SO SÁNH MỘT SỐ PHƯƠNG PHÁP XỬ LÝ DỮ LIỆU THIẾU CHO CHUỖI DỮ LIỆU THỜI GIAN MỘT CHIỀU

Phan Thị Thu Hồng

Khoa Công nghệ thông tin, Học viện Nông nghiệp Việt Nam

Tác giả liên hệ: ptthong@vnua.edu.vn

Ngày nhận bài: 20.07.2020 Ngày chấp nhận đăng: 10.09.2020

TÓM TẮT

Chuỗi thời gian chứa các giá trị thiếu xảy ra trong hầu hết mọi lĩnh vực khoa học ứng dụng. Bỏ qua các giá trị thiếu có thể dẫn đến giảm hiệu năng của hệ thống và kết quả không đáng tin cậy, đặc biệt là khi dữ liệu mất theo khoảng lớn. Do đó, xử lý dữ liệu thiếu là một bước rất quan trọng để thực hiện các công việc tiếp như phân lớp, phân tích dữ liệu... Bài viết này trước tiên nhằm giới thiệu các phương pháp xử lý dữ liệu thiếu. Tiếp theo một framework cho phép điền đầy dữ liệu mất mát cho chuỗi thời gian đơn biến được xây dựng. Cuối cùng, chúng tôi thực hiện so sánh hiệu suất của các phương pháp ước lượng giá trị thiếu trên ba chuỗi dữ liệu thời gian thực sử dụng bốn chỉ số đánh giá. Thông qua kết quả thử nghiệm, phương pháp DTWBI và eDTWBI đạt được kết quả vượt trội hơn các phương pháp khác khi dữ liệu có tính chất mùa vụ và không có thành phần xu hướng, trong khi đó thì na.interp tốt hơn các phương pháp khi dữ liệu có cả hai tính chất mùa vụ và xu hướng.

Từ khóa: Chuỗi thời gian một chiều, dữ liệu thiếu, ước lượng giá trị thiếu, độ tương tự.

An Empirical Study of Imputation Methods for Univariate Time Series

ABSTRACT

Time series with missing values occur in almost areas of applied science. Ignoring missing values can lead to a reduction of system performance and unreliable results, especially in case of large missing values. Therefore, handling missing data is an important task to effectively perform further purposes such as classification, data analysis, etc. This article aims first to introduce approaches for dealing with missing data. Next a framework is built to fill the incomplete data in univariate time series and then to compare the performance of various imputation methods. Four indices are used to evaluate the ability of imputation methods on 3 different real-time data series. Through experimental results, the DTWBI and eDTWBI methods achieve better results with data having seasonality component and without trend factor, while na.interp is more superior as the data have both seasonality and trend components.

Keywords: Univariate time series, missing data, imputation, similarity.

1. ĐẶT VẤN ĐỀ

Ngày nay, với tiến bộ vượt bậc của hệ thống giám sát, sự phát triển công nghệ lưu trữ dữ liệu, sự sẵn sàng của các bộ cảm biến với chi phí thấp và việc triển khai các hệ thống viễn thám, gần như tất cả dữ liệu mà con người dùng để phục vụ cho cuộc sống của mình đã được ghi nhận một cách tự động. Các dữ liệu này được lưu trữ (trong máy tính) để con người có thể truy xuất khi cần thiết. Chúng tồn tại trong nhiều ứng dụng thực tế thuộc nhiều lĩnh vực khác nhau như: kinh tế, tài chính, y tế, giáo dục, môi trường, địa lý, sinh học… và tồn tại ở nhiều dạng thức khác như: số liệu, văn bản, hình ảnh, âm thanh, đoạn phim… Tuy nhiên, các dữ liệu thu thập được thường không đầy đủ, vì nhiều lý do lỗi của một hay nhiều thiết bị cảm biến, các sai sót xảy ra trong quá trình trao đổi/ truyền tải dữ liệu, lỗi của dụng cụ đo đạc không

chính xác, điều kiện thời tiết xấu (cảm biến ngoài trời), hoặc thiếu các tác động của con người như việc thực hiện lấy mẫu nước biển (Rousseeuw & cs., 2013).

Mặt khác, hầu hết các mô hình dự báo hoặc các mô hình phân tích chuỗi thời gian (đơn biến hay đa biến) thường gặp khó khăn khi xử lý các bộ dữ liệu không đầy đủ, mặc dù đó là những kĩ thuật mạnh (như mạng nơron, mô hình Markov ẩn, rừng ngẫu nhiên...). Những mô hình này đòi hỏi dữ liệu phải đầy đủ trong quá trình học (xây dựng mô hình dự đoán). Hơn nữa việc thiếu dữ liệu tạo ra sự mất thông tin và có thể là nguyên nhân dẫn đến việc giải thích dữ liệu không chính xác, sai lệch.

Thiếu dữ liệu hoặc thiếu giá trị nghĩa là có sự tồn tại của các quan sát nhưng giá trị không được thu thập hoặc mất sau khi thu thập hoặc tương ứng với các giá trị sai (nằm ngoài phạm vi cảm biến) trong cơ sở dữ liệu. Việc tìm ra hoặc hiểu các nguyên nhân gây ra dữ liệu bị thiếu là rất quan trọng. Việc này giúp phát triển, đề xuất hoặc tìm ra một phương pháp xử lý dữ liệu thiếu thích hợp (Moritz & cs., 2015). Nhưng trong thực tế, việc hiểu nguyên nhân vẫn là một nhiệm vụ đầy thách thức khi thiếu dữ liệu hoàn toàn không thể biết được hoặc khi những dữ liệu này có phân phối phức tạp (Molenberghs & cs., 2014). Theo các nhà thống kê học, nguyên nhân của việc xuất hiện các dữ liệu thiếu có thể phân thành 3 trường hợp: "Thiếu dữ liệu hoàn toàn ngẫu nhiên" (Missing Completely At Random, MCAR), "Thiếu dữ liệu là ngẫu nhiên" (Missing At Random, MAR) và "Thiếu dữ liệu không phải là ngẫu nhiên" (Not Missing At Random, NMAR) (Little & cs., 2014).

Thiếu dữ liệu hoàn toàn ngẫu nhiên (Missing Completely At Random, MCAR)

Dữ liệu bị thiếu được coi là MCAR khi sự thiếu dữ liệu không liên quan đến bất kỳ giá trị nào của chính biến bị thiếu hoặc các giá trị của bất kỳ biến nào khác. Điều này có nghĩa là các điểm dữ liệu bị thiếu này tạo thành một tập hợp con ngẫu nhiên của dữ liệu và hoàn toàn không có hệ thống. Ví dụ, khi một người từ chối tiết lộ thu nhập của mình, điều này không ảnh hưởng đến thu nhập thực tế của anh ta cũng như thu nhập của gia đình anh ta. Do đó, bỏ qua các giá trị thiếu MCAR không làm cho phân tích dữ liệu bị sai lệch nhưng sẽ làm tăng sai số chuẩn của các ước tính mẫu do kích thước mẫu giảm (Dong & cs., 2013).

Thiếu dữ liệu ngẫu nhiên (Missing At Random, MAR)

Thiếu dữ liệu ngẫu nhiên là kiểu thiếu dữ liệu mà xác suất của giá trị thiếu chỉ phụ thuộc vào dữ liệu được quan sát, chứ không phụ thuộc vào phần dữ liệu bị thiếu. Hay nói cách khác, các giá trị thiếu của một biến phụ thuộc vào các giá trị có sẵn của chính nó và các biến khác. Điều này cho phép có thể ước tính dữ liệu thiếu dựa trên các biến khác. Ví dụ, đánh giá học sinh tham gia một môn học bao gồm hai bài kiểm tra: bài kiểm tra giữa kỳ và bài kiểm tra cuối kỳ. Để làm bài kiểm tra cuối kỳ, học sinh phải vượt qua bài kiểm tra giữa kỳ. Giả sử rằng một sinh viên trượt kỳ thi giữa kỳ và sinh viên ấy bỏ học. Vì vậy, việc thiếu điểm kỳ thi cuối của sinh viên này là MAR.

Thiếu dữ liệu không ngẫu nhiên (Not Missing At Random, NMAR)

Dữ liệu bị thiếu là kiểu thiếu dữ liệu ngẫu nhiên nếu xác suất xuất hiện của các giá trị bị thiếu phụ thuộc vào các giá trị bị thiếu khác. Do đó, với loại dữ liệu bị thiếu này, chúng ta không thể ước tính dữ liệu không đầy đủ từ các dữ liệu hiện có.

Lưu ý rằng các nguyên nhân gây ra sự thiếu dữ liệu này chỉ là giả định về lý do thiếu dữ liệu trong ngữ cảnh phân tích. Do đó, theo quan điểm giả thuyết (các nhà thống kê học), chúng không thể được xác minh (ngoại trừ giả thuyết MCAR) các giả thuyết này. Vì vậy, việc gán nguyên nhân các giá trị bị thiếu cho một loại dữ liệu thiếu ở trên là không rõ ràng và chắc chắn (Moritz & cs., 2015). Hiện nay hầu hết các nghiên cứu tập trung vào ba loại dữ liệu bị thiếu ở trên để nghiên cứu đề xuất hoặc lựa chọn phương pháp điền dữ liệu tương ứng. Tuy nhiên Molenberghs & cs. (2014) khuyên rằng sẽ luôn luôn tốt hơn khi chúng ta kiểm kiểm tra mức độ chính xác của kết quả phân tích đối với các giả định khác nhau.

Do đó, xử lý các dữ liệu mất mát nói chung và mất mát dữ liệu trong chuỗi dữ liệu thời gian nói riêng đóng vai trò đặc biệt quan trọng trong học máy, khai phá và xử lý dữ liệu thống kê, là tiền đề để thực hiện tiếp các mục đích khác như phân tích, phân lớp, dự báo... Trong bài báo này, chúng tôi trình bày một số tiếp cận xử lý dữ liệu thiếu và thực hiện so sánh khả năng điền đầy dữ liệu thiếu của một số phương pháp cho các chuỗi dữ liệu thời gian một chiều khác nhau. Điều này cho phép người dùng lựa chọn phương pháp điền đầy dữ liệu phù hợp với tính chất của dữ liệu chuỗi thời gian một chiều.

2. PHƯƠNG PHÁP XỬ LÝ DỮ LIỆU THIẾU

Trên thực tế, các lĩnh vực khác nhau có kiểu dữ liệu đặc trưng và cách thức lưu trữ khác nhau. Do đó không có phương pháp chuyên dụng nào thực sự thỏa đáng được khuyên dùng cho việc xử lý dữ liệu thiếu mà phải tùy thuộc vào kiểu dữ liệu và loại dữ liệu thiếu để từ đó quyết định áp dụng hoặc đề xuất các phương pháp phù hợp (dựa vào các phân tích để kết quả có sai số nhỏ nhất có thể). Đặc biệt, là các phương pháp trả lời câu hỏi "Làm thế nào để có thể xử lý được dữ liệu thiếu trong các cơ sở dữ liệu lớn?". Trong phần này chúng tôi trình bày một số phương pháp cơ bản xử lý dữ liệu thiếu đó là: (1) Bỏ qua các giá trị thiếu (2) Ước lượng các giá trị bị thiếu.

2.1. Bỏ qua các giá trị thiếu (Deletion method)

Bỏ qua tất cả những quan sát không có dữ liệu được xem như phân tích trường hợp đầy đủ và là một trong những phương pháp phổ biến nhất (Horton & Kleinman, 2007). Có hai phương pháp thường được áp dụng đó là:

2.1.1. Listwise Deletion

Phương pháp Xóa Listwise thực hiện xóa mọi trường hợp dữ liệu thiếu giá trị cho một hoặc nhiều biến (Gelman & Hill, 2006). Cách tiếp cận này loại bỏ tất cả các trường hợp dữ liệu có giá trị bị thiếu, dẫn đến một tập dữ liệu chỉ quan sát đầy đủ. Phương pháp này phổ biến do tính đơn giản và dễ thực hiện. Một mặt, nó đảm bảo rằng dữ liệu không chứa các giá trị bị thiếu và không có giá trị ngẫu nhiên nào được thêm vào. Tuy nhiên, nó làm giảm kích thước của tập dữ liệu. Bằng cách loại bỏ các quan sát với bất kỳ giá trị bị thiếu nào, một số thông tin về bộ dữ liệu sẽ bị mất, điều này có thể dẫn đến kết quả sai lệch.

Bảng 2 cho thấy một ví dụ về Xóa Listwise. Tập dữ liệu với các giá trị bị thiếu được hiển thị trong bảng 1. có 10 bản ghi và sau khi thực hiện Xóa Listwise chỉ còn 6 trường hợp. Đối với các bộ dữ liệu có nhiều giá trị bị thiếu, lượng dữ liệu bị loại bỏ lớn và do đó, dữ liệu có ý nghĩa sẽ bị mất.

Bảng 1. Tập dữ liệu mẫu chứa các giá trị thiếu

STT	Ngày	Giờ	Mực nước	Lưu lượng
1	1/1/2008	1	130	612
2	1/1/2008	7	112	?
3	1/1/2008	13	115	542
4	1/1/2008	19	?	574
5	1/2/2008	1	118	556
6	1/2/2008	7	116	546
7	1/2/2008	13	?	546
8	1/2/2008	19	116	546
9	1/3/2008	1	118	556
10	1/3/2008	7	?	?

Bảng 2. Tập dữ liệu đầy đủ sau khi xóa các giá trị thiếu

STT	Ngày	Giờ	Mực nước	Lưu lượng
1	1/1/2008	1	130	612
2	1/1/2008	13	115	542
3	1/2/2008	1	118	556
4	1/2/2008	7	116	546
5	1/2/2008	19	116	546
6	1/3/2008	1	118	556

Bảng 3. Kết quả sử dụng phương pháp Pairwise

(a) Các biến được sử dụng phân tích				(b) Kết quả phương pháp Pairwise			
STT	Ngày	Giờ	Mực nước	STT	Ngày	Giờ	Mực nước
1	1/1/2008	1	130	1	1/1/2008	1	130
2	1/1/2008	7	112	2	1/1/2008	7	112
3	1/1/2008	13	115	3	1/1/2008	13	115
4	1/1/2008	19	?	4	1/2/2008	1	118
5	1/2/2008	1	118	5	1/2/2008	7	116
6	/2/2008	7	116	6	1/2/2008	19	116
7	1/2/2008	13	?	7	1//2008	1	118
8	1/2/2008	19	116				
9	1/32008	1	118				
10	1/3/2008	7	?				

2.1.2. Pairwise method (Available-Case Analysis, ACA)

Phương pháp này chỉ loại bỏ những trường hợp có các giá trị bị thiếu trong số các biến được phân tích (Gelman & Hill, 2006). Thông thường, phương pháp này loại bỏ ít trường hợp hơn phương pháp xóa Listwise. ACA vẫn có một nhược điểm tương tự như Listwise, đặc biệt là các kết quả sai lệch, được chỉ ra trong nghiên cứu của Ghosh và Pahwa.

Bảng 3 minh họa phương pháp Pairwisse. Trong ví dụ này, các biến được sử dụng để phân tích là Ngày, giờ và mực nước. Bảng 3a. là tập dữ liệu con của bộ dữ liệu thiếu ban đầu chỉ có ba biến này. Khi áp dụng phương pháp Pairwise, các hàng có giá trị bị thiếu sẽ bị xóa, điều này tạo ra tập dữ liệu được thể hiển thị ở bảng 3. Trong trường hợp này, 7 trong số 10 trường hợp vẫn còn, làm giảm kích thước của tập dữ liệu. Khi sử dụng phương pháp này, kích thước của tập dữ liệu cuối cùng sẽ khác nhau vì nó phụ thuộc vào các biến sử dụng trong mỗi phân tích.

2.2. Ước lượng các giá trị thiếu

Khác với cách tiếp cận trên là loại bỏ các bản ghi chứa dữ liệu thiếu, cách tiếp cận sẽ tìm cách thay thế các giá trị thiếu bằng các giá trị ước lượng sử dụng phương pháp khác nhau. Trong phần này chúng tôi trình bày một số phương pháp phổ biến và cập nhật:

2.2.1. Thay thế dữ liệu bằng giá trị trung bình/trung vị (Mean/median substitution)

Allison (2001) và Bishop (2006) đề xuất phương pháp thay thế giá trị trung bình hoặc trung vị của các giá trị quan sát của biến cho mỗi giá trị còn thiếu. Các thuật toán sử dụng cùng một giá trị (trung bình hoặc trung vị) thay thế tất cả các giá trị bị thiếu dẫn đến kết quả sai lệch và lỗi về độ lệch chuẩn (undervalue

standard derivation) (Crawford & cs., 1995; Sterne & cs., 2009).

2.2.2. Phương pháp sử dụng giá trị quan sát cuối cùng (Last Value Carried Forward, LVCF)

Phương pháp này dùng giá trị quan sát cuối cùng để điền vào các giá trị còn thiếu. Cách tiếp cận này có thể dẫn đến kết quả sai và dưới hoặc quá mức của các giá trị thực. Về mặt lý thuyết, phương pháp này giả định rằng kết quả sẽ không thay đổi sau giá trị quan sát cuối cùng.

2.2.3 Thay thế giá trị bằng phương pháp nội suy (Interpolation)

Phương pháp nội suy sẽ tạo ra các điểm dữ liệu mới từ một tập hợp các điểm dữ liệu đã biết. Đây là phương pháp cho kết quả khá tốt khi dữ liệu thiếu từng điểm.

Bảng 4. Kết quả thay thế bằng giá trị trung bình hoặc trung vị

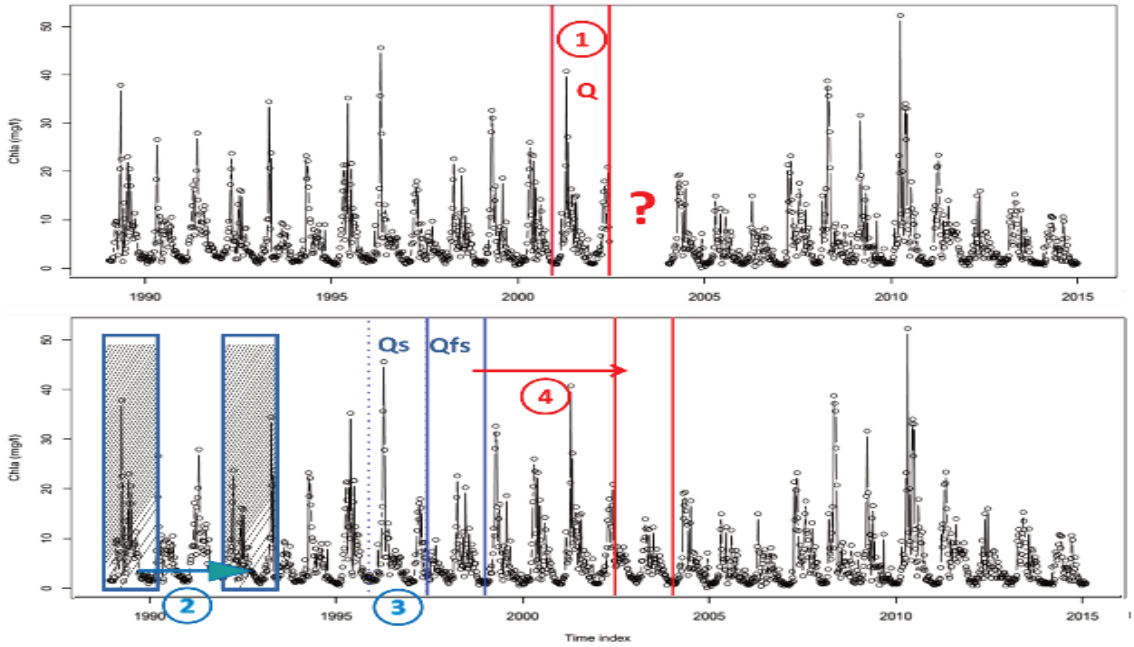
STT	Ngày	Giờ	Mực nước	Lưu lượng
1	1/1/2008	1	130	612
2	1/1/2008	7	112	560/551
3	1/1/2008	13	115	542
4	1/1/2008	19	118/116	574
5	1/2/2008	1	118	556
6	1/2/2008	7	116	546
7	1/2/2008	13	118/116	546
8	1/2/2008	19	116	546
9	1/3/2008	1	118	556
10	1/3/2008	7	118/116	560/551

Bảng 5. Kết quả điền đầy giá trị thiếu sử dụng phương pháp LCVF

STT	Ngày	Giờ	Mực nước	Lưu lượng
1	1/1/2008	1	130	612
2	1/1/2008	7	112	612
3	1/1/2008	13	115	542
4	1/1/2008	19	115	574
5	1/2/2008	1	118	556
6	1/2/2008	7	116	546
7	1/2/2008	13	116	546
8	1/2/2008	19	116	546
9	1/3/2008	1	118	556
10	1/3/2008	7	118	556

Bảng 6. Kết quả điền đầy giá trị thiếu sử dụng phương pháp nội suy

STT	Ngày	Giờ	Mực nước	Lưu lượng
1	1/1/2008	1	130	612
2	1/1/2008	7	112	577
3	1/1/2008	13	115	542
4	1/1/2008	19	116.5	574
5	1/2/2008	1	118	556



Hình 1. (1) Xây dựng cửa sổ Q trước dữ liệu thiếu;
(2) Dịch chuyển từng cửa sổ để tìm các cửa sổ tương tự với cửa sổ Q;
(3) Chọn cửa sổ tương tự nhất Qs với cửa sổ Q;
(4) Thay thế giá trị thiếu bằng giá trị cửa sổ Qfs

2.2.4. Các phương pháp ước lượng giá trị thiếu trực tiếp dựa vào dữ liệu có sẵn

- Phương pháp DTWBI (Phan & cs., 2017)

Phương pháp này cho phép điền đầy khoảng dữ liệu thiếu lớn của dữ liệu chuỗi thời gian đơn biến. Hình 1 mô tả các bước thực hiện ước lượng giá trị thiếu của thuật toán DTWBI. Phương pháp này thay thế khoảng giá trị thiếu bằng cách tìm chuỗi con tương tự nhất (Qs, ③-Hình 1) với chuỗi con trước (hoặc sau) các giá trị bị thiếu (Q - ①-Hình 1), sau đó lấp đầy khoảng dữ liệu trống bằng cách sao chép chuỗi con ngay sau (tương ứng ngay trước) chuỗi con tương tự tiếp (Qfs - ④-Hình 1). Để tìm ra các chuỗi con tương tự với cửa sổ Q, từng cửa sổ (cùng kích thước với với cửa sổ Q) được dịch chuyển trên chuỗi dữ liệu (②-Hình 1) để tìm ra các chuỗi tương tự với Q dựa trên độ tương tự toàn cục (Phan & cs., 2016) và độ tương tự cục bộ DTW (Sakoe và Chiba, 1978). Sau đó, Qs, chuỗi có độ khác biệt ít nhất được chọn ra từ tập các chuỗi tương tự vừa tìm được.

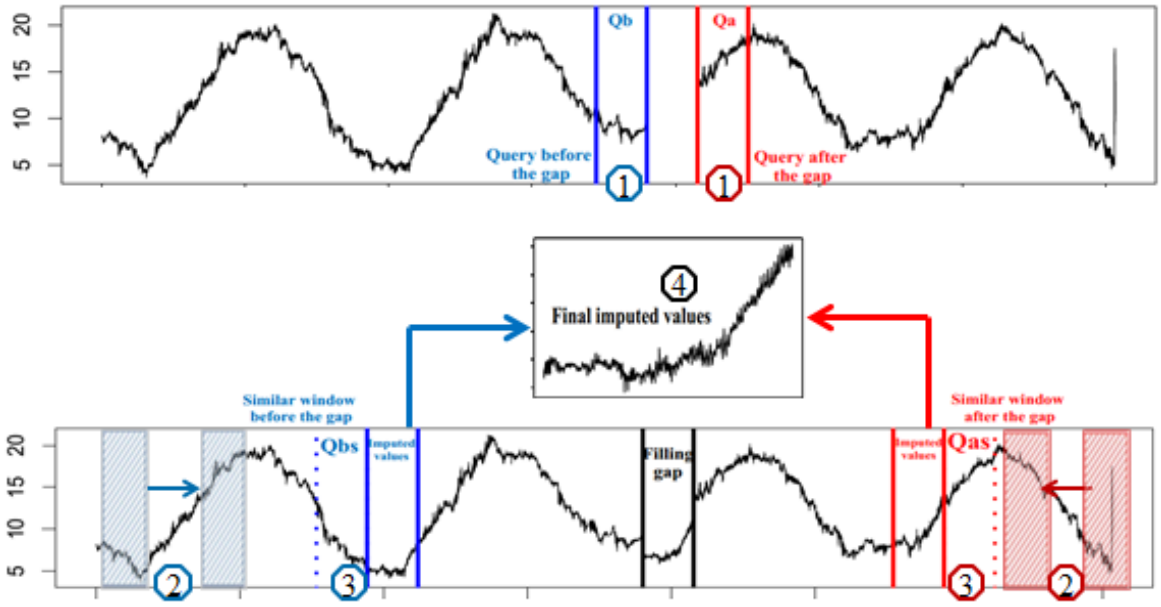
- Phương pháp eDTWBI (Phan & cs., 2020)

eDTWBI là phương pháp mở rộng của phương pháp DTWBI. Hình 2 mô tả các bước ước lượng giá trị thiếu trên chuỗi dữ liệu thời gian đơn biến. Ở phương pháp này, với mỗi khoảng trống dữ liệu, dữ liệu trước và dữ liệu sau khoảng trống này được xem xét như hai chuỗi dữ liệu thời gian riêng biệt. Từ đó phương pháp DTWBI được áp dụng trên từng chuỗi dữ liệu đơn lẻ để ước tính vector giá trị thiếu. Kết quả cuối cùng để điền đầy khoảng trống dữ liệu là giá trị trung bình của 2 vector ước tính trước đó.

3. THỰC NGHIỆM

3.1. Miêu tả dữ liệu

Chúng tôi phân tích 3 bộ dữ liệu để đánh giá hiệu suất phương pháp điền đầy giá trị thiếu. Trong đó có bộ dữ liệu Khách hàng hàng không (Airpassenger) đến từ gói R-TSA (Chan & Ripley, 2018). Bộ dữ liệu này được chọn vì chúng thường được sử dụng trong các tài liệu nghiên cứu. Ngoài ra, chúng tôi cũng chọn thêm hai bộ dữ liệu khác đến từ lĩnh vực khác nhau ở những địa điểm khác nhau bao gồm:



Hình 2. (1) Xây dựng cửa sổ Qb, Qa trước và sau dữ liệu thiếu;
(2) Dịch chuyển từng cửa sổ để tìm các cửa sổ tương tự
với cửa sổ Q trên dữ liệu trước và sau dữ liệu thiếu;
(3) Chọn cửa sổ tương tự nhất Qbs và Qas với cửa sổ Q;
(4) Thay thế giá trị thiếu bằng giá trị trung bình cửa sổ trước Qbs và sau cửa sổ Qas

- Khách hàng hàng không (Airpassenger): Số khách hàng trung bình đi máy bay hàng tháng. Dữ liệu được thu thập từ tháng 1/1960 đến tháng 12/1971.

- Nhiệt độ không khí Phù Liễn (Temperature): Bộ dữ liệu này bao gồm nhiệt độ không khí trung bình hàng tháng tại trạm khí tượng Phù Liễn tại Việt Nam từ 1/1961 đến 12/2014.

- Mực nước tại trạm Hưng Yên (Water level): Bộ dữ liệu mực nước hàng giờ được thu thập tại trạm Hưng Yên từ 1/1/2008 đến 30/04/2008 (2904 bản ghi).

3.2. Các bước tiến hành thực nghiệm

Trên thực tế, việc đánh giá hiệu suất của các phương pháp điền đầy dữ liệu không thể thực hiện được do các giá trị thực bị thiếu. Vì vậy, chúng ta phải tạo dữ liệu thiếu giả lập trên chuỗi thời gian đầy đủ để so sánh khả năng của các phương pháp ước lượng giá trị thiếu. Trong nghiên cứu này, một kỹ thuật gồm ba bước được sử dụng để tiến hành các thí nghiệm được mô tả chi tiết như sau:

Bước 1: Dữ liệu thiếu giả lập được tạo ra bằng cách xóa các phân đoạn (gồm các giá trị liên tiếp) khỏi mỗi chuỗi thời gian với kích thước khác nhau.

Bước 2: Sử dụng các thuật toán điền đầy để ước tính các giá trị thiếu.

Bước 3: Đánh giá hiệu quả của các phương pháp điền đầy giá trị thiếu.

Ở đây, chúng tôi thực hiện tạo 5 mức dữ liệu thiếu trên 3 bộ dữ liệu. Đối với bộ dữ liệu khách hàng hàng không, và nhiệt độ không khí Phù Liễn, kích thước dữ liệu thiếu lần lượt là 6%, 7,5%, 10%, 12,5% và 15% chiều dài bộ dữ liệu. Đối với bộ dữ liệu mực nước Hưng Yên, đây là một tập dữ liệu khá lớn, do đó, các khoảng trống được tạo ra với kích thước 3%, 3,75%, 5%, 6,25% và 7,5% chiều dài bộ dữ liệu.

3.3. Các chỉ số đánh giá hiệu suất

Sau khi thực hiện điền đầy các giá trị thiếu, chúng tôi đánh giá hiệu suất của phương pháp của dựa trên bốn chỉ số khác nhau được mô tả như sau:

- Độ tương tự (Similarity) - Sim (y, x) cho biết độ tương tự nhau giữa giá trị thực (*x*) và giá trị ước lượng (*y*) được tính bởi công thức sau:

$$\mathrm{Sim}(y,x)=\frac{1}{T}\sum_{i=1}^{T}\frac{1}{1+\dfrac{|y_i-x_i|}{\max(x)-\min(x)}}$$

Trong đó, T là kích thước dữ liệu thiếu, độ tương tự nằm trong [0,1]. Độ tương tự cao hơn cho thấy phương pháp điền đầy dữ liệu thiếu có khả năng ước lượng giá trị thiếu tốt hơn.

- NMAE (Normalized Mean Absolute Error): Sai số tuyệt đối trung bình chuẩn hóa giữa giá trị thực (*x*) và giá trị ước lượng (*y*) được tính như sau:

$$\mathrm{NMAE}(y,x)=\frac{1}{T}\sum_{i=1}^{T}\frac{|y_i-x_i|}{V_{max}-V_{min}}$$

Trong đó, V_{max}, V_{min} là giá trị max và min của chuỗi thời gian ban đầu. Kết quả NMAE nhỏ hơn cho thấy phương pháp điền đầy dữ liệu thiếu cho kết quả sát với giá trị thực hơn.

- RMSE (Root Mean Square Error): Lỗi trung bình bình phương giữa giá trị thực (x) và giá trị ước lượng (y) được định nghĩa như sau:

$$\mathrm{RMSE}(y,x)=\sqrt{\frac{1}{T}\sum_{i=1}^{T}(y_i-x_i)^2}$$

Chỉ số này rất hữu ích để đo độ chính xác tổng thể của phương pháp ước tính dữ liệu thiếu. Phương pháp hiệu quả hơn khi giá trị RMSE thấp hơn.

- FSD (Fractional Standard Deviation): Tỷ lệ lệch chuẩn nhau giữa giá trị thực (*x*) và giá trị ước lượng (*y*) được tính bởi công thức:

$$\mathrm{FSD}(y,x)=2*\frac{|\mathrm{SD}(y)-\mathrm{SD}(x)|}{\mathrm{SD}(y)+\mathrm{SD}(x)}$$

Tỷ lệ này cho biết liệu một phương pháp xử lý dữ liệu thiếu có được chấp nhận hay không? Giá trị của FSD càng gần 0 thì các giá trị ước lượng càng gần với giá trị thực.

4. KẾT QUẢ VÀ THẢO LUẬN

Chúng tôi tiến hành so sánh hiệu năng của các phương pháp nội suy (na.interp, Hyndman & Khandakar, 2008), phương pháp sử dụng giá trị quan sát cuối (na.locf, Zeileis & Grothendieck, 2018), phương pháp thay thế bởi giá trị trung bình (na.aggregate, Zeileis & Grothendieck, 2018), phương pháp DTWBI (Phan & cs., 2017), và eDTWBI (Phan & cs., 2020). Bảng 7 trình bày kết quả trung bình của các phương pháp điền đầy giá trị thiếu ở trên áp dụng trên 3 bộ dữ liệu sử dụng 4 tiêu chí để đánh giá kết quả: độ tương tự (Sim), NMAE, RMSE, FSD. Các kết quả tốt nhất cho mỗi tỷ lệ thiếu dữ liệu được in đậm. Những kết quả này cho thấy eDTWBI có khả năng ước lượng dữ liệu thiếu tốt hơn những phương pháp điền đầy dữ liệu thiếu trong bài báo này.

Hai bộ dữ liệu nhiệt độ Phù Liễn và mực nước Hưng Yên có đặc điểm là chỉ có thành phần mùa vụ mà không có thành phần xu hướng. Trên hai bộ dữ liệu này, chúng ta thấy rằng eDTWBI cho giá trị lớn nhất về độ tương tự (Sim), giá trị nhỏ nhất ở mức độ sai số (NMAE và RMSE) ở hầu hết các mức dữ liệu thiếu. Điều này cho thấy giá trị ước lượng dữ liệu thiếu sinh bởi phương pháp eDTWBI là gần với giá trị thực. FSD là chỉ số so sánh hình dáng của dữ liệu dự đoán và dữ liệu thực. Ở chỉ số FSD, so với các chỉ số so sánh định lượng thì eDTWBI không còn cho kết quả tốt như những chỉ số như Sim, NMAE và RMSE, nó chỉ cho kết quả tốt ở một số mức dữ liệu thiếu trên bộ dữ liệu mực nước Hưng Yên (3%, 3,75% và 6,25%). Ở các mức dữ liệu thiếu còn lại trên bộ dữ liệu mực nước Hưng Yên và trên toàn bộ các khoảng dữ liệu của bộ dữ liệu nhiệt độ Phù Liễn, phương pháp eDTWBI đứng sau DTWBI.

Bộ dữ liệu khách hàng hàng không vừa có tính chất mùa vụ, vừa có xu hướng tăng dần. Hai phương pháp DTWBI và eDTWBI hoạt động tốt với giả thuyết tồn tại "mẫu" (pattern) ở vị trí nào đó trên dữ liệu, nên hai phương pháp này chỉ cho kết tốt hơn các phương pháp khác ở những mức dữ liệu thiếu nhỏ trên bộ dữ liệu này. Ở những khoảng dữ liệu thiếu lớn hơn, na.interp là phương pháp nội suy kết hợp với xử lý tính chất mùa vụ của dữ liệu, cho kết quả tốt hơn ở các chỉ số Sim, NMAE, và RMSE. Mặc dù vậy, ở chỉ số so sánh hình dáng của dữ liệu dự đoán, DTWBI vẫn chứng tỏ được thế mạnh của mình khi kết quả chỉ số FSD có giá trị nhỏ nhất tại 4/5 mức dữ liệu thiếu.

Bảng 7. Kết quả so sánh các phương pháp điền đầy dữ liệu thiếu trên 3 bộ dữ liệu

Phương pháp	Kích thước	Khách hàng hàng không				Nhiệt độ Phù Liễn				Kích thước	Mực nước Hưng Yên			
		Sim	NMAE	RMSE	FSD	Sim	NMAE	RMSE	FSD		Sim	NMAE	RMSE	FSD
DTWBI	6%	0,73	0,07	45,39	0,26	0,88	0,11	2,43	0,04	3%	0,78	0,13	27,89	0,30
eDTWBI		0,81	0,04	28,01	0,11	0,93	0,06	1,35	0,07		0,83	0,10	20,12	0,30
na,interp		0,75	0,06	34,17	0,86	0,79	0,22	4,94	1,30		0,80	0,12	26,03	0,54
na,locf		0,75	0,06	38,38	2	0,78	0,24	5,28	2		0,79	0,12	25,93	2
na,aggregate		0,56	0,14	75,70	2	0,79	0,21	4,27	2		0,79	0,13	26,17	2
DTWBI	7,50%	0,81	0,06	37,13	0,10	0,89	0,11	2,44	0,06	3,75%	0,81	0,15	27,89	0,13
eDTWBI		0,85	0,04	21,97	0,21	0,89	0,10	2,22	0,04		0,81	0,14	27,70	0,11
na,interp		0,78	0,07	41,69	1,33	0,79	0,25	5,41	1,19		0,82	0,13	27,49	0,87
na,locf		0,80	0,06	40,33	2	0,79	0,25	5,42	2		0,78	0,17	34,24	2
na,aggregate		0,64	0,13	77,82	2	0,79	0,22	4,49	2		0,77	0,19	37,78	2
DTWBI	10%	0,73	0,11	67,03	0,12	0,90	0,10	2,21	0,02	5%	0,84	0,13	27,26	0,14
eDTWBI		0,80	0,07	45,00	0,41	0,92	0,07	1,72	0,04		0,85	0,11	24,52	0,61
na,interp		0,81	0,07	42,26	1,01	0,79	0,24	4,96	0,91		0,84	0,12	25,83	0,73
na,locf		0,78	0,08	51,19	2	0,79	0,25	5,71	2		0,80	0,17	36,23	2
na,aggregate		0,71	0,12	70,35	2	0,80	0,22	4,48	2		0,83	0,14	29,11	2
DTWBI	12,5%	0,69	0,17	105,81	0,30	0,88	0,11	2,61	0,07	6,25%	0,83	0,14	29,08	0,22
eDTWBI		0,81	0,10	64,38	0,42	0,90	0,09	2,08	0,09		0,85	0,12	24,91	0,22
na,interp		0,81	0,09	61,30	1,59	0,79	0,25	5,52	1,03		0,80	0,17	35,92	0,99
na,locf		0,82	0,09	60,18	2	0,75	0,31	6,71	2		0,76	0,23	47,90	2
na,aggregate		0,76	0,13	79,64	2	0,79	0,22	4,46	2		0,83	0,15	31,35	2
DTWBI	15%	0,74	0,14	80,65	0,28	0,89	0,11	2,53	0,06	7,5%	0,87	0,11	23,98	0,14
eDTWBI		0,77	0,13	72,32	0,25	0,91	0,08	1,95	0,10		0,89	0,10	20,28	0,18
na,interp		0,83	0,09	62,29	1,17	0,78	0,26	5,87	1,38		0,84	0,15	31,57	1,44
na,locf		0,80	0,11	76,09	2	0,79	0,26	5,97	2		0,82	0,18	36,76	2
na,aggregate		0,70	0,19	114,63	2	0,80	0,22	4,38	2		0,84	0,15	30,99	2

5. KẾT LUẬN

Trong bài viết này, chúng tôi đã trình bày các hướng tiếp cận xử lý dữ liệu thiếu cho dữ liệu chuỗi thời gian một chiều bao gồm hai nhóm phương pháp: i) Nhóm phương pháp bỏ qua dữ liệu thiếu và ii) Nhóm phương pháp ước lượng giá trị thiếu. Kết quả thực nghiệm trên 3 bộ dữ liệu thực tế cho thấy, phương pháp eDTWBI và DTWBI cho kết quả ước lượng khá chính xác trong trường hợp dữ liệu có tính chất mùa vụ nhưng không có xu hướng. Phương pháp na.interp cho kết quả dự báo tốt hơn trong trường hợp dữ liệu vừa có tính chất mùa vụ và có xu hướng. Bước tiếp theo chúng tôi dự định sẽ tiếp tục mở rộng nghiên cứu này cho dữ liệu chuỗi thời gian nhiều chiều.

TÀI LIỆU THAM KHẢO

Allison P.D. (2001). Missing Data, Quantitative Applications in the Social Sciences, 136. Sage Publication.

Buuren S. & Groothuis-Oudshoorn K. (2011). Mice: Multivariate imputation by chained equations in R. Journal of statistical software. 45(3).

Bishop C.M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Chan K.S. & Ripley B. (2020). TSA: Time Series Analysis. R package version 1.3. Retrieved from https://CRAN.R-project.org/package=TSA, on March 10, 2020.

Crawford S.L., Tennstedt S.L. & McKinlay J.B. (1995). A comparison of anlaytic methods for non-random missingness of outcome data. J. Clin. Epidemiol. 48(2): 209-219.

Dong Y. & Peng J. (2013). Principled missing data methods for researchers. SpringerPlus. 2: 222.

Gelman A. & Hill J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press.

Ghosh S. & Pahwa P. (2008). Assessing bias associated with missing data from joint Canada/U.S. survey of health: An application, JSM Biometrics.

Horton N.J. & Kleinman K.P. (2007). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. American Statistical Association. 61. 79-90.

Hyndman R. & Khandakar Y. (2008). Automatic time series forecasting: the forecast package for R., used package in 2020. J. Stat. Softw. pp. 1-22.

Little R.J.A. & Rubin D.B. (2014). Statistical Analysis with Missing Data. John Wiley & Sons. Google-Books-ID: AyVeBAAAQBAJ.

Moritz S., Sardá A., Bartz-Beielstein T., Zaefferer M. & Stork J. (2015). Comparison of different Methods for Univariate Time Series Imputation in R. arXivpreprint arXiv:1510.03924.

Molenberghs G., Fitzmaurice G., Kenward M.G., Verbeke G. & Tsiatis A. (2014). Handbook of missing data methodology. CRC Press.

Phan T.T.H., Caillault E.P. & Bigand A. (2016). Comparative study on supervised learning methods for identifying phytoplankton species, in 2016 IEEE Sixth International Conference on Communications and Electronics (ICCE). pp. 283-288, doi: 10.1109/CCE.2016.7562650.

Phan T.T.H., Poisson Caillault E., Lefebvre A. & Bigand A. (2017). Dynamic Time Warping-based imputation for univariate time series data, Pattern Recognition Letters.

Rousseeuw K., Caillault ÉP., Lefebvre A. & Hamad D. (2013). Monitoring system of phytoplankton blooms by using unsupervised classifier and time modeling. In 2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS. pp. 3962-3965.

Stekhoven D.J. & Bühlmann P. (2012). MissForest-non-parametric missing value imputation for mixed-type data. Bioinformatics. 28(1): 112-118.

Sterne J.A.C., White I.R., Carlin J.B., Spratt M., Royston P., Kenward M.G., Wood A.M. & Carpenter J.R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ (Clin. Resear. ed.).

Sakoe H. & Chiba S. (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Transactions On Acoustics, Speech, And Signal Processing. 16: 43-49.

Zeileis A. & Gabor Grothendieck (2005). zoo: S3 infrastructure for regular and irregular time series. Journal of Statistical Software. 14(6): 1-27.