VNU Journal of Science: Earth and Environmental Sciences

Journal homepage: https://js.vnu.edu.vn/EES

Original Article

Application of Machine Learning Algorithms in Studying Water Quality in the Red River System

Nguyen Quoc Son[1,*], Nguyen Cam Linh[1], Le Thi Phuong Quynh[2], Le Phuong Thu[1]

[1]*University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam*

[2]*Institute of Natural Product Chemistry, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam*

Received 07 March 2023
Revised 20 April 2023; Accepted 12 May 2023

Abstract: The Red River system plays an important role in the socio-economic development of the Northern of Vietnam. Therefore, regular monitoring and evaluation of water quality parameters in the Red River system are important in water resources management and protection. However, current monitoring methods are often quite expensive and time-consuming. To predict the downstream water quality, this study uses multiple machine learning algorithms to understand the correlation between environmental parameters measured at upstream and downstream stations of the Red River system. The environmental parameters that are chosen for this study include suspended sediment concentration (SSC), inorganic nitrogen content (total N), phosphorus content (total P), and dissolved silicon (DSi). The results show that machine learning algorithms can estimate the downstream DSi and sediment concentrations based on combining values of three upstream stations with relatively high efficiency (R^2 equals 0.75 and 0.66, respectively). Meanwhile, these algorithms have limited performance in estimating total N and P content, due to the influence of many exogenous factors. The study introduces a new direction for applying machine learning algorithms in water quality research in the Red River system with the potential application in other river systems in Vietnam.

Keywords: Water quality, nutrients, suspended sediment, machine learning, Red River system.

* Corresponding author.
E-mail address: nguyen-quoc.son@usth.edu.vn

https://doi.org/10.25073/2588-1094/vnuees.4936

Ứng dụng các thuật toán học máy trong nghiên cứu chất lượng nước tại hệ thống sông Hồng

Nguyễn Quốc Sơn[1,*], Nguyễn Cẩm Linh[1], Lê Thị Phương Quỳnh[2], Lê Phương Thu[1]

[1]*Trường Đại học Khoa học và Công Nghệ Hà Nội, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, 18 Hoàng Quốc Việt, Cầu Giấy, Hà Nội, Việt Nam*

[2]*Viện Hóa học các Hợp chất thiên nhiên, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, 18 Hoàng Quốc Việt, Cầu Giấy, Hà Nội, Việt Nam*

Nhận ngày 07 tháng 3 năm 2023
Chỉnh sửa ngày 20 tháng 4 năm 2023; Chấp nhận đăng ngày 12 tháng 5 năm 2023

Tóm tắt: Hệ thống sông Hồng có vai trò quan trọng trong phát triển kinh tế xã hội của đồng bằng Bắc Bộ. Vì vậy, theo dõi các thông số chất lượng nước thường xuyên tại hệ thống sông Hồng có ý nghĩa quan trọng trong công tác quản lý và bảo vệ nguồn tài nguyên nước. Tuy nhiên, các phương pháp quan trắc hiện nay thường khá tốn kém. Nghiên cứu này sử dụng một số thuật toán học máy trong nghiên cứu mối tương quan giữa số liệu đo đạc các thông số môi trường nước như nồng độ bùn cát lơ lửng (SSC), hàm lượng Ni tơ vô cơ tổng số (tổng N), hàm lượng Phốt Pho tổng số (tổng P) và hàm lượng Silic hòa tan (DSi) tại các trạm thượng nguồn và hạ nguồn sông Hồng, qua đó ước tính các thông số này tại các trạm hạ nguồn. Kết quả cho thấy các mô hình hồi quy đa biến có thể ước tính nồng độ DSi và bùn cát tại trạm hạ nguồn dựa trên kết hợp các giá trị của ba trạm thượng nguồn với hiệu suất tương đối cao (R^2 lần lượt bằng 0,75 và 0,66). Trong khi đó, các thuật toán học máy đã thử nghiệm có hiệu suất hạn chế trong việc ước tính hàm lượng tổng N và P, do sự tác động của nhiều yếu tố ngoại sinh. Nghiên cứu đồng thời mở ra hướng nghiên cứu áp dụng các mô hình học máy trong nghiên cứu chất lượng nước tại hệ thống sông Hồng và các hệ thống sông khác tại Việt Nam.

Từ khóa: Chất lượng nước, bùn cát lơ lửng, dinh dưỡng, học máy, hệ thống sông Hồng.

1. Mở đầu

Theo dõi và đánh giá chất lượng nước rất quan trọng trong quản lý tài nguyên nước. Các phương pháp truyền thống thường được sử dụng hiện nay để quan trắc chất lượng nước thường dựa trên các phép đo tại chỗ, thu thập mẫu và phân tích trong phòng thí nghiệm để đo các chỉ số liên quan đến các đặc tính vật lý, hóa học và sinh học của vùng nước. Các phương pháp tại chỗ nhìn chung rất chính xác, tuy nhiên khá tốn thời gian, tốn kém cả về thời gian và chi phí, yêu cầu các thiết bị đặc biệt và nhân lực được đào tạo [1].

Tải lượng bùn cát và dinh dưỡng là hai yếu tố quan trọng đóng vai trò quyết định chất lượng nước. Tải lượng bùn cát và dinh dưỡng trong môi trường nước thường được đánh giá qua nồng độ bùn cát lơ lửng (SSC), hàm lượng Ni tơ vô cơ tổng số (tổng N), hàm lượng Phốt Pho tổng số (tổng P) và hàm lượng Silic hòa tan (DSi) trong nước. Các yếu tố này góp phần vào quá trình sinh địa hóa trong hệ sinh thái dưới nước. Lượng bùn

* Tác giả liên hệ.
Địa chỉ email: nguyen-quoc.son@usth.edu.vn

https://doi.org/10.25073/2588-1094/vnuees.4936

cát và chất dinh dưỡng quá cao trong nước có thể có tác động bất lợi đến hệ sinh thái dưới nước, dẫn đến suy giảm chất lượng nước và tác động tiêu cực đến các sinh vật dưới nước. Nồng độ bùn cát lơ lửng cao có thể làm tăng độ đục của nước, làm giảm lượng ánh sáng mặt trời có thể xuyên qua nước và hạn chế sự phát triển của thực vật thủy sinh. Các nguyên tố dinh dưỡng như N, P hay Si là những yếu tố chính có thể gây ra sự phát triển quá mức của tảo và các loại thực vật thủy sinh khác, dẫn đến hiện tượng phú dưỡng và cạn kiệt oxy trong nước. Tỷ lệ N:P:Si trong môi trường nước cần phải gần bằng 16:1:16 để tránh hiện tượng phú dưỡng khi hiện tượng này có thể trở nên nghiêm trọng hơn do sự gia tăng của N và P cũng như sự suy giảm của Si [2, 3]. Hơn nữa, vận chuyển bùn cát cũng có thể liên quan đến tải trọng dinh dưỡng. Một số nghiên cứu trên sông Hồng [4, 5] cũng cho thấy tải lượng bùn cát lơ lửng có mối quan hệ chặt chẽ với các nguyên tố dinh dưỡng như P. Do đó, quản lý lượng bùn cát và chất dinh dưỡng là rất quan trọng để duy trì chất lượng nước và bảo tồn hệ sinh thái dưới nước.

Hiện nay, việc áp dụng các kỹ thuật học máy (Machine Learning) vào các lĩnh vực khoa học khác nhau đang trở thành xu thế nhằm tăng độ chính xác của các phép toán ước lượng và phân loại. Học máy là một ứng dụng của trí tuệ nhân tạo, cho phép các hệ thống tự động học và cải thiện mà không cần phải lập trình phức tạp. Các phương pháp học máy học các mối quan hệ thống kê bậc cao, do đó thường vượt trội so với các phương pháp mô hình hóa và ước tính truyền thống. Các thuật toán học máy có thể được áp dụng trong ước lượng và giám sát các tham số chất lượng nước vì nó có thể phát hiện không chỉ mối quan hệ tuyến tính mà còn cả mối quan hệ phi tuyến tính giữa các tham số.

Các thuật toán học máy ngày càng được ứng dụng nhiều trong quan trắc môi trường nước. Một nghiên cứu về hồ chứa Trị An ở Việt Nam [6] đã dự đoán về hiện tượng phú dưỡng nước mặt từ các thông số chất lượng nước thay thế khác bằng cách sử dụng công cụ hồi quy rừng ngẫu nhiên. Cụ thể, nhóm nghiên cứu đã ước tính được hàm lượng nitrit (NO_2^-), nitrat (NO_3^-) và photphat (PO_4^{3-}) từ 6 thông số sinh lý hóa bao gồm độ đục, độ dẫn điện (EC), tổng bùn cát lơ lửng (TSS), tổng bùn cát hòa tan (TDS), nhu cầu oxi hóa học (COD) và nhu cầu oxi sinh học (BOD) từ bộ dữ liệu được thu thập tại 9 điểm lấy mẫu trong hồ Trị An từ năm 2009 đến 2014. Sự phát triển của vi khuẩn trong nước, một chỉ số khác về ô nhiễm nước, cũng có thể được mô hình hóa bằng học máy. Nghiên cứu được thực hiện bởi Mohammed và cộng sự (2018) [7] đã sử dụng một số thuật toán học máy để dự đoán số lượng vi khuẩn chỉ thị chất thải trong nước từ một số thông số hóa lý sinh học, bao gồm pH, nhiệt độ, độ dẫn điện, độ đục, màu sắc, độ kiềm, coliform và E.coli.

Trong hệ thống nước chảy, các thuật toán học máy cũng có thể được áp dụng để tìm mối quan hệ giữa các thông số khác nhau của chất lượng nước. Haghiabi và cộng sự (2018) đã phân tích một bộ dữ liệu chứa các phép đo hàng tháng bao gồm nhiệt độ (T), pH, độ dẫn điện (EC), bicacbonat (HCO_3^-), sunfat (SO_4^{2-}), clorua (Cl^-), TDS, natri (Na^+), magie (Mg^{2+}), canxi (Ca^{2+}) từ năm 1960 tại 14 trạm trên dòng chính sông Tireh-Silakhor (Iran) [8]. Mỗi tham số này được dự đoán bằng cách sử dụng các tham số còn lại bằng hai thuật toán bao gồm mạng nơ-ron nhân tạo và máy vectơ hỗ trợ. Một nghiên cứu khác được thực hiện bởi Kurniawan và cộng sự (2021) được thực hiện trên hệ thống phức hợp của sông Kelantan (Malaysia) từ tháng 9 năm 2005 đến tháng 12 năm 2017 [9]. Bộ dữ liệu chứa phép đo hàng tháng của một số thông số chất lượng oxy hòa tan (DO), BOD, COD, pH, nitơ amoniac (NH_3-N) và bùn cát lơ lửng (SS). Nghiên cứu cho thấy giá trị của từng thông số được đo trong trạm ở đầu ra có thể được dự đoán bằng cách sử dụng tất cả các thông số có sẵn trong các trạm ở đầu vào của hệ thống tại cùng một thời điểm đo, trong đó thuật toán máy vector hỗ trợ cho hiệu suất tính toán cao.

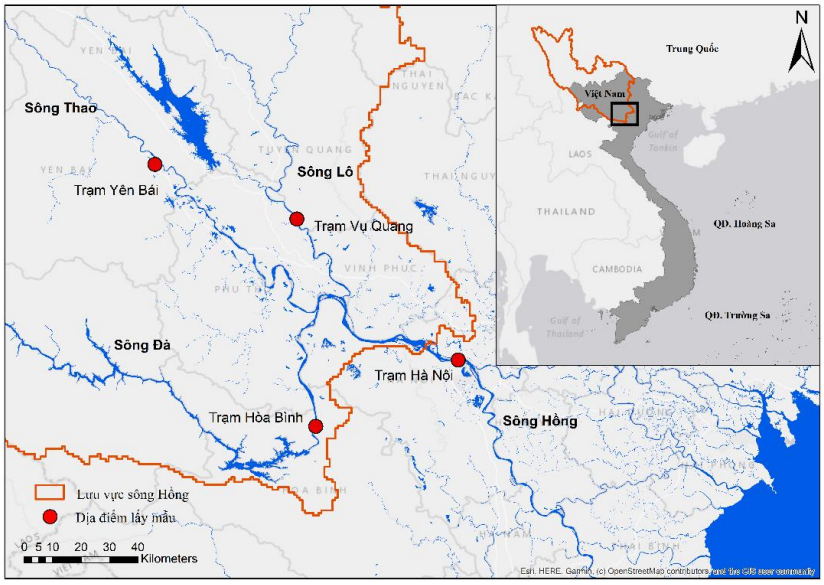
Sông Hồng là một trong những con sông chính của Việt Nam, là nguồn cung cấp nước chính cho đồng bằng châu thổ sông Hồng và các vùng lân cận. Nước sông Hồng được sử dụng cho các mục đích chính như tưới tiêu, thuỷ sản, vận chuyển hàng hóa và du lịch. Vậy nên, việc theo

dõi và đánh giá chất lượng nước tại hệ thống sông Hồng là rất cần thiết. Các phương pháp học máy đã được áp dụng để quan trắc và dự đoán chất lượng nước tại nhiều lưu vực trên thế giới với kết quả tương đối khả quan. Tuy nhiên phương pháp này chưa được ứng dụng rộng rãi trên hệ thống sông Hồng. Bên cạnh đó, các phương pháp mô hình hóa cũng đã được áp dụng trên hệ thống sông Hồng tuy nhiên chưa đáp ứng được kỳ vọng [10, 11]. Từ những lý do trên, nghiên cứu được tiến hành với mục đích sử dụng một số thuật toán học máy trong nghiên cứu mối tương quan giữa số liệu đo đạc một số thông số chất lượng nước như tải lượng bùn cát và dinh dưỡng tại các trạm thượng nguồn và hạ nguồn sông Hồng, qua đó ước tính các thông số này tại các trạm hạ nguồn. Phương pháp này hy vọng sẽ mang lại một giải pháp hữu ích cho công tác quan trắc và dự đoán chất lượng nước trong tương lai.

2. Phương pháp và khu vực nghiên cứu

2.1. Khu vực nghiên cứu

Sông Hồng bắt nguồn từ vùng núi Vân Nam, Trung Quốc, chảy qua miền Bắc Việt Nam đến Vịnh Bắc Bộ. Ở Việt Nam, đoạn sông chảy từ biên giới lãnh thổ, qua các tỉnh Lào Cai, Yên Bái, Phú Thọ được gọi là sông Thao. Sông Thao hợp lưu với hai phụ lưu chính là sông Đà và sông Lô gần Việt Trì, tỉnh Phú Thọ để hợp thành sông Hồng. Các trạm Yên Bái, Hòa Bình, Vụ Quang lần lượt nằm trên sông Thao, sông Đà và sông Lô. Từ đó, sông Hồng bắt đầu phân phối nước cho mạng lưới các sông phân lưu, bao gồm hệ thống sông Đuống, sông Đáy, sông Thái Bình.



Hình 1. Khu vực nghiên cứu.

Trong hệ thống sông Hồng, chất lượng và số lượng nước bị ảnh hưởng lớn bởi tác động của con người. Tất cả các nhánh thượng nguồn đều bị ảnh hưởng bởi các đập thủy điện. Trong hệ thống sông Hồng, đất cho cây công nghiệp chiếm ưu thế ở lưu vực sông Lô (58,1%); đất rừng chiếm ưu thế ở lưu vực sông Đà (74,4%), ruộng lúa chiếm ưu thế ở khu vực đồng bằng (66,3%), trong khi lưu vực sông Thao được đặc trưng bởi sự đa dạng lớn hơn về sử dụng đất (rừng: 54,2%; lúa nước 18,7%; cây công nghiệp 12,8%) [5].

Nghiên cứu này tập trung phân tích và dự đoán khu vực hạ lưu lưu vực sông Hồng trên lãnh thổ Việt Nam trên cơ sở số liệu thu thập tại bốn

trạm: Vụ Quang (sông Lô), Yên Bái (sông Thao), Hòa Bình (sông Đà) và Hà Nội (sông Hồng). Việc lấy mẫu được tiến hành mỗi tháng một lần từ tháng 01 năm 2012 đến tháng 12 năm 2013 tại bốn trạm đo trên [12]. Tất cả các mẫu nước sau khi thu thập được lưu giữ ở 4–10 ºC trong quá trình vận chuyển đến phòng thí nghiệm. Sau đó chúng được lọc trong 10 giờ sau khi thu thập và được bảo quản đông lạnh cho đến khi phân tích. Phân tích mẫu trong phòng thí nghiệm: hàm lượng các chỉ tiêu như NH_4^+, NO_3^-, NO_2^-, tổng P, Si hòa tan được xác định bằng phương pháp so màu trên máy đo quang UV-VIS V-630 (JASCO, Nhật Bản) theo các phương pháp tiêu chuẩn của APHA – EPA trong khi bùn cát lơ lửng được xác định dựa trên phương pháp khối lượng với giấy lọc (Whatman GF/F) theo phương pháp tiêu chuẩn của Mỹ [13]. Giá trị hàm lượng Ni tơ vô cơ tổng số trong nghiên cứu này được tính bằng tổng các giá trị của nồng độ NH_4^+, NO_3^-, NO_2^- trong nước.

Dữ liệu sau khi thu thập và xử lý gồm 24 mẫu trên 1 trạm (1 mẫu/tháng trong vòng 24 tháng), tổng cộng 4 trạm là 96 mẫu nước được thu thập và phân tích. Trong đó, các thông số đo được của mẫu lấy từ các trạm thượng nguồn (Vụ Quang, Yên Bái, Hòa Bình) sẽ được sử dụng làm biến đầu vào, và các thông số tương ứng đo được ở mẫu lấy từ trạm Hà Nội sẽ được sử dụng làm biến đầu ra. Do đó, dữ liệu cuối cùng gồm 24 điểm dữ liệu với các biến đầu ra và đầu vào được định nghĩa như trên.

2.2. Phương pháp nghiên cứu

2.2.1. Phân tích thống kê

Trong nghiên cứu này, các mẫu nước được lấy từ thượng nguồn và hạ nguồn vào cùng một thời điểm mỗi tháng được coi là một cặp. Như vậy, chúng ta có ba cặp giá trị cho mỗi tháng, bao gồm Yên Bái – Hà Nội, Vụ Quang – Hà Nội và Hòa Bình – Hà Nội. Sự khác biệt về giá trị đo được của từng thông số chất lượng nước giữa các vị trí thượng nguồn và hạ nguồn được đánh giá bằng cách sử dụng kiểm tra dấu hạng Wilcoxon [14]. Giả thuyết H_0 đặt ra là không có sự khác biệt đáng kể về các thông số chất lượng nước giữa vị trí thượng nguồn và hạ nguồn trong hệ thống sông Hồng. Giả thuyết H_0 bị bác bỏ khi giá trị p nhỏ hơn 0,05. Ngoài ra, chúng tôi cũng thực hiện kiểm tra tương quan Spearman [15] để đánh giá mức độ thay đổi phụ thuộc của các giá trị của các thông số chất lượng nước giữa các vị trí thượng nguồn và hạ nguồn.

2.2.2. Hồi quy tuyến tính

Hồi quy tuyến tính đa biến (Multiple Linear Regression - MLR) [16] ước lượng một biến phụ thuộc theo các biến độc lập, với công thức như sau:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon \quad (1)$$

Trong đó, y là biến phụ thuộc, còn $x_1, \ldots x_p$ là các biến độc lập, β_0 là giao tuyến, $\beta_1, \ \ldots \ \beta_p$ là hệ số góc tương ứng với từng biến độc lập, còn ϵ là sai số. Thuật toán của MLR xác định bộ hệ số tối ưu bằng cách tìm cực tiểu của tổng bình phương phần dư. Hồi quy tuyến tính đơn biến (Simple Linear Regression – SLR) là dạng đơn giản nhất của MLR, trong đó chỉ một biến độc lập được sử dụng để ước lượng biến phụ thuộc. Trong nghiên cứu này, mô hình hồi quy tuyến tính được thực hiện bằng thư viện cơ bản (Rbase) trong lập trình R.

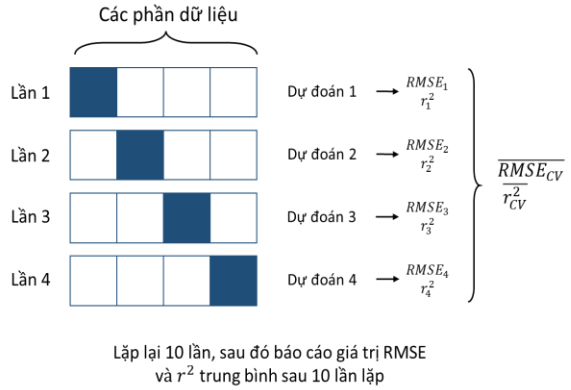
2.2.3. Rừng ngẫu nhiên

Rừng ngẫu nhiên (Random Forest - RF) là một phương pháp học máy đồng bộ dùng cho cả mục đích phân loại và hồi quy, hoạt động dựa trên việc xây dựng nhiều cây quyết định (decision tree) tại thời điểm huấn luyện mô hình [17]. Mỗi cây quyết định được xây dựng trên một tập huấn luyện được lấy mẫu ngẫu nhiên có hoàn lại từ tập huấn luyện ban đầu. Tại mỗi lần phân chia nhánh của cây quyết định, một tập hợp ngẫu nhiên gồm m trong số p biến ban đầu được chọn làm ứng cử viên để phân nhánh, trong đó chỉ một trong số m biến đó được chọn. Đối với nhiệm vụ hồi quy, giá trị trung bình của dự đoán tạo bởi tất cả các cây riêng lẻ được lựa chọn. RF được thực hiện bằng thư viện 'randomForest' trong lập trình R (phiên bản 4.6-14).

2.2.4. Đánh giá hiệu suất của các mô hình học máy

Mô hình được huấn luyện và kiểm định thông qua phương pháp kiểm định chéo bốn

nhóm. Kiểm định chéo bốn nhóm là phương pháp lấy mẫu để ước tính hiệu suất của các mô hình học máy, đặc biệt trong trường hợp dữ liệu bị giới hạn. Theo phương pháp này, dữ liệu sẽ được xáo trộn ngẫu nhiên và chia thành bốn nhóm với kích thước tương đương gần bằng nhau. Mỗi lần, một trong bốn nhóm sẽ đóng vai trò là tập kiểm định còn ba nhóm còn lại sẽ đóng vai trò tập huấn luyện để huấn luyện mô hình. Mô hình học máy được huấn luyện trên tập huấn luyện (gồm 3/4 dữ liệu) và hiệu suất sẽ được tính toán trên tập kiểm định gồm 1/4 dữ liệu còn lại. Sau bốn lần, mỗi 1/4 dữ liệu sẽ đóng vai trò tập kiểm định một lần, và hiệu suất chung của mô hình sẽ được tính toán bằng cách lấy giá trị trung bình của bốn giá trị hiệu suất (Hình 2). Phương pháp kiểm định chéo không chỉ giúp tận dụng tất cả dữ liệu có sẵn, mà nó còn giúp việc ước tính hiệu suất ít sai lệch và biến động hơn so với việc chia dữ liệu đơn giản thành hai tập huấn luyện và kiểm định. Trong nghiên cứu này, chúng tôi chọn chia dữ liệu thành 4 nhóm để: i) Tránh việc phải làm tròn số mẫu trong mỗi tập huấn luyện và kiểm định; và ii) Giúp mô hình có cả độ sai lệch và biến thiên không quá lớn [18].



Hình 2. Huấn luyện và kiểm định mô hình thông qua phương pháp kiểm định chéo 4 nhóm.

Hiệu suất ước lượng của các mô hình học máy được tính toán dựa trên tập kiểm định, nói cách khác, dựa trên các điểm dữ liệu không được dung để huấn luyện mô hình. Hiệu suất được thể hiện qua căn của sai số bình phương trung bình (RMSE) và hệ số xác định R^2. Quá trình này được lặp lại 10 lần, mỗi lần sử dụng một bộ sinh số ngẫu nhiên trong quá trình phân bổ dữ liệu trong kiểm định chéo, nhằm mục đích đánh giá hiệu suất và tính ổn định của các mô hình học máy (Hình 2).

Kết quả đánh giá chéo dựa vào hai tiêu chí là hệ số xác định R^2 và căn của sai số bình phương trung bình (RMSE). RMSE là một giá trị biểu hiện hiệu suất thường dung để mô tả sai số giữa giá trị đo được và giá trị ước lượng bởi mô hình. RMSE được tính theo công thức:

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

Trong đó: y_i là giá trị đo được và $\hat{y}_i$ là giá trị ước lượng bởi mô hình của biến phụ thuộc ở lần quan sát thứ i. RMSE càng nhỏ thì mô hình dự đoán càng đúng so với thực tế, và có độ chính xác càng cao.

Hệ số xác định (R^2) là một giá trị biểu hiện hiệu suất khác, được định nghĩa là phần biến đông của y có thể được ước lượng bởi x, được tính bởi công thức:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

R^2 chỉ nhận giá trị từ 0 đến 1, trong đó, R^2 càng gần 1 thì phần biến động của biến phụ thuộc có thể được giải thích bởi mô hình càng cao, do đó hiệu suất của mô hình càng cao.

3. Kết quả

3.1. Sự khác biệt thống kê giữa giá trị thượng nguồn và hạ nguồn của các thông số chất lượng nước trong hệ thống sông Hồng

Nghiên cứu này tập trung phân tích các thông số chất lượng nước gồm SSC, tổng N, tổng P và DSi được đo tại bốn trạm thuộc hệ thống sông Hồng, trong đó có ba trạm thượng lưu Yên Bái (sông Thao), Vụ Quang (sông Lô), Hòa Bình (sông Đà). Bảng 1 tóm tắt các giá trị trung bình của các thông số này, sự khác biệt giữa các giá trị trung bình và mối tương quan giữa các giá trị của các thông số ở từng trạm thượng nguồn với

giá trị của trạm hạ nguồn (Hà Nội). Mối tương quan được tính toán bằng phương pháp của Spearman. Giá trị p được tính toán thông qua kiểm tra dấu hạng Wilcoxon để kiểm tra có sự khác biệt về mặt thống kê của các thông số này ở các trạm thượng nguồn và hạ nguồn hay không. Kết quả cho thấy đồng thời cả bốn tham số chất lượng nước SSC, tổng N, tổng P và DSi đều khác biệt có tính thống kê giữa từng trạm đầu nguồn và trạm Hà Nội ($p < 0,05$). Bên cạnh đó, các giá trị đo đạc thể hiện mối tương quan tương đối chặt chẽ giữa các thông số ở thượng nguồn và hạ nguồn ($r > 0,6$), ngoại trừ giữa nồng độ SSC đo được ở trạm Vụ Quang và Hà Nội ($r = 0,25$), tổng N của Yên Bái và Hà Nội ($r = 0,21$), và tổng N của Vụ Quang và Hà Nội ($r = 0,54$). Kết quả cũng chỉ ra rằng tổng P có giá trị tương quan cao nhất trong số các thông số chất lượng nước (r lần lượt 0,87, 0,84, và 0,75 ở 3 trạm Hòa Bình, Vụ Quang và Yên Bái) và Hòa Bình là trạm có mối tương quan với trạm Hà Nội cao nhất.

Bảng 1. Thống kê các giá trị thông số chất lượng nước ở 4 trạm đo

Thông số	Đơn vị	Hà Nội	Yên Bái				Vụ Quang				Hòa Bình			
			TB	Sai khác*	Giá trị p**	r	TB	Sai khác*	Giá trị p**	r	TB	Sai khác*	Giá trị p**	r
SSC	mg/L	62,17	156,20	94,03	3,36E-04	0,70	33,80	-28,37	7,42E-04	0,25	10,48	-51,69	1,19E-07	0,69
Tổng N	mgN/L	0,66	0,78	0,12	7,92E-03	0,21	0,74	0,08	3,15E-02	0,54	0,49	-0,17	2,50E-05	0,81
Tổng P	mgP/L	0,18	0,30	0,12	4,94E-04	0,75	0,16	-0,02	2,29E-02	0,84	0,12	-0,06	5,68E-04	0,87
Si	mg/L	5,98	6,74	0,76	2,78E-04	0,64	5,16	-0,82	1,19E-07	0,78	6,39	0,41	8,34E-06	0,82

* Sai khác = Giá trị TB của trạm thượng nguồn – giá trị TB tại trạm Hà Nội;
** Giá trị p được tính bởi kiểm định dấu hạng Wilcoxon.

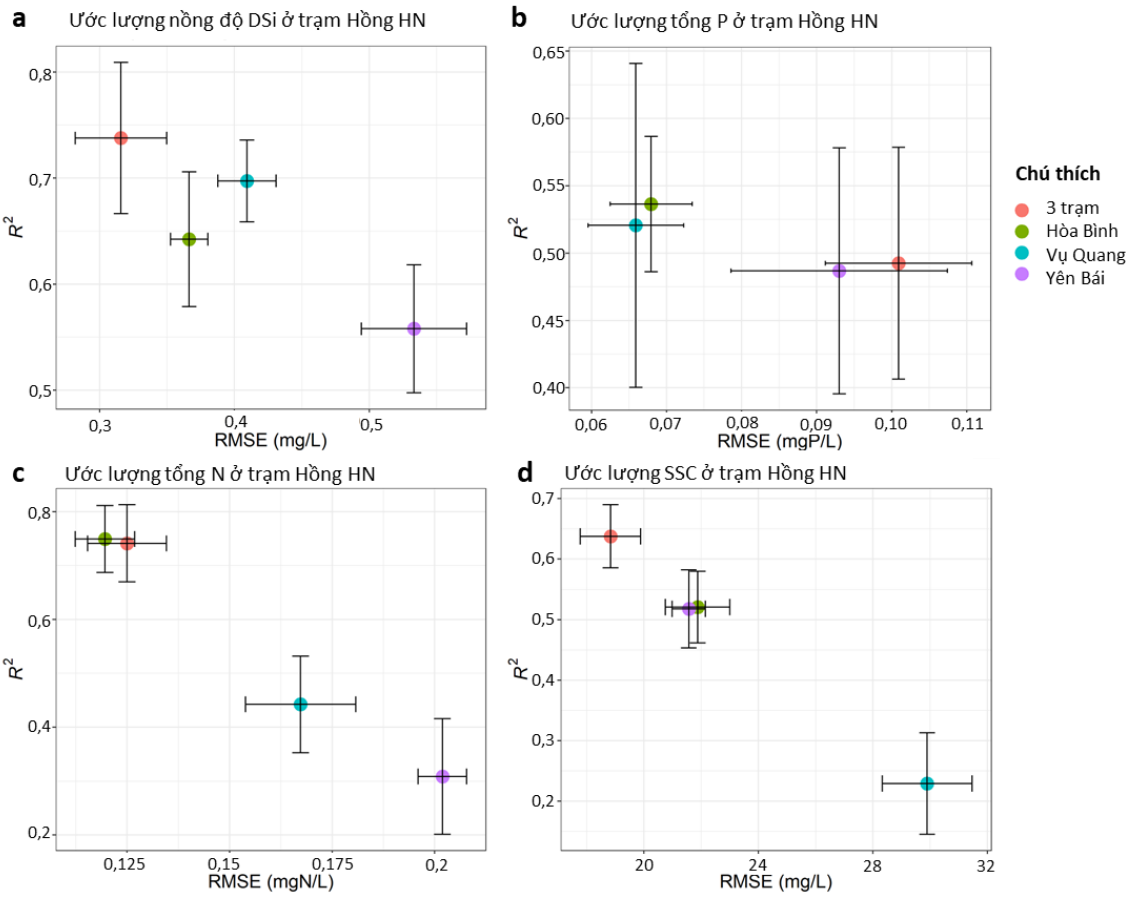
3.2. Ước tính giá trị các thông số chất lượng nước tại trạm Hà Nội bằng hồi quy tuyến tính

Sau khi phân tích thống kê, thuật toán hồi quy tuyến tính được sử dụng để ước tính giá trị của các tham số chất lượng nước ở hạ lưu bằng cách sử dụng các giá trị của tham số đó của một vị trí đầu nguồn (SLR) hoặc giá trị kết hợp của ba vị trí đầu nguồn (MLR). Nói cách khác, các giá trị của từng tham số chất lượng nước tại trạm Hà Nội được tính toán bằng hàm tuyến tính, hàm này lấy các giá trị tương ứng của một trạm thượng nguồn làm biến. Sau đó, các giá trị ước tính từ ba hàm tuyến tính (mỗi hàm được huấn luyện bằng tham số chất lượng nước từ một trong ba trạm) được so sánh với các giá trị đo được để đánh giá hiệu suất của các mô hình này.

3.2.1. Hồi quy đơn biến

Kết quả của các mô hình SLR của mỗi trong số ba trạm ngược dòng cho mỗi trong số bốn tham số quan tâm được thể hiện ở Hình 3 (xem các chấm màu xanh lục, màu xanh lam và tím đại diện cho RMSE và R^2 trung bình của SLR sau 10 lần lặp lại kiểm định chéo bốn nhóm). Đối với ước tính DSi, mô hình SLR sử dụng Si tại trạm Vụ Quang có hệ số xác định (R^2) lớn nhất, trong khi sử dụng DSi tại trạm Hòa Bình có sai số thấp nhất (RMSE nhỏ nhất) (Hình 3a). SLR sử dụng tổng P của trạm Hòa Bình có thể giải thích tốt nhất cho hệ số xác định tổng P tại trạm Hà Nội, trong khi mô hình sử dụng tổng P của trạm Vụ Quang có thể dự đoán tổng P của trạm Hà Nội với độ chính xác cao nhất (Hình 3b). Bên cạnh đó, SLR sử dụng tổng N của trạm Hòa Bình có hệ số xác định cao nhất và ước tính tổng N của trạm Hà Nội với sai số thấp nhất (Hình 3c). SLR sử dụng số liệu SSC tại trạm Hòa Bình và Yên Bái có thể dự đoán cho trạm Hà Nội với hiệu suất tương tự (Hình 3d). Kết quả cũng cho thấy rằng nếu sử dụng thuật toán SLR để dự đoán một thông số chất lượng nước hạ nguồn bằng chính giá trị của thông số đó tại thượng nguồn, chỉ có tổng N tại trạm Hòa Bình đáp ứng được yêu cầu ($R^2 > 0,7$). Phương trình hồi quy tuyến tính đơn biến ước lượng tổng N tại trạm Hà Nội từ tổng N của trạm Hòa Bình như sau:

$$N_{Hà\ Nội} = 0,8 \times N_{Hòa\ Bình} - 2,1e^{-16} \quad (4)$$



Hình 3. Hiệu suất của mô hình hồi quy tuyến tính nhằm ước tính các thông số chất lượng nước tại trạm Hà Nội, bao gồm mô hình dựa trên thông số tương ứng của trạm Hòa Bình (màu xanh lục), Vụ Quang (màu xanh lam) và Yên Bái (màu tím) và mô hình sử dụng tổng hợp thông số của 3 trạm thượng nguồn (màu cam). a) Dsi; b) tổng P; c) tổng N; d) SSC.

3.2.2. Hồi quy tuyến tính đa biến

Do trạm Hà Nội nằm ở vị trí hợp lưu của dòng chảy qua cả ba trạm thượng nguồn, nên nhóm nghiên cứu đưa ra giả thuyết các rằng thông số chất lượng nước tại trạm này bị ảnh hưởng bởi cả ba trạm Hòa Bình, Vụ Quang và Yên Bái. Do đó, chúng tôi xây dựng một mô hình MLR, lấy ba giá trị thông số chất lượng nước của các trạm thượng nguồn làm biến, để ước tính thông số chất lượng nước tương ứng của trạm Hà Nội. Hiệu suất của MLR cũng được trình bày trong Hình 3. Có thể thấy rằng, MLR đạt được hiệu suất tốt hơn so với SLR trong việc dự đoán DSi và SSC của trạm Hà Nội. Cụ thể, với trường hợp của DSi, giá trị trung bình của R^2 đã tăng từ 0,69 lên 0,74 và RMSE đã giảm từ 0,36 xuống 0,32 mg/L; với trường hợp của SSC, R^2 đã tăng từ 0,52 lên 0,64 và RMSE đã giảm từ 21,88 xuống 18,83 mg/L (xem các điểm màu cam trên Hình 3). Tuy nhiên trong trường hợp của tổng N và P, thuật toán này không tốt hơn kết quả tốt nhất trong các trường hợp với SLR. Phương trình hồi quy tuyến tính đa biến biểu hiện mối quan hệ giữa thông số DSi và SSC ở trạm Hà Nội với 3 trạm trên lần lượt như sau:

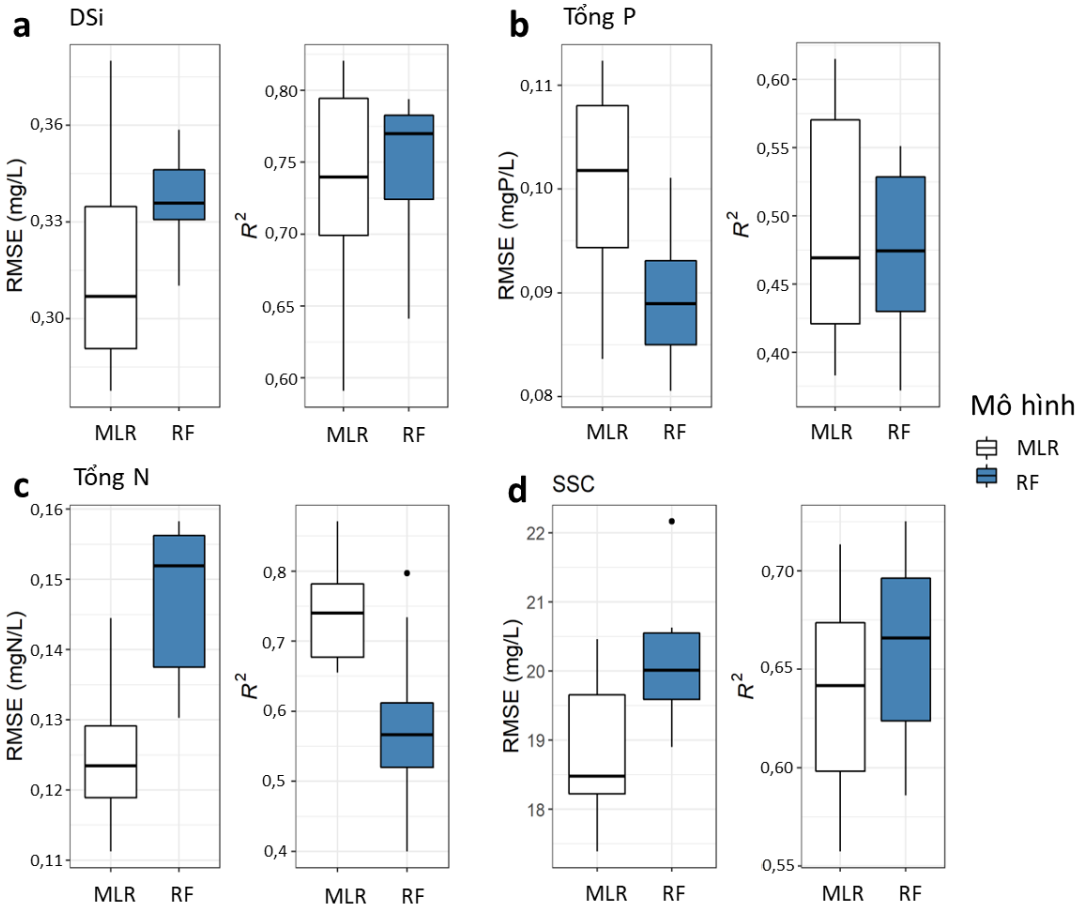
$$Si_{Hà\ Nội} = 0{,}27 \times Si_{Yên\ Bái} + 0{,}28 \times Si_{Vụ\ Quang} + 0{,}52 \times Si_{Hòa\ Bình} + 5{,}5e^{-16} \quad (5)$$

$$SSC_{Hà\ Nội} = 0{,}58 \times SSC_{Yên\ Bái} - 0{,}29 \times SSC_{Vụ\ Quang} + 0{,}49 \times SSC_{Hòa\ Bình} + 4{,}2e^{-17} \quad (6)$$

3.4. Ước tính giá trị các thông số chất lượng nước tại trạm Hà Nội bằng thuật toán RF

Trong bước tiếp theo, chúng tôi áp dụng thuật toán RF, một thuật toán hồi quy phi tuyến tính nhằm thử nghiệm khả năng nâng cao hiệu suất mô hình ước lượng thông số chất lượng nước của trạm Hà Nội, thông qua tìm kiếm mối quan hệ phi tuyến tính giữa các giá trị thông số nước thượng và hạ nguồn. Do đó, chúng tôi đã huấn luyện một mô hình RF bằng cách sử dụng các giá trị của ba trạm thượng nguồn để ước tính các giá trị thông số tương ứng của trạm Hà Nội. Kết quả của mô hình MLR và RF được trình bày và so sánh trong Hình 4.

Trong ước tính tổng P, mô hình RF đã đạt sai số nhỏ hơn trong khi có R^2 tương đương với MLR (Hình 4b). Trong ước tính DSI, mô hình RF đạt hệ số xác định R^2 cao hơn MLR, nhưng lại có sai số lớn hơn (Hình 4a). Tương tự, RF cũng đạt được hệ số xác định lớn hơn, nhưng sai số cũng lớn hơn MLR trong dự đoán SSC (Hình 4d). Tuy nhiên, các sai số của RF trong hai trường hợp này dường như có phân phối hẹp hơn (IQR nhỏ hơn), cho thấy RF có độ tin cậy cao hơn mô hình MLR trong các trường hợp này. Trong khi đó, RF đã không cho thấy bất kỳ sự cải thiện về hiệu suất nào so với MLR trong việc ước lượng tổng N của trạm Hà Nội (Hình 4).



Hình 4. So sánh hiệu suất của mô hình hồi quy tuyến tính đa biến (MLR) và Rừng ngẫu nhiên (RF) trong ước tính các thông số chất lượng nước của trạm Hà Nội dựa trên tổng hợp thông số 3 trạm thượng nguồn. a) Dsi; b) tổng P; c) tổng N, d) SSC.

4. Thảo luận

Phân tích tương quan cho thấy mối tương quan giữa các giá trị của một trạm thương nguồn với trạm Hà Nội, trong đó hàm lượng tổng N, tổng P và DSi tại trạm Hòa Bình có mối tương quan cao nhất. Điều này trước hết có thể giải thích là do lượng nước chảy từ sông Đà sang sông Hồng chiếm tỷ trọng cao nhất trong 3 nhánh [10]. Quan trọng hơn, thành phần các thông số chất lượng nước sông có thể bị ảnh hưởng lớn bởi các hoạt động nông nghiệp và công nghiệp dọc theo bờ sông. Thật vậy, một nghiên cứu do nhóm tác giả T. P. Q. Le và cộng sự trên hệ thống sông Hồng [5] chỉ ra rằng một phần lớn đất dọc theo bờ sông Thao và sông Lô được sử dụng cho mục đích nông nghiệp và công nghiệp, trong khi các hoạt động này ít xảy ra hơn ở sông Đà. Do đó, dẫn đến sự khác biệt về mức độ tương quan giữa hai trạm này và trạm Hà Nội. Trong khi đó, chúng tôi quan sát thấy xu hướng ngược lại đối với SSC, trong đó SSC của trạm Yên Bái có tương quan cao nhất với SSC của trạm Hà Nội. Không giống như các thông số khác, SSC không hòa tan và do đó bị giữ lại bởi các công trình nhân tạo trên sông như đập thủy điện. Vì các trạm quan trắc trên nhánh sông Đà và sông Lô sử dụng trong nghiên cứu này đều nằm hạ nguồn các đập Hòa Bình, Tuyên Quang và Thác Bà không xa, trong khi những đập này có khả năng giữ lại đến hơn 70% bùn cát lơ lửng [11] nên nồng độ SSC tại hai trạm này là tương đối thấp. Trong khi đó, những đập gần nhất trên sông Thao cũng cách trạm Yên Bái cả trăm cây số, tạo điều kiện cho khả năng hình thành và vận chuyển bùn cát. Như vậy, trong ba phụ lưu, sông Thao đóng góp lượng bùn cát lơ lửng lớn nhất vào sông Hồng, điều này giải thích cho mối tương quan cao nhất giữa trạm Yên Bái và Hà Nội.

Dựa trên mức độ tương quan cao giữa các trạm thượng nguồn và trạm Hà Nội, chúng tôi tiến hành các mô hình hồi quy nhằm mục đích dự đoán thông số chất lượng nước ở hạ nguồn dựa trên thông số tương ứng ở thượng nguồn. Kết quả cho thấy các mô hình hồi quy đơn biến hầu hết cho kết quả không tốt, thể hiện qua hệ số xác định thấp hoặc sai số lớn. Điều này có nghĩa là mô hình hồi quy tuyến tính đơn biến chưa phù hợp để ước lượng thông số ở hạ nguồn từ thông số ở thượng nguồn. Bên cạnh đó, biết rằng nồng độ của một thông số cụ thể có thể bị ảnh hưởng bởi lượng tương ứng của nó trong cả ba nhánh, chúng tôi đã xây dựng một mô hình hồi quy tuyến tính đa biến để ước tính các giá trị của một thông số chất lượng nước cụ thể trong trạm Hà Nội bằng cách sử dụng các giá trị tương ứng của nó tại ba trạm thượng nguồn làm giá trị biến độc lập. Kết quả cho thấy rằng phương pháp này chỉ hoạt động tốt với một số thông số nhất định, trong khi có hiệu suất kém hơn đối với các thông số khác. Trong khi SLR sử dụng tổng N và tổng P của trạm Hòa Bình thực tế cho kết quả tốt nhất trong trường hợp này, phương pháp MLR đã có thể dự đoán nồng độ DSi và SSC với sai số thấp hơn và hệ số xác định cao hơn.

Tương tự, khi nhóm nghiên cứu thử áp dụng một thuật toán hồi quy phi tuyến là RF vào ước tính thông số chất lượng nước ở trạm hạ nguồn, kết quả cho thấy thuật toán phi tuyến này nâng cao hiệu suất mô hình chỉ trong trường hợp ước tính DSi và SSC, và chỉ cải thiện được một trong hai chỉ số là RMSE hoặc R^2. Điều này chứng tỏ hiệu suất của các thuật toán hồi quy phụ thuộc rất lớn vào hoàn cảnh nghiên cứu và loại dữ liệu, tuân theo đúng định lý No-free-lunch trong Học máy và Trí tuệ nhân tạo [19]. Do đó, một cách để nâng cao hiệu suất ước lượng là tiếp tục thử nghiệm các mô hình hồi quy phi tuyến khác, ví dụ như Máy vectơ hỗ trợ (Support Vector Machines), lưới đàn hồi (Elastic Net) hoặc các thuật toán Học sâu (Deep Learning), tuy nhiên, cần có một bộ dữ liệu lớn hơn để phục vụ cho việc tối ưu hóa tham số, cũng như huấn luyện và kiểm định các mô hình dựa trên thuật toán phức tạp này.

Một nguyên nhân khác có thể dẫn đến việc các mô hình hồi quy đa biến ước tính thông số chất lượng nước của trạm Hà Nội chưa đạt hiệu suất mong muốn là thực tế là trên đoạn sông từ vị trí hợp lưu đến điểm lấy mẫu ở Hà Nội, sông Hồng có một phân lưu là sông Đuống. Do đó, nước từ nhánh chính của sông Hồng chảy sang sông Đuống có thể dẫn đến thất thoát một phần lượng N, P, DSi và bùn cát, và lượng này không được đo đạc. Vì thế, hiệu suất của mô hình đa

biến có thể được cải thiệu nếu các thông số chất lượng nước được đo trên nhánh sông này và được đưa vào mô hình.

Tóm lại, các mô hình đa biến, bao gồm cả tuyến tính và phi tuyến, đã đạt hiệu suất cao trong ước tính DSi và SSC hơn so với ước tính hai thông số còn lại tại trạm Hà Nội. Hiện tượng này có thể là hậu quả của thực tế là N và P là hai trong số các thông số chất lượng nước bị ảnh hưởng nhiều nhất bởi các hoạt động của con người. Thật vậy, N và P ven sông là kết quả của nhiều yếu tố ngoại sinh, bao gồm nước thải từ các hoạt động công nghiệp và sinh hoạt của con người, chăn thả gia súc và phân bón cho cây trồng [5]. Bên cạnh đó, cũng khó có thể tính toán lượng N và P biến đổi do các quá trình khác, chẳng hạn như lắng đọng N_2 trong khí quyển cũng như hiện tượng cố định N và P. Các nguồn đầu vào N và P vì vậy: i) Không phụ thuộc vào nồng độ N và P đo được ở các trạm thượng nguồn; ii) Có mức độ khác nhau ở mỗi nhánh; và iii) thực tế không dễ ước tính. DSi ít bị ảnh hưởng bởi các hoạt động nhân tạo hơn nhiều so với hai thông số trên [20], do đó, các giá trị của DSi ở trạm hạ lưu có thể phụ thuộc nhiều hơn vào các giá trị thượng nguồn. Bối cảnh tương tự được áp dụng cho bùn cát lơ lửng, cũng không bị ảnh hưởng nhiều bởi các hoạt động nông nghiệp và công nghiệp, ngoại trừ khai thác cát.

5. Kết luận

Nghiên cứu này đã tìm hiểu về mối quan hệ giữa các thông số chất lượng nước, bao gồm tổng N, tổng P, Si và SSC, tại các trạm thượng nguồn và hạ nguồn trong hệ thống sông Hồng, từ đó, phát hiện rằng một số thông số nhất định có thể được ước tính bằng các thuật toán hồi quy, trong đó kết hợp các thông số đo được ở các trạm thượng nguồn. Cụ thể, mô hình hồi quy đa biến sử dụng thuật toán RF đã ước tính được nồng độ Si và SSC tại trạm hạ nguồn với hệ số xác định lần lượt là 0,75 và 0,66. Điều này cũng mở ra hướng nghiên cứu mới trong việc áp dụng các thuật toán hồi quy khác nhau vào nghiên cứu và ước lượng các thông số chất lượng nước tại hệ thống sông Hồng.

Lời cảm ơn

Nghiên cứu này đã được hỗ trợ bởi dự án nghiên cứu cơ bản từ Trường Đại học Khoa học và Công nghệ Hà Nội, với mã số tài trợ USTH.WEO.01/22 cho TS. Nguyễn Quốc Sơn.

Tài liệu tham khảo

[1] W. Duan, K. Takara, B. He, P. Luo, D. Nover, Y. Yamashiki, Spatial and Temporal Trends in Estimates of Nutrient and Suspended Sediment Loads in the Ishikari River, Japan, 1985 to 2010, Sci. Total Environ., Vol. 461-462, 2013, pp. 499-508, https://doi.org/10.1016/j.scitotenv.2013.05.022.

[2] G. Billen et al., A Long-term View of Nutrient Transfers Through the Seine River Continuum, Sci. Total Environ., Vol. 375, No. 1-3, 2007, pp. 80-97, https://doi.org/10.1016/j.scitotenv.2006.12.005.

[3] R. E. Turner, N. N. Rabalais, D. Justic, Predicting Summer Hypoxia in the Northern Gulf of Mexico: Riverine N, P, and Si Loading, Mar. Pollut. Bull., Vol. 52, No. 2, 2006, pp. 139-148, https://doi.org/10.1016/j.marpolbul.2005.08.012.

[4] T. P. Q. Le, J. Garnier, G. Billen, S. Théry, C. Minh, The Changing Flow Regime and Sediment Load of the Red River, Vietnam, J. Hydrol., Vol. 334, 2007, pp. 199-214, https://doi.org/10.1016/j.jhydrol.2006.10.020.

[5] T. P. Q. Le, G. Billen, J. Garnier, S. Théry, C. Fezard, C. Minh, Nutrient (N, P) Budgets for the Red River Basin (Vietnam and China), Glob. Biogeochem. Cycles, Vol. 19, 2005, pp. 1-16, https://doi.org/10.1029/2004gb002405.

[6] H. Thang, Q. Nguyen Hao, N. Truong, L. Le, V. Thai, T. L. Pham, Estimation of Nitrogen and Phosphorus Concentrations from Water Quality Surrogates Using Machine Learning in the Tri An Reservoir, Vietnam, Environ. Monit. Assess., Vol. 192, 2020, https://doi.org/10.1007/s10661-020-08731-2.

[7] H. Mohammed, A. Longva, R. Seidu, Predictive Analysis of Microbial Water Quality Using Machine-Learning Algorithms, Environ. Res. Eng. Manag., Vol. 74, No. 1, 2018, https://doi.org/10.5755/j01.erem.74.1.20083.

[8] A. H. Haghiabi, A. H. Nasrolahi, A. Parsaie, Water Quality Prediction Using Machine Learning Methods, Water Qual. Res. J., Vol. 53, No. 1, 2018, pp. 3-13, https://doi.org/10.2166/wqrj.2018.025.

[9] I. Kurniawan, G. Hayder, H. M. Mustafa, Predicting Water Quality Parameters in A Complex River System, J. Ecol. Eng., Vol. 22, No. 1, 2021, pp. 250-257, https://doi.org/10.12911/22998993/129579.

[10] X. Wei et al., A Modeling Approach to Diagnose the Impacts of Global Changes on Discharge and Suspended Sediment Concentration Within the Red River Basin, Water, Vol. 11, No. 5, 2019, https://doi.org/10.3390/w11050958.

[11] X. Wei et al., A Modelling-Based Assessment of Suspended Sediment Transport Related to New Damming in the Red River Basin from 2000 to 2013, Catena, Vol. 197, 2021, pp. 104958, https://doi.org/10.1016/j.catena.2020.104958.

[12] T. P. Q. Le et al., Water Quality of the Red River System in the Period 2012 - 2013, J. Vietnam, Environ., Vol. 6, 2014, pp. 191-195, https://doi.org/10.13141/jve.vol6.no3.pp191-195.

[13] Apha, Awwa, and Wef, Standard Methods for the Examination of Water and Wastewater - 22nd Edition, American Public Health Association, 2012.

[14] W. J. Conover, Practical Nonparametric Statistics, 3rd Edition, Wiley. Wiley, 1999, [Online], Available: https://www.wiley.com/en-us/practical+nonparametric+statistics%2c+3rd+edition-p-9780471160687 (accessed on: March 2nd, 2023).

[15] J. L. Myers, A. D. Well, Research Design and Statistical Analysis, 2nd Ed., Lawrence Erlbaum Associates Publishers, 2003.

[16] G. James, D. Witten, T. Hastie, R. Tibshirani, Linear Regression, in An Introduction to Statistical Learning: with Applications in R, G. James, D. Witten, T. Hastie, R. Tibshirani, Eds., in Springer Texts in Statistics, New York, Ny: Springer, 2013, pp. 59-126, https://doi.org/10.1007/978-1-4614-7138-7_3.

[17] L. Breiman, Random Forests, Mach. Learn., Vol. 45, No. 1, 2001, pp. 5-32, https://doi.org/10.1023/a:1010933404324.

[18] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, 2nd Ed, 2021 Edition, New York Ny: Springer, 2021.

[19] D. Wolpert, The Supervised Learning No-free-lunch Theorems, 2001, https://doi.org/10.1007/978-1-4471-0123-9_3.

[20] T. N. M. Luu et al., N, P, Si Budgets for the Red River Delta (Northern Vietnam): How the Delta Affects River Nutrient Delivery to the Sea, Biogeochemistry, Vol. 107, No. 1, 2012, pp. 241-259, https://doi.org/10.1007/s10533-010-9549-8.