基於深度學習技術應用於空氣品質 PM2.5預測

林澔[a]、陳源安[a]、楊朝棟[a,b,*]、姜自強[c]
a.東海大學資訊工程學系、b.東海大學電子計算機中心、
c.東海大學資訊管理學系
mike840724@gmail.com,a961309@gmail.com,
ctyang@thu.edu.tw,steve312kimo@thu.edu.tw

摘要

由於氣候劇烈變化、生態環境的破壞以及工業、能源產業之需求逐年上升，空氣污染對於人民生活品質及健康的影響成為無法忽視的重大議題。預測空氣汙染將能使人民及行政單位在生活以及政令安排時有明確有力的量化依據。本研究中將使用深度學習技術、卷積神經網路、遞歸神經網路以及長短期記憶模型，利用序列化的空氣污染因子如 O3、SO2、CO2等等來訓練模型，以此對短期未來之空氣汙染程度做出預測。透過皮爾森相關係數(Pearson Correlation)和斯皮爾曼相關係數(Spearman Correlation)量化空氣汙染因子的相關性,找出相關性最高的空氣汙染因子。K Nearest Neighbor 方法應用於遺漏值的填補，並與線性、平均值等補值方法進行實驗與討論，尋找最能符合真實數據的補值方法。透過以上技術來提升深度學習模型的精準度以及泛化能力。實驗中將使用平均絕對誤差百分比(MAPE)值作為實驗調整之依據以及模型預測能力的量化標準。由於視覺化對於開發者及使用者幫助是非常顯著的，將使用 PHP 及 Matplotlib 視覺化實驗成果，將深度學習模型預測視覺化成果呈現於網頁中，協助開發者進行實驗以及方便使用者使用。

關鍵詞：空氣汙染、深度學習、卷積神經網路、遞歸神經網路、長短期記憶模型、K Nearest Neighbor

Abstract

Due to the dramatic changes in the climate, the destruction of the ecological environment, and the increasing demand for industrial and energy industries, the impact of air pollution on people's quality of life and health has become a major issue that cannot be ignored. Predicting air pollution will enable people and administrative units to have a clear and powerful basis for quantification of their lives and decrees. In this study, deep learning techniques, convolutional neural networks, recurrent neural networks, and long- and short-term memory models will be used to train models using serialized air pollution factors such as O3, SO2, CO2, etc., for the short-term future. The degree of air pollution is predicted. The Pearson correlation coefficient and the Spearman Correlation coefficient were used to quantify the correlation of air pollution factors to find the most relevant air pollution factor. The K Nearest Neighbor method is applied to the filling of missing values, and experiments and discussions with linear, average and other complement methods to find the best complement method that can match the real data. Through the above techniques to improve the accuracy and generalization ability of the deep learning model. The mean absolute error percentage (MAPE) value will be used as the basis for the experimental adjustment and the quantitative criteria for the model prediction ability. Since visualization is very helpful for developers and users, PHP and Matplotlib will be used to visualize experimental results, and the deep learning model predictive visualization results will be presented on the webpage to assist developers in experimenting and user-friendly.

Keywords: Air pollution, Deep learning, Convolutional neural networks, Recurrent neural networks, Long- short-term memory models, K Nearest Neighbor

1. 前言

近年來因為氣候變遷自然資源短缺和生活環境被破壞等原因,加上越來越多國家現代化等因素,使得各地空氣汙染日趨嚴重,同時也因為空氣汙染造就許多環境災害,影響國民的生存空間與生活品質。隨著環境保護意識的提升,人們對於有良好的空氣品質也越來越重視,許多的人意識到空氣汙染對於環境的破壞及人體健康的影響是很巨大的,因此,空氣汙染已經成為現今社會最重要的議題。

影響空氣污染的原因有很多,可能是工廠廢氣,亦或是汽機車排氣等等,在此我們選擇使用 PM2.5 作為觀察重點,以此來判斷空氣污染的影響,其中,影響 PM2.5的空氣汙染因子有很多,在此實驗中所收集的數據集中的空氣資訊有 TEMP、CO、NO、NO2、NOx、O3、PM10、PM2.5、SO2...等 等,此實驗將會在數據集中選擇適合的空氣因子來使用。

近年來機器學習慢慢崛起,ML 的基本定義是「在不經過程式導引的前提下,機器就具備學習的能力」,它是透過樣本數據對機器進行訓練,而不是使用特定的規則來編成模型,因此,在此實驗中將會運用機器學習與大數據的結合來完成,使用大數據之技術處理所收集到的數據集,在將其運用於機器學習中建置模型。

時間序列資料預測是一種具有挑戰性的預測模型研究,因為時間序列需要考慮到輸入變量之間的時間依賴性。遞歸神經網路(RNN)能夠處理時間序列的問題,因為他們可以通過使用自己的輸出作為下一步的輸入來保持上一次迭代到下次迭代的狀態,他的循環能力已經證明 RNN 是專門用於時間序列數據的強大引擎。然而,空氣汙染數據集是具有時間排序及連續數據的時間序列數據,因此 RNN 可用於空氣汙染數據集的時間序列建模及預測。但是,RNN 在訓練的過程中會產生梯度爆炸及梯度

消失的問題,為了解決此種問題的出現,此實驗採用長短期記憶模型(LSTM),他是一種 RNN 的模型,作為此實驗的主要工具。

在本研究中,我們的目標是使用長短期記憶模型於分析和自動化空氣污染預測。具體目標是:

1. 使用 Pearson Correlation 及 Spearman Correlation 對數據進行測量,以了解 PM2.5與變數之間的相關性。
2. 使用 KNN 及 Linear 的方式對數據及進行補值實驗,並經由實驗結果來選擇此數據集較為適用的補值方式。
3. 使用 Matplotlab 對預測結果進行可視化,以優化預測模型的參數。
4. 分析及比較 CNN 及 LSTM 的預測結果,並使用平均絕對百分誤差(MAPE)值測量預測精準度。

2. 文獻探討

2.1. Anaconda

Anaconda 是一個 Python 和 R 語言的免費開源平台,除了有眾多使用者及企業用戶外,目前也有超過1000種的數據包可供使用,它適用於 Windows,Linux 和 MacOS 等不同的作業系統環境下的 Conda 軟體包和虛擬環境管理器,對於在安裝、執行複雜的機器學習環境上變得更加簡單快速。

2.2. MySQL

MySQL 是一個開源的小型關聯式資料庫管理系統,目前 MySQL 被廣泛運用於許多小型網站中,由於其總體成本較低,並且擁有開放原始碼的特點,使其被許多為了降低成本的中小型網站選擇使用,然而,MySQL 也有許多不足的地方,最明顯的缺點是規模較小、功能有限等。

目前網路上流行的網站架設方式有 LAMP、WAMP 及 MAMP 等,其中 LAMP 便是此研究中我們使用的方式,其代表的是 Linux、Apache、MySQL 和 PHP 的縮寫,即是使用 Linux 作為作業系統,Apache 做為網站伺服器,MySQL 為資料庫,而 PHP 則是網站開發語言,這些資源都是開放原始碼的,因此架設起來十分方便、迅速,另外 WAMP 及 MAMP 則是分別是以 Windows 及 MacOS 作為作業系統的架設方式。

2.3. 皮爾森相關係數(Pearson correlation)&斯皮爾曼相關係數(Spearman correlation)

在做研究中,時常需要針對幾個不同的變數來做相關性的實驗,需要檢驗其與預測目標是否具有足夠的相關性,以及了解變量與目標之間的相關性為正向相關,亦或是反向相關,然而,最常被使用來呈現相關性的指標即為皮爾森相關係數(Pearson correlation)和斯皮爾曼相關係數(Spearman correlation),而這兩個指標的適用情形有所不同。

皮爾森相關係數時常用來呈現連續型變數之間的關聯性,尤其是在變數呈現常態分佈時,其結果最為精確,而斯皮爾曼相關係數則不需要符合常態分布,僅需要變數的資料型態為有序的即可,另外斯皮爾曼相關係數是以排序值(rank)來計算相關系數,因此其不會受到離群值(outliers)的影響。

2.4. TensorFlow

TensorFlow 是一個利用資料流圖(Data Flow Graphs)來表達數值運算的開放式原始碼函式庫,主要用於機器學習與深度神經網路方面研究,TensorFlow 可被用於語音識別或圖像識別等多項機器學習和深度學習領域,TensorFlow 它可以支持分散式運算,能夠在各個平台上自動運行,從單個 CPU 到多個 GPU 組成的系統都可以運行,另外,它也支援多種程式語言,python 及 C++都可以使用於 TensorFlow 上。

2.5. Keras

Keras 是一個完全由 Python 編寫,並且兼容 Theano 和 TensorFlow 的神經網路庫,使用其來組建神經網路將會更加簡單、快速,其支持 CNN 及 RNN 的使用,也能夠支持任意方案,包括多輸入和多輸出訓練,還能夠在 CPU 和 GPU 之間切換使用,而且其廣泛的兼容性能使 Keras 在 Windows、Linux 和 MacOS 上皆可正常使用。

2.6. 最近鄰算法(KNN)

最近鄰算法(K Nearest Neighbor, KNN),這個算法是機器學習中相對較為容易理解的分類算法之一。

KNN 原理是,將一個沒有標籤的新數據輸入進一個已經將每個數據皆存有標籤的樣本數據集中,將新數據的每個特徵與樣本集中的數據對應特徵進行比較,經過計算後,提取樣本集中特徵最為相似的數據分類標籤做為此新數據於樣本集中的標籤。

一般來說,KNN 只會選擇樣本集中前 K 個相似的數據進行比較分類,而不是使用樣本集中的全部數據,然而,鄰近性用距離度量的話,距離愈大,則代表兩個數據之間愈不相似。

2.7. 卷積神經網路(Convolution Neural Network)

卷積神經網路(Convolution Neural Network, CNN)一直以來都是 deep learning 中極為重要的一部份，其在圖片、影像辨識中的威力非常強大，甚至可以說是比人眼辨識還要厲害，許多影像辨識的模型也都是從 CNN 的架構延伸而形成的。

CNN 含有幾種不同的層，包括卷積層、池化層及全連接層等，卷積層的作用是對圖片做擷取特徵的動作，找出最好的特徵後，再進行分類，而在卷積層之間工常會加入池化層，它是用來壓縮圖片，並保留重要資訊的一種方式，最後，在模型最後面會加入全連接層，而全連接層的作用便是要用來實現分類，這就是 CNN 的主要架構組成。

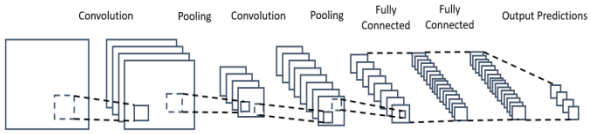

圖1 CNN 架構示意圖

2. 8. 長短期記憶網路(Long Short-Term Memory)

長短期記憶(Long Short-Term Memory, LSTM)是一種時間遞歸神經網路(RNN)，LSTM 解決了梯度消失(gradient vanishing)的問題，而能夠解決梯度消失問題的關鍵在於它多了 gate，它的功用是針對模型所獲得的信息流做過濾的用途。

在 LSTM 中，總共建構了 3 個 gate 來控制信息流的過濾，分別為:

輸入門 i(t)：決定當前有多少信息可以流入 memory cell c(t)。

遺忘門 f(t): 決定上一層的 memory cell c(t)中的信息可以累積多少到當前的 memory cell c(t)中。

輸出門 o(t)：當前時刻的 memory cell c(t)有多少可以流入當前隱藏狀態 h(t)中。

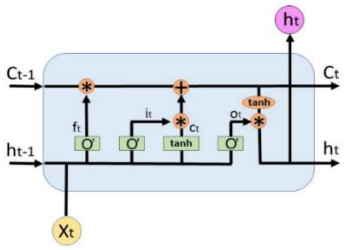

圖2LSTM 架構示意圖

$$
\begin{aligned}
i_t &= \sigma(W_i h_{t-1} + U_i x_t + b_i) \\
f_t &= \sigma(W_f h_{t-1} + U_f x_t + b_f) \\
o_t &= \sigma(W_o h_{t-1} + U_o x_t + b_o) \\
\tilde{c}_t &= tanh(W h_{t-1} + U x_t + b) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \\
h_t &= o_t \circ \tanh(c_t) \\
y_t &= h_t
\end{aligned}
\qquad (1)
$$

3. 系統設計與實作

在此研究中，我們將透過 Tensorflow 與 Keras 進行深度學習，建立空氣汙染預測模型，並對 PM2,5進行預測，以提供人民一個可以快速了解空氣汙染程度的平台。

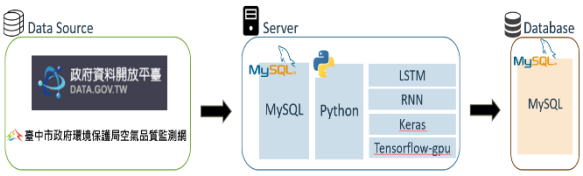

圖3 系統架構示意圖

3.1 資料蒐集

在此研究中，我們將使用 Python 自動從政府開放資料平台進行抓取，並將其儲存至 MySQL 資料庫中，以方便後續研究使用。

圖4 政府開放資料平台

SiteName	County	AQI	Pollutant	Status	SO2	CO	CO_8hr	O3	O3_8hr	PM10	PM25
三義	苗栗縣	143	細懸浮微粒	對敏感族群不健康	3.3	0.45	0.5	64	39	78	57
三義	苗栗縣	137	細懸浮微粒	對敏感族群不健康	3.9	0.52	0.5	68	34	91	62
三義	苗栗縣	128	細懸浮微粒	對敏感族群不健康	3.4	0.49	0.5	53	26	83	56
三義	苗栗縣	122	細懸浮微粒	對敏感族群不健康	4.1	0.45	0.5	35	22	81	51
三義	苗栗縣	118	細懸浮微粒	對敏感族群不健康	4.1	0.41	0.5	36	19	63	46

圖5 Mysql 資料庫

3.2 資料修補

在原始資料中，可能會因為某些無法避免的因素導致數據集中有些許的缺失值，這些缺失值會導致機器在學習的過程中有誤差產生。

然而，為了盡量避免這種情況發生，在此研究中我們分別使用 KNN 及 Linear 的方式，對數據進行缺失值補值，並比較此二種補值方式的誤差。

3.3 資料切割

將從政府開放資料平台所收集到的空氣品質數據集切割成訓練集(training data)、測試集(testing data)及驗證集(validation data)，本研究將49%數據做為訓練集，21%數據做為測試集，最後的30%做為驗證集，在訓練完成後，將未參與訓練過程的測試集放入模型中進行預測，並比較真實的 PM2.5 與預測結果之誤差，以此來當作此預測模型精準度評估的參考之一。

預測的時間單位則是選擇使用空氣品質數據集中欲預測時間的前300個小時的多項空氣汙染因子當作此模型的輸入資料，其中包括 SO2、CO、PM10等等，輸出的預測結果則是預測 PM2.5未來8個小時的濃度數據。

3.4 相關性

在此研究中，我們將各變數與 PM2.5去做 Pearson correlation 及 Spearman correlation 的實驗，並且將部分幾個表現出較好關聯性的數據加入預測模型中使用，以期望會得到較好的預測結果。

圖6為部分幾個空氣汙染因子與 PM2.5之間的相關性點狀分布圖，我們可以看出 PM2.5與 SO2、CO 及 PM10的相關性表現是較好的，因此可以判端出他們之間的相關性較高，然而，點狀圖只是作為初步的參考與判斷，在實驗結果的部分會有更多的汙染因子相關性比較及更詳細的數據。

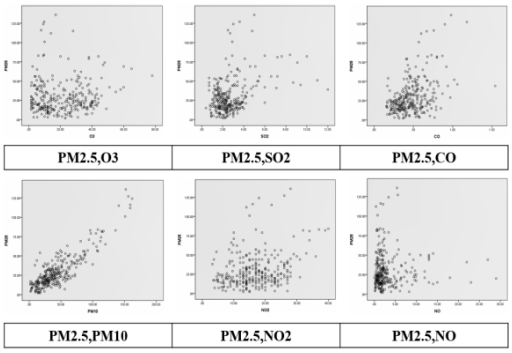

圖6 變數相關性點狀分布圖

3.5 Lagtime

使用 Pearson 及 Spearman correlation 對各個時段的變數與 PM2.5進行實驗，找出各時滯中與PM2.5相關性表現最好的時段，並加入模型中使用。

表1 不同時滯的變數相關性(Pearson correlation)

	O3	SO2	CO
0 hr	0.059	0.389	0.435
1 hr	0.097	0.371	0.379
2 hr	0.132	0.359	0.315
3 hr	0.149	0.355	0.255
4 hr	0.140	0.345	0.207
5 hr	0.121	0.315	0.167
6 hr	0.094	0.284	0.146
7 hr	0.065	0.257	0.144
8 hr	0.041	0.236	0.132
	PM10	NO2	NO
0 hr	0.869	0.279	0.020
1 hr	0.842	0.255	0.006
2 hr	0.792	0.220	-0.017
3 hr	0.737	0.193	-0.051
4 hr	0.686	0.184	-0.083
5 hr	0.642	0.186	-0.111
6 hr	0.605	0.186	-0.125
7 hr	0.571	0.190	-0.120
8 hr	0.547	0.181	-0.122

表2 不同時滯的變數相關性(Spearman correlation)

	O3	SO2	CO
0 hr	0.063	0.238	0.364
1 hr	0.109	0.219	0.303
2 hr	0.148	0.212	0.235
3 hr	0.164	0.193	0.179
4 hr	0.158	0.172	0.140
5 hr	0.144	0.150	0.102
6 hr	0.135	0.135	0.083
7 hr	0.130	0.123	0.080
8 hr	0.130	0.120	0.075
	PM10	NO2	NO
0 hr	0.750	0.229	0.068
1 hr	0.748	0.181	0.031
2 hr	0.726	0.134	-0.026
3 hr	0.711	0.094	-0.086
4 hr	0.696	0.078	-0.130
5 hr	0.679	0.066	-0.151
6 hr	0.658	0.059	-0.163
7 hr	0.626	0.058	-0.168
8 hr	0.600	0.042	-0.185

3.6 建立及訓練預測模型

在將此研究中所需要的數據及進行處理完成後，便可以開始進行模型的建置，首先需要設定此模型所需要的神經網路層類別與該神經網路層的層數，以及每層所擁有的神經元個數。

將模型的架構堆疊完成後，則是需要針對超參數進行選擇，而每個超參數代表著不同的意義，Epoch 代表在此訓練過程中，數據被使用了多少次，Keras 中參數更新是按批次進行的，就是小批度下降算法，把數據分為若干組，稱為 Batch，而 Batch_size 則是每組數據的樣本數量。此研究的超參數選擇:

另外，LSTM 只解決了梯度消失問題，並沒有解決梯度爆炸的問題，所以我們必須在編寫架構時進行測試，並選擇適用於此數據及模型的激活函數，以避免發生梯度爆炸而沒有學習效果的情況。

3.7 模型評估

在完成神經網路模組訓練後，進行模型評估以了解此模型學習的成效，藉此作為實驗進行的依據。

在此研究中，我們使用平均絕對誤差 MAE(mean absolute error)作為訓練模型時的 loss_function，並使用平均絕對百分比誤差 MAPE(mean absolute percentage error)做為模型評估指標，利用 MAPE 評估此模組的預測效果。以下第(2)式及第(3)式分別為 MAE 及 MAPE 的公式:

$$MAE = \frac{\sum_{t=1}^{n} |Y_t - \bar{Y}_t|}{n} \qquad (2)$$

$$MAPE = \sum_{t=1}^{n} \frac{\frac{|Y_t - \bar{Y}_t|}{Y_t} * 100\%}{n} \quad (3)$$

4. 實驗結果

4.1. 輸入變量相關性測量

輸入變量的選擇將會影響到此空氣汙染預模型的準確度，選擇相關性較低的數據輸入進模型中進行預測，將會使機器難以掌握數據間相互的關係，因而倒是此模型的學習狀況不佳，更會造成此模型的預測結果的不精確，

因此，為了盡量避免造成此種問題，在此實驗中，會先對空氣汙染因子的選擇進行皮爾森與斯皮爾曼相關性測量，利用兩種不同的方法相互比較，並從中挑選幾個汙染因子作為此預測模型的輸入，以期望得到較好的預測結果。

表3 輸入變量相關性

	Pearson correlation	Spearman correlation
ρ(PM2.5,SO2)	0.369	0.337
ρ(PM2.5,CO)	0.453	0.449
ρ(PM2.5,O3)	0.059	0.063
ρ(PM2.5,PM10)	0.765	0.673
ρ(PM2.5,NOx)	0.299	0.320
ρ(PM2.5,NO)	0.174	0.186
ρ(PM2.5,NO2)	0.349	0.338
ρ(PM2.5,THC)	0.379	0.380
ρ(PM2.5,NMHC)	0.379	0.373
ρ(PM2.5,WS)	-0.200	-0.171
ρ(PM2.5,WD)	0.132	0.096
ρ(PM2.5,TEMP)	0.197	0.174
ρ(PM2.5,RAINFALL)	-0.143	-0.221
ρ(PM2.5,CH4)	0.324	0.314
ρ(PM2.5,UVB)	0.067	0.082
ρ(PM2.5,RH)	0.056	-0.006
ρ(PM2.5,WS_HR)	-0.216	-0.179
ρ(PM2.5,WD_HR)	0.150	0.091

4.2. 資料缺失值補值

為了找到更適合此研究中的實驗數據的補值方式，我們實驗了 KNN 及線性(Linear)的方式，在實驗中，我們看出 KNN 在短時間的缺失值的誤差值表現較大，而線性補值的缺失值的誤差值表現則是在長時間缺失值較大。

因此，在實驗中我們得知 KNN 適用於長時間的缺失值補值，而線性補值則是適用於短時間的缺失值補值。

表4 不同補值方式誤差值表現

	KNN	Linear
隨機50小時	4.7	3.3
連續6小時	3.3	3
連續12小時	2.5	4.2
連續24小時	2.8	3.3

4.3. 模型訓練及評估

在此研究的實驗中，我們分別使用 CNN 及 LSTM 神經網路建置深度學習預測模型，分別加入了多個不同的變量進入模型中進行訓練，並針對不同的預測時間建置模型及預測。

以下**圖7**為使用 CNN 訓練預測模型的 loss 值變化的視覺化。

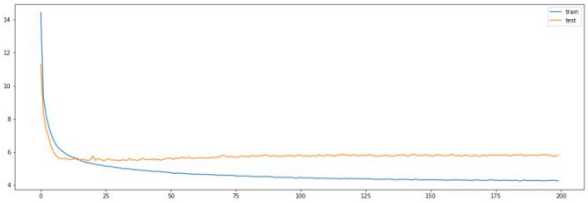

圖7 CNN 訓練 loss 值變化視覺化

以下**表5**為使用 CNN 建置預測模型的預測結果評估，而**圖7**為預測結果評估視覺化。

表5 CNN 每小時預測結果評估

Time	MAPE	Time	MAPE
1 hr	0.23	5 hr	0.50
2 hr	0.31	6 hr	0.56
3 hr	0.40	7 hr	0.58
4 hr	0.43	8 hr	0.60

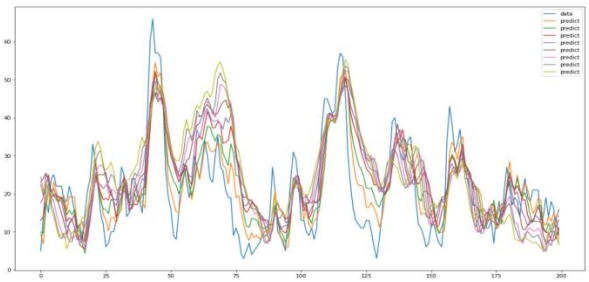

圖8 CNN 預測結果視覺化

以下**圖9**為使用 LSTM 訓練預測模型的 loss 值變化的視覺化。

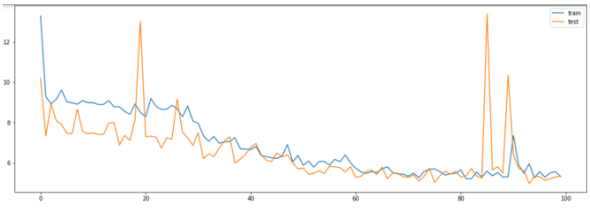

圖9 LSTM 訓練 loss 值變化視覺化

以下**表6**為使用 LSTM 建置預測模型的預測結果評估，而**圖10**為預測結果評估視覺化。

表6 LSTM 每小時預測結果評估

Time	MAPE	Time	MAPE
1 hr	0.21	5 hr	0.40
2 hr	0.29	6 hr	0.43
3 hr	0.33	7 hr	0.46
4 hr	0.37	8 hr	0.48

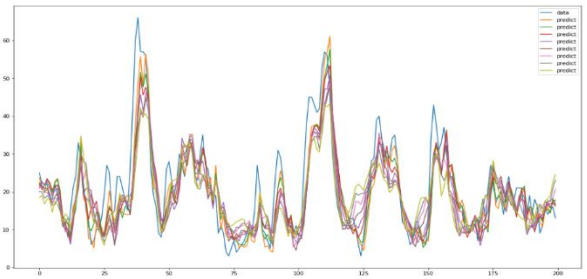

圖10LSTM 預測結果視覺化

以下**表7**為 CNN 及 LSTM 預測未來8個小時的 PM2.5濃度的預測結果比較，其中包括每個小時的真實值、預測值，及誤差值等等，從下**表7**我們可以看出 LSTM 預測結果的平均誤差值是比 CNN 預測結果的平均誤差值小，因此，我們可以認為 LSTM 比 CNN 更適合使用於具有時序性數據集的深度學習預測模型的建置。

表7 CNN 與 LSTM 預測結果比較

	Real	CNN	誤差	LSTM	誤差
1 hr	25	23	2	23	2
2 hr	22	20	2	22	0
3 hr	22	19	3	22	0
4 hr	22	20	2	21	1
5 hr	18	21	3	21	3
6 hr	18	18	0	21	3
7 hr	22	15	7	20	2
8 hr	19	14	5	18	1
平均誤差		CNN	3/hr	LSTM	1.5/hr

5. 結論與展望

此研究使用 Pearson Correlation 及 Spearman correlation 兩種不同的關聯性計算方式，能夠提供明確量化標準，對於空氣汙染因子的選擇可以更加的快速。並且結合相關性計算方法與 Lag time 機制，能夠找出汙染因子之間具有高度相關性的時間點，提供更加精確的數據以利模型訓練。

另外，透過線性及 KNN 兩種補值技術的使用，理解並比較此兩種補值方式是否適用於此研究中的空氣汙染數據集，選擇使用較適合的補值方式，能夠提供更加接近真實情況的數據進行訓練，以提供模型預測效果。

在此研究的實驗數據處理的部分，我們分別數據級切割為 training data、testing data，以及 validation data，在模型訓練中，可以使用 training data 和 validation data 來確認該模型的學習情況，而在模型訓練後，可以使用 testing data 進行模型評估及測試，如此便能夠更加了解此預測模型的預測效果，以方便後續修改及優化。

在此研究中，選擇使用 CNN 及 LSTM 來建置深度學習模型，使用此兩種神經網路來處理時間序列的空氣汙染數據，有效的預測短期的未來空氣品質狀況，並使用 MAPE 作為模型評估的方式，比較兩種不同的神經網路預測模型的預測精準度及誤差值。

未來本研究將會選擇結合 CNN 及 LSTM 兩種神經網路，使用 C-LSTM 深度學習模型進行建置，並且另外使用其他迴歸分析的方法進行預測，再將幾種不同的預測方法的預測效果進行比較。也會在空氣汙染因子的選擇上繼續找尋更加適合的變量加入模型中訓練，以期望能夠提升預測模型的精準度。在完善此預測模型後，期望可以建立空氣汙染預測平台，並將數據視覺化於此平台上，提供給有需求的使用者參考。另外，未來也會針對沒有測站的地區，使用擴散的方式進行空氣汙染品質的預測，讓沒有測站的地區也能夠藉由平台理解所在地區的空氣污染狀況。

6. 致謝

本研究論文資料感謝行政院環境保護署空氣品質指標開放資料平台提供之開放資料，感謝中華民國科技部計畫經費支持，感謝科技部研究計畫: 台灣空氣品質大數據監測平台、機器學習與政策模擬之跨領域研究--台灣空氣品質大數據監測平台、機器學習與政策模擬之跨領域研究，計畫編號為 MOST 106-3114-M-029-001-A、MOST 108-2119-M029-001-A。

參考文獻

[1] Congcong Wen, Shufu Liu, Xiaojing Yao, Ling Peng, Xiang Li, Yuan Hu,and Tianhe Chi, “A novel spatiotemporal convolutional long short-term neural network for air pollution prediction”, Science of The Total Environment, vol 654, pp. 1091-1099, March 2019.

[2] Chic-Fong Tsai, Fu-Yu Chang,” Combining instance selection for better missing value imputation”, Journal of Systems and Software, vol 122, pp. 63-71, December 2016.

[3] Yagmur Gizem Cinar , Hamid Mirisaee , Parantapa Goswami, Eric Gaussier, Ali Aït-Bachir, ” Period-aware content attention RNNs for time series forecasting with missing values”, Neurocomputing, vol 312, pp. 177-186, October 2018.

[4] Haydar Demirhan, Zoe Renwick,”Missing value imputation for short to mid-term horizontal solar irradiance data”, Applied Energy, vol 225, pp. 998-1012, September 2018.

[5] Yanlai Zhou, Fi-John Chang, Li-Chiu Chang, I-Feng Kao, Yi-ShinWang, “Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts”, Journal of Cleaner Production, vol 209, pp. 134-145, February 2019.

[6] Lau, J.T., Griffiths, S., Choi, K.-c., and Lin, C., “Prevalence of preventive behaviors and associated fac-tors during early phase of the H1N1 influenza epidem-ic”, American Journal of Infection Control, vol. 38, pp. 374-380, 2010