收集得到，利用這些資料建立一套模型以作為估計，配適並且預測交通網路中的交通狀況之用，這些資訊將為交通號誌控制或是交通車隊管理的關鍵。

關鍵詞：參數導向模式、蒙地卡羅 EM 演算法、改良型 EM 演算法、計數時間序列資料、交通流量預測

Abstract

This paper describes the methodology for analyzing of parameter-driven models for time series of count data generated from latent process that characterize the correlation structure. These models result in very complex likelihoods. A modified EM algorithm is proposed which we replace the marginal expectation with the conditional expectation given *y* in the *E* step of the EM algorithm. We illustrate our method by traffic flow data. Through the data collected by the detectors mounted on the road, we can estimate, smooth and predict the traffic condition about the network, this information is critical to signal control and traffic queue management.

Keywords: Parameter-driven model; Monte Carlo Expectation-Maximization algorithm; Time series count Data; Traffic flow forecast.

一、研究目的

塞車已為現代人日常生活揮之不去的問題，而都會區尖峰時間之交通瓶頸路口號運作不良，是造成交通癱瘓的問題之一，因此號誌控制系統運作之良窳也是目前交通相關單位之重要課題之一。因此要如何有效的運用號誌控制，首先必須要有預測交通行為的能力，針對交通狀況的改變進行調整，將交通擁塞的情況縮減至最少，一直都是交通工程師的目標。本研究將建立一套模式以預測具有時間序列特徵的計數資料，而流量亦為具有時間序列特徵的計數資料，故本模式將試著建立一套流量預測模式。

二、研究內容

對於時間序列資料這部分，Cox 在 1981 年提出將時間序列模式分成兩類，分別為參數導向(parameter-driven) 與觀察導向(observation-driven) 兩種模式。在觀察導向模式中，觀察物的條件分配是建構在過去的資料上；而參數導向模式中，用一些無法直接觀測到的潛在的程序來描述序列資料彼此的相關性。在處理資料序列與相關問題方面，這兩個方法都越來越普遍被使用。近期，對於參數導向模式的發展與改良有著顯著的進步，首先 Cox 所提出的模式在處理時間序列資料上有很好的效果；此外，Chan 和 Ledolter(1995)提出利用一階自我迴歸高斯過程產生一組觀測值，並利用蒙地卡羅演算法對該觀測值估計以求得一組參數的方法。但是單純利用蒙地卡羅法模擬出一組條件期望值是相當耗費時間的。因此，本研究將介紹更有效率的演算法，內容概述如下：第三章當中說明了模型架構以及改良型的演算法並且建立了兩種平滑與配適資料矩陣的方法；第四章將利用本研究的方法應用於實際交通流量資料；最後在第五章將針對本方法論提出結論。

三、研究方法

3.1 模型

假設 $\{Y_t\}_{t=1}^{N}$ 代表一組非負整數的計數時間序列，這些資料代表實際的反應過程。本研究建立參數導向(parameter-driven)模式以求得該序列與歷史資料之間的關係，並且求得其他由該序列與歷史資料之間的共變異數。如序列的相關係數，而相關係數是由潛在過程(latent process)所產生的。假設 $\{W_t\}$ 是一組由平穩一階自我回歸高斯過程的潛在過程所產生的值。已知潛在過程 $\{W_t\}$、觀測值 $\{Y_t\}$ 為兩組由期望值為 λ_t 的卜瓦松分配所產生的獨立變數，並且關係如下：

$$\log \lambda_t = W_t + \alpha' U_t \qquad (1)$$

其中，α 代表一組 $p\times 1$ 的參數向量; 而 $\{U_t\}$ 代表一組 $p\times 1$ 共變異數向量。因此 $\{W_t\}$ 為一組滿足卜瓦松分配而期望值為 λ_t 的獨立變數，因此 $\{W_t\}$ 的條件式可寫成

$$y_t | \lambda_t \sim Poiss(\lambda_t) \qquad (2)$$

其中 $\{\lambda_t\}$ 為一組滿足 $\lambda_t = \exp(W_t + \alpha' U_t)$ 的變數，且

$$W_t = \rho W_{t-1} + \varepsilon_t) \qquad (3)$$

其中$\{\varepsilon_t\}$為一組滿足 i.i.d. 的常態分配變數$N\left(0,\sigma_t^2\right)$。$\theta$為一組參數向量，包含了在方程式(1)、(3)中的係數，其關係如下$\theta=\left(\rho,\sigma_t^2,\alpha'\right)$。

雖然參數導向模式可以將反應變數與共變異數之間的關係以較簡單且直接的方向呈現，然而該模式比較無法簡易地分析該模式所產生的概似函數。因此現今大多仰賴電腦的能力來處理這類複雜的問題。

3.2 方法論 – 改良型 EM 演算法(The Modified EM Algorithm)

設$X_t=(y_t,W_t)$代表一組具有資料完整的隨機向量，該資料隨時間 *t* 而改變。

$X=(X_1,X_2,...,X_N)$ 代表一組實際獲得的完整資料，

$Y=(Y_1,Y_2,...,Y_N)$ 代表所觀測的資料，

$W=(W_1,W_2,...,W_N)$ 代表潛在過程所包括的資料，可視為遺失資料(missing values)。(y_t,W_t)的聯合機率密度函數可以表示為

$$f_x(y_t,W_t|W_{t-1})=\exp(-\exp(W_t+\alpha'U_t))\times(\exp(W_t+\alpha'U_t))^{y_t}/y_t! \qquad (4)$$

而 **X** 受到W_0的影響所產生的對數概似函數(log-likelihood)可以表示成下式：

$$\begin{aligned} l_x(\theta) &= \sum_{t=1}^{N}\log f_x(y_t,W_t|W_{t-1}) \\ &= \sum_{t=1}^{N}(-\exp(W_t+\alpha'U_t)+y_tW_t+y_t\alpha'U_t)-n\log(\sigma_\varepsilon^2)/2 \\ &\quad -\sum_{t=1}^{n}(W_t+\rho W_{t-1})^2/2\sigma_\varepsilon^2 \end{aligned} \qquad (5)$$

其中W_0代表初始的潛在變數，該組資料滿足常態分配$(W_0 \sim N(0,\sigma_\varepsilon^2/(1-\rho^2)))$ 。

當潛在過程發生時，可以將該組資料視為遺失資料(missing data)，該情形下唯一可觀察求得的資料，即為對數概似函數 $l_x(\theta|y)=\log\int f_x(y,W|\theta)dW$ 。然而欲求得 $l_x(\theta|y)$ 的最大值並不是一件簡單的事，因此本研究引入了 EM 演算法，透過反覆地求取 $l_x(\theta|y)$ 的最大值最後即可得到 $l_x(\theta)$ 最大值。在 EM 演算法中的每一次的重複運算過程皆包含了以下兩個步驟：

E step: From $Q(\theta'|\theta)=E_\theta(l_x(\theta_\cdot)|y)$;

M step: Maximize $Q(\bullet|\theta)$.

$E_\theta(\bullet|y)$ 表示當 $Y=y$ 時的條件期望值；而 θ 為一個真實的參數且 $l_x(\theta')$ 為 X 的對數概似函數。當 $Y=y$ 給定時，欲知 $l_x(\theta|y)$ 需先得到 y 的條件期望值 $E_\theta(\bullet|y)$ ，但由於 y 的條件期望值不易求得，所以本研究進一步利用改良型 EM 演算法進行求解，其調整的方法為利用給定 y 的邊際期望值取代條件期望值。在改良型 EM 演算法運算的過程中，其變動的誤差在我們可以接受的範圍內，則可以停止運算。以下詳敘整個演算法的運算流程：

步驟 1　以給定的初始值 $\theta_0=(\rho_0,\sigma^2_{\varepsilon 0},\alpha_0')'$ 做初始化。

步驟 2　The *E step*: 加入我們所需要的期望值

$$Q(\bullet\,|\,\theta^{(r)})=\sum_{t=1}^{n}(-\exp(\sigma_\varepsilon^{2(r)}/2(1-\rho^{(r)2}))\exp(\alpha'U_t)+y_t\alpha'U_t)-n\log(\sigma_\varepsilon^2)/2 \qquad (6)$$

$$-(\sum_{t=1}^{n}(1-\rho^{(r)2(t+1)})\sigma_\varepsilon^{2(r)}/(1-\rho^{(r)2})-2\rho\sum_{t=1}^{n}\rho^{(r)}\sigma_\varepsilon^{2(r)}/(1-\rho^{(r)2})$$

$$+\rho^2\sum_{t=1}^{n}(1-\rho^{(r)2(t)})\sigma_\varepsilon^{2(r)}/(1-\rho^{(r)2}))/2\sigma_\varepsilon^2\,.$$

步驟 3　The *M step*: 求 $Q(\bullet\,|\theta^{(r)})$ 最大值，即，

$\rho^{(r+1)}$ 和 $\sigma_\varepsilon^{2\,(r+1)}$ 可以構成 $Q(\bullet\,|\theta^{(r)})$ 的最大值，表成

$$\rho^{(r+1)}=[\sum_{t=1}^{n}\rho^{(r)}\sigma_\varepsilon^{2(r)}/(1-\rho^{(r)2})]/[\sum_{t=1}^{n}(1-\rho^{(r)2(t)})\sigma_\varepsilon^{2(r)}/(1-\rho^{(r)2})]\,. \qquad (7)$$

與

$$\sigma_{\varepsilon}^{2(r+1)} = (\sum_{t=1}^{n}(1-\rho^{(r)2(t+1)})\sigma_{\varepsilon}^{2(r)} \big/ (1-\rho^{(r)2}) \qquad (8)$$

$$-[\sum_{t=1}^{n}\rho^{(r)}\sigma_{\varepsilon}^{2(r)}/(1-\rho^{(r)2})]^2 / [\sum_{t=1}^{n}(1-\rho^{(r)2})\sigma_{\varepsilon}^{2(r)}/(1-\rho^{(r)2})])/n$$

$\alpha^{(r+1)}$ 組成 $Q(\bullet|\theta^{(r)})$ 的最大值，且這將用 Iterative Reweighed Least Square (IRLS) 來估計。

步驟 4 若 $|\rho^{(r)}-\rho^{(r+1)}|<error$, $|\sigma_{\varepsilon}^{2(r)}-\sigma_{\varepsilon}^{2(r+1)}|<error$ and $|\alpha^{(r)\prime}-\alpha^{(r+1)\prime}|<error$ ，則停止。

步驟 5 更替參數，使得 $(\rho^{(r)},\sigma_{\varepsilon}^{2(r)},\alpha^{(r)\prime})' = (\rho^{(r+1)},\sigma_{\varepsilon}^{2(r+1)},\alpha^{(r+1)\prime})'$.

步驟 6 回到步驟 2

為了要準確地找到參數的估計，我們計算被估計的資訊矩陣由 $I(\hat{\theta}) = E(-\partial^2 l(\theta)/\partial\theta\,\partial\theta')_{\theta=\hat{\theta}}$ 所給定，且被估計的變異數-共變數矩陣為 $\hat{Var}(\hat{\theta}) = I^{-1}(\hat{\theta})$。

3.3 比對與預測

由 3.1 節的方程式(2)和(3)，以及對任一時間點 t 而言，期望值所求如下：

$$E_{\theta}(\lambda_t|y) = \exp(U_t E_{\theta}) \quad (\text{We}|y) \quad \text{for all } t \qquad (9)$$

因此可以利用 EM 演算法求得參數建立模式。當需要進行比對或是預測資料時，可以利用下列的方程式預測第 $N+l$ 的時間點下的期望值，該值如下式所示：

$$E_{\theta}(\lambda_{N+l}|y) = \exp(U_{N+l}' + -\rho^{2l}(1\sigma_{\varepsilon}^{2}/) - \rho 2^{2}(E1 \quad \rho^{l}W_N(|\text{ex} \qquad (10)$$

當有資料遺失發生時，也可以利用上式進行處理。

本研究提出下列兩個方法對於反應變數進行配適及預測。在方法 1 中，在潛在過程利用非條件期望值做為期望值；而在方法 2 中，利用模擬的結果估計在潛在過程下的條件期望值，以下將詳述這兩個方法的流程。

方法 1 (由 $E_\theta(\lambda_t)$ 來取期望值)

步驟 1 (比對 Y_t , $t=1,\ldots,N$)

$$\hat{Y}_t = \hat{E}_\theta(\lambda_t) = \exp(\hat{\alpha}'U_t + \hat{\sigma}_\varepsilon^2 / 2(1-\hat{\rho}^2)) \quad \text{for } t=1,\ldots,N. \tag{11}$$

步驟 2 (預測 Y_{N+l}, $l \geq 1$)

$$\hat{E}_\theta(\lambda_{N+l}) = \exp(\hat{\alpha}'U_{N+l} + \hat{\sigma}_\varepsilon^2 / 2(1-\hat{\rho}^2)) \tag{12}$$

方法 2 (用方程式(2)和(3)來產生模型)

步驟 1 利用 $W_0 \sim N\left(0, \hat{\sigma}_\varepsilon^2 / (1-\hat{\rho}^2)\right)$ 和 $\varepsilon_t \sim N(0, \hat{\sigma}_\varepsilon^2)$，且令

$$W_t = \hat{\rho} W_{t-1} + \varepsilon_t, \quad t=1,...,N$$

計算出 W_t, $t=1,...,N$ ，再代入上式

步驟 2 利用 $Y_t \sim Poisson(\exp(W_t + \hat{\alpha}' U_t))$ for $t=1,...,N$ 來比對 $Y_t, t=1,...,N$ 。

步驟 3 利用 $\varepsilon_t \sim N(0, \hat{\sigma}_\varepsilon^2)$, 來計算 $W_t = \hat{\rho} W_{t-1} + \varepsilon_t$, for $t = N+1,\ldots,N+l$,

且 $Y_t \sim Poisson(\exp(W_{N+l} + \hat{\alpha}' U_{N+l}))$ ，以此來預測 Y_{N+l}, $l \geq 1$ 。

四、實際案例分析：交通流量

4.1 資料描述

對大多數的都會區而言，每天都有交通擁塞的現象發生，這不僅對於市區交通造成嚴重的問題，對附近鄰近郊區而言也造成相當的困擾。因此為了要使市區路網的運作可以更為順暢，如何紓解擁塞的道路以及預先避免交通的擁塞，交通號誌的控制被視為一個交通網絡不可或缺的元素，號誌控制也是 ITS 的發展中重要議題之一。為了達到更有效率的交通管理，提供更好的服務水準。

以下本研究將分析的流量資料是由 RTMS 偵測器(Remote Traffic Microwave Sensor)所收集得到，觀測地點為新竹市的中華路和警光路相交的十字路口，觀測時間為 2005 年 4 月 25 號下午 14:13:28 到 16:59:57 之間的流量資訊。本案例的目標為研究在尖峰時刻前交通流量的改變現象，該資料是以每 10 秒作一個間隔。由於存在一些塞車或未知的因素所引起的無法偵測到的未知情況，因此本研究把潛在過程納入模型中進行分析。

下列列出模型的記號意義：

Y：交通流量，一段時間通過的車輛數。

u： 佔有率(Occupancy)，計算一段時間下所有車輛通過偵測點所花的時間佔該段時間的百分比 。

N： 所收集資料的數量。

圖 1 為不同的佔有率對應的流量分佈圖，如圖所示，資料分佈主要集中於低佔有率時，並且該圖大略呈現出交通流量跟佔有率有正相關的現象。

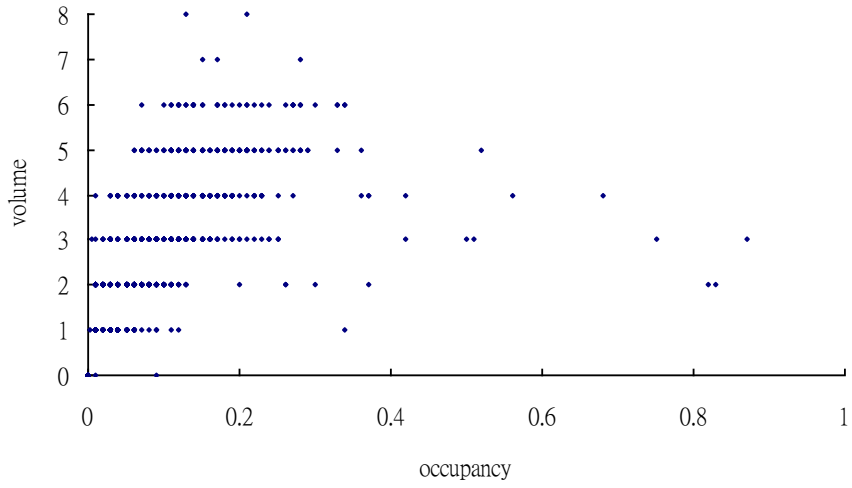


圖 1 佔有率-流量分佈圖

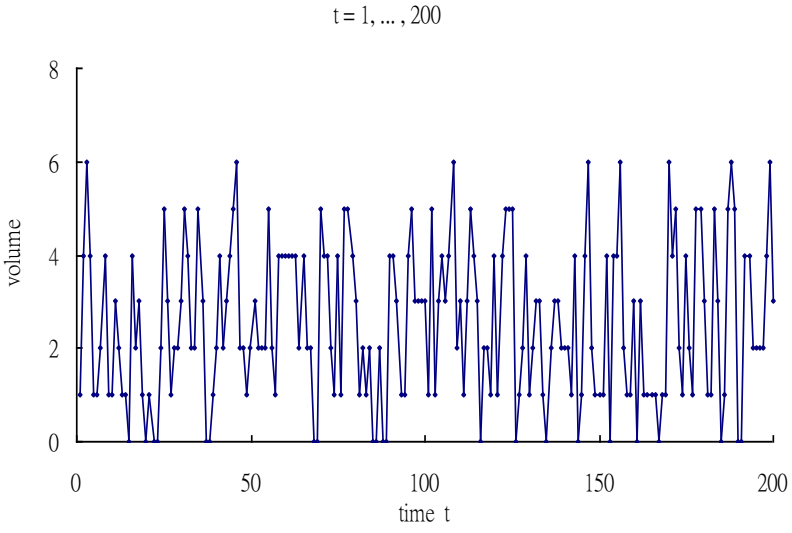


圖 2 交通流量圖 (一開始的 200 筆完整資料)

4.2 資料分析

本研究試著將參數導向模式應用到交通流量的分析的實際案例進行討論。根據方程式(1)中的共變異數已知為$U_t=(1,u_t)'$。本研究試著針對潛在過程建立一階自我回歸模式，用修正後的 EM 演算法來處理估計。而ρ和σ_ε^2的起始值皆為 0.5。而α的起始值是由資料比對的對數-線形模型所獲得。假設在去除時間效應(temporal dependence)下，可求得$(\alpha_0,\alpha_1)'$的值為 $(0.682636\ ,2.550254\ )'$。

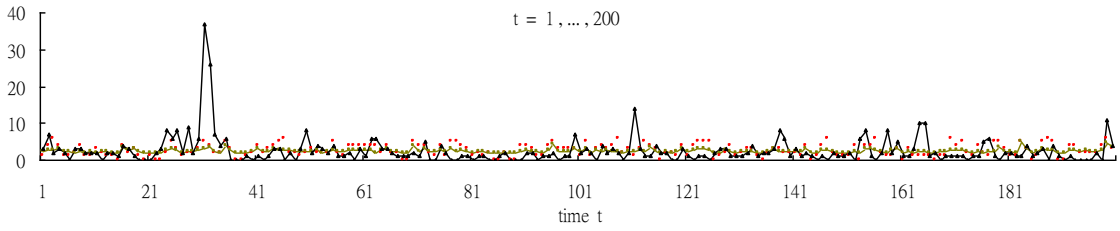
$Q(\bullet|\theta)$的最大值和近似的最大概似估計(approximate ML estimates)可由下列之參數值求得，

$$\hat{P}=0.707559\text{，}\quad \hat{\sigma}_\varepsilon^2=0.332942\text{，}\quad \hat{\alpha}'=(0.350308\ ,2.580916\ )$$

(0.022361)　(0.014899)　　(0.025593) (0.129316)

括號內的數值為估計的標準差，由方程式(9)所得。

比對結果顯示如圖 3、4 所示，其中各點代表觀察值。顏色較淺的實線代表由方法 1 模擬的比對資料，而觀察值和比對資料間的平均誤差為$\sum_{t=1}^{n}\frac{|y_t-\hat{y}_t|}{n}=1.181$；顏色較深的實線代表由方法 2 模擬的比對資料，而觀察值和比對資料的平均誤差為$\sum_{t=1}^{n}\frac{|y_t-\hat{y}_t|}{n}=2.277$。因此可以發現到由方法 1 所得的平均誤差小於較方法 2 的平均誤差，這個不穩定的情況可能來自於潛在過程W_t的變化。



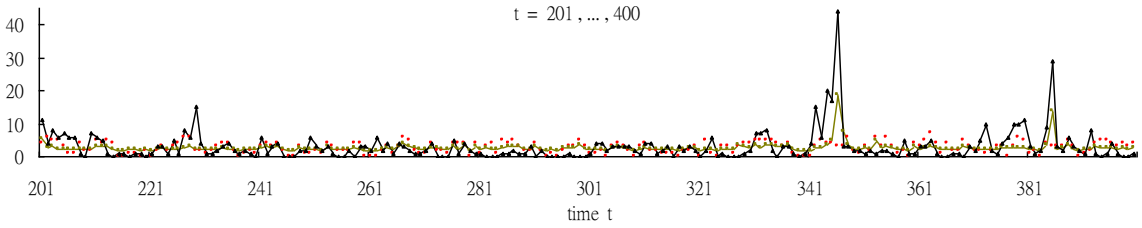


圖 3 交通流量的時間序列 (一開始的 400 筆完整資料)

圖 3 上各點代表觀察流量值。顏色較淺的實線代表著從方法 1 模擬的比對資料；顏色較深的實線代表著從方法 2 模擬的比對資料。

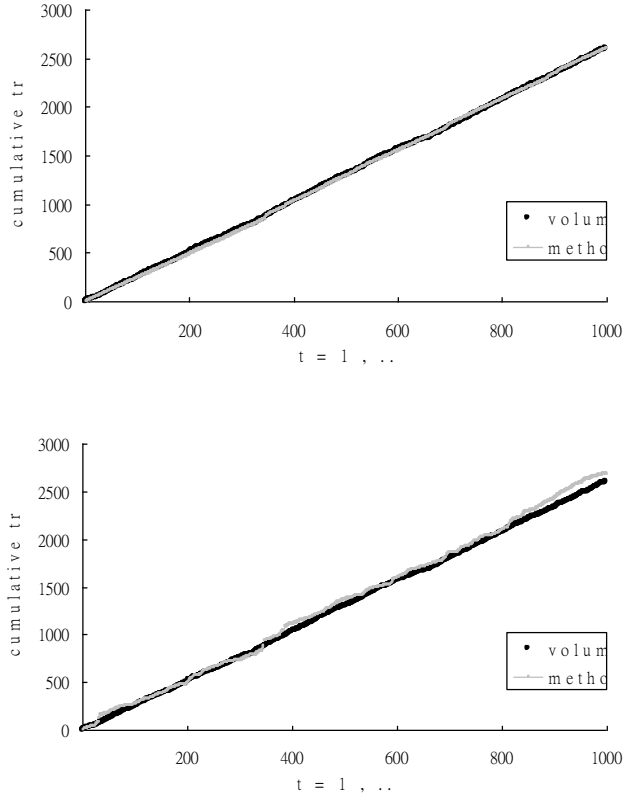


圖 4 累積交通流量分佈圖

圖 4 上各點代表各時間點的流量累積觀察值，左圖表示從資料 1 模擬的比對資料；右圖表示從方法 2 模擬的比對資料。可以發現方法 1 的結果會比方法 2 的更接近觀察到的交通流量，而這個不穩定情況可能來自潛在過程 W_t 的產生。

五、 結論

本研究將記數時間序列的資料上運用於參數導向模式，並利用改良型的 EM 演算法進行求解。假設潛在過程為一組一階自我回歸高斯過程，利用本模分析交通流量和佔有率之間的關係。本演算法所得到的預測值與實際值的平均誤差不大，但是無論在計算的時間花費或記憶儲存空間上都有經濟上節省的優點，因此在交通控制的即時控制與判斷上，可以得到相當大的效果。

參考資料

1. Cox, D.R., Statistical analysis of time series: some recent developments (with discussion), Scandinavian Journal of Statistics, Vol. 8, 1981, pp. 93-115.

2. Cho, H.J., and M.T. Tseng, A novel computational algorithm for traffic signal control SoC, WSEAS Transactions on Mathematics, Issue 1, Vol. 5, 2006, pp. 123-128.

3. Chan, K.S., and J. Ledolter, Monte Carlo EM estimation for time Series models involving counts, Journal of the American Statistical Association, Vol. 90, No. 429, 1995, pp. 242-252.

4. Freeland, R.K., and B.P.M. McCabe, Analysis of low count time series data by Poisson autoregression, Journal of Time Series Analysis, Vol. 25, No.5, 2004, pp. 701-722.

5. Louis, T.A., Finding the observed information matrix when using the EM algorithm, Journal of the Royal Statistical Society Series B (Methodological), Vol. 44, No. 2, 1982, pp. 109-286.

6. Meng, X.L., and D.B. Rubin, Using EM to obtain asymptotic variance-covariance matrices: The SEM Algorithm, Journal of the American Statistical Association, Vol. 86, No. 416, 1991, pp. 899-909.

7. Zeger, S.L., A regression model for time series of counts, Biometrika, Vol. 75, 1988, pp. 621-629.