Текстуальный анализ ценообразования на рынке московской жилой недвижимости

Гончаров Г.И., Натхов Т.В.

В данной работе применяется текстуальный анализ для оценки параметров гедонистической модели ценообразования на рынке вторичной недвижимости г. Москвы. Для проведения исследования был собран уникальный массив данных – активные в июле 2019 г. объявления о продаже жилой недвижимости на сайте ЦИАН. Для сбора информации была написана специальная программа-парсер на языке Python. Всего было собрано около 60 тыс. объявлений, которые представляют все районы Москвы. На основе этого массива данных и разработанного авторами алгоритма анализа текстов определены слова (униграммы) и словосочетания (биграммы), которые являются наиболее значимыми предикторами цены. Преимущество данного подхода в том, что подбор объясняющих переменных для эконометрической модели опирается на выявленные предпочтения участников рынка – алгоритм определяет характеристики жилья, которые указывают сами владельцы, заинтересованные в успешной продаже. Таким образом, мы выявляем важные субъективные факторы ценообразования на рынке московской недвижимости. Показано, что использование текстуального анализа позволяет заметно улучшить предсказательную силу эконометрической модели ценообразования. В частности, благодаря использованию униграмм мы можем сократить среднеквадратичную ошибку на 15%. Механизм этого улучшения заключается в учете факторов ценообразования, которые трудно измерить количественным образом. К примеру, биграммы «очистка воды», «охрана консьерж», «клубный дом», «система видеонаблюдение» и им подобные отвечают за факторы благоустройства самого жилья и его окрестностей, безопасность и другие общественные блага локального уровня, которые практически не поддаются количественному измерению по единой методике.

***Ключевые слова*:** гедонистическая модель; ЛАССО; недвижимость.

Авторы выражают благодарность анонимному рецензенту за подробные комментарии, которые помогли существенно улучшить работу.

Гончаров Глеб Игоревич – студент 4 курса, ОП Прикладная математика, МИЭМ Национального исследовательского университета «Высшая школа экономики». E-mail: gigoncharov@edu.hse.ru
Натхов Тимур Владимирович – старший научный сотрудник ИНИИ ВШЭ, доцент факультета экономических наук Национального исследовательского университета «Высшая школа экономики». E-mail: timur.natkhov@hse.ru

Статья поступила: 16.12.2019/Статья принята: 05.03.2020.

DOI: 10.17323/1813-8691-2020-24-1-101-116

Для цитирования: Гончаров Г.И., Натхов Т.В. Текстуальный анализ ценообразования на рынке московской жилой недвижимости. *Экономический журнал ВШЭ*. 2020; 24(1): 101–116.

For citation: Goncharov G.I., Natkhov T.V. Textual Analysis of Pricing in the Moscow Residential Real Estate Market. *HSE Economic Journal.* 2020; 24(1): 101–116. (In Russ.)

1. Введение

Жилая городская недвижимость – один из важнейших активов в современном мире. Для многих россиян приобретение жилья является одним из ключевых событий в жизни. Об этом говорит и рост цен на жилую недвижимость в последние десятилетия, и рост предложения квадратных метров, и стабильное увеличение объемов ипотечного кредитования.

Объяснению и прогнозированию цен на жилую недвижимость посвящена обширная научная литература. Обычно предполагается, что цена квартиры зависит от ряда показателей, которые можно разделить на два типа. Первый тип – внутренние характеристики квартиры: общая площадь, количество комнат, наличие современного ремонта и другие. Второй тип – внешние факторы, к которым можно отнести транспортную и социальную инфраструктуру, экологическую обстановку в районе и т.п. Одной из самых распространенных моделей для изучения зависимости стоимости жилой недвижимости от различных характеристик и определения ценности внешних факторов является гедонистическая модель ценообразования. В этой модели ценность внешних факторов определяется как разница в стоимости жилья с одинаковыми внутренними характеристиками, расположенного в разных географических локациях.

Исследование [Magnus, Peresetsky, 2010] – одна из первых работ с гедонистической моделью для российского рынка недвижимости. В своей работе ученые исследуют московский рынок 2003 г. В их модели, помимо внутренних факторов (площадь, количество комнат и т.д.), используются два внешних фактора: удаленность от метро и от центра города. Показано, что эти факторы имеют значимое влияние на цену жилья. Последующие исследования, используя эту статью как основу, идут либо по пути усложнения модели, добавляя новые внешние характеристики, либо по пути расширения географии исследования. Так, в работе [Красильников, Щербакова, 2011] сделана попытка распространить модель [Magnus, Peresetsky, 2010] на другие города России: Санкт-Петербург, Новосибирск, Екатеринбург. Авторы находят некоторые региональные особенности, одной из которых является низкая эластичность цены по расстоянию до метро в Екатеринбурге. В статье [Сидоровых, 2015] исследуется вариация цен на квартиры в Перми в зависимости от транспортной доступности. Автор выявил, что число маршрутов в микрорайоне положительно влияет на цену квартиры. Возможное влияние окружающей среды на ценообразование квартир было проанализировано в работе [Катышев, Хакимова, 2012]. Для этой цели исследователи рассматривают среднегодовые уровни CO, NO, NO_2, расстояние до ближайшего промышленного предприятия в качестве факторов окружающей среды. В результате исследования они обнаруживают положительную связь удаленности от за-

вода и цены, а также отрицательную зависимость цены от концентрации угарного газа. Исследование [Ожегов, Косолапов, Позолотина, 2017] пытается связать стоимость квартир в Перми с уровнем школьного образования, предполагая, что квартиры в районах с хорошими школами имеют бóльшую цену. Авторы обнаруживают, что увеличение среднего балла ЕГЭ близлежащих школ увеличивает цену квартиры только для квартир со стоимостью выше медианной; на цену дешевых квартир этот показатель не оказывает статистически значимого влияния.

Таким образом, исследователи объясняют ценообразование на жилую недвижимость с помощью множества различных факторов. Можно утверждать, что не существует единого мнения о наборе переменных, объясняющих цену квартир. Поэтому выбор и обоснование независимых переменных остается актуальной исследовательской задачей. И если количество внутренних факторов – это сравнительно небольшое множество, то количество внешних факторов ограничено только воображением исследователя.

В данной работе мы делаем попытку решить эту проблему, применяя (впервые в отечественной литературе) текстуальный анализ в модели ценообразования на рынке жилой недвижимости в Москве в 2019 г. Мы используем алгоритмы анализа текстов объявлений о продаже недвижимости для выявления наиболее значимых предикторов цены. Преимущество этого подхода в том, что мы не опираемся на интуицию при подборе объясняющих переменных, а выявляем факторы ценообразования, которые считают важными непосредственные участники рынка, в данном случае продавцы квартир. Мы предполагаем, что продавец, заинтересованный в успешной продаже, указывает в объявлении большинство факторов, положительно влияющих на цену. Однако мы допускаем, что продавцы могут не упоминать некоторые недостатки квартиры, поэтому объявления будут несколько завышать цену квартиры. Имея это в виду, мы проводим текстуальный анализ объявлений о продаже недвижимости, подобный изложенному в работе [Nowak, Smith, 2017]. В этой статье авторы анализировали записи о продажах на массиве данных GAMLS – записях о продаже домов в Атланте, штат Джорджия[1]. Мы будем следовать этой работе, однако в ходе собственного исследования сравним разные алгоритмы обработки объявлений.

В работе [Hausler, Ruscheinsky, Lang, 2018] авторы впервые используют текстуальный и сентимент-анализ для предсказания динамики цен на американском рынке жилой недвижимости. Результаты показывают важную роль настроений (*sentiments*) в предсказании цен, даже после учета макроэкономических и других факторов. В работе [Pryce, Oates, 2008] авторы показывают, что содержание объявлений о продаже недвижимости меняется в зависимости от стадии бизнес-цикла, времени года и повестки общенациональных новостей. Авторы работы [Goodwin et al., 2019] показывают, что коннотация объявления (количество слов с положительным окрасом) влияет на конечную цену сделки. Наконец, исследователи [Lawani, Reed, 2018] анализируют рынок временной аренды жилья на Airbnb и показывают, что включение слов из объявлений в регрессионное уравнение уменьшает ошибки в предсказании цен и может быть полезным для учета пропущенных показателей качества в гедонистических ценовых моделях.

[1] База состоит из более чем 500 тыс. описаний домов, выставленных на продажу с 1 января 2000 г. до 31 декабря 2014 г., и включает в себя информацию о физических параметрах дома, местоположении, цене и других характеристиках.

В следующем разделе статьи рассматриваются основные способы работы с текстом и модель регрессии с регуляризацией. В разделе 3 описывается база данных, в разделе 4 эконометрические модели. В разделе 5 показаны результаты анализа, в разделе 6 приведены выводы работы.

2. Методы исследования

2.1. Алгоритмы работы с текстами

В текстуальном анализе текст представляют как множество слов или фраз, каждую из которых называют токеном (от английского слова *token* – знак, символ). Соответственно процесс разбиения текста на токены называется токенизацией. Перед токенизацией следует очистить текст от предлогов и стоп-слов. А также мы убирали и имена собственные, которые иногда включаются в объявления в качестве контактного лица для покупки квартиры. Существуют несколько подходов к токензиации, которые отличаются способом определения токенов. Первый подход определяет токены как отдельные слова (униграммы): каждое слово, встречающееся в тексте, считается отдельным токеном. Такой подход, описывающий текст как набор слов, может иногда приводить к «потере контекста».

Альтернативным способом является использование в качестве токена устойчивых словосочетаний (биграммы). К примеру, если мы хотим, чтобы словосочетание «теплый пол» целиком было отдельным параметром в модели, то мы должны использовать биграммы. При таком подходе каждая пара слов является токеном. В теории мы можем строить наши токены на основании N-грамм, где N – количество последовательных слов для токенизации. Однако при включении большего количества слов в один токен возникает риск ухудшения обобщающей способности модели.

В нашей работе мы рассматриваем модели, основанные на униграммах и биграммах. Создание токенов на необработанном тексте может приводить к искажениям, когда несколько форм одного слова или фразы будут соответствовать разным токенам. Например, биграмы «теплый пол» и «теплым полом», обозначающие одно и то же, в такой модели будут разными токенами. Во избежание подобных расхождений исходные тексты объявлений обрабатываются алгоритмами стемминга (*stemming*) и лемматизации (*lemmatization*). Стемминг выделяет основы слова, преобразовывая словосочетание «теплый пол» и «теплым полом» в «тепл пол», которое в итоге сводит разные формы к одному токену. Лемматизиция приводит словоформы к лемме – начальной форме слова. Таким образом, «теплый пол» и «теплым полом» станут одним словосочетанием «теплый пол»[2]. Поскольку токены после лемматизации легче воспринимаются, мы решили использовать лемматизацию[3].

[2] В нашей работе мы сравнили SnowBall стеммер и лемматизатор из пакетов nltk и pymorphy2 для обработки текста и пришли к выводу, что алгоритмы оказывают примерно одинаковое влияние на результат.

[3] Униграммы и биграммы могут быть использованы вместе. Мы пробовали применить такой подход, но получили, что множество отобранных моделями токенов состоит в основном из униграмм, при этом качество модели (R-квадрат) улучшается всего на 0,02. В топ-20 положительных

Мы также ограничиваем используемые токены по причине низкой информативности токенов с очень высокой и очень низкой частотой. Для пояснения рассмотрим крайние случаи. С одной стороны, токены, которые используются почти во всех текстах, не содержат уникальной информации, влияющей на цену. С другой стороны, очень редкие токены, встречающиеся лишь в нескольких объявлениях, вряд ли будут полезны для выявления закономерностей. Исходя из этого мы убрали токены, которые встречаются более чем в 90% объявлений и взяли 2000 следующих за этим порогом токенов.

После токенизации текстов необходимо сопоставить каждому токену некую численную характеристику. Самый простой способ – это использовать функцию-индикатор: $f(t_i, a_j) = 1$, если токен t_i содержится в объявлении a_j, и $f(t_i, a_j) = 0$ в противоположном случае. Данный способ отличается простотой интерпретации: появление слова в объявлении влияет на цену с определенным коэффициентом. С другой стороны, переменная-индикатор игнорирует такие количественные характеристики токена, как количество использования токена в объявлении. Например, во фразе «5 минут пешком до метро, 15 минут пешком до жд станции» модель учтет словосочетание *«минут пешком»* только один раз, хотя приближенность и к метро, и к железнодорожной станции может влиять на цену.

Альтернативный способ сопоставления – каждому токену соответствует количество его вхождений в объявление: $f(t_i, a_j) = n$, где n – количество вхождений токена t_i в a_j. Этот подход позволяет сохранить количественную зависимость токенов в объявлении. Однако при таком подходе токены в больших по размеру объявлениях будут иметь больший вес. Третья модель, которую мы использовали в нашей работе, сопоставляет каждому токену его TF-IDF характеристику. Эта характеристика токена рассчитывается как произведение TF и IDF характеристик. *TF* (*term frequency*) – частота вхождения слова в документ – определяется по формуле:

$$TF = \frac{n_i}{\sum_k^N n_k},$$

где N – количество токенов в документе; n_i – количество вхождений i-го токена в документ. *IDF* (*inverse document frequency*) – инверсия частоты документов (объявлений), в которых встречается токен, среди всей коллекции документов:

$$IDF = log\frac{D}{D_i},$$

где D – количество документов в коллекции; D_i – количество документов, содержащих i-й токен. Таким образом, данная мера будет уменьшать значения токенов, встречающихся во многих документах, и увеличивать значения редких.

и отрицательных результатов попадают 17 униграмм и только 3 биграммы. По этой причине мы не показываем эти результаты.

Для выбора наилучшего способа сопоставления мы используем LASSO-регрессию. Наилучшую модель определяем по наименьшей ошибке на отложенной выборке. Коэффициент λ определяется с помощью кросс-валидации[4].

Ниже представлены модели обработки текста.

$$p_i = const^{id} + \beta^{id} v_i^{id} + \varepsilon_i,$$
$$p_i = const^{id} + \beta^{id} w_i^{id} + \varepsilon_i,$$
$$p_i = const^{count} + \beta^{count} v_i^{count} + \varepsilon_i,$$
$$p_i = const^{count} + \beta^{count} w_i^{count} + \varepsilon_i,$$
$$p_i = const^{tfidf} + \beta^{tfidf} v_i^{tfidf} + \varepsilon_i,$$
$$p_i = const^{tfidf} + \beta^{tfidf} w_i^{tfidf} + \varepsilon_i.$$

В моделях v^{model_i} – набор униграмм соответствующей модели для *i*-го объявления; w^{model_i} – набор биграмм соответствующей модели для *i*-го объявления; p_i – логарифм цены *i*-го объявления; ε_i – ошибка, а $const^{fi}$ – константа, где f_i – способ сопоставления, и β – коэффициент регрессии, выбранной нами модели соответствия.

Таблица 1.

Сравнение моделей предобработки текста

	Count uni	Count bi	I uni	I bi	TFIDF uni	TFIDF bi
Параметр λ_{cv}	0,00001	0,0001	0,0001	0,0001	0,00005	0,0001
N переменных	1398	1476	1372	1446	974	1121
RMSE	0,41	0,43	0,39	0,41	0,38	0,4

Примечания. Сравнение с помощью кросс-валидации на 5 частях. RMSE измерялась на отложенной выборке.

В табл. 1 представлены результаты сравнения моделей. Модели, использующие TF-IDF-характеристику, имеют наименьшую ошибку как в случае униграмм, так и в случае биграмм. Поэтому далее в работе мы использовали только такой способ, несмотря на его не столь прозрачную интерпретацию, как в случае функции индикатора.

2.2. Регрессия с регуляризацией

Включая в модель токены, мы сильно повышаем размерность модели. Тем не менее логично предположить, что далеко не все параметры нашей исходной модели будут влиять на цену. Следовательно, возникает вопрос корректной спецификации модели. Стан-

[4] Подробнее о кросс-валидации и LASSO-регрессии в подразделе 2.2.

дартные способы постепенного включения или исключения нам плохо подходят: у нас имеется 2^N потенциальных моделей, где N – количество параметров. Если N, как в нашем случае, велико, то перебрать все модели не представляется возможным. Мы можем попытаться сделать это с помощью ЛАССО-регрессии, решая следующую оптимизационную задачу:

$$\min_d \sum_i (p_i - x_i d)^2 + \lambda |d|_{l1},$$

где x_i – вектор параметров i-го объявления; d – вектор весов; $|d|_{l1}$ – $l1$-норма вектора d. Многое зависит от выбора гиперпараметра λ. Если положить λ = 0, то мы получим уравнение линейной регрессии. Выбор λ > 0 из-за геометрической формы $l1$-нормы способствует обнулению коэффициентов регрессии. Таким образом мы можем производить отбор коэффициентов. Один из самых простых способов выбрать коэффициенты – кросс-валидация. Этот процесс включает в себя следующие шаги.

1. Разбиваем наблюдения на M групп с одинаковым количеством наблюдений, стараясь сохранить распределение цены, идентичное всему массиву данных.
2. Для каждой из групп M обучаем ЛАССО-регрессию на оставшихся (M − 1) группах вместе и оцениваем полученную среднеквадратичную ошибку для группы m_i, получая $MSE(\lambda, m_i)$
3. Усредняем значение $MSE(\lambda) = \frac{1}{|M|}\sum_{i=1}^{M} MSE(\lambda, m_i)$.
4. Выбираем λ: $\lambda_{cv} = argmin MSE(\lambda)$.

Полученный коэффициент часто бывает маленьким, поэтому в модели отбирается большое количество атрибутов. Данные по кросс-валидации и коэффициенту λ_{cv} приведены в табл. 1. К тому же, данный подход не гарантирует, что у модели будет правильная спецификация.

Существуют другие способы выбора λ, которые могут дать лучший с точки зрения спецификации модели результат, которые подробнее изложены в статье [Nowak, Smith, 2017]. Однако мы остановимся на вышеизложенном методе на 5 частях кросс-валидации из-за его простоты. В дополнение к вышеописанной проблеме существует еще одна: ЛАССО дает нам смещенные оценки на коэффициенты. Для получения несмещенных оценок мы применяем линейную регрессию на отобранных ЛАССО признаках. Эта процедура называется post-LASSO и подробно описана в работе [Belloni, Chernozhukov, 2013]. Ее суть заключается в том, что если с помощью ЛАССО были отобраны «истинные» параметры, тогда наша модель при применении обычной регрессии будет иметь правильную спецификацию и ее оценки будут несмещенными. Другими словами, мы сначала применяем ЛАССО для отбора текстовых признаков, а потом добавляем отобранные токены к нашим классическим моделям. Однако повторимся, что валидация не гарантирует получения «правильных» текстовых признаков, поэтому итоговая модель может иметь неправильную спецификацию.

3. Описание данных

Для проведения исследования был собран уникальный массив данных – активные в июле 2019 г. объявления о продаже жилой недвижимости с Москве на сайте ЦИАН. Для сбора информации была написана специальная программа-парсер на языке Python. Всего было собрано около 60 тыс. объявлений, которые представляют все районы Москвы. Из выборки были исключены три административных округа, которые, на наш взгляд, сильно отличаются от московских районов, хотя формально относятся к Москве – Зеленоградский, Троицкий и Новомосковский административные округа. Также из выборки исключены нестандартные квартиры: со свободной планировкой, студии, многокомнатные квартиры (больше 4-х комнат). Кроме того, исключены объявления о продаже квартир в домах, со сроком сдачи после 2017 г. Такого рода объявления обычно заполняются по стандартному шаблону от компании-застройщика, поэтому несут мало уникальной информации. Итоговая база данных состоит из 32889 объявлений. Описательные статистики численных переменных приведены в табл. 2. Распределение логарифма цены представлено на гистограмме (рис. 1).

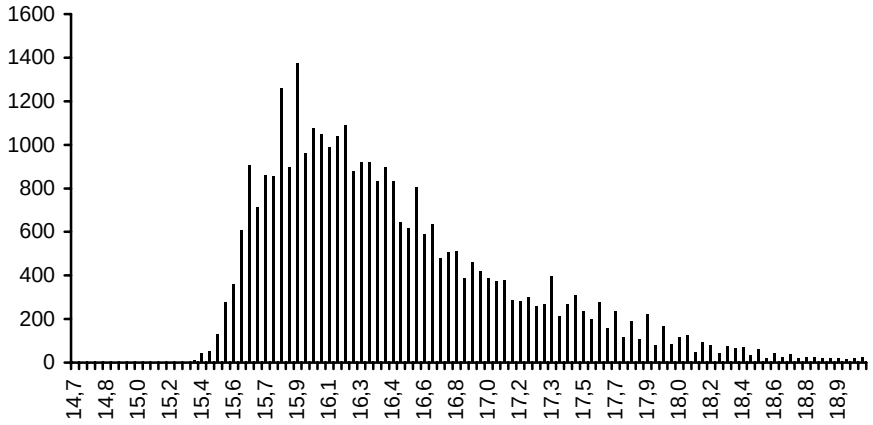

Рис. 1. Распределение логарифма цены

Таблица 2.

Описательные статистики

	Среднее	Медиана	Стандартное отклонение	Минимум	Максимум
Цена, млн руб.	19,4	12,2	20,9	2,35	200
Общая площадь, м^2	70,3	60,3	34,7	11,8	290
Год постройки	1985,1	1985	23,8	1910	2017
Расстояние до Кремля, км	10,9	11,2	5,8	0,7	25,8
Расстояние до метро, км	0,99	0,81	0,74	0,04	5,85
Кухня/общая площадь	0,16	0,16	0,06	0,08	0,7

В наших моделях мы использовали следующие категориальные переменные:

- Апартаменты – бинарная переменная, равная единице, если квартира является апартаментами;
- Комната *i* – бинарная переменная, равная единице, если в квартире *i* комнат;
- Первый этаж – бинарная переменная, равная единице, если квартира на первом этаже;
- Последний этаж – бинарная переменная, равная единице, если квартира на последнем этаже;
- Название района Москвы – бинарная переменная, равная единице, если квартира относится к этому району;
- Линия *i* – бинарная переменная, равная единице, если у квартиры в пешей доступности есть станция метро *i*-й линии.

Распределение квартир по количеству комнат представлено на рис. 2.

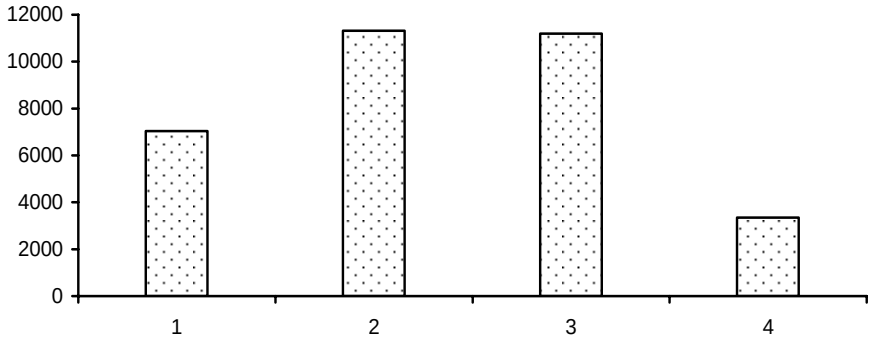

Рис. 2. Распределение квартир по количеству комнат

Для определения расстояния до ближайшей станции метро и определения ближайшей ветки мы перевели адреса домов в координаты и рассчитали расстояния до станций метро. Мы считаем, что станция метрополитена находится в пешей доступности от квартиры, если расстояние до нее не превышает километр по прямой. Такое расстояние не учитывает рельеф местности, поэтому, вероятно, оно меньше реального пешего расстояния до объекта.

4. Модели

Мы оцениваем параметры следующих моделей:

$$p_i = \beta_0 + \beta_x x_i + \varepsilon_i (Base),$$
$$p_i = \beta_0 + \beta_x x_i + \beta_a a_i + \varepsilon_i (Areas),$$
$$p_i = \beta_0 + \beta_x x_i + \beta_a a_i + \beta_s s_i + \varepsilon_i (Subs),$$
$$p_i = \beta_0 + \beta_x x_i + \beta_a a_i + \beta_s s_i + \beta_u u_i + \varepsilon_i (Uni),$$
$$p_i = \beta_0 + \beta_x x_i + \beta_a a_i + \beta_s s_i + \beta_w w_i + \varepsilon_i (Bi),$$

где для i-го объекта p_i – логарифм цены; x_i – вектор параметров (см. табл. 3); a_i – вектор бинарных переменных на районы, кроме ЦАО; s_i – вектор бинарных переменных на пешую доступность до метро; u_i – вектор униграмм; b_i – вектор биграмм; ε_i – ошибка.

Модель Base – типичная модель для подобного рода исследований. Ее образцом послужила модель из работы [Magnus, Peresetsky, 2010], хотя в нашей модели отсутствует ряд переменных. В целом, мы ожидаем подтверждения результатов этой модели на наших данных. Так как базовая квартира нашей модели – четырехкомнатная, то мы ожидаем, что коэффициенты перед другими типами квартир будут отрицательными.

В модели Areas мы пытаемся оценить влияние округа Москвы на цену, добавив бинарные переменные на округа и определив в качестве базового Центральный административный округ (ЦАО). Поскольку ЦАО – самый дорогой район Москвы, мы ожидаем, что коэффициенты перед остальными районами будут иметь отрицательный знак.

Модель Subs учитывает пешую доступность до линий метро. Мы ожидаем, что наличие метро будет положительно влиять на цену, однако, возможно, новые линии (МЦК и прочие) будут оказывать меньшее влияние на цену.

В моделях Uni и Bi используются униграммы и биграммы соответственно. Основная цель этих моделей – проверить, несут ли токены дополнительную информацию для оценки квартиры.

Для проверки качества моделей мы специальным образом разбили нашу выборку на две части: выборку для обучения и отложенную выборку. Для этого мы выделили 5 условных ценовых категорий квартир и разбили выборку с сохранением пропорций этих категорий. Таким образом мы стараемся сохранить первоначальное ценовое распределение в наших выборках. Лучшую модель мы выбираем по среднеквадратичной ошибке на отложенной выборке.

5. Результаты анализа

Результаты оценки параметров регрессии первых трех моделей приведены в табл. 3. Модель Areas полностью согласуется с нашими ожиданиями. Как мы видим, все коэффициенты перед районами имеют отрицательный знак, что говорит нам о том, что ЦАО – самый дорогой район Москвы. Западные районы, подтверждая распространенное мнение, имеют наибольшие коэффициенты среди оставшихся районов. Модель Subs дает противоречивые результаты. Мы ожидали, что любая станция метро, находящаяся в пешей доступности, будет положительно влиять на цену квартиры. Выяснилось, что такое влияние оказывают только более старые линии, а новые оказывают негативное влияние.

Как видно из табл. 4, объясняющая способность нашей модели сильно улучшается при добавлении информации о районе квартиры. Она также становится немного лучше при добавлении информации о линиях метро.

Использование текстуального анализа позволяет заметно улучшить объясняющую способность модели. В частности, благодаря использованию униграмм мы можем сократить среднеквадратичную ошибку на 15%. Результат модели с биграммами немного хуже результата униграмм, что соответствует предыдущим опытам, показанным в табл. 1. Видимо, это вызвано тем, что биграммы более «специфичны» и хуже обобщают информацию.

Таблица 3.

Результаты регрессионного анализа

Коэффициент	Base	Areas	Subs
Апартаменты	–0,077 (0,012)	–0,1 (0,011)	–0,11 (0,01)
Общая площадь	0,015 (0,000)	0,013 (0,000)	0,013 (0,000)
Год	0,0034 (0,000)	0,0039 (0,000)	0,0042 (0,000)
Этаж/макс. этаж	0,024 (0,007)	0,035 (0,006)	0,036 (0,000)
Расстояние до Кремля	–0,049 (0,000)	–0,038 (0,000)	–0,038 (0,000)
Расстояние до метро	–0,032 (0,002)	–0,035 (0,002)	–0,025 (0,002)
Кухня/площадь	0,71 (0,035)	0,61 (0,031)	0,59 (0,03)
Последний этаж	–0,054 (0,006)	–0,049 (0,006)	–0,044 (0,005)
Первый этаж	–0,11 (0,007)	–0,95 (0,006)	–0,96 (0,006)
Комната 1	–0,024 (0,01)*	–0,08 (0,009)	–0,083 (0,009)
Комната 2	0,1 (0,008)	0,058 (0,007)	0,054 (0,007)
Комната 3	0,11 (0,006)	0,082 (0,006)	0,08 (0,005)
САО	–	–0,32 (0,005)	–0,28 (0,007)
ЮВАО	–	–0,49 (0,006)	–0,43 (0,007)
СЗАО	–	–0,21 (0,006)	–0,17 (0,008)
СВАО	–	–0,35 (0,006)	–0,31 (0,007)
ВАО	–	–0,41 (0,006)	–0,37 (0,007)
ЮАО	–	–0,34 (0,006)	–0,3 (0,007)
ЮЗАО	–	–0,23 (0,006)	–0,20 (0,007)
ЗАО	–	–0,19 (0,005)	–0,17 (0,007)
Линия 1	–	–	0,15 (0,005)
Линия 2	–	–	0,08 (0,005)
Линия 3	–	–	0,0042 (0,005)*
Линия 4	–	–	0,08 (0,007)
Линия 6	–	–	0,044 (0,005)
Линия 7	–	–	0,012 (0,005)*
Линия 8	–	–	–0,062 (0,008)
Линия 8А	–	–	–0,052 (0,008)
Линия 9	–	–	0,01 (0,005)*
Линия 10	–	–	–0,046 (0,005)
Линия 11А	–	–	–0,092 (0,013)
Линия 12	–	–	0,076 (0,011)
МЦК	–	–	–0,036 (0,006)
Линия 15	–	–	–0,019 (0,008)*
Константа	9,16 (0,2)	8,4 (0,14)	7,8 (0,139)

Примечание. * – значимость на 5-процентном уровне, все остальные коэффициенты значимы на 1-процентном уровне, в скобках указана стандартная ошибка.

Таблица 4.

Сравнение моделей

	Base	Areas	Subs	Uni	Bi
RMSE	0,248	0,218	0,214	0,178	0,187
R2	0,871	0,9	0,904	0,933	0,926
F-статистика	15110	12130	7561	417	328
Количество переменных	12	20	34	1003	1159
Количество переменных (0,05)	12	20	34	375	391
Количество переменных (0,01)	11	20	30	269	264

Примечание. RMSE и R^2 измерялись на отложенной выборке.

Рисунки 3 и 4 суммируют основной результат нашей работы. На них представлены токены с наибольшими коэффициентами, где размер шрифта соответствует размеру коэффициента.

Как мы и предполагали, токены позволяют нам находить зависимость цены от различных параметров. В частности, биграммы «очистка воды», «охрана консьерж», «клубный дом», «система видеонаблюдение» и им подобные отвечают за факторы благоустройства самого жилья, которые довольно трудно учесть, если бы мы старались построить нашу модель без анализа текста объявлений. Также, используя текстуальный анализ, мы смогли добиться более мелкой гранулярности топонимов: «покровский бульвар», «христос спаситель», «трубецкой», «патриарший». Однако есть ряд токенов, которые, в противоположность ожиданиям, оказались отрицательными: «дом премиум», «бизнес класс», «пентхаус», «ЦАО».

В целом можно заключить, что использование текстуального анализа для предсказания цен помогает улучшить качество модели и находить менее очевидные взаимосвязи.

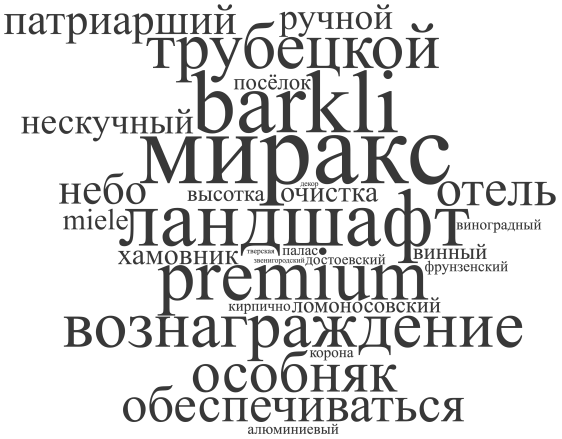

Рис. 3. Положительные униграммы

Рис. 4. Отрицательные униграммы

Рис. 5. Положительные биграммы

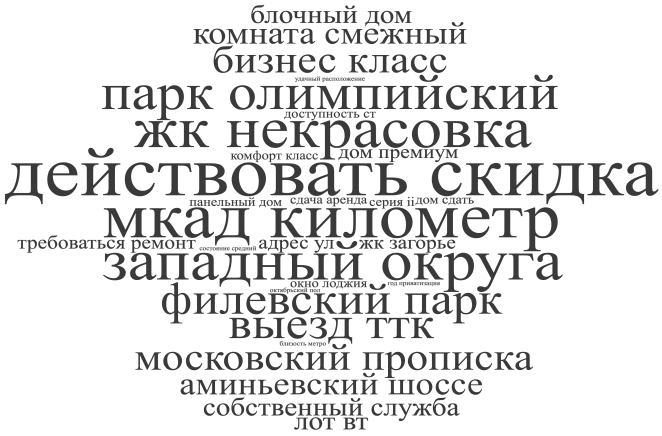

Рис. 6. Отрицательные биграммы

Примечание. На рис. 3–6 размер шрифта соответствует размеру коэффициента.

6. Выводы

В данной работе мы впервые в русскоязычной литературе провели анализ ценообразования на московскую недвижимость с использованием текстуального анализа. Мы используем алгоритмы анализа текстов объявлений о продаже недвижимости для выявления наиболее значимых предикторов цены. Этот подход позволяет включить в анализ те факторы ценообразования, которые считают важными непосредственные участники рынка, а именно продавцы квартир. Таким образом, мы выявляем важные субъективные факторы ценообразования на рынке московской недвижимости. Показано, что данный метод заметно улучшает предсказательную силу модели. Разработка алгоритма анализа текстов объявлений является важным методологическим вкладом данной работы, помимо количественных результатов регрессионного анализа.

* *
*

СПИСОК ЛИТЕРАТУРЫ

Катышев П., Хакимова Ю. Экологические факторы и ценообразование на рынке недвижимости (на примере г. Москвы) // Прикладная эконометрика. 2012. 4 (28).

Красильников А., Щербакова А. Ценообразование на вторичном рынке жилья в мегаполисах России // Экономические науки. 2011. 12. С. 103–106.

Ожегов Е., Косолапов Н., Позолотина Ю. О взаимосвязи между стоимостью жилья и характеристиками близлежащих школ // Прикладная эконометрика. 2017. 3 (28).

Сидоровых А. Оценка влияния транспортной доступности на цены недвижимости // Прикладная эконометрика. 2015. 1(37).

Belloni A., Chernozhukov V. Least Squares after Model Selection in High-dimensional Sparse Models // Bernoulli. 2013. 19 (2). P. 521–547.

Goodwin K.R., Waller B.D., Weeks H.S. Connotation and Textual Analysis in Real Estate Listings // Journal of Housing Research. 2019. 27 (2). P. 93–106.

Hausler J., Ruscheinsky J., Lang M. Newsbased Sentiment Analysis in Real Estate: A Machine Learning Approach // Journal of Property Research. 2018. 35 (4). P. 344–371.

Lawani A., Reed M. Textual Analysis and Omitted Variable Bias in hedonic Price Models Applied to Short-term Apartment Rental Market. Mimeo. 2018.

Magnus J., Peresetsky A. The Price of Moscow Apartments // Прикладная эконометрика. 2010. 1 (17).

Nowak A., Smith P. Textual Analysis in Real Estate // Journal of Applied Econometrics. 2017. 32 (4). P. 896–918.

Pryce G., Oates S. Rhetoric in the Language of Real Estate Marketing // Housing Studies. 2008. 23 (2). P. 319–348.

Textual Analysis of Pricing in the Moscow Residential Real Estate Market

Gleb Goncharov[1], **Timur Natkhov**[2]

[1] National Research University Higher School of Economics,
34, Tallinskaya st., Moscow, 123458, Russian Federation.
E-mail: gigoncharov@edu.hse.ru

[2] National Research University Higher School of Economics,
11, Pokrovsky blv., Moscow, 109028, Russian Federation.
E-mail: timur.natkhov@hse.ru

In this paper, we apply textual analysis to the hedonic pricing model in the residential real estate market of Moscow. We collect data on 60 thousand sale ads in July 2019 on the CIAN web-site (one of the largest web-sites on residential real estate market in Russia). A special parser program was written in Python to gather the data. The text analyzing algorithm developed by authors chooses words (unigrams) and phrases (bigrams) that are the most significant predictors of price. The advantage of this approach is that the selection of explanatory variables for the econometric model is based on the revealed preferences of market participants – the algorithm determines tokens indicated by apartment owners interested in a successful sale. Thus, we identify important subjective pricing factors in the Moscow real estate market. It is shown that the use of text analysis can significantly improve the predictable power of the pricing model. In particular, inclusion of unigrams reduces the standard error of estimation by 15%. The mechanism of this improvement is the inclusion of pricing factors that are difficult to quantify. For example, «water purification», «concierge guard», «club house», «video surveillance system» and similar bigrams reflect the safety, location type and other local public goods that are difficult to measure.

***Key words*:** hedonic model; LASSO; real estate.

JEL Classification: C01, C21.

* *
*

References

Belloni A., Chernozhukov V. (2013) Least Squares after Model Selection in High-dimensional Sparse Models. *Bernoulli*, 19 (2), pp. 521–547.

Goodwin K.R., Waller B.D., Weeks H.S. (2019) Connotation and Textual Analysis in Real Estate Listings. *Journal of Housing Research*, 27 (2), pp. 93–106.

Hausler J., Ruscheinsky J., Lang M. (2018) Newsbased Sentiment Analysis in Real Estate: A Machine Learning Approach. *Journal of Property Research*, 35 (4), pp. 344–371.

Katyshev P., Khakimova Y. (2012) Ekologicheskie faktory i cenoobrazovanie na rynke nedvizhimosti (na primere g. Moskvy) [Ecological Factors and the Price of Moscow Apartments]. *Applied Econometrics*, 4 (28).

Krasilnikov A., Sherbakova A. (2011) Cenoobrazovanie na vtorichnom rynke zhil'ya v megapolisah Rossii [Pricing in the Secondary Housing Market in Cities of Russia]. *Economic Sciences*, 12, pp. 103–106.

Lawani A., Reed M. (2018) *Textual Analysis and Omitted Variable Bias in hedonic Price Models Applied to Short-term Apartment Rental Market*. Mimeo.

Magnus J., Peresetsky A. (2010) The Price of Moscow Apartments. *Applied Econometrics*, 1 (17).

Nowak A., Smith P. (2017) Textual Analysis in Real Estate. *Journal of Applied Econometrics*, 32 (4), pp. 896–918.

Ozhegov E., Kosolapov N., Pozolotina Y. (2017) O vzaimosvyazi mezhdu stoimost'yu zhil'ya i harakteristikami blizlezhashchih shkol [On Dependence between Housing Value and School Characteristics]. *Applied Econometrics*, 3 (28).

Pryce G., Oates S. (2008) Rhetoric in the Language of Real Estate Marketing. *Housing Studies*, 23 (2), pp. 319–348.

Sidorovyh A. (2015) Ocenka vliyaniya transportnoj dostupnosti na ceny nedvizhimosti [Estimation of Effects of Transport Accessibility on Housing Prices]. *Applied Econometrics*, 1(37).