

MÔ HÌNH DỰ BÁO PHỤ TẢI NGẮN HẠN DỰA TRÊN PHƯƠNG PHÁP XỬ LÝ DỮ LIỆU NHÓM

A SHORT-TERM LOAD FORECASTING MODEL BASED ON GROUP METHOD OF DATA HANDLING

Lê Đình Dương

Trường Đại học Bách khoa - Đại học Đà Nẵng; ldduong@dut.udn.vn

Tóm tắt - Dự báo phụ tải nói chung và dự báo phụ tải ngắn hạn nói riêng đóng một vai trò rất quan trọng trong việc vận hành hiệu quả và tin cậy hệ thống điện. Hiện nay, có nhiều phương pháp dự báo phụ tải đã được đề xuất và mỗi phương pháp có ưu nhược điểm riêng. Bài báo trình bày phương pháp dự báo phụ tải ngắn hạn dựa trên kỹ thuật xử lý dữ liệu nhóm kết hợp với các kỹ thuật tiền xử lý trong xác suất thống kê. Phương pháp đề xuất cho kết quả có độ chính xác cao và linh hoạt do đó dễ sử dụng với các chuỗi số liệu trong thực tế. Kết quả đạt được từ phương pháp đề xuất còn được so sánh với các phương pháp phổ biến khác để chứng tỏ ưu điểm của phương pháp được đề xuất.

Từ khóa - Dự báo phụ tải; ngắn hạn; phương pháp xử lý dữ liệu nhóm; mô hình chuỗi thời gian; tự hồi quy tích hợp trung bình trượt

Abstract - Load forecasting in general and short-term load forecasting in particular plays a very important role in efficient and reliable operation of power systems. Currently, there are many load forecasting methods that have been proposed and each method has its own advantages and disadvantages. This paper presents a short-term load forecasting method based on Group Method of Data Handling (GMDH) combined with pre-processing techniques in probability and statistics. The proposed method can give high accurate results and is very flexible so it is easy to use with time series data in practice. The results obtained from the proposed method are also compared with those obtained from other popular methods to demonstrate the advantages of the proposed method.

Key words - Load forecasting; Short-term; Group Method of Data Handling; time series model; ARIMA

1. Đặt vấn đề

Dự báo phụ tải là một trong những lĩnh vực rất được quan tâm bởi các nhà nghiên cứu cũng như các đơn vị quản lý, vận hành hệ thống điện. Dự báo phụ tải cung cấp những thông tin rất quan trọng cho quy hoạch và phân phối điện năng. Dự báo phụ tải quyết định kế hoạch vận hành, kế hoạch sản xuất và hướng đầu tư phát triển trong tương lai. Kết quả dự báo sẽ ảnh hưởng rất lớn đến vận hành hệ thống: Nếu dự báo cao hơn nhiều so với thực tế thì phải huy động vốn để xây dựng nhiều nguồn phát điện nhưng thực tế không dùng hết công suất sẽ gây lãng phí; ngược lại, kết quả dự báo thấp hơn nhiều so với thực tế sẽ giảm độ an toàn cung cấp điện như thiếu hụt nguồn điện gây cắt điện diện rộng thiệt hại về kinh tế - xã hội, an ninh quốc phòng.

Hiện nay, có rất nhiều phương pháp dự báo phụ tải. Một trong những phương pháp cho kết quả chính xác đó là phương pháp Persistence sử dụng giả thuyết đơn giản rằng giá trị được dự báo tại thời điểm tương lai sẽ bằng với giá trị thực tế trước đó một bước thời gian và sau một bước thời gian thì giá trị đo được sẽ được cập nhật để phục vụ cho dự báo cho bước tiếp theo. Phương pháp này cho kết quả chính xác khi áp dụng trong miền thời gian dự báo cực ngắn (vài phút đến một vài giờ). Tuy nhiên, tính chính xác sẽ giảm một cách nhanh chóng khi miền thời gian dự báo tăng lên. Nhóm phương pháp dự báo rất phổ biến là phương pháp chuỗi thời gian [1]. Các mô hình này dễ mô hình hóa và dễ phát triển so với các mô hình khác. Phương pháp chuỗi thời gian bao gồm nhóm phương pháp dự báo như tự hồi quy (AR – Auto-Regressive), trung bình trượt (MA – Moving Average), trung bình trượt kết hợp tự hồi quy (ARMA – Auto-Regressive Moving Average), tự hồi quy tích hợp trung bình trượt (ARIMA – Auto-Regressive Integrated Moving Average). Phương pháp thống kê thích hợp dự báo cho miền thời gian ngắn. Với sự phát triển của trí tuệ nhân tạo nhiều

phương pháp mới dự báo phụ tải đã được phát triển. Mạng nơron nhân tạo [2, 3] có thể giải quyết các vấn đề không tuyến tính và phức tạp về dự báo, tuy nhiên mô hình đòi hỏi nguồn dữ liệu đầy đủ cho việc huấn luyện mô hình (bao gồm cả số liệu về phụ tải và các số liệu về môi trường...). Một nhóm phương pháp khác đó là sự kết hợp (hybrid) giữa các phương pháp khác nhau để tận dụng được ưu điểm của từng phương pháp riêng lẻ. Tuy nhiên, phương pháp này thường phức tạp hơn và việc kết hợp phải thực hiện một cách hiệu quả mới phát huy tác dụng của từng phương pháp [4]. Nhìn chung, mỗi phương pháp dự báo có ưu nhược điểm và đặc điểm sử dụng riêng, tùy theo số liệu cụ thể thu thập được, đặc trưng của số liệu cũng như yêu cầu ứng dụng của bài toán dự báo, miền thời gian tương lai yêu cầu mà lựa chọn phương pháp dự báo phù hợp.

Theo miền thời gian tương lai, dự báo phụ tải có thể được phân thành ba loại: dự báo ngắn hạn (vài phút đến vài giờ, một ngày), trung hạn (một vài ngày đến một tuần) và dài hạn (một tuần, một tháng, một năm đến vài năm). So với dự báo trung và dài hạn, dự báo tải ngắn hạn đóng vai trò quan trọng trong công tác vận hành. Bài báo này tập trung vào dự báo ngắn hạn. Trong miền dự báo này, ngoài các phương pháp đề cập như trên thì phương pháp xử lý dữ liệu nhóm (GMDH) [5-7] được đề xuất sử dụng trong bài báo. Phương pháp GMDH là một nhóm các mô hình toán học và thuật toán hồi quy phi tuyến, ban đầu được đề xuất bởi Ivakhnenko, một nhà khoa học và nhà toán học người Ukraine vào năm 1968. Phương pháp này có cách tiếp cận như mạng nơron đa thức, sử dụng ý tưởng chọn lọc tự nhiên để kiểm soát kích thước, độ phức tạp và độ chính xác của mạng. Phương pháp này cho kết quả dự báo có độ chính xác cao. Ngoài ra, kết quả dự báo đạt được từ GMDH còn được so sánh với nhóm phương pháp chuỗi thời gian sử dụng phổ biến đối với dự báo phụ tải ngắn hạn hiện nay.

2. Mô hình chuỗi thời gian

2.1. Mô hình tự hồi quy (AR)

Một quá trình ngẫu nhiên $\{X_t, t \in \mathbb{Z}\}$ là một quá trình tự hồi quy cấp p , viết là $X_t \sim AR(p)$ là một quá trình dừng khi $\{X_t, t \in \mathbb{Z}\}$ thỏa mãn [1]:

$$X_t = a_0 + a_1 X_{t-1} + a_2 X_{t-2} + \dots + a_p X_{t-p} + \varepsilon_t \quad (1)$$

Với:

- $\{\varepsilon_t\}$: Nhiễu trắng [1] (white noise). Một chuỗi thời gian là nhiễu trắng nếu các biến độc lập và phân phối giống hệt nhau với giá trị trung bình bằng không.

- $a_0, a_1, a_2 \dots a_p$: Các hệ số phân tích hồi quy.

Ta có thể viết biểu thức của quá trình tự hồi quy ở trên bởi công thức:

$$X_t - a_0 - a_1 X_{t-1} - a_2 X_{t-2} - \dots - a_p X_{t-p} = \varepsilon_t \quad (2)$$

Ví dụ:

+ Mô hình AR(1): $X_t = a_0 + a_1 X_{t-1} + \varepsilon_t$

+ Mô hình AR(2): $X_t = a_0 + a_1 X_{t-1} + a_2 X_{t-2} + \varepsilon_t$

2.2. Mô hình trung bình trượt (MA)

Quá trình $\{X_t, t \in \mathbb{Z}\}$ được gọi là quá trình trung bình trượt cấp q , ký hiệu $X_t \sim MA(q)$, là một quá trình $\{X_t, t \in \mathbb{Z}\}$ thỏa mãn biểu thức [1]:

$$X_t = \varepsilon_t + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q} \quad (3)$$

Với: $b_1, b_2, \dots, b_q$ là các hệ số của mô hình.

Ví dụ:

+ Mô hình MA(1): $X_t = \varepsilon_t + b_1 \varepsilon_{t-1}$

+ Mô hình MA(2): $X_t = \varepsilon_t + b_1 \varepsilon_{t-1} + b_2 \varepsilon_{t-2}$

2.3. Mô hình tự hồi quy trung bình trượt (ARMA)

Một quá trình $\{X_t, t \in \mathbb{Z}\}$ được gọi là quá trình tự hồi quy trung bình trượt cấp p, q , ký hiệu $X_t \sim ARMA(p, q)$ là một quá trình $\{X_t, t \in \mathbb{Z}\}$ thỏa mãn [1]:

$$X_t = a_0 + a_1 X_{t-1} + \dots + a_p X_{t-p} + \varepsilon_t + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q} \quad (4)$$

Trong đó:

$a_0, a_1 \dots a_p$; $b_1, b_2, \dots, b_q$: Các hệ số của mô hình.

$A(z)$ và $B(z)$ lần lượt là đa thức tự hồi quy và đa thức trung bình trượt có bậc tương ứng là p và q [1].

$$A(z) = 1 + a_1 z + a_2 z^2 + \dots + a_p z^p \quad (5)$$

$$B(z) = 1 + b_1 z + b_2 z^2 + \dots + b_q z^q \quad (6)$$

Khi đó ta có thể viết quá trình ARMA(p, q) ở dạng toán tử:

$$A(z).X_t = B(z).\varepsilon_t \quad (7)$$

Ví dụ mô hình ARMA(2,1):

$$X_t = a_0 + a_1 X_{t-1} + a_2 X_{t-2} + \varepsilon_t + b_1 \varepsilon_{t-1}$$

2.4. Mô hình tự hồi quy tích hợp trung bình trượt (ARIMA)

Các mô hình trên được sử dụng cho các chuỗi số liệu đơn có tính dừng (stationary). Khi dữ liệu không có tính dừng (nonstationary), trước khi áp dụng các mô hình trên, chuỗi

số liệu phải được xử lý để thỏa mãn điều kiện dừng. Nếu dùng phương pháp tích hợp (I - Integrated) để biến chuỗi thời gian không dừng thành chuỗi dừng rồi áp dụng mô hình ARMA lúc đó ta sẽ có mô hình ARIMA [1]. Như vậy, mô hình phân tích và mô phỏng một chuỗi thời gian ARIMA gồm các quá trình sau: tự hồi quy (AR), tích hợp (I) và trung bình trượt (MA). Nếu chuỗi tích hợp bậc d (ký hiệu là $I(d)$) thì sau khi lấy sai phân d lần thì chuỗi sẽ dừng. Trong thực tế với chuỗi không dừng thì thường d chỉ bằng 1 hoặc bằng 2. Mô hình ARIMA được ký hiệu ARIMA(p, d, q).

Như vậy, mô hình ARIMA là mô hình tổng quát nhất của chuỗi thời gian. Để dùng mô hình ARIMA cho việc dự báo cần thực hiện các bước sau đây [1]:

Bước 1: Nhận dạng mô hình

+ Bước này để xác định chuỗi thời gian có dạng nào trong các dạng AR, MA, ARMA, ARIMA và xác định các thành phần p, d và q của mô hình.

+ Thành phần sai phân d của mô hình được nhận dạng thông qua việc xác định tính dừng của chuỗi thời gian: Nếu chuỗi thời gian dừng ở bậc không ta có $d=0$, nếu sai phân bậc 1 của chuỗi dừng ta có $d=1$, nếu sai phân bậc 2 của chuỗi dừng ta có $d=2 \dots$ Các phương pháp phổ biến kiểm tra tính dừng là Dickey-fuller [8], Kwiatkowski-Phillips-Schmidt-Shin (KPSS) [9].

+ Xác định bậc p, q của thành phần AR và MA:

Có thể sử dụng hàm tự tương quan (Autocorrelation function - ACF) [1] và tự tương quan riêng phần (Partial autocorrelation function - PACF) [1] như sau:

- Nếu ACF có các hệ số tương quan sau bậc q giảm nhanh về không và PACF có các hệ số tương quan riêng phần giảm dần về không, khi đó q chính là bậc của MA;

- Nếu PACF có các hệ số tương quan riêng phần sau bậc p giảm nhanh về không và ACF có các hệ số tương quan giảm dần về không, khi đó p chính là bậc của AR;

Bước 2: Ước lượng các tham số của mô hình

Có thể sử dụng phương pháp ước lượng hợp lý cực đại (Maximum Likelihood Estimation - MLE) [10] hoặc phương pháp bình phương cực tiểu (Ordinary Least Squares - OLS) [11].

Bước 3: Kiểm định mô hình

Để đảm bảo mô hình là phù hợp, sai số của mô hình phải là nhiễu trắng. Có thể sử dụng hàm tự tương quan ACF để kiểm tra.

Bước 4: Sử dụng mô hình để dự báo.

3. Phương pháp xử lý dữ liệu nhóm

Thuật toán GMDH có thể được biểu diễn dưới dạng một tập hợp các nơron trong đó các cặp khác nhau trong mỗi lớp được kết nối thông qua đa thức bậc hai và do đó tạo ra các nơron mới ở lớp tiếp theo.

Mối quan hệ giữa các biến đầu ra và đầu vào có thể được biểu thị dưới dạng rời rạc phức tạp bằng chuỗi Volterra [5, 6]:

$$y = a_0 + \sum_{i=1}^r a_i x_i + \sum_{i=1}^r \sum_{j=1}^r a_{ij} x_i x_j + \sum_{i=1}^r \sum_{j=1}^r \sum_{k=1}^r a_{ijk} x_i x_j x_k + \dots \quad (8)$$

Biểu thức trên được biết đến là đa thức Kolmogorov–Gabor, trong đó $X = (x_1, x_2, \dots, x_r)$ là vector đầu vào, y là biến đầu ra.

Dạng (8) là mô tả toán học đầy đủ, nó có thể được biểu diễn bằng một hệ thống đa thức bậc hai bán phần chỉ gồm hai biến (nơron) dưới dạng:

$$\hat{y} = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2 \quad (9)$$

Mục đích của biểu thức trên là ước lượng sao cho giá trị $\hat{y}$ càng gần với giá trị thực tế y thì mô hình dự báo càng chính xác.

Thuật toán GMDH gồm các bước chính sau:

Bước 1: Xác định tất cả nơron (ước lượng các vector tham số dùng tập dữ liệu huấn luyện) với đầu vào bao gồm tất cả các cặp giá trị biến đầu vào. Do đó, bao gồm $r(r-1)/2$ cặp (nơron).

Bước 2: Dùng tập hợp số liệu xác nhận để chọn các nơron phù hợp nhất với tiêu chí lựa chọn.

Bước 3: Nếu điều kiện dừng thỏa mãn (mạng phù hợp với dữ liệu với độ chính xác mong muốn hoặc việc giới thiệu các nơron mới không làm tăng đáng kể khả năng xấp xỉ của mạng nơron), sau đó sẽ cho dừng, nếu không thì sử dụng đầu ra của các nơron tốt nhất (được chọn trong Bước 2) để hình thành vector đầu vào cho lớp tiếp theo, rồi sau đó chuyển sang Bước 1.

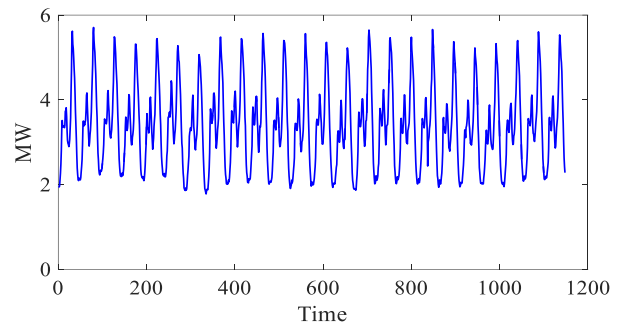
GMDH hoạt động bằng cách xây dựng các lớp liên tiếp với các liên kết (hoặc kết nối) phức tạp là các mục của một đa thức. Lớp ban đầu chỉ đơn giản là lớp đầu vào. Lớp đầu tiên được thực hiện bằng cách hồi quy của các biến đầu vào và sau đó chọn các biến tốt nhất. Lớp thứ hai được tạo bằng cách tính hồi quy của các giá trị trong lớp đầu tiên cùng với các biến đầu vào. Điều này có nghĩa là thuật toán về cơ bản xây dựng các đa thức của đa thức.

4. Áp dụng

Trong phần này, các phương pháp trên được thực hiện để dự báo cho các phụ tải trong thực tế. Trong số liệu đo đếm thực tế thường tồn tại những số liệu lỗi và dữ liệu bị mất (missing data). Nếu các số liệu này không được xử lý thì sẽ rất khó trong việc xây dựng mô hình dự báo cũng như ảnh hưởng rất lớn đến độ chính xác của bài toán dự báo. Do đó, các số liệu lỗi phải được loại ra trước khi xây dựng mô hình. Các số liệu này thường do lỗi thiết bị đo đếm hoặc việc truyền và lưu trữ dữ liệu gây ra. Có nhiều phương pháp để loại trừ dữ liệu như vậy, trong bài báo này tác giả sử dụng phương pháp được đề xuất bởi Grubbs [12] trong đó các số liệu quá lớn hoặc quá khác so với phần còn lại sẽ được xác định và loại bỏ ra khỏi tập số liệu. Sau khi loại trừ dữ liệu lỗi, bước tiếp theo là xác định và xử lý dữ liệu bị mất. Tương tự như số liệu lỗi, các số liệu bị mất do lỗi thiết bị đo đếm hoặc việc truyền và lưu trữ dữ liệu gây ra. Các số liệu này có thể được suy ra từ các số liệu lân cận bằng các phương pháp nội suy [13].

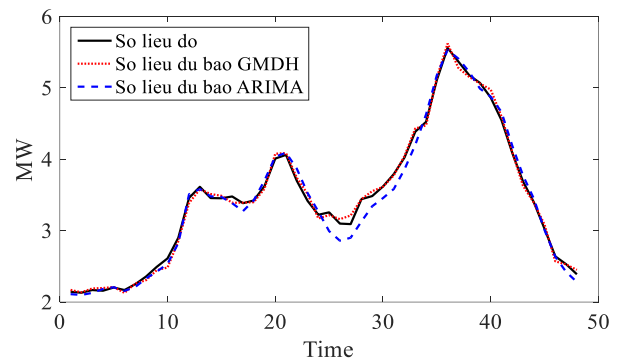
Hình 1 vẽ số liệu thu thập được ở 1 điểm đo phụ tải trong thực tế, trong đó số liệu được thu thập theo bước thời gian là 30 phút (mỗi ngày có 48 số liệu được thu thập). Tổng số liệu có được là 30 ngày trong đó số liệu 24 ngày đầu dùng để ước lượng và huấn luyện các mô hình, số liệu

các ngày còn lại để kiểm định độ chính xác của các mô hình. Trên Hình 1 ta có thể thấy rằng, số liệu có đặc tính ngày theo sự biến động của phụ tải.



Hình 1. Số liệu thu thập tại phụ tải 1 (Bước thời gian 30 phút)

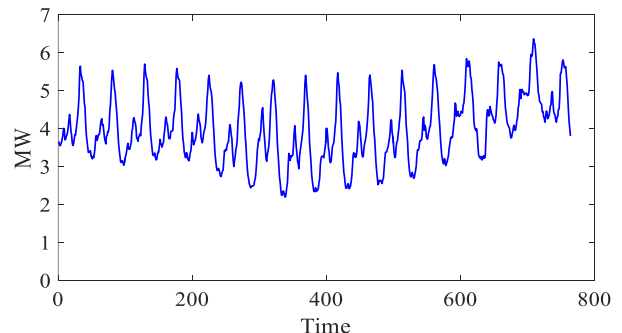
Thực hiện các bước của ARIMA cho dự báo của ngày tiếp theo (ngày thứ 25) như trình bày trong Mục 2.4 thu được kết quả như trong Hình 2. Mô hình có dạng ARIMA(2,0,0) với các hệ số ước lượng được là $\{a_1 = 0,806394, a_2 = -0,0145791\}$. Tương tự, thực hiện dự báo áp dụng thuật toán GMDH thu được kết quả trên Hình 2. Trên Hình 2, so sánh kết quả ta thấy rằng, cả hai phương pháp đều cho kết quả tốt, trong đó phương pháp GMDH cho kết quả chính xác hơn so với ARIMA. Sai số tuyệt đối trung bình (Mean Absolute Error - MAE) [14] lần lượt là 0,0332 cho GMDH và 0,0844 cho ARIMA.



Hình 2. Kết quả dự báo cho phụ tải 1 (Bước thời gian 30 phút)

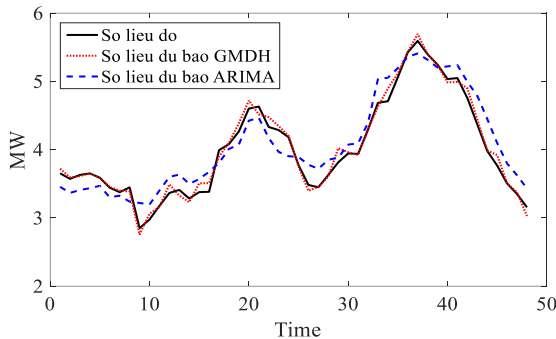
Tương tự, áp dụng các phương pháp trên cho chuỗi số liệu thu thập được ở điểm đo thứ 2 của một phụ tải trong thực tế. Số liệu thu thập được 20 ngày trong đó số liệu 16 ngày để ước lượng mô hình và 4 ngày còn lại để kiểm tra độ chính xác.

Hình 3 số liệu của tập dữ liệu dùng để ước lượng. Hình 3 cho thấy, phụ tải có đặc tính ngày và ở đây có độ thay đổi nhiều hơn so với phụ tải tại điểm đo 1.



Hình 3. Số liệu thu thập tại phụ tải 2

Tương tự, ta ước lượng được mô hình chuỗi thời gian có dạng ARIMA(4,0,0) với các hệ số ước lượng được là $\{a_1 = 1,06281, a_2 = -0,114748, a_3 = 0,06088, a_4 = -0,06263\}$ và mô hình GMDH. Kết quả so sánh trên Hình 2 (dự báo cho ngày thứ 17) cho thấy các phương pháp có độ chính xác giảm hơn so với trường hợp trên. Tuy nhiên, phương pháp GMDH vẫn cho kết quả rất chính xác còn phương pháp ARIMA bị ảnh hưởng nhiều bởi tính phức tạp của số liệu đầu vào. Sai số tuyệt đối trung bình MAE [14] lần lượt là 0,0646 cho GMDH và 0,2108 cho ARIMA.



Hình 4. Kết quả dự báo cho phụ tải 2 (Bước thời gian 30 phút)

5. Kết luận

Bài báo trình bày các phương pháp dự báo phụ tải trong miền thời gian ngắn hạn. Phương pháp dùng mô hình chuỗi thời gian được sử dụng khá phổ biến vì vừa chính xác vừa dễ sử dụng trong thực tế. Trong bài báo này, phương pháp xử lý dữ liệu nhóm GMDH được đề xuất sử dụng vì đặc điểm nổi trội về độ chính xác và tính linh hoạt để thích ứng với các chuỗi số liệu trong thực tế. Điều này được thể hiện qua các ví dụ áp dụng khác nhau. Hướng nghiên cứu tiếp theo là áp dụng phương pháp GMDH cho nhiều loại phụ tải khác nhau trong thực tế.

(BBT nhận bài: 02/6/2019, hoàn tất thủ tục phản biện: 29/6/2019)

TÀI LIỆU THAM KHẢO

- [1] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, San Francisco, CA: Holden Day, 1976.
- [2] Luis Hernández, Carlos Baladrón, Javier M. Aguiar, Lorena Calavia, Belén Carro, Antonio Sánchez-Esguevillas, Francisco Pérez, Ángel Fernández, Jaime Lloret, Artificial Neural Network for Short-Term Load Forecasting in Distribution Systems, *Energies* 2014, 7, 1576-1598.
- [3] Papia Ray, Debani Prasad Mishra, Rajesh Kumar Lenka, Short term load forecasting by artificial neural network, *International Conference on Next Generation Intelligent Systems (ICNGIS)*, 2016.
- [4] Jinliang Zhang, Yi-Ming Wei, Dezhi Li, Zhongfu Tan, Jianhua Zhou, Short term electricity load forecasting using a hybrid model, *Energy*, vol. 158, pp. 774-781, 2018.
- [5] A.G. Ivakhnenko, The group method of data handling in prediction problems, *Soviet Automatic Control c/c of Avtomatika*, vol.9, no.6, pp.21-30, 1976.
- [6] J.A. Muller, A.G. Ivachnenko, F. Lemke, GMDH algorithms for complex systems modelling, *Mathematical and Computer Modelling of Dynamical Systems*, vol. 4, no. 4, pp. 275-316, 1998.
- [7] Wei Liu, Zhenhai Dou, Weiguo Wang, Yueyu Liu, Hao Zou, Bo Zhang, Shoujun Hou, Short-Term Load Forecasting Based on Elastic NetImproved GMDH and Difference DegreeWeighting Optimization, *Appl. Sci.*, 8, 1603, 2018.
- [8] D. A. Dickey, W. A. Fuller, Distribution of the Estimators for Autoregressive Time Series with a Unit Root, *Journal of the American Statistical Association*, 74 (366), 427-431, 1979.
- [9] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, Y. Shin, Testing the null hypothesis of stationarity against the alternative of a unit root, *Journal of Econometrics*, 54 (1-3), 159-178, 1992.
- [10] Online: https://en.wikipedia.org/wiki/Maximum_likelihood_estimation
- [11] Online: https://en.wikipedia.org/wiki/Ordinary_least_squares
- [12] F. E. Grubbs, Procedures for Detecting Outlying Observations in Samples, *Technometrics*, Feb. 1969.
- [13] P. J. Davis, *Interpolation and approximation*, Dover, New York, 1976.
- [14] Cort J. Willmott, Kenji Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, *Climate Research*, 30: 79-82, 2005.