https://doi.org/10.71573/eemab371

Prediction of nitrate in different catchments using domain adaptation for regression method

Mehran Janmohammadi [1], Baiqian Shi [2] & David McCarthy[3]*

[1] BoSL Water Monitoring and Control, Department of Civil Engineering, Monash University, Wellington Road, Clayton 3800, Australia
[2] School of Civil and Environmental Engineering, Queensland University of Technology (QUT), Brisbane, Australia
[3] Canada Excellence Research Chair (CERC) in Waterborne Pathogens, School of Environmental Sciences, University of Guelph, Ontario, Canada
**Corresponding author email: David.McCarthy@uoguelph.ca*

Abstract

Surface water quality is increasingly at risk due to anthropogenic activities and climate change, leading to issues such as eutrophication that threaten aquatic ecosystems and human well-being. This study harnesses the power of Artificial Intelligence (AI), specifically deep learning and domain adaptation techniques, to predict nitrate concentrations using readily measurable parameters such as electrical conductivity (EC), pH, and temperature. We propose the Multi-Domain Adaptation for Regression under Conditional Shift (DARC) framework, designed to tackle data scarcity and marginal shifts between catchments. By incorporating a Modified Pairwise Similarity Preserver (MPSP) loss function, our model achieved an NSE value of 0.44 using only seven data points from the target dataset, outperforming traditional linear regression, which failed to reach comparable performance even with more than 20 data points. This study highlights the potential of AI-based domain adaptation methods as cost-effective, scalable solutions for water quality monitoring, addressing global environmental challenges through improved prediction and management of surface water resources.

Highlights

- DARC method achieved an NSE of o.44 with only 7 training data points.
- DARC method successfully adapted the source and target domain in two different catchments.
- DARC method with 7 data points outperformed linear regression with 19 data points.

Introduction

Anthropogenic activities and climate change have significantly degraded one of our ecosystem's most valuable resources: surface water. This deterioration can lead to harmful phenomena such as eutrophication which threatens aquatic ecosystems and ultimately our ability to utilise these waterways effectively. To address these challenges, it is essential to monitor surface water quality to facilitate the prediction and early detection of pollution events (Chen et al., 2022). However, monitoring specific chemical parameters, such as nitrate, presents significant challenges. While commercial sensors for nutrient measurement are available, their widespread application is hindered by high maintenance requirements, substantial costs, and considerable power consumption, limiting their effectiveness in diverse monitoring contexts (Castrillo & García, 2020).

AI offers a promising alternative for predicting complex relationships between water quality parameters. In particular, deep learning methods have shown significant potential in predicting intricate water quality parameters using easily measurable inputs such as EC, pH, and temperature (Peng et al., 2022). However, these methods face challenges in scenarios with limited data availability

and model transferability (Foumani, Miller, et al., 2024). One approach to overcome this limitation is using contrastive learning which, that improves the performance of model on data limited scenarios, this method has been used in various tasks such as timeseries classification (Foumani, Tan, et al., 2024). The aim of this study is to leverage the data collected from two Melbourne catchments (Yarra River and Dandenong Creek) to predict nitrate levels in two other catchments (Bass River and Merri Creek), employing a contrastive learning method to address conditional and marginal shifts. The approach utilizes a framework which creates a shared feature space with lower dimensions so the linear regression on top of the shared space generalizes to all domains. We combined this method with a modified Pairwise Similarity Preserver (MPSP) loss function to optimize performance. The MPSP loss function enables mapping the differences in water quality characteristics and nitrate ranges between the catchments into a shared feature space while preserving the distinct difference in nitrate levels of each catchment.

Methodology

The domain adaptation across different catchments was conducted in three steps: 1) Generate sample pairs from source data (Yarra River and Dandenong Creek) and a subset of the target data (Bass River or Merri Creek), 2) Train a feature extractor (FE) on these paired samples to map the water quality characteristics into a shared space while preserving the respective nitrate level differences between the pairs, and 3) Train a linear regression model on the mapped results and validate its performance using the validation dataset from the target data. The primary objective of the FE was to map data from two different catchments into a new shared space, ensuring that samples with similar nitrate levels are positioned closer to each other, while samples with differing nitrate levels remain distinct. This was achieved using the MPSP.

Sample pair generation

We randomly selected between 5 and 20 samples from the target dataset, pairing each selected sample with all 2772 records from the source dataset. For each pair, the difference in nitrate levels between the source and target samples was also computed. For example, if 20 samples were selected from the target dataset, this resulted in 55,440 pairs (each target data pairs with all source data) comprising source data, target data, and nitrate difference values. These pairs were then used to train the FE.

Feature extractor

The FE used in this study is a Neural Network designed to process six water quality parameters: temperature, dissolved oxygen saturation (DO saturated), EC, pH, turbidity, and suspended solids. The network consists of an input layer corresponding to these six parameters, followed by three hidden layers with 64, 32, and 8 nodes, and a final output layer with two nodes. The FE transforms the input features from either the source or target datasets into a two-dimensional representation. However, the transformation needs to be applied simultaneously to both sets of data. To achieve this, a Siamese Neural Network was implemented, consisting of two parallel branches for processing source and target datasets. The Siamese model takes paired samples along with their nitrate differences as inputs and outputs the distances between their transformed representations. The MPSP loss function was applied to the Siamese model to enhance the transformation of paired data into a shared feature space, ensuring that the mapped features reflect the nitrate level differences effectively.

Modified Pairwise Similarity Preserver (MPSP) loss function

The output of the FE needed to satisfy two key criteria: First, pairs with small nitrate differences (Δy) should be positioned closer together in the shared space, while pairs with large nitrate differences should be positioned further apart. Second, the distance between pairs in the shared space should be proportional to their actual nitrate difference. To achieve this, we introduced positive and negative sets for each batch during the training of the Siamese model. A pair was assigned to the positive set if

its Δy was smaller than the mean Δy of all pairs in the batch; otherwise, it was assigned to the negative set. The loss function was designed to minimize the distances between pairs in the positive set while simultaneously increasing the distances between pairs in the negative set by a value proportional to the Δy of the batch (Figure 1).

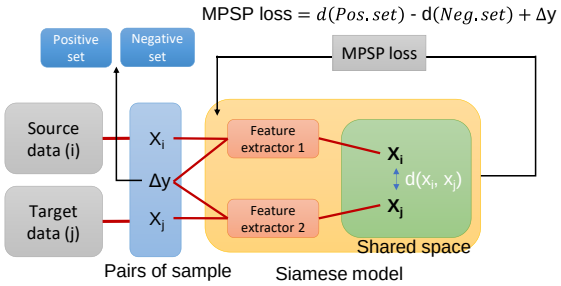

Figure 1: The schematic architecture of DARC method explained with the MPSP loss function

Model Evaluation

A linear regression model was applied to correlate the transformed features (X1 and X2) with the actual nitrate concentrations. To evaluate the performance of the DARC method, the linear regression model was trained and tested under several scenarios: we trained the linear regression using the raw source data, the raw target data used for training the FE, the transformed source data, and the transformed target data used for training the FE. The performance of these models was then compared using Nash-Sutcliffe efficiency (NSE)(1) on the validation dataset, which consisted of target data that had not been used in training the FE.

$$NSE=1-\frac{\sum_{i=1}^{n}(O_i-P_i)^2}{\sum_{i=1}^{n}(O_i-\overline{O}_i)^2} \quad (1)$$

Where $O_i$is the observed, $P_i$is the predicted water quality parameter, and $\overline{O}_i$is the mean value of observed water quality parameters.

Results and discussion

To compare the distribution of the source and target data, two parameters, EC and temperature, were plotted against nitrate concentrations in Figure 2 (a, b, e, and f). In both cases, a stronger correlation between temperature (indicating seasonality) and nitrate concentration is evident. After transforming the source and target data using the FE (Figure 2 c and g), it is clear that the FE effectively created a shared feature space with an improved correlation between the transformed features and nitrate concentrations, applicable to both the source and target datasets.

Figure 2 (d and h) illustrates the distribution of the transformed features in relation to nitrate concentrations. We are forcing any two samples to be placed in the shared space based on their nitrate difference and therefore, they form a linear line with higher nitrate concentrations located toward the bottom left of the plots and lower concentrations toward the top right. However, in both cases, certain outliers were not successfully transformed. This is likely due to their large nitrate differences, which caused them to be consistently categorized as part of the negative set during training. This limitation highlights the challenge of handling extreme outliers within the feature transformation process.

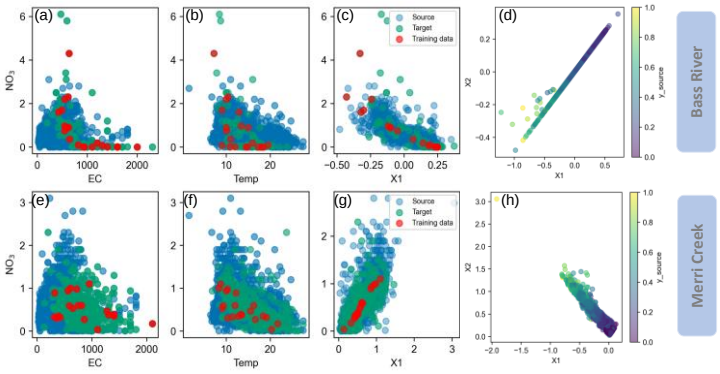

Figure 2. EC, temperature, and first transformed feature (X_1) against nitrate for Bass River (a, b, and c) and Merri Creek (e, f, and g). Distribution of nitrate (y target) across X_1 and X_2 showed in a heatmap for Bass River (d), and Merri Creek (h).

The performance of all linear regression models on the validation dataset, using 5 to 20 data points from the target dataset, is presented in Figure 3. The results reflect the median NSE values across 20 replicates. In Bass River (Figure 3a), the model trained on transformed source data outperformed other models when using up to 7 data points. However, beyond 7 data points, the model trained on transformed target data demonstrated superior performance. Notably, models trained on raw source data and transformed source data showed minimal improvement as more data points were added, whereas the performance of models trained on raw target data and transformed target data improved steadily with an increasing number of training points.

For Merri Creek (Figure 3b), a similar pattern was observed for the transformed source and raw source models, although the performance of the transformed source model was slightly lower compared to the transformed target model. These results validate the effectiveness of the DARC method, which consistently outperformed models trained on raw data from either source or target datasets. The findings demonstrate the utility of the DARC approach in scenarios with limited data availability and in addressing conditional and marginal shifts in datasets, making it a successful example of model transferability.

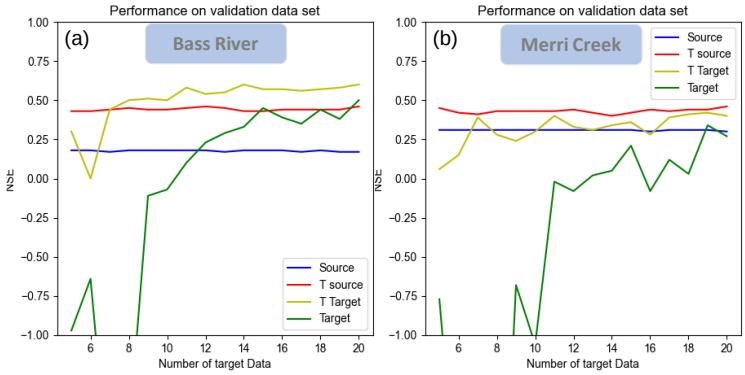

Figure 3: The performance of a linear regression fitted on raw Yarra River and Dandenong Creek (source) data, raw data (5-20) from (a)Bass River or (b)Merri Creek (target), all source data transformed to the new feature space (T source), and all target data (5 - 20), (a)Bass River or (b)Merri Creek, transformed to the new feature space (T target).

Conclusions and future work

In conclusion, this study demonstrates the effectiveness of the DARC method in addressing marginal and conditional shifts in data-limited scenarios across two different catchments. The DARC method outperformed traditional linear regression models with as few as 7 data points and successfully transformed the source and target data into a shared feature space, enabling the development of a generalized model applicable to both datasets. Future research should focus on refining this approach to further minimize the impact of outliers during data transformation, enhancing its robustness and reliability.

References

Castrillo, M., & García, Á. L. (2020). Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods. *Water Research*, *172*. https://doi.org/10.1016/j.watres.2020.115490

Chen, S., Zhang, Z., Lin, J., & Huang, J. (2022). Machine learning-based estimation of riverine nutrient concentrations and associated uncertainties caused by sampling frequencies. *PLOS ONE*, *17*(7), e0271458. https://doi.org/10.1371/JOURNAL.PONE.0271458

Foumani, N. M., Miller, L., Tan, C. W., Webb, G. I., Forestier, G., & Salehi, M. (2024). Deep Learning for Time Series Classification and Extrinsic Regression: A Current Survey. *ACM Computing Surveys*, *56*(9). https://doi.org/10.1145/3649448

Foumani, N. M., Tan, C. W., Webb, G. I., Rezatofighi, H., & Salehi, M. (2024). Series2vec: similarity-based self-supervised representation learning for time series classification. *Data Mining and Knowledge Discovery*, *38*(4), 2520–2544. https://doi.org/10.1007/s10618-024-01043-w

Peng, L., Wu, H., Gao, M., Yi, H., Xiong, Q., Yang, L., & Cheng, S. (2022). TLT: Recurrent fine-tuning transfer learning for water quality long-term prediction. *Water Research*, *225*. https://doi.org/10.1016/j.watres.2022.119171