В. А. Балаш, О. С. Балаш, А. В. Харламов

Эконометрический анализ геокодированных данных о ценах на жилую недвижимость

В представленной статье рассматриваются проблемы регрессионного анализа пространственных данных на примере моделирования цен вторичного рынка жилья в городе Саратове методом географически взвешенной регрессии.

Ключевые слова: географически взвешенная регрессия, ценообразование на вторичном рынке жилья, регрессионный анализ пространственных данных.

JEL classification: C21, R21.

1. Введение

В последние десятилетия существенно повысилась доступность информации, имеющей пространственную привязку, в том числе, данных геоинформационных систем (ГИС). Использование геокодированной информации значительно обогащает возможности статистического анализа, т. к. позволяет явно или неявно учитывать взаимное расположение объектов либо изменчивость изучаемого явления в пространстве.

Традиционные приемы анализа статистических данных, как правило, не используют информацию об упорядоченности объектов. В случае временных рядов естественным является представление результатов как последовательности наблюдаемых значений во времени. Соответственно, методы обработки нацелены на выявление и моделирование временной зависимости последовательных значений. Геокодированные данные отражают расположение объектов внутри некоторой области или территории. Специализированные методы и модели анализа территориально-распределенной информации позволяют учитывать изменчивость изучаемого процесса по территории или взаимосвязь значений показателей для соседних объектов или смежных областей. Спектр таких методов достаточно разнообразен. В настоящее время сформировалось научное направление пространственной статистики и эконометрики, см., например, (Anselin, 1988, 2006; Fotheringham et al., 2002; Haining, 2004; Lloyd, 2007; Schabenberger, Gotway, 2005).

В данной работе кратко обсуждаются некоторые положения, используемые при анализе пространственных данных, а также приводятся результаты применения метода географически взвешенной регрессии для моделирования цен на жилую недвижимость.

2. Представление пространственных данных при построении эконометрических моделей

При построении эконометрических моделей данные обычно представляют в виде таблицы «объект — признак», либо организуют в виде панели. Пространственные модели, кроме этого, используют информацию о взаимном расположении объектов. В общем случае ме-

тоды представления пространственных данных зависят от задач исследования и особенностей объектов наблюдения.

При анализе точечных процессов исследователь располагает результатами наблюдений для некоторого числа точек — мест проявления изучаемого явления в пространстве. Такой тип данных более характерен для естественных наук: астрономии, геологии, эпидемиологии, экологии и т. д., чем для экономических приложений. Например, это могут быть данные о местах жительства пациентов, страдающих тем или иным заболеванием, местах возникновения лесных пожаров, эпицентрах стихийных бедствий, расположении населенных пунктов и т. п. При этом географические координаты наблюдений известны и их можно внести в матрицу данных в виде дополнительных столбцов. Задачи анализа имеют определенную специфику. Примером может быть проверка, является ли расположение наблюдений на изучаемой территории случайным или имеет место какая-либо закономерность, кластеризация, и если да, то какие факторы определяют места сгущения.

В случае геостатистических данных анализу подвергаются результаты выборочного наблюдения в ограниченном числе точек. Например, результаты сделок с объектами недвижимости, совершенных в заданный период, сведения о загрязнении воздуха, получаемые от некоторого числа стационарных метеостанций и пр. Типичная задача состоит в прогнозировании уровня одного или нескольких показателей для других точек изучаемой территории — модельной цены для вновь появившегося на рынке объекта недвижимости, уровня загрязнения воздуха в выбранной точке и т. д. Координаты наблюдений известны и могут быть добавлены в таблицу данных, также возможно определить координаты точек, в которых требуется построить прогноз, например, задана равномерная сетка внутри заданной территории и т. п.

Важными аспектами при построении регрессионных моделей по геостатистическим данным является отражение пространственной зависимости и пространственной неоднородности. О пространственной зависимости говорят, если значения показателей у близлежащих объектов положительно или отрицательно коррелированы. Под пространственной неоднородностью (эффект местоположения) имеют в виду зависимость проявления изучаемого процесса или явления от уникальных, связанных с расположением, характеристик. Если каждому местоположению свойственны некоторые уникальные особенности, отличающие его от прочих, то регрессионная модель, не учитывающая пространственной неоднородности, может неадекватно описать процесс в заданной точке.

Применение стандартных методов регрессионного анализа к пространственно зависимым или неоднородным данным сопровождается рядом проблем. Среди них: неустойчивость коэффициентов модели, неправильно вычисленные стандартные ошибки коэффициентов, границы доверительных интервалов и т. д. Существует ряд способов учета пространственной зависимости в регрессионных моделях, таких как модели пространственного лага зависимой переменной, независимых переменных или случайного члена. Если полагать, что сочетание ненаблюдаемых факторов зависит от местоположения, но достаточно плавно изменяется по территории, то для отражения пространственной зависимости или неоднородности можно использовать модели с переменной структурой (переменными коэффициентами). Далее рассмотрим более подробно подход географически взвешенной регрессии.

Региональные данные представляют собой совокупность показателей, относящихся к заранее определенным территориям — странам, регионам, районам. В отличие от предыдущих случаев, наблюдаемые значения относятся к областям, т. е. протяженным объектам.

Поэтому для точного отображения взаимного расположения объектов недостаточно указать две координаты. Важным аспектом исследования может быть то, как взаимосвязаны уровни изучаемого явления в различных регионах, существует ли их пространственная автокорреляция либо пространственные экстерналии. При положительном ответе может ставиться задача оценки влияния изменения уровня развития изучаемого явления в одной из областей на показатели в других, смежных, удаленных и т. п. районах.

Общей проблемой при построении эконометрических моделей для геокодированных данных является учет пространственной неоднородности и взаимозависимости, т. е. того, как различаются и как связаны между собой наблюдаемые значения в соседних точках или областях.

Ряд методов, учитывающих взаимное расположение объектов, предполагает при построении эконометрических моделей использование пространственной матрицы весов $W(n \times n)$. Элементы весовой матрицы отражают силу потенциальных взаимодействий между объектами. Выбор способа формирования весовой матрицы — ключевой, наиболее важный и трудный этап применения большинства методов анализа территориально-распределенных данных.

В общем случае матрица пространственных весов определяется как симметричная матрица смежности, которая иногда может быть построена на основе топологической информации, представленной в ГИС, т. е. информации о близости, сопредельности объектов или расстояний между ними. Выбор метода построения весовой матрицы зависит от целей исследования. По-видимому, не существует универсального способа определения весов, который может использоваться во всех задачах. В случае геокодированных данных за основу могут браться географические расстояния, затраты времени на достижение объекта i из объекта j и т. п. При анализе региональных данных расстояния между объектами трудно определить однозначно, поэтому элементы весовой матрицы определяют, учитывая близость или сопредельность областей (Bavaud, 1998; Cliff, Ord, 1973, 1981). При решении конкретной проблемы полезно сравнить результаты, полученные при разных вариантах определения весов, а затем выбрать из них наиболее адекватный.

Многие методы расчета весовых матриц реализованы как встроенные в ряде геоинформационных систем, а также в специализированных пакетах статистического и эконометрического анализа.

3. Модели пространственной авторегрессии и географически взвешенной регрессии

Регрессионные модели с переменной структурой широко используются в практике эконометрических исследований. Например, выборка может быть разбита на несколько групп по признаку местоположения объекта в том или ином районе города, а модель расширена за счет дополнительных (фиктивных) переменных. При пересечении границ районов один или несколько коэффициентов модели меняются скачкообразно (Айвазян, Мхитарян, 2001; Магнус и др., 2006). Адаптивные методы анализа временных рядов допускают непрерывную трансформацию коэффициентов во времени. Географически взвешенная регрессия (Fotheringham et al., 2002; LeSage, 1999, 2001) может интерпретироваться как частный случай регрессионных моделей с переменной структурой при предположении, что коэффициенты модели не являются постоянными, а плавно изменяются по территории.

Не претендуя на полноту изложения, приведем несколько частных случаев пространственных эконометрических моделей с постоянными и переменными коэффициентами.

Пусть n — число наблюдений; Y — вектор значений зависимой переменных размерности n; W — заданная матрица весов размерности $n \times n$, диагональные элементы которой равны 0; (u_i, v_i) — географические координаты объектов, $i = 1,...,n$; X — матрица значений независимых переменных:

$$X = \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ ... & ... & ... & ... \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix}.$$

Определим вектор значений пространственной лаговой переменной как $Y^* = WY$.

Значения пространственной лаговой переменной часто поддаются прозрачной интерпретации. Допустим, P_i — цены объектов недвижимости. Если элементы матрицы W заданы по следующему правилу:

$$w_{ij} = \begin{cases} 1, \text{ если } j \text{ один из } k \text{ ближайших соседей } i, \\ 0, \text{ в противном случае,} \end{cases}$$

то значение $P_i^* = \dfrac{\sum_{j=1}^{n} w_{ij} P_j}{\sum_{j=1}^{n} w_{ij}}$ равно средней цене k ближайших объектов. При ином способе определения весовой матрицы результат вычисления величины пространственного лага будет равен средневзвешенному значению переменной для всех или части объектов.

Иногда удобно работать с нормированной по строкам матрицей весов W^*. Если принять, что $w_{ij}^* = \dfrac{w_{ij}}{\sum_{t=1}^{n} w_{it}}$, $i = 1,...,n$, $j = 1,...,n$, то сумма элементов каждой строки матрицы W^* равна 1, и $P_i^* = \sum_{j=1}^{n} w_{ij}^* P_j$.

Заметим, что пространственный лаг можно определить и для независимых переменных: $X^* = WX$.

Простейшая модель пространственной авторегрессии первого порядка описывает зависимость цены объекта от значений пространственного лага:

$$P_i = \alpha + \rho P_i^* + \varepsilon_i, \quad i = 1,...,n,$$

где ε_i — независимые случайные ошибки, α, ρ — неизвестные коэффициенты. В рассматриваемом примере модель предполагает зависимость цены объекта от средневзвешенных цен соседних объектов:

$$P_i = \alpha + \rho \sum_{j=1}^{n} w_{ij}^* P_j + \varepsilon_i, \; i = 1,...,n.$$

Матричная запись имеет следующий вид:

$$Y = ai_n + \rho W(Y - ai_n) + \varepsilon,$$
$$\varepsilon \sim N(0, \sigma^2 I_n),$$

где i_n — вектор размерности n, все элементы которого равны 1. Для упрощения записи обычно предполагают, что зависимая переменная центрирована, т. е. $a = 0$. Такую модель обычно называют Spatial Autoregressive Model (SAR):

$$Y = \rho WY + \varepsilon,$$
$$\varepsilon \sim N(0, \sigma^2 I_n).$$

Могут быть определены модели пространственной авторегрессии второго, третьего и более порядков, модели пространственного скользящего среднего (Spatial Moving Average Model, SMA) и т. д.

При включении дополнительных регрессоров получим смешанную модель пространственной авторегрессии (Mixed Regressive Spatial Autoregressive Model):

$$Y = \rho WY + X\beta + \varepsilon,$$
$$\varepsilon \sim N(0, I_n),$$

где β — вектор коэффициентов регрессии:

$$\beta = (\beta_0, \beta_1, ..., \beta_p)^T.$$

Пространственную автокорреляцию ошибок модели формализуют с использованием Spatial Error Model (SEM):

$$Y = X\beta + u,$$
$$u = \lambda Wu + \varepsilon,$$
$$\varepsilon \sim N(0, \sigma^2 I_n).$$

Обобщение рассмотренных случаев приводит к модели вида

$$Y = \rho W_1 Y + X\beta + u,$$
$$u = \lambda W_2 u + \varepsilon,$$
$$\varepsilon \sim N(0, \sigma^2 I_n),$$

где W_1, W_2 — заданные матрицы весов. Однако при этом необходимо определить две различные весовые матрицы. Если, например, принять $W_1 = W_2$, то возникает проблема идентификации параметров.

Кроме этого, модель может включать пространственные лаги независимых переменных

$$Y = X\beta + WX\theta + \varepsilon,$$

где W — матрица смежности, элементы на главной диагонали которой равны 0. Такие модели естественно возникают при исследовании региональных взаимодействий. На изучаемый показатель, например, уровень цен, темпы экономического роста, уровень преступности и т. д., влияют как собственные факторы, так и факторы соседних регионов.

Напомним, что в случае моделей с постоянными коэффициентами полагают, что для всех объектов изучаемой совокупности верна одна и та же «глобальная» модель. Условием ее применения является территориальная однородность изучаемой совокупности. Под однородностью совокупности имеется в виду, что коэффициенты модели одинаковы во всех подобластях.

В случае территориальной неоднородности качество модели частично удается повысить, разбив территорию на районы и построив серию локальных моделей. В предельном случае вместо задачи оценки параметров глобальной зависимости приходим к проблеме оценивания серии локально линейных моделей, коэффициенты которых зависят от местоположения объекта. Например, модели пространственной авторегрессии первого порядка с переменными коэффициентами $y_i = \alpha(u_i, v_i) + \rho(u_i, v_i) y_i^* + \varepsilon_i$ или модели с переменными коэффициентами при регрессорах $y_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i) x_{i1} + \beta_2(u_i, v_i) x_{i2} + ... + \beta_p(u_i, v_i) x_{ip} + \varepsilon_i$.

Закономерности изменения коэффициентов по территории могут быть определены исследователем, например, как заданные функции координат.

Альтернативный подход, состоящий в построении отдельной модели для каждого объекта на основании подвыборки близлежащих наблюдений, получил названия географически взвешенной регрессии (Geographically Weighted Regression, GWR).

Модель географически взвешенной регрессии имеет вид:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ik} + \varepsilon_i, \qquad (1)$$

где пара переменных (u_i, v_i) представляет координаты точки (местоположение) i, $i = 1,...,n$, y_i — значение наблюдаемой зависимой переменной; $x_{i1},\ldots,x_{ip}$ — независимые детерминированные регрессоры, p — число регрессоров; $\beta_k(u_i, v_i)$ — неизвестные коэффициенты, подлежащие оценке, $k = 0, 1,..., p$; ε_i — случайные ошибки.

Предполагается, что регрессионные модели для соседних точек схожи, но могут варьироваться по территории. Допустим, что коэффициенты регрессии $\beta_0(u,v), \beta_1(u,v),..., \beta_k(u,v)$ являются непрерывными функциями координат (u,v). Если эти функции достаточно гладкие, то коэффициенты регрессии для близлежащих объектов приблизительно равны между собой. Тогда в некоторой окрестности точки наблюдения с координатами (u_i, v_i) исследуемая зависимость с переменными коэффициентами может быть приближена локальной линейной моделью с постоянными коэффициентами:

$$y_j = \gamma_0 + \sum_{k=1}^{p} \gamma_k x_{jk} + \varepsilon_j,$$

где $\gamma_0 = \beta_0(u_i, v_i)$, $\gamma_1 = \beta_1(u_i, v_i), \ldots, \gamma_k = \beta_k(u_i, v_i)$. Для нахождения оценок коэффициентов локальной модели используют взвешенный метод наименьших квадратов, при этом ближайшие объекты учитываются с большим, а отдаленные с меньшим (нулевым) весом:

$$\sum_{j=1}^{n} w_{ij} \left(y_j - \gamma_0 - \sum_{k=1}^{p} \gamma_k x_{jk} \right)^2 \to \min_{\gamma_0, \ldots, \gamma_p},$$

где w_{ij} — вес j-го наблюдения при построении локальной модели в точке с координатами (u_i, v_i).

Локальная модель может быть представлена в матричном виде следующим образом:

$$V(u_i, v_i)Y = V(u_i, v_i)X\beta(u_i, v_i) + V(u_i, v_i)\varepsilon,$$
$$\varepsilon \sim N(0, \sigma^2 I_n),$$

где Y — вектор значений зависимых переменных размерности n; X — матрица значений независимых переменных; $\beta(u_i, v_i)$ — вектор коэффициентов регрессии в местоположении i; (u_i, v_i) — географические координаты объектов, $i = 1, \ldots, n$, $V(u_i, v_i) = W(u_i, v_i)^{1/2}$, $W(u_i, v_i)$ — диагональная матрица весовых коэффициентов размерности $n \times n$:

$$Y = \begin{pmatrix} y_0 \\ y_1 \\ \ldots \\ y_p \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \; \varepsilon = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \ldots \\ \varepsilon_n \end{pmatrix}, \; \beta(u_i, v_i) = \begin{pmatrix} \beta_o(u_i, v_i) \\ \beta_1(u_i, v_i) \\ \ldots \\ \beta_p(u_i, v_i) \end{pmatrix},$$

$$W(u_i, v_i) = \begin{bmatrix} w_{i1} & 0 & \ldots & 0 \\ 0 & w_{i2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & w_{in} \end{bmatrix}.$$

Элемент матрицы $\{w_{ij}\}$, $i, j = 1, \ldots, n$ определяет степень влияния соседей j на зависимости в местоположении i. Матрица весовых коэффициентов вычисляется для каждого местоположения.

Вектор оценок коэффициентов для каждого местоположения i вычисляется по формуле:

$$\hat{\beta}(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) Y.$$

Расчеты коэффициентов проводятся для всех измерений, в результате получают матрицу оценок параметров:

$$\hat{B} = \begin{bmatrix} \hat{\beta}_0(u_1, v_1) & \hat{\beta}_1(u_1, v_1) & \ldots & \hat{\beta}_p(u_1, v_1) \\ \hat{\beta}_0(u_2, v_2) & \hat{\beta}_1(u_2, v_2) & \ldots & \hat{\beta}_p(u_2, v_2) \\ \ldots & \ldots & \ldots & \ldots \\ \hat{\beta}_0(u_n, v_n) & \hat{\beta}_1(u_n, v_n) & \ldots & \hat{\beta}_p(u_n, v_n) \end{bmatrix},$$

где i-ая строка представляет собой вектор оценок коэффициентов в точке (u_i, v_i), $i = 1, \ldots, n$.

Так как каждому местоположению *i* соответствуют координаты (u_i, v_i), то явный вид зависимости от координат можно опустить:

$$\hat{\beta}(i) = (X^T W(i) X)^{-1} X^T W(i) Y.$$

Для проверки гипотез о значимости локальной модели рассчитывают ковариационную матрицу оценок.

Пусть $C = (X^T W(i) X)^{-1} X^T W(i)$. Вектор прогнозных значений можно представить в виде: $\hat{Y} = SY$, где $S = XC$ — матрица линейного преобразования наблюдений Y в вектор прогнозных значений $\hat{Y}$. Тогда ковариационная матрица оценок:

$$Var(\hat{\beta}(i)) = CC^T \hat{\sigma}^2,$$

где $\hat{\sigma}^2 = \sum_i \frac{(y_i - \hat{y}_i)^2}{n - 2\nu_1 + \nu_2}$ — несмещенная оценка дисперсии, $\nu_1 = tr(S)$, $\nu_2 = tr(S^T S)$, tr — след матрицы.

По аналогии с классическим случаем величину $n - 2\nu_1 + \nu_2$ можно трактовать как число степеней свободы, а $2\nu_1 - \nu_2$ — число эффективных параметров для данной точки построения регрессии. Можно отметить, что значения $\nu_1 = tr(S)$ и $\nu_2 = tr(S^T S)$ практически не различаются, поэтому число параметров считают равным $\nu_1 = tr(S)$.

Стандартные ошибки оценок коэффициентов вычисляют по формуле:

$$s(\hat{\beta}(i)) = \sqrt{Var(\hat{\beta}(i))}.$$

Опишем способы построения матрицы весовых коэффициентов.

При определении элементов матрицы применяют естественный принцип: более близкие соседи оказывают наибольшее влияние. Наиболее употребляемые методы вычисления весовых коэффициентов: административно-территориальное деление, метод движущегося окна, фиксированные и адаптивные ядра.

Часто исследуемая территория разделена на районы, например, по административному принципу. Если такое административное деление раскрывает специфические закономерности, присущие некоторым или всем административным единицам, то это учитывается в весовых коэффициентах. Для точек, принадлежащих району *A* с местоположением *i*, элемент весовой матрицы принимаем равным единице, в противном случае полагаем его равным нулю:

$$w_{ij} = 1, \text{ если } (i, j) \in A;$$
$$w_{ij} = 0, \text{ если } (i, j) \notin A.$$

Если административные районы сформированы исторически и не отражают естественное расслоение объектов, то дискретные веса определяются с учетом расстояния между исследуемыми объектами. В этом случае применяют метод движущегося окна. При этом задают предельно допустимую удаленность, т. е. некоторое фиксированное расстояние *b*, относительно которого определяют категорию ближайшего соседа. Вес принимают равным единице, если расстояние d_{ij} между объектами *i* и *j* не превосходит заданного расстояния *b*, и равным нулю в противном случае:

$$w_{ij} = 1, \text{ если } d_{ij} < b;$$
$$w_{ij} = 0, \text{ если } d_{ij} \geq b.$$

Расстояние между исследуемыми объектами находят как расстояние между точками на плоскости. Величина b фиксирована и называется шириной окна (или полосы пропускания).

Использование дискретного подхода при определении весов позволяет учесть территориальную неоднородность, но при этом модели для каждого района не связаны друг с другом. Кроме того, влияние всех соседей, попавших в полосу пропускания, считается одинаковым. Однако в большинстве случаев влияние соседей уменьшается с увеличением расстояния. Поэтому имеет смысл более близким соседям придавать больший вес, чем дальним.

Подход, в котором веса строятся с учетом непрерывного изменения расстояния между исследуемыми объектами, называют ядерным, а веса, которые являются убывающими функциями расстояния — ядрами.

Наиболее часто применяют ядра Гаусса:

$$w_{ij} = \exp\left(-\frac{\alpha}{2}\left(\frac{d_{ij}}{b}\right)^2\right),$$

где b — фиксированная ширина полосы пропускания, α — масштабный коэффициент. В местоположении i вес равен единице, а при удалении объектов исследования от него быстро уменьшается.

Как альтернативу можно использовать ядро би-квадрат:

$$w_{ij} = \begin{cases} \left(1-\left(\frac{d_{ij}}{b}\right)^2\right)^2, & \text{если } d_{ij} < b; \\ 0, & \text{если } d_{ij} \geq b. \end{cases}$$

Би-квадрат обеспечивает непрерывное изменение веса в пределах полосы пропускания и ноль за ее границей.

Еще одним примером вычисления непрерывно меняющегося веса может служить ядро три-куб:

$$w_{ij} = \begin{cases} \left(1-\left(\frac{d_{ij}}{b}\right)^3\right)^3, & \text{если } d_{ij} < b; \\ 0, & \text{если } d_{ij} \geq b. \end{cases} \quad (2)$$

Здесь убывание происходит «более круто», чем в предыдущих случаях. Больший вес возникает у ближайшего окружения заданной точки и быстро убывает при приближении к границе полосы пропускания.

Если измерения проводились на равномерной решетке, то ядра с постоянной шириной полосы пропускания дают хороший результат. Но во многих практических задачах наблюдения неравномерно расположены по территории. В этом случае использование фиксиро-

ванной ширины полосы пропускания может привести как к недостатку данных в слабо заполненных районах, и вследствие этого неустойчивости оценок коэффициентов, так и к огрублению зависимости в районах с высокой плотностью наблюдений. Чтобы избежать указанных недостатков, прибегают к использованию адаптивных ядер. Рассмотрим некоторые методы их построения.

Часто веса рассчитывают с учетом рангов. Ближайшим соседям присваивают нулевой ранг и вес, равный единице. При удалении объектов от местоположения ранг, как и расстояние, увеличивается, а вес уменьшается.

Если ширину полосы пропускания определить как расстояние до m-го соседа, то получим ядро с изменяющейся шириной полосы пропускания. В таком случае полоса автоматически меняется в зависимости от скученности точек измерения. В более густых местах — сужается, а в более разреженных — увеличивается.

Оптимальное число ближайших соседей m можно определить с помощью итеративной процедуры, сравнивая качество моделей для разных значений параметра. Для полученного оптимального числа соседей проводится расчет весов с ядром би-квадрат или три-куб. Положительные веса получают только m ближайших соседей, для остальных веса равны нулю. Например,

$$w_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{b}\right)^2\right)^2, & \text{если } j \text{ один из } m \text{ соседей;} \\ 0, & \text{иначе,} \end{cases}$$

при этом величина параметра b задается расстоянием до самого дальнего из m ближайших соседей.

Более сложный подход к построению адаптивного ядра состоит в том, что для каждого местоположения i число соседей определяется таким образом, чтобы сумма весов соседних точек измерения была постоянной:

$$\sum_j w_{ij} = c.$$

При этом веса могут быть вычислены с помощью какого-либо непрерывного ядра, например, ядра Гаусса. Так же, как в предыдущем случае, в более плотных областях ядра будут сжиматься, а в разреженных — растягиваться.

Для определения оптимального значения параметра A можно использовать итерационные процедуры, где для различных значений параметра вычисляют статистики качества (адекватности) модели, из которых выбирают наилучшую.

Очевидно, что оценки коэффициентов регрессионной модели зависят от способа расчета весов. Рассмотрим методы вычисления оптимальных значений параметров весовых функций. Так, при достаточно больших значениях ширины полосы пропускания b можно получить такие же оценки коэффициентов модели, как и в случае классической регрессии. При этом все индивидуальные местные особенности могут быть нивелированы, и тем самым необходимый эффект географического подхода может не проявиться. Напротив, при малых значениях b возникнет опасность получения незначимых и неэффективных оценок коэф-

фициентов регрессии. Следовательно, необходимо подбирать оптимальные значения параметров весовой функции.

Для определения оптимальных значений естественным подходом, на первый взгляд, является применение метода наименьших квадратов. Действительно, оценки коэффициентов модели зависят от параметров функций, используемых при расчете весов, в частности, от ширины полосы пропускания b. Поэтому прогнозные значения можно рассматривать как функцию параметра b. Оптимальное значение b можно получить, минимизируя значение функционала

$$Z = \sum_{i=1}^{n} (y_i - \hat{y}_i(b))^2.$$

Но при данном способе минимум может достигаться для малых значений b. В этом случае величина Z будет близка к нулю, и в качестве оптимального значения может быть также выбрано $b = 0$, что, естественно, противоречит здравому смыслу. Поэтому прибегают к методам взаимной ратификации и обобщенной взаимной ратификации.

Метод взаимной ратификации состоит в том, что при построении оценок коэффициентов в местоположении i саму эту точку исключают из рассмотрения. Оптимальное значение параметра b выбирается исходя из задачи минимизации функционала CV:

$$CV = \sum_{i=1}^{n} (y_i - \hat{y}_{\neq i}(b))^2 \to \min.$$

В методе обобщенной взаимной ратификации оптимальное значение b выбирается исходя из следующей задачи минимизации:

$$GCV = \frac{n}{(n - \nu_1)^2} \sum_{i=1}^{n} (y_i - \hat{y}_{\neq i}(b))^2 \to \min,$$

где величина $\nu_1 = tr(S)$, а S — матрица линейного преобразования вектора зависимой переменной Y в вектор прогнозных значений $\hat{Y}$. При этом производится коррекция на число используемых параметров в каждой точке построения регрессии.

Выбор оптимальных параметров полосы пропускания может основываться на использовании информационного критерия Акаике (AIC).

В качестве оптимального значения параметра b берется решение задачи на минимум AIC:

$$AIC = 2n \ln \hat{\sigma} + n \ln(2\pi) + n \frac{n + \nu_1}{n - 2 - \nu_1} \to \min, \qquad (3)$$

где $\hat{\sigma}$ — оценка стандартного отклонения, $\nu_1 = tr(S)$.

Альтернативой является использование байесовского информационного критерия. Оптимальное значение ширины полосы пропускания определяется минимизацией величины

$$BIC = -2 \ln L + (p + 1) \ln n,$$

где L — значение функции правдоподобия; $p + 1$ — число оцениваемых коэффициентов.

Описанная процедура основана на предположении, что все коэффициенты модели меняются по территории. Отметим, что существует класс смешанных регрессионных моделей, позволяющих учитывать, что некоторые коэффициенты регрессии одинаковы во всей совокупности, а другие являются функциями координат. Такие модели являются обобщением географического подхода (Anselin, 1988; Fotheringham et al., 2002).

4. Эмпирические результаты

Метод географически взвешенной регрессии был применен для построения модели ценообразования на рынке недвижимости на примере стоимости однокомнатных квартир города Саратова.

Информационной базой послужили данные о продажах однокомнатных квартир на вторичном рынке жилья[1] за январь 2006 года. Численность выборки составила 1813 объектов.

На карте (рис. 1) четко прослеживаются направления «вытянутости» расположения города вдоль реки Волги и в перпендикулярном ей направлении. То есть город имеет достаточно сложную географическую структуру. Жирно выделены границы, внутри которых рассчитаны показатели географически взвешенной регрессии.

Рис. 1. Карта Саратова

Зависимой переменной является y — цена квартиры (тыс. руб.); а регрессорами: x_1 — жилая площадь ($м^2$), x_2 — площадь кухни ($м^2$), x_3 — дополнительная площадь ($м^2$), x_4 — логарифм расстояния от центра города (ln (км)), x_5 — расположение на первом этаже, x_6 — расположение на последнем этаже, x_7 — дом малой этажности, x_8 — пятиэтажка, x_9 — кирпичный дом, x_{10} — квартира в хорошем или отличном состоянии, x_{11} — наличие балкона или лоджии.

Для применения географически взвешенной регрессии к исходным данным были добавлены условные координаты объектов, полученные с помощью электронной базы данных «Все города России». Переменная x_4 включалась в стандартную регрессионную модель, а в географически взвешенном подходе не использовалась.

При построении весовой матрицы использовалась функция три-куб (2), в качестве критерия оптимизации ширины окна — критерий Акаике (3).

Географически взвешенный метод дал следующие результаты.

[1] http://www.ks.sarbc.ru/.

Оптимальное число ближайших соседей, дающее минимум критерия Акаике, равно 295. Коэффициент детерминации $R^2 = 0.8$.

Схематично зависимость критерия от числа точек регрессии (числа ближайших соседей) изображена на рис. 2.

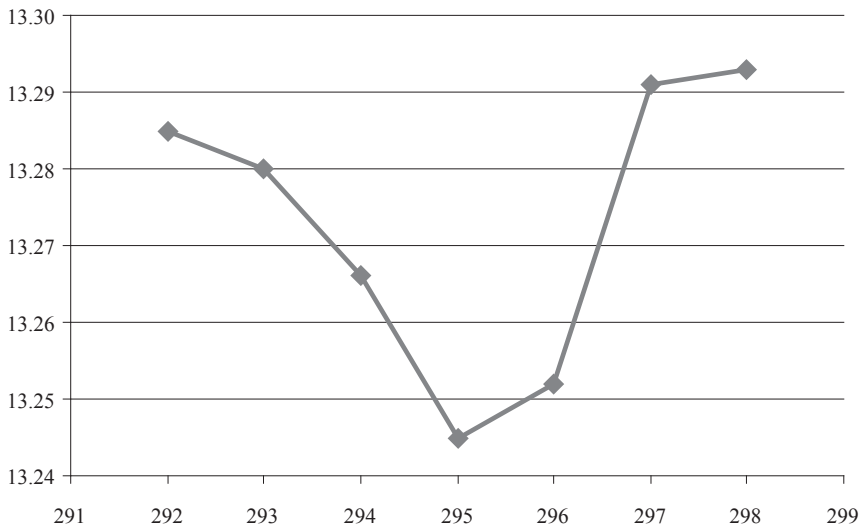


Рис. 2. Зависимость значения критерия CV от числа ближайших соседей

При определении этого параметра пришлось учитывать специфику фиктивных переменных и определять условный минимум, при котором матрица регрессоров является обратимой.

Проанализируем значения полученных оценок коэффициентов при каждом регрессоре. Для удобства анализа представим результаты в виде табл. 1 и 2, в которых значения оценок коэффициентов усреднены по целым значениям координат, а также представлены в виде диаграмм. Центр города располагается в квадрате $X = 61,\ Y = 32$.

В таблице 1 и на рисунке 3 представлены усредненные оценки коэффициента регрессии при переменной «жилая площадь».

В центральной части города выделяется квадрат (координаты $X = 60,\ Y = 32$) с самыми дорогими квартирами — около 30 тыс. рублей за квадратный метр. Вокруг него стоимость метра жилой площади превышает 20 тыс. руб. Четко выделяются окраины города, где цена квадратного метра жилой площади составляет около 10 тыс. рублей. Прослеживается дрейф убывающей цены от центра в направлении Ленинского и Заводского районов («левого верхнего» и «левого нижнего» углов табл. 1 и рис. 3). Линии на рис. 4 соответствуют уровням оценок коэффициентов при переменной «жилая площадь».

Оценки коэффициента при регрессоре «площадь кухни» представлены в табл. 2 и на рис. 4.

Анализ коэффициентов позволяет локализовать районы с наиболее высокой оценкой квадратного метра площади кухни. Наиболее высокая стоимость в квадрате $X = 59,\ Y = 31$ и прилегающих к нему зонах.

Таблица 1. Зависимость оценок коэффициента регрессии при переменной «жилая площадь» от координат *X, Y*

Y	*X*											
	53	54	55	56	57	58	59	60	61	62	63	64
36		9.1 (1.23)	8.6 (1.28)		12.2 (1.48)	11.2 (1.47)	8.9 (1.26)	8.6 (1.22)				
35		9.7 (1.24)	10.9 (1.25)	11.9 (1.29)	12.9 (1.51)	9.8 (1.39)			12.4 (1.55)			16.7 (3.29)
34				11.9 (1.29)	11.5 (1.37)	11.8 (1.35)	14.7 (1.48)		16.6 (2.24)	15.2 (2.40)	16.6 (3.25)	16.8 (3.51)
33				11.2 (1.17)		12 (1.37)	15.7 (1.75)	20.3 (2.68)	19.1 (2.95)		15.7 (4.34)	
32					14.1 (1.60)	16.6 (1.83)	23.2 (2.67)	28.4 (3.47)	21.4 (3.78)	14.1 (4.10)	14.1 (4.46)	
31				13 (1.87)	15.7 (1.85)	17.2 (2.11)	19.7 (3.05)	22.4 (3.40)	17.4 (3.71)	11.6 (4.06)	13.5 (4.21)	
30	10.3 (1.48)	12.9 (1.53)	14.2 (1.61)	13.6 (1.70)	14.7 (1.85)	16.4 (2.11)	18.6 (2.88)					
29	9.6 (1.51)	11 (1.51)	13.3 (1.60)	13.7 (1.62)	14 (1.77)	12.7 (2.06)						
28	9.8 (1.53)	10 (1.54)	10.3 (1.61)									
27		9.8 (1.54)										

Примечание. Все оценки коэффициентов значимы на 5%-ном уровне.

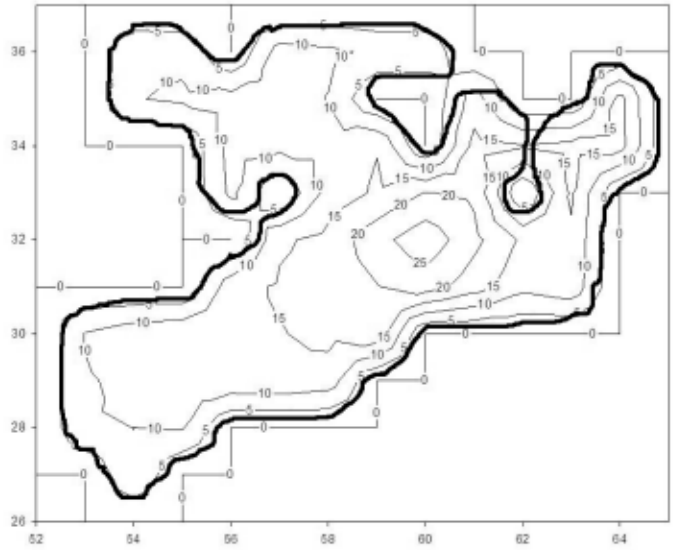
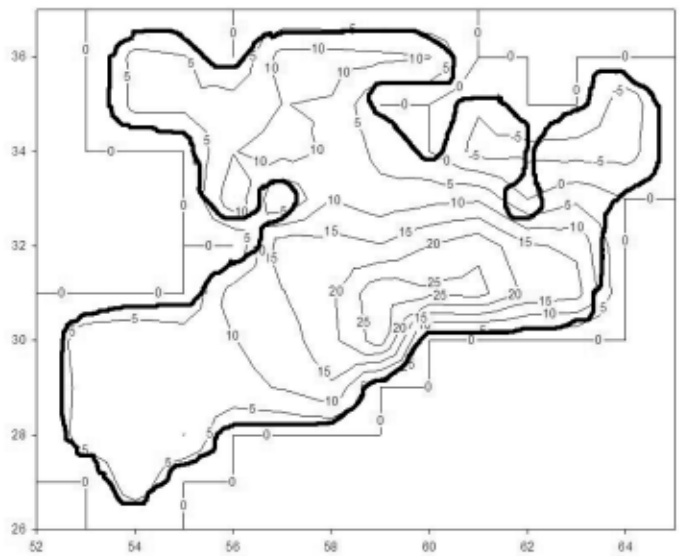


Рис. 3. Линии уровней значений оценок коэффициентов при переменной «жилая площадь»

Таблица 2. Зависимость оценок коэффициентов регрессии при переменной «площадь кухни» от координат *X, Y*

Y	*X*											
	53	54	55	56	57	58	59	60	61	62	63	64
36		6 (2.2)	5.2 (2.24)		12.2 (2.18)	12.1 (2.05)	11.2 (1.79)	10.2 (1.72)				
35		6 (2.20)	6 (2.22)	7 (2.28)	10.1 (2.33)	9.6 (2.03)			–4.4 (2.2)			–7.7 (3.04)
34				9.8 (2.23)	11.7 (2.38)	9.4 (2.15)	3.5 (2.20)		–6.6 (3.15)	–6.4 (2.66)	–7 (2.99)	–7 (3.19)
33				13.5 (2.18)		9.9 (2.22)	6.8 (2.67)	8.4 (3.26)	9.3 (3.87)		3.9 (4.22)	
32					18.2 (2.26)	17.3 (2.05)	15.4 (3.15)	20.1 (4.03)	23.7 (4.79)	15.7 (4.53)	14 (4.58)	
31				12.5 (2.22)	14.7 (2.19)	19.3 (2.29)	29.1 (3.04)	26.1 (3.55)	26.8 (4.23)	18.7 (4.24)	16.9 (4.36)	
30	7.8 (1.95)	9 (1.82)	7.8 (1.87)	9.8 (1.96)	12.3 (2.17)	18.5 (2.29)	27.7 (2.75)					
29	6.9 (2.04)	8.6 (1.92)	8.3 (1.88)	8.6 (1.89)	10.4 (2.04)	14.3 (2.32)						
28	7.4 (2.03)	8.1 (2.04)	10 (2.04)									
27		8.2 (2.02)										



Площадь кухни

Рис. 4. Линии уровней значений оценок коэффициентов при переменной «площадь кухни»

Отметим, что квадраты с наиболее дорогими площадями комнат и кухни не совпадают. На наш взгляд, это отражает специфику застройки разных районов города, которые не удается адекватно измерить и включить в состав регрессоров.

Еще одной особенностью этих центральных квадратов является то, что дополнительный метр кухни стоит дороже дополнительного метра жилой площади, так, для квадрата $X = 57,\ Y = 32$ эти стоимости равны 18.2 и 14.1 тыс. руб. соответственно. На окраинах города расположены зоны с относительно дешевыми кухнями. Кухонный метр по сравнению с жилым здесь стоит меньше или столько же.

Выделяется квадрат $X = 63,\ Y = 33$, в котором коэффициент регрессии отрицательный и незначимый. Это можно объяснить типичностью застройки данного района — для всех домов размеры кухни практически одинаковы и не являются определяющим параметрами в цене.

5. Заключение

Геокодированные данные существенно расширяют возможности экономического исследования пространственно распределенных явлений и процессов. Для моделирования цен на вторичном рынке жилья Саратова в работе был использован подход географически взвешенной регрессии. Переменные коэффициенты модели, плавно изменяющиеся по территории, позволяют в агрегированной форме отразить закономерности и локальные особенности ценообразования на вторичном рынке жилья, которые трудно воспроизвести стандартными методами.

Список литературы

Айвазян С. А., Мхитарян В. С. (2001). *Прикладная статистика и основы эконометрики*. М.: ЮНИТИ.

Магнус Я. Р., Катышев П. К., Пересецкий А. А. (2006). *Эконометрика. Начальный курс*. М.: Дело.

Anselin L. (1988). *Spatial econometrics: Methods and models*. Dordrecht: Kluwer Academic.

Anselin L. (2006). Spatial Econometrics. In: *Palgrave handbook of econometrics: Volume 1. Econometrics theory*, 901 – 941. Basingstoke: Palgrave Macmillan.

Bavaud F. (1998). Models for spatial weights: A systematic look. *Geographical Analysis*, 30, 153 – 171.

Cliff A., Ord J. K. (1973). *Spatial autocorrelation*. London: Pion.

Cliff A. D., Ord J. K. (1981). *Spatial processes: Models and applications*. London: Pion Limited.

Fotheringham A. S., Brunsdon C., Charrlton M. (2002). *Geographically weighted regression the analysis of spatially varying relationships*. New York: Wiley.

Haining R. (2004) *Spatial data analysis: Theory and practice.* Cambridge: Cambridge University Press.

LeSage J. P. (1999). *The theory and practice of spatial econometrics*. Department of Economics, University of Toledo.

LeSage J. P. (2001). Econometrics toolbox for MATLAB. http://www.spatial-econometrics.com/.

Lloyd C. D. (2007). *Local models for spatial analysis*. CRC Press.

Schabenberger O., Gotway C. A. (2005). *Statistical methods for spatial data analysis.* CRC Press.