Tạp chí Khoa học Kỹ thuật Mỏ - Địa chất

Trang điện tử: http://tapchi.humg.edu.vn

Phát hiện và xử lý ngoại lai cho dữ liệu nhiệt độ tại các trạm quan trắc 3h của Việt Nam

Đặng Văn Nam [1,*], Nông Thị Oanh [1], Nguyễn Xuân Hoài [2] , Ngô Văn Mạnh [3], Nguyễn Thị Hiền [4]

[1] *Khoa Công nghệ Thông tin, Trường Đại học Mỏ - Địa chất, Việt Nam*
[2] *Viện Trí tuệ Nhân tạo, Việt Nam*
[3] *Trung tâm Thông tin và Dữ liệu khí tượng thủy văn, Việt Nam*
[4] *Học viện Kỹ thuật quân sự, , Việt Nam*

THÔNG TIN BÀI BÁO

Quá trình:
Nhận bài 15/11/2019
Chấp nhận 06/01/2020
Đăng online 28/02/2020

Từ khóa:
Dữ liệu ngoại lai,
Outliers,
Anomalies,
Z - Score,
Box - plot.

TÓM TẮT

Trong bất kỳ một dự án khoa học dữ liệu nào thì chuẩn bị dữ liệu (Data preparation) là công đoạn bắt buộc và không thể thiếu. Kết quả của nhiều nghiên cứu đã chỉ ra rằng, chuẩn bị dữ liệu là công đoạn chiếm tới 80% thời gian, công sức và nguồn lực của một dự án khoa học dữ liệu. Chuẩn bị dữ liệu bao gồm rất nhiều bước xử lý, với nhiều nghiệp vụ khác nhau và phụ thuộc vào từng bài toán, từng loại dữ liệu cụ thể. Phát hiện và xử lý dữ liệu ngoại lai (Outliers) là một trong những bước tiền xử lý quan trọng, đặc biệt là các dữ liệu số dạng chuỗi thời gian (Time series) (Hermine N.Akouemo and et al., 2014). Trong nội dung của bài báo này, tác giả sẽ nghiên cứu hai phương pháp hiệu quả đang được sử dụng để phát hiện ngoại lai cho dữ liệu có số chiều thấp là Z - Score và biểu đồ Box - plot, cũng như các phương pháp để xử lý dữ liệu ngoại lai nói chung. Sau đó tiến hành thực nghiệm, áp dụng những phương pháp phát hiện và xử lý này cho dữ liệu nhiệt độ thu thập được từ 43 trạm quan trắc 3 giờ của Việt Nam trong giai đoạn 6 năm gần đây từ năm 2014 đến năm 2019.

© 2020 Trường Đại học Mỏ - Địa chất. Tất cả các quyền được bảo đảm.

1. Mở đầu

Dữ liệu khí tượng thủy văn trong đó có dữ liệu nhiệt độ, được thu thập, xử lý và lưu trữ tại cơ sở dữ liệu của Trung tâm thông tin và Dữ liệu khí tượng thủy văn. Đây là dữ liệu dạng chuỗi thời gian được thu thập định kỳ theo từng khoảng thời gian nhất định (3 giờ hoặc 6 giờ) tùy thuộc vào từng trạm cụ thể. Hiện tại ở Việt Nam, có 43 trạm quan trắc dữ liệu với tần suất 3 giờ một lần, chi tiết các trạm quan trắc và dữ liệu được trình bày cụ thể trong phần 2 của bài báo. Quá trình đo đạc, xử lý, tổng hợp, truyền và lưu trữ dữ liệu quan trắc từ các trạm bị ảnh hưởng bởi các yếu tố chủ quan và khách quan dẫn đến mất mát dữ liệu và/hoặc tác động đến độ chính xác của dữ liệu. Do đó yêu cầu bắt buộc là dữ liệu cần phải được chuẩn bị (Data

*Tác giả liên hệ
E - mail: dangvannam@humg. edu. vn

preparation) trước khi sử dụng cho bất kỳ mục đích gì.

Trong (Davy Cielen and et al., 2016) đã chỉ ra rằng chuẩn bị dữ liệu được đánh giá 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 chuẩn bị dữ liệu. 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 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).

Khi nghiên cứu và làm việc với dữ liệu khí tượng thủy văn nói chung, dữ liệu nhiệt độ nói riêng tác giả thấy rằng việc chuẩn bị dữ liệu cho dữ liệu nhiệt độ tập trung chủ yếu vào 4 vấn đề chính dưới đây:

1) Kết hợp và sắp xếp dữ liệu quan trắc theo chuỗi thời gian và theo vị trí địa lý của các trạm.

2) Phát hiện và xử lý các dữ liệu ngoại lai (Outliers) trong tập dữ liệu quan trắc.

3) Phát hiện và xử lý các dữ liệu thiếu (Missing data) trong tập dữ liệu quan trắc.

4) Chuyển đổi, định dạng và xuất dữ liệu đã xử lý để lưu trữ theo yêu cầu.

Các điểm dữ liệu ngoại lai hay còn được gọi là các dữ liệu bất thường (Anomalies) có ảnh hưởng lớn đến độ chính xác của các mô hình dự đoán. Phát hiện và xử lý ngoại lai là thao tác quan trọng trong quá trình làm sạch dữ liệu. Việc phát hiện ngoại lai giúp phát hiện ra những điểm dữ liệu không phù hợp hay bất thường hơn so với phần còn lại của tập dữ liệu (C.Aggarwal, 2017).

Phát hiện ngoại lai không chỉ được ứng dụng trong việc làm sạch dữ liệu mà nó còn được ứng dụng vào nhiều bài toán thực tế như: Phát hiện lỗi (fraud detection); Giám sát (surveillance); Chuẩn đoán (diagnosis), Dự đoán bảo trì (predictive maintaence),… Tuy nhiên việc phát hiện các điểm dữ liệu ngoại lai không phải là một công việc đơn giản, nó yêu cầu phải có những hiểu biết sâu sắc về tập dữ liệu, cũng như nắm vững các phương pháp hiệu quả để thực hiện việc này.

Trong nội dung của bài báo, nhóm tác giả chỉ tập trung vào giải quyết một trong số bốn vấn đề chính đã chỉ ra ở trên, đó là nhiên cứu các phương pháp phát hiện và xử lý ngoại lai, trên cơ sở đó áp dụng các phương pháp này vào việc xử lý ngoại lai cho dữ liệu nhiệt độ thu thập được tại toàn bộ 43 trạm quan trắc 3 giờ của Việt Nam trong khoảng thời gian 6 năm gần đây, từ ngày 1 tháng 1 năm 2014 tới hết ngày 31 tháng 12 năm 2019. Toàn bộ dữ liệu sử dụng trong bài báo đều là dữ liệu thực tế được cung cấp bởi Trung tâm thông tin và dữ liệu khí tượng thủy văn.

2. Dữ liệu nhiệt độ tại các trạm quan trắc 3h

Trạm quan trắc khí tượng thủy văn được lắp đặt tại các vị trí khác nhau để thực hiện đo các thông số khí tượng như: Nhiệt độ, tốc độ gió, hướng gió, lượng mưa, độ ẩm,… các trạm này định kỳ sau một khoảng thời gian cố định được thiết lập thực hiện việc đo các thông số này. Với các trạm quan trắc 3h sẽ thực hiện thu thập dữ liệu 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. Ở nước ta hiện nay có tổng số 43 trạm quan trắc khí tượng thủy văn với thời gian quan trắc là 3h một lần. Danh sách 43 trạm này được cho trong Bảng 1 dưới đây, vị trí đặt các trạm được thể hiện trong Hình 2.

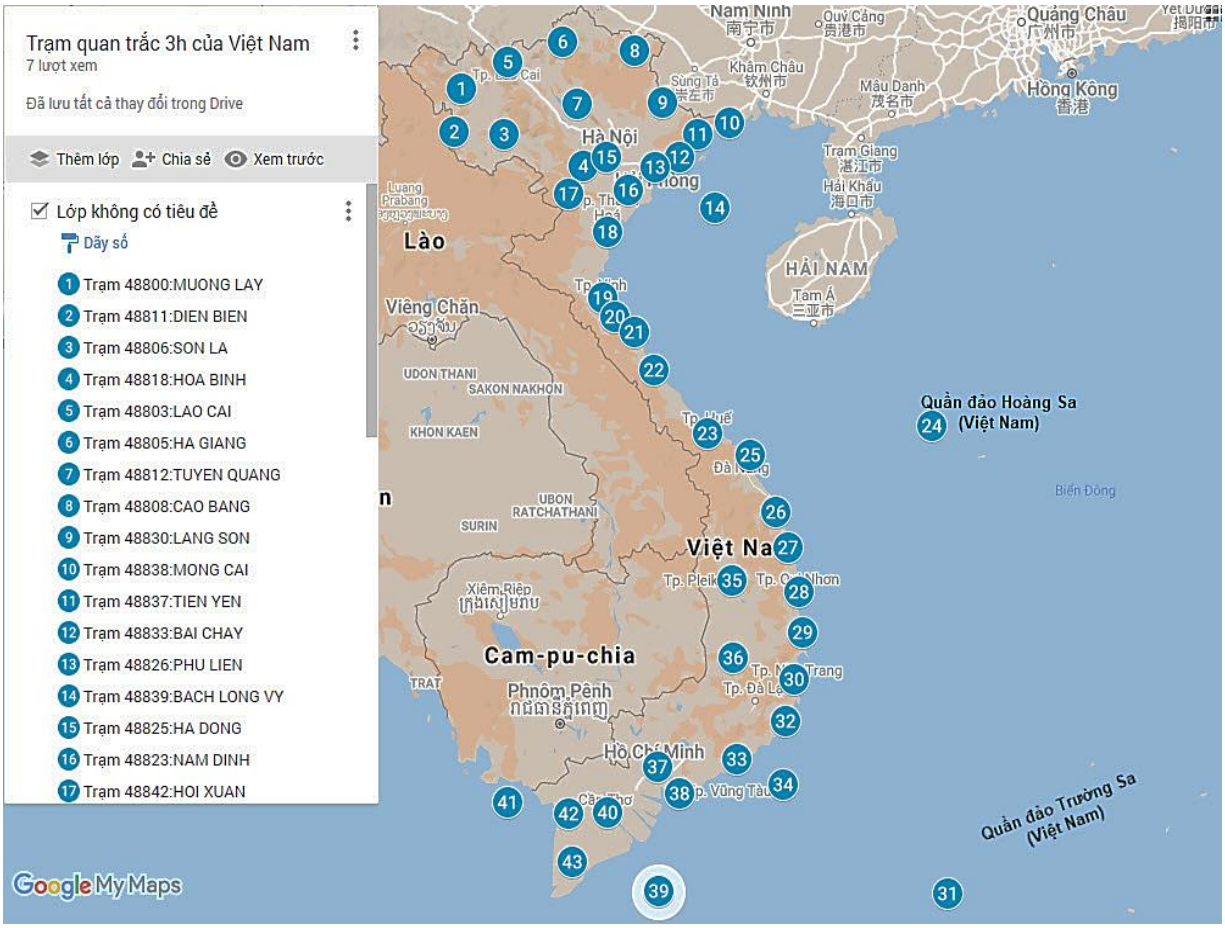
Dữ liệu khí tượng thủy văn nói chung, dữ liệu nhiệt độ nói riêng, sau khi được đo đạc từ các trạm quan trắc sẽ được gửi về Trung tâm thông tin và dữ liệu khí tượng thủy văn.

Dữ liệu được lưu trữ trong cơ sở dữ liệu MongoDB, tiến hành kết nối tới máy chủ cơ sở dữ liệu và truy xuất thông số nhiệt độ của 43 trạm trong khoảng thời gian từ 01h ngày 01/01/2014 tới 22h ngày 31/12/2019.

Các dữ liệu nhiệt độ sau đó được lưu trữ ra tệp định dạng .CSV (Comma - separated values) có tên Data_Temp43_ Original.csv (Hình 2) để thuận tiện cho việc xử lý các bước tiếp theo. Cột đầu tiên trong tệp có tên "TimeVN" cho biết thời điểm quan trắc dữ liệu, các cột còn lại (tiêu đề mỗi cột tương ứng với mã trạm quan trắc) là dữ liệu nhiệt độ của từng trạm ứng với mốc thời gian của cột "TimeVN". Đây là tệp dữ liệu gốc (dữ liệu thô - Raw dataset) được tổng hợp khi các trạm gửi về, quá trình thu thập dữ liệu, truyền nhận và lưu trữ có thể do các nguyên nhân chủ quan và khách quan dẫn đến dữ liệu có thể bị mất mát, bị sai lệch,… Do đó trước khi sử dụng các số liệu này cần phải được xử lý.

Bảng 1. Danh sách 43 trạm quan trắc 3h của Việt Nam.

STT	Mã trạm	Tên quốc tế	Tên Việt Nam	Trạm đảo	Tỉnh/Thành phố
1	48800	MUONG LAY	Mường Lay		Điện Biên
2	48811	DIEN BIEN	Điện Biên		Điện Biên
3	48806	SON LA	Sơn La		Sơn La
4	48818	HOA BINH	Hòa Bình		Hòa Bình
5	48803	LAO CAI	Lào Cai		Lào Cai
6	48805	HA GIANG	Hà Giang		Hà Giang
7	48812	TUYEN QUANG	Tuyên Quang		Tuyên Quang
8	48808	CAO BANG	Cao Bằng		Cao Bằng
9	48830	LANG SON	Lạng Sơn		Lạng Sơn
10	48838	MONG CAI	Móng Cái		Quảng Ninh
11	48837	TIEN YEN	Tiên Yên		Quảng Ninh
12	48833	BAI CHAY	Bãi Cháy		Quảng Ninh
13	48826	PHU LIEN	Phù Liễn		Hải Phòng
14	48839	BACH LONG VI	Bạch Long Vĩ	X	Hải Phòng
15	48825	HA DONG	Hà Đông		Hà Nội
16	48823	NAM DINH	Nam Định		Nam Định
17	48842	HOI XUAN	Hồi Xuân		Thanh Hóa
18	48840	THANH HOA	Thanh Hóa		Thanh Hóa
19	48845	VINH	Vinh		Nghệ An
20	48846	HA TINH	Hà Tĩnh		Hà Tĩnh
21	48/86	KY ANH	Kỳ Anh		Hà Tĩnh
22	48848	DONG HOI	Đồng Hới		Quảng Bình
23	48852	HUE	Huế		Thừa Thiên Huế
24	48860	HOANG SA	Hoàng Sa		Đà Nẵng
25	48855	DA NANG	Đà Nẵng		Đà Nẵng
26	48863	QUANG NGAI	Quảng Ngãi		Quảng Ngãi
27	48/96	HOAI NHON	Hoài Nhơn		Bình Định
28	48870	QUY NHON	Quy Nhơn		Bình Định
29	48873	TUY HOA	Tuy Hòa		Phú Yên
30	48877	NHA TRANG	Nha Trang		Khánh Hòa
31	48920	TRUONG SA	Trường Sa	X	Khánh Hòa
32	48890	PHAN RANG	Phan Rang		Ninh Thuận
33	48887	PHAN THIET	Phan Thiết		Bình Thuận
34	48889	PHU QUY	Phú Quý	X	Bình Thuận
35	48866	PLEIKU	Pleiku		Gia Lai
36	48875	BUON MA THUAT	Buôn Ma Thuật		Đắc Lắc
37	48894	NHA BE	Nhà Bè		Hồ Chí Minh
38	48903	VUNG TAU	Vũng Tàu		Bà Rịa - Vũng Tàu
39	48918	CON DAO	Côn Đảo	X	Bà Rịa - Vũng Tàu
40	48910	CAN THO	Cần Thơ		Cần Thơ
41	48917	PHU QUOC	Phú Quốc	X	Kiên Giang
42	48907	RACH GIA	Rạch Giá		Kiên Giang
43	48914	CA MAU	Cà Mau		Cà Mau



Hình 1. Vị trí các trạm quan trắc 3h trên bản đồ Google Maps.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**TimeVN**	**48800**	**48811**	**48806**	**48818**	**48803**	**48805**	**48812**	**48808**	**48830**	**48838**	**48837**	**48833**	**48826**
2	2014-01-01 1:00	15.4	14	7.6	10.2	10.9	10.1	10.4	6.7	5.3	8.8	9.2	10.8	13.8
3	2014-01-01 4:00	15	13.8	8.6	9.4	11.2	9.1	9.2	6	3.8	8.1	8.2	10	13.4
4	2014-01-01 7:00	15.1	13.4	9.4	8.6	11.6	8.6	8.4	5.6	3.3	7	7.2	9.6	13.2
5	2014-01-01 10:00	17.4	15	11.2	16.1	13.6	13.7	10.8	7.3	11.3	17.8	15.5	14.8	16.4
6	2014-01-01 13:00	21.4	19	14.7	22.4	19.4	20.7	20.2	18.6	20.1	20.7	19.7	19.6	20.4
7	2014-01-01 16:00	21.2	21.4	17	22.2	21	21.1	21.1	21.2	20.6	18.2	19.9	18.9	20.4
8	2014-01-01 19:00	17.3	17.5	13	16.7	15.8	16.2	15.9	13.5	12.7	12.2	12.6	16.1	16.1
9	2014-01-01 22:00	15.1	14.1	10.4	12.9	12.5	11.8	13.2	9.6	9.4	9.9	9	13.3	15.4
10	2014-01-02 1:00	14.6	12.3	9.4	11.3	11	10.1	11.4	7.9	7.7	8.6	8.6	11.6	14.6
11	2014-01-02 4:00	13.6	11.8	8.6	10.2	9.6	9.2	10.6	7	6.4	8.4	7.6	10.9	14.3
12	2014-01-02 7:00	13.2	11.8	8	10	9	9.9	9.1	6.4	5.9	7.7	7.6	10.7	13.8
13	2014-01-02 10:00	15.2	15.2	17.3	16.4	12.4	15.2	12.5	8.7	17.4	18.8	16.2	15	18.2
14	2014-01-02 13:00	21.1	20.2	22	22.6	20.4	21.6	20.5	20.9	20.3	20.8	22.6	20.4	21.6
15	2014-01-02 16:00	25.6	24.6	23.4	24.2	23	21.8	22.3	23.8	21.3	19.4	21.4	19	20.8
16	2014-01-02 19:00	19.6	18.6	16.4	17.8									
17	2014-01-02 22:00	17.6	14.3	13.5	15									
18	2014-01-03 1:00	15.1	14.1	11.5	13.7									
19	2014-01-03 4:00	14.7	13.7	10.6	14.5									
20	2014-01-03 7:00	16.2	12.5	9.8	12.6	11	10.3	12.6	10.2	14.7	16.4	15	17.3	18.2
21	2014-01-03 10:00	15.3	14.6	19.4	17.4	13	16.3	16.3	12.9	17.1	20.6	19.8	19.6	20.8
22	2014-01-03 13:00	20.8	17.9	25.2	19.4	19.6	23.6	22.8	21.7	20	22.6	20.4	21.1	23.4
23	2014-01-03 16:00	24.9	24.5	23.4	22.6	21.1	22.3	22.6	22.2	20.4	20.9	22	19.5	20.8
24	2014-01-03 19:00	19.3	19.7	17.4	19.6	17.1	19.6	19	15.3	14.4	17.4	16.2	18	17.7
25	2014-01-03 22:00	16.4	16	15	18.8	16.4	17.2	16.3	13.2	11.8	15.1	14.3	16.5	17.1

Data_Temp43_Original

Ready

100%

Hình 2. Dữ liệu nhiệt độ thu thập được tại 43 trạm quan trắc 3h.

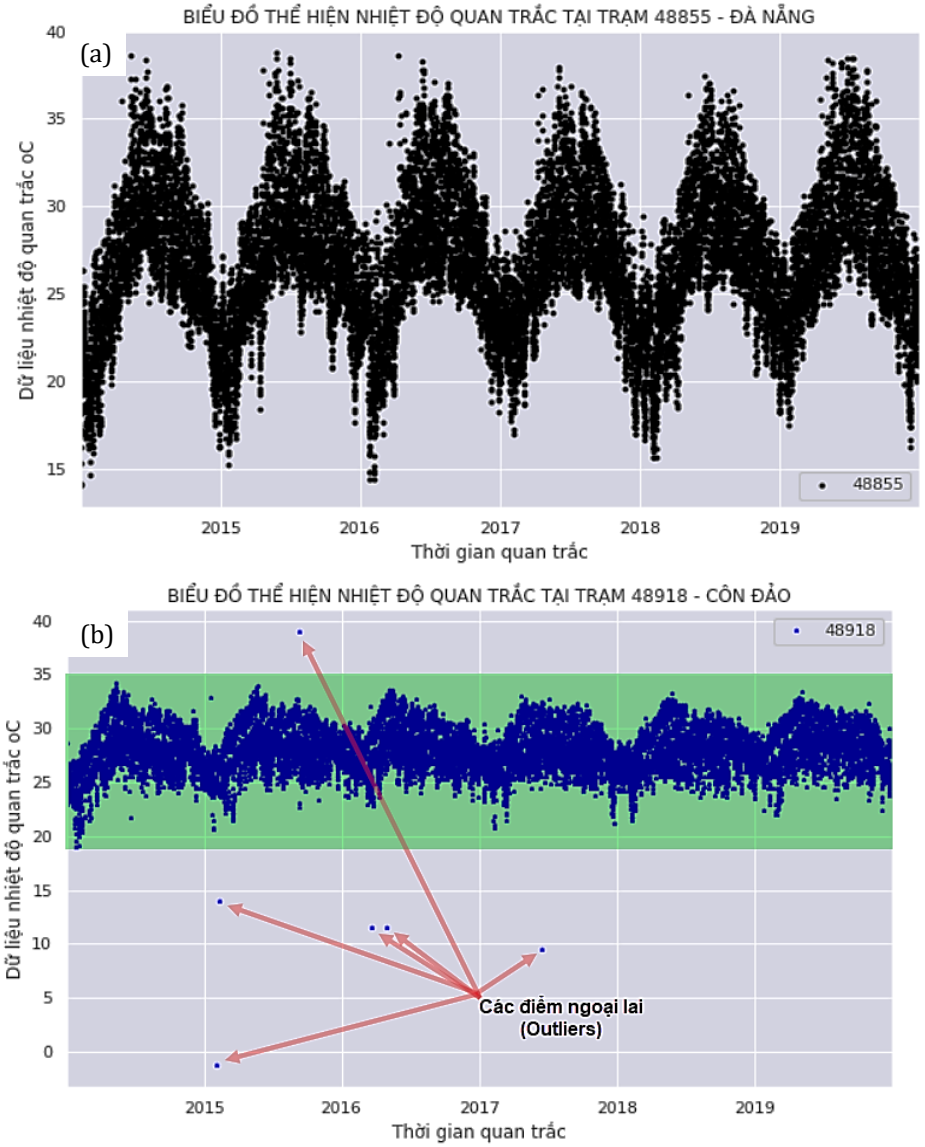
Như đã trình bày trong phần 1, có rất nhiều yêu cầu cần phải thực hiện cho bước chuẩn bị dữ liệu, tuy nhiên trong nội dung của bài báo tác giả chỉ tập trung vào phát hiện và xử lý các ngoại lai cho dữ liệu nhiệt độ tại 43 trạm này. Trong phần 3 dưới đây, sẽ trình bày những nội dung cơ bản về phát hiện và xử lý ngoại lai, trong đó có 2 phương pháp được sử dụng để phát hiện ngoại lai cho dữ liệu có số chiều thấp là Z - Score và Box - plot. Đây cũng là 2 phương pháp mà tác giả sử dụng cho việc phát hiện ngoại lai trong tệp dữ liệu nhiệt độ ở trên.

3. Phát hiện và xử lý ngoại lai

3.1. Giới thiệu về dữ liệu ngoại lai

Một điểm ngoại lai là một điểm dữ liệu khác biệt đáng kể so với phần còn lại của tập dữ liệu (C.Aggarwal, 2017). Ta thường xem các giá trị ngoại lai như là các mẫu dữ liệu đặc biệt, cách xa khỏi phần lớn dữ liệu khác trong tập dữ liệu (N.N.R Ranga Suri and et al., 2018).

Hình 3(a) thể hiện tập dữ liệu nhiệt độ quan trắc được của trạm 48855 - Đà Nẵng, dữ liệu này không chứa giá trị ngoại lai. Hình 3(b) thể hiện dữ liệu nhiệt độ quan trắc của trạm 48918 - Côn Đảo, dữ liệu này có chứa một số giá trị ngoại lai. Các điểm dữ liệu này cách xa khỏi phần lớn các phần tử khác trong tập dữ liệu đã được chỉ ra cụ thể trong hình.



Hình 3. (a) Minh họa tập dữ liệu không chứa dữ liệu ngoại lai; (b) Minh họa tập dữ liệu chứa các điểm dữ liệu ngoại lai.

Có rất nhiều nguyên nhân chủ quan và khách quan dẫn tới sự xuất hiện của các điểm ngoại lai trong tập dữ liệu như: Các lỗi nhập dữ liệu do con người gây ra; Các lỗi đo lường do thiết bị, dụng cụ lấy mẫu, thí nghiệm gây ra; Do cố ý tạo ra để phục vụ việc kiểm tra các phương pháp phát hiện; Các lỗi xử lý dữ liệu phát sinh trong quá trình thao tác dữ liệu; Các lỗi do lấy mẫu được trích xuất hoặc trộn dữ liệu từ các nguồn sai khác nhau; Do tự nhiên gây ra, đây không phải là lỗi mà là các giá trị quan sát thật tuy nhiên rất hiếm khi xuất hiện (N.N.R Ranga Suri and et al., 2018).

Trong khai phá dữ liệu và trong các tài liệu thống kê, dữ liệu ngoại lai còn được gọi là dữ liệu bất thường (anomalies), lệch lạc (deviants)… Trong hầu hết các ứng dụng, dữ liệu được tạo ra bởi quá trình sinh dữ liệu, phản ánh hoạt động của hệ thống hoặc các quan sát thu thập về các thực thể. Khi quá trình tạo ra có những vấn đề bất thường kết quả sẽ tạo ra các ngoại lai. Do đó, các giá trị ngoại lai thường chứa đựng những thông tin hữu ích về những đặc điểm bất thường của hệ thống và thực thể ảnh hưởng tới quá trình sinh dữ liệu. Việc phát hiện dữ liệu bất thường giúp chúng ta có những hiểu biết sâu sắc về từng ứng dụng cụ thể. Một số ứng dụng của dữ liệu ngoại lai trong thực tế (C.Aggarwal, 2017) có thể chỉ ra như:

- Hệ thống phát hiện xâm nhập (Intrusion detection systems)
- Phát hiện gian lận tín dụng (Credit card fraud)
- Các sự kiện cảm biến quan tâm (Interesing sensor events)
- Trong chuẩn đoán y tế (Medical diagnosis)
- Trong thực thi pháp luật (Law enforcement)
- Trong khoa học trái đất (Earth science)

Có nhiều phương pháp để phát hiện các điểm dữ liệu ngoại lại, Trong (C.Aggarwal, 2017) đã liệt kê một số phương pháp cơ bản được sử dụng bao gồm:

- Phân tích giá trị cực trị (Extreme Value Analysis): Đây là phương pháp cơ bản nhất được sử dụng để phát hiện các điểm ngoại lai, áp dụng tốt cho dữ liệu một chiều.

- Các mô hình xác suất và thống kê (Probabilistic and Statistical Models): Phương pháp này áp đặt một phân bố cụ thể trên tập dữ liệu như phân bố đều, phân bố Bernoulli, phân bố Poisson…Sau đó tính xác suất cho các phần tử thuộc tập dữ liệu ban đầu, các phần tử nào có xác suất thấp sẽ được cho là điểm ngoại lai.

- Các mô hình tuyến tính (Linear Models): Với phương pháp này, sẽ phải chuyển đổi tập dữ liệu ban đầu sang không gian ít chiều hơn bằng cách sử dụng tương quan tuyến tính. Sau đó, khoảng cách của từng điểm dữ liệu đến mặt phẳng ở không gian mới sẽ được tính toán và khoảng cách này sẽ được dùng để tìm ra các điểm ngoại lai.

- Các mô hình dựa trên lân cận (Proximity - based Models): Phương pháp này dựa trên ý tưởng là mô hình hóa các điểm ngoại lai sao cho chúng hoàn toàn tách biệt khỏi toàn bộ các điểm dữ liệu còn lại. Phân cụm, phân tích dựa trên mật độ, phân tích dựa trên người hàng xóm gần nhất là các hướng tiếp cận chính của phương pháp này.

- Các mô hình dựa trên lý thuyết thông tin (Information Theoretic Models): Phương pháp này dựa trên nguyên lý các điểm ngoại lai sẽ làm tăng giá trị minimum code length khi mô tả tập dữ liệu.

Dữ liệu nhiệt độ thu thập được từ các trạm quan trắc đều là các dữ liệu một chiều. Quá trình làm việc với dữ liệu này, có 2 dạng ngoại lai chủ yếu được phát hiện và xử lý bao gồm:

- Ngoại lai trái (Left outlier): Là các điểm ngoại lai có giá trị cực tiểu (Extreamly low) trong tập mẫu quan sát (C.Aggarwal, 2017).

- Ngoại lai phải (Right outlier): Là các điểm ngoại lai có giá trị cực đại (Extreamly large) trong tập mẫu quan sát (C.Aggarwal, 2017).

Do đặc điểm của tập dữ liệu, phương pháp được dùng để phát hiện dữ liệu ngoại lai áp dụng cho 43 trạm quan trắc của Việt Nam thuộc nhóm đầu tiên đã chỉ ra ở trên là phân tích giá trị cực trị, trong đó 2 phương pháp chính là phương pháp sử dụng Z - Score và phương pháp sử dụng đồ thị Box - plot. Chi tiết của hai phương pháp này được trình bày trong phần 3.2 dưới đây.

3.2. Phát hiện ngoại lai cho dữ liệu một chiều

3.2.1. Phương pháp sử dụng Z - Score

Điểm tiêu chuẩn hay Z - Score chỉ ra một thành phần chênh lệch so với trung bình là bao nhiêu độ lệch chuẩn (C.Aggarwal, 2017). Z - Score của bất kỳ một điểm dữ liệu nào được tính theo công thức:

$$z = \frac{(x - \mu)}{\sigma}$$

Trong đó: x là giá trị của điểm dữ liệu cần tính Z - Score; μ là giá trị trung bình của tập dữ liệu; σ là độ lệch chuẩn của tập dữ liệu. (Nếu z < 0 thể hiện

điểm dữ liệu đó nhỏ hơn giá trị trung bình; Nếu z > 0 thể hiện điểm dữ liệu đó lớn hơn giá trị trung bình; Nếu z = 0 thể hiện điểm dữ liệu đó bằng với giá trị trung bình).

Sau khi tính toán Z - Score cho từng điểm trong tập dữ liệu, một ngưỡng (threshold) sẽ được thiết lập để lọc các điểm này so với giá trị trung bình. Nếu tập dữ liệu theo phân phối chuẩn như chỉ ra trong Hình 4, cho thấy:

- Với ngưỡng 2.5 (- 2.5 < *Z* - Score < +2.5) có 99% điểm dữ liệu nằm trong phạm vi 2.5 lần độ lệch chuẩn.
- Với ngưỡng 3.0 (- 3.0 < *Z* - Score < +3.0) có 99.8% điểm dữ liệu nằm trong phạm vi 3.0 lần độ lệch chuẩn.
- Với ngưỡng 5.0 (- 5.0 < *Z* - Score < +5.0) có 99.9999426% điểm dữ liệu nằm trong phạm vi 5.0 lần độ lệch chuẩn.

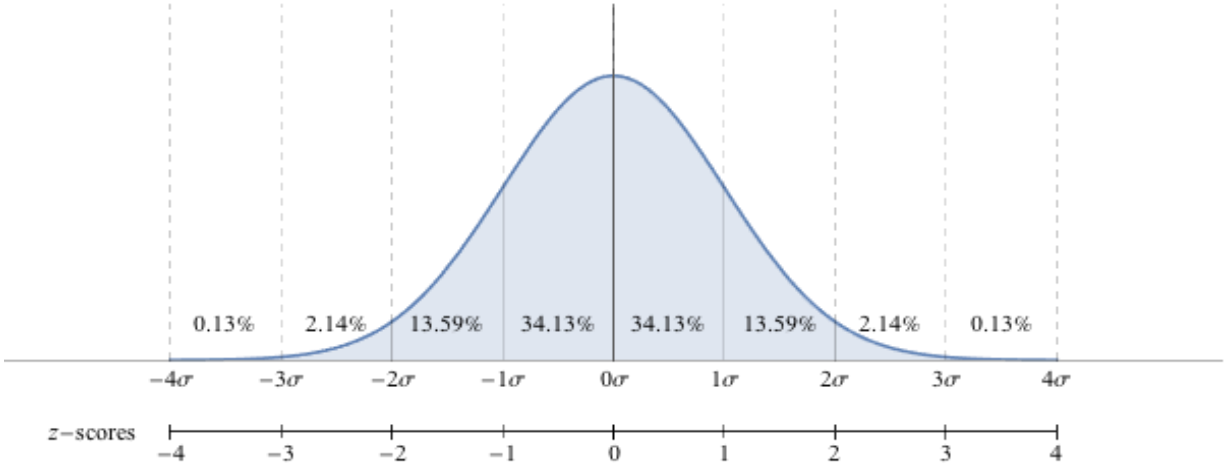
Như vậy bằng cách gắn thẻ, hoặc lọc các điểm dữ liệu nằm ngoài ngưỡng nhất định, chúng ta có thể phân loại các điểm dữ liệu thành ngoại lai và không ngoại lai. Z - Score là một phương pháp đơn giản nhưng khá mạnh mẽ để phát hiện các điểm ngoại lai trong một tập dữ liệu. Tuy nhiên, phương pháp này chỉ tốt đối với dữ liệu có số chiều thấp và có phân phối chuẩn.

3.2.2. Phương pháp sử dụng biểu đồ Box - plot

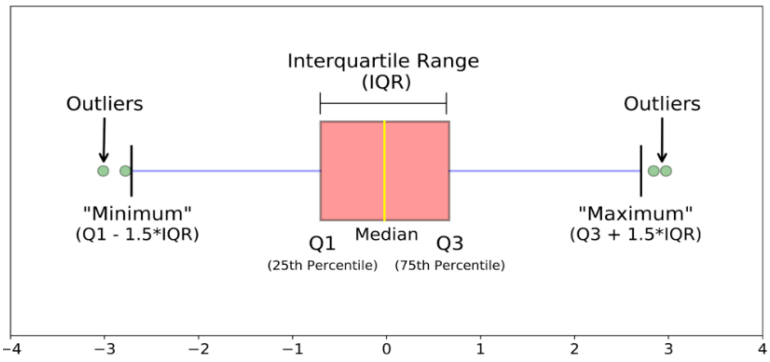
Biểu đồ Box - plot được sử dụng để đo khuynh hướng phân tán và xác định các giá trị ngoại lai của tập dữ liệu. Biểu đồ Box - plot chia tập dữ liệu thành các khoảng phần tư, phần thân của biểu đồ bao gồm một chiếc hộp, biểu đồ thể hiện 5 giá trị của tập dữ liệu (Hình 5) bao gồm:

- Giá trị bé nhất (Minimum) của tập dữ liệu được xác định bằng Q1 - 1.5 * IQR;
- Tứ phân vị thứ nhất (Q1) của tập dữ liệu.
- Tứ phân vị thứ hai (Q2) chính là giá trị trung vị (Median) của tập dữ liệu.
- Tứ phân vị thứ ba (Q3) của tập dữ liệu.
- Giá trị lớn nhất (Maximum) của tập dữ liệu có giá trị bằng Q3 + 1.5* IQR.

Nếu tập dữ liệu có chứa các giá trị ngoại lai thì chiều dài tối đa của 2 râu tính từ mỗi cạnh hộp



Hình 4. Tỷ lệ điểm dữ liệu nằm trong phạm vi theo ngưỡng Z - Score với phân phối chuẩn.



Hình 5. Hình dạng và các giá trị của tập dữ liệu thể hiện trên biểu đồ Box - plot.

sẽ được xác định bằng 1.5 lần độ trải giữa (IQR - Interquatile Range). Các điểm dữ liệu nằm ngoài râu Minimum được xem xét là các điểm ngoại lai trái (Left outlier), các điểm dữ liệu nằm ngoài râu Maximum được xem xét là các điểm ngoại lai phải (Right outlier). Các điểm dữ liệu ngoại lai này được thể hiện bằng dấu chấm tròn trên biểu đồ Box - plot . Như trong hình 5 ở trên thể hiện 2 điểm ngoại lai trái và 2 điểm ngoại lai phải. Ngoài ra, biểu đồ Box - plot còn cung cấp thông tin về hình dạng của tập dữ liệu. Nếu đường trung vị (Median) chia hộp thành 2 nửa đều nhau, thì tập dữ liệu này đối xứng; Nếu nửa phải lớn hơn nửa trái thì tập dữ liệu bị lệch phải, và ngược lại, nếu nửa trái lớn hơn nửa phải thì tập dữ liệu bị lệch trái (Munzer, 2014).

Box - plot là đồ thị trực quan thường được các nhà phân tích, thống kê, nhà khoa học dữ liệu sử dụng để tóm tắt thông tin về một biến dữ liệu định lượng bất kỳ phục vụ cho nhiều giai đoạn trong quá trình khai thác và tiền xử lý dữ liệu (Nguyễn Văn Tuấn, 2014).

3.3. Xử lý dữ liệu ngoại lai

Việc phát hiện các điểm dữ liệu ngoại lai có thể thực hiện bằng nhiều phương pháp khác nhau, sau khi phát hiện được các điểm ngoại lai yêu cầu đặt ra là phải xử lý chúng. Các điểm dữ liệu ngoại lai có ảnh hưởng rất lớn đến độ chính xác của các mô hình, việc lựa chọn được phương pháp nào để xử lý sao cho phù hợp với từng loại dữ liệu cụ thể thường khó hơn rất nhiều so với việc phát hiện ra chúng (N.N.R Ranga Suri and et al., 2018).

Cũng tương tự như việc phát hiện, để xử lý các điểm ngoại lai cũng có nhiều phương pháp. Mỗi một phương pháp lại có ưu và nhược điểm riêng. Việc chọn phương pháp xử lý nào tùy thuộc vào yêu cầu phân tích dữ liệu của từng bài toán cụ thể đặt ra. Dưới đây là tổng hợp các phương pháp xử lý ngoại lai chung cho tập dữ liệu:

- Loại bỏ các dòng chứa ngoại lai khỏi tập dữ liệu: Đây là cách xử lý ngoại lai đơn giản và dễ thực hiện nhất. Sau khi phát hiện các điểm ngoại lai thực hiện xóa các dòng dữ liệu chứa giá trị ngoại lai này khỏi tập dữ liệu. Tuy nhiên, phương pháp này chỉ áp dụng cho tập dữ liệu chứa các biến độc lập. Với dữ liệu dạng chuỗi thời gian (Time series data), chúng ta không thể sử dụng phương pháp này để loại bỏ một điểm ngoại lai tại một vị trí vì các điểm dữ liệu trong chuỗi thời gian có mối quan hệ tương quan với nhau. Ngoài ra, với dữ liệu có nhiều thuộc tính khác nhau, nếu xóa cả dòng dữ liệu chứa một thuộc tính có giá trị ngoại lai sẽ làm mất thông tin trên các cột khác nếu cột này cần cho phân tích.

- Thay thế bằng một giá trị khác: Thay thế giá trị của các điểm ngoại lai bằng một giá trị khác phù hợp hơn với tập dữ liệu. Với phương pháp này vấn đề khó khăn gặp phải đó là lựa chọn giá trị nào để thay thế cho giá trị của điểm ngoại lai? Câu trả lời là tùy thuộc vào từng loại dữ liệu, kiểu dữ liệu và trong những ngữ cảnh cụ thể để xác định được giá trị thay thế phù hợp nhất. Trong một số trường hợp có thể thay thế các giá trị ngoại lại bằng giá trị trung bình (mean) của tập dữ liệu, hoặc thay thế bằng một giá trị cụ thể (specific value) do các nhà phân tích dữ liệu, chuyên gia đề xuất.

- Thay thế giá trị của các điểm ngoại lai bằng NULL (empty): Việc thực hiện này sẽ chuyển đổi các điểm ngoại lai thành các điểm thiếu dữ liệu (missing value). Các điểm ngoại lai bây giờ được xem xét như là một điểm dữ liệu thiếu trong tập dữ liệu để xử lý.

Không có một phương pháp, cách thức xử lý ngoại lai chung nào áp dụng cho tất cả các bài toán, các kiểu dữ liệu khác nhau (N.N.R Ranga Suri and et al., 2018). Vì vậy, để lựa chọn được phương pháp phù hợp cần có những hiểu biết sâu sắc về tập dữ liệu, về bài toán đang giải quyết, có thể sử dụng chỉ một phương pháp xử lý ngoại lai và/hoặc kết hợp cả 3 nhóm phương pháp đã chỉ ra ở trên để xử lý ngoại lai cho cùng một tập dữ liệu.

4. Áp dụng cho dữ liệu nhiệt độ tại các trạm quan trắc 3h của Việt Nam

Trong phần 2 và 3 của bài báo, đã trình bày về tập dữ liệu nhiệt độ thu thập được từ 43 trạm quan trắc 3h của Việt Nam, hai phương pháp được sử dụng phổ biến và hiệu quả trong việc phát hiện ngoại lai cho dữ liệu có số chiều thấp là Z - Score và Box - plot. Ở phần này, tác giả sẽ sử dụng kỹ năng lập trình kết hợp với các thư viện, hệ thống mã nguồn mở để áp dụng các phương pháp đó cho việc phát hiện và xử lý ngoại lai với dữ liệu nhiệt độ tại 43 trạm quan trắc 3h của Việt Nam.

Dữ liệu nhiệt độ phụ thuộc rất lớn vào vị trí địa lý, với các tỉnh biên giới phía bắc nước ta nhiệt độ có thể hạ thấp xuống 0^0C hoặc thậm chí là âm vẫn có thể coi là bình thường; Tuy nhiên với các tỉnh phía Nam, nếu dữ liệu nhiệt độ ghi nhận được có

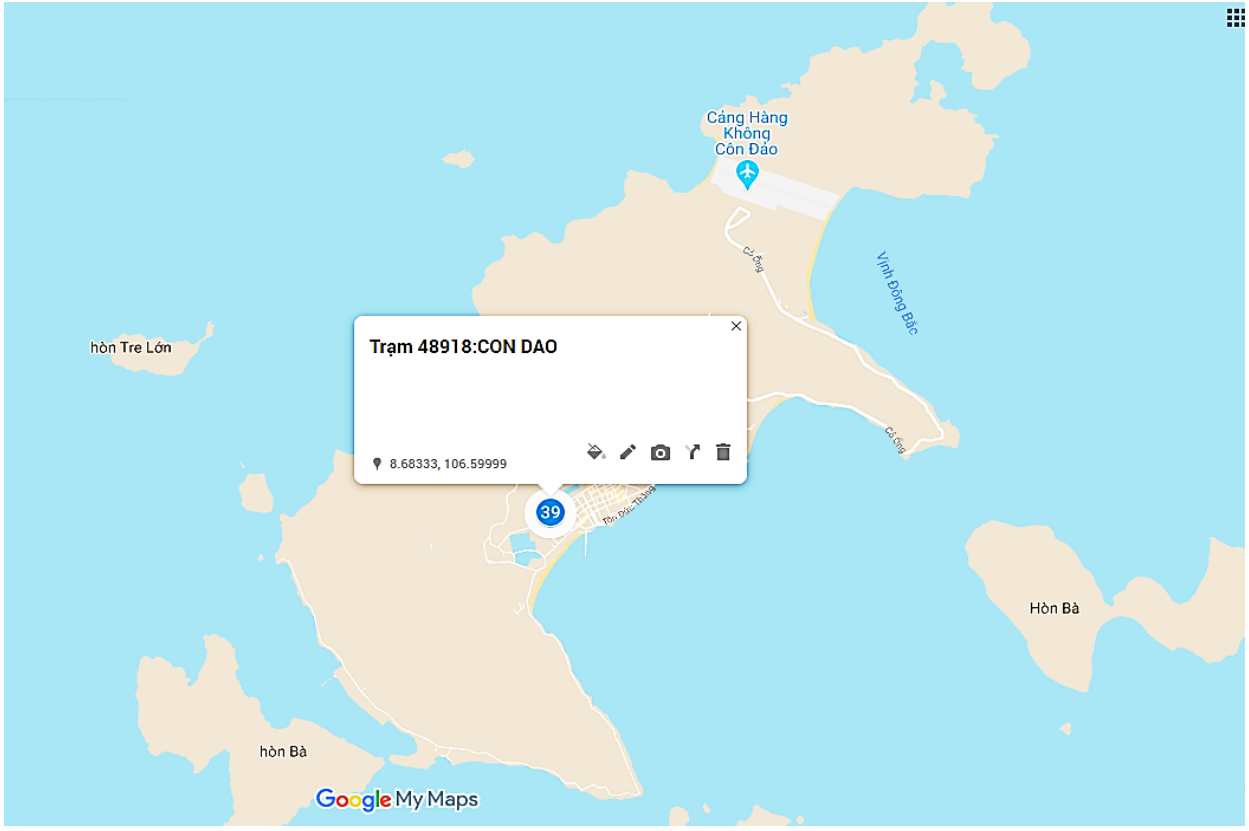
giá trị thấp dưới 15^{0}C có thể xem xét nó là các điểm ngoại lai cần phải được kiểm tra và xử lý. Do vậy, việc phát hiện và xử lý dữ liệu ngoại lai được thực hiện lần lượt cho từng trạm và không có một ngưỡng chung nào được áp dụng cho tất cả các trạm.

Trong khuôn khổ của bài báo, tác giả chọn một trạm điển hình (Trạm 48918: Côn Đảo) trong số 43 trạm để trình bày, minh họa chi tiết việc phát hiện và xử lý các điểm ngoại lai. Các trạm còn lại cũng sẽ được xử lý lần lượt theo các bước tương tự như với trạm này. Tác giả lựa chọn trạm 48918 trình bày trong bài báo vì đây là trạm nằm trên Đảo có điều kiện khí hậu khắc nghiệt, việc truyền dữ liệu gặp nhiều khó khăn,... có nhiều nguyên nhân dẫn đến các điểm ngoại lai trong dữ liệu quan trắc. Trạm 48918 có số thứ tự 39 trong Bảng 1, vị trí của trạm này được thể hiện tương đối trong Hình 1 ở trên và chi tiết trong Hình 6.

Để lập trình chúng tôi lựa chọn ngôn ngữ lập trình Python, mã nguồn được viết trên hệ thống Google Colab, sử dụng 3 thư viện nguồn mở để tính toán và trực quan hóa bao gồm: Pandas, Matplotlib, Seaborn. Tiến hành đọc và trích xuất dữ liệu quan trắc của trạm 48918 trong tập dữ liệu thô Data_Temp43_Original.csv. Bảng 2 chỉ ra thông số của tập dữ liệu và Hình 7 thể hiện biểu đồ tần suất (histogram) của dữ liệu nhiệt độ trạm 48918.

Bảng 2. Thông số tập dữ liệu quan trắc của trạm 48918.

TT	Thời điểm bắt đầu dữ liệu	01:00:00 01 - 01 - 2014
1	Thời điểm kết thúc dữ liệu	22:00:00 31 - 12 - 2019
2	Tổng số điểm dữ liệu	17 528
3	Số điểm có dữ liệu	17 495
4	Số điểm dữ liệu thiếu	33
5	Giá trị trung bình của tập dữ liệu	27.8478
6	Độ lệch chuẩn của tập dữ liệu	2.0407
7	Giá trị cực tiểu	- 1.3
8	Tứ phân vị thứ nhất (Q1)	26.4
9	Tứ phân vị thứ hai (Q2)	27.8
10	Tứ phân vị thứ ba (Q3)	29.1
11	Giá trị cực đại	39.0



Hình 6. Vị trí trạm 48918:CON DAO trên Google Maps.

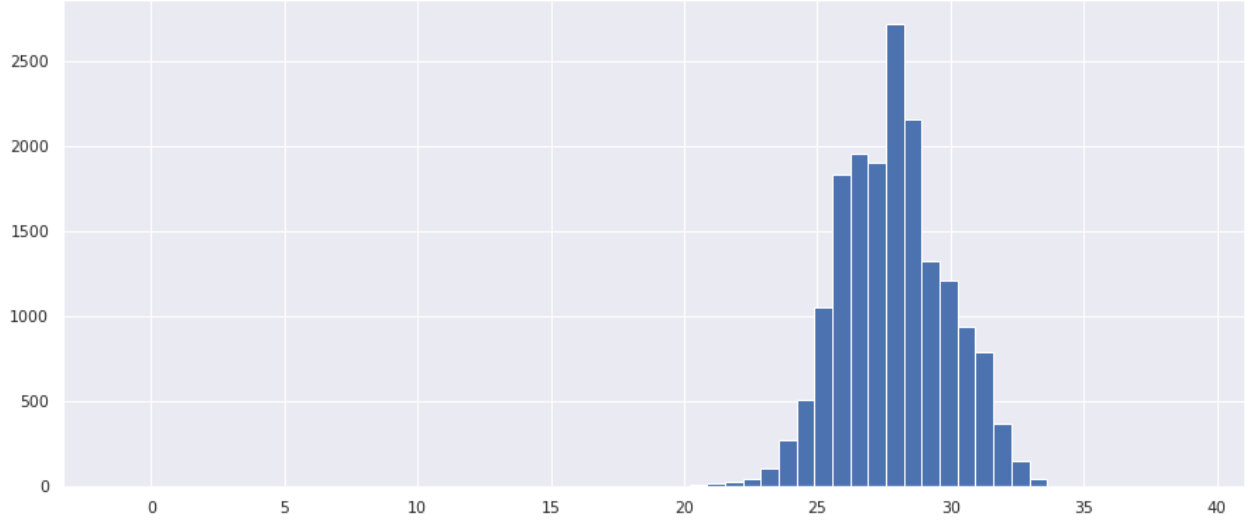
4.1. Sử dụng Z - Score phát hiện ngoại lai trong dữ liệu nhiệt độ của trạm 48918

Thực hiện tính giá trị Z - Score theo công thức (*) đã trình bày trong phần 3.2.1 của tất cả các điểm trong tập dữ liệu trạm 48918. Mã lập trình việc tính toán và kết quả được thể hiện như trong Hình 8.

Để xem xét và xác định ngoại lai, sau khi tính được giá trị *Z* - score đó là phải chọn một ngưỡng (threshold) phù hợp. Khi trao đổi với chuyên gia khí tượng thủy văn, cùng với các số liệu thống kê nhiệt độ tại khu vực phía Nam nói chung, Côn Đảo nói riêng thì nhiệt độ ở đây nằm trong khoảng [170C , 380C]. Như vậy các thông số quan trắc thấp hơn 170C và cao hơn 380C sẽ được xem xét là ngoại lai.

Theo như Bảng 3, với ngưỡng *Z* = 5, thỏa mãn điều kiện giới hạn nhiệt độ trong khoảng [170C , 380C]. Vì vậy, giá trị 5 được chọn là ngưỡng để lọc các điểm xem xét ngoại lai. Kết quả lọc các điểm có Z - Score nằm ngoài ngưỡng 5 cho trạm 48918 như trong Hình 9.

Như vậy, theo phương pháp *Z* - Score với ngưỡng lọc chọn bằng 5 có tất cả 6 điểm dữ liệu được xem xét là ngoại lai, trong đó có 5 điểm ngoại lai trái (zscore < 0) và 1 điểm ngoại lai phải (zscore>0).



Hình 7. Biểu đồ histogram của tập dữ liệu nhiệt độ trạm 48918.

```
#Tính Zscores cho từng điểm dữ liệu của trạm 48918 lưu vào cột zscore
# zscore = (x -mean)/std
mean_48918 = df_48918['48918'].mean()
std_48918 = df_48918['48918'].std(ddof=0)
df_48918['zscore'] = (df_48918['48918'] - mean_48918)/std_48918

#-------------------Hiển thị 5 dòng dữ liệu đầu tiên ----------------
df_48918.head()
```

TimeVN	48918	zscore
2014-01-01 01:00:00	24.3	-1.738529
2014-01-01 04:00:00	24.0	-1.885539
2014-01-01 07:00:00	24.2	-1.787532
2014-01-01 10:00:00	25.6	-1.101487
2014-01-01 13:00:00	26.6	-0.611455

Hình 8. Kết quả tính Z - Score cho các điểm quan trắc của trạm 48918.

Bảng 3. Ngưỡng và khoảng nhiệt tương ứng với ngưỡng thiết lập của trạm 48918.

TT	Ngưỡng (threshold)	Giới hạn theo ngưỡng Z	Khoảng nhiệt độ nằm trong giới hạn ngưỡng Z(mean: 27.85 |std:2.04)
1	3.0	- 3.0 <= Z<= 3.0	[21.73^{0}C - 33.97^{0}C]
2	4.0	- 4.0 <= Z<= 4.0	[19.69^{0}C - 36.01^{0}C]
3	5.0	- 5.0 <= Z<= 5.0	[17.65^{0}C - 38.05^{0}C]
4	5.5	- 5.5 <= Z<= 5.5	[16.63^{0}C - 39.07^{0}C]

4.2. Sử dụng biểu đồ Box - plot phát hiện ngoại lai trong dữ liệu nhiệt độ của trạm 48918

Sử dụng thư viện Matplotlib và Seaborn để dựng biểu đồ Box - Plot cho dữ liệu nhiệt độ trạm 48918. Theo như biểu đồ Box - plot trong Hình 10(a) ta có thể nhận thấy có khá nhiều điểm dữ liệu nằm trên và dưới hai râu minimum và maximum của biểu đồ, về nguyên tắc các điểm này đều được xem xét là các điểm dữ liệu ngoại lai. Tuy nhiên, như đã trình bày trong phần a, thông số nhiệt độ ở Côn Đảo thường nằm trong khoảng [170C, 380C]. Do đó, từ biểu đồ Box - plot ta có thể lọc các điểm ngoại lai trái với ngưỡng 170C, các điểm ngoại lai phải với ngưỡng 380C.

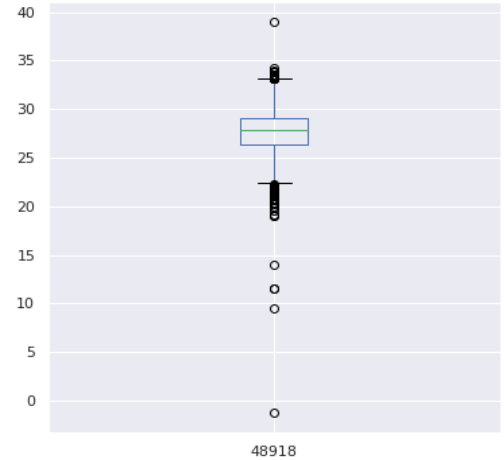
```
#So sánh Zscore với một ngưỡng (=5), điểm nào có zscore>5 được xem xét ngoai lai
#Cột outlier cho biết giá trị zscore có lớn hơn ngưỡng hay không
#(0: Không | 1: Có)
df_48918['outlier'] = (abs(df_48918['zscore'])>5).astype(int)

#Hiển thị những thời điểm xem xét ngoại lai (Zscore>5)
df_48918.loc[df_48918.outlier==1]
```

TimeVN	48918	zscore	outlier
2015-02-01 16:00:00	-1.3	-14.283356	1
2015-02-10 19:00:00	14.0	-6.785862	1
2015-09-10 10:00:00	39.0	5.464945	1
2016-03-20 13:00:00	11.5	-8.010942	1
2016-04-30 22:00:00	11.5	-8.010942	1
2017-06-15 16:00:00	9.5	-8.991007	1

Hình 9. Phát hiện dữ liệu ngoại lai sử dụng Z - Score với trạm 48918.

```
#Dựng biểu đồ Box-plot của dữ liệu nhiệt độ trạm 48918
df_48918.loc[:,['48918']].boxplot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f66e17ef0b8>
```



(a)

```
#Hiển thị danh sách ngoại lai trái (left outlier)
df_48918[df_48918['48918']<17]['48918']
```

```
TimeVN
2015-02-01 16:00:00    -1.3
2015-02-10 19:00:00    14.0
2016-03-20 13:00:00    11.5
2016-04-30 22:00:00    11.5
2017-06-15 16:00:00     9.5
Name: 48918, dtype: float64
```

```
#Hiển thị danh sách ngoại lai phải (right outlier)
df_48918[df_48918['48918']>38]['48918']
```

```
TimeVN
2015-09-10 10:00:00    39.0
Name: 48918, dtype: float64
```

(b)

Hình 10. Biểu đồ box - plot và các điểm xem xét ngoại lai của trạm 48918.

Kết quả tách các điểm ngoại lai trái - phải được thể hiện trong Hình 10(b).

Theo như kết quả thu được cả hai phương pháp sử dụng Z - Score và Biểu đồ Box - plot đều cho cùng một danh sách các điểm ngoại lai với 6 điểm dữ liệu chi tiết như trong Bảng 4.

Bảng 4. Thời điểm và giá trị quan trắc xem xét ngoại lai của trạm 48918.

TT	Thời điểm	Giá trị quan trắc	Loại ngoại lai
1	2015 - 02 - 01 16:00:00	- 1.3	Ngoại lai trái (Left outlier)
2	2015 - 02 - 10 19:00:00	14.0	
3	2016 - 03 - 20 13:00:00	11.5	
4	2016 - 04 - 30 22:00:00	11.5	
5	2017 - 06 - 15 16:00:00	9.5	
6	2015 - 09 - 10 10:00:00	39.0	Ngoại lai phải (Right outlier)

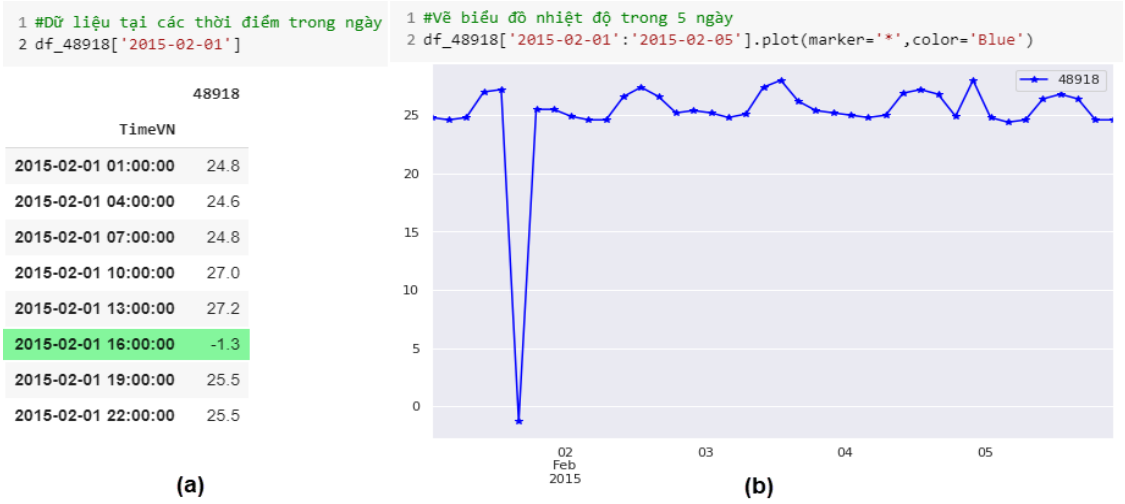
Đây chỉ là các điểm xem xét ngoại lai, để khẳng định có phải là ngoại lai thật hay không thì cần phải được kiểm chứng.

4.3. Kiểm chứng các điểm ngoại lai phát hiện được

Dữ liệu nhiệt độ thu nhận được từ các trạm quan trắc như đã trình bày có dạng chuỗi thời gian, sau mỗi khoảng thời gian 3h sẽ có một điểm dữ liệu mới. Do vậy, để khẳng định đây là các điểm ngoại lai, cần xem xét điểm dữ liệu này trong một chuỗi dữ liệu tương ứng với điểm đó. Trong phần này, Chúng tôi sẽ thực hiện kiểm chứng tại hai thời điểm có giá trị nhỏ nhất (- 1.3) và lớn nhất (39.0), các vị trí khác được kiểm chứng tương tự và mô tả cụ thể trong phần mã nguồn của bài báo (địa chỉ mã nguồn xử lý được cung cấp ở phần cuối của bài báo).

Với kết quả kiểm chứng dữ liệu nhiệt độ của trạm 48918 tại thời điểm 16h ngày 01/02/2015 như thể hiện trong Hình 11, có thể khẳng định dữ liệu quan trắc thu thập được tại thời điểm này là hoàn toàn sai lệch. Nhiệt độ ghi nhận tại thời điểm 16h phải có mối tương quan với nhiệt độ tại thời điểm trước đó lúc 13h và sau đó lúc 19h; Ngoài ra đồ thị biểu diễn nhiệt độ của trạm trong khoảng thời gian 5 ngày từ 1h ngày 01/02/2015 đến 22h ngày 05/02/2015 (Hình 11(b)) cũng thể hiện rõ mức độ sai khác dữ liệu tại thời điểm này.

Tương tự như vậy, Hình 12 thể hiện kết quả kiểm chứng dữ liệu quan trắc tại thời điểm 10h ngày 10/09/2015. Hình 12(a) hiển thị toàn bộ số liệu quan trắc trong ngày 10/09/2015 tại các thời điểm 1h, 4h, 7h, 10h, 13h, 16h, 19h và 22h. Hình 12(b) thể hiện đồ thị nhiệt độ quan trắc trong khoảng thời gian 5 ngày từ 1h ngày 10/09/2015 đến 22h ngày 15/09/2015. Chúng ta cũng dễ dàng nhận thấy dữ liệu quan trắc tại thời điểm 10h ngày 10/09 có mức độ sai khác tương đối lớn so với mặt bằng chung của các điểm đo. Hơn nữa, nhiệt độ tại thời điểm 10h có giá trị là 39^{0}C cao hơn nhiệt độ lúc 7h là 27.8^{0}C (chênh lệch + 11.2^{0}C) và cao hơn nhiệt độ ghi nhận lúc 13h là 30.7^{0}C (chênh lệch +8.3^{0}C).



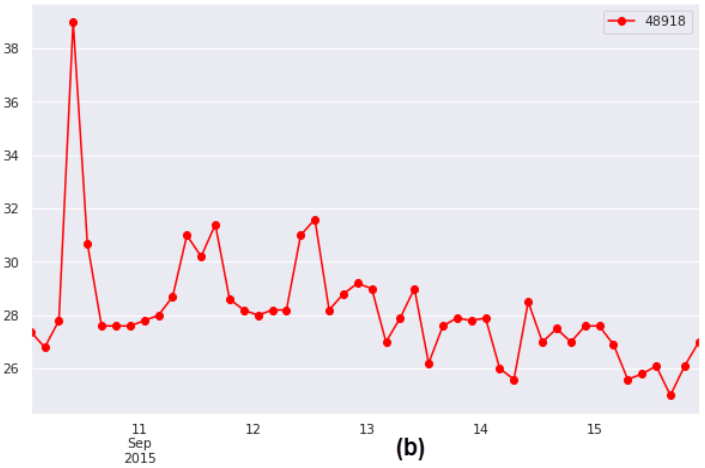
Hình 11. Kiểm chứng điểm ngoại lai trái có giá trị thấp nhất tại trạm 48918.

```
#Dữ liệu tại các thời điểm trong ngày
df_48918['2015-09-10']
```

TimeVN	48918
2015-09-10 01:00:00	27.4
2015-09-10 04:00:00	26.8
2015-09-10 07:00:00	27.8
2015-09-10 10:00:00	39.0
2015-09-10 13:00:00	30.7
2015-09-10 16:00:00	27.6
2015-09-10 19:00:00	27.6
2015-09-10 22:00:00	27.6

(a)

```
#Vẽ biểu đồ nhiệt độ trong 5 ngày
df_48918['2015-09-10':'2015-09-15'].plot(marker='o',color='Red')
```



(b)

Hình 12. Kiểm chứng điểm ngoại lai phải có giá trị cao nhất tại trạm 48918.

Điều này trong thực tế là phí lý khi mức độ thay đổi trong khoảng 3h là rất lớn, và thời điểm nhiệt độ cao nhất trong ngày không phải là thời điểm 13h như bình thường.

Từ các kết quả kiểm chứng có thể khẳng định các điểm này đều là các điểm dữ liệu ngoại lai, có giá trị sai khác rất lớn so với giá trị thực tế. Do vậy, dữ liệu tại các điểm này cần phải được xử lý trước khi sử dụng cho bất kỳ mục đích nào

4.3. Xử lý các điểm ngoại lai cho trạm 48918

Trong phần 3.3, đã chỉ ra các phương pháp để xử lý ngoại lai nói chung, như đã trình bày dữ liệu nhiệt độ quan trắc thu thập được là dữ liệu dạng chuỗi thời gian do vậy không thể sử dụng phương pháp loại bỏ các điểm này ra khỏi tập dữ liệu. Trong thực tế khi xử lý các điểm ngoại lai, tác giả chọn phương pháp thay thế các điểm ngoại lai về giá trị NULL (ứng với None trong Python - Hình 13), xem các điểm ngoại lai là điểm dữ liệu thiếu (missing data). Sau đó sẽ sử dụng phương pháp xử lý dữ liệu thiếu cho toàn bộ tập dữ liệu. Trong khuôn khổ nội dung của bài báo này tác giả không đề cập đến việc xử lý giá trị thiếu.

```
#Xử lý dữ liệu ngoại lai của trạm 48918
df_48918.loc['2015-02-01 16:00:00',['48918']] = None
df_48918.loc['2015-02-10 19:00:00',['48918']] = None
df_48918.loc['2016-03-20 13:00:00',['48918']] = None
df_48918.loc['2016-04-30 22:00:00',['48918']] = None
df_48918.loc['2017-06-15 16:00:00',['48918']] = None
df_48918.loc['2015-09-10 10:00:00',['48918']] = None
```

Hình 13. Chuyển đổi các điểm ngoai lai về giá trị NULL để xử lý cho trạm 48918.

Quá trình phát hiện và xử lý ngoại lai tại 42 trạm còn lại được thực hiện lần lượt theo các bước như đã trình bày với trạm 48918. Toàn bộ mã nguồn trình bày trong bài báo và xử lý ngoại lai cho các trạm còn lại tham khảo tại: https://colab.research.google.com/drive/1eI_yc3mZ-UlxWfxofa4B_hfB2_ULS5w

5. Kết luận

Phát hiện và xử lý dữ liệu ngoại lai là yêu cầu bắt buộc và rất quan trọng trong quá trình chuẩn bị dữ liệu. Các điểm ngoại lai có ảnh hưởng rất lớn tới độ chính xác của các mô hình dự đoán, dự báo. Trong nội dung của bài báo này, đã trình bày chi tiết về dữ liệu nhiệt độ thu thập được tại 43 trạm quan trắc 3h của Việt Nam; Tổng quan về dữ liệu ngoại lai nói chung và hai phương pháp xử lý điển hiển để phát hiện các điểm ngoại lai với dữ liệu có số chiều thấp là Z - Score và Box - plot. Kết quả chính của bài báo, thể hiện ở phần thực nghiệm, áp dụng các phương pháp Z - Score và Box - plot để phát hiện các điểm ngoại lai cho một trạm điển hình đó là trạm 48918 - Côn Đảo, các điểm ngoại lai sau khi phát hiện sẽ được kiểm chứng để sàng lọc một cách chính xác và xử lý về dạng dữ liệu thiếu (missing data). Sau khi thực hiện với toàn toàn bộ dữ liệu của các trạm, sẽ thu được một tập dữ liệu mới đã xử lý ngoại lai. Tập dữ liệu này tiếp tục được làm sạch với các yêu cầu khác như xử lý giá trị thiếu, chuẩn hóa...và sẽ được sử dụng làm dữ liệu đầu vào cho các mô hình dự báo liên quan.

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."

Tài liệu tham khảo

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

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

Hermine N., Akouemo, Richard J. Povinelli, (2014). Time series outlier detection and imputation, IEEE.

Nguyễn Văn Tuấn, (2014). Phân tích dữ liệu với R,Nhà xuất bản tổng hợp Thành phố Hồ Chí Minh.

Ranga Suri, N. N. R , Narasimha Murty M., Athithan, G., (2018). Outlier Detection: Techniques and Applications, Springer Nature Switzerland AG, Cham.

Tamara Munzer, (2014). Visualization Analysis and Design,CRC Press.

ABSTRACT

Detect and process outliers for temperature data at 3h monitoring stations in Vietnam

Nam Van Dang [1], Oanh Thi Nong [1], Hoai Xuan Nguyen [2], Manh Van Ngo [3], Hien Thi Nguyen [4]

[1] *Faculty of Information Technology, Hanoi University of Mining and Geology, Vietnam*
[2] *AI Academy Vietnam, Vietnam*
[3] *Center for Hydro - Meteorological Data and Information, Vietnam*
[4] *Falculty of Information Technology Technical University, Vietnam*

Data preparation is a compulsory process in any data science project. Many research have shown that it constitutes 80% of the time, effort and resources of a data science project. Depending on the particular project and data type, Data preparation step may required different methods/steps. Detecting and processing outlier data is one of the important preprocessing steps in data preparation , especially for time series data. This paper reviews two methods for detecting outliers for low dimensional data, namely Z - Score and Box - plot charts. We also present results of experiments which applied these methods for temperature data collected from 43 monitoring stations in 3 - hour in Vietnam over the last 6 years from 01/01/2014 to 31/12/2019.