EARTH SCIENCES AND NATURAL RESOURCES FOR SUSTAINABLE DEVELOPMENT

HỘI NGHỊ TOÀN QUỐC KHOA HỌC TRÁI ĐẤT VÀ TÀI NGUYÊN VỚI PHÁT TRIỂN BỀN VỮNG (ERSD2020 – 12/11/2020)

Đánh giá độ chính xác của một số phương pháp xử lý dữ liệu thiếu cho dữ liệu chuỗi thời gian (time series data)

Nguyễn Thị Phương Bắc, Đặng Văn Nam
Trường Đại học Mỏ - Địa chất

TÓM TẮT

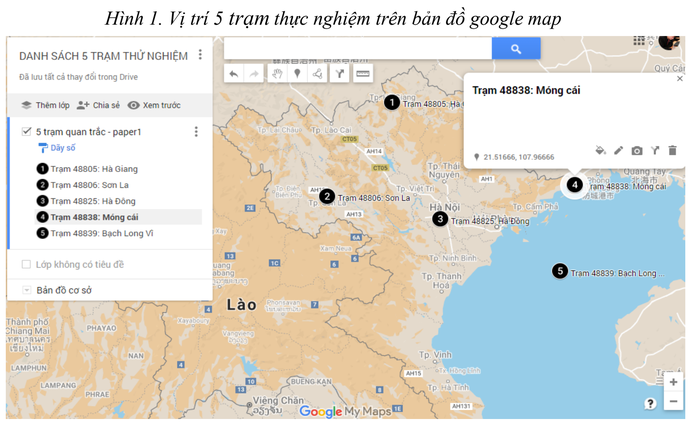
Xử lý dữ liệu thiếu (missing values) là một trong những công đoạn không thể thiếu trong quá trình làm sạch dữ liệu (Data cleansing). Có hai nhóm phương pháp xử lý dữ liệu thiếu đó là loại bỏ các giá trị thiếu và/hoặc thay thế dữ liệu thiếu bằng một giá trị mới. Việc lựa chọn sử dụng nhóm phương pháp 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.

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).

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 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.

Hình 1. Vị trí 5 trạm thực nghiệm trên bản đồ google map



Hình 2. File dữ liệu nhiệt độ quan trắc trong năm 2019

	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

Một số phương pháp xử lý giá trị thiếu

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:

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

- Thay thế giá trị thiếu bằng giá trị liền sau (NOCB-Next observation carried backward)
- Thay thế giá trị thiếu bằng phương pháp nội suy tuyến tính (Linear interpolation)
- Thay thế giá trị thiếu bằng phương pháp nội suy spline (Spline interpolation)

- 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.
- Kết quả thay thế các giá trị thiếu bằng 4 phương pháp cho trạm 48806 thời điểm ngày 10-08-2019

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

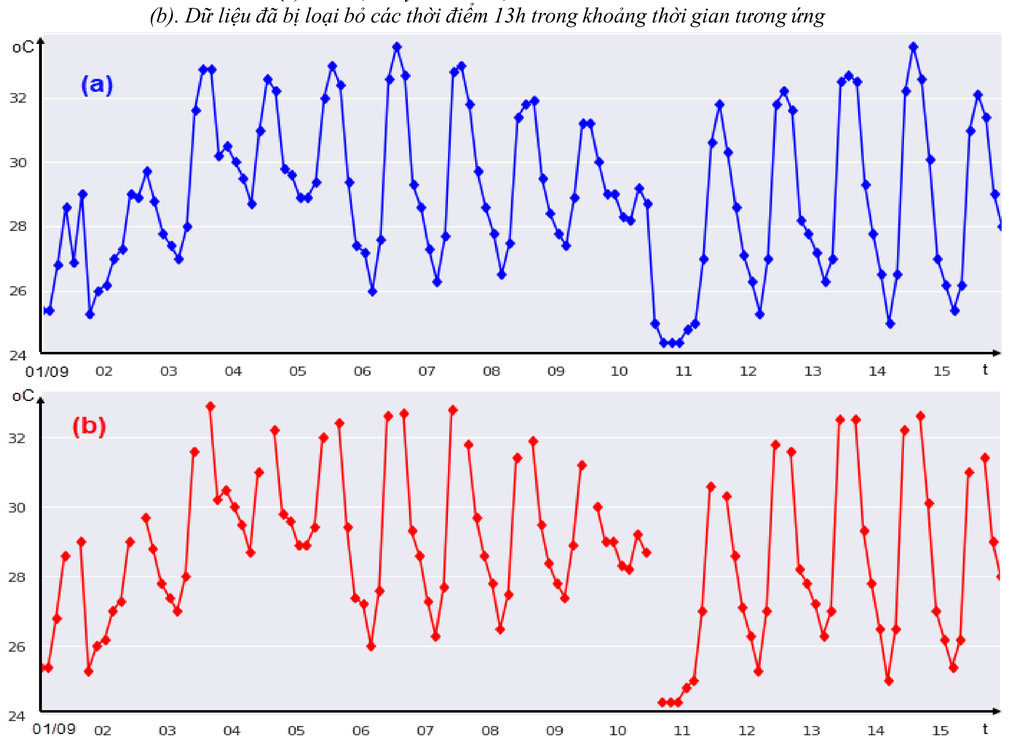
Bảng 2. Thông số thông kế cho tập dữ liệu chứa giá trị thiếu (Data_Missing)

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

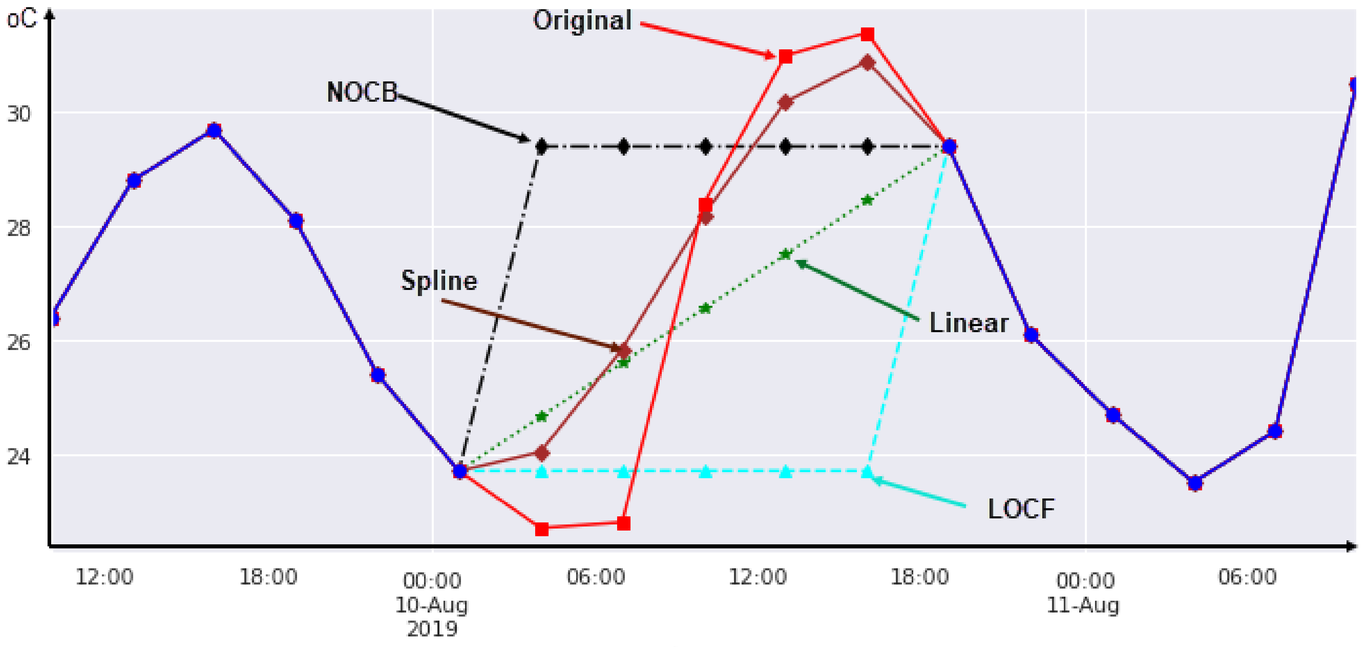
Bảng 3. Kết quả xử lý giá trị thiếu theo các phương pháp khác nhau

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



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

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.