DOI: 10.36335/VNJHM.2020(714).18-29

CÁC PHƯƠNG PHÁP CHUẨN HÓA DỮ LIỆU THỦY VĂN ÁP DỤNG CHO TRẠM 74129 - YÊN BÁI

Đặng Văn Nam[1], Hoàng Quý Nhân[2], Ngô Văn Mạnh[3], Nguyễn Thị Hiền[4]

Tóm tắt: *Dữ liệu mực nước (water level) tại các trạm trên sông hiện nay chủ yếu được thu thập bằng phương pháp quan trắc thủ công với tần suất thu thập khác nhau tùy thuộc vào từng thời điểm trong năm. Các dữ liệu này cần phải được làm sạch để loại bỏ các điểm bất thường (Outliers), các giá trị thiếu (Missing values), chuẩn hóa về dạng chuỗi thời gian (Time series)…. Trong nội dung của bài báo này, nhóm tác giả sẽ chỉ ra hiện trạng của dữ liệu mực nước thu thập được tại trạm 74129 - Yên Bái trong giai đoạn 9 năm từ 01/01/2011 đến 31/12/2019; Đây là các dữ liệu thực tế, được cung cấp bởi Trung tâm thông tin và Dữ liệu khí tượng thủy văn. Trên cơ sở hiện trạng của tập dữ liệu này, sẽ tiến hành thực nghiệm các phương pháp làm sạch dữ liệu để loại bỏ ngoại lai, thay thế giá trị thiếu bằng phương pháp nội suy và chuẩn hóa dữ liệu về dạng chuỗi thời gian với khoảng thời gian cách đều nhau 3h. Dữ liệu sau khi đã được chuẩn hóa, làm sạch, đảm bảo tính đầy đủ và độ tin cậy sẽ là yếu tố quyết định tới độ chính xác của các mô hình dự đoán, dự báo.*

Từ khóa: *Mực nước, ngoại lai, dữ liệu thiếu, chuỗi thời gian.*

Ban Biên tập nhận bài: 12/04/2020 Ngày phản biện xong: 20/06/2020 Ngày đăng bài: 25/06/2020

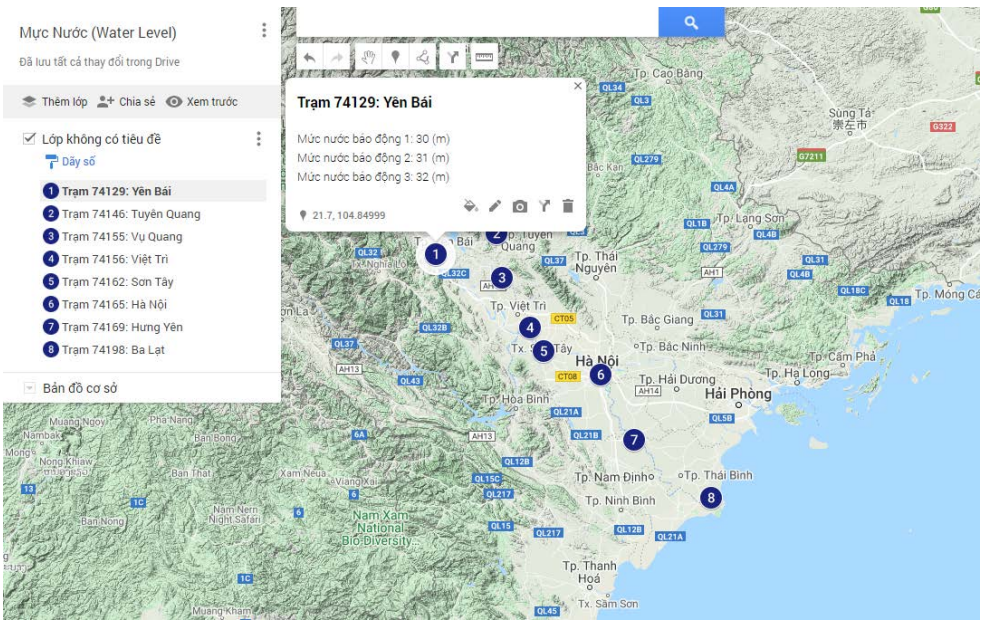
1. Đặt vấn đề

Dữ liệu mực nước thu thập từ các trạm quan trắc trên sông có thể được thực hiện thông qua quan trắc thủ công (ghi nhận trực tiếp giá trị của yếu tố đo trên thiết bị quan trắc) hoặc quan trắc tự động (ghi nhận giá trị của yếu tố đo bằng thiết bị tự động và truyền về người sử dụng theo nhu cầu) [1]. Hiện nay, việc quan trắc mực nước trên các hệ thống sông chủ yếu vẫn sử dụng phương pháp quan trắc thủ công, người quan trắc sẽ ghi nhận giá trị trên thước đo mực nước sau đó gửi dữ liệu này về trung tâm để lưu trữ, xử lý. Do nhiều yếu tố chủ quan và khách quan, dẫn đến quá trình ghi nhận giá trị và gửi số liệu quan trắc về trung tâm bị sai sót, nhầm lẫn, mất mát so với giá trị thực tế. Hơn nữa, tùy vào từng thời điểm, mùa vụ trong năm mà chế độ quan trắc mực nước cũng khác nhau có thể chỉ 2 lần/ngày (7h, 19h), 4 lần/ngày (1h, 7h, 13h, 19h) hoặc 8 lần/ngày (1h, 4h, 7h, 10h, 13h, 16h, 19h, 21h) vào thời điểm mùa cạn, hoặc thời kỳ đầu mùa lũ khi biên độ mực nước trong ngày nhỏ; nhưng có thể tăng lên 12 lần/ngày (1h, 3h, 5h, 7h, 9h, 11h, 13h, 15h, 17h, 19h, 21h, 23h), hoặc 24 lần/ngày (0h, 1h, 2h,…., 22h, 23h)… được áp dụng trong mùa lũ khi mực nước biến đổi trong ngày lớn [1]. Vì vậy, dữ liệu thu thập được bị ngắt quãng và không liên tục, thời điểm lấy dữ liệu khác nhau tùy thuộc vào từng mùa trong năm, đặc điểm lưu vực, đặc điểm trận mưa, thời gian lũ…Đây là các dữ liệu được ghi nhận và lưu trữ theo thời gian, nhưng lại không phải là dữ liệu chuỗi thời gian (*Time series data*). Do đó không thể áp dụng các mô hình dự báo chuỗi thời gian như: MA, ARMA, ARIMA, PARMA, GARMA…hay các mô hình học máy, học sâu khác trong việc xây dựng mô hình dự báo mực nước tại các trạm quan trắc, phục vụ việc cảnh báo lũ hoặc các bài toán liên quan khác [2-4].

Có thể thấy, các dữ liệu quan trắc mực nước thu thập và lưu trữ hiện tại là các dữ liệu thô (*Raw data*), các dữ liệu này cần phải được chuẩn hóa và làm sạch (*Data preparation*) trước khi sử dụng cho bất kỳ mục đích gì, đây là công đoạn bắt buộc và không thể thiếu [5,6]. Kết quả của nhiều nghiên cứu đã chỉ ra rằng, 80% thời gian,

[1]*Đại học Mỏ-Địa Chất,*
[2]*Đại học Nông lâm Thái Nguyên,*
[3]*Trung tâm Thông tin và Dữ liệu khí tượng thủy văn,*
[4]*Học viện Kỹ thuật quân sự*
Email: dangvannam@humg.edu.vn

công sức và nguồn lực của một dự án khoa học dữ liệu là nằm ở khâu chuẩn bị dữ liệu. Trong các phần tiếp theo của bài báo, nhóm tác giả sẽ tìm hiểu về phương pháp thu thập và hiện trạng dữ liệu thủy văn tại trạm 74129 - Yên Bái trong giai đoạn 9 năm từ ngày 01/01/2011 đến hết ngày 31/12/2019, từ đó xác định được những phương pháp chuẩn hóa dữ liệu cần thiết, phù hợp với tập dữ liệu này. Nhóm tác giả sử dụng các thư viện, kỹ thuật lập trình để xây dựng các module thực hiện việc loại bỏ các điểm ngoại lai, các điểm thiếu dữ liệu và chuẩn hóa dữ liệu mực nước về dạng chuỗi thời gian. Các phương pháp tiền xử lý dữ liệu áp dụng cho trạm 74129 sẽ làm cơ sở áp dụng với các trạm quan trắc thủy văn khác trên hệ thống sông Hồng nói chung.



Hình 1. Vị trí của trạm 74129 trên bảng đồ Google Maps

2. Phương pháp thu thập và hiện trạng dữ liệu thủy văn trạm 74129 - Yên Bái

2.1. Phương pháp thu thập dữ liệu mực nước

Dữ liệu mực nước tại các trạm quan trắc thủy văn trên sông Hồng nói chung và trạm 74129 nói riêng được thu thập bằng phương pháp quan trắc thủ công. Hàng ngày, vào các thời gian quy định người quan trắc sẽ ghi nhận trực tiếp giá trị mực nước trên thiết bị quan trắc sau đó gửi giá trị này về Trung tâm Thông tin và Khí tượng thủy văn để lưu trữ và xử lý, phục vụ cho các mục đích cụ thể. Hình 1 thể hiện vị trí của một số trạm trên hệ thống sông Hồng trong đó có trạm 74129 - Yên Bái.

Chế độ quan trắc mực nước phải đảm bảo phản ánh được quá trình diễn biến mực nước một cách đầy đủ, khách quan và phải có tính khả thi [1]. Theo TCVN 12636-2:2019 với quan trắc thủ công có 8 chế độ:

• Chế độ 1: Mỗi ngày quan trắc 2 lần vào các thời điểm: 7h, 19h; được áp dụng trong mùa cạn ở các sông vùng không ảnh hưởng thủy triều, thời kỳ biên độ mực nước trong ngày nhỏ hơn hoặc bằng 5cm ($\Delta H \leq 5cm$)

• Chế độ 2: Mỗi ngày quan trắc 4 lần vào các thời điểm: 1h, 7h, 13h, 19h; được áp dụng trong thời kỳ biên độ mực nước trong ngày lớn hơn 5 cm nhưng nhỏ hơn hoặc bằng 10cm ($5 < \Delta H \leq 10cm$), như đầu và cuối mùa cạn ở các sông thuộc vùng không ảnh hưởng thủy triều.

• Chế độ 3: Mỗi ngày quan trắc 8 lần vào các thời điểm: 1h, 4h, 7h, 10h, 13h, 16h, 19h, 22h; được áp dụng trong thời kỳ mực nước biến đổi rõ rệt trong ngày, như thời kỳ đầu mùa lũ ở các sông vừa và lớn thuộc vùng không ảnh hưởng thủy triều.

• Chế độ 4: Mỗi ngày quan trắc 12 lần vào các thời điểm: 1h, 3h, 5h, 7h, 9h, 11h, 13h, 15h, 17h,

19h, 21h, 23h; được áp dụng trong thời kỳ mực nước biến đổi lớn trong ngày, như mùa lũ ở các sông vừa và lớn, những nơi chịu ảnh hưởng nhật triều có biên độ nhỏ hơn 1m.

• Chế độ 5: Mỗi ngày quan trắc vào các thời điểm: 1h, 3h, 5h, 7h, 9h, 11h, 13h, 15h, 17h, 19h, 21h, 23h. Ngoài ra trước, sau chân, đỉnh (triều hoặc lũ) mỗi giờ quan trắc 1 lần, được áp dụng ở những trạm chịu ảnh hưởng nhật triều có biên độ triều khá lớn ($\Delta H \geq 1m$) và những ngày có lũ lớn ở sông vừa và lớn.

• Chế độ 6: Mỗi ngày quan trắc 24 lần vào các thời điểm: 0h, 1h, 2h …, 22h, 23h; được áp dụng trong thời kỳ lũ của các con sông, ở các tuyến quan trắc chịu ảnh hưởng nhật triều và ảnh hưởng khá lớn của bán nhật triều.

• Chế độ 7: Mỗi ngày quan trắc 24 lần vào các thời điểm: 0h, 1h, 2h, …, 22h, 23h. Ngoài ra chân, đỉnh (triều hoặc lũ) cách 5, 10, 15 hoặc 30 phút quan trắc thêm 1 lần. Khoảng thời gian quan trắc được xác định theo sự biến đổi mực nước, nhằm quan trắc chính xác trị số mực nước và thời gian xuất hiện của mực nước và thời gian xuất hiện của mực nước chân, đỉnh được áp dụng tại những nơi mực nước chịu ảnh hưởng triều mạnh và tại các sông, suối nhỏ trong thời kỳ lũ.

• Chế độ 8: Cách 5 phút, 10 phút, 15 phút hoặc 20 phút quan trắc một lần, từ khi lũ lên đến hết trận lũ. Tại chân, đỉnh lũ quan trắc dày hơn, sườn lũ lên quan trắc dày hơn sườn lũ xuống. Khoảng cách thời gian quan trắc được xác định theo sự biến đổi của cường suất mực nước và thời gian kéo dài của trận lũ. Cường suất mực nước biến đổi càng lớn, thời gian lũ càng ngắn, để đảm bảo quan trắc chính xác trị số mực nước chân, đỉnh lũ và các điểm chuyển tiếp của trận lũ. Cần nắm vững đặc điểm lưu vực, đặc điểm trận mưa (cường độ mưa, trung tâm mưa…) để bố trí thời gian quan trắc [1].

Với trạm 74129 thực hiện theo các chế độ quan trắc từ 1 đến 6 tùy thuộc vào từng điều kiện cụ thể theo mùa, theo trận lũ…. Dữ liệu sau khi được ghi nhận sẽ được gửi về lưu trữ trong cơ sở dữ liệu của Trung tâm Thông tin và Dữ liệu khí tượng thủy văn. Để thuận lợi cho việc phân tích chúng tôi đã truy xuất các dữ liệu thủy văn được lưu trữ trong MongoDB và tách để lấy số liệu trong giai đoạn 9 năm gần đây (2011 - 2019); Dữ liệu sau đó được lưu trữ trong file theo định dạng .CSV *(Comma Separated Values)* có tên Data_waterlevel_74129, bao gồm thuộc tính TimeVN: Cho biết thời điểm quan trắc mực nước định dạng YYYY-MM-DD hh:mm; và thuộc tính 74129: Giá trị quan trắc mực nước (*Water level*) của trạm 74129 tương ứng với thời điểm quan trắc, đơn vị cm. Hình 2 minh họa 12 dòng dữ liệu đầu tiên trong tập dữ liệu.

TimeVN	74129
2011-01-01 7:00	2573
2011-01-01 19:00	2557
2011-01-02 1:00	2542
2011-01-02 7:00	2537
2011-01-02 13:00	2535
2011-01-02 19:00	2533
2011-01-03 7:00	2535
2011-01-03 19:00	2549
2011-01-04 7:00	2543
2011-01-04 19:00	2543
2011-01-05 7:00	2544
2011-01-05 19:00	2546

Hình 2. Cấu trúc file Data_waterlevel_74129.csv

2.2. Khám phá dữ liệu mực nước tại trạm 74129

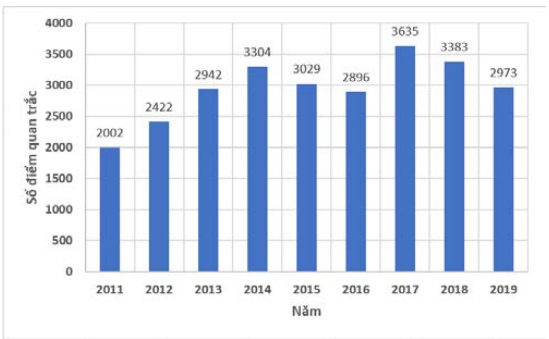
Trước khi đưa ra các phương pháp xử lý và chuẩn hóa dữ liệu thủy văn cho trạm 74129, ta cần phải khám phá và hiểu được chi tiết hiện trạng của các số liệu này. Bảng 1 cho biết những thông số tổng quan nhất của tập dữ liệu quan trắc.

Bảng 1. Thống kê thông số quan trắc tại trạm 74129

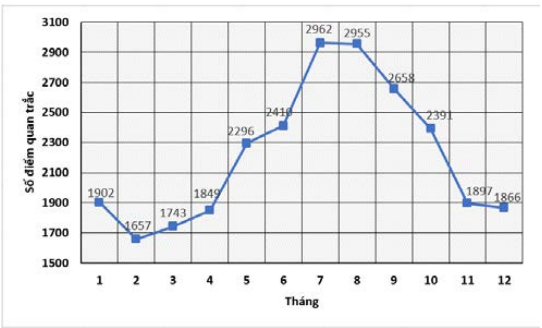
Thông số	Giá trị
Thời điểm bắt đầu (starttime)	2011-01-01 7:00
Thời điểm kết thúc (endtime)	2019-12-31 19:00
Tổng số điểm quan trắc (number)	26 586 điểm
Mực nước trung bình (mean)	2668.25 cm
Độ lệch chuẩn (std)	176.04 cm
Mực nước thấp nhất (min)	1.0 cm
Mực nước cao nhất (max)	3312.0 cm

Hình 3 thể hiện biểu đồ thống kê số điểm quan trắc theo từng năm, qua đó ta có thể thấy rằng số thời điểm quan trắc thay đổi theo từng năm cao nhất là năm 2017 với 3635 thời điểm quan trắc, thấp nhất là năm 2011 với 2002 thời điểm. Mức chênh lệch lên tới 1633 điểm dữ liệu quan trắc.

Hình 4 thể hiện số liệu thống kê số điểm quan trắc theo từng tháng, chúng ta có thể nhận thấy tần suất quan trắc dữ liệu mực nước thay đổi theo từng tháng trong năm, tần suất cao trong giai đoạn từ tháng 5 đến tháng 10 hàng năm, cao nhất tập trung vào tháng 7 và 8; Nó cũng phản ánh đúng thời tiết chung của khu vực khi giai đoạn này là vào mùa lũ và cao điểm mưa lũ chủ yếu rơi vào tháng 7, 8.

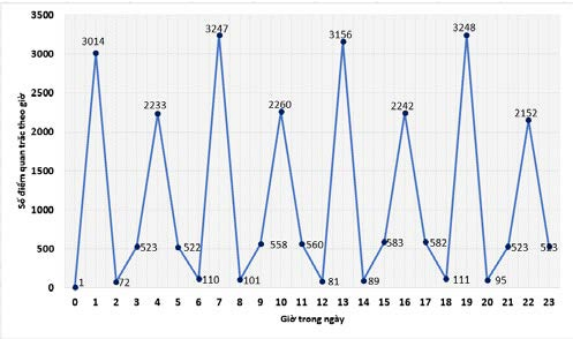


Hình 3. Biểu đồ thống kê số điểm quan trắc theo năm



Hình 4. Biểu đồ thống kê số điểm quan trắc theo tháng

Hình 5 thể hiện số liệu thống kê số điểm quan trắc mực nước theo từng giờ trong ngày. Dễ dàng nhận thấy tần suất lấy số liệu chủ yếu tập trung vào các thời điểm 1h, 4h, 7h, 10h, 13h, 16h, 19h, 22h; Các thời điểm 0h, 2h, 6h, 8h, 12h, 14h, 18h, 20h rất ít số liệu quan trắc. Số liệu này có ý nghĩa quan trọng trong phần tiếp theo khi thực hiện chuẩn hóa nó về dạng chuỗi thời gian sẽ được trình bày trong phần 3 của bài báo này.



Hình 5. Biểu đồ thống kê số điểm quan trắc theo giờ

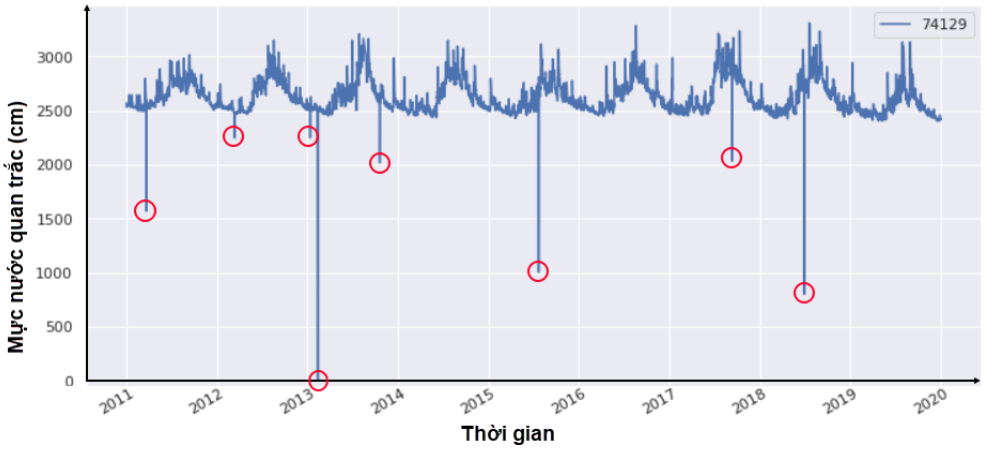
3. Chuẩn hóa dữ liệu thủy văn trạm 74129

3.1. Phát hiện và xử lý các điểm dữ liệu bất thường

Như đã trình bày trong nội dung 2.1, dữ liệu mực nước tại trạm 74129 được thu thập theo phương pháp quan trắc thủ công, vì vậy trong quá trình ghi nhận dữ liệu và truyền về trung tâm lưu trữ do các nguyên nhân chủ quan và khách quan có thể xảy ra các sai sót làm cho số liệu bị sai lệch, bất thường. Các điểm dữ liệu này được gọi là ngoại lai (*Outliers*).

Một điểm ngoại lai là một điểm dữ liệu khác biệt đáng kể so với phần còn lại của tập dữ liệu. Các dữ liệu ngoại lai thường được xem như là các mẫu dữ liệu đặc biệt, cách xa khỏi phần lớn dữ liệu khác trong tập dữ liệu [7]. Có nhiều phương pháp để phát hiện các điểm ngoại lai như: Phân tích giá trị cực trị (*Extreme Value Analysis*); Các mô hình xác suất và thống kê (*Probabilistic and Statistical Models*); Các mô hình tuyến tính (*Linear Models*); Các mô hình dựa trên lân cận (*Proximity - based Models*); Các mô hình dựa trên lý thuyết thông tin (*Information Theoretic Models*) [7,8,9].

Hình 6 là đồ thị biểu diễn giá trị mực nước quan trắc từ năm 2011 đến năm 2019, trực quan bằng mắt có thể dễ dàng nhận thấy có khá nhiều điểm dữ liệu ngoại lai trái *(Left outliers)* - các điểm được đánh dấu bằng các hình tròn màu đỏ. Đây là các giá trị xem xét và kiểm tra ngoại lai trong tập dữ liệu.

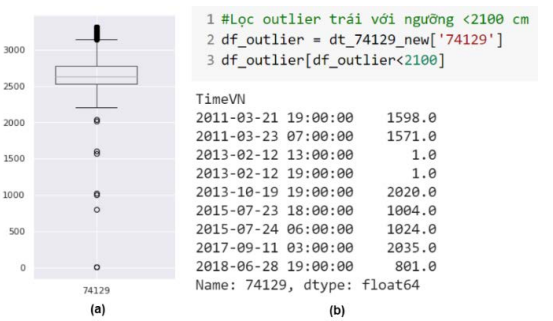


Hình 6. Đồ thị thể hiện số liệu mực nước quan trắc của trạm 74129 trong gian đoạn từ 2011-2019

Dữ liệu mực nước thu thập được là các dữ liệu một chiều, nên phương pháp đơn giản và hiệu quả để có thể phát hiện những điểm dữ liệu ngoại lai này là sử dụng phân tích giá trị cự trị. Hai phương pháp hiệu quả để phát hiện giá trị cực trị bao gồm Z-Scores và đồ thị Box-plot [10].

Trong nội dung thực nghiệm cho trạm 74129, nhóm tác giả sử dụng ngôn ngữ lập trình Python, kết hợp với một số thư viện mã nguồn mở hỗ trợ trong việc phân tích, xử lý và trực quan hóa bao gồm: Pandas, Numpy và Matplotlib, toàn bộ mã nguồn được viết trên hệ thống Google Colab.

Để phát hiện ngoại lai cho tập dữ liệu mực nước quan trắc, nhóm tác giả sử dụng biểu đồ Box-plot. Biểu đồ Box-plot được sử dụng để đo khuynh hướng phân tán và xác định ngoại lai của tập dữ liệu [10]. Hình 7(a) là biểu đồ Box-plot của tập dữ liệu. Các điểm dữ liệu nằm ngoài vạch ngang thấp nhất trong biểu đồ Box-plot được xem xét là các điểm ngoại lai trái. Hình 7(b) liệt kê danh sách 9 điểm quan trắc có giá trị nhỏ nhất trong tập dữ liệu cách xa khỏi phần lớn các điểm khác. Để có thể khẳng định đây có phải là các điểm dữ liệu ngoại lai không? Cũng như đưa ra được phương án xử lý phù hợp với các điểm này, chúng ta cần phải thực hiện kiểm chứng. Trong phần dưới đây nhóm tác giả thực hiện kiểm chứng cho 2 điểm dữ liệu xem xét ngoại lai ghi nhận vào 19h ngày 21/03/2011 và 7h ngày 23/03/2011, kiểm chứng ngoại lai cho các điểm khác sẽ được thực hiện tương tự.



Hình 7. Biểu đồ box-plot của tập dữ liệu (a); Danh sách các điểm quan trắc xem xét ngoại lai trái (b)

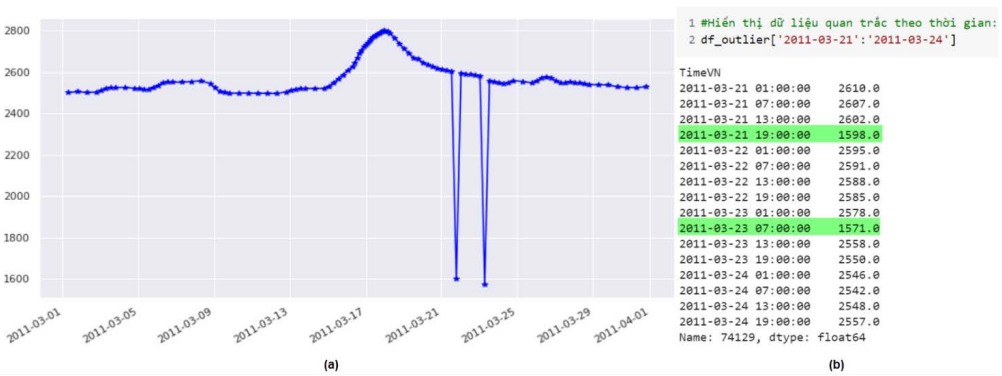
Theo như hình 8(a) có thể thấy ngay rằng mực nước tại trạm Yên Bái trong giai đoạn tháng 03/2011 có 2 điểm quan trắc có giá trị biến thiên đột ngột. Hình 8b thể hiện mức độ thay đổi mực nước của 2 điểm quan trắc này so với các điểm quan trắc lân cận chênh nhau rất lớn; Thời điểm 19h ngày 21/03/2011 dữ liệu mực nước ghi nhận 1598cm trong khi tại thời điểm quan trắc liền trước nó lúc 13h ngày 21/03/2011 là 2602cm (mức độ chênh lệch giảm giữa hai thời điểm quan trắc là -1004cm) và thời điểm liền sau lúc 1h ngày 22/03/2011 là 2595cm (mức độ chênh lệch tăng giữa hai thời điểm quan trắc là +997 cm). Mức độ thay đổi đột ngột cũng xảy ra tương tự với thời điểm lúc 7h ngày 23/03/2011. Tháng 3 là giai đoạn mùa khô, theo như dữ liệu cho thấy chế độ quan trắc đang thực hiện theo chế độ 2 (6 tiếng một lần vào các thời điểm 1h, 7h, 13h, 19h), Do đó có thể khẳng định đây là các điểm ngoại lai, dữ liệu ghi nhận và lưu trữ đã bị sai

lệch hoàn toàn so với dữ liệu thực tế.

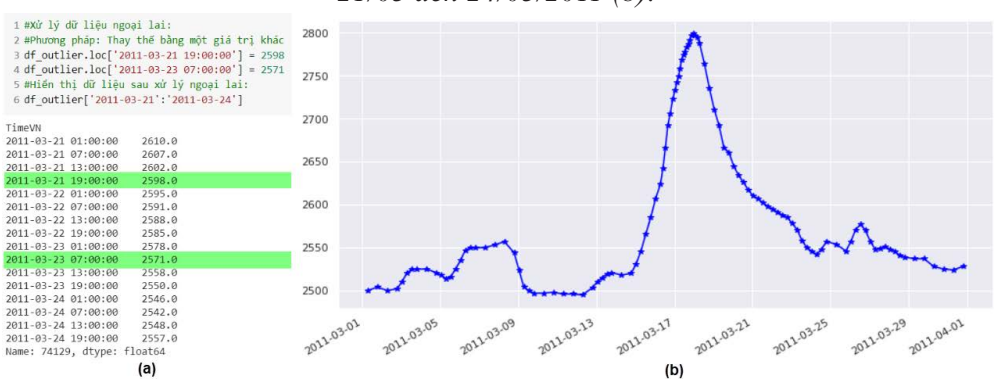
Các điểm dữ liệu ngoại lai có ảnh hưởng rất lớn đến độ chính xác của các mô hình dự đoán, dự báo. Do đó, yêu cầu bắt buộc là cần phải được phát hiện và xử lý chúng. Phần trên đã chỉ ra cách để phát hiện các điểm này, câu hỏi đặt ra là sẽ xử lý các điểm ngoại lai này như thế nào?

Có 3 phương pháp được sử dụng để xử lý dữ liệu ngoại lai bao gồm: Loại bỏ các dòng chứa điểm ngoại lai khỏi tập dữ liệu; Thay thế các giá trị ngoại lai bằng một giá trị khác phù hợp hơn; Thay thế giá trị ngoại lai bằng giá trị NULL (*empty*), xem xét đây như là một điểm dữ liệu thiếu *(missing value)* [11]. Không có một phương pháp xử lý dữ liệu ngoại lai chung nào được áp dụng cho tất cả các bài toán [12], vì vậy để lựa chọn được phương pháp phù hợp cần có những hiểu biết sâu sắc về tập dữ liệu, về bài toán giải quyết, có thể sử dụng chỉ một phương pháp và/hoặc kết hợp cả 3 nhóm phương pháp ở trên. Và thực tế với dữ liệu thủy văn của trạm 74129, để xử lý dữ liệu ngoại lai nhóm tác giả đã sử dụng cả 3 phương pháp này trong từng trường hợp cụ thể. Trong trường hợp điểm ngoại lai ghi nhận lúc 19h ngày 21/03/2011 và lúc 7h ngày 23/03/2011 có thể thấy rằng điểm ngoại lai này gây ra bởi yếu tố chủ quan của con người trong khi ghi nhận và gửi dữ liệu về trung tâm lưu trữ. Đây là tháng mùa khô, mực nước đang có xu hướng giảm và cường độ thay đổi thấp. Giá trị thực tế trong trường hợp này là 2598cm và 2571cm nhưng đã bị sai lệch thành 1598cm và 1571cm. Do đó, với trường hợp này sẽ sử dụng phương pháp xử lý là thay thế giá trị ngoại lai bằng giá trị mới phù hợp hơn. Hình 9 minh họa phương pháp thay thế và kết quả sau khi xử lý 2 điểm ngoại lai này.

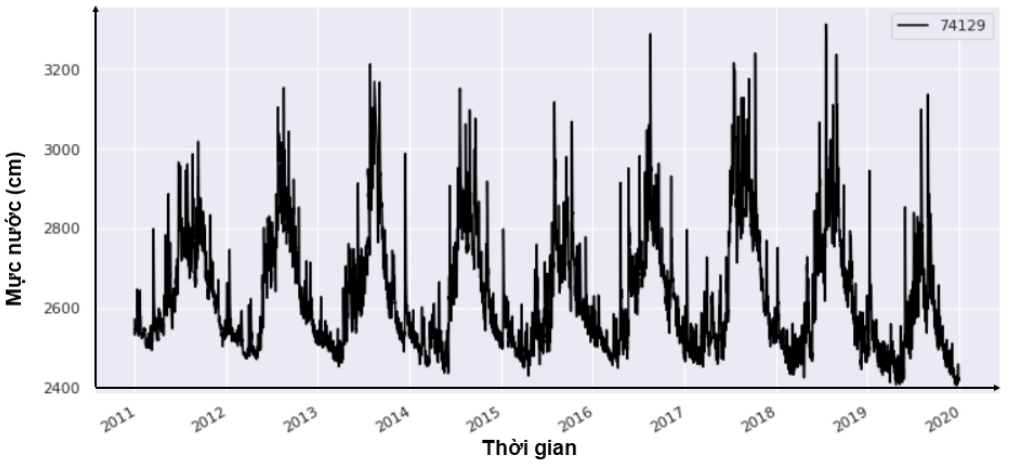
Trên cơ sở phương pháp và cách thức như trình bày ở trên, sẽ thực hiện việc kiểm chứng và xử lý ngoại lai cho toàn bộ tập dữ liệu. Sau bước này các điểm ngoại lai trong tập dữ liệu thủy văn của trạm 74129 đã được xử lý. Hình 10 là đồ thị thể hiện dữ liệu mực nước sau khi đã xử lý các giá trị ngoại lai.



Hình 8. Biểu đồ thể hiện giá trị mực nước quan trắc của trạm 74129 trong thời gian tháng 03/2011 (a); Danh sách thời điểm quan trắc và giá trị mực nước ghi nhận trong thời gian từ 21/03 đến 24/03/2011 (b).



Hình 9. Xử lý ngoại lai theo phương pháp thay thế bằng giá trị mới (a); Đồ thị biểu diễn dữ liệu mực nước tháng 03/2011 sau khi đã xử lý điểm ngoại lai (b).



Hình 10. Dữ liệu mực nước thủy văn trạm 74129 sau khi đã xử lý ngoại lai

TimeVN		TimeVN		TimeVN		TimeVN		TimeVN	
2011-01-03 07:00:00	2535.0	2011-03-22 01:00:00	2595.0	2011-10-04 01:00:00	2752.0	2012-07-24 01:00:00	2817.0	2014-08-29 01:00:00	3094.0
2011-01-03 19:00:00	2549.0	2011-03-22 07:00:00	2591.0	2011-10-04 04:00:00	2751.0	2012-07-24 03:00:00	2821.0	2014-08-29 02:00:00	3095.0
2011-01-04 07:00:00	2543.0	2011-03-22 13:00:00	2588.0	2011-10-04 07:00:00	2753.0	2012-07-24 05:00:00	2823.0	2014-08-29 03:00:00	3096.0
2011-01-04 19:00:00	2543.0	2011-03-22 19:00:00	2585.0	2011-10-04 10:00:00	2753.0	2012-07-24 07:00:00	2823.0	2014-08-29 04:00:00	3097.0
2011-01-05 07:00:00	2544.0	2011-03-23 01:00:00	2578.0	2011-10-04 13:00:00	2754.0	2012-07-24 09:00:00	2822.0	2014-08-29 05:00:00	3097.0
2011-01-05 19:00:00	2546.0	2011-03-23 07:00:00	2571.0	2011-10-04 16:00:00	2750.0	2012-07-24 11:00:00	2821.0	2014-08-29 06:00:00	3096.0
2011-01-06 07:00:00	2543.0	2011-03-23 13:00:00	2558.0	2011-10-04 19:00:00	2746.0	2012-07-24 13:00:00	2817.0	2014-08-29 07:00:00	3095.0
2011-01-06 19:00:00	2546.0	2011-03-23 19:00:00	2550.0	2011-10-04 22:00:00	2744.0	2012-07-24 15:00:00	2811.0	2014-08-29 08:00:00	3094.0
2011-01-07 07:00:00	2546.0	2011-03-24 01:00:00	2546.0	2011-10-05 01:00:00	2747.0	2012-07-24 17:00:00	2802.0	2014-08-29 09:00:00	3089.0
2011-01-07 19:00:00	2547.0	2011-03-24 07:00:00	2542.0	2011-10-05 04:00:00	2757.0	2012-07-24 19:00:00	2799.0	2014-08-29 10:00:00	3086.0
Chế độ 1 (2 lần/ngày)		Chế độ 2 (4 lần/ngày)		Chế độ 3 (8 lần/ngày)		Chế độ 4,5 (12 lần/ngày)		Chế độ 6 (24 lần/ngày)	

Hình 11. Các chế độ quan trắc mực nước tại trạm 74129

3.2. Chuẩn hóa dữ liệu về dạng chuỗi thời gian

Dữ liệu chuỗi thời gian (*time series data*) là chuỗi các điểm dữ liệu được đo theo từng khoảng thời gian liền nhau, khoảng cách giữa các lần đo bằng nhau [2]. Dữ liệu mực nước trạm 74129 thu thập trong khoảng thời gian từ 1h ngày 01/01/2011 đến 23h ngày 31/12/2019. Tuy nhiên, như đã trình bày trong phần đặt vấn đề tần suất thu thập dữ liệu mực nước rất khác nhau tùy thuộc vào từng khoảng thời gian trong năm, cũng như phụ thuộc vào cường độ và mức độ của từng cơn lũ, đợt lũ. Với trạm 74129, thực hiện thu thập dữ liệu theo 6 chế độ khác nhau từ chế độ 1 đến chế độ 6. Hình 11 thể hiện dữ liệu thu thập tại một số thời gian tương ứng với các chế độ quan trắc khác nhau. Qua biểu đồ hình 4 cho thấy tháng 7 và tháng 8 hàng năm là hai tháng có số lượng điểm quan trắc nhiều nhất. Đây là 2 tháng cao điểm trong mùa lũ, chế độ quan trắc chủ yếu theo chế độ 5, 6.

Như vậy, có thể thấy rằng dữ liệu quan trắc thủy văn được thu thập theo mốc thời gian cụ thể theo giờ, nhưng đây không phải là dữ liệu dạng chuỗi thời gian vì khoảng cách giữa các lần quan trắc không cách đều nhau, tùy vào từng điều kiện cụ thể (mùa khô khoảng cách thưa hơn mùa lũ rất nhiều). Do không phải là dữ liệu chuỗi thời gian nên không thể sử dụng các mô hình dự báo chuỗi thời gian như: MA, ARMA, ARIMA…[4]. Vì vậy, cần chuẩn hóa dữ liệu này về dạng chuỗi thời gian để có thể áp dụng được các mô hình dự đoán, dự báo như trên.

Nhóm tác giả đưa ra phương án chuẩn hóa tập dữ liệu này về dạng chuỗi thời gian như sau:

- Bước 1: Xác định khoảng thời gian t cách đều nhau giữa các lần quan trắc. Tham số t sử dụng làm cơ sở để chuẩn hóa dữ liệu về dạng chuỗi thời gian với các thời điểm quan trắc cách đều nhau một khoảng t. Với dữ liệu thủy văn trạm 74129, tham số t lựa chọn theo giờ, có thể là 1h, 2h, 3h…Theo số liệu thống kê được thể

hiện trong biểu đồ Hình 5, chúng ta thấy rằng trong giai đoạn thời gian từ năm 2011 đến 2019, thời điểm quan trắc tập trung chủ yếu vào các mốc thời gian 1h, 4h, 7h, 10h, 13h, 16h, 19h, 22h trong ngày (> 2000 quan trắc), các thời điểm quan trắc khác còn lại trong ngày 0h, 2h, 3h, 5h, 6h, 8h, 9h, 11h, 12h, 14h, 15h, 17h, 18h, 20h, 21h, 23h có số lượng điểm rất ít (<600 quan trắc); Do đó với tập dữ liệu này, chúng ta sẽ chọn tham số t = 3, nghĩa là chúng ta sẽ chuẩn hóa dữ liệu thủy văn thu thập được về dạng chuỗi thời gian với khoảng cách lấy mẫu cách đều nhau là 3h (chế độ 3: 8 lần/ngày).

- Bước 2: Thực hiện việc lọc các mốc thời gian lấy mẫu còn thiếu trong tập dữ liệu tương ứng với khoảng thời gian t = 3h vào các thời điểm 1h, 4h, 7h, 10h, 13h, 16h, 19h, 22h trong ngày. Hình 11 minh họa đoạn mã nguồn thực hiện việc thống kê số điểm và danh sách điểm chưa có dữ liệu. Theo như thống kê cho thấy nếu chuẩn hóa về dạng chuỗi thời gian theo chế độ 3 thì tập dữ liệu thiếu 4725 điểm quan trắc. Thực hiện việc chèn các các thời điểm lấy mẫu thiếu này vào trong file dữ liệu của trạm 74129 với giá trị NULL (xem xét đây như là các điểm dữ liệu thiếu - missing values).

```
#Thiết lập khoảng thời gian kiểm tra dữ liệu quan trắc
#Từ 01:00:00 01-01-2011 đến 22:00:00 31-12-2019
starts = dt.datetime(2011,1,1,1,0,0)
ends = dt.datetime(2019,12,31,22,0,0)

#Tạo một index kiểu datetime từ thời điểm start tới ends cách nhau 3h
index_ref =pd.date_range(start= starts, end=ends, freq='3H')

#Lọc các vị trí thiếu dữ liệu trong chuỗi thời gian
ga = index_ref[~index_ref.isin(dt_74129_new.index)]
print('Số điểm quan trắc thiếu theo chế độ 3:',len(ga))
print('Các vị trí thiếu:\n', ga)
```

```
Số điểm quan trắc thiếu theo chế độ 3: 4745
Các vị trí thiếu:
 DatetimeIndex(['2011-01-01 01:00:00', '2011-01-01 04:00:00',
                '2011-01-01 10:00:00', '2011-01-01 13:00:00',
                '2011-01-01 16:00:00', '2011-01-01 22:00:00',
                '2011-01-02 04:00:00', '2011-01-02 10:00:00',
                '2011-01-02 16:00:00', '2011-01-02 22:00:00',
                ...
                '2019-12-25 16:00:00', '2019-12-25 22:00:00',
                '2019-12-26 04:00:00', '2019-12-26 10:00:00',
                '2019-12-26 16:00:00', '2019-12-26 22:00:00',
                '2019-12-31 04:00:00', '2019-12-31 10:00:00',
                '2019-12-31 16:00:00', '2019-12-31 22:00:00'],
               dtype='datetime64[ns]', length=4745, freq=None)
```

Hình 11. Thống kê số điểm quan trắc thiếu và danh sách các điểm này trong tập dữ liệu

- Bước 3: thời điểm không thực hiện quan trắc đã bổ sung thêm vào tập dữ liệu trong bước 2. Có rất nhiều phương pháp xử lý dữ liệu thiếu, nội dung bước 3 sẽ được trình bày chi tiết trong phần 3.3 dưới đây.

- Bước 4: Chuẩn hóa tập dữ liệu về dạng chuỗi thời gian; Kết thúc bước 3 tập dữ liệu thủy văn trạm 74129 đã được xử lý các dữ liệu thiếu. Tuy nhiên, tập dữ liệu này còn chứa rất nhiều thời điểm quan trắc khác ngoài 8 thời điểm ở trên ứng với các khoảng thời kỳ quan trắc theo chế độ 4, 5 và 6. Do đó, nhóm tác giả thực hiện việc trích lọc dữ liệu từ tập gốc ra tại các vị trí 1h, 4h, 7h, 10h, 13h, 16h, 19h, 22h hàng ngày để thu được dữ liệu thủy văn dạng chuỗi thời gian với khoảng cách 3h.

3.3. Xử lý giá trị thiếu trong tập dữ liệu

Xử lý giá trị thiếu (*missing values*) luôn là một bước quan trọng và bắt buộc trong quá trình làm sạch dữ liệu [5]. Do nhiều nguyên nhân chủ quan và khách quan trong quá trình thu thập dữ liệu có thể dẫn tới giá trị thiếu. Như đã mô tả trong phần 3.2, để chuẩn hóa về dạng chuỗi thời gian với khoảng cách lấy mẫu 3h cần phải thực hiện chèn thêm vào tập dữ liệu những thời điểm chưa quan trắc này với dữ liệu Null (coi đây là các điểm dữ liệu thiếu - missing values). Do đó, yêu cầu đặt ra là phải xử lý các dữ liệu thiếu này.

Có rất nhiều phương pháp để xử lý dữ liệu thiếu [13-14], có thể gom các phương pháp này vào 2 nhóm chính đó là: Loại bỏ các dòng hoặc các cột dữ liệu chứa giá trị thiếu ra khỏi tập dữ liệu; Thay thế các điểm dữ liệu thiếu bằng một giá trị mới theo từng thuật toán cụ thể. Với dữ liệu chuỗi thời gian ta không thể xóa bỏ các dòng dữ liệu thiếu mà chỉ có thể sử dụng nhóm phương pháp thứ 2 là thay thế bằng một giá trị mới. Với các dữ liệu dạng chuỗi thời gian, các điểm dữ liệu sẽ có mối quan hệ với các điểm phía trước và phía sau nó, cũng như tuân theo xu hướng và mùa vụ. Có 4 giải pháp đơn giản nhưng hiệu quả để xử lý dữ liệu thiếu cho chuỗi thời gian bao gồm:

- Thay thế giá trị thiếu bằng giá trị liền trước (*Last observation carried forward - LOCF*);

- Thay thế giá trị thiếu bằng giá trị liền sau (*Next observation carried backward - NOCB*);

- Thay thế giá trị thiếu bằng phương nội suy tuyến tính (*Linear interpolation*);

- Thay thế giá trị thiếu bằng phương nội suy spline (*Spline interpolation*).

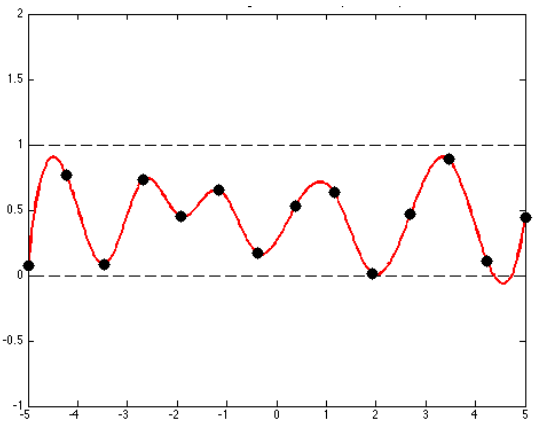
Với đặc điểm dữ liệu thủy văn trạm 74129, nhóm tác giả sử dụng phương pháp nội suy Spline bậc 3 để xử lý giá trị thiếu.

Nội suy Spline là phương pháp xây dựng các đường cong trơn đi qua n + 1 điểm dữ liệu đã biết $(x_0, y_0),..., (x_n, y_n)$. Thực thế là đi tìm một hàm f(x) sao cho $f(x_i) = yi$ với mọi i. Chúng ta sẽ xác định n đa thức bậc $p_0,…., p_{n-1}$ sao cho $f(x) = p_i(x)$ với mọi x trong khoảng $[x_i, x_{i+1}]$ [15]. Trong thực tế nhóm tác giả sử dụng nội suy spline với đa thức bậc 3 khi đó pi(x) được định nghĩa như sau:

$p_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i$ [16]

Hình 12 minh họa việc xây dựng các đường cong bậc 3 (đường màu đỏ) đi qua 14 điểm đã biết (điểm chấm đen).

Áp dụng cho dữ liệu thủy văn của trạm 74129, trong hình 13a thể hiện 10 điểm dữ liệu đầu tiên trong tháng 01/2012 chứa các điểm giá trị thiếu tại 4h, 10h, 16h, 22h (trong Pandas giá trị thiếu ký hiệu là NaN) và đồ thị biểu diễn các giá trị mực nước quan trắc trong tháng 01/2012 - Hình 13b. Hình 14a là kết quả sau khi xử lý giá trị thiếu với phương pháp nội suy Spline bậc 3 cho các điểm dữ liệu mô tả trong hình 13a cũng như là đồ thị thể hiện toàn bộ dữ liệu của trạm 74129 trong tháng 01/2012 bao gồm cả dữ liệu quan trắc và dữ liệu nội suy cho các điểm thiếu (Hình 14b).

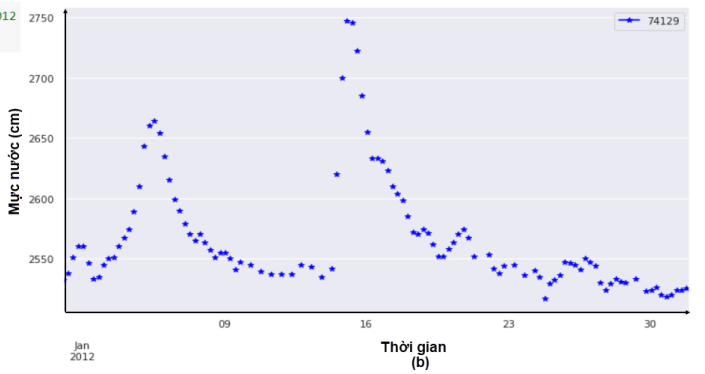


Hình 12. Nội suy Spline bậc 3 qua 14 điểm đã biết

```
1 # 10 điểm dữ liệu đầu tiên trong tháng 01-2012
2 data_74129_missing_3h['2012-01'].head(10)
```

	74129
2012-01-01 01:00:00	2532.0
2012-01-01 04:00:00	NaN
2012-01-01 07:00:00	2538.0
2012-01-01 10:00:00	NaN
2012-01-01 13:00:00	2551.0
2012-01-01 16:00:00	NaN
2012-01-01 19:00:00	2560.0
2012-01-01 22:00:00	NaN
2012-01-02 01:00:00	2560.0
2012-01-02 04:00:00	NaN

(a)

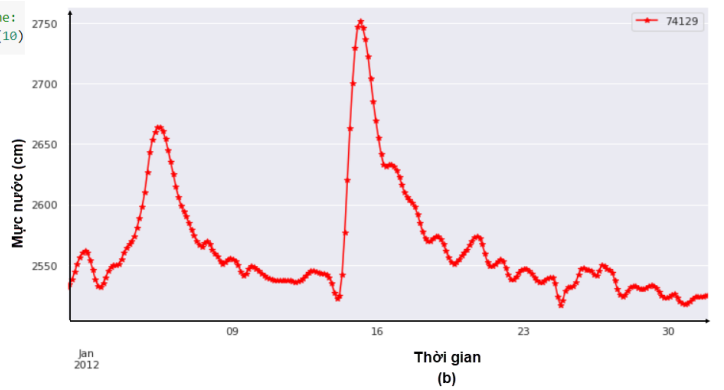


Hình 13. Dữ liệu trước khi xử lý giá trị thiếu (a) và Đồ thị biểu diễn dữ liệu trong tháng 01/2012 (b)

```
1 #Dữ liệu sau khi sử lý missing với Spline:
2 data_74129_nomissing_3h['2012-01'].head(10)
```

	74129
2012-01-01 01:00:00	2532.000000
2012-01-01 04:00:00	2534.500751
2012-01-01 07:00:00	2538.000000
2012-01-01 10:00:00	2544.117816
2012-01-01 13:00:00	2551.000000
2012-01-01 16:00:00	2556.402984
2012-01-01 19:00:00	2560.000000
2012-01-01 22:00:00	2561.645249
2012-01-02 01:00:00	2560.000000
2012-01-02 04:00:00	2554.141021

(a)



Hình 14. Dữ liệu sau khi xử lý giá trị thiếu bằng phương pháp nội suy spline(a) và Đồ thị biểu diễn dữ liệu sau xử lý trong tháng 01/2012(b)

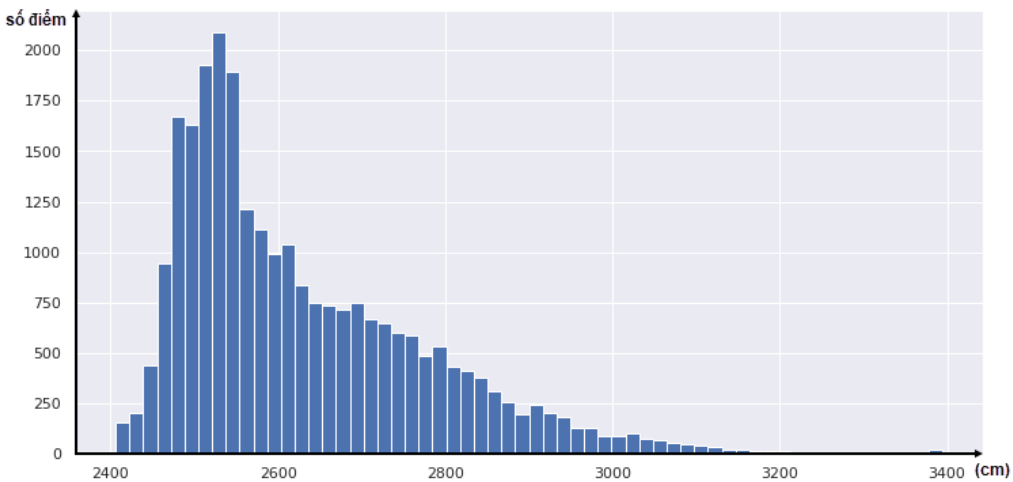
4. Kết quả chuẩn hóa dữ liệu trạm 74129

Sau khi thực hiện các bước tiền xử lý và chuẩn hóa dữ liệu đã trình bày trong phần 3, sẽ thu được tập dữ liệu thủy văn mới của trạm 74129 - Yên Bái được lưu với tên Data_processed_74129.csv tập dữ liệu này cũng có cấu trúc như tập dữ liệu thô ban đầu với 2 cột là TimeVN cho biết thời điểm quan trắc và cột 74129 cho biết giá trị mực nước tương ứng với từng thời điểm quan trắc. Tập dữ liệu sau chuẩn hóa đã xử lý được các điểm ngoại lai, xử lý các điểm dữ liệu thiếu và đưa về dạng chuỗi thời gian với khoảng thời gian cách nhau t = 3h. Bảng 2 mô tả các đặc trưng thống kê chính và Hình 15 thể hiện biểu đồ Histogram của tập dữ liệu mực nước trạm 74129 sau khi đã chuẩn hóa.

Bảng 2. Thống kê thông số tập dữ liệu Data_processed_74129

Thông số	Giá trị
Thời điểm bắt đầu (starttime)	2011-01-01 01:00
Thời điểm kết thúc (endtime)	2019-12-31 22:00
Tổng số điểm dữ liệu (number)	26 296 điểm
Mực nước trung bình (mean)	2631.13 cm
Độ lệch chuẩn (std)	151.19 cm
Mực nước thấp nhất (min)	2406.0 cm
Mực nước cao nhất (max)	3394.47 cm

Tập dữ liệu chuẩn hóa này có thể được sử dụng để làm đầu vào (*input*) cho các mô hình dự đoán, dự báo chuỗi thời gian như MR, ARMA, ARIMA…hoặc làm dữ liệu đầu vào cho các mô hình học máy, học sâu.



Hình 15. Biểu đồ Histogram tập dữ liệu đã xử lý Data_processed_74129

5. Kết luận

Dữ liệu mực nước thu thập được đều là các dữ liệu thô, cần phải được chuẩn hóa và làm sạch để loại bỏ được các điểm ngoại lai ra khỏi tập dữ liệu, các điểm ngoại lai có ảnh hưởng rất lớn tới độ chính xác của các mô hình dự đoán, dự báo. Xử lý các giá trị thiếu cũng là yêu cầu bắt buộc trong quá trình làm sạch dữ liệu, với mỗi một bài toán, một loại dữ liệu cụ thể lại áp dụng những phương pháp xử lý riêng. Đồng thời để có thể sử dụng được các mô hình dự báo chuỗi thời gian thì dữ liệu đầu vào phải được chuẩn hóa về dạng này. Bài báo đã phân tích chi tiết phương pháp thu thập và hiện trạng dữ liệu thủy văn của trạm 74129 - Yên Bái, từ đó thực hiện việc chuẩn hóa dữ liệu này bằng việc giải quyết 3 vấn đề chính bao gồm: Phát hiện và xử lý ngoại lai; Chuẩn hóa về dạng chuỗi thời gian; Xử lý giá trị thiếu. Kết quả sau khi thực hiện toàn bộ quá trình này là một tập dữ liệu đã được chuẩn hóa và làm sạch, có thể sử dụng tập dữ liệu này làm đầu vào cho các mô hình dự báo chuỗi thời gian, học máy, học sâu. Các phương pháp và kỹ thuật xử lý áp dụng với dữ liệu trạm 74129 có thể được sử dụng đối với các trạm thủy văn khác trên hệ thống sông Hồng nói chung.

Lời cảm ơn: *Nghiên cứu này được hỗ trợ bởi đề tài "Nghiên cứu cơ sở khoa học và giải pháp ứng dụng trí tuệ nhân tạo để nhận dạng, hỗ trợ dự báo và cảnh báo một số hiện tượng khí tượng thủy văn nguy hiểm trong bối cảnh biến đổi khí hậu tại Việt Nam", mã số BĐKH.34/16-20." thuộc chương trình Khoa học và công nghệ ứng phó với biến đổi khí hậu, quản lý tài nguyên và môi trường giai đoạn 2016 - 2020.*

Tài liệu tham khảo

1. Tiêu chuẩn quốc gia (2019), *TCVN 12636-2:2019 "Quan trắc khí tượng thủy văn-Phần 2: Quan trắc mực nước và nhiệt độ nước sông"*.

2. Shumway, R.H., Stoffer, D.S. (2017), *Time Series Analysis and Its Applications: With R Examples.* Cham, Switzerland: Springer, 562 p.

3. Brockwell, P.J., Davis, R.A. (2016), *Introduction to Time Series and Forecasting*. Basel, Switzerland: Springer.

4. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M. (2015), *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley.

5. Wang, X., Wang, C. (2019), *Time Series Data Cleaning: A Survey, IEEE Access, 1866-1881*.

6. Song, S., Cao, Y., Wang, J. (2016), *Cleaning timestamps with temporal constraints.* Proc. PVLDB, 9 (10), 708-719.

7. Aggarwal, C.C. (2017), *Outlier Analysis, Springer International Publishing AG, New York*.

8. Akouemo, H.N., Povinelli, R.J. (2014), *Time series outlier detection and imputation*. 2014 IEEE PES General Meeting | Conference & Exposition. Doi:10.1109/pesgm.2014.6939802.

9. Ranga Suri, N.N.R., Murty, N.M, Athithan, G. (2018), *Outlier Detection: Techniques and Applications*, IJCSI International Journal of Computer Science Issues, 9 (1), 307-323.

10. Munzer, T. (2014), *Visualization Analysis and Design*, CRC Press, 428 p.

11. Đặng Văn Nam, Nông Thị Oanh, Ngô Văn Mạnh, Nguyễn Xuân Hoài, Nguyễn Thị Hiền (2020), *Phát hiện và xử lý ngoại lai cho dữ liệu nhiệt độ tại các trạm quan trắc 3h của Việt Nam*. Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất, 61 (1), 132-146.

12. Zhang, A., Song, S., Wang, J., Yu, P.S. (2017), *Time series data cleaning: From anomaly detection to anomaly repairing.* Proc. VLDB Endownment, 10 (10), 1046-1057.

13. Choi, J., Dekkers, O.M., le Cessie, S. (2018), *A comparison of different methods to handle missing data in the context of propensity score analysis*. European Journal of Epidemiology, 34 (1), 23-36.

14. Bonander, C., Strömberg, U. (2018), *Methods to handle missing values and missing individuals.* European Journal of Epidemiology, 34, 5-7.

15. Erdogan KAYA. *Spline Interpolation Techniques.* Journal of Technical Science and Technologies, 2 (1), 47-52.

16. Ajao, I.O., Ibraheem, A.G., Ayoola, F.J. (2012), *Cubic spline interpolation: A robust method of disaggregating annual data to quarterly series.* Journal of Physical Sciens and Environmental Safety, 2 (1), 1-8.

RESULTS OF APPLYING STANDARDIZED METHODS OF HYDROGRAPHIC DATA FOR STATIONS 74129 - YEN BAI

Dang Van Nam[1], Hoang Quy Nhan[2], Ngo Van Manh[3], Nguyen Thi Hien[4]

[1]Hanoi University of Mining and Geology

[2]Thai Nguyen University of Agriculture and Forestry

[3]Center for Hydro-Meteorological Data and Information

[4]Le Quy Don Technical University

Abstract: *Water level data at river stations in Viet Nam are collected by manual observation method with frequency of collection depending on the time of year. These data need to be cleaned to eliminate outliers, missing values ; standardized form of time series .... In the research of this paper, the authors will indicate the current status of water level data collected at the station 74129 - Yen Bai over a period of 9 years from January 1, 2011 to December 31, 2019; These are actual data, provided by the National Center for Hydrometeorological Forecasting. Based on the current status of this data set, experimental methods of Data processing to replace missing values with the method of interpolation and normalization of data in time series form shall be carried out with time spaced 3 hours apart. When there is complete data, ensuring the completeness and reliability will be the decisive factor to the accuracy of the prediction and forecast models.*

Keywords: *Water level, Outliers, Missing values, Time series.*