Application of Text Mining and Machine Learning Algorithms in Describing Nutrient Transport in Groundwater-Surface Water Interactions

Ahmed Elsayed[a], Sarah Rixon[a], Christina Zeuner[a], Jana Levison[a], Andrew Binns[a], Pradeep Goel[b]

[a]School of Engineering, University of Guelph, Ontario, Canada
[b] Ministry of the Environment, Conservation and Parks (MECP), Ontario, Canada

Introduction

• Excessive release of nutrients (e.g., nitrogen and phosphorus) can cause eutrophication in lakes and rivers.
• Agricultural activities are one of the most critical non-point sources of nutrient and sediment loss due to the application of synthetic fertilizers and manure.
• There are multiple factors that can affect nutrient transport in agricultural watersheds including hydrological and field conditions.
• The Great Lakes Basin (GLB), known for intensive agricultural production, is one of the regions that suffers from extensive eutrophication due to excess release of nutrients (especially phosphorus) from agricultural fields.
• The Upper Parkhill is a clay-dominant agricultural watershed with an area of 127 km^2 within the Lake Huron Basin.

Fig. 1: Eutrophication in the Laurentian Great Lakes
(Brouwer and Liu, 2018)

Text Mining Approach and Objectives

• Conduct a meta-research on previous publications related to nutrient transport in surface and subsurface water using text mining.
• Identify and investigate the most frequent (latent) topics in the literature using quantitative and qualitative analyses.
• Determine important research gaps based on the study results.

Methods

Quantitative Analysis

Extracting relevant publications using certain keywords (e.g., nutrient transport, surface water and groundwater)
↓
Filtering the extracted publications (e.g., elimination of irrelevant publications)
↓
Pre-processing of the text datasets from the extracted publications (e.g., transformation)
↓
Building the text mining model
↓
Evaluating the performance of the developed model
↓
Determining the optimal number of topics (K)
↓
Examining the selected topics through a qualitative analysis

Fig. 2: Sequence of topic modelling approach
(Elsayed et al., ready for submission)

Main Findings

Topic Identification → Word Frequency

Nutrient transport associated with lakes
Modelling of nutrient transport
Agricultural watershed management
Sediment transport in watersheds
Nutrient transport in forest watersheds
Nitrogen transport in groundwater
Phosphorus transport in surface water
Nitrogen transport in surface water
Nutrient transport in subsurface drainage
Effect of soil litter on nutrient transport

Fig. 3: The ten identified topics
(Elsayed et al., ready for submission)

Machine Learning Approach and Objectives

• Apply different classification machine learning (ML) algorithms on a dataset of an agricultural watershed (i.e., Upper Parkhill watershed) in southern Ontario, Canada.
• Evaluate the performance of the different ML algorithms in categorizing nutrient concentrations using four evaluation metrics.
• Determine the optimal ML algorithm based on these evaluation metrics.

Methods

Input variables: Water temperature, Water DO, Water pH, Water EC, Water ORP, Surface flow, Precipitation, Crop use

Machine learning algorithms:
○ Decision Trees
○ K-nearest neighbors
○ Support vector machine
○ Discriminant models
○ Ensemble models
○ Linear regression

Evaluation metrics:
○ Classification Accuracy
○ Precision
○ Recall
○ F1 - Score

Output variables: Surface NO$_3$-N, Surface TP, Surface SRP, Surface TDP

(May et al., 2023)
Fig. 4: The Upper Parkhill Watershed

(Elsayed et al., ready for submission)
Fig. 5: Structure of the classification process

Main Findings

• The ensemble bagged trees was the optimal algorithm to classify nitrate and phosphorus concentrations in surface water.
• The highest classification accuracy was 91% while the F1 – Score was 94% using the ensemble bagged trees.
• The interdependence between the process variables was investigated.

Fig. 6: Nitrate concentrations in surface water
(Elsayed et al., ready for submission)

Conclusion and Summary

• ML and text mining are suitable algorithms for describing nutrient concentrations in surface water and groundwater.
• Using text mining, ten topics were identified as the most common topics in the field of nutrient dynamics in surface water and groundwater.
• 24 classification ML algorithms were used to categorize nutrient concentrations in surface water into two classes using a group of input variables.