ТЕХНИЧЕСКИЕ НАУКИ
И ТЕХНОЛОГИИ

АВТОМАТТАНДЫРУ ЖӘНЕ БАСҚАРУ
АВТОМАТИЗАЦИЯ И УПРАВЛЕНИЯ
AUTOMATION AND CONTROL

DOI 10.51885/1561-4212_2022_2_24
МРНТИ 50.49.35

Б.А. Амиев[1], А.Г. Тюлепбердинова[2], А.С. Өмірзақ[3], А.М. Ахмедов[4]
Әл-Фараби атындағы Қазақ ұлттық университеті, г. Алматы, Казахстан
[1]*E-mail: bolatbekamiev@gmail.com**
[2]*E-mail: tyulepberdinova@gmail.com*
[3]*E-mail: u.azamat.1997@gmail.com*
[4]*E-mail: askar.1230@mail.ru*

ПРИМЕНЕНИЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ДЛЯ РАСЧЕТА ПОТЕРИ ЭЛЕКТРОЭНЕРГИИ

ДЕРЕКТЕРДІ ИНТЕЛЛЕКТУАЛДЫ ТАЛДАУ АРҚЫЛЫ ЭЛЕКТРЭНЕРГИЯНЫҢ ЖОҒАЛУЫН ЕСЕПТЕУ

APPLICATION OF DATA MINING TO CALCULATE POWER LOSS

Аннотация. *В этом исследовании описываются три различных метода интеллектуального анализа данных для обнаружения аномального энергопотребления освещения с использованием ежечасно регистрируемых данных о потреблении энергии и пиковом спросе (максимальной мощности). Для обнаружения аномального потребления в одном и том же наборе данных к каждому классу и кластеру применяются два метода обнаружения выбросов. В каждом классе и кластере с аномальным потреблением величина отклонения от нормы определяется с использованием модифицированных стандартных оценок. Результаты исследования будут полезны при создании систем управления энергопотреблением, чтобы сократить эксплуатационные расходы и время за счет отсутствия необходимости обнаруживать неисправности вручную или диагностировать ложные предупреждения. Кроме того, это будет полезно для разработки модели обнаружения неисправностей и диагностики энергопотребления всего здания.*

Ключевые слова: *кластеризация, выбросы, классификация, методы, к-средние.*

Аңдатпа. *Бұл зерттеу энергияны тұтыну және ең жоғары сұраныс (максималды қуат) туралы сағат сайын тіркелген деректерді қолдана отырып, жарықтандырудың шамадан тыс энергия шығынын анықтау үшін деректерді іздеудің үш түрлі әдісін сипаттайды. Бір деректер жиынтығында аномалды тұтынуды анықтау үшін әр сынып пен кластерге шығарындыларды анықтаудың екі әдісі қолданылады. Қалыпты емес тұтынуы бар әрбір сынып пен кластерде нормадан ауытқу шамасы түрлендірілген стандартты бағалауды пайдалана отырып айқындалады. Зерттеу қолмен ақаулықтарды анықтаудың немесе жалған ескертулерді диагностикалаудың қажеті жоқ болғандықтан, пайдалану шығындары мен уақытты азайту үшін энергияны басқару жүйелерін құру үшін пайдалы болады. Сонымен қатар, бұл ақауларды анықтау моделін жасау және бүкіл ғимараттың энергия тұтынуын диагностикалау үшін пайдалы болады.*

Түйін сөздер: *кластерлеу, шығарындылар, жіктеу, әдістер, к-орташа.*

Abstract: *This study describes three different data retrieval methods to determine excessive energy consumption of lighting using hourly recorded data on energy consumption and maximum demand (maximum power). To determine abnormal consumption in a single data set, two methods are used to determine emissions for each class and cluster. In each class and cluster with abnormal consumption, the*

value of deviations from the norm is determined using a modified standard estimate. Since the study does not need to manually detect malfunctions or diagnose false warnings, it will be useful for creating energy management systems to reduce operating costs and time. In addition, it will be useful for developing a fault detection model and diagnosing the energy consumption of the entire building.

Keywords: *Clustering, outliers, classification, methods, k-means.*

Введение. Потребление энергии как жилыми, так и коммерческими зданиями неуклонно растет, достигая показателей до 40 % в развитых странах. Растущий спрос на строительные услуги и высокий уровень теплового комфорта, а также количество времени, проводимого в помещении, увеличат потребность в энергии в будущем [1]. Растет понимание того, что многие здания работают не так, как задумали их проектировщики. Типичные здания потребляют на 20 % больше энергии, чем необходимо, из-за неправильной конструкции, неисправного оборудования, неправильно настроенных систем управления и несоответствующих рабочих процедур [2] и [3]. Для оптимизации энергопотребления оценка данных о потреблении энергии в здании в режиме реального времени является востребованной и развивающейся областью энергетического анализа зданий. Несмотря на проверку эффективности с использованием различных доступных передовых моделей, системы здания могут не соответствовать ожидаемым характеристикам из-за различных неисправностей. Плохо обслуживаемое, изношенное и неправильно контролируемое оборудование тратит, по оценкам, от 15 % до 30 % энергии, используемой в коммерческих зданиях [4] и [5]. Таким образом, существует большой потенциал для разработки автоматических, быстро реагирующих, точных и надежных схем обнаружения и диагностики неисправностей (ОДН) для обеспечения оптимальной работы систем в целях экономии энергии. Системы управления и контроля энергопотреблением могут собирать и хранить огромное количество данных о потреблении энергии. Поэтому требуются мощные и эффективные инструменты для извлечения ценной информации из огромных объемов доступных данных и преобразования ее в организованные знания. Было опубликовано несколько исследований, посвященных методам автоматического обнаружения аномальных данных о потреблении энергии в зданиях. В [6] представлен метод преобразования данных о потреблении энергии в информацию и учет еженедельных изменений в потреблении энергии путем группировки дней недели с аналогичным потреблением энергии. Надежный статистический метод используется для определения того, значительно ли потребление энергии отличается от предыдущего потребления энергии.

Чтобы найти закономерности в наборе данных, важно классифицировать данные, прежде чем обнаруживать выбросы в энергопотреблении здания. Кластеризация и классификация – это два распространенных метода, используемых в интеллектуальном анализе данных для поиска скрытых закономерностей в наборах данных. Несколько методов классификации для решения проблемы ОДН в данных, генерируемых моделью ПВУ ПОВ, обсуждаются в [7]. В некоторых исследовательских работах [8] [9] и [10] были представлены методы классификации, включая подход с использованием коробочных графиков и алгоритм распознавания образов. Лью и др. [11] использовали надежный статистический алгоритм для обнаружения аномального потребления электроэнергии в здании и добились хороших результатов.

Выбросы – это случаи, в которых значения данных сильно отличаются от значений данных для большинства случаев в наборе данных. В последнее время в основном обсуждаются методы, основанные на статистике, на расстоянии, на отклонениях и на плотности. Методы, основанные на статистической теории, используют алгоритм экстремального стьюдентифицированного отклонения (ЭСО) для обнаружения аномального потребления

энергии, достигающий хороших результатов [12] и [11].

В данном исследовании три различных метода интеллектуального анализа данных используются для анализа в режиме реального времени ежечасно регистрируемых данных об энергии и энергопотреблении для освещения в офисном здании. Была проведена классификация и кластеризация ежечасно записанных данных с использованием алгоритмов ДКиР, k-средних и ПКПШОП соответственно. С помощью методов ДКиР и к-средних обнаружение очевидных выбросов в каждом классе и кластере было выполнено с использованием алгоритма обобщенного экстремального стьюдентизированного отклонения (ОЭСО) и статистического метода боксплот (boxplot). В методе ПКПШОП выбросы обнаруживаются непосредственно при анализе конкретного кластера, в котором они изолированы. Было проведено сравнение вышеупомянутых методов распознавания образов и кластеризации, подчеркнув потенциал и пределы каждого подхода для анализа обнаружения неисправностей. Экспериментальные результаты показывают эффективность предложенных подходов в автоматическом обнаружении аномального потребления энергии, что может повысить производительность труда операторов зданий за счет сокращения времени на обнаружение неисправностей.

Материалы и методы исследования.

1. Построение и описание данных. Тематическое исследование, выбранное для анализа обнаружения неисправностей, представляет собой офисное здание, расположенное в Алматы, Казахстан. Здание состоит из трех этажей и подвала, соединенного через большую сторону со вторым зданием. Здание оборудовано системой мониторинга, направленной на сбор данных о потреблении энергии (электрической и тепловой) и состоянии окружающей среды. Кроме того, каждая комната/офис в здании были оснащены датчиком присутствия. Были проведены эксперименты с набором данных, относящимся к потреблению энергии для искусственного освещения только для первого этажа. На этом этаже расположены 13 офисов разного размера площадью от 15 до 37 м^2 и две карточные комнаты площадью около 21 м^2 каждая. В каждом офисе/помещении установлено различное количество люминесцентных ламп (каждая мощностью 55 Вт) в диапазоне от 4 до 8. В двух карточных комнатах установлены 12 ламп мощностью 55 Вт каждая. Чтобы определить ненормальное энергопотребление освещения, в качестве зависимых переменных для моделей рассматриваются такие характеристики, как среднее почасовое потребление энергии и пиковая потребность (максимальная мощность). Энергия офисного освещения и энергопотребление были проанализированы за декабрь и январь.

Кроме того, независимыми переменными, которые были записаны с часовым шагом по времени, являются: присутствие людей, количество активных комнат (комната считается активной, если присутствует хотя бы один человек), глобальная солнечная радиация, время, дата и день недели. Для проверки надежности и эффективности предложенных методов 24 и 25 января были созданы два искусственных сбоя. В эти дни в конце рабочего времени с меньшим количеством людей между 17:30 и 18:00 включалось все искусственное освещение офисов на первом этаже, создавая пик спроса на энергию.

В следующем разделе представлено краткое теоретическое описание методов классификации, кластеризации и обнаружения выбросов, используемых в работе. Во второй части анализ обнаружения неисправностей для потребления энергии освещения выполняется с использованием трех различных методов с целью сравнения возможностей каждого метода в обнаружении в основном созданных искусственных неисправностей.

2. Дерево классификации и регрессии (ДКиР). Алгоритм ДКиР основан на деревьях классификации и регрессии. ДКиР – это двоичное дерево решений, которое строится путем

многократного разделения родительского узла на два дочерних узла, начиная с корневого узла, содержащего всю обучающую выборку. ДКиР может легко обрабатывать как числовые, так и категориальные переменные и полезен для надежного обнаружения выбросов. Из записанных данных строится дерево решений, которое может быть легко преобразовано в правила классификации. Методология ДКиР обычно состоит из трех частей [13]:

1) построение максимального дерева. Классификационное дерево строится в соответствии с правилом разбиения. Каждый раз данные должны быть разделены на две части с максимальной однородностью. Мера примеси Джини в узле t определяется как

$$i(t) = \Sigma_{k \neq l} p\left(\frac{k}{t}\right) p(\frac{l}{t}) , \quad (1)$$

где k – индекс класса; $p\left(\frac{k}{t}\right)$ – условная вероятность класса k при условии, что мы находимся в узле t.

2) выбор дерева правильного размера. Оптимизация размера дерева важна, потому что структура деревьев может оказаться очень сложной и состоять из сотен уровней. На практике можно использовать два алгоритма обрезки: оптимизацию по количеству точек в каждом узле и перекрестную проверку.

3) классификация новых данных. По набору вопросов в дереве каждое из новых наблюдений попадет в один из конечных узлов дерева. Новому наблюдению присваивается доминирующее значение класса/ответа терминального узла, к которому принадлежит это наблюдение.

3. Кластеризация. Выбранные алгоритмы можно разделить на две категории: методы разделения и методы, основанные на плотности. Эти методы требуют определения метрики для вычисления расстояний между объектами в наборе данных. В анализируемом примере расстояния между объектами измеряются с помощью евклидова расстояния, вычисленного по нормализованным данным. Методы секционирования подразделяют набор данных из n объектов на k непересекающихся разделов, где k < n. Общий критерий для выполнения секционирования назначает объекты одному и тому же кластеру, когда они находятся близко, и разным кластерам, когда они находятся далеко друг от друга по отношению к определенной метрике. Методы разделения способны находить только кластеры сферической формы, если только кластеры не разделены хорошо, и чувствительны к наличию выбросов. k-средние [14] – популярный метод, который относится к этой категории. Методы, основанные на плотности, предназначены для работы с кластерами несферической формы и менее чувствительны к наличию выбросов. Целью этих методов является идентификация участков пространства данных, характеризующихся высокой плотностью объектов. Плотность определяется как количество объектов, находящихся в определенной области n-мерного пространства. Общая стратегия заключается в исследовании пространства данных путем увеличения существующих кластеров до тех пор, пока количество объектов в их окрестностях не превысит заданный порог. ПКПШОП [15] – это метод, основанный на плотности, рассмотренный в нашем тематическом исследовании.

3.1. Метод разделения (k-среднее значение). K-среднее [14] требует в качестве входного параметра количество разделов k, на которые должен быть разделен набор данных. Он представляет каждый кластер со средним значением объектов, которые он объединяет, называемым центроидом. Алгоритм основан на итеративной процедуре, которой предшествует этап настройки, где k объектов набора данных случайным образом выбираются в качестве начальных центроидов. Каждая итерация выполняет два шага. На первом шаге каждый объект присваивается кластеру, центр тяжести которого находится ближе всего к

этому объекту. На втором этапе центроиды перемещаются путем вычисления среднего значения объектов внутри каждого кластера. Итерации продолжаются до тех пор, пока k центроидов не изменятся. K-среднее значение эффективно для кластеров сферической формы. Различные формы кластеров обнаруживаются только в том случае, если кластеры хорошо разделены. Подобно другим методам разбиения, k-среднее чувствителен к выбросам и требует предварительного знания количества кластеров.

3.2. На основе плотности (ПКПШОП). ПКПШОП [15] требует двух входных параметров, используемых для определения порога плотности в пространстве данных: действительного числа r и целого числа минут. Область высокой плотности в пространстве данных представляет собой n-мерную сферу с радиусом r, которая содержит по меньшей мере объекты minPts. ПКПШОП – это итеративный алгоритм, который перебирает объекты в наборе данных, анализируя их окрестности. Если существует более 10 объектов, расстояние от которых до рассматриваемого объекта меньше r, то объект и его окрестности создают новый кластер. ПКПШОП эффективен при поиске кластеров произвольной формы и способен идентифицировать выбросы как области с низкой плотностью в пространстве данных. На эффективность алгоритма сильно влияет настройка параметров r и minPts.

4. Методы обнаружения выбросов. Удаленное наблюдение или выброс – это наблюдение, которое, по-видимому, заметно отличается от других элементов выборки, в которой оно встречается, или выброс – это наблюдение (или подмножество наблюдений), которое, по-видимому, не согласуется с остальной частью этого набора данных. Выбросы возникают из-за человеческой ошибки, ошибки прибора, изменений в поведении систем или сбоев в системах. В этом исследовании для обнаружения аномального потребления энергии использовались алгоритмы боксплот и обобщенные экстремальные стьюдентизированные отклонения (ОЭСО) с множеством выбросов. Боксплот – это распространенный статистический метод для выявления скрытых закономерностей в наборе данных, который также может быть легко уточнен для выявления отклоняющихся значений данных.

Процедура ОЭСО с множеством выбросов представляет собой модифицированную версию теста с экстремальным отклонением от нормы (ЭСО), предложенного [16], который позволяет находить множественные выбросы. Для выполнения метода необходимо установить два параметра: вероятность α неправильного объявления одного или нескольких ложных выбросов и верхний предел nu ожидаемого числа потенциальных выбросов. На основе указаний [17] было оценено потенциальное число потенциальных выбросов, найдя наибольшее целое число, удовлетворяющее неравенству n <0,5(n – 1), где n - количество наблюдений в наборе данных X : {x1, x2, x3, . . ., xn}. Метод позволяет обнаруживать значения выбросов в наборе данных путем вычисления и сравнения двух следующих важных параметров:

1) *ith*-е экстремальное студенческое отклонение Ri, определяемое из равенства

$$R_i = \frac{|x_{e,i} - \overline{x}|}{s}, \tag{2}$$

где $x_{e,i}$ – крайний элемент в наборе X, который наиболее удален от среднего $\overline{x}$ элементов в наборе X;

2) *ith*-е критическое значение λ_i, определяемое из (3):

$$\lambda_i = \frac{t_{n-i-1,p}(n-i)}{\sqrt{\left(n-i-1+t^2_{n-i-1,p}\right)(n-i+1)}}, \tag{3}$$

где $t_{n-i-1,p}$ – распределение Стьюдента с $(n-i-1)$ степенями свободы, p – вероятность

хвостовой области, определяемая из (4):

$$\mathrm{p} = 1 - [\alpha/2(n - i + 1)] \,. \quad (4)$$

5. Z-баллы и модифицированные z-баллы. Стандартные баллы использовались для анализа выбросов, а модифицированные z-баллы использовались для количественной оценки того, насколько далеко и в каком направлении находится выброс от среднего значения типичных наблюдений. Измененные z-баллы определяются как:

$$z_m = \frac{x_{outlier} - \overline{x_{robust}}}{s_{robust}} \,, \quad (5)$$

где z_m – модифицированный стандартный балл, $x_{outlier}$ – исходное значение выброса, x_{robust} – среднее значение не-выбросов в наборе данных, s_{robust} – стандартное отклонение отсутствия выбросов в наборе данных

Результаты и их обсуждения.

6. Классификация и анализ. Метод дерева классификации и регрессии был использован для анализа данных об энергии освещения и энергопотреблении в офисном здании. В данных максимальное количество людей и активных комнат составляют 21 и 14 соответственно. Значения солнечной радиации варьируются от 0 до 476,45 Вт/м^2. В среднем количество людей и активных комнат в выходные дни и после 21:00 в будние дни равно нулю. Количество людей выше в период с 09:00 до 16:00, то есть 16 и более человек. Анализ чувствительности данных показывает, что как энергия, так и мощность связаны с другими переменными, т. е. люди, солнечная радиация, дневные и активные помещения, поэтому можно сделать вывод, что экстремальные значения как на графиках энергии, так и на графиках последовательности мощности не являются определенными выбросами, поскольку существуют и другие переменные, которые могут повлиять на потребление. Поэтому важно классифицировать набор данных в аналогичных условиях, прежде чем обнаруживать выбросы.

Как для энергии, так и для мощности были разработаны отдельные деревья решений с учетом дня, времени, активного количества комнат, количества людей и солнечной радиации в качестве независимых переменных. Алгоритм дерева классификации и регрессии (ДКиР) был использован для процесса выращивания деревьев с максимальной глубиной дерева 5 с использованием обоих методов обрезки, описанных в разделе 2. Данные были разделены на 9 и 10 классов по энергии и мощности соответственно. Классы, построенные для энергии и мощности, были проанализированы отдельно. Краткое описание ключевых характеристик приведено в табл. 1 и 2 соответственно, при этом значения солнечной радиации менее 150 Вт/м^2 считаются более низкими, а более 150 Вт/м^2 – более высокими.

Таблица 1. Энергетические классы и краткое описание характеристик каждого класса

Класс	Время	Присутствие людей	Активная комната	Солнечное излучение	День
3	В основном ранним утром	Ноль или один	Ноль или один	В основном ноль	Чт-Пт
6	18:00-21:00	В основном < 7	В основном < 7	Ноль	Будни
8	Выходные – весь день Вечера в будние дни	Ноль или один	Ноль или один	Ноль, кроме выходных	Выходные и Пн-Ср
9	07:00, 08:00	80 % ≤10	80 % ≤10	Более низкие значения	Будни
11	06:00-07:00	Ноль	Ноль	В основном	Выходные

				ноль	
12	06:00-07:00	Ноль	Ноль	Ноль	Пн-Ср

Окончание таблицы 1

Класс	Время	Присутствие людей	Активная комната	Солнечное излучение	День
13	Разное время	В основном ≥ 7	В основном ≥ 7	Более низкие значения	Будни
15	12:00-16:00	≥ 10	≥ 10	Более высокие значения	Будни
16	09:00-11:00 и 17:00	≥ 10	≥ 10	Более высокие значения	Будни

Таблица 2. Классы мощности и краткое описание характеристик каждого класса

Класс	Время	Присутствие людей	Активная комната	Солнечное излучение	День
6	18:00-21:00	80% ≤ 5	80% ≤5	Ноль	Будни
7	06:00-08:00	Ноль	Ноль	Нулевые или более низкие значения	Чт-Пт
9	Вечером и ранним утром	В основном ноль	В основном ноль	В основном ноль	Чт-Пт
10	Выходные – весь день, в будние дни – раннее утро	Ноль	Ноль	Ноль, кроме выходных	Выходные и Пн-Ср
11	07:00, 08:00	В основном ≤ 10	В основном ≤ 10	Более низкие значения	Будни
13	06:00-08:00	Ноль	Ноль	В основном ноль	Выходные
14	06:00-08:00	Ноль	Ноль	В основном ноль	Пн-Ср
15	Разное время	Почти 60 % ≥ 10	Почти 60 % ≥ 10	Более низкие значения	Будни
17	09:00-17:00	Почти 70 % ≥ 10	Почти 60 % ≥ 10	Средний диапазон	Будни
18	09:00-15:00	В основном ≥ 10	В основном ≥ 10	Более высокие значения	Будни

Точечные диаграммы для всех энергетических классов показывают, что классы 6, 9, 11, 13 и 15 являются чистыми. Аналогично для классов мощности чисты 6, 7, 10, 11, 13, 14 и 18. Точечные диаграммы только для класса 3 (энергия) и класса 17 (мощность) представлены на рис. 1. Из этих графиков можно предположить, что ненормальное потребление энергии должно существовать в обоих классах.

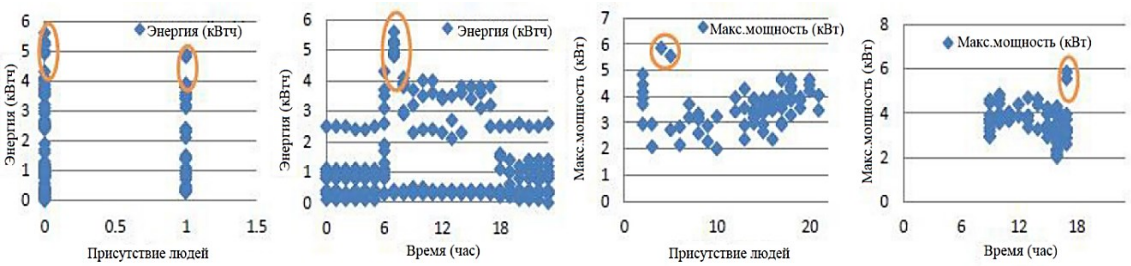

Рисунок 1. Точечные диаграммы для энергии (класс 3) и мощности (класс 17) – ненормальное потребление обведено кружком

Два различных метода обнаружения выбросов, описанных в разделе 4, были применены к каждому классу как для энергии, так и для мощности. Величина отклонения от нормы была определена с использованием надежных оценок среднего и стандартного отклонения (модифицированные z-баллы), чтобы показать результаты. Выбросы, обнаруженные для энергии и энергопотребления, подтверждают правильность построения классов и алгоритмов обнаружения выбросов. Например, в классе 3 (энергия) потребление выше при нулевом количестве людей и активных помещениях. Большинство обнаруженных аномальных потреблений приходится на раннее утро, т.е. с 06:00 до 07:00, когда потребление энергии почти равно тому, которое наблюдается у 15 или более человек в рабочее время. Аналогично в классе 8 (энергия) наиболее ненормальное потребление приходится на 08:00, когда в здании нет или мало людей. При анализе среднего часового потребления энергии оба метода обнаружения выбросов не смогли обнаружить искусственные неисправности, даже если они были вставлены через ДКиР в том же классе, т.е. 16. Результаты, полученные с помощью классов мощности и алгоритма обнаружения выбросов, были довольно хорошими, поскольку они смогли обнаружить искусственные неисправности, присутствующие в классе 17 (мощность). Также можно сделать вывод, что искусственные сбои были связаны с ненормальным максимальным энергопотреблением, а не с потреблением энергии.

На рис. 2 показан график последовательности почасового зарегистрированного энергопотребления и модифицированный график z-балла для класса 17 (мощность) с двумя искусственными выбросами. Из рисунка видно, что обнаружение выбросов затруднено только при использовании графика последовательностей. Выбросы в отдельных классах в основном являются пиковыми значениями и могут быть легко обнаружены, в то время как в последовательных данных то же самое невозможно.

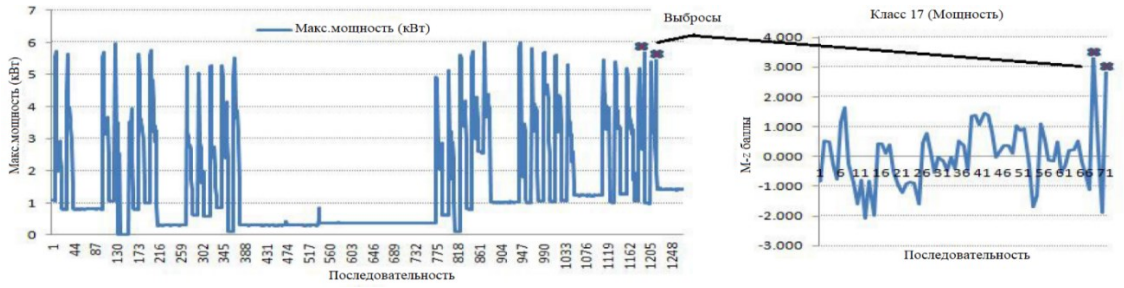

Рисунок 2. График последовательности степеней и график оценок M-z для класса 17, показывающий искусственные выбросы

7. Кластеризация и анализ. Алгоритмы кластеризации могут быть использованы для обнаружения интересных корреляций между отслеживаемыми показателями (например, энергией, присутствием людей, активным помещением и солнечным излучением), чтобы автоматически определять, какие записи отражают ненормальное потребление энергии. Как время, так и день не могут использоваться в качестве независимых переменных в методах кластеризации, следовательно была исследована возможность обнаружения неисправности без информации о временной структуре с помощью k-среднее и ПКПШОП. Поскольку мы фокусируемся только на подмножестве показателей, используемых для

управления методом классификации, записанные реальные данные как по энергии, так и по энергопотреблению были разделены на дневное, ночное время и выходные дни для управления алгоритмами кластеризации. Среди доступных подходов к кластеризации мы сосредоточились на алгоритме k-средних и подходе ПКПШОП. Перед выполнением кластеризационного анализа записанные реальные данные были нормализованы с помощью стандартного метода оценки (z-баллы).

7.1. Применение k-средних. Алгоритм k-средних является популярным алгоритмом кластеризации данных. Однако одним из его недостатков является требование, чтобы перед применением алгоритма было указано количество кластеров k. В этом исследовании была проведена иерархическая кластеризация для определения количества кластеров с применением метода Уорда для оценки основных компонентов. Из графика агломерации путем определения шага, на котором «коэффициенты расстояния» делают больший скачок, было выбрано количество кластеров k. Для формирования энергетических кластеров использовались нормированные значения энергии, присутствия людей, активного помещения и солнечной радиации, а также аналогичные значения для мощности. Для обнаружения выбросов в каждом кластере два метода обнаружения выбросов (ОЭСО и боксплот), описанные ранее, были применены к каждому кластеру как для энергии, так и для мощности. Исходя из результатов, полученных с помощью классификационного анализа, краткое резюме кластеров для энергии и для мощности только для дневных данных приведено в табл. 3 и 4 соответственно.

Таблица 3. Краткое описание дневных кластеров для энергии

№	Присутствие людей	Активная комната	Солнечное излучение	Стандартное отклонение энергии [кВтч]
1	Почти 60 % ≥ 10	Почти 60 % ≥ 10	Более высокие значения	0.6203
2	Почти 80 % ≤ 5	Почти 80 % ≤ 5	Более низкие значения	0.9578
3	Всегда ≥ 7	Всегда ≥ 7	Различный диапазон	0.9111
4	Почти 85 % ≤ 3	Почти 85 % ≤ 3	В основном ноль	0.4387

Таблица 4. Краткое описание дневных кластеров для мощности

№	Присутствие людей	Активная комната	Солнечное излучение	Стандартное отклонение энергии [кВтч]
1	Почти 60 % ≤ 5	Почти 60 % ≤ 5	Более высокие значения	0.5944
2	Почти 95 % ≤ 5	Почти 95 % ≤ 5	Более низкие значения	1.8910
3	Всегда ≥ 10	Всегда ≥ 10	Более высокие значения	0.6361
4	Почти 60 % ≥ 15	Почти 65 % ≥ 10	Более низкие значения	1.1503

Результаты показывают, что большинство кластеров не были чистыми, и аномальные значения были распространены. В энергетической кластеризации искусственные выбросы находятся в кластере 2, в то время как в кластеризации для мощности они присутствуют в кластерах 2 и 4. Кластер 4 нечист в обеих кластеризациях. В энергетическом кластере 4 наибольшее количество положительных ложных сигналов приходится на вечер и мало на раннее утро, и наоборот для энергетического кластера 4. Причиной этих положительных значений false может быть то, что некоторые переменные в наборе данных могут иметь те же значения, что и при реальных сбоях. Оба метода обнаружения выбросов не смогли

обнаружить искусственные сбои, присутствующие в энергетическом кластере 2. На рис. 3 были представлены графики последовательной мощности для данных за дневное время и модифицированные z-баллы для кластера 4 (мощность) с выделенными выбросами. Методы обнаружения выбросов смогли обнаружить искусственную ошибку в кластере 4, но не смогли обнаружить в кластере 2. После тщательного анализа результатов сделан вывод, что, хотя алгоритм k-средних может быть полезен для обнаружения аномальных значений, он не подходит для надежного обнаружения выбросов.

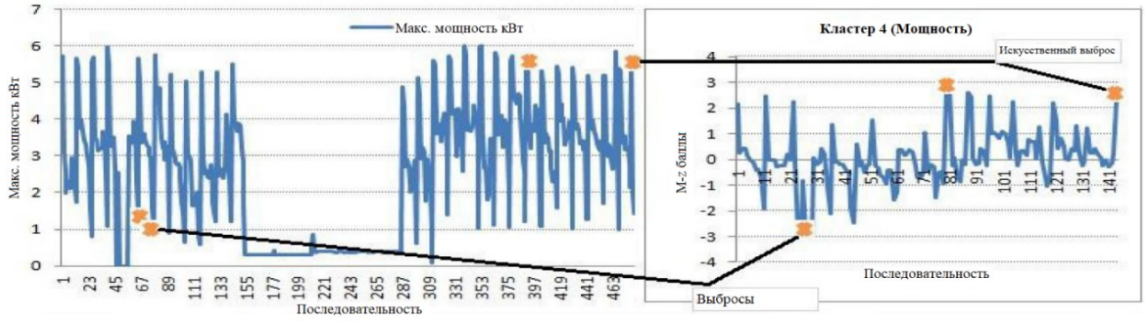

Рисунок 3. График последовательности мощности для дневного времени и график оценок M-z для кластера 4 с выбросами

7.2. Приложение ПКПШОП

Как представлено в разделе 3, ПКПШОП требует двух входных параметров, настройка которых сильно влияет на эффективность алгоритма: r и minPts. В отличие от k-средних, этот метод эффективен при выделении выбросов. Во всех обнаруженных кластерах нулевая метка кластера содержит все точки, идентифицированные как выбросы или шум. Для установки входных параметров были проведены различные тесты для всех данных (энергии и мощности) с использованием разных значений для этих параметров. Результаты показывают, что при сохранении значения одного параметра постоянным и изменении значения другого параметра обнаруженные кластеры отличаются. Например, если значение minPts остается постоянным, а значение r уменьшается, то количество кластеров и выбросов увеличивается. Результаты, полученные в результате этих тестов, были проанализированы, и были исследованы сходства внутри кластеров, чтобы выбрать соответствующие значения обоих входных параметров.

Результаты показывают, что ПКПШОП способен идентифицировать кластеры с одинаковой плотностью и с очень похожими данными. Наибольшим преимуществом ПКПШОП является его способность группировать все выбросы (включая шум) в один кластер, помеченный как нулевой кластер. В табл. 5 приведены фактические значения показателей, таких как энергия и т.д., включенные в нулевой кластер для данных об энергии в дневное время, при значениях входных параметров minPts = 5 и r = 0,7. Эти записи можно было бы рассматривать как ненормальное потребление энергии. Например, при меньшем количестве людей и активных помещениях (случаи 1, 5, 6, 10) потребление энергии является высоким и противоположным для случаев 2, 3, 4 и 9. Алгоритм ПКПШОП был эффективен при обнаружении реальных выбросов или шума, но не смог обнаружить две искусственные ошибки. Это может быть связано с природой искусственных сбоев, которые сильно связаны с переменной времени.

Выбросы, обнаруженные всеми тремя методами, анализируются и сравниваются. В

табл. 6 приведены некоторые выбросы с другими показателями (время, дата, мощность, присутствие людей, активное помещение и солнечная радиация). Эти выбросы являются общими по крайней мере для двух из трех предложенных методов интеллектуального анализа данных, включая один искусственный выброс (случай 7). Из таблицы видно, что обнаруженные выбросы являются реальными неисправностями, т.е. значение максимальной мощности велико для меньшего количества людей и активного помещения и противоположно для большего количества людей и активного помещения. Эти результаты показывают, что в целом предложенные методы были способны обнаруживать реальные неисправности или шумы.

Таблица 5. Фактические значения различных показателей, включенных в нулевой кластер (minPts =5, r = 0,7)

№	Энергия (кВтч)	Присутствие людей	Активная комната	Солнечное излучение (Вт/м^2)
1	4.3	2	2	343.6
2	1.4	13	10	190.3
3	1.8	19	14	218.83
4	2.1	15	12	445
5	2.9	6	5	476.45
6	3.3	8	6	429.32
7	3.0	13	12	413.75
8	3.1	15	13	408.08
9	1.9	10	7	330.12
10	2.4	7	6	209.6

Таблица 6. Данные о некоторых распространенных выбросах

№	День	Дата	Время	Макс.мощ (кВт)	Присутствие людей	Активная комната	Солнечное излучение (Ват/м^2)
1	Пятница	07.12.2020	08:00	5.42	7	6	38.5
2	Вторник	11.12.2020	08:00	5.48	10	9	0.00
3	Пятница	14.12.2020	11:00	4.84	2	2	343.6
4	Понедельник	14.01.2021	07:00	5.98	4	4	1.83
5	Понедельник	14.01.2021	14:00	1.95	19	14	218.83
6	Вторник	15.01.2021	07:00	5.82	5	5	3.8
7	Пятница	25.01.2021	17:00	5.55	5	4	1.75

Заключение. В этой статье были проанализированы ежечасно записанные данные энергетической системы здания с использованием трех различных методов интеллектуального анализа данных для обнаружения аномального потребления энергии. При классификации и кластеризации k-средних два алгоритма обнаружения выбросов были применены к каждому классу и кластеру соответственно для обнаружения аномального потребления в одном и том же наборе данных. В то время как в ПКПШОП нулевая метка кластера содержит все значения, идентифицируемые как выбросы или шум. Были проанализированы результаты трех подходов, и было проведено сравнение с точки зрения их потенциала и пределов для анализа обнаружения неисправностей. Результаты этого исследования заключаются в следующем:

– Дерево классификации и регрессии с алгоритмом обнаружения выбросов ОЭСО отличается высокой точностью и корректностью. Метод способен обнаруживать две

искусственные неисправности и более правильно определяет, значительно ли потребление энергии отличается от предыдущего потребления с аналогичными данными.

– Экспериментальные результаты с использованием подхода k-средних показывают, что, хотя метод способен обнаруживать некоторое ненормальное потребление, включая одну искусственную неисправность, он не является наиболее подходящим методом для надежного обнаружения выбросов. В обнаруженных кластерах большинство из них загрязнены, и распространены аномальные значения энергопотребления.

– С помощью алгоритма ПКПШОП оба искусственных сбоя не обнаруживаются, но метод способен идентифицировать кластеры с одинаковой плотностью и с очень похожими данными. Наибольшим преимуществом ПКПШОП является его способность группировать все выбросы (включая шум) в один кластер, помеченный как нулевой кластер.

В заключение следует отметить, что подход дерева классификации и регрессии с использованием метода выбросов ОЭСО более эффективен при автоматическом обнаружении аномального потребления энергии. Методы кластеризации не способны обнаруживать ошибки, сильно связанные с переменной времени. Исследование поможет системам энергоменеджмента зданий (СЭЗ) в профилактическом обслуживании путем отслеживания и обнаружения аномального потребления энергии в общей энергетической системе здания. Кроме того, это сделает работу энергетиков зданий более продуктивной, поскольку им не придется вручную обнаруживать неисправности или шумы.

Список литературы

1. Data for the modelling of the future power system with a high share of variable renewable energy, Data in Brief 42 (2022) 108095.
2. Research on computer interactive optimization design of power system based on genetic algorithm, Energy Reports 7 (2021) 1-13.
3. Frequency control strategy for coordinated energy storage system and flexible load in isolated power system, Energy Reports 8 (2022) 966–979.
4. Lili H, Bo Z, Hongtao C, et al. Research on optimization model for overhaul of electric power system transmission under the condition of smart grid. Bol Tec/Tech Bull 2017;55(20): 393-8.
5. Prasad, A.; Belwin, E.J.; Ravi, K. A review on fault classification methodologies in power transmission systems: Part-II. J. Electr. Syst. Inf. Technol. 2019, 5, 61-67.
6. LI Jianning, MA Xiaoli, TAN Huamin, JIANG Chen, et al. A nomaly Diagnosis System for Low-Voltage Area Line Loss Based on Wireless Communication and Big Data Technology. Electricity and energy, 2019; 40(1) : 36-40.
7. Bao Y, Xu J, Liao S, et al. Field verification of frequency control by energy-intensive loads for isolated power systems with high penetration of wind power. IEEE Trans Power Syst 2018;33(6): 6098-108.
8. Ge, Z.; Song, Z.; Ding, S.X.; Huang, B. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. IEEE Access 2018, 5, 20590-20616.
9. S Katipamula, Michael R. Brambley, Methods for fault detection, diagnostics, and prognostics for building systems, A Review,Part I, HVAC&R Research 2005; 11 (1): 3-25.
10. Zhang Q, Ding J, Zhang D, et al. Reactive power optimization of high-penetration distributed generation system based on clusters partition. Dianli Xitong Zidonghua/Autom Electr Power Syst 2019;43(3):130-7.
11. D Liu, Q. Chen, K. Mori and Y. Kida, A Method for detecting abnormal electricity energy consumption in buildings, Journal of Computational Information Systems, 2010, 6 (14) 4887-4895.
12. Aftab MA, Hussain S, Ali I, et al. Dynamic protection of power systems with high penetration of renewables: a review of the traveling wave based fault location techniques. Int J Electr Power Energy Syst 2020;144(1): 1-13.
13. Valtierra-Rodriguez, M. Fractal dimension and data mining for detection of short-circuited turns in transformers from vibration signals. Meas. Sci. Technol. 2019, 31, 025902.
14. B.H. Juang, L.R. Rabiner, The segmental K-Means algorithm for estimating parameters of hidden Markov models. IEEE Transactions on acoustics, speech, and signal Processing, 1990,38(9), 1639-1641.

15. M. Ester, H.P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, proceedings of the 2nd international conference on knowledge discovery and data mining, 226-231.
16. B. Rosner, Percentage points for generalized esd many-outlier procedure, Technometrics.
17. S. Wu, Jian Q. Sun, Cross-level fault detection and diagnosis of building HVAC systems, Building and Environment, 46(2011) 1558-1566