УДК 332.63
DOI: 10.17223/19988648/32/12

А.Л. Богданов

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ВЛИЯНИЯ РАЗЛИЧНЫХ МЕТОДОВ УЧЁТА ПРОСТРАНСТВЕННОЙ ИНФОРМАЦИИ НА ТОЧНОСТЬ МОДЕЛЕЙ ОЦЕНКИ СТОИМОСТИ ЖИЛОЙ НЕДВИЖИМОСТИ НА ПРИМЕРЕ ДВУХКОМНАТНЫХ КВАРТИР г. ТОМСКА

Проводится исследование влияния различных методов учёта информации о пространственном расположении объектов недвижимости на точность оценки их стоимости. Показано, что использование такой информации любым из способов, рассмотренных в работе, позволяет повысить точность оценки. Сравнение эффективности методов проводилось по методу t-q-кросс-валидации. Показано, что к наибольшему увеличению точности прогноза приводит использование метода K ближайших соседей, для которого исследовано распределение оптимального значения параметра K. Установлено, что оно имеет бимодальный характер.
Ключевые слова: *ценообразование на вторичном рынке жилья, регрессионный анализ пространственных данных, метод K ближайших соседей.*

1. Введение

В настоящее время благодаря сети Интернет, развитию множества коммерческих интернет-сервисов, правительственным инициативам многих государств по раскрытию данных и популяризации движения open data стали доступными большие наборы данных и сервисы, которые позволяют проводить исследования учёным и создавать новые сервисы разработчикам программного обеспечения. Развитие геоинформационных систем и сервисов открыло путь к получению различной информации о географических характеристиках тех или иных объектов, её изучения, анализа, прогнозирования и построения на её основе новых приложений и сервисов. Направление экономической науки, занимающееся изучением пространственных зависимостей, называется *пространственная статистика и эконометрика* (Spatial Statistics and Econometrics) [1–4]. В рамках этого направления разрабатываются и исследуются математические модели и методы анализа территориально распределённой информации, позволяющие учитывать пространственную изменчивость свойств объектов и влияние их окружения.

В задачах оценки стоимости недвижимости возникает естественное желание учесть ставшую в настоящее время доступной информацию о географическом расположении объектов. В данной работе рассматривается рынок вторичного жилья г. Томска. Фокус внимания сосредоточен на достаточно небольшой и сравнительно однородной группе двухкомнатных квартир, расположенных в кирпичных и панельных домах. Основной целью исследования является сравнительный анализ влияния различных способов учёта пространственной информации на точность моделей оценки стоимости жилой недвижимости.

2. Описание исходных данных

Основными источниками данных, на основе которых проводилось исследование, стали сайты информационно-поисковых систем ru09 (www.tomsk.ru09.ru) и Яндекс (yandex.ru). Первый сайт являлся поставщиком основных данных о квартирах, таких как цена, площадь, тип материала, из которого сделан дом, этаж, на котором расположена квартира, и этажность дома, район города, адрес (улица, номер дома). Выбор этого сайта был обусловлен, во-первых, большим размером базы данных о продаваемых квартирах (на момент проведения исследования (апрель 2015 г.) на сайте было зарегистрировано 14 212 объявлений, из которых 5 072 касались продажи двухкомнатных квартир); во-вторых, информация, представленная на сайте, оказалась в форме, удобной для автоматической загрузки. Второй сайт (yandex.ru) использовался для обогащения загруженных с сайта ru09 данных, а именно для проведения *геокодирования* – сопоставления каждой квартире из выборки координат дома (широты и долготы), в котором она расположена. Выбор этого сайта был обусловлен тем, что его результаты оказались точнее, чем результаты аналогичных сервисов сайтов google.com и openstreetmap.org.

Загрузка данных с сайта ru09 осуществлялась с помощью интернет-сервиса import.io, который позволяет настроить процедуру автоматической последовательной загрузки страниц с целевого сайта с последующим автоматическим разбором содержимого страниц на составляющие и сохранением результатов в csv-файл (текстовый файл с разделителями в виде табуляции) для дальнейшей обработки в аналитических системах, таких как, например, Excel, Deductor, R, SPSS и др. Загружались данные, удовлетворяющие следующим критериям:

- количество комнат – 2;
- дом – кирпичный или панельный;
- расположение – в черте города.

Загруженные данные подвергались предварительной обработке, включающей в себя: фильтрацию данных, поиск и устранение дублирующихся и противоречивых записей, обогащение данных, создание новых переменных. В частности, к загруженному набору данных были добавлены следующие переменные:

- *долгота* и *широта* – координаты расположения дома, в котором расположена квартира;
- *материал* – фиктивная переменная, принимающая значение 1, если квартира расположена в кирпичном доме, и значение 0, если квартира располагается в панельном доме;
- *этаж* – фиктивная переменная, принимающая значение 1, если квартира расположена либо на первом, либо на последнем этаже, и принимающая значение 0 в противном случае;
- *расстояние до центра* – переменная, равная расстоянию (длина кратчайшей дуги большого круга, соединяющего точки на сфере, являющегося аналогом евклидова расстояния между точками на плоскости) от дома, в котором расположена квартира, до центра города (площадь им. Ленина, широта: 56,48814107, долгота: 84,94861284).

Очищенная и подготовленная для дальнейшей обработки выборка содержала данные о 1 656 квартирах, из которых 354 находились в Ленинском, 662 – в Октябрьском, 269 – в Кировском и 371 – в Советском районах города. На рис. 1 изображена карта города с указанием границ районов.

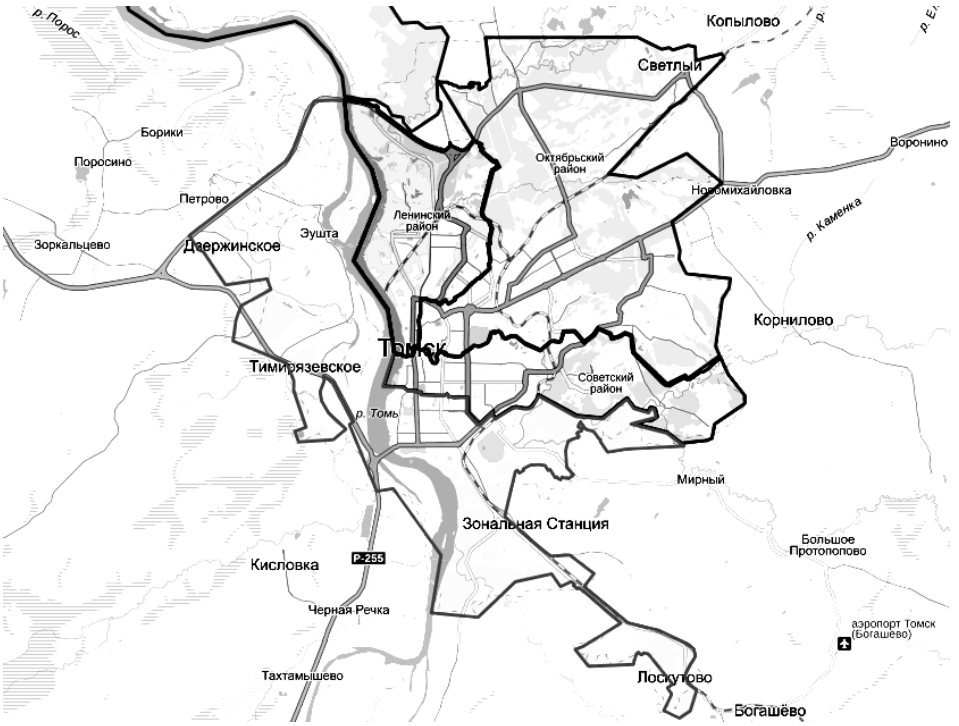

Рис. 1. Карта г. Томска с границами районов
(источник изображения: http://neagent.info/tomsk/karta/)

В табл. 1–7 даны основные статистические характеристики (минимальное, максимальное, среднее и медианное значения, величина стандартного отклонения) и выборочная корреляционная матрица, а на рис. 1–10 представлены гистограммы распределения каждой из доступных переменных, вычисленные по данным всей выборки и по каждому району в отдельности. На рис. 12 изображена карта города, на которой точками обозначено расположение квартир, попавших в выборку; размер каждой точки пропорционален цене продаваемой квартиры.

Из данных описательной статистики и выборочной корреляционной матрицы можно сделать следующие выводы:

- средние цены квартир Ленинского и Октябрьского районов ниже средней цены по городу, в то время как средние цены квартир Кировского и Советского районов выше;
- во всех районах подавляющее большинство квартир имеют цену в интервале от 2 до 4 млн руб.;

- в каждом районе имеется некоторое количество квартир с ценой более 5 млн руб., гистограммы распределения цен скошены вправо;
- наименьший разброс цен на квартиры наблюдается в Октябрьском районе, а наибольший – в Советском районе города;
- доля квартир, расположенных на первом/последнем этаже, составляет примерно треть от всего объема выборки; почти такая же пропорция наблюдается в каждом из районов;
- квартиры в панельных и кирпичных домах представлены в выборке в примерно одинаковых пропорциях; в Ленинском районе пропорции также примерно равны, в Октябрьском районе большая часть квартир расположена в панельных домах, в Кировском и Советском районах – в кирпичных;
- среднее расстояние до центра города равно примерно 4 км; наиболее удалёнными являются квартиры Октябрьского района;
- наиболее дорогие квартиры расположены в центре города, в районе, ограниченном ул. Сибирской, пр. Комсомольским, пр. Кирова и пр. Ленина, а также вдоль центральной части пр. Ленина и улиц, расположенных вдоль берега р. Томи;
- наиболее сильная по величине связь наблюдается между ценой квартиры и её площадью (значение выборочного коэффициента корреляции равно 0,79); направление связи положительное, т.е. большему значению площади соответствует большее значение цены квартиры;
- следующая по силе связь наблюдается между ценой квартиры и её удалением от центра (значение выборочного коэффициента корреляции равно –0,34); направление связи отрицательное, т.е. при удалении от центра города цены на квартиры снижаются;
- отрицательное значение выборочного коэффициента корреляции (–0,17) между переменными *цена* и *этаж* говорит о том, что квартиры, расположенные не на первом и не на последнем этажах, ценятся выше, чем квартиры, расположенные на этих этажах;
- положительное значение выборочного коэффициента корреляции (0,21) между переменными *цена* и *материал* свидетельствует о том, что квартиры в кирпичных домах ценятся выше, чем квартиры в панельных домах;
- отрицательное значение коэффициента корреляции (–0,3) между переменными *расстояние* и *материал* говорит о том, что в исследуемой выборке кирпичные дома располагаются ближе к центру города чаще, чем панельные (среднее расстояние до центра города для квартир в кирпичных домах равно 3,7 км, а для квартир в панельных домах – 4,78 км);
- малые значения коэффициентов корреляции (0,02; 0,01 и 0,00) между парами переменных соответственно *материал – площадь*, *материал – этаж* и *расстояние – этаж* свидетельствуют об отсутствии линейной взаимосвязи между этими переменными.

Таблица 1. **Описательная статистика: распределение квартир по районам, кол-во**

Характеристика	Город	Район			
		Ленинский	Октябрьский	Кировский	Советский
Количество	1656	354	662	269	371

Таблица 2. **Описательная статистика: цена, тыс. руб.**

Характеристика	Город	Район			
		Ленинский	Октябрьский	Кировский	Советский
Среднее	2824	2731	2630	2893	3206
Медиана	2600	2550	2620	2550	2800
Минимум	1200	1200	1300	1530	1220
Максимум	9850	6370	8500	7500	9850
Ст.отклонение	935.56	803.51	647.85	1017.34	1254.82

Таблица 3. **Описательная статистика: площадь, кв.м**

Характеристика	Город	Район			
		Ленинский	Октябрьский	Кировский	Советский
Среднее	52,3	51,47	52,91	49,59	53,95
Медиана	52	51,5	54	45	53,1
Минимум	22,5	31	30	22,5	29
Максимум	126	100	85	110	126
Ст.отклонение	10,5	9,83	8,87	11,47	12,45

Таблица 4. **Описательная статистика: этаж (1 – первый или последний, иначе 0)**

Характеристика	Город	Район			
		Ленинский	Октябрьский	Кировский	Советский
Количество 1	525	120	193	105	107
Количество 0	1131	234	469	164	264

Таблица 5. **Описательная статистика: тип дома (1 – кирпичный, 0 – панельный)**

Характеристика	Город	Район			
		Ленинский	Октябрьский	Кировский	Советский
Количество 1	873	170	253	215	235
Количество 0	783	184	409	54	136

Таблица 6. **Описательная статистика: расстояние до центра города, км**

Характеристика	Город	Район			
		Ленинский	Октябрьский	Кировский	Советский
Среднее	4,211	3,336	5,651	3,531	2,970
Медиана	4,042	3,516	6,025	3,476	2,713
Минимум	0,367	0,367	0,831	1,659	0,509
Максимум	8,896	6,597	8,896	5,865	6,635
Ст.отклонение	1,8	1,33	1,6	0,9	1,18

Таблица 7. **Выборочная корреляционная матрица**

	Цена	Площадь	Этаж	Материал	Расстояние
Цена	1	0,79	–0,17	0,21	–0,34
Площадь	0,79	1	–0,12	0,02	–0,11
Этаж	–0,17	–0,12	1	0,01	0,00
Материал	0,21	0,02	0,01	1	–0,3
Расстояние	–0,34	–0,11	0,00	–0,3	1

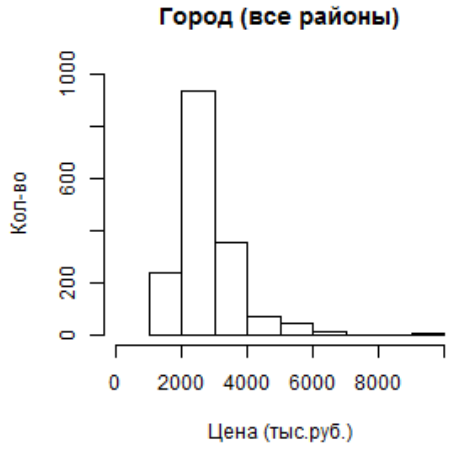
Город (все районы)
Кол-во
0
200
600
1000
0
2000
4000
6000
8000
Цена (тыс.руб.)

Рис. 2

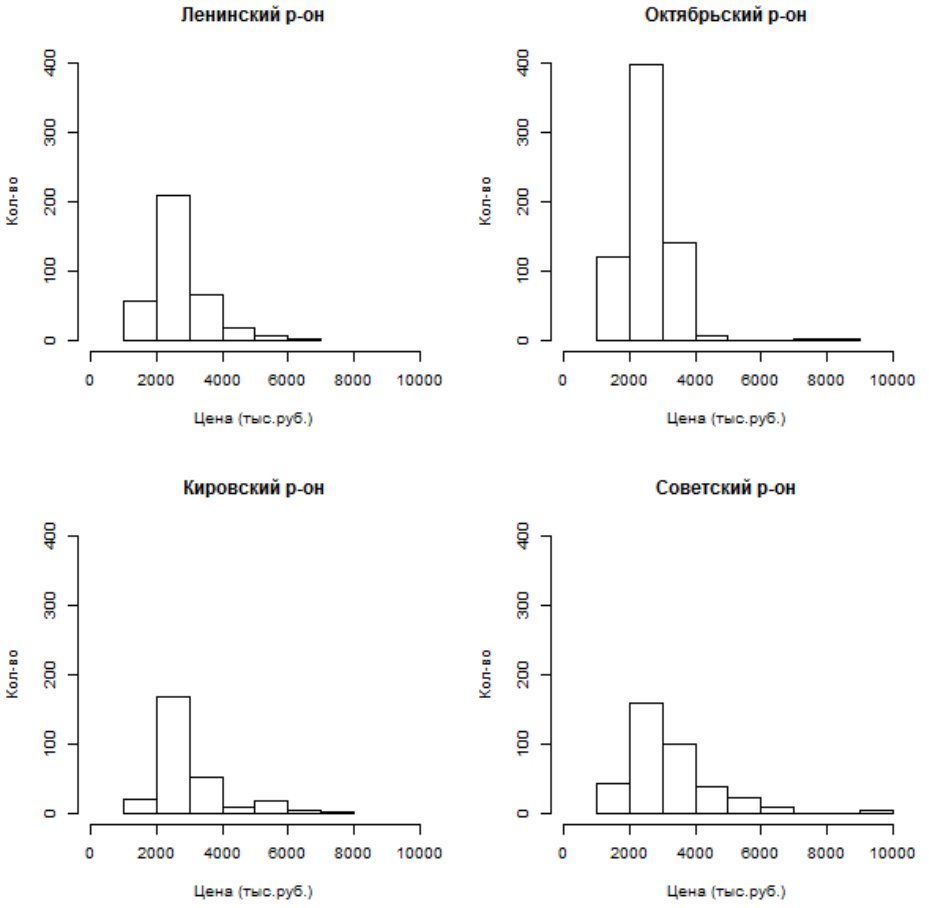
Ленинский р-он
Октябрьский р-он
Кировский р-он
Советский р-он
Кол-во
0
100
200
300
400
0
2000
4000
6000
8000
10000
Цена (тыс.руб.)

Рис. 3

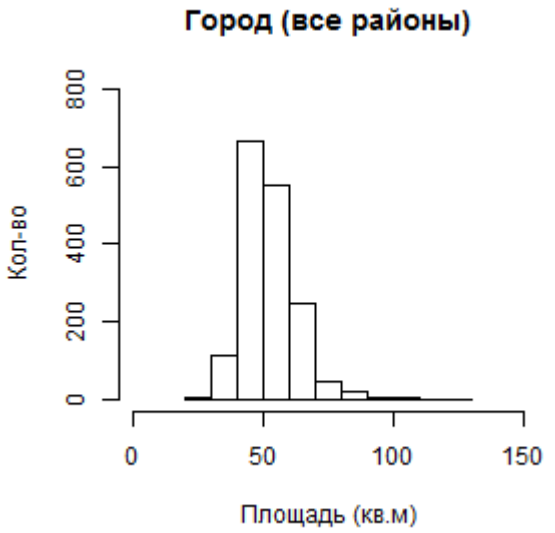
Город (все районы)
Кол-во
0
200
400
600
800
0
50
100
150
Площадь (кв.м)

Рис. 4

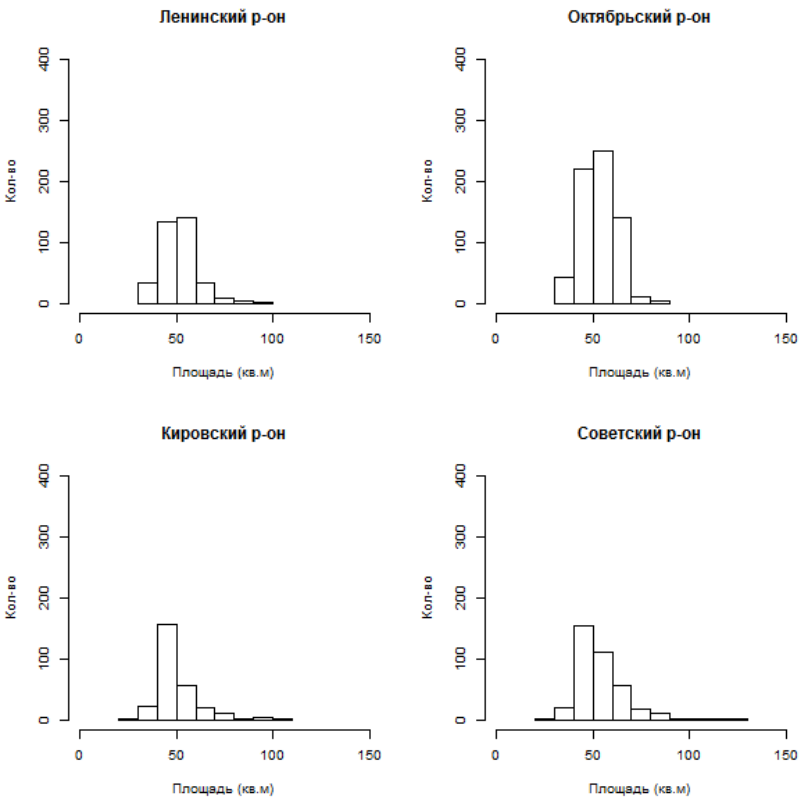
Ленинский р-он
Октябрьский р-он
Кировский р-он
Советский р-он
Кол-во
Площадь (кв.м)
0
100
200
300
400
0
50
100
150

Рис. 5


Город (все районы)
Кол-во
0
200
600
1000
не первый/
не последний
первый/
последний
Этаж

Рис. 6

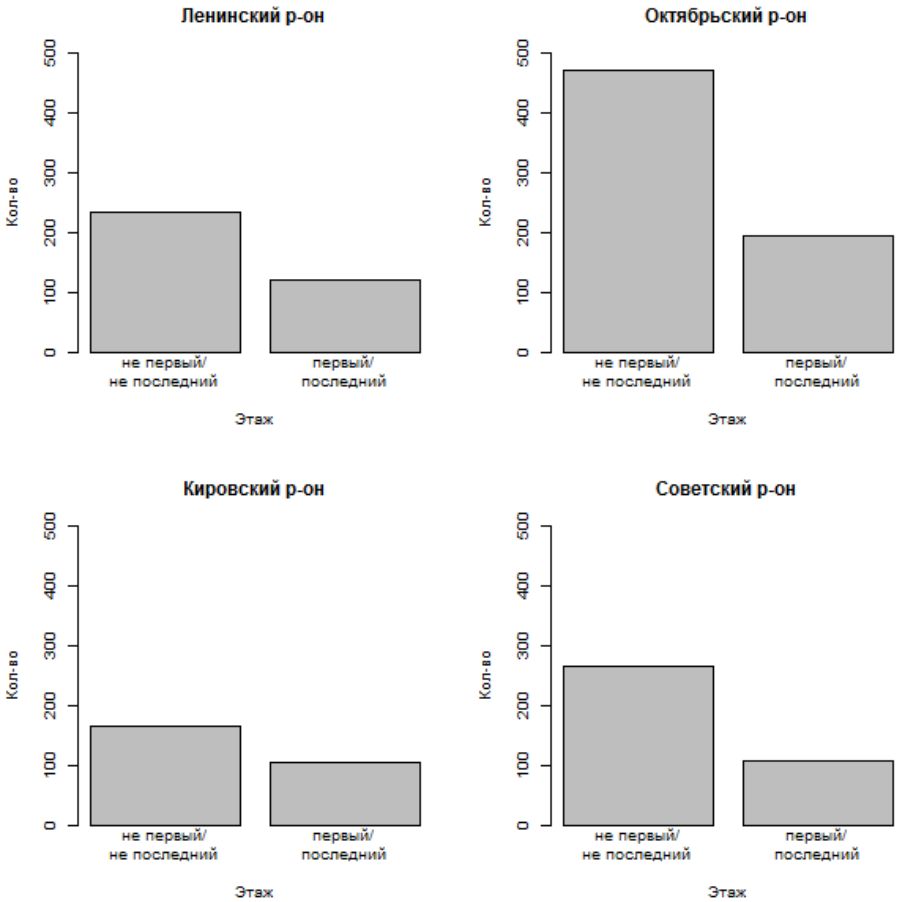
Ленинский р-он
Октябрьский р-он
Кировский р-он
Советский р-он
Кол-во
0
100
200
300
400
500
не первый/
не последний
первый/
последний
Этаж

Рис. 7

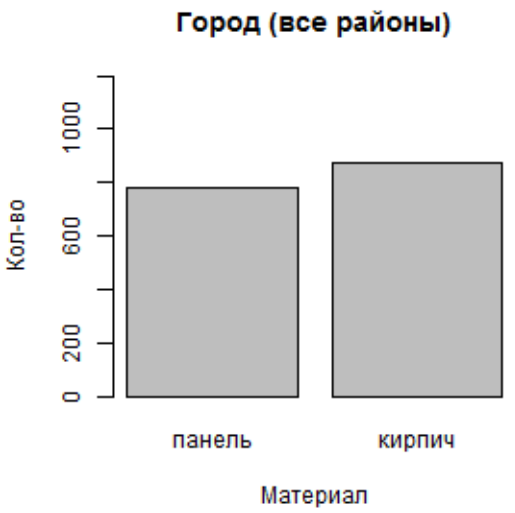
Город (все районы)
Кол-во
0
200
600
1000
панель
кирпич
Материал

Рис. 8

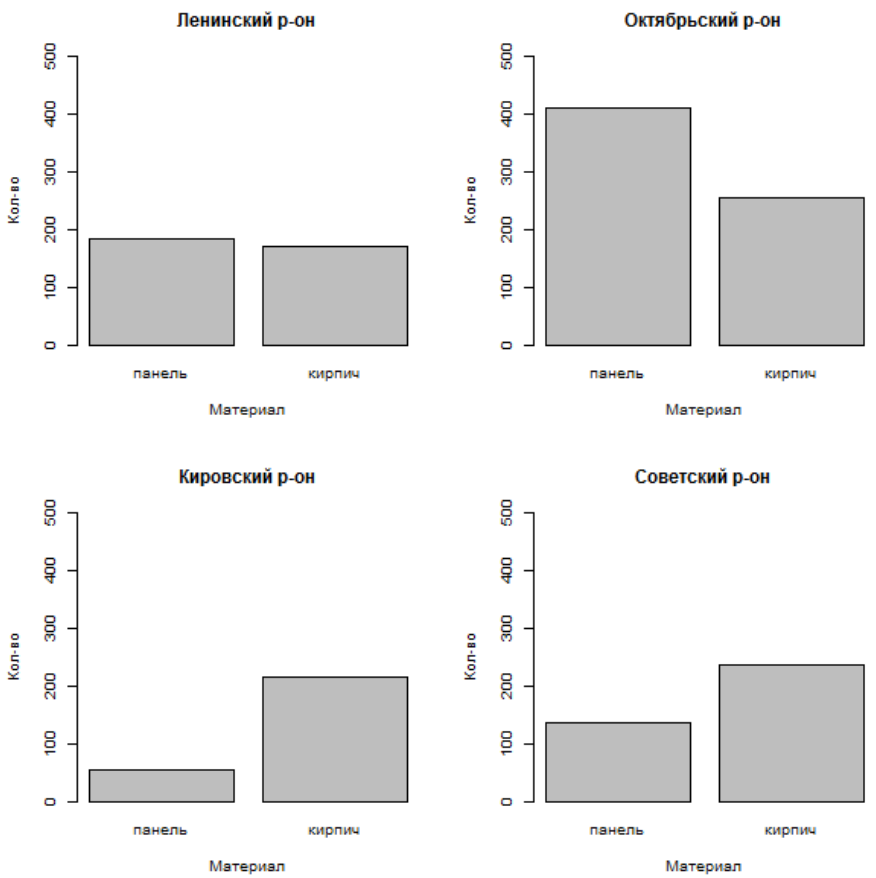
Ленинский р-он
Октябрьский р-он
Кировский р-он
Советский р-он
Кол-во
0
100
200
300
400
500
панель
кирпич
Материал

Рис. 9

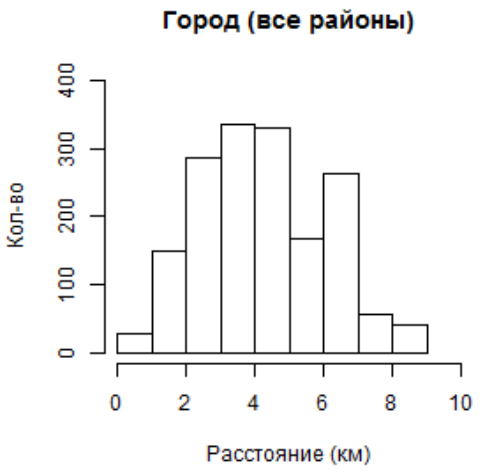
Город (все районы)
Кол-во
0
100
200
300
400
0
2
4
6
8
10
Расстояние (км)

Рис. 10

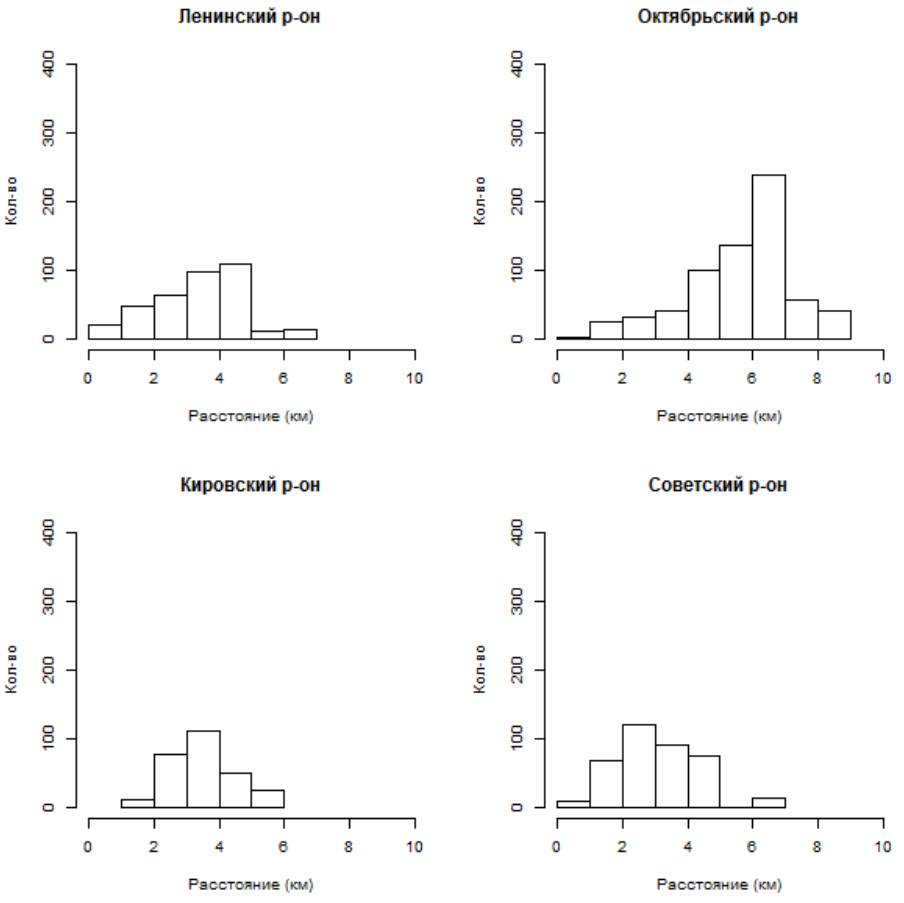
Ленинский р-он
Октябрьский р-он
Кировский р-он
Советский р-он
Кол-во
Расстояние (км)

Рис. 11

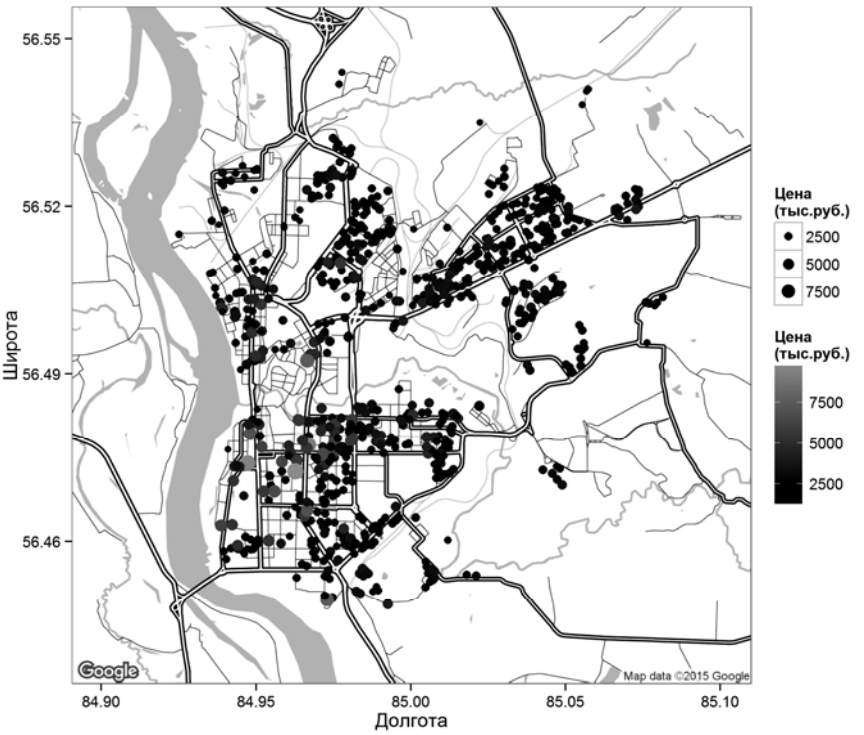

Рис. 12

3. Описание исследуемых моделей стоимости квартир

Для изучения влияния информации о местоположении квартиры на точность построения прогноза рассматривалось 8 моделей. Прежде чем переходить к их описанию, введём следующие обозначения:

- y – *цена* квартиры (тыс. руб.);
- x_1 – *площадь* квартиры (кв. м);
- x_2 – *этаж* – фиктивная переменная, принимающая значение 1, если квартира расположена или на первом или на последнем этаже, и значение 0 в противном случае;
- x_3 – *материал* – фиктивная переменная, принимающая значение 1, если квартира расположена в кирпичном доме, и значение 0, если квартира расположена в панельном доме;
- x_4 – фиктивная переменная, принимающая значение 1, если квартира расположена в Октябрьском районе, и значение 0 в противном случае;
- x_5 – фиктивная переменная, принимающая значение 1, если квартира расположена в Кировском районе, и значение 0 в противном случае;
- x_6 – фиктивная переменная, принимающая значение 1, если квартира расположена в Советском районе, и значение 0 в противном случае;
- x_7 – *расстояние* до центра (км);
- ε – прочие, не учтённые в модели факторы.

При таком способе кодирования районов города в ситуации, когда квартира располагается в Ленинском районе, фиктивные переменные x_4, x_5 и x_6 равны нулю одновременно, т.е. значение *Ленинский район* качественного признака *район* является *опорным*. При анализе результатов, касающихся районов города, все выводы будут делаться в сравнении с Ленинским районом.

Рассматривались следующие модели:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon, \quad (1)$$

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\varepsilon, \quad (2)$$

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_7x_7+\varepsilon, \quad (3)$$

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\varepsilon, \quad (4)$$

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+(\beta_{41}x_4+\beta_{51}x_5+\beta_{61}x_6)x_1+\varepsilon, \quad (5)$$

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_7x_7+\beta_{71}x_7x_1+\varepsilon, \quad (6)$$

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+(\beta_{41}x_4+\beta_{51}x_5+\beta_{61}x_6+\beta_{71}x_7)x_1+\varepsilon. \quad (7)$$

Числа β_1, β_2, ... называются *неизвестными коэффициентами* или *параметрами* модели и подлежат оцениванию на основе имеющих данных.

Гипотеза, лежащая в основе модели (1), предполагает, что цена квартиры зависит только от переменных *площадь*, *этаж* и *материал* и не зависит от расположения квартиры. Данную модель будем называть *базовой* и использовать в качестве опорной модели, с которой в дальнейшем будем сравнивать остальные модели. Коэффициентам модели можно дать следующую интерпретацию:

- коэффициент β_1 – показывает, на сколько в среднем отличаются цены двух однотипных (в смысле совпадения значений переменных *этаж* и *материал*) квартир, площадь которых отличается на 1 кв. м, т.е. можно сказать, что коэффициент β_1 – это цена 1 кв.м;
- коэффициент β_2 – показывает, на сколько в среднем отличаются цены двух однотипных (см. замечание выше) квартир, одна из которых расположена либо на первом, либо на последнем этаже, а вторая нет;
- коэффициент β_3 – показывает, на сколько в среднем отличаются цены двух однотипных (см. замечание выше) квартир, расположенных в кирпичном и панельном домах.

В модели (2) предполагается, что цена однотипных (в смысле совпадения значений переменных *площадь*, *этаж* и *материал*) квартир в разных районах города различается на некоторую фиксированную величину. Например, пусть имеется две однотипные квартиры, расположенные в Ленинском и Октябрьском районах города, тогда с учётом того, что для квартир Ленинского района переменные $x_4 = x_5 = x_6 = 0$, а для квартир Октябрьского района $x_4 = 1$ и $x_5 =$

$x_6 = 0$, модель (2) для каждой из них фактически будет выглядеть соответственно:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \qquad \text{и} \qquad y = (\beta_0 + \beta_4) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

Таким образом, коэффициент β_4 показывает, на сколько в среднем отличается цена квартиры в Октябрьском районе от цены на аналогичную квартиру в Ленинском районе. Коэффициенты β_5 и β_6 интерпретируются аналогичным образом.

В модели (3) предполагается, что цена квартиры кроме основных факторов (*площадь*, *этаж* и *материал*) в явном виде зависит от *расстояния* до центра города. В этой модели коэффициент β_7 показывает, на сколько в среднем изменяется цена квартиры с удалением от центра города на один километр.

Модель (4) является обобщением моделей (2) и (3). В ней учитывается информация как о *районе* расположения квартиры, так и о *расстоянии* до центра города.

Модели (5), (6) и (7) отличаются от моделей (2), (3) и (4) наличием дополнительных слагаемых, содержащих произведения переменных x_4, x_5, x_6 и x_7 на переменную x_1, т.е. в моделях делается предположение, что влияние переменной *площадь* на цену квартиры изменяется в зависимости от того, в каком районе города и как далеко от центра города расположена квартира. В модели (5) предполагается, что влияние переменной *площадь* зависит только от *района* города, а в модели (6) – только от *расстояния* до центра города. Модель (7) является обобщением моделей (5) и (6).

В дополнение к перечисленным моделям рассматривался подход на основе оценивания параметров модели (1) по данным K ближайших соседей. В этом случае коэффициенты модели зависят от расположения (широты p и долготы q) дома, в котором расположена квартира, на карте города:

$$y = \beta_0(p, q) + \beta_1(p, q)x_1 + \beta_2(p, q)x_2 + \beta_3(p, q)x_3 + \varepsilon. \qquad (8)$$

Данный метод оценивания называется методом *K ближайших соседей* [5], поэтому модель (8) будем называть *KNN*-моделью (от англ. K Nearest Neighbors).

4. Метод оценивания параметров моделей и метод сравнения моделей

Оценивание параметров первых семи моделей осуществлялось по методу наименьших квадратов (МНК) [5, 6]. Суть метода заключается в следующем. Пусть $b = [b_0, b_1, \ldots]$ – *вектор оценок* неизвестных коэффициентов $\beta_0, \beta_1, \ldots$ и $h(b)$ – *прогноз* значения переменной y, тогда величина $e = y - h(b)$ называется *остатком* или *ошибкой прогноза*. В методе наименьших квадратов выбор оценок неизвестных коэффициентов осуществляется в соответствии со следующим критерием:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \to \min_b , \qquad (9)$$

где e_i – ошибка прогноза для i-го наблюдения; n – размер выборки.

Оценивание параметров KNN-модели также осуществлялось методом наименьших квадратов, но в отличие от моделей (1)–(7) при построении прогноза цены i-го наблюдения использовались не все имеющиеся данные, а только данные K ближайших соседей. Метод выбора значения K описывается в следующем разделе. В ситуациях, когда матрица регрессоров оказывалась вырожденной из-за нулевой вариации одной из переменных, модель корректировалась путём удаления этой переменной; если же матрица регрессоров оказывалась вырожденной в результате полного совпадения значений переменных x_2 и x_3, то из модели удалялись обе переменные.

Оценку качества модели можно выполнять множеством различных способов: вычислением некоторого интегрального показателя качества, проведением анализа остатков, проверкой статистических гипотез о значимости оценок коэффициентов модели и т.д. В данном исследовании в качестве основного инструмента для оценки качества использовалась величина MSE, которая называется *среднеквадратичной ошибкой* (от англ. Mean Squared Error) и является естественным для метода наименьших квадратов интегральным показателем точности прогноза модели. При сравнении двух моделей по данному показателю следует выбирать модель, у которой этот показатель меньше.

В данном исследовании для повышения объективности оценки качества оцениваемых моделей, а также в процедуре поиска оптимального значения параметра K модели (8) применялся метод *кросс-валидации* [5, 6] (от англ. Cross-Validation, часто также используется термин «скользящий контроль»). Существует несколько способов реализации этого метода. Дадим краткое описание идеи применяемого в данном исследовании варианта этого метода, известного как *t-q-кросс-валидация* или *контроль по t×q блокам*.

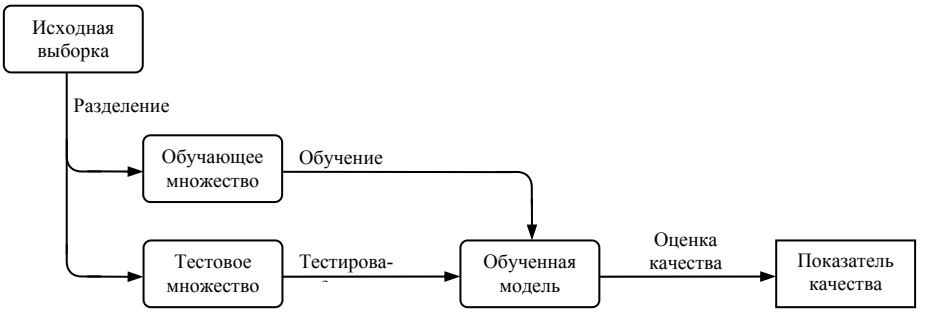

Рис. 13

Пусть выборка разбита на две части: первая называется *обучающим множеством* и используется для оценки неизвестных параметров модели,

вторая называется *тестовым множеством* и используется для оценки качества модели. Процедуру оценки параметров модели часто называют *обучением модели*, а процедуру оценки качества модели – *контролем* или *тестированием модели*. Блок-схема процесса *обучение – тестирование* изображена на рис. 13.

Зависимость рассчитываемого показателя качества от способа разбиения исходной выборки на тестовое и обучающее множества является недостатком описанного процесса, для его устранения исходную выборку разбивают на q блоков, после чего, объявляя по очереди каждый из блоков тестовым множеством, а оставшиеся блоки – обучающим множеством, процедуру *обучение – тестирование* повторяют q раз. В результате имеют q значений показателей качества, после усреднения которых получают одно усреднённое значение показателя качества модели. Данный подход называется *контролем по q-блокам* или *q-кросс-валидацией*, его блок-схема изображена на рис. 14.

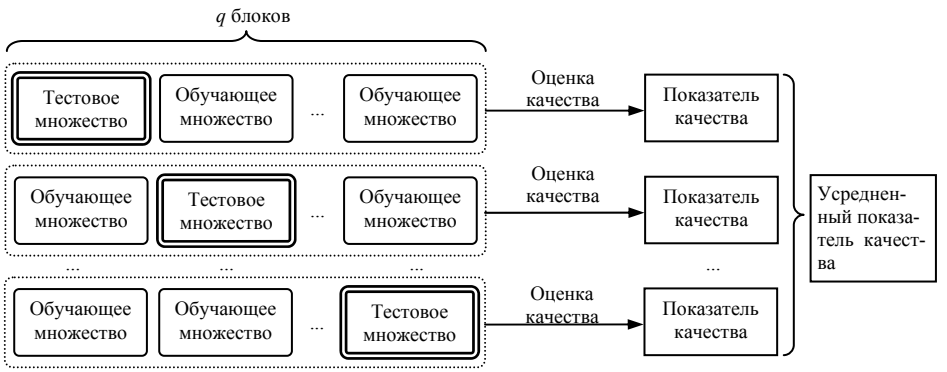

Рис. 14

В методе *контроль по t×q блокам* процедура q-кросс-валидации повторяется t раз, при этом каждый раз используется новое разбиение исходной выборки на q блоков. После завершения процедуры получается усреднённый показатель качества модели, рассчитанный по $t \times q$ промежуточным показателям качества. При таком подходе каждое наблюдение исходной выборки принимает участие в процедуре контроля ровно t раз. При сравнении моделей по методу t-q-кросс-валидации предпочтение следует отдавать той, у которой усреднённое значение показателя MSE меньше.

5. Описание и результаты вычислительного эксперимента

Вычислительный эксперимент для оценки и сравнения качества моделей (1)–(8) проводился по методу t-q-кросс-валидации со значениями параметров $t = 100$ и $q = 4$. Псевдокод эксперимента имеет следующий вид:

Делать *для каждого значения параметра t,* $t \in [1, 100]$:
Разбить *случайным образом выборку на 4 блока.*
Делать *для каждого блока с номером i, где* $i \in [1, 4]$:

Оценить *по блокам с номерами j*, где $j \in [1, 4], j \neq i$:
- *параметры моделей* (1)–(7),
- *оптимальное значение параметра K модели* (8),
- *параметры модели* (8),
- *показатели качества MSE для каждой из моделей* (1)–(8).

Поиск оптимального значения параметра K проводился в интервале значений [10, 200] с шагом 5 методом t-q-кросс-валидации со значениями параметров $t = 5$ и $q = 4$. На рис. 15 изображен график зависимости показателя качества MSE от выбранного значения K. Жирная линия изображает значение показателя MSE, усреднённое по всем $t \times q$ экспериментам ($t \in [1, 100], q \in [1, 4]$); тонкие вертикальные линии изображают диапазон изменений минимальных значений показателя MSE для каждого из значений параметра K.

На рис. 16 представлена гистограмма распределения оптимальных значений параметра K, которые были получены в каждом из $t \times q$ экспериментов ($t \in [1, 100], q \in [1, 4]$). На гистограмме хорошо видно, что распределение оптимальных значений параметра K имеет явно выраженный бимодальный характер. Это неожиданный результат, объяснение которому предстоит ещё найти. Из гистограммы видно, что значение 60 делит множество оптимальных значений параметра K на два примерно одинаковых по размеру подмножества: в первом подмножестве оказалось 177 значений и среднее значение K равно 34, во втором подмножестве оказалось 223 значения и среднее значение K равно 82. Среднее значение радиуса круга, накрывающего K ближайших соседей, в первом подмножестве равно 2,6 км, во втором подмножестве – 3,1 км; среднее значение радиуса по всем наблюдениям – 2,9 км.

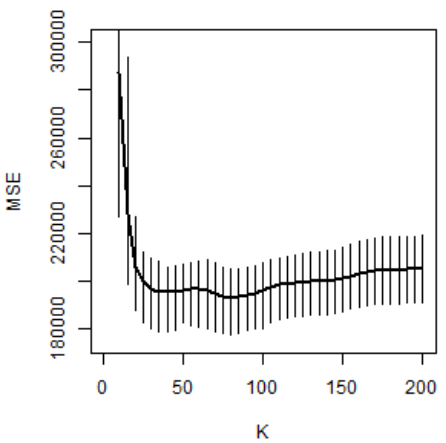

Рис. 15

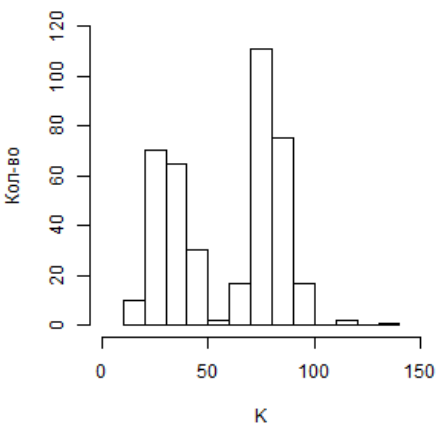

Рис. 16

Основные результаты экспериментов представлены в табл. 8 и на рис. 17–20. Из данных табл. 8 следует, что усреднённое значение показателя MSE

уменьшается с увеличением номера модели (исключение составляет модель (5), значение показателя которой незначительно (на уровне статистической погрешности) превысило значение показателя модели (4)), а также уменьшается стандартное отклонение этого показателя, т.е. можно сделать вывод, что каждая следующая модель оказывается точнее предыдущей. Из третьей графы табл. 8 можно сделать следующие выводы:

- модели (2) и (3) примерно одинаково улучшают базовую модель (модель (3) оказывается точнее на 1,55%), т.е. добавление в модель информации о районе расположения квартиры улучшает базовую модель так же, как добавление в модель информации о расстоянии до центра города;
- модель (4), обобщающая модели (2) и (3), оказывается точнее базовой модели на 15,7% и, если сравнивать с моделями (2) и (3), не приводит к значительному улучшению – её точность по сравнению с моделью (3) увеличилась на 3,06%;
- модель (5), предполагающая зависимость стоимости одного квадратного метра жилья от района расположения квартиры, оказывается сопоставимой по точности с моделью (4) – увеличение точности по сравнению с базовой моделью составляет 15,5%, что лишь на 0,02% меньше точности модели (4);
- модель (6), предполагающая изменение стоимости одного квадратного метра жилья в зависимости от удаления квартиры от центра города, повышает точность базовой модели на 22,31%, что на 6,81% больше, чем точность модели (5);
- модель (7), являющаяся обобщением моделей (5) и (6), оказывается точнее модели (6) на 3,62% и точнее базовой модели на 25,93%;
- модель (8), построенная на основе данных K ближайших соседей, где значение параметра K определялось методом t-q-кросс-валидации при $t = 5$ и $q = 4$, оказалась точнее всех моделей (1)–(7) – улучшение по сравнению с базовой моделью составило 32,11%, что на 6,15% превышает точность модели (7).

Таблица 8. **Показатели качества моделей (1)–(8)**

Модель	MSE		Улучшение	
	среднее значение	стандартное отклонение	в % к опорной модели	в % к предыдущей модели
1	2	3	4	5
1	288117,1	38389,75	–	–
2	256285,8	35959,66	11,09	11,09
3	251666,7	34638,11	12,64	1,55
4	242875,1	33710,74	15,7	3,06
5	243504,4	36077,86	15,5	–0,2
6	223552,2	31186,13	22,31	6,81
7	213101,0	29607,19	25,93	3,62
8 (KNN)	195320,7	28691,52	32,11	6,18

На рис. 17 изображены «ящики с усами» (буквальный перевод «Box and Whiskers Diagram» – распространённый способ отображения распределения значений некоторого показателя), позволяющие визуально оценить диапазон изменений показателей MSE для каждой из моделей (1)–(8):

- прямоугольник показывает диапазон, в котором сосредоточена центральная половина значений показателя; его границы являются соответствен-

но 25 и 75% квантилями; жирная линия внутри прямоугольника представляет положение медианы показателя или 50% квантили;

• пунктирные линии («усы») – это минимальное и максимальное значения показателя, не выходящие за границы 1,5 межквартильного расстояния (длины прямоугольника); наблюдения, выходящие за эти границы, принято считать выбросами и изображать точками.

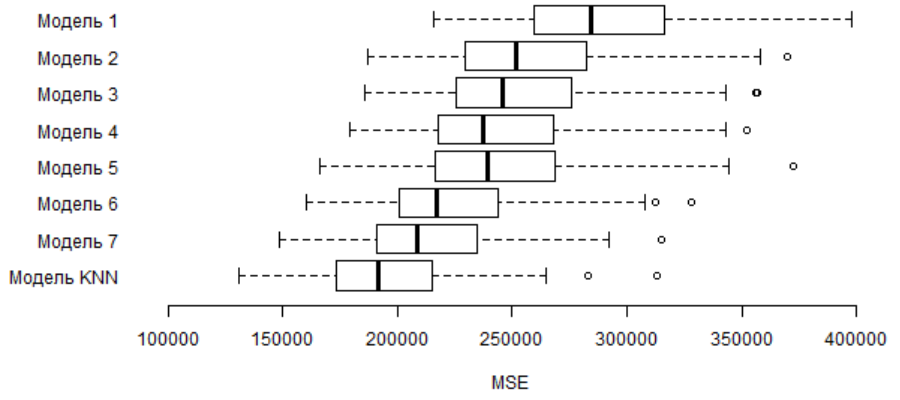

Рис. 17

На рис. 18–20 изображены графики эмпирической плотности распределения значений показателя MSE моделей (1)–(8): на первом рисунке даны графики для первой группы моделей (2)–(4), на втором – графики для второй группы моделей (5)–(7) и третьем – графики наилучших представителей каждой из групп (модели (4) и (7)) и KNN-модели; также на каждом рисунке изображен график эмпирической плотности распределения значений показателя MSE базовой модели.

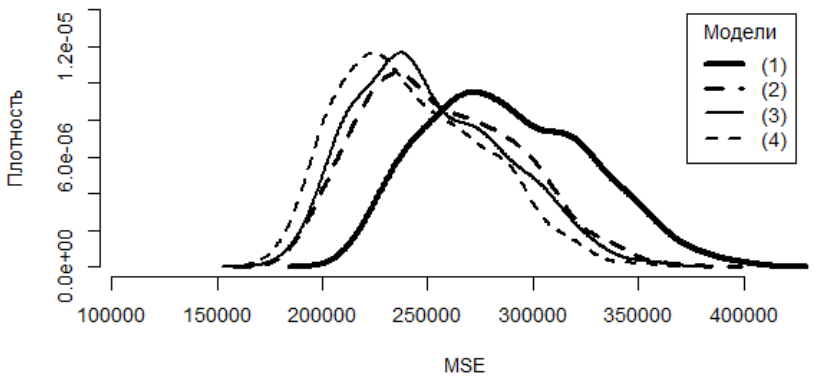

Рис. 18

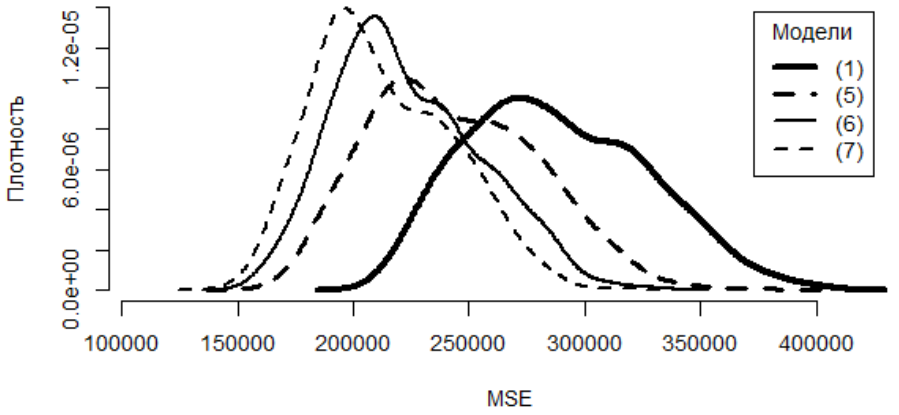

Рис. 19

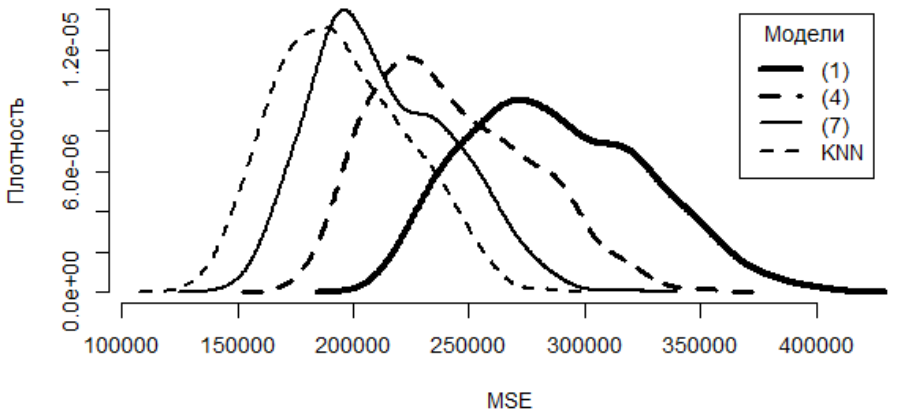

Рис. 20

В табл. 9 представлены оценки (усреднённые по всем $t = 100$ экспериментам) коэффициентов моделей (1)–(8) и их стандартные отклонения. Дадим краткую интерпретацию полученных результатов (как обычно, каждая интерпретация коэффициента перед некоторой переменной даётся при условии постоянства значений остальных переменных):

- коэффициент β_0 в данной задаче не имеет экономической интерпретации в любой из моделей (1)–(8);
- коэффициент β_2 имеет одинаковую интерпретацию в каждой из моделей (1)–(8): квартира, расположенная на первом или последнем этаже, оценивается ниже (так как все оценки коэффициентов отрицательны), чем квартира, расположенная на других этажах; значения средней величины оценок коэффициентов моделей колеблются в диапазоне от –179,4 тыс. руб. (модель (7)) до –123,7 тыс. руб. (модель (8));

• коэффициент β_3 также имеет одинаковую интерпретацию для каждой из моделей (1)–(8): квартира, расположенная в кирпичном доме, оценивается выше (так как все оценки коэффициентов положительны), чем квартира в панельном доме; значения средней величины оценок коэффициентов моделей лежат в интервале от 179,45 тыс. руб. (модель (7)) до 361,3 тыс. руб. (модель (1));

• коэффициент β_1 имеет простую интерпретацию в моделях (1)–(4) и (8), а именно, – это цена квадратного метра площади квартиры; в моделях (1)–(4) средняя цена квадратного метра колеблется в интервале от 67,6 до 70,1 тыс. руб.; в модели (8) – наиболее точной из всех рассмотренных моделей и в то же время простой, как базовая модель, среднее значение оценки этого коэффициента равно 63,6 тыс. руб.;

• интерпретация коэффициента β_1 в моделях (5)–(7) более сложная и должна рассматриваться совместно с другими коэффициентами:

– в модели (5) коэффициент β_1 – это средняя цена квадратного метра площади квартиры, расположенной в Ленинском районе города (65,2 тыс. руб.); если же квартира расположена в другом районе города, то средняя цена квадратного метра определяется как сумма коэффициента β_1 и одного из коэффициентов β_{41}, β_{51} и β_{61}: так, для квартиры в Октябрьском районе имеем $\beta_1 + \beta_{41} = 65{,}2 - 10{,}3 = 54{,}9$ тыс. руб., для квартиры в Кировском районе $\beta_1 + \beta_{51} = 65{,}2 + 8{,}36 = 73{,}56$ тыс. руб., для квартиры в Советском районе имеем $\beta_1 + \beta_{61} = 65{,}2 + 19{,}8 = 85$ тыс. руб.;

– в модели (6) коэффициент β_1 – это средняя цена квадратного метра гипотетической квартиры, расположенной в самом центре города; по мере удаления от центра цена квадратного метра корректируется на величину $\beta_{71} = -8{,}5$ тыс. руб. на каждый километр;

– в модели (7) коэффициент β_1 – это средняя цена квадратного метра гипотетической квартиры Ленинского района, расположенной в самом центре города; интерпретация коэффициентов β_{41}, β_{51}, β_{61} и β_{71} такая же, как в моделях (5) и (6);

– интерпретация коэффициентов β_4, β_5 и β_6 во всех моделях, где они используются, следующая: они показывают, на сколько в среднем цена квартиры, расположенной соответственно в Октябрьском, Кировском и Советском районах, отличается от цены квартиры Ленинского района.

На рис. 21–23 представлены графики эмпирической плотности распределения значений коэффициентов β_1, β_2 и β_3 базовой и KNN-моделей, рассчитанных по результатам всех $t = 100$ экспериментов; вертикальными линиями обозначены средние значения соответствующих коэффициентов. На рис. 24–29 представлены карты г. Томска, на которых точками обозначены значения оценки коэффициента β_1 (стоимость 1 кв. м): на рис. 24 – для модели (1), на рис. 25–27 – для моделей (5)–(7) соответственно, на рис. 28 – для модели (8) при значении параметра $K = 34$ и на рис. 29 – для модели (8) с параметром $K = 82$.

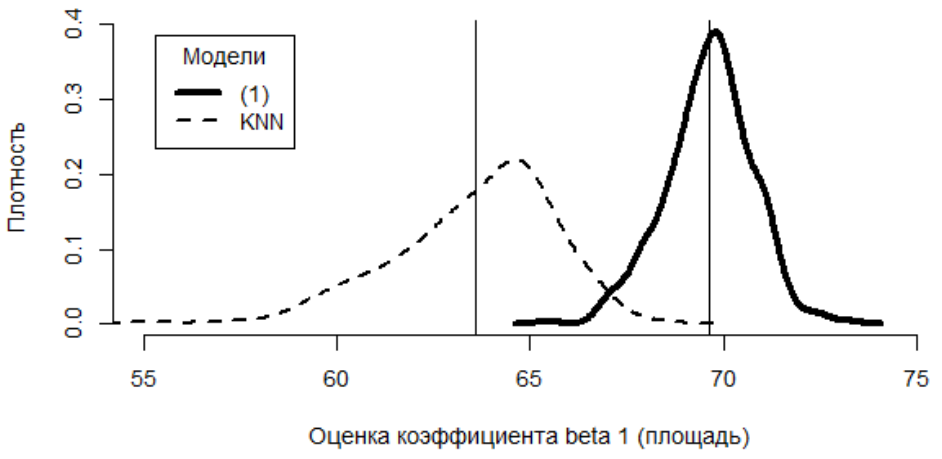
Модели
(1)
KNN
Плотность
0.0
0.1
0.2
0.3
0.4
55
60
65
70
75
Оценка коэффициента beta 1 (площадь)

Рис. 21

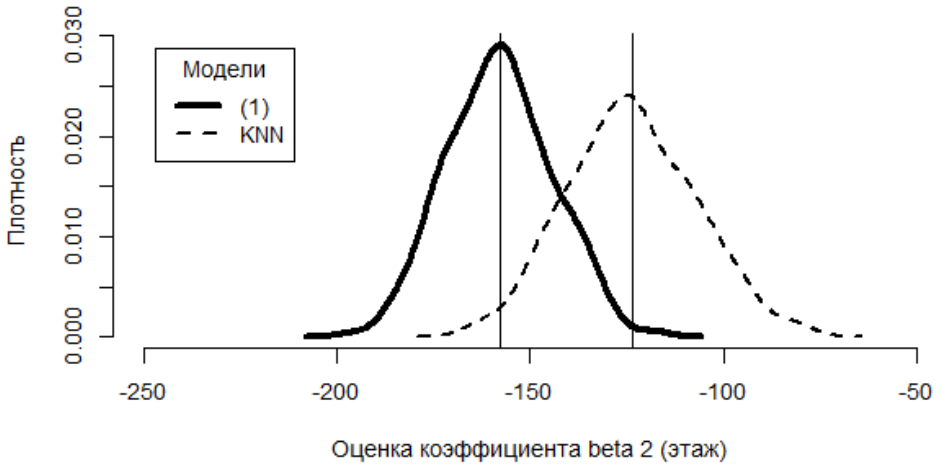
Модели
(1)
KNN
Плотность
0.000
0.010
0.020
0.030
-250
-200
-150
-100
-50
Оценка коэффициента beta 2 (этаж)

Рис. 22

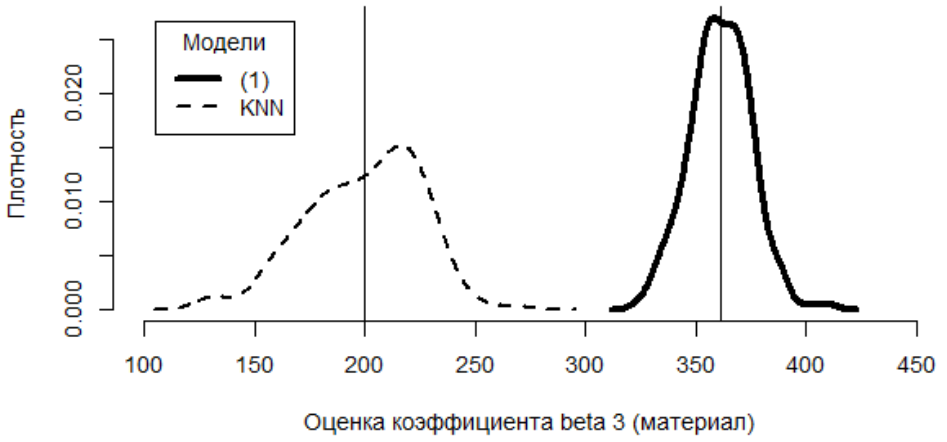

Рис. 23

Таблица 9. **Оценки коэффициентов моделей (1)–(8)**

Модель	Оценки коэффициентов: среднее значение (стандартное отклонение)											
	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_{41}	β_{51}	β_{61}	β_{71}
(1)	–958.6 (58.9)	69.6 (1.1)	– 158.0 (13.7)	361.3 (14.0)	–	–	–	–	–	–	–	–
(2)	–942.5 (59.0)	70.1 (1.1)	– 168.4 (12.8)	255.0 (13.6)	– 185.2 (17.3)	220.4 (23.9)	252.8 (21.7)		–	–	–	–
(3)	–316.1 (62.2)	67.6 (1.05)	– 160.8 (13.0)	242.6 4 (13.7)	–	–	–	– 111.6 (4.7)	–	–	–	–
(4)	–519.6 (67.1)	68.2 (1.04)	– 164.6 (12.7)	204.1 (13.4)	20.3 (23.2)	250.1 (23.7)	232.8 (21.5)	–89.6 (6.7)	–	–	–	–
(5)	–672.4 (99.4)	65.2 (1.97)	– 174.3 (12.7)	224.0 (15.0)	366.8 (153.1)	–193.0 (160.3)	–797.1 (150.8)	–	– 10.3 (3.1)	8.36 (3.4)	19.8 (3.0)	–
(6)	–1918.3 (122.8)	97.9 (2.4)	– 172.3 (12.9)	223.3 (12.0 5)	–	–	–	335.9 (24.9)	–	–	–	–8.5 (0.5)
(7)	–1565.5 (184.6)	88.0 (3.56)	–179.4 (12.86)	179.45 (12.6)	–564.9 (193.1)	–513.6 (177.2)	–588.9 (148.3)	364.3 (46.6)	11.7 (3.85)	16.0 (3.7)	15.3 (2.9)	– 8.74 (0.92)
(8)	–561.6 (119.1)	63.6 (2.1)	–123.7 (16.6)	199.6 (26.1)	–	–	–	–	–	–	–	–

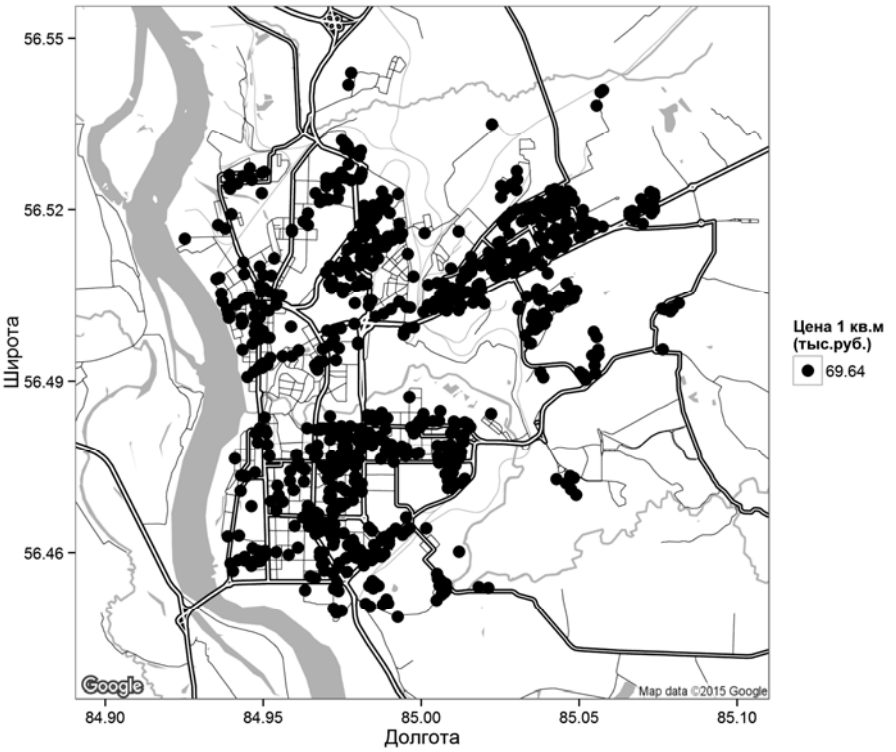
56.55
56.52
56.49
56.46
Широта
84.90
84.95
85.00
85.05
85.10
Долгота
Google
Map data ©2015 Google
Цена 1 кв.м
(тыс.руб.)
69.64

Рис. 24

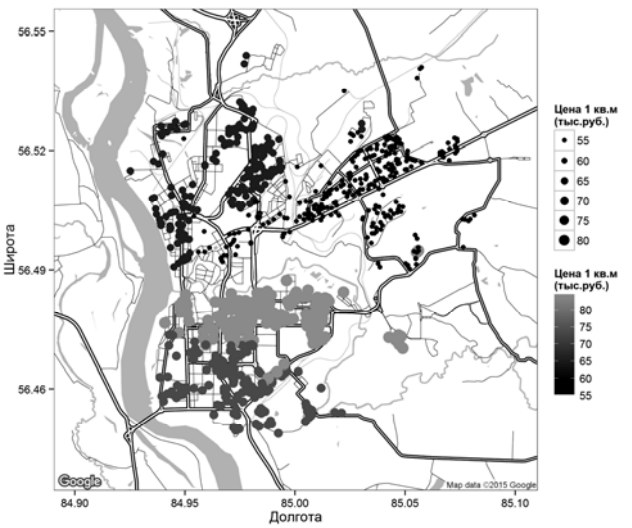
56.55
56.52
56.49
56.46
Широта
84.90
84.95
85.00
85.05
85.10
Долгота
Google
Map data ©2015 Google
Цена 1 кв.м
(тыс.руб.)
55
60
65
70
75
80
Цена 1 кв.м
(тыс.руб.)
80
75
70
65
60
55

Рис. 25

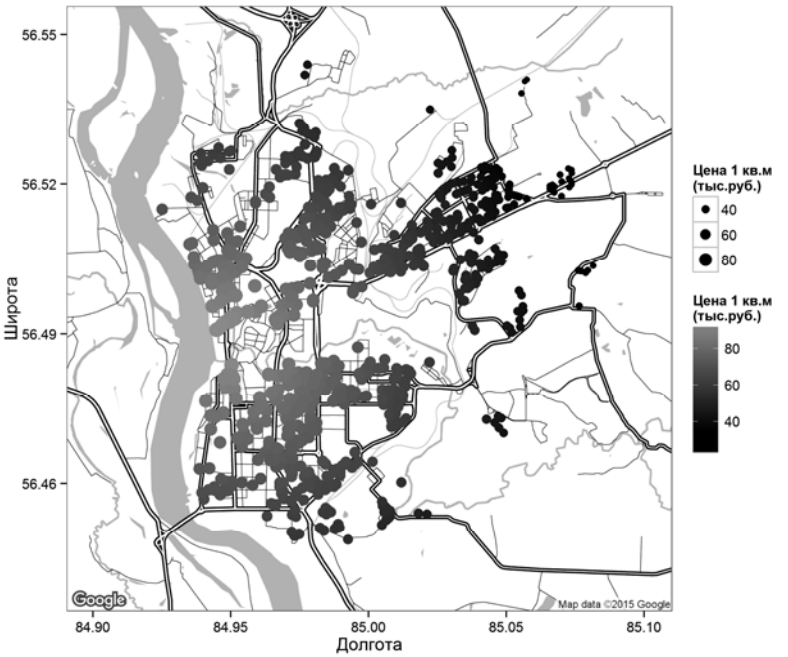
56.55
56.52
56.49
56.46
Широта
84.90
84.95
85.00
85.05
85.10
Долгота
Google
Map data ©2015 Google
Цена 1 кв.м (тыс.руб.)
40
60
80
Цена 1 кв.м (тыс.руб.)
80
60
40

Рис. 26

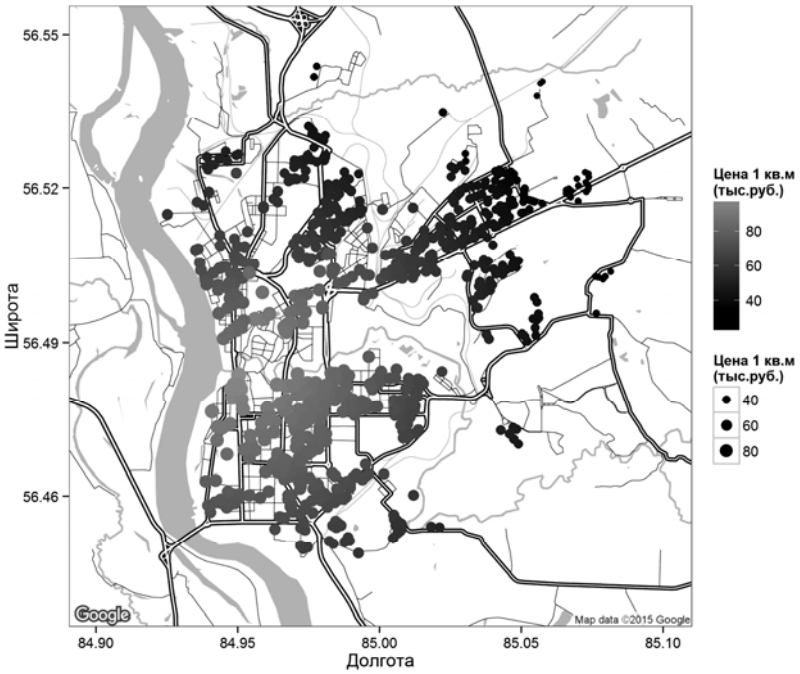
56.55
56.52
56.49
56.46
Широта
84.90
84.95
85.00
85.05
85.10
Долгота
Google
Map data ©2015 Google
Цена 1 кв.м (тыс.руб.)
80
60
40
Цена 1 кв.м (тыс.руб.)
40
60
80

Рис. 27

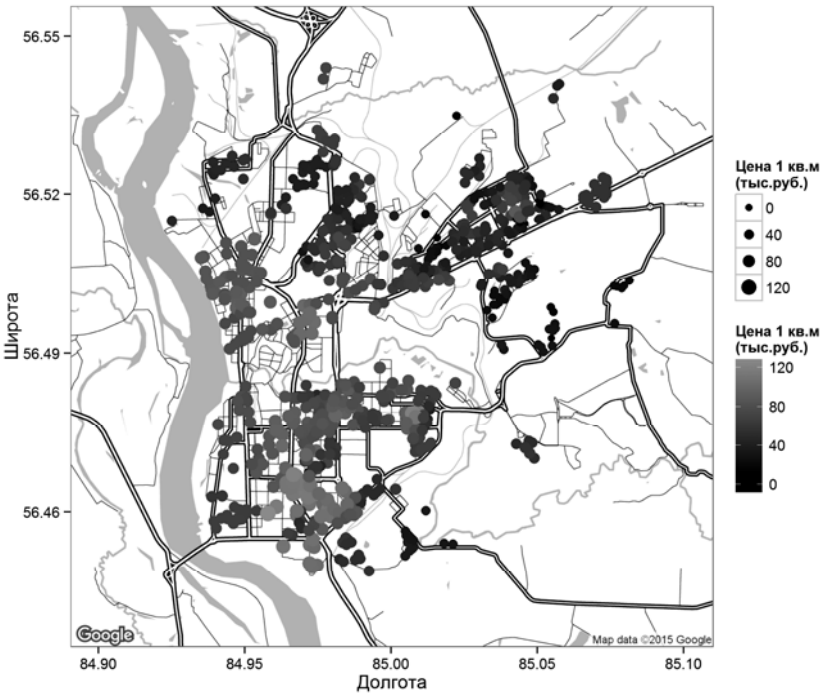
56.55
56.52
56.49
56.46
Широта
84.90
84.95
85.00
85.05
85.10
Долгота
Google
Map data ©2015 Google
Цена 1 кв.м (тыс.руб.)
0
40
80
120
Цена 1 кв.м (тыс.руб.)
120
80
40
0

Рис. 28

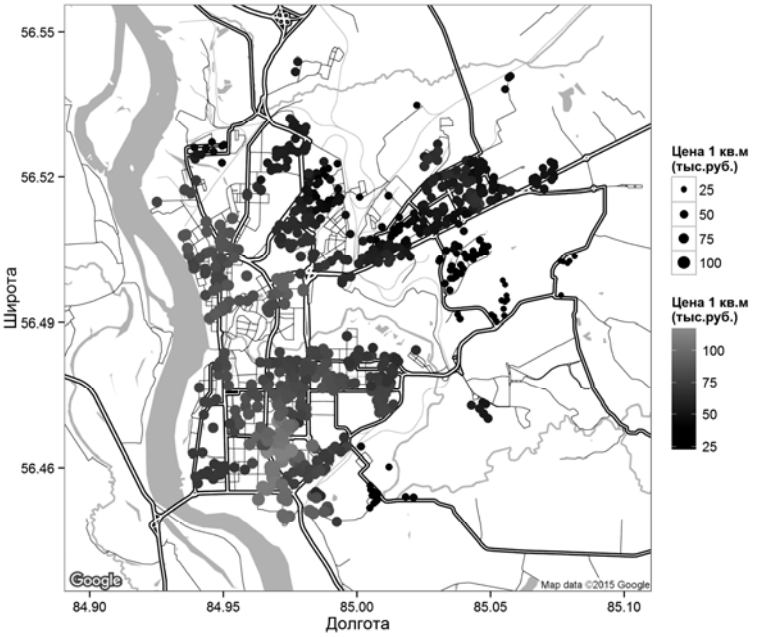
56.55
56.52
56.49
56.46
Широта
84.90
84.95
85.00
85.05
85.10
Долгота
Google
Map data ©2015 Google
Цена 1 кв.м (тыс.руб.)
25
50
75
100
Цена 1 кв.м (тыс.руб.)
100
75
50
25

Рис. 29

6. Заключение

Выполненное исследование показало, что включение в регрессионные модели информации о расположении объектов недвижимости повышает точность оценки их стоимости. Из рассмотренных методов наибольшие успехи продемонстрировал метод *K* ближайших соседей – его применение позволило повысить точность прогноза на 32% по сравнению с базовой моделью. Было установлено, что распределение оптимального значения параметра *K* имеет бимодальный характер, наиболее часто встречающиеся оптимальные значения параметра равны 34 и 82; соответствующие им средние размеры расстояний до наиболее удалённого из *K* ближайших соседей равны 2,6 и 3,1 км.

Литература

1. *Anselin L.* Spatial Econometrics: Methods and Models. Dordrecht: Kluwer Academic Publishers, 1988. 284 p.

2. *LeSage J.P.; Pace R.K.* Introduction to spatial econometrics. Boca Raton, FL: CRC Press, 2009. 374 p.

3. *Paelinck J.; Klaassen L.* Spatial econometrics. Farnborough: Saxon House, 1979. 211 p.

4. *Griffith D.A., Paelinck, J.H.P.* Non-standard Spatial Statistics and Spatial Econometrics. Heidelberg: Springer, 2011, 300 p.

5. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Heidelderg: Springer, 2013. 764 p.

6. *Magnus J.R., Neudecker H.* Matrix Differential Calculus with Applications in Statistics and Econometrics. New York: John Willey and Sons Publ., 1999. 422 p.

A.L. Bogdanov
Department of Mathematical Methods and Information Technologies in Economics, National Research Tomsk State University, Tomsk, Russia
E-mail: bogdanov.al@mail.ru
DOI: 10.17223/19988648/32/12

COMPARATIVE ANALYSIS OF THE EFFECT OF VARIOUS SPATIAL INFORMATION ACCOUNTING METHODS ON THE ACCURACY OF VALUATION MODELS OF REAL ESTATE: THE CASE OF TOMSK TWO-ROOM APARTMENTS

Keywords: Pricing in the secondary housing market; Spatial regression analysis; K-nearest neighbors.

This paper considers different methods of accounting information on the spatial arrangement of real estate items to assess the accuracy of cost estimation. The study is focused on a relatively homogeneous group of two-room apartments in brick and bearing-wall houses in the city of Tomsk. Basic data (cost, area, etc.) on the apartments was obtained from ru09.ru. Yandex.ru was used to perform geocoding.

The obtained data was filtered, cleaned and enriched. Duplicates and conflicting data were removed from the sample; values of the new variables were calculated: initial data set consisted of 5072 records; cleaned data set consisted of 1656 records. This was followed by a descriptive analysis of the downloaded data. Basic statistical characteristics were calculated and the distribution histograms were built for each variable.

We considered several ways of accounting information on the spatial arrangement of the apartments on the city map: based on models with variable structure and based on k-nearest neighbor method. The effectiveness of each model was evaluated in comparison with the efficiency of the basic model that ignored spatial information. Comparison of the models was carried out using tq-Cross Validation method.

It was established that the inclusion of information about the spatial arrangement of an apartment enables to increase the accuracy of the forecast. The approach based on k-nearest neighbor method was considered the most effective. The search for optimal parameter values was conducted in the range of [10, 200] in increments of 5 using tq-CrossValidation method. It was found that the distribution of the optimal values of the parameter K had a bimodal character. The most common parameter K equaled to

34 and 82. The comparative analysis of the models performance was conducted. Histograms of MSE parameter values distribution for each model were built. A comparative analysis of estimates of models coefficients was conducted.

References

1. Anselin L., *Spatial Econometrics: Methods and Models*. Dordrecht, Kluwer Academic Publishers, 1988. 284 p.

2. LeSage J.P., Pace R.K. *Introduction to Spatial Econometrics*. Boca Raton, FL, CRC Press, 2009. 331 p.

3. Paelinck J.H.P, Klaassen L. *Spatial Econometrics*. Farnborough, Saxon House, 1979.

4. Griffith D.A., Paelinck J.H.P. *Non-standard Spatial Statistics and Spatial Econometrics*. Heidelberg, Springer, 2011.

5. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Heidelberg, Springer, 2009. 763 p.

6. Magnus J.R., Neudecker H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York, John Willey & Sons Publ., 1999. 468 p.

For referencing:

Bogdanov A.L. Sravnitel'nyy analiz vliyaniya razlichnykh metodov ucheta prostranstvennoy informatsii na tochnost' modeley otsenki stoimosti zhiloy nedvizhimosti (na primere dvukhkomnatnykh kvartir g. Tomska) [Comparative analysis of the effect of various spatial information accounting methods on the accuracy of valuation models of real estate: the case of Tomsk two-room apartments]. *Vestnik Tomskogo gosudarstvennogo universiteta. Ekonomika – Tomsk State University Journal of Economics*, 2015, no. 4 (32), pp. 171-197.