主題：國道縱向車禍之事故車數量預測

組員：王冠勛、蘇柏洋、張承暐、李世澤

開課老師：陳婉淑教授

課程名稱：預測分析

開課系所：統計學系

開課學年：112 學年度 第 2 學期

目錄

摘要

國道車禍事故乃是影響道路交通安全的重要因素，本研究旨在利用 SAS 軟體，基於國道一號北向車禍歷史數據，採用多種時間序列分析方法對未來車禍車輛數量進行預測，為交通管理決策提供數據支持，以減少事故發生率。本報告採用 113 年國道智慧交通管理創意競賽所提供之民國 112 年 1 月至 10 月的國道一號北向車禍數據，在進行了必要的數據清洗和整理後，分別應用 ARIMA 模型、時間序列迴歸模型、指數平滑法模型、UCM 模型對數據進行了分析和建模，並根據均方誤差(MSE)、平均絕對誤差(MAE)、均絕對百分比誤差(MAPE)等評估指標對預測結果的準確性進行了綜合評估。

從分析結果可知 ARIMA 模型能較好捕捉數據趨勢和季節性變化；時間序列迴歸模型通過引入節日虛擬變數等外生變量提高了預測準確性；而指數平滑法則能降低數據隨機波動的影響。然而，在節假日等特殊時期，所有模型的預測結果均存在較大偏差，這可能是由於節日期間車流量和人員流動的劇烈變化所導致。通過評估指標對比，我們發現 ARIMA 和時間序列迴歸模型在均方誤差和平均絕對誤差上表現優異，而指數平滑法則在均絕對百分比誤差方面較為出色。

關鍵字：ARIMA 模型、時間序列迴歸模型、指數平滑法模型、UCM 模型

Abstract

This study aims to utilize SAS software to forecast the number of vehicles involved in freeway longitudinal traffic accidents based on historical data, adopting multiple time series analysis methods to provide data support for traffic management decisions.

In the research process, we first collected accident data from January to October 2022 for the northbound directions of National Highway No.1 in Taiwan. After performing necessary data cleaning and processing, we applied the ARIMA model, time series regression model, exponential smoothing method and UCM Procedure to analyze the data. The prediction results were comprehensively evaluated based on evaluation metrics such as mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

Specifically, the ARIMA model effectively captured data trends and seasonality; the time series regression model improved prediction accuracy by incorporating exogenous variables like holiday dummy variables; and the exponential smoothing method mitigated the impact of random fluctuations in the data. However, during special periods such as holidays, all models exhibited substantial deviations in their predictions, potentially due to the drastic changes in traffic flow and population movement during these periods.

By comparing the four evaluation metrices, we found that the ARIMA model and time series regression model performed excellently in terms of MSE and MAE, while the exponential smoothing method had a better performance in MAPE, reflecting the differences in accuracy and robustness among the models.

Key word：ARIMA model, time series regression model, exponential smoothing method, UCM

一、 簡介

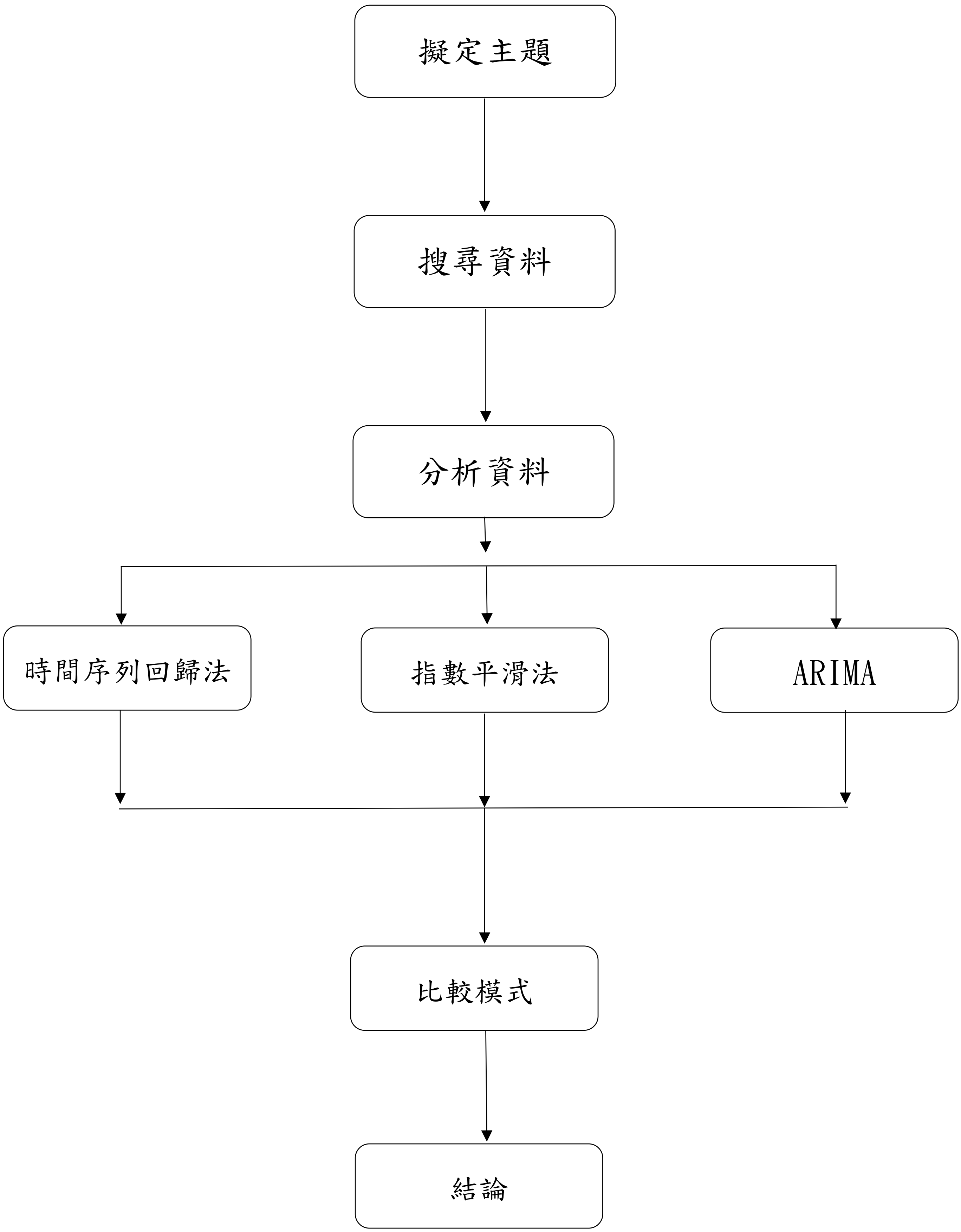
作為臺灣的現代化交通命脈，國道網絡的暢通直接關係到整體社會經濟運行的效率。據統計，國道每年發生上萬起各類交通事故，不僅影響著國道的通暢效率，進而加劇交通擁堵、耗能增加等連鎖反應，對經濟發展和生態環境造成負面影響。在此背景下，準確預測交通事故發生情況，對於制定科學有效的管控措施至關重要，故本報告專注於預測國道縱向車禍事故車輛數量變化趨勢，旨在為交通主管部門提供決策支持,助力交通安全管理水平的提升。

近年來，隨著大數據技術的迅速發展，數據預測的智慧交通管理模式日益受到重視。大數據分析和預測可以為決策者提供可靠的先期指引，使管理措施更加主動有為、精準高效，而在國道車禍預防領域，透過建立適當的時間序列預測模型，對歷史車禍數據進行深入分析並預測未來發展趨勢，將為制定針對性的管控舉措提供重要參考。

本報告使用 113 年國道智慧交通管理創意競賽所提供之民國 112 年 1 月至 10 月的國道一號北向車禍數據，保留最後 7 筆數據做為預測準確度參考，使用了四種分析模型:(1)ARIMA 模型(2)時間序列迴歸模型(3)指數平滑法模型(4)UCM 模型預測未來車禍車輛數量變化情況，並以均方誤差(MSE)、平均絕對誤差(MAE)、均絕對百分比誤差(MAPE)等評估指標對預測結果的準確性進行綜合評估。

本研究為國道車禍事故預測提供了有益的模型參考和技術支持，有望為提升國家交通運行效率貢獻綿力，未來我們可以將預測分析範圍拓展至其他道路領域，並持續探索時間序列分析在智慧交通等更廣泛場景中的應用前景。隨著相關技術的不斷完善，數據驅動的交通管理將為構建安全、便捷、高效的現代化綜合交通體系注入新的動能。

研究流程



二、 資料來源

本報告使用 113 年國道智慧交通管理創意競賽所提供之 112 年 1 月至 10 月的交通事故資料，將原始國道一號北向肇事車輛數整理成日資料做分析，保留七筆數據測試該模型預測的優劣（圖(二)），用來預測未來七天(黃色區塊)的發生事故車輛。

國道一號北向數據

月	日	星期	肇事車輛數
1	1	星期日	34
1	2	星期一	58
1	3	星期二	26
1	4	星期三	39
1	5	星期四	43
1	6	星期五	48
1	7	星期六	16
1	8	星期日	16
1	9	星期一	36
1	10	星期二	43
1	11	星期三	47
1	12	星期四	43
1	13	星期五	99
1	14	星期六	73
1	15	星期日	26

圖（一）國道一號北向數據

10	19	星期四	52
10	20	星期五	65
10	21	星期六	40
10	22	星期日	55
10	23	星期一	37
10	24	星期二	54
10	25	星期三	35
10	26	星期四	31
10	27	星期五	43
10	28	星期六	57
10	29	星期日	32
10	30	星期一	29
10	31	星期二	17

圖（二）數據保留的最後十二筆資

三、 分析方法

（一）分析方法

1. ARIMA

ARIMA（Autoregressive Integrated Moving Average）是一種用於時間序列預測的統計模型，通過分析數據的自相關和偏自相關特徵，確定最佳的模型參數（p, d, q），適合處理非平穩時間序列數據，並且可以進行短期預測。

- **AR 模型（自迴歸模型）**：以前期的資料來預測本期的資料，而且越接近本期的資料，對預測結果的影響力就越大，模型定義如下：

$$(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)y_t = a_t \quad , a_t \sim iid\, N(o,\sigma^2)$$

可逆條件:無

平穩條件：$(\phi_1+\cdots+\phi_p)<1, (\phi_p-\cdots-\phi_1)<1, |\phi_p|<1$

- **MA 模型（移動平均模型）**：根據時間序列資料、逐項推移，依次計算包含一定項數的序時平均值，以反映長期趨勢的方法，其模型如下：

$$y_t = (1-\theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p)a_t \quad , \quad a_t \sim iid\, N(o,\sigma^2)$$

平穩條件：無

其可逆條件: $(\theta_1 + \cdots + \theta_p) < 1, (\theta_p - \cdots - \theta_{1)} < 1, |\theta_p| < 1$

- **ARMA 模型（自回歸滑動平均模型）**：由自回歸模型（簡稱 AR 模型）與滑動平均模型（簡稱 MA 模型）為基礎“混合”構成，其模型如下

$$\left(1-\phi_1 B - \cdots - \phi_p B^p\right)(1-B)y_t = (1-\theta_1 B - \cdots - \theta_p B^p)a_t$$
$$, a_t \sim iid\, N(o,\sigma^2)$$

其平穩條件和可逆條件為: $|\phi_P| < 1$、$|\theta_P| < 1$

2. Time series regression

時間序列迴歸（Time Series Regression）是一種分析和建模時間序列數據的方法，用來探索和描述一個或多個變數（自變數）與時間序列變量（因變數）之間的關係，與傳統迴歸分析不同，時間序列迴歸考慮了數據的時間依賴性和結構，其模型如下：

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

其中，y_t = 時間序列在時間 t 的觀察值

TR_t = 時間 t 的趨勢項

SN_t = 時間 t 的季節因子

ε_t=時間 t 的誤差項

季節虛擬變數模型中$M_1...M_{11}$為季節虛擬變數，其模型如下：

$$y_t = \beta_0 + \beta_1 t + \beta_2 M_1 + \beta_3 M_2 + \beta_4 M_3 + \beta_5 M_4 + \beta_6 M_5 + \beta_7 M_6 + \beta_8 M_7$$
$$+ \beta_9 M_8 + \beta_{10} M_9 + \beta_{11} M_{10} + \beta_{12} M_{11} + \beta_{13} P_1 + \beta_{14} P_2 + \varepsilon_t$$

$$M_1 = \begin{cases} 1 & \text{if period t is January} \\ 0 & \text{otherwise} \end{cases} \cdots M_{11} = \begin{cases} 1 & \text{if period t is November} \\ 0 & \text{otherwise} \end{cases}$$

P_1, P_2 為重大事件影響點

$$P_1 = \begin{cases} 1 & \text{if period t is September 1999} \\ 0 & \text{otherwise} \end{cases} \quad P_2 = \begin{cases} 1 & \text{if period t is April 2004} \\ 0 & \text{otherwise} \end{cases}$$

3. Exponential smoothing method

Exponential smoothing method(指數平滑法)是由 Robert Goodell Brown 於 1956 年首次提出，隨後由 Charles Holt 於 1957 年進一步發展，指數平滑法是在移動平均法基礎上發展起來的一種時間序列分析預測法，它是通過計算平滑值，配合一定的時間序列預測模型對現象的未來進行預測，其原理是任一期的平滑指數值都是本期實際觀察值和前一期平滑指數值的加權平均。

- **簡單指數平滑（SES）**：又稱為單指數平滑，是指數平滑的最簡單形式，假設時間序列沒有趨勢或季節性，其模型如下：

$$F_t = \alpha A_{t-1} + (1-\alpha)F_{t-1}$$

F_t: 未來 t 時段的預測值
A_{t-1}: t-1 時段的實際值
F_{t-1}:t-1 時段的預測值
α:平滑參數在[0,1]之間

- **Holt(雙指數平滑)**：用於預測具有線性趨勢但沒有季節性模式的時間序列資料，其模型如下：

$$F_t = \alpha A_{t-1} + (1-\alpha)(F_{t-1} + \beta_{t-1})$$
$$b_t = \beta(F_t - F_{t-1}) + (1-\beta)b_{t-1}$$

b_t: 是時間 t 時趨勢的斜率和最佳估計
α:是資料的平滑參數（0 < α < 1）
β:是趨勢的平滑參數（0 < β < 1）

4. UCM

UCM 使用未觀測成分模型來分析和預測等距單變量時間序列資料，在時間序列文獻中也稱為結構模型。 UCM 將反應序列分解為趨勢、季節、週期以及預測變數序列等組成部分，其水平μ_t和斜率β_t會隨時間變化，且變化由擾動項η_t及其ε_t各自方程式的變異數控制。

Observation Equation

$$y_t = \mu_t + \gamma_t + c_t + \epsilon_t,$$

State Equations

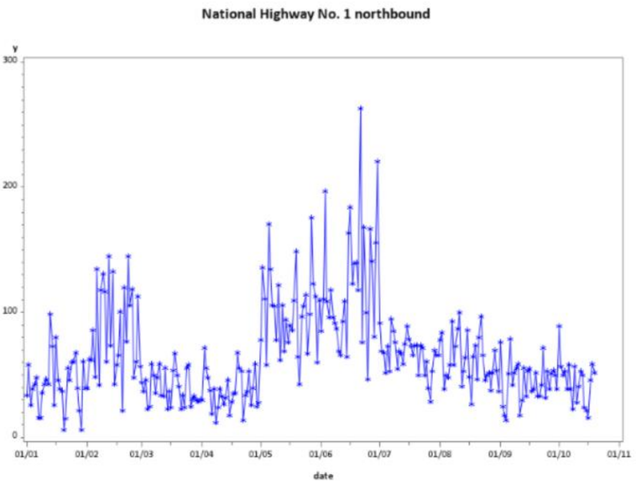
- Trend Component:
  $$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t,$$
  $$\beta_t = \beta_{t-1} + \zeta_t,$$
- Seasonal Component:
  $$\gamma_t = -\sum_{i=1}^{s-1} \gamma_{t-i} + \omega_t,$$
- Cyclical Component:
  $$c_t = c_{t-1}\cos(\lambda) + c_{t-1}^{*}\sin(\lambda) + \kappa_t,$$
  $$c_t^{*} = -c_{t-1}\sin(\lambda) + c_{t-1}^{*}\cos(\lambda) + \kappa_t^{*},$$
- Irregular Component:
  $$\epsilon_t \sim N(0, \sigma_\epsilon^2).$$

（二） 研究準則

公式:	準則：
MAE = $\sum(\frac{\lvert y_t-\hat{y}\rvert}{n})$	MAE>=0，在 0 時最好
MSE = $\sum(\frac{(y_t-\hat{y})^2}{n})$	MSE 越接近 O 越好
MPE = $\sum(\frac{\frac{y_t-\hat{y}}{y_t}}{n} * 100\%)$	MPE(平均百分比誤差)越小越好
MAPE = $\sum(\frac{\frac{\lvert y_t-\hat{y}\rvert}{y_t}}{n} * 100\%)$	MAPE(平均絕對百分比誤差)越小越好

四、分析結果

（一）時間序列圖



圖（三）國道一號北向時間序列圖

從上面圖(三)的時間序列圖中，可以看出在二月到三月之間有較為突出的現象產生，這是因為春節連假所造成的連鎖效應，而五月到七月之間其肇事車輛數也有較為突出的表現，根據我們的調查這可能是因為五月至七月是台灣梅雨季節，並且六月至七月是學生放假時間導致國內旅遊人數增多，所以使得該時段的肇事車禍數增加，而該時段在六月七日時政府開始放寬疫情限制，所以也可能是導致該時段人數遽增的原因之一。

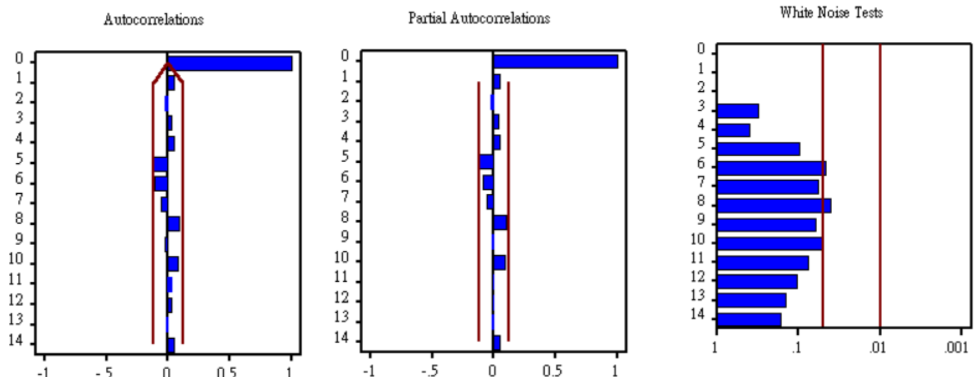
（二）ARIMA model

對此時間序列圖我們發現在 2 月 6 日時其數據比左右兩天的數據來的大上許多，這是因為該日期是在元宵節的後一天，許多人為盡快回其工作崗位而造成較多的車禍產生，而在 5 月 2 日開始其數據有較為增加的表現到 7 月 1 日，所以對於該時段我們使用 step function 來去處理這段時間的數據，而在這段時間內又有兩點其數據比起左右兩天來的大上許多，分別是 6 月 21 日以及 30 日，這是因為 6 月 21 日是端午連假的前一天，而 6 月 30 日是由於政府實施新政策在該天實施，所以最終我們使用介入以及 step function 來去配試模型。

	配試模型
國道 1 號北向 ARMA(1, 1)	$y_t = 26.1895 + 61.8624P_1 + 146.4935P_2 + 123.0248P_3 + 57.2511S_t + \frac{(1-0.17838B_1)}{(1-0.9815B_1)}a_t$ $a_t \sim iid\ N(0,\sigma^2)$、$\hat{\sigma}^2 = 679.9402$

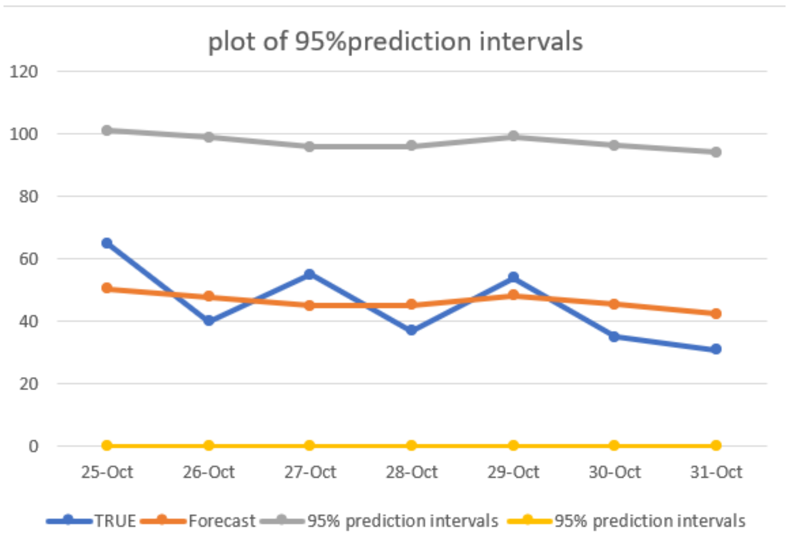
表（一） ARIMA 最終配試模型

從圖(三)中我們可以發現到有一個點較為特別，而該點是由於端午節的存在而發生與其他點相離甚遠的數據，且從該圖中發現數據並不平穩，於是我們將其進行一次差分。



圖（四）殘差圖

從圖(四)來看，可以看出其 ACF 和 PACF 都在兩倍標準差以內，而白噪音檢定也是不顯著的，表示殘差無自我相關。



圖（五）95%預測區間圖

從 95%預測區間（圖(五)）可以看出我們預測的其實還不錯，實際值都在95%信賴區間內，且預測值跟實際值的趨勢雷同。

（三） Time series regression model

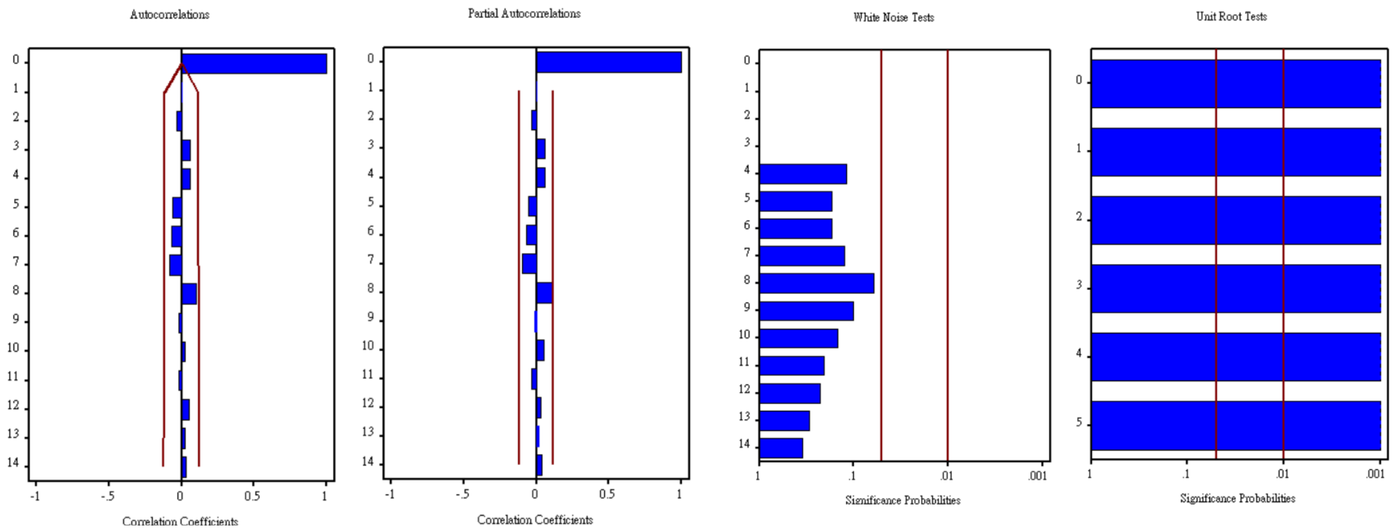
同 ARIMA 判斷，於此設置了 step function

$$S_t = \begin{cases} 0, & if\ t < 02May05 \\ 1, & if\ \ 02may05 < t < 01JUL07 \\ 0, & if\ t > 01JUL07 \end{cases}$$

最終配試模型：

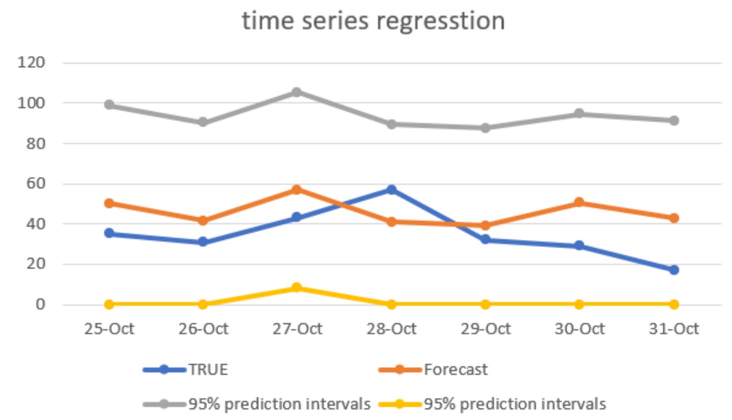
	配試模型
國道 1 號北向	$y_t = 49.2979 + 0.00284t_1 - 5.6616M_1 + 6.4239M_2 + 2.537M_3 + 2.8251M_4 - 2.9787M_5 + 13.7928M_6 + 58.7507P_1 + 143.9241P_2 + 89.7401P_3 + 56.1899S_t + \varepsilon_t$, $\varepsilon_t = 0.9016\varepsilon_{t-1} + a_t - 0.7153a_{t-1}, a_t \sim iid\ N(0,\sigma^2)$ 、$\hat{\sigma}^2 = 616.9619$

表（二）時間序列最終配適模型



圖（六）殘差圖

從圖(六)可以看出其 ACF 與 PACF 皆位於兩倍標準差內，其 White Noise 也大於 0.05，表示我們的模型是合適的。



圖（七）95%預測區間圖

從 95%預測區間圖（圖(七)）可以看出時間序列圖的預測有限，沒辦法很準確地抓到實際點的資料，但由於日資料本身起伏會來的比月資料還大，所以我們認為該預測還在可接受的範圍。

（四）Exponential smoothing model

本報告採用 Exponential smoothing 模型中 Simple Exponential Smoothing 和 Holt's linear trend model 去做挑選，最終選擇 Holt's linear trend model 並加上移動視窗的手法去配試模型。

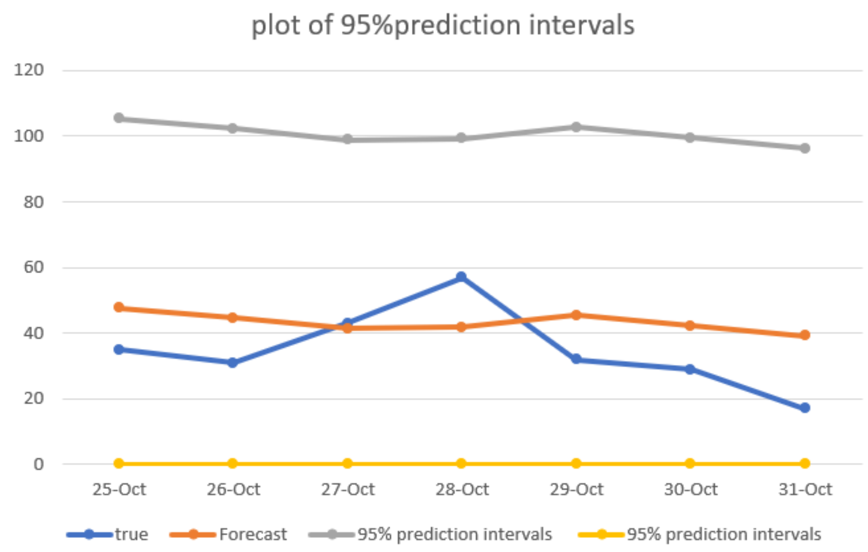
	最終配試模型
國道 1 號北向	$L_t = 0.233Y_T + (1 - 0.233)(L_{t-1} + b_{t-1})$ $b_t = 0.001(L_t - L_{t-1}) + (1 - 0.001)b_{t-1}$ $F_{t+m} = L_t + b_t m$

表（三）Exponential smoothing model 最終配適模型

Exponential smoothing			
MSE	MAE	MPE(%)	MAPE(%)
161.7017	12.7162	-36.3320	36.3320
188.1835	13.7180	-44.2516	44.2516
2.2222	1.4907	3.4667	3.4667
228.6719	15.1219	26.5296	26.5296
180.4267	13.4323	-41.9759	41.9759
175.7587	13.2574	-45.7152	45.7152
490.8086	17.0000	-130.3188	130.3188
203.9676	12.3909	-38.3710	46.9414

圖（八） MAE、MSE、MPE、MAPE 圖

從圖(八)來看，對於每個地方的數值都算中規中矩，不算太好，但也不會到太差。



圖（九）95%預測區間圖

從 95%預測區間圖（圖(九)）可以看出有抓到實際值的趨勢，且實際值皆落在 95%信賴區間，但還是能看出該模型對於預測的不足。

（五） UCM

Final Estimates of the Free Parameters					
Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	Error Variance	662.38053	65.93569	10.05	<.0001
Level	Error Variance	51.52915	18.84493	2.73	0.0062

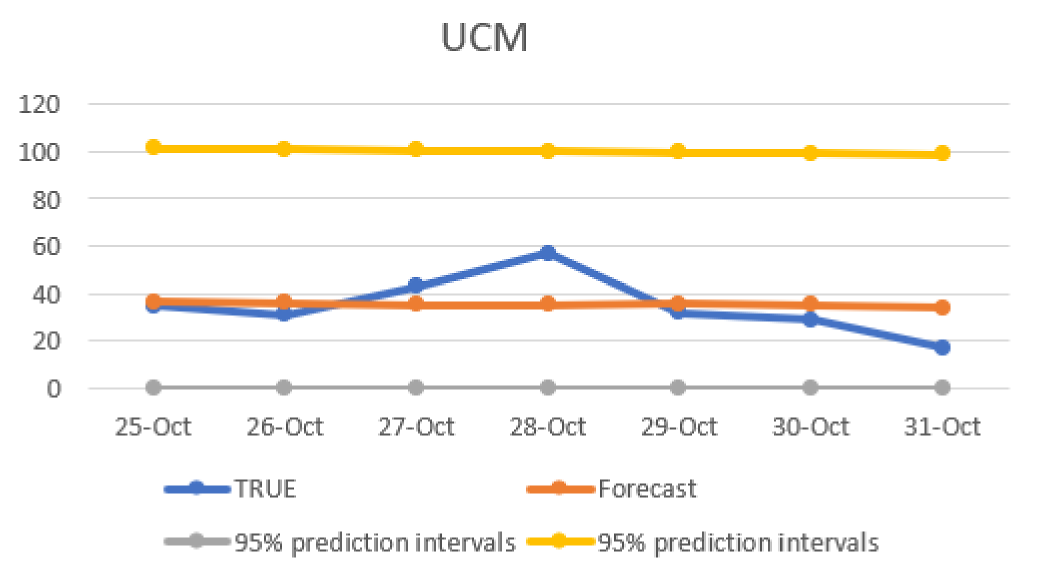
圖（十）參數估計表

圖(十)顯示 UCM 模型的自由參數的最終估計值，可以看到不規則參數(Irregular Component)呈現顯著，而水準參數則呈現略為顯著。

Significance Analysis of Components (Based on the Final State)			
Component	DF	Chi-Square	Pr > ChiSq
Irregular	1	0.46	0.4966
Level	1	11.51	0.0007
Slope	1	0.00	0.9607

圖（十一）參數卡方表

圖(十一)顯示 UCM 模型的自由參數的最終估計值，可以看到斜率參數(Slope Component)呈現顯著；不規則參數呈現略為顯著，水準參數則呈現不顯示狀態。



圖（十二）95%信賴區間圖

從圖(十二)可以看出我們預測的結果還不錯，與實際值相差不多，並且有抓到實際值的趨勢，只可惜在 10 月 28 日時的預測較為不準。

五、MAE、MSE、MPE、MAPE 圖比較表

	arima			
國道一號北向	MSE	MAE	MPE(%)	MAPE(%)
	101.6270	9.7396	-8.2631	22.8210
	Time series regression			
	MSE	MAE	MPE(%)	MAPE(%)
	280.3498	14.4572	-47.0087	55.0735
	Exponential smoothing			
	MSE	MAE	MPE(%)	MAPE(%)
	203.9676	12.3909	-38.3710	46.9414
	UCM			
	MSE	MAE	MPE(%)	MAPE(%)
	129.0416	8.9774	-14.0392	29.9867

表（四）綜合比較圖

從表(四)綜合比較表可以看到每個數據集所做的四種模型的 MAE、MSE、MPE、MAPE 的大小，而最後我們可以發現國道一號北向是 ARIMA 模型表現得比較好，其他模型的預測相對來說誤差都稍嫌太大。

六、結論

綜觀各種分析結果，我們發現在節假日等特殊時期，所有模型的預測結果均存在較大偏差，這可能是由於節假日期間車流量和人員流動的巨大變化所導致的異常情況。

綜合考慮 MSE、MAE、MPE 及 MAPE 四種評估指標，我們發現 ARIMA 表現相對優異，而 UCM 則在 MAE 有不錯的表現，這反映了不同模型在預測準確性和穩健性方面的差異特點。此外，於 95%預測模型中可以發現 UCM 預測相對較佳，惟 10/28（週五）有明顯低估及 10/30（周二）有明顯高估。

七、參考文獻

[1] Bowerman, O’Connell and Koehler, Forecasting, Time series, and Regression, Thomson,4th,2005

[2] O.D. Adubisi , C. Ezenweke ,C.E. Adubisi .(2020) Unobserved component model with multi-interval input interventions: An application to crude oil production

[3] Anpalaki J. Ragavan, George C. Fernandez.(2006)Modeling Water Quality Trend in Long Term Time Series

[4] Cathy W. S. Chen , Ming Chieh Cheng and Songsak Sriboonchitta(2018) Predictive analytics of Taiwan inbound tourism from ASEAN 5