

УДК 004.94

МОДЕЛИ ПОТРЕБЛЕНИЯ ЭЛЕКТРОЭНЕРГИИ В МНОГОКВАРТИРНЫХ ЖИЛЫХ ДОМАХ НА ОСНОВЕ МНОГОМЕРНЫХ АДАПТИВНЫХ РЕГРЕССИОННЫХ СПЛАЙНОВ

Федосин А.С.

ФГБОУ ВПО «Мордовский государственный университет им. Н.П. Огарева», Саранск,
e-mail: nsdfxela@gmail.com

Энергетический бенчмаркинг – эффективный и относительно простой способ оценки потенциала энергосбережения жилого сектора. Несмотря на наличие более точных и достоверных технических методов, статистический анализ данных об энергопотреблении может быть применен как начальный шаг при решении поставленной задачи в условиях ограничений по времени или бюджету. Данная статья описывает способ проведения бенчмаркинга на основе многомерных адаптивных регрессионных сплайнов. Для исследования создан набор данных, использующий информацию АСКУЭ и биллинговой системы г. Саранска. Разработаны модели уровня квартиры и МКД. Для оценки качества моделей различных уровней иерархии определены значения коэффициента детерминации. Полученные результаты показывают, что модели уровня квартиры лучше аппроксимируют исходные данные при максимальной степени взаимодействия признаков, равной двум. Проведена кластеризация 122 МКД на основе значений регрессионных остатков с помощью метода К-средних. Каждый дом отнесен к одной из групп, отмеченных оценками от «неудовлетворительно» до «отлично». Такие оценки характеризуют эффективность использования электроэнергии в отдельных МКД.

Ключевые слова: многомерные адаптивные регрессионные сплайны, электроэнергия, энергосбережение, регрессия, АСКУЭ, бенчмаркинг

APARTMENT HOUSES RESIDENTAL ELECTRICITY CONSUMPTION MODELS BASED ON MULTIVARIATE ADAPTIVE REGRESSION SPLINES

Fedosin A.S.

Ogarev Mordovia State University, Saransk, e-mail: nsdfxela@gmail.com

Energy efficiency benchmarking is effective and simple way to evaluate residential sector's energy saving potential. Although many confidential and more precise engineering methods are known today, statistical analysis of energy consumption data is a very good starting point in case of limited time or budget. This article provides a new method of benchmarking that uses multivariate adaptive regression splines. SCADA-timeseries and information from billing system of Saransk are combined to form the dataset. Then set apartment-level and house-level models built. For evaluating regression's goodness of fit the study provides R-squared values for models built on different hierarchy levels. The obtained results show that apartment-level models with degree of interaction equal to 2 better approximate the data. K-means clustering of house level model's residuals has been used for benchmarking of 122 apartment houses. Each house is assigned to one of groups with label from «poor» to «very good» which correspond to relative level of energy efficiency. These labels might be useful for policy making in energy management.

Keywords: multivariate adaptive regression splines, electricity, energy saving, regression, SCADA, benchmarking

Тема энергоэффективности Российской экономики на протяжении уже длительно-го времени является актуальной. Работы над повышением этого показателя ведутся в разных областях: принимаются законы, тарифы, исследуются новые материалы и технологии. Одной из отраслей, которые, по мнению исследователей, обладают значительным потенциалом энергосбережения, является ЖКХ [1].

Среди шагов, предпринимаемых в направлении повышения эффективности потребления энергии многоквартирными домами, действенным и важным является установка систем АСКУЭ. В то же время необходимо понимать, что организация достоверного и оперативного учета энергопотребления сама по себе не является залогом успеха. Применение АСКУЭ для коммерческого учета расхода электроэнергии в МКД, конечно же, удобно для энергосбытовых

компаний, управляющих компаниями и добросовестных конечных потребителей. В то же время, большой объем данных, который агрегируют АСКУЭ, по мнению автора, может быть использован для анализа с целью выявления закономерностей и правил, которые могут обеспечить информационно-аналитическую поддержку мер повышения энергоэффективности многоквартирных домов.

Другой важной мерой является составление энергетических паспортов зданий и определение классов энергетической эффективности объектов. Вне сомнений, такой энергоаудит мог бы значительно прояснить реальную картину, помочь определить многоквартирные дома, которые в первую очередь нуждаются в принятии мер по повышению энергоэффективности. Однако необходимо учесть, что полноценный энергоаудит – достаточно дорогостоящая

операция, и ее проведение даже в масштабах среднего города, не говоря уже о городах-миллионниках, может стоить бюджету значительных сумм. Так, по сообщениям дирекции коммунального хозяйства и благоустройства г. Саранска, сделанным в 2011 году, в ближайшей (на тот момент) перспективе на фасадах всех многоэтажных зданий должны были появиться таблички с указанием класса энергетической эффективности. На момент проведения данного исследования (начало 2016 года) такие таблички установлены не были [3]. Это может подтверждать тезис о том, что проведение полноценного энергоаудита в масштабах целого города – сложная и дорогостоящая процедура.

Более оптимальным и гораздо менее дорогостоящим решением, пригодным для первичной оценки ситуации, изучения общей картины в масштабах района или города, может стать проведение процедуры бенчмаркинга. Бенчмаркинг – практика сравнения измеренной фактической производительности устройства, процесса, объекта или организации по отношению к своим более ранним состояниям, аналогичным объектам или установленным нормам с целью информирования и мотивации к повышению производительности [2]. В контексте задачи повышения энергоэффективности жилого сектора российских городов бенчмаркинг – процедура количественного определения потенциала энергосбережения многоквартирных жилых домов, а также ранжирования некоторого множества таких домов на основе совокупности выявленных значений.

Процедура бенчмаркинга требует определения некоторых индикаторов потенциа-

ла энергосбережения как максимально инвариантных к внешним условиям величин. Прежде всего, исследователи стремятся некоторым способом исключить зависимость величины потребляемой электроэнергии от площади жилья (корреляция между этими величинами очевидна). Как следствие, в рамках исследований часто анализируется удельная величина, такая как $\text{кВт}\cdot\text{ч}/\text{м}^2$, или же $\text{кВт}\cdot\text{ч}/\text{фут}^2$. Данную величину в зарубежных исследованиях называют Energy Use Intensity (EUI), или Интенсивность Энергопользования. При проведении исследований, как правило, считается, что объем потребляемой энергии линейно зависит от площади изучаемого объекта (будь то жилое или коммерческое здание) [4]. Действительная же ситуация, по мнению автора, может не соответствовать подобному утверждению. Этот тезис проиллюстрирован на рис. 1: данные, изображенные на левом графике только с большим трудом можно назвать линейно зависимыми. Кроме того, изображение справа свидетельствует о возможном наличии связи между площадью дома и средним потреблением в отдельных квартирах, чего при линейной зависимости суммарного потребленного объема энергии в доме от общей площади наблюдаться не должно.

Применение столь простого индикатора, как EUI, по мнению ряда исследователей, является достаточно спорным приемом, поэтому зачастую также учитывается влияние метеофакторов: вычисление показателей для зданий на основе значений, собранных при разных погодных условиях, может привести к неправильным выводам. Для коммерческих зданий важно учитывать часы работы.

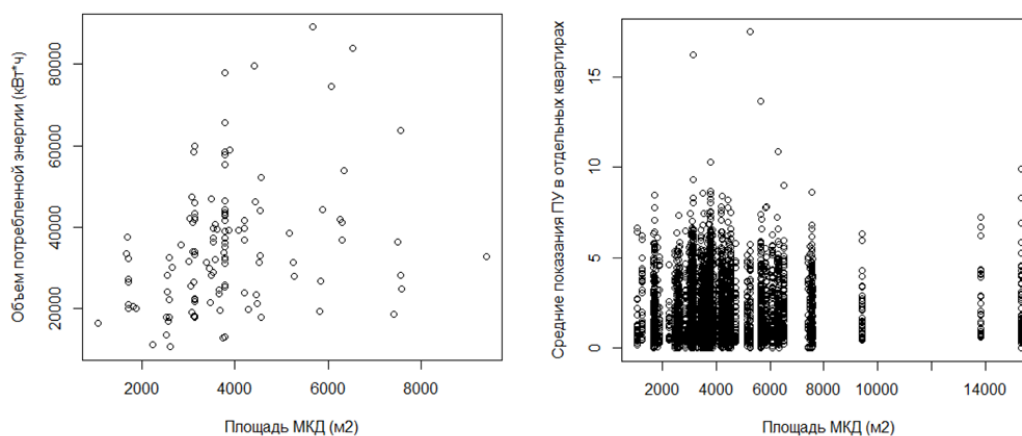


Рис. 1. Зависимость суммарного объема электроэнергии в исследуемой группе МКД от площади (на основе данных биллинговой системы и АСКУЭ для МКД г. Саранска)

Моделирование величины потребляемой электроэнергии – один из эффективных способов информационно-аналитической поддержки мер повышения энергоэффективности многоквартирных жилых домов. Несмотря на заметное внимание к проблемам энергосбережения в различных отраслях российской экономики, меры статистического наблюдения за уровнем эффективности использования энергии в жилых домах на сегодняшний день не получили должного распространения [1]. В то же время многие зарубежные ученые уделяют значительное внимание упомянутым моделям в своих исследованиях.

Согласно [9], модели потребления электроэнергии могут быть отнесены к двум категориям: нисходящие модели и восходящие. Первые исследуют особенности потребления энергии в рамках целых регионов и влияние на эту величину общих макроэкономических показателей, таких как ВВП. Вторые ставят объектом своих исследований конечных потребителей – отдельные домовладения или же многоквартирные дома. В рамках восходящих моделей, в свою очередь, выделяют инженерные и статистические методы. Инженерные методы базируются прежде всего на данных технического характера, описывающих исследуемые объекты, т.е. строения. Статистические методы используют, помимо прочего, исторические данные о потреблении электроэнергии.

Задачу моделирования величины потребляемой энергии можно представить следующим образом:

Заданы временные ряды, состоящие из матрицы признаков X (где x_{ij} представляет собой вектор значений j -х признаков в день t для объекта учета, в котором установлен i -й прибор учета) и измеренных значений потребленной электроэнергии $y_i(t) \in R$ – показания i -го прибора учета за день t ($i \in I$, I – множество приборов учета, $t \in N$ – натуральное число). Требуется восстановить регрессионную зависимость $y = f(X)$.

Используемая модель, с точки зрения дисперсионного анализа, может быть представлена в следующем виде:

(Наблюдаемое значение) = $\sum$ [Параметры, описывающие определяемые эффекты] + $\sum$ [Случайные величины, описывающие неопределяемые (остаточные) эффекты].

С учетом особенностей предметной области, мы можем заметить, что случайные величины, описывающие остаточные эффекты, в общем могут быть отнесены к двум группам: не известные нам в процессе факторы, характеризующие конструктивные особенности здания, и факторы, описывающие особенности потребления энергии жильцами МКД

(т.е. конкретными людьми, имеющими свои привычки, потребности и возможности, во многом определяющие общую картину).

Самым простым, но в то же время получившим широкое распространение подходом стала линейная регрессия вида:

$$\sum_{i=1}^n \epsilon_i^2 EUI_i = a + b_1 x_{1,i} + \dots + b_j x_{j,i} + \epsilon_i,$$

где $x(j, i)$ – факторы, влияющие на моделируемую величину, a и b – некоторые коэффициенты, ϵ – случайная ошибка [8]. В рамках данного метода также делается предположение о линейной зависимости EUI от исследуемых факторов, однако в реальных данных зависимость может иметь другой вид, что проиллюстрировано рис. 1.

Описаны способы построения регрессионных моделей энергопотребления на основе методов машинного обучения [8]. Целесообразность их применения к данной задаче обусловлена нелинейной зависимостью между регрессорами и переменной отклика, а также наличием других сложных и неочевидных взаимосвязей в данных.

Следует отметить, что построение регрессионной модели как таковой не является конечной целью анализа – выводы позволяет делать исследование остатков:

$$\epsilon_i = y_i - f_i,$$

где f_i – отклики, посчитанные по модели, y_i – эмпирические остатки [10]. На основе анализа величин остатков для различных зданий производится процедура кластеризации, в конечном итоге определяющая каждое из зданий в некоторую группу. Каждая группа характеризуется оценкой (от неудовлетворительной до отличной), позволяющей сделать вывод о потенциале энергосбережения данного строения.

Предложенные Джеромом Фридманом многомерные адаптивные регрессионные сплайны (MARS) [6] обладают рядом преимуществ перед другими регрессионными методами:

- модели MARS – более гибкие, нежели модели, построенные при помощи линейной регрессии;

- модели MARS являются более простыми для понимания, их результаты проще интерпретировать (по сравнению, например, с нейронными сетями или SVM);

- MARS позволяет работать с численными и категориальными признаками;

- благодаря разделению исходных данных на области базисными функциями, MARS позволяет определять выбросы;

- кусочно-линейные базисные функции с узлом в некоторой точке, применяемые

в MARS, позволяют лучше аппроксимировать численные переменные по сравнению с кусочно-константными функциями, используемыми в регрессионных деревьях;

- MARS не требует значительных мер по подготовке входных данных;

- для моделей MARS, как правило, характерен оптимальный баланс смещения/дисперсии (bias-variance tradeoff).

Перечисленные преимущества являются существенными для проведения энергетического бенчмаркинга, требующего учета большого числа разнородных факторов.

MARS относится к разряду непараметрических процедур (в противоположность параметрическим, таким, как например, линейная регрессия). Параметрический подход заключается в предположении, что функция отклика Y имеет некоторую предписанную функциональную форму. Непараметрические процедуры в свою очередь подразумевают использование модели, не описываемой конечным числом параметров. Несмотря на то, что параметрические методы просты в плане построения и интерпретации результатов, их «гибкость» бывает зачастую недостаточной при решении реальных задач. Для того, чтобы справиться с этой проблемой, используют подход, при котором непараметрические аппроксимирующие функции строятся отдельно на некоторых областях исходных данных. При этом сначала решается задача нахождения оптимального числа таких областей (в русскоязычной литературе их иногда называют гильотинными разбиениями).

Одним из известных непараметрических подходов является рекурсивная разделяющая регрессия (Recursive partitioning regression, RPR). RPR – адаптивный алгоритм построения аппроксимирующей функции, позволяющий работать с большим числом регрессоров.

Пусть дан некоторый набор регрессоров $X = \{x_1, x_2, \dots, x_n\}$, определенный в пространстве $D \subset \mathbb{R}_n$, зависимая переменная y , а также ряд из N наблюдений $\{y_i, x_{i1}, x_{i2}, \dots, x_{in}\}_{i=1}^N$. Фактическая взаимосвязь между y и X может быть выражена следующей формулой:

$$y = f(x_1, x_2, \dots, x_n) + \varepsilon,$$

где y – некоторая неизвестная функция, а ε – ошибка. RPR аппроксимирует $f(x)$ следующим образом:

$$\hat{f}(X) = \sum_{m=1}^M a_m B_m(X), \quad (1)$$

где $\{a_i\}_1^M$ – коэффициенты модели, подобранные с целью максимального соответствия исходным данным. M – число областей, на которые разделен массив данных $R_m \subset D$

и в то же время число базисных функций B_m модели. Базисная функция определяется выражением

$$B_m(X) = \begin{cases} 1, & X \in R_m \\ 0 & \text{в остальных случаях} \end{cases}$$

RPR является относительно мощным методом, при этом обладая рядом недостатков. Одним из них является наличие разрывов на границах разбиений. Многомерные адаптивные регрессионные сплайны представляют собой обобщение RPR, позволяющее избежать многих ограничений последнего. В основе метода MARS лежат кусочно-линейные базисные функции (получившие в западной литературе название «hockey stick», благодаря форме графика), отражающие связь между x и x^* :

$$x^* = \max(x - c, 0), \quad (2)$$

где c называют узлом базисной функции. Отметим, что подобным образом может быть использована функция

$$x^* = \max(-(x - c), 0), \quad (3)$$

график которой зеркально отражен по сравнению с графиком функции (2). Пример подобного графика представлен на рис. 2.

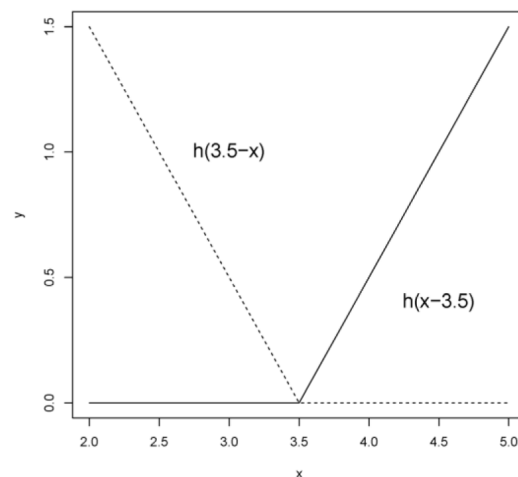


Рис. 2. Пример графика hinge-функции

Метод многомерных адаптивных регрессионных сплайнов видоизменяет модель (1), представляя её в следующем виде:

$$\hat{f}(X) = a_0 + \sum_{m=1}^M a_m H_m(X),$$

где $H_m(X)$ – базисные функции, задаваемые выражением

$$H_m(X) = \prod_{k=1}^{K_m} \left[\max(s_{k,m}(x_{v(k,m)} - t_{k,m}, 0)) \right].$$

Здесь $v(k, m)$ – регрессоры, соответствующие базисной функции H_m , K_m – степень взаимодействия между переменными $v(k, m)$, $t_{k,m}^{kp}$ задает местоположение узла функции H_m , а $s_{k,m}$ – коэффициент принимающий значения -1 и $+1$ и задающий форму базисной функции (2) или (3). Метод MARS осуществляет построение модели в два этапа: прямая шаговая и обратная шаговая регрессии. На первом этапе (прямая шаговая регрессия) вся выборка D разделяется на ряд областей, для каждой из которых производится подбор параметров модели. Если число таких областей не задано, то на этом этапе мы получим по одной базисной функции для каждого значения. Кроме того, придется учитывать все возможные взаимосвязи между регрессорами [11].

На втором этапе отбрасываются те из базисных функций, которые меньше всего влияют на точность модели. Для того, чтобы обеспечить приемлемое время работы алгоритма, степень взаимодействия между регрессорами и максимальное количество базисных функций предлагается задавать пользователю.

Для оценки обобщающей способности алгоритма MARS используется критерий кросс-проверки (MGCV):

$$MGCV = \frac{1}{N} \sum_{i=1}^N [y_i - \hat{f}(X_i)]^2 / \left[1 - \frac{\hat{C}(M)}{N} \right],$$

где $\left[1 - \frac{\hat{C}(M)}{N} \right]$ – штраф, учитывающий изменение дисперсии значений, а $\hat{C}(M)$ определяется следующим образом:

$$\hat{C}(M) = C(M) + d \times M.$$

Здесь $C(M)$ – число параметров модели, а d – еще один штраф, значение которого, как правило принимают равным 3 [11].

Методу MARS свойственны некоторые недостатки:

- по сравнению с MARS, регрессионные деревья обладают значительным преимуществом в скорости;
- как и для других методов непараметрической регрессии, для MARS не представляется возможным непосредственно определить значения доверительных интервалов параметров (в отличие от линейной регрессии), и для проверки модели приходится использовать кросс-валидацию;
- модели MARS не обеспечивают столь же качественного приближения данных, как деревья решений, построенные методом градиентного роста, однако их построение

занимает значительно меньшее время, а результаты проще интерпретировать;

• открытые программные реализации MARS (такие, как, например, пакет earth для языка R) не позволяют обучать модели на основе данных, содержащих пропущенные значения, в то же время регрессионные деревья могут быть использованы для решения подобной задачи.

Следует заметить, что подготовка данных (включающая очистку и замещение пропущенных значений [3]), подбор наиболее оптимальных гиперпараметров для MARS-моделей, а также осуществление перекрестной проверки могут позволить существенно уменьшить влияние недостатков выбранного метода на процедуру моделирования и конечный результат. Именно этой задаче и посвящена данная работа.

Для обучения MARS-моделей в рамках данного исследования был использован набор данных о многоквартирных домах г. Саранска в совокупности с историей показаний автоматизированной системы коммерческого учета потребления энергоресурсов (АСКУПЭ) и сведениями, характеризующими температуру воздуха и длину светового дня на рассматриваемых временных отрезках. Отдельно следует упомянуть, что данные организованы иерархически: некоторые признаки соответствуют уровню дома (например дата постройки или тип кровли), другие – уровню квартиры (средний возраст проживающих, наличие льгот и т.д.), отдельные признаки определены для всего множества исследуемых объектов, то есть для целого города (среднесуточная температура, длительность светового дня). Для обучения были использованы следующие признаки.

Приборы АСКУПЭ фиксируют расход электрической энергии в квартирах дважды в сутки (днем и ночью), передавая значения на единый сервер при помощи ЛВС. Обрабатываемые значения представляют собой набор временных рядов $y(t_i)$, где $t_i \in R$. К сожалению, как и для любых сложных технических систем, для АСКУЭ характерны ошибки. Грамотная эксплуатация и настройка системы позволяют минимизировать количество проблемных данных, однако не позволяют избежать их совсем. Использование ошибочных, некорректных сведений крайне нежелательно при проведении технического или коммерческого учета. Успех работы алгоритмов регрессионного анализа в то же время зависит от качества данных. При наличии большого числа ошибок наблюдения процедура восстановления аппроксимации неизвестной функции отклика может оказаться неуспешной, что

не позволит сконцентрировать внимание на важных особенностях зависимости. Это обуславливает необходимость проведения операции очистки исходных данных.

Большинство зарубежных публикаций в качестве объектов исследований описывают или частные домовладения, или крупные коммерческие здания (офисы, торговые центры, моллы и т.д.). Отечественные урбанистические реалии значительно отличаются от зарубежных. Жилой фонд в городах нашей страны представлен преимущественно многоквартирными зданиями. Более того, по-прежнему не получили массового распространения системы кондиционирования

(об изначально предусмотренных в зданиях системах говорить не приходится). Преимущественное использование центрального отопления также определяет иные паттерны потребления электроэнергии. Кроме того, многие особенности взаиморасчетов между потребителями и поставщиками абсолютно не характерны для развитых стран дальнего зарубежья (примером тому может служить распространенность льгот среди различных групп населения) [7]. Поэтому в рамках данного исследования впервые делается попытка применения статистических инструментов для изучения особенностей энергопотребления в МКД.

Таблица 1

Список признаков, использованных для обучения модели

Признак	Тип	Возможные значения	Псевдоним	Уровень иерархии
Площадь дома	Вещественная	8120, 5300, 4100 ...	totalArea	Дом
Этаж	Целочисленная	1, 3, 10 ...	Floor	Квартира
Площадь квартиры	Вещественная	42.0, 72.5, 101.1 ...	area	Квартира
Количество проживающих	Целочисленная	0, 2, 5 ...	residents	Квартира
Средний возраст проживающих	Вещественная	35.7, 78.3, 54.7 ...	avgAge	Квартира
Время суток	Категориальная	День/Ночь	timeOfDay	Город
Дата постройки здания	Целочисленная	1974, 1982, 2001 ...	yearBegin	Дом
Среднесуточная температура	Целочисленная	+ 3, – 20, 0 ...	avgTemp	Город
Продолжительность светового дня	Целочисленная	9, 8, 11...	dayLength	Город
Тип кровли	Категориальная	Металлическая Мягкая Шатровая Шифер Безрулонная	roofType	Дом
Тип фундамента	Категориальная	Бетонный ленточный Бутовый Бутовый ленточный Железобетонный Железобетонные блоки Керамзитобетонные блоки Сборный ж/б	foundationType	Дом
Тип отопления	Категориальная	Централизованное Автономное	heatingType	Дом
Тип стен	Категориальная	Кирпичные Панельные Сборный железобетон Монолитные	wallType	Дом
Широта	Вещественная	54.11	attitude	Дом
Долгота	Вещественная	45.11	longitude	Дом
Выходной или праздничный день	Булева	TRUE, FALSE	weekend	Город
Тип оборудования, используемого для обогрева пищи	Категориальная	Электроплиты, газовые плиты	eqType	Дом
Наличие льгот на оплату электроэнергии у проживающих	Булева	TRUE, FALSE	isBenefit	Квартира

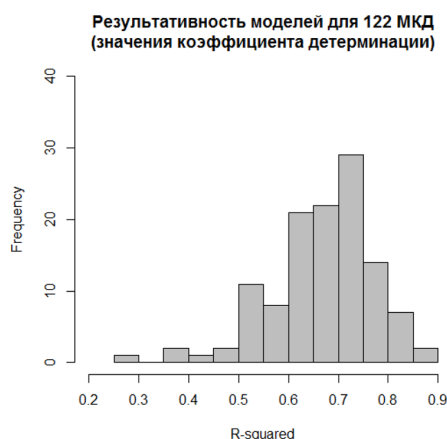


Рис. 3. Распределение значений RSS восходящих моделей MARS уровня дома (максимальная степень взаимодействия $D = 1$)

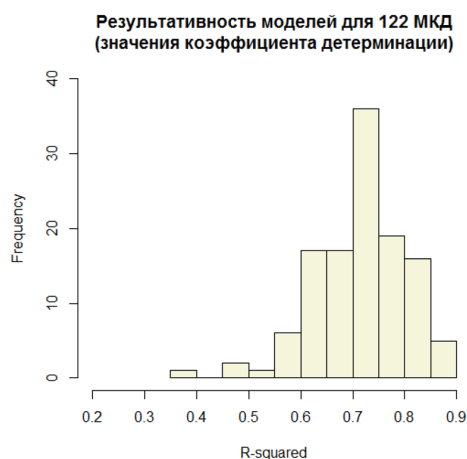


Рис. 4. Распределение значений RSS восходящих моделей MARS уровня дома (максимальная степень взаимодействия $D = 2$)

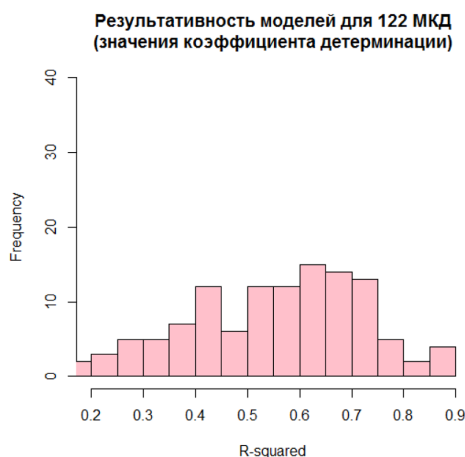


Рис. 5. Распределение значений RSS для восходящих моделей MARS уровня дома (максимальная степень взаимодействия $D = 5$)

Исходные данные, как упоминалось выше, детализированы на уровнях квартир и домов. На начальном этапе для каждого многоквартирного дома была построена отдельная модель уровня квартиры (т.е. позволяющая оценивать значения потребляемой электроэнергии в квартире с заданными параметрами).

Целями вычислительных экспериментов являются:

1. Исследование возможности применения MARS в задачах моделирования энергопотребления и энергетического бенчмаркинга.

2. Выявление наиболее оптимальных гиперпараметров для MARS-моделей.

3. Определения наиболее значимых факторов, влияющих на моделируемую величину (на основе совокупности данных о множестве МКД г. Саранска).

Ряд гиперпараметров MARS-моделей задается пользователем. Среди них – пороговое значение RSS, при котором завершается первый проход алгоритма, некоторые характеристики процедуры автоматической кросс-валидации, а также максимальная степень взаимодействия различных переменных в рамках одной базис-функции (иными словами, максимальное число переменных, которые могут быть задействованы в одной базисной функции). Последний параметр позволяет сделать некоторые выводы о природе данных, в частности, о взаимосвязи некоторых регрессоров.

Распределение значений коэффициента детерминации

$$R^2 = 1 - \frac{\sum (Y - X)^2}{\sum (Y - Y_{\text{сред.}})^2}$$

для таких моделей со значением параметра максимальной степени взаимодействия D , равным 1, представлено на рис. 3. Распределение для значений $D = 2$ и $D = 5$ отображено соответственно на рис. 4 и 5.

Сравнение распределений коэффициентов детерминации для моделей с различными значениями D показывает, что значение зависимой переменной лучше всего определяется при условии вхождения максимум двух членов с различными регрессорами в одну базисную функцию.

Профили суммарного удельного объема потребленной электроэнергии двух МКД, модели которых имеют соответственно лучший ($> 0,9$) и худший ($< 0,4$) коэффициенты детерминации представлены на рис. 6, а и б.

Алгоритм MARS позволяет определять степень влияния различных факторов на переменную отклика [6]. Для этого могут быть использованы три оценки.

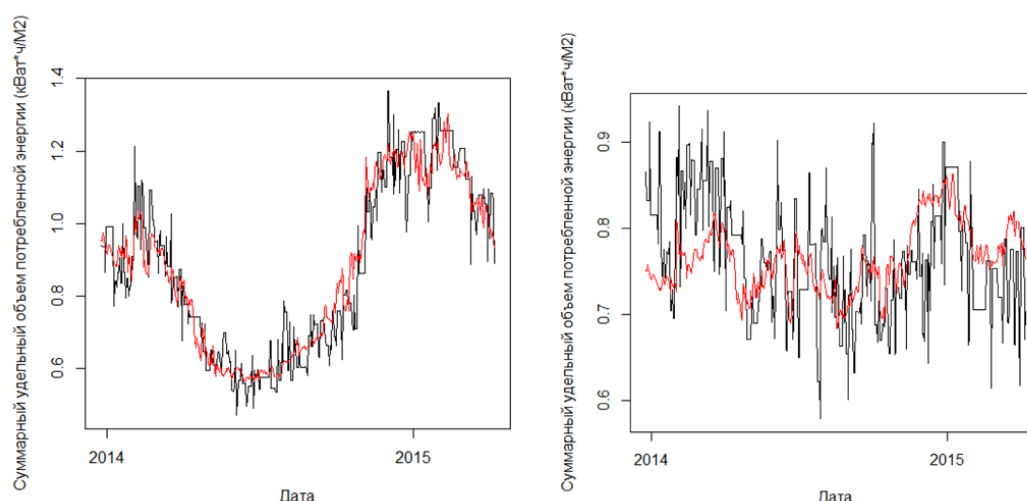


Рис. 6. Смоделированные (светлая линия) и эмпирические (темная линия) профили суммарного удельного объема потребленной электроэнергии. Модели, показавшие результаты $R^2 > 0,9$ (а) и $R^2 > 0,4$ (б) ($D = 2$). https://www.youtube.com/watch?v=17QbQF_9XM

Таблица 2

Оценки значимости самых важных регрессоров для четырех моделей с разными коэффициентами детерминации ($D = 2$)

№ модели	Размерность	R^2	Регрессор «Residents» (число зарегистрированных жильцов)			Регрессор «AvgAge» (средний возраст жильцов)			Регрессор «IsBenefit» (Наличие льгот)			Регрессор «DayLength» (Длина светового дня)			Регрессор «T» (Средняя дневная температура)		
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
10	30	0,74	20	45,2	45,4	27	100	100	17	33,8	34,0	12	20,7	21,1	—	—	—
51	37	0,62	29	58,8	59,5	31	100	100	28	58,6	59,2	20	25,9	27,3	17	21,2	22,7
84	38	0,59	33	100	100	30	93,8	93,7	31	96,4	96,3	30	89,2	89,5	23	43,0	44,2
120	15	0,46	10	53,6	54,1	13	100	100	—	—	—	5	13,8	15,4	3	7,5	9,3

Первая основывается на подсчете числа вхождений переменной в подмножества S_i множества базисных функций, конструируемого на втором (обратном) этапе работы алгоритма. Такие подмножества имеют мощность c_i^s , где i изменяется в пределах 1 до n , а n – размерность итоговой модели. Для каждого i подмножество S – наилучшее (с точки зрения приближения функции) из всех возможных подмножеств данной мощности.

RSS-оценка вычисляется следующим образом: сначала определяется величина, на которую уменьшает показатель RSS каждое из подмножеств S_i . Затем эти величины суммируются на всех подмножествах, включающих нужную переменную.

GCV-оценка работает схожим образом, за исключением того, что вместо суммы квадратов остатков вычисляется упомянутый выше обобщенный критерий кросс-проверки.

Все три оценки для простоты интерпретации их значений масштабируют в интервале $[1, 100]$. Примеры значений описанных оценок для ряда моделей представлены в табл. 2, где для каждого из пяти самых значимых регрессоров («Residents», «AvgAge», «IsBenefit», «DayLength», «T») в колонках «1», «2» и «3» указаны значения соответствующих оценок.

Отдельный интерес представляют факторы уровня МКД. Для изучения их влияния на зависимую переменную (измеряемую в кВт·ч/м²) построена MARS-модель, основанная на агрегированных данных. Значение R^2 для данной модели составило 0,52. В табл. 3 представлены оценки значимости факторов (каждая из трех описанных оценок для каждого из регрессоров указана в соответствующем столбце «1», «2» и «3»).

Таблица 3

Оценки значимости самых важных регрессоров для модели уровня МКД

Регрессоры	Оценки		
	1	2	3
floor (Этажность)	31	100,0	100,0
houseAge (Год постройки)	30	91,3	91,3
attitude (Координаты широты)	28	89,7	89,7
longitude (Координаты долготы)	28	81,9	81,9
numberOfBenefits (Процент льготников среди жильцов)	18	54,8	54,8
roofType (Тип крыши)	13	42,2	42,2
avgAge (Средний возраст жильцов)	8	20,7	20,8

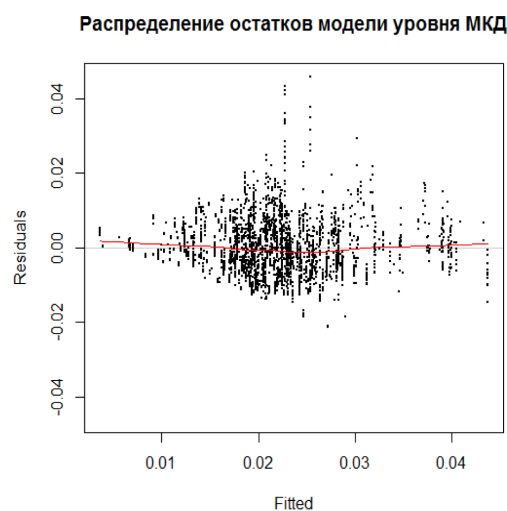


Рис. 7. Распределение остатков модели уровня МКД

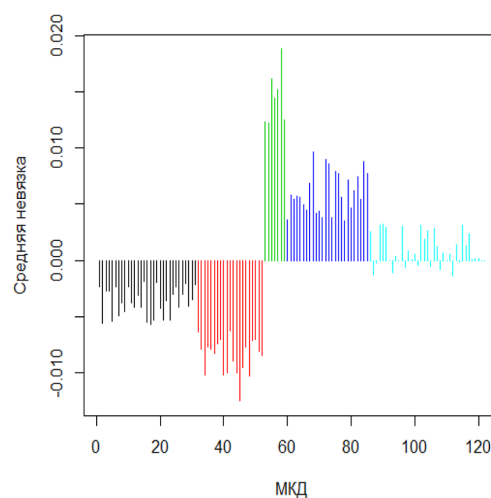


Рис. 8. Кластеризация 122 МКД на основе сведений о средних невязках

Из рис. 7, отражающего распределение остатков построенной модели, можно сделать выводы о гетероскедстичности модели. В то же время следует отметить, что важными критериями качества моделей многомерных регрессионных сплайнов являются прежде всего независимость остатков и их симметричность.

Для наглядной интерпретации модели представляется возможным осуществить кластеризацию остатков при помощи метода k-средних, который стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров:

$$\arg \min \sum_{i=1}^k \sum_{x_j \in S_i} x_j - \mu_i^2,$$

где k – число кластеров, S_i – полученные кластеры, $i = 1, 2, 3, \dots, k$ и μ_i – центры масс векторов $x_j \in S_i$.

Результат проведения данной операции (при $k = 5$) представлен на рис. 8. Каждому МКД соответствует отдельный столбец. Дома, принадлежащие разным кластерам, отмечены различными цветами (и упорядочены по индексу кластера). На оси ординат отражена средняя невязка, полученная в результате применения описанной выше регрессионной модели. Каждому из полученных кластеров можно дать некоторую относительную оценку, характеризующую ситуацию с эффективностью использования электроэнергии: от «неудовлетворительно» (черные столбцы на графике) до «отлично» (зеленые столбцы).

Выводы

В рамках данной работы было проведено исследование свойств моделей энергопотребления на основе многомерных адаптивных регрессионных сплайнов. Были построены

модели различного уровня, использующие в основе данные разной степени детализации. Выявлены факторы, которые наиболее значительно среди прочих исследованных влияют на моделируемую величину удельного энергопотребления на уровне квартиры и уровне многоквартирного дома, а также экспериментально установлено наличие взаимосвязей между регрессорами.

Среди факторов уровня МКД наиболее значимыми оказались этажность, год постройки, координаты (широта и долгота), процент льготников среди жильцов, тип крыши, средний возраст жильцов, среди факторов уровня квартир – число зарегистрированных жильцов, средний возраст, наличие льгот, а также длина светового дня и среднесуточная температура.

Наилучшим образом аппроксимируют данные модели, для которых значение максимальной степени взаимодействия регрессоров $D = 2$. Значения коэффициента детерминации R^2 для моделей уровня квартиры в среднем составили 0,716. Значение R^2 для модели уровня дома составило 0,52.

Предложен способ проведения первичного энергоаудита, основанный на статистическом анализе данных. Преимуществом данного способа является высокая степень независимости результатов от часто меняющихся факторов, таких как длина светового дня, средний возраст жильцов или среднесуточная температура. Результаты такой процедуры могут стать основой для принятия решения о необходимости детального обследования тех или иных многоквартирных зданий.

В целом результаты данного исследования позволяют сделать вывод о целесообразности применения многомерных адаптивных регрессионных сплайнов для

анализа данных об энергопотреблении в МКД.

Список литературы

1. Башмаков И.А. Тридцать мер политики для ускорения повышения энергоэффективности в российских зданиях [Электронный ресурс] // Центр по эффективному использованию энергии: сайт. – URL: http://www.cenef.ru/file/30_Policies.pdf (дата обращения: 05.06.2016).
2. Башмаков И.А. Рейтинг жилых и общественных зданий по уровню энергоэффективности / Башмаков И.А. // Энергосбережение. – 2016. – № 1. – С. 46–51.
3. В Саранске многоквартирные дома получают энергетический паспорт [Электронный ресурс] / Информационное агентство REGNUM. – Режим доступа: <http://regnum.ru/news/1437507.html>, свободный. – Заглавие с экрана. – Яз. рус. (дата обращения: 05.06.2016).
4. Федосин А.С., Савкина А.В. Проблемы качества данных в автоматизированных системах коммерческого учета потребления энергоресурсов // Прикаспийский журнал: управление и высокие технологии. – 2014. – № 2. – С. 158–164.
5. Chung W. Benchmarking the energy efficiency of commercial buildings / W. Chung, Y.V. Hui, Y.M. Lam // Applied Energy. – 2006. – № 83. – P. 1–14.
6. Friedman J.H. Multivariate adaptive regression splines / J.H. Friedman // The Annals of Statistics. – 1991. – № 19. – P. 1–141.
7. Kavousian A.A. Method to Analyze Large Data Sets of Residential Electricity Consumption to Inform Data-Driven Energy Efficiency [Electronic resource] / A. Kavousian, R. Rajagopal, M. Fischer. – Режим доступа: http://cife.stanford.edu/sites/default/files/WP130_0.pdf (дата обращения: 15.06.2016), свободный – Загл. с экрана. – Яз. Англ.
8. Shabunko V. Developing building benchmarking for Brunei Darussalam / V. Shabunko, C.M. Lim, S. Brahim, S. Mathew // Energy and Buildings. – 2014. – № 85. – P. 79–85.
9. Swan L.G. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques / L.G. Swan, V.I. Ugursal // Renewable and Sustainable Energy Reviews. – 2009. – № 13. – P. 1819–1835.
10. Wang E. Benchmarking energy performance of building envelopes through a selective residual-clustering approach using high dimensional dataset / E. Wanga, Z. Shenb, K. Grosskopf // Energy and Buildings. – 2014. – № 75. – P. 10–22.
11. Zareipour H. Forecasting the Hourly Ontario Energy Price by Multivariate Adaptive Regression Splines / H. Zareipour, K. Bhattacharya, C.A. Canizares // IEEE Transactions on power systems. – 2006. – № 20. – P. 1035–1042.