使用 GAN 進行資料補遺改善在不同深度學習之地下水位預測

Using GAN for Imputation of Missing Recorded Data to Improve Groundwater Level Prediction Based on Deep Learning Methods

國立成功大學

特聘教授 羅偉誠 Wei-Cheng Lo

碩士 陳欣妤 Hsin-Yu Chen

摘　要

文明的發展與環境生態系統運作皆與水資源相關，受地理環境限制與持續都市化、工業化影響，致使地面水資源供給不足之部分，須由地下水支應。而地下水屬須長年累積之資源，無法經過短時間的補注即恢復，因此基於世代間公平性而言，維持長期水資源的經營與管理，成為永續發展中的機制。以天然環境特質為基礎，準確預測地下水位為估算水資源總量的第一步，以利後續進行水資源的分配。現今，使用人工神經網絡來預測地下水位趨近於成熟，可成為管理地下水的重要工具。

然而在蒐集數據的過程中，會因為各種原因而導致遺漏，因此資料缺失是任何研究領域均須克服的問題，基於維持數據完整性，選擇遺漏值填補（Missing Value Imputation, MVI）來作應對，並證實以模型主的機器學習方法較為優異。因此，本研究主旨採用生成對抗網絡,以生成模型(Generative Model)和鑑別模型(Discriminative Model）相互抗衡的方式，對地下水位資料缺失部分進行補遺，以期望改善水資源度量的品質。

並使用兩種跨領域深度學習方法，發展其於 Univariate 和 Seq2val 模式中地下水位之預測性能，探討參數條件之意義，比較兩種模型於水文方面模擬之優劣。最後發展 Seq2seq 模式，展示兩模型於長時間水位模擬的極限，期望能降低水利研究工具選擇之侷限性，不僅能使用專門為序列資料開發的模型，亦能將在跨領域運用良好的工具列入考量，提供水資源研究更多可能性。

關鍵詞：生成對抗網路，卷積神經網路，長短期記憶，補遺，地下水位預測

Abstract

Using GAN for Imputation of Missing Recorded Data to Improve Groundwater Level Prediction Based on Deep Learning Methods

author: Hsin-Yu Chen
advisor: Wei-Cheng Lo
Department of Hydraulics and Ocean Engineering, National Cheng Kung University

SUMMARY

The development of civilization and the preservation of environmental ecosystems are strongly dependent on water resources. Typically, the insufficient supply of surface water resources for domestic, industrial, and agriculture needs is often supplemented by the ground water resources. However, the groundwater is a natural resource that must be accumulated over many years and cannot be recovered after a short period of recharge. Therefore, the long-term management of groundwater resources is an important issue for the sustainable development. The accurate prediction of groundwater levels is the first step to evaluate the total water resources and its allocation.

However, in the process of data collection, data may be missing due to various factors. Thus, retracting the missing data is a main problem which any research field must deal with. It has been well known that to maintain the data integrity, one of the effective approaches is to choose missing value imputation (MVI) for tackling the problem. In addition, it has been demonstrated that the method of the machine learning may be a better tool. Therefore, the main purpose of this study is to utilize a generative adversarial network (GAN) that consists of a generative model and a discriminative model for imputation. Our result shows that GAN can improve the accuracy of water resource evaluations.

In the current study, two interdisciplinary deep learning methods, Univariate and Seq2val, are used for groundwater level estimation. In addition to addressing the significance of the parameter conditions, the advantages and disadvantages of these two models in hydrological simulations are also discussed and compared. Finally, Seq2seq is employed to examine the limit of the models in long-term water level simulations. Our result suggests that the interdisciplinary deep learning approach may be beneficial for providing a better evaluation of water resources.

Keywords: GAN，CNN，LSTM，Imputation，Groundwater prediction

INTRODUCTION

Due to geographical and hydrological environments, the problem that water resources are not sufficient has been a central issue in Taiwan. Such insufficient supply of surface water is conventionally supported by groundwater. However, the groundwater is a resource that must be accumulated over many years. Furthermore, excessive extraction of groundwater for a long time would lead to the land subsidence. Hence there remains a need to maintain the long-term water management. In order to be able to effectively allocate water resources, it requires to accurately predict the groundwater levels. Nowadays, artificial neural networks (ANNs) have been widely used in hydrological research as an important tool for the groundwater management. However, in recent decades, the development of artificial intelligence has gradually changed from machine learning to deep learning.

Therefore, in order to address the problem of missing hydrological observation data, the purpose of the present study is to use the generative adversarial network (GAN) for imputation. The complete hydrological data could improve the accuracy of the prediction of groundwater level. Accordingly, the performance of groundwater level prediction is compared each other using two interdisciplinary methods. One is the long short-term memory (LSTM) specially developed for sequence data while the second model is the convolutional neural network (CNN), which is good at processing image information.

MATERIALS AND METHODS

In order to investigate the long-term groundwater variability, the measured groundwater level in the alluvial fan of Choushui River from 2002 to 2020 is used as the examined data. First, the imputation of missing data is performed. For this purpose, five years of data are used to train GAN. After our preliminary results are successfully verified, the missing part of the 20-year groundwater level data could be further imputed. In the current study, Univariate and Seq2val are employed using the groundwater level and rainfall as hydrological parameters, which are inputted to CNN and LSTM, respectively. Then, the meaning of parameters as well as the pros and cons of the model are discussed. Finally, the Encoder-Decoder is incorporated as to explore the limit of long-term simulated groundwater level in Seq2seq.

RESULT AND DISCUSSION

The imputation method using a generative adversarial network (GAN) is composed of a generative model and a discriminative model to fill in the missing data. Our results show that the trend of the sequence region could be reasonably simulated in the smooth section. Although some of the extreme values could not be captured in the undulating section of the groundwater level curve, there still bears a certain trend so that the feasibility of the model could be determined. Then CNN and LSTM are found to be excellent for the hydrological estimation. The coefficient of determination for both is calculated to be around 0.99, as indicated in Table 1 that also lists the value of RMSE and MAE. On closer inspection, as compared to LSTM, CNN is shown be to slightly better in the evaluation index. It is also revealed that both models underestimate the prediction of the long-term performance. Indeed, CNN outperforms LSTM for a shorter stride, but for a longer stride, the drop of accuracy in CNN is observed to be faster than that in LSTM.

Table 1.Evaluation index of CNN and LSTM

	RMSE	MAE	R2
Univariate-CNN	0.007	0.005	0.998
Univariate-LSTM	0.008	0.005	0.997
Seq2val-CNN	0.0321	0.0194	0.9981
Seq2val-LSTM	0.0508	0.0342	0.9955

CONCLUSION

Although the GAN could not capture the groundwater level endpoints in the violent section, the overall simulation performance still is excellent to some extent. We note that if the characteristics of other models can be incorporated into the basis of the GAN architecture, the performance thus can be improved to be faster. From the perspective of the hydrological simulation, CNN, which is good at processing the two-dimensional image information, is shown to be better in two aspects of accuracy and speed. We also observe that the model originally developed for 2D data still has a good performance when applied to 1D data. Our result suggests that the interdisciplinary deep learning approach may be beneficial for providing a better evaluation of water resources.

研究成果

水文觀測資料因各種原因而缺乏完整性，本研究主旨以深度學習的方法改善填補資料的品質，有利於提高地下水位預測的準確度，並使用兩種跨領域的深度學習方法對地下水水位預測的性能進行比較，降低未來相關研究選擇工具之侷限性。

首先，使用深度學習領域的巨擘 GAN，創建無監督式學習模型框架，結合生成模型（Generative Model）和鑑別模型（Discriminative Model）構成的生成對抗網絡（Generative Adversarial Network, GAN）對缺失數據進行填補，隨機取樣作為輸入，透過兩網絡相互對抗、不斷調整參數，使結果盡量模仿真實樣本，最終使鑑別模型無法判斷生成模型的輸出結果是否真實，大幅減少資料量需求且不依賴任何預先假設，以此改善填補資料的品質，以利提高地下水位預測的準確度。而研究結果表明，依照土壤性質與含水特性而區分出的扇頂、扇央、扇尾中，各站日水位均受到不同程度的噪音影響，然而由於模型以加入高斯隨機噪音進行訓練為主，因此具有一定的抗噪能力，於樣本不多的情形下，以非監督式的方法增強模型的學習能力。記憶時序列變化特徵進行補遺，於平滑、無大幅度變化的區段皆能保有序列趨勢，且評估指標 R2 表表現優異，代表模型能有效補遺於該特性區段，且於日水位曲線起伏明顯區，各站序列變化雖然有相似度，仍有局部變化差異使得補遺誤差稍大，不過整體而言，雖然並無與原始資料完全相同，但補遺後仍然有一定趨勢，沒有破壞整體時序列變化，因此能確定該模型的可行性。

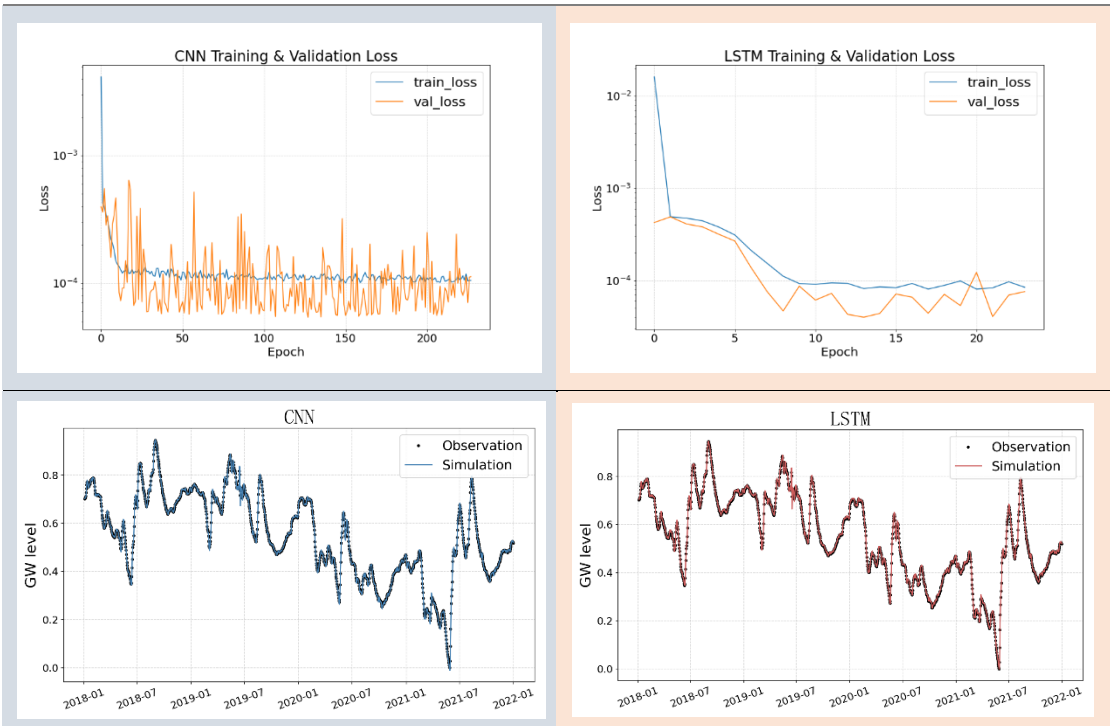

圖 1. Univariate-CNN、Univariate-LSTM

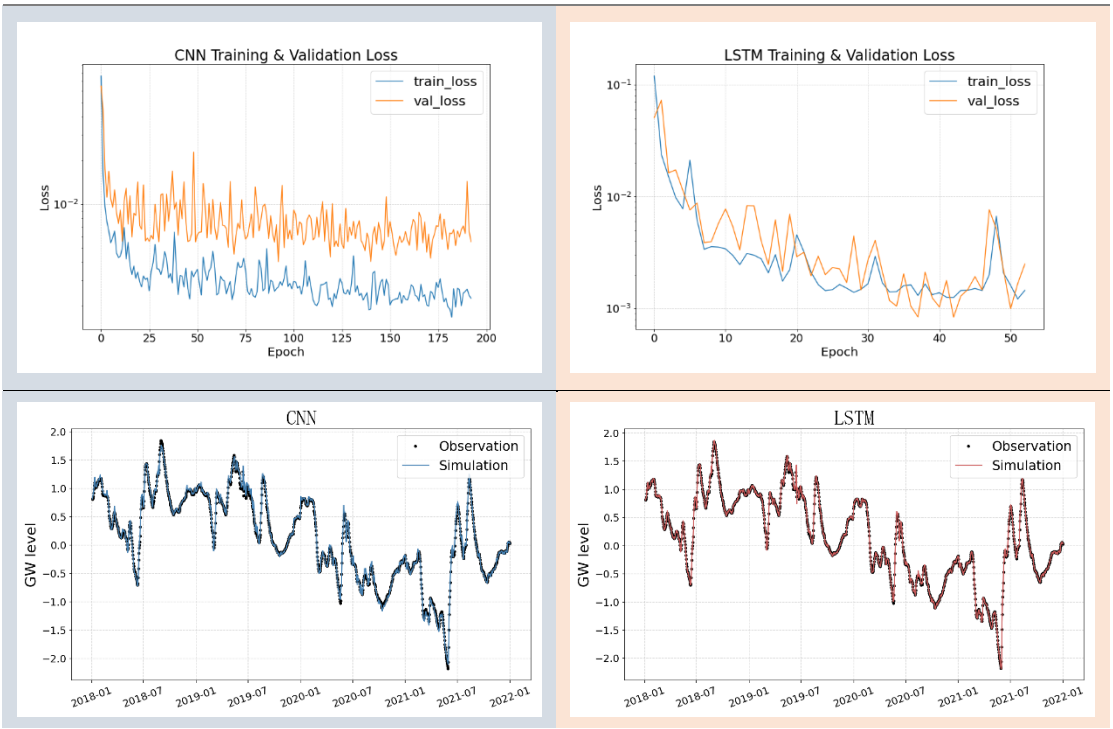

圖 2. Seq2val-CNN、Seq2val-LSTM

再評估兩種跨領域的深度學習方法 CNN 和 LSTM 模型，發展對地下水位的模擬並對精度與性能進行比較，探討輸入參數意義與兩模型於長時間預測的極限。雖然整體形式與模型均表現優異，但依循細微差異，能於 Univariate 和 Seq2val 預測的情況探討參數意義，輸入自相關強的地下水位模擬效果較好，且能察覺 CNN 的表現於評估指標上均略優於 LSTM，性能方面以 Seq2seq 進行測試，CNN 於短步長內評估指標優於 LSTM，當預測時間步長為 10 時精度的下降比 LSTM 還快，兩者均於長步長預測有平均低估的現象，以此表示 CNN 和 LSTM 預測性能之極限。結論，普遍認為適用於序列資料的 LSTM，擅長於處理跨領域資訊的 CNN 模擬更勝一籌，且模型運行速度也更快，因此能夠證明，即使原本專門為二維影像資料開發的模型，將其運用於一維資料中，也能表現出相當不錯的性能，增加序列資料研究方法的選擇，未來水文相關研究，不僅能使用專門為序列資料開發的模型，也能將在跨領域運用良好的工具列入考量，提供水資源研究更多可能性。