

逢甲大學學生報告 ePaper

高鐵每月進站旅客人數之時間序列分析

Time series analysis of the number of passengers entering the Taiwan High-Speed Rail station each month.

作者：楊侑錚、范芸翎、王雨柔、李振維、熊墨樺、陳羽均、張梓淇

系級：統計三甲

學號：D0680595、D0680565、D0609436、D0680387、D0636331、D0636314、D0680429

開課老師：李仁佑

課程名稱：預測分析

開課系所：統計系

開課學年：108 學年度 第二學期

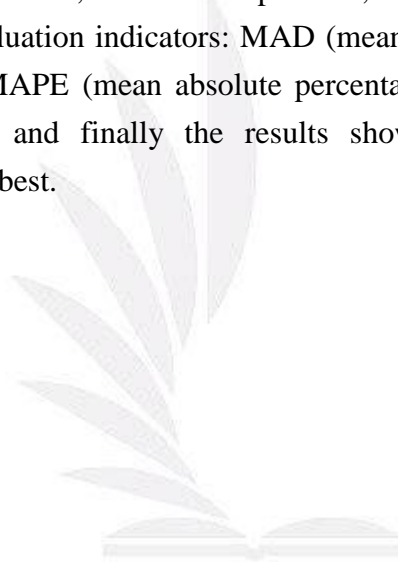
中文摘要

在科技越來越進步的趨勢下，各種交通工具的出現，提供民眾許多選擇。此外對於大眾運輸服務水準以及乘坐時間要求也逐漸提高，所以選擇搭乘高鐵的人逐年遞增，我們想知道未來高鐵成長趨勢會如何，因此希望透過課程學習的統計預測方法，找出一個最佳模型來預測未來一年高鐵每月進站人數。並期待這份報告能夠提供給高鐵單位作為參考。本報告主要研究 2007 年 1 月至 2018 年 12 月臺灣高鐵每月進站旅客人數，資料來源為台灣高鐵網站上提供的數據。此研究將透過三個分析方法:ARIMA 法、時間序列迴歸法和分解法進行分析，並建立了四個模型，再利用三種評估指標: MAD(mean absolute deviation)、MSE(mean square error)及 MAPE(mean absolute percentage error)作為選出最佳模型的依據，最後分析結果顯示，乘法分解模型為最佳模型。

關鍵字：台灣高鐵，ARIMA 法，時間序列迴歸法，分解法

Abstract

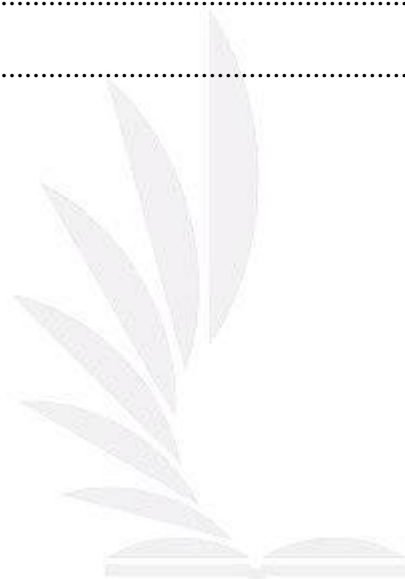
With the increasing progress of science and technology, the emergence of various modes of transport provides many choices for the public. Besides, the level of public transport services and travel time requirements are gradually increasing, so the number of people choosing to take high-speed rail is increasing year by year. We would like to know the future growth trend of high-speed rail, so we hope to learn through the course of statistical forecasting methods, to find out the best model to predict the number of high-speed rail monthly inbound in the coming year. It is also expected that this report can be provided to high-speed rail units for reference. This report focuses on the number of monthly passengers entering Taiwan's high-speed rail from January 2007 to December 2018, based on data available on the Taiwan high-speed rail website. This study will be analyzed by three analytical methods: ARIMA, time series regression, and decomposition, and four models will be established, using three evaluation indicators: MAD (mean absolute deviation), MSE (mean square error), and MAPE (mean absolute percentage error), as the basis for selecting the best model, and finally the results show that the multiplication decomposition model is the best.



Keywords : Taiwan High Speed Rail, ARIMA, Time series regression, Decomposition

目 次

第一章、緒論.....	4
第二章、資料描述.....	6
第三章、分析方法.....	8
第一節、ARIMA 分析法.....	9
第二節、時間序列迴歸法.....	13
第三節、分解法.....	17
第四章、模型的預測能力及比較.....	21
第五章、結論.....	錯誤! 尚未定義書籤。
參考文獻.....	25



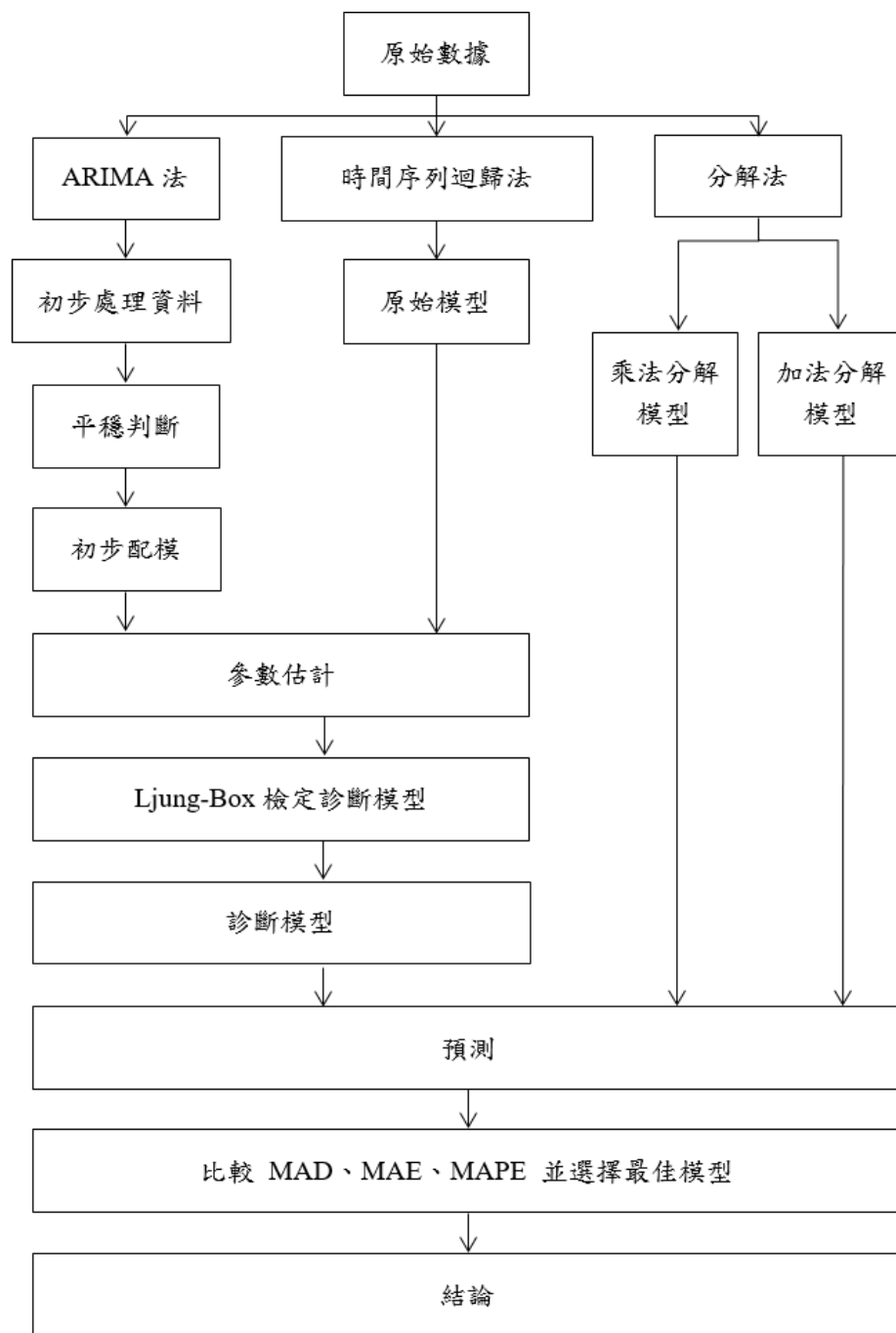
第一章、緒論

一、研究背景、動機及目的

臺灣高鐵(Taiwan High Speed Rail)於 2007 年 1 月 5 日通車，有著準點、運輸量大、安全等各方面優勢，不僅縮短了縣市間的距離也節省許多通車時間。憑藉著多方面的優勢，高鐵的旅運量逐年遞增。對於改變交通結構的高鐵，不乏許多學者研究、探討高鐵的成長趨勢能到達何種程度。以淡江大學范俊海教授團隊的類應用神經網路預測高鐵車站之運量[1]為例，此研究以 2008 年 1 月至 2014 年 12 月高鐵左營站的運量作為研究對象，樣本數共 84 筆，將資料以類神經網路以及多元迴歸模式兩種方法進行預測並比較兩者的預測能力。而逢甲大學葉名山教授指導的專題研究[2]也有對高鐵運量進行預測，此研究以 2007 年 1 月至 2018 年 3 月的旅客人數作為研究對象，是以樣本不足而採用灰預測對旅客人數進行迴歸分析，其分析結果顯示預測準確率高達 92.52%，並提到高鐵運量具有時間序列的影響。透過上述兩篇文獻，本研究將對高鐵每月的進站人數進行時間序列分析，並增加樣本數至 156 筆，資料取自台灣高鐵網站[3]，藉由以往的資料預測未來一年內每月搭乘高鐵的人數，並且將預測所得結果給予高鐵單位作為參考，在不同旅客量的月份可以選擇增加班次來預防旅客量過多影響乘車品質，或減少班次將剩餘的人力做其他安排，使得顧客與高鐵皆能雙贏。

二、研究流程

本研究以 2007 年 1 月至 2018 年 12 月臺灣高鐵每月進站旅客人數作為研究對象，並使用三個分析方法:ARIMA 法、時間序列迴歸法和分解法進行分析以及預測。研究架構圖如圖 1-1 所示。主要分別對三個分析方法配適出適當的模型，並利用模型進行預測，接著再進行比較決定出最佳模型。最後，根據每個部分所得到的結果歸納出結論。



< 圖 1-1 >、研究架構圖

第二章、資料描述

此筆研究資料為臺灣高鐵每月進站旅客人數，其單位為萬人次。資料時間自 2007 年 1 月至 2019 年 12 月，共 156 筆資料，且保留民國 2019 年的 12 筆數值做預測分析，觀察自臺灣高鐵官方網站蒐集的表 2-1 可知，初期人次持續成長直至營運兩年後，中期的成長幅度漸漸下降且後期趨近零，現象持續到 2016 獲得改善，人次成長改變幅度較中前期劇烈。因此我們試著找出成長幅度發生改變的主要原因：

一、特殊節日

台灣人遵循傳統習俗，有時須回家鄉與家人團圓或祭祀，如中元節、端午節及中秋節，國家因此給予放假；有時如西洋情人節及七夕，更甚是兒童節、母親節及父親節等...，在假日只有兩天時間的情況下，民眾較容易傾向搭乘快速省時的高鐵，藉以享有更多的節日時間。除此，在蒐集的 12 個月分裡，我們猜測九月雖有教師節，因沒有放假，與沒有節日的十一月在人次表現上，在資料中除了初期兩年皆為一年中的低點位置。

二、連假

除了帶來家庭旅遊商機的寒暑假和春節之外，每年許多不同節日可能連著周末而有超過兩天以上的假期形成連假，民眾可體驗較難得的三天兩夜等規劃，容易傾向搭乘快速省時的高鐵，藉以享有更多的假期；每年固定至少放三天的春節更是家庭外出旅遊首選，透過下表特別標記為除夕所在月份，觀察到人數的確有因春節而增加人次，且大約一周的春節過後，返回工作崗位的人次也相當可觀。最終將資料對照連假起始時間及結束時間皆符合，可知上述猜測連假對於人次的影響是相當有可信度的。

三、投票

自高鐵營運共經歷 2008 年 1 月及 3 月的兩次公投、2018 年 11 月的公投；以及 2012 年、2016 年及 2020 年三次總統大選，六次都造成人次突破性的成長。

〈表 2-1〉、2007 至 2019 高鐵每月旅客進站人數

年份 (西元)	1 月	2 月	3 月	4 月	5 月	6 月
	7 月	8 月	9 月	10 月	11 月	12 月
2007	116	72	91	107	115	124
	142	137	136	144	165	200
2008	195	209	231	254	260	253
	281	297	248	285	274	265
2009	278	239	264	258	275	243
	295	271	244	283	281	297
2010	287	304	293	309	321	298
	316	318	286	311	331	335
2011	315	343	325	347	331	338
	369	350	347	370	354	369
2012	394	340	361	356	357	348
	390	373	369	382	376	402
2013	364	385	424	391	390	389
	405	413	397	388	386	401
2014	379	383	400	398	402	391
	407	413	394	415	401	411
2015	395	414	419	417	421	408
	421	419	404	424	417	492
2016	460	484	461	465	461	456
	473	476	442	492	468	516
2017	517	479	494	509	514	456
	497	522	465	537	500	561
2018	495	515	530	547	521	518
	528	541	525	531	551	587
2019	532	542	564	567	547	561
	562	555	550	580	568	607

*標記為當年度初一所在月份

第三章、分析方法

此章節將分別介紹三個分析方法並運用這三個方法對資料進行分析。

ARIMA 模型的全名為自迴歸移動平均模型(Autoregressive Integrated Moving Average Model)，又稱為 Box-Jenkins 模型。此模型建模的基本條件為資料是平穩的時間序列，若為不平穩的時間序列可採用差分來處理。其分析的步驟為配適模型、參數估計、模式診斷及預測。

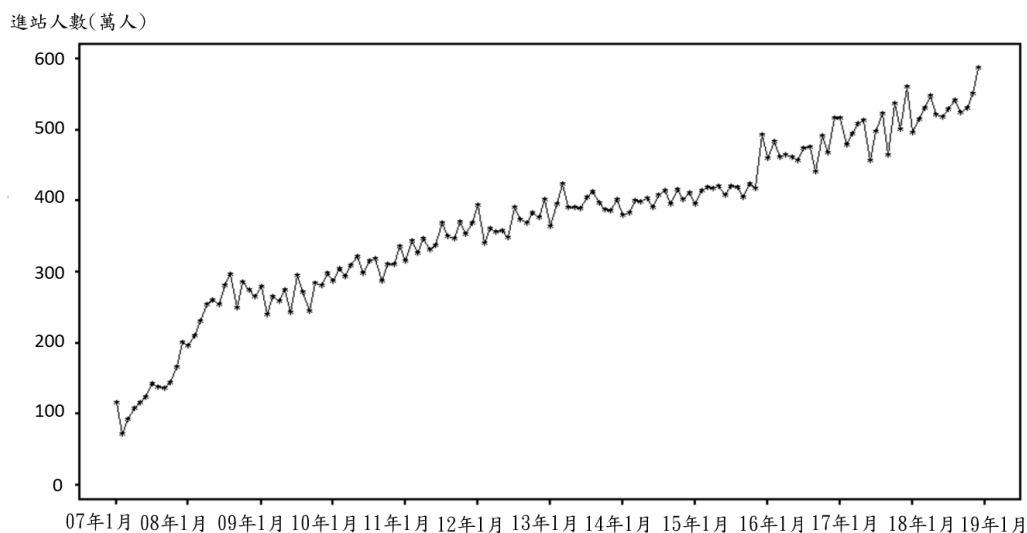
時間序列迴歸模型通常以趨勢、季節與誤差項來描述時間序列，其參數不會隨著時間改變。此模型的誤差項通常會有自我相關，因此可透過配適 ARIMA 模型來處理。時間序列迴歸模型與一般迴歸模型的差別在於其觀察值之間存在自我相關，與迴歸模型中觀察值獨立的假設不同。

時間序列分解法是起源於 20 世紀的一種分析方法，包括譜分析、時間序列分析和傅里葉級數分析等。使用時間序列分解可以長期追蹤數據，在指標出現上漲和下跌時能夠直接地觀察到並且去調查原因，以及透過指標的波動來判斷未來趨勢。其分析步驟也相對簡單，是其他時間序列分解法的基礎。分解法可以將時間序列表示成趨勢、季節、循環及不規則項這四個因素的函數，其常見的方法有兩種，加法模型(additive model)和乘法模型(multiplicative model)。

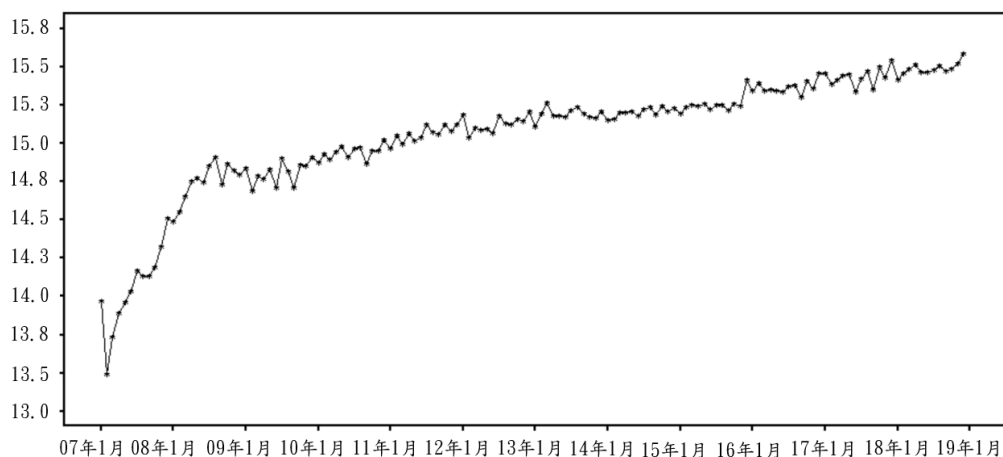
第一節、ARIMA 分析法

(一) 確認資料平穩

在 ARIMA 分析中資料的振幅一致而且資料為平穩的時間序列才能建立模型，若資料的振幅不一致可將資料做轉換來處理；若資料為不平穩的時間序列則可用差分來處理。由圖 3-1 可以得知資料的振幅並非一致，因此將資料取自然對數(nature log)做轉換。由圖 3-2 可以得知取自然對數的資料振幅為一致了。再來確認資料是否為平穩的時間序列，此處可透過自我相關係數(SAC)圖來判斷，若自我相關係數圖快速的截斷(cuts off fairly quickly)或快速的下降(dies down fairly quickly)，則表示這筆資料是平穩的。

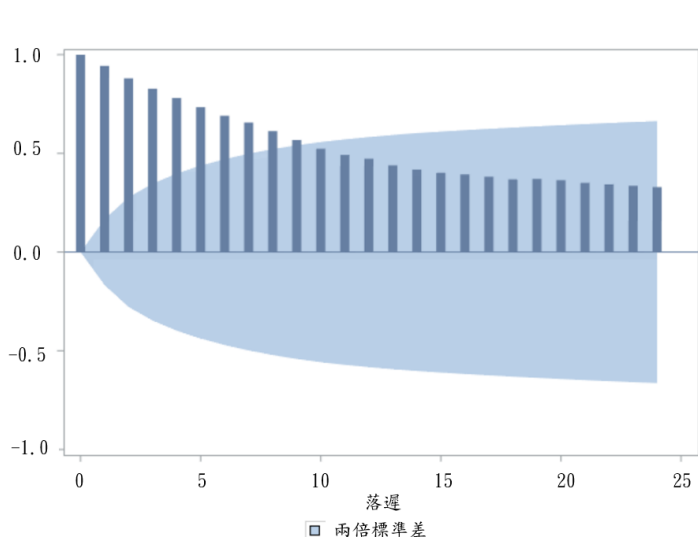


< 圖 3-1 >、高鐵每月旅客進站人數原始資料的時間序列圖

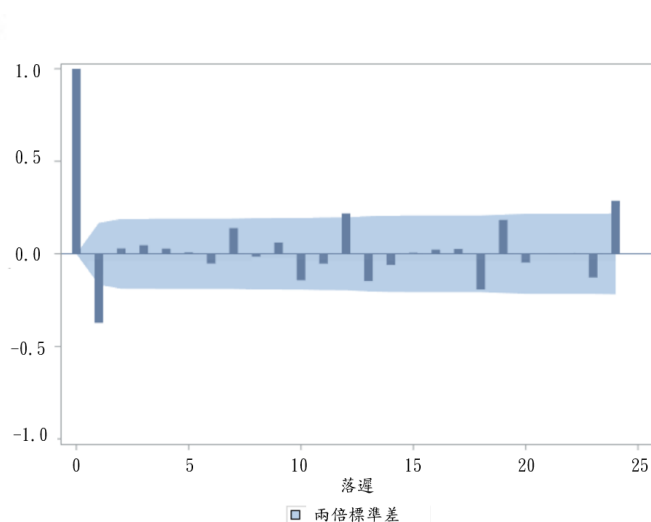


< 圖 3-2 >、原始資料取自然對數的時間序列圖

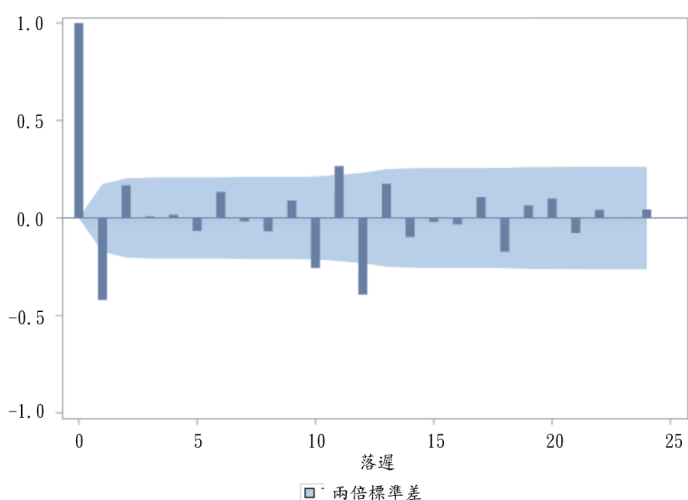
由圖 3-3 可觀察到在非季節性的部分自我相關係數圖非常緩慢的下降 (dies down extremely slowly)，表示資料在非季節性的部分不平穩，因此將取對數轉換的資料做一次差分來處理。做完處理後由圖 3-4 可以發現在非季節性的部分自我相關係數圖快速的截斷，表示在非季節性的部分為平穩。同時也發現在季節性的部分落遲 12 期(lag12)、落遲 24 期(lag24)都有凸起(spike)，表示資料在季節性的部分不平穩，因此將取對數轉換及一次差分的資料做季節差分來處理。由圖 3-5 可觀察到在非季節性及季節性的部分自我相關係數圖都是快速的截斷，表示資料在非季節性及季節性的部分都是平穩。因此可將取自然對數、一次差分及季節差分後的資料建立模型。



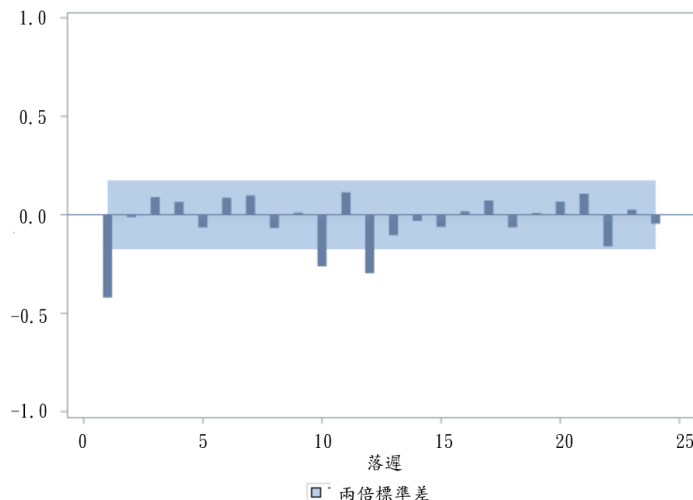
< 圖 3-3 >、取對數(log)轉換資料的自我相關係數圖



< 圖 3-4 >、取對數轉換資料做一次差分的自我相關係數圖



< 圖 3-5 >、取對數轉換資料做一次差分及季節差分的自我相關係數圖



< 圖 3-6 >、取對數轉換資料做一次差分及季節差分的偏自我相關係數圖

(二)配適模型

此處將透過自我相關係數圖及偏自我相關係數(SPAC)圖來判斷要配適的模型。在 Bowerman et al.[4]這本書中建議使用三階段步驟(three-step procedure)來初步確定模型，因此我們使用這個方法來判斷要配適的模型。

步驟一：非季節性的部分自我相關係數圖在落遲 1 期有凸起而且在落遲 1 期之後截斷(cuts off)，樣本偏自我相關係數圖下降(dies down)。因此配適非季節性移動平均模型(nonseasonal moving average model)。

步驟二：在季節性的部分自我相關係數圖在落遲 10 期、11 期和 12 期有凸起而且在落遲 12 期之後截斷，偏自我相關係數圖下降。因此配適季節性移動平均模型(seasonal moving average model)。

步驟三：結合在步驟一和步驟二得到的模型，因此配適移動平均模型。

由表 3-1 可以看到透過三階段步驟我們嘗試配了三個模型，最後選擇配的模型是 MA(1,12)。

〈表 3-1〉、移動平均模型比較表

模型	不選擇的原因	最後配適的模型
MA(1,10,11,12)	因為落遲 11 期不顯著，不使用此模型	X
MA(1,10,12)	基於模式精簡原則，不選擇參數多的模型	X
MA(1,12)	無	O

註:以上皆為經過取自然對數、一次差分及一次季節差分後所配適的模型

(三)參數檢定

在建立模型後可以得到此模型的參數估計值，此處將透過 p 值(p-value)來判斷各個參數是否皆顯著。假設顯著水準為 0.05，從表 3-2 可以得知參數估計的 p 值皆小於 0.05，表示此模型的參數都顯著。

< 表 3-2 > 、MA(1,12)模型的參數估計分析

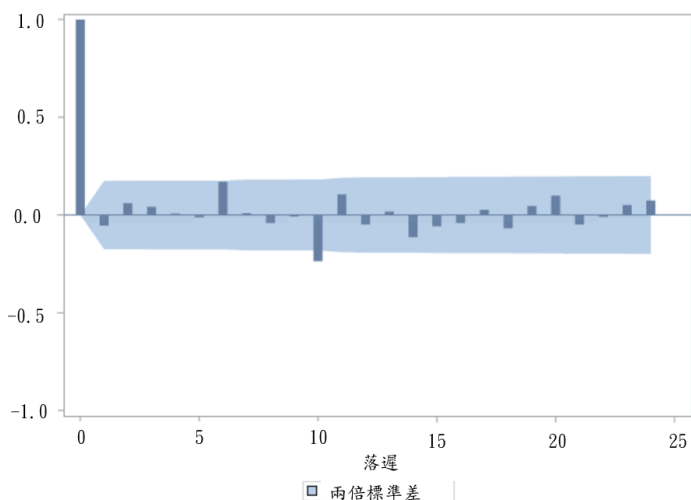
條件最小平方估計					
參數	估計值	標準差	t 值	p 值	落遲
MA1,1	0.34092	0.08285	4.12	<.0001	1
MA2,1	0.38107	0.08250	4.62	<.0001	12

(四)模式診斷

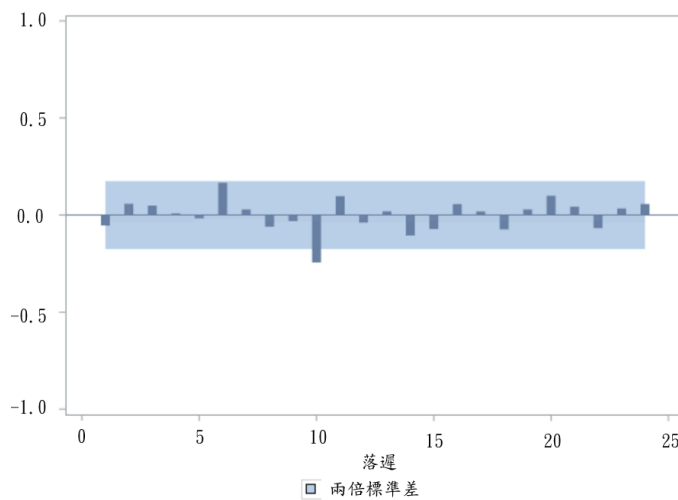
此處將透過殘差是否有自我相關來判斷模型的好壞，若殘差間有自我相關，表示這個模型不好；如果殘差間無自我相關，則表示這個模型很好。以下我們將對殘差進行分析，由 Ljung-Box 檢定及觀察殘差的自我相關係數圖、偏自我相關係數圖可以判斷殘差是否有自我相關。假設顯著水準為 0.05，從表 3-3 可以得知 p 值皆大於 0.05，表示殘差間無自我相關。由圖 3-7、圖 3-8 可觀察到大部分的落遲期數皆落在兩倍標準差以內，其中在落遲 10 期(lag10)有一點突出，但還在我們可接受的範圍內，因此殘差並無自我相關。

< 表 3-3 > 、殘差的 Ljung-Box 分析

殘差的自我相關檢查									
落遲	卡方值	自由度	p 值	自我相關係數					
6	6.19	4	0.1855	-0.036	0.078	0.059	0.024	0.004	0.182
12	15.49	10	0.1152	0.025	-0.026	0.007	-0.220	0.115	-0.038
18	18.44	16	0.2989	0.027	-0.103	-0.049	-0.033	0.033	-0.060
24	22.27	22	0.4436	0.050	0.103	-0.042	-0.004	0.055	0.078



〈圖 3-7〉、殘差的自我相關係數圖



〈圖 3-8〉、殘差的偏自我相關係數圖

第二節、時間序列迴歸法

(一) 模式建立

透過原始資料的時間序列圖可觀察到進站人數有隨著時間長期上升的趨勢，又於 ARIMA 分析中得知時間序列具有季節性，因此以趨勢及季節性兩個部分建立以下模型。

原始模型：

$$y_t = TR_t + SN_t + \varepsilon_t$$

$$= \beta_0 + \beta_1 t + \beta_2 M_1 + \beta_3 M_2 + \dots + \beta_{12} M_{11} + \varepsilon_t, \varepsilon_t \stackrel{iid}{\sim} N(0, \sigma^2)$$

其中 M_i 代表季節影響的虛擬變數，舉例來說：

$$M_1 = \begin{cases} 1, & \text{如果第 } t \text{ 期是一月} \\ 0, & \text{其他} \end{cases}$$

假設顯著水準為 0.05，從表 3-4 可以得知虛擬變數的 p 值皆大於 0.05，表示虛擬變數並不顯著，因此將季節性的部分拿掉後再重新配適模型。表 3-5 為重新配模後的參數估計分析，在此模型中參數估計的 p 值皆小於 0.05，表示參數估計都顯著，因此決定採用此模型。

最終模型：

$$y_t = TR_t + \varepsilon_t = \beta_0 + \beta_1 t + \varepsilon_t, \varepsilon_t \stackrel{iid}{\sim} N(0, \sigma^2)$$

< 表 3-4 >、原始模型的參數估計分析

參數估計					
變數	自由度	參數估計值	標準差	t 值	p 值
截距	1	14.52817	0.06257	232.17	<.0001
t	1	0.00818	0.00038312	21.36	<.0001
M_1	1	-0.7779	0.07786	-1.00	0.3196
M_2	1	-0.11991	0.07784	-1.54	0.1259
M_3	1	-0.08286	0.07783	-1.06	0.2900
M_4	1	-0.06491	0.07781	-0.83	0.4057
M_5	1	-0.06297	0.07780	-0.81	0.4198
M_6	1	-0.10087	0.07778	-1.30	0.1970
M_7	1	-0.02991	0.07777	-0.38	0.7012
M_8	1	-0.04199	0.07777	-0.54	0.5902
M_9	1	-0.11325	0.07776	-1.46	0.1477
M_{10}	1	-0.04928	0.07775	-0.63	0.5273
M_{11}	1	-0.06700	0.07775	-0.86	0.3904

< 表 3-5 >、最終模型的參數估計分析

參數估計					
變數	自由度	參數估計值	標準差	t 值	p 值
截距	1	1797415	52381	34.31	<.0001
t	1	25857	626.78622	41.25	<.0001

(二)模式診斷與調整

此處先透過 DW 值來分析殘差，從表 3-6 可以得知 DW 值為 0.633，表示殘差有一階自我相關。但是藉由 DW 值只能確定殘差在落遲 1 期時有問題，無法確定其他的落遲期數有沒有問題，以下將透過 Ljung-Box 檢定、殘差的自我相關係數圖及偏自我相關係數圖來判斷殘差是否有自我相關。假設顯著水準為 0.05，從表 3-7 中可以看到 p 值都很小，代表殘差有自我相關。再透過圖 3-12、圖 3-13 可以觀察到並非所有落遲期數皆落在兩倍標準差之內，因此我們將對殘差配適 ARIMA 模型來進行修正。因為殘差的自我相關係數圖下降，偏自我相關係數圖在落遲 1 期、落遲 2 期及落遲 12 期有突起，故對殘差配適自迴歸模型。透過圖 3-14 和圖 3-15 可以觀察到修正後模型所有的落遲期數皆落在兩倍標準差之內，表示殘差沒有自我相關。此處可確定為最後的模型。

修正後模型：

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t ,$$

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_{1,12} \varepsilon_{t-12} - \phi_1 \phi_{1,12} \varepsilon_{t-13} + \phi_2 \phi_{1,12} \varepsilon_{t-14} + a_t ,$$

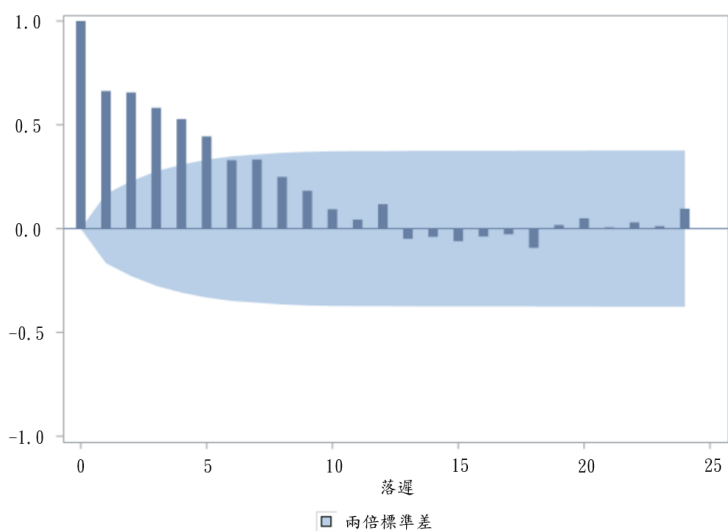
$$a_t \stackrel{iid}{\sim} N(0, \sigma^2)$$

< 表 3-6 > 、DW 統計量表

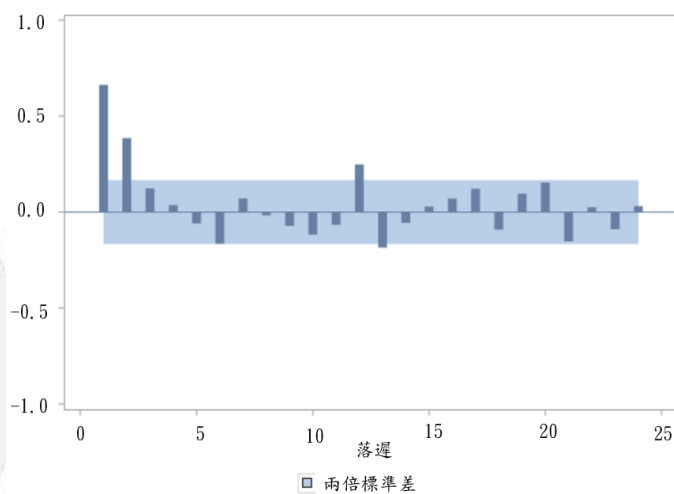
Durbin-Watson	0.633
觀察值個數	144
一階自我相關係數	0.663

〈表 3-7〉、Ljung-Box 統計量表

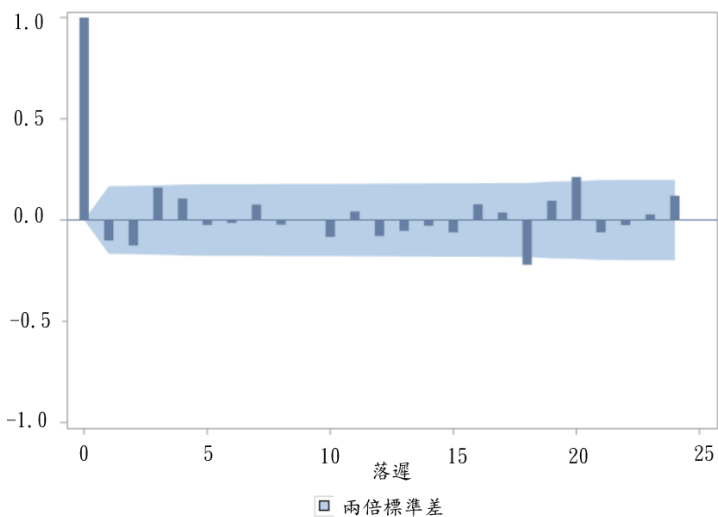
殘差的自我相關檢查									
落遲	卡方值	自由度	p 值	自我相關係數					
6	266.95	6	<.0001	0.663	0.656	0.582	0.528	0.444	0.329
12	302.56	12	<.0001	0.333	0.249	0.183	0.093	0.044	0.118
18	305.59	18	<.0001	-0.049	-0.040	-0.061	-0.038	-0.027	-0.092
24	307.86	24	<.0001	0.017	0.050	0.006	0.030	0.012	0.096



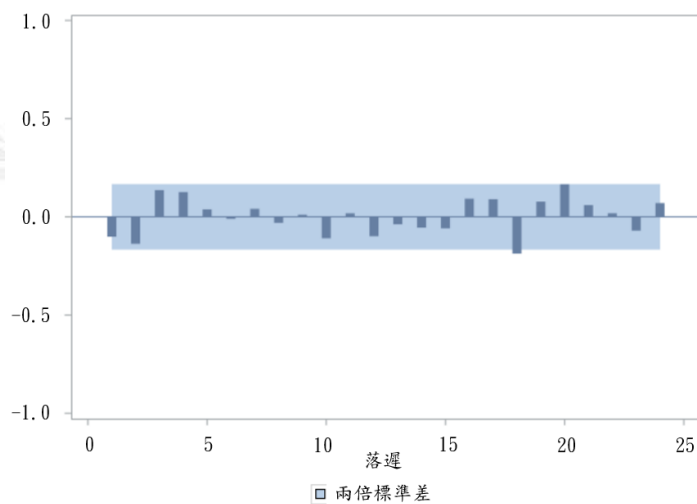
〈圖 3-9〉、殘差的自我相關係數圖



〈圖 3-10〉、殘差的偏自我相關係數圖



〈圖 3-11〉、修正後模型殘差的自我相關係數圖



〈圖 3-12〉、修正後模型殘差的偏自我相關係數圖

第三節、分解法

時間序列分解法可以分成兩種常用模型：一為乘法分解模型，另一個為加法分解模型，本章節會個別使用乘法以及加法分解模型對我們的資料進行預測。

·乘法模型： $y_t = TR_t \times SN_t \times CL_t \times IR_t$

·加法模型： $y_t = TR_t + SN_t + CL_t + IR_t$

TR_t : Trend 趨勢

SN_t : Season 季節

CL_t : Cyclical 循環

IR_t : Irregular 不規則項

乘法分解法

(一)預測模型

Bowerman et al.[4]於第七章說明分解法是一種沒有理論依據的方法也就是一種直覺的方法，如何決定哪些因子(component)會影響我的時間 t 中時間序列的觀測值(y_t)，首先我從原始資料的時間序列圖進行觀察如圖 3-1，發現時間序列有上升的趨勢，所以會將趨勢因子納入考量，回顧第三章 ARIMA 分析，發現 ARIMA 模型有季節性，所以也會將季節因子納入考量，從表 3-8 看出循環因子前幾期係數並不是皆為 1(在乘法的概念裡，係數為 1 代表此因子對時間(t)中時間序列的觀測值不會有影響)，所以也會將循環因子納入考量，不規則變動因子不存在，因為是隨機誤差對於新的期數是沒有影響的。最後就可以選擇我們乘法模型的因子如下：

$$y_t = TR_t \times SN_t \times CL_t$$

根據 Bowerman et al.[4]於第七章所提到各個因子(component)的計算方式，我們先採用乘法模型，其中表頭符號意義如下：

y_t : 在時間區間 t 下觀察到時間序列

sn_t : 在時間區間 t 下點預測的季節因子

d_t : 在時間區間 t 下的去季節因子

tr_t : 在時間區間 t 下點預測的趨勢因子

cl_t : 在時間區間 t 下點預測的循環因子

ir_t : 在時間區間 t 下點預測的不規則變動因子

由表 3-8 Excel 分解法輸出結果，得出的乘法分解法點預測 \hat{y}_t 如下：

$$\hat{y}_t = tr_t \times sn_t \times cl_t$$

$$tr_t = 1797437 + 25860t$$

< 表 3-8 > 、Excel 乘法分解法輸出結果

t	y_t	sn_t	d_t	tr_t	cl_t	ir_t
1	1161047	0.996216899	1165456.038	1823297	0.506726888	-
2	724784	0.991880319	730717.1908	1849157	0.506726888	-210303
3	919455	1.009377506	910912.9089	1875017	0.479746254	12565.7
4	1076413	1.014350215	1061184.77	1900877	0.545871059	26086.02
5	1155098	1.010063749	1143589.205	1926737	0.603885702	-2498.43
6	1241227	0.963356348	1288440.153	1952597	0.649513835	-32750.2
7	1425755	1.036675592	1375314.526	1978457	0.676258869	78832.06
8	1373686	1.017204333	1350452.368	2004317	0.695252682	-15233.6
9	1367236	0.939478043	1455314.48	2030177	0.697723349	-29195.6
10	1448553	1.002817464	1444483.221	2056037	0.745314697	-43254.9
11	1659506	0.976200417	1699964.445	2081897	0.810239789	-44100.3
12	2002896	1.042379117	1921465.969	2107757	0.883113203	129374.7
13	1958004	0.996216899	1965439.456	2133617	0.936991932	-60678.3
14	2095210	0.991880319	2112361.704	2159477	0.982469909	-26460.4
15	2311821	1.009377506	2290343.293	2185337	1.05371496	-5703.63
16	2545527	1.014350215	2509514.921	2211197	1.111708872	58607.69
17	2603395	1.010063749	2577456.129	2237057	1.150374842	41197.23
18	2537620	0.963356348	2634144.681	2262917	1.167116318	-113335
19	2811997	1.036675592	2712513.946	2288777	1.203986474	37707.72
20	2973150	1.017204333	2922864.074	2314637	1.193225007	215215.8

(二)預測後的數據比對

< 表 3-9 > 、乘法分解法真實值與預測值

時間	真實值 y_t	預測值 \hat{y}_t	95%預測區間下界	95%預測區間上界
2019/01	5326693	5464464.281	4869261.281	6059667.28
2019/02	5425372	5426168.063	4830795.563	6021540.56
2019/03	5648453	5606632.989	5011088.489	6202177.49
2019/04	5678567	5618091.831	5022372.831	6213810.83
2019/05	5475139	5670627.501	5074732.001	6266523
2019/06	5610936	5352390.465	4756315.965	5948464.96
2019/07	5621665	5775662.420	5179406.420	6371918.42
2019/08	5559008	5678083.894	5081644.394	6274523.39
2019/09	5505456	5357776.072	4761151.072	5954401.07
2019/10	5801186	5838914.779	5242101.779	6435727.78
2019/11	5683553	5673438.171	5076434.671	6270441.67
2019/12	6075220	6085036.464	5487840.964	6682231.96

加法分解法

(一)預測模型

如何決定哪些因子(component)會影響我的時間(t)中時間序列的觀測值，如同乘法模型挑選的原因，所以最後就可以選擇我們加法模型的因子如下：

$$y_t = TR_t + SN_t + CL_t$$

由表 3-10 Excel 分解法輸出結果，得出的加法分解法點預測 \hat{y}_t 如下：

$$\hat{y}_t = tr_t + sn_t + cl_t$$

$$tr_t = 1797415 + 25857t$$

< 表 3-10 > 、Excel 加法分解法輸出結果

t	y_t	sn_t	d_t	tr_t	cl_t	ir_t
1	1161047	-4.17961789	1161051.2	1823272	-914034.747	-
2	724784	-6.85518276	724790.86	1849129	-914034.748	-210303.39
3	919455	14.2749116	919440.73	1874986	-968110.974	12565.699
4	1076413	19.5034743	1076393.5	1900843	-850535.523	26086.019
5	1155098	9.79024538	1155088.2	1926700	-769113.356	-2498.434

6	1241227	-51.2260031	1241278.2	1952557	-678528.606	-32750.168
7	1425755	36.2530814	1425718.7	1978414	-631527.311	78832.058
8	1373686	22.9053557	1373663.1	2004271	-615374.323	-15233.582
9	1367236	-72.1893985	1367308.2	2030128	-633624.222	-29195.589
10	1448553	12.9497243	1448540.1	2055985	-564190.004	-43254.946
11	1659506	-30.7489178	1659536.7	2081842	-378204.908	-44100.343
12	2002896	61.5223274	2002834.5	2107699	-234239.198	129374.68
13	1958004	-4.17961789	1958008.2	2133556	-114869.496	-60678.325
14	2095210	-6.85518276	2095216.9	2159413	-37735.7467	-26460.398
15	2311821	14.2749116	2311806.7	2185270	132240.3589	-5703.634
16	2545527	19.5034743	2545507.5	2211127	275772.8105	58607.686
17	2603395	9.79024538	2603385.2	2236984	325203.9774	41197.232
18	2537620	-51.2260031	2537671.2	2262841	388164.7276	-113334.50
19	2811997	36.2530814	2811960.7	2288698	485555.0225	37707.724
20	2973150	22.9053557	2973127.1	2314555	443356.3437	215215.75

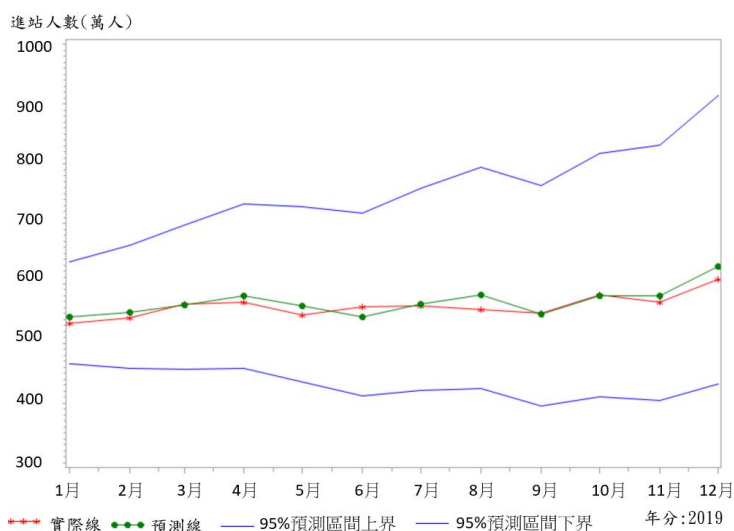
(二)預測後的數據比對

< 表 3-11 > 、加法分解法真實值與預測值

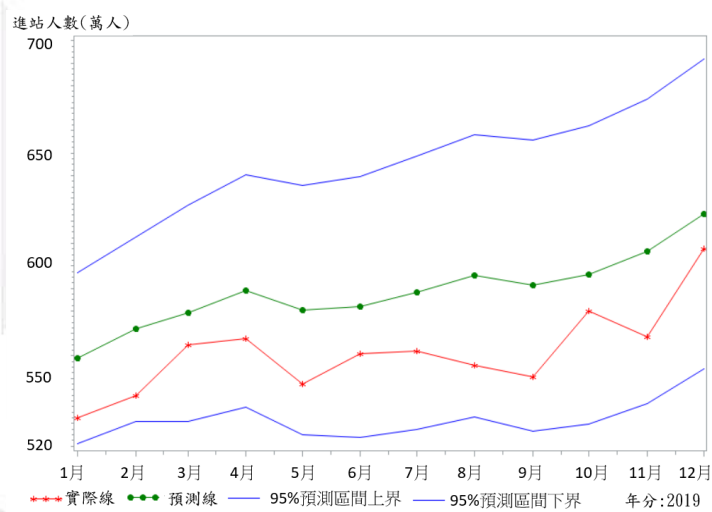
時間	真實值 y_t	預測值 \hat{y}_t	95%預測區間下界	95%預測區間上界
2019/01	5326693	5542948.991	4916302.491	6169595.491
2019/02	5425372	5466831.398	4840006.398	6093656.398
2019/03	5648453	5584135.967	4957129.467	6211142.467
2019/04	5678567	5600724.647	4973534.647	6227914.647
2019/05	5475139	5588231.101	4960855.101	6215607.101
2019/06	5610936	5569197.168	4941632.668	6196761.668
2019/07	5621665	5597236.609	4969481.609	6224991.609
2019/08	5559008	5562070.249	4934121.749	6190018.749
2019/09	5505456	5621823.255	4993679.255	6249967.255
2019/10	5801186	5663441.279	5035099.279	6291783.279
2019/11	5683553	5853274.343	5224732.343	6481816.343
2019/12	6075220	5879223.615	5250478.615	6507968.615

第四章、模型的預測能力及比較

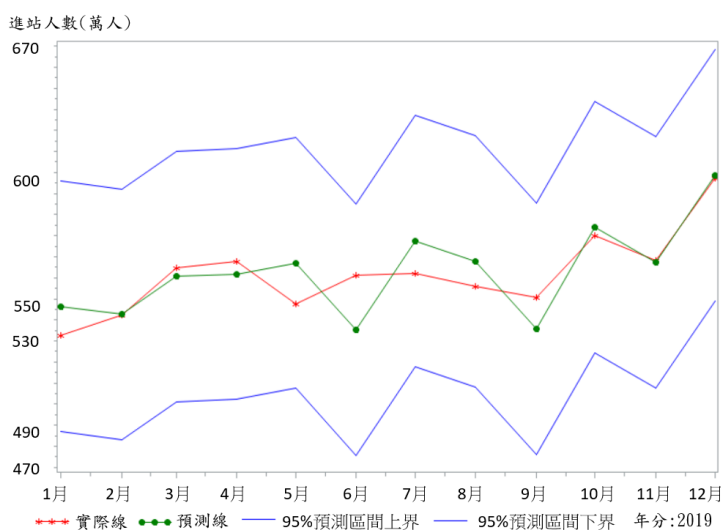
在第三章我們透過三個分析方法建構了四個模型，此章將討論這四個模型的預測結果，我們先藉由預測圖來看模型的預測能力。在四張預測圖中，紅色線是2019年12個月份高鐵旅客進站的實際人數，而綠色線是透過模型估計出來的預測值，藍色線則是95%預測區間的上下界。由圖中可觀察到2019年12筆的預測值跟真實值皆落在95%預測區間之內，表示這四個模型的預測結果還算不錯。而透過比較四張預測圖，可觀察到ARIMA模型預測圖的實際值及真實值很相近，因此以預測圖作為判斷標準我們認為ARIMA模型為最佳模型。



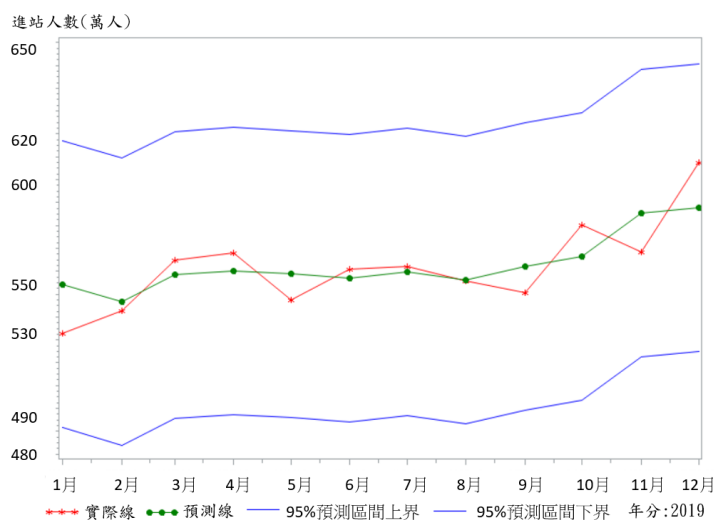
< 圖 4-1 >、ARIMA 模型的預測圖



< 圖 4-2 >、時間序列迴歸模型的預測圖



< 圖 4-3 >、乘法分解模型的預測圖



< 圖 4-4 >、加法分解模型的預測圖

若只將預測圖作為判斷最佳模型的標準並不精確，因此以下使用了三個常用的評估指標來衡量各個模型的預測能力。我們將透過平均絕對誤差(mean absolute deviation)、均方誤差(mean square error)及平均絕對百分比誤差(mean absolute percentage error)選擇最佳模型。

三個評估指標的定義：

$$\text{平均絕對誤差} = \frac{\sum |y_t - \hat{y}_t|}{n}$$

$$\text{均方誤差} = \frac{\sum (y_t - \hat{y}_t)^2}{n}$$

$$\text{平均絕對百分比誤差} = \frac{\sum \frac{|y_t - \hat{y}_t|}{y_t} \times 100}{n}$$

均方誤差為三種評估指標中容易計算的一種，其計算方式為誤差取平方後加總再除以樣本數。透過公式可以得知，此評估指標因平方的計算會給予較大的誤差更多權重；而平均絕對誤差同樣為三種評估指標中容易計算的一種，其計算方式為誤差取絕對值後加總再除以樣本數。透過公式可以得知，此評估指標給予全部誤差相同的權重；平均絕對百分比誤差是由平均絕對誤差衍生而得，所以這兩種評估指標會有相同的結果。觀察公式後可以發現，當真實值為 0 時此公式並不成立，因此在使用平均絕對百分比誤差作為評估指標時須注意真實值是否為 0。

< 表 4-1 >、各個模型的比較

分析方法	平均絕對誤差	均方誤差	平均絕對百分比誤差
ARIMA 模型	106223.87	17172513325.71	1.8874
時間序列迴歸模型	266961.58	79688653943.33	4.7845
乘法分解模型	97775.82	15897919599.21	1.7615
加法的分解模型	100168.83	14462882370.32	1.7778
最佳模型	乘法的分解模型	加法的分解模型	乘法的分解模型

將四個模型 2019 年 12 個月份進站人數的預測值，代入上述這三種評估指標的公式做運算，以相同的真實值進行比較，誤差越小即預測值越接近真實值，意指其模型配適的越好。故看到下方表 4-1，可見不同指標下預測能力最好的有乘法分解模型及加法分解模型，而乘法分解模型在平均絕對誤差、平均絕對百分比誤差之下為最佳，因此以評估指標作為判斷標準我們選擇乘法分解模型為最佳模型。



第五章、結論

我們從各個模型可得知，所有預測模型的進站人數皆為穩定上升。在最終模型的預測可發現，除了6月、9月和12月，大致上旅客數皆有增加，不過，實際上那三個月與其他月的差異沒有預測那麼大的差異；再去比較原始資料，發現早期幾年，有幾年三個月確實是下降趨勢，因此我們猜測高鐵營運之前，國內只有客運與國內飛機等，價錢和花費時間極端反比的選擇造成流行國外旅遊。高鐵的出現增加了國內旅遊交通工具的選擇，間接提升大家國內旅遊的意願，這同時解釋為甚麼後幾年那三個月的差異較小。

那麼高鐵方面可以根據我們上面猜測，暑假、春節及投票日時，在旅客多的車站需有應對的措施，如與當地政府協商增設高鐵附近公車火接駁車，減少民怨又能增加更多商機，淡季的部分減少班次或增加一些限時優惠票等，以保持旅客數，就如高鐵的企業願景一樣，既可以提升內部營業額，同時帶動社會旅遊風氣，提升當地觀光業至新的高度。

參考文獻

1. 范俊海、顏志宇。「應用類神經網路預測高鐵車站之運量以左營站為例」。第二十三屆海峽兩岸都市交通學術研討會論文集。民 104。上網日期：109 年 9 月 30 日，<<http://tkuir.lib.tku.edu.tw:8080/dspace/handle/987654321/103637>>。
2. 柯旻嬋、邱煜勝、劉秋蘭、馬偉誠。「台灣高鐵運輸需求預測及對運輸市場之衝擊分析」。逢甲大學優質學生報告，台中市：逢甲大學。民 97。上網日期：109 年 9 月 30 日，<<http://dspace.lib.fcu.edu.tw/handle/2377/11035>>。
3. 台灣高鐵，營運資訊－客運運輸情形，上網日期：2020 年 09 月 30 日，擷取自：
<https://www.thsrc.com.tw/ArticleContent/117f6de2-ed8b-403a-ab4a-820d123e37bf>
4. Bowerman, B. L., O'Connell, R., and Koehler, A. (2004) *Forecasting, Time Series, and Regression*, 4th ed, Cengage Learning.