УДК 330.43

О статистическом методе построения прогноза цены недвижимости по неоднородным данным

А. В. Харламов

Харламов Александр Владимирович, кандидат экономических наук, доцент кафедры основ математики и информатики, Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского, harlamovav@info.sgu.ru

Введение. В статье рассмотрены вопросы построения прогнозов на рынке недвижимости по неоднородным данным. Установление «справедливой» цены жилья является актуальной задачей при назначении залога, в целях страхования, определения эффективности инвестпроектов и т.д. Для решения этой задачи применяют эконометрические модели ценообразования, специфицированные по всей обследуемой совокупности. В случае значительной неоднородности обследуемой совокупности получаемые по этим моделям прогнозы могут содержать существенные ошибки. **Теоретический анализ**. На сегодняшний день существуют разнообразные методы и модели анализа неоднородных, в том числе пространственно распределенных данных. Для преодоления неоднородности исходных данных применяют зонирование обследуемой совокупности или строят модели переменной структуры, что сопряжено с рядом проблем. Дается обзор подходов, реализующих эти методы, перечисляются их плюсы и минусы. Для повышения качества прогноза предложен новый метод построения зон однородности на основе результатов построения оценок глобальной модели. Описан соответствующий алгоритм вычисления поправочного локального коэффициента, позволяющего корректировать прогноз глобальной модели. **Эмпирический анализ**. Для демонстрации эффективности работы предложенного метода по эмпирическим данным регионального рынка недвижимости рассчитаны прогнозы стоимости жилья, дан анализ результатов прогнозирования. **Результаты**. Предложенный новый метод определения зон однородности по результатам прогнозов с помощью расчета поправочного локального позволяет избежать ряда проблем, возникающих при использовании других подходов, и представляет эффективный инструмент прогнозирования.

Ключевые слова: пространственное моделирование, неоднородные данные, геокодированные данные, модели переменной структуры, зонирование.

DOI: https://doi.org/10.18500/1994-2540-2019-19-2-189-193

Введение

Как правило, многие статистические исследования социально-экономических явлений заканчиваются выводами о тенденциях в развитии этих явлений, т. е. представлением прогнозов. Причем достаточно часто прогнозирование является самостоятельной и порой единственной задачей подобных исследований. Например, анализ рынка жилой недвижимости преследует практически единственную цель – уточнение «справедливой рыночной» цены квартиры для продажи, для определения залога или для целей страхования. Для решения данной задачи успешно применяются эконометрические модели [1].

Построение и оценка классических эконометрических моделей, таких как модель множественной линейной регрессии, требуют соблюдения достаточно жестких исходных предпосылок, таких как гомоскедастичность, независимость регрессоров, нормальное распределение ошибок. Что, как правило, выполняется при анализе однородных данных.

Применение классической модели регрессии для анализа процессов и явлений на пространственно неоднородных территориях может неверно описывать реальную ситуацию. Например, стоимость объектов недвижимости может сильно отличаться в разных районах города.

Поэтому для анализа неоднородных данных используют специальные методы и строят модели, учитывающие специфику данных.

Теоретический анализ

Исходные статистические данные считаются однородными, если все они зарегистрированы при одних и тех же значениях сопутствующих переменных, в противном случае – неоднородными.

Учет территориальной неоднородности может рассматриваться в контексте общей проблемы построения регрессионных моделей по неоднородным данным [1, 2]. Можно отметить два подхода решения проблемы неоднородности. Это выделение тем или иным образом однородных зон или построение моделей переменной структуры, учитывающих имеющиеся неоднородности. Иногда эти подходы совмещаются.

При анализе неоднородной совокупности исходные данные можно разделить на однородные зоны и провести моделирование в каждой из них. При этом однородные совокупности малых объемов не позволят оценить модель, что является существенным недостатком данного подхода. Поэтому для анализа неоднородных

данных применяют регрессионные модели переменной структуры [2]. Разнообразие этих моделей достаточно велико, и выбор модели определяется конкретной ситуацией. Например, принадлежность выделенным зонам можно определять, вводя соответствующие бинарные фиктивные переменные, которые отражают эти различия [3, 4].

Разделение обследуемой территории на зоны связано с задачей выбора размеров и границ зон. Различные решения данной задачи могут оказывать значительное влияние на результаты моделирования. Различают четкие и нечеткие границы. Обычно четкие границы определяются принадлежностью к административному району. Прогнозная модель в этом случае может быть специфицирована следующим образом:

$$y_{i\in M} = \alpha_{0M} + \alpha_{1M} x_{1i\in M} + \alpha_{2M} x_{2i\in M} + \ldots + \alpha_{kM} x_{ki\in M} + \varepsilon_{i\in M},$$

где M – номер зоны. В результате получают отличающиеся между собой наборы оценок коэффициентов для каждой зоны. При эмпирическом определении зон произвольное определение границ может приводить к совершенно противоположным результатам в оценках коэффициентов и их ошибочной интерпретации [5]. Поэтому в процессе зонирования используют нечеткие границы. В этом случае оценивается модель вида

$$y_i = \alpha_0(u,v) + \alpha_1(u,v)x_{1i} + \alpha_2(u,v)x_{2i} + \ldots + \alpha_k(u,v)x_{ki} + \varepsilon_i.$$

Пара (u,v) определяет местоположение центра зоны. Границы нечеткие, и каждый объект имеет возможность попасть в любую зону с той или иной вероятностью. При этом возникает задача вычисления оптимального размера зон.

Изменения структуры модели могут носить различный характер. Например, значения коэффициентов могут меняться скачкообразно на границах построенных интервалов, как в случае кусочно-линейных или сплайн моделей [1]. Такие модели называют моделями с переключениями, при их построении считают, что момент скачка зависит от значений какой-либо управляющей переменной или происходит в определенный момент времени, и фактор времени является косвенной объясняющей переменной.

Таким образом, кусочно-линейная регрессия может быть представлена в виде системы линейных моделей:

$$y_t^l = \alpha_0^l + \alpha_1^l x_{1t}^l + \ldots + \alpha_n^l x_{nt}^l + \varepsilon_t^l,\ l = \overline{1,r},$$

оцениваемых на каждом сегменте. В случае сплайн-моделей в систему добавляют условия на границах сегментов. Изменения коэффициентов могут иметь и эволюционный характер, обусловленный постоянными переменами во внешней среде [2]. Модель с эволюционно меняющимися коэффициентами имеет вид

$$y_t = \alpha_0(t) + \alpha_1(t)x_{1t} + \ldots + \alpha_n(t)x_{nt} + \varepsilon_t,$$

где коэффициенты модели меняются во времени.

Построение оценок коэффициентов достаточно трудоемко и в общем случае практически невозможно, поэтому выдвигаются предположения о закономерностях их изменения. Например, предполагается, что они являются линейными функциями некоторого внешнего фактора [6]:

$$\alpha_i(t) = \beta_{i0} + \beta_{i1} z_t + u_{it},$$

где β_{i0}, β_{i1} – неизвестные коэффициенты; z_t – известные значения фактора в момент t; u_{it} – ошибки аппроксимации.

После подстановки коэффициентов получают линейную модель вида

$$y_t = \alpha_0 + \alpha_1 x_{1t} + \ldots + \alpha_n x_{nt} + \alpha_{n+1} x_{n+1,t} + \ldots \alpha_{2n+1} x_{2n+1,t} + \xi_t,$$

где $\alpha_0 = \beta_{00}$, $\alpha_1 = \beta_{10}$, …, $\alpha_n = \beta_{n0}$, $\alpha_{n+1} = \beta_{01}$, …, $\alpha_{2n+1} = \beta_{n1}$, $x_{n+1,t} = z_t$, $x_{n+2,t} = z_t x_{1t}$, …, $x_{2n+1,t} = z_t x_{nt}$, $\xi_t = \varepsilon_t + u_{0t} + x_{1t}u_{1t} + \ldots + x_{nt}u_{nt}$.

В зависимости от свойств ошибок применяют тот или иной метод построения оценок.

Спецификация моделей переменной структуры в пространстве еще больше усугубляет описанные проблемы, так как изменению подвержены уже два параметра – координаты местоположения объектов.

Эти проблемы решаются при использовании метода географически взвешенной регрессии [3, 4], когда применяются методы оптимизации по величине зоны и числу объектов в ней.

Модель географически взвешенной регрессии имеет вид

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) \cdot x_{ik} + \varepsilon_i,$$

где (u_i, v_i) – координаты точки i, $i = \overline{1,n}$ и неизвестные коэффициенты зависят от координат. Оценки коэффициентов модели вычисляются в каждой точке, в которой проводились измерения, и также являются функциями координат. В целях выявления местных особенностей используются не все данные, а только «ближайшие соседи». Предполагается, что регрессионные модели для соседних объектов схожи, но могут варьироваться по территории, степень близости объектов учитывается с помощью весов. Матрица весовых коэффициентов вычисляется для местоположения каждого объекта.

Классическую (глобальную) модель можно рассматривать как частный случай географического подхода для ситуации, когда оценки

коэффициентов не меняются при изменении местоположения и остаются постоянными на всей территории.

Вычислительные процедуры при данном подходе достаточно громоздки, теряется наглядность и возможна ситуация, когда оптимальные размеры «скользящей зоны» практически соизмеримы с обследуемой областью, что полностью сглаживает эффект применения этого метода.

Рассмотрим другой подход решения перечисленных проблем при прогнозировании по неоднородным данным [7]. Предлагается использовать классическую модель множественной линейной регрессии, специфицированную по всей обследуемой совокупности, а однородные зоны определять относительно точки (объекта) прогноза. Такой подход позволит избежать многих сформулированных выше проблем, в частности проблемы спецификации модели по малой выборке.

Единая (глобальная) модель, специфицированная по всей обследуемой совокупности, усредняет прогнозируемые значения, сглаживая специфические особенности местности, как, впрочем, и все модели переменной структуры. Такие модели не могут определять экстремальные территориальные особенности, существенно влияющие на цену объекта. Предлагаемый подход лишен этого недостатка и позволяет выявлять подобные ситуации.

Приведем следующий алгоритм построения прогноза с помощью выделения зон однородности и определения на их основе локальных коэффициентов ($K_{\text{лок}}$) коррекции прогноза.

1. Строится классическая модель множественной линейной регрессии по всей обследуемой территории. При этом используются геокодированные данные.

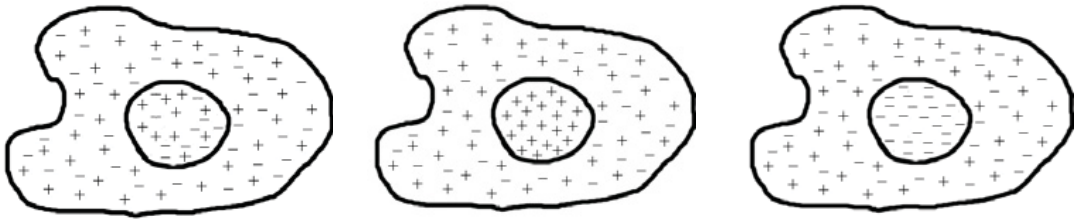
2. Выбирается объект прогнозирования и определяется его местоположение на обследуемой территории.

3. Вокруг объекта определяются «ближайшие соседи», участвовавшие в построении модели, и анализируются модельные «остатки» этих соседних объектов. Число ближайших соседей можно определять из условия пространственной однородности анализируемых остатков.

4. Возможны несколько вариантов значений остатков в выделенной зоне (рисунок): а) остатки внутри зоны имеют однородный характер, как и по всей территории, и вычисление локального коэффициента не повысит точности прогноза; б) остатки внутри зоны имеют преимущественно положительные значения, и необходимо применять повышающий локальный коэффициент; в) остатки внутри зоны имеют преимущественно отрицательное значение, и необходимо применять понижающий локальный коэффициент.

5. Прогнозное значение, вычисленное по глобальной модели, умножается на локальный коэффициент.

Значение локального коэффициента можно рассчитывать, например, как отношение средней, возможно, взвешенной по «близости», стоимости «соседей» к среднему прогнозу, полученному по модели для этих «соседей».



а *б* *в*

Варианты однородности остатков в зоне

Variants of Homogeneity of Residues in the Zone

Эмпирический анализ

Апробация предложенного метода прогнозирования была проведена на рынке однокомнатных квартир г. Саратова (исходные данные являются выборкой с сайта https://kvadrat64.ru/, полученной в январе 2019 г.). Глобальная модель множественной линейной регрессии имеет вид

$$y = 2928 + 41,3x_1 + 64,1x_2 + 21,4x_3 - 81,2x_4 - 72,1x_5 - 463,1x_6 - 102,2x_7 +$$
$$+42,3x_8 + 175,4x_9 + 56,1x_{10} - 48,3x_{11} - 99,1x_{12} + 176,2x_{13} - 298,3x_{14}$$

Здесь y – цена квартиры, тыс. руб.; x_1 – жилая площадь, м^2; x_2 – площадь кухни, м^2; x_3 – дополнительная площадь, м^2; x_4 – квартира расположена на первом этаже; x_5 – квартира расположена на последнем этаже; x_6 – дом малой этажности; x_7 – пятиэтажный дом; x_8 – дом выше девяти этажей; x_9 – кирпичный дом; x_{10} – монолитный дом; x_{11} – раздельный санузел; x_{12} – планировка гостиничного типа; x_{13} – планировка студия; x_{14} – логарифм расстояния до центра города, ln(м). Коэффициент детерминации $R^2 = 0{,}72$, все факторы статистически значимы на уровне 5%.

Для спецификации модели использовано 95% исходных данных. Из оставшихся 5% данных был выбран объект для расчета прогнозной цены. Прогноз, рассчитанный по модели, равен 1 867 534 руб. Визуально были выделены 18 «ближайших соседей» с положительными остатками (случай б). Значение локального коэффициента равно $K_{лок}$=1,118. Исправленное прогнозное значение равно 2 087 903 руб., что всего на 2,9% ниже заявленной в объявлении стоимости квартиры, равной 2 150 000 руб.

Результаты

Предложенный метод прогнозирования цены жилой недвижимости по неоднородным исходным данным позволяет избежать ряда существенных проблем, возникающих при использовании известных подходов, связанных с зонированием обследуемой территории и построением моделей переменной структуры.

Так как модель оценивается по совокупности данных, то учитываются все ценообразующие факторы, потеря значимости которых возможна при локализации обследуемой территории.

Анализ остатков глобальной модели в зоне положения объекта прогнозирования позволяет учитывать специфические особенности влияния территории на цену недвижимости, что, как правило, сглаживается в других случаях.

Апробация метода по эмпирическим данным позволила получить удовлетворительный результат по прогнозу.

Хотя еще некоторые вопросы не нашли окончательного ответа, например, критерий оптимального размера зоны однородности и число «ближайших соседей» или способ расчета локального коэффициента – по стоимости объектов или стоимости квадратного метра, тем не менее, данный метод можно считать некоторой техникой оценивания объектов жилой недвижимости, достаточно близкой к построению оценок по «аналогам», используемым в оценочной практике, но основанным на массовой оценке.

Список литературы

1. Эконометрика : учебник / под ред. В. С. Мхитаряна. М. : Проспект, 2009. 384 с.
2. *Айвазян С. А., Мхитарян В. С.* Прикладная статистика и основы эконометрики. М. : ЮНИТИ, 1998. 1000 с.
3. *Fotheringham A. S., Brunsdon C., Charlton M.* Geographically Weighted Regression the Analysis of Spatially Varying Relationships. University of Newcastle, UK : John Wiley & Sons Ltd, 2002. 284 p.
4. *Балаш О. С., Харламов А. В.* Эконометрическое моделирование пространственных данных. Саратов : Научная книга, 2010. 112 с.
5. *Appleton D. R., French J. M., Vanderpump M. P.* Ignoring a Covariate : An Example of Simpson's Paradox // The American Statistician. 1996. Vol. 50, iss. 4. P. 340–341.
6. *Anselin L.* Spatial Externalities, Spatial Multipliers and Spatial Econometrics // International Regional Science Review. 2003. Vol. 26, iss. 2. P. 153–166.
7. *Харламов А. В.* Применение пространственного коэффициента при прогнозировании по геокодированным данным // Математическое и компьютерное моделирование в экономике, страховании и управлении рисками : материалы VII Междунар. молод. науч.-практ. конф. Саратов : Научная книга, 2018. С. 157–160.

Образец для цитирования:

Харламов А. В. О статистическом методе построения прогноза цены недвижимости по неоднородным данным // Изв. Сарат. ун-та. Нов. сер. Сер. Экономика. Управление. Право. 2019. Т. 19, вып. 2. С. 189–193. DOI: https://doi.org/10.18500/1994-2540-2019-19-2-189-193

The Statistical Method of Constructing a Forecast for the Real Estate Price Using Heterogeneous Data

A. V. Harlamov

Alexander V. Harlamov, https://orcid.org/0000-0002-1709-6518, Saratov State University, 83 Astrakhanskaya St., Saratov 410012, Russia, harlamovav@info.sgu.ru

Introduction. The article deals with the issues of constructing forecasts in the real estate market using heterogeneous data. Establishing a "fair" price of housing is an urgent task for collateral assigning, for insurance purposes, for determining the investment projects effectiveness, etc. To solve this problem, econometric pricing models are used, which are specified for the entire surveyed population. In case of significant heterogeneity of the surveyed population, the predictions obtained from these

models may contain significant errors. **Theoretical analysis.** Now, there is a variety of methods and models for the analysis of heterogeneous, spatially distributed data. Population zoning or a variable structure model is used to overcome data heterogeneity. It is connected with a number of problems. An overview of the approaches that implement these methods is given, their advantages and disadvantages are listed. A new method for constructing homogeneity zones, based on the results of the global model estimates building, is proposed to improve the forecast quality. The corresponding algorithm for calculating the local correction factor is described, which makes it possible to correct the global model forecast. **Empirical analysis.** To demonstrate the effectiveness of the proposed method in action, a forecast for the real estate price, based on the empirical data of the regional real estate market, was calculated, and an analysis of the forecast results was given. **Results.** The proposed new method for determining homogeneity zones based on the results of forecasts using the local correction calculation makes it possible to avoid a number of problems arising from the use of other approaches and represents an effective forecasting tool.

Keywords: spatial modeling, heterogeneous data, geocoded data, variable structure models, zoning.

References:

1. *Ekonometrika* [Econometrics. Ed. by V. S. Mkhitarian]. Moscow, Prospekt Publ., 2009. 384 p. (in Russian).
2. Aivazian S. A., Mkhitarian V. S. *Prikladnaia statistika i osnovy ekonometriki* [Applied Statistics and the Basics of Econometrics]. Moscow, IuNITI Publ., 1998. 1000 p. (in Russian).
3. Fotheringham A. S., Brunsdon C., Charlton M. *Geographically Weighted Regression the Analysis of Spatially Varying Relationships*. University of Newcastle, UK, John Wiley & Sons Ltd, 2002. 284 p.
4. Balash O. S., Kharlamov A. V. *Ekonometricheskoe modelirovanie prostranstvennykh dannykh* [Econometric Modeling of Spatial Data]. Saratov, Nauchnaia kniga Publ., 2010. 112 p. (in Russian).
5. Appleton D. R., French J. M., Vanderpump M. P. Ignoring a Covariate: An Example of Simpson's Paradox. *The American Statistician,* 1996, vol. 50, iss. 4, pp. 340–341.
6. Anselin L. Spatial Externalities, Spatial Multipliers and Spatial Econometrics. *International Regional Science Review,* 2003, vol. 26, iss. 2, pp. 153–166
7. Kharlamov A. V. Primenenie prostranstvennogo koeffitsienta pri prognozirovanii po geokodirovannym dannym [Applying of Spatial Coefficient in Geocoded Data Forecast]. In: *Matematicheskoe i komp'iuternoe modelirovanie v ekonomike, strakhovanii i upravlenii riskami: materialy VII Mezhdunar. molod. nauch.-prakt. konf.* [Mathematical and Computer Modeling in Economics, Insurance and Risk Management: Proceedings of the VII Intern. young sci.-pract. conf.]. Saratov, Nauchnaia kniga Publ., 2018, pp.157–160 (in Russian).

Cite this article as:

Harlamov A. V. The Statistical Method of Constructing a Forecast for the Real Estate Price Using Heterogeneous Data. *Izv. Saratov Univ. (N. S.), Ser. Economics. Management. Law,* 2019, vol. 19, iss. 2, pp. 189–193 (in Russian). DOI: https://doi.org/10.18500/1994-2540-2019-19-2-189-193