福祉科技與服務管理學刊
Journal of Gerontechnology and Service Management

研究論文

改良式 ARIMA 演算法預測台灣未來五年之癌症死亡率

莊政宏[1,2] 呂威甫[1,3] *陳瑞奇[1]
[1]亞洲大學 資訊工程學系
[2]中國醫藥大學 醫學研究部
[3]亞洲大學 生物資訊與醫學工程學系

摘要

近年來，65 歲以上癌症死亡數占比呈上升趨勢。死因之統計與分析有助於公共衛生政策之規劃與全民健康之提升。趨勢延伸演算法（時間序列分析）常被用於疾病定量預測，其中之一就是自迴歸整合移動平均(ARIMA)模型，該模型可以用來對時間序列資料進行預測，尤其針對隨機過程特徵隨時間變化、而且導致時間序列非平穩的原因是隨機的問題上特別有用。目前台灣各縣市歷年癌症死亡統計資料大都屬於所謂成因不明確、非平穩的時間序列資料集，適合使用趨勢延伸演算法對未來癌症死亡趨勢執行預測。本研究為了進一步提升預測的準確性，對傳統 ARIMA 演算法進行改良，我們先將截至目前為止衛福部已公布的台灣各縣市 1992 年至 2017 年共計 26 年癌症死亡率，分為訓練及測試資料，實施幾種不同方法的預測效能比較，其中一項是以平均絕對百分比誤差(MAPE)評估各方法之預測準確度，最後以最佳改良式 ARIMA 演算法預估台灣未來五年之癌症死亡率，提供政府相關單位事先了解癌症死亡的可能趨勢及其政策規劃上的參考，以便民眾（尤其是銀髮族）擁有適當的癌症篩檢機制、罹癌者都能獲得妥善治療，降低癌症死亡率，並提高生活品質。

關鍵詞：銀髮健康、癌症死亡率、趨勢延伸演算法、ARIMA 模型

1. 前言

行政院衛生福利部（以下簡稱衛福部）統計處(2018)於去年 6 月公佈的 2017 年國人主要死因分析，台灣地區十大死因死亡人數占總死亡人數之 76.8%，其中惡性腫瘤（癌症，cancer）死亡率位居首位（占 28.0%），年長者的死因首位也是癌症。就長期變化趨勢觀察，隨人口成長及高齡人口比重增加，死亡人數呈上升趨勢，以 WHO 西元 2000 年世界標準人口年齡結構計算之標準化死亡率為每十萬人口 424.3 人，受人口結構快速高齡化影響，65 歲以上死亡者占總死亡人數比率呈現逐年遞增趨勢，主要係因 85 歲以上死亡人數快速增加所致。自 1982 年起，連續這三十六年來癌症一直是國內十大死因之首位，2017 年國人癌症死亡率為每十萬人口 203.9 人，如圖 1 所示。依死亡率排序，2017 年十大癌症死因順位依序為：(1)氣管、支氣管和肺癌（死亡率：每十萬人口

*通訊作者：rikki@asia.edu.tw

39.2 人)；(2)肝和肝內膽管癌（35.7 人)；(3)結腸、直腸和肛門癌（24.7 人)；(4)女性乳癌（20.1 人)；(5)口腔癌(12.1 人)；(6)前列腺（攝護腺）癌（11.9 人)；(7)胃癌（9.8 人)；(8)胰臟癌（8.8 人)；(9)食道癌（7.6 人)；(10)子宮頸及部位未明示子宮癌（5.5 人)。而 65 歲以上癌症總死亡數占比呈上升趨勢，如圖 2 所示。死因之統計與分析有助於公共衛生政策之規劃與全民健康之提升。(衛生福利部統計處，2018)

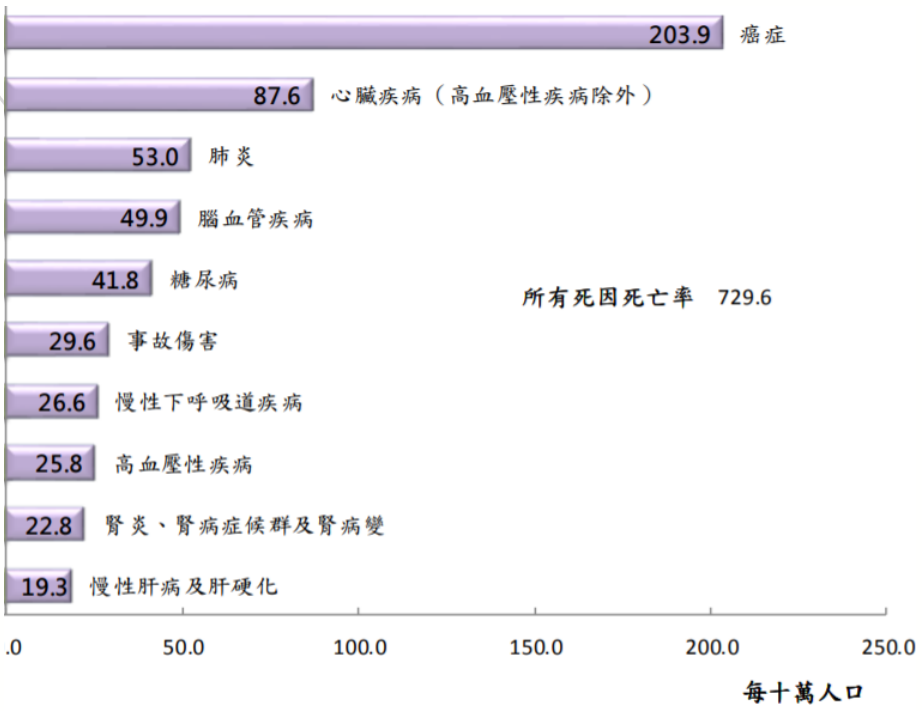

圖 1. 2017 年國人十大死因死亡率（衛生福利部統計處，2018）

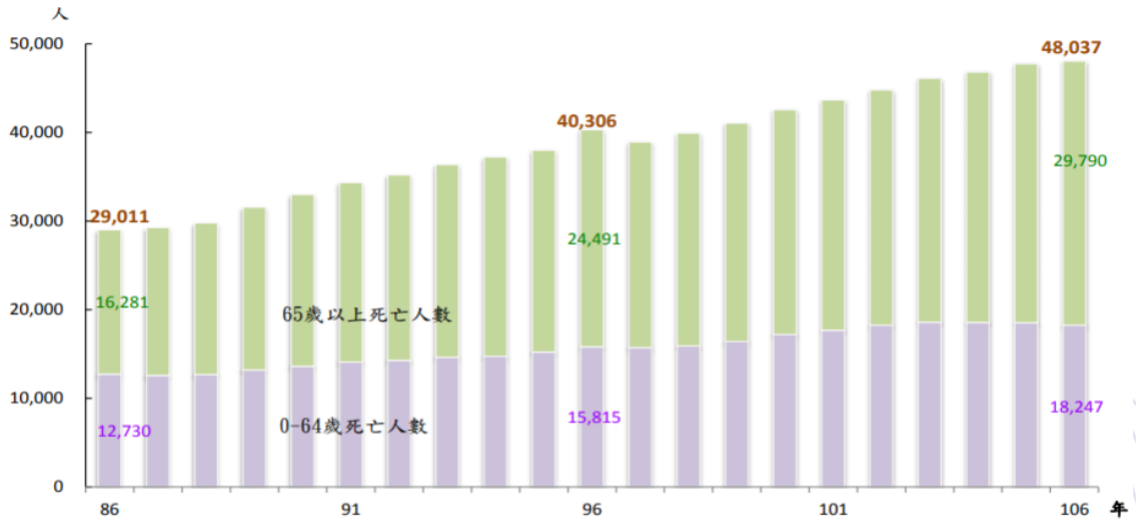

圖 2. 歷年癌症死亡人數（衛生福利部統計處，2018）

近年來，自迴歸整合移動平均(Autoregressive integrated moving average, ARIMA)趨勢延伸演算法常被應用在疾病預測上，利用疾病歷史資料建立隨時間變化的動態模型，是一種預測疾病的實用

演算法（Xu, 2015; He & Tao, 2018；時照華等人，2013）。政府開放資料(open data)提供台灣地區歷年癌症相關統計的資料集(國家發展委員會，2018)，其中，各縣市歷年癌症死亡統計資料，大部分都屬於所謂成因不明確、非平穩(non-stationary)、非季節性(non-seasonal)的時間序列(time series)資料集，確實也可嘗試著運用 ARIMA 預測演算法，對台灣各縣市的癌症死亡趨勢進行預測。

臺灣大學流行病學與預防醫學研究所劉祥雯(2016)發表「以統計模型分析肺癌發生率及死亡率之長期趨勢」論文，應用一系列統計分析方法，探討過去 30 年來臺灣地區肺癌死亡率的時間趨勢，認為肺癌的死亡率及其組織型態，皆因時間演進而改變，其方法之一就是採用 ARIMA (0, 1, 0)隨機漫步(random walk)模型，預測不同性別整體及個別組織型態肺癌死亡率。

本研究為了進一步提升預測的準確性，對傳統 ARIMA 演算法進行改良。我們先將截至目前為止，衛福部已公布的台灣各縣市 1992 年至 2017 年、共計 26 年癌症死亡率，分為訓練及測試資料，實施幾種不同方法的預測效能比較，其中一項是以平均絕對百分比誤差(MAPE)評估各方法之預測準確度，這些方法包括：系統自動配適 ARIMA 模型、最小赤池參數 ARIMA 模型、經前處理以及後處理的幾種改良式 ARIMA 趨勢延伸演算法，最後以最佳改良式 ARIMA 演算法，去預估台灣未來五年之癌症死亡率，提供政府相關單位事先了解癌症死亡的可能趨勢，及其政策規劃上的參考。

例如某縣市某項癌症死亡率長期、乃至未來預測走勢比其他縣市都還要偏高，那麼政府相關單位應該進一步著手研究，有甚麼樣的原因或什麼樣的政策，導致這個結果。該縣市是否癌症篩檢機制哪裡不夠周全，以致無法早期發現、早期治療，尤其針對高齡者的篩檢，使得死亡率提高；是否民眾癌症防治衛教宣導不足、或與其他縣市間有政策上的差異等等。是否應該加強該項致癌因子的預防工作，包括推行「現代國民營養計畫」以利防治像肥胖、飲食與運動不足等這類的新興致癌因子；或者考慮透過醫院或政府衛生單位聯絡通知，乃至結合民間團體，主動走入社區推廣菸、酒、及不健康飲食等致癌風險的宣導，建立健康生活型態，降低罹癌風險。是否應該針對該縣市地區環境的加強監測與改善，例如某些癌症常與水質或空氣品質有很大的關聯性，地方政府對機動車輛及工廠排放是否應該有更好的監測與管理機制，乃至大眾交通工具的完善規劃與改進。甚至應該參考死亡率趨勢較低的縣市做法，增進市民福祉、改善生活品質。（衛福部國健署，2019）

另外，縣市政府也可以推出所謂的「癌友導航計畫」，引導癌症病人到癌症品質認證醫院，不錯失任何一位可治療之癌症病人，使早期病人得以治癒，有效降低癌症死亡率。對晚期病人（還是以高齡的銀髮族佔多數）則提供安寧療護，減少病友迷航，讓每位病人都能夠得到有品質、有尊嚴的治療與照護。（衛福部國健署，2019）

2. 材料與方法

2.1 材料

政府為提升民眾參與公共政策議題，由國家發展委員會(2018)管理政府開放資料共用平臺，主動提供各機關各式資料集。本研究主要是以台灣地區各縣市歷年癌症 WHO，2000 年人口年齡標準化死亡率做為預測研究對象，所採用的「癌症死亡統計」資料集，是由衛福部國民健康署提供之我國癌症死亡統計資料，資料更新日期為 2018/09/23，資料檔案格式為純文字檔(.txt)，文字編碼：UTF-8，每年記錄一檔，從 1992 年到 2017 年共有 26 個檔案、計 60,996 筆資料，以 ZIP 格式壓縮成一個資料包提供下載，每個純文字檔主要欄位有年度別(year)、性別(sex)、年齡層代碼(age code)、鄉鎮市區代碼(township code)、死亡原因(cause of death)、以及死亡數(N)。（國家發展委員會，2018）

為了計算癌症死亡率，還需下載「人口數性別概況—按區域別分」資料集，由內政部統計處所提供之我國及各縣市人口數，資料更新日期為 2018/09/04，資料檔案格式為以逗號分隔欄位的純文字檔(.csv)，文字編碼：UTF-8，全部記錄成一檔，提供從 1982 年到 2017 年之各縣市人口數統計資料（行政院內政部統計處，2018）。

本研究預測模型之訓練資料與測試資料，採用 1992 年到 2017 年前後總計 26 年期間的死亡率，分別以第 1 年及第 2 年為起點各做 100 組資料（5 類癌症別搭配 20 個地區別）之訓練與預測，（後來經審稿委員提醒，兩個年度起點的資料幾乎重覆取樣，意義不大，其實只做一個年度為起點即可）。利用每組資料的前 20 年死亡率進行訓練、以其後續 5 年死亡率做為預測準確性的測試資料；地理區域包含台灣全區及 19 縣市（台中市、台北市、台東縣、台南市、宜蘭縣、花蓮縣、南投縣、屏東縣、苗栗縣、桃園縣、高雄市、基隆市、雲林縣、新北市、新竹市、新竹縣、嘉義市、嘉義縣、彰化縣），但不包含人口數較少、死亡樣本數少、而死亡率變異較大的外島三縣市（澎湖縣、金門縣、連江縣）；癌症原發部位則採計全癌症及前四大癌症，所謂的前四大是指衛福部 2018 年 6 月公布的 2017 年國人十大癌症依死亡率（男女合計）排行前五名中的肺癌、肝癌、大腸癌以及口腔癌，因乳癌絕大部分為女性，暫不列入訓練預測項目。

2.2 方法

本研究將台灣各縣市歷年癌症死亡率分為訓練及測試資料，進行幾種不同趨勢預測改良方法的實驗比較：

傳統單變量 ARIMA 趨勢延伸演算法(Auto.ARIMA)

傳統 ARIMA (p, d, q)模型是從歷史的資料中學到隨時間變化的函數，以此預測未來，AR (p)是自迴歸成分，p 為其自迴歸項數；MA (q)為移動平均成分，q 為其移動平均項數；d 是差分的階數，用以得到平穩序列，ARIMA 模型參考公式如下：

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t = \delta + \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t, \quad (1)$$

其中，有 p 個自我迴歸項及 q 個移動平均項，X_t 為第 t 期之實際值，L 是滯後運算子(Lag operator)，L 滯後運算一次定義為 $LX_t = X_{t-1}$，重複 L 滯後運算 k 次定義為 $L^k X_t = X_{t-k}$，ϕ_i 及 θ_i 分別為自迴歸項與移動平均項之係數，δ 是一個常數值，ε_t 是一個隨機誤差值，需符合白噪音(White noise)程序。(Wikipedia, 2019)

ARIMA 模型取得 *p, d, q* 參數最快的方法之一，就是直接利用 R 語言之時間序列 forecast 套件 auto.arima 函式(Hyndman et al., 2017)，由系統自動配適最佳之 ARIMA(*p, d, q*)模型。

<u>選擇最小赤池參數的 ARIMA 預測模型(MinAIC.ARIMA)</u>

統計學家們認為較佳預測模型的選擇，在於模型複雜度與模型對資料集描述能力之間取得一個最佳平衡點，過去曾提出一些信息準則，透過加入模型複雜度懲罰項(penalty term)來避免過度配適(overfitting)的問題，其中最常用的有赤池訊息準則(Akaike information critera with correction, AICc) (Akaike, 1974)，當樣本數增加時，赤池參數 AICc 會收斂成 AIC，AICc 可以應用在任何樣本大小的情況下(Burnham & Anderson, 2002)；以及貝葉斯信息準則(Bayesian information criterion, BIC)或稱施瓦茨準則(Schwarz information criterion, SIC) (Kass & Raftery, 1995; Schwarz, 1978)。

當我們進行模型選擇時，AIC 及 BIC 值越小的模型越合適。如前所述，auto.arima 函式是由系統自動配適最佳之 ARIMA (*p, d, q*)模型，在選擇模型參數時，如未設定訊息準則(Information critera)，主要考量是以赤池準則 AICc 最低數值為優先，但我們發現 auto.arima 並非直接找到 AICc 真正最低的數值，原因不明，也許另有考量，亦未再深入追究。因此，我們使用 R 語言之時間序列 forecast 套件中 Arima (MortalityTimeSeries, order = c(*p, d, q*))函式，針對癌症死亡率訓練資料，以窮舉法尋找 ARIMA (0, 0, 0) ~ ARIMA (5, 3, 5)之間擁有最小 AICc 值的 ARIMA (*p*, *d*, *q*)模型，請參考已發表論文(Chuang et al., 2018)。

<u>原始資料經 BSWMA 一次前處理的 ARIMA 預測模型(BSWMA-ARIMA)</u>

有關本研究 ARIMA 預測前處理及後處理所使用的技術，是我們過去曾經發表所謂平衡式加權移動平均線法(Balanced weighted moving average, BWMA)的概念(Chuang et al., 2018)，一則可以利用移動平均技術去平滑癌症死亡率數據的上下波動，再則可以解決使用傳統算術移動平均數(Arithmetic moving average, AMA)會有數值滯後(Lag)的問題，在大多數的情況下可以得到很好的平滑效果，同時在改善滯後問題上也有很大的貢獻。BWMA 是採用每年(*y*)的癌症死亡率 *n* 年平衡式 *k* 次加權移動平均數(BWMA of power *k*, BWMAP_k)產生移動平均新數據，其公式如下：

$$n\mathrm{BWMAP}_k(y) = \left[\left\lfloor\frac{n}{2}+1\right\rfloor^k \cdot I_y + \sum_{j=1}^{\lfloor\frac{n}{2}\rfloor}\left(\left\lfloor\frac{n}{2}+1\right\rfloor - j\right)^k (I_{y-j} + I_{y+j})\right] \Big/ \left(\left\lfloor\frac{n}{2}+1\right\rfloor^k + 2\cdot\sum_{i=1}^{\lfloor\frac{n}{2}\rfloor} i^k\right), \quad (2)$$

$$where\ k, n \in \mathrm{N}, \mathrm{n} \geq 2, and\ I_j = mortality\ in\ the\ year\ j.$$

其中，當接近最大年度端點之移動平均值時，後面$\lfloor \frac{n}{2} \rfloor$年採取截斷式計算的做法（未來是可以利用每年預測值回溯填補及替代計算），亦即，不可使用到必須假設未知的年度資料，例如，假設已知最大年度的那一年 BWMAP 只計算當年度死亡率及往前$\lfloor \frac{n}{2} \rfloor$年的加權平均，而把訓練資料中假設應該無數據的後$\lfloor \frac{n}{2} \rfloor$年截斷（不列入平均計算），其最大年度那一年的公式如下：

$$n\text{BWMAP}_k(y)[\text{最大年度}] = \left[\left\lfloor \tfrac{n}{2}+1 \right\rfloor^k \cdot I_y + \sum_{j=1}^{\lfloor \frac{n}{2} \rfloor} \left(\left\lfloor \tfrac{n}{2}+1 \right\rfloor - j\right)^k (I_{y-j})\right] \Big/ \left(\left\lfloor \tfrac{n}{2}+1 \right\rfloor^k + \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} i^k\right), \quad (3)$$

$$where\ k, n \in \mathrm{N}, \mathrm{n} \geq 2, and\ I_j = mortality\ in\ the\ year\ j.$$

若當公式(2)中的 $k=2$ 時，BWMAP_k 就會等於 BSWMA。再者，本研究以使用 9 年為計算移動平均之週期，亦即 $n=9$，所以每年(y)癌症死亡率的 9BSWMA 公式變為：

$$9\text{BSWMA}(y) = \left[25I_y + \sum_{j=1}^{4}(5-j)^2(I_{y-j} + I_{y+j})\right] \Big/ (25 + 2\sum_{i=1}^{4} i^2), \quad (4)$$

$$where\ \ I_j = mortality\ in\ the\ year\ j.$$

BSWMA 接近最大年度端點之移動平均值採取截斷式計算的做法，例如訓練資料中最大年度的 9BSWMA 只計算當年度死亡率及往前四年的加權平均，而把無數據的後四年截斷（不列入平均計算），其參考公式如下：

$$9\text{BSWMA}(y)[\text{最大年度}] = \left[25I_y + \sum_{j=1}^{4}(5-j)^2(I_{y-j})\right] \Big/ (25 + \sum_{i=1}^{4} i^2), \quad (5)$$

$$where\ \ I_j = mortality\ in\ the\ year\ j.$$

本方法先將原始資料經過一次 9BSWMA 前處理之後，再應用 ARIMA 進行配適(Fitting)及預測，用以預測未來 m 年死亡率，我們稱之為「BSWMA 前處理 ARIMA 預測演算法」(BSWMA-ARIMA)，其演算法如下：

D：原始資料（死亡率）
t：原始資料總筆數（年）
m：未來預測總筆數（年）
MA：配適最佳 ARIMA 模型
OP：未來預測值（死亡率）
▷：註解

BSWMA-ARIMA (D, t, m)

(1) for each year $u \in [1, t]$
(2) do $F[u] \leftarrow$ 9BSWMA($D[u]$) ▷一次 9BSWMA 前處理
(3) $MA \leftarrow$ Auto.ARIMA(F)
(4) $OP \leftarrow$ Forecast(MA, m) ▷預測未來 m 年
(5) for each year $v \in [1, m]$
(6) do print $OP[v]$ ▷輸出最後預測值

原始資料經 BSWMA 前後處理一次的 ARIMA 預測演算法(BSWMA-1-ARIMA)

本方法先將原始資料經過一次 9BSWMA 前處理之後，再應用 ARIMA 進行配適及預測，預測未來 m 年趨勢後、還需將原始資料及 m 年預測值再經過一次 9BSWMA 後處理，後處理之後所得的未來 m 年死亡率即為最後預測值，我們稱之為「BSWMA 前後處理 1 次 ARIMA 預測演算法」(BSWMA-1-ARIMA)，其演算法如下：

```
BSWMA-1-ARIMA(D, t, m)
(1)  for each year u ∈ [1, t]
(2)       do F[u] ← 9BSWMA(D[u])        ▷一次前處理
(3)  MA ← Auto.ARIMA(F)
(4)  OP ← Forecast(MA, m)               ▷預測未來 m 年
(5)  for each year v ∈ [1, m]
(6)       do D[t + v] ← OP[v]
(7)  for each year u ∈ [1, t + m]
(8)       do F[u] ← 9BSWMA(D[u])        ▷一次後處理
(9)  for each year v ∈ [1, m]
(10)      do print F[t + v]             ▷輸出最後預測值
```

原始資料經 BSWMA 前後處理 m 次的 ARIMA 預測演算法(BSWMA-m-ARIMA)

本方法先將原始資料經過一次 9BSWMA 前處理之後，再應用 ARIMA 進行配適及預測，每次先預測未來 1 年趨勢，再將原始資料及前幾年預測值，透過一次 9BSWMA 後處理，可獲得最新 1 年預測值，並轉當成新加入的原始資料，再繼續進行下一次前處理及 ARIMA 配適。如此逐年預測，經過 m 次前後處理之後，所得未來 m 年死亡率即為最後預測值，我們稱之為「BSWMA 前後處理 m 次 ARIMA 預測演算法」(BSWMA-m-ARIMA)，其演算法如下：

```
BSWMA-m-ARIMA(D, t, m)
(1)  for each year v ∈ [1, m]
(2)       for each year u ∈ [1, t + v - 1]
(3)            do F[u] ← 9BSWMA(D[u])      ▷總共 m 次前處理
(4)       MA ← Auto.ARIMA(F)
(5)       OP ← Forecast(MA, 1)             ▷每次預測未來 1 年
(6)       D[t + v] ← OP[1]
(7)            for each year u ∈ [1, t + v]
(8)                 do F[u] ← 9BSWMA(D[u])  ▷總共 m 次後處理
(9)       D[t + v] ← F[t + v]
(10) for each year v ∈ [1, m]
(11)      do print D[t + v]                ▷輸出最後預測值
```

本研究實驗比較所採用的預測準確度評估方法(performance measures)為均方根誤差(root mean squared error, RMSE)、平均絕對誤差(mean absolute error, MAE)、以及平均絕對百分比誤差(mean absolute percent error, MAPE)等(Hyndman & Athanasopoulos, 2014; Hill, 2011; Lewis, 1982)，若評估的預測誤差數值越小，表示預測模型愈好。

從公式上來看，MSE, RMSE, MAE 及 MAPE 對誤差值的計算，都是針對預測值與實際值之間絕對距離的評估，利用上述不同評估方法、進行不同演算法預測誤差的比較時，排名順序不會因此而改變，只是評估數值變異大小不同。故於文後大部分結果的顯示，考慮以百分比率的 MAPE 做為預測效能的比較，利於讀者快速理解，另外便於參考學者 Lewis (1982)所提出 MAPE 按預測能力分為四種等級，如表 1 所示，MAPE 公式參考如下：

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}|(y_i - p_i)/y_i| \times 100\%, \text{ 其中 } p_i \text{是預測值}, y_i \text{是實際值。} \quad (6)$$

表 1. MAPE 預測能力等級表(Lewis, 1982)

MAPE	預測能力
<10%	高精確度的預測
10%~20%	良好的預測
10%~20%	合理的預測
>50%	不正確的預測

3. 結果與討論

我們先以台灣 19 縣市 1993 年到 2012 年共 20 年、全癌症及前四大癌症死亡率為訓練資料，以 2013 年到 2017 年死亡率為測試資料，進行簡單預測與比較，例如圖 3 及圖 4 所示，橫軸是癌症死亡年，縱軸為 WHO 每十萬人口年齡標準化死亡率，圓圈標示紅色虛線代表原始死亡率（含前 20 年訓練資料及後 5 年測試資料），綠色粗線代表 9BSWMA 前處理後之移動平均死亡率（1993-2012 年），藍色粗線代表 9BSWMA 前處理後資料配適之「BSWMA-ARIMA」演算法、所得到的預測結果（2013-2017 年）。

圖 3 及圖 4 是本文所提出第一種改良方法的預測範例圖示，先將原始資料進行 9BSWMA 前處理後、再用傳統 Auto.ARIMA 函數做預測，而且只是一組數據範例（單一縣市、單一癌症、及單一年度週期）預測結果的觀察，不足以做為不同演算法預測效能的比較基礎。有關應用大樣本資料集、以及比較幾種改良方法與原始資料直接用 Auto.ARIMA 函數預測的效能差異，將於表 2 及下個段落中再予以說明。

圖 3(a)南投縣全癌症及圖 3(b)台東縣大腸癌的預測結果非常接近紅色圓圈之原始數據，但圖 4 苗栗縣口腔癌的預測效果就令人有點失望，從平衡式加權移動平均（綠色粗線）的趨勢，直覺上自

動配適模型的預測（藍色粗線）似乎非常合理，然而未來死亡率本來就有不確定性，也不能以單一數據預測結果加以論斷。

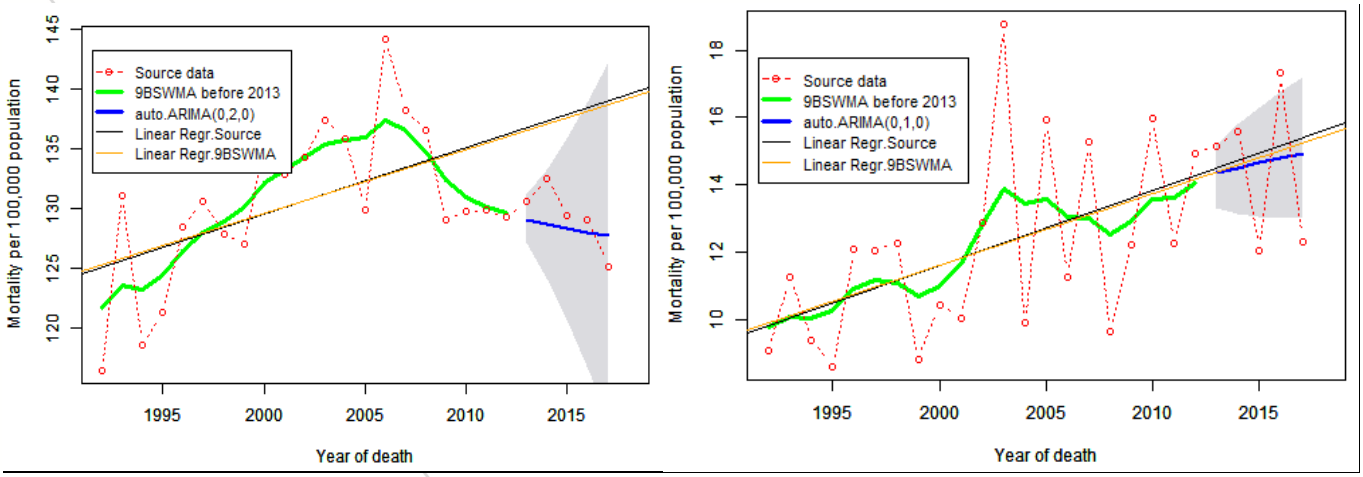

圖 3(a)(b). BSWMA-ARIMA 進行南投及台東縣癌症死亡率預測結果之比較

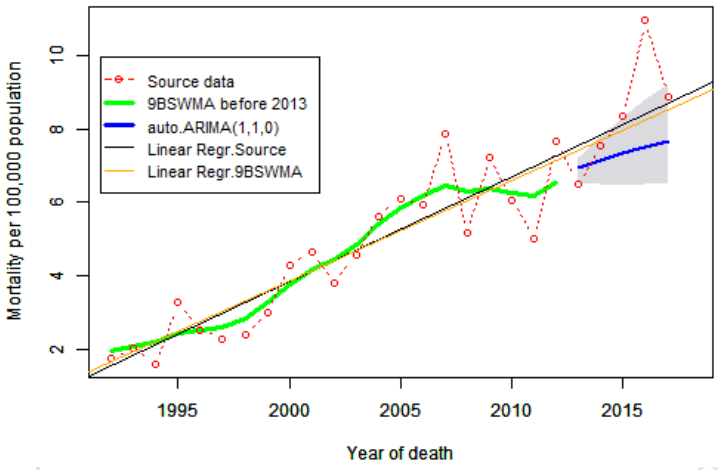

圖 4. BSWMA-ARIMA 進行苗栗縣歷年口腔癌死亡率預測結果之比較

我們進一步分台灣各縣市、針對全癌症及前四大癌症死亡率共 100 組資料集進行訓練、預測及效能評估，從 1992-2011 年到 1993-2012 年連續二個周期（每周期 20 年）為訓練資料，分別對應以預測 2012-2016 年到 2013-2017 年連續二個周期（每周期 5 年）死亡率做為效能評估，計算每周期 5 年預測誤差(MAPE)、再全部平均進行以下 6 種預測方法之效能比較：「傳統系統自動配適 ARIMA 模型(Auto.ARIMA)」、「最小赤池參數 ARIMA 模型(MinAIC.ARIMA)」、「BSWMA 前處理之 ARIMA 預測演算法(BSWMA-ARIMA)」、「BSWMA 前後處理 1 次之 ARIMA 預測演算法(BSWMA-1-ARIMA)」、「BSWMA 前後處理 *m* 次之 ARIMA 預測演算法(BSWMA-*m*-ARIMA)」、「線性迴歸模型(Linear Regression)」等六種預測方法。全部總計 20 個地理區域、5 項癌症別、連續 2 個周期、每周期預測 5 年死亡率共計 1,000 年（次）預測，預測結果準確度之評估為 MAPE 總平均誤差，誤差越小代表準確度越高。

預測結果之比較如表 2 所示，所有各地理區域以傳統自動配適模型 Auto.ARIMA 預測準確度，都有比線性迴歸模型簡單預測好很多，而最小的赤池參數模型 MinAIC.ARIMA 也不一定有更好預測效能，其他有經過 BSWMA 前處理的 ARIMA 預測演算法大部分都比傳統 Auto.ARIMA 還要再好一些。若以最下方全部 20 個區域死亡率所預測結果之 MAPE 總平均值來做整體比較，線性迴歸模型簡單預測(12.32%)最差，其次是傳統自動配適模型 Auto.ARIMA 預測準確度(MAPE=9.75%)，但尚符合 Lewis 高精確度等級之內，其中以 BSWMA-*m*-ARIMA 預測準確率最高、全部 20 個區域死亡率預測總平均誤差 MAPE 為 7.91%。

另外，實驗發現死亡率較高者則不見得準確度就會較高，且死亡率是以每十萬人計或以每百萬人計完全沒有影響效能評估的準確性。後續以「Auto.ARIMA」、「BSWMA-*m*-ARIMA」及「Linear Regression」等三種預測方法，倆倆進行上述台灣全區及 19 縣市預測全癌症及前四大癌症死亡率準確度之成對 *t* 檢定(Paired *t*-test)，檢定結果，改良式趨勢延伸演算法「BSWMA-*m*-ARIMA」預測準確度最高，傳統模型「Auto.ARIMA」其次，簡單線性迴歸「linear Regression」最不好，而且這三種演算法倆倆之間皆有統計上的顯著差異；由此可知，若使用一些技巧和改良，預測誤差(MAPE)可以從線性預測的 12.32%降低到改良式趨勢延伸演算法的 7.91%，顯然高精確度的預測並非全然不可達到。採用良好準確度的預測技術才使得後續對未來預測趨勢的解釋較為合理化。

表 2. 台灣及 19 縣市使用六種方法預測全癌症及前四大癌症死亡率準確度(MAPE)之比較

Forecasting Algorithm	台灣	台中市	台北市	台東縣	台南市	宜蘭縣	花蓮縣	南投縣	屏東縣	苗栗縣
Auto.ARIMA	5.8%	7.5%	8.8%	10.4%	11.7%	10.9%	8.1%	6.6%	7.7%	9.7%
MinAIC.ARIMA	4.4%	5.6%	7.0%	10.9%	10.9%	10.5%	9.3%	6.3%	7.6%	7.6%
BSWMA-ARIMA	4.6%	6.1%	6.8%	11.1%	9.7%	8.6%	7.8%	4.6%	5.4%	10.0%
BSWMA-1-ARIMA	4.5%	6.0%	6.9%	11.0%	9.4%	8.6%	7.8%	4.5%	5.3%	9.9%
BSWMA-***m***-ARIMA	4.4%	5.6%	8.1%	10.3%	7.5%	9.0%	7.5%	4.3%	4.8%	9.1%
Linear Regression	11.6%	11.9%	8.6%	11.6%	14.2%	9.1%	10.3%	9.0%	13.3%	9.4%

Forecasting Algorithm	桃園市	高雄市	基隆市	雲林縣	新北市	新竹市	新竹縣	嘉義市	嘉義縣	彰化縣
Auto.ARIMA	9.6%	7.1%	15.1%	7.5%	6.5%	13.8%	16.3%	13.3%	8.8%	9.8%
MinAIC.ARIMA	9.6%	6.5%	13.0%	9.6%	6.8%	12.0%	16.5%	19.1%	10.1%	7.7%
BSWMA-ARIMA	9.1%	6.2%	10.8%	6.2%	5.1%	13.2%	15.9%	15.2%	7.5%	6.8%
BSWMA-1-ARIMA	9.0%	6.0%	10.9%	6.1%	5.1%	12.9%	15.6%	14.7%	7.6%	6.8%
BSWMA-***m***-ARIMA	7.7%	5.3%	10.8%	6.3%	4.9%	11.0%	14.6%	12.9%	7.6%	6.3%
Linear Regression	12.6%	13.0%	11.1%	16.8%	8.2%	11.4%	20.0%	14.4%	13.8%	16.3%

Forecasting Algorithm	MAPE in 20 regions
Auto.ARIMA	9.75%
MinAIC.ARIMA	9.53%
BSWMA-ARIMA	8.53%
BSWMA-1-ARIMA	8.43%
BSWMA-***m***-ARIMA	7.91%
Linear Regression	12.32%

另外我們再針對上述 100 組資料集及六種預測方法、分癌症別計算 20 組（不同地理區域）五年死亡率預測誤差(MAPE)再平均，得到表 3 預測結果之比較，不同預測方法在不同癌症別的表現

各有千秋，大致上「BSWMA-m-ARIMA」預測準確度還是優於「Auto.ARIMA」，不過，特定癌症別(不分縣市別)之預測困難度較高，例如表 3 第七列各癌症別預測平均 MAPE，以口腔癌和肝癌最難預測(MAPE>10%)，大腸癌及肺癌預測準確度稍高(MAPE<10%)，尚符合 Lewis 高精確度等級之內，全癌症預測準確度最高(MAPE=3.8%)。

表 3. 台灣地區全癌症及前四大癌症死亡率、使用六種預測方法準確性之比較

Forecasting Algorithm	全癌症	肺癌	肝癌	大腸癌	口腔癌	Mean MAPE
Auto.ARIMA	4.0%	9.1%	13.0%	8.3%	14.3%	9.75%
MinAIC.ARIMA	3.9%	9.0%	12.7%	8.2%	13.9%	9.53%
BSWMA-ARIMA	3.8%	8.6%	10.5%	8.3%	11.4%	8.53%
BSWMA-1-ARIMA	3.8%	8.5%	10.5%	8.2%	11.2%	8.43%
BSWMA-*m*-ARIMA	3.6%	8.1%	9.9%	7.6%	10.3%	7.91%
Average	3.8%	8.6%	11.3%	8.1%	12.2%	
Linear Regression	7.6%	11.2%	13.4%	11.1%	18.3%	12.32%

最後，模型配適資料採用 1992 年到 2017 年前後總計 26 年期間的癌症死亡率，以改良式趨勢延伸演算法「BSWMA-*m*-ARIMA」進行未來 2018-2022 年全癌症死亡率的預測與推估，在此以台灣六都為例，得到表 4 預測之結果。並繪製六都全癌症死亡率預測趨勢圖，如圖 5 所示，橫軸是癌症死亡年（1993-2022 年），縱軸為 WHO 每十萬人口年齡標準化死亡率，圖中實線代表 1993-2017 年原始資料經 9BSWMA 計算處理之移動平均死亡率，虛線代表未來 2018-2022 年死亡率之預測值。

表 4. 以「BSWMA-*m*-ARIMA」演算法預測台灣六都未來五年全癌症死亡率

癌症別	區域別	預測每十萬人死亡率(年)				
		2018	2019	2020	2021	2022
全癌症	台　灣	124.18	124.04	123.86	123.88	123.88
	台北市	100.50	100.60	100.11	100.12	100.12
	新北市	114.31	114.07	113.94	113.95	113.96
	桃園市	110.03	109.79	109.66	109.67	109.70
	台中市	124.35	124.24	124.19	124.05	123.88
	台南市	137.00	136.83	136.95	137.22	137.31
	高雄市	136.29	136.17	136.12	136.06	136.12

根據預測結果，我們想提供相關單位及地方政府一些思考與後續研究空間。若仔細觀察圖 5 台灣六都全癌症發生趨勢，墨綠及綠線的高雄及台南近二十幾年來都一直在六都之冠，除了受平均壽命增加的影響之外，是否有其他長期大環境的因素存在，需要搭配政策積極改善。反倒是人口密度更高的台北市標準化死亡率（圖 5 橙線）從二十年前開始急速下降，並與桃園市和新北市都還維持在台灣全區平均趨勢線（圖 5 淺藍線）以下，近幾十年是否在公共政策或大環境上有甚麼樣的改變與進步，得以讓癌症死亡率快速下降，確實值得相關單位後續的探討，並做為其他縣市的借鏡。

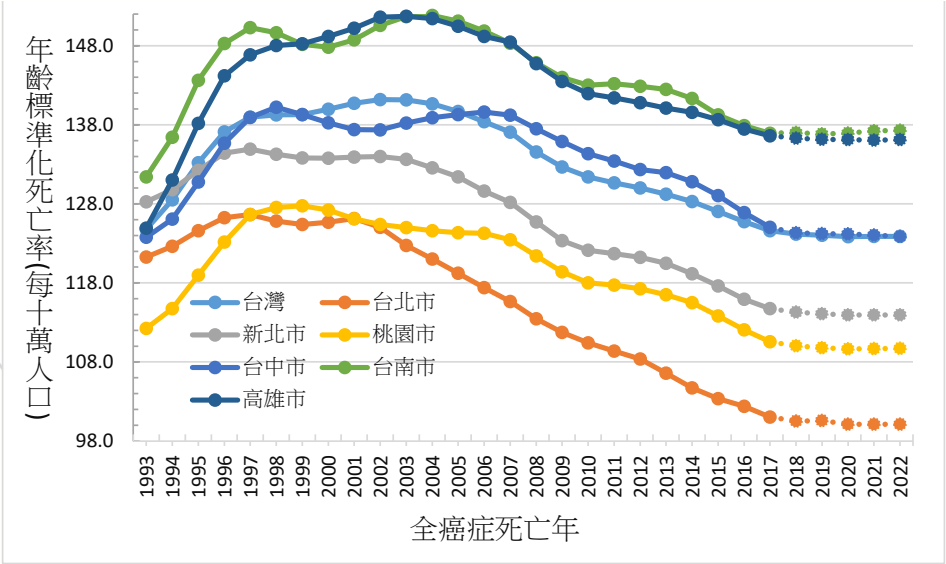

圖 5. 台灣六都 2018-2022 年全癌症死亡率預測趨勢圖

雖然從圖 5 的線形來看，近十幾年台灣地區及六都全癌症年齡標準化死亡率都有和緩下來的趨勢，其實，若將歷年六個直轄市未進行年齡標準化的粗死亡率（如圖 6）拿出來作比較，便可知道台灣及六都的粗死亡率幾乎是線性上升的，這是因為人口老化的問題，平均壽命（零歲平均餘命）的延長，使得年長者死於癌症的機率升高。從圖 7 台灣歷年全癌症粗死亡率與年齡標準化的關係圖中，可以看出，粗死亡率幾乎與癌症死亡平均年齡呈現等比例上升，而年齡標準化死亡率卻沒有增長的趨勢、甚至下降，這意謂老年人口相較於年輕族群、死於癌症的比例是逐年提高的，相關單位更應該重視銀髮族群癌症的問題，我們是否對年長者忽略了甚麼應該做的事。

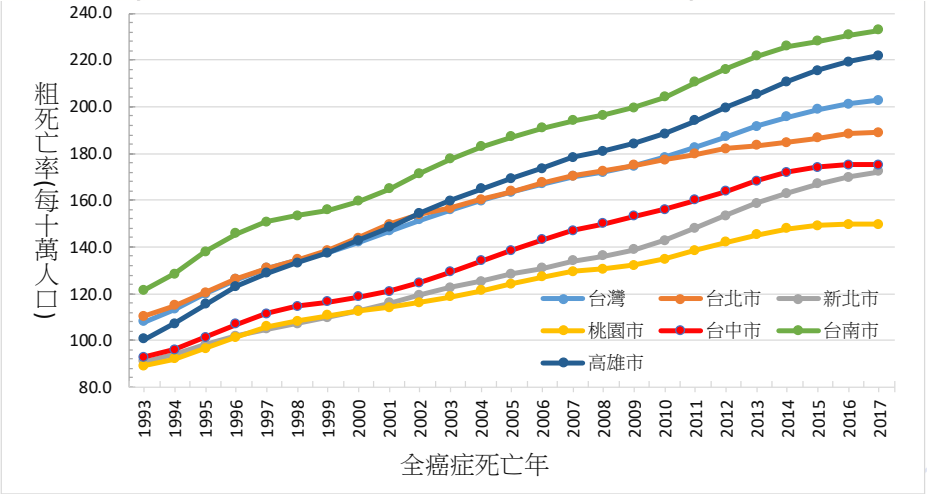

圖 6. 台灣六都歷年全癌症粗死亡率趨勢圖（經 9BSWMA 平滑處理）

仔細比較圖 5 與圖 6，有一個比較特別的現象，在圖 5 中台北市全癌症年齡標準化死亡率的趨勢遠低於其他五都，但在圖 6 的粗死亡率比較中，可以看到台北遠高於台中、新北及桃園，誠如上述，台北市老年人口相較於年輕族群、死於癌症的比例是逐年提高、而且反差更大，儘管可能因老年人口增多所致，但是否更應該重視老人的健康宣導、癌症篩檢、早期與妥善治療的福利為依歸。

另外，我們特別針對女性乳癌及男性口腔癌死亡趨勢進行預測，圖 8(a)(b)所示，圓圈標示紅色虛線為原始死亡率，綠色粗線代表移動平均後的死亡率（1993-2012 年），藍色粗線則是預測結果（2013-2017 年），儘管台灣大部分癌症年齡標準化死亡率都有逐漸和緩的趨勢，但女性乳癌及男性口腔癌死亡率卻是節節升高，有待相關單位加以重視及預防。

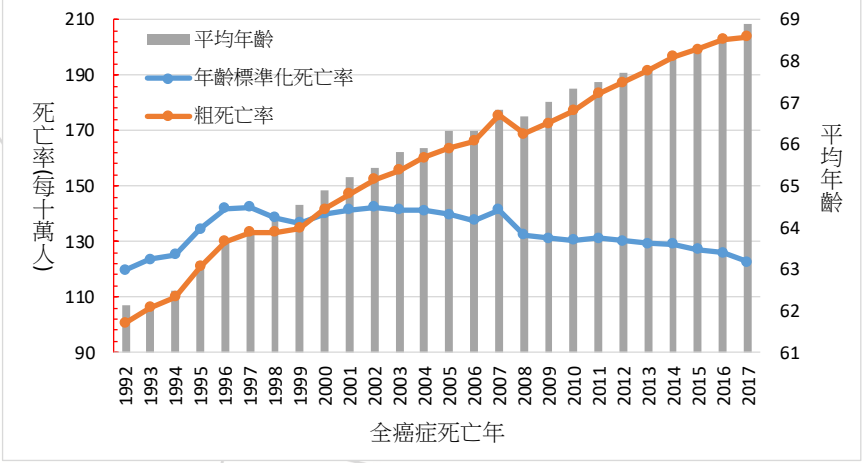

圖 7. 台灣歷年全癌症粗死亡率與年齡標準化關係圖

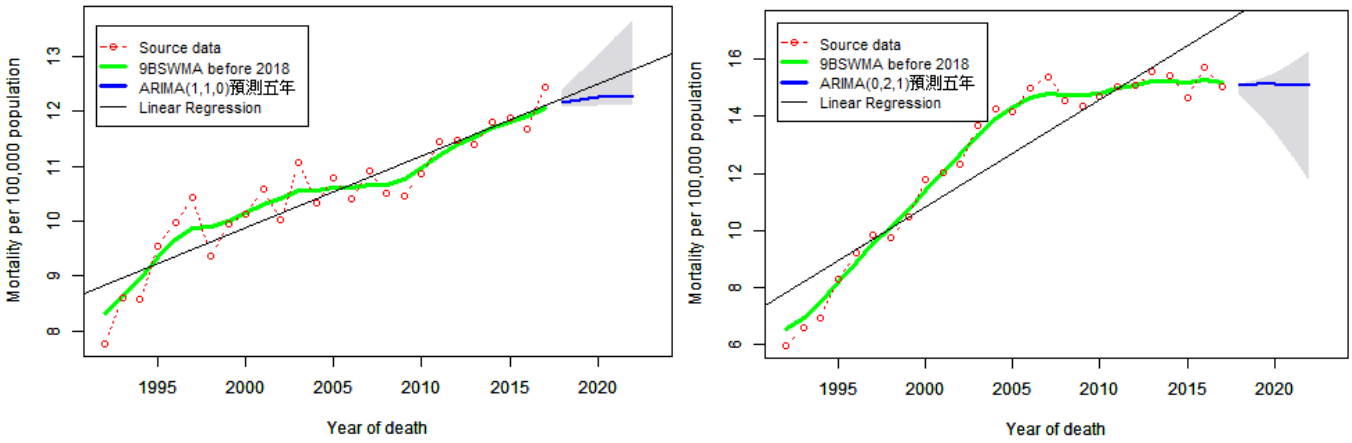

圖 8(a)(b). 台灣 2018-2022 年女性乳癌及男性口腔癌死亡率預測趨勢圖

4. 結論

本研究對傳統趨勢延伸演算法 ARIMA 模型進行改良，我們將台灣各縣市歷年癌症死亡率分為訓練及測試資料，比較幾種不同方法的預測效能，預測準確性的評估則是以 MAPE 為主。透過擴大樣本測試與統計檢定，改良式趨勢延伸演算法 BSWMA-*m*-ARIMA 絕大多數的情況可獲得高精確度的預測能力(MAPE<10%)，以及最低的平均誤差，確實可以提升預測準確性，且與傳統 ARIMA 演算法之間有統計上的顯著差異。因此，雖然癌症死亡趨勢難以預測，但若善用一些技巧，更精準的預測並非全然不可得。

最後，我們以 BSWMA-*m*-ARIMA 預測台灣六都 2018 年至 2022 年未來五年全癌症死亡率，後續也可以進一步用來對其他縣市及個別癌症進行趨勢預測，提供各項癌症死亡的可能趨勢與變化，給政府相關單位事先了解及其政策規劃上的參考，例如：加強民眾對健康飲食、健康生活型態、及癌症防治宣導，事前預防、避免罹癌；癌症篩檢機制的改善與精進，早期發現、早期治療，有效降低癌症死亡率；針對高癌症死亡率的地區環境，加強各式汙染監測，改善地區民眾的生活環境；對於罹癌者都能以同理心，獲得相關單位在身心各方面的支持與妥善治療，以提高癌後的存活率與生活品質。

到目前為止，人類對於癌症的了解仍存在著很多疑問，尚有許多不明的致癌因子。本研究的限制在於尚未考慮癌症發生的可能原因，癌症死亡率預測結果僅能提供最保守的參考意見，寧願過度警示、提醒有關單位的重視。未來期待更多專家學者能夠明確找到特定癌症致病或保護的真正原因，預防重於治療，促進人類福祉。

參考文獻

1. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.
2. Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical-Theoretic Approach (The 2nd edition)*. Springer-Verlag.
3. Chuang, C.-H., Lu, W.-F., Lin, Y.-C., & Chen, J.-C. (2018). Visual Exploration Using Improved Moving Average Methods for Time Series Datasets. *International Journal of Electronics and Information Engineering (IJEIE)*, *9*(1), 46-60.
4. Chuang, C.-H., Mong, M.-C., Wang, C.-Y., & Chen, J.-C. (2018). Study on Parameter Optimization for a ARIMA Model in R Packages Using Government Open Data. *Proc. of the 2018 International Conference on Information Technology and Industrial Application (ITIA 2018)*, Session S03-3: IT and Applications, 96-103.
5. He, Z., & Tao, H. (2018). Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study. *International Journal of Infectious Diseases*, *74*, 61-70.
6. Hill, A. V. (2011). *The Encyclopedia of Operations Management: A Field Manual and Glossary of Operations Management Terms and Concepts*. FT Press.
7. Hyndman, R. J., & Athanasopoulos, G. (2014). Measuring forecast accuracy, Section 2.5 of Forecasting: principles and practice. 2018 年 1 月 5 日取自 http://www.otexts.org/fpp/2/5
8. Hyndman, R. J., O'Hara-Wild, M., Bergmeir, C., Razbash, S., & Wang, E. et al. (2017). Package 'forecast'. 2018 年 7 月 18 日取自 https://cran.r-project.org/web/packages/forecast/forecast.pdf
9. Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773-795.
10. Lewis, C. D. (1982). *Industrial and Business Forecasting Methods*. Butterworths, London.
11. Schwarz, G. E. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461-464.

12. Wikipedia (2019). Autoregressive integrated moving average. 2019 年 9 月 25 日取自 https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average
13. Xu, Y.-T. (2015). Application of ARIMA model in forecasting the mortality of mumps. *South China Journal of Preventive Medicine*, *41*(3), 255-259.
14. 行政院內政部統計處(2018)。統計資料動態查詢主目錄：土地與人口概況、地區別平均餘命、我國生命表統計。2018 年 9 月 8 日取自 http://statis.moi.gov.tw/micst/stmain.jsp?sys=100
15. 時照華、蘇虹、秦鳳雲(2013)。ARIMA 模型在常見呼吸道傳染病疫情預測中的應用。*安徽醫科大學學報*，*48*(7)，783-786。
16. 國家發展委員會(2018)。政府資料開放平臺(Taiwan Open Government Data)。2018 年 9 月 1 日取自 https://data.gov.tw/
17. 劉祥雯(2016)。以統計模型分析肺癌發生率及死亡率之長期趨勢。國立臺灣大學流行病學與預防醫學研究所碩士論文，台北市。
18. 衛生福利部統計處(2018)。2017 年死因統計結果。2018 年 9 月 1 日取自 https://www.mohw.gov.tw/cp-3795-41794-1.html
19. 衛福部國健署(2019)。癌症防治，降低癌症死亡率。2019 年 2 月 5 日取自 https://www.hpa.gov.tw/Pages/List.aspx?nodeid=47,/Pages/Detail.aspx?nodeid=1125&pid=1688

Using an Improved ARIMA Algorithm to Forecast the Cancer Mortality in Taiwan for the Next Five Years

Chuang, C.-H.[1, 2], Lu, W.-F.[1, 3], *Chen, J.-C.[1]
[1] Department of Computer Science and Information Engineering, Asia University
[2] Department of Medical Research, China Medical University Hospital, Taiwan
[3] Department of Bioinformatics and Medical Engineering, Asia University

Abstract

The first cause of death in the elderly is also cancer, and unfortunately, the percentage of such death in people over 65 years old is increasing with days. The statistics and analysis of the cause of death have become significant for planning of public health policies and improving the overall health of the people of Taiwan. Trend extension algorithm is often used for quantitative disease forecasting. One such model is the Autoregressive Integrated Moving Average (ARIMA) model. This model can be used to forecast time series data, especially for problems where random process characteristics change over time and causes the time series to be non-stationary and random. At present, most of the datasets on cancer mortality in various cities and counties in Taiwan belong to non-stationary and non-seasonal time-series data. They are suitable to use in the trend extension algorithm to forecast future trends. In order to further improve the forecast accuracy, the present study improved the traditional ARIMA algorithm. First, the cancer mortality data announced by the Ministry of Health and Welfare, Taiwan, for the past 26 years from 1992 to 2017 were divided into training and test data to compare the accuracy of different improved forecast algorithms. One of the forecasting performance evaluation methods is to estimate the forecast accuracy of each algorithm via Mean Absolute Percentage Error (MAPE). Finally, the best improved ARIMA algorithm was then used to forecast the cancer mortality in Taiwan for the next five years. The forecast results will provide the relevant government agencies with prior knowledge of the possible trends of cancer mortality and act as a reference for policy planning. These would allow people (especially the elderly) to receive appropriate cancer screening mechanisms, and those who already have cancer can get proper treatment, reduce cancer mortality, and improve their quality of life.

Keywords: elder health, cancer mortality, trend extension algorithm, ARIMA model