УДК 621.311:681.3 DOI: https://doi.org/10.15407/publishing2021.60.051

АНАЛІЗ МЕТОДІВ ДОСТОВІРИЗАЦІЇ ДАНИХ ДЛЯ ЗАДАЧ КОРОТКОСТРОКОВОГО ПРОГНОЗУВАННЯ ВУЗЛОВИХ ЕЛЕКТРИЧНИХ НАВАНТАЖЕНЬ

В.О. Мірошник[*], **П.В. Шиманюк**[**]
Інститут електродинаміки НАН України,
пр. Перемоги, 56, Київ, 03057, Україна
e-mail: miroshnyk.volodymyr@gmail.com, shymanp@ied.org.ua

В умовах лібералізованого ринку електроенергії України у його учасників виникають стимули до підвищення ефективності операційної діяльності, у тому числі в операторів систем розподілу. Одним з аспектів їхньої діяльності є прогнозування втрат у мережах для купівлі відповідних обсягів електричної енергії на оптовому ринку. Перспективним підходом є прогнозування вузлового навантаження та розрахунок втрат з урахуванням топології мережі. До того ж точний прогноз вузлового навантаження необхідний для оцінювання запасу стійкості енергосистеми. Під час побудови прогностичних моделей вирішальне значення має якість даних, на яких оцінюються параметри моделі. У статті запропоновано метод виявлення та заміни аномальних значень часового ряду на основі кластеризації. Проведено порівняльний аналіз методів кластеризації для виявлення пропусків та аномальних значень у часових рядах погодинного споживання електричної енергії. Для оцінки методів достовіризації використано дані північно-західного регіону США. За результатами аналізу було виявлено, що використання методу DBSCAN призводить до значно меншої кількості хибно позитивних спрацювань. Бібл.9, рис. 3., табл. 3.
Ключові слова: достовіризація даних**,** Smart Grid, алгоритм кластеризації, прогнозування, енергосистема

У новому ринку електричної енергії [1] оператори систем розподілу (ОСР) та системи передачі (ОСП) повинні купувати електроенергію для покриття втрат у власних мережах. Логіка роботи ринку вказує на те, що купівля електричної енергії (укладення договору на постачання) якомога раніше до дати споживання дає змогу знизити вартість електроенергії [2]. Це призводить до необхідності прогнозування майбутніх втрат електричної енергії. Отже, в учасників ринку з'являються прямі економічні стимули до підвищення точності прогнозів. На практиці для подання заявок на ринку «на добу наперед» використовуються короткострокові прогнози з горизонтом упередження від 12 до 36 годин.

Розрахунок економічного ефекту від зниження похибки короткострокових прогнозів обсягу втрат енергорозподільчих компаній [3] показує, що зниження похибки на 5 % дасть змогу знизити сумарні витрати операторів мереж для компенсації небалансів на 184 млн грн на рік за середньої ціни похибки у 225 грн/МВт·год, що дозволить знизити тарифи на розподіл та тариф на передачу електричної енергії для всіх кінцевих споживачів.

Найбільш розповсюдженим підходом є прямий прогноз часового ряду втрат, але за таких умов не враховується топологія мережі, за зміни якої в часовому ряді втрат виникають аномальні викиди, які ускладнюють побудову адекватної моделі та знижують точність прогнозування. Більш коректним буде розрахунок втрат на основі прогнозу вузлових навантажень з урахуванням параметрів режиму та прямого розрахунку втрат з огляду на топологію мережі.

Крім того, на основі прогнозу вузлового навантаження можна ефективно вирішувати низку технологічних задач, пов'язаних з управлінням режимом роботи

ORCID ID: * https://orcid.org/0000-0001-9036-7268, ** https://orcid.org/0000-0002-7585-7493

енергосистеми. Так, зі зміною порядку формування графіку покриття навантаження, яка пов'язана з переходом від централізованого планування (що здійснювало ДП «Енергоринок») до використання комерційних графіків, суттєво збільшилось навантаження на диспетчерську службу ОЕС України, спричинене частою зміною режимів роботи генеруючих агрегатів. Водночас виникає необхідність оцінювання запасу динамічної та статичної стійкості енергосистеми, надійність яких залежить від точності прогнозів навантаження в вузлах електричної мережі.

Під час побудови статистичних моделей електричного навантаження вирішальне значення має якість вихідної інформації. Спотворені дані та пропуски завжди присутні під час звичайної роботи енергосистеми. Вони можуть бути пов'язані з помилками та похибками вимірювань, шумами в інформаційному тракті або людським фактором. За структурою їх можна розділити на одиничні та групові. Під час побудови математичних моделей для прогнозування, різкі зміни навантаження, викликані аварійними відключеннями або раптовими змінами режимів роботи великих споживачів, також слід відносити до аномальних значень, оскільки вони вносять похибку в модель взаємозв'язку між навантаженням та зовнішніми факторами.

З огляду на тенденції щодо зростання автоматизації енергосистеми, втілення ідей концепції Smart Grid [4,5] на практиці та підвищення кількості активних споживачів виникає необхідність в ускладненні інструментів аналізу даних та визначення аномальних значень.

Метою цієї роботи є порівняння методів кластеризації, які використовуються для виявлення та заміни аномальних значень у часових рядах електричного навантаження на етапі побудови статистичних моделей прогнозування вузлового навантаження.

Основою будь-якого методу визначення аномальних значень є застосування критерію аномальності, що може бути інтерпретовано як кластеризація даних на два класи: нормальні та аномальні значення. У роботі проведено аналіз ефективності найбільш розповсюджених методів кластеризації, які можуть використовуватись для цієї задачі.

Метод DBSCAN(Density-Based Spatial Clustering of Applications with Noise) [6] є широко поширеним методом кластеризації на основі щільності даних. Такий підхід був представлений М. Естером у 1996 році. Особливістю цього методу є мінімальна кількість кластерів які використовуються як параметри. Кількість потенційних кластерів регулюється параметром мінімальної щільності.

Метод «Ізольованих лісів» IF-Isolation forest [7] може чітко виявляти аномальні значення порівняно з іншими методами, які будують профілі даних та ідентифікують їх на відповідність нормальному профілю. До таких методів можна віднести методи класифікації та кластеризації. Цей підхід використовує алгоритм машинного навчання. Деревоподібна структура алгоритму дає змогу ізолювати кожен окремий екземпляр.

Метод локального рівня викиду (LOF – Local outlier factor) [8] для виявлення аномальних даних використовує підхід щільності даних. За своєю концепцією він подібний до методу DBSCAN, оскільки так само використовує допустиму відстань для оцінки локальної щільності. LOF досить непогано виявляє аномальні значення, хоча в деяких випадках алгоритм розпізнає добові провали та піки навантаження як аномальні. Також цей алгоритм не здатний виявляти явні пропуски даних, що можна виправити за допомогою попередньої заміни та інтерполяції пропущених даних.

Метод Еліптичної обвідної (EE – Elliptic Envelope) [9] для виявлення аномальних даних використовує розподіл Гауса. Алгоритм формує еліпс навколо даних на основі надійної коваріації, що дає змогу ідентифікувати вихідні дані без впливу викидів. Відповідно будь-які значення, що знаходяться поза межею даного еліпсу, вважаються викидами. Метод EE моделює дані, як високорівневий розподіл Гауса з можливою коваріацією між розмірними ознаками. Тобто, такий метод має намір знайти межу еліпса з найбільшою кількістю даних в ньому. Метод EE здатний виявляти як аномальні, так і

пропущені дані, але при великих об'ємах ретроспективних даних алгоритм спотворює нормальні дані за усередненим значенням усієї вибірки.

Порівняльний аналіз результатів роботи методів виявлення аномальних даних показав, що більшість методів здатні виявляти аномальні та в деяких випадках пропущені значення. Метод DBSCAN добре справляється зі значними аномальними викидами, але за тривалих спотворень даних алгоритм ігнорує ці значення. Для покращення роботи цього алгоритму необхідно масштабувати дані та змінювати розмірність вхідних даних для більш точної їх ідентифікації. Метод «Ізольованих лісів» показав один з найгірших результатів. Метод ЕЕ здатний виявляти як аномальні, так і пропущені дані, але при великих об'ємах ретроспективних даних алгоритм спотворює нормальні дані за усередненим значенням усієї вибірки. Метод локального рівня викиду LOF демонструє також хороші результати, хоча він не здатний виявляти пропущені дані, які можна вирішити завдяки попередньому виділенню пропущених даних та їхній інтерполяції.

Для перевірки ефективності методів достовіризації було використано дані з відкритих джерел про споживання електричної енергії північно-західного регіону енергосистеми США, які включають погодинні значення для 10 вузлів за період з 2015 до 2019 року.

Аналіз такого набору даних показав наявність аномальних значень, як одиничних, так і групових. За природою виникнення їх можна розділити на два класи: пропущені дані та помилки (від'ємні та аномально великі значення), які виникли в процесі передачі даних. У табл. 1 наведено статистичні показники даних США до застосування алгоритму достовіризації.

Таблиця 1

Вузли	1	2	3	4	5	6	7	8	9	10
Середнє значення, МВт·год	2735	2347	6295	554	577	1376	3396	1104	230	182
Стандартне відхилення, МВт·год	36836	447	978	112	80	289	640	221	96	46
Max, МВт·год	3340061	21556	11827	998	982	11297	5504	11583	591	397
Min, МВт·год	-90964	630	2600	-12	362	748	679	0	-819	0

Подальший аналіз виявив, що алгоритми на базі кластеризації спричиняють значну кількість хибно позитивних спрацювань у вузлах зі значною добовою та тижневою періодичністю. Для розв'язання цієї задачі запропоновано використовувати метод декомпозиції часових рядів за допомогою двосторонніх ковзних середніх з різною глибиною усереднення.

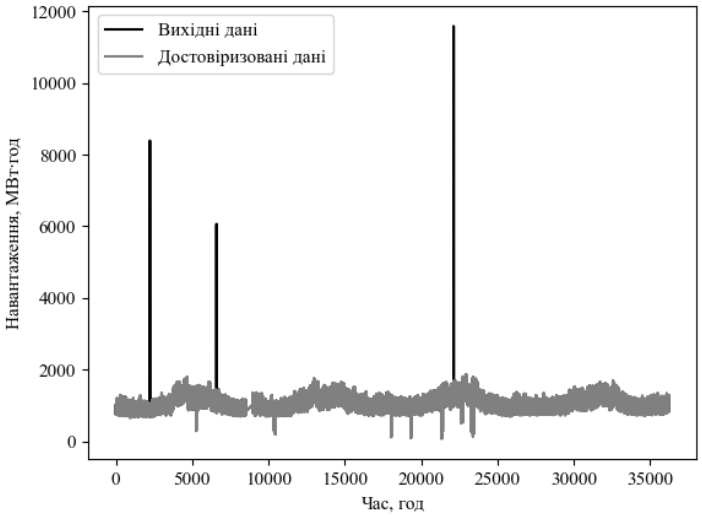

Рис.1

Кінцевий алгоритм виявлення та заміни для одного вузла складається з таких етапів:

1. Виділення часових зрізів із суцільного часового ряду навантаження $\mathbf{R}^{n\times 1} \rightarrow \mathbf{R}^{\frac{n}{24}\times 24}$.
2. Виявлення в часових зрізах грубих аномальних значень за допомогою

методу кластеризації DBSCAN. Аномальними приймаються значення, які не належать до першого кластеру.

3. Заміна аномальних значень за допомогою інтерполяції (рис. 1).

4. Розгортка часових зрізів у суцільний ряд навантаження $\mathbf{R}^{\frac{n}{24}\times 24} \rightarrow \mathbf{R}^{n\times 1}$.

5. Декомпозиція часового ряду на трендову, сезонну та залишкову складові (рис. 2).

6. Виявлення аномальних значень (пункт 2) у часовому ряді залишкової складової.

7. Заміна виявлених значень за допомогою лінійної інтерполяції (рис. 3).

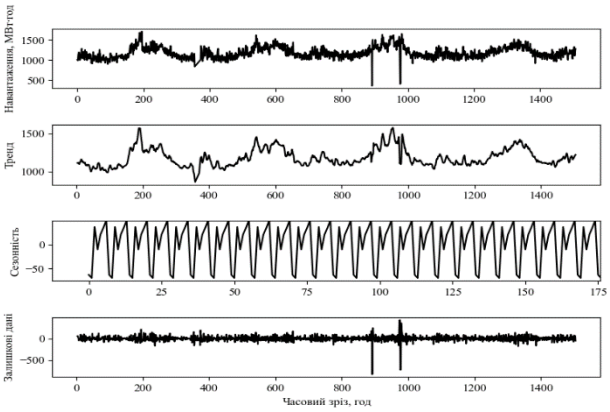

Рис. 2

На рисунках 1–3 наведено приклад двоетапної достовіризації даних для вузла № 8.даних США.

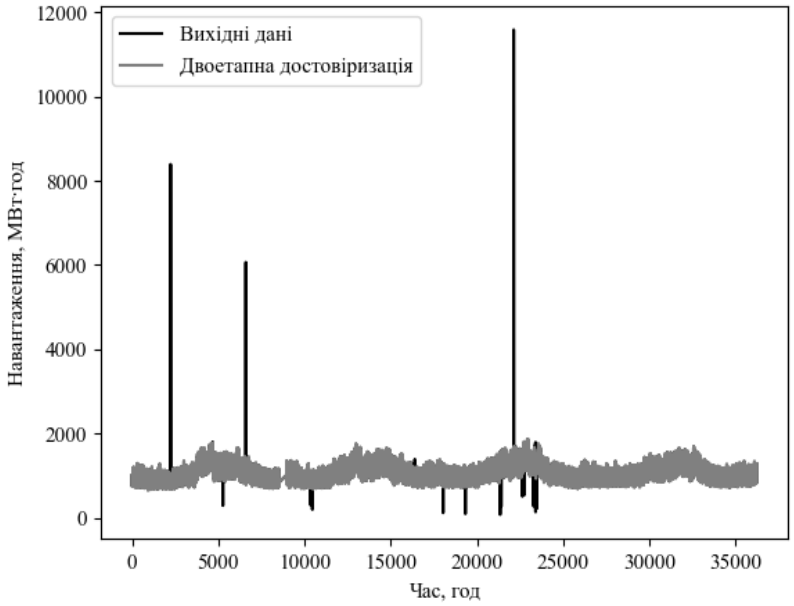

Рис. 3

Особливістю задачі виявлення аномальних значень є розбалансованість вибірки: зазвичай кількість аномальних значень значно нижча за кількість нормальних. Це призводить до неадекватності оцінки алгоритму класифікації за показником точності класифікації:

$$A = \frac{(TP + TN)}{n}, \quad (1)$$

де A – точність; TP – кількість правильно виявлених аномальних значень; TN – кількість правильно виявлених нормальних значень; n – загальна кількість значень у вибірці. У цьому випадку, якщо алгоритм класифікує всі значення ряду як нормальні, то значення точності буде 0,99884, але практична цінність такого підходу відсутня.

Для оцінки якості класифікатора на незбалансованих даних використовується показники Precision (коректність виявлення аномальних значень) та Recall (чутливість):

$$Presicion = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN},$$

де FP – кількість нормальних значень класифікованих як аномальні; FN – кількість аномальних значень класифікованих як нормальні. Для можливості однозначного ранжування методів класифікації використовують показник F-score, який є середнім гармонічним показників Precision та Recall:

$$F = \frac{2}{Recall^{-1} + Precision^{-1}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}. \quad (2)$$

Для розрахунку *TP, FP, TN, FN* фактичні аномальні значення було визначено на підставі суджень експерта. Усього виявлено 420 аномальних значень з 362 160. У табл. 2 наведено значення показників якості класифікації для всіх розглянутих методів.

Таблиця 2

	TN	FN	FP	TP	Presicion	Recall	F
DBSCAN	360 961	66	779	354	0,312	0,843	0,456
IF	292 440	26	69 300	394	0,006	0,938	0,011
Lof	294 999	125	66 741	295	0,004	0,702	0,009
EE	290 714	27	71 026	393	0,006	0,936	0,011

Ураховуючи переваги та недоліки наведених алгоритмів кластеризації, для подальшого використання вирішено застосовувати метод DBSCAN, який об'єднує в собі високу точність, специфічність визначення аномальних викидів та ефективність з погляду використання обчислювальних ресурсів.

Для оцінки похибки прогнозу використовувались ті ж самі дані США. Всі дані було розділено на навчальну та тестові вибірки. Останні 744 погодинні значення, які не використовувались для навчання моделі, віднесено в тестову вибірку, на якій розраховувались значення показника похибки прогнозу.

Для порівняння результатів прогнозування використано розроблену нейронну мережу глибинного навчання типу LSTM (long- short term memory). Нейронна мережа складається з рекурентного модуля LSTM після якого йдуть два прихованих повнозв'язних шари. Як активаційна функція прихованого шару використовується функція selu (scaled exponential linear unit). Навчання нейронної мережі виконується за допомогою алгоритму оптимізації ADAM.

У табл. 3 наведено результати прогнозування за допомогою нейронної мережі на даних до та після усунення аномальних значень. Для оцінки середньої похибки прогнозу використовується MAPE.

Наведені результати показують, що усунення аномальних значень значно знижує середню похибку прогнозу з 17,3 до 5,8 % та підвищує стабільність прогнозів, зниження стандартного відхилення похибки з 35,1 до 1,6 %.

Таблиця 3

Вузли	з аномальними значеннями	без аномальних значень
Мінімальна похибка, %	4,1	3,6
Максимальна похибка, %	117,2	7,8
Середня похибка, %	17,3	5,8
Стандартне відхилення похибки, %	35,1	1,6

Висновки. Метод достовіризації на базі алгоритму DBSCAN ефективно визначає значні одиничні аномальні викиди, але за тривалих групових викидів такий алгоритм

ігнорує ці значення. Для покращення роботи алгоритму необхідно масштабувати та змінювати розмірність вхідних даних.

Використання алгоритмів на базі кластеризації IF, LOF, EE призводять до більшої кількості хибно позитивних спрацювань у вузлах зі значною добовою та тижневою періодичністю, ніж використання методу DBSCAN. Для розв'язання цієї задачі запропоновано використовувати метод декомпозиції часових рядів за допомогою двосторонніх ковзних середніх із різною глибиною усереднення.

Метод двоетапної достовіризації даних за допомогою методу DBSCAN та методу декомпозиції часових рядів дає змогу зменшити кількість хибно позитивних спрацювань, подальше використання достовіризованих даних для прогнозу дозволяє зменшити середню похибку з 17.3 до 5.8 %.

За результатами досліджень для задач виявлення аномальних даних та пропусків найбільш доцільним є метод DBSCAN, використання та розвиток якого дасть змогу збільшити ефективність виявлення аномальних значень.

Фінансується за держбюджетною темою «Науково-технічні засади розвитку та керованості сегменту розосереджених джерел енергії в структурі генеруючих потужностей електроенергетичних систем» (шифр «СЕГМЕНТ»), що виконується за Постановою Бюро ВФТПЕ НАН України, протокол №11 від 04.07.2017. Державний реєстраційний номер роботи 0117U007711. КПКВК 6541030.

1. Блінов І.В. Проблеми функціонування та розвитку нової моделі ринку електричної енергії в Україні (за матеріалами наукової доповіді на засіданні Президії НАН України 3 лютого 2021 р.). *Вісник Національної академії наук України*. 2021. № 3. С. 20–28. DOI: https://doi.org/10.15407/visn2021.03.020
2. Блінов І.В., Мірошник В.О., Шиманюк П.В. Короткостроковий інтервальний прогноз сумарного відпуску електроенергії виробниками з відновлювальних джерел енергії. *Праці Інституту електродинаміки НАН України.* 2019. Вип. 54. С. 5–12. DOI: https://doi.org/10.15407/publishing2019.54.005
3. Блінов І. В., Мірошник В. О., Шиманюк П.В. Оцінка вартості похибки прогнозу «на добу наперед» технологічних втрат в електричних мережах України. *Технічна електродинаміка*. 2020. № 5. С. 70–73. DOI: https://doi.org/10.15407/techned2020.05.070.
4. IEC 63097/TR/Ed1: Smart Grid Roadmap. International Electrotechnical Commission, 2016, 308 p.
5. Кириленко О.В., Блінов І.В., Танкевич С.Є. Smart Grid та організація інформаційного обміну в електроенергетичних системах. Технічна електродинаміка. 2012. № 3. С.47–48.
6. Ester M., Kriegel H.P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, 1996. Pp. 226–231.
7. Liu F.T., Ting K.M., Zhou Z. Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy. 2008. Pp. 413–422. DOI: https://doi.org/10.1109/ICDM.2008.17
8. Breunig M.M., Kriegel H.P., Ng R.T., Sander J. LOF: identifying density-based local outliers. ACM Sigmod Record 29. 2000. Pp. 93–104. DOI: https://doi.org/10.1145/342009.335388.
9. Rousseeuw P.J., Van Driessen K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics.* 1999. Vol. 41 (3). Pp. 212–23. DOI: https://doi.org/10.1080/00401706.1999.10485670.

ANALYSIS OF METHODS OF INCREASING DATA RELIABILITY FOR PROBLEMS OF SHORT TERM FORECASTING OF NODAL LOAD

V. O. Miroshnyk , **P. V. Shymaniuk**
Institute of Electrodynamics of the National Academy of Sciences of Ukraine,
pr. Peremohy, 56, Kyiv, 03680, Ukraine
e-mail: miroshnyk.volodymyr@gmail.com, shymanp@ied.org.ua

A comparative analysis of clustering methods was performed to identify gaps and anomalous values in the data. Data from the northwestern region of the United States were used for evaluation. According to the analysis results, it was found that the use of the DBSCAN method leads to a much smaller number of false positives. An algorithm for two-stage data validation using clustering and time series decomposition methods is proposed. Ref. 9, fig. 3, tables 3.
Keywords: anomaly detection**,** Smart Grid, clustering algorithm, forecasting, power system.

1. Blinov I.V. Problems of functioning and development of a new electricity market model in Ukraine (According to the scientific report at the meeting of the Presidium of NAS of Ukraine, February 3, 2021) Visn. Nac. Acad. Nauk Ukr. 2021. No 3. Pp. 20–28. DOI: https://doi.org/10.15407/visn2021.03.020 (Ukr)
2. Blinov I., Miroshnyk V., Shymaniuk P. Short-term interval forecast of total electricity generation by renewable energy sources producers. Pratsi Instytutu elektrodynamiky NAN Ukrainy . 2019. No 54. Pp. 5–12. DOI: https://doi.org/10.15407/publishing2019.54.005 (Ukr)
3. Blinov I., Miroshnyk V., Shymaniuk P. The cost of error of "day ahead" forecast of technological losses of electrical energy. Tekhnichna elektrodynamika. 2020. No 5. Pp. 70–73. DOI: https://doi.org/10.15407/techned2020.05.070 (Ukr)
4. IEC 63097/TR/Ed1: Smart Grid Roadmap. International Electrotechnical Commission, 2016, 308 p.
5. Kyrylenko, O.V., Blinov, I.V., Tankevych, S.E. Smart grid and organization of information exchange in electric power systems. Tekhnichna elektrodynamika. 2012. No 3. Pp. 47–48. (Ukr)
6. Ester M., Kriegel H.P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, 1996. Pp. 226–231.
7. Liu F.T., Ting K.M., Zhou Z. Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008. Pp. 413–422. DOI: https://doi.org/10.1109/ICDM.2008.17 .
8. Breunig M.M., Kriegel H.P., Ng R.T., Sander J. LOF: identifying density-based local outliers. ACM Sigmod Record 29, 2000. Pp. 93–104. DOI: https://doi.org/10.1145/342009.335388 .
9. Rousseeuw P.J., Van Driessen K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics.* 1999. Vol. 41(3). Pp. 212–223. DOI: https://doi.org/10.1080/00401706.1999.10485670 .

Надійшла: 11.06. 2021
Received: 11.06. 2021