等距特徵映射演算法於台灣氣候分群之研究

蔡明衡、曾建翰
中央氣象局氣象科技研究中心

摘　要

本研究利用等距特徵映射(Isometric Feature Mapping, ISOMAP)為基底的機器學習模型進行台灣測站的氣候分群，使用台灣地區測站，包含局屬站和自動站的溫度資料。同時引入深度學習特徵提取及多層類神經網路的概念，將一種模式中整合了降尺度及資料後處理等技術。

在資料處理方面，利用多層類神經網路的概念，提取測站的特徵，再以此來進行模式辨識與分群，在提取特徵的過程中會使用ISOMAP與主成分分析(Principal Component Analysis, PCA)兩種方法。

結果顯示，採用ISOMAP和PCA的兩種不同降維方法下，能找到相似的特徵，且兩者都確實能將具有相似氣候變化的測站辨識為同一種類別，但是ISOMAP方法能更有效地將兩個相差較大的數據分離，更有利於分群。透過初步的研究結果可知，以此方法建立的分類模型，對於氣候分群和模式辨識具有一定的掌握能力。

關鍵字：機器學習、氣候分群、等距特徵映射

一、前言

近年來隨著科技的進步，人工智慧的應用有越來越多的突破，而辨識與分群方面更是已經進展到商用階段。在氣象研究上，辨識與分群更是眾人研究的焦點，包含聖嬰現象的模式辨識等等，也依舊是熱門的研究主題。

台灣是一個多山的海島國家，日夜溫差大。在包含氣候變遷，農業技術甚至包含健康相關等問題時，日溫差時常成為討論的重點[6,7]。因此，這項研究會將辨識的重點放在日溫差上面。

為了達成辨識的目的，特徵的提取扮演至關重要的角色。這項研究使用了早期針對圖像的研究方法[3]，先從數據學習特徵，再加以分類，進而達到辨識的機器學習方法。在學習特徵的過程會使用線性降維的PCA方法和非線性降維的ISOMAP方法，目的在於分析兩個不同的特徵學習方法在處裡氣象資料上有何異同。

二、研究方法

(一) 資料蒐集

測站資料使用中央氣象局從2011至2020年全島六百餘個測站之溫度數據，包含局屬站和自動站。其中因為測站建立時間各有不同，有部分測站缺少較早年分的資料。

(二) ISOMAP方法介紹

ISOMAP是一種非線性降維方法，此種方法相比PCA更適合處理具有流形性質的數據[2,4]。

先定義一個圖，每個頂點都是一個數據集，成對的點之間存在邊。當找到這個圖中最短的路徑距離，即形成一個鄰域。也可以說是找到這個圖中任兩頂點的測地距離。

藉由ISOMAP方法，可以將數據降維。藉此來進行特徵提取。

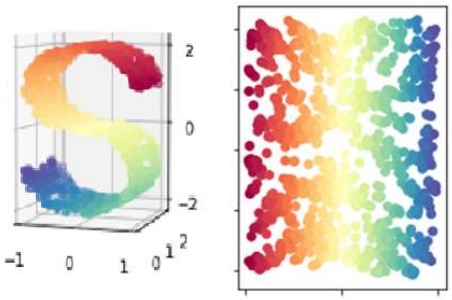

圖 1 ISOMAP 鄰域示意圖

(三) 測站資料處理

測站資料有部分缺損，有些測站是近年才設立的就缺少較早年份的資料。如果是缺少小時資料，則以內插方法填補資料，缺少太多資料則會不採用這個測站。

(四) 測站資料分析

1. 提取每月日變化特徵

輸入資料維度為(n*24,n=當月日數)，降維後得到(2*24)，這表示這個月每日24小時變化的特徵。

PCA方法則取用PC1和PC2。這個過程濃縮了每個月中，日與日之間的變化。

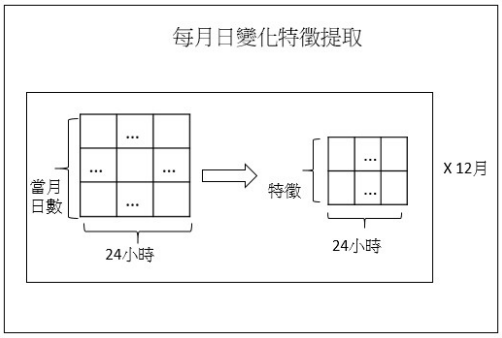

圖 2 每月日變化特徵提取

2. 提取測站日變化特徵

輸入資料為每個月提取第一個特徵排列成維度(12*24)的矩陣，降維後得到(2*24)的矩陣，這表示每年每日24小時變化的特徵。這個過程濃縮了每年月與月之間的變化。

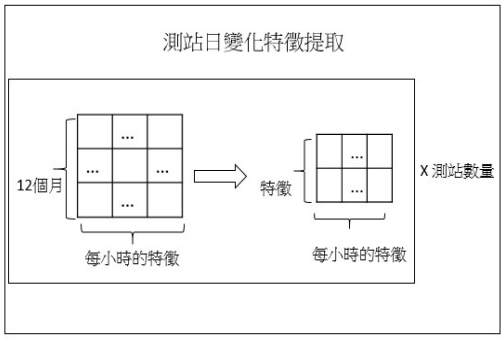

圖 3 測站日變化特徵提取

3. 提取測站特徵

每個測站都進行前兩個步驟，取用第一個特徵排列成(n*24,n=可用測站數量)的矩陣，降維後得到(n*3,n=可用測站數量) 的矩陣。這個過程濃縮了每個測站的小時特徵

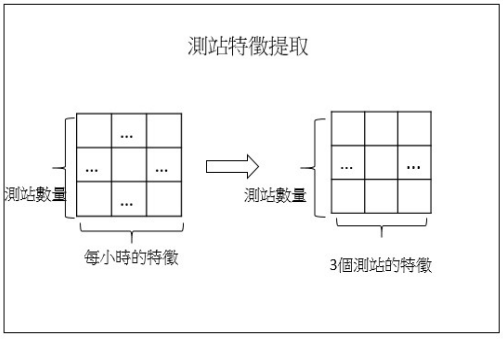

圖 4 測站特徵提取

(五) 氣候分群

1. 非監督式分類－K-means分類

利用降維所得到的每個測站的特徵1和特徵2以及經度、緯度和高度以K-means方法進行分類，分成5，7，13類

2. 監督式分類－貝氏分類

對每年的資料進行同樣的步驟，以貝氏分類方法學習每個測站經度、緯度和高度的分類結果，再依據2公里網格的經度、緯度和高度預測每個格點的分類結果。

三、研究結果

(一) 特徵提取

此項研究分別以PCA和ISOMAP方法進行特徵提取，以2020年溫度資料為例，2020年測站溫度資料以ISOMAP方法的特徵提取結果如圖5，在特徵一中突顯出有較高值的南部內陸地區和較低值的離島地區；特徵二中突顯出有較低值的南部沿海地區和較高值的山區地帶。

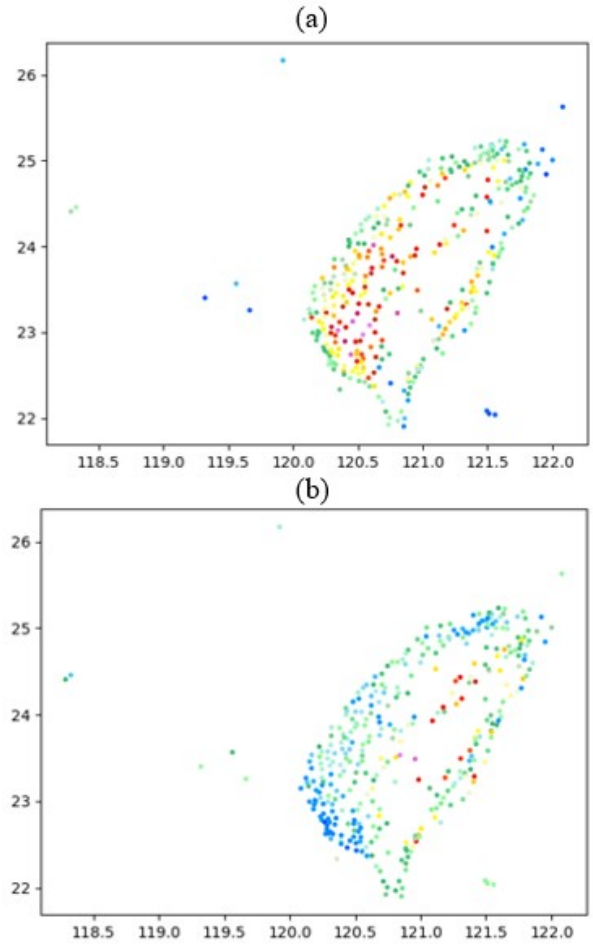

圖 5 2020測站溫度資料ISOMAP降維結果(a)特徵一(b)特徵二

以PCA方法的特徵提取結果如圖6，在特徵一中突顯出有較高值的南部沿海地區；特徵二中突顯

出有較低值的北部沿海地區和較高值的山區地帶。帶。

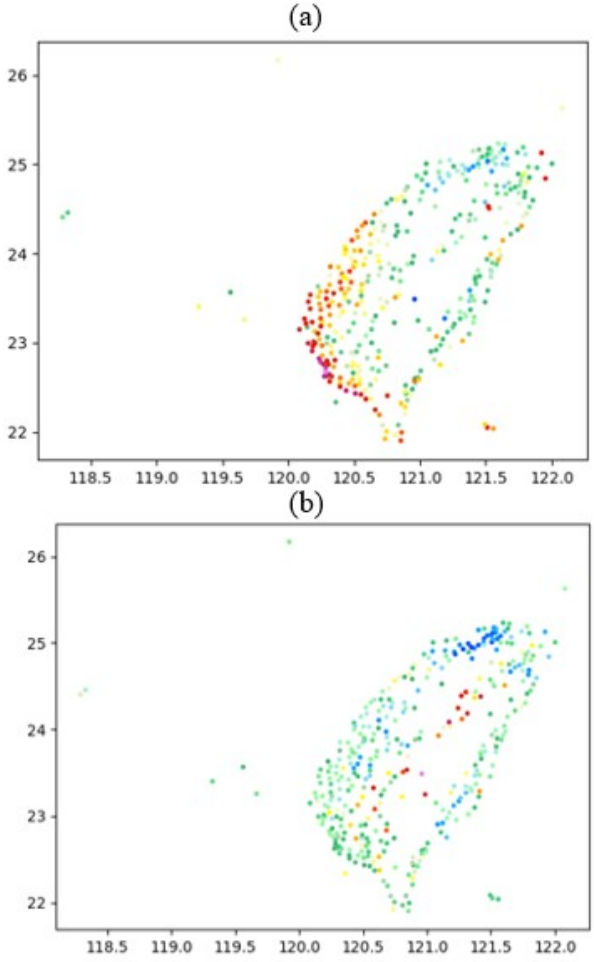

圖 6 2020測站溫度資料PCA降維結果(a)特徵一(b)特徵二

2020年測站資料以ISOMAP方法所建立的測站空間如圖7，圖中的子圖代表測站的年平均日溫度變化。經由ISOMAP方法提取的第一個特徵(X軸)可以看出日溫度變化大小，由左至右呈現由小到大的分布；而第二個特徵(Y軸)可以看出平均溫度的變化大小，由上至下呈現由低到高的分布。

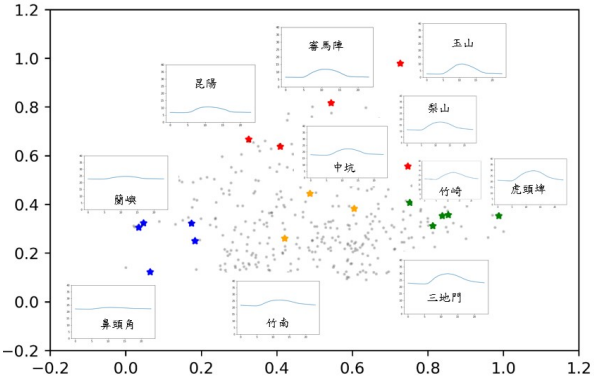

圖 7 2020年測站資料以ISOMAP方法所建立的測站空間

2020年測站資料以PCA方法所建立的測站空間如圖8。特徵一(X軸)可以看出日溫度變化大小，由右至左呈現由小到大的分布，特徵二(Y軸)可以看出平均溫度的大小變化，由上至下呈現由低到高的分布。

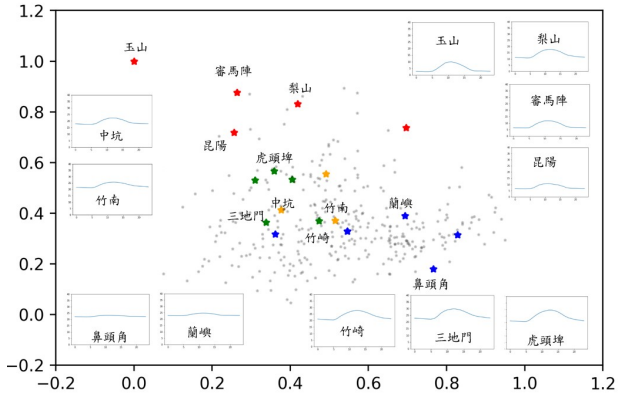

圖 8 2020年測站資料以PCA方法所建立的測站空間

兩者相比較下，ISOMAP與PCA方法皆找到相似的特徵，但是可以明顯看出PCA方法的測站空間中，測站分布的較為密集。竹南和竹崎兩個測站具有明顯的日溫度變化差異，在ISOMAP測站空間中被明顯分開而在PCA測站空間中則否。因此，在溫度資料上，ISOMAP方法相較於PCA方法能更好的將數據分散。

(二) 氣候分群

將所有的測站的特徵結合經度、緯度以及高度以K-means進行分類結果如圖9，撇除山區以外，大致與行政區域分布相似。

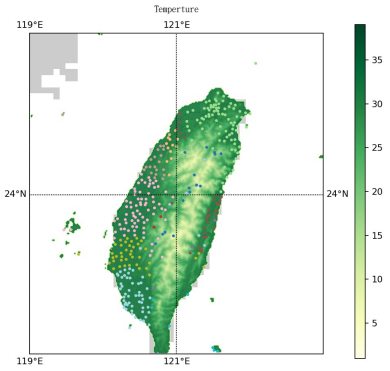

圖 9 2011-2020 ISOMAP測站溫度資料分群結果

表1 測站分類結果與各分類所含測站之RMSE

測站	分類結果	分類01	分類02	分類03	分類04	分類05	分類06	分類07	分類08	分類09	分類10	分類11	分類12	分類13
彰佳嶼	4	9.689	4.941	1.707	1.618	1.787	4.154	1.447	*0.960*	1.561	1.259	1.809	2.522	2.800
臺北	4	10.987	6.206	2.387	2.176	2.689	5.412	2.303	*1.179*	1.283	1.960	1.436	2.534	2.068
竹子湖	4	6.132	1.895	3.135	3.266	3.061	*1.431*	3.108	4.099	4.747	3.710	5.123	5.042	6.281
新竹	7	10.556	5.802	1.904	1.939	2.195	5.040	2.015	*0.810*	1.252	1.687	1.519	2.640	2.347
臺中	9	11.143	6.368	2.701	2.630	3.204	5.520	2.513	1.479	0.915	1.769	*0.775*	2.046	1.189
臺南	13	11.935	7.156	3.469	3.355	3.950	6.291	3.259	2.229	1.595	2.551	1.229	2.183	*0.736*
花蓮	7	11.027	6.244	2.712	2.480	3.174	5.374	2.312	1.563	1.203	1.817	*1.083*	1.871	1.437
阿里山	1	*3.642*	6.996	10.822	11.005	10.863	7.814	10.934	12.051	12.523	11.434	12.882	12.518	13.950
玉山	1	*8.485*	13.281	17.134	17.294	17.067	14.147	17.254	18.367	18.870	17.786	19.234	18.887	20.318
恆春	13	12.633	7.890	4.456	4.283	5.020	6.962	4.079	3.303	2.635	3.418	2.176	2.238	*1.672*
澎湖	9	10.829	6.046	2.422	2.256	2.819	5.206	2.167	1.186	*0.832*	1.540	0.896	2.103	1.590
馬祖	5	7.437	3.335	3.105	3.173	*2.441*	3.073	3.143	3.602	4.413	3.652	4.809	5.098	5.971

表1中數字表示該測站與該分類所包含之各測站的年平均日溫的RMSE平均值，藍色粗斜體表示最低RMSE所在分類。呈現出在山區，離島，以及南部地區有最好的分類結果。與非同分類之測站的RMSE遠大於與同分類之測站的RMSE。最差的結果則是北部地區。

若只針對測站資料分群則難以將結果推展到沒有測站的地方，因此進行貝氏分類來預測沒有測站的位置其分群結果。

ISOMAP將溫度分13群結果如圖10、圖11和圖12，顏色表示平均溫度的高低排名，數字越大代表平均溫度越大。涵蓋的年份越多分類的結果也越穩定，範圍變動較小。大致上可分為：澎湖、雲嘉南、高屏、竹苗、山地、鄰近山地以及夾雜在地型交界中的零碎地區。

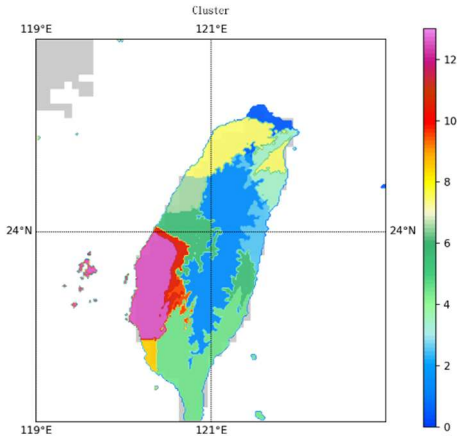

圖 10 2019-2020 ISOMAP測站溫度資料分群結果

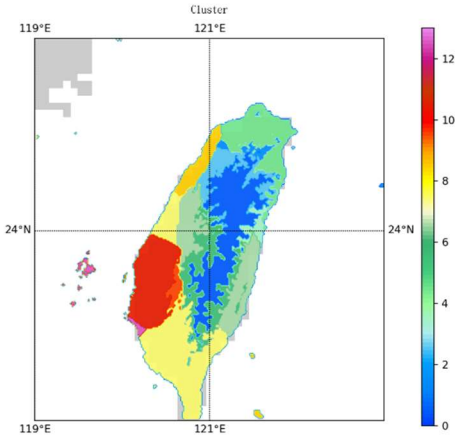

圖 11 2015-2020 ISOMAP測站溫度資料分群結果

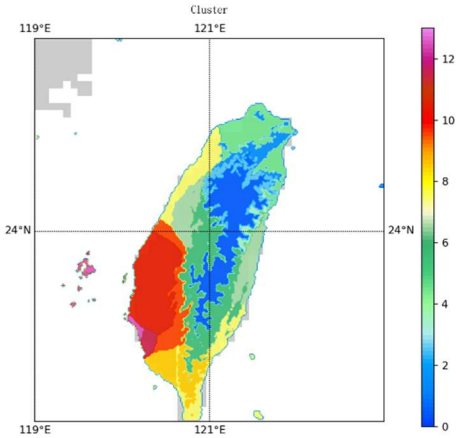

圖 12 2011-2020 ISOMAP測站溫度資料分群結果

PCA將溫度分13群結果如圖13、圖14和圖15，分類的範圍變動較大且有較多的破碎和不連續的情形，尤其是在資料較少的情況，分類結果沒有ISOMAP穩定。

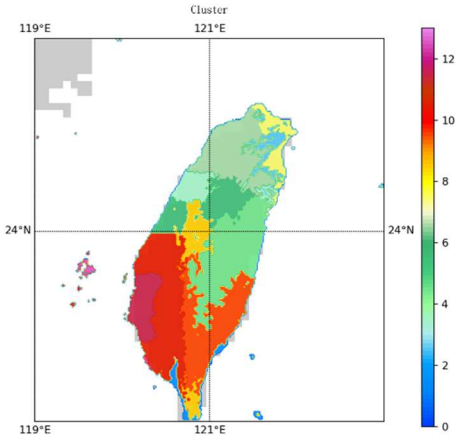

圖 13 2019-2020 PCA測站溫度資料分群結果

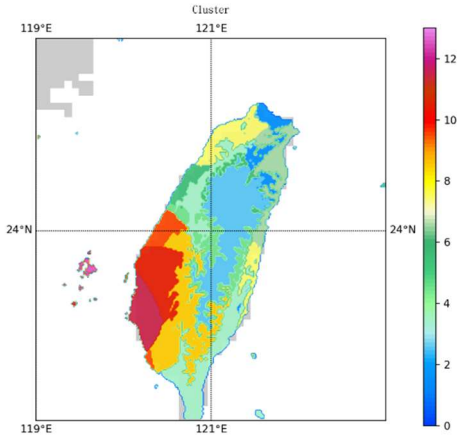

圖 14 2015-2020 PCA測站溫度資料分群結果

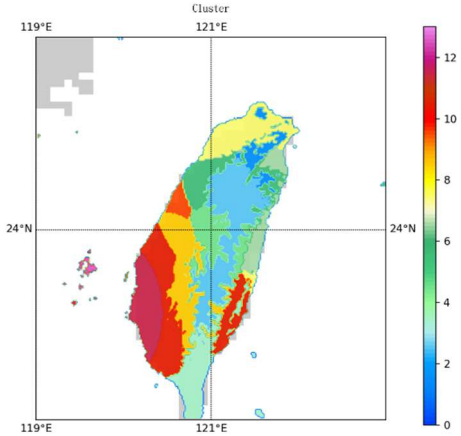

圖 15 2011-2020 PCA測站溫度資料分群

不論是ISOMAP還是PCA，在山區，南部和離島地區有最穩定的分群結果。跟測站空間中所看到的結果相符合，山地因為較低溫被分離；離島因為溫差較小被分離而南部則是因為高溫。

同時，北部是分類穩定度最差的地區。在測站空間中，不論是平均溫度或是日溫度變化上，皆沒有明顯的特徵。

PCA和ISOMAP最大的差異在於PCA方法中呈現較明顯的不連續、破碎等情形。因為PCA方法相較於ISOMAP方法較不易將數據分散，使分類的穩定度也跟著下降。

ISOMAP將溫度分5群結果如圖16。缺少了山地地區、依舊有明顯的雲嘉南分群。

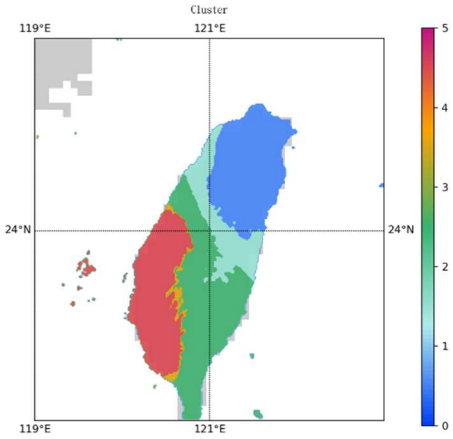

圖 16 2011-2020 ISOMAP測站溫度資料分群

ISOMAP將溫度分7類結果如圖17。可看出中南部沿海、中南部、山地、西北、東北等分群。

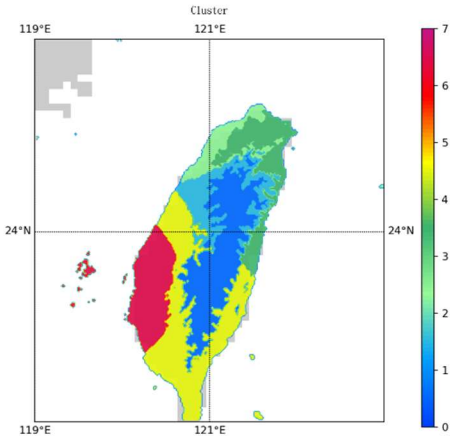

圖 17 2011-2020 ISOMAP測站溫度資料分群

在分類的數量差異上，分7類與13類的結果大致相同，在測站空間上看到的日溫變化，均溫差異，皆有呈現出來。在5類時分類結果相比之下較差，只有南部地區有著與其他分類較接近的結果。

四、結論與展望

在溫度的特徵學習中ISOMAP方法能分離出較多位處極端的族群，更有助於分群。在分群中兩種分法在南部地區都能看到顯著的分群。而同一種分群代表具有相似的氣候狀況，以台南、高雄和澎湖地區為例，三者溫度的日夜變化都較小且均溫較高因此被分為同一類。

氣候分群的困難點在於很難有一個客觀的標準來判斷，因此，分群的結果會因為目的不同而不

同，這項研究考量日夜變化對農業的影響極大，甚至也會影響人體的健康，以日夜變化為最主要的分群特徵。

綜合以上測試和分析，可以明顯看到ISOMAP方法在分離數據上與PCA方法相比具有明顯的優勢，在分析極端氣候時，也能有極大的作用，在比較事件的相似程度時，會有比PCA更好的結果。

目前在氣候的分析上有許多降維過程使用PCA方法，此項研究證明非線性降維在處裡測站資料上能更好的分散數據，有助於辨識事件。同時在預測聖嬰現象時也能有不錯的結果[20]，因此在做分類或是預測氣候現象時，非線性降維會是一個值得參考的選項。

五、參考文獻

[1] Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, *40*, 100378

[2] Gámez, A. J., Zhou, C. S., Timmermann, A., & Kurths, J. (2004). Nonlinear dimensionality reduction in climate data. *Nonlinear Processes in Geophysics*, *11*(3), 393-398

[3] Pless, R. (2003, October). Image Spaces and Video Trajectories: Using Isomap to Explore Video Sequences. In ICCV (Vol. 3, No. 2, pp. 1433-1440).

[4] Shi, J., & Luo, Z. (2010). Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. *Computers in biology and medicine*, *40*(8), 723-732.

[5] Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, *10*(66-71), 13

[6] 談珮華, & 曾雅琪. (2014). 臺灣局部地區溫度分布的影響因子. *地理學報*, (73), 1-27.

[7] 盧虎生, & 劉韻華. (2006). 臺灣優質水稻栽培之環境挑戰與因應措施. *作物, 環境與生物資訊*, *3*(4), 297-306.