Research Paper Review *06/12/2024*

Spatial-Temporal Graph Neural Networks (ST-GNNs) for Groundwater Data

*Submitted as part of the **CSCI E-82** course offered by **Harvard Extension School**.*

By: Daud Waqas

Abstract:

This paper introduces a novel application of spatial-temporal graph neural networks (ST-GNNs) to predict groundwater levels. Groundwater level prediction is inherently complex, influenced by various hydrological, meteorological, and anthropogenic factors. Traditional prediction models often struggle with the nonlinearity and non-stationary characteristics of groundwater data. Our study leverages the capabilities of ST-GNNs to address these challenges in the Overbetuwe area, Netherlands. We utilise a comprehensive dataset encompassing 395 groundwater level time series and auxiliary data such as precipitation, evaporation, river stages, and pumping well data. The graph-based framework of our ST-GNN model facilitates the integration of spatial interconnectivity and temporal dynamics, capturing the complex interactions within the groundwater system. Our modified Multivariate Time Graph Neural Network model shows significant improvements over traditional methods, particularly in handling missing data and forecasting future groundwater levels with minimal bias. The model's performance is rigorously evaluated when trained and applied with both synthetic and measured data, demonstrating superior accuracy and robustness in comparison to traditional numerical models in long-term forecasting. The study's findings highlight the potential of ST-GNNs in environmental modelling, offering a significant step forward in predictive modelling of groundwater levels.

Link: https://www.nature.com/articles/s41598-024-75385-2

Context: Since this paper has to do with one of my key interests in data science (spatio-temporal modelling), I decided on exploring a type of graph-based NN that would be optimised for spatio-temporal datasets, especially those that may have temporally varying feature-sets. Though I have not yet finalised the exact set of methods which I am to cover in the final project, I have decided that I am at least going to cover STARIMA and ST-GNNs. In terms of complexity, ST-GNNs have a much more complex architecture behind them, which is why I thought they would be most relevant to cover for this paper extension. I did originally decide to do the official ST-GNN paper (refer to the first link in the alternative papers), but it was too mathematical so I instead pivoted to a case study focused on a hydrological use case (groundwater monitoring); this case study is relatively similar to what will be covered in my final project, so it'll be good to learn ST-GNNs from an "environmental modelling" perspective.

Alternative papers:

(Since there are many papers on spatio-temporal time series modelling, there were some other papers which I had in mind as well):

https://arxiv.org/abs/2110.02880 (the original; too mathematical for this extension)
https://arxiv.org/abs/2001.02250
https://arxiv.org/abs/2312.12396
https://arxiv.org/abs/2205.13504

(This YouTube video also helped in getting some of the fundamentals figured out):

https://www.youtube.com/watch?v=RRMU8kJH60Q

Extension *(from last paragraph onwards)*:

The application of spatio-temporal graph neural networks (ST-GNNs) did manage to achieve a significant advancement versus MODFLOW (a traditional numerical model) in addressing the *nonlinear and non-stationary challenges* of groundwater data modelling *(refer to Figure 6)*. The study demonstrated the robustness of ST-GNNs in handling *sparse, noisy, and hybrid datasets* through *predefined adjacency matrices* and *masked loss functions*, both of which quickly became a key point of interest when it comes to the accurate modelling of spatial and temporal dynamics. However, before addressing these key components, I believe its relevant to first and foremost cover the usage of *auxiliary data*, as it addressed one of the largest concerns in regard to environmental modelling: external influences. Groundwater systems, due to their hydrological nature, are bound to be effected by factors such as precipitation, evaporation, river stages and pumping well metrics *(within the study, datasets of all 4 were collected and processed as "auxiliary" spatial datasets, at different spatial locations from the groundwater systems; due to their different feature-sets and spatial locations, the combination of all 4 alongside the data of the groundwater systems meant that the compiled model ended up being a highly bespoke and hybrid approach; refer to Figure 1 for more details)*. Since these "external influences" wouldn't typically be measured at the same measuring stations for groundwater metrics *(due to the real-world differences in how different metrics are measured, and the conditions at which they can be measured)*, being able to combine data in this regard *(without having to line up the spatial points between auxiliary and groundwater data)* ends up being crucial for the sake of this study (and other similar environmental studies) where the collection and nature of the data, from a purely spatial perspective, will almost always be imperfect. The *predefined adjacency matrices* linked groundwater data collection points *(sometimes referred to as "wells" in the study)* to not only one another but also to the auxiliary data collection points. The *masked loss functions*, on the other hand, excluded missing or unreliable data points during training, focusing the model on robust observations, and by extension allowing it to generalise across noise and sparseness within the dataset. While this study does seem to be conducted properly, with highly useful results, it may be worth exploring the effect of assigning "edge weights" between spatial points in an attempt to improve model performance (though, then again, this would be highly taxing in terms of training computation). It would have also helped to cover more methodologies than just MODFLOW, as STARMA and well-designed feature engineering could also have the potential to produce reliable results.

Potential Improvements *(how I would take this study forward)*:

Other than the potential improvements implied within the study *(specifically "edge weights" and comparison to methodologies other than MODFLOW)*, I had an idea on an alternative method to approach ST-GNNs: decomposition. There are many ways that decomposition can be taken into account, but I believe that it will be most beneficial for identifying high-level characteristics of the time series that regular neural networks would struggle to identify. Simply put, because environmental data would typically have some metrics collected by the year and some by the minute, NNs may struggle to assign anything of "meaning" to the trends on metrics collected by the year. Providing a high-level view of *level, trend and seasonality (though Holt-Winters or another similar decomposition method)*, we could provide NNs with the "general characteristics" of that localised metric. This would also help in the handling of auxiliary data, since decompositions of those supplementary feature-sets could provide relevant information in how they correlate with the main feature-sets. Alternatively, numerical values of the decompositions could be used in conjunction to one another for *assigning "edge weights" based on level, trend and/or seasonality (once again, through Holt-Winters or a similar method)*; however, this approach would be somewhat controversial and would need some further study.

NOTE: I will attempt to apply this "decomposition + ST-GNN" idea onto my project, but this would depend on how well NN's interpret this data my approach within the final could change.