DOI: 10.17323/2587-814X.2022.4.7.18

Особенности применения методов, основанных на деревьях решений, в задачах оценки недвижимого имущества

М.Б. Ласкин[a]
E-mail: laskinmb@yahoo.com

Л.В.Гадасина[b]
E-mail: l.gadasina@spbu.ru

[a] Санкт-Петербургский ФИЦ РАН (СПИИРАН)
Адрес: Россия, 199178, Санкт-Петербург, Васильевский остров,14 линия, д. 39

[b] Санкт-Петербургский государственный университет, Центр эконометрики и бизнес аналитики
Адрес: Россия, 199034, Санкт-Петербург, Университетская наб., д. 7/9

Аннотация

Значительный рост интереса исследователей к применению в оценке недвижимого имущества методов, основанных на деревьях решений, обусловлен увеличивающимися потоками доступной рыночной информации, развитием методов машинного обучения, искусственного интеллекта, и ограниченными возможностями традиционных методов оценки недвижимого имущества. В то же время, распределение цен на недвижимость хорошо приближается логарифмически нормальным распределением, что проявляется в завышении предсказаний такими методами в области меньше средних значений имеющегося набора данных и занижении предсказаний в области выше средних значений. В статье показаны причины этих особенностей и предложен адаптивный алгоритм случайного леса, основной идеей которого является корректировка результатов работы базового алгоритма с помощью устранения смещения предсказанных значений. Результаты апробировались на ценах предложений объектов недвижимости в Санкт-Петербурге.

Ключевые слова: дерево решений, случайный лес, рынок недвижимости, ценообразующие факторы, оценка рыночной стоимости

Цитирование: Ласкин М.Б., Гадасина Л.В. Особенности применения методов, основанных на деревьях решений, в задачах оценки недвижимого имущества // Бизнес-информатика. 2022. Т. 16. № 4. С. 7–18. DOI: 10.17323/2587-814X.2022.4.7.18

Введение

В последнее время наблюдается значительный рост числа публикаций, посвященных нетрадиционным методам в оценке недвижимого имущества, ориентированным на большие выборки данных, и, в частности, методам машинного обучения. Интерес исследователей к данной тематике понятен: изменившаяся информационная среда и широкий выбор специализированных прикладных пакетов позволяет для целей оценки рассматривать методы, недоступные ранее по очевидным причинам, см., например [1–6]. Рассматриваются, например, метод гедонистического ценообразования, модели линейной регрессии, логарифмической или частично-логарифмической зависимости [7–10]. Предлагаются методы интеллектуального анализа данных, такие, как нейронные сети [11–15], метод опорных векторов [16], сравниваются результаты применения таких методов, как деревья решений, наивный байесовский классификатор и ансамблевый алгоритм AdaBoost [17]. Настоящая статья посвящена обсуждению смещений предсказаний, возникающих при применении методов, основанных на деревьях решений, в оценке недвижимого имущества и предлагает алгоритм устранения этих смещений. Интерес к этой группе методов подтверждается такими работами как [18–21]. Исследователи обращаются к методам машинного обучения, в частности методам, основанным на деревьях решений, в условиях, когда имеется обширный набор исходных данных, при этом нет априорных предположений о виде функции $F(\cdot)$, описывающей зависимость $V = F(x_1, x_2, ..., x_n)$ между выходной или зависимой переменной V, как правило являющейся ценой, и предикторами $x_1, x_2, ..., x_n$, являющимися ценообразующими факторами.

Преимуществом методов, основанных на деревьях решений, является то, что знание вида функции $F(\cdot)$ не требуется. Это, однако, не означает, что специфические особенности распределений цен на объекты недвижимого имущества не оказывают влияния на результаты работы таких алгоритмов. Метод построения одного решающего дерева заключается в последовательном разбиении всей области определения предикторов на подмножества меньшего размера. Каждому элементу такого подмножества присваивается значение среднего арифметического значений зависимой переменной на таком подмножестве. Это итеративная процедура, известна как рекурсивное разбиение:

1. На каждом шаге проводится разбиение на подмножества.

2. Каждое из полученных на предыдущем шаге подмножеств, в свою очередь, разбивается. В целом будет построено разбиение пространства независимых переменных (предикторов) $x_1, x_2, ..., x_n$ на какое-то количество (например, m) непересекающихся областей $R_1, R_2, ..., R_m$.

3. Для всех наблюдений, попавших в область R_j, $j = 1, 2, ..., m$, устанавливается одинаковое предсказание значения зависимой переменной V, которое равно среднему значению всех откликов, попавших в R_j. Основное правило разбиения на каждом шаге, это разбиение, обеспечивающее минимум среднеквадратичных отклонений RSS (residual sum of squares)

$$RSS = \sum_{j=1}^{m} \sum_{i \in R_j} (V_i - \hat{V}_{R_j})^2,$$

где $\hat{V}_{R_j}$ – среднее значение отклика у обучающих наблюдений из множества R_j.

С вычислительной точки зрения рассмотреть все комбинации разбиений на возможно большую глубину не представляется осуществимым. Поэтому используется основной принцип «жадных» алгоритмов: оптимальное разбиение (с точки зрения минимума RSS) определяется только на текущем шаге. Глубина разбиения может быть настолько большой, насколько позволяет объем данных. Возможные правила остановки: по количеству шагов, по количеству элементов в подмножествах (листьях дерева), по достижении заранее заданного улучшения результата на последующем шаге. Интерес представляет такое разбиение, при котором получаемые множества $R_1, R_2, ..., R_m$ содержат «достаточно много» элементов и стандартное отклонение значений отклика в каждом множестве от своего среднего значения не слишком велико и находится в пределах заранее заданной точности.

1. Методика проведения исследования

В задачах оценки объектов недвижимости применение метода дерева решений порождает дополнительную возможность – он позволяет разбить множество объектов на подмножества с меньшей дисперсией с более однородными объектами внутри каждого множества, которые можно исследовать отдельно.

Преимуществами применения деревьев решений являются:

1. Простая интерпретируемость модели.
2. Считается, что такой алгоритм отражает процесс принятия решений людьми.
3. Для одного дерева решений существует наглядное графическое представление.
4. Решающие деревья легко справляются с факторными и ранговыми переменными.

Недостатком таких алгоритмов, является невысокая точность прогноза (не достаточно мала дисперсия внутри каждого множества разбиения – на листьях). Этот недостаток можно устранить, применив ансамблевые методы, основанные на деревьях решений, например, случайный лес, градиентный или стохастический бустинг. Такие алгоритмы не позволяют получить интерпретируемый результат, однако позволяют проанализировать важность предикторов для предсказания отклика и позволяют работать с факторными переменными. В данной работе рассматривается метод случайный лес.

В задачах оценки недвижимого имущества алгоритмы, основанные на деревьях решений, следует применять с учетом особенностей выходной переменной V, поскольку существуют основания предполагать ее случайной величиной, распределенной логарифмически нормально. По-видимому, впервые на этот факт обратили внимание Айчинсон и Браун [22], и впоследствии это наблюдение было подтверждено в современных работах, как, например, [4]. В работах [23, 24] дано теоретическое обоснование причин появления такого вида распределения.

Случайная величина V называется распределенной логарифмически нормально с параметрами μ и σ, если случайная величина $\ln(V)$ распределена нормально с теми же параметрами. При этом математическое ожидание $E(V) = e^{\mu+\sigma^2/2}$, медиана $Median(V) = e^{\mu}$, мода $Mode(V) = e^{\mu-\sigma^2}$.

Доля значений эмпирического распределения, находящихся слева от $E(V)$ (меньше математического ожидания) может быть оценена как

$$\frac{1}{2} + \Phi\left(\frac{\sigma}{2}\right),$$

где $\Phi(\cdot)$ – функция Лапласа. Доля таких значений зависит от стандартного отклонения и увеличивается с его ростом.

Таким образом, если рассмотреть гипотезу о том, что результаты предсказаний алгоритмов, основанных на деревьях решений, также подчиняются логарифмически нормальному распределению, и, более того, образуют с наблюденными значениями совместное логарифмически нормальное распределение в смысле двумерного нормального распределения логарифмов, то становится понятно, что методы, основанные на деревьях решений, лучше применять не к ценам объектов недвижимости, а к их логарифмам. Поскольку на каждом шаге минимизируются среднеквадратичные отклонения RSS, от средних значений в подмножествах R_1, R_2, ..., R_m, внутри которых может сохраняться принцип логарифмически нормального распределения зависимой переменной V, то в случае применения ансамблевых алгоритмов, основанных на деревьях решений, предсказание результатов на тестовом множестве будет довольно точным только в области значений близких к средним значениям откликов. В области значений ниже среднего предсказания такими алгоритмами будут завышаться, выше среднего – занижаться, отклонение будет увеличиваться по мере приближения к границам эмпирических распределений. На соответствующей диаграмме, отражающей зависимость между истинными и предсказанными значениями, при приемлемой точности предсказаний будут наблюдаться облака рассеяния, вытянутые вдоль некоторой прямой линии, несколько смещенной относительно биссектрисы первого координатного угла путем поворота вокруг точки, с координатами среднего значения отклика и среднего значения предсказания. При этом большая часть результатов будет завышена, т.к. при логарифмически нормальном распределении большая часть возможных значений находится слева от среднего (меньше среднего). При этом для величины $\ln(V)$, имеющей нормальное распределение, области завышения и занижения предсказаний будут примерно одинаковыми по количеству элементов. При описанном подходе результаты предсказания остаются смещенными относительно истинных значений. Это наблюдается, например, в работе [20].

В данной работе предлагается применять адаптивный метод, основанный на корректировке результатов предсказания ансамблевого алгоритма случайный лес. Адаптация состоит в следующей процедуре. Множество исходных данных разбивается на три части: обучающая, валидационная и тестовая. На первом множестве проводится процедура обучения (подбора параметров) алгоритма случайный лес. Далее анализируется зависимость предсказанных значений для валидационного множества от их истинных значений. Проводится корректировка предсказания путем поворота облака рассеяния в координатах (значение отклика – предсказание) на

некоторый угол, устраняющий смещение от биссектрисы первого координатного угла, и пересчета предсказанных значений. На третьем шаге качество предсказаний алгоритма случайный лес с учетом корректировки проверяется на тестовом множестве.

Следует отметить, что в любом случае будет определена оценка рыночной стоимости (РС) не как наиболее вероятная цена, а как средняя цена (если алгоритм применялся к исходным ценам) или средняя геометрическая цена (если алгоритм применялся к логарифмам цен).

2. Применение методики к реальным рыночным данным

Апробируем описанную процедуру на следующем примере. Рассмотрим цены на вторичную жилую недвижимость в секторе масс-маркета в Санкт-Петербурге в феврале 2017 года, взятые из открытого источника (рекламное издание Бюллетень недвижимости №1765, февраль 2017 г., в наукометрических базах не индексируется), общее число записей в наборе данных после удаления некорректных объявлений – 4294. Построим предсказательную модель методом случайный лес. Зависимой (целевой) переменной является цена за 1 кв.м. вторичной жилой недвижимости в секторе масс-маркет в Санкт-Петербурге в феврале 2017 года или ее логарифм. Предикторами являются:

- количество комнат в квартире – количественная переменная,
- административный район места положения – факторная переменная,
- этаж – факторная переменная,
- количество этажей в доме – факторная переменная,
- жилая площадь – количественная переменная,
- общая площадь – количественная переменная,
- доступность метро – бинарная переменная,
- тип дома – факторная переменная,
- количество санузлов, их тип – факторная переменная.

Расчеты проводились в свободно распространяемом программном продукте R. Прежде всего, обратим внимание на несимметричность распределения цен (*рис. 1*).

Проверка нулевой гипотезы о следовании эмпирического распределения цен за 1 кв.м. логарифмически нормальному распределению связана с определенными трудностями. Объем выборки – 4294. Как справедливо отмечалось в [25], большинство распространенных и часто используемых критериев оказываются неработающими для выборок порядка даже одной тысячи наблюдений, так как их статистики существенно зависят от объема выборки. Поэтому возникает вопрос о поиске, помимо визуального соответствия, показанного на *рисунке 1*, дополнительных аргументов в пользу того или иного вида распределения. В этом контексте отметим работу [26], в которой предлагается для выдвижения нулевой гипотезы о том или ином виде распределения изучение соотношений между коэффициентом асимметрии и эксцесса наблюдаемой выборки. В данной работе для проверки соответствующих гипотез использовался метод, предложенный в работе [27]. Полученные значений *p*-value дают основания предполагать, что наблюдаемая выборка следует логарифмически нормальному закону, следовательно, есть основания решать предсказательную задачу в логарифмах.

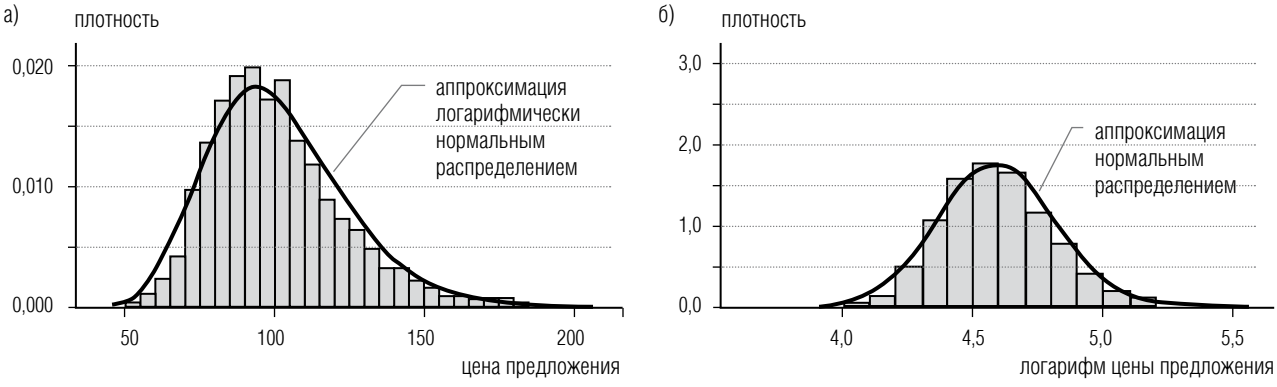

Рис.1. Распределение цен (слева) и их логарифмов (справа) вторичной жилой недвижимости в Санкт–Петербурге в феврале 2017 года.

Последовательно рассмотрим, какими получаются предсказания по одному дереву решений, по алгоритму случайный лес, затем проведем процедуру корректировки получаемого предсказания. С этой целью из исходной выборки (объем 4294 записей) методом случайного отбора формируется обучающее множество объемом 2000 записей, валидационное множество объемом 1000 записей, оставшиеся 1294 записи образуют тестовое множество, на котором проводится оценка качества модели.

Обученная на случайной выборке, состоящей из 2000 записей, модель дерева, дает диаграмму предсказаний цен (дерево решений, обрезано до 11 листов), показанную на *рисунке 2*.

Подобное разбиение в задачах оценки не лишено смысла, т.к. позволяет формировать группы с разным набором ценообразующих факторов, в которых ожидается свое, характерное для группы среднее значение и стандартное отклонение наблюденных значений от группового среднего. Предикторами, оказавшими существенное влияние на формирование дерева (по мере убывания влияния), оказались:

- административный район,
- тип дома,
- количество этажей в доме,
- количество санузлов, их тип,
- общая площадь,
- жилая площадь,
- доступность метро,
- этаж.

Качество предсказаний, показанных на *рисунке 2*, неудовлетворительное – слишком большим диапазонам значений предсказывается одинаковая цена (идеальными были бы предсказания, находящиеся на *рисунке 2* вблизи биссектрисы первого координатного угла). Эффективным средством борьбы с одинаковыми предсказаниями является алгоритм случайный лес – построение большого количества деревьев и усреднение результатов по каждому объекту. На *рисунке 3* показан результат работы алгоритма случайный лес для предсказания цены за 1 кв.м. вторичной жилой недвижимости для тестового множества. Каждое дерево в алгоритме строилось на основе 4-х предикторов, общее количество деревьев – 200.

Аналогичный рисунок, с характерным смещением предсказаний, можно увидеть в статье [20], посвященной алгоритмам, основанным на решающих деревьях при анализе рынка недвижимости в Анкаре. На *рисунке 3* виден характерный для совместного логарифмически нормального распределения, увеличивающийся с ростом цены разброс предсказаний, а также тот факт, что большая часть предсказаний (выше биссектрисы первого координатного угла, отмеченной черной линией) оказывается завышенной, что и следовало ожидать, учитывая несим-

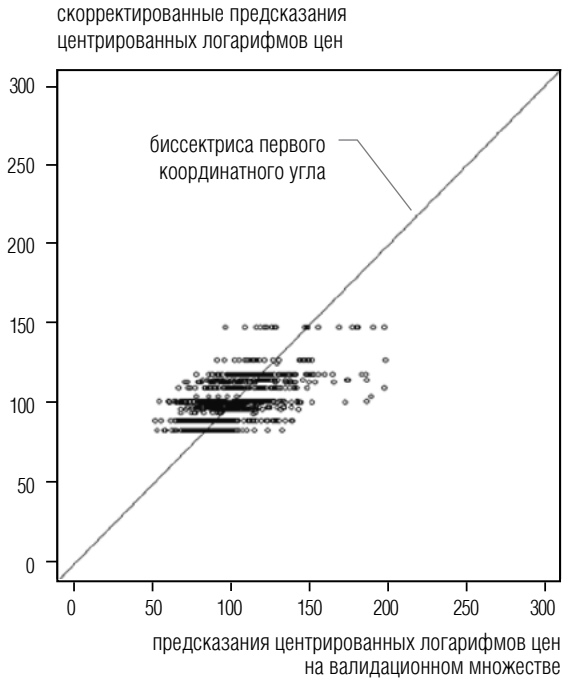

Рис. 2. Диаграмма предсказаний
на тестовом множестве по одному дереву решений.

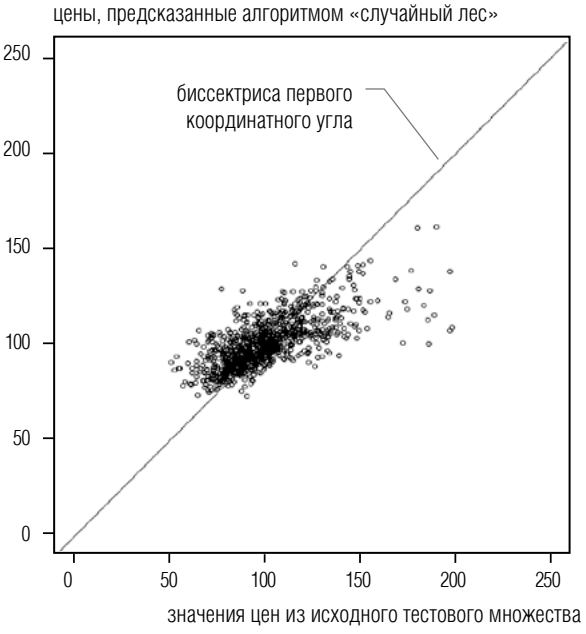

Рис. 3. Диаграмма предсказаний
на тестовом множестве по алгоритму «случайный лес».

метричность распределения цен в исходном множестве. Также отметим, что полученные таким образом предсказания являются предсказаниями средних значений (математических ожиданий в подмножествах R_1, R_2, ..., R_{21}), а не рыночной стоимости, как наиболее вероятной цены. Оценка рыночной стоимости, в действительности, несколько ниже.

Применим адаптивный метод случайного леса к логарифмам цен.

На *рисунке 4* показан результат работы алгоритма случайный лес для предсказания логарифма цены за 1 кв.м. вторичной жилой недвижимости (количество деревьев – 200, случайный отбор по 4 предиктора на каждом дереве, обучающая случайная выборка из 2000 записей).

На *рисунке 4* видно, что области завышения и занижения предсказаний приблизительно одинаковы, однако ось облака рассеяния имеет характерный тренд, несовпадающий с биссектрисой первого координатного угла. Следует также отметить, что полученные при потенцировании результатов предсказания величины являются предсказаниями медианных значений (средних геометрических в подмножествах R_1, R_2, ..., R_{21}), а не рыночной стоимости, как наиболее вероятной цены. Оценка рыночной стоимости и в этом случае находится несколько ниже.

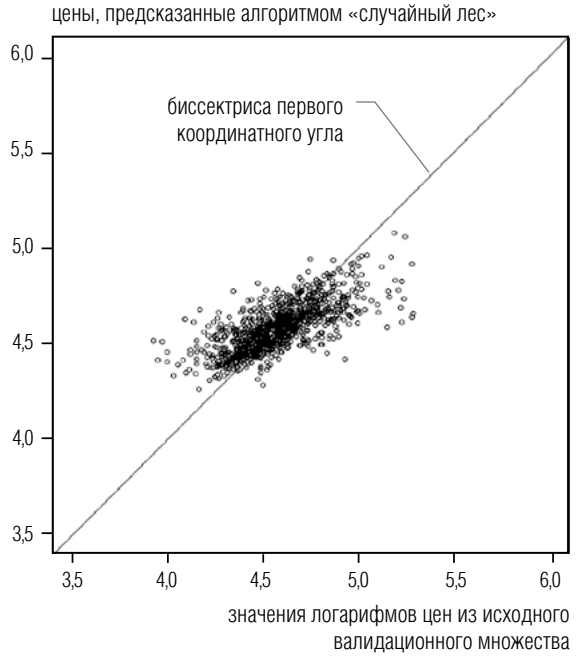

Рис. 4. Диаграмма предсказаний на валидационном множестве по алгоритму «случайный лес».

Результаты предсказаний алгоритма случайный лес можно поправить следующими преобразованиями результатов. С помощью метода линейной регрессии определим линейный тренд облака рассеяния, показанного на *рисунке 4*. Результат показан на *рисунке 5*.

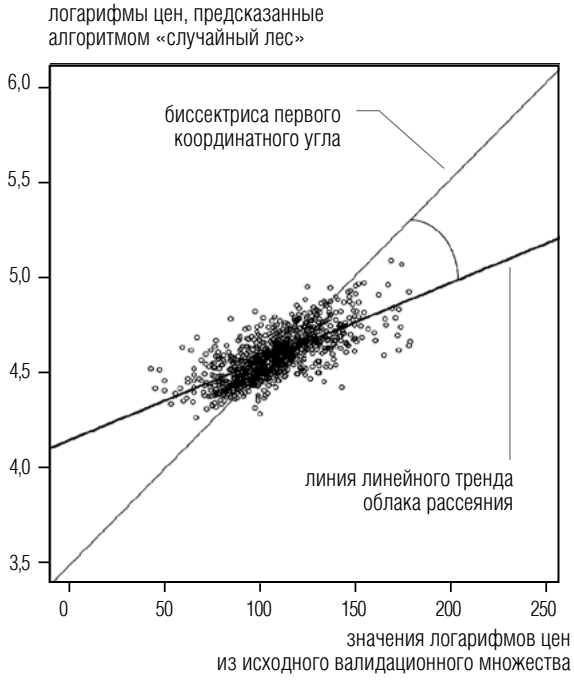

Рис. 5. Диаграмма предсказаний на валидационном множестве по алгоритму «случайный лес».

В данном примере уравнением линии тренда является

$$\ln(\hat{V}) = 0,40791 \cdot \ln(V) + 2,71920, \qquad (1)$$

где $\ln(V)$ – наблюденное значение логарифма цены; $\ln(\hat{V})$ – предсказанное значение логарифма цены.

Статистические характеристики полученной линии тренда: *p*-value теста Стьюдента для коэффициентов модели и критерия Фишера для модели в целом – машинный ноль. Стандартная ошибка – 0,086, т.е. разброс значений с вероятностью 0,99 находится в интервале ~ +/–26%. Относительно невысокое значение $R^2 = 0,5053$ не портит картины, т.к. в данном случае нет строгого следования линейной зависимости между $\ln(\hat{V})$ и $\ln(V)$, наши ожидания связаны с совместным нормальным распределением величин $\ln(\hat{V})$ и $\ln(V)$, для которого линейный тренд совпадает с главной осью эллипса рассеяния, подробнее о многомерном логарифми-

чески нормальном распределении см. [27]. Уравнение (1) соответствует линии, показанной на *рисунке 5* жирным черным цветом.

Для показанного на *рисунке 5* облака рассеяния получено значение среднеквадратического отклонения наблюденных значений от предсказанных равно 0,168. Остается откорректировать предсказания, показанные на *рисунках 3, 4* и *5*. Для этого все значения центрируются (вычитаются средние значения по горизонтальной и вертикальной осям), в новой системе координат, выполняется поворот облака рассеяния против часовой стрелки на угол, составляющий разницу между линией, заданной уравнением (1) (*рис. 5*, жирная линия) и биссектрисой первого координатного угла. Такой угол φ равен

$$\frac{\pi}{4} - \text{arctg}(0,40791).$$

Результат поворота показан на *рисунке 6*.

Для показанного на *рисунке 6* облака рассеяния получено значение среднеквадратического отклонения наблюденных значений от предсказанных равно 0,118. Таким образом, проведенная корректировка дает лучшее качество предсказания.

Теперь мы имеем два вектора значений: вектор предсказаний центрированных логарифмов цен на валидационном множестве (обозначим его y^*) и вектор скорректированных предсказаний центрированных логарифмов цен (обозначим его y^+).

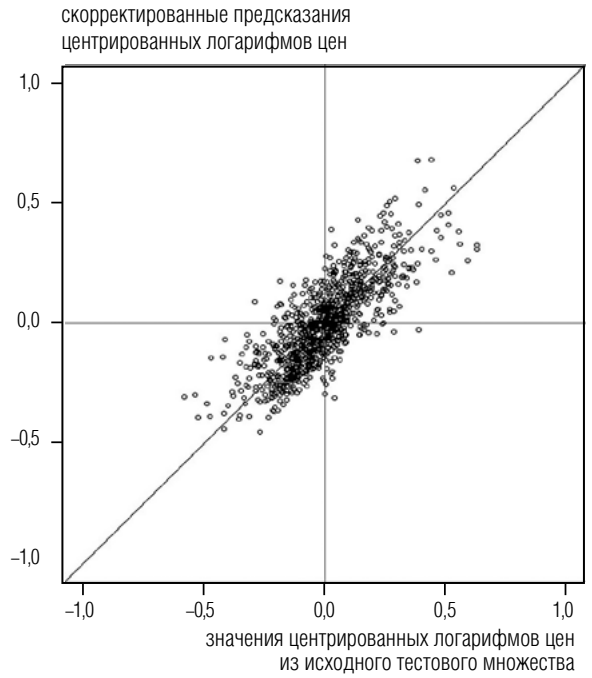

Рис. 6. Скорректированные предсказания на валидационном множестве по алгоритму «случайный лес».

На *рисунке 7* по горизонтальной оси отмечены значения компонент вектора y^*, по вертикальной – значения компонент вектора y^+.

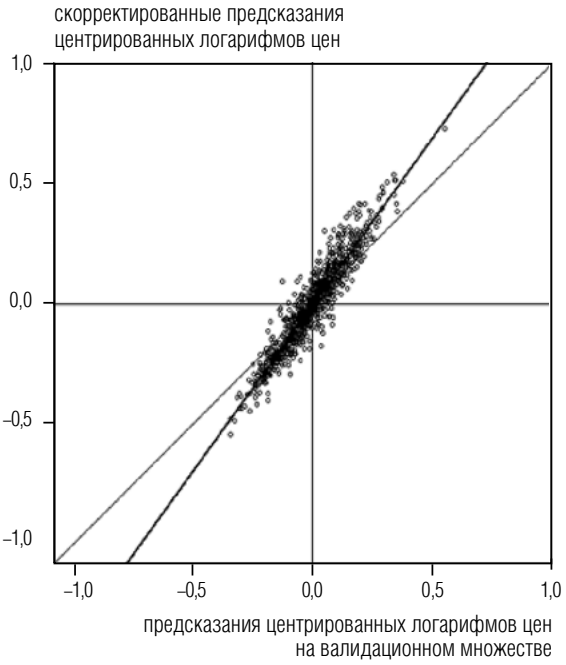

Рис. 7. Соотношение предсказанных и скорректированных значений.

Показанное на *рисунке 7* множество имеет характерный линейный тренд вида $y^+ = \alpha \cdot y^*$, который легко определяется применением библиотечной функции lm статистического пакета R, что дает в этом примере $y^+ = 1{,}388 \cdot y^*$. Теперь применим последовательно уже полученную на обучающей выборке модель случайного леса и корректировку предсказания, полученную на валидационном множестве к тестовому множеству (1294 записи).

На *рисунке 8* показаны предсказания центрированных логарифмов цен на тестовом множестве (применена модель случайного леса, полученная на обучающем множестве).

На *рисунке 9* показаны скорректированные предсказания центрированных логарифмов цен на тестовом множестве (применена модель случайного леса, полученная на обучающем множестве и корректировка, полученной на валидационном множестве).

Приведем необходимые формулы и последовательность действий:

1. Пусть $\overrightarrow{\ln(V)} = (\ln(V_1), \ln(V_2), \ln(V_3), ..., \ln(V_m))$ – наблюденные значения логарифмов цен со средним α, $\overrightarrow{\ln(\hat{V})} = (\ln(\hat{V}_1), \ln(\hat{V}_2), \ln(\hat{V}_3), ..., \ln(\hat{V}_m))$ – предска-

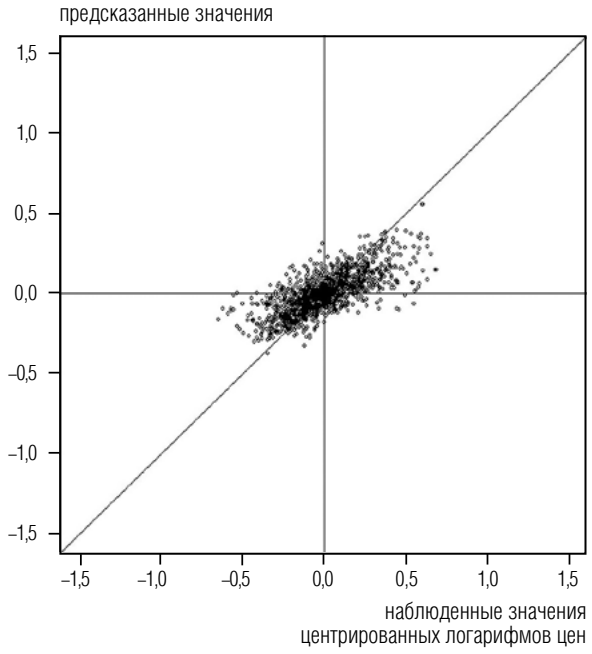

Рис. 8. Соотношение значений наблюденных центрированных логарифмов цен и их предсказанных значений алгоритмом «случайный лес».

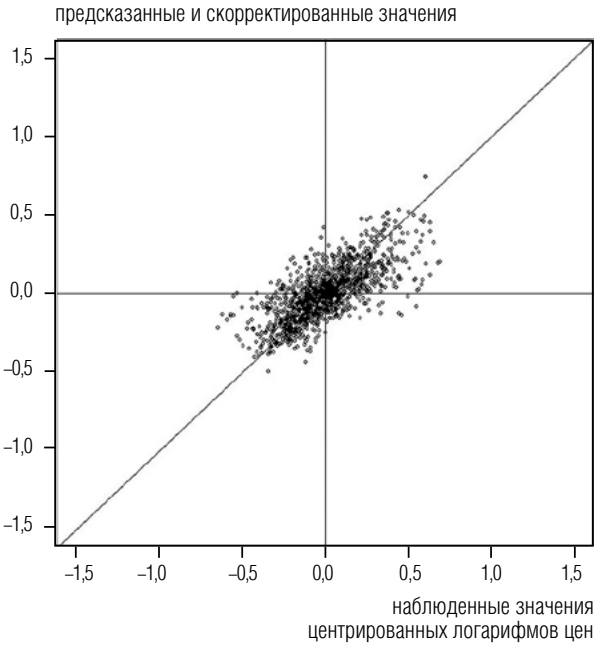

Рис. 9. Соотношение значений наблюденных центрированных логарифмов цен и их предсказанных значений алгоритмом «случайный лес» и скорректированных по формуле, полученной на валидационном множестве.

занные на обучающем множестве (тестовом, верифицирующем, m – может принимать разные значения) значения логарифмов цен со средним β. Тогда

$$y_i^* = \ln(\hat{V}_i) - \beta,\ x_i^* = \ln(V_i) - \alpha \text{ и}$$

$$\begin{pmatrix} x_i^+ \\ y_i^+ \end{pmatrix} = (x_i^+, y_i^+) \cdot \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix},$$

где угол поворота $\varphi = \frac{\pi}{4} - \text{arctg}(0,40791)$ (под знаком арктангенса – тангенс угла наклона линейного тренда облака рассеяния). См. *рис. 4, 5, 6.*

2. Сравниваем предсказанные и скорректированные на валидационной выборке значения центрированных логарифмов (*рис. 7*), определяем угловой коэффициент α тренда $y^+ = \alpha \cdot y^*$.

3. Предсказанное и скорректированное значение y^+ на тестовой выборке сравниваем с наблюденными значениями центрированных логарифмов цен x^* (*рис. 8, 9*).

Для исследуемого набора данных наиболее значимыми (при проведении оценки относительной значимости с помощью метода перестановки признаков) оказались такие предикторы как район города, тип дома, количество этажей в доме и общая площадь помещения. Наименее важными оказались: этаж расположения помещения, близость метро, количество комнат в помещении.

Выводы

1. Алгоритмы, основанные на деревьях решений, предсказывают средние значения (при построении для цен) и медианные значения (при построении для логарифмов цен), а не наиболее вероятные. Предсказанная по ним оценка рыночной стоимости несколько выше, чем оценка рыночной стоимости по наиболее вероятному значению.

2. В силу логарифмически нормальных распределений цен в исходных множествах, предсказания, построенные с использованием методов, основанных на деревьях решений, требуют корректировки. Предложенная в работе процедура позволяет провести такую корректировку с помощью двойной кросс-валидации, когда выделяется промежуточное подмножество исходного набора данных, на котором проводится адаптация алгоритма. Затем проводится оценка результатов на тестовом множестве. Проведенная апробация показала эффективность предложенного подхода. ■

Благодарности

Исследование для данной статьи выполнено при поддержке гранта РНФ (проект № 20-18-00365) и гранта РФФИ (проект № 20-01-00646 А).

Литература

1. Gu S., Kelly B., Xiu D. Empirical asset pricing via machine learning // Chicago Booth Research Paper No. 18-04, 31st Australasian Finance and Banking Conference 2018, Yale ICF Working Paper No. 2018-09. 2019. https://doi.org/10.2139/ssrn.3159577
2. Jim C.Y., Chen W.Y. Impacts of urban environmental elements on residential housing prices in Guangzhou (China) // Landscape and Urban Planning. 2006. Vol. 78. No. 4. P. 422–434. https://doi.org/10.1016/j.landurbplan.2005.12.003
3. Kok N., Koponen E.-L., Martínez-Barbosa C.A. Big data in real estate? From manual appraisal to automated valuation // The Journal of Portfolio Management. 2017. Vol. 43. No. 6. P. 202–211. https://doi.org/10.3905/jpm.2017.43.6.202
4. Ohnishi T., Mizuno T., Shimizu C., Watanabe T. On the evolution of the house price distribution // Columbia Business School. Center of Japanese Economy and Business. 2011. Working Paper Series. 296. https://doi.org/10.7916/D8794CJJ
5. Steurer M., Hill R.J., Pfeifer N. Metrics for evaluating the performance of machine learning based automated valuation models // Journal of Property Research. 2021. Vol. 38. No. 2. P. 99–129. https://doi.org/10.1080/09599916.2020.1858937
6. Tay D.P., Ho D.K. Artificial intelligence and the mass appraisal of residential apartments // Journal of Property Valuation and Investment. 1992. Vol. 10. No. 2. P. 525–540. https://doi.org/10.1108/14635789210031181
7. Anselin L., Lozano-Gracia N. Errors in variables and spatial effects in hedonic house price models of ambient air quality // Empirical Economics. 2008. Vol. 34. No. 1. P. 5–34. https://doi.org/10.1007/s00181-007-0152-3
8. Benson E.D., Hansen J.L., Schwartz Jr. A.L., Smersh G.T. Pricing residential amenities: The value of a view // The Journal of Real Estate Finance and Economics. 1998. Vol. 16. No. 1. P. 55–73. https://doi.org/10.1023/A:1007785315925
9. Debrezion G., Pels E., Rietveld P. The impact of rail transport on real estate prices: an empirical analysis of the Dutch housing market // Urban Studies. 2011. Vol. 48. No. 5. P. 997–1015. https://doi.org/10.1177/0042098010371395
10. Wena H., Zhanga Y., Zhang L. Assessing amenity effects of urban landscapes on housing price in Hangzhou, China // Urban Forestry & Urban Greening. 2015. Vol. 14. P. 1017–1026. https://doi.org/10.1016/j.ufug.2015.09.013
11. Do A.Q., Grudnitski G. A neural network approach to residential property appraisal // The Real Estate Appraiser. 1992. Vol. 58. No. 3. P. 38–45.
12. Evans A., James H., Collins A. Artificial neural networks: An application to residential valuation in the UK. University of Portsmouth, Department of Economics, 1992.
13. McGreal S., Adair A., McBurney D., Patterson D. Neural networks: the prediction of residential values // Journal of Property Valuation and Investment. 1998. Vol. 16. No. 1. P. 57–70. https://doi.org/10.1108/14635789810205128
14. Peterson S., Flanagan A. Neural network hedonic pricing models in mass real estate appraisal // Journal of Real Estate Research. 2009. Vol. 31. No. 2. P. 147–164. https://doi.org/10.1080/10835547.2009.12091245
15. Worzala E., Lenk M., Silva A. An exploration of neural networks and its application to real estate valuation // Journal of Real Estate Research. 1995. Vol. 10. No. 2. P. 185–201. https://doi.org/10.1080/10835547.1995.12090782
16. Kontrimas V., Verikas A. The mass appraisal of the real estate by computational intelligence // Applied Soft Computing. 2011. Vol. 11. No. 1. P. 443–448. https://doi.org/10.1016/j.asoc.2009.12.003
17. Park B., Bae J.K. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data // Expert Systems with Applications. 2015. Vol. 42. No. 6. P. 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040
18. Chen T., Guestrin C. XGBoost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 785–794. https://doi.org/10.1145/2939672.2939785
19. Cordoba M., Carranza J.P., Piumetto M., Monzani F., Balzarini M. A spatially based quantile regression forest model for mapping rural land values // Journal of Environmental Management. 2021. Vol. 289. No. 1. Article ID 112509. https://doi.org/10.1016/j.jenvman.2021.112509
20. Yilmazer S., Kocman S. A mass appraisal assessment study using machine leaning based on multiple regression and random forest // Land Use Policy. 2020. Vol. 99. Article ID 104889. https://doi.org/10.1016/j.landusepol.2020.104889
21. Webb G.I. Multiboosting: A technique for combining boosting and wagging // Machine Learning. 2000. Vol. 40. No. 2. P. 159–196. https://doi.org/10.1023/A:1007659514849
22. Aitchinson J., Brown J.A.C. The Lognormal distribution with special references to its uses in economics. Cambridge: At the University Press, 1963.
23. Rusakov O., Laskin M., Jaksumbaeva O. Pricing in the real estate market as a stochastic limit. Lognormal approximation // International Journal of the Mathematical models and methods in applied sciences. 2016. Vol. 10. P. 229–236.
24. Rusakov O., Laskin M., Jaksumbaeva O., Ivakina A. Pricing in real estate market as a stochastic limit. Lognormal approximation // Second International Conference on Mathematics and Computers in Sciences and in Industry. Malta. 2015. https://doi.org/10.1109/MCSI.2015.48
25. Лемешко Б.Ю., Лемешко С.Б., Семенова М.А. К вопросу статистического анализа больших данных // Вестник Томского Университета. 2018. Т. 44. С. 40–49. https://doi.org/10.17223/19988605/44/5

26. Жукова Г.Н. Идентификация вероятностного распределения по коэффициентам асимметрии и эксцесса // Автоматизация. Современные технологии. 2016. Т. 5. С. 26–33.

27. Ласкин М.Б. Многомерное логарифмически нормальное распределение в оценке недвижимого имущества // Бизнес-информатика. 2020. Т. 14. № 2. С. 48–63. https://doi.org/10.17323/2587-814X.2020.2.48.63

Об авторах

Ласкин Михаил Борисович
кандидат физико-математических наук;
доцент, с.н.с., Санкт-Петербургский ФИЦ РАН (СПИИРАН), 199178, Санкт-Петербург, В.О., 14 линия, д.39;
E-mail: laskinmb@yahoo.com
ORCID: 0000-0002-0143-4164

Гадасина Людмила Викторовна
кандидат физико-математических наук;
доцент, Санкт-Петербургский государственный университет, 199034, Россия, Санкт-Петербург, Университетская наб., д.7/9;
E-mail: l.gadasina@spbu.ru
ORCID: 0000-0002-4758-6104

Peculiarities of applying methods based on decision trees in the problems of real estate valuation

Mikhail B. Laskin [a]
E-mail: laskinmb@yahoo.com

Lyudmila V. Gadasina [b]
E-mail: l.gadasina@spbu.ru

[a] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS)
Address: 39, 14 line, Vasilevskiy Island, St. Petersburg 199178, Russia

[b] St. Petersburg State University, Center for econometrics and business analytics (CEBA)
Address: 7/9, Universitetskaya emb., Saint-Petersburg 199034, Russia

Abstract

The increasing flow of available market information, the development of methods of machine learning, artificial intelligence and the limited capabilities of traditional methods of real estate valuation are leading to a significant increase of researchers' interest in real estate valuation by applying methods based on decision trees. At the same time, the distribution of real estate prices is well approximated by a lognormal distribution. Therefore, traditional methods overestimate the predicted values in the region below the average of the available data set and underestimate the predicted values in the region above the average. This article shows the reasons for these features and proposes an adaptive random forest algorithm which corrects the results of the basic algorithm prediction by revising the bias of these predicted values. The results were tested on the real estate offer prices in St. Petersburg.

Keywords: decision trees, random forest, real estate market, price-forming factors, market value apprising

Citation: Laskin M.B., Gadasina L.V. (2022) Peculiarities of applying methods based on decision trees in the problems of real estate valuation. *Business Informatics*, vol. 16, no. 4, pp. 7–18. DOI: 10.17323/2587-814X.2022.4.7.18

References

1. Gu S., Kelly B., Xiu D. (2019) Empirical asset pricing via machine learning. *Chicago Booth Research Paper No. 18-04, 31st Australasian Finance and Banking Conference 2018, Yale ICF Working Paper No. 2018-09*. https://doi.org/10.2139/ssrn.3159577

2. Jim C.Y., Chen W.Y. (2006) Impacts of urban environmental elements on residential housing prices in Guangzhou (China). *Landscape and Urban Planning*, vol. 78, no. 4, pp. 422–434. https://doi.org/10.1016/j.landurbplan.2005.12.003

3. Kok N., Koponen E.-L., Martínez-Barbosa C.A. (2017) Big data in real estate? From manual appraisal to automated valuation. *The Journal of Portfolio Management,* vol. 43, no. 6, pp. 202–211. https://doi.org/10.3905/jpm.2017.43.6.202

4. Ohnishi T., Mizuno T., Shimizu C., Watanabe T. (2011) On the evolution of the house price distribution. *Columbia Business School. Center of Japanese Economy and Business, Working Paper Series, 296*. https://doi.org/10.7916/D8794CJJ

5. Steurer M., Hill R.J., Pfeifer N. (2021) Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research*, vol. 38, no. 2, pp. 99–129. https://doi.org/10.1080/09599916.2020.1858937

6. Tay D.P., Ho D.K. (1992) Artificial intelligence and the mass appraisal of residential apartments. *Journal of Property Valuation and Investment*, vol. 10, no. 2, pp. 525–540. https://doi.org/10.1108/14635789210031181

7. Anselin L., Lozano-Gracia N. (2008) Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical Economics*, vol. 34, no. 1, pp. 5–34. https://doi.org/10.1007/s00181-007-0152-3

8. Benson E.D., Hansen J.L., Schwartz Jr. A.L., Smersh G.T. (1998) Pricing residential amenities: The value of a view. *The Journal of Real Estate Finance and Economics*, vol. 16, no. 1, pp. 55–73. https://doi.org/10.1023/A:1007785315925

9. Debrezion G., Pels E., Rietveld P. (2011) The impact of rail transport on real estate prices: an empirical analysis of the Dutch housing market. *Urban Studies*, vol. 48, no. 5, pp. 997–1015. https://doi.org/10.1177/0042098010371395

10. Wena H., Zhanga Y., Zhang L. (2015) Assessing amenity effects of urban landscapes on housing price in Hangzhou, China. *Urban Forestry & Urban Greening*, vol. 14, pp. 1017–1026. https://doi.org/10.1016/j.ufug.2015.09.013

11. Do A.Q., Grudnitski G. (1992) A neural network approach to residential property appraisal. *The Real Estate Appraiser*, vol. 58, no. 3, pp. 38–45.

12. Evans A., James H., Collins A. (1992) *Artificial neural networks: An application to residential valuation in the UK*. University of Portsmouth, Department of Economics.

13. McGreal S., Adair A., McBurney D., Patterson D. (1998) Neural networks: the prediction of residential values. *Journal of Property Valuation and Investment*, vol. 16, no. 1, pp. 57–70. https://doi.org/10.1108/14635789810205128

14. Peterson S., Flanagan A. (2009) Neural network hedonic pricing models in mass real estate appraisal. *Journal of Real Estate Research*, vol. 31, no. 2, pp. 147–164. https://doi.org/10.1080/10835547.2009.12091245

15. Worzala E., Lenk M., Silva A. (1995) An exploration of neural networks and its application to real estate valuation. *Journal of Real Estate Research*, vol. 10, no. 2, pp. 185–201. https://doi.org/10.1080/10835547.1995.12090782

16. Kontrimas V., Verikas A. (2011) The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing*, vol. 11, no. 1, pp. 443–448. https://doi.org/10.1016/j.asoc.2009.12.003

17. Park B., Bae J.K. (2015) Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040

18. Chen T., Guestrin C. (2016) XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. https://doi.org/10.1145/2939672.2939785

19. Cordoba M., Carranza J.P., Piumetto M., Monzani F., Balzarini M. (2021) A spatially based quantile regression forest model for mapping rural land values. *Journal of Environmental Management*, vol. 289, no. 1, 112509. https://doi.org/10.1016/j.jenvman.2021.112509

20. Yilmazer S., Kocman S. (2020) A mass appraisal assessment study using machine leaning based on multiple regression and random forest. *Land Use Policy*, vol. 99, 104889. https://doi.org/10.1016/j.landusepol.2020.104889

21. Webb G.I. (2000) Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, vol. 40, no. 2, pp. 159–196. https://doi.org/10.1023/A:1007659514849

22. Aitchinson J., Brown J.A.C. (1963) *The Lognormal distribution with special references to its uses in economics*. Cambridge: At the University Press.

23. Rusakov O., Laskin M., Jaksumbaeva O. (2016) Pricing in the real estate market as a stochastic limit. Log Normal approximation. *International Journal of the Mathematical models and methods in applied sciences*, vol. 10, pp. 229–236.

24. Rusakov O., Laskin M., Jaksumbaeva O., Ivakina A. (2015) Pricing in real estate market as a stochastic limit. Lognormal approximation. *Second International Conference on Mathematics and Computers in Sciences and in Industry. Malta, 2015*. https://doi.org/10.1109/MCSI.2015.48

25. Lemeshko B. Yu., Lemeshko S.B., Semenova M.A. (2018) On the issue of statistical analysis of big data. *Bulletin of Tomsk University*, vol. 44, pp. 40–49 (in Russian). https://doi.org/10.17223/19988605/44/5

26. Zhukova G.N. (2016) Identification of the probability distribution by the coefficients of asymmetry and kurtosis. *Automation. Modern Technologies*, vol. 5, pp. 26–33 (in Russian).

27. Laskin M.B. (2020) Multidimensional lognormal distribution in real estate appraisals. *Business Informatics*, vol. 14, no. 2, pp. 48–63. https://doi.org/10.17323/2587-814X.2020.2.48.63

About the authors

Mikhail B. Laskin

Cand. Sci. (Phys.-Math.);

Associate Professor, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), 39, 14 line, Vasilevskiy Island, St. Petersburg 199178, Russia;

E-mail: laskinmb@yahoo.com

ORCID: 0000-0002-0143-4164

Lyudmila V. Gadasina

Cand. Sci. (Phys.-Math.);

Associate Professor, St. Petersburg State University, 7/9, Universitetskaya emb., Saint-Petersburg 199034, Russia;

E-mail: l.gadasina@spbu.ru

ORCID: 0000-0002-4758-6104