國立臺灣大學工學院土木工程學研究所

碩士論文

Department of Civil Engineering

College of Engineering

National Taiwan University

Master's Thesis

基於深度學習之臺北市緊急醫療服務需求之時空網格化機率預測：以 DeepAR 與 TFT 為例

Probabilistic Grid-Based EMS Demand Forecasting in Taipei: DeepAR and TFT

陳羿璇

Yi-Syuan Chen

指導教授：陳俊杉博士

Advisor: Chuin-Shan Chen, Ph.D.

中華民國 115 年 3 月

2026, March

doi:10.6342/NTU202600877

國立臺灣大學碩士學位論文
口試委員會審定書

NATIONAL TAIWAN UNIVERSITY
MASTER'S THESIS ACCEPTANCE CERTIFICATE

基於深度學習之臺北市緊急醫療服務需求之時空網格化機率預測：以 DeepAR 與 TFT 為例

Probabilistic Grid-Based EMS Demand Forecasting in Taipei: DeepAR and TFT

本論文係 **陳羿璇** (R12521607) 在國立臺灣大學土木工程學系電腦輔助工程組完成之碩士學位論文，於民國115年03月04日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Civil Engineering Computer-Aided Engineering on March 4, 2026 have examined a Master's Thesis entitled above presented by Chen, Yi-Syuan R12521607 candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

陳俊杉
（指導教授 Advisor）

汪立本

林祺皓

系主管Director： 游景雲

doi:10.6342/NTU202600877

誌謝

感謝陳俊杉老師與林祺皓博士多年以來在學術研究上的教導與支持，使得本研究能夠順利完成並在過程中持續精進；也非常感謝汪立本老師在碩士班期間的照顧與鼓勵。特別感謝陳翊翔學長在碩士就讀期間的照顧與支持，也感謝地理所陳妍儒學妹協助資料特徵處理與產製、城鄉所劉祐任學長在論文寫作上給予很大的協助和精神支持。

doi:10.6342/NTU202600877

摘要

緊急醫療服務（Emergency Medical Services, EMS）需求具高度隨機性與時空異質性，若能準確掌握需求於不同時間與空間單元的變化，將有助於救護資源之前置部署與動態調度。本研究以臺北市為研究區，建構「時空網格化 × 機率式多步預測」之 EMS 需求預測框架，目的在於高解析度情境下同時提供點預測與不確定性資訊，以支援風險導向之資源規劃。

本研究以規則網格（1,000m × 1,000m）為空間單元，並以 4 小時為一時段進行時間切分，彙整各網格之需求量形成多序列時間序列資料；整合時間日曆特徵、氣象（降雨、氣溫）與人口結構（總人口、高齡人口）等共變數，其中降雨以 Kriging 內插至網格尺度，人口則由行政區轉換至網格以維持尺度一致。資料期間選取民國 107、108、112、113 年以降低非典型事件干擾。

模型採用 DeepAR 與 Temporal Fusion Transformer（TFT）進行機率預測，並以 MAE、RMSE、預測區間覆蓋率（PICP@80%）衡量準確性與覆蓋表現；另納入容忍誤差率（TRE，τ=±1）評估可操作性，並以分位數校準曲線檢核分位數輸出之校準程度。本研究貢獻在於建立高解析度網格需求之機率預測流程與多源資料整合方法，並比較 DeepAR 與 TFT 之系統性權衡，提供救護部署與決策支援之實證基礎。

關鍵字：EMS、需求預測、時空網格、機率式預測、DeepAR、TFT

doi:10.6342/NTU202600877

Abstract

Emergency Medical Services (EMS) demand is highly stochastic and spatiotemporally heterogeneous. Accurate forecasting can support ambulance pre-positioning and dynamic dispatching. This study proposes a grid-based probabilistic multi-horizon forecasting framework for EMS demand in Taipei City, providing both point forecasts and uncertainty information for risk-aware planning.

Taipei is partitioned into regular 1,000 m × 1,000 m grids, with demand aggregated into 4-hour intervals to form multiple time series. Covariates include calendar/time features, meteorological variables (rainfall and temperature), and demographics (total and older population). Station rainfall is interpolated to grids using Kriging, and administrative-area population is converted to grids to ensure spatial consistency. Data from 2018, 2019, 2023, and 2024 (ROC years 107, 108, 112, 113) are used to reduce atypical effects.

We implement DeepAR and the Temporal Fusion Transformer (TFT) for probabilistic multi-horizon forecasting. Performance is evaluated by MAE, RMSE, and the 80% prediction interval coverage probability (PICP@80%), together with the tolerant rate error (TRE, $\tau=\pm1$) to reflect operational usability. We further assess probabilistic reliability using quantile calibration curves. The results provide a reproducible high-resolution forecasting pipeline and an empirical comparison of DeepAR and TFT to inform EMS deployment and decision support.

Keywords: EMS, demand forecasting, spatiotemporal grid, probabilistic forecasting, DeepAR, TFT

doi:10.6342/NTU202600877

目次

doi:10.6342/NTU202600877

doi:10.6342/NTU202600877

圖次

doi:10.6342/NTU202600877

doi:10.6342/NTU202600877

表次

doi:10.6342/NTU202600877

第一章 簡介

緊急醫療服務（Emergency Medical Services, EMS）為公共安全與公共衛生體系中的關鍵環節，其核心任務在於以有限資源回應高度不確定且具急迫性的院前緊急事件。對於到院前心跳停止（Out-of-Hospital Cardiac Arrest, OHCA）等高危情境而言，救護車反應時間與存活率及神經學預後具有顯著關聯，因此如何透過更精準的需求預測以支援救護資源之前置部署與動態調度，長期以來皆為實務與研究共同關注的議題。

EMS 需求具備典型的時空異質性（spatio-temporal heterogeneity）：同一城市中，不同區域因人口結構、土地使用、活動強度與交通可達性差異而呈現不同的需求基底；同一區域亦會因日內通勤與活動節奏、週期性（日／週／季）行為，以及外生因素（如天候條件）而波動。若僅以行政區為分析單元，容易受制於邊界固定、尺度不一致與區內差異被平均化等限制；因此，近年研究與實務逐漸採用規則網格（grid-based）作為空間單元，以更一致的尺度刻畫需求分布，並利於特徵融合、模型建構與結果視覺化。

在預測方法上，傳統統計與機器學習模型多以點預測（point forecast）為主，雖可提供一定程度的平均趨勢判讀，但在高解析度網格與短時間尺度情境下，EMS 需求常呈現計數型資料特性與大量零值（zero-inflation），使需求分布更具離散性與不確定性。對救護部署而言，決策者往往不僅需要了解預測值是多少，更需要需求可能落在何種範圍以進行風險導向的備援與容錯規劃；同時，公共服務決策亦高度重視可說明性與可追溯性，以提升模型輸出被採納的可行性。因此，本研究以高解析度需求刻畫、不確定性量化與決策支援為核心目標，導入機率式深度學習架構進行 EMS 網格需求預測。

本研究以臺北市為研究區域，採用規則網格作為空間單元，並以較短時間解析度刻畫日內需求波動，同時整合時間日曆、氣象與人口等多源共變數；預測模型則採用機率式深度學習架構—深度自回歸循環網絡（Deep Autoregressive

doi:10.6342/NTU202600877

Recurrent Network, DeepAR)與時間融合轉換器(Temporal Fusion Transformer, TFT)—執行多步（multi-horizon）需求預測，以同時輸出點預測與預測區間／分位數等不確定性資訊，提供更貼近救護資源規劃情境的決策依據。

1.1 研究背景與動機

城市 EMS 需求的變動具有高隨機性與強週期性並存的特徵：一方面，緊急事件本質上難以完全預測，且在短時間尺度下可能出現突發尖峰；另一方面，需求又常與日內活動節奏（白天活動、夜間低活動）、週末與平日差異、節假日型態改變等呈現規律。加上降雨、氣溫等天候條件可能改變外出風險與交通狀況，使需求與到達時間同時受影響，進一步提升調度難度。若缺乏對需求時空分布的有效掌握，容易造成資源配置失衡：高需求區域出現供給不足、低需求區域則可能資源閒置，最終影響系統整體反應效率與服務品質。

因此，本研究動機在於建立一套能於高解析度網格 × 短時間尺度下運作、並能輸出不確定性資訊的 EMS 需求預測框架，使救護部署能從單一預測值驅動進一步提升為風險型規劃與容錯型決策，以更貼近實務需求。

1.2 研究問題

綜合上述背景，本研究關注以下核心問題：

1. 在高解析度網格與短時間尺度情境下，如何以一致且可重現的流程，整合 EMS 派遣資料與多源外生資料（氣象、人口等），建立可供模型學習的時空面板資料。
2. 相較於以點預測為主的模型，機率式深度學習模型是否能在 EMS 需求預測中提供更具決策價值的不確定性資訊（如分位數、預測區間）。
3. 在不同預測視野（如 1 日與 7 日）下，DeepAR 與 TFT 兩類機率式深度學習架構的表現差異為何，其優劣是否反映不同決策偏好（偏向覆蓋需求 vs. 偏向精準降低閒置）。

doi:10.6342/NTU202600877

4. 在公共服務決策脈絡中，如何以更貼近調度容錯的方式設計評估指標，以衡量模型輸出在實務部署上的可操作性。

1.3 研究目的與研究範圍

本研究之目的為建構一套針對臺北市 EMS 需求之時空網格化機率預測方法，並比較 DeepAR 與 TFT 在不同預測視野下之效能與適用情境。具體目標如下：

1. 建立臺北市 EMS 需求的網格化時空資料架構：以規則網格為空間單元，並以日內固定時段切分形成多序列時間序列資料。
2. 建構多源共變數的尺度一致整合流程：將測站型氣象資料轉換為網格尺度特徵，並將行政區人口統計轉換至網格單元，形成可重現的特徵工程流程。
3. 導入機率式深度學習模型進行多步預測：採用 DeepAR 與 TFT 進行 multi-horizon 預測，以同時產生點預測與不確定性資訊。
4. 建立兼具「準確性—不確定性—實務可用性」的評估框架：除傳統點預測誤差外，納入預測區間覆蓋與容忍誤差等指標，以更貼近救護調度決策情境。

研究範圍方面，本研究以臺北市為研究區，資料期間選定民國 107、108、112、113 年，以降低疫情期間非典型行為模式與政策衝擊對需求學習之干擾；空間單元採規則網格，時間單元採日內固定時段切分，以支援短時間尺度的需求變動刻畫與部署應用。

1.4 研究方法概述

本研究流程可概分為四個步驟：

1. **資料蒐集與前處理**：匯整 EMS 派遣紀錄、氣象觀測（降雨、氣溫）與人口統計（總人口、高齡人口）資料，進行清理與欄位標準化，建立一致的時間索引。
2. **網格化與特徵工程**：將 EMS 事件點位彙整至網格×時段的案件計數需求；

doi:10.6342/NTU202600877

將測站降雨資料以空間內插方法轉換為網格尺度特徵；將行政區人口統計以面積加權方式網格化，並加入日曆事件與日內尖離峰等時間特徵。

3. **模型建構與訓練**：以 DeepAR 與 TFT 建構跨網格共享學習的機率式預測模型，進行多步需求預測；並依預測視野設定訓練／測試切分方式以避免未來資訊洩漏。
4. **模型評估與解釋**：以點預測指標衡量準確性，以預測區間覆蓋衡量不確定性校準程度，並以容忍誤差概念評估部署可操作性；同時透過特徵重要性或模型內建機制提供解釋線索。

1.5 論文架構

本論文各章內容安排如下：

- **第一章（簡介）**：說明研究背景、研究問題、研究目的、方法概述與論文架構。
- **第二章（文獻回顧）**：回顧 EMS 需求預測相關研究，涵蓋時空網格化方法、傳統與機器學習模型演進，以及機率式深度學習模型（DeepAR、TFT）與不確定性評估概念，並據此界定研究缺口與定位。
- **第三章（研究方法）**：詳細說明研究架構、資料來源與範圍、資料前處理與特徵工程、模型輸入資料產製、模型訓練設定與評估指標。
- **第四章（實驗結果與討論）**：呈現 DeepAR 與 TFT 在不同預測視野下之效能比較，並從點預測、不確定性可靠性與決策可用性等面向進行分析與討論。
- **第五章（結論與未來研究方向）**：總結研究發現與貢獻，提出研究限制與後續可行之延伸方向。

doi:10.6342/NTU202600877

第二章 文獻回顧

2.1 EMS 需求預測經典與作業決策脈絡

2.1.1 EMS 需求預測之作業決策意義與研究背景

緊急醫療服務（Emergency Medical Services, EMS）需求預測之研究目的，並非僅在於估計未來案件量，而在於支援救護車預置（pre-positioning）、動態重定位（redeployment）、派遣規則調整、人力配置與醫療資源協調等作業決策。由於 EMS 系統同時面對高不確定性、服務時效壓力與有限資源約束，預測誤差往往會透過派遣與配置流程被放大，進而影響平均反應時間、服務涵蓋能力與系統整體效率。因此，需求預測應視為 EMS 作業管理與決策設計的重要資訊基礎，而非單純統計估計問題。[1-3]

此外，在高時效急症（如院外心跳停止（out-of-hospital cardiac arrest, OHCA））情境下，系統層級之辨識、派遣與到院前處置流程與患者預後具有高度關聯，因而使 EMS 需求預測同時具有公共衛生與急救醫療系統韌性之政策意涵。本文涉及高時效處置之敘述時，係以心肺復甦術（cardiopulmonary resuscitation, CPR）／OHCA 指引及急救醫學實證研究作為引用依據，以維持論述之嚴謹性與可驗證性。[4-7]

從需求生成機制觀點而言，EMS 需求通常同時受到人口結構、土地使用、社會經濟條件、醫療資源可近性、氣象條件、日夜週期與特殊事件等因素影響，且不同因素在不同時間解析度與空間單元下的作用強度並不一致。此一特性使 EMS 需求預測本質上成為兼具時間相依、空間異質與多源特徵整合需求之時空建模問題。[8-11]

2.1.2 經典統計預測與比較研究之發展

EMS 需求預測之早期研究，多以統計模型與比較研究為主，其研究重點在於辨識呼叫量或案件量之時間序列規律，並建立可供營運管理使用之基本預測架

doi:10.6342/NTU202600877

構。Channouf 等人（2007）以加拿大 Calgary 之 EMS 資料比較多種預測技術，為 EMS 需求預測之實證研究奠定重要基礎；Setzler 等人（2009）則進一步比較多類模型於 EMS 呼叫量預測中的表現，指出模型效能會受到時間粒度與資料聚合方式影響，顯示 EMS 需求預測具有明顯之尺度敏感性。[12, 13]

然而，當研究目標由較粗粒度之總量預估，轉向高解析度之短期需求預測時，資料常出現大量零值、尖峰事件與變異不穩定等現象，使傳統線性模型或簡單時間序列方法在模型彈性與分布假設上逐漸受到限制。此一發展脈絡促使後續研究開始轉向更能處理非線性關係、異質特徵與稀疏資料的建模方法。[14-16]

2.1.3 細尺度需求建模之經典脈絡

為回應細尺度 EMS 需求預測之建模需求，研究逐步導入時空統計模型，以同時處理時間動態、空間分布與需求稀疏性。Zhou 等人（2015）將救護需求建模為時空點過程（spatio-temporal point process），強調 EMS 需求應被視為隨時間變動之空間事件過程，而非僅以總量時間序列表徵。Zhou 與 Matteson（2015）則提出時空核密度方法（spatio-temporal kernel approach），透過對歷史事件給予不同時間與空間權重，以提升短期需求空間分布之預測表現。[17-19]

此一研究脈絡對後續地理網格預測、熱點辨識與動態重定位之發展具有深遠影響，因其已建立「何時、何地可能發生需求」之分析框架。然而，當研究問題進一步擴展至多源特徵整合（例如人口、土地使用、氣象與醫療資源等）與跨區域泛化時，傳統點過程與核密度方法在非線性擬合與高維特徵表徵上的限制亦逐漸浮現。[8, 10, 20]

2.1.4 EMS 需求預測與預配置／重定位決策之連結

EMS 需求預測研究之價值，最終仍須回到作業決策層面加以檢驗。相關作業研究文獻指出，救護車配置、站點位置、動態重定位與派遣策略之效能，往往高度依賴需求預測是否能正確反映時空風險分布。若預測模型未能捕捉局部熱點、尖峰時段或異常事件風險，即便後續最佳化模型形式完整，仍可能在實際運作中

doi:10.6342/NTU202600877

產生資源錯置或反應時間惡化等問題。 [1-3]

因此，EMS 需求預測在本質上應被理解為預測模型—作業決策之耦合問題，而非單一演算法競賽。此一觀點亦構成本研究後續強調機率式預測、多步期預測與決策導向評估架構之理論基礎。[21-23]

2.1.5 小結

本節整理 EMS 需求預測之經典研究與作業決策脈絡，指出 EMS 需求預測不應僅被視為數值估計問題，而應作為支援預配置、重定位與派遣決策的核心資訊基礎。文獻亦顯示，隨研究尺度由總量預測轉向細尺度時空預測，傳統統計模型與核密度方法雖具奠基性價值，但在高維特徵整合與非線性表徵方面逐漸面臨限制，並促成後續資料驅動方法之發展。[1, 18, 22]

2.2 現代 ML／DL 的時空 EMS 預測

2.2.1 由傳統統計模型轉向資料驅動方法之背景

隨著 EMS 資料來源日益多元（例如歷史事件序列、人口特徵、氣象變數、地理資訊與活動訊號）以及運算能力提升，EMS 需求預測研究逐步由傳統統計模型轉向機器學習（machine learning, ML）與深度學習（deep learning, DL）方法。其主要原因在於，EMS 需求常呈現非線性關係、跨變數交互作用與空間異質性，傳統方法較難完整捕捉，而 ML／DL 在高維特徵整合與複雜模式擬合方面具有較高彈性。[10, 20, 23]

另一方面，研究問題本身亦由總量預估逐漸轉向時空細尺度預測，使模型不僅需回答未來需求有多少，更需辨識需求將發生於何時、何地，以及形成局部熱點之可能性。此一任務轉變促使空間特徵工程、聚類分區、固定網格表示與圖結構建模方法逐步成為 EMS 需求預測的重要技術構成。[11, 22, 24]

2.2.2 特徵工程與集成學習在 EMS 需求預測之應用

在 ML 方法中，特徵工程與集成學習（ensemble learning）被證實為提升 EMS

doi:10.6342/NTU202600877

需求預測效能的重要因素。Lin 等人（2020）於國家尺度區域救護需求預測中，透過多類型特徵工程（multi-nature features）與機器學習模型整合時間、區域與環境資訊，顯示多源特徵設計可有效提升預測表現。Martin 等人（2021）亦指出，空間分區策略與模型設計需一併考量，方能使時空機器學習方法在 EMS 場域中展現穩定效能。[9, 10]

近年研究進一步發展至堆疊式與元學習集成架構。Megouo 與 Pierre（2024）提出 emergency call forecasting 之 stacking ensemble 模型，顯示多基學習器之互補性可提升預測穩健性；Garg 等人則提出兼具穩健性與可解釋性的 meta-learning ensemble framework，強調 EMS 場域模型不應僅追求準確度，亦需兼顧部署穩定性與解釋能力。[22, 25, 26]

在區域實證研究方面，Hermansen 與 Mengshoel（2021）之 Oslo 案例以及 Neira-Rodado 等人（2025）之 Barranquilla 案例均顯示，EMS 需求預測效能高度依賴資料特性與特徵設計品質。尤其 Neira-Rodado 等人（2025）透過聚類式空間切分、主成分分析（principal component analysis, PCA）降維與多源時空特徵整合，展示 XGBoost 等模型在稀疏資料情境下之潛力，進一步凸顯「空間劃分策略＋特徵工程」對模型成效之關鍵影響。[11, 20, 27]

2.2.3 時空特徵、空間劃分與網格化建模

在時空 EMS 需求預測任務中，空間單元的定義方式（如行政區、固定網格或聚類分區）會直接影響資料稀疏性、區域同質性與模型可解釋性。固定網格有助於標準化空間表示並與外部地理圖層對齊，但當網格尺度過細時，容易產生大量零值與噪音；聚類分區雖可提升區域內部同質性，然其結果易受分群方法、參數與資料期間影響，在跨期間或跨城市比較時可能較不穩定。此議題可與可變面積單元問題（modifiable areal unit problem, MAUP）相互對照理解。[10, 11, 28]

都市時空網格預測文獻（例如 DeepMeshCity）對本研究之空間網格化表示與城市尺度時空建模具有重要參考價值；然而，其研究目標並非 EMS 需求預測，

doi:10.6342/NTU202600877

因此於本文中係定位為相關方法參照，而非核心基礎模型。就方法論脈絡而言，此類研究有助於說明固定網格單元下跨網格資訊傳遞與多尺度表示學習之可行性。[29-32]

2.2.4 圖神經網路與異質拓撲關係建模

相較於以網格或行政區作為相對獨立的空間單元，圖神經網路（graph neural network, GNN）可更自然地處理非歐幾里得空間關聯與異質節點互動。Jin 等人（2021）提出二分圖卷積網路（Bipartite Graph Convolutional Network, BiGCN），以醫院與區域作為異質節點建構二分圖，將 EMS 需求與醫療資源之交互關係納入同一圖框架中建模。此一方向之方法論意義在於，研究焦點已由單純需求端預測延伸至供需雙端結構表徵，更接近 EMS 實際作業中的資源互動情境。[24, 33, 34]

雖然本研究之核心模型未直接採用 GNN，然相關文獻仍提供重要之方法論補充：其一，作為空間關聯與供需結構建模之理論依據；其二，作為後續特徵工程與模型延伸之參照，以增強模型輸出與作業決策（例如跨區支援與醫院負荷）之連結。[1, 22, 24]

2.2.5 小節

本節說明 EMS 需求預測方法由傳統統計模型轉向 ML／DL 之背景，並整理特徵工程、集成學習、空間劃分、網格化建模與 GNN 方法在 EMS 場域中的應用脈絡。整體文獻顯示，現代方法之優勢不僅在於提升預測準確度，更在於強化時空關聯表徵、多源特徵整合與決策可用性；惟模型選擇仍須與任務目標、空間單元設計及資料特性相互對齊。[10, 11, 22, 24]

2.3 機率預測與評估方法

2.3.1 點預測與機率預測之差異

既有 EMS 需求預測研究多以點預測（point forecasting）為主，即輸出單一

doi:10.6342/NTU202600877

預測值（例如未來某時段案件數）。此類方法在總量規劃、班表估算與容量管理層面具有實用性，惟在高不確定性且需風險管理之 EMS 調度情境下，單一點估計往往不足以反映需求分布的不確定性。特別是在零膨脹、尖峰事件與局部熱點情境下，僅依點預測進行資源預置，可能導致高風險區域被低估，進而降低系統應變能力。[13, 15, 21]

相較之下，機率預測（probabilistic forecasting）可提供需求分布、分位數區間或事件發生機率，使決策者得以依風險偏好設定不同預置門檻與重定位策略。此一轉向不僅是模型輸出形式的調整，更代表研究問題由預測多少進一步延伸至需求將於何時何地，以多大不確定性發生，因此與 EMS 之風險導向決策本質更為一致。[35-37]

2.3.2 分類任務之評估指標選擇

在細空間網格與短時間窗設定下，EMS 資料常呈現大量零值與少數非零事件，形成典型不平衡（imbalanced）與零膨脹（zero-inflated）資料結構。若在此情境下僅以 Accuracy 作為評估指標，模型即使偏向預測多數類別（無事件），仍可能取得表面上較高之準確率，但對少數且作業上重要之事件樣本辨識能力不足。[38-40]

因此，若研究任務涉及事件發生預測、熱點辨識或門檻式風險分類，除 Accuracy 之外，宜納入 F1 score、Balanced Accuracy 與 PR-AUC 等指標，以更完整反映模型對少數類別之辨識能力與實務價值。特別是 PR-AUC 在不平衡資料情境下通常較 ROC-AUC 更具資訊性，而 Balanced Accuracy 則有助於降低類別比例偏斜造成之評估偏誤。[38, 39, 41]

在資料處理策略方面，不平衡與零膨脹問題可透過 zero-inflated 模型、重抽樣（如 SMOTE）與成本敏感學習等方法處理；惟於時空序列任務中仍須審慎評估其對時間依賴與空間結構之影響。因此，本文之方法設計係以評估架構補足不平衡指標為主要策略，並將資料不平衡視為任務本質之一部分，而非假設其可透

doi:10.6342/NTU202600877

過前處理完全消除。[40, 42-44]

2.3.3 機率預測之評估：校準、覆蓋率與銳利度

機率預測之評估不應僅關注排序能力或分類能力，亦須檢驗模型輸出之機率值是否可信，即是否具備良好校準（calibration）。若模型預測某事件發生機率為0.7，理想情況下長期觀察應約有 70%實際發生；此一性質對 EMS 調度決策尤為重要，因車輛預置與增援門檻常直接依賴預測機率值設定。[42, 45-47]

若研究採用區間預測或分位數預測，則除校準外，尚應同時評估覆蓋率（coverage）與區間寬度（sharpness）。覆蓋率可檢驗名目信賴水準與實際涵蓋比例是否一致，而區間寬度則反映模型是否過度保守。進一步而言，proper scoring rules（如 CRPS）與分位數損失（pinball loss）亦可作為整體機率預測品質之評估工具。[35, 48-50]

2.3.4 本研究評估架構之設計原則

基於 EMS 時空需求預測之任務特性，評估設計可從三個面向加以檢視：第一，點預測誤差（例如 MAE、RMSE），用以衡量需求量估計能力；第二，機率輸出之可靠性（例如校準、預測區間覆蓋率與區間寬度），用以檢驗模型輸出能否作為風險導向決策之可信資訊來源；第三，若任務進一步擴展為事件發生預測／熱點偵測／門檻式風險分類，則可額外納入事件／熱點辨識指標（例如 F1、Balanced Accuracy、PR-AUC），以評估模型在不平衡資料下對高風險樣本之識別能力。本研究聚焦於計數型需求量之機率預測與調度支援，因此以 MAE/RMSE 評估點預測準確性，並以 PICP、區間寬度與分位數校準曲線檢核機率可靠性，同時以容忍誤差率（TRE，$\tau=\pm1$）反映操作上可接受的誤差尺度；事件／熱點辨識指標則列為未來研究擴充方向。[14, 21, 47]

此種多面向評估架構有助於避免模型僅在單一指標上表現良好，卻無法有效支援實務決策之情形，亦可使本研究之模型比較更貼近 EMS 管理需求，而非停留於一般機器學習模型績效比較。[22, 23, 35, 42]

doi:10.6342/NTU202600877

2.3.5 小結

本節將點預測與機率預測之差異獨立說明，並補充不平衡資料與機率預測評估之方法論基礎。整體文獻顯示，若研究目標為支援 EMS 風險導向調度決策，則評估架構不宜僅停留於 Accuracy 或單一誤差指標；若研究任務涉及事件／熱點辨識，可納入 F1、Balanced Accuracy、PR-AUC 等指標，並搭配校準相關指標，以提高模型比較之實務意義。[35, 38, 47]

2.4 本研究之方法基礎模型與研究定位

2.4.1 以 DeepAR 為代表之機率式自回歸時間序列建模基礎

本研究之第一項方法基礎建立於機率式時間序列預測之建模原則，即 EMS 需求不應僅以單一點估計表示，而應以分布、區間或事件發生機率呈現，以支援風險導向決策。DeepAR 為代表性之機率式自回歸神經網路框架，其核心特性在於可透過多序列共同訓練學習跨區域共享之時間動態結構，並輸出未來需求之機率分布參數，適合用於多時空單元、需求稀疏且波動顯著之場景。[21, 35, 36]

從模型邏輯觀點而言，DeepAR 延續自回歸時間序列建模思想，並結合循環神經網路（recurrent neural network, RNN）／長短期記憶網路（long short-term memory, LSTM）對歷史資訊之遞迴表徵能力，使模型得以同時捕捉週期性、短期波動與跨序列共享模式。此特性對 EMS 需求預測尤為重要，因不同區域或網格雖具空間異質性，仍可能在日夜週期、週間規律與整體環境變化下呈現可共享之時間結構。[9, 36, 51]

就本研究之方法定位而言，DeepAR 的重要性不僅在於預測效能，更在於其機率式輸出可作為後續風險排序、預置門檻設定與動態重定位策略之資訊基礎。亦即，本文採納 DeepAR 所代表之方法論原則，強調以需求不確定性量化取代單一點值預測，並將其作為 EMS 決策支援之主要資訊訊號。[22, 23, 36, 42]

doi:10.6342/NTU202600877

2.4.2 以 TFT 為代表之多步期預測與可解釋性特徵融合基礎

本研究之第二項方法基礎為多步期預測（multi-horizon forecasting）與可解釋性特徵融合架構。EMS 調度與資源配置決策通常不僅針對下一個時間窗進行反應，而需同時考量未來多個時段之風險變化。因此，模型需具備多步期輸出能力，並能整合靜態特徵（如區域屬性）、已知未來特徵（如時間與行事曆變數、部分天氣預報）與歷史觀測特徵。Temporal Fusion Transformer（TFT）在此方面提供了與本研究問題高度對齊之代表性架構。[20, 37, 52]

TFT 透過變數選擇機制、門控殘差網路與注意力機制，提升模型對特徵重要性與關鍵時間依賴之表徵能力，使模型除具備預測功能外，亦可提供一定程度之可解釋資訊。對 EMS 管理場域而言，此特性具有實質意義，因模型若能說明高風險判斷主要由哪些因素驅動（例如時段、天候、區域屬性或近期需求變化），將有助於管理者理解模型行為並提高採納意願。[22, 37, 42, 46]

因此，本文採納 TFT 所代表之方法論原則，以多來源特徵融合與多步期預測能力支援 EMS 短期規劃與決策判讀，並提升模型輸出與實務管理語境之對接性，而非僅停留於黑箱式預測結果。[21, 23, 37, 42]

2.4.3 DeepMeshCity 與 GNN 文獻之角色

DeepMeshCity 屬於都市網格預測（urban grid prediction）之通用深度學習模型，對本研究在空間網格化表示與城市尺度時空建模上具有方法參照價值；然其研究任務並非 EMS 需求預測，故本文不將其列為方法基礎模型。本文之核心方法基礎仍以 DeepAR 與 TFT 所代表之機率式與多步期時間序列預測架構為主，而 DeepMeshCity 之角色係作為跨領域網格化建模文獻，用以支持本研究之空間表示合理性與方法討論。[32, 36, 37]

另一方面，GNN／BiGCN 相關研究雖非本研究現階段核心模型，但其對於 EMS 問題中非歐幾里得關聯、供需雙端結構與異質節點互動建模提供重要理論補充。此類文獻可作為本研究在特徵工程設計與後續模型延伸之參照，尤其在需

doi:10.6342/NTU202600877

將區域需求、醫療資源量能與跨區服務外溢納入同一分析框架時，具有明確方法論價值。[24, 33, 34, 42]

綜合而言，本研究之方法定位可概括為：以 EMS 作業決策需求為導向，在地理網格化時空單元下，採用機率式與多步期深度時間序列模型作為核心架構，並透過時空特徵工程、資料不平衡評估指標與機率校準觀點建立完整評估架構，以提升模型對於風險辨識、預置決策與動態重定位之實務可用性。此一定位亦可視為對既有文獻中點預測偏重、校準評估不足、方法與決策連結不夠明確等問題之回應。[21-23, 35, 42, 47]

2.4.4 小結

本節界定本研究之方法基礎與研究定位，明確指出核心基礎模型為 DeepAR 與 TFT 所代表之機率式與多步期時間序列預測架構；DeepMeshCity 則作為網格化建模之跨領域參照，GNN／BiGCN 文獻則作為空間關聯與供需結構建模之理論補充。此一定位使本研究在方法選擇上得以同時回應 EMS 時空需求預測之預測效能、風險量化與決策可用性需求。[24, 32, 36, 37]

2.5 文獻評述與研究缺口

2.5.1 現有文獻之整體評述

綜合前述文獻可知，EMS 需求預測研究已由早期以總量預測與統計比較為主之研究脈絡，逐步發展至兼具時空建模、多源特徵整合與深度學習架構之資料驅動方法。此一演進顯示研究焦點已不再侷限於是否能提升預測誤差指標，而是進一步延伸至模型能否捕捉時空異質性、支援細尺度風險辨識，並與實際調度與資源配置需求相連結。從研究方向而言，這代表 EMS 需求預測正由單純預測問題，轉向兼具作業管理與公共衛生決策意涵之應用型時空預測問題。[10, 13, 18, 22]

然而，現有文獻在方法評估與研究定位上仍存在若干結構性問題。首先，部

doi:10.6342/NTU202600877

分研究雖報告模型準確度提升，但對於模型輸出如何實際轉化為預配置、重定位或派遣策略之決策規則，說明仍相對有限。其次，不同研究於時間粒度、空間單元與特徵集設定差異甚大，導致跨研究間之比較困難，亦限制了方法結論之一般化。換言之，現階段文獻雖已累積豐富之模型實作成果，但在可比較性與決策可轉譯性方面仍有待進一步強化。[1, 2, 21, 23]

2.5.2 方法與資料表徵層面之研究缺口

在方法與資料表徵層面，現有研究的重要缺口之一，在於空間單元設計與空間結構建模策略尚未形成一致且具可比較性之方法框架。固定網格、行政區與聚類分區各具優勢與限制：固定網格有利於標準化與外部資料對接，但易導致細尺度稀疏與噪音問題；聚類分區雖可提升區域同質性，卻可能因資料期間與分群設定不同而降低跨期間穩定性；行政區則具政策與治理上的可解釋性，但未必符合需求生成之自然邊界。此一差異使研究成果常受空間切分方式主導，而非純粹反映模型本身能力。[10, 11, 28]

此外，雖已有 GNN／BiGCN 等研究嘗試將供需雙端結構與空間關聯納入模型，但多數 EMS 需求預測文獻仍以需求端單向建模為主，對於醫療資源容量、跨區支援關係與醫院負荷外溢等結構性因素之納入程度有限。此一現象使模型雖可提升需求預測精度，卻未必能完整對應 EMS 系統之實際服務拓撲與資源互動機制。相關文獻顯示，異質圖建模具有補足此缺口之潛力，但其在不同城市情境與不同資料可得性條件下之穩健性與部署成本，仍需更多實證驗證。[24, 33, 34, 42]

再者，多數研究之實證設計仍以單城市、單資料集為主，外部驗證（external validation）與跨區域泛化能力分析相對不足。即使近期研究開始強調基準模型與外部驗證的重要性，現有文獻仍普遍缺乏在不同城市結構、不同需求密度與不同資料品質條件下之系統性比較。此一缺口意味著模型在特定場域中表現優良，並不必然代表其可直接移轉至其他城市或行政區。[20-22, 42]

doi:10.6342/NTU202600877

2.5.3 預測任務定義與評估層面之研究缺口

在預測任務定義與評估方法方面，現有文獻之主要缺口在於仍偏重點預測（point forecasting）與傳統誤差指標（如 MAE、RMSE），而較少將機率預測與風險辨識作為主要研究目標。對 EMS 調度而言，單一點估計雖可支援總量規劃，但對於高風險時段與高風險區域之辨識而言，其決策資訊密度仍有限。相對地，若能提供事件發生機率、分位數區間或風險排序結果，將更能支援預置門檻設定與重定位決策之彈性化設計。[35-37]

其次，在細尺度時空任務常見之零膨脹與類別不平衡情境下，部分研究仍過度依賴 Accuracy 或整體平均誤差，導致模型在多數類別（無事件）上取得高分時，掩蓋其對少數關鍵事件樣本之辨識不足。由此可見，若研究目標涉及事件發生預測、熱點偵測或門檻式風險分類，則評估指標應更明確納入 F1 score、Balanced Accuracy 與 PR-AUC，以反映模型對少數類別之實際辨識能力。[38-42]

另一項關鍵缺口則在於機率輸出之校準（calibration）評估不足。即使研究採用可輸出機率或區間之模型，亦常將焦點集中於排序能力或平均誤差，而未充分檢驗模型機率值是否可被決策端可信地使用。對 EMS 而言，預測機率若未經校準，將可能導致預置門檻設定失真，進而造成資源過度保守或不足配置。因此，校準、覆蓋率與區間寬度等指標應被納入核心評估架構，而非僅作為附加分析。[42, 45-47]

2.5.4 本研究之回應策略與研究缺口對應

綜合上述文獻評述，本研究之主要研究缺口對應可歸納為三點。第一，在任務定義層面，現有 EMS 需求預測文獻仍偏重點預測與平均誤差最小化，較少系統性強調機率式風險輸出與決策門檻設計之連結；因此，本研究將機率預測納入核心方法設計，以回應 EMS 調度在風險辨識與預置決策上的實際需求。[21, 35, 36, 42]

第二，在方法架構層面，現有研究雖已廣泛使用 ML／DL 與時空特徵工程，

doi:10.6342/NTU202600877

但在多步期預測、特徵融合與決策可解釋性之整合上仍有進一步發展空間。基於此，本研究以 DeepAR 與 TFT 所代表之機率式與多步期時間序列建模原則作為核心方法基礎，並以網格化時空單元與多源特徵整合支援短期風險評估與作業判讀；DeepMeshCity 則僅作為空間網格建模之方法參照，而非核心基礎模型。[32, 36, 37, 42]

第三，在評估設計層面，現有文獻對不平衡資料與機率校準之處理仍不充分，致使模型評估結果與決策可用性之對應關係不足。為回應此一缺口，本研究之評估架構以點預測準確性與機率輸出可靠性為核心，並進一步納入容忍誤差率（TRE，$\tau=\pm1$）以反映在可操作容錯尺度下之決策可用性；至於事件／熱點辨識指標（如 F1、Balanced Accuracy、PR-AUC）屬於任務擴展為事件分類或熱點偵測時之補充評估工具，故於本研究中列為後續可延伸之方向。[14, 22, 38, 42, 47]

整體而言，本研究之研究定位並非僅追求單一模型在單一指標上的最佳化表現，而是試圖建立一套更貼近 EMS 作業決策需求之時空需求預測框架，使模型輸出可用於風險辨識、預置決策與資源調度之實務情境。此一方向亦可視為對現有文獻在任務定義、評估設計與決策連結三方面不足之系統性回應。[1, 2, 22, 23, 42]

2.5.5 小結

本節綜整既有文獻之方法發展與應用限制，指出目前 EMS 需求預測研究仍存在三項主要缺口：其一，點預測偏重且機率式風險輸出不足；其二，空間單元設計與空間結構建模之可比較性與泛化驗證不足；其三，不平衡資料與機率校準之評估架構尚未充分納入。基於上述缺口，本研究以機率式、多步期且決策導向之時空預測架構作為主要研究方向。[21, 36, 37, 42, 47]

doi:10.6342/NTU202600877

第三章 研究方法

本章旨在系統性闡述本研究之研究設計、資料來源與前處理、特徵工程、模型輸入資料產製、深度學習模型訓練流程，以及模型評估指標。研究目標為針對臺北市緊急救護服務（Emergency Medical Services, EMS）需求，建立具備空間解析度與時間解析度之時空序列預測框架；並選用民國 107 年、108 年、112 年、113 年資料（排除新冠疫情期間資料）進行建模與驗證，以降低非典型社會行為、就醫模式與政策衝擊對需求學習之干擾。

本研究以 1000 公尺×1000 公尺網格作為空間分析單元，並在時間維度上將一天劃分為 6 個 4 小時時段（Time_Bucket），整合時間、氣象與人口統計等多源共變數，最終以 DeepAR 與 Temporal Fusion Transformer（TFT）兩類深度學習模型進行機率式多步預測（multi-horizon forecasting），建立可重現之資料處理與訓練流程。

3.1 研究架構

為確保研究流程具備一致性、透明性與可複製性，本研究之整體架構可歸納為四個階段（如圖 3-1 所示），由原始資料處理至模型評估形成完整分析流程。

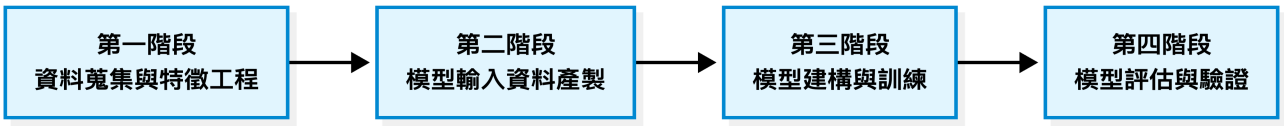

圖 3-1 研究流程圖

第一階段為資料蒐集與特徵工程，主要匯整 EMS 派遣紀錄、氣象觀測資料與人口統計資料，並完成清理、欄位標準化與基礎特徵建構。

第二階段為模型輸入資料產製，重點在於多源資料於時間與空間索引之對齊：將點位事件彙整至標準網格、將測站資料轉換至網格尺度，並整合時間特徵與人口背景變數，形成一致的時空面板資料。

doi:10.6342/NTU202600877

第三階段為模型建構與訓練，分別訓練 DeepAR 與 TFT 兩模型，使其學習跨網格（多序列）共享的需求生成機制，同時保留各網格之異質性。

第四階段為模型評估與驗證，利用保留之測試資料，以點預測誤差與預測區間可靠性等指標量化模型表現，並比較不同模型在需求尖離峰、空間差異與不確定性刻畫能力之差異。

3.2 研究資料與研究範圍

本節界定研究之空間、時間邊界與資料來源，以確保分析結果具有效度與可解釋性。

3.2.1 研究區域與空間單元

本研究以臺北市行政範圍為主要研究區。原始 EMS 案件以地理座標點位呈現，為利於空間統計、特徵融合與模型輸入之結構化處理，需將不規則點位事件轉換為規則空間單元。本研究比較 250 公尺與 1000 公尺兩種網格尺度後，選定 1000 公尺×1000 公尺之規則網格作為基本地理單元（如圖 3-2）。其主要考量在於 EMS 事件於短時間尺度（例如單一 4 小時時段）具有高度隨機與波動特性；若採用較細尺度網格，將出現大量零案件網格（zero-inflation），不利於時序模型穩定學習。相對地，1000 公尺網格可透過適度聚合平滑單點事件突發性，提升訊號穩健性，使需求密度更能反映區域活動圈與需求強度。本研究依此網格化後，臺北市共劃分為 334 個網格，如圖 3-3，作為後續所有特徵彙整與建模之空間索引（grid_id）。

doi:10.6342/NTU202600877

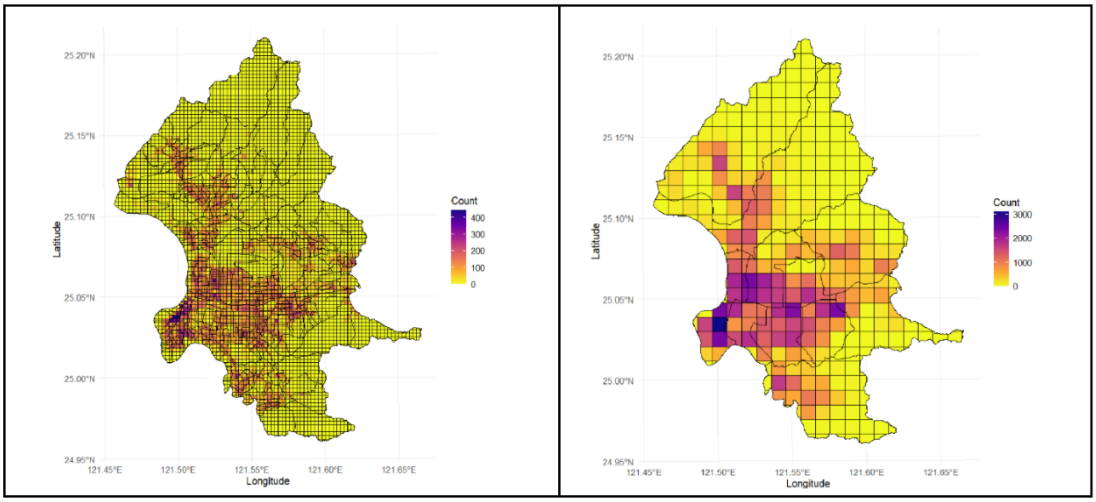

圖 3-2 臺北市 250 公尺與 1000 公尺網格比較圖

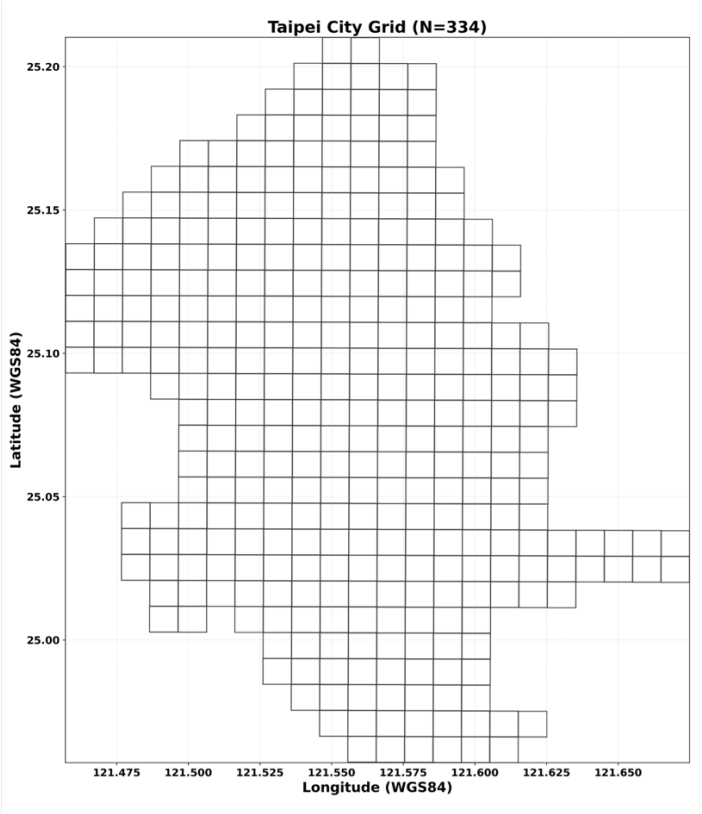

圖 3-3 臺北市網格切分圖

3.2.2 資料來源與時間區間

本研究整合以下資料以刻畫需求生成機制：(1) EMS 派遣紀錄：包含案件發生時間與地點，作為目標變數（需求量）之主要來源；(2) 氣象觀測資料：包含降雨量與氣溫，用以衡量天氣條件對需求之影響；(3) 人口統計資料：以各月

doi:10.6342/NTU202600877

份村里層級總人口與 65 歲以上人口，作為區域結構性背景因子。時間範圍選定民國 107 年、108 年、112 年及 113 年，並排除新冠疫情期間資料，以降低疫情造成之行為改變、政策管制與醫療需求異常波動對模型常態性學習之干擾。

3.3 資料前處理與特徵工程

原始資料需經系統性處理與特徵工程，方能轉化為模型可識別且具預測力之量化變數。本研究之特徵工程聚焦於：時間特徵（勤務與週期行為）、空間對位與目標變數定義、氣象特徵（降雨量與氣溫）以及人口統計特徵之網格化處理。此章節的部分資料由地理所碩二黃妍儒同學產製。

3.3.1 時間特徵（Time Features）

為捕捉 EMS 需求在年度、季節、週期與日內尺度之變動，本研究自時間戳記衍生年（Year）、月（Month）、日（Day）與星期（Day_of_Week）等基礎特徵，並建構與勤務排程相符之時段特徵 Time_Bucket：將每日 24 小時以 4 小時為單位劃分為 6 個區間（00:00–03:59、04:00–07:59、08:00–11:59、12:00–15:59、16:00–19:59、20:00–23:59），以兼顧日內模式辨識與避免過細時間解析度造成之資料稀疏問題。

另建立 Rush_Hour 二元指標，定義 08:00 至 20:00 為活動尖峰時段，以區分日間高活動強度與夜間相對低活動之差異，各時間段的緊急救護需求統計如圖 3-4；同時納入 Weekend 與 Holiday 標記，以捕捉休假日與國定假日之行為模式改變對需求之影響。此外，考量臺灣制度下補班日可能造成「非週期性」的通勤與活動模式，本研究亦納入 make_up_workday 作為外部日曆特徵，以提升模型對特殊日程之辨識能力（時間特徵樣式可參考表 3-1）。

doi:10.6342/NTU202600877

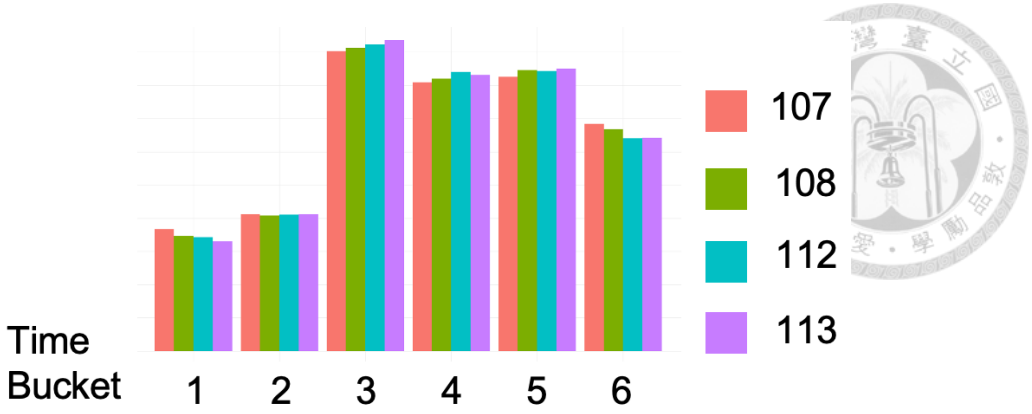

圖 3-4 臺北市四年各時段的緊急救護需求統計圖

表 3-1 2018 年時間特徵之表格樣式

Year	Month	Day	Day_of_Week	Time_Bucket	Rush_Hour	Weekend	Holiday	make_up_workday	index
2018	1	1	1	1	0	0	1	0	2018010101
2018	1	1	1	2	0	0	1	0	2018010102
2018	1	1	1	3	1	0	1	0	2018010103
…	…	…	…	…	…	…	…	…	…	…

3.3.2 空間對位與目標變數定義（Demand Construction）

本研究將每一筆可定位之 EMS 案件依其地理座標對應至 1000 公尺網格，並賦予網格識別碼 grid_id，以完成點位事件之網格化。為避免案件點位落於網格邊界造成重複計算或歸屬歧義，採用閉開區間之邊界規則：左邊界 ≤ 點 < 右邊界、下邊界 ≤ 點 < 上邊界，使每一點位能唯一對應至單一網格。目標變數定義為需求量（demand）：在每一個 Time_Bucket 與 grid_id 之組合下，累計該時段內網格所發生之案件總數。藉此將離散點位事件轉換為結構化之時空面板資料（統計表範例可參考表 3-2）。

表 3-2 2018 年網格與 EMS 需求數量統計表

index	grid_271	grid_300	grid_301	grid_188	grid_302	…
2018010101	6	3	0	0	0	…
2018010102	0	0	1	0	0	…
2018010103	2	2	3	0	0	…
…	…	…	…	…	…	…

doi:10.6342/NTU202600877

3.3.3 氣象特徵（Meteorological Features）

本研究納入降雨量與氣溫作為影響 EMS 需求之外部環境因子。由於氣象資料之空間形態與模型網格需求不一致（測站點資料對應至網格面資料），需經空間化處理後方可整合至時空面板。

3.3.3.1 降雨量：克利金法空間內插（Kriging）

降雨量資料來自 41 處離散測站，且部分測站存在觀測缺漏，無法直接對應至 334 個網格之面資料需求；因此本研究採用空間內插將點位觀測轉換為網格尺度之連續降雨特徵。經比較反距離加權法（Inverse Distance Weighting, IDW）與克利金法（Kriging）後，考量 IDW 在測站外圍易產生不自然之數值驟降，而 Kriging 可在統計模型框架下生成較平滑且符合空間連續性之預測曲面，故選擇 Kriging 作為主要內插方法，比較圖如圖 3-5。Kriging 奠基於空間自相關假設，透過變異圖（semivariogram）量化距離與屬性差異之關係，並以數學模型（如球狀或指數模型）擬合經驗半變異圖，以估計空間結構參數；其中變程（range）描述自相關趨近於零之距離尺度，基台（sill）對應變異穩定後之上限，而塊金效應（nugget）反映量測誤差與微尺度變異。完成變異圖建模後，Kriging 依據擬合模型計算最佳權重，對未知位置進行加權估計，並可提供預測不確定性之量化資訊。

實務上，本研究因鄰近區域歷史降雨資料不完整，採用臺北市測站進行內插，並於研究區外設置 4 公里緩衝區以降低邊界效應並填補測站稀疏之郊山區域，如圖 3-6。內插作業依月份獨立進行，以因應測站月別缺測情形；此外，對於可能因邊界外推或模型設定導致之多組估計情境，採用一致化整合策略（如取算術平均）以維持網格降雨特徵之可比性，經處理後實際使用之數據如表 3-3。為檢核外推策略對研究核心區之影響，本研究並檢視研究期間從未發生 EMS 案件之網格分布，如圖 3-7，結果顯示其多位於市郊山區與緩衝範圍，顯示外推對與案件發生地相關之特徵影響有限，具可行性。

doi:10.6342/NTU202600877

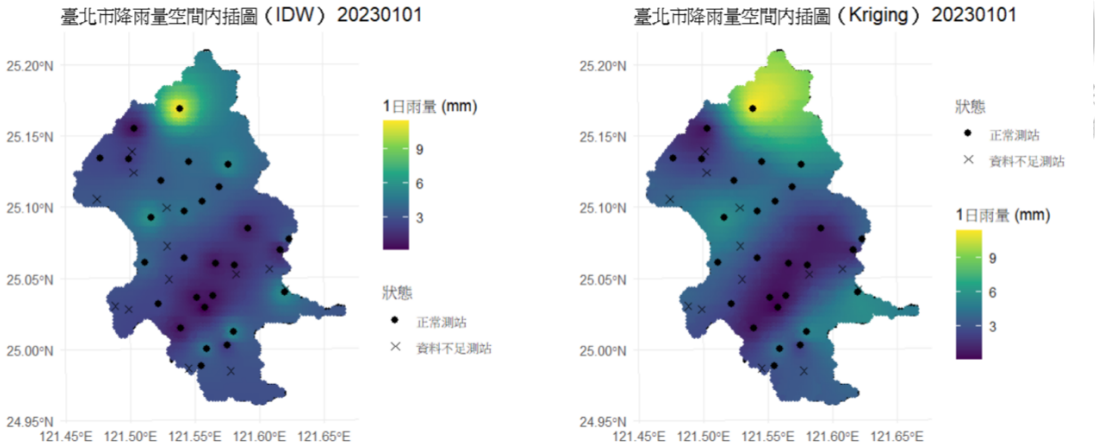

圖 3-5 反距離加權法與克利金法之內差比較圖

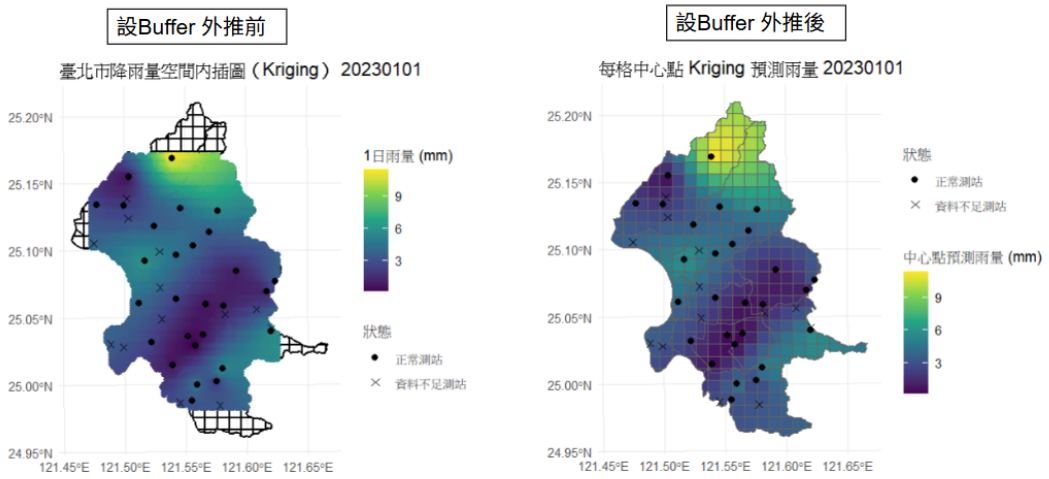

圖 3-6 克利金法設定四公里緩衝區外推前後比較圖

doi:10.6342/NTU202600877

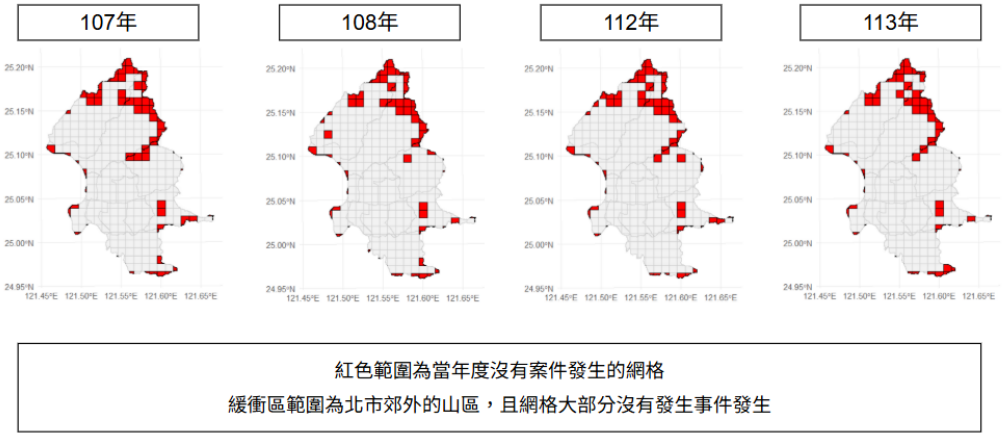

圖 3-7 研究期間從未發生 EMS 案件之網格分布

表 3-3 2018 年經克利金法處理之降雨量數據表

index	grid_9	grid_10	grid_11	grid_12	grid_13	…
20180101	0.1	0.1	0	0	0	…
20180102	0.05	0.05	0	0	0	…
20180103	0	0	0	0	0	…
…	…	…	…	…	…	…

3.3.3.2 氣溫：臺北測站代表值（Taipei Station Proxy）

除降雨量外，本研究另納入氣溫作為環境條件之重要解釋變數，以捕捉溫度變化可能對緊急救護需求造成之影響。氣溫資料來源採中央氣象署（CWA）之氣象觀測資料，下載研究年度（民國 107、108、112、113 年）「臺北測站」之每日氣溫觀測值（測站編號：422920），資料單位為攝氏溫度（°C），並以 CSV 格式取得。資料前處理包含日期欄位標準化與數值欄位型別轉換，以確保各年度資料可一致串接並與案件資料正確對齊。

在特徵對齊與合併方面，本研究依據 EMS 案件資料之索引（index）進行資料整合，採用左連接（left join）方式將每日氣溫併入案件資料，以維持所有 EMS 觀測紀錄之完整性並避免因氣象資料缺漏而造成樣本流失。由於本研究之案件資

doi:10.6342/NTU202600877

料最終係以時間區段（Time_Bucket）進行彙整，故同一日期內不同時間區段之觀測值將對應相同之日尺度氣溫值；亦即，氣溫特徵在日內不隨 Time_Bucket 變動，而作為反映當日整體熱環境條件之共同外生特徵，合併後之表格範例如表 3-4。

考量臺北市都市環境具高度複雜性，且市郊周邊多為山區地形、溫度垂直差異與局地變異明顯，若採多測站內插可能增加資料處理與模型解釋之複雜度。本研究基於方法一致性與分析簡化之需求，採用臺北測站（422920）之日氣溫作為研究區域之代表值，作為各網格共同之外生氣溫特徵。此作法可在確保資料可得性與時序連續性之前提下，提供穩定且可重現之溫度指標；惟亦需注意其隱含以單點測站近似整體都市區域熱環境之假設，可能低估市郊與高程差異所造成之空間溫度異質性。

表 3-4 2018 年時間特徵結合溫度之表格

Year	Month	Day	Day_of_Week	Time_Bucket	Rush_Hour	Weekend	Holiday	make_up_workday	TX01	index
2018	1	1	1	1	0	0	1	0	17.3	2018010101
2018	1	1	1	2	0	0	1	0	17.3	2018010102
2018	1	1	1	3	1	0	1	0	17.3	2018010103
…	…	…	…	…	…	…	…	…	…	…

3.3.4 人口統計特徵：村里人口網格化（Demographic Gridding）

本研究納入人口統計資料作為影響緊急救護需求之結構性背景因子，包含各月份村里層級之總人口與 65 歲以上人口。惟原始人口資料以村里行政區為統計單元，其面積尺度與形狀存在顯著差異，若直接以村里數值對應至網格，可能因村里合併統計或村里被切分至多個網格而導致單一網格人口被高估或低估，進而影響模型對區域需求基數的學習與解釋。此一尺度不一致問題，亦可由網格化前後之人口空間分布對照觀察：不同村里面積大小將造成網格化後人口量在局部區域出現相對增加或相對減少之現象，如圖 3-8。

doi:10.6342/NTU202600877

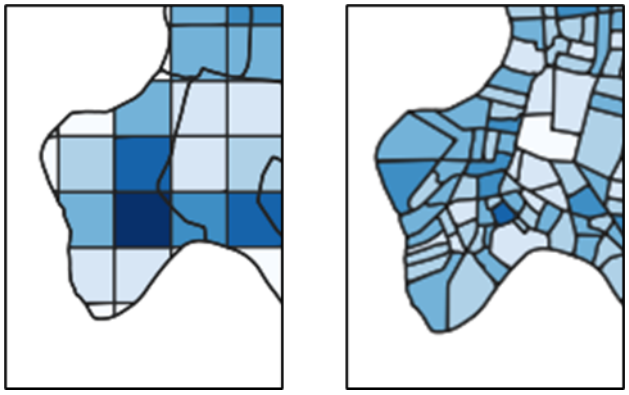

圖 3-8 臺北市部分村里與網格化人口面量比較圖

為使人口特徵可與本研究之 1000 公尺×1000 公尺網格分析單元一致，本研究採用「面積加權配置」（areal-weighted apportionment）的網格化策略，將村里人口由行政區單元轉換為網格單元。其處理流程如下：第一步，計算各月份村里人口密度（人口數除以村里面積），以近似表示該村里人口於其範圍內之平均分布強度；第二步，透過村里邊界與網格邊界之疊合（overlay），計算每一網格內所涵蓋之各村里面積；第三步，以「人口密度 × 對應涵蓋面積」估算該村里在特定網格中的人口量，並將同一網格內所有村里之估計值加總，得到該月份之網格人口（總人口與 65 歲以上人口分別計算），網格化之前後比較圖如圖 3-9。

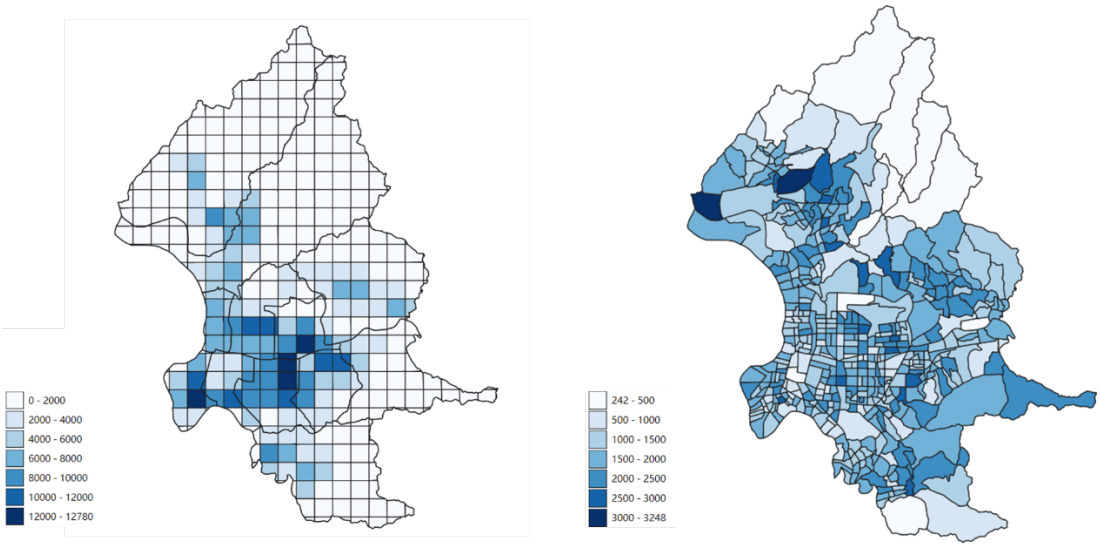

圖 3-9 臺北市 112 年 1 月村里 65 歲以上人口網格化前後比較圖

上述方法可形式化表示為：對於月份 t與村里 v，令 $P_{v,t}$為人口數、A_v 為村里面積、$d_{v,t} = P_{v,t}/A_v$ 為人口密度；對於網格 g，令 $A_{v\cap g}$為村里 v與網格 g

doi:10.6342/NTU202600877

之交集面積，則網格人口估計為

$$P_{g,t} = \sum_{v} d_{v,t}\ A_{v \cap g}.$$

同理可得高齡人口 $E_{g,t}$，經處理後的資料表如表 3-5 所示。此流程在操作上亦具備可重現性與一致性：交集面積 $A_{v \cap g}$僅由行政區與網格邊界決定，可預先計算並於各月份重複套用；而面積加權配置在研究區域內具備總量守恆特性（各網格加總可回復至村里總量），有助於降低尺度轉換造成之系統性偏差。需注意的是，該方法隱含「人口於村里內均勻分布」之假設，故其目的在於提供與網格尺度一致、可供模型學習之近似人口基數特徵，而非取代更細緻的人口微觀分布推估。

表 3-5 2018 年各網格人口統計表

index	grid_9	grid_10	grid_11	grid_12	grid_13	…
2018010101	182.91139	2419.47457	3811.67172	1460.23726	14.5817011	…
2018010102	182.91139	2419.47457	3811.67172	1460.23726	14.5817011	…
2018010103	182.91139	2419.47457	3811.67172	1460.23726	14.5817011	…
…	…	…	…	…	…	…

3.3.5 資料一致性與缺失值處理（Consistency & Missing Values）

為維持時間序列之完整性，對於特定 Time_Bucket 下某網格未發生案件之情形，需求量 demand 以 0 填補，使每一網格於每一時段皆具有連續觀測，避免因缺值造成序列斷裂。類別型二元特徵（如 Holiday、Weekend、make_up_workday、Rush_Hour）統一轉為 0/1 數值表示，以利模型訓練。另因需求量屬計數資料且理論上不為負，模型輸出與後處理亦採用非負限制（例如將負值截斷為 0）以符合問題定義。

3.4 模型輸入資料產製

本研究之模型訓練採用多序列全球模型（global model）設定：每一個 grid_id 對應一條時間序列，模型在共享參數下同時學習 334 條序列之共同結構與各網格

doi:10.6342/NTU202600877

差異。本節說明資料如何由長格式面板表轉換為模型可直接訓練之序列資料結構，並定義訓練/測試切分策略。

3.4.1 長格式時空面板資料（Long-format Panel）

完成特徵建構後，本研究以 grid_id 與 datetime 作為主索引，將需求量 demand、時間特徵、降雨量、氣溫與人口統計等資料進行時空對齊與整併，形成長格式面板資料，如表 3-6：每一列代表某網格在某一 Time_Bucket 之觀測，包含目標變數與對應之動態/靜態共變數。此長格式資料主要用於檢核資料完整性、進行特徵合併與後續彙整分析。

3.4.2 序列化輸入格式（GluonTS ListDataset）

在模型訓練階段，本研究將長格式面板資料依 grid_id 分組，排序後轉換為序列化輸入。對每一網格建立一筆資料項（item），包含：（1）target：需求量序列（demand）；（2）start：序列起始時間（以 4 小時為頻率）；（3）feat_dynamic_real：動態數值共變數序列，用於提供隨時間變動之外部資訊；（4）feat_static_real：靜態數值特徵（例如 population、population_65），用於提供跨期間相對穩定之區域背景差異。針對 TFT 模型，動態共變數概念上可區分為可預先得知之未來特徵與僅於過去可觀測之特徵；惟本研究離線比較採完美預報假設，故除日曆與時間結構特徵外，降雨量與氣溫亦視為可取得之未來外生資訊，併入 feat_dynamic_real，以在推論時提供預測視野內對應時間窗之天氣序列；實務部署時則以天氣預報或情境模擬序列取代觀測值。若未來改採較保守資訊情境（無可靠預報），亦可將天氣特徵改置於 past_feat_dynamic_real 僅提供歷史觀測並另行評估其影響。

3.4.3 訓練/測試切分策略（Per-series Hold-out）

為評估模型在未來時段之預測能力，本研究採用各序列末端保留（hold-out）策略：對每一網格序列，保留最後 prediction_length 個時間步作為測試區間，其餘作為訓練區間。本研究以 4 小時為頻率，設定 prediction_length＝6，對應「預

doi:10.6342/NTU202600877

測未來一天（24 小時）」；並設定 context_length＝42，對應「使用過去七天（7×24/4＝42 步）」作為主要歷史脈絡。若個別網格序列長度過短導致訓練段不足，則以比例切分方式確保測試段至少保留 prediction_length 步，以維持訓練與評估之可行性。

doi:10.6342/NTU202600877

表 3-6 輸入模型之長資料示意表

grid_id	datetime	index	Year	Month	Day	Day_of_Week	Time_Bucket	Rush_Hour	Weekend	Holiday	make_up_workday	demand	population	population_65	rainfall	TX01
grid_10	2018/1/1 0:00	2018010101	2018	1	1	1	1	0	0	1	0	0	2419.47457	392.20142	0.1	17.3
grid_10	2018/1/1 4:00	2018010102	2018	1	1	1	2	0	0	1	0	0	2419.47457	392.20142	0.1	17.3
grid_10	2018/1/1 8:00	2018010103	2018	1	1	1	3	1	0	1	0	1	2419.47457	392.20142	0.1	17.3
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
grid_10	2024/12/31 20:00	2024123106	2024	12	31	2	6	0	0	0	0	0	2166.29557	468.922024	0	16.4
grid_103	2018/1/1 0:00	2018010101	2018	1	1	1	1	0	0	1	0	0	4948.09886	961.374897	0	17.3
grid_103	2018/1/1 4:00	2018010102	2018	1	1	1	2	0	0	1	0	0	4948.09886	961.374897	0	17.3
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
grid_253	2024/12/31 20:00	2024123106	2024	12	31	2	6	0	0	0	0	0	14.8771391	4.28281276	0	16.4

doi:10.6342/NTU202600877

3.5 預測模型與訓練設定（Modeling & Training）

本研究採用 DeepAR 與 TFT 兩類機率式深度學習模型。兩模型皆可在多序列情境下共享參數，適用於跨網格之需求預測；並能輸出分位數（quantile）或預測分佈，以支援不確定性評估。本節說明模型架構、核心超參數與訓練流程設定。

為確保兩模型比較之資訊使用一致性，本研究於離線評估採完美預報假設：推論視野內之天氣變數視為可事先取得的未來外生資訊（future-known covariates），因此在推論時對 DeepAR 與 TFT 均提供預測視野內對應時間窗之天氣序列；離線評估以實際觀測值替代完美預報以進行方法驗證，實務部署時則可改以天氣預報或情境模擬輸入取代觀測值。

3.5.1 DeepAR：深度自回歸循環網絡

DeepAR（Deep Autoregressive Recurrent Network）屬機率式自回歸循環神經網絡模型，透過 RNN（如 LSTM/GRU）學習需求量之序列依賴，並以分佈參數化方式輸出預測分佈。本研究將需求量視為計數資料，採用負二項分佈（Negative Binomial）作為輸出分佈，以同時刻畫平均與離散程度。模型訓練以 prediction_length＝6、context_length＝42 為核心設定，並使用多組延遲項（lags_seq＝[1, 2, 3, 4, 6, 8, 12, 18, 24]）以捕捉短期與日內週期型依賴。訓練階段採用批次大小 batch_size＝64、學習率 learning_rate＝1e-3、最大訓練回合 max_epochs＝30，並以 num_parallel_samples＝1000 於推論時進行平行抽樣以穩定估計分位數與預測區間。由於 DeepAR 之 feat_dynamic_real 在推論視野中需提供對應特徵，本研究在離線評估階段以實際觀測之天氣特徵補齊預測視野，以評估在可取得天氣預報（或以情境模擬方式提供外部天氣輸入）下之需求預測能力；實務部署時可改以天氣預報或外部情境輸入取代觀測值，以避免資訊洩漏。

3.5.2 TFT：時間融合轉換器

Temporal Fusion Transformer（TFT）為多視野時間序列預測架構，結合序列

doi:10.6342/NTU202600877

表徵、門控機制與自注意力（self-attention），可同時處理靜態特徵、過去可觀測特徵與可預先得知之未來特徵，並以注意力權重提供一定程度可解釋性。本研究之 TFT 設定 prediction_length＝6、context_length＝42，並將時間結構特徵（如 Year、Month、Day、Day_of_Week、Time_Bucket、Holiday、Weekend、make_up_workday、Rush_Hour）作為可預先得知之動態特徵（feat_dynamic_real），並將降雨量與氣溫等天氣變數亦納入可預先得知之動態特徵（feat_dynamic_real），使模型在推論時得以使用預測視野內之天氣序列；離線評估以實際觀測值填補預測視野以對應完美預報假設，實務部署時則以天氣預報或情境模擬序列取代觀測值，以維持資訊使用一致性並避免資訊洩漏。模型訓練設定包含隱藏維度 hidden_dim＝32、注意力頭數 num_heads＝4、dropout＝0.1、批次大小 batch_size＝64、學習率 learning_rate＝1e-3 與最大回合 max_epochs＝50，以提升對非線性與多尺度動態之擬合能力。

3.5.3 訓練流程與輸出後處理

兩模型訓練皆以固定隨機種子進行，以提升可重現性。推論階段輸出多分位數預測（例如 0.05、0.1、0.25、0.5、0.75、0.9、0.95），並以中位數（q0.5）作為點預測代表值。考量需求量不為負，模型輸出於後處理階段對負值採截斷為 0 之限制，以符合計數需求之定義。為利於後續分析與繪圖，本研究亦將各網格在測試視野之分位數預測與真實值整併輸出為 CSV 檔，作為第四章實驗結果呈現與比較之資料基礎。

3.6 模型評估指標（Evaluation Metrics）

為客觀且量化地評估模型之預測表現，本研究自「點預測準確性」與「機率式預測可靠性」兩面向建立評估架構。設測試集中共有 N筆評估樣本，真實需求量為 y_i，點預測值為 $\hat{y}_i$（本研究以中位數分位數 $q_{0.5,i}$作為點預測），預測分位數為 $q_{\tau,i}$，其中 $\tau \in (0,1)$。指示函數 $\mathbb{I}(\cdot)$於條件成立時取 1，否則取 0。

doi:10.6342/NTU202600877

3.6.1 點預測準確性：MAE 與 RMSE

平均絕對誤差（Mean Absolute Error, MAE）定義為

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} | \hat{y}_i - y_i |$$

均方根誤差（Root Mean Squared Error, RMSE）定義為

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

其中 $\hat{y}_i = q_{0.5,i}$為模型於樣本 i的中位數預測值。

3.6.2 預測區間覆蓋率（Coverage / PICP）

由於本研究採用機率式預測模型，除點預測誤差外，亦需評估其不確定性刻畫之可靠性。本研究以 **80% 中央預測區間**作為區間評估基準，亦即以第 0.1 與第 0.9 分位數形成預測區間：

$$[q_{0.1,i},\ q_{0.9,i}]$$

預測區間覆蓋率（Prediction Interval Coverage Probability, PICP）定義為真實觀測值落入預測區間之比例：

$$\text{PICP}_{80} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(y_i \in [q_{0.1,i},\ q_{0.9,i}]\right)$$

當模型之不確定性估計具良好校準（calibration）時，PICP_{80} 應接近 0.8。

3.6.3 容忍誤差率：Tolerant Rate Error（TRE）

為貼近 EMS 資源規劃之決策情境，本研究另以容忍誤差帶衡量點預測之可操作性。令容忍誤差閾值為 τ（例如 $\tau = 1$件），則 命中率（Hit Rate）定義為

$$\text{Hit}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(| \hat{y}_i - y_i | \leq \tau\right)$$

進一步定義容忍誤差率（Tolerant Rate Error, TRE）為未命中比例：

doi:10.6342/NTU202600877

$$\mathrm{TRE}(\tau) = 1 - \mathrm{Hit}(\tau)$$

其中 $\hat{y}_i = q_{0.5,i}$為樣本 i的點預測值。

3.7 置換法特徵重要性（Permutation Feature Importance）

為提升深度學習時空序列模型於實務應用情境中的可解釋性，本研究於模型訓練完成後，進一步採用置換法特徵重要性（Permutation Feature Importance, PFI）評估各輸入特徵對預測效能之相對貢獻。PFI 屬於模型不可知（model-agnostic）之解釋方法，其核心概念為：若某一特徵對模型預測具有關鍵影響，當該特徵之資訊被隨機打亂、使其與目標變數之關聯遭破壞時，模型在測試集上的預測誤差將明顯上升。相反地，若置換後誤差變化有限，則可推論該特徵對模型的邊際貢獻相對較小。透過此方法，本研究得以以一致且可比較的方式，量化不同類型特徵（時間結構、氣象因子、人口結構等）對需求預測的影響程度，作為後續結果詮釋與特徵設計合理性檢核之依據。

3.7.1 方法定義與重要性量化指標

本研究以測試資料集上的平均絕對誤差（MAE）作為 PFI 的衡量基準，並以「置換前後 MAE 差值」定義特徵重要性。令測試集共有 N個評估樣本（可視為所有網格在預測視野內的時間步集合），真實值為 y_i，點預測值為 $\hat{y}_i$。本研究於推論時採用模型的中位數分位數 $q_{0.5,i}$作為點預測，以避免以平均值作為點估計時可能出現的警告或不穩定情形，並將預測值限制為非負以符合案件量之定義，即 $\hat{y}_i = \max(0, q_{0.5,i})$。則基準 MAE 可表示為：

$$\mathrm{MAE}_{\mathrm{base}} = \frac{1}{N}\sum_{i=1}^{N} | \hat{y}_i - y_i |.$$

對任一特徵 X_j進行置換後得到置換資料集，重新推論並計算置換後 MAE：

$$\mathrm{MAE}_{\mathrm{perm}}(j) = \frac{1}{N}\sum_{i=1}^{N} | \hat{y}_i^{(j)} - y_i |.$$

本研究定義特徵 X_j的置換重要性為：

doi:10.6342/NTU202600877

$$I_j = \Delta\text{MAE}_j = \text{MAE}_{\text{perm}}(j) - \text{MAE}_{\text{base}}.$$

其中 ΔMAE_j越大，表示置換該特徵後造成誤差上升越顯著，該特徵對模型預測之貢獻亦越重要。

3.7.2 評估資料與推論設定

PFI 之評估採用與模型效能驗證一致之**測試資料集（test set）**，其切分方式為各網格序列末端保留（hold-out）：每一序列保留最後 $prediction_length = 6$個時間步作為測試預測視野，對應未來一天（24 小時）之 6 個 4 小時時段；並以 $context_length = 42$代表主要歷史脈絡（7 日）。在推論設定上，本研究以固定隨機種子確保流程可重現，並採用與模型評估一致的抽樣數 $num_samples = 1000$產生預測分佈，再由其中位數分位數 $q_{0.5}$取得點預測以計算 MAE。上述設計確保 PFI 的誤差差異主要反映特徵資訊被破壞所造成的影響，而非來自資料切分或推論設定不一致所導致的偏差。

3.7.3 置換策略與特徵類型處理

由於本研究同時包含動態特徵（隨時間變動）與靜態特徵（相對穩定），且 TFT 模型另區分「未來可得之已知特徵」與「僅過去可觀測之未知特徵」，因此置換策略依特徵類型進行差異化設計，以符合模型輸入結構並避免不當資訊使用。

（一）DeepAR 模型之置換設計

DeepAR 之輸入包含動態實數特徵（feat_dynamic_real）與靜態實數特徵（feat_static_real）。本研究對動態特徵（包含降雨量、氣溫以及時間結構與日曆類特徵等）採用「序列內時間置換」：對每一網格序列，針對單一特徵之時間序列進行隨機打亂（random permutation），以破壞其與需求量在時間上的對應關係，但保留該特徵在該序列內的邊際分佈；其餘特徵與目標序列保持不變。對靜態特徵（例如人口與高齡人口）則採用「跨序列置換」：在不同網格之間隨機交換該特徵值，以破壞其與特定網格需求基數之關聯，同時維持整體樣本分佈一致。每完成一項特徵之置換，即建立對應之置換測試資料集並重新推論計算

doi:10.6342/NTU202600877

$MAE_{perm}(j)$，最後得到各特徵之 ΔMAE_j排序。

（二）TFT 模型之置換設計

本研究離線評估採完美預報假設，因此 TFT 之動態特徵中，日曆／時間結構特徵與天氣變數（降雨量、氣溫）皆視為可預先取得之未來外生資訊，並以 feat_dynamic_real 形式提供推論視野內對應序列。基於此設定，本研究之置換流程分為兩類：

1. **動態特徵置換（feat_dynamic_real）**：於各網格序列內，針對單一特徵之時間序列進行隨機打亂（random permutation），以破壞其與需求量之時間對應關係，但保留該特徵在該序列內的邊際分佈；為利於解讀，動態特徵可再區分為（a）時間日曆類（Time_Bucket、週末/假日、補班、尖峰等）與（b）天氣類（降雨量、氣溫）兩群，分別檢視其敏感度。
2. **靜態特徵置換（feat_static_real）**：與 DeepAR 一致，對人口與高齡人口等背景因子採跨序列置換（跨網格隨機交換），以破壞其與網格需求基數之關聯並維持整體分佈一致。

上述置換後皆以相同推論設定重新計算 MAE，並以 ΔMAE 作為重要性度量。另需說明：若未來改採無可靠預報之保守情境，天氣特徵可改視為 unknown（僅過去可觀測），並以 past_feat_dynamic_real 表示，此時置換設計亦可依 unknown 特徵另行定義。

3.7.4 輸出與呈現方式

本研究將 PFI 結果彙整為表格，包含特徵類型（dynamic：時間日曆類／天氣類；static）、置換後 MAE 與 ΔMAE等欄位，並輸出為 CSV 檔以利後續視覺化與章節呈現。同時，為直觀比較各特徵之影響程度，本研究依 ΔMAE由大到小繪製水平長條圖，作為第四章結果討論中特徵貢獻與模型行為之實證依據。需注意的是，置換法之重要性估計可能受單次隨機置換之抽樣波動影響；因此在後續延伸研究中，可進一步對每項特徵進行多次置換並以平均與標準差呈現，以提升重

doi:10.6342/NTU202600877

要性排序的穩健性與可重複性。

doi:10.6342/NTU202600877

第四章 實驗結果與討論

本章旨在基於第三章所建立之研究方法，系統性呈現並比較深度自回歸循環網絡（Deep Autoregressive Recurrent Network, DeepAR）與時間融合轉換器（Temporal Fusion Transformer, TFT）兩種深度學習模型，在臺北市緊急救護服務（Emergency Medical Services, EMS）需求預測任務上的表現。為全面剖析各模型的性能特性與適用情境，本研究將從「點預測準確性」、「機率式預測可靠性」與「實務應用效益」三大面向進行評估，並同時涵蓋 **1 日**與 **7 日**兩種預測視野，以檢視模型在不同時間尺度下的穩健性與差異。

4.1 實驗設計總覽與評估框架

4.1.1 資料切分策略與預測視野設定

本研究以時間序列之因果性為前提進行評估，資料期間涵蓋民國 107、108、112、113 年之 EMS 派遣紀錄（排除 COVID-19 顯著干擾年度），空間單元採 1,000 公尺 ×1,000 公尺網格，時間解析度為每 4 小時一時段（Time_Bucket）之案件計數。

與一般固定比例切分（例如 80%/20%）不同，本研究採用以**資料末端預測視野作為留出測試窗（hold-out forecast window）**的方式，以更貼近實務用目前所有歷史資料去預測最近未來一段時間的情境。具體作法如下：

- 1 日預測（6 個 4 小時時段）：將資料最後需預測的 6 個時段保留作為測試窗；其餘所有更早的時段資料全部用於訓練。模型完成訓練後，直接輸出該測試窗的 1 日需求預測結果並計算各項指標。
- 7 日預測（42 個 4 小時時段）：將資料最後需預測的 42 個時段保留作為測試窗；其餘所有更早的時段資料全部用於訓練。模型完成訓練後，輸出該測試窗的 7 日需求預測結果並計算各項指標。

此策略確保模型在訓練階段僅使用測試窗開始之前的歷史資訊，避免未來資

doi:10.6342/NTU202600877

訊洩漏，同時也使評估結果能直接對應到臨近未來（1 日/7 日）需求預測的實務使用情境。

需要注意的是，1 日與 7 日預測在本研究中屬於兩個不同的預測任務定義：兩者測試窗長度不同（6 步 vs 42 步），且在高度零膨脹的資料下，較長視野的測試窗可能涵蓋較多「零需求」狀態，造成 MAE/RMSE 在跨視野比較時出現不可直接對照的現象。因此，本研究之主要結論以「同一預測視野內的模型比較」為主；跨視野之指標差異僅作為現象描述，後續可透過 rolling forecast（多個末端測試窗）或依月份／季節分段的穩健性檢驗，確認該現象是否具普遍性並釐清其成因。

4.1.2 評估指標與三維度評估架構

為建立一致且可比較的評估基準，本研究採用四項核心指標，分別對應點預測、機率式預測與實務容錯效益三個面向，如表 4-1 所示。所有指標皆於各預測視野對應的「留出測試窗」上計算。

表 4-1 本研究核心評估指標定義與評估面向

指標	全名	評估維度	衡量目標
MAE	Mean Absolute Error	點預測準確性	預測值與真實值之平均絕對誤差
RMSE	Root Mean Squared Error	點預測準確性	對較大誤差賦予更高權重之離散程度衡量
PICP@80%	Prediction Interval Coverage Probability	機率式預測可靠性	真實值落入 80% 預測區間的比例（理想值接近名目水準）
TRE（τ=±1）	Tolerant Rate Error（±1 case）	實務應用效益	點預測誤差超出 ±1 件容忍範圍之比例（越低越好）

4.2 整體效能綜合比較（Overall Performance Comparison）

本節提供高層次之整體視圖，直接比較 DeepAR 與 TFT 在 1 日與 7 日預測視野下的核心指標表現，以建立後續深入分析之基礎。表 4-2 彙整兩模型於四種情境下之 MAE、RMSE、PICP@80% 與 TRE（τ=±1）結果，其中粗體為該

doi:10.6342/NTU202600877

預測時長下之最佳表現。

表 4-2 兩模型於不同預測時長下之整體效能比較

模型	預測時長	MAE	RMSE	PICP@80%	TRE
TFT	1 日	0.1735	0.5382	76.65%	4.89%
DeepAR	1 日	0.1742	0.5700	97.16%	2.89%
TFT	7 日	**0.1634**	**0.4895**	72.92%	5.25%
DeepAR	7 日	0.1713	0.5245	**97.35%**	**2.77%**

由表 4-2 可歸納三點主要趨勢。第一，在點預測準確性方面，TFT 整體略優於 DeepAR，且在 7 日視野之優勢更為明顯（MAE 與 RMSE 均為四情境最低），顯示其在較長預測視野下捕捉趨勢與週期性之能力相對較佳。惟不同預測視野之指標不宜直接以絕對值互比，其比較基準以同一視野內模型差異為主。

第二，在機率式預測可靠性方面，兩模型呈現顯著差異：DeepAR 之 PICP@80% 遠高於 80% 目標值，顯示其預測區間偏向保守；相對地，TFT 之 PICP@80% 略低於目標值，反映其區間較窄且可能低估尾端風險。第三，從更貼近調度決策之 TRE 指標觀之，DeepAR 在 1 日與 7 日情境皆低於 TFT，顯示其在是否落於可接受誤差帶之實務尺度上更具穩健性。基於上述發現，以下將依評估維度進一步深入分析。

4.3 點預測準確性深度分析（In-depth Analysis of Point Prediction Accuracy）

僅以整體 MAE/RMSE 進行比較，可能因資料高度稀疏而掩蓋模型在高需求事件之表現差異。本節因此從「資料稀疏性與誤差分布」及「不同案件量級之分組誤差」兩方面，剖析模型點預測能力之細節。

4.3.1 資料稀疏性與誤差分布

測試資料顯示，約 87.3%～87.5% 的觀測值（網格－時段）為零案件，反映 EMS 需求在高空間解析度與短時間尺度下具有明顯零膨脹現象（圖 4-1、圖

doi:10.6342/NTU202600877

4-2)。由於大量樣本屬「無事件發生」狀態，模型在多數時間皆需預測接近零之需求量(圖 4-3、圖 4-4)，使得整體 MAE 得以維持在約 0.16～0.17 的低水準。然而，低 MAE 並不必然代表模型對罕見之高需求事件亦具同等可靠性，因此需搭配分層分析，以評估模型在非典型情境下之誤差表現。

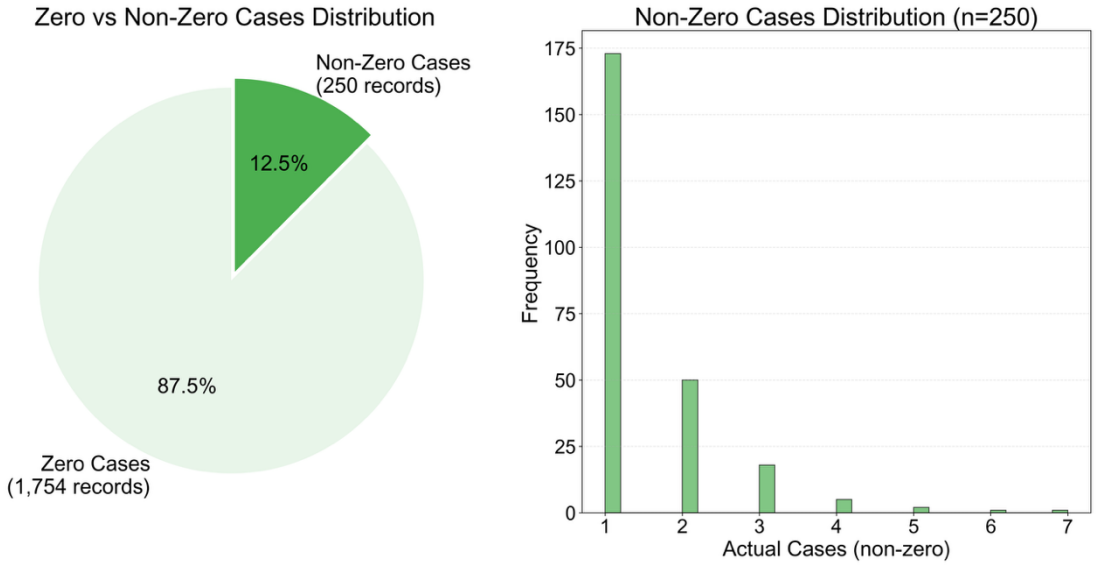

圖 4-1 預測 1 日之零案件比例與非零案件分布

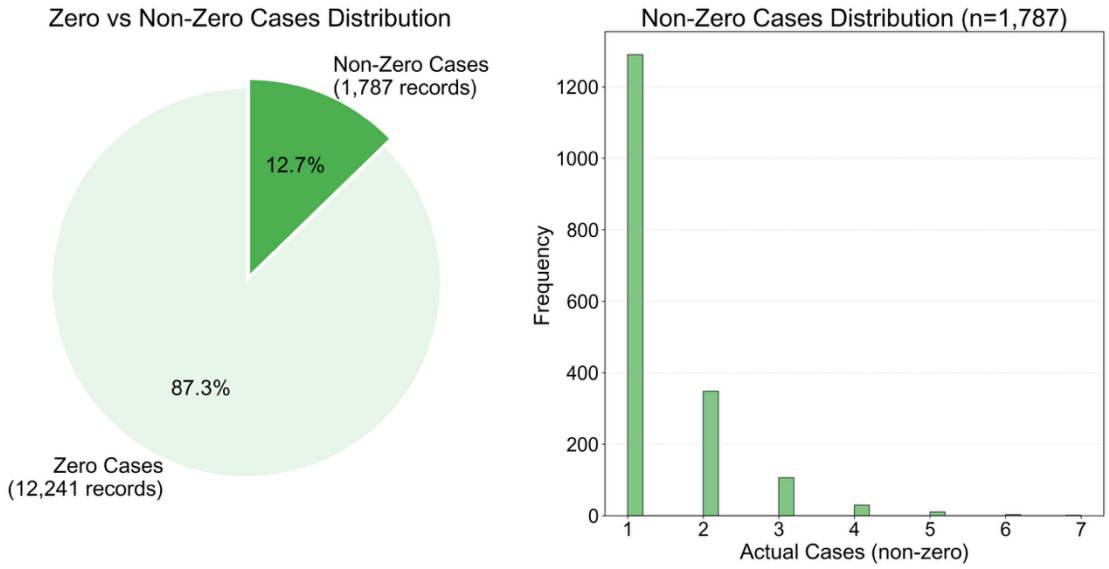

圖 4-2 預測 7 日之零案件比例與非零案件分布

doi:10.6342/NTU202600877

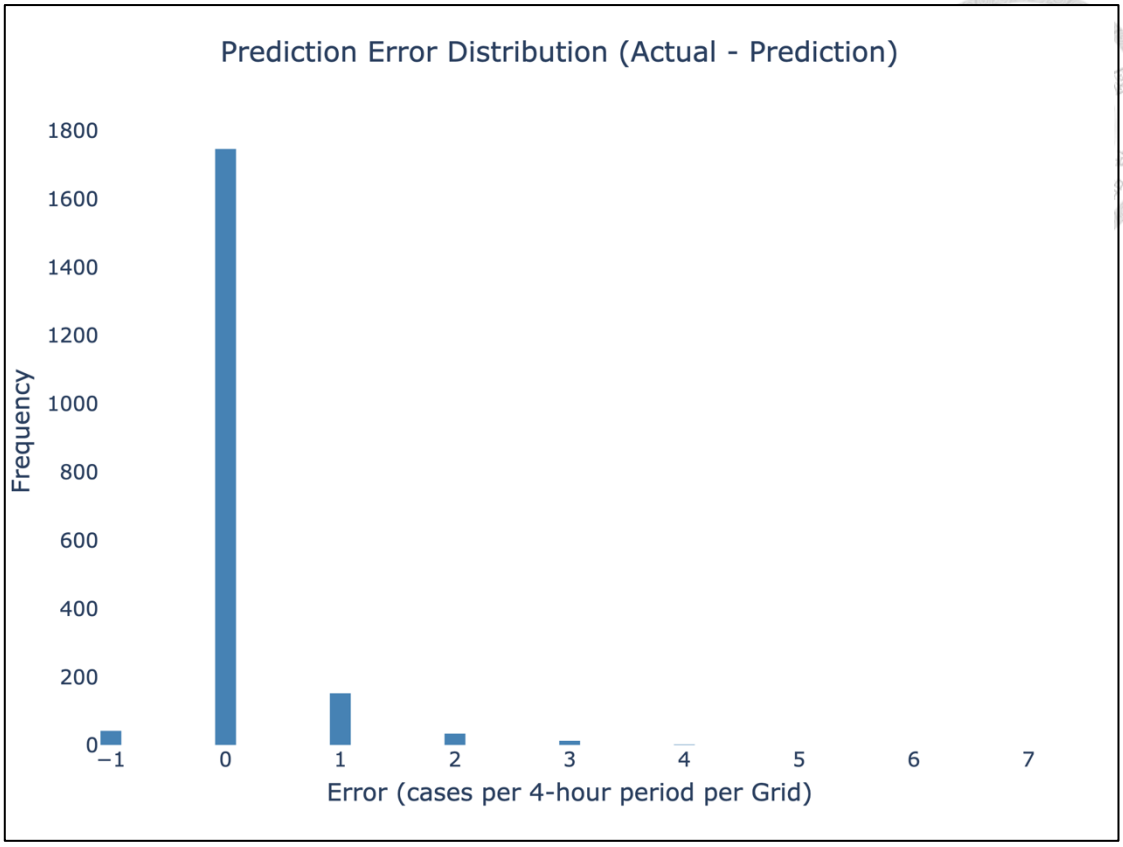

圖 4-3 DeepAR 在 1 日預測下之誤差分布直方圖

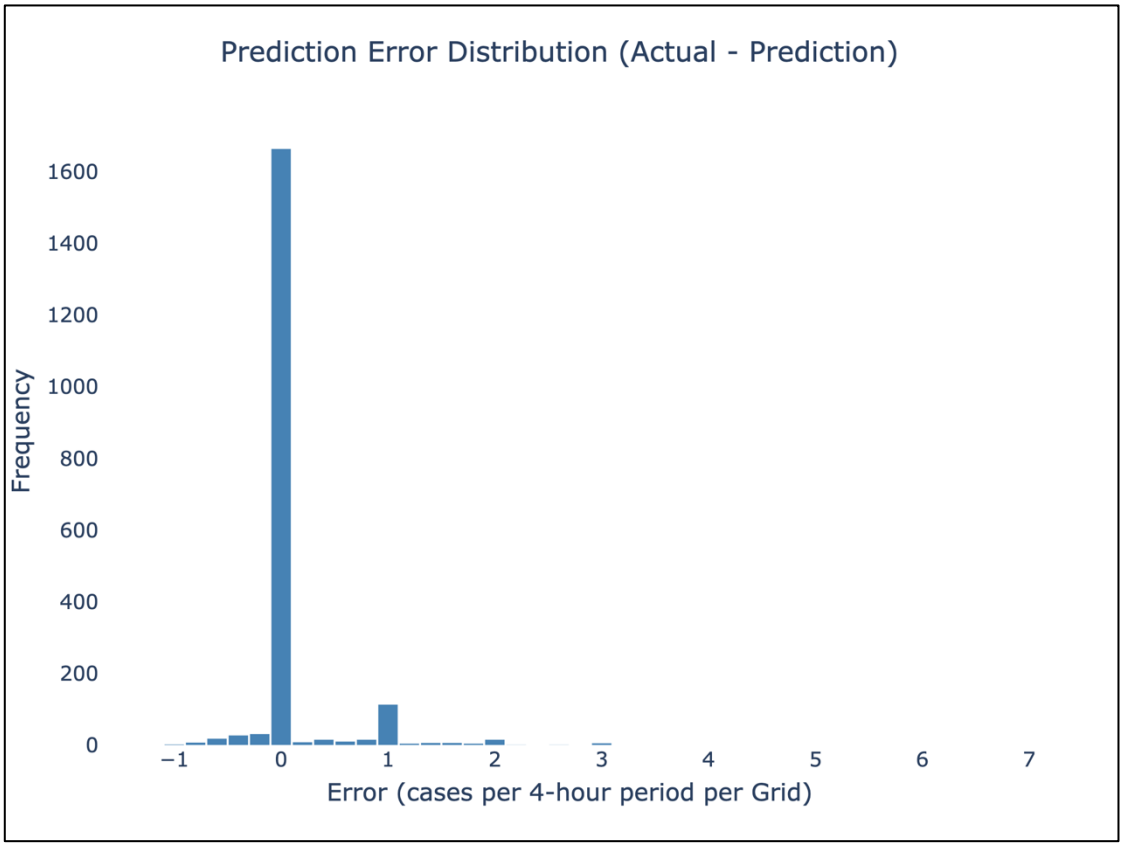

圖 4-4 TFT 在 1 日預測下之誤差分布直方圖

doi:10.6342/NTU202600877

4.3.2 不同案件量級之分組誤差比較

為更細緻評估模型在不同需求量級下之表現，本研究將測試樣本依實際案件量分組，計算各組 MAE，結果如表 4-3 。整體而言，兩模型在 0 案件情境下皆能維持極低誤差，顯示對常態低需求狀態之辨識能力良好；然而，當實際案件量上升，誤差亦迅速增加，反映高需求事件本身之稀少性與不確定性，仍是兩模型共同面臨之核心挑戰。相較之下，TFT 在中高需求群組（≥2 件）多呈現較低 MAE，且在 7 日預測視野下此優勢更為明顯，暗示 TFT 對高需求變化之擬合能力相對較佳。

表 4-3 不同案件量群組之 MAE 比較

實際案件量群組	DeepAR（1 日）MAE	TFT（1 日）MAE	DeepAR（7 日）MAE	TFT（7 日）MAE
0	**0.025**	0.027	**0.025**	0.028
1	0.780	0.800	0.826	**0.778**
2	1.620	**1.546**	1.690	1.552
3–4	2.913	2.733	2.800	**2.595**
≥5	5.500	5.393	4.571	**4.338**

4.4 機率式預測可靠性評估（Evaluation of Probabilistic Forecast Reliability）

本節聚焦評估模型機率式預測能力，即其不確定性量化之品質。對 EMS 調度而言，提供合理的需求波動範圍（例如 80% 機率落在某區間）通常比單一點預測更具決策價值，因其有助於風險管理與備援資源配置。

4.4.1 PICP@80% 偏差分析

理想情況下，80% 預測區間之 PICP@80% 應接近 80%。然而，整體結果顯示兩模型存在系統性偏差。DeepAR 之覆蓋率顯著高於目標值（1 日：97.16%；7 日：97.35%），顯示其區間過度保守、偏寬，雖提高安全性但降低資訊量；相對地，TFT 覆蓋率低於目標值（1 日：76.65%；7 日：72.92%），反映其區間偏

doi:10.6342/NTU202600877

窄且略顯自信，可能低估尾端風險。

4.4.2 預測區間寬度分析

區間寬度反映模型對需求不確定性之估計幅度。表 4-4 比較兩模型在不同預測視野下之 80% 預測區間平均寬度。結合 PICP@80% 與區間寬度可知，DeepAR 透過較保守之區間策略取得極高覆蓋率；TFT 則提供較貼近其中位數之區間，資訊量較高但覆蓋率偏低。

表 4-4 80% 預測區間平均寬度比較

預測時長	模型	平均區間寬度（80%）
1 日	DeepAR	0.511 件
1 日	TFT	0.525 件
7 日	DeepAR	0.497 件
7 日	TFT	0.500 件

4.4.3 分位數校準曲線（Quantile Calibration Curve）

為更全面檢驗模型機率輸出之可靠性，本研究除比較單一名目水準之 PICP@80% 外，進一步繪製分位數校準曲線（quantile calibration curve），用以檢查不同分位數（q05、q10、q25、q50、q75、q90、q95）在長期觀測下是否符合校準性質。具體而言，對任一分位數水準 q，若模型輸出之第 q 分位數預測為 $\hat{y}(q)$，則其觀測涵蓋率（observed coverage）定義為測試集中滿足 $y \leq \hat{y}(q)$ 的比例。理想校準情況下，觀測涵蓋率應接近名目分位數水準，即曲線應貼近 45 度對角線（Perfect Calibration）。

在解讀上，若某分位數點落在對角線上方（observed > nominal），代表該分位數預測值偏高；反之若落在對角線下方（observed < nominal），則代表該分位數預測值偏低。需注意本研究目標變數為高比例零案件之離散計數資料，低分位數端可能因大量機率集中於 0 而呈現階梯狀（例如多個低分位數對應到相同的預測值），因此本研究除觀察低分位數外，亦特別關注中高分位數與其對應之決策

doi:10.6342/NTU202600877

意涵。

DeepAR 的校準結果如圖 4-5、圖 4-6 所示，1 日與 7 日兩種視野下的曲線皆明顯偏離對角線；其中低分位數端呈現明顯階梯現象，反映在大量零需求背景下，模型於低端分位數的解析度有限。此外，於中高分位數端（q50～q95）亦普遍高於對角線，顯示 DeepAR 的分位數輸出整體偏保守，與前述 PICP@80% 顯著高於 80% 之結果一致，表示其傾向以較寬或較保守的分佈來確保覆蓋，雖可降低漏接尖峰的風險，但可能降低機率輸出的資訊密度並導致過度預備。

TFT 的校準結果如圖 4-7、圖 4-8 所示，呈現較為複雜的分位數偏差型態。以 1 日視野為例，高分位數端（特別是 q90、q95）落在對角線下方，顯示上尾分位數可能偏低，對應到需求突升時存在風險低估的可能；而在部分中位數附近分位數則可能出現偏高現象，顯示其分佈形狀並非單向偏差。至 7 日視野時，曲線整體型態改變：低至中分位數端多偏離對角線，可能導致以中央區間（例如 80% 區間）構造之下界過高或上界不足，與 TFT 之 PICP@80% 低於目標值的現象相互呼應。

綜合而言，分位數校準曲線提供了比單一 PICP@80% 更細緻的檢核視角，使本研究能辨識偏差主要集中於哪些分位數區間，並直接對應到風險導向決策中常用的高分位數門檻設定（例如以 q75/q90 作為預備資源依據）。

doi:10.6342/NTU202600877

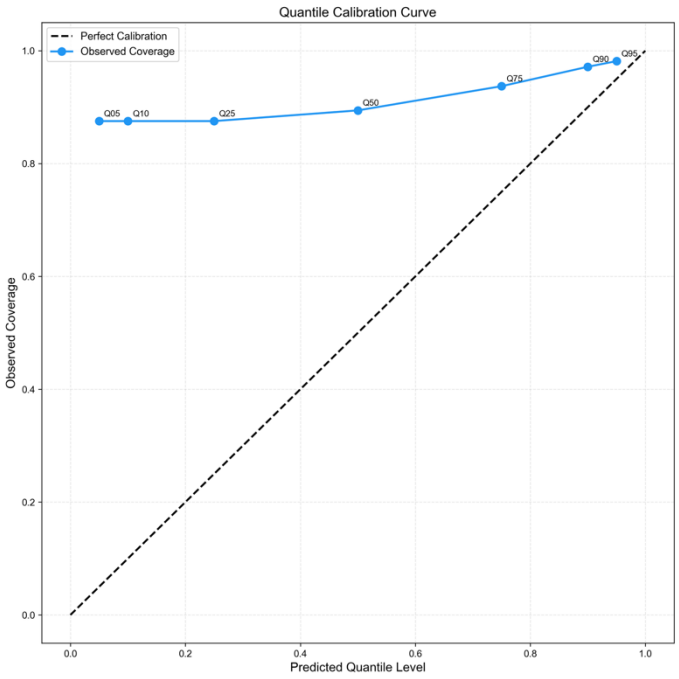

圖 4-5 DeepAR–1 日分位數校準曲線

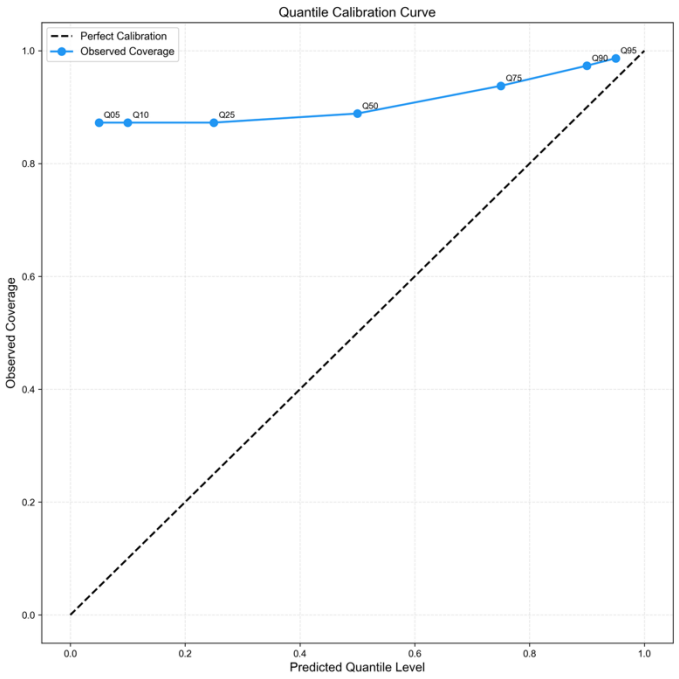

圖 4-6 DeepAR–7 日分位數校準曲線

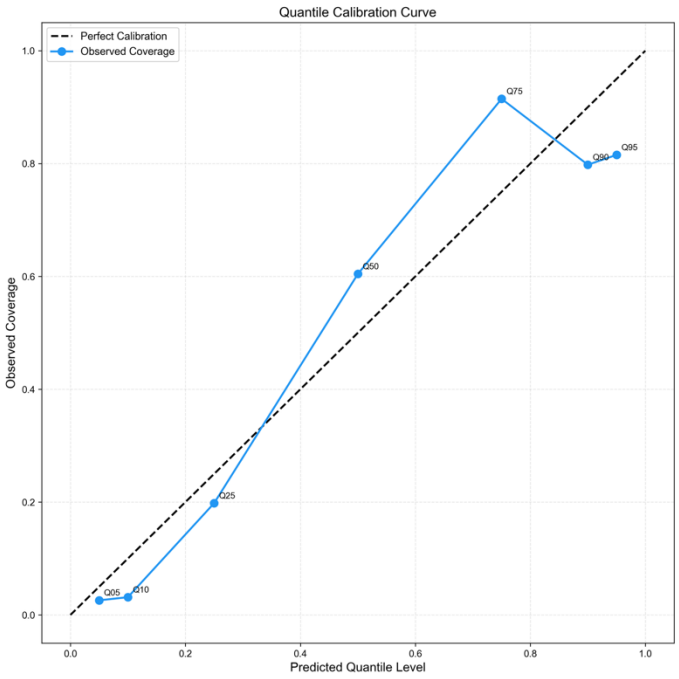

圖 4-7 TFT–1 日分位數校準曲線

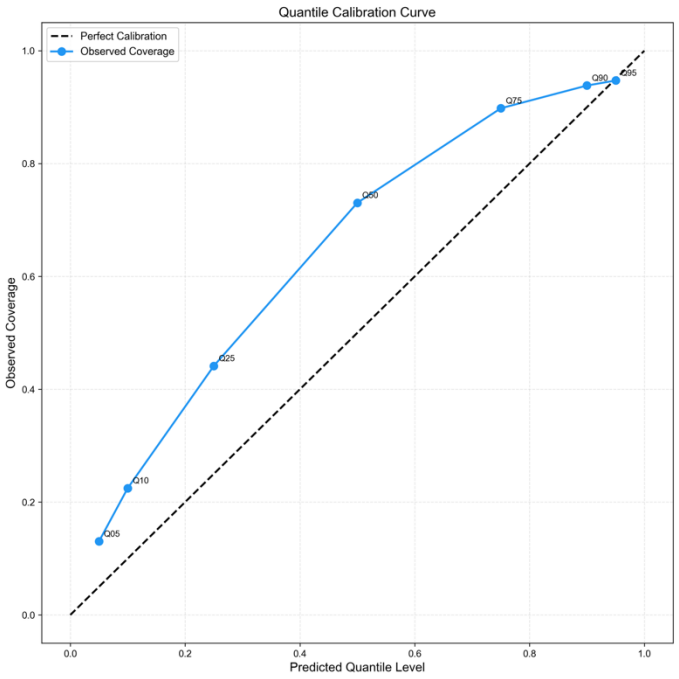

圖 4-8 TFT–7 日分位數校準曲線

4.4.4 代表性網格之預測示例

為直觀呈現差異，本研究選取代表性網格展示預測結果，如圖 4-9、圖 4-10。

doi:10.6342/NTU202600877

整體而言，DeepAR 之區間較能涵蓋真實值之尖峰波動，而 TFT 之區間較緊密、資訊量較高，但在需求突升時偶有真實值落於區間之外之情形。

從決策使用角度而言，PICP 與分位數校準曲線的結果意味著：若將高分位數（例如 q75、q90）直接作為預置/增援門檻，模型的校準偏差會直接轉化為資源配置的偏差。DeepAR 較保守的分佈刻畫可降低漏接風險，但可能提高閒置與過度預備成本；相對地，TFT 在特定情境下的上尾分位數偏低或中央區間覆蓋不足，可能使突升需求下出現低配風險。因此，若要將分位數輸出用於門檻式調度規則，除比較 MAE/RMSE 外，亦需同時檢視其校準程度，並可考慮導入分位數校準或 conformal prediction 等後處理，使高分位數風險訊號在長期統計意義下更可被信任地使用。

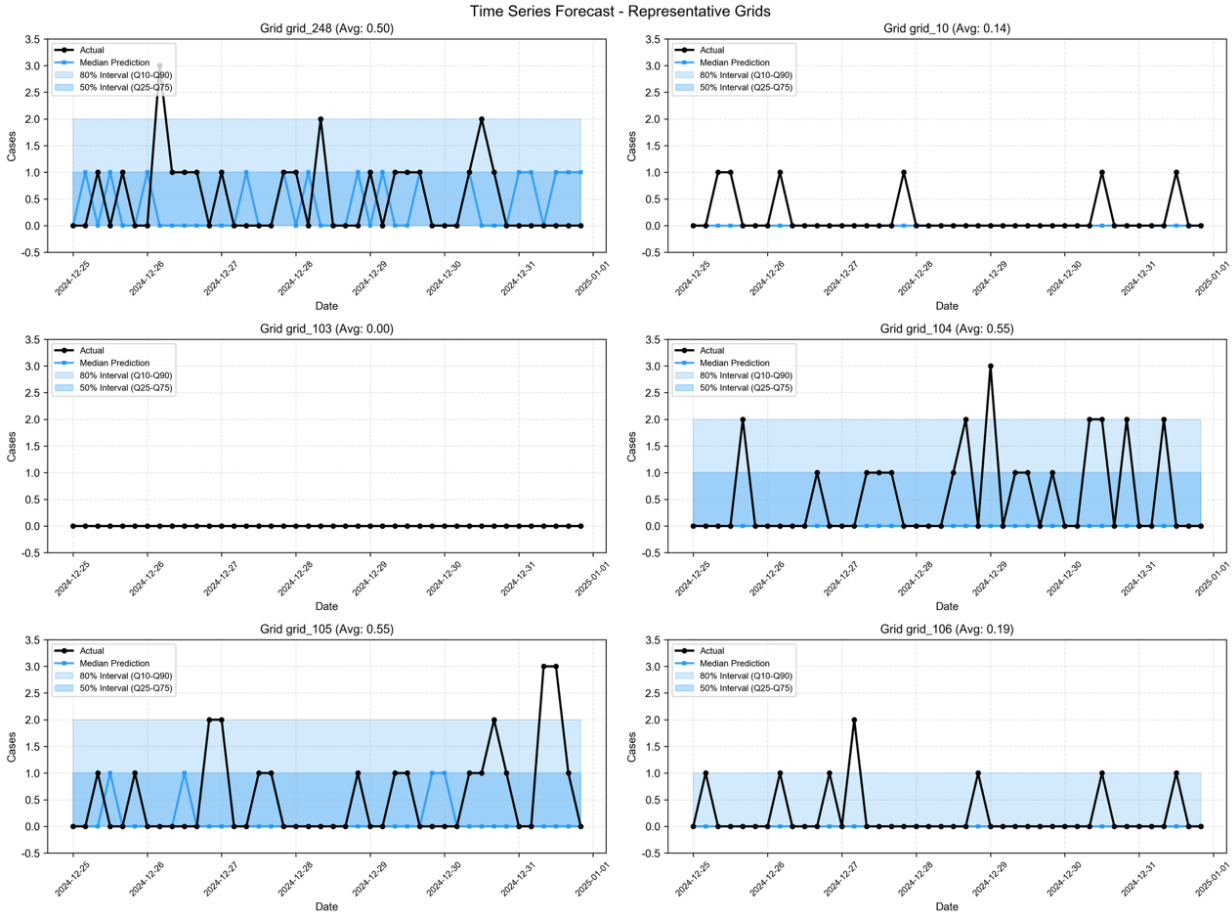

圖 4-9 DeepAR 於代表性網格之 7 日預測

doi:10.6342/NTU202600877

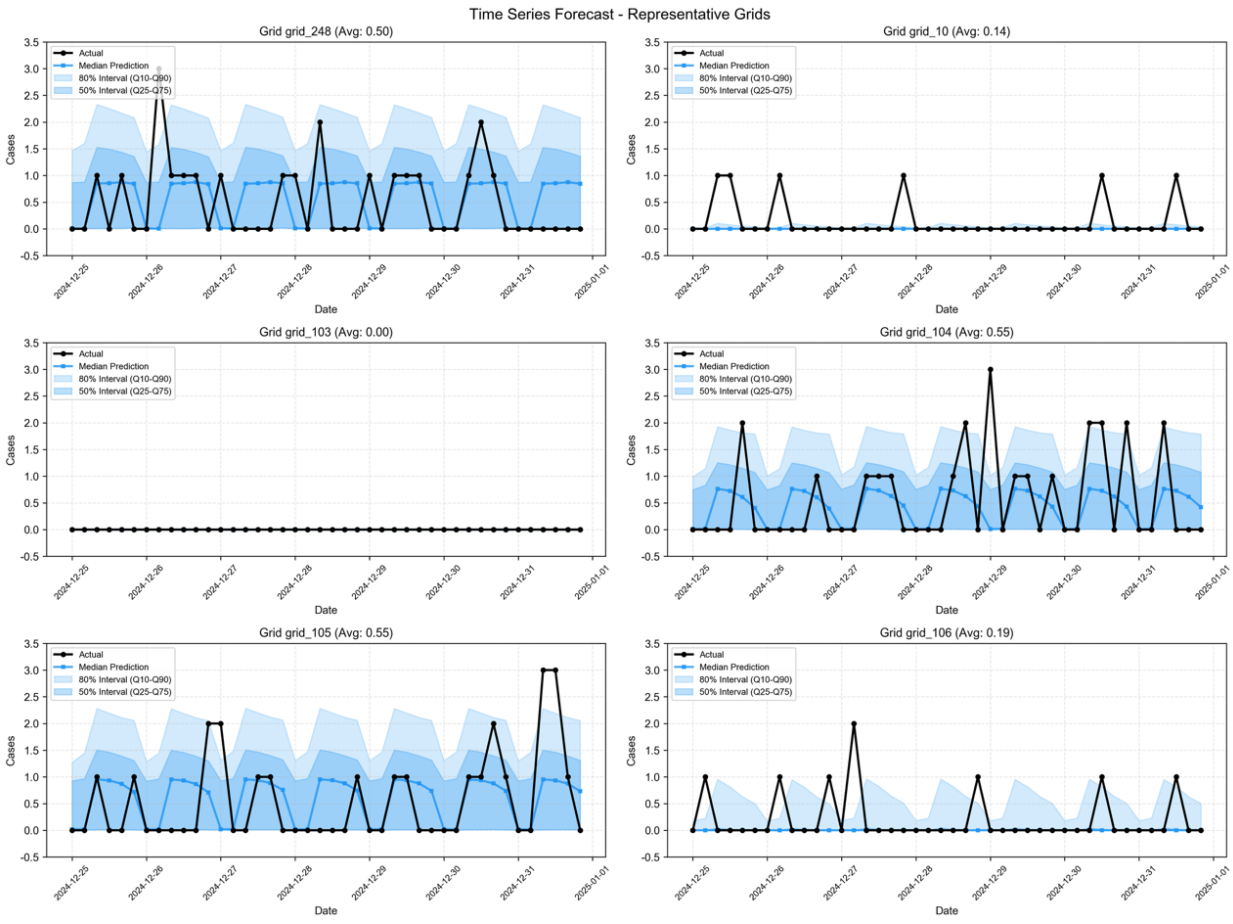

圖 4-10 TFT 於代表性網格之 7 日預測

4.5 實務應用效益分析：容忍誤差率（TRE）

本節以更貼近 EMS 調度實務之指標——容忍誤差率（Tolerant Rate Error, TRE）——評估模型可用性。當預測值與真實值差距在 ±1 件之內，通常可被視為可接受之操作誤差；因此 TRE 用以衡量超出容忍範圍之比例，TRE 越低代表模型越具實務可操作性。

整體而言，DeepAR 在 1 日與 7 日預測中皆呈現較低 TRE（1 日：2.89%；7 日：2.77%），優於 TFT（1 日：4.89%；7 日：5.25%），顯示 DeepAR 在可操作尺度上更為穩健。進一步就不同分位數作為決策基準之情境分析，若採用較保守之高分位數（例如 q75）作為資源預備依據，DeepAR 的 TRE 可進一步下降，反映其在風險規避型決策下具有潛在優勢；相對地，TFT 於高分位數設定下

doi:10.6342/NTU202600877

TRE 可能顯著惡化，顯示其高分位數預測在保守決策情境之直接可用性相對較低，如圖 4-11、圖 4-12。

圖 4-11 DeepAR 預測 1 日在不同分位數下 TRE 變化

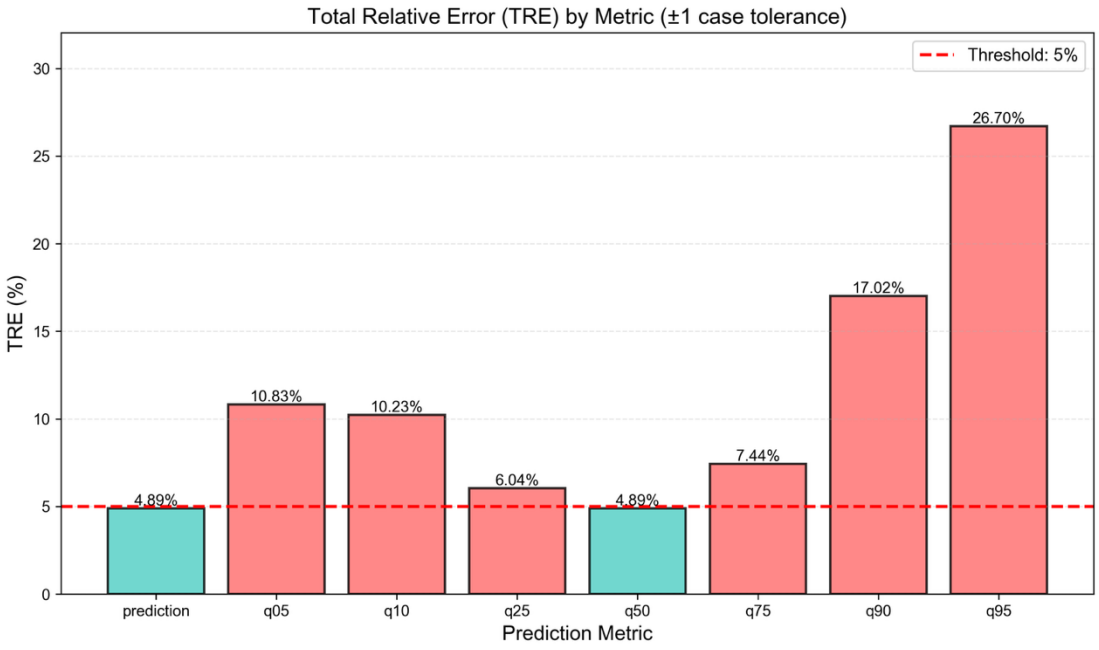

圖 4-12 TFT 預測 1 日在不同分位數下 TRE 變化

從指標關聯性角度解釋，DeepAR 在 MAE 略遜於 TFT 的情況下仍能取得

doi:10.6342/NTU202600877

較佳 TRE，可能與其對計數資料之分佈假設與輸出特性有關，使其點預測較常落於 0 或 1 等較接近整數之區間，提高落在 ±1 容忍帶內之機率；而 TFT 雖能取得更低平均誤差，但在部分情境下可能輸出較多非整數的小幅偏差，進而在 TRE 指標下累積較多未命中樣本（圖 4-13、圖 4-14）。整體而言，以實務容錯角度評估，DeepAR 具有較高可操作性與穩健性。

為使 TRE 與調度流程具體連結，本研究將分位數預測視為可調式的風險控制參數：決策端先依風險偏好選定分位數，再以該分位數預測值設定預置或增援門檻。以單一網格在某一 4 小時時段為例，可形成下列可操作規則：(1) 常態配置（效率導向）：以 q50 作為主要配置依據；(2) 增援觸發（風險導向）：以較高分位數（例如 q75 或 q90）作為增援門檻，當預測值超過容量臨界值時觸發跨分隊支援或動態重定位；(3) 尖峰保護（極端風險）：以 q95 作為極端風險訊號，用於特殊節日、惡劣天候或大型活動等高不確定性情境。此時 TRE 可被解讀為在指定容錯尺度（±1）下，採用某分位數作門檻時，預測落在可接受範圍內的穩健程度，使模型比較能直接回扣到操作決策需求。

doi:10.6342/NTU202600877

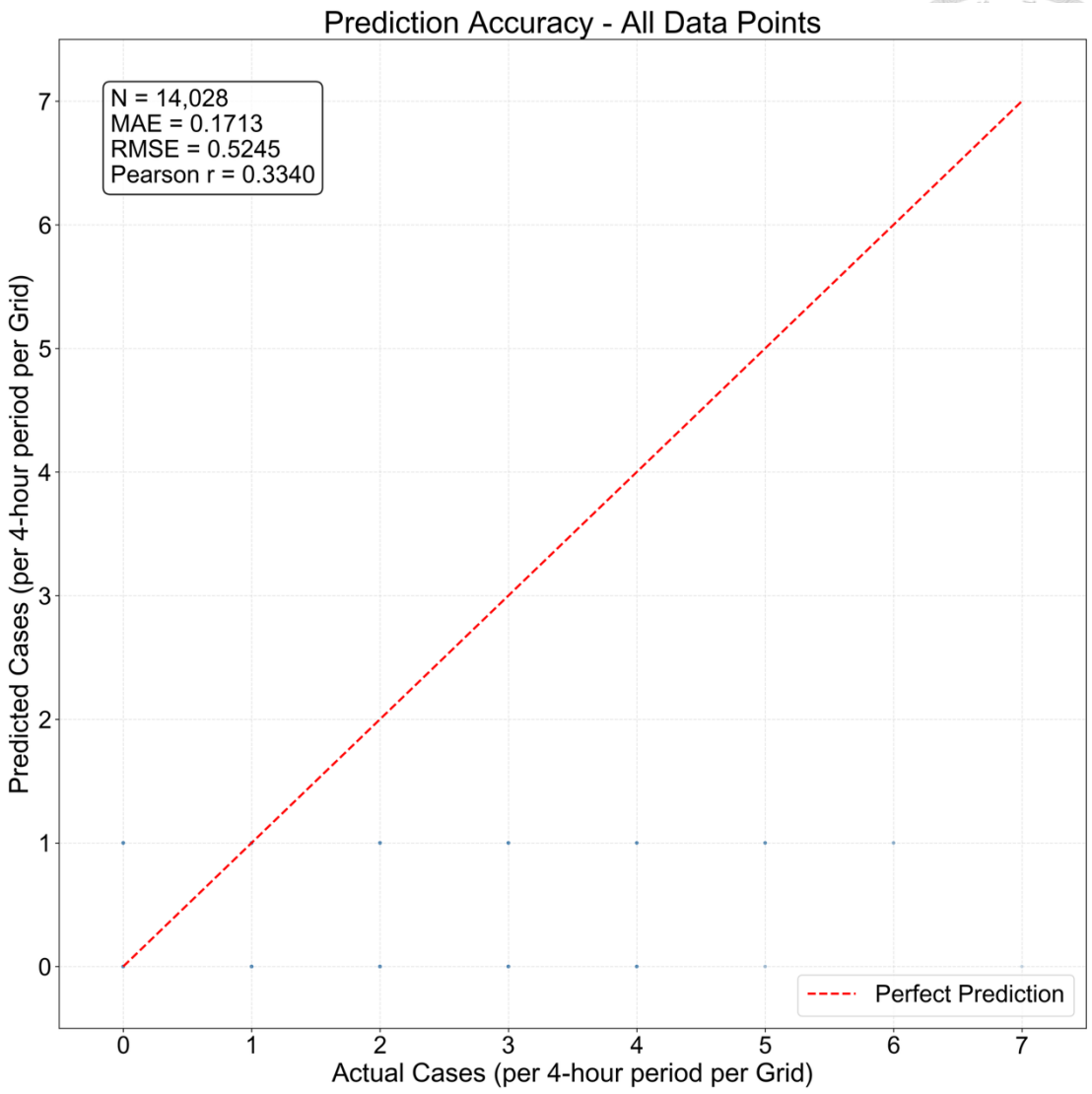

圖 4-13 DeepAR 預測 7 日正確率分布圖

doi:10.6342/NTU202600877

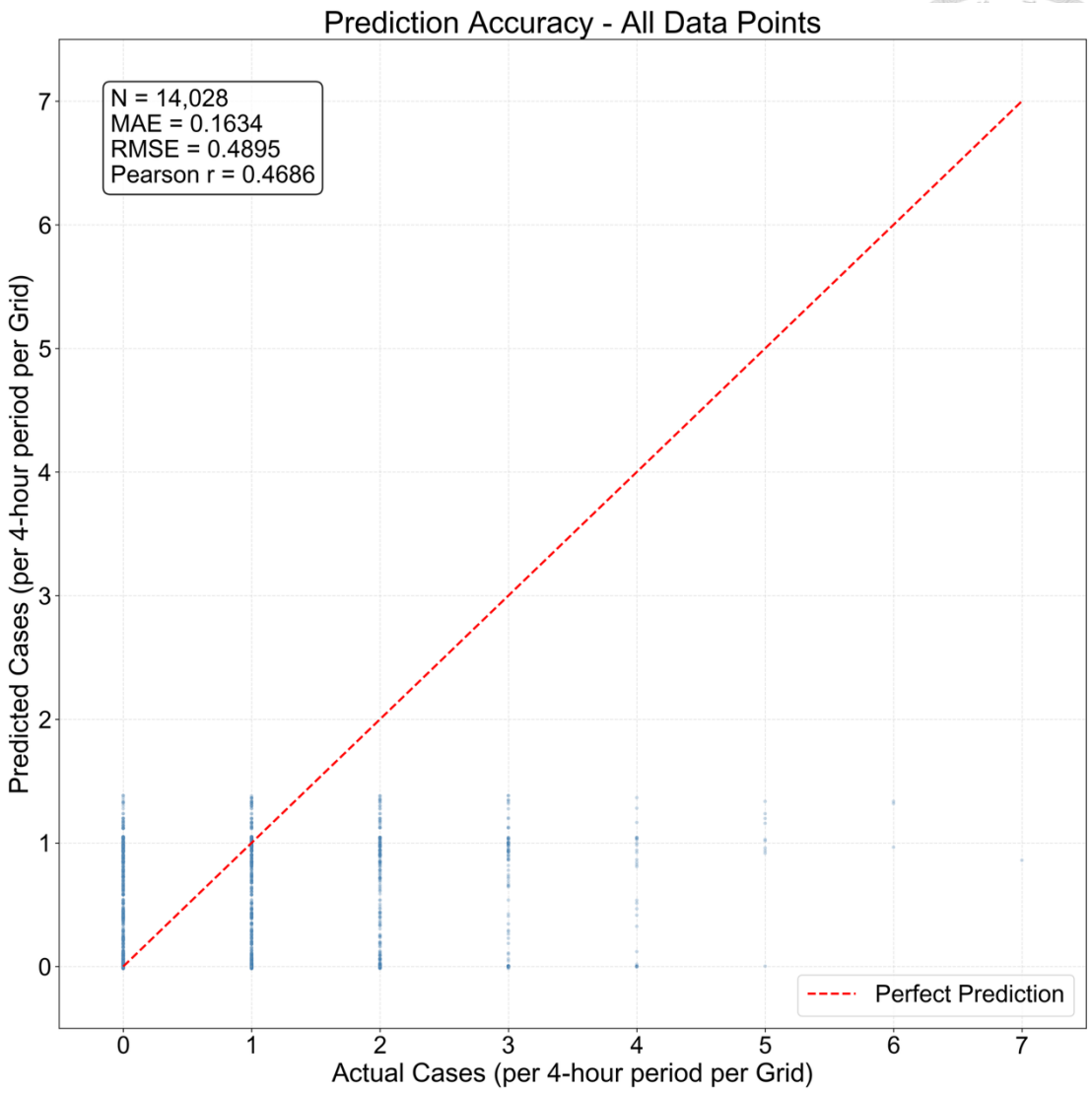

圖 4-14 TFT 預測 7 日正確率分布圖

4.6 置換法特徵重要性（Permutation Feature Importance）

為補足深度學習模型於實務應用情境中的可解釋性，本研究在完成第四章前述效能比較後，進一步採用置換法特徵重要性（Permutation Feature Importance, PFI）分析模型對各輸入特徵的依賴程度。PFI 的核心概念為：在測試資料上固定模型參數不變，針對單一特徵進行隨機置換（shuffle），以破壞其與目標變數之關聯，並觀察置換前後預測誤差的變化幅度。若某特徵被置換後導致模型誤差顯著上升，則表示該特徵對模型預測具有較高貢獻；反之，若誤差變化有限，則代表該特徵之邊際貢獻相對較小。本研究以測試集 MAE 作為衡量基準，定義

doi:10.6342/NTU202600877

$$\Delta\text{MAE} = \text{MAE}(\text{shuffled}) - \text{MAE}(\text{baseline}),$$

並以 ΔMAE 量化重要性大小，據以排序比較不同特徵之相對影響。

需要注意的是，置換法可能受單次隨機置換的抽樣波動與特徵間冗餘影響，部分特徵可能出現 $\Delta\text{MAE} < 0$（即置換後誤差反而下降）之情形；因此本節主要以 $|\Delta\text{MAE}|$的量級與排序作為解讀依據，並將負值視為重要性偏低或不具穩健貢獻的可能訊號，後續可透過多次置換取平均與標準差以提升估計穩健性。

另需強調，PFI 衡量的是在固定模型參數與既有輸入結構下，單一外生特徵被破壞時對誤差的邊際影響，不應被解讀為模型所有資訊來源的重要性總排名。由於兩模型皆大量利用歷史需求序列作為主要訊號來源，而本節置換的對象以外生共變數（時間日曆、氣象、人口等）為主，故 PFI 更接近回答在已能利用歷史序列的前提下，外生特徵額外提供多少資訊。此外，DeepAR 與 TFT 的重要性量級與排序差異，亦可能反映兩者的特徵融合機制不同：DeepAR 於序列生成過程中持續使用共變數訊號，對外生特徵擾動較敏感；TFT 則透過門控與注意力機制選擇性整合特徵，使單一特徵被置換不一定造成等量級誤差惡化。因此本研究以排序與相對量級作為行為理解依據，而非將其視為絕對因果影響估計。而由於本研究未區分不同 EMS 需求成因（如事故、急病等），若各成因對天候敏感度不同，混合後可能稀釋氣象特徵在整體資料上的邊際貢獻，後續可透過案件類型分層或分群評估以驗證此推論。

4.6.1 DeepAR 之 PFI 結果

圖 4-15 與圖 4-16 分別呈現 DeepAR 在 1 日與 7 日預測時長下之 PFI 結果。整體而言，DeepAR 在兩種視野下均呈現較明確的特徵重要性排序，其中人口與氣象因子在多數情境下具有較高的影響幅度。

doi:10.6342/NTU202600877

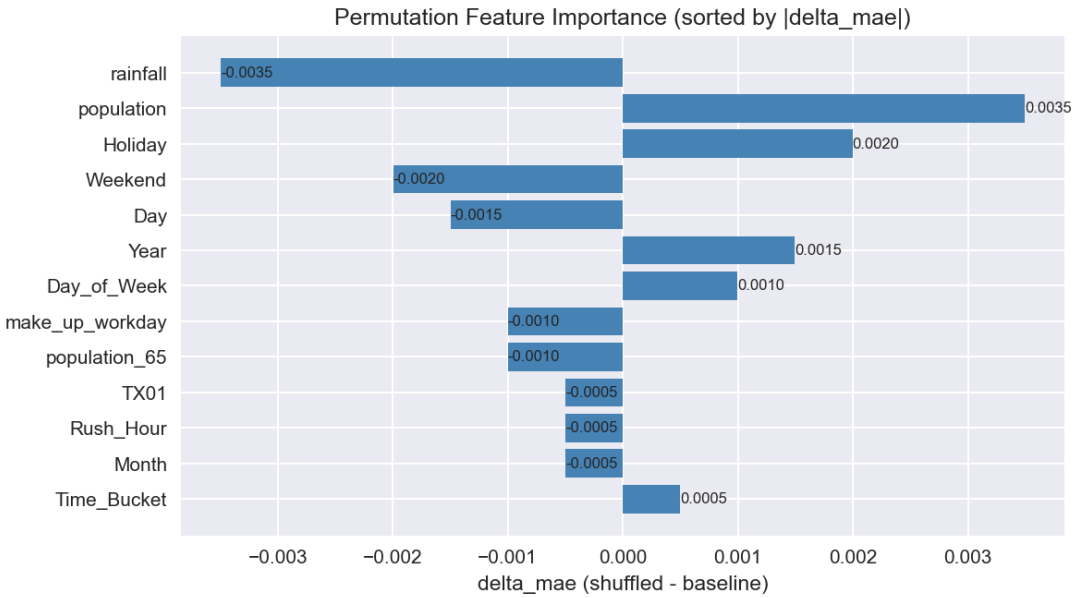

圖 4-15 DeepAR 在 1 日預測時長下之 PFI 結果

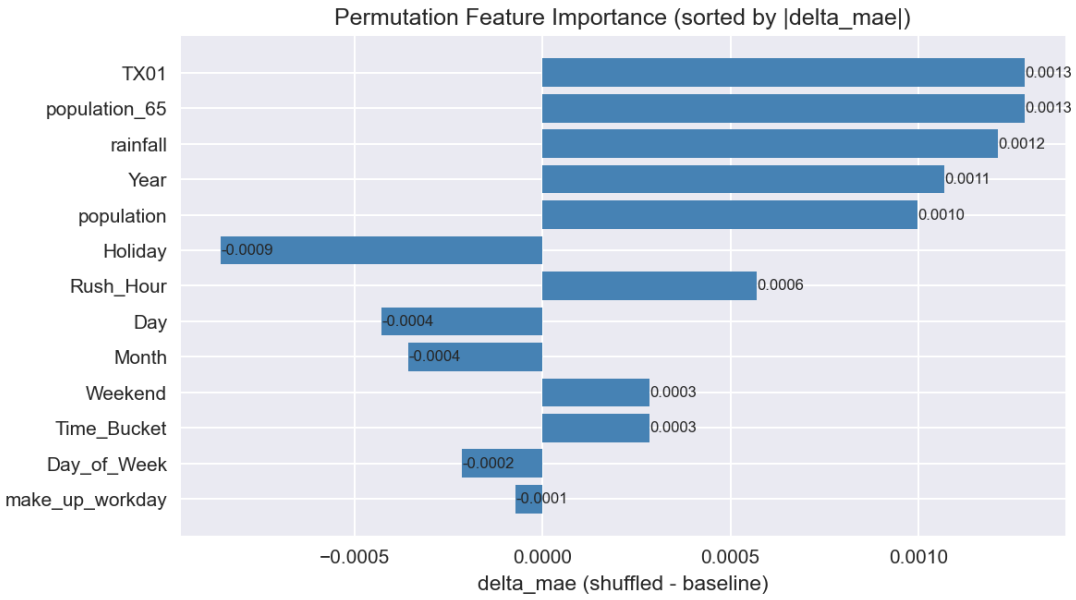

圖 4-16 DeepAR 在 7 日預測時長下之 PFI 結果

在 1 日預測中，模型對人口規模（population）與降雨量（rainfall）最為敏感，其次為假日（Holiday）、週末（Weekend）與年度／日期等時間結構特徵。此結果顯示短期 EMS 需求預測除仰賴歷史需求型態外，仍會受到空間基數差異（人口結構）與即時環境條件（降雨）之共同影響。此外，假日與週末的重要性

doi:10.6342/NTU202600877

亦反映休假日活動型態改變可能對 EMS 需求造成可辨識的擾動。

在 7 日預測中，DeepAR 的重要性排序則更偏向慢變背景特徵與季節性訊號，例如氣溫（TX01）、高齡人口（population_65）、降雨量（rainfall）與年度（Year）等。此一差異可解釋為：當預測視野拉長至週尺度時，模型更需要倚賴反映季節性與人口結構的變數，以刻畫跨日尺度之需求趨勢與穩定波動，而非僅依賴短期內的日曆事件訊號。

4.6.2 TFT 之 PFI 結果

圖 4-17 與圖 4-18 分別呈現 TFT 在 1 日與 7 日預測視野下之 PFI 結果。相較於 DeepAR，TFT 的 ΔMAE 整體量級偏小，且重要性排序多由人口相關靜態特徵主導，而部分時間與氣象特徵對整體誤差的邊際影響相對有限。此現象可能表示：在本研究的資料結構與模型設定下，TFT 更傾向以歷史需求訊號與靜態背景建立網格的需求基準，而單一外生特徵被置換時，對整體預測誤差的敏感度較低。

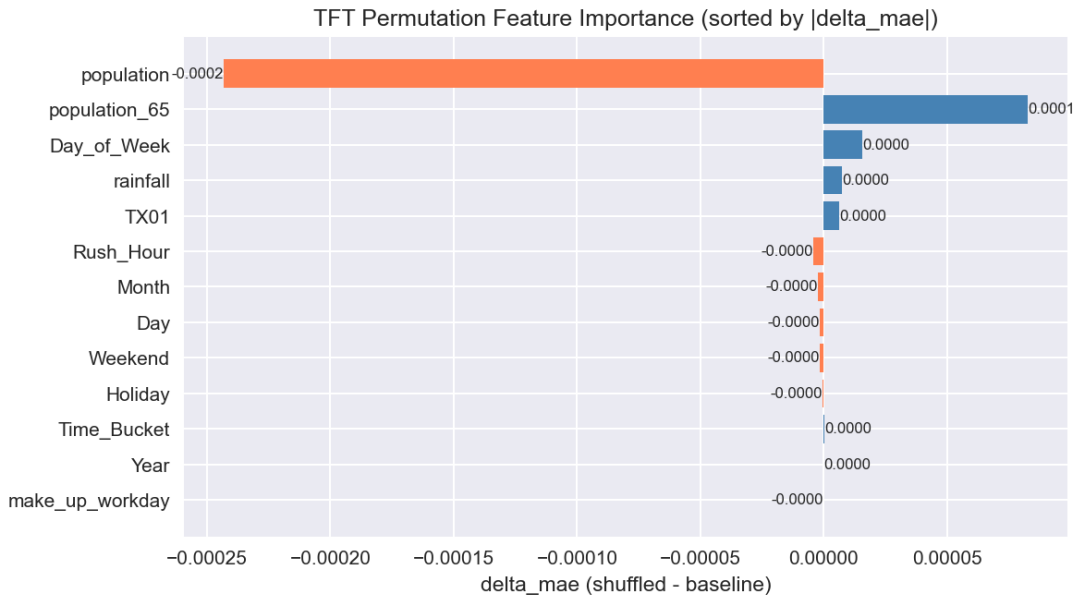

圖 4-17 TFT 在 1 日預測時長下之 PFI 結果

doi:10.6342/NTU202600877

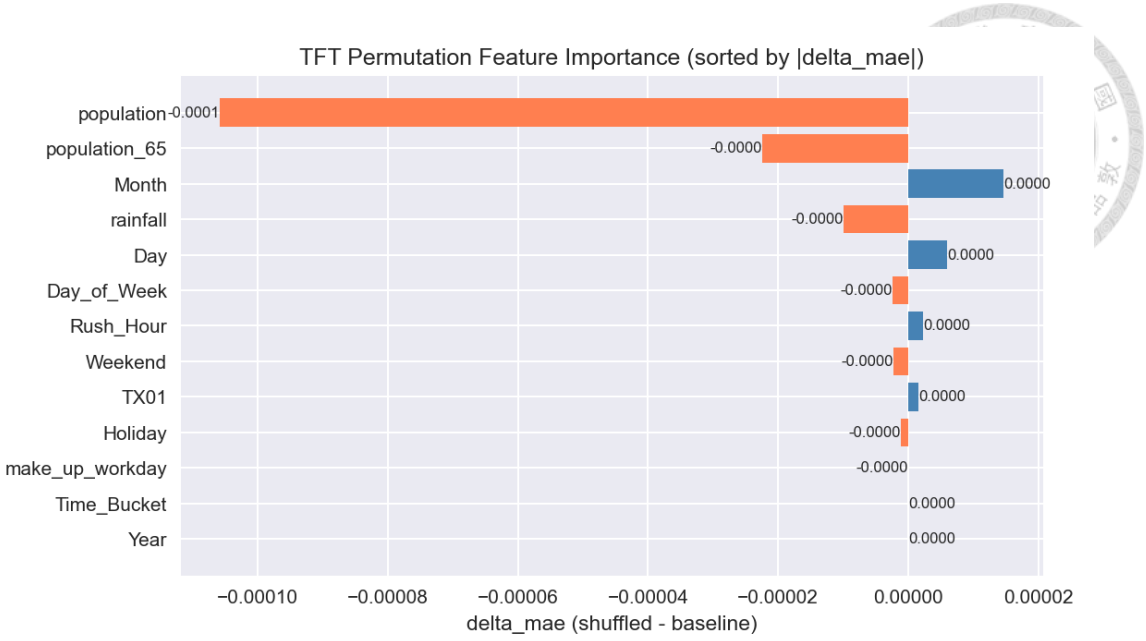

圖 4-18 TFT 在 7 日預測時長下之 PFI 結果

在 1 日預測中，PFI 顯示人口規模（population）與高齡人口（population_65）仍是相對重要的解釋來源；時間結構（如 Day_of_Week）與氣象因子（rainfall、TX01）的重要性幅度較小。至 7 日預測時，人口與高齡人口的重要性仍維持在前段，並可觀察到月份／日期等時間尺度特徵具有一定影響，但整體 ΔMAE 仍偏小，顯示 TFT 對單一特徵置換的敏感度有限。值得注意的是，部分特徵出現 ΔMAE < 0，可能源自特徵間資訊重疊、置換造成偶然改善或單次置換波動；因此本研究以 |ΔMAE|作為主要排序依據，後續亦建議採多次置換取平均以確認排序穩健性。

4.6.3 四種情境之重要性摘要比較

為便於跨模型與跨視野比較，表 4-5 彙整四種情境下 PFI 前五名之重要特徵（依|ΔMAE|由大至小排序）。整體結果顯示，人口（population）與降雨量（rainfall）在多數情境下皆名列前茅，指出需求預測除既有時序依賴外，亦顯著受人口規模與天候條件驅動；此外，TFT 在不同視野下較頻繁依賴時間結構特徵（如 Day_of_Week、Month），反映其對週期性與季節性訊號之捕捉。

doi:10.6342/NTU202600877

表 4-5 四種情境之 PFI 前五名特徵摘要（依 |ΔMAE|排序）

情境	前五名重要特徵（依 |ΔMAE| 排序）
DeepAR（1 日）	population、rainfall、Holiday、Weekend、Year
DeepAR（7 日）	TX01、population_65、rainfall、Year、population
TFT（1 日）	population、population_65、Day_of_Week、rainfall、TX01
TFT（7 日）	population、population_65、Month、rainfall、Day

綜合而言，PFI 結果提供了對前述效能差異的行為層面解讀：DeepAR 在不確定性覆蓋與容忍誤差表現較佳的同時，對外生變數的擾動亦較敏感；TFT 雖在點預測誤差上略占優勢，但其在本研究設定下對單一特徵置換的敏感度較低，可能呈現更依賴序列表示的預測機制。下一節將整合各項評估結果，對兩模型特性進行總結並提出應用情境建議。

4.7 綜合討論與模型特性總結

本節彙整前述各評估維度之發現，對 DeepAR 與 TFT 在臺北市 EMS 需求預測任務之表現進行整體權衡分析。整體而言，TFT 在點預測精度（MAE/RMSE）上具有微弱優勢，且在較長預測視野下對高需求事件之誤差控制相對較佳；DeepAR 則在不確定性量化上呈現較高覆蓋率與更低之 TRE，顯示其在保守調度與風險規避情境中具備較高之可操作性與穩健性。表 4-6 綜合整理兩模型之主要特性差異。

表 4-6 模型特性綜合比較

特性維度	DeepAR	TFT
點預測準確性（MAE/RMSE）	表現良好，但整體略遜於 TFT	整體表現較佳，尤以 7 日視野更顯著
不確定性校準（PICP@80%）	過度保守：PICP 顯著高於 80%，區間偏寬、資訊量較低	略顯激進：PICP 低於 80%，區間偏窄、可能低估尾端風險
實務應用效益（TRE，τ=±1）	表現更優：TRE 較低（約 2.8%），可操作性較佳	表現良好（約 5.3%），但略遜於 DeepAR

doi:10.6342/NTU202600877

高需求事件預測能力	仍具挑戰，誤差隨案件量提升而上升	同樣具挑戰，但高需求區間之誤差控制略優
核心優勢	覆蓋率高、TRE 穩健，適用風險規避決策	點預測精準、可整合多類共變數，適用資源優化情境
核心劣勢	區間過寬、資訊量較低	覆蓋率偏低，高分位數決策下可用性需審慎

基於上述特性，本研究建議模型選用應回扣實務目標與風險偏好：若調度目標偏向「預留餘裕、避免漏接需求」，則 DeepAR 之高覆蓋率與較低 TRE 可提供較穩健之決策基準；若決策情境更重視「提升資源利用效率、降低閒置成本」，則 TFT 之較高點預測精度可作為更具成本效益之部署依據。綜上所述，兩模型各有所長，並不存在絕對的單一最佳解；後續應用宜依場域特性、資源限制與可接受風險水準，選擇最契合之模型與決策分位數設定。

doi:10.6342/NTU202600877

第五章 結論與未來研究方向

5.1 研究結論

本研究以臺北市為研究區域，針對緊急救護服務（EMS）需求之時空變動特性，建構一套具可重現性的資料處理、特徵工程與深度學習預測流程。研究以 1000 公尺 × 1000 公尺網格作為空間單元，並將一天切分為 6 個 4 小時時段（Time_Bucket）以形成多序列時間序列資料架構；同時整合時間日曆特徵、降雨量（以克利金法由測站內插至網格）、氣溫（臺北測站代表值）與人口／高齡人口等多源共變數，進一步比較 DeepAR 與 TFT 兩種機率式深度學習模型於 1 日（6 步）與 7 日（42 步）預測視野下的表現，並以 MAE、RMSE、PICP@80%、TRE（τ=±1）建立兼具「準確性—不確定性—實務可用性」的評估框架。

整體結果顯示，兩模型各具優勢且呈現明顯權衡。就點預測準確性而言，TFT 整體略優於 DeepAR，且在 7 日預測中取得較低誤差（MAE=0.1634、RMSE=0.4895），顯示 TFT 在捕捉較長時間尺度之趨勢與週期訊號方面具一定優勢；相對地，DeepAR 在點預測誤差上雖略高（例如 1 日 RMSE=0.5700），但仍維持在相近水準（MAE 約 0.1742）。此外，本研究觀察到 7 日預測在 MAE/RMSE 上略低於 1 日預測之現象，推測與較長視野有助於模型以更平滑方式反映週期性結構、降低短期隨機波動干擾有關，惟其機制仍需搭配更多情境分析與資料切分檢驗以確認其普遍性。

就機率式預測可靠性（PICP80%）而言，DeepAR 的覆蓋率明顯高於名目水準（1 日 97.16%、7 日 97.35%），反映其預測區間偏向過度保守（區間過寬、資訊量較低）；TFT 的覆蓋率則低於 80%（1 日 76.65%、7 日 72.92%），呈現略顯激進（區間偏窄、可能低估尾端風險）的特性。此一差異顯示兩模型在不確定性刻畫策略上存在系統性偏差：DeepAR 以安全覆蓋為主，而 TFT 更接近區間緊密、資訊較集中的取向。

就實務應用效益（TRE，τ=±1）而言，DeepAR 在兩種預測視野皆呈現較低

doi:10.6342/NTU202600877

之 TRE（1 日 2.89%、7 日 2.77%），優於 TFT（1 日 4.89%、7 日 5.25%）。此結果表示，即便 TFT 在平均誤差指標（MAE/RMSE）略具優勢，DeepAR 的預測在是否落於可接受容錯範圍的操作尺度上更穩健，對應 EMS 資源調度與排班等情境下的可用性較高。

綜合而言，若應用目標偏向風險規避或確保供給安全，DeepAR 的區間覆蓋與 TRE 優勢更具吸引力；若目標偏向資源效率與常態配置精準化，TFT 的點預測優勢則較能發揮。

最後，本研究亦針對模型可解釋性需求引入置換法特徵重要性（PFI）作為輔助分析工具，以檢視各類特徵對預測效能之邊際貢獻，並作為特徵工程合理性與模型行為理解之依據。該方法提供一個可與深度模型訓練流程相容、且能跨模型比較的解釋途徑，有助於後續在政策或實務溝通中說明模型判斷所倚賴的資訊來源。

5.2 研究貢獻

本研究之主要貢獻可歸納如下：

1. 提出可重現之時空序列建模流程：從事件點位網格化、時段彙整、多源特徵融合、資料序列化到模型訓練與評估，建立完整且可複製的端到端架構。
2. 建立兼顧準確性、可靠性與實務性的評估框架：除 MAE/RMSE 外，納入 PICP@80% 與 TRE（±1）以貼近 EMS 調度的實際需求，使模型比較不僅停留在平均誤差，而能反映風險與可操作性。
3. 比較兩類機率式深度模型於不同視野下的系統性差異：揭示 TFT 在點預測精度上具優勢、DeepAR 在覆蓋安全性與容忍誤差表現上更佳的互補特性，並提供對應的應用情境建議。
4. 引入模型解釋分析（PFI）以支援特徵與模型行為理解：補足深度學習模型在實務應用時常被質疑的黑箱性問題，提升結果詮釋與後續精進方向的依據。

doi:10.6342/NTU202600877

5.3 研究限制

本研究仍存在若干限制，後續解讀結果時需審慎看待：

1. 氣溫採單一測站代表值：以臺北測站日尺度氣溫近似全市熱環境，可能低估市郊山區與都市核心之溫度異質性。
2. 降雨量內插與人口網格化存在方法假設：克利金法仰賴空間自相關結構與參數擬合；面積加權人口配置隱含村里內均勻分布假設，可能造成局部網格人口估計偏差。
3. 零膨脹與高峰事件仍是主要挑戰：整體誤差受大量零需求樣本主導，模型於罕見高需求事件的誤差仍顯著上升，對尖峰應變的保障程度有限。
4. 未區分 EMS 需求成因：本研究以整體 EMS 案件量作為目標變數，未依事故、急病等成因分層。若不同成因對天候敏感度不同，混合後可能稀釋氣象特徵在整體資料上的邊際影響，並降低對特定高風險情境的辨識能力。後續可依案類分層建模或採多任務學習以驗證差異。
5. 不確定性校準仍未完善：DeepAR 的覆蓋率過高、TFT 的覆蓋率偏低，顯示兩者皆可能存在校準偏差；僅以 PICP 評估仍不足以全面刻畫區間品質（例如區間寬度—覆蓋率之效率權衡）。本研究已以分位數校準曲線補充檢核，但尚未進一步進行校準後處理，故仍需審慎解讀分位數在決策門檻上的直接使用。
6. 外生變數在預測視野的可得性議題：為使模型比較具可比性，本研究於離線評估採完美預報假設，將預測視野內天氣變數視為可取得資訊，並以實際觀測值替代完美預報進行驗證；然而真實部署時需以天氣預報或情境模擬輸入取代觀測值，且預報誤差可能影響模型表現。本研究尚未納入不同預報誤差情境之敏感度分析，故結果外推至真實部署時仍需審慎解讀。
7. 基準模型（baseline）比較仍不足：本研究主要比較兩類深度機率模型（DeepAR 與 TFT），尚未納入傳統時間序列或常見機器學習之基準模型

doi:10.6342/NTU202600877

（例如 seasonal naive、Poisson/NegBin GLM、XGBoost 等）作為對照，因此難以量化深度模型相對於簡單基準的增益幅度。後續可於相同資料切分與資訊假設下補充 baseline 實驗，以提升結論的可比性與說服力。

5.4 未來研究方向

基於上述發現與限制，本研究建議未來可從以下方向延伸：

1. 不確定性校準與區間品質優化：導入分位數校準（quantile calibration）、保形（conformal prediction）或後驗校準方法，使 PICP 更貼近名目水準，同時評估區間寬度與覆蓋率之效率；亦可加入區間相關損失函數，直接在訓練目標中反映校準需求。
2. 更符合計數與零膨脹特性的分佈建模：針對大量零需求與少數高峰事件，可嘗試零膨脹模型（zero-inflated / hurdle）、混合分佈、或以兩階段模型先判斷事件發生機率再估計需求量，以提升高需求情境辨識。
3. 強化空間相依結構的建模：現行以多序列全球模型共享參數，尚未顯式利用鄰近網格關係；後續可結合圖神經網路（GNN）、空間注意力或圖式 Transformer，將鄰接關係、道路可達性與跨區域擴散效應納入。
4. 導入更多具機制意涵之外生變數：例如土地使用、醫療資源分布、交通壅塞、人口流動（通勤/活動）、大型活動與節慶、空氣品質與熱浪指標等，以提升模型對高峰需求的解釋力與預測力。
5. 氣象特徵之解析度與資料來源升級：未來可將降雨特徵由測站內插改採氣象格點降雨產品（如雷達與雨量站校正之 QPE/QPESUMS），直接對齊 1km 網格以降低內插假設；氣溫亦可由單站代表值改為多測站/格點溫度，並以小時尺度彙整至 4 小時 Time_Bucket，以更貼近日內變化與空間異質性。
6. 以可用的未來資訊進行更貼近實務的預測設定：將氣象觀測改為天氣預報或情境模擬輸入，並建立不同預報誤差情境下的敏感度分析，提升研究結

doi:10.6342/NTU202600877

果對部署情境的外推性。

7. 擴充評估尺度與決策導向驗證：除統計指標外，可進一步建立資源調度模擬（例如以不同預測策略推估車輛需求、待命配置與反應時間變化），並以成本—效益或風險函數評估模型在政策與操作上的真實價值。

8. 跨年度與跨城市之可移植性檢驗：可嘗試跨年驗證、跨行政區（含新北交界）或跨城市移轉學習，以檢驗模型與特徵工程流程的泛化能力；亦可將疫情期間作為領域轉移（domain shift）案例，探索韌性與自適應更新策略。

doi:10.6342/NTU202600877

參考文獻

1. Aringhieri, R., et al., *Emergency medical services and beyond: Addressing new challenges through a wide literature review.* Computers & Operations Research, 2017. **78**: p. 349–368.
2. Bélanger, V., A. Ruiz, and P. Soriano, *Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles.* European Journal of Operational Research, 2019. **272**(1): p. 1–23.
3. Brotcorne, L., G. Laporte, and F. Semet, *Ambulance location and relocation models.* European Journal of Operational Research, 2003. **147**(3): p. 451–463.
4. Olasveengen, T.M., et al., *European Resuscitation Council Guidelines 2021: Basic Life Support.* Resuscitation, 2021. **161**: p. 98–114.
5. Panchal, A.R., et al., *Part 3: adult basic and advanced life support: 2020 American Heart Association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care.* Circulation, 2020. **142**(16_Suppl_2): p. S366–S468.
6. Pons, P.T., et al., *Paramedic response time: does it affect patient survival?* Academic Emergency Medicine, 2005. **12**(7): p. 594–600.
7. Soar, J., et al., *European Resuscitation Council Guidelines 2021: Adult advanced life support.* Resuscitation, 2021. **161**: p. 115–151.
8. Chen, A.Y., et al., *Demand Forecast Using Data Analytics for the Preallocation of Ambulances.* IEEE Journal of Biomedical and Health Informatics, 2016. **20**(4): p. 1178–1187.
9. Lin, A.X., et al. *Leveraging Machine Learning Techniques and Engineering of Multi-Nature Features for National Daily Regional Ambulance Demand Prediction*. International Journal of Environmental Research and Public Health, 2020. **17**, 4179 DOI: 10.3390/ijerph17114179.
10. Martin, R.J., R. Mousavi, and C. Saydam, *Predicting emergency medical service call demand: A modern spatiotemporal machine learning approach.* Operations Research for Health Care, 2021. **28**: p. 100285.
11. Neira-Rodado, D., et al., *A novel machine learning approach for spatiotemporal prediction of EMS events: A case study from Barranquilla, Colombia.* Heliyon, 2025. **11**(2): p. e41904.
12. Channouf, N., et al., *The application of forecasting techniques to*

doi:10.6342/NTU202600877

modeling emergency medical system calls in Calgary, Alberta. Health Care Management Science, 2007. **10**(1): p. 25–45.
13. Setzler, H., C. Saydam, and S. Park, *EMS call volume predictions: A comparative study.* Computers & Operations Research, 2009. **36**(6): p. 1843–1851.
14. Hyndman, R.J. and A.B. Koehler, *Another look at measures of forecast accuracy.* International Journal of Forecasting, 2006. **22**(4): p. 679–688.
15. Lee, H. and T. Lee, *Demand modelling for emergency medical service system with multiple casualties cases: k-inflated mixture regression model.* Flexible Services and Manufacturing Journal, 2021. **33**(4): p. 1090–1115.
16. Matteson, D.S., et al., *Forecasting emergency medical service call arrival rates.* 2011.
17. Cressie, N. and C.K. Wikle, *Statistics for spatio-temporal data*. 2011: John Wiley & Sons.
18. Zhou, Z., et al., *A spatio-temporal point process model for ambulance demand.* Journal of the American Statistical Association, 2015. **110**(509): p. 6–15.
19. Zhou, Z. and D.S. Matteson, *Predicting Ambulance Demand: a Spatio-Temporal Kernel Approach*, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, Association for Computing Machinery: Sydney, NSW, Australia. p. 2297–2303.
20. Abreu, P., D. Santos, and A. Barbosa-Povoa, *Data-driven forecasting for operational planning of emergency medical services.* Socio-Economic Planning Sciences, 2023. **86**: p. 101492.
21. Monks, T., et al., *Forecasting the daily demand for emergency medical ambulances in England and Wales: a benchmark model and external validation.* BMC Medical Informatics and Decision Making, 2023. **23**(1): p. 117.
22. Tluli, R., et al., *A Survey of Machine Learning Innovations in Ambulance Services: Allocation, Routing, and Demand Estimation.* IEEE Open Journal of Intelligent Transportation Systems, 2024. **5**: p. 842–872.
23. Shahidian, N., et al., *Short-term forecasting of emergency medical services demand exploring machine learning.* Computers & Industrial Engineering, 2025. **200**: p. 110765.
24. Jin, R., et al., *Predicting Emergency Medical Service Demand With*

doi:10.6342/NTU202600877

Bipartite Graph Convolutional Networks. IEEE Access, 2021. **9**: p. 9903–9915.
25. Megouo, T.G.P. and S. Pierre, *A Stacking Ensemble Machine Learning Model for Emergency Call Forecasting.* IEEE Access, 2024. **12**: p. 115820–115837.
26. Garg, T., D. Toshniwal, and M. Parida, *A meta-learning ensemble framework for robust and interpretable prediction of emergency medical services demand.* Scientific Reports, 2025. **16**(1): p. 2132.
27. Hermansen, A.H. and O.J. Mengshoel. *Forecasting Ambulance Demand using Machine Learning: A Case Study from Oslo, Norway*. in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. 2021.
28. Fotheringham, A.S. and D.W.S. Wong, *The Modifiable Areal Unit Problem in Multivariate Statistical Analysis.* Environment and Planning A: Economy and Space, 1991. **23**(7): p. 1025–1044.
29. Shi, X., et al., *Convolutional LSTM network: A machine learning approach for precipitation nowcasting.* Advances in neural information processing systems, 2015. **28**.
30. Li, Y., et al., *Diffusion convolutional recurrent neural network: Data-driven traffic forecasting.* arXiv preprint arXiv:1707.01926, 2017.
31. Wu, Z., et al., *Graph wavenet for deep spatial-temporal graph modeling.* arXiv preprint arXiv:1906.00121, 2019.
32. Zhang, C., et al., *Deepmeshcity: A deep learning model for urban grid prediction.* ACM Transactions on Knowledge Discovery from Data, 2024. **18**(6): p. 1–26.
33. Kipf, T.N. and M. Welling, *Semi-supervised classification with graph convolutional networks.* arXiv preprint arXiv:1609.02907, 2016.
34. Hamilton, W., Z. Ying, and J. Leskovec, *Inductive representation learning on large graphs.* Advances in neural information processing systems, 2017. **30**.
35. Gneiting, T. and A.E. Raftery, *Strictly Proper Scoring Rules, Prediction, and Estimation.* Journal of the American Statistical Association, 2007. **102**(477): p. 359–378.
36. Salinas, D., et al., *DeepAR: Probabilistic forecasting with autoregressive recurrent networks.* International Journal of Forecasting, 2020. **36**(3): p. 1181–1191.
37. Lim, B., et al., *Temporal Fusion Transformers for interpretable multi-horizon time series forecasting.* International Journal of Forecasting,

doi:10.6342/NTU202600877

2021. **37**(4): p. 1748–1764.
38. Saito, T. and M. Rehmsmeier, *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.* PLOS ONE, 2015. **10**(3): p. e0118432.
39. Brodersen, K.H., et al. *The Balanced Accuracy and Its Posterior Distribution*. in *2010 20th International Conference on Pattern Recognition*. 2010.
40. He, H. and E.A. Garcia, *Learning from Imbalanced Data.* IEEE Transactions on Knowledge and Data Engineering, 2009. **21**(9): p. 1263–1284.
41. Davis, J. and M. Goadrich. *The relationship between Precision-Recall and ROC curves*. in *Proceedings of the 23rd international conference on Machine learning*. 2006.
42. Glenn, W.B., *Verification of forecasts expressed in terms of probability.* Monthly weather review, 1950. **78**(1): p. 1–3.
43. Lambert, D., *Zero-inflated Poisson regression, with an application to defects in manufacturing.* Technometrics, 1992. **34**(1): p. 1–14.
44. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. **16**: p. 321–357.
45. Niculescu-Mizil, A. and R. Caruana, *Predicting good probabilities with supervised learning*, in *Proceedings of the 22nd international conference on Machine learning*. 2005, Association for Computing Machinery: Bonn, Germany. p. 625–632.
46. Guo, C., et al., *On Calibration of Modern Neural Networks*, in *Proceedings of the 34th International Conference on Machine Learning*, P. Doina and T. Yee Whye, Editors. 2017, PMLR: Proceedings of Machine Learning Research. p. 1321—1330.
47. Van Calster, B., et al., *Calibration: the Achilles heel of predictive analytics.* BMC Medicine, 2019. **17**(1): p. 230.
48. Murphy, A.H., *A New Vector Partition of the Probability Score.* Journal of Applied Meteorology and Climatology, 1973. **12**(4): p. 595–600.
49. Koenker, R. and G. Bassett, *Regression Quantiles.* Econometrica, 1978. **46**(1): p. 33–50.
50. Khosravi, A., et al., *Comprehensive Review of Neural Network-Based Prediction Intervals and New Advances.* IEEE Transactions on Neural Networks, 2011. **22**(9): p. 1341–1356.
51. Hochreiter, S. and J. Schmidhuber, *Long Short-Term Memory.* Neural

doi:10.6342/NTU202600877

Computation, 1997. **9**(8): p. 1735–1780.
52. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.

doi:10.6342/NTU202600877