統計學一下
期末報告

第三組

資管二 B09705010 吳和謙
資管二 B09705014 吳欣紘
資管二 B09705025 徐懷山
資管二 B09705026 詹景棠
資管二 B09705030 柯師為
資管二 B09705031 李奕杰
資管二 B09705039 劉惟恩
資管二 B09705044 王裕勳

研究主題

翡翠水庫未來進水量預測分析

研究動機

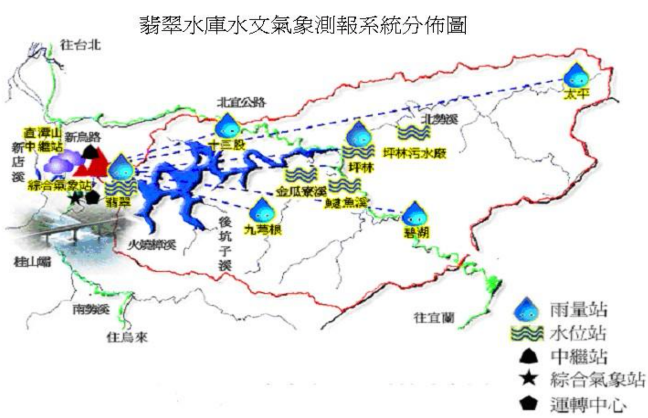
因全球氣候變遷，導致台灣近10年來極端降雨頻傳，未來極端降雨發生的頻率和強度都有可能持續攀升。面對極端氣候可能產生的極端降雨，翡翠水庫供應大台北及新北地區的用水，必須確保安全及穩定的供水，因此我們決定運用水庫進流量的歷史資料，預測未來水庫的進流量，並利用預測資料以精進水庫颱洪操作、降低下游洪災或缺水的風險。

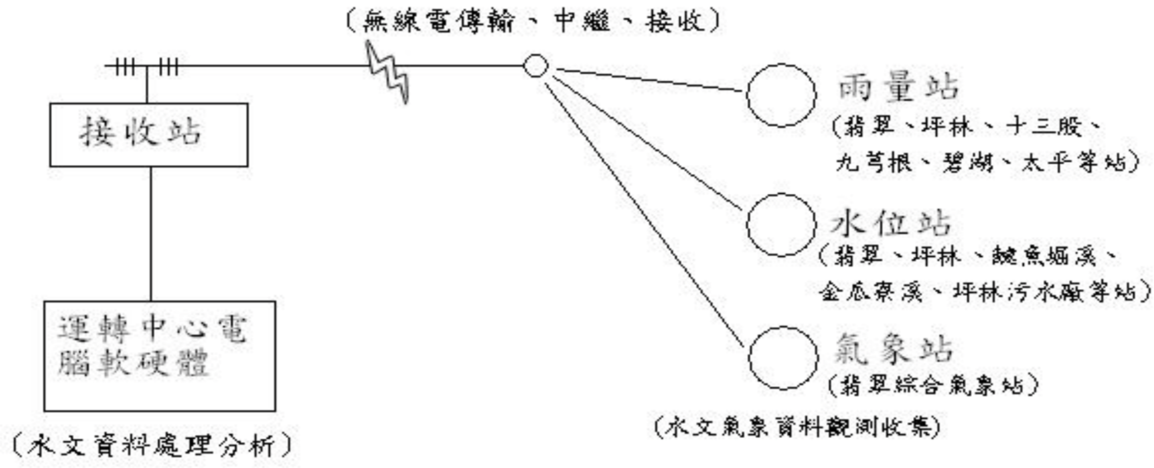
研究方法

我們搜集了 108, 109, 110 三年的翡翠水庫進流量的資料，並將資料分別整理成日進流量和月進流量兩種形式。我們用日進流量的資料進行時間序列的預測分析，以及以日為單位的季節性分析，試圖預測未來一日（111 年 1 月 1 日）的進流量，並且用月進流量的資料，試圖進行季節性的分析，預測未來一月（111 年 1 月）的平均月進流量。

數據來源

翡管局為了隨時獲得翡翠水庫集水區及上游河川區域的氣象、雨量及水位等資訊，因此設置了水文氣象測報系統。
本系統分別建置了6個雨量站、5個水位站及1個綜合氣象站，當各水文測站的雨量或氣象、水位變化值達到設定值或設定時間後，便會透過無線電自動將測值傳回翡管局運轉中心，做為水庫平時及颱洪期間水位調蓄操作參考。



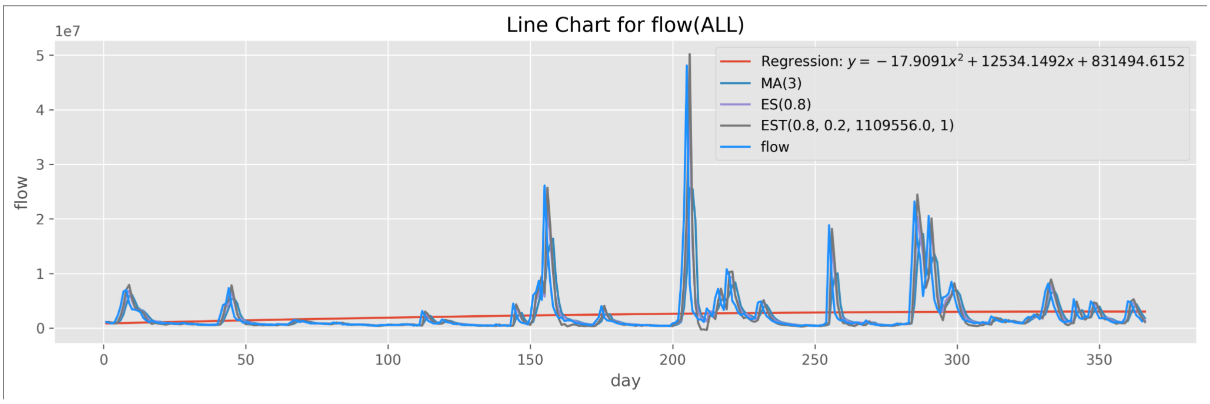


研究方法一：時間序列

第一種研究方法是時間序列的分析方法，我們採用了三種不同時段來進行時間序列分析：110 年一整年、109 和 110 兩年、108、109 及 110 三年，並相互比較何者最佳。

一、110 年

我們首先進行了以 110 年整年度每一天的進流量來做時間序列的迴歸分析，所得到的結果如下：



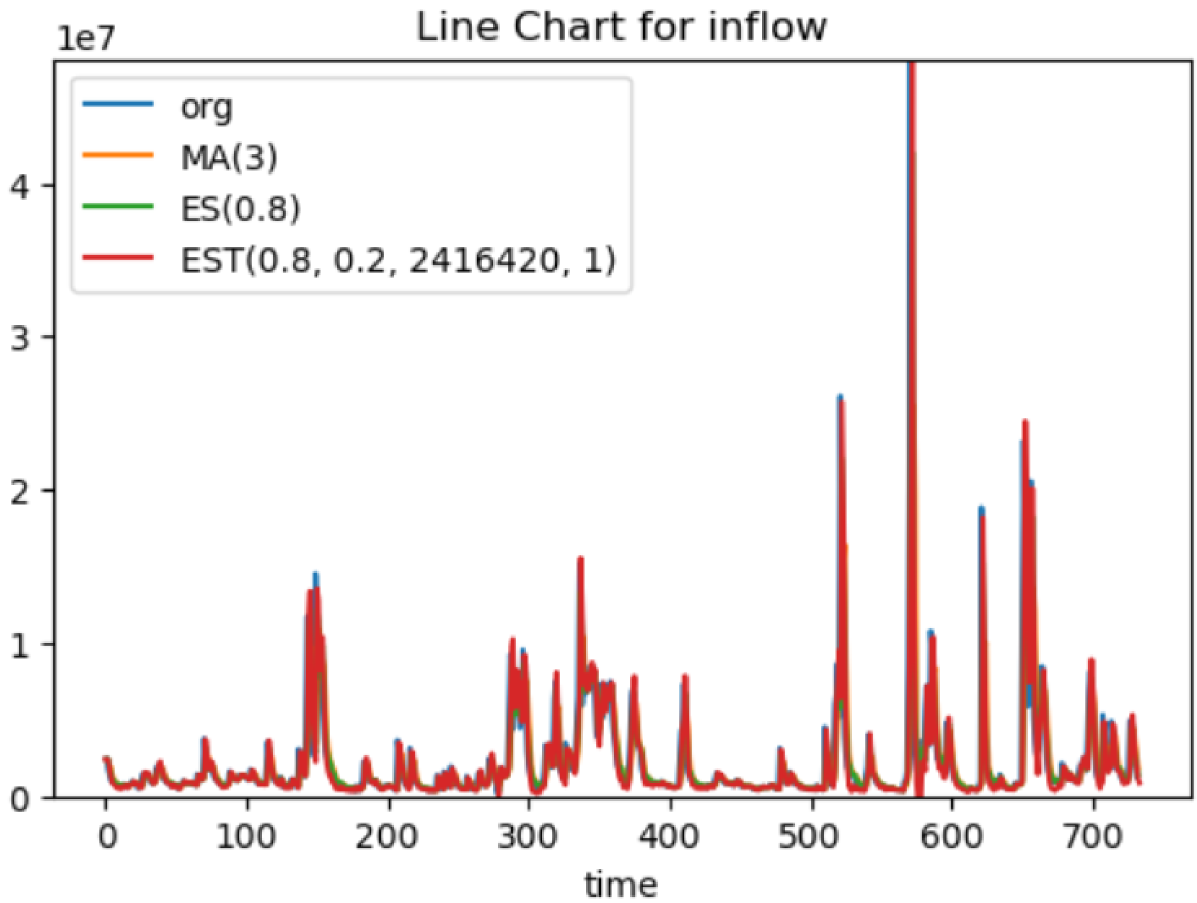
	MA(3)	ES(0.8)	EST(0.8, 0.2, 1109556.0, 1)	Regression
MAD	1.456729e+06	1.210197e+06	1.336593e+06	2.066870e+06
MSE	1.499028e+13	1.237267e+13	1.461905e+13	1.610129e+13
RMSE	3.871728e+06	3.517480e+06	3.823487e+06	4.012641e+06
MAPE	4.457701e+01	3.169461e+01	3.642625e+01	1.563212e+02

	day	flow	MA(3)	ES(0.8)	EST(0.8, 0.2, 1109556.0, 1)	Regression
0	1	1109556.0	NaN	1.109556e+06	1.109557e+06	8.440109e+05
1	2	974520.0	NaN	1.109556e+06	1.109557e+06	8.564913e+05
2	3	817460.0	NaN	1.001527e+06	9.799223e+05	8.689359e+05
3	4	1168992.0	9.671787e+05	8.542734e+05	8.023534e+05	8.813447e+05
4	5	2306772.0	9.869907e+05	1.106048e+06	1.106727e+06	8.937176e+05
...	...	...	...	...	...	...
361	362	3473152.0	3.821808e+06	4.636274e+06	5.287576e+06	3.021971e+06
362	363	2310048.0	4.371589e+06	3.705776e+06	4.085927e+06	3.021521e+06
363	364	1857808.0	3.504287e+06	2.589194e+06	2.630973e+06	3.021035e+06
364	365	1244456.0	2.547003e+06	2.004085e+06	1.854484e+06	3.020513e+06
365	366	NaN	1.804104e+06	1.396382e+06	1.110900e+06	3.019956e+06

折線圖包含四種模型和實際數值，error則是因為原始數值數量級太大，之後統一用MAPE比較。上圖可看出ES的方法擁有最低的MAPE，而除去MAPE過高的regression，其他的MAPE大致都落在30 - 40%。

二、**109** 和 **110** 年

我們接下來進行了以 109 和 110 年兩年每一天的進流量來做時間序列的迴歸分析，所得到的結果如下：

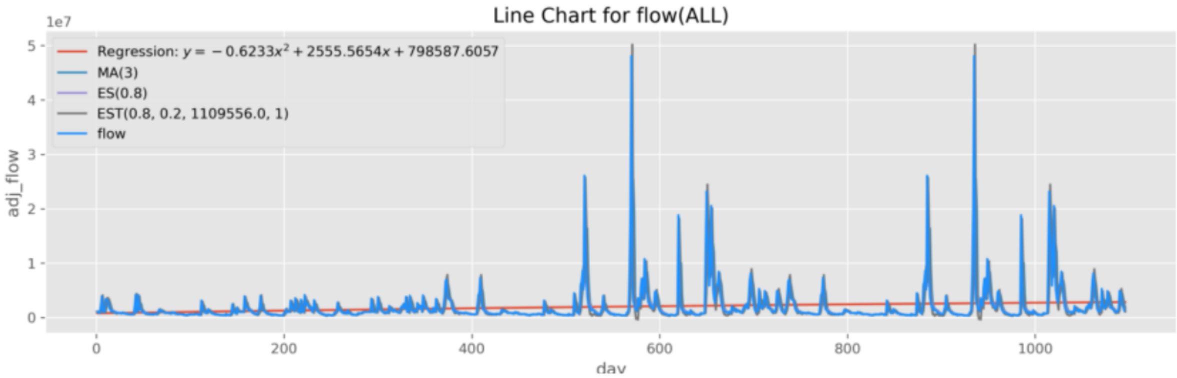


	day	flow	MA(3)	ES(0.8)	EST(0.8, 0.2, 2416420, 1)
0	1	2416420	NaN	2.416420e+06	2.416421e+06
1	2	1718880	NaN	2.416420e+06	2.416421e+06
2	3	1391012	NaN	1.858388e+06	2.416421e+06
3	4	1002260	1.842104e+06	1.484487e+06	1.746782e+06
4	5	1018292	1.370717e+06	1.098705e+06	1.293637e+06
...	...	...	...	...	...
729	730	1857808	3.504287e+06	2.589194e+06	4.085927e+06
730	731	1244456	2.547003e+06	2.004085e+06	2.630973e+06
731	732	NaN	1.804104e+06	1.396382e+06	1.854484e+06
732	733	NaN	NaN	NaN	1.110900e+06
733	734	NaN	NaN	NaN	8.553388e+05

	ErrM	MA03	ES08	ESMab
0	MAD	1132187.113553	925584.768585	1016819.195054
1	MSE	8731710304428.945	7119298526821.251	8400354186597.145
2	RMSE	2954946.751539	2668201.365493	2898336.451587
3	MAPE	40.50731	29.437464	33.091837

三、108、109 和 110 年

我們最後進行了以 108 到 110 年三年年每一天的進流量來做時間序列的迴歸分析，所得到的結果如下：

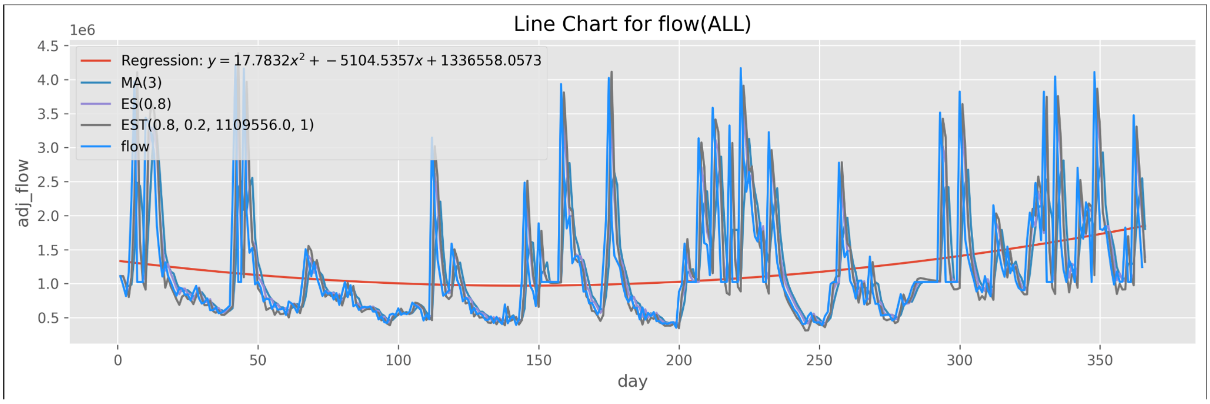


	day	adj_flow	MA(3)	flow	ES(0.8)	EST(0.8, 0.2, 1109556.0, 1)	Regression
0	1	1.10956e+06	NaN	1.10956e+06	1.109556e+06	1.109557e+06	8.011425e+05
1	2	974520	NaN	974520	1.109556e+06	1.109557e+06	8.036962e+05
2	3	817460	NaN	817460	1.001527e+06	9.799223e+05	8.062487e+05
3	4	1.16899e+06	9.671787e+05	1.16899e+06	8.542734e+05	8.023534e+05	8.087999e+05
4	5	2.30677e+06	9.869907e+05	2.30677e+06	1.106048e+06	1.106727e+06	8.113498e+05
...	...	...	...	...	...	...	...
1091	1092	3.47315e+06	3.821808e+06	3.47315e+06	4.636274e+06	5.287576e+06	2.845978e+06
1092	1093	2.31005e+06	4.371589e+06	2.31005e+06	3.705776e+06	4.085927e+06	2.847172e+06
1093	1094	1.85781e+06	3.504287e+06	1.85781e+06	2.589194e+06	2.630973e+06	2.848364e+06
1094	1095	1.24446e+06	2.547003e+06	1.24446e+06	2.004085e+06	1.854484e+06	2.849555e+06
1095	1096	NaN	1.804104e+06	NaN	1.396382e+06	1.110900e+06	2.850745e+06

	ErrM	MA03	ES08	ESMab
0	MAD	1213886.708143	980013.599847	1072020.317264
1	MSE	9803266634856.246	8032041348512.574	9464265281920.486
2	RMSE	3131016.869143	2834085.628296	3076404.603091
3	MAPE	42.809036	30.696658	34.723904

四、**110** 年（調整）

由於上述的誤差數值都不盡理想，因此我們將 110 年進流量中的極端值替換為中位數，再進行一次預測，嘗試獲得更好的結果。



	MA(3)	ES(0.8)	EST(0.8, 0.2, 1109556.0, 1)	Regression
MAD	4.454673e+05	3.943312e+05	4.378451e+05	5.419790e+05
MSE	5.728891e+11	5.503311e+11	6.506094e+11	6.018484e+11
RMSE	7.568944e+05	7.418430e+05	8.066036e+05	7.757889e+05
MAPE	3.067154e+01	2.580860e+01	2.844501e+01	5.202787e+01

	day	adj_flow	MA(3)	flow	ES(0.8)	EST(0.8, 0.2, 1109556.0, 1)	Regression
0	1	1109556.0	NaN	1109556.0	1.109556e+06	1.109557e+06	1.331471e+06
1	2	974520.0	NaN	974520.0	1.109556e+06	1.109557e+06	1.326420e+06
2	3	817460.0	NaN	817460.0	1.001527e+06	9.799223e+05	1.321404e+06
3	4	1168992.0	9.671787e+05	1168992.0	8.542734e+05	8.023534e+05	1.316424e+06
4	5	2306772.0	9.869907e+05	2306772.0	1.106048e+06	1.106727e+06	1.311480e+06
...	...	...	...	...	...	...	...
361	362	3473152.0	1.291264e+06	3473152.0	1.050625e+06	9.756904e+05	1.819093e+06
362	363	2310048.0	1.841045e+06	2310048.0	2.988647e+06	3.307582e+06	1.826881e+06
363	364	1857808.0	2.269397e+06	1857808.0	2.445768e+06	2.683871e+06	1.834705e+06
364	365	1244456.0	2.547003e+06	1244456.0	1.975400e+06	2.065167e+06	1.842564e+06
365	366	NaN	1.804104e+06	NaN	1.390645e+06	1.319431e+06	1.850459e+06

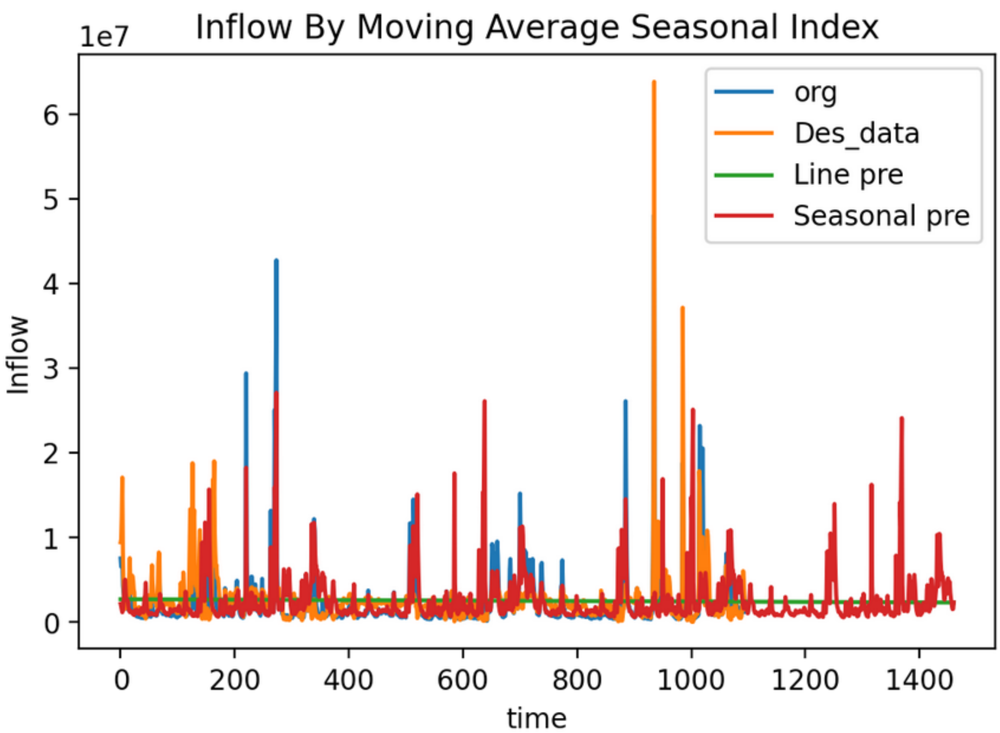
從MAPE可看出誤差值相較於前三個要少，大致下降了10%，因此新的模型理論上可達到較準確的預測。

研究方法二 : 以日為單位的季節性回歸

由於進水量似乎可以以一整年當作一次循環，我們嘗試以 108、109 及 110 三年各 365 天為單位進行的季節性回歸。

一、三年日進流量的季節性 Smoothing by Centered Moving Average

預測折線圖與表格 :



	time	org	Des_data	Line pre	Seasonal pre
0	0.0	7543380.0	9.448437e+06	2.730245e+06	2.179755e+06
1	1.0	6572620.0	9.848032e+06	2.729978e+06	1.821999e+06
2	2.0	7019088.0	1.254411e+07	2.729710e+06	1.527416e+06
3	3.0	7374848.0	1.709533e+07	2.729442e+06	1.177469e+06
4	4.0	4338772.0	8.335019e+06	2.729174e+06	1.420664e+06
...	...	...	...	...	...
1456	1456.0	NaN	NaN	2.340369e+06	2.444691e+06
1457	1457.0	NaN	NaN	2.340102e+06	1.828982e+06
1458	1458.0	NaN	NaN	2.339834e+06	1.550185e+06
1459	1459.0	NaN	NaN	2.339566e+06	1.586998e+06
1460	1460.0	NaN	NaN	2.339298e+06	2.350376e+06

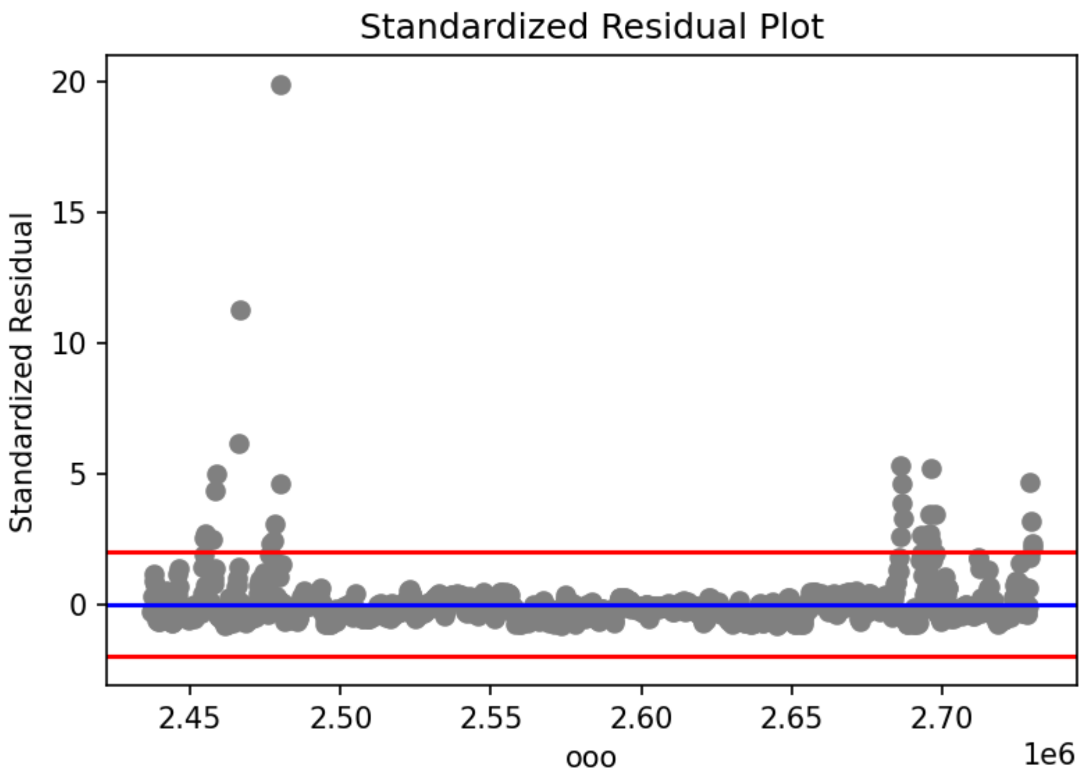
殘差分析：

```
Shapiro Test
Statistics=0.454, p=0.000
```

```
runs =  159
n1 =  549
n2 =  548
runs_exp =  549.4995442114858
stan_dev =  16.5529246995655
z =  -23.59096965032023
pval_z =  4.771199460866353e-123
p_value for Z-statistic=  4.771199460866353e-123
```

```
x_square_sum =  1096.6565029778735
size =  1096
x_d =  [0. 0. 0. ... 0. 0. 0.]
x_d =  [ 0.          0.12971404  0.87474338 ... -0.59582245 -0.50485788
 -0.09409877]
d =  0.8881689541207388
0.8881689541207388
```

For n = 1100, k = 1, dL = 1.899, dU = 1.903.



可以看到除了同異質性以外的所有殘差分析都沒有通過，但是我們還是來看一下 error matrix 的表現：

Error Matrix:

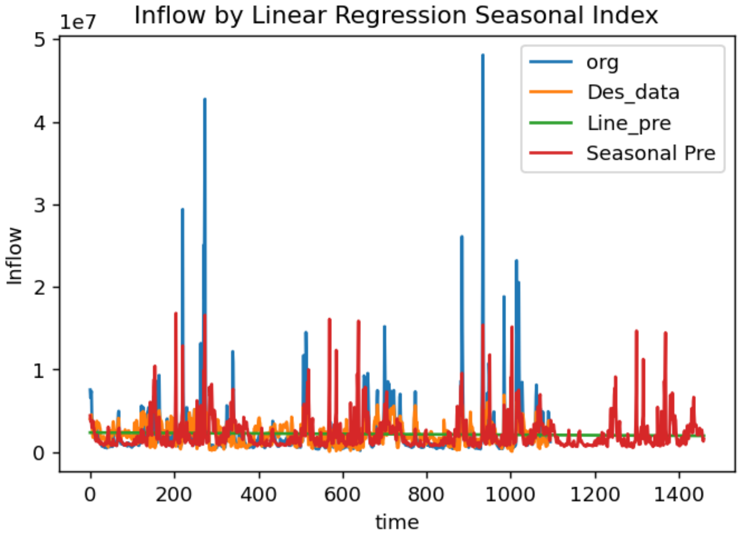
	ErrM	MA365
0	MAD	1522689.899903
1	MSE	10643584270201.549
2	RMSE	3262450.654064
3	MAPE	114.327154

比較上而言，四種error的值看起來都不太理想，數值都不小，再加上前面殘差分析也幾乎失敗，因此這種方法比較不可行。

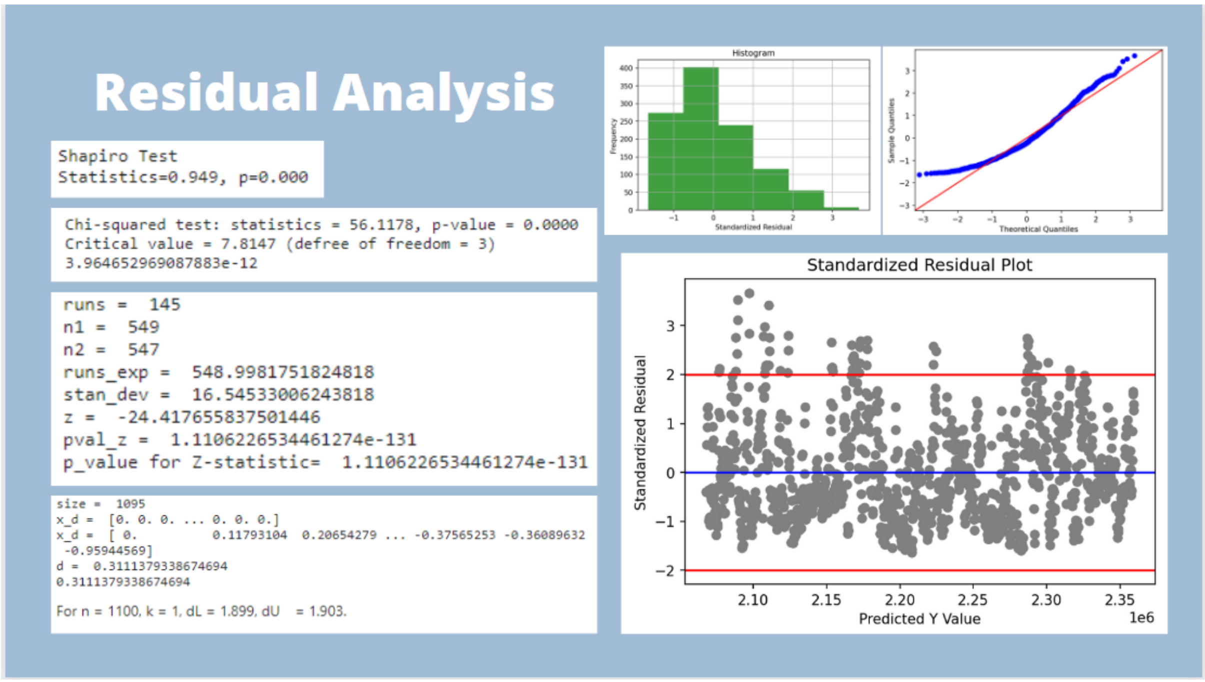
二、三年日進流量的季節性 Smoothing by Linear Regression Model

預測數據與折線圖：

	time	org	Des_data	Line_pre	Seasonal Pre
0	0.0	7543380.0	4.030803e+06	2.359633e+06	4.415895e+06
1	1.0	6572620.0	4.185165e+06	2.359366e+06	3.705282e+06
2	2.0	7019088.0	4.455707e+06	2.359099e+06	3.716296e+06
3	3.0	7374848.0	4.534867e+06	2.358833e+06	3.836062e+06
4	4.0	4338772.0	3.465800e+06	2.358566e+06	2.952646e+06
...	...	...	...	...	...
1455	1455.0	NaN	NaN	1.971609e+06	2.438915e+06
1456	1456.0	NaN	NaN	1.971343e+06	1.810687e+06
1457	1457.0	NaN	NaN	1.971076e+06	1.384601e+06
1458	1458.0	NaN	NaN	1.970809e+06	1.300637e+06
1459	1459.0	NaN	NaN	1.970543e+06	1.575111e+06



殘差分析的結果 :



Normality test 的 p-value，Runs test 的 p-value 都小於 0.05，因此殘差不是常態分佈也不隨機；Durbin Watson test 的 d 值小於 dL，有 first-order auto-correlation 的問題；右邊的散佈圖也有很多點大於 2 因此也拒絕了 null hypothesis，因此可以看到所有的殘差分析都沒有通過，但是我們還是來看一下 error matrix 的表現：

```
Error Metrics for Seasonal Index by LR
MAD =  1384335.280174
MSE =  7892040429938.033
RMSE =  2809277.563705
MAPE =  95.075719 %
```

可以看到這樣的差距也很不理想。

三、三年日進流量為單位進行的 **Regression Model by Indicator Variables**

以日為單位進行的 Regression Model by Indicator Variables 因為需要生成 365 個 indicator variables，模型的結果很容易出現偏差，因此我們沒有使用此方法。

四、比較

稍微將 CMA 和 Smoothing by Linear Regression 的 error matrix 進行比較可以發現，Linear Regression 的數值都較小，因此模型較好。

而以 111 年 1 月 1 日的實際進水量 1020824 立方公尺與兩者的預測進水量相比，CMA 的預測量 1945450.0055立方公尺，相較於 Linear Regression 的預測量 3869403.9367928立方公尺更為接近實際值，但是兩者的預測值都與實際值相去甚遠。

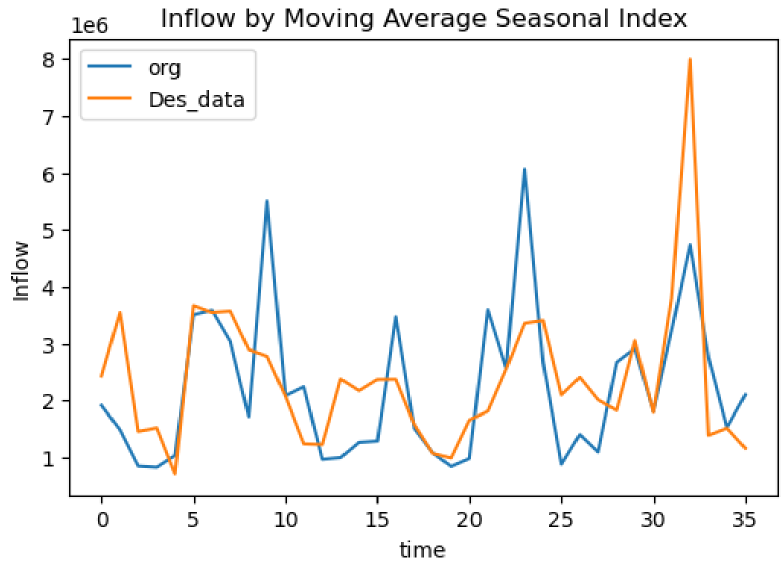
五、結論

雖然在 error 的比較上 Linear Regression 的模型看起來比較合適，但是其實兩者的殘差分析都沒過，再加上實際預測值的結果皆不符預期，因此研究方法二的模型可靠性都不高。

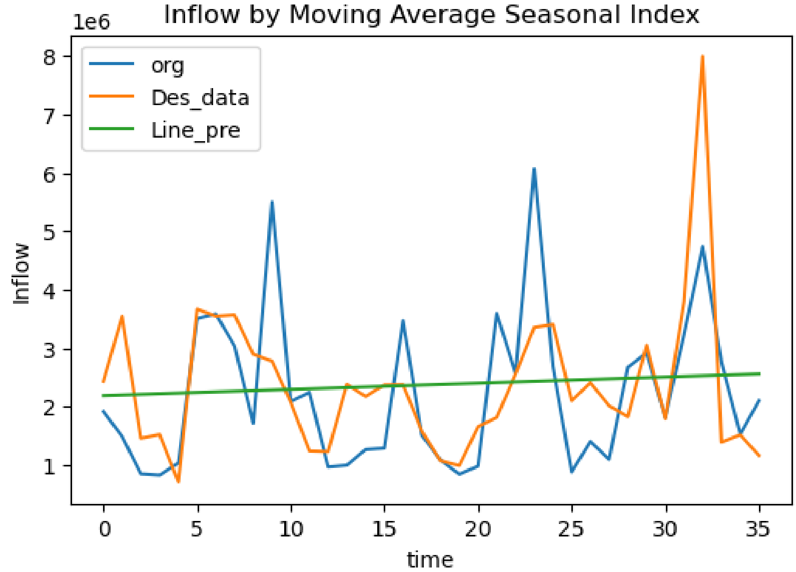
研究方法三：以月為單位的季節性回歸

我們先將 108、109 及 110 年每月的進流量平均，取得月平均之後，以十二個月為單位嘗試進行季節性的回歸。

一、以月平均進行的 **Smoothing by Centered Moving Average**



我們以移動平均法的方式去平滑原始資料，並以此做出de-seasonalize data，可以發現即使原始資料似乎有季節性變化的趨勢，但de-seasonalize data 的去季節化效果並不好。

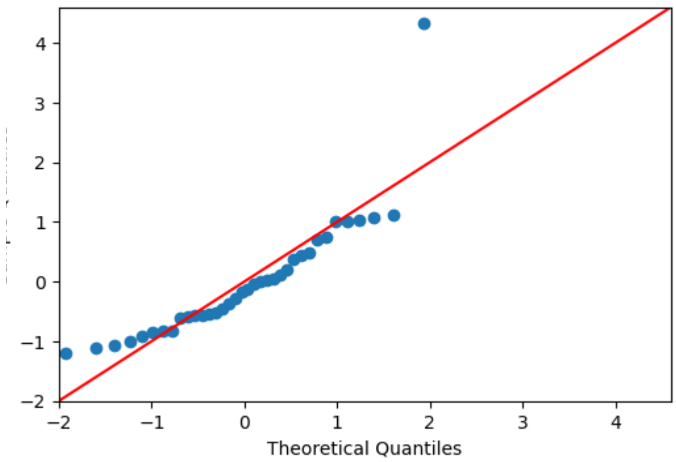


我們以de-seasonalize data 得到的迴歸直線從圖上可以發現其預測能力並不足。

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Des_D   R-squared:                       0.007
Model:                            OLS   Adj. R-squared:                 -0.022
Method:                 Least Squares   F-statistic:                    0.2551
Date:                Wed, 18 May 2022   Prob (F-statistic):              0.617
Time:                        17:43:47   Log-Likelihood:                -557.04
No. Observations:                  36   AIC:                             1118.
Df Residuals:                      34   BIC:                             1121.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2.189e+06   4.27e+05      5.130      0.000    1.32e+06    3.06e+06
t           1.059e+04    2.1e+04      0.505      0.617    -3.2e+04    5.32e+04
==============================================================================
Omnibus:                       34.047   Durbin-Watson:                   1.568
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              100.065
Skew:                           2.145   Prob(JB):                     1.87e-22
Kurtosis:                       9.950   Cond. No.                         39.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

迴歸分析的結果發現，R-squared和coefficient的值都相當小，表示其迴歸線的代表力並不足。

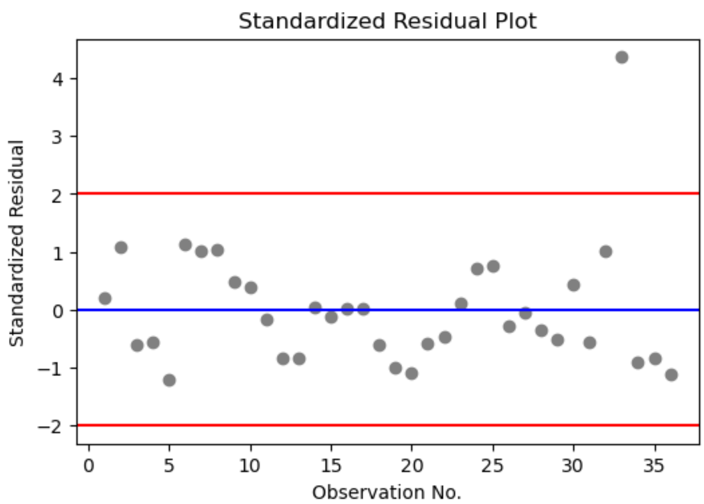
Normality test




```
Shapiro Test
p_value for Z-statistic=  0.09083702152945593
```

透過qq plot 和 shapiro test 的結果我們可以推知殘差的分佈為常態。

Homoscedasticity



從上面的分佈我們可以推論 variable of error 為相同而平均 0，因此可認定其據有同質性。

Dependence of the Error Variable

```
runs =  14
n1 =  18
n2 =  18
runs_exp =  19.0
stan_dev =  2.9568322818274866
z =  -1.6909988539863081
pval_z =  0.09083702152945593
p_value for Z-statistic=  0.09083702152945593
```

Because n1, n2 < 20, Lr = 12 < R = 14 < Ur = 26, do not reject H_0. There is no evidence to infer that the sample is not random.

由上述檢驗可證明此迴歸通過所有的殘差分析

Durbin Waston Test

```
The Durbin Watson test
x_square_sum =  36.67482871939605
size =  36
x_d =  [ 0.          0.88771418 -1.68759304  0.04471343 -0.65362713  2.34168332
 -0.10770544  0.00704716 -0.54167356 -0.1013166  -0.55886467 -0.65805598
 -0.01676306  0.88776299 -0.16762035  0.14446894 -0.00568119 -0.62399378
 -0.40138059 -0.07199578  0.50725344  0.11478195  0.56545411  0.61727843
  0.03193954 -1.02818958  0.23348043 -0.31801896 -0.15628584  0.95651488
 -1.00563809  1.58316249  3.34629656 -5.28208926  0.08824531 -0.29887391]
d =  1.5771008430844435
d value = 1.5771008430844435
```

n=36, k=2, α = 0.05

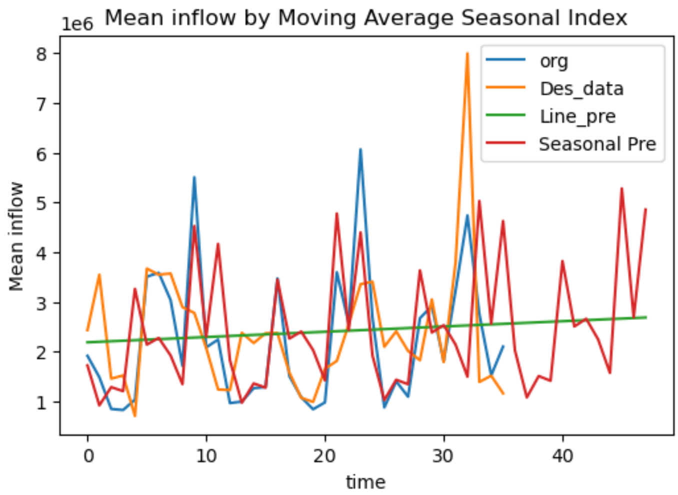
From the Durbin-Watson table we have:

dL = 1.354, dU = 1.587, 4-dU = 2.413, 4-dL = 2.646.

The statistic d is between the dL and dU, so we can not reject the H_0, and conclude that the test is inconclusive.

由於 d value 落在 dL 和 dU 之間，我們無法由 Durbin Waston Test 判定是否有自相關的性質。

Forecast



雖然迴歸直線的代表性不足，但其預測效果從圖上判斷似乎不差。

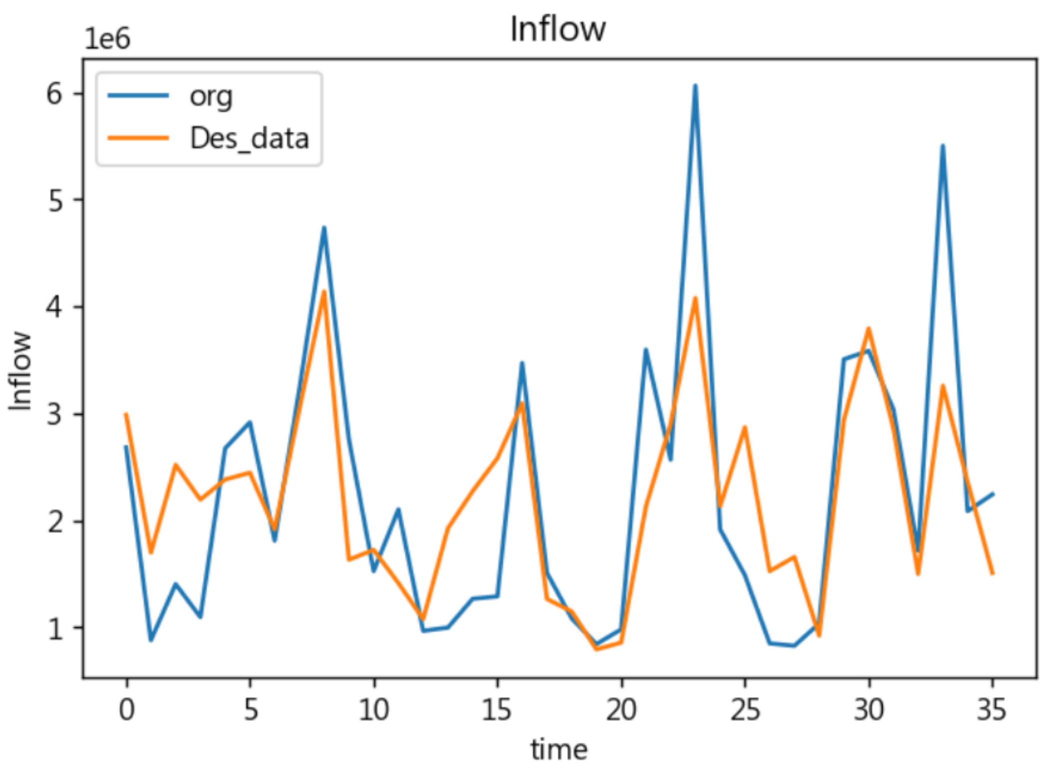
	Year/month	Predict Inflow
36	111/01	2.024442e+06
37	111/02	1.082822e+06
38	111/03	1.511199e+06
39	111/04	1.416061e+06
40	111/05	3.819943e+06
41	111/06	2.507366e+06
42	111/07	2.661643e+06
43	111/08	2.249644e+06
44	111/09	1.573755e+06
45	111/10	5.278628e+06
46	111/11	2.691396e+06
47	111/12	4.853787e+06

二、以月平均進行的 **Smoothing by Linear Regression Model**

1.De-Seasonalized Data

先把3年36個月的資料De-seasonalized，這邊只放第一年的資料作為範例，以免版面太亂

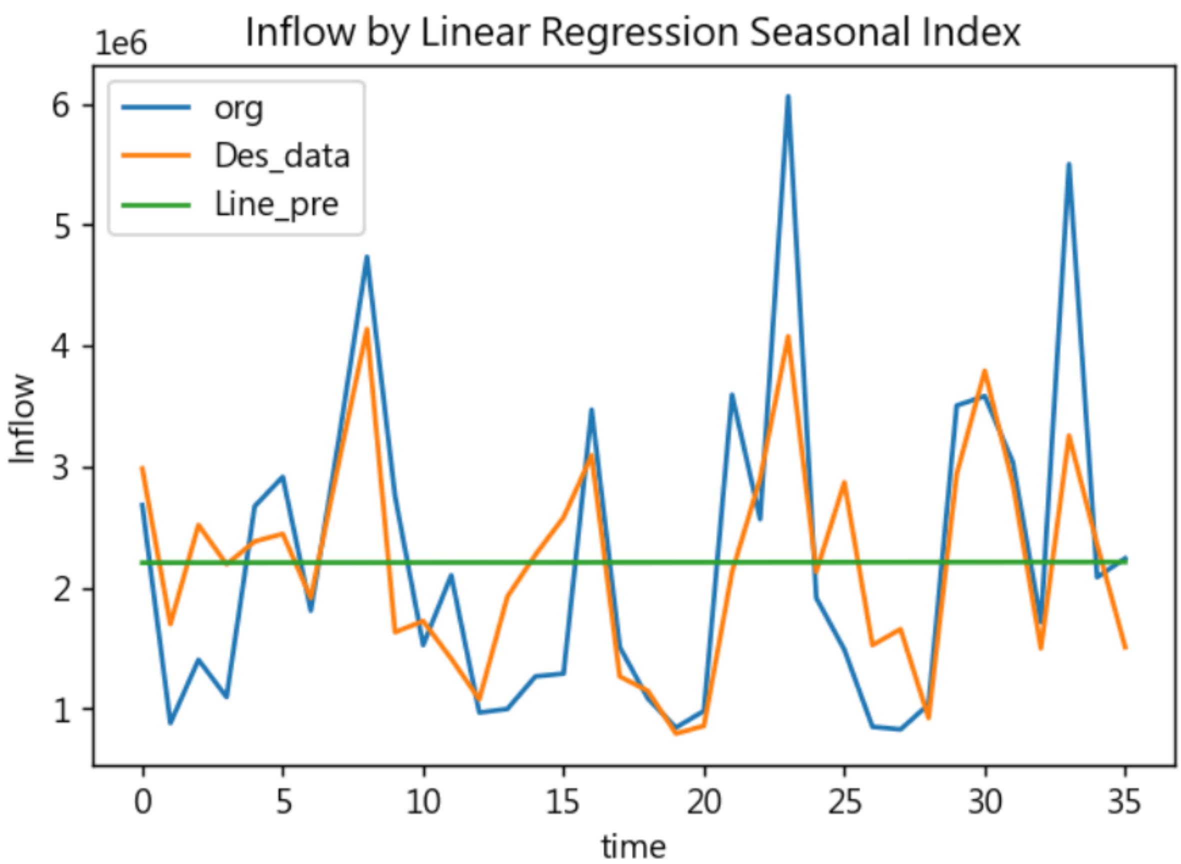
	SID	SeaIdx	orig	Des_D
0	1.0	0.898547	2683037.0	2.985972e+06
1	2.0	0.518363	881138.0	1.699847e+06
2	3.0	0.557710	1405088.0	2.519390e+06
3	4.0	0.499758	1096764.0	2.194591e+06
4	5.0	1.122025	2673796.0	2.383009e+06
5	6.0	1.192688	2917327.0	2.446010e+06
6	7.0	0.944780	1811000.0	1.916847e+06
7	8.0	1.061965	3228925.0	3.040520e+06
8	9.0	1.144253	4737023.0	4.139841e+06
9	10.0	1.688389	2757452.0	1.633185e+06
10	11.0	0.884868	1526410.0	1.725014e+06
11	12.0	1.486654	2103162.0	1.414695e+06



2. New Regression

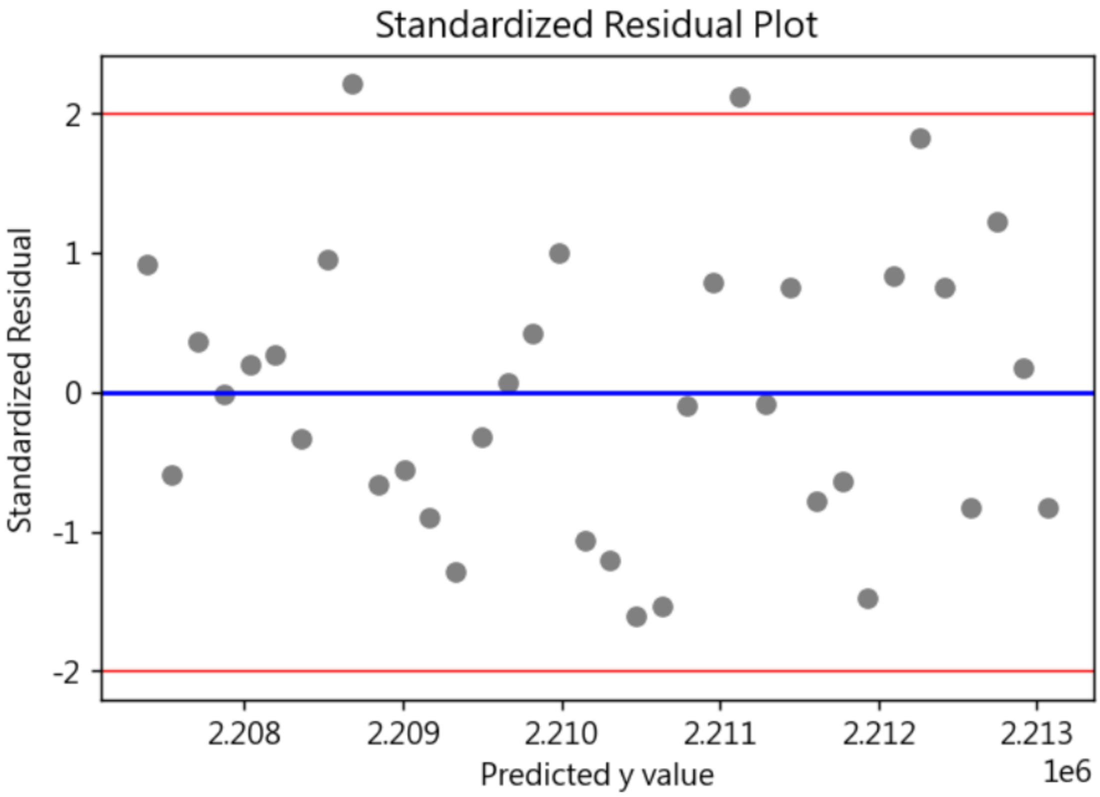
把De-seasonalized的資料做成新的Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Des_D   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.029
Method:                 Least Squares   F-statistic:                 0.0001277
Date:                Thu, 19 May 2022   Prob (F-statistic):              0.991
Time:                        21:41:26   Log-Likelihood:                -543.46
No. Observations:                  36   AIC:                             1091.
Df Residuals:                      34   BIC:                             1094.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2.207e+06   2.93e+05      7.544      0.000    1.61e+06     2.8e+06
t            162.4628   1.44e+04      0.011      0.991   -2.91e+04    2.94e+04
==============================================================================
Omnibus:                        1.266   Durbin-Watson:                   1.422
Prob(Omnibus):                  0.531   Jarque-Bera (JB):                1.247
Skew:                           0.371   Prob(JB):                        0.536
Kurtosis:                       2.471   Cond. No.                         39.9
==============================================================================
```



3.Residual Analysis

Residual相關的檢定，Shapiro的結果看起來是常態，RunTest的結果也是隨機，Durbin Watson結果為inconclusive，Standardized Residual的結果為同質變異




```
Shapiro Test
Statistics=0.966, p=0.334
runs =  16
n1 =  18
n2 =  18
runs_exp =  19.0
stan_dev =  2.9568322818274866
z =  -1.014599312391785
pval_z =  0.3102968672638191
p_value for Z-statistic=  0.3102968672638191


The Durbin Watson test
x_square_sum =  35.783926362059596
size =  36
x_d =  [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
x_d =  [ 0.         -1.51563834  0.96097267 -0.37997792  0.2184428   0.07190145
 -0.61071289  1.29145593  1.25645156 -2.86986844  0.10593549 -0.35231829
 -0.38152444  0.96408747  0.39436665  0.34883329  0.57924443 -2.06825063
 -0.13358174 -0.40129046  0.07063841  1.43975943  0.87652652  1.33735354
 -2.21095545  0.84109253 -1.53543001  0.14944828 -0.84544383  2.31803111
  0.99032962 -1.07699237 -1.58191439  2.05317747 -1.05211957 -1.00180468]
d =  1.4313915532378143
d value = 1.4313915532378143
```

n = 36, k = 1, alpha = 0.05

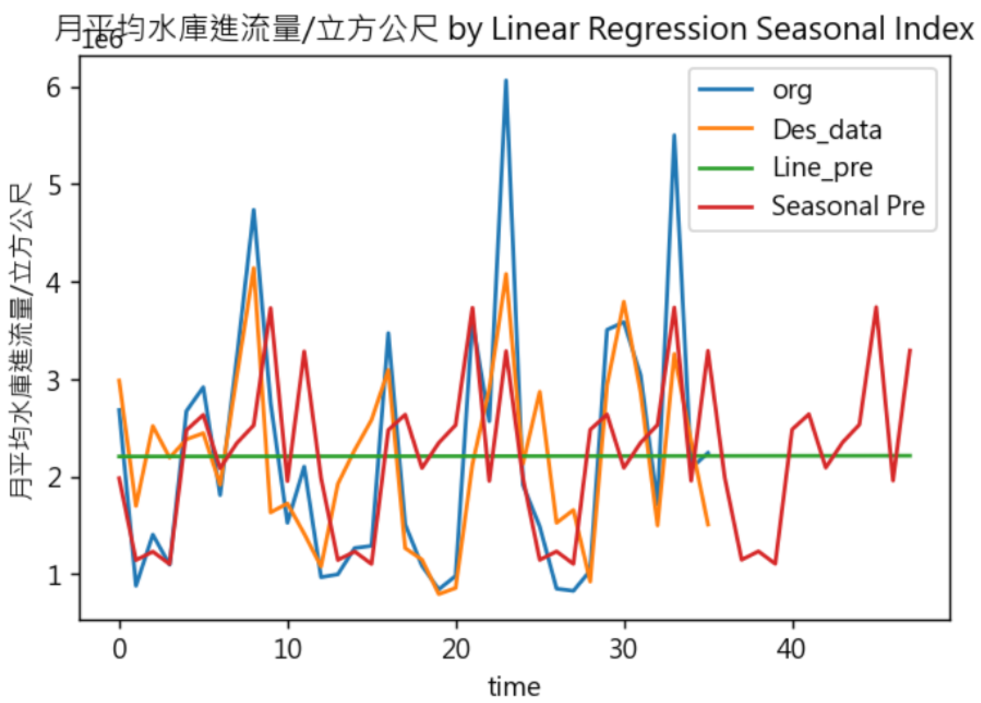
dL=1.411, dU=1.525 Because dL = 1.411 < d value 1.4313915532378143 < dU=1.525 , it is inconclusive whether the first order correlation exists.

because n1, n2 < 20, Lr = 12 < R = 16 < Ur = 26, do not reject H0, There is no evidence to infer that the sample is not random.

4.Forecasting

最後將第四年也就是第36個月到47個月的預測值預測出來

	time	org	Des_data	Line_pre	Seasonal Pre
36	36.0	NaN	NaN	2.213233e+06	1.988694e+06
37	37.0	NaN	NaN	2.213395e+06	1.147343e+06
38	38.0	NaN	NaN	2.213558e+06	1.234522e+06
39	39.0	NaN	NaN	2.213720e+06	1.106324e+06
40	40.0	NaN	NaN	2.213883e+06	2.484031e+06
41	41.0	NaN	NaN	2.214045e+06	2.640666e+06
42	42.0	NaN	NaN	2.214207e+06	2.091940e+06
43	43.0	NaN	NaN	2.214370e+06	2.351583e+06
44	44.0	NaN	NaN	2.214532e+06	2.533984e+06
45	45.0	NaN	NaN	2.214695e+06	3.739266e+06
46	46.0	NaN	NaN	2.214857e+06	1.959857e+06
47	47.0	NaN	NaN	2.215020e+06	3.292969e+06



三、以月平均進行的 **Regression Model by Indicator Variables**

1. Regression Model:

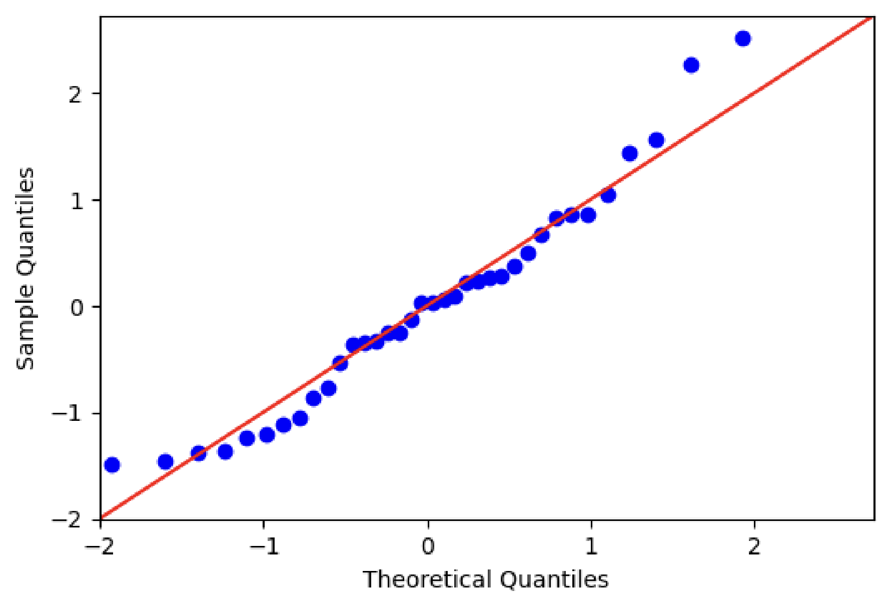
```
                            OLS Regression Results
==============================================================================
Dep. Variable:               water_in   R-squared:                       0.415
Model:                            OLS   Adj. R-squared:                  0.110
Method:                 Least Squares   F-statistic:                     1.362
Date:                Wed, 18 May 2022   Prob (F-statistic):              0.253
Time:                        17:27:57   Log-Likelihood:                -548.68
No. Observations:                  36   AIC:                             1123.
Df Residuals:                      23   BIC:                             1144.
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3.472e+06   8.78e+05      3.953      0.001    1.65e+06    5.29e+06
t            -69.5788   2.14e+04     -0.003      0.997   -4.44e+04    4.42e+04
month_1    -1.615e+06   1.05e+06     -1.531      0.139    -3.8e+06    5.67e+05
month_2    -2.348e+06   1.05e+06     -2.236      0.035   -4.52e+06   -1.76e+05
month_3    -2.296e+06   1.05e+06     -2.195      0.039   -4.46e+06   -1.32e+05
month_4    -2.399e+06   1.04e+06     -2.301      0.031   -4.56e+06   -2.42e+05
month_5    -1.077e+06   1.04e+06     -1.036      0.311   -3.23e+06    1.07e+06
month_6    -8.257e+05   1.04e+06     -0.797      0.434   -2.97e+06    1.32e+06
month_7     -1.31e+06   1.03e+06     -1.267      0.218   -3.45e+06    8.29e+05
month_8      -1.1e+06   1.03e+06     -1.066      0.297   -3.23e+06    1.03e+06
month_9     -9.92e+05   1.03e+06     -0.963      0.346   -3.12e+06    1.14e+06
month_10    4.814e+05   1.03e+06      0.468      0.644   -1.65e+06    2.61e+06
month_11    -1.41e+06   1.03e+06     -1.371      0.184   -3.54e+06    7.18e+05
==============================================================================
Omnibus:                        2.278   Durbin-Watson:                   1.556
Prob(Omnibus):                  0.320   Jarque-Bera (JB):                1.702
Skew:                           0.533   Prob(JB):                        0.427
Kurtosis:                       2.977   Cond. No.                         259.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Fitted Values [1856900.2360215  1123435.46950845 1175619.63387097 1072820.94569893
 2394894.38655914 2645864.1234767  2161432.2360215  2371251.41881721
 2479356.45681004 3952609.91344086 2061418.67903226 3471108.10698925
 1856065.29032258 1122600.52380952 1174784.68817204 1071986.
 2394059.44086021 2645029.17777778 2160597.29032258 2370416.47311828
 2478521.51111111 3951774.96774194 2060583.73333333 3470273.16129032
 1855230.34462366 1121765.5781106  1173949.74247312 1071151.05430108
 2393224.49516129 2644194.23207885 2159762.34462366 2369581.52741936
 2477686.56541219 3950940.02204301 2059748.78763441 3469438.2155914 ]
```

We can see that 41.5% of variations are explained by this model. However, overfitting exists.

2. Residual Analysis:

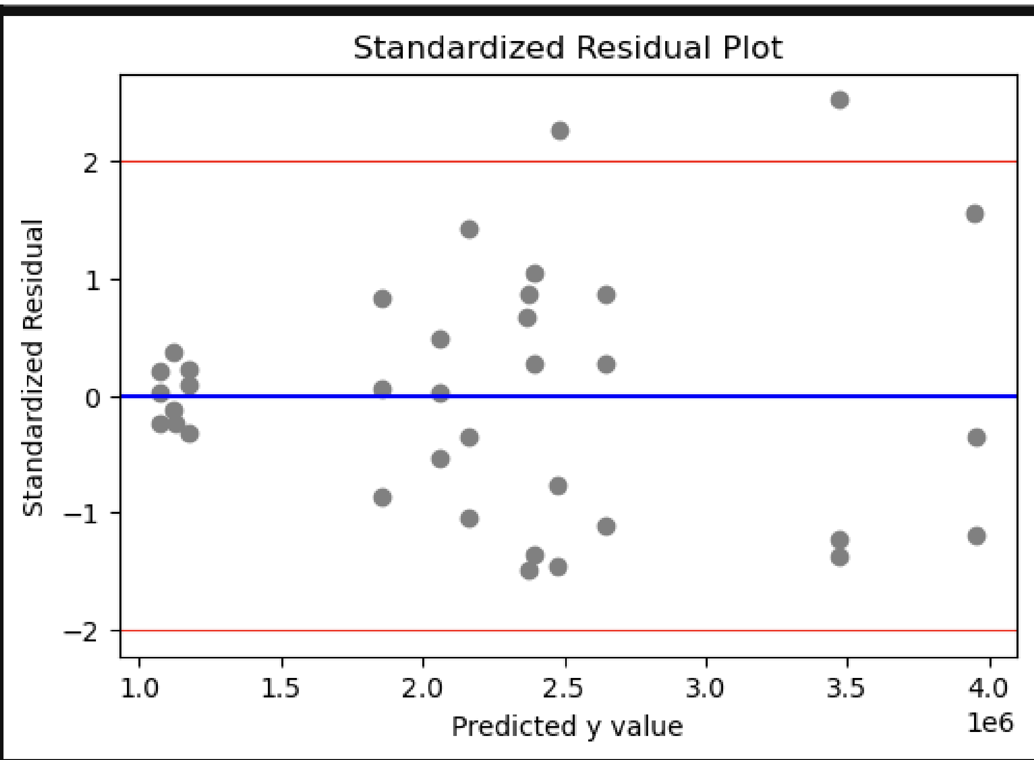
Normality Test :

```
Shapiro Test
Statistics=0.957, p=0.180
/Users/liuweien/opt/anaconda3/lib/python3.9/site-packages/statsmodels/graphics/gofplots.py:
g "bo" (-> marker='o'). The keyword argument will take precedence.
  ax.plot(x, y, fmt, **plot_style)
```



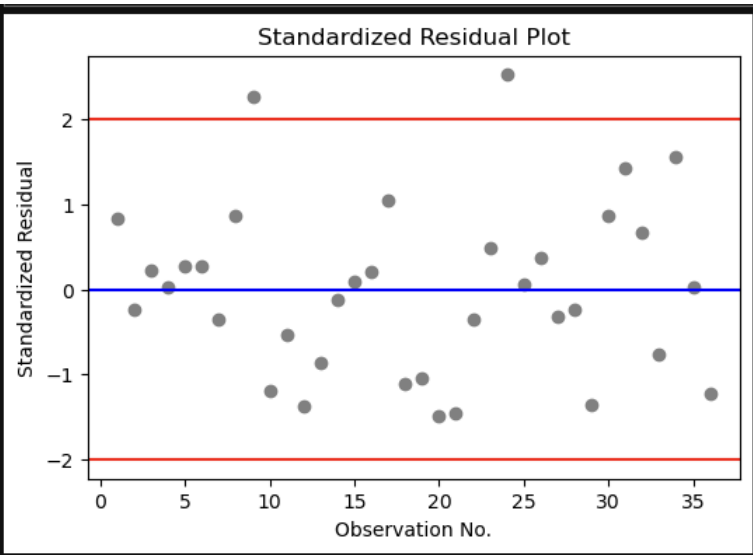
Because p_value = 0.180 > a = 0.05, do not rejected H0. We can assume that errors are normally distributed.

Homoscedasticity and Heteroscedasticity Test



Do not rejected H0. We can assume that the variation is constant and the mean is around 0.

Dependence of the Error Variable Test:




```
runs =  16
n1 =  18
n2 =  18
runs_exp =  19.0
stan_dev =  2.9568322818274866
z =  -1.014599312391785
pval_z =  0.3102968672638191
p_value for Z-statistic=  0.3102968672638191
```

Because n1, n2 < 20, Lr = 12 < R = 16 < Ur = 26, do not reject H0. There is no evidence to infer that the sample is not random.

Durbin Watson Test:

```
x_square_sum =  35.78236961701012
size =  36
x_d =  [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
x_d =  [ 0.         -1.07313834  0.47384275 -0.20643065  0.25608129 -0.00747103
 -0.62463315  1.21342467  1.40615775 -3.46802759  0.66305591 -0.83660509
  0.51076711  0.74241981  0.21157326  0.12161196  0.83554961 -2.1507051
  0.05735659 -0.43874108  0.02890915  1.10904361  0.83990245  2.02824913
 -2.4602304   0.30637045 -0.69235468  0.08083036 -1.11903322  2.22870971
  0.56539552 -0.76029483 -1.43601499  2.32261227 -1.53050343 -1.25816161]
d =  1.5667996478416744
1.5667996478416744
```

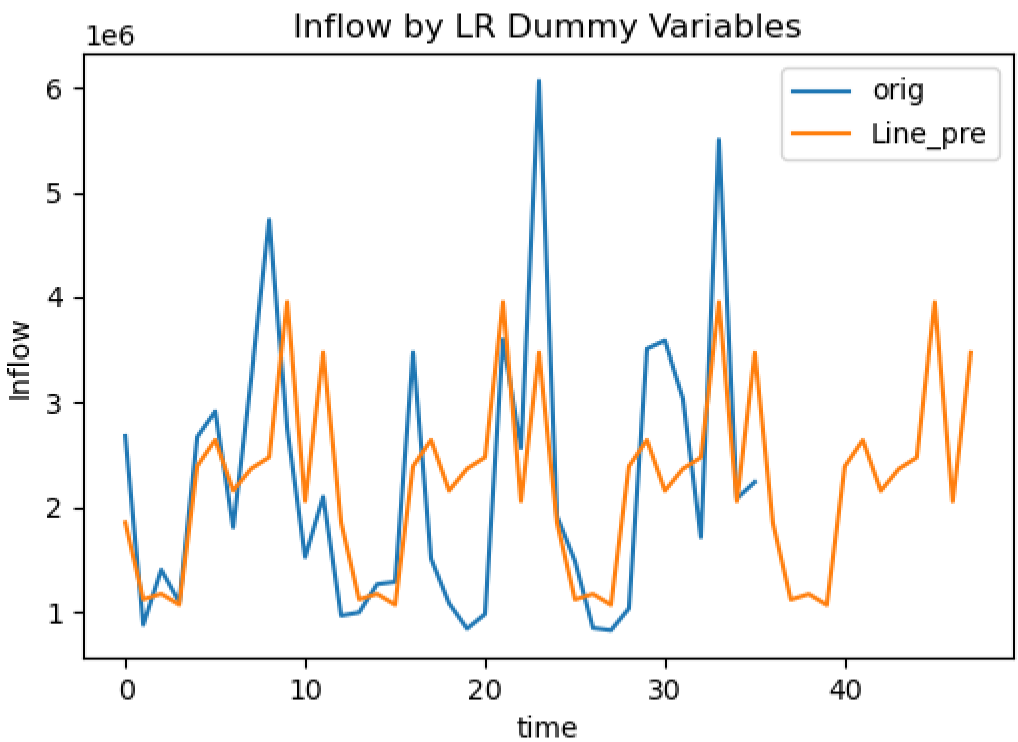
(T = 36, K = 12, alpha = 0.05): dl = 0.748 , du = 2.398

dl < d = 1.5667996478416744 < du. Thus, it is inconcusive whether first order auto-correlation exists.

3. Forecasting (111/01 ~ 111/12):

	Year/month	Predict Inflow
0	111/01	1.854395e+06
1	111/02	1.120931e+06
2	111/03	1.173115e+06
3	111/04	1.070316e+06
4	111/05	2.392390e+06
5	111/06	2.643359e+06
6	111/07	2.158927e+06
7	111/08	2.368747e+06
8	111/09	2.476852e+06
9	111/10	3.950105e+06
10	111/11	2.058914e+06
11	111/12	3.468603e+06

:



四、比較

相互比較的結果如下：

dataset:

	ErrM	SIMA	SILR	Dummy
0	MAD	897899.420924	777168.888083	788789.261086
1	MSE	1410625116976.3325	1023939149413.8843	1013279558234.3698
2	RMSE	1187697.401267	1011898.784175	1006617.880943
3	MAPE	46.281122	41.716465	42.033599

111/ 01 三種方式預測與實際值做比較：

	Year/month	Predict Inflow	Seasonal Pre	Predict Inflow	actual value
36	111/01	2.024442e+06	1.988694e+06	1.854395e+06	2.718889e+06

五、結論

根據以上結果，我們可以看到三種方式的 residual analysis test 都通過檢定。因此我們可以推論，使用月平均資料做出來的季節性回歸相較於前面的模型有比較好的預測結果。
然而，比較了 111/01 實際資料與預測資料後，發現他們仍然存在一定的誤差。因此我們推測，也許以月為單位季節性因素帶來的影響並沒有預期中的顯著。

總結：

透過以上的預測與模型之間的比較，在預測每日進流量的情形下，時間序列的預測相較於季節性回歸可以得到更準確的結果。而在我們預測月平均進流量的情形下，則是以 SILR 的方法可以得到更準確的結果。在我們可以準確預測日進流量以及月平均進流量的前提下，我們即可使用預測資料超前部署，防範水庫下游出現洪災或缺水的情形，以保證翡翠水庫能夠安全及穩定的供水，使大台北及新北地區不會因極端氣候，出現缺水的困境，也能促進經濟發展，創造人民福祉。