ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH VÀ TRỰC QUAN DỮ LIỆU CHẤT LƯỢNG NƯỚC TẠI ÚC VÀ XÂY DỰNG MÔ HÌNH ĐIỀN KHUYẾT DỮ LIỆU TIME SERIES

Sinh viên thực hiện		
STT	Họ tên	MSSV
1	Nguyễn Thanh Thiện Quá	20521783
2	Huỳnh Lê Phương Vy	20520951
3	Nguyễn Hiếu Nghĩa	20521654
4	Ngô Thị Phúc	20521765

TP. HỒ CHÍ MINH – 12/2022

1. GIỚI THIỆU

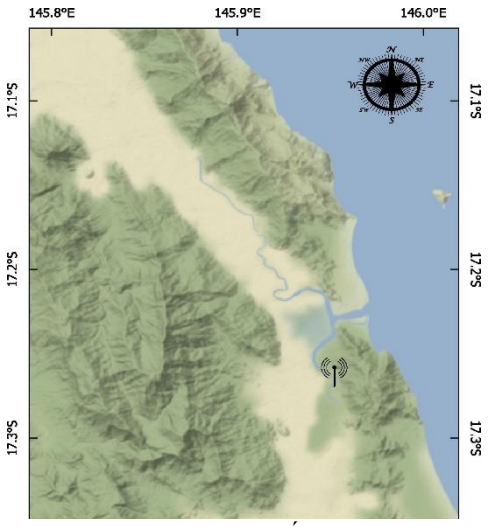
Time series (chuỗi thời gian) là chuỗi các điểm dữ liệu được đo theo từng khoảnh khắc thời gian liền nhau theo một tần suất thời gian thống nhất. Chuỗi dữ liệu thuộc loại này phần lớn được sử dụng trong Tài chính, Khoa học Tự nhiên,… và nói chung trong bất kỳ lĩnh vực nào. Ngày nay, các động lực để phân tích chuỗi thời gian và cố gắng dự đoán các giá trị tương lai của chúng có thể rất đa dạng, bao gồm cả việc tìm kiếm các xu hướng trong dữ liệu và mô hình hóa các hiện tượng tự nhiên hoặc con người. Trong đồ án môn học này, chúng tôi sẽ giải quyết hai vấn đề cơ bản trong dữ liệu chuỗi thời gian. Đầu tiên, chúng tôi sẽ phân tích và trực quan hóa bộ dữ liệu chuỗi thời gian về chất lượng nước được thu thập sông Mulgrave, Deeral, Queensland, Úc trong năm 2019, và tiến hành mở rộng bộ dữ liệu bằng cách thu thập thêm dữ liệu về thời tiết và nhiệt độ để giải thích các vấn đề đang diễn ra trong bộ dữ liệu. Tiếp theo, chúng tôi nhận thấy bộ dữ liệu này tồn tại một số điểm dữ liệu bị khuyết liên tục. Do đó, chúng tôi sẽ xây dựng mô hình điền khuyết các giá trị này sao cho gần đúng với giá trị thực tế nhất có thể. Cách tiếp cận của chúng tôi dựa trên kĩ thuật Deep Learning và Gated Recurent Units (GRUs) hai chiều. Kết quả thực nghiệm đã chứng minh mô hình của chúng tôi hoạt động khá tốt, đạt hiệu quả gấp 14 lần so với mô hình dùng kĩ thuật KNN đối với 6 giá trị khuyết liên tục.

2. NỘI DUNG

Mội dụng bài báo cáo này được trình bày theo qui trình sau: Chúng tôi sẽ giới thiệu bộ dữ liệu chất lượng nước được thu thập từ sông Mulgrave (Deeral, Queensland, Úc) và tiền xử lý trong phần 2.1, Phân tích, trực quan các yếu tố ảnh hưởng đến chất lượng nước từ bộ dữ liệu trên được trình bày trong phần 2.2, Tiến hành mở rộng dữ liệu và phân tích bằng cách thu thập thêm các bộ dữ liệu liên quan được trình bày trong phần 2.3, Xây dựng mô hình điền khuyết time series và kết quả thực nghiệm sẽ trình bày trong phần 2.4.

2.1. Giới thiệu bộ dữ liệu

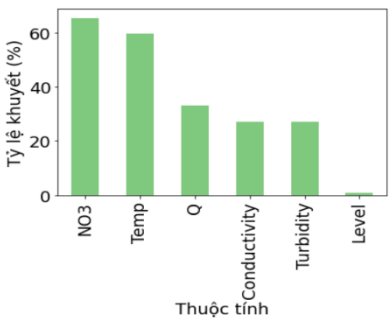
Bộ dữ liệu chất lượng nước được thu thập từ các cảm biến tự động trên sông Mulgrave (Deeral, Queensland, Úc), vị trí cụ thể thể hiện trong hình 1, gồm nhiều thuộc tính khác nhau được công bố trên trang Australian Government[1] phục vụ cho mục đích nghiên cứu. Bộ dữ liệu bao gồm 9 thuộc tính tất cả, được trình bày trong bảng 1.



Hình 1: Trạm giám sát chất lượng nước ở lưu vực Mulgrave, Australia. Biểu tượng đại diện cho trạm giám sát tại chỗ nằm ở thượng nguồn sông Mulgrave.

Tiền xử lý dữ liệu: Bộ dữ liệu thô ban đầu mà chúng tôi thu thập được là tập hợp gồm có 29051 điểm dữ liệu và 9 thuộc tính, mỗi thuộc tính là kết quả thu thập từ 1 cảm biến khác nhau được lưu trữ trong các bộ dữ liệu khác nhau. Các điểm dữ liệu mô tả giá trị được thu thập liên tục cách nhau một giờ trong vòng 5 năm liên tiếp từ 2016-2020. Tập hợp bộ dữ liệu này khuyết khá nhiều giá trị (Hình 2), như NO3 và nhiệt độ nước chiếm đến 65% và 59% giá trị khuyết, một số thuộc tính còn lại tỷ lệ khuyết vẫn rất cao. Vì thế, xử lý giá trị thiếu là việc cần thiết. Chúng tôi nhận thấy các giá trị này tập

[1] https://data.gov.au/data/



Hình 2: Trực quan dữ liệu bị khuyết

trung chủ yếu trong năm 2016-2017, nên chúng tôi sẽ không sử dụng dữ liệu trong những năm này, tuy nhiên dữ liệu này vẫn còn giá trị phân tích, sẽ được sử dụng khi cần thiết. Nguồn dữ liệu chính sẽ được sử dụng trong phân tích sẽ là dữ liệu trong năm 2019 vì chất lượng dữ liệu tốt, giá trị bị thiếu khá ít nên sẽ được xử lý bằng phương pháp thay bằng giá trị trung bình của thuộc tính đó. Sau khi đã có bộ dữ liệu sạch, chúng tôi tiến hành tách thành 2 bộ dữ liệu, bao gồm dữ liệu theo giờ và dữ liệu theo ngày để thuận tiện cho quá trình phân tích ở phần tiếp theo.

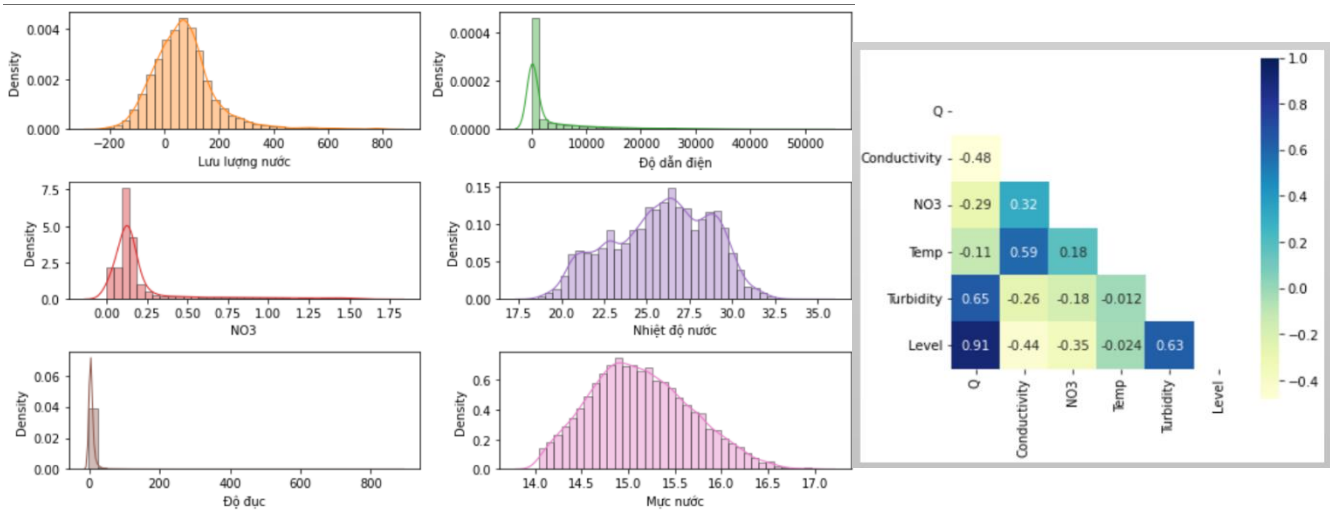
Bảng 1: Mô tả các thuộc tính của bộ dữ liệu

STT	Tên thuộc tính	Giải thích	Đơn vị
1	Temp	Nhiệt độ nước	^oC
2	NO3	Nồng độ Nitrat có trong nước	mg/l
3	Timestamp	Mốc thời gian mà các giá trị được ghi nhận	
4	Turbidity	Độ đục: Mức độ mất đi sự trong suốt của nước	NTU[2]
5	Conductivity	Độ dẫn điện: Khả năng dẫn truyền dòng điện của nước	µS/cm
6	Level	Mực nước: Độ cao của mặt nước so với 1 chuẩn đơn vị	m
7	Q	Lưu lượng nước: Lượng nước chảy qua trong 1 đv thời gian	m^3/s
8	Dayofweek	Ngày thứ mấy trong tuần đang được ghi nhận dữ liệu	
9	Moth	Số tháng trong năm mà dữ liệu được ghi nhận trong năm đó	

2.2. Phân tích và trực quan các yếu tố chất lượng nước

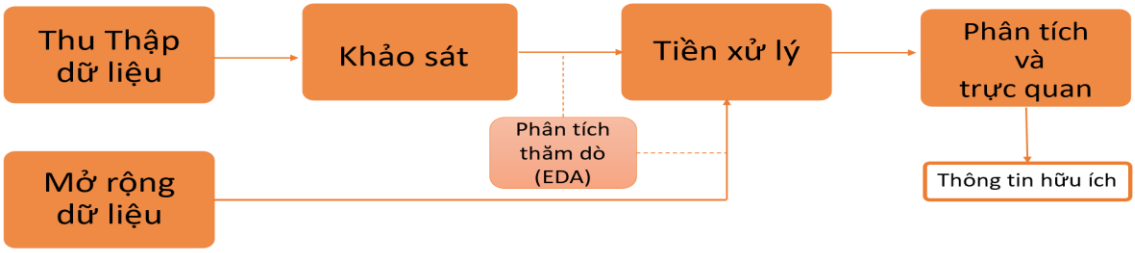
2.2.1. Tổng quan các thuộc tính

Đầu tiên, chúng tôi sẽ tổng quan sự phân bố dữ liệu và sự tương quan giữa các thuộc tính hiện tại. Quan sát hình 3a, ta thấy các biến về lưu lượng nước, nhiệt độ nước và mực nước có dữ liệu phân bố rộng khắp, đều đặn khá giống với phân phối chuẩn, các biến còn lại phân



Hình 3: a) Sự phân phối dữ liệu của các thuộc tính. b) Mối tương quan giữa các thuộc tính

[2] NTU (Nephelometric Turbidity Units): đơn vị đo độ đục khuếch tán



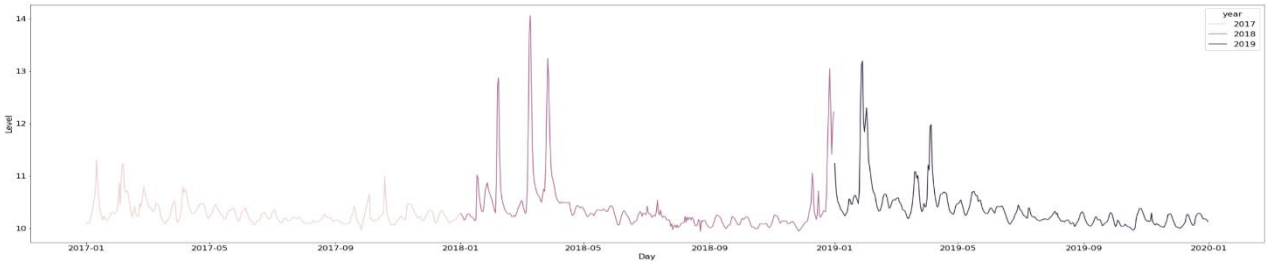
Qui trình phân tích

phối không đều, chủ yếu lệch về phía bên phải. Ngoài ra, giữa giá trị lớn nhất và nhỏ nhất của các thuộc tính cũng rất chênh lệch (bảng 4), điều này phát sinh ra một vấn đề: Liệu có giá trị ngoại lai hay không? Quan sát boxplot ta nhận thấy rằng, quả nhiên có tồn tại giá trị ngoại lại ở một số thuộc tính, tuy nhiên, để đảm bảo tính chân thực của dữ liệu, chúng tôi vẫn giữ nguyên giá trị này và phân tích chúng . Bên cạnh đó, các thuộc tính có mối tương quan khá mạnh mẽ (Hình 3b), chẳng hạn như giữa lưu lượng nước và mực nước có mối tương quan dương rất mạnh mẽ - điều này cũng dễ hiểu và gần gũi với tự nhiên khi dòng nước chảy nhanh thì mực nước khá cao. Ngoài ra một số cặp thuộc tính phụ thuộc nhau mạnh mẽ như mực nước – độ đục, lưu lượng – độ đục, nhiệt độ nước và độ dẫn điện,…

2.2.2. Tiến hành phân tích chi tiết các thuộc tính

Về nhiệt độ của nước (Temp) quanh năm luôn lớn hơn 19°C và không quá 31°C (hình), có sự thay đổi theo các tháng trong năm, khá cao vào tháng 10 cho đến tháng 4 năm sau, những tháng còn lại thì nhiệt độ nước thấp hơn nhưng không đáng kể. Giá trị trung bình và trung vị gần bằng nhau (xấp xỉ 20 °C), hình dao động nhiệt cho thấy nhiệt độ nước phân bố khá đồng đều, không có sự thay đổi bất thường nào xảy ra. So với tiêu chuẩn chất lượng nước loại II theo Quy định số 22 năm 2021 của Chính phủ Úc, hiện trạng chất lượng nước sông về các thông số nhiệt độ vẫn nằm trong giới hạn quy chuẩn chất lượng nước theo quy định.

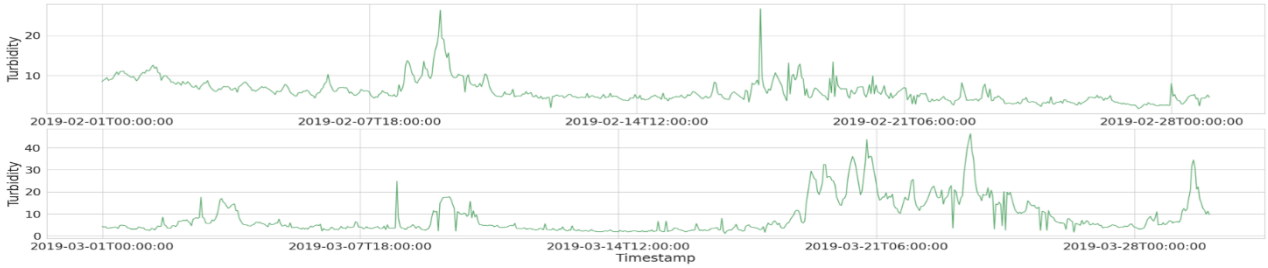
Về lưu lượng nước (Q) và **mực nước** (Level) có sự ảnh hưởng tỷ lệ thuận với nhau (hình dao động) và sự phân bố dữ liệu cũng khá tương đồng. Lưu lượng nước trung bình năm 2019 là 75.1 m^3/s, con số khá nhỏ so với lưu lượng trung bình chuẩn vì đây là thường nguồn sông, và duy trì khá đều đặn trong năm, không xuất hiện dòng chảy bất thường. Mực nước thì có sự thay đổi theo chu kì trong năm-hình dưới, mực nước dâng cao bất thường ở 4 tháng đầu năm có thể do mưa và đây là mùa lũ và duy trì đều đặn mực nước trong các tháng còn lại, tuy nhiên 4 tháng đầu năm 2017 mực nước ít hơn thất thường so nới các năm còn lại. Nhìn chung, mực nước có sự thay đổi theo mùa, theo tháng, và không có xu hướng giảm hay tăng mà vẫn giữ đều đặn qua các năm.



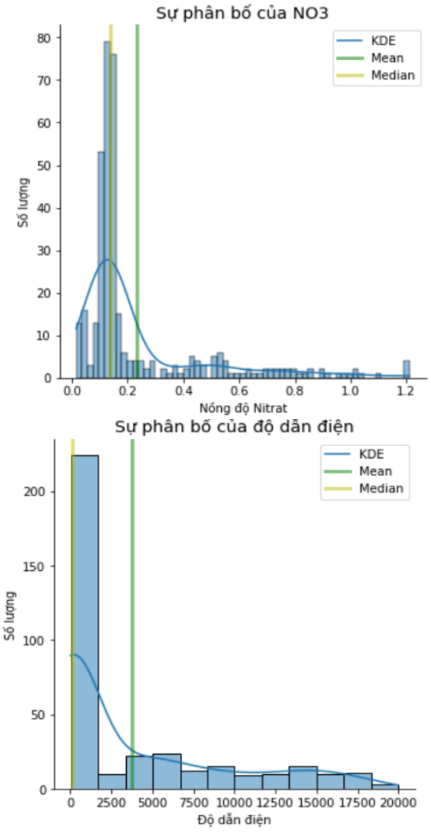
Hình 4: Mực nước sông từ năm 2017-2019

Về độ đục của nước (Turbidity) là việc nước giữ các hạt lơ lững trong nước thay vì chúng lắng xuống đáy. Vì vậy, trong các con sông trong môi trường dòng chảy tự nhiên thì độ đục

luôn hiện diện, có thể phá vỡ sự chuyển động, giảm tầm nhìn và làm tổn hại đến đời sống thủy sinh. Thời tiết và lượng mưa cũng ảnh hưởng đến dòng chảy từ đó ảnh hưởng đên độ đục được phân tích trong phần sau. Độ đục ở sông Mulgrave thấp nhất là 1.185 NTU và cao nhất là 32.59 NTU, đối với nước uống thì độ đục phải nhỏ hơn 5 NTU, tốt nhất là là dưới 1NTU và lý tưởng là dưới 0.1 NTU, từ khảo sát đó cho thấy nước sông này cần phải xử lý kĩ trước khi đưa vào tiêu dùng. Hình dưới là ví dụ của độ đục trong tháng 2 và tháng 3 trong tập dữ liệu, độ đục có sự thay đổi theo ngày và theo tháng, tầm tuần thứ 3 trong tháng thì độ đục tăng mạnh hơn những ngày còn lại.



Hình 5: Sự dao động của độ đục nước trong tháng 2 và tháng 3 năm 2019



Hình 6:Nồng độ nitrat và độ dẫn điện của nước

Về **độ dẫn điện** (Conductivity) và **nồng độ nitrat** (NO3). Quan sát biểu đồ phân bố dữ liệu hình bên, nồng độ nitrat phân bố chủ yếu từ 0.0015 đên 1,2 mg/l, tập trung chủ yếu ở 0.1 đến 0.15, và bị lệch phải. Nồng độ hợp chất Nito cao gây ảnh hưởng xấu đến hệ thủy sinh và có thể làm ảnh hưởng tới sức khỏe của con người thông qua chuỗi thức ăn. Nồng độ này khá nhỏ so với nồng độ trung bình trong các con sông, vì vậy so với tiêu chuẩn chất lượng nước cấp II theo Quy định số 22 năm 2021 của Chính phủ, nồng độ nitrat trên sông Mulgrave năm 2019 vẫn đạt chuẩn chất lượng. Các thông số tương tự với độ dẫn điện của nước. Cả độ dẫn điện và nồng độ nitrat đều có sự thay đổi lớn trong năm và có sự ảnh hưởng lẫn nhau đáng kể. Biểu đồ dao động cho ta thấy các chỉ số này khá cao vào 4 tháng cuối năm, do lượng mưa khá ít vào các tháng này , phần sau. Khi nồng độ nitrat của nước tăng dẫn đến độ dẫn điện trong nước tăng vọt theo, bên cạnh đó, độ dẫn điện cũng phụ thuộc vào nhiệt độ nước. Độ dẫn điện trong nước không gây ảnh hưởng đến con người nhưng nó ảnh hưởng đến quá trình hòa tan một số chất trong nước, ảnh hưởng đến việc nuôi trồng hải sản và chăn nuôi.

2.3. Mở rộng bộ dữ liệu và các phân tích liên quan

2.3.1. Giới thiệu 3 bộ dữ liệu mới

Qua quan sát, chúng tôi thấy rằng bộ dữ liệu chất lượng nước trên có khá ít thuộc tính và không giải thích được 1 số yếu tố ảnh hưởng bên ngoài, vì thế chúng tôi đã sưu tầm thêm các bộ dữ liệu sau:

1. Bộ dữ liệu lượng mưa trong ngày ở khu vực Deeral.
2. Bộ dữ liệu nhiệt độ cao nhất trong ngày ở khu vực Deeral.
3. Bộ dữ liệu nhiệt độ thấp nhất trong ngày ở khu vực Deeral.

Tất cả bộ dữ liệu đều được thu thập từ trang Australian Government[3]. Các bộ dữ liệu đều có cùng cách thu thập, đo lường tự động bởi các trạm thời tiết (AWSs) ở thị trấn Deeral (Queensland, Úc). Các thuộc tính trong bộ dữ liệu được trình bày trong bảng 2:

Bảng 2: Các thuộc tính trong bộ dữ liệu mở rộng

STT	Tên thuộc tính	Mô tả	Đơn vi
1. Bộ dữ liệu lượng mưa trong ngày ở khu vực Deeral (Úc)			
1	Product code	Mã sản phẩm	
2	Station number	Số hiệu trạm cục khí tượng	
3	Year	Năm mà dữ liệu được ghi lại	
4	Month	Tháng mà dữ liệu được ghi lại	
5	Day	Ngày mà dữ liệu được ghi lại	
6	Rainfall amount	Lượng mưa trong ngày	mm
7	Period	Khoảng thời gian đo	days
8	Quality	Chất lượng công việc	
2. Bộ dữ liệu nhiệt độ cao nhất trong ngày ở khu vực Deeral.			
1	Product code	Mã sản phẩm	
2	Station number	Số hiệu trạm cục khí tượng	
3	Year	Năm mà dữ liệu được ghi lại	
4	Month	Tháng mà dữ liệu được ghi lại	
5	Day	Ngày mà dữ liệu được ghi lại	
6	Maximun temperature	Nhiệt độ tối đa trong ngày	^oC
7	Period	Khoảng thời gian đo	days
8	Quality	Chất lượng công việc	
3. Bộ dữ liệu nhiệt độ thấp nhất trong ngày tại khu vực Deeral.			
1	Product code	Mã sản phẩm	
2	Station number	Số hiệu trạm cục khí tượng	
3	Year	Năm mà dữ liệu được ghi lại	
4	Month	Tháng mà dữ liệu được ghi lại	
5	Day	Ngày mà dữ liệu được ghi lại	
6	Minimun temperature	Nhiệt độ thiểu trong ngày	^oC
7	Period	Khoảng thời gian đo	days
8	Quality	Chất lượng công việc	

2.3.2. Tiền xử lý dữ liệu

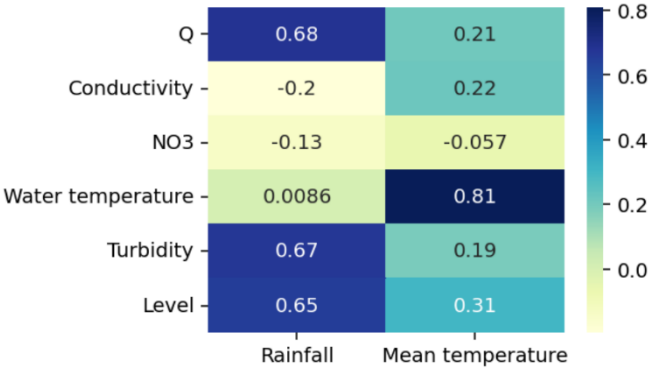
Để có thể kết hợp với data ban đầu, chúng tôi chỉ lấy ra dữ liệu từ năm 2017-2019. Dữ liệu ban đầu khá chất lượng, giá trị bị thiếu khá ít nên việc xử lý không gây ra ảnh hưởng đến phân tích. Chúng tôi kết hợp 3 thuộc tính 'Day', 'Month', 'Year' thành cột với tên mới là

[3] https://data.gov.au/data

'Datetime' và bỏ đi các thuộc tính không cần thiết cho việc phân tích. Kết quả cuối cùng chúng tôi thu được 2 bộ dữ liệu mới bao gồm: Bộ dữ liệu lượng mưa và bộ dữ liệu nhiệt độ thấp nhất + cao nhất và nhiệt độ trung bình ngày bằng trung bình cộng của nhiệt độ cao nhất và thấp nhất ở khu vụ Deeral, Úc. Sẽ kết hợp với bộ dữ liệu đã phân tích ở phần 2.2 để mở rộng việc phân tích và trực quan.

2.3.3. Phân tích và trực quan các bộ dữ liệu tổng hợp

Để tổng quát việc phân tích, chúng tôi dựa vào biểu đồ headmap (hình 7) để quan sát độ tương quan giữa các bộ dữ liệu với nhau. Qua đó, chúng tôi xem xét mức độ ảnh hưởng của các thuộc tính lượng mưa (Rainfall) trong ngày và thuộc tính nhiệt độ trung bình (Mean temperature) mỗi ngày đối với các yếu tố lượng nước. Qua quan sát, dễ thấy rằng lưu lượng nước, độ đục và mực nước bị ảnh hưởng mạnh mẽ bởi lượng mưa, nhiệt độ nước bị ảnh hưởng bởi nhiệt độ không khí. Và hình regplot một lần nữa khẳng định rằng các mối tương quan mạnh mẽ và tương quan yếu giữa các thuộc tính.

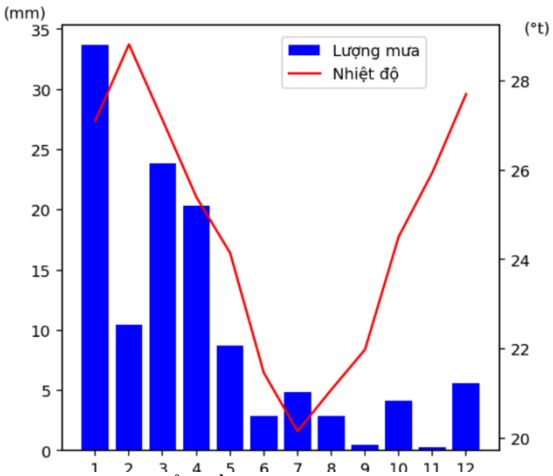


Hình 7: Mối tương quan giữa các thuộc tính tổng hợp

Bảng 3: Dữ liệu thống kê lượng mưa và nhiệt độ trung bình

STT	Thuộc tính	Min	Max	Mean	SD[4]
1	Rainfall	0.0	253.0	9.89	25.59
2	Mean temperature	13.3	37.5	24.6	3.2

Dữ liệu nhiệt độ trong ngày được phân phối đối xứng (mean ~ median) hình a. Độ lệch nghiêng của dữ liệu so với giá trị trung bình gần như bằng 0. Dựa vào đường cong trên biểu đồ, có thể thấy dữ liệu tuân theo một phân phối chuẩn và không bị ảnh hưởng nhiều bởi các giá trị ngoại lệ. Nhìn vào biểu đồ b, ta thấy được dữ liệu lượng mưa mỗi ngày được phân phối tích cực (median < mean). Độ lệch nghiêng của dữ liệu so với giá trị trung bình theo xu thế tích cực. Dựa vào đường cong trên biểu đồ có thể thấy dữ liệu đang có các giá trị chênh lệch khá lớn do ảnh hưởng của các giá trị ngoại lệ. Dựa vào biểu đồ lượng mưa tại vùng Deeral 2019 (hình 8) và kết hợp biểu đồ dao động, phần nào ta lý giải được tại sao các thông số của nước như nhiệt độ nước, nồng độ nitrat, độ dẫn điện và độ đục thay đổi thất thường. Mùa mưa vùng Deeral tập trung theo mùa, từ tháng 1 đến tháng 5 và giảm dần về cuối năm, đều này lý giải tại sao nồng độ NO3 và độ dẫn điện

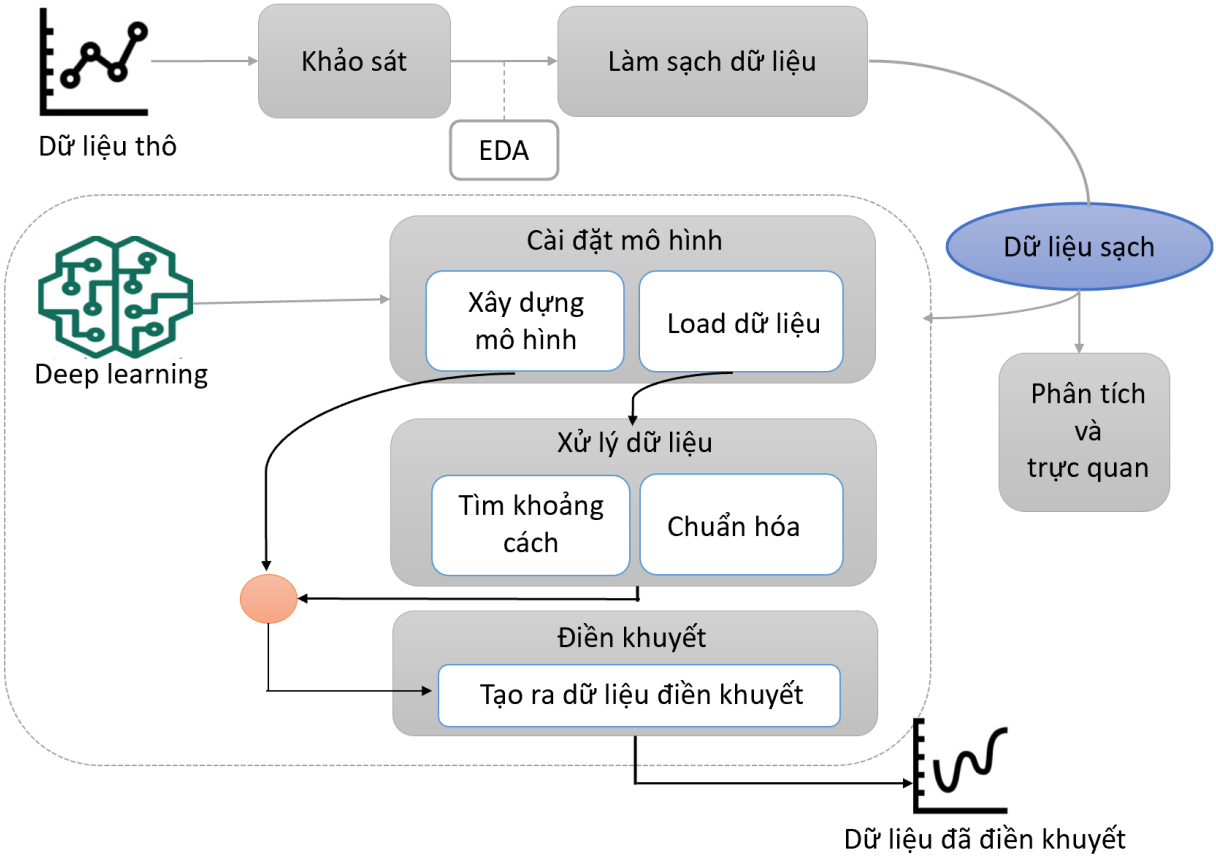


Hình 8: Biểu đồ lượng mưa và nhiệt độ tại Deeral 2019

[4] Độ lệch chuẩn

vào các tháng đầu năm rất thấp do lượng mưa lớn làm loãng các nồng độ trong nước, về cuối năm, lượng mưa giảm dần nên các nồng độ này lại tăng đột biến trở lại. Tương tự vậy, nhiệt độ nước cũng thay đổi theo nhiệt độ không khí. Về mực nước và lưu lượng nước cũng bị ảnh hưởng và dao động theo lượng mưa, điều này khá tự nhiên khi mưa lớn, một lượng nước lớn đổ vào các con sông làm các giá trị này biến động và kéo theo một lớp bùn đất trên bề mặt đổ xuống các con sông gần đó góp phần làm độ đục của nước tăng đáng kể.

2.4. Xây dựng mô hình điền khuyết time series và kết quả thực nghiệm

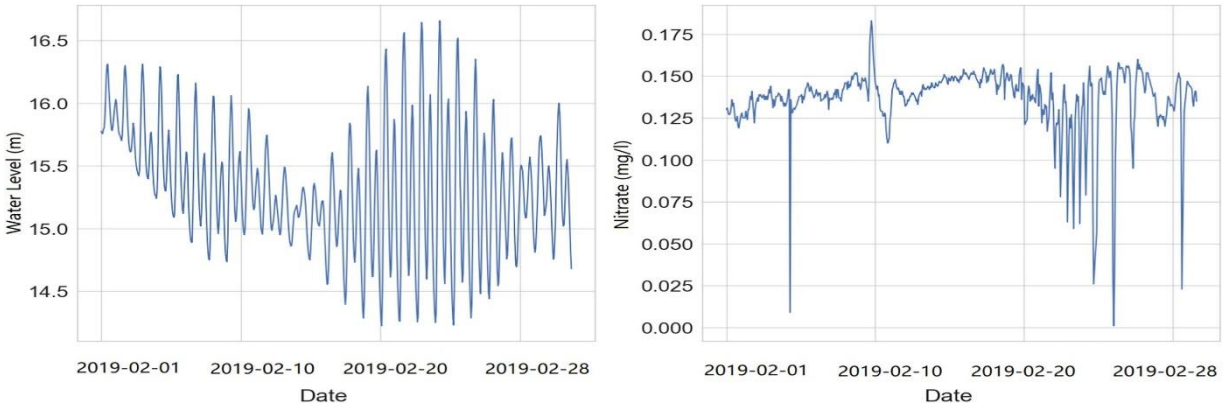


Toàn bộ quá trình phân tích và xây dựng mô hình

2.4.1. Dữ liệu chất lượng nước

Dữ liệu chất lượng nước bị ảnh hưởng bởi các quá trình tự nhiên hoặc sự can thiệp của con người, chất lượng nước không đồng nhất đối với các biến số khác nhau [1]. Chẳn hạn, độ dẫn điện trong bộ dữ liệu nằm trong một phạm vi rất rộng, giá trị tối thiểu gần bằng 0 và giá trị tối đa là hơn 50000 (µS/cm). Các tính huống tương tự có thể được tìm thấy trên các thuộc tính khác như nồng độ nitrat và độ đục, bảng 4. Do đó, chuẩn hóa dữ liệu là điều cần thiết trước khi đưa vào mô hình điền khuyết. Trong trường hợp này, chúng tôi đã tiến hành thay đổi tỷ lệ tất cả dữ liệu trong phạm vi [0,1], và loại bỏ giá trị ngoại lại tránh ảnh hưởng xấu đến kết quả mô hình. Qua quan sát (hình 9) cho thấy,mặc dù mực nước lên xuống trong tháng nhưng mực nước thay đổi rõ rệt theo ngày. Ngược lại, dữ liệu theo thời gian của nồng

độ nitrat không thể xác định được sự thay đổi theo ngày hay tuần. Việc này mang lại nhiều thách thức lớn trong việc thiết kế mô hình điền khuyết dựa trên loại dữ liệu đó.



Hình 9: Sự dao động của mực nước và nitrat theo giờ trong tháng 2-2019

Bảng 4: Dữ liệu chất lượng nước theo giờ năm 2019

Thông số	Đơn vị	Min	Max	Mean	SD
Nhiệt độ nước	^oC	18.6	32.2	24.9	2.8
Mực nước	m	14.0	17.0	15.1	0.5
Lưu lượng nước	m^3/s	247.7	670.2	75.1	108.6
Độ dẫn điện	µS/cm	0.1	50825.8	3740.4	7607.9
Độ đục	NTU	0.5	124.3	5.8	5.6
NO3	mg/l	0.001	1.7	0.2	0.3

Việc xử lý 1 điểm dữ liệu bị khuyết trong chuỗi thời gian khá đơn giản, như đã được chứng minh trong nghiên cứu[2], phương pháp hồi quy tuyến tính có thể đặt được kết quả rất tốt. Nhưng những thách thức xuất hiện khi dữ liệu bị khuyết liên tục và kéo dài. Khi kích thước khoảng trống tăng lên, hiệu suất của nhiều phương pháp quy nạp bị giảm đáng kể [3]. Hiện nay, việc thu thập dữ liệu với tần suất cao có khả năng dẫn đến những lỗ hổng lớn do nhiều yếu tố gây ra chẳng hạn như trục trặc cảm biến, lỗi truyền hoặc lỗi lưu trữ.

2.4.2. Mô hình điền khuyết dữ liệu time series hai đầu

Trong chuỗi thời gian đa biến, một số điểm của 1 thuộc tính nào đó bị thiếu liên tiếp. Mô hình chúng tôi xây dựng dựa trên mạng nơ-ron, bộ mã hóa kép GRU hai chiều sẽ khôi phục các điểm dữ liệu bị thiếu (biểu diễn bởi dấu gạch nối) với sự trợ giúp của tất cả dữ liệu có sẵn ở 2 phía nơi dữ liệu thiếu được thể hiện khái quát trong hình 10. p, k, q lần lượt là số điểm dữ liệu phía bên trái, dữ liệu thiếu và dữ liệu bên phải khoảng trống. Để xây dựng mô hình, chúng tôi xác định chuỗi thời gian đa biến **X**:

$$X = \{x^1, \dots, x^n\}^T = \{x_1, \dots, x_T\} \epsilon \mathbb{R}^{n\,x\,T} \quad (1)$$

Trong đó $x^i = \{x_1^i, \dots, x_T^i\}^T \epsilon \mathbb{R}^T$ là chuỗi thời gian thứ i và $x_t = \{x_t^1, \dots, x_T^n\}^T \epsilon \mathbb{R}^n$ là vector của n chuỗi thời gian tại thời điểm t.

Gọi M là dữ liệu bị thiếu bắt đầu từ thời điểm $p + 1$ cho đến $p + k$

$$M = \{x_{p+1}^1, \dots, x_{p+k}^1\} \epsilon \mathbb{R}^{1\,x\,k} \quad (2)$$

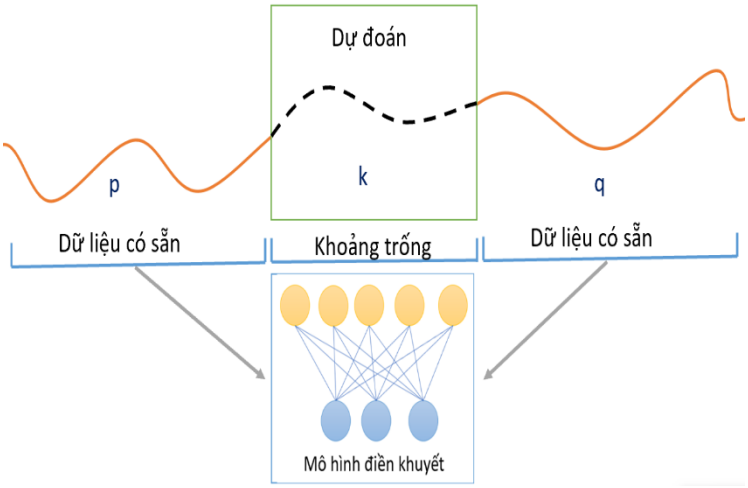
Minh họa như hình 10: dữ liệu xung quanh khoảng trống bị thiếu bao gồm thông tin có giá trị để hỗ trợ dự đoán các điểm dữ liệu bị khuyết. Đặt $L_{available}$ và $R_{available}$ là những điểm dữ liệu có sẵn ở bên trái và bên phải được thể hiện bằng phương trình sau:

$$L_{available} = \{x_1, \dots, x_p\} \in \mathbb{R}^{n\,x\,p} \quad (3)$$

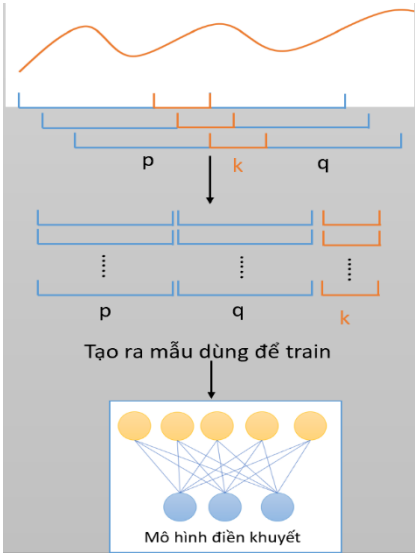
$$R_{available} = \{x_{p+k+1}, \dots, x_{p+k+q}\} \in \mathbb{R}^{n\,x\,q} \quad (4)$$

Do đó, cần có một mô hình quy nạp dự đoán các giá trị còn thiếu dựa trên các dữ liệu có sẵn

$$\widehat{M} = Model(L_{available} \cup R_{available}) \in \mathbb{R}^{1\,x\,k} \quad (5)$$



Hình 10: Minh họa bài toán điền khuyết liên tục có dữ liệu 2 đầu



Hình 11: Quá trình tạo ra mẫu đào tạo

Đối với vấn đề quy nạp trên, chúng tôi tuân thủ theo mô hình học tập có giám sát. Chiến lược cửa sổ trượt, hình 11, được áp dụng để tạo ra dữ liệu huấn luyện. Mỗi mẫu huấn luyện có thể thay đổi linh hoạt p, q và k:

Input: Chuỗi thời gian với kích thước p và q.

Biến mục tiêu: Chuỗi thời gian với kích thước k.

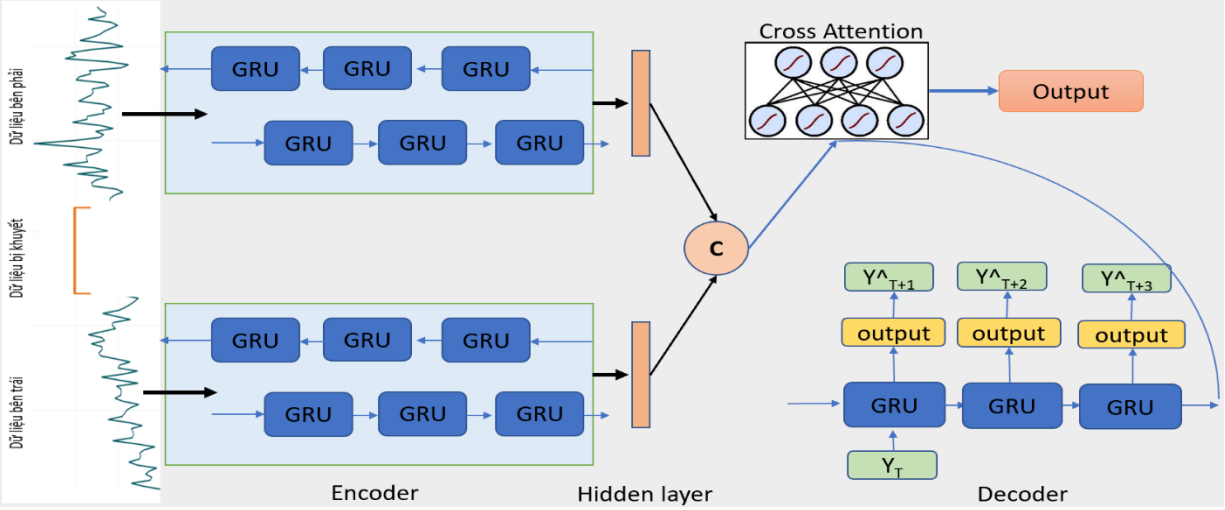
Output: Chuỗi thời gian có kích thước $p + k' + q$ với k' là *kết quả dự đoán*

Hình 12 mô tả cấu trúc mô hình chúng tôi xây dựng gồm 3 phần:

Encoder: Các bộ mã hóa kép GRU xử lý thông tin từ mỗi phía khoảng trông một cách riêng biệt. Theo cách tiếp cận này, vị trí của khoảng cách trong đầu vào chuỗi thời gian có thể được xác định một cách tự nhiên.

Cross Attention: Do mô hình có 2 bộ encoder riêng biệt nên cần thiết phải có cross attention, để nhận thông tin chuỗi đầu vào từ các bộ mã hóa. Lấy ý tưởng từ [4], hỗ trợ xử lý chuỗi thời gian từ 2 bộ mã hóa khác nhau.

Decoder: Trong bộ giải mã, các dự đoán về giá trị bị khuyết được tạo ra liên tiếp. Tại mỗi thời điểm t, GRU sẽ được cập nhật trạng thái mới dựa trên trạng thái trước đó. Tại đầu ra sẽ có 1 lớp tuyến tính để tạo ra giá trị số



Hình 12: Kiến trúc mô hình: Mỗi Encoder 2 chiều chịu trách nhiệm nắm bắt ngữ cảnh và xu hướng về dữ liệu từ 1 phía của khoảng trống, các lớp ẩn (Hidden layer) của 2 bộ mã hóa được ghép nối với nhau và xử lý bởi Cross Attention. 1 hàm softmax để dữ đoán đầu ra của Decoder.

2.4.3. *Thiết kế thực nghiệm và kết quả thực nghiệm*

Trong thí nghiệm này, chúng tôi đánh giá kết quả đạt được bằng các độ đo sau: Root Mean Square Error (RMSE), Mean Absolute Error (MAE) và Dynamic Time Warping [5](DTW). Bên cạnh đó, chúng tôi sẽ tiến hành so sánh với mô hình cơ bản KNN (K-nearest neighbour)[5] được thiết kế để tìm k điểm dữ liệu gần nhất với dữ liệu quan sát. Các siêu tham số để tối ưu hóa mô hình được trình bày trong bảng 5. Bên cạnh đó, số lượng dữ liệu có sẵn xung quanh giá trị bị thiếu cũng có thể điều chỉnh linh hoạt. Trong thí nghiệm này, để nhất quán và công bằng trong quá trình so sánh, chúng tôi chọn kích thước dữ liệu cố định xung quanh khoảng trống. Dựa trên phân tích ở bảng 4, hơn 90% các giá trị bị thiếu liên tục có kích thước nhỏ hơn 6. Hơn nữa, để bao quát thông tin hữu ít gần giá trị thiếu, chúng tôi sử dụng 10 dữ liệu có sẵn ở bên trái khoảng trống và tương tự ở bên phải làm đầu vào cho toàn bộ mô hình. Với cách cài đặt trên, mô hình được chúng tôi đánh giá là phù hợp để khôi phục dữ liệu.

Bảng 5: Các siêu tham số của mô hình

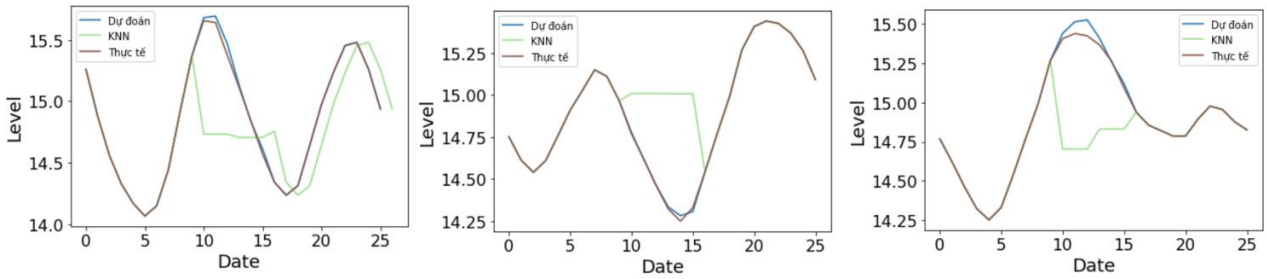
Siêu tham số	Giá trị
Số lớp ẩn cho bộ mã hóa kép	1
Số đơn vị GRU trên bộ mã hóa	50
Số lớp ẩn cho bộ giải mã	1
Số đơn vị GRU trên bộ giải mã	50
Trình tối ưu hóa	AdamW
Batch size	10

Vì công cụ chạy mô hình rất tốn kém, mất nhiều thời gian và vượt ngoài khả năng của nhóm về chi phí, nên chúng tôi chỉ tiến hành khôi phục dữ liệu cho các thuộc tính Level và NO3, với khoảng cách thiếu liên tục lần lượt là 6, 10, 20. Các thuộc tính còn lại vẫn có thể khôi phục được nếu có đủ thời gian hơn. Bảng 4 minh họa hiệu suất mô hình cho 2 thuộc tính mực nước và nồng độ nitrat.

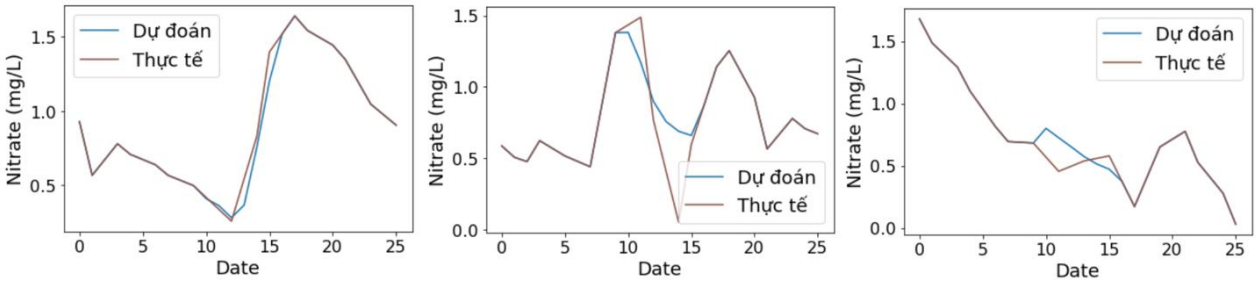
[5] https://en.wikipedia.org/wiki/Dynamic_time_warping

Bảng 3: Kết quả các độ đo đánh giá mô hình

Độ đo	Level				NO3			
Khoảng trống bị khuyết	6	KNN-6[6]	10	20	6	KNN-6	10	20
RMSE	**0.032**	0.44	0.041	0.051	**0.100**	0.552	0.147	0.185
MAE	**0.029**	0.39	0.036	0.042	**0.087**	0.483	0.125	0.153
DTW	**0.071**	1.01	0.100	0.161	**0.188**	1.206	0.344	0.712



Hình 13: Kết quả khôi phục dữ liệu Mực nước với số lượng missing data là 6



Hình 14: Kết quả dự khôi phục dữu liệu nồng độ nitrat với số lượng missing data là 6

Nhận xét: Chúng tôi tự đánh giá mô hình đã xây dựng hoạt động khá tôt để khôi phục dữ liệu bị khuyết liên tục. Dựa trên các độ đo hoạt động trên điền khuyết 6 giá trị liên tục, thuật toán của chúng tôi hiệu quả hơn gần 14 lần so với thuật toán KNN (k=10). Mô hình chúng tôi đã xây dựng xác định tốt xu hướng của dữ liệu, trong khi KNN thì không nắm bắt được. Đối với thuộc tính mực nước (Level), dữ liệu có tính chu kì dễ dàng xác định xu hướng, ngược lại đối với nồng độ nitrat (NO3) dữ liệu rất khó xác định được xu hướng (hình 9-NO3) nên mô hình không hoạt động tốt bằng. Một điểm hạn chế của mô hình là khi dữ liệu bị khuyết liên tục càng nhiều thì hiệu suất của mô hình hoạt động càng kém.

3. KẾT LUẬN

Sau quá trình thu thập dữ liệu, khảo sát và phân tích trên bộ dữ liệu chất lượng nước kết hợp với các bộ dữ liệu bổ xung có thể kết luận rằng, lượng mưa trong ngày có ảnh hưởng đến các yếu tố như mực nước, lưu lượng và độ đục của nước, giữa chúng đều có mối quan hệ tuyến tính đồng thuận. Khá thực tế, khi những ngày có lượng mưa cao thì mực nước và lưu lượng của nước tăng cao đáng kể kéo theo đó là độ đục cũng tăng đột biến. Chúng tôi đã đúc trích được một số kiến thức về chất lượng nước như lượng mưa có ảnh hưởng rất lớn đến các nồng độ trong nước, lượng mưa lớn thì nồng độ nitrat và độ dẫn điện của nước cũng giảm đi đáng kể và ngược lại, lượng mưa cũng gây sự tăng đột biến về lưu lượng chảy của nước và mực nước tăng cao, khi có mưa lớn theo mùa nên có các biện pháp đề phòng lũ.

[6] Thuật toán KNN (k=10) với dữ liệu bị thiếu liên tục là 6

Bên cạnh đó, chúng tôi đã xây dựng thành công mô hình điền khuyết dữ liệu chuỗi thời gian và đã chúng minh mô hình hoạt động hiệu quả qua thực nghiệm. Kết quả tốt nhất mà mô hình đạt được là **0.032, 0.029, 0.071** tương ứng lần lượt với các độ đo RMSE, MSE, DTW, khi điền khuyết 6 giá trị liên tiếp của thuộc tính mực nước và hoạt động tốt hơn gấp 14 lần mô hình xây dụng bằng kỹ thuật KNN với k=10. Mô hình của chúng tôi có thể mở rộng để áp dụng trên các tác vụ tương tự

TÀI LIỆU THAM KHẢO

[1] E. Ishaq S., O. Agada P., and S. Rufus, “Spatial and Temporal Variation in Water Quality of River Benue, Nigeria,” *J. Environ. Prot.*, vol. 2012, Aug. 2012, doi: 10.4236/jep.2012.328106.

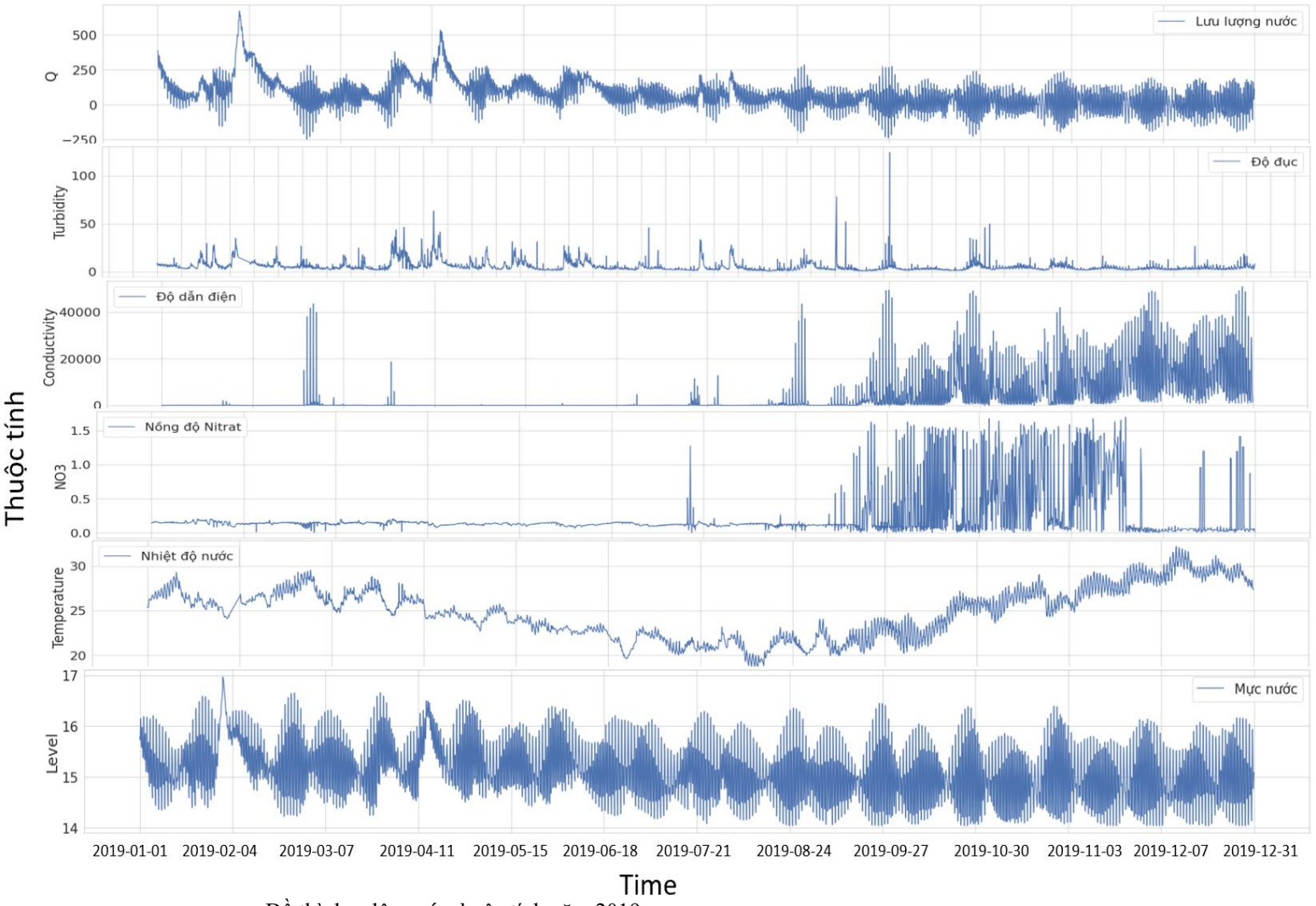
[2] “Applied Sciences | Free Full-Text | Improved Interpolation and Anomaly Detection for Personal PM2.5 Measurement.” https://www.mdpi.com/2076-3417/10/2/543 (accessed Dec. 01, 2022).

[3] A. M. Moffat *et al.*, “Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes,” *Agric. For. Meteorol.*, vol. 147, no. 3, pp. 209–232, Dec. 2007, doi: 10.1016/j.agrformet.2007.08.011.

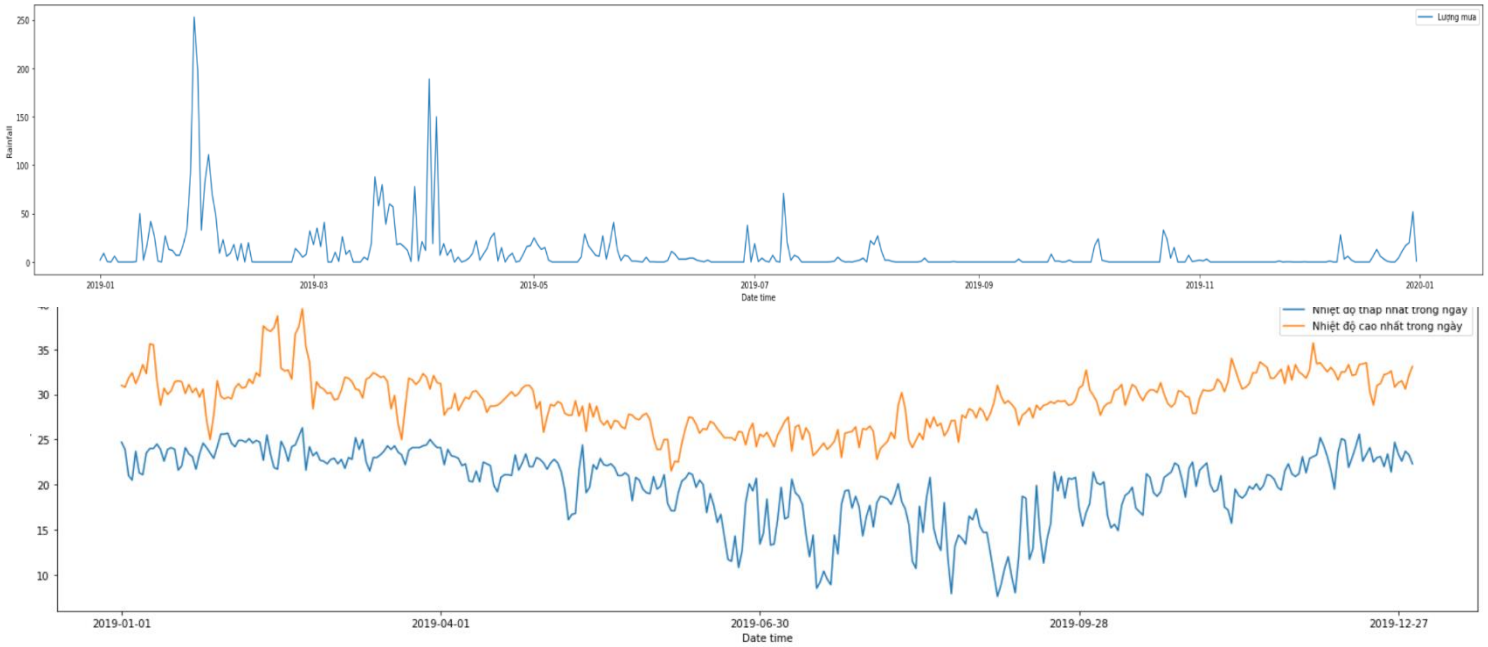
[4] M.-T. Luong, H. Pham, and C. D. Manning, “Effective Approaches to Attention-based Neural Machine Translation.” arXiv, Sep. 20, 2015. doi: 10.48550/arXiv.1508.04025.

[5] L. Beretta and A. Santaniello, “Nearest neighbor imputation algorithms: a critical evaluation,” *BMC Med. Inform. Decis. Mak.*, vol. 16, no. 3, p. 74, Jul. 2016, doi: 10.1186/s12911-016-0318-z.

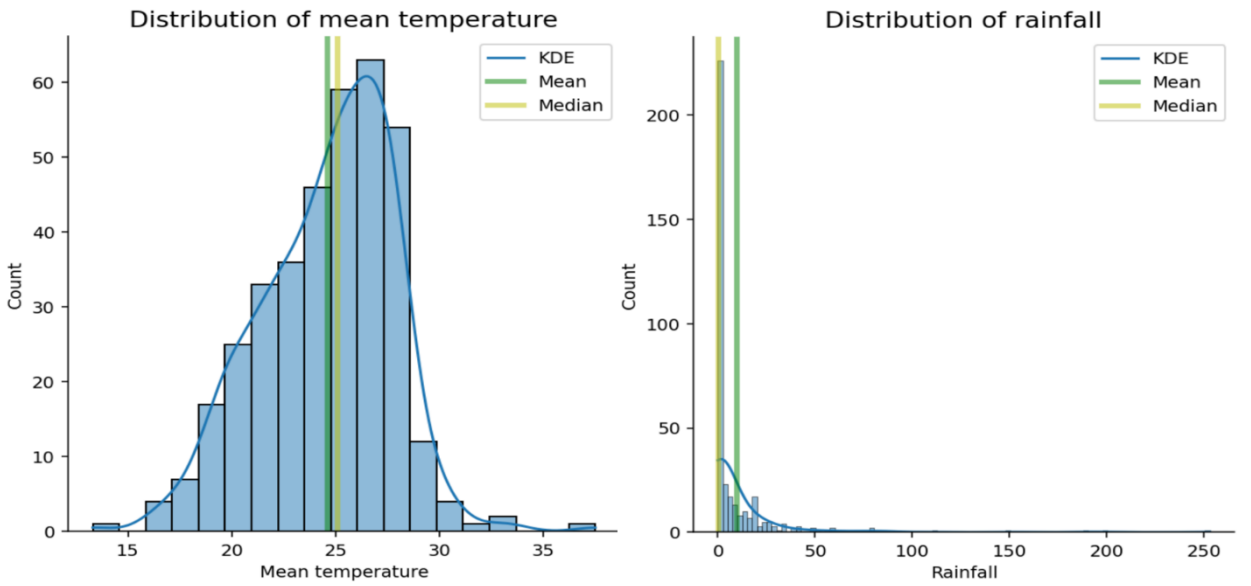
CÁC BIỂU ĐỒ PHỤ TRỢ VIỆC PHÂN TÍCH



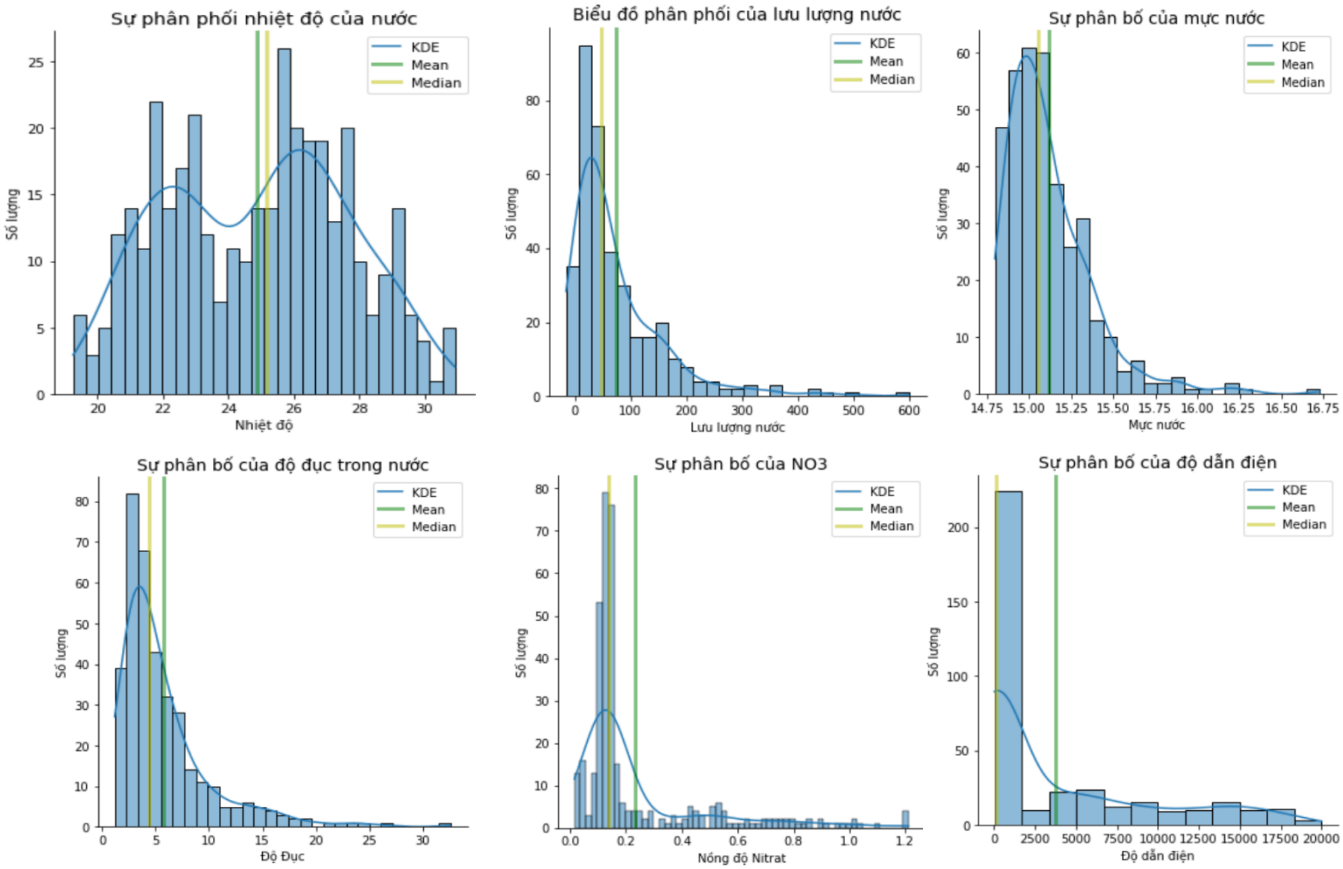
Đồ thì dao động các thuộc tính năm 2019



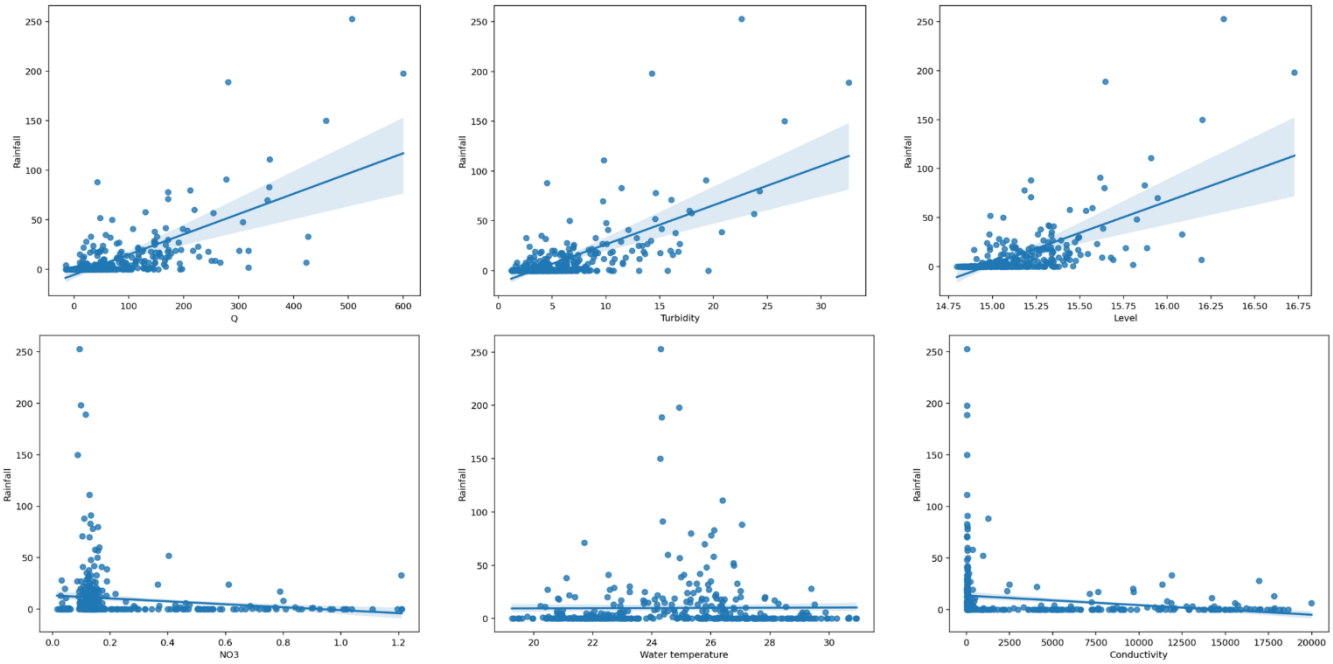
Biểu đồ dao động của nhiệt độ trung bình và lượng mưa 2019



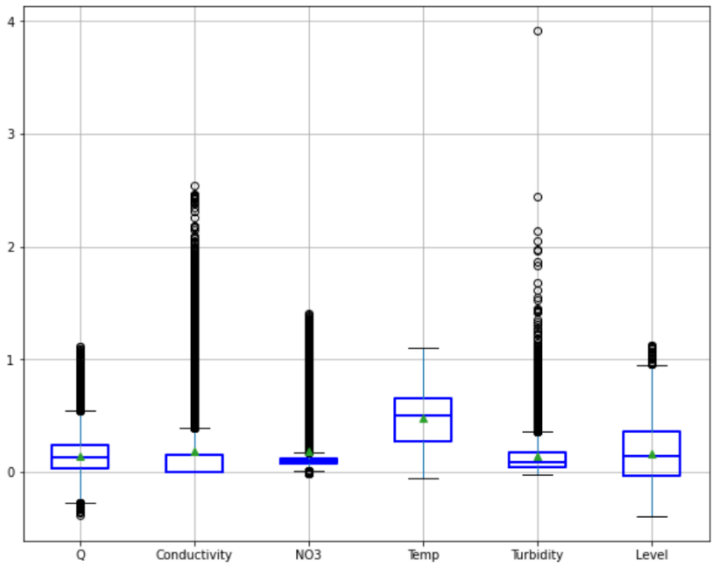
Sự phân phối nhiệt độ trung bình và lượng mưa



Sự phân bố dữ liệu của các thuộc tính



Mối tương quan giữa lượng mưa và các thông số chất lượng nước



Boxplot các thuộc tính cơ bản

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Nguyễn Thanh Thiện Quá	Sưu tầm dữ liệu, phân tích thăm dò và trực quan, xây dựng mô hình, tổng hợp báo cáo.
2	Huỳnh Lê Phương Vy	Thu thập dữ liệu mở rộng, phân tích, trực quan dữ liệu, thiết kế mô hình, tạo dashboard, thiết kế slide.
3	Nguyễn Hiếu Nghĩa	Sưu tầm dữ liệu mở rộng, phân tích thăm dò dữ liệu mở rộng, chạy thuật toán train mô hình, viết báo cáo.
4	Ngô Thị Phúc	Thu thập dữ liệu mở rộng, phân tích, trực quan dữ liệu, thiết kế mô hình, tạo dashboard.