Tạp chí Phát triển Khoa học và Công nghệ – Kĩ thuật và Công nghệ, 2(4):223-239

Open Access Full Text Article

Bài nghiên cứu

Ứng dụng phân tích thống kê để đánh giá độ tin cậy của nguồn dữ liệu đầu vào nhằm nâng cao chất lượng dự báo phụ tải điện ngắn hạn trên lưới điện TP.HCM

Lê Duy Phúc[1,2,*], Bùi Minh Dương[2], Phạm Anh Duy[3], Nguyễn Thanh Hoan[1], Bành Đức Hoài[1], Nguyễn Minh Tùng[1], Nguyễn Minh Khôi[1], Đoàn Ngọc Minh[1], Nguyễn Việt Dũng[1]

TÓM TẮT

Công tác dự báo phụ tải điện ngắn hạn đóng vai trò quan trọng trong việc vận hành hệ thống điện, đặc biệt là trên lưới điện Tp.HCM – Thành phố có sản lượng điện thương phẩm cũng như nhu cầu cung ứng điện cao nhất cả nước trong những năm qua. Qua khảo sát, phụ tải điện thuộc khu vực Tp.HCM thường xuyên xuất hiện những thay đổi đột biến và tạo nên những nhiễu động khi quan sát bộ cơ sở dữ liệu quá khứ. Theo đó, việc đánh giá độ tin cậy của bộ dữ liệu này sẽ rất cần thiết trong giai đoạn xử lý dữ liệu (còn gọi là khâu lọc dữ liệu) trước khi đưa vào các mô hình dự báo phụ tải điện để xuất kết quả dự báo. Nghiên cứu này trình bày một phương pháp lọc dữ liệu có xem xét đến độ tin cậy của nguồn dữ liệu bằng cách phân tích trên nhiều mức độ tin cậy khác nhau và có thực hiện đối chiếu, so sánh kết quả với các phương pháp lọc dữ liệu trước đây (chẳng hạn như các phương pháp lọc Kalman, DBSCAN, Wavelet Transform và SSA). Nguồn dữ liệu sử dụng trong nghiên cứu này được thu thập từ hơn 50 trạm trung gian thuộc lưới điện Tp.HCM và được đưa vào mô hình dự báo mạng nơ-ron ANN (Artificial Neural Network) và mô hình dự báo ARIMA (Autoregressive Integrated Moving Average) để chứng minh hiệu quả của phương pháp lọc dữ liệu đề xuất. Các kết quả mô phỏng xuất ra từ mô hình dự báo ANN và ARIMA cho thấy sự hiệu quả của phương pháp đề xuất, cụ thể, độ tin cậy dữ liệu của lưới điện Tp. Hồ Chí Minh ở mức 95% thì kết quả dự báo phụ tải tốt hơn so với khi không có áp dụng phương pháp lọc và khi sử dụng những phương pháp lọc khác.

Từ khoá: Dự báo phụ tải điện ngắn hạn, lọc dữ liệu, phân tích thống kê, độ tin cậy, mạng nơ-ron và ARIMA

[1]*Tổng công ty Điện lực Tp.HCM*

[2]*Viện Kỹ thuật, Trường Đại học Công nghệ TP.HCM*

[3]*Khoa Kỹ thuật, Trường Đại học Việt – Đức*

Liên hệ

Lê Duy Phúc, Tổng công ty Điện lực Tp.HCM

Viện Kỹ thuật, Trường Đại học Công nghệ TP.HCM

Email: phucld@hcmpc.com.vn

Lịch sử

- Ngày nhận: 15-10-2019
- Ngày chấp nhận: 25-11-2019
- Ngày đăng: 31-12-2019

DOI :10.32508/stdjet.v2i4.614

Check for updates

TỔNG QUAN

Hiện nay, hệ thống lưới điện phân phối ở Việt Nam đang bám sát lộ trình định hướng phát triển bền vững đã hoạch định sẵn. Tuy nhiên, sự xuất hiện của các nguồn năng lượng tái tạo (*Renewable Energy Source - RES*) cùng với sự đa dạng của các mô hình phụ tải đã ảnh hưởng đến nhiều mặt công tác như dự báo, quy hoạch và vận hành hệ thống điện. Trong một cụm khu vực gồm nhiều dạng tải khác nhau, phụ tải điện công nghiệp thường cao hơn nhiều so với phụ tải khu dân cư. Điều này dẫn đến các phụ tải dân cư có thể bị hiểu nhầm là nhiễu trong các thuật toán lọc dữ liệu. Bên cạnh đó, công suất phát từ các nguồn phát RES có thể thay đổi đột ngột do yếu tố tự nhiên. Chính vì vậy, việc cải thiện chất lượng dự báo phụ tải ngắn hạn là một vấn đề nghiên cứu cấp thiết đối với lưới điện phân phối, nơi có chứa các phụ tải thường xuyên biến động đột ngột và có tích hợp các nguồn RES.

Dự báo phụ tải ngắn hạn (*Short-time Load Forecasting - STLF*) trong lưới phân phối có thể được thực hiện bằng các thuật toán học máy thông thường hoặc phức tạp. Đã có rất nhiều công trình nghiên cứu về các phương pháp dự báo phụ tải ngắn hạn, nhưng chỉ một số ít nghiên cứu có đề cập hoặc tập trung vào việc phát triển giải thuật/thuật toán lọc dữ liệu trước khi áp dụng mô hình dự báo phụ tải [1–6]. Nguyên nhân xuất phát từ việc một số tác giả cho rằng dữ liệu đầu vào là hoàn hảo hoặc đã được lọc trước khi được khai thác trong mô hình dự báo phụ tải. Trong tài liệu [7], một mô hình ARIMA được kết hợp với một phương pháp ngày tương tự để dự báo phụ tải trong ngày. Trong các nghiên cứu [8,9], mạng nơ-ron hàm RB (*Radial Basis*) được sử dụng để phục vụ công tác dự báo phụ tải ngắn hạn. Trong các tài liệu [10–13], các phương pháp học máy thông thường, cụ thể là mạng nơ-ron nhân tạo, máy học ELM (*Ensembled Extreme Learning Machine*) và KNN (*K-nearest-neighbor*), cũng cho kết quả tốt khi thực hiện dự báo phụ tải ngắn hạn. Trong giai đoạn gần đây, phương pháp học sâu đã được áp dụng rộng rãi cho lĩnh vực nghiên cứu dự báo phụ tải. Cụ

Trích dẫn bài báo này: Phúc L D, Dương B M, Anh Duy P, Thanh Hoan N, Đức Hoài B, Minh Tùng N, Minh Khôi N, Ngọc Minh D, Việt Dũng N. **Ứng dụng phân tích thống kê để đánh giá độ tin cậy của nguồn dữ liệu đầu vào nhằm nâng cao chất lượng dự báo phụ tải điện ngắn hạn trên lưới điện TP.HCM**. *Sci. Tech. Dev. J. - Eng. Tech.;* 2(4):223-239.

thể, các mạng nơ-ron học sâu mạnh mẽ, chẳng hạn như mạng nơ-ron CNN (*Convolutional Neural Network)* và LSTM (*Long-Short Term Memory*) đã được triển khai thành công và thu được kết quả tốt trong dự báo tải[14,15]. Theo đó, các mô hình CNN được sử dụng để nhận ra mẫu tổng thể, trong khi mô hình LSTM được sử dụng để trích xuất mối quan hệ giữa các bước thời gian. Bài báo[15] đã cho thấy những tác động của nhiễu theo cấp số khi sử dụng trong mạng nơ-ron lan truyền ngược (BP - back propagation neural network) và bài báo cũng đã đưa ra phương pháp giảm thiểu nhiễu bằng cách sử dụng hàm Wavelets để tiền xử lý trước khi thực hiện mô hình dự báo sử dụng SARIMA và BP. Theo khảo sát hiện tại, rất ít nghiên cứu để cập đến thuật toán lọc hoặc các phương pháp phát hiện bất thường của nguồn dữ liệu để phục vụ cho công tác dự báo phụ tải ngắn hạn. Chỉ một vài thuật toán phổ biến thường được sử dụng như bộ lọc Kalman, DBSCAN, bộ lọc rời rạc dựa trên biến đổi Wavelet (*Discrete Wavelet-transform - DW*) và phân tích SSA (*Singular Spectrum Analysis*).

Bộ lọc Kalman hoạt động như một thuật toán hiệu chỉnh tín hiệu thời gian thực khi các hệ số của nó thích ứng để thay đổi tín hiệu lõi và loại bỏ nhiễu ngẫu nhiên trong bộ dữ liệu. Do đó, bộ lọc Kalman cơ bản được áp dụng như một công cụ ước tính trong STLF hoặc dự báo phụ tải siêu ngắn hạn (*Very Short-term Load Forecasting - VSTLF*) thay vì sử dụng nó như một bộ lọc thông thường để điều chỉnh nhiễu trong quá trình thu thập dữ liệu. Công cụ dự báo dựa trên bộ lọc Kalman có thể là một công cụ dự báo độc lập[16,17] hoặc được kết hợp với thuật toán ước tính lõi để tạo thành công cụ ước tính lai[18–20]. Tóm lại, các công trình nêu trên cho thấy ứng dụng hiệu quả của bộ lọc Kalman để dự báo nhu cầu tải trong tương lai; tuy nhiên, cách tiếp cận bộ lọc Kalman này yêu cầu phải thiết lập mô hình trạng thái; trong đó, các tham số của nó phải được mô hình hóa triệt để trong không gian trạng thái.

Phân cụm không gian dựa trên mật độ của các đối tượng có nhiễu (*Density-Based Spatial Clustering of Applications with Noise - DBSCAN)* là thuật toán phân cụm theo vùng tối thiểu dựa trên các tham số được xác định trước, trong khi nó có thể khai thác theo các cụm hình dạng tùy ý với hiệu quả tốt thông qua việc xác định mật độ tối đa các điểm kết nối[21]. DBSCAN không yêu cầu xác định trước số lượng cụm, nhưng thay vào đó, nó cần xác định trước ba tham số: khoảng cách, số tối thiểu các mẫu trong một cụm và loại đo khoảng cách. Trong tham khảo[1], DBSCAN đã được sử dụng để phát hiện các dấu hiệu bất thường trong lịch sử tiêu thụ năng lượng nhằm xác định mức độ nhất quán của phụ tải, tuy nhiên, nghiên cứu này không đề xuất bất kỳ giải pháp lọc thích hợp nào để xử lý các bất thường được phát hiện.

Chuỗi dữ liệu phụ tải được coi là một hệ thống đa thành phần được hình thành từ các yếu tố phi tuyến tính và các thành phần chu kỳ. Trong những năm gần đây, biến đổi Wavelet đã được thực hiện một cách hiệu quả để tách chuỗi ngẫu nhiên thành các mức phân giải tương ứng với các nhóm tín hiệu có đặc trưng khác nhau. Nó đã mang lại một số kết quả đầy hứa hẹn trong việc giải quyết các vấn đề liên quan đến dự báo phụ tải trên lưới phân phối[22–24]. Cụ thể, độ tin cậy của mô hình dự báo sử dụng Wavelet có thể tăng lên rất nhiều bởi vì tạo ra tính xác định và độ ổn định của tín hiệu đầu vào. Tuy nhiên, khi sử dụng biến đổi Wavelet, người dùng cần quan tâm đến vấn đề đánh đổi giữa độ sâu độ phân giải và số lượng điểm dữ liệu đầu vào. Do đó, mức độ phân giải phải được xác định theo cách cân bằng số lượng điểm dữ liệu trong mỗi cấp do tính chất của việc thu thập dữ liệu tải. Tương tự, phân tích SSA cũng là một kỹ thuật phân rã tín hiệu được sử dụng như một kỹ thuật phân tích chuỗi thời gian đáng tin cậy, để phát hiện và trích xuất các xu hướng, các thành phần định kỳ và nhiễu. SSA, một công cụ mạnh mẽ về kỹ thuật không tham số trong dự báo phụ tải khi xem xét các yếu tố của phân tích chuỗi thời gian, thống kê đa biến, hình học, hệ thống động lực và xử lý tín hiệu[25]. Tuy nhiên, kỹ thuật này có thể không được sử dụng phổ biến cho dự báo theo chuỗi thời gian.

Nhìn chung, các phương pháp lọc dữ liệu được để cập ở trên, về cơ bản được phân thành ba dạng: hiệu chỉnh đo lường, phân cụm dữ liệu và phân tách tín hiệu. Ba loại này được áp dụng trực tiếp vào dữ liệu thô, có độ tin cậy thấp và độ ổn định không cao do các lỗi thu thập dữ liệu từ nhiều điểm đo lường. Do đó, tính linh hoạt và tính thích ứng đối với các hệ thống thu thập dữ liệu khác nhau là thấp, bởi vì việc lọc dữ liệu được thực hiện bất kể độ tin cậy của dữ liệu như thế nào. Đây là một khuyết điểm có thể được khắc phục bằng cách phát triển một phương pháp lọc dữ liệu dựa trên kỹ thuật phân tích thống kê từ phân phối độ lệch của dữ liệu tải theo giờ trong hai ngày liên tiếp. Điều này làm tăng tính ổn định của chuỗi đầu vào khi xảy ra việc dữ liệu mất xu hướng. Ngoài ra, phương pháp dựa trên kỹ thuật phân tích thống kê còn xem xét đến độ tin cậy của các hệ thống thu thập dữ liệu bằng cách thực hiện tìm kiếm theo kinh nghiệm hoặc nhập thông tin độ tin cậy để xác định khoảng tin cậy hiệu quả nhất của nguồn tín hiệu đầu vào đã thu thập. Trong bài báo này, dữ liệu phụ tải theo chuỗi thời gian trước tiên được xây dựng thành cơ sở 24 giờ và được chuyển đổi thành dữ liệu chênh lệch ngày tới bằng cách tính hiệu theo cặp dữ liệu của hai ngày liên tiếp. Sau đó, việc

tìm kiếm theo kinh nghiệm được áp dụng trên đầu dữ liệu này để xác định mức độ tin cậy hiệu quả nhất từ việc tính toán tất cả các mức độ tin cậy có thể áp dụng đối với STLF. Mục tiêu chính của bài viết là xây dựng một mô hình học thống kê trong quá trình huấn luyện và lấy mức độ tin cậy tốt nhất, dẫn đến độ chính xác cao nhất trong các mô hình dự báo phụ tải. Theo đó, bảng so sánh sai số MAPE (*Mean Absolute Percentage Error* - sai số phần trăm trung bình tuyệt đối) giữa kết quả dự báo có tích hợp phương pháp lọc đề xuất với kết quả dự báo sử dụng các phương pháp lọc hiện nay sẽ được trình bày để chứng minh sự hiệu quả của phương pháp đề xuất.

Mặt khác, thông qua nguồn dữ liệu thời gian thực được thu thập từ hơn 50 hệ thống SCADA tại các trạm trung gian thuộc lưới điện Tp. HCM, cùng với các mô hình dự báo phụ tải ngắn hạn dựa trên phương pháp ANN và ARIMA, các thí nghiệm số sẽ được thực hiện để xác định tính hiệu quả của phương pháp lọc dữ liệu đề xuất. Theo đó, các kết quả mô phỏng chứng minh rằng giá trị phụ tải dự báo được xử lý trước bằng phương pháp lọc đề xuất ở mức độ tin cậy 95% đối với dữ liệu tải của Tp. HCM sẽ vượt trội so với các phương pháp lọc khác thông qua việc quan sát bảng đánh giá sai số phần trăm trung bình tuyệt đối MAPE. Phần còn lại của bài nghiên cứu được tổ chức như sau: Phần 2 trình bày phương pháp lọc dữ liệu dựa trên phương pháp phân tích thống kê để nâng cao chất lượng kết quả dự báo phụ tải ngắn hạn. Tiếp theo, Phần 3 trình bày các số liệu thí nghiệm số, hình ảnh và bảng so sánh, đối chiếu giữa các phương pháp lọc trước đây với phương pháp đề xuất. Sau đó, một số đánh giá và kết luận sẽ được thể hiện trong Phần 4 của bài báo này.

PHƯƠNG PHÁP LỌC DỮ LIỆU DỰA TRÊN PHÂN TÍCH THỐNG KÊ ÁP DỤNG CHO MÔ HÌNH DỰ BÁO PHỤ TẢI NGẮN HẠN

Biến đổi dữ liệu

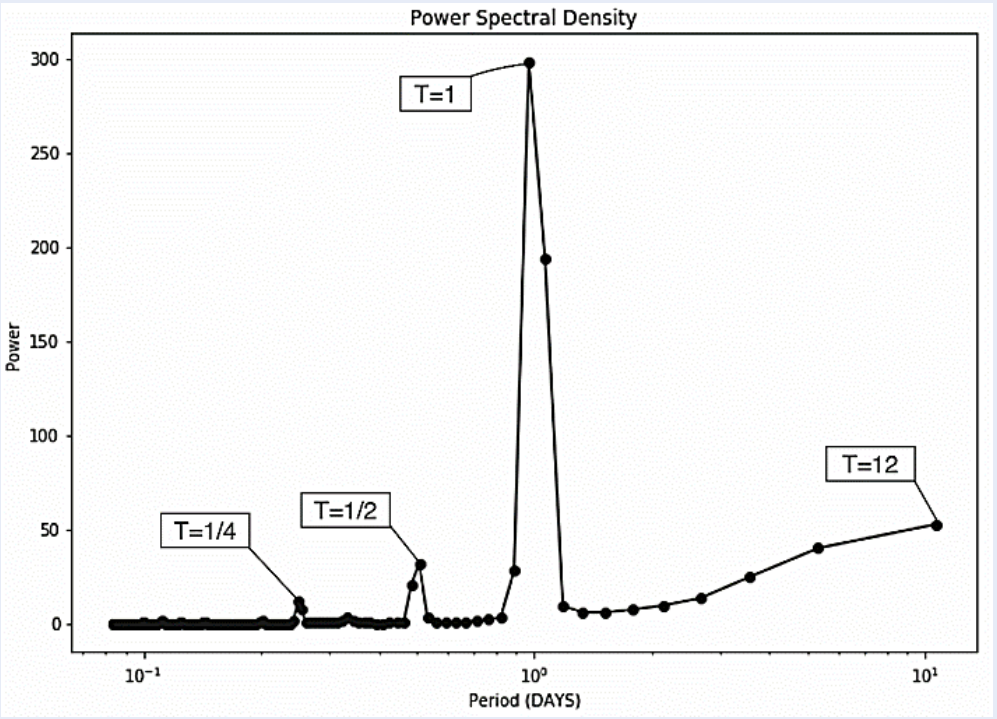
Chuỗi dữ liệu phụ tải là kết quả thể hiện hành vi tiêu thụ năng lượng từ các nhóm phụ tải điện khác nhau. Do đó, dữ liệu phụ tải phải có mẫu tần số đặc trưng và có thể hữu dụng khi dùng để mô tả dữ liệu. Như **Hình 1**, mật độ phổ năng lượng (*Power Spectral Density - PSD*) có thể cung cấp cái nhìn tổng quan về các hành vi của chuỗi phụ tải. Phân phối trong **Hình 1** cho thấy đóng góp chính vào việc mô tả dữ liệu dựa trên nguồn khảo sát chính là đặc trưng "hàng ngày" (T = 1). Số lượng dữ liệu được lấy từ nhiều năm (từ năm 2014 đến năm 2018) để chứng minh mật độ phổ là tin cậy. Trục hoành "Period (DAYS)" thể hiện các điểm dữ liệu trong chuỗi dữ liệu phụ tải T theo giờ qua các ngày, cụ thể, T=12 tương đương 12 ngày nếu dự báo phụ tải ngắn hạn theo tháng, T=1 tương đương 1 ngày và mật độ phổ theo ngày phù hợp nhất với đơn vị là hàng giờ. Hơn nữa, trong **Hình 1**, T =1/2 tương đương nửa ngày và T =1/4 tương đương một phần tư ngày. Điều này cho thấy rằng chuỗi phụ tải nên được chia thành các phần cơ sở 24 giờ để giảm bớt quá trình huấn luyện của các mô hình hồi quy khi thực hiện dự báo ngắn hạn. Các mô hình hồi quy không chỉ học các đặc điểm của từng điểm dữ liệu mà còn học được các mẫu tương đối giống nhau của dữ liệu từ ngày này sang ngày khác. Tuy nhiên, các yếu tố khác như chỉ số giờ, chỉ số ngày, chỉ số tuần, chỉ số tháng, chỉ số quý và chỉ số ngày lễ cũng cần được xem xét và xây dựng trong chuỗi tải (tham khảo ở **Bảng 1**). Trong bảng này, chuỗi phụ tải có kích thước 133 chỉ số phân phối vị trí bao gồm dữ liệu phụ tải và dãy giá trị bit đại diện cho các chỉ số theo thời gian mô tả đầy đủ các đặc điểm của từng đặc trưng trong bộ dữ liệu.

Ổn định dữ liệu

Dữ liệu chuỗi thu thập trong khoảng thời gian dài có thể dẫn đến sự mất ổn định do ảnh hưởng của những thay đổi và xu hướng phát triển. Xu hướng trong chuỗi dài hạn chủ yếu là xu hướng tăng tuyến tính, có thể dễ dàng được loại bỏ bằng các phương pháp đơn giản, cụ thể là các phương pháp ước lượng trung bình động, hồi quy tuyến tính và so lệch. Việc loại bỏ xu hướng dữ liệu nhằm mục đích làm cho chuỗi tải ổn định để đưa vào mô hình dự báo. Phương pháp so lệch là phương pháp đơn giản nhất được định nghĩa từ phương trình (1).

$$\begin{aligned}\mathrm{diff}(d,h0:h23) &= T(d,h0:h23,0)\\ &-T(d-1,h0:h23,0)\end{aligned} \quad (1)$$

Trong đó *diff* là chuỗi dữ liệu so lệch đối với một ngày trước đó, được tính bằng hiệu các giá trị tải tương ứng theo giờ của hai ngày lân cận, ở chỉ số bit thứ *0* của chuỗi phụ tải T với tất cả các điểm dữ liệu 24 giờ, thể hiện từ *h0* đến *h23* (*giờ thứ* 0 *đến giờ thứ 23*); và biến d là chỉ số thứ tự của ngày đang xem xét/khảo sát và *d-1* là chỉ số thứ tự của ngày liền kề trước đó so với ngày đang khảo sát trong bộ dữ liệu. Cần lưu ý rằng, các chỉ số thứ tự ngày được đánh số liên tục đối với tất cả các ngày được khảo sát trong bộ dữ liệu ngõ vào. Như được chỉ ra trong công thức (1), mảng so lệch được xác định bằng cách trừ điểm dữ liệu của ngày hiện tại vào điểm dữ liệu của ngày liền kề trước đó (*theo từng giờ*) trong chuỗi dữ liệu phụ tải đầu vào. Điều này dẫn đến dữ liệu ngày đầu tiên trong bộ dữ liệu được xem là chuỗi dữ liệu so lệch tham chiếu đầu tiên. Hơn nữa, chuỗi dữ liệu so lệch còn thể hiện đặc trưng của phụ tải qua các điểm thời gian theo giờ của các ngày.



Hình 1: **Mật độ phổ năng lượng của chuỗi phụ tải thuộc lưới điện Tp. HCM.**

Bảng 1: Bộ giá trị bit gán cho các chỉ số trong chuỗi phụ tải

Chỉ số	Phân phối vị trí bit	Mô tả
Dữ liệu phụ tải	0	Giá trị tuyệt đối của phụ tải
Chỉ số giờ *(Hour)*	1-24	Biến mã hóa của giờ hiện tại trong một ngày *(24 giờ)*
Chỉ số ngày trong tuần *(Weekday)*	25-31	Biến mã hóa của loại ngày hiện tại trong tuần
Chỉ số ngày trong tháng *(Day)*	32-62	Biến mã hóa của ngày hiện tại trong tháng *(31 ngày)*
Chỉ số tuần *(Week)*	63-115	Biến mã hóa của tuần hiện tại trong năm *(53 tuần)*
Chỉ số tháng *(Month)*	116-127	Biến mã hóa của tháng hiện tại trong năm *(12 tháng)*
Chỉ số Quý *(Quarter)*	128-131	Biến mã hóa của quý hiện tại trong năm *(4 quý)*
Chỉ số ngày lễ *(Holiday)*	132	Biến nhị phân cho biết ngày hiện tại có là ngày nghỉ hay không

Phân tích dữ liệu

Sau khi tiền xử lý chuỗi phụ tải bằng cách biến đổi và loại bỏ xu hướng trong bộ dữ liệu, nhóm tác giả áp dụng tính toán hàm mật độ xác suất đối với chuỗi dữ liệu so lệch một ngày liền kề trước đó, $diff(d, h0 : h23)$ theo biến thời gian giờ. **Hình 2** cho thấy chuỗi dữ liệu so lệch một ngày liền kề trước đó có dạng trực quan của hàm mật độ xác suất (*Probability Density Function - PDF*), với giả định phù hợp rằng chuỗi phụ tải là loại biến ngẫu nhiên liên tục bởi vì đường cong phụ tải của các loại khách hàng khác nhau sẽ là khác nhau và thay đổi ngẫu nhiên theo thời gian mà không theo bất kỳ quy luật sử dụng nào. Hàm mật độ xác suất (PDF) của chuỗi dữ liệu so lệch một ngày nên là hàm phân phối chuẩn Gauss dựa theo kết quả đánh giá histogram và Chi-square goodness of fit test của chuỗi dữ liệu. Khi đánh giá độ tin cậy của chuỗi dữ liệu thì việc thể hiện đặc tính thông qua hàm mật độ

xác suất sẽ giúp đánh giá dễ dàng xác suất của các chuỗi dữ liệu này. Đầu tiên, chúng tôi xem xét đặc tính giới hạn trung tâm, về cơ bản nó sẽ liên quan đến việc ước tính giá trị trung bình của các biến độc lập với bất kỳ phân phối tùy ý nào để tuân theo phân phối Gaussian. Điều này quan trọng bởi vì trong các mẫu trong dữ liệu thực, chúng tôi thấy rằng dữ liệu chuỗi phụ tải thực tế là tổng hợp của nhiều yếu tố cơ bản và cho thấy các tổ hợp tuyến tính của các biến độc lập tạo ra một biến tổng hợp có xu hướng thể hiện phân phối Gaussian. Tuy nhiên, **Hình 2** thể hiện phân phối có hình dạng phân phối fat-tailed (không cân đối) của chuỗi phụ tải so lệch một ngày. Điều đó có nghĩa là phân phối này có nhiều thành phần bên trong và cần được xem xét, đánh giá, phân tích, và phân loại bằng phương pháp PCA (*Principal Components Analysis*) để xác định số lượng thành phần và tách chúng thành các chuỗi phụ tải khác nhau[26]. Phương pháp PCA được dùng dựa trên hàm tính pca (X) trong Matlab, với chuỗi dữ liệu X là chuỗi so lệch có được từ bước ổn định dữ liệu.

Có thể nhận thấy từ **Hình 3**, ba chuỗi phụ tải được tách thành ba phân phối khác nhau. Trong các bảng phân phối này, chỉ có chuỗi chênh lệch phụ tải của các ngày còn lại *(không phải từ Ch*); đặc tính phân phối phi tuyến phức tạp cho thấy dữ liệu chuỗi phụ tải trong các ngày này thể hiện các biến độc lập theo giờ với độ tương quan không đồng nhất. Do vậy chúng tôi cần xem xét phân tách dữ liệu chuỗi phụ tải theo các nhóm giờ với độ tương quan phù hợp.

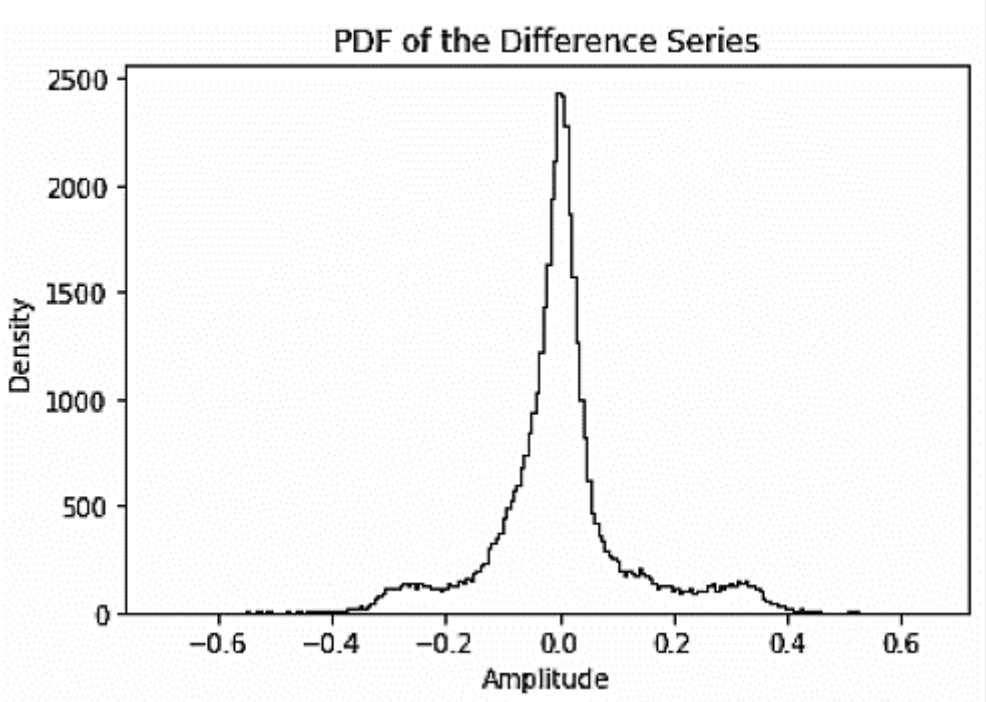
Hơn nữa, để xác định các mẫu ẩn được phân phối giữa các tải hàng giờ trong các ngày thể hiện ở **Hình 4b, c**, chúng tôi thực hiện phương pháp sơ đồ cây - **dendrogram** trong toàn bộ chuỗi để phân tích mối quan hệ giữa các giờ và phân cụm chúng thành các nhóm giờ có độ tương quan cao, như được minh họa trong **Hình 5**.

Trong **Hình 5**, ba bộ tải hàng giờ có khoảng cách gần nhất được nhóm lại với nhau trong cả hai chuỗi tải gồm : đối với nhóm tải từ Thứ Bảy đến Chủ Nhật [0,1,2,3,4] giờ, [5,6,16,17,18,19,20,21,22,23] giờ và [7,8,9,10,11,12,13,14,15] giờ; đối với nhóm tải từ Chủ Nhật đến Thứ Hai [0,1,2,3,4] giờ, [5,6,17,18,19,20,21,22,23] giờ và [7,8,9,10,11,12,13,14,15,16] giờ. Để đảm bảo tính liên tục về thời gian trong mỗi bộ dữ liệu, bộ tải [5,6] giờ sẽ được chia thành bộ thứ tư trong cả hai nhóm phụ tải. Sự tách biệt theo các nhóm giờ tương quan này đóng vai trò quan trọng trong phương pháp lọc thống kê đề xuất bởi nhóm tác giả. Dựa vào phân phối xác xuất thể hiện đặc trưng của các biến phụ tải theo thời gian như trên, thì theo mức tin cậy dữ liệu được chọn, dữ liệu sẽ được tính độ tin cậy tương ứng bằng việc áp dụng độ tương quan với các phân phối theo nhóm ngày và nhóm giờ như đã phân tích. **Hình 6** cho thấy các bảng phân phối chuỗi kết quả của các nhóm tải theo giờ không theo phân phối chuẩn Gauss.

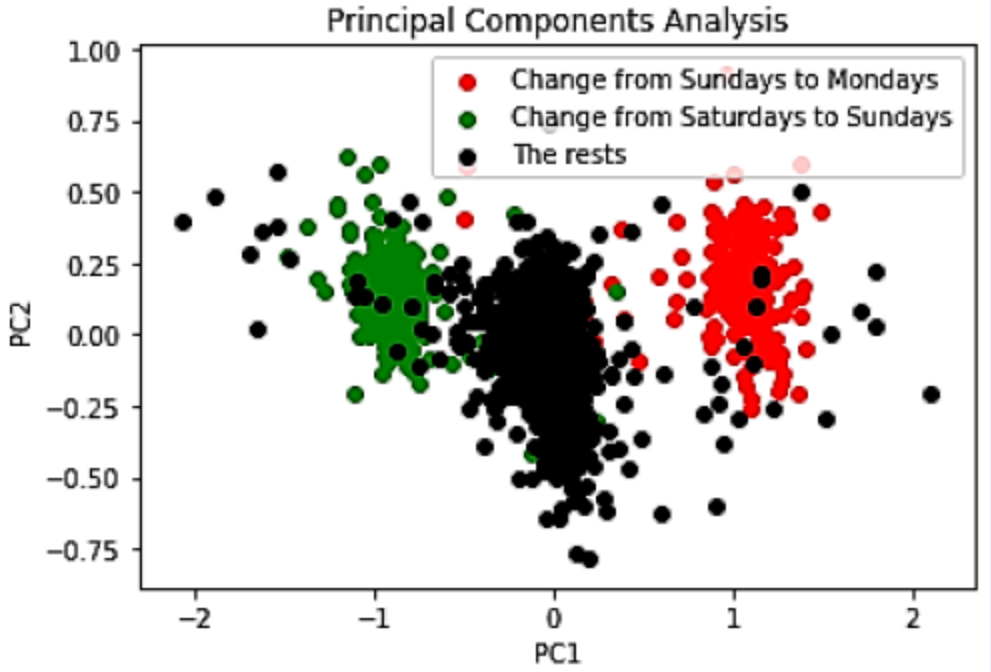
Phương pháp lọc dữ liệu dựa trên phân tích thống kê để phục vụ công tác dự báo ngắn hạn

Sau các quá trình tiền xử lý tập dữ liệu bằng cách phân cụm các điểm dữ liệu thành ba nhóm phụ tải khác nhau: *(i)* bộ dữ liệu chênh lệch công suất tải theo giờ tương ứng từ Thứ Hai đến Thứ Bảy; *(ii)* bộ dữ liệu chênh lệch công suất tải theo giờ tương ứng từ Thứ Bảy đến Chủ Nhật và *(iii)* bộ dữ liệu chênh lệch công suất tải theo giờ tương ứng từ Chủ Nhật đến Thứ Hai, và có kết hợp phân tích những đặc trưng của phụ tải theo giờ, nhóm tác giả sẽ để cập đến phương pháp lọc dữ liệu dựa trên phân tích thống kê để phục vụ công tác dự báo ngắn hạn trong mục này. Mặc dù, độ chính xác của các giá trị dữ liệu thu thập từ các hệ thống SCADA trạm đã được cải thiện tốt nhưng việc xuất hiện các lỗi ngẫu nhiên là không thể tránh khỏi. Hơn nữa, các giá trị dữ liệu tải trong bộ dữ liệu được thu thập từ nhiều điểm đo khác nhau - vốn chưa qua xử lý, quy đổi, hiệu chỉnh về một hệ quy chiếu - có thể dẫn đến việc đánh giá độ tin cậy ở cấp độ hệ thống là không chính xác. Dữ liệu thực từ hệ thống SCADA được sử dụng cho quá trình dự báo phụ tải, như trong bài báo này chúng tôi áp dụng hai mô hình dự báo gồm ANN và ARIMA. Trong nghiên cứu này, dữ liệu SCADA được lấy mẫu định kỳ theo giờ. Khi đó, những dữ liệu có chứa các lỗi ngẫu nhiên, các số liệu nhiễu này trở thành các yếu tố trong tính toán trọng số hay hàm phức hợp trong giải thuật dùng ANN hoặc ARIMA. Do vậy, việc dữ liệu không tin cậy và chứa các lỗi ngẫu nhiên sẽ không chỉ làm sai lệch kết quả dự báo, mà các lỗi ngẫu nhiên này có thể tác động không tốt theo bội số đến kết quả tính toán cuối cùng.

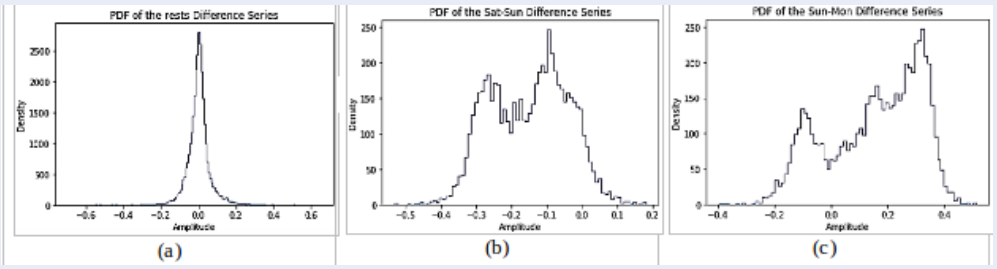
Để khắc phục các vấn đề trên, phương pháp lọc dữ liệu để xuất sẽ được thực hiện trên nhiều mức độ tin cậy trước khi lựa chọn một chỉ số độ tin cậy đại diện cho toàn bộ dữ liệu nguồn của hệ thống. Dữ liệu thu thập từ hệ thống SCADA sẽ được tính toán theo các bước như thể hiện ở **Hình 7**. Dữ liệu đầu vào sau khi được biến đổi thành các chuỗi phụ tải như trình bày ở mục **Biến đổi dữ liệu**; dữ liệu sẽ được loại bỏ tính xu hướng tuyến tính bằng phương pháp ổn định ở mục **Ổn định dữ liệu**. Sau đó, dữ liệu sẽ được tính toán hàm mật độ xác suất (PDF) để đánh giá dạng dữ liệu có phân phối chuẩn hoặc phân tích Dendrogram, tham khảo mục



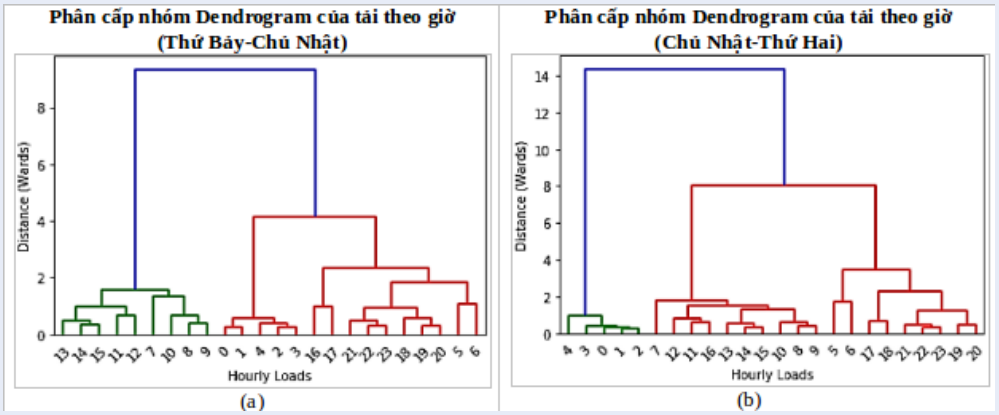
Hình 2: **Biểu đồ PDF của chuỗi phụ tải so lệch một ngày.**



Hình 3: **Biểu đồ PCA thể hiện sự phân cụm của ba chuỗiphụ tải khác nhau.**



Hình 4: **PDF của ba chuỗi chênh lệch phụ tải: (a) thay đổi công suất tải giữa các ngày trong tuần từ Thứ Hai đến Thứ Bảy; (b) từ Thứ Bảy đến Chủ Nhật; và (c) từ Chủ Nhật đến Thứ Hai.**



Hình 5: **Phân tích Dendrogram về tải hàng giờ trong hai nhóm phụ tải: (a) chuỗi chênh lệch công suất tải theo giờ từ Thứ Bảy đến Chủ Nhật và (b) từ Chủ Nhật đến Thứ Hai**

Phân tích dữ liệu. Tiếp theo, các dữ liệu có hàm phân phối chuẩn sẽ được đưa vào thuật toán ANN để có các kết quả dự báo S_{pred}, sau đó chỉ số MAPE được tính dựa vào S_s và S_{pred} như ở biểu thức (2). Giá trị của chỉ số sai số phần trăm trung bình tuyệt đối (MAPE) sẽ cho biết dữ liệu với độ tin cậy nào là tốt nhất.

Như vậy, mục tiêu của phương pháp lọc dữ liệu dựa trên phân tích thống kê là nhằm tìm kiếm mức độ tin cậy tốt nhất nhưng vẫn đảm bảo giá trị sai số MAPE là thấp nhất nhờ vào việc tính toán biểu thức (2) sau khi giá trị dự báo của ngày tới được tính ra từ mô hình toán học ANN. Giá trị sai số MAPE được tính toán ứng với các trường hợp mức độ tin cậy khác nhau, để có thể được áp dụng cho phương pháp lọc dữ liệu mang tính thống kê như đã đề cập. Mục đích của tính toán MAPE sẽ giúp chọn ra trường hợp mức độ tin cậy tốt nhất với giá trị MAPE là thấp nhất. Công thức tính toán MAPE giữa hai giá trị S_s và S_{pred} như sau:

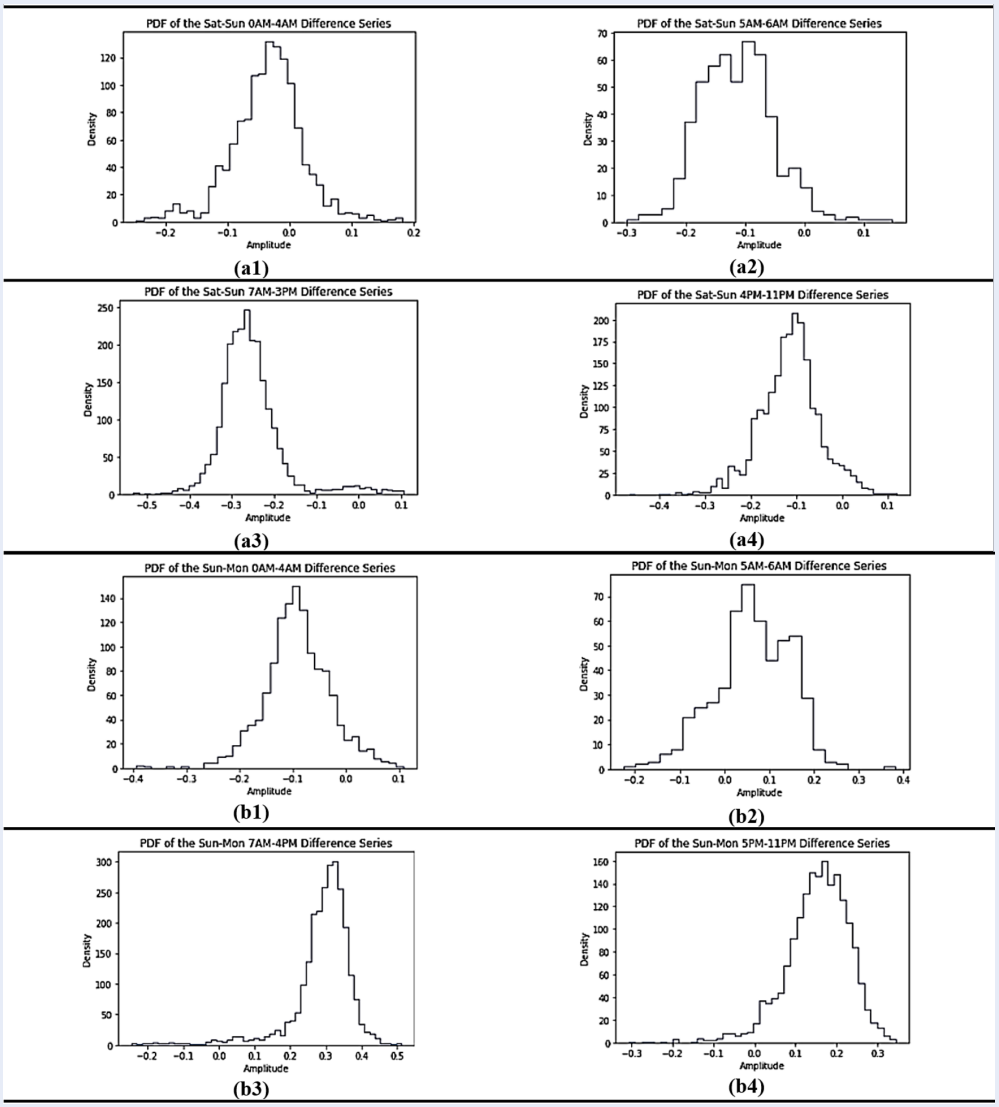
$$MAPE = \frac{1}{N-1}\int^{N-1}\frac{S_{pred} - S_S}{S_S} \qquad (2)$$

Trong đó: S_s là giá trị gốc thu thập được và S_{pred} là giá trị dự báo sau khi sử dụng mô hình toán ANN; N là số chu kỳ tính toán giá trị dự báo, trong bài báo này tương ứng là số lần dự báo theo giờ tới liên tiếp.

Độ tin cậy của bộ dữ liệu phụ tải được giả định là lớn hơn 90% vì các hệ thống SCADA hầu hết có độ chính xác cao. Điều này cho phép một phạm vi tin cậy nhất định gồm mười ba mức độ từ 90% đến 99%, 4,5-sigma *(~99,73%)*, 5,5-sigma *(~99,9937%)* và 6-sigma *(~99,99966%)*. Theo đó, các mô hình kết quả sẽ được đánh giá bằng một bộ dữ liệu thực để tìm ra mức độ tin cậy tốt nhất bằng cách lựa chọn độ chính xác cao nhất.

Các mô hình dự báo thông thường

Một điểm cần lưu ý rằng bài báo này tập trung vào phát triển phương pháp lọc dữ liệu cho STLF, nên chỉ sử dụng các phương pháp/mô hình dự báo phụ tải thông thường, như mô hình ANN và ARIMA để chứng minh tính hiệu quả của phương pháp lọc đề xuất.



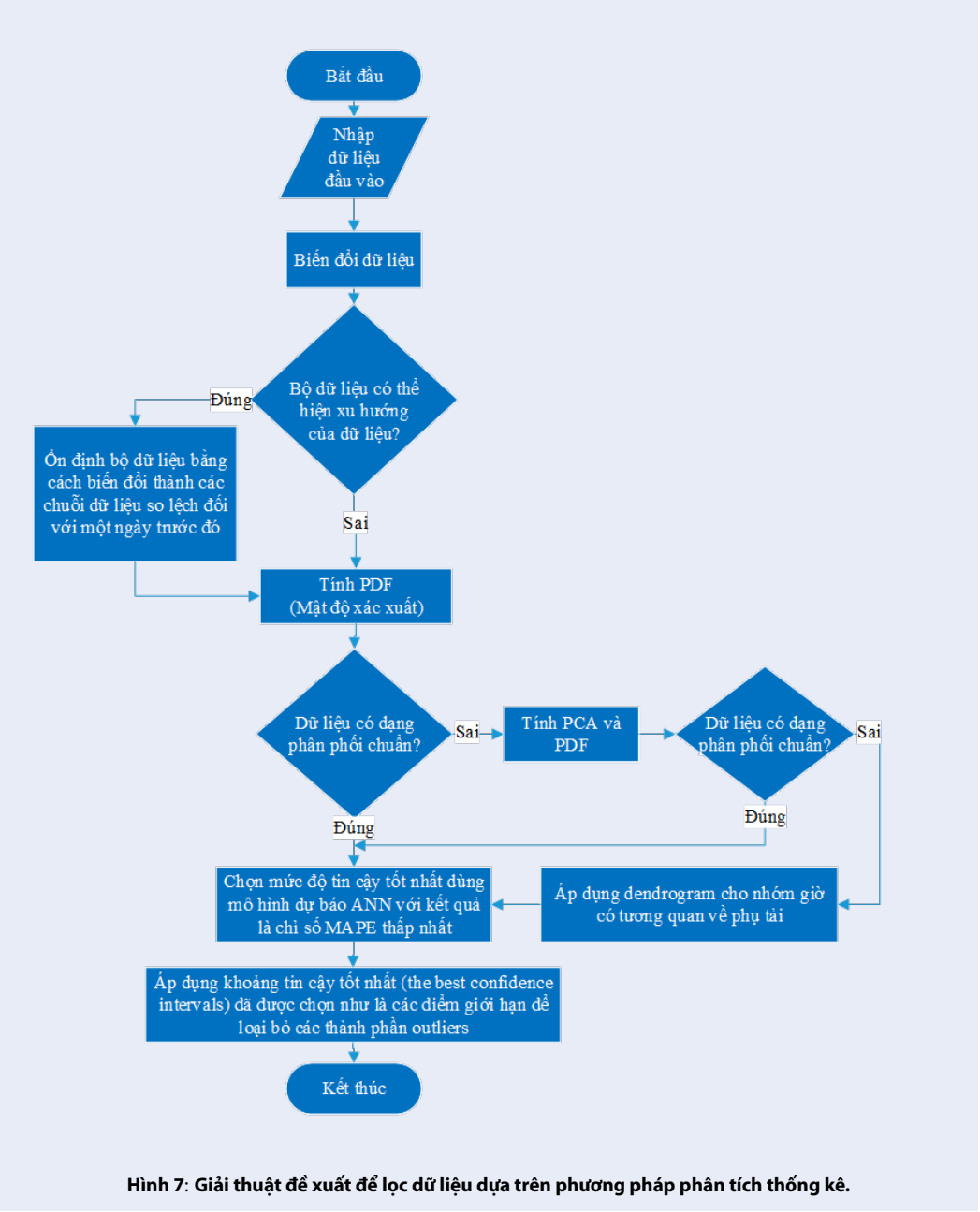
Hình 6: **Phân phối của các chuỗi riêng biệt tương ứng theo các khung giờ: (a1) 0AM đến 4AM, (a2) 5AM đến 6AM, (a3) 7AM đến 3PM, (a4) 4PM đến 11PM từ Thứ Bảy đến Chủ Nhật; và (b1) 0AM đến 4AM, (b2) 5AM đến 6AM, (b3) 7AM đến 4PM, (b4) 5PM đến 11PM từ Chủ nhật đến thứ Hai.**

Mạng nơ-ron nhân tạo - Artificial Neural Network (ANN)

a) Kiến trúc của một mạng nơ-ron nhân tạo

Kiến trúc ANN cơ bản được triển khai trong bài báo này bao gồm một lớp đầu vào với kích thước phù hợp với dạng dữ liệu đầu vào, hai lớp ẩn với 100 nút nơ-ron trên mỗi lớp và một lớp đầu ra có cùng kích thước với lớp đầu vào, được thể hiện trong **Hình 8**. Như được mô tả trong **Bảng 2**, mô hình lấy nguồn dữ liệu tải đầu vào tương ứng với số giờ trong tập dữ liệu đầu vào, trong đó mỗi giờ chứa 133 đặc tính đã được liệt kê trong **Bảng 1**. Kết quả là các giá trị dự báo tương ứng với số nguồn dữ liệu đầu vào được ánh xạ theo giờ tương ứng của ngày hôm sau.

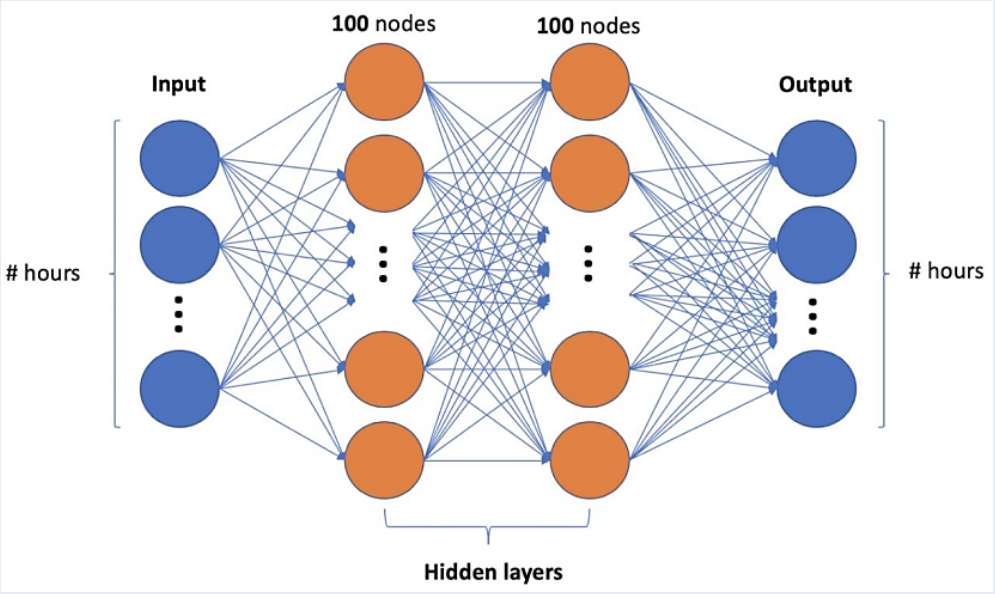
Theo **Bảng 2** mô tả, lớp ngõ vào có số lượng nút nơ-ron là số lượng giờ. Dữ liệu vào theo chuỗi tải T được chọn ở mục **Biến đổi dữ liệu**. Trong nghiên cứu này, số lượng giờ liên tiếp ký hiệu là #hours và được biến đổi thành chuỗi tải T với các chỉ số thông tin khác có kích thước 133; đầu ra của lớp này có kích thước 100. Lớp ẩn 1 và lớp ẩn 2 có số lượng 100 nơ-ron đối với kích thước ngõ vào và ngõ ra là 100. Lớp ngõ ra có số lượng nơ-ron tương ứng là số giờ ở ngõ vào, kích thước ngõ ra theo mảng 100 từ kết quả lớp ẩn, và kích thước ngõ ra là mảng giá trị dự báo theo số lượng



Hình 7: **Giải thuật đề xuất để lọc dữ liệu dựa trên phương pháp phân tích thống kê.**

Bảng 2: Số lượng nút và kích thước của ANN

Cấu trúc mạng nơ-ron	Lớp ngõ vào (Input layer)	Lớp ẩn 1 (Hidden layer 1)	Lớp ẩn 2 (Hidden layer 2)	Lớp ngõ ra (Output layer)
Số lượng nút nơ-ron	#hours	100	100	#hours
Kích thước ngõ vào mỗi lớp	(#hours, 133)	(#hours, 100)	(#hours, 100)	(#hours, 100)
Kích thước ngõ ra mỗi lớp	(#hours, 100)	(#hours, 100)	(#hours, 100)	(#hours, 1)



Hình 8: Kiến trúc của ANN đơn giản.

#hours.

b) Huấn luyện một mạng nơ-ron nhân tạo

Theo kiến trúc ANN, một vectơ kết quả của lớp đầu ra được hình thành bởi các khuôn mẫu của dữ liệu đầu vào cùng với các giá trị mục tiêu trong mạng nơ-ron. Nói chung, trọng số mạng W_{ij} trong liên kết giữa mỗi cặp nút mạng sẽ được cập nhật để thể hiện tính liên quan, sự khác biệt giữa các đầu ra được tạo và ngõ ra mong muốn dựa vào dữ liệu sai số tính toán *MAE*, như được minh họa bằng công thức **(3)**

$$MAE = \frac{1}{N}\left|y_{\text{actual}} - y_{\text{predict}}\right| \quad (3)$$

Sai số MAE được xem là hàm trung bình mục tiêu. Nghiên cứu này xem xét tối thiểu sai lệch giữa ngõ ra chương trình dự báo với giá trị thực tế so sánh. Sai số lớp ngõ ra này được truyền ngược qua tất cả các lớp ẩn sang lớp đầu vào bằng cách lấy đạo hàm hàm suy hao[27] và tính toán các trọng số dựa trên trạng thái nơ-ron của chúng. Ngoài ra, để nhận ra tính phi tuyến trong bộ dữ liệu, các hàm kích hoạt được sử dụng sau mỗi lớp, sao cho các chức năng của bộ điều chỉnh tuyến tính đơn vị (*Rectifier-Linear-Unit - RELU*), như thể hiện trong công thức **(4)**, và bộ thuần tuyến tính (*Pure-Linear*), như được hiển thị ở công thức **(5)**, được áp dụng cho các lớp ẩn và lớp đầu ra, tương ứng.

$$R(z) = \max(0, z) \quad (4)$$

$$P(z) = z \quad (5)$$

Trong đó, z là trạng thái đầu ra của một lớp.

Phương pháp ARIMA - Autoregressive Integrated Moving Average

Phương pháp ARIMA là một thuật toán dự báo dựa trên chuỗi thời gian, thường được sử dụng để giải quyết nhiều vấn đề dự báo do cấu trúc toán học đơn giản của nó và chỉ cần quan sát lịch sử[28]. Phương trình tổng quát của công thức ARIMA được trình bày như sau:

$$D^D X_t = (1 - B)^D X_t \quad (6)$$

Trong đó, D là thứ tự so lệch và thường có giá trị 1 hoặc 2, và B là toán tử backshift.

KẾT QUẢ MÔ PHỎNG VÀ NHỮNG THẢO LUẬN

Trong phần này, nhóm tác giả sẽ sử dụng bộ dữ liệu phụ tải của lưới điện Tp.HCM để thử nghiệm phương pháp lọc dữ liệu dựa trên phân tích thống kê được đề xuất.

Giới thiệu bộ dữ liệu thu thập thực tế

Từ cuối năm 2009, dữ liệu phụ tải điện của lưới điện Tp. HCM bắt đầu được thu thập đầy đủ và định kỳ theo từng giờ. Tuy nhiên, công việc này được thực

hiện một cách thủ công và lỗi thu thập dữ liệu cũng như sự chậm trễ là điều không thể tránh khỏi. Cho đến năm 2014, bộ dữ liệu thời gian thực của lưới điện này đã được cải thiện về độ tin cậy thông qua việc thu thập từ hệ thống SCADA của hơn 50 trạm trung gian. Phương pháp lọc đề xuất sẽ được áp dụng để tính toán mức độ tin cậy của tất cả các phân phối dựa vào bộ dữ liệu phụ tải đã thu thập nêu trên. Theo đó, các mô phỏng thí nghiệm sẽ được thực hiện để minh họa tính khả thi của hai phương pháp tính toán và để ước lượng tất cả mức độ tin cậy của bộ dữ liệu. Hai phương pháp tính toán bao gồm: i) lấy trung bình tất cả các mức độ tin cậy đã tính toán có sai số MAPE thấp nhất hoặc ii) lấy trung bình tất cả các sai số MAPE và chọn giá trị thấp nhất.

Thử nghiệm và kết quả

Các thử nghiệm được thực hiện trên tất cả các bảng phân phối. Trong đó, dữ liệu được lọc theo nhiều mức độ tin cậy và nhập vào cùng một mô hình ANN đơn giản *(mô tả qua **Hình 8**)* để được huấn luyện và đánh giá bằng chỉ số MAPE giữa các giá trị so lệch trước một ngày thực tế và những giá trị dự báo. Kết quả sai số MAPE tính toán được hiển thị trong **Bảng 3** thể hiện từng độ tin cậy của 9 phân phối liệt kê trong **Hình 4a** và **Hình 6**. Hệ số tin cậy của dữ liệu có thể được ước tính theo hai cách tính toán:

- Tìm kiếm mức độ tin cậy tương ứng với giá trị MAPE thấp nhất cho mỗi phân phối, lấy trung bình các mức độ tin cậy của tất cả các phân phối để chọn mức độ tin cậy phù hợp và áp dụng cho toàn hệ thống;
- Lấy giá trị trung bình của tất cả giá trị MAPE trong từng mức độ tin cậy và tìm kiếm mức độ thấp nhất trong số các giá trị trung bình;

Việc sử dụng hai phương pháp này là để i) tạo điều kiện thuận lợi trong việc đánh giá độ chính xác của phương pháp lọc dữ liệu đề xuất và ii) để có được mức độ tin cậy phù hợp nhất của toàn hệ thống.

Mỗi phân phối trong **Bảng 3** có giá trị sai số MAPE tương ứng với các mức độ tin cậy cụ thể. Có thể thấy rằng, các giá trị MAPE tối thiểu ở mức 99,73%, 97%, 90%, 90%, 99%, 96%, 99%, 95% và 92% cho tất cả các bảng phân phối, lần lượt như sau : Thứ Ba- Thứ Bảy-0AM-11PM, Chủ Nhật-0AM-4AM, Chủ Nhật-5AM-6AM, Chủ Nhật-7AM-3PM, Chủ Nhật-4PM-11PM, Thứ Hai-0AM-4AM, Thứ Hai-5AM-6AM, Thứ Hai-7AM-4PM và Thứ Hai-5PM-11PM. Áp dụng phương pháp tính toán đầu tiên, mức độ tin cậy **95,303%** được xác định là giá trị đáng tin cậy và có thể đại diện cho toàn hệ thống. Hàng " Trung bình " trong **Bảng 3** cho thấy kết quả của cách tính toán thứ hai là **5,69%** ở mức độ tin cậy **95%**. Điều đáng chú ý là kết quả của hai phương pháp tính toán là nhất quán. Do đó, mức độ tin cậy **95%** của bộ dữ liệu phụ tải theo thời gian thực sẽ được sử dụng để so sánh độ hiệu quả giữa phương pháp lọc đề xuất với các phương pháp lọc đã đề cập trong **Phần tổng quan.**

Trong **Bảng 4** và **Bảng 5**, phương pháp lọc dữ liệu đề xuất cho STLF trong lưới phân phối vượt trội so với các phương pháp khác sau khi chạy hai mô hình dự báo phổ biến là ANN và ARIMA. **Bảng 4** chỉ ra các MAPE (%) của ANN khi không có bộ lọc hoặc khi áp dụng các phương pháp lọc khác trên bộ dữ liệu có độ tin cậy 95% như Kalman, DBSCAN, DWT và SSA. Đối với cả ba nhóm dữ liệu của bộ dữ liệu phụ tải sau phân tách, cụ thể là Chủ nhật, Thứ Hai và các ngày còn lại trong tuần, kết quả của mô hình dự báo ANN cho thấy rằng các giá trị sai số MAPE của phương pháp lọc dữ liệu đề xuất là thấp nhất so với các phương pháp lọc khác (6,02% cho nhóm phụ tải Chủ Nhật, 5,87% cho nhóm phụ tải Thứ Hai và 5,87% cho nhóm phụ tải còn lại). Cần lưu ý rằng, bài báo này tập trung vào việc xây dựng phương pháp lọc dữ liệu mới nên các giá trị sai số tính toán MAPE tương đối cao. Tương tự, **Bảng 5** cho thấy các giá trị sai số MAPE của mô hình dự báo ARIMA sử dụng phương pháp lọc dữ liệu đề xuất cao hơn việc sử dụng các bộ lọc khác (9,89% cho nhóm phụ tải Chủ Nhật, 15,32% cho nhóm phụ tải Thứ Hai và 4,9% cho nhóm phụ tải còn lại) khi cùng dựa trên bộ dữ liệu với mức tin cậy là 95%.

Hình 9 thể hiện biểu đồ dữ liệu phụ tải thực tế so với dữ liệu phụ tải dự báo từ hai mô hình ANN và ARIMA sử dụng phương pháp lọc dữ liệu đề xuất với ba nhóm phụ tải khác nhau. Chi tiết hơn, **Hình 9**a thể hiện cho nhóm chênh lệch phụ tải tương ứng theo giờ Thứ Bảy-Chủ Nhật, **Hình 9**b cho nhóm chênh lệch phụ tải tương ứng theo giờ Chủ Nhật-Thứ Hai và Hình 9c cho nhóm chênh lệch phụ tải tương ứng theo giờ của các ngày liền kề còn lại trong tuần. Các mô hình dự báo được huấn luyện với bộ dữ liệu trong khoảng thời gian từ ngày 01 tháng 01 năm 2014 đến ngày 01 tháng 11 năm 2018 trước khi được thử nghiệm, đánh giá và so sánh với bộ dữ liệu được thu thập trong hai tháng tiếp theo. Như minh họa trên **Hình 9** là kết quả dự báo từ hai mô hình ANN và ARIMA so với dữ liệu thực tế thu thập trong một tuần tiếp theo, tương ứng chuỗi dữ liệu 168 (=24x7) giờ. Mô hình dự báo thực hiện chạy dự báo theo chu kỳ cho kết quả liên tục với số giờ tương đương hai tháng tiếp theo. Có thể thấy rằng từ **Hình 9**a và **Hình 9**b: i) mức độ tin cậy 95% của tập dữ liệu dự báo đường cong tải từ mô hình ANN gần với đường cong tải thực tế hơn so với đường cong từ mô hình ARIMA; ii) khi nhóm phụ tải thô không

Bảng 3: MAPE (%) của 09 phân phối khác nhau với phạm vi độ tin cậy cho phép

Phân phối	90%	91%	92%	93%	94%	95%	96%	97%	98%	99%	99,73%	99,9936%	99,99932%
Thứ Ba – Thứ Bảy 0AM-11PM	3,78	3,77	3,72	3,67	3,92	3,73	3,63	3,75	3,84	4,10	3,56	3,85	3,83
Chủ Nhật 0AM-4AM	5,58	5,25	5,02	5,27	5,40	4,94	4,96	4,71	4,89	5,17	4,91	5,17	5,17
Chủ Nhật 5AM-6AM	5,90	7,12	7,00	6,50	6,66	6,38	7,05	6,41	6,17	7,43	6,11	7,65	7,42
Chủ Nhật 7AM-3PM	5,15	5,96	6,07	6,00	6,55	5,71	5,83	5,93	5,68	6,06	5,81	6,49	7,21
Chủ Nhật 4PM-11PM	5,37	5,34	5,97	5,56	5,26	5,23	5,38	6,17	5,6	5,15	5,20	5,20	5,20
Thứ Hai 0AM-4AM	5,64	5,36	5,47	5,49	6,03	5,50	5,18	5,69	5,21	5,43	5,65	5,81	5,81
Thứ Hai 5AM-6AM	9,44	8,74	10,6	9,70	9,24	9,36	10,36	9,47	10,9	8,57	9,36	10,96	10,96
Thứ Hai 7AM-4PM	5,94	5,65	6,39	6,18	6,29	5,47	6,14	6,00	5,99	6,45	6,73	7,17	6,07
Thứ Hai 5PM-11PM	5,31	4,98	4,64	4,96	4,92	4,88	5,55	4,77	5,49	4,82	4,67	5,25	4,85
Trung bình	5,79	5,80	6,10	5,93	6,03	**5,69**	6,02	5,88	5,98	5,91	6,00	6,39	6,28

Bảng 4: Sai số tính toán MAPE (%) của mô hình ANN khi không có hoặc áp dụng phương pháp lọc khác nhau

	Không lọc	Bộ lọc đề xuất với mức độ tin cậy 95%	Bộ lọc Kalman	Bộ lọc DBSCAN	Bộ lọc DWT (loại bỏ tín hiệu tần số cao)	Bộ lọc SSA
Chủ Nhật	7,92	6,02	6,45	7,72	6,88	7,44
Thứ Hai	9,65	5,87	8,22	8,45	9,14	9,00
Những ngày còn lại	7,06	5,87	6,72	6,87	7,00	6,70

Bảng 5: Sai số tính toán MAPE (%) của mô hình ARIMA khi không có hoặc áp dụng phương pháp lọc khác nhau

	Không lọc	Bộ lọc đề xuất với mức độ tin cậy 95%	Bộ lọc Kalman	Bộ lọc DBSCAN	Bộ lọc DWT (loại bỏ tín hiệu tần số cao)	Bộ lọc SSA
Chủ Nhật	12,83	9,89	17,47	12,65	16,73	12,92
Thứ Hai	24	15,32	24,21	24,22	24,75	24,15
Những ngày còn lại	21,76	4,9	19,99	21,76	20,00	20,84

tuân theo phân phối chuẩn nhưng được hình thành từ các phân phối chuẩn con (như được minh họa trong **Hình 6**) thì phương pháp lọc dữ liệu để xuất hiệu quả hơn trong việc cải thiện mô hình dự báo phụ tải từ mô hình ANN so với mô hình ARIMA.

Hình 9c giải thích rằng các mô hình ANN và ARIMA có thể cho kết quả dự báo phụ tải chính xác với việc sử dụng phương pháp lọc dữ liệu để xuất qua quan sát các đường cong tải đều bám sát với đường cong phụ tải thực tế. Tuy nhiên, khi xem xét kỹ lưỡng hơn, giá trị MAPE khi sử dụng mô hình ARIMA nhỏ hơn so với mô hình ANN, tương ứng 4,9% so với 5,87%. Điều này cho thấy nếu dữ liệu thô có phân phối chuẩn như trong **Hình 4**a, với phương pháp lọc dữ liệu đề xuất, mô hình dự báo tải chuỗi thời gian như ARIMA hoạt động tốt hơn mô hình học máy như ANN.

ĐÁNH GIÁ VÀ KẾT LUẬN

Một phương pháp lọc dữ liệu dựa trên phương pháp phân tích thống kê để phục vụ cho các mô hình dự báo phụ tải ngắn hạn đã được giới thiệu trong bài báo này. Phương pháp này tìm kiếm mức độ tin cậy phù hợp nhất của bộ dữ liệu phụ tải thô bằng cách tính toán tất cả các mức độ tin cậy có thể. Do đó, dữ liệu phụ tải sau khi lọc sẽ đáng tin cậy hơn, thể hiện rõ đặc trưng của dữ liệu thu thập, và có thể đưa vào các mô hình dự báo phụ tải điện hiện nay. Kết quả của bài báo này đã cho thấy rằng:

i) Với việc đưa ra các mức độ tin cậy thông qua phương pháp để xuất, độ chính xác của dữ liệu dự báo phụ tải dựa trên hai phương pháp ANN và ARIMA được cải thiện rõ rệt, và vượt trội so với khi áp dụng các thuật toán lọc khác như Kalman, DBSCAN, DWT và SSA;

ii) Trong trường hợp dữ liệu của một nhóm phụ tải có dạng phân phối chuẩn, mô hình ARIMA cho kết quả tốt hơn so với mô hình ANN khi sử dụng phương pháp lọc dữ liệu để xuất tại cùng mức độ tin cậy;

iii) Mặt khác, phương pháp phân cụm PCA có thể được áp dụng để phân chia dữ liệu tải thô thành các nhóm phụ tải con có dạng phân phối chuẩn trước khi áp dụng phương pháp lọc để xuất cho STLF; trong trường hợp đó, nhóm tác giả khuyến nghị sử dụng mô hình dự báo dựa trên mạng nơ-ron;

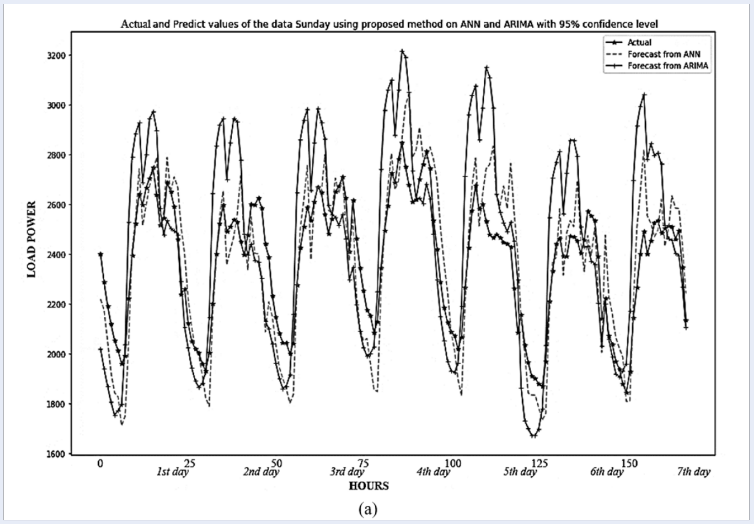
iv) Nếu số liệu bất thường trong thời gian dài liên tục trong chuỗi dữ liệu đầu vào, giả sử dữ liệu này xem xét là biến ngẫu nhiên, thì khi đó việc áp dụng phân phối chuẩn gauss theo phương pháp lọc thống kê để xuất sẽ hạn chế các tác động của số liệu bất thường này; và

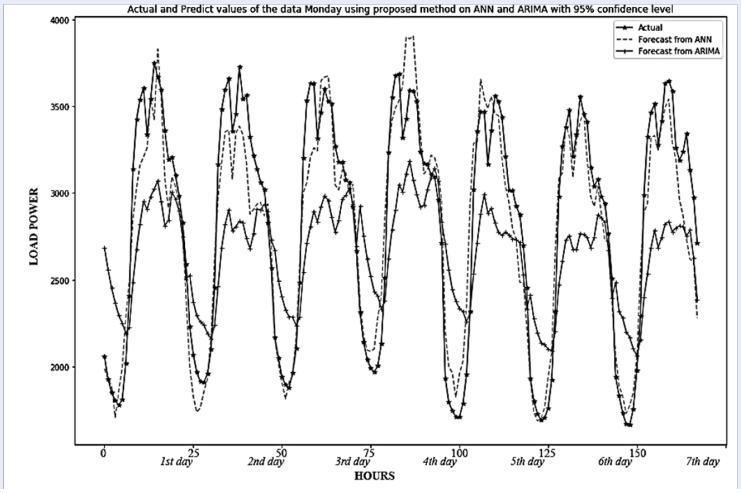
v) Trong trường hợp, số liệu bất thường lại không ngẫu nhiên mà thể hiện tính xu hướng hoặc dữ liệu hằng số như trong trường hợp mất tín hiệu trong thời gian dài thì phương pháp lọc thống kê để xuất sẽ cần cải tiến để nhận biết các trường hợp này.

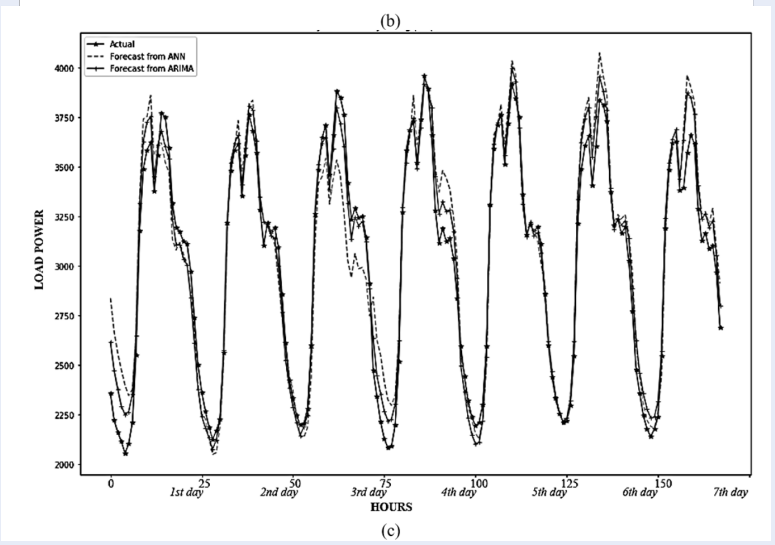
Cuối cùng, phương pháp lọc để xuất của nhóm tác giả có thể ứng dụng trong công tác STLF trên lưới điện Tp.HCM hoặc lưới điện có hiện tượng nhiễu động với mật độ dày đặc cũng như tại khu vực có sự đa dạng về phụ tải điện, khiến cho độ tin cậy của nguồn dữ liệu không cao.

DANH MỤC CÁC TỪ VIẾT TẮT:

RES: nguồn năng lượng tái tạo – Renewable Energy Source.







Hình 9: **Kết quả dự báo phụ tải của các mô hình ANN và ARIMA theo ba nhóm phụ tải; (a) nhóm chênh lệch-phụ tải tương ứng theo giờ Thứ Bảy-Chủ Nhật; (b) nhóm chênh lệch phụ tải tương ứng theo giờ ChủNhật-Thứ Hai; (c) Nhóm chênh lệch phụ tải tương ứng theo giờ của những ngày liền kề còn lại trong tuần.**

ANN: phương pháp mạng thần kinh nhân tạo – Artificial Neural Network.

ARIMA : phương pháp tự hồi qui tích hợp trung bình trượt – Autoregressive Integrated Moving Average.

MAPE: sai số phần trăm tuyệt đối trung bình - Mean Absolute Percentage Error

SCADA : hệ thống điều khiển giám sát và thu thập dữ liệu - Supervisory Control And Data Acquisition

STLF: dự báo phụ tải ngắn hạn - Short-time Load Forecasting

RBF: hàm cơ sở bán kính - Radial Basis Function

ELM: máy học cực kết hợp - Ensembled Extreme Learning Machines

KNN : hệ số K - K-nearest-neighbor

CNN: mạng nơ-ron chuyển đổi - Convolutional Neural Network

LSTM: Long-Short Term Memory

DW: bộ lọc rời rạc dựa trên biến đổi Wavelet - Discrete Wavelet-transform

SSA: phân tích phổ đơn - Singular Spectrum Analysis

VSTLF: dự báo tải siêu ngắn hạn - Very Short-term Load Forecasting

DBSCAN: phân cụm không gian dựa trên mật độ của các đối tượng có nhiễu - Density-Based Spatial Clustering of Applications with Noise

PSD: mật độ phổ năng lượng - Power Spectral Density

PDF: hàm mật độ xác suất - Probability Density Function

PCA: phương pháp Phân tích thành phần chính - Principal Components Analysis

MAE: sai số trung bình tuyệt đối - Mean Absolute Error

RELU: bộ chỉnh tuyến tính đơn vị - Rectifier- Linear-Unit

XUNG ĐỘT LỢI ÍCH

Nhóm tác giả xin cam đoan rằng không có bất kỳ xung đột lợi ích nào trong công bố bài báo.

ĐÓNG GÓP CỦA TÁC GIẢ

Bùi Minh Dương, Phạm Anh Duy và Lê Duy Phúc đưa ra ý tưởng viết bài, đóng góp diễn giải phương pháp thực hiện, kết quả mô phỏng, những phân tích, thảo luận của nghiên cứu và viết bản thảo.

Bành Đức Hoài, Nguyễn Minh Tùng, Nguyễn Minh Khôi và Nguyễn Việt Dũng tham gia hỗ trợ thu thập dữ liệu, kiểm tra lại bài viết, đóng góp phần tổng quan và kết luận của bài viết.

Phạm Anh Duy, Đoàn Ngọc Minh và Nguyễn Thanh Hoan tham gia thu thập dữ liệu, chạy kết quả mô phỏng và kiểm tra lại chính tả, kết quả của bài viết.

TÀI LIỆU THAM KHẢO

1. Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y. Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Transactions on Smart Grid. 2017 Sep 18;.
2. Hippert HS, Pedreira C, Souza RC. Neural networks for short-term load forecasting: a review and evaluation. IEEE Transactions on Power Systems. Feb 2001;16(1):44–55. Available from: 10.1109/59910780.
3. Pandey AS, Singh D, Sinha SK. Intelligent Hybrid Wavelet Models for Short-Term Load Forecasting. IEEE Transactions on Power Systems. Aug 2010;25(3):1266–1273. Available from: 10.1109/TPWRS2010.2042471.
4. Qiang S, Pu Y. Short-term power load forecasting based on support vector machine and particle swarm optimization. Journal of Algorithms &t Computational Technology. 2019;doi: 10.1177174830181879706l.
5. Wang, Jujie Wang, Jianzhou Li, Yaning Zhu, Suling Zhao, Jing. Techniques of applying wavelet de-noising into a combined model for short-term load forecasting. International Journal of Electrical Power Energy Systems. 2014;62:816–824. Available from: 10.1016/.ijepes2014.05.038.
6. Garcia M, Valero S, Senabre C, Marin AG. Short-Term Predictability of Load Series Characterization of Load Data Bases. in IEEE Transactions on Power Systems. Aug 2013;28(3):2466–2474. Available from: 10.119/TPWRS20132250317.
7. Cao X, Dong S, Wu Z, Jing Y. A Data-Driven Hybrid Optimization Model for Short-Term Residential Load Forecasting. in Computer and Information Technology: Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing: Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on. 2015;p. 283–287.
8. Yun Z, Quan Z, Caixin S, Shaolan L, Yuming L, Yang S. RBF Neural Network and ANFIS-Based Short-Term Load Forecasting Approach in Real-Time Price Environment. IEEE Transactions on Power Systems. 2008;23:853–858.
9. Li H, Zhao Y, Zhang Z, Hu X. Short-term load forecasting based on the grid method and the time series fuzzy load forecasting method. in International Conference on Renewable Power Generation (RPG 2015). 2015;.
10. Qingle P, Min Z. Very Short-Term Load Forecasting Based on Neural Network and Rough Set. in Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on. 2010;p. 1132–1135.
11. Zhang R, Dong ZY, Xu Y, Meng K, Wong KP. Short-term load forecasting of Australian National Electricity Market by an ensemble model of extreme learning machine. IET Generation, Transmission Distribution. 2013;7:391–397.
12. Zhang R, Xu Y, Dong ZY, Kong W. A Composite k-Nearest Neighbor Model for Day-Ahead Load Forecasting with Limited Temperature Forecasts. presented at the IEEE General Meeting Boston. 2016;.
13. Al-Qahtani FH, Crone SF. Multivariate k-nearest neighbour regression for time series data-A novel algorithm for forecasting UK electricity demand. in Neural Networks (UJCNN), The 2013 International Joint Conference on. 2013;.
14. Tian C, Ma J, Zhang C, Zhan P. A Deep Neural Network Model for Short-Term Load Forecast Based on Long Short-Term Memory Network and Convolutional Neural Network. Energies. 2018 Dec;11(12):3493.
15. Ryu S, Noh J, Kim H. Deep neural network based demand side short term load forecasting. in 2016 IEEE International Conference on Smart Grid Communications (Smart Grid Comm). 2016;p. 308–313.
16. Gastaldi M, Lamedica R, Nardecchia A, Prudenzi A. Short-term forecasting of municipal load hrough a Kalman filtering based approach. IEEE PES Power Systems Conference and Exposition, 2004, New York, NY. 2004;3:1453–1458. Available from: 101109/PSCE2004.1397538.
17. Al-Hamadi HM, Soliman SA. Fuzzy short-term electric load forecasting using Kalman filter. IEE Proceedings-Generation, Transmission and Distribution. 16 March 2006;153(2):217–227. Available from: 101049ip-gtd20008.

18. Guan C, Luh PB, Michel LD, Chi Z. Hybrid Kalman Filters for Very Short-Term Load Forecasting and Prediction Interval Estimation. in IEEE Transactions on Power Systems. Nov 2013;28(4):3806–3817. Available from: 10.1109/TwRs20132264488.
19. Ghofrani M, Hassanzadeh M, Etezadi-Amoli MS, Fadali MS. Smart meter based short-term load forecasting for residential customers. 2011 North American Power Symposium, Boston, MA. 2011;p. 1–5. Available from: 10.1109NAPs2011.6025124.
20. Nengling T, Stenzel J, Hongxiao W. Techniques of applying wavelet transform into combine model for short-term load forecasting. Electric Power Systems Research. 2006;76(6-7):525–533. ISSN 0378-7796,. Available from: https://doi.org/10.1016/j.epsr.2005.07.003.
21. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. in Kdd. 1996;p. 226–231.
22. Chao-Ming H, Yann-Chang H. Combined wavelet-based networks and game-theoretical decision approach for real-time power dis- patch. IEEE Trans Power Syst. 2002;17(8):38.
23. Oonsivilai A, Ei-Hawary ME. Wavelet neural network based short term load forecasting of electric power system commercial load. in: Proceed- ings of IEEE Canadian Conf Elect Comput Eng. May 1999;3:1223–1228. Edmonton, Canada.
24. Kim CI, Yu LK, Song YH. Kohonen neural networks and wavelet transform based approach to short-term load forecasting. Elect Electr Power Syst Res. 2002;63(3):169–176.
25. Zhang X, Wang J, Zhang K. Short-term electric load forecasting based on singular spectrum analysis and support vector machine optimized by Cuckoo search algorithm,. Electric Power Systems Research. 2017;146:270–285. ISSN 0378-7796. Available from: https://doi.org/10.1016/jepsr.2017.01.035.
26. Sehgal S, et al.. Data analysis using principal component analysis. in 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), DOI: 10.1109/MedCom.2014.7005973.
27. Yang L, Yang H. Analysis of Different Neural Networks and a New Architecture for Short-Term Load Forecasting. Energies . 2019;12:1433. Available from: 10.3390/en12081433.
28. Khashei M, Bijari M. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Appl Soft Comput. 2011;11:2664–2675.

Science & Technology Development Journal – Engineering and Technology, 2(4):223-239

Open Access Full Text Article

Research Article

Applying statistical analysis for assessing the reliability of input data to improve the quality of short-term load forecasting for a Ho Chi Minh City distribution network

Phuc Le Duy[1,2,*], Duong Minh Bui[2], Anh Pham Duy[3], Hoan Nguyen Thanh[1], Hoai Banh Duc[1], Tung Nguyen Minh[1], Khoi Nguyen Minh[1], Minh Doan Ngoc[1], Dung Nguyen Viet[1]

ABSTRACT

Short-term load forecasting has an extremely important role in the design, operation and planning of power system, especially on apower grid of Ho Chi Minh City (HCMC) - anactive city hasthe highest power demand in Vietnam. Through the data survey, the load power in the HCMC area changes suddenly so that it causes disturbances in the load data. Accordingly, the reliability assessment of the load data will be essential in the processing stage of data-filtering before implementing load forecasting models. This study introduces a novel statistical data-filtering method that takes into account the reliability of the input-data source by analyzing many different confidence levels. Results of the proposed data-filtering method will be compared to previous data -iltering methods (such as Kalman, DBSCAN, Wavelet Transform and SSA filtering methods). The data source used in this study was collected from more than 50 substations uisng the SCADA system in Ho Chi Minh City's distribution network and was put into a neural network prediction model - ANN (Artificial Neural Network) and a ARIMA model (Autoregressive Integrated Moving Average), to demonstrate the effectiveness of the proposed data-filtering method. Numerical results derived from ANN and ARIMA predictive models show the effectiveness of the proposed data-filtering method, particularly, when the reliability of real data from the Ho Chi Minh city distribution network is determined at the 95% level, the forecasting results of ANN and ARIMA models using the proposed data-filtering method are obviously better than that without filtering method or using other data-filtering methods.

Key words: Short-term load forecast, data filtering, statistical analysis, confidence level, neural network and ARIMA

[1]*Ho Chi Minh City Power Corporation*

[2]*Institute of Engineering, Ho Chi Minh University of Technology (HUTECH)*

[3]*Faculty of Engineering, Vietnamese-German University*

Correspondence

Phuc Le Duy, Ho Chi Minh City Power Corporation

Institute of Engineering, Ho Chi Minh University of Technology (HUTECH)

Email: phucld@hcmpc.com.vn

History

- Received: 5-10-2019
- Accepted: 25-11-2019
- Published: 31-12-2019

DOI : 10.32508/stdjet.v2i4.614

Check for updates

VNU-HCM Press

Cite this article : Le Duy P, Minh Bui D, Pham Duy A, Nguyen Thanh H, Banh Duc H, Nguyen Minh T, Nguyen Minh K, Doan Ngoc M, Nguyen Viet D. **Applying statistical analysis for assessing the reliability of input data to improve the quality of short-term load forecasting for a Ho Chi Minh City distribution network**. *Sci. Tech. Dev. J. – Engineering and Technology;* 2(4):223-239.