HỘI NGHỊ TOÀN QUỐC KHOA HỌC TRÁI ĐẤT
VÀ TÀI NGUYÊN VỚI PHÁT TRIỂN BỀN VỮNG (ERSD 2020)

Đánh giá độ chính xác của một số phương pháp xử lý dữ liệu thiếu cho dữ liệu chuỗi thời gian (time series data); Thực nghiệm cho dữ liệu nhiệt độ tại 5 trạm quan trắc khu vực miền Bắc

Nguyễn Thị Phương Bắc[1*], Đặng Văn Nam
[1]*Trường Đại học Mỏ - Địa chất*

TÓM TẮT

Xử lý dữ liệu thiếu (missing values) là một trong những công đoạn không thể thiếu trong quá trình làm sạch dữ liệu (Data cleansing). Có hai nhóm phương pháp xử lý dữ liệu thiếu đó là loại bỏ các giá trị thiếu và/hoặc thay thế dữ liệu thiếu bằng một giá trị mới. Việc lựa chọn sử dụng nhóm phương pháp nào hoặc đồng thời cả hai phụ thuộc vào từng bài toán, từng loại và kiểu dữ liệu cụ thể…Dữ liệu chuỗi thời gian (time series data) là một chuỗi các điểm dữ liệu được đo theo từng khoảng thời gian liền nhau theo một tần suất thời gian thống nhất. Việc xử lý giá trị thiếu trong dữ liệu chuỗi thời gian cũng có những khác biệt lớn so với việc xử lý giá trị thiếu của các dạng dữ liệu khác. Trong nội dung của bài báo này, chúng tôi sẽ giới thiệu 4 phương pháp chính được sử dụng để xử lý giá trị thiếu đối với dữ liệu chuỗi thời gian bao gồm: LOCF, NOCB, nội suy tuyến tính, nội suy Spline; Cũng trong bài báo, chúng tôi sẽ trình bày việc áp dụng các phương pháp này và đánh giá độ chính xác của mỗi phương pháp sử dụng độ đo MAE, RMSE cho dữ liệu nhiệt độ trong năm 2019 tại 5 trạm khu vực phía Bắc.

Từ khóa: Dữ liệu thiếu, Chuỗi thời gian, MAE, RMSE

1. Mở đầu

Chuẩn bị dữ liệu (Data preparation) là công đoạn bắt buộc, là khâu chiếm nhiều thời gian, công sức và nguồn lực nhất của bất kỳ một dự án khoa học dữ liệu nào. Các kết quả nghiên cứu cho thấy 80% thời gian, công sức và nguồn lực của một dự án khoa học dữ liệu là cho việc này. Chuẩn bị dữ liệu bao gồm rất nhiều thao tác, nghiệp vụ, kỹ thuật và yêu cầu khác nhau, phụ thuộc vào từng loại dữ liệu và từng dự án cụ thể. Tuy nhiên, chúng ta có thể tổng hợp vào ba nhóm thao tác chính: Làm sạch dữ liệu (Data cleansing); Chuyển đổi dữ liệu (Data transformation) và Tích hợp dữ liệu (Combining data). (Davy Cielen and et al., 2016)

Như vậy, làm sạch dữ liệu là bước đầu tiên cần phải thực hiện sau khi thu thập dữ liệu, các dữ liệu ban đầu khi thu thập được từ các nguồn khác nhau được gọi là dữ liệu thô (raw data). Dữ liệu này thường chứa rất nhiều "nhiễu", các dữ liệu không liên quan, chưa theo một chuẩn chung thống nhất, hoặc dữ liệu không đầy đủ, bị mất mát, thiếu thông tin…đây là những vấn đề luôn hiện hữu trong mọi bộ dữ liệu.

Một trong những thao tác quan trọng trong quá trình làm sạch dữ liệu đó là xử lý các giá trị thiếu, mất mát (missing values). Có rất nhiều phương pháp khác nhau để xử lý giá trị thiếu, mỗi một phương pháp lại phù hợp với một bài toán, một lĩnh vực và một loại dữ liệu cụ thể. Không có một phương pháp xử lý nào là tốt cho tất cả mọi trường hợp.

Những dữ liệu quan sát liên tục cho một hiện tượng (vật lý, kinh tế ...) trong một khoảng thời gian sẽ tạo nên một chuỗi thời gian như: Doanh số của công ty trong 20 năm gần đây, hoặc nhiệt độ ghi nhận tại một trạm quan trắc khí tượng, hoặc công suất điện năng tiêu thụ trong một nhà máy, đó là các ví dụ điển hình cho một chuỗi thời gian. Dữ liệu chuỗi thời gian (time series data) là một trong những loại dữ liệu đang rất phổ biến và quan trọng ngày nay. Được ứng dụng vào trong rất nhiều lĩnh vực khác nhau, đặc biệt cho các bài toán dự đoán, dự báo. Cũng giống như các loại dữ liệu khác, dữ liệu chuỗi thời gian hoàn toàn có thể bị thiếu, mất mát. Tuy nhiên, việc xử lý dữ liệu thiếu trong chuỗi thời gian lại có những phương pháp và cách thức riêng.

Trong nội dung của bài báo này, nhóm tác giả sẽ nghiên cứu về dữ liệu chuỗi thời gian, các phương pháp cơ bản để xử lý dữ liệu thiếu trong chuỗi thời gian bao gồm: Thay thế giá trị thiếu bằng giá trị liền trước (LOCF); Thay thế giá trị thiếu bằng giá trị liền sau (NOCB); Nội suy tuyến tính (Linear), Nội suy Spline. Nhóm tác giả cũng tiến hành thực nghiệm các phương pháp này trên bộ dữ liệu nhiệt độ tại 5 trạm quan trắc: 48805 – Hà Giang, 48806 – Sơn La, 48825 – Hà Đông, 48838 – Móng Cái và 48839 – Bạch Long Vĩ

* *Tác giả liên hệ:*
Email: nguyenthiphuongbac@humg.edu.vn

trong năm 2019, sau đó sử dụng độ đo MAE và RMSE để đánh giá độ chính xác của 4 phương pháp xử lý dữ liệu thiếu nêu trên.

2. Một số phương pháp xử lý giá trị thiếu cho dữ liệu chuỗi thời gian

2.1 Dữ liệu chuỗi thời gian (time series data)

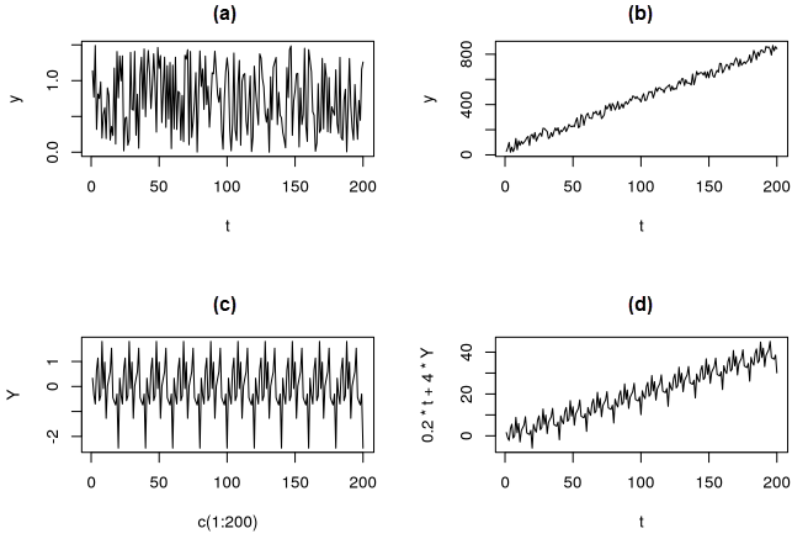
Chuỗi thời gian (time series) trong thống kê, xử lý tín hiệu, kinh tế lượng và toán tài chính là một chuỗi các điểm dữ liệu, được đo theo từng khoảng thời gian liền nhau và theo một tần suất thời gian thống nhất. Dữ liệu chuỗi thời gian thường bao gồm các phép đo liên tiếp thực hiện từ cùng một nguồn trong một khoảng thời gian. Phân tích chuỗi thời gian có mục đích nhận dạng và tập hợp lại các yếu tố, những biến đổi theo thời gian mà nó ảnh hưởng tới giá trị của biến quan sát.

Nếu phân tích dữ liệu chuỗi thời gian theo mô hình hồi quy tuyến tính, giá trị tại thời điểm t là y(t) được xác định như sau:

$$y(t) = m(t) + s(t) + \varepsilon(t) \qquad (1)$$

Trong đó: m(t) là xu hướng (trend), s(t) là tính thời vụ (seasonality) và $\varepsilon(t)$ là biến ngẫu nhiên. Dựa trên phương trình 1, có thể chia dữ liệu chuỗi thời gian thành 4 loại:

- Dữ liệu chuỗi thời gian không có xu hướng và mùa vụ (còn gọi là nhiễu trắng – white nosie) – Hình 1(a)
- Dữ liệu chuỗi thời gian có tính xu hướng nhưng không có tính mùa vụ - Hình 1(b)
- Dữ liệu chuỗi thời gian không có tính xu hướng nhưng có tính mùa vụ - Hình 1(c)
- Dữ liệu chuỗi thời gian có cả tính xu hướng và tính mùa vụ - Hình 1(d)



Hình 1. Các dạng dữ liệu chuỗi thời gian

Dữ liệu chuỗi thời gian được ứng dụng rất rộng rãi trong nhiều lĩnh vực như: Dự báo kinh tế; Phân tích thời gian thực; Tính toán doanh số bán hàng; Phân tích thị trường; Dự báo thời tiết….(Shumway and et al., 2017). Quá trình thu thập dữ liệu chuỗi thời gian có thể do nhiều nguyên nhân chủ quan, khách quan một số thời điểm trong chuỗi bị thiếu, mất mát. Do đó việc xử lý giá trị thiếu là vấn đề bắt buộc trước khi có thể sử dụng được dữ liệu này cho bất kỳ mục đích nào (Xi Wang, Chen Wang, 2019).

2.2 Một số phương pháp xử lý giá trị thiếu

Xử lý giá trị thiếu (missing values) luôn là một bước quan trọng và bắt buộc trong quá trình làm sạch dữ liệu. Có rất nhiều phương pháp để xử lý dữ liệu thiếu, 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 (Choi J and et al., 2018). Trong thực tế, dữ liệu thiếu được chia thành 3 dạng, bao gồm:

- Dữ liệu thiếu hoàn toàn ngẫu nhiên (Missing Completely at Random – MCAR): Sự mất mát dữ liệu tại các điểm quan sát là hoàn toàn ngẫu nhiên, và không có bất kỳ một mối quan hệ hay sự liên hệ nào giữa dữ liệu thiếu với các dữ liệu quan sát khác.
- Dữ liệu thiếu ngẫu nhiên (Missing at Random – MAR): Sự mất mát dữ liệu ở đây là ngẫu nhiên, tuy nhiên vẫn có mối quan hệ hệ thống giữa dữ liệu bị mất và dữ liệu được quan sát.

- Dữ liệu thiếu không ngẫu nhiên (Missing Not at Random – MNAR): Sự mất mát dữ liệu không phải là ngẫu nhiên mà có một mối quan hệ xu hướng giữa giá trị bị thiếu và giá trị không bị thiếu trong một biến.

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 (Carl Bonander, Ulf Stromberg, 2018). Có 4 giải pháp đơn giản nhưng hiệu quả để xử lý dữ liệu thiếu cho chuỗi thời gian bao gồm:

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

Phương pháp này đơn giản là thay thế các điểm dữ liệu thiếu bằng các dữ liệu của điểm liền trước nó. Hình 1 minh họa việc xử lý giá trị thiếu bằng phương pháp LOCF. Hình 1(a) thể hiện một đoạn dữ liệu chuỗi thời gian theo ngày, trong đó có 3 ngày dữ liệu bị thiếu là các ngày 05/01, 07/01 và 08/01 (ký hiệu N/A). Hình 1(b) thể hiện kết quả khi sử dụng phương pháp LOCF để xử lý giá trị thiếu. Dữ liệu thiếu ngày 05/01 sẽ được thay thế bằng dữ liệu của ngày liền trước đó là ngày 04/01.

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	N/A	86%
6	6-Jan	155	87%
7	7-Jan	N/A	89%
8	8-Jan	N/A	90%
9	9-Jan	180	92%

(a)

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	90	86%
6	6-Jan	155	87%
7	7-Jan	155	89%
8	8-Jan	155	90%
9	9-Jan	180	92%

(b)

Hình 1(a). Dữ liệu gốc ban đầu chứa các dữ liệu thiếu (N/A)
(b). Kết quả thay thế giá trị thiếu bằng phương pháp LOCF

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

Phương pháp này thực hiện tương tự như với phương pháp LOCF chỉ khác biệt là sử dụng các giá liền sau để thay thế cho giá trị thiếu. Hình 2 minh họa việc sử dụng phương pháp NOCB để xử lý giá trị thiếu. Dữ liệu thiếu ngày 05/01 sẽ được thay thế bằng dữ liệu của ngày liền sau đó là ngày 06/01.

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	N/A	86%
6	6-Jan	155	87%
7	7-Jan	N/A	89%
8	8-Jan	N/A	90%
9	9-Jan	180	92%

(a)

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	155	86%
6	6-Jan	155	87%
7	7-Jan	180	89%
8	8-Jan	180	90%
9	9-Jan	180	92%

(b)

Hình 2(a). Dữ liệu gốc ban đầu chứa các dữ liệu thiếu (N/A)
(b). Kết quả thay thế giá trị thiếu bằng phương pháp NOCB

c) Thay thế giá trị thiếu bằng phương nội suy tuyến tính (Linear interpolation):

Nội suy là phương pháp ước tính giá trị của các điểm dữ liệu chưa biết trong phạm vi của một tập hợp rời rạc chứa một số điểm dữ liệu đã biết. Nội suy tuyến tính là nội suy đơn giản nhất, được thực hiện bằng cách lấy giá trị trung bình giữa giá trị của điểm dữ liệu trước và sau điểm dữ liệu bị thiếu. Hình 3 minh họa phương pháp xử lý giá trị thiếu bằng phương pháp nội suy tuyến tính. Dữ liệu thiếu ngày 05/01 được xác định dựa trên dữ liệu ngày trước đó 04/01 và ngày sau đó 06/01. Bản chất của nội suy tuyến tính là xây dựng một đường thẳng đi qua 2 điểm trước và sau của điểm dữ liệu bị thiếu để sau đó xác định giá trị của điểm dữ liệu thiếu này.

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	N/A	86%
6	6-Jan	150	87%
7	7-Jan	160	89%
8	8-Jan	N/A	90%
9	9-Jan	180	92%

(a)

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	120	86%
6	6-Jan	150	87%
7	7-Jan	160	89%
8	8-Jan	170	90%
9	9-Jan	180	92%

(90+150)/2 = 120

(160+180)/2 = 170

(b)

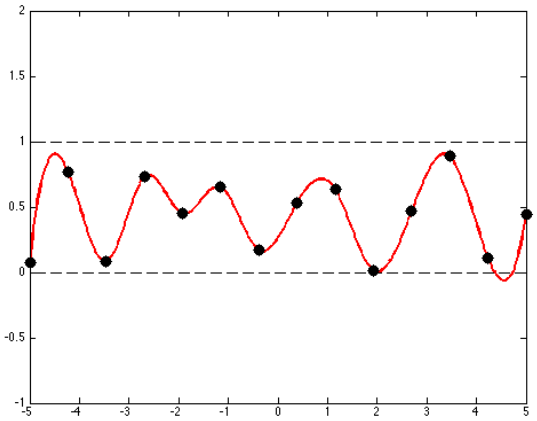
Hình 3(a). Dữ liệu gốc ban đầu chứa các dữ liệu thiếu (N/A)
(b) Kết quả thay thế giá trị thiếu bằng phương pháp nội suy tuyến tính

d) Thay thế giá trị thiếu bằng phương nội suy spline (Spline interpolation):

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 sao $f(x_i) = y_i$ 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 $[x_i, x_{i+1}]$ (Erdogan KAYA, 2014). Trong thực tế nhóm tác giả sử dụng nội suy spline với đa thức bậc 3 khi đó $p_i(x)$ được định nghĩa như sau:

$$p_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i \quad \text{(Ajao and et al., 2012)}$$

Hình 4 minh họa việc xây dựng các đường cong bậc 3 (đường màu đỏ) đi qua 14 điểm đã biết (điểm chấm đen). Dựa vào các đường cong đi qua các điểm đã biết này để xác định giá trị của các điểm dữ liệu bị thiếu.



Hình 4. Nội suy Spline bậc 3 qua 14 điểm đã biết

Các phương pháp này áp dụng cho dữ liệu chuỗi thời gian với giả thiết là dữ liệu liên tục, các điểm dữ liệu có mối quan hệ tương quan với nhau. Dữ liệu nhiệt độ tại các thời điểm trong ngày là dữ liệu dạng này. Dữ liệu tại một thời điểm sẽ có mối quan hệ với dữ liệu tại điểm trước và sau đó. Vì vậy trong phần thực nghiệm khi đánh giá độ chính xác của các phương pháp thay thế giá trị thiếu cho dữ liệu nhiệt độ tại 5 trạm khu vực phía bắc, nhóm tác giả sẽ thực hiện cả 4 phương pháp đã trình bày ở trên.

3. Chỉ số xác định sai số

Để đánh giá độ chính xác của dữ liệu thay thế với giá trị quan sát, trong thống kê người ta thường sử dụng 4 chỉ số bao gồm: Sai số trung bình (ME – Mean Error); Sai số tuyệt đối trung bình (MAE – Mean Absolute Error); Sai số bình phương trung bình (MSE – Mean Square Error) và sai số bình phương trung bình quân phương (RMSE – Root Mean Square Error). Các chỉ số đánh giá này được xác định như sau:

a) Sai số trung bình (ME):

$$ME = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)$$

Trong đó:

- F_i là giá trị thay thế (trong phần thực nghiệm F_i là giá trị thay thế có được bằng một trong các phương như LOCF, NOCB, nội suy Linear, nội suy Spline).
- O_i là giá trị thật của biến (trong phần thực nghiệm O_i là giá trị nhiệt độ quan trắc).
- N là tổng số các điểm xác định sai số

Các thông số F_i, O_i và N tương tự cho MAE, MSE và RMSE

Sai số trung bình có giá trị nằm trong khoảng (-∞, +∞). ME cho biết xu hướng lệch trung bình của

giá trị thay thế so với giá trị quan trắc, nhưng không phản ánh độ lớn của sai số. ME dương cho biết giá trị thay thế vượt quá giá trị quan trắc và ngược lại. Nếu ME = 0 được xem là "hoàn hảo" khi đó dữ liệu thay thế trùng với dữ liệu quan trắc.

b) Sai số tuyệt đối trung bình (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|F_i - O_i|$$

Sai số tuyệt đối trung bình nằm trong khoảng (0,+∞). MAE biểu thị biên độ trung bình của sai số mô hình nhưng không nói lên xu hướng lệch của giá trị thay thế và giá trị quan trắc. Khi MAE = 0, các giá trị thay thế hoàn toàn trùng khớp với các giá trị quan trắc, khi đó phương pháp xử lý giá trị thiếu được xem là "lý tưởng". Thông thường, MAE được sử dụng cùng với ME để đánh giá độ tin cậy.

c) Sai số bình phương trung bình (MSE):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2$$

Sai số bình phương trung bình nằm trong khoảng (0,+∞), MSE phản ánh mức độ dao động giữa giá trị thay thế với giá trị quan trắc.

d) Sai số bình phương trung bình quân phương (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2}$$

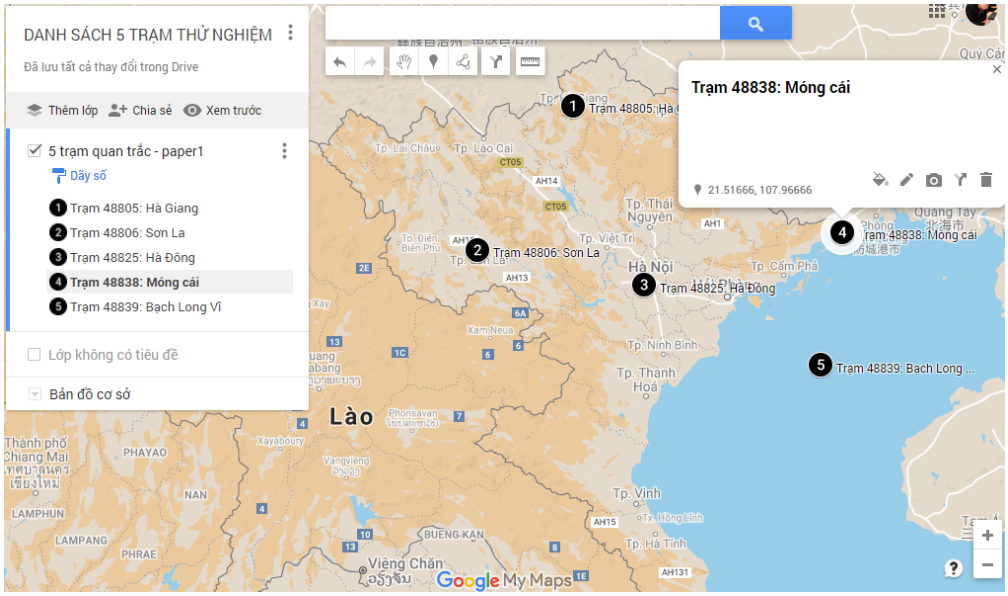
Sai số bình phương trung bình là một trong những đại lượng cơ bản và thường được sử dụng phổ biến cho việc đánh giá kết quả giữa giá trị thay thế và giá trị quan trắc. Người ta thường hay sử dụng RMSE biểu thị độ lớn trung bình của sai số. Đặc biệt RMSE rất nhạy với những giá trị sai số lớn. Giống như MAE, RMSE không chỉ ra độ lệch giữa giá trị dự báo và giá trị quan trắc. Giá trị của RMSE nằm trong khoảng (0,+ ∞)

Trong phần thực nghiệm, nhóm tác giả sử dụng 2 chỉ số là MAE và RMSE để xác định sai số giữa giá trị thay thế và giá trị quan trắc thực tế, qua đó xác định phương pháp thay thế nào cho độ chính xác cao hơn (nghĩa là có sai số nhỏ hơn).

4. Thực nghiệm trên dữ liệu nhiệt độ tại 5 trạm quan trắc

4.1. Mô tả dữ liệu

Để đánh giá độ chính xác của các phương pháp xử lý dữ liệu thiếu đã trình bày trong phần 2, nhóm tác giả sử dụng tập dữ liệu nhiệt độ thu thập được từ 5 trạm quan trắc bao gồm: Trạm 48805 – Hà Giang; Trạm 48806 – Sơn La; Trạm 48825 – Hà Đông; Trạm 48838 – Móng Cái và Trạm 48839 – Bạch Long Vĩ. Vị trí của các trạm trên Google Map như hình 5. Cả 5 trạm này đều là các trạm quan trắc 3h, nghĩa là sẽ thực hiện thu thập dữ liệu khí tượng 8 lần mỗi ngày, mỗi lần cách nhau 3 giờ tại các thời điểm 00h, 03h, 06h, 09h, 12h, 15h, 18h, 21h theo giờ GMT, tương ứng với 01h, 04h, 07h, 10h, 13h, 16h, 19h, 22h giờ Việt Nam. Đây là các dữ liệu thực tế được cung cấp bởi Trung tâm thông tin và dữ liệu khí tượng thủy văn; Trong phần thực nghiệm này, nhóm tác giả sử dụng dữ liệu nhiệt độ quan trắc trong năm 2019 từ thời điểm 01h ngày 01 tháng 01 đến 22h ngày 31 tháng 12. Dữ liệu được lưu trữ trong file .CSV bao gồm các thông tin về thời điểm quan trắc và giá trị nhiệt độ (°C) của 5 trạm này. Hình 6 mô tả chi tiết file dữ liệu, trong đó dòng đầu tiên là tiêu đề của file dữ liệu, bao gồm 6 cột. Cột đầu tiên cho biết thời điểm quan trắc, 5 cột còn lại tương ứng với giá trị nhiệt độ tại 5 trạm lấy theo mã trạm.



Hình 5. Vị trí 5 trạm quan trắc trên Google Maps

	A	B	C	D	E	F
1	TimeVN	48805	48806	48825	48838	48839
2	2019-01-01 1:00	12.4	9.4	11.8	9.3	12.2
3	2019-01-01 4:00	12.2	9.3	11.4	9.3	12.1
4	2019-01-01 7:00	12.2	9.1	11	9.4	12.2
5	2019-01-01 10:00	13.2	11.1	11.6	11.2	12
6	2019-01-01 13:00	14.8	13.1	12.2	13	12.6
7	2019-01-01 16:00	14.6	13.7	13	12	12.8
8	2019-01-01 19:00	13.2	12	12.3	10.4	13
9	2019-01-01 22:00	12.2	11.2	12	10.4	12.8
10	2019-01-02 1:00	12.1	10.6	12	10.6	13
11	2019-01-02 4:00	12.2	10.4	12	10.4	13
12	2019-01-02 7:00	11.8	10.4	11.9	10.4	12.8
13	2019-01-02 10:00	14	12	13.8	12.6	13.1
14	2019-01-02 13:00	16.2	13.4	16.1	16.2	14.1
15	2019-01-02 16:00	16.1	13.1	16.2	15	14.2
16	2019-01-02 19:00	14.6	12.1	15.3	13.4	14.3
17	2019-01-02 22:00	14	11.9	15.2	13	14.1
18	2019-01-03 1:00	13.6	11.6	14.6	12.8	13.2
19	2019-01-03 4:00	13.4	11.2	13.6	11.8	12.9

Data_Original

Hình 6. File dữ liệu của 5 trạm thực nghiệm

Bảng 1 dưới đây tổng hợp các thông số và một số đặc trưng thống kê liên quan đến dữ liệu của 5 trạm này, qua đó thấy được cái nhìn tổng quan về tập dữ liệu. Dữ liệu của 5 trạm này trong năm 2019 là dữ liệu đầy đủ, không bị thiếu giá trị nào. Đây là file dữ liệu chuẩn sẽ được sử dụng để đánh giá độ chính xác của các phương pháp xử lý giá trị thiếu.

Bảng 1. Thông số thông kế cho tập dữ liệu ban đầu (Data_Original)

Thông số	Trạm quan trắc				
	48805	48806	48825	48838	48839
Tổng số điểm dữ liệu quan trắc	2920	2920	2920	2920	2920
Số điểm có dữ liệu	2920	2920	2920	2920	2920
Số điểm dữ liệu thiếu	0	0	0	0	0
Nhiệt độ thấp nhất (°C)	6.9	2.9	10.3	7.4	12
Nhiệt độ cao nhất (°C)	38.5	36.9	39.7	35.5	34.4
Nhiệt độ trung bình (°C)	24.405	22.939	25.482	24.133	25.341
Độ lệch chuẩn	5.237	5.329	5.435	5.424	4.502

Để đánh giá độ chính xác của các phương pháp xử lý dữ liệu thiếu, nhóm tác giả sẽ thực hiện loại bỏ đi một số giá trị trong tập dữ liệu gốc với cả 3 dạng mất mát dữ liệu khác nhau bao gồm MAR, MCAR và MNAR cho dữ liệu nhiệt độ của 5 trạm này. Tổng số điểm dữ liệu bị loại bỏ tại mỗi trạm là 35 điểm trong

tổng số 2920 điểm, tương ứng với 1.2% tổng số điểm dữ liệu. Mỗi một trạm sẽ có một kiểu mất mát dữ liệu khác nhau cụ thể như sau:

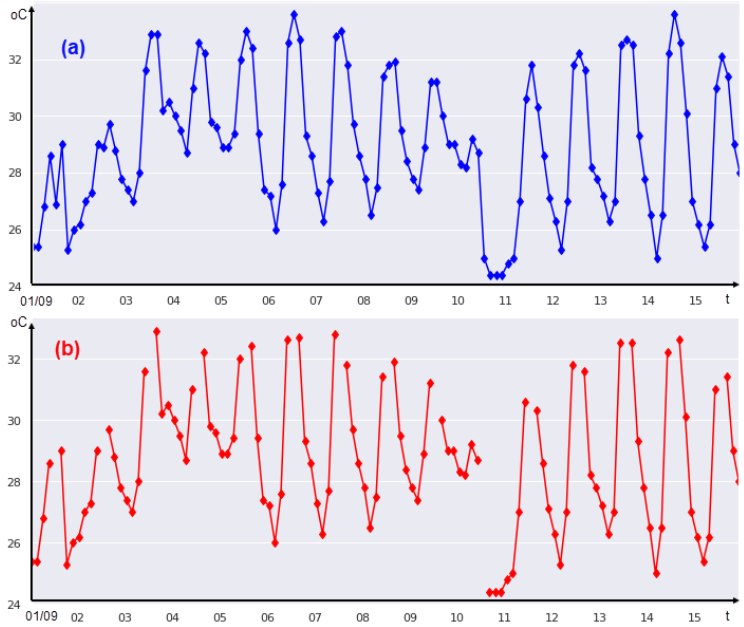
- Trạm 48805: Các điểm mất mát dữ liệu là hoàn toàn ngẫu nhiên; Dữ liệu gốc bị loại bỏ từng điểm quan trắc riêng rẽ hoàn toàn ngẫu nhiên.
- Trạm 48806: Các điểm mất mát dữ liệu là hoàn toàn ngẫu nhiên; Dữ liệu gốc bị loại bỏ một số điểm liền nhau có thể là 2 điểm, 3 điểm hoặc một ngày liên tục hoàn toàn ngẫu nhiên.
- Trạm 48825: Dữ liệu mất mát không ngẫu nhiên; Dữ liệu gốc bị loại bỏ tại thời điểm 07h (thời điểm nhiệt độ trung bình thấp nhất trong ngày) của 35 ngày liên tiếp.
- Trạm 48838: Dữ liệu mất mát không ngẫu nhiên; Dữ liệu gốc bị loại bỏ tại thời điểm 13h (thời điểm nhiệt độ trung bình cao nhất trong ngày) của 35 ngày liên tiếp.
- Trạm 48839: Dữ liệu mất mát không ngẫu nhiên; Dữ liệu gốc bị loại bỏ tại các thời điểm 10h và 19h các ngày 01 và 15 hàng tháng.

Bảng 2 mô tả thông số của tập dữ liệu sau khi đã loại bỏ đi một số giá trị để trở thành tập dữ liệu thiếu theo các nguyên tắc ở trên.

Bảng 2. Thông số thông kế cho tập dữ liệu chứa giá trị thiếu (Data_Missing)

Thông số	Trạm quan trắc				
	48805	48806	48825	48838	48839
Tổng số điểm dữ liệu quan trắc	2920	2920	2920	2920	2920
Số điểm có dữ liệu	2885	2885	2885	2885	2885
Số điểm dữ liệu thiếu	35	35	35	35	35
Nhiệt độ thấp nhất (°C)	6.9	2.9	10.3	7.4	12
Nhiệt độ cao nhất (°C)	38.5	36.9	39.7	35.5	34.4
Nhiệt độ trung bình (°C)	24.397	22.899	25.450	24.040	25.329
Độ lệch chuẩn	5.245	5.337	5.458	5.385	4.508

Như vậy, chúng ta đã có 2 tập dữ liệu; Tập dữ liệu Data_Original là tập dữ liệu gốc, chứa dữ liệu nhiệt độ quan trắc tại 5 trạm trong năm 2019 với 2920 thời điểm, mỗi thời điểm cách nhau 3h. Đây là tập dữ liệu đầy đủ, không bị thiếu giá trị nào; Tập dữ liệu Data_Missing là tập dữ liệu lấy từ tập Data_Original nhưng đã bị loại bỏ đi 35 điểm quan trắc cho mỗi trạm (dữ liệu thiếu chiếm ~1.2% tổng số điểm dữ liệu), Dữ liệu trong mỗi trạm sẽ bị loại bỏ đi theo những dạng mất mát khác nhau bao gồm MCAR và MNAR như đã trình bày ở trên.



Hình 7(a). Dữ liệu đầy đủ của trạm 48838 từ 01/09 đến 15/09
(b). Dữ liệu đã bị loại bỏ các thời điểm 13h trong khoảng thời gian tương ứng

Hình 7 minh họa dữ liệu trạm 48838 trong khoảng thời gian từ 01/09 đến 15/09/2019. Hình 7(a) thể hiện

dữ liệu trong tập Data_Origninal, dữ liệu gốc thu tập được không bị mất mát. Hình 7(b) thể hiện dữ liệu trong tập Data_Missing của trạm 48838, Dữ liệu này đã bị loại bỏ tại thời điểm lúc 13h trong 35 ngày liên tiếp.

4.2. Áp dụng các phương pháp xử lý giá trị thiếu

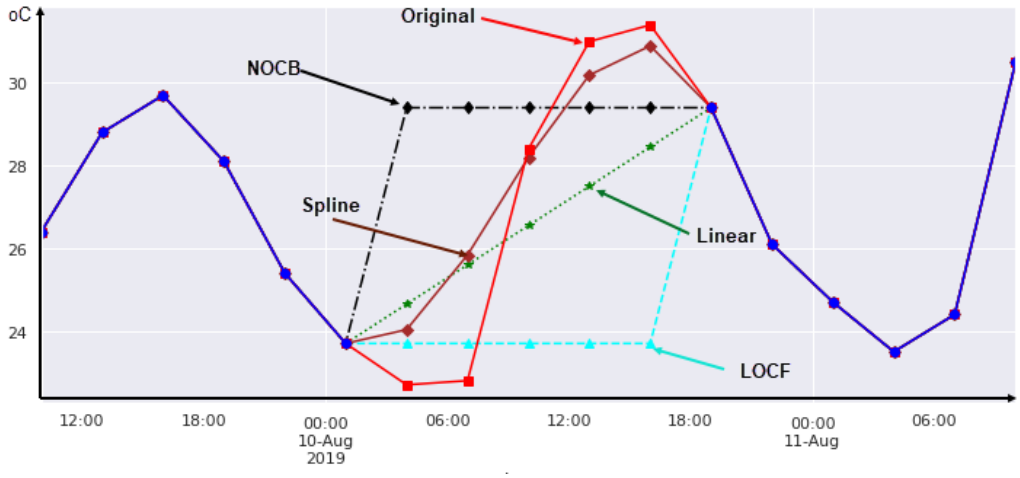
Để xử lý giá trị thiếu theo các phương pháp đã trình bày trong phần 2, nhóm tác giả sử dụng ngôn ngữ lập trình Python, thư viện hỗ trợ phân tích và xử lý số liệu Pandas. Pandas là một trong những thư viện rất mạnh mẽ trong việc xử lý dữ liệu, đặc biệt là dữ liệu chuỗi thời gian.

Nhóm tác giả sẽ tiến hành lần lượt việc thay thế các điểm dữ liệu thiếu theo 4 phương pháp LOCF, NOCF, nội suy tuyến tính (linear) và nội suy Spline bậc 3. Bảng 3 thể hiện kết quả thay thế các giá trị thiếu bằng 4 phương pháp cho trạm 48806 thời điểm ngày 10-08-2019. Cột đầu tiên thể hiện thời điểm quan trắc dữ liệu, với tần suất 3h một lần, trong một ngày sẽ thu thập dữ liệu tại 8 thời điểm. Cột 48806_mis là dữ liệu lấy từ tập Data_Missing, như vậy trong ngày 10/08/2019 có 5 điểm dữ liệu quan trắc liền nhau (4h, 7h, 10h, 13h và 16h) bị mất mát (trong Pandas dữ liệu thiếu ký hiệu NaN). Cột 48806_ori là dữ liệu lấy từ tập Data_Original, đây là các giá trị quan trắc chính xác tại các thời điểm tương ứng, các giá trị đúng này sẽ làm cơ sở để đánh giá sai số cho các phương pháp xử lý dữ liệu thiếu được trình bày trong phần 4.3 dưới đây. Cột LOCF - kết quả thay thế các giá trị thiếu bằng phương pháp thay thế giá trị tại thời điểm liền trước. Cột NOCB - kết quả thay thế giá trị thiếu bằng phương pháp thay thế giá trị tại thời điểm liền sau. Cột Linear - kết quả thay thế giá trị thiếu bằng phương pháp nội suy tuyến tính. Cột Spline – kết quả thay thế giá trị thiếu bằng phương pháp nội suy spline bậc 3.

Bảng 3. Kết quả xử lý giá trị thiếu theo các phương pháp khác nhau

Thời điểm	48806_mis	48806_ori	LOCF	NOCB	Linear	Spline
2019-08-10 01:00:00	23.7	23.7	23.7	23.7	23.7	23.7
2019-08-10 04:00:00	NaN	22.7	23.7	29.4	24.7	24.0
2019-08-10 07:00:00	NaN	22.8	23.7	29.4	25.6	25.8
2019-08-10 10:00:00	NaN	28.4	23.7	29.4	26.6	28.2
2019-08-10 13:00:00	NaN	31.0	23.7	29.4	27.5	30.2
2019-08-10 16:00:00	NaN	31.4	23.7	29.4	28.5	30.9
2019-08-10 19:00:00	29.4	29.4	29.4	29.4	29.4	29.4
2019-08-10 22:00:00	26.1	26.1	26.1	26.1	26.1	26.1

Hình 8 là đồ thị thể hiện trực quan giữa dữ liệu quan trắc thực (các điểm màu đỏ) và dữ liệu tại các điểm thay thế sử dụng các phương pháp xử lý giá trị thiếu khác nhau.



Hình 8. Trực quan hóa dữ liệu thực và các giá trị thay thế dữ liệu thiếu theo 4 phương pháp

4.3. Đánh giá độ chính xác của các phương pháp

Để đánh giá phương pháp xử lý giá trị thiếu nào trong 4 phương pháp LOCF, NOCB, Linear và Spline cho kết quả gần với giá trị quan trắc nhất (Sai số giữa giá trị thay thế và giá trị quan trắc nhỏ nhất) nhóm tác giả sử dụng 4 độ đo bao gồm: ME, MAE, MSE, RMSE. Hình 9 là mã nguồn các hàm tính toán sai số theo các công thức đã được trình bày trong phần 3 của bài báo.

```python
#Xây dựng các hàm tính sai số MA, MAE, MSE, RMSE
import numpy as np
def me(actual,predicted):
    """ Mean Error """
    error = np.mean(actual -  predicted)
    return round(error,4)

def mae(actual, predicted):
    """ Mean Absolute Error """
    error = np.mean(abs(actual -  predicted))
    return round(error,4)

def mse(actual, predicted):
    """ Mean Squared Error """
    error = np.mean(np.square(actual -  predicted))
    return round(error,4)

def rmse(actual, predicted):
    """ Root Mean Squared Error """
    error = np.sqrt(np.mean(np.square(actual -  predicted)))
    return round(error,4)
```

Hình 9. Xây dựng các hàm tính sai số ME, MAE, MSE, RMSE

Các bảng 4,5 thể hiện kết quả sai số MAE và RMSE giữa giá trị thay thế với giá trị quan trắc tại 4 trạm ứng với 4 phương pháp xử lý giá trị thiếu.

Bảng 4: Tổng hợp sai số MAE của 4 phương pháp xử lý giá trị thiếu

Trạm	MAE_LOCF	MAE_NOCB	MAE_Linear	MAE_Spline
48805	0.0201	0.0188	0.0105	**0.0078**
48806	0.0344	0.0471	0.0301	**0.0287**
48825	**0.0076**	0.0316	0.0135	0.0086
48838	0.0175	0.0156	0.0157	**0.0123**
48839	0.0221	**0.0100**	0.0106	0.0106

Bảng 5: Tổng hợp sai số RMSE của 4 phương pháp xử lý giá trị thiếu

Trạm	RMSE_LOCF	RMSE_NOCB	RMSE_Linear	RMSE_Spline
48805	0.2712	0.2138	0.1275	**0.1034**
48806	0.3902	0.5156	**0.3416**	0.4086
48825	**0.0823**	0.3167	0.1359	0.0977
48838	0.1967	0.2124	0.1790	**0.1436**
48839	0.2523	**0.1068**	0.1229	0.1242

Qua kết quả tổng hợp sai số giữa giá trị thay thế và giá trị quan trắc có thể nhận thấy không có một phương pháp xử lý giá trị thiếu nào là tốt cho mọi loại mất mát.

- Nếu mất mát là hoàn toàn ngẫu nhiên nhưng chỉ bị mất ứng với từng thời điểm riêng lẻ (như của trạm 48805), hoặc mất mát không ngẫu nhiên, thời điểm mất là đơn lẻ, tại thời điểm nhiệt độ cao nhất trong ngày (như của trạm 48838) thì phương pháp sử dụng nội suy Spline bậc 3 cho độ chính xác cao hơn (sai số nhỏ nhất).
- Nếu mất mát không ngẫu nhiên, thời điểm mất mát là đơn lẻ, mất mát tại thời điểm nhiệt độ thấp nhất trong ngày 07h (như của trạm 48825) thì phương pháp LOCF cho độ chính xác cao hơn.

Tuy nhiên về tổng thể, với 5 dạng mất mát khác nhau ứng với dữ liệu thiếu tại 5 trạm, phương pháp nội suy Spline bậc 3 cho sai số thấp hơn, nghĩa là độ chính xác của dữ liệu được thay thế gần với giá trị quan trắc hơn.

5. Kết luận

Chuẩn bị dữ liệu là giai đoạn bắt buộc trong bất kỳ một dự án khoa học dữ liệu nào, dữ liệu được chuẩn bị tốt sẽ giúp cho việc xây dựng các mô hình dự đoán, dự báo được chính xác. Chuẩn bị dữ liệu bao gồm rất nhiều công đoạn và yêu cầu khác nhau, trong đó xử lý giá trị thiếu (giá trị mất mát – missing values) là yêu cầu bắt buộc; Có nhiều phương pháp khác nhau để xử lý giá trị thiếu, việc lựa chọn phương pháp nào hay kết hợp nhiều phương pháp phụ thuộc bào từng bài toán, từng loại dữ liệu cụ thể. Với dữ liệu chuỗi thời gian (time series) chúng ta không thể sử dụng phương pháp loại bỏ các điểm giá trị thiếu, mà chỉ có thể thay thế các điểm dữ liệu thiếu bằng một giá trị khác phù hợp. Có rất nhiều phương pháp thay thế giá trị thiếu, tuy nhiên không có một phương pháp pháp thay thế nào là tối ưu cho mọi loại dữ liệu, mọi bài

toán. Trong nội dung của bài báo này, nhóm tác giả đã giới thiệu 4 phương pháp thay thế giá trị thiếu thường được áp dụng cho dữ liệu chuỗi thời gian mà ở đó giá trị tại mỗi điểm có mối quan hệ tương quan với các điểm phía trước và phía sau nó. Nhóm tác giả cũng thực nghiệm 4 phương pháp thay thế giá trị thiếu LOCF, NOCB, nội suy Linear và nội suy Spline trên tập dữ liệu nhiệt độ tại 5 trạm quan trắc 48805, 48806, 48825, 48838 và 48839 với các dạng mất mát MCAR và MNAR. Các kết quả của bài báo làm cơ sở cho nhóm tác giả lựa chọn được phương pháp xử lý giá trị thiếu phù hợp với loại dữ liệu nhiệt độ, và áp dụng phương pháp xử lý đó cho dữ liệu tại các trạm quan trắc khác.

Tài liệu tham khảo

Davy Cielen, Arno D. B., Meysman, Mohamed Ali, (2016). *Introducing Data Science*, Manning Publications Co.

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

Xi Wang, Chen Wang, *Time Series Data Cleaning: A Survey*, IEEE Access (12/2019) , page 1866-1881

Choi J, Dekkers OM, le Cessie S. *A comparison of different methods to handle missing data in the context of propensity score analysis*. Eur J Epidemiol, 2018.

Carl Bonander, Ulf Stromberg. *Methods to handle missing values and missing individuals*. European Journal of Epidemiology, 2018.

Erdogan KAYA. *Spline Interpolation Techniques. Journal of Technical Science and Technologies*, ISSN 2298-0032, 2014.

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

ABSTRACT

Evaluate the accuracy of some missing value processing methods for time series data; conduct temperature data experiments at 5 monitoring stations in the north

Nguyen Thi Phuong Bac[1], Dang Van Nam[1]*

[1]Hanoi University of Mining and Geology

Dealing with missing data (missing values) is one of the important steps in the data cleansing process. There are two ways to deal with missing data: delete missing values and/or replace missing data with new values. The choice of which set of methods or both to use depends on each question, specific data type and data type... Time series data is a series of data points, with a uniform time frequency for each continuous time. The treatment of missing values in time series data is also very different from the treatment of missing values of other data types. In this article, we will introduce 4 main methods for dealing with missing values of time series data, including: LOCF, NOCB, linear interpolation, spline interpolation... The results of using applying these methods to different missing data types MAE and RMSE measurements to evaluate the accuracy of each method in the 2019 temperature observation data of 5 stations in northern Vietnam.

Keywords: Missing values, Time series data, MAE, RMSE