

ПРИМЕНЕНИЕ МЕТОДОВ ГЕОСТАТИСТИЧЕСКОГО АНАЛИЗА В МОДЕЛИРОВАНИИ СТОИМОСТИ ЖИЛОЙ НЕДВИЖИМОСТИ

А.А. Бычков, Н.В. Петкова

Южный федеральный университет, г. Ростов-на-Дону
bychkov@sfedu.ru; petkova@sfedu.ru

Аннотация. В статье рассматриваются вопросы применения геостатистических методов в разработке модели оценки стоимости жилой недвижимости в рамках сравнительного подхода. Высказывается идея о необходимости учета фактора местоположения объектов недвижимости на стадии формирования выборки для статистических исследований. Приводятся результаты сравнительного анализа регрессионных моделей, построенных на выборке объектов, сгруппированных классическим способом, и модели, основанной на выборке дополнительно локализованных объектов. В качестве примера возможных способов локализации объектов дается анализ горячих точек. Обсуждается влияние геоинформационных исследований с точки зрения повышения качества расчетных моделей оценки стоимости недвижимости.

Ключевые слова: геостатистический анализ, оценка стоимости недвижимости, регрессионная модель.

В рамках сравнительного подхода к оценке недвижимости широко используются регрессионные модели, назначение которых состоит в установлении зависимости некоторой единицы сравнения (абсолютной стоимости объекта недвижимости, удельной стоимости на кв. м, куб. м и т. п.) от ценообразующих факторов на основе статистических данных. Одним из основных условий построения корректной регрессионной модели является однородность объектов сравнения в исходной выборке, что означает принадлежность этих объектов к одному сегменту рынка недвижимости, одинаковый набор ценообразующих факторов, а также единообразие в отношении влияния этих факторов на зависимую переменную. Известно, что среди ценообразующих факторов местоположение объекта недвижимости играет особую роль. Однако считается, что в моделях массовой оценки жилья, например, влияние данного фактора уже отражено в стоимости объекта недвижимости [1]. Другими словами, чем лучше местоположение, тем выше рыночная цена продажи конкретного объекта. По этой причине в оценочных моделях данный фактор зачастую либо не рассматривается в качестве независимой переменной, либо учитывается путем введения поправочных коэффициентов, значение которых может установить эксперт.

Геоинформационные системы располагают широким спектром инструментов геостатистического анализа, которые можно использовать на

разных этапах регрессионного моделирования в целях выявления структурных закономерностей в пространственном распределении объектов, комплексной оценки местоположения и в других задачах.

В данной работе рассмотрен вопрос совместного применения методов классической статистики и геоинформационного анализа для оценки недвижимости на примере данных о сделках купли-продажи квартир в многоквартирных домах г. Ростова-на-Дону.

Исходная база данных содержит записи о более чем 38 тыс. объектах. Для того чтобы оценить влияние именно фактора местоположения, исходная выборка отфильтрована по однокомнатным квартирам, материалу стен, окон, этажу/этажности, времени постройки и состоянию отделки. Остальные возможные ценообразующие факторы не рассматривались. В результате отбора в исходной выборке осталось около 3600 объектов (рис. 1).

№	Объект	Цена	Этаж	Комн.	Этаж	Стены	Соб.	Ст.	Окн.	Домаш.	Этаж	Возраст
81	квартира	1650	МДР	1	2	5 кир	31,5	20	6	м/пласт	хор	2000
82	квартира	2100	Чаловский	1	3	5 кир	34	19	9	м/пласт	хор	2000
801	квартира	2150	МДР	1	2	5 кир	30	17	6	м/пласт	хор	2000
996	квартира	1900	Чаловский	1	3	5 кир	31,5	17	6	м/пласт	хор	2000
1213	квартира	2150	Ромаш	1	4	5 кир	30	18	5	м/пласт	хор	2000
1342	квартира	2300	Донная	1	3	5 кир	33	16	6	м/пласт	хор	2000
1863	квартира	2800	Центр	1	4	5 кир	32	17	5	м/пласт	хор	2000
4014	квартира	1900	Чаловский	1	3	5 кир	34	19	9	м/пласт	хор	2000
4412	квартира	2180	МДР	1	3	5 кир	25,8	11	8	м/пласт	хор	2000
4613	квартира	1950	СММ	1	2	5 кир	26,4	13	6	м/пласт	хор	2000
4646	квартира	1850	СММ	1	2	5 кир	31,2	18	6	м/пласт	хор	2000
5499	квартира	2200	Александр.	1	4	5 кир	31	18	6	м/пласт	хор	2000
6636	квартира	2250	СММ	1	3	5 кир	32	17	9,3	м/пласт	хор	2000
6880	квартира	2150	СММ	1	3	5 кир	32	17	6	м/пласт	хор	2000
7128	квартира	1900	МДР	1	3	5 кир	39	22	7	м/пласт	хор	2000
7210	квартира	3700	Центр	1	2	5 кир	44	22	9	м/пласт	хор	2000
7216	квартира	2370	МДР	1	3	5 кир	38	20	9	м/пласт	хор	2000
7850	квартира	1750	Радвановка	1	2	5 кир	29	18	5	м/пласт	хор	2000
9050	квартира	2100	Александр.	1	4	5 кир	31	18	6	м/пласт	хор	2000
9275	квартира	2700	ЦБ	1	3	5 кир	31,5	17	6,5	м/пласт	хор	2000

Рис. 1. Исходная выборка сведений о купле-продаже однокомнатных квартир в г. Ростове-на-Дону (данные из открытых источников)

Анализ структурных закономерностей – один из видов геостатистического анализа, который дает ответ на вопрос, проявляют пространственные объекты (или значения, которые связаны с объектами) статистически значимую кластеризацию или они распределены хаотично. Большинство популярных геоинформационных систем поддерживают этот вид анализа и предлагают пользователям инструменты, с помощью которых можно принять или отклонить нулевую гипотезу о полной пространственной хаотичности объектов или их характеристик.

Использование инструментов анализа структурных закономерностей связано с вычислением величин p -значения и z -оценки, которые показывают, можно отклонить нулевую гипотезу или нет. Величина вероятности p -значение – это вероятность того, что наблюдаемые пространственные объекты распределены случайным образом. Когда p -значение мало, то случайность в структурных закономерностях маловероятна и можно отклонить нулевую гипотезу. Значения z -оценки являются стандартными отклонениями. Z -оценки и p -значения связаны со стандартным нормальным распределением.

Инструменты анализа локальных пространственных закономерностей работают для каждого объекта в контексте окружающих объектов и определяют, отличается ли локальная закономерность (целевой объект и его окружение) от глобальной (все объекты набора данных). Результаты вычислений z -оценки и p -значения, связанные с каждым объектом, позволяют определить, является различие статистически значимым или нет. Пространственную кластеризацию можно выполнить разными способами. Остановимся на анализе горячих точек, с помощью которого можно выделять области с устойчиво высокими (низкими) значениями характеристик.

Getis-Ord G_i^* – это статистика Гетиса – Орда, также известная как G_i^* . В ГИС данная статистика обычно используется для анализа горячих точек. С ее помощью можно определить, где объекты с высокими или низкими значениями пространственно сгруппированы статистически значимым образом [2].

Рассмотрим инструмент Hot Spot Analysis («Анализ горячих точек») из набора инструментов пространственной статистики ArcGIS, который рассчитывает статистический показатель Getis-Ord G_i^* для каждого объекта во входном наборе

данных. Статистика горячих точек для расчета G_i^* использует следующую формулу [3]:

$$G_i^* = \frac{\sum_{j=1}^n w_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}}{S \sqrt{\frac{[n \sum_{j=1}^n w_{i,j}^2 - (\sum_{j=1}^n w_{i,j})^2]}{n-1}}},$$

где x_j – атрибутивное значение для объекта j ; $w_{i,j}$ – пространственный вес между объектами i и j ; n – общее число объектов выборки. Здесь значения \bar{X} и S определяются по формулам:

$$\bar{X} = \frac{\sum_{j=1}^n x_j}{n},$$
$$S = \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - (\bar{X})^2}.$$

В данном случае G_i^* статистика является z -оценкой и дополнительные расчеты не требуются.

Инструмент Hot Spot Analysis исследует данные путем анализа каждого объекта в контексте соседних объектов. Сам по себе объект с высоким значением атрибута хоть и представляет интерес для исследования, но может не быть статистически значимой горячей точкой. Чтобы стать статистически значимой горячей точкой, объект должен не только иметь высокое значение, но и быть окруженным другими объектами с такими же высокими значениями. Локальная сумма для объекта и его соседей сравнивается пропорционально с суммой всех объектов. Когда локальная сумма существенно отличается от ожидаемой локальной суммы и когда это отличие является слишком большим, чтобы быть результатом случайного процесса, получается статистически значимая z -оценка.

Статистическая величина G_i^* , возвращенная в качестве атрибута для каждого объекта в наборе входных данных, является z -оценкой. Для статистически значимых положительных z -оценок чем больше z -оценка, тем более интенсивна кластеризация высоких значений (горячая точка). Для статистически значимых негативных z -оценок чем меньше z -оценка, тем более интенсивна кластеризация низких значений (холодная точка). Таким образом, итоговые z -оценки и p -значения указывают, в какой области пространства кластеризуются объекты с высокими или низкими значениями.

Это дает основание рассмотреть применение данного инструмента в определении территорий города с устойчиво дорогими или дешевыми объектами недвижимости. В среде ArcGIS 10.8 на примере данных об удельных стоимостях квартир

в многоквартирных домах рассмотрено, как пространственный анализ горячих точек может быть использован на первом шаге определения границ локальной ценовой зоны, соответствующей сегменту дорогого жилья в Ростове-на-Дону.

В качестве исходных данных собраны сведения о сделках купли-продажи однокомнатных квартир на вторичном рынке жилой недвижимости. Для проведения анализа в общем случае рекомендуется объем выборки не менее тридцати объектов. В данном примере для выполнения анализа было геокодировано около 500 объектов. После удаления выбросов во *Входном классе* (исходные данные) осталось 443 объекта с целевым числовым атрибутом – удельной стоимостью в тыс. руб. квадратного метра общей площади однокомнатной квартиры (рис. 2).

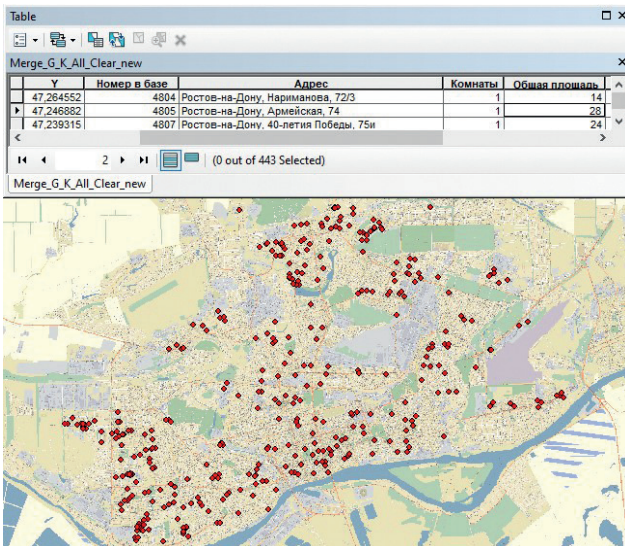


Рис. 2. Исходные данные. *Входной класс* пространственных объектов

Инструмент Hot Spot Analysis создает новый Выходной класс объектов с z-оценкой, p-значением и уровнем достоверности (G_i^*) для каждого объекта во *Входном классе* объектов. Если есть выборка, относящаяся к *Входному классу* объектов, только отобранные объекты будут проанализированы и только выбранные объекты появятся в *Выходном классе* объектов. Таким образом, изначально не требуется фильтровать исходные данные в целях типизации объектов недвижимости. Это можно сделать в процессе анализа. Отметим, что для определения пространственных взаимоотношений предлагается набор методов, из которых для данного примера выбран наиболее подходящий – Фиксированный диапазон расстояний.

В результате работы инструмента *Выходной класс* объектов автоматически добавляется к таблице содержания. Атрибутивная таблица *Выходного класса* содержит для каждого объекта z-оценку, p-значение и величину G_i^* , значение которой характеризует пространственное распределение объектов по целевому атрибуту. Из 443 проанализированных объектов всего 244 квартиры определены как статистически значимые для группировки в кластер дорогого жилья. Из них 142 имеют наибольшее значение статистики G_i^* и доверительный интервал 99 % (рис. 3).

Территориально эти объекты в основном расположены в центре города, а также на территории Западного жилого массива и Левенцовского микрорайона (рис. 4).

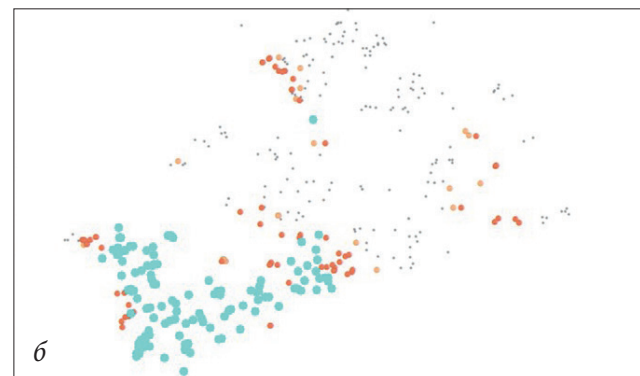
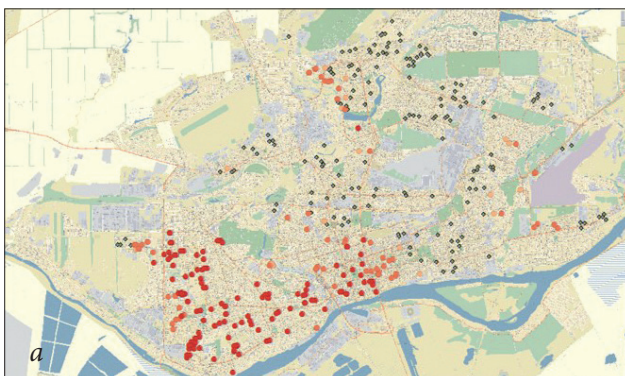


Рис. 3. *Выходной класс* объектов: а) горячие и холодные точки на карте города; б) выборка объектов с наибольшим значением G_i^*

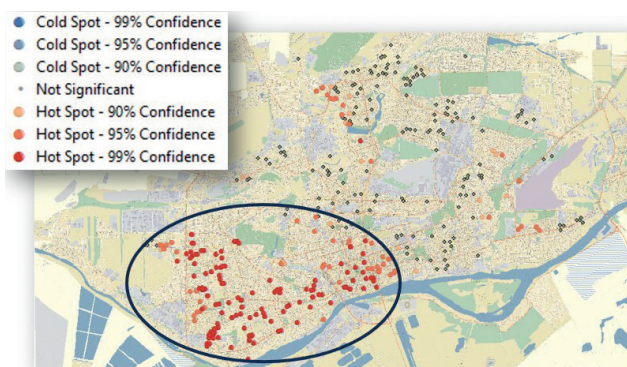


Рис. 4. Горячие точки на территории города

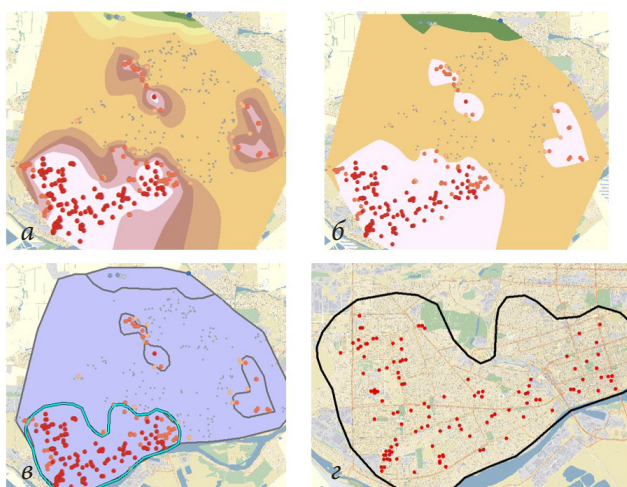


Рис. 5. Области кластеризации по результатам интерполяции: а) непрерывный растр; б) целочисленный растр; в) границы полигонов; г) область «горячего кластера»

Статистически значимых холодных точек всего 8, причем половина из них имеет наименьшее значение G_i^* . Территориально эти точки расположены на окраинах города, что согласуется с экспертным мнением риелторов. Для остальных объектов $G_i^*=0$, и они не являются статистически

значимыми для кластеризации, что говорит о произвольном пространственном распределении показателя удельной стоимости этих объектов.

В качестве примера поиска локальной ценовой зоны с дорогими квартирами выбран достоверный интервал 99 %, которому соответствует 142 объекта (рис. 3).

Для установления границ территории с дорогими квартирами (территории «горячего кластера») можно использовать различные методы ГИС. Для сравнения рассмотрены два варианта – интерполяция растров методом Natural Neighbor, а также метод построения минимальной геометрии. Входным набором данных в обоих случаях являются объекты выявленного горячего кластера.

Чтобы получить векторные границы «горячего кластера», использовано преобразование непрерывного раstra в целочисленный растр с последующей конвертацией в векторный полигональный слой. Наложение полученного в результате таких преобразований полигона «горячего кластера» на карту города дает возможность в первом приближении определить границы района, соответствующего дороговому сегменту вторичного рынка однокомнатных квартир в многоквартирных домах (рис. 5).

Метод построения минимальной ограничивающей геометрии дает менее сглаженные границы рассматриваемой области. Очевидно, что оба метода определения границ нуждаются в дополнительной ручной корректировке границ (рис. 6).

Отметим, что выявленные границы локальных ценовых зон не совпадают ни с границами административных районов, ни с границами территориальных зон города, хотя многие классические методы статистического анализа ориентированы именно на данные виды территориального деления.

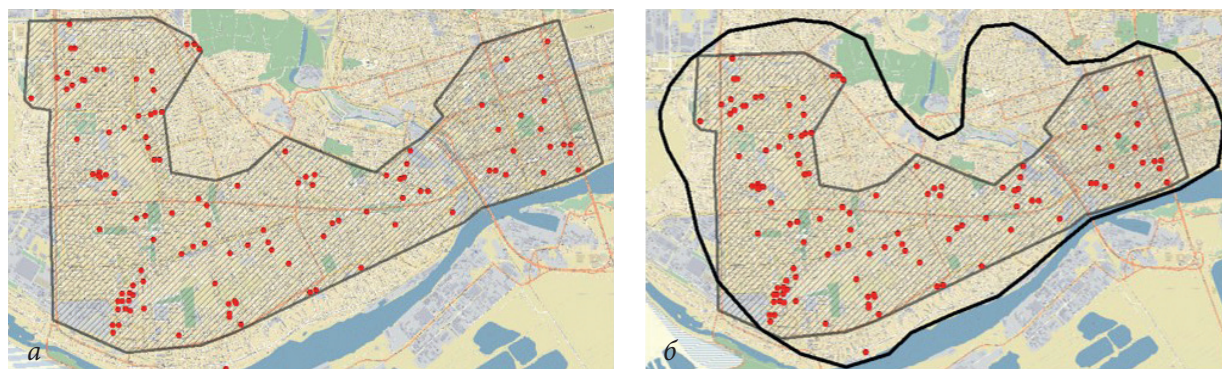


Рис. 6. Области локализации дорогих однокомнатных квартир в г. Ростове-на-Дону: а) метод Minimum Bounding Geometry; б) сравнение способов выделения территории

С помощью запроса Select by Location из исходного множества однокомнатных квартир выбраны те квартиры, которые расположены внутри выделенных областей. Всего в область «горячего кластера» было локализовано около 200 объектов.

Для того чтобы оценить влияние местоположения на качество оценочных моделей, рассмотрены три регрессионные линейные модели, построенные на разных выборках объектов недвижимости.

1. «Общая модель». Исходная (общая) выборка квартир содержит более 400 однотипных однокомнатных квартир, распределенных по всей территории города.

2. Модель «дорогие квартиры». Из исходной выборки удалены дешевые и средние по цене квартиры.

3. «Модель горячего кластера». Выборка содержит только локализованные на территории «горячего кластера» объекты.

Построение и анализ регрессионных моделей проведены с помощью пакета TIBCO Statistica 13.5 [3]. В качестве зависимой переменной рассмотрена стоимость квартиры как функция от площадных характеристик: общей площади (Soб), жилой площади (Sж) и площади кухни (Sk). Результат анализа независимых переменных на основе общей выборки квартир показывает, что независимые переменные коррелируют друг с другом (рис. 7; 8).

Это означает, что в качестве независимой переменной целесообразно рассмотреть только

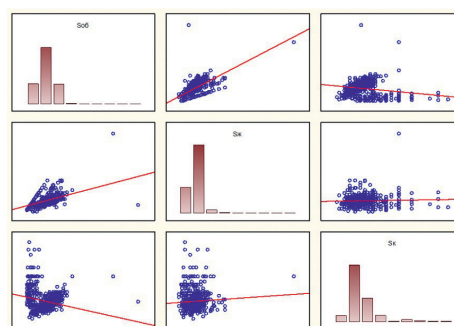


Рис. 7. Графики корреляции независимых переменных

Variable	Correlations		
	Soб	Sж	Sk
Soб	1,000000	0,541964	-0,211636
Sж	0,541964	1,000000	0,047827
Sk	-0,211636	0,047827	1,000000

Рис. 8. Матрица коэффициентов корреляции независимых переменных

одну из перечисленных характеристик, например общую площадь квартиры.

Анализ «общей модели», основанной на выборке всех однокомнатных квартир, показывает коэффициент детерминации 0,478, а коэффициент корреляции 0,69. На рисунках 9–11 приведены результаты построения и исследования данной модели.

Model is: $v_3 = b_0 + b_1 * v_9$						
Dep. Var. : Цена						
Level of confidence: 95.0% (alpha=0.050)						
	Estimate	Standard error	t-value df = 4744	p-value	Lo. Conf Limit	Up. Conf Limit
0	217,5018	31,26236	6,95730	0,000000	156,2130	278,7905
1	58,5730	0,88894	65,89090	0,000000	56,8303	60,3158

Рис. 9. t-критерий Стьюдента. Статистические характеристики коэффициентов уравнения

Effect	Model is: $v_3 = b_0 + b_1 * v_9$				
	Dep. Var.: Цена				
	1 Sum of Squares	2 DF	3 Mean Squares	4 F-value	5 p-value
Regression	2,433966E+10	2,000	1,216983E+10	20901,17	0,00
Residual	2,762222E+09	4744,000	5,822558E+05		
Total	2,710188E+10	4746,000			
Corrected Total	5,290150E+09	4745,000			
Regression vs. Corrected Total	2,433966E+10	2,000	1,216983E+10	10915,73	0,00

Рис. 10. F-критерий Фишера, определяющий статистическую значимость уравнения в целом; p-значение уравнения

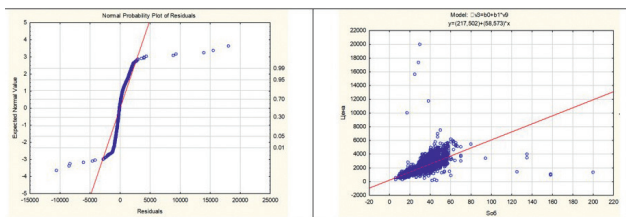


Рис. 11. Диаграмма вероятностей нормального распределения остатков и график уравнения

Анализ модели показывает, что коэффициенты полученного уравнения статистически значимы, также статистически значимо и уравнение в целом.

Нормальный вероятностный график остатков показывает, насколько близко к нормальному распределению остатков (ошибок). График строится следующим образом. Остатки (разности наблюдаемых и предсказанных значений) ранжируются. Полученные ранги используются для вычисления z-значений (т. е. значений стандартного нормального распределения) в предположении, что данные исходят из нормального вероятностного распределения, z-значения откладываются вдоль оси y на графике. В случае, когда откладываемые по оси x наблюдаемые остатки распределены нормально, соответствующие точки располагаются вдоль прямой линии. Иначе точки на графике будут иметь существенные отклонения от прямой линии. На данном

графике также становятся заметны выбросы. Если модель плохо согласуется с наблюдениями и данные располагаются особым образом (имеют, например, S-образный вид), то можно предположить, что требуется нелинейное преобразование зависимой переменной (как вариант, логарифмирование для подтягивания хвостов распределения).

Аналогичное исследование для выборки «дорогие квартиры», полученной путем отсеивания средних по цене и дешевых квартир, показывает, что коэффициент при независимой переменной не является статистически значимым. Таким образом, зависимость цены от площади квартиры не может быть описана как линейная – и в этом случае, вероятно, требуется рассматривать другой вид зависимости.

Исследование выборки для «модели горячего кластера» дает коэффициент детерминации 0,49, а коэффициент корреляции 0,7. На рисунках 12–14 приведены результаты построения и исследования данной модели.

Коэффициенты полученного уравнения статистически значимы, также статистически значимо и уравнение в целом.

Нормальный вероятностный график остатков показывает, что распределение остатков близко к нормальному, соответствующие точки располагаются вдоль прямой линии, что говорит о высоком качестве построенной модели.

Model is: $v_4 = b_0 + b_1 * v_3$ Dep. Var. : Цена Level of confidence: 95.0% ($\alpha = 0.050$)						
	Estimate	Standard error	t-value df = 193	p-value	Lo. Conf Limit	Up. Conf Limit
0	370,6766	157,4677	2,35398	0,019579	60,09804	681,2552
1	54,3184	3,9590	13,72019	0,000000	46,50987	62,1268

Рис. 12. t-критерий Стьюдента; статистические характеристики коэффициентов уравнения

Effect	Model is: $v_4 = b_0 + b_1 * v_3$ Dep. Var.: Цена				
	1 Sum of Squares	2 DF	3 Mean Squares	4 F-value	5 p-value
Regression	1,239396E+09	2,0000	619697788	2048,604	0,00
Residual	5,838204E+07	193,0000	302498		
Total	1,297778E+09	195,0000			
Corrected Total	1,153252E+08	194,0000			
Regression vs. Corrected Total	1,239396E+09	2,0000	619697788	042,455	0,00

Рис. 13. F-критерий Фишера, определяющий статистическую значимость уравнения в целом; p-значение уравнения

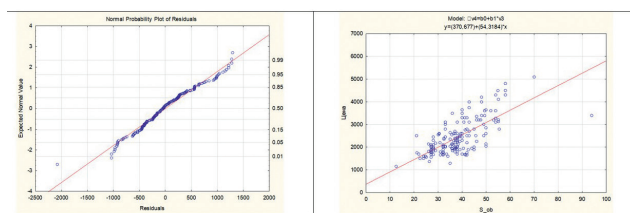


Рис. 14. Диаграмма вероятностей нормального распределения остатков и график уравнения

Как видно из полученных результатов, «модель горячего кластера» оказывается наиболее адекватной для описания линейной зависимости цены дорогих однокомнатных квартир от их площади.

В целях приблизительного сравнения построенных моделей были определены средние цены

наиболее типичных однокомнатных квартир, расположенных в центре и на окраинах города, которые имеют общую площадь от 30 до 40 кв. м. Расчеты погрешностей по «общей модели» показали, что данная модель дает существенные расхождения средней цены и расчетной стоимости объектов для квартир, которые расположены в менее престижных районах города. Модель занижает цены на квартиры в центре города и существенно завышает цены на окраинах. То есть в «общей модели» фактически не отразился пространственный фактор, хотя и предполагалось, что местоположение уже должно быть заложено в ценах на объекты. В то же время «модель горячего кластера» показала более точные результаты для объектов, расположенных на территории «горячего кластера», то есть в центре города (рис. 15).

Средняя Sob	Центр		Окраины
	"Общая" модель%	Модель "Гор.кластер" %	"Общая" модель%
30	21,01%	-4,01%	-31,65%
35	24,41%	0,94%	-22,57%
40	26,85%	4,47%	-28,02%

Рис. 15. Сравнение погрешностей моделей для объектов разной локализации

Подводя итог сравнению моделей, полученных на основе различных исходных данных, отметим, что применение традиционных статистических подходов не всегда на практике является оправданным, даже в тех случаях, когда построенные модели удовлетворяют статистическим критериям качества. Методы геоинформационного анализа помогают

правильно подготовить пространственные данные для исследования и таким образом существенно повысить качество моделей. Рассмотренный частный пример может служить основанием для разработки более общего подхода к использованию геостатистического анализа в моделировании цен на недвижимость на основе рыночных сведений.

Список литературы

1. Иванова Е.Н. Оценка стоимости недвижимости: учебное пособие для студентов ... специальности «финансы и кредит». 4-е изд. М.: КноРус, 2019. 344 с.
2. Анализ горячих точек (Getis-Ord Gi*). URL: <https://desktop.arcgis.com/ru/arcmap/latest/tools/spatial->

[statistics-toolbox/hot-spot-analysis.htm](https://desktop.arcgis.com/ru/arcmap/latest/tools/spatial-statistics-toolbox/hot-spot-analysis.htm) (дата обращения: 15.07.2024).

3. Пакет TIBCO Statistica 13.5. URL: <https://support.tibco.com/s/search/All/Home/%40uri#t=All&sort=relevancy> (дата обращения: 15.07.2024).

APPLICATION OF GEOSTATISTICAL ANALYSIS METHODS IN MODELING THE VALUE OF RESIDENTIAL REAL ESTATE

A.A. Bychkov, N.V. Petkova

Southern Federal University, Rostov-on-Don
aabychkov@sfnedu.ru; petkova@sfnedu.ru

Abstract. The paper discusses the application of geostatistical methods in the development of models for estimating the value of residential real estate within the framework of a comparative approach. The idea is expressed, that it is necessary to consider the factor of the location of real estate objects at the stage of sampling for statistical research. The results of a comparative analysis of regression models based on a sample of objects grouped in the classical way and a model based on a sample of additionally localized objects are presented. The analysis of hot spots is considered as an example of possible ways to localize objects. The influence of geoinformation research on improving the quality of computational models for estimating the value of real estate is discussed.

Keywords: geostatistical analysis, valuation of real estate, regression model.

References

1. Ivanova E.N. 2019. *Otsenka stoimosti nedvizhimosti: uchebnoe posobie dlya studentov ... spetsial'nosti "Finansy i kredit"*. [Real estate valuation: a textbook for students ... specialty "Finance and credit"]. 4th ed. Moscow, "KnoRus": 344 p. (In Russian).
2. Hot Spot Analysis (Getis-Ord Gi*). URL: <https://desktop.arcgis.com/ru/arcmap/latest/tools/spatial-statistics-toolbox/hot-spot-analysis.htm> (accessed 15 July 2024).
3. TIBCOStatistica 13.5. URL: <https://support.tibco.com/s/search/All/Home/%40uri#t=All&sort=relevancy> (accessed 15 July 2024).