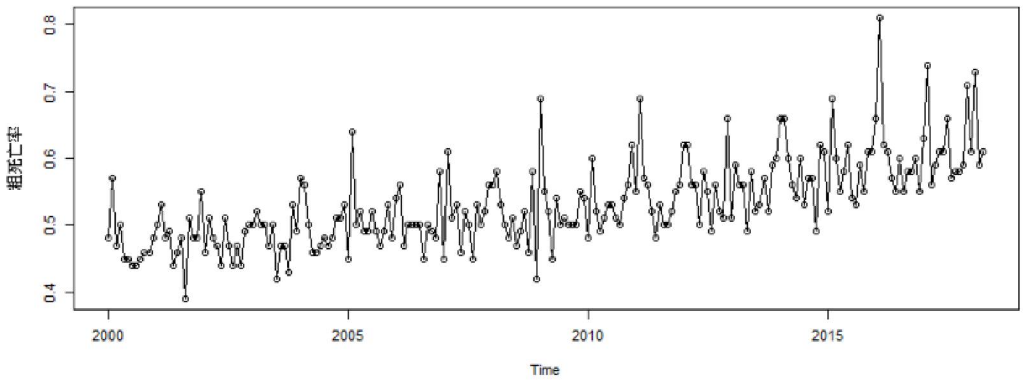
時間數列_粗死亡率

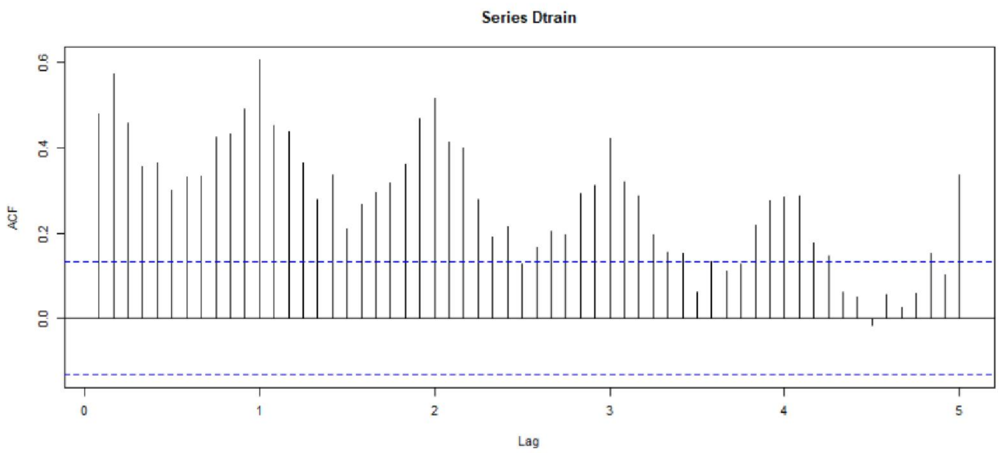
1. 「粗死亡率」定義 ：平均每千人口中的死亡數

統計資料從 2000 年一月至 2019 年四月，這裡將資料切為兩個部份 ：

Training : 2000 年一月至 2018 四月，Testing : 2018 年五月至 2019 四月

下圖是粗死亡率對時間的對照圖及 ACF(Auto-Correlation Function)，由對照圖可看出粗死亡率對時間有些微的上升趨勢，而由 ACF 可看出粗死亡率的自我相關具有季節性趨勢(死亡可能受氣候影響)，因此可得知想要讓資料從 non-stationary 變為 Stationary，需要做削去趨勢的工作(de-trend)，而這裡需要消除的趨勢就是線性趨勢及季節趨勢。





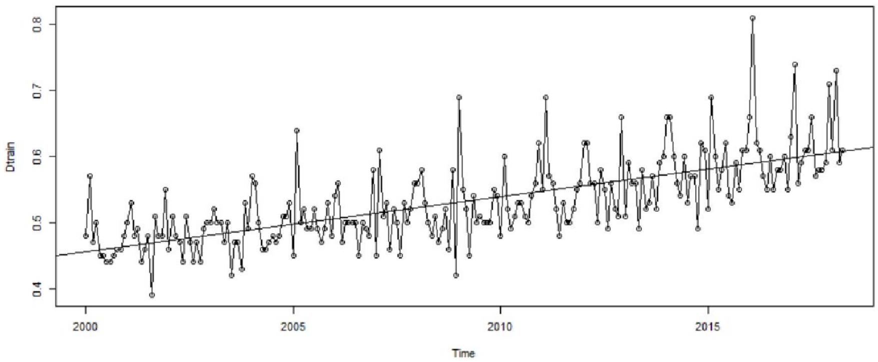
2. De-trend

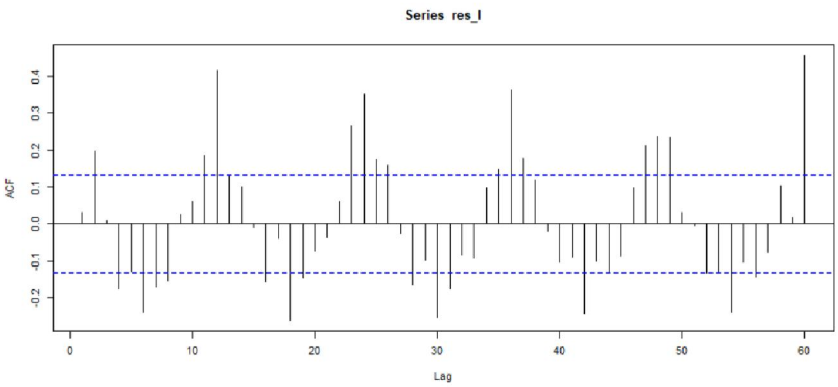
2-1. Linear Trend Model

$$\mu_t = \beta_0 + \beta_1 t$$

從下圖表可看出線性模型的斜率項 p-value 極低，代表這個線性趨勢是極為顯著的，但觀察殘差的 ACF 可發現自我相關超出範圍，不符合 White Noise 的假設，代表粗死亡率的時間數列並不能被單單一個線性模型解釋完全。但從殘差的 ACF 可得知去除線性趨勢後的殘差的自我相關有季節性趨勢，這個現象與前面的假設不謀而合。

```
Call:
lm(formula = Dtrain ~ time(Dtrain))

Residuals:
      Min        1Q    Median        3Q       Max 
-0.110853 -0.032628 -0.006659  0.019340  0.219354 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.623e+01  1.208e+00  -13.43   <2e-16 ***
time(Dtrain)  8.343e-03  6.014e-04   13.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04721 on 218 degrees of freedom
Multiple R-squared:  0.4689,	Adjusted R-squared:  0.4664 
F-statistic: 192.4 on 1 and 218 DF,  p-value: < 2.2e-16
```





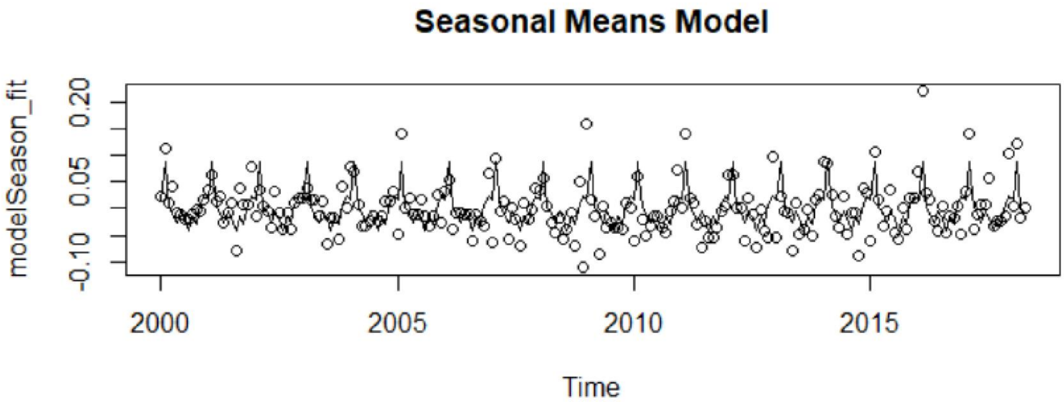
2.2 Seasonal-Trend

$$Y_t = \mu_t + X_t \ \ ;where\ \ E(X_t) = 0\ \ for\ \ all\ \ t$$

$$\mu_t = \begin{cases} \beta_1\ ;\ for\ \ t = 1,13,25 \\ \beta_2\ ;\ for\ \ t = 2,14,26 \\ \quad\quad\quad\vdots \\ \beta_{12};\ for\ \ t = 12,24,36 \end{cases}$$

將 2.1 去除過 linear trend 的殘差去配適 Seasonal-Trend Model，可發現在二月時係數為正，代表二月在平均上是死亡的高峰期，而八月是負最多，代表八月份是死亡的低谷。

```
Call:
lm(formula = res_l ~ month.)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.134162 -0.018648  0.001443  0.017052  0.140526 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      0.017926   0.008037   2.230 0.026796 *  
month.February   0.068778   0.011367   6.051 6.61e-09 ***
month.March     -0.017180   0.011367  -1.511 0.132190    
month.April     -0.024717   0.011367  -2.175 0.030790 *  
month.May       -0.045305   0.011523  -3.932 0.000115 ***
month.June      -0.023223   0.011523  -2.015 0.045161 *  
month.July      -0.035585   0.011523  -3.088 0.002289 ** 
month.August    -0.059058   0.011523  -5.125 6.78e-07 ***
month.September -0.031420   0.011523  -2.727 0.006945 ** 
month.October   -0.048226   0.011523  -4.185 4.21e-05 ***
month.November  -0.010033   0.011523  -0.871 0.384960    
month.December   0.005383   0.011523   0.467 0.640871    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03503 on 208 degrees of freedom
Multiple R-squared:  0.4746,	Adjusted R-squared:  0.4468 
F-statistic: 17.08 on 11 and 208 DF,  p-value: < 2.2e-16
```

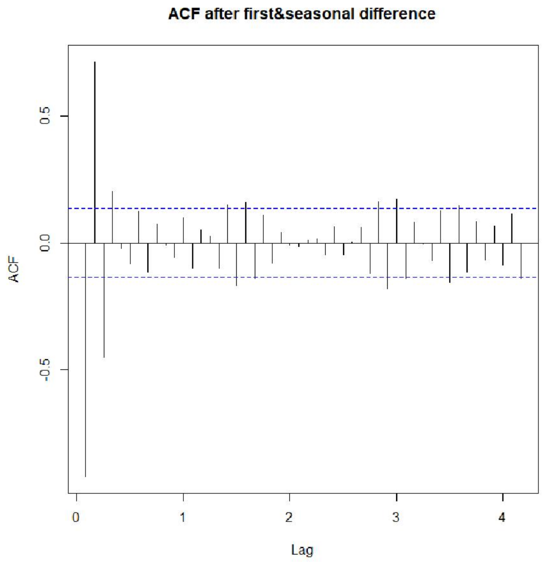


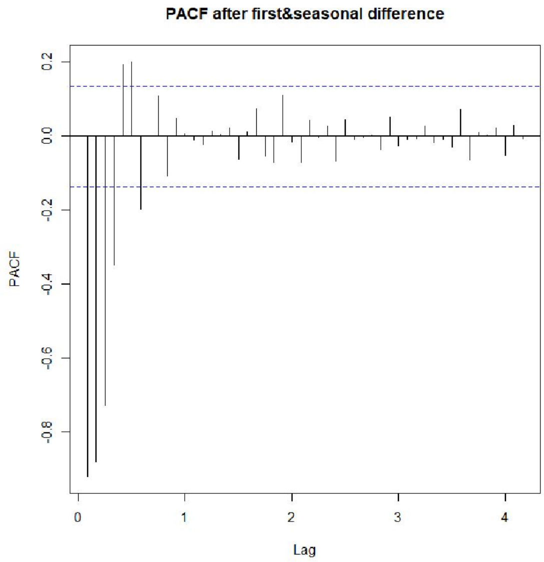
3. ARMA

消除線性趨勢及季節性趨勢，與對時間數列做一次差分及一次季節性差分為等

價的動作，因此將兩次 De-trend 模型簡化為一次$SARIMA(0,1,0) \times (0,1,0)_{12}$

再觀察殘差的 ACF 及 PACF：



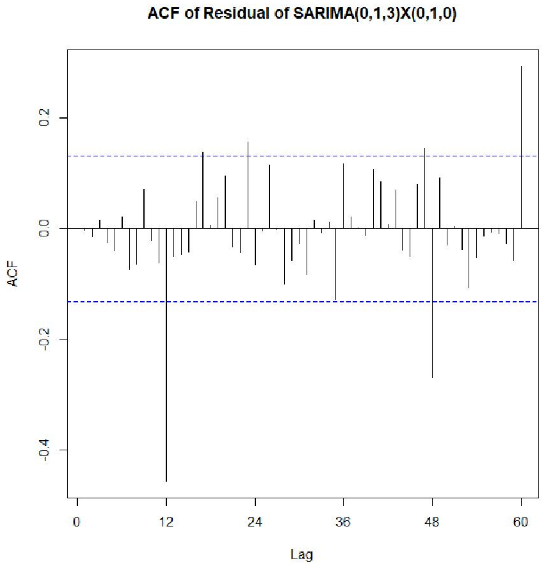


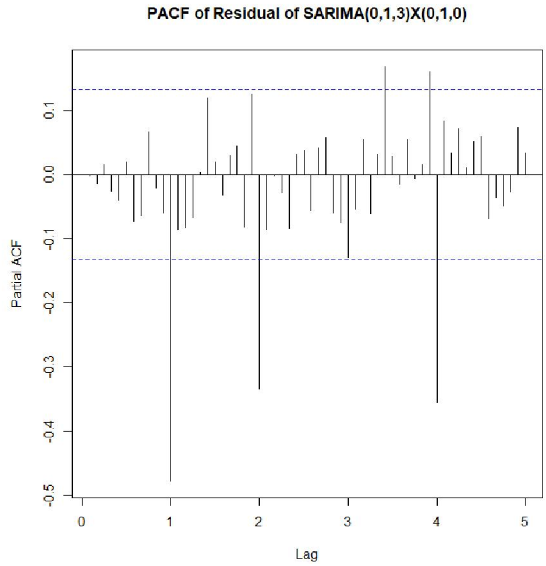
不管是從 ACF 還是 PACF 來看都可看出前三期有高度的自我相關，ACF 看的出

來從第三期之後有切斷(Cut-off)的趨勢，PACF 看的出來從第三期以後有漸近

消失的趨勢，符合 MA(3)的特性，因此嘗試配適$SARIMA(0,1,3) \times (0,1,0)_{12}$

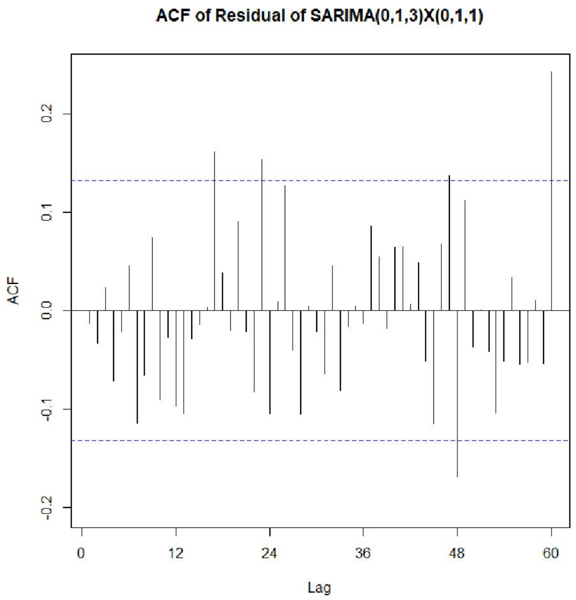
```
Call:
arima(x = Dtrain, order = c(0, 1, 3), seasonal = list(order = c(0, 1, 0), period = 12))

Coefficients:
          ma1     ma2      ma3
      -1.2215  0.4516  -0.2301
s.e.   0.0699  0.1052   0.0689

sigma^2 estimated as 0.002287:  log likelihood = 332.88,  aic = -659.75

Training set error measures:
              ME RMSE MAE MPE MAPE
Training set NaN  NaN NaN NaN  NaN
```

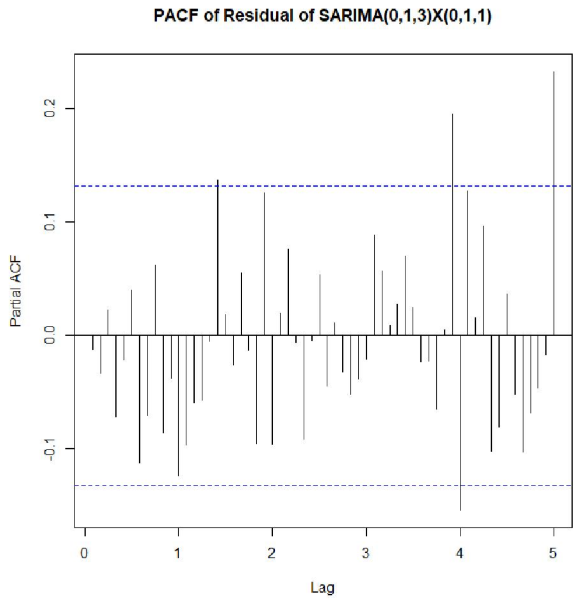


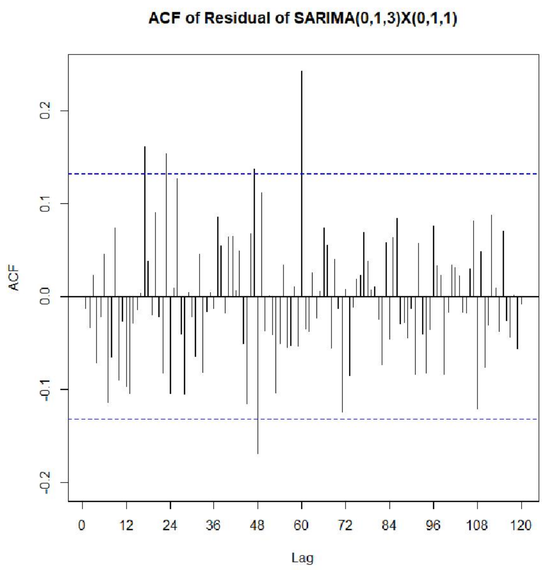


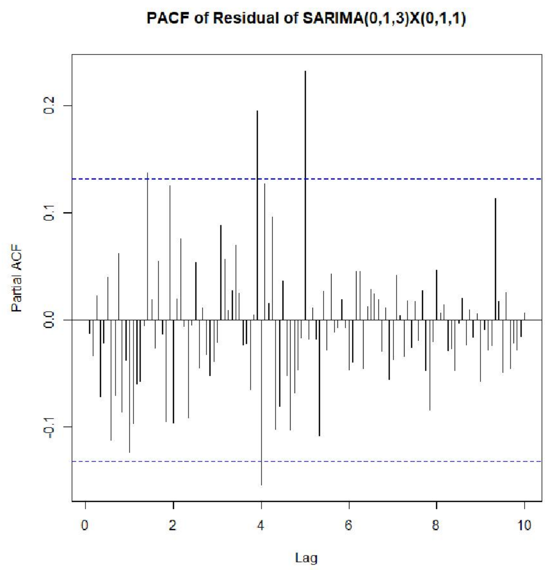
觀察殘差的 ACF 及 PACF，可發現在整年的期數有高度的自我相關，代表季節性趨勢並未被解釋完全，故 Seasonal 應有 AR(1)或 MA(1)，而 ACF 圖比起 PACF 較有切斷(Cut-off)的趨勢，因此嘗試配適$SARIMA(0,1,3) \times (0,1,1)_{12}$

```
Call:
arima(x = Dtrain, order = c(0, 1, 3), seasonal = list(order = c(0, 1, 1), period = 12))

Coefficients:
          ma1     ma2      ma3     sma1
      -1.2313  0.4649  -0.1904  -0.9366
s.e.   0.0689  0.1075   0.0706   0.0895

sigma^2 estimated as 0.001181:  log likelihood = 389.94,  aic = -771.89

Training set error measures:
              ME RMSE MAE MPE MAPE
Training set NaN  NaN NaN NaN  NaN
```



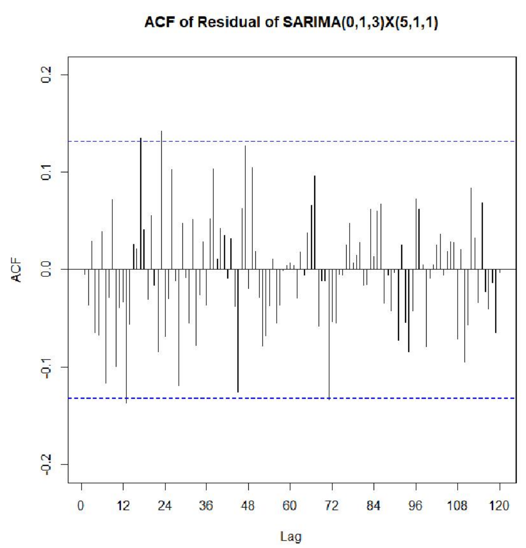


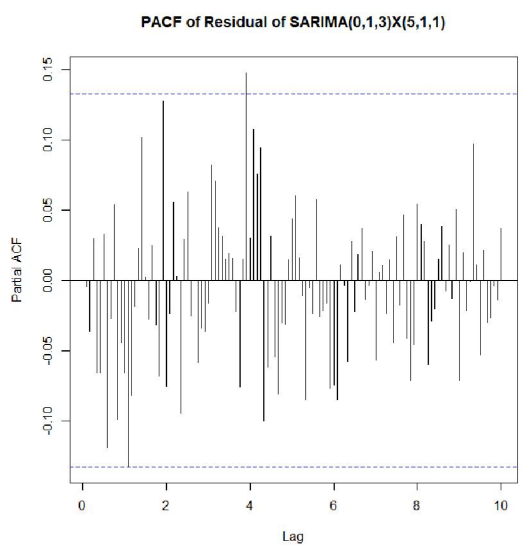




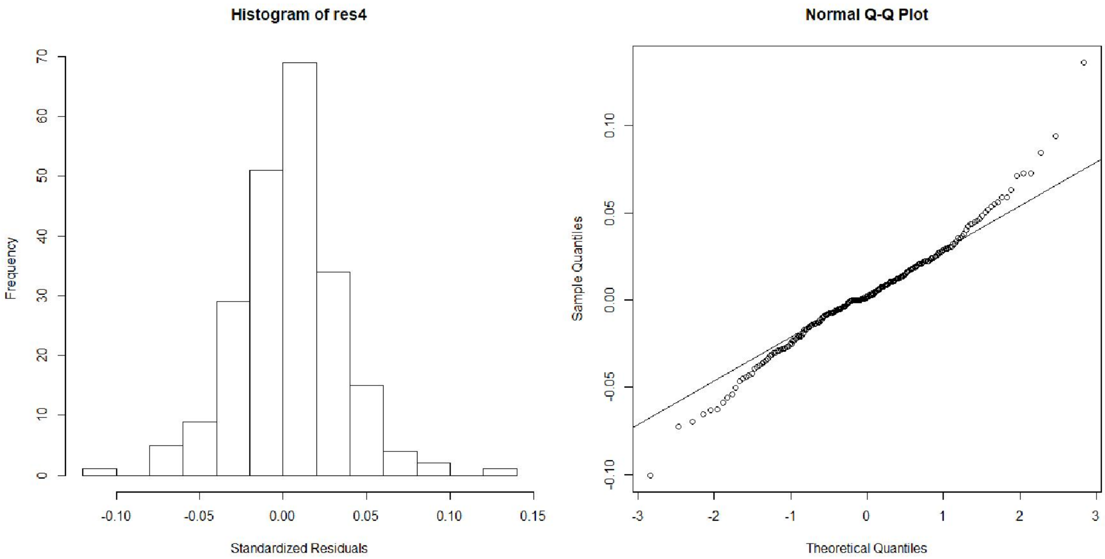
接下來的 ACF 及 PACF 走勢相當的耐人尋味，在前面的期數超出範圍的個數不多，即使超過程度也不大，但在第四年開始超出一些，第五年則有顯著的突出，將期數上限拉到 120 期(10 年)也是得到相同的結果，除了第五年之外並無顯著的突出，因此決定配適$SARIMA(0,1,3)\times(5,1,1)_{12}$

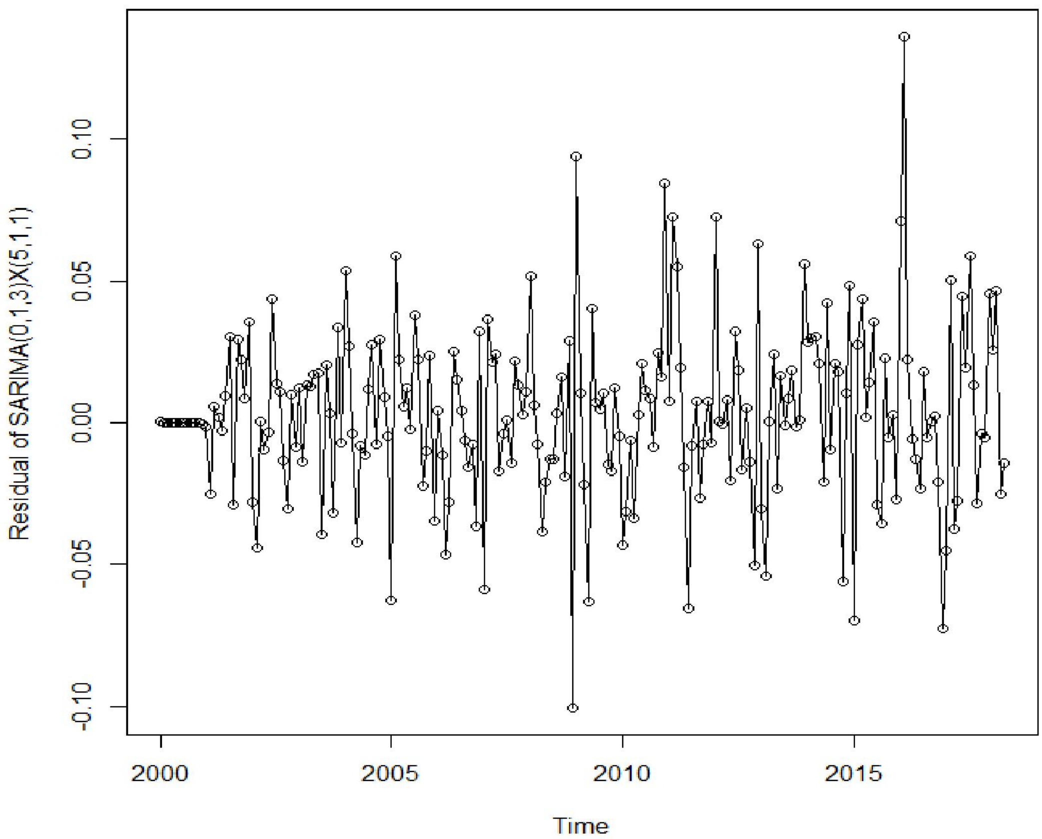
```
Call:
arima(x = Dtrain, order = c(0, 1, 3), seasonal = list(order = c(5, 1, 1), period = 12))

Coefficients:
          ma1     ma2      ma3     sar1     sar2     sar3     sar4    sar5     sma1
      -1.1406  0.3805  -0.1870  -0.2118  -0.2996  -0.1466  -0.3024  0.2014  -0.7040
s.e.   0.0760  0.1109   0.0725   0.2006   0.1686   0.1691   0.1360  0.1416   0.1819

sigma^2 estimated as 0.001001:  log likelihood = 405.54,  aic = -793.07

Training set error measures:
              ME RMSE MAE MPE MAPE
Training set NaN  NaN NaN NaN  NaN
```





由 ACF 及 PACF 我們可知$SARIMA(0,1,3) \times (5,1,1)_{12}$的殘差幾乎全數都在範圍內，也就是自我相關不顯著，做到這裡已滿足 White Noise 的獨立假設，從分佈圖及 qqpolt 來看，分佈及分位數大致上與常態分配相似，故可認定$SARIMA(0,1,3) \times (5,1,1)_{12}$的殘差符合 White Noise





4. 預測

用以 2000 年 1 月至 2018 年 4 月的資料所建立的$SARIMA(0,1,3) \times (5,1,1)_{12}$來預測往後一年(2018 年 5 月至 2019 年 4 月)的粗死亡率，並對真實值進行比較

時間	預測值	真實值	預測誤差絕對值	95%下界	95%上界
May 2018	0.5752602	0.61	0.0347398	0.5132460	0.6372744
Jun 2018	0.6267129	0.55	0.0767129	0.5640885	0.6893372
Jul 2018	0.5975384	0.61	0.0124616	0.5331721	0.6619048
Aug 2018	0.5788529	0.6	0.0211471	0.5144033	0.6433026
Sep 2018	0.6169314	0.55	0.0669314	0.5523985	0.6814642
Oct 2018	0.5902818	0.63	0.0397182	0.5256659	0.6548977
Nov 2018	0.6324991	0.58	0.0524991	0.5678003	0.6971980
Dec 2018	0.6689096	0.55	0.1189096	0.6041278	0.7336913
Jan 2019	0.6845001	0.7	0.0154999	0.6196356	0.7493646
Feb 2019	0.7249033	0.57	0.1549033	0.6599562	0.7898505
Mar 2019	0.6451538	0.64	0.0051538	0.5801241	0.7101835
Apr 2019	0.6252999	0.64	0.0147001	0.5601877	0.6904120

從上表可知大部分的預測值與真實值相差不大，都有在 95%預測區間之內，但明顯有差距的是在 2018 年的 12 月及 2019 年的 2 月，預測值與真實值分別是 (0.6689096 , 0.55)及(0.7249033 , 0.57)真實值比預測值低了許多，從 Seasonal Trend Model 我們可以從係數看出二月及十二月是死亡的高峰，而 $SARIMA(0,1,3) \times (5,1,1)_{12}$預測出來的預測值也是偏高的，相比之下 2018/12 及 2019/2 的粗死亡率真實值反而是異常的低，除了可能還有某種趨勢是 $SARIMA(0,1,3) \times (5,1,1)_{12}$無法解釋的之外，也有可能是出現無法被時間解釋的變異，但目前無法找出發生變異的原因。

5. 結論

我們用 2000/1~2018/4 的粗死亡率去配適一個最適合的時間數列模型，從一連串的殘差分析，最終找到使殘差最接近 White Noise 的$SARIMA(0,1,3) \times (5,1,1)_{12}$，除了可以做為未來的預測外，模型本身也是有可以解釋的地方，由於人口數穩定成長，因此即使醫療科技在進步，粗死亡率也會隨時間慢慢攀升，而粗死亡率的季節性趨勢有可能是受氣候影響，在氣溫驟降的冬天粗死亡率是比較高的(在地球暖化的往後，可能反而是夏天的高溫會成為致命的殺手)，MA(3)可能是因為季節在三個月內氣候較相似，因此會受前三期的變化所影響，至於 seasonal 的 AR(5)及 MA(1)就較難解釋。另外在做死亡的時間數列研究時發現死亡與氣溫息息相關，或許往後研究也可以將氣溫的時間數列作為自變數解釋粗死亡率。

R Code

```
library(data.table)
library(TSA)
library(timeSeries)
library(forecast)
library(locfit)
library(tseries)
D<-read.csv("C:/Users/User/Desktop/NCCU/TimeSeries/死亡.csv",header = T)
colnames(D)<-c("Time","CDR")
Dtrain<-ts(D$CDR,start=c(2000,1),end = c(2018,4),frequency = 12)
win.graph(width=2.5,height=2.5,pointsize=8)
plot(Dtrain,ylab='粗死亡率',type='o')
acf(Dtrain,lag.max = 60)

#Building linear model for CDR-time series####
model_l=lm(Dtrain~time(Dtrain))
win.graph(width=2.5,height=2.5,pointsize=8)
plot(Dtrain,type='o')
abline(model_l)
res_l=residuals(model_l)
Acf(rstudent(model_l),lag.max = 60)
Acf(res_l,lag.max = 60)

#Building seasonal mean model for CDR-time series( With intercept )####
res_l<-ts(res_l,start=c(2000,1),frequency = 12)
month.=season(res_l) # period added to improve table display
modelSeason=lm(res_l~month.) # January is dropped automatically

res_ls=residuals(modelSeason)
Acf(res_ls,lag.max = 60)
pacf(res_ls,lag.max = 60)

adf.test(res_ls, alternative = c("stationary"),k = 1)

modelSeason_fit = ts(fitted(modelSeason),start = c(2000,1),freq = 12)

ts.plot(modelSeason_fit, main = 'Seasonal Means Model',
        ylim = c(min(res_l),max(res_l)));points(res_l, col = 'black')
```

```r
#Building SARIMA Model####
par(mfrow=c(1,2))
model1 <-arima(Dtrain,order=c(0,1,0),seasonal=list(order=c(0,1,0),period=12))
res1<-residuals(model1)
Acf(res1,lag.max = 60,main="ACF of Residual of SARIMA(0,1,0)X(0,1,0)")
pacf(res1,lag.max = 60,main="PACF of Residual of SARIMA(0,1,0)X(0,1,0)")

de_trend = diff(diff(Dtrain),differences = 12)
par(mfrow=c(1,2))
acf(de_trend,lag.max=50,xlab="Lag",ylab="ACF",
    main="ACF after first&seasonal difference")
acf(de_trend,lag.max=50,xlab="Lag",ylab="PACF",
    type="partial",
    main="PACF after first&seasonal difference")

model2 <-arima(Dtrain,order=c(0,1,3),seasonal=list(order=c(0,1,0),period=12))
res2<-residuals(model2)
Acf(res2,lag.max = 60,main="ACF of Residual of SARIMA(0,1,3)X(0,1,0)")
pacf(res2,lag.max = 60,main="PACF of Residual of SARIMA(0,1,3)X(0,1,0)")

model3 <-arima(Dtrain,order=c(0,1,3),seasonal=list(order=c(0,1,1),period=12))
res3<-residuals(model3)
Acf(res3,lag.max = 120,main="ACF of Residual of SARIMA(0,1,3)X(0,1,1)")
pacf(res3,lag.max = 120,main="PACF of Residual of SARIMA(0,1,3)X(0,1,1)")

model4 <-arima(Dtrain,order=c(0,1,3),seasonal=list(order=c(5,1,1),period=12))
res4<-residuals(model4)
Acf(res4,lag.max = 120,main="ACF of Residual of SARIMA(0,1,3)X(5,1,1)")
pacf(res4,lag.max = 120,main="PACF of Residual of SARIMA(0,1,3)X(5,1,1)")

#Forecast####
D_forecast = as.data.frame(forecast(Dtrain,model = model4,12))
```