

Филиал Кемеровского государственного университета
в г. Анжеро-Судженске
Национальный исследовательский
Томский государственный университет
Кемеровский государственный университет
Институт проблем управления им. В. А. Трапезникова РАН
Институт вычислительных технологий СО РАН

**ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
И МАТЕМАТИЧЕСКОЕ
МОДЕЛИРОВАНИЕ
(ИТММ–2015)**

**Материалы XIV Международной конференции
имени А. Ф. Терпугова
18–22 ноября 2015 г.
Часть 2**

Издательство Томского университета

2015

3. Ukkonen E. Constructing Suffix-trees On-Line in Linear Time // Algorithms, Software, Architecture: Information Processing. 1992. № 1(92). P. 484–492.
4. Kärkkäinen J., Sanders P. Simple linear work suffix array construction // Baeten et al. (Eds.) // J.C.M.: ICALP 2003, LNCS 2719. 2003. P. 943–955.
5. Хмелёв Д. В. Классификация и разметка текстов с использованием методов сжатия данных. Краткое введение. 2003 [Электронный ресурс] // URL: <http://compression.graphicon.ru/download/articles/classif/intro.html>
6. Khmelev D. V., Teahan W. J. Verification of text collections for text categorization and natural language processing // Technical Report AIIA 03.1. School of Informatics, University of Wales. Bangor, 2003.
7. Ашуров М. Ф., Поддубный В. В. Метод классификации текстов художественной литературы на основе R-меры // Новые информационные технологии в исследовании сложных структур: матер. десятой Рос. конф. с междунар. участием. Томск: Изд. Дом Том. гос. ун-та, 2014. С. 63–64.
8. Ашуров М. Ф. Сравнение потоковых методов классификации текстов художественной литературы на основе сжатия информации и подсчета подстрок // Вестник Том. гос. ун-та. Управление, вычислит. техника и информатика. 2014. № 4(29). С. 16–22.
9. Ашуров М. Ф., Поддубный В. В. Поточный метод классификации текстов художественной литературы на основе C-меры // Информационные технологии и математическое моделирование (ИТММ-2013): матер. XII Всерос. науч.-практ. конф. с междунар. участием им. А. Ф. Терпугова (29–30 ноября 2013 г.). Томск: Изд-во Том. ун-та, 2013. Ч. 2. С. 85–89.
10. Shevelyov O. G., Poddubnyj V. V. Complex investigation of texts with the system «StyleAnalyzer» / Ed by P. Grzyber, E. Kelih, J. Macutek // Text and Language. Wien: Praesens Verlag, 2010. P. 207–212.
11. Van Rijsbergen C. J. Information Retrieval. London: Butterworths, 1979.

ИЗУЧЕНИЕ СТЕПЕНИ ВЛИЯНИЯ ИНФОРМАЦИИ О ГЕОГРАФИЧЕСКОМ РАСПОЛОЖЕНИИ ОБЪЕКТА НЕДВИЖИМОСТИ НА ТОЧНОСТЬ ОЦЕНКИ ЕГО РЫНОЧНОЙ СТОИМОСТИ

А. Л. Богданов

*Национальный исследовательский
Томский государственный университет, Томск, Россия*

Целью исследования является изучение степени влияния информации о географическом расположении объекта недвижимости на точность оценки его рыночной стоимости. Объектом исследования является рынок жилой недвижимости г. Томска, а именно двухкомнатные квартиры, расположенные в кирпичных и панельных домах. Так как исходные данные о ценах на недвижимость, использованные при проведении исследования, получены из открытых источников и не являются случайной выборкой, то полученные результаты, строго говоря, не могут быть обобщены на рынок недвижимости г. Томска, поэтому можно считать, что данное исследование носит демонстрационный характер.

Основными источниками данных, на основе которых проводилось исследование, являются 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, материал, этаж, этажность дома, район города, адрес (улица, номер дома). Второй сайт использовался для определения точных географических координат (широта и долгота) домов, в которых располагаются квартиры. На момент проведения исследования (апрель 2015 г.) на сайте ru09 было зарегистрировано 14 212 объявлений, из которых 5 072 касались продажи двухкомнатных квартир. Удовлетворяющими критериям отбора в выборку (дом: кирпичный или панельный, расположение: в черте города) после удаления дубликатов оказались 1 656 объявлений.

В качестве базовой модели формирования цены квартиры, с которой в дальнейшем будут сравниваться другие модели, была выбрана модель

$$y = \beta_0 + \beta_{sq}x_{sq} + \beta_{fl}x_{fl} + \beta_{br}x_{br} + \varepsilon, \quad (1)$$

где y – это цена квартиры; x_{sq} – площадь квартиры (кв. м); x_{fl} – фиктивная переменная, равная 1, если квартира расположена на первом или последнем этаже, и равная 0 в противном случае; x_{br} – фиктивная переменная, равная 1, если квартира расположена в кирпичном доме, и равная 0, если в панельном; ε – случайная составляющая. Базовая модель соответствует гипотезе о том, что на цену квартиры влияют её площадь, этаж, на котором располагается квартира (квартиры, расположенные на первом или последнем этажах, как правило, не пользуются спросом), и тип дома (кирпичный или панельный).

Для описания местоположения дома, в котором расположена квартира, использовались следующие способы:

- 1) указание района города;
- 2) указание расстояния от дома до центра города;
- 3) указание координат (широта и долгота) дома.

В соответствии с перечисленными способами рассматривались следующие конкурирующие модели:

$$y = \beta_0 + \beta_{sq}x_{sq} + \beta_{fl}x_{fl} + \beta_{br}x_{br} + \beta_{do}x_{do} + \beta_{dk}x_{dk} + \beta_{ds}x_{ds} + \varepsilon, \quad (2)$$

$$y = \beta_0 + \beta_{sq}x_{sq} + \beta_{fl}x_{fl} + \beta_{br}x_{br} + \beta_{ce}x_{ce} + \varepsilon, \quad (3)$$

где x_{do} , x_{dk} , x_{ds} – фиктивные переменные, равные 1, если квартира расположена в Октябрьском, Кировском или Советском районах г. Томска соответственно, и одновременно равные 0 в случае расположения квартиры в Ленинском районе; x_{ce} – расстояние от квартиры до центра города (в качестве центра бала выбрана пл. Ленина). Также рассматривался подход на основе оценки параметров базовой модели (1) по данным K ближайших соседей (KNN – K-nearest neighborhood).

В качестве метода оценки неизвестных параметров моделей использовался метод наименьших квадратов. При оценке параметров базовой модели на основе данных о K ближайших соседях состав регрессоров подвергался корректировке при необходимости (из модели исключались переменные, вариация которых была равна нулю, так как в этом случае матрица регрессоров оказывается вырожденной и система нормальных уравнений не имеет единственного решения).

В качестве основного показателя, используемого для оценки качества модели и сравнения моделей, был выбран средний квадрат ошибки

$$MSE = \frac{1}{n} \sum_{i=1}^n e_i^2, \quad (4)$$

где $e_i = y_i - \hat{y}_i$ – ошибка прогноза в i -м наблюдении ($y_i, \hat{y}_i$ – истинное и предсказанное значения цены квартиры соответственно), n – общее количество наблюдений.

Оценка качества моделей проводилась по методу *tqCrossValidation* [1] со следующими значениями параметров: $t = 100, q = 4$. Суть метода заключается в следующем. Исходная выборка разбивается случайным образом на q равных частей. Каждая из частей по очереди объявляется *тестовым множеством*, а данные, не попавшие в тестовое множество, – *обучающим множеством*. Данные обучающего множества используются для оценки неизвестных параметров модели, а данные тестового – для оценки её качества. В результате получается q значений показателя качества, которые после усреднения дают «объективную» оценку обобщающей способности модели, зависящую от разбиения исходного множества на тестовое и обучающее множества. Для устранения этой зависимости *эксперимент* (процедуру разбиения/оценки) повторяют t раз, в результате получают t усреднённых значений показателя качества, после усреднения которых получается итоговое значение оценки обобщающей способности модели, не зависящее от разбиения исходного набора данных на тестовое и обучающие множества.

Поиск оптимального (в смысле минимума показателя MSE) количества ближайших соседей – параметра K – выполнялся в интервале от 10 до 200 с шагом 5 методом *tqCrossValidation* с параметрами $t = 5, q = 4$. Зависимость усредненного показателя MSE при различных значениях параметра K показана на рис. 1. Распределение оптимальных значений параметра K в различных экспериментах имеет бимодальный вид (рис. 2).

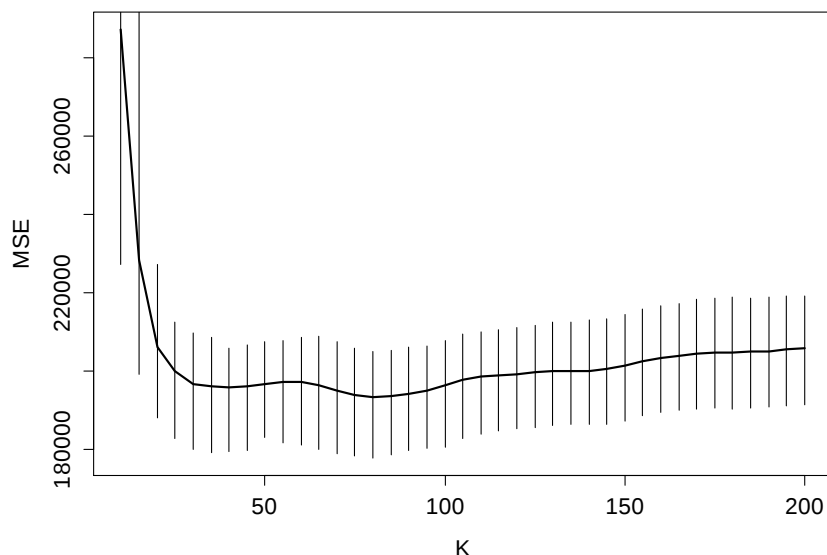


Рис. 1. Зависимость показателя MSE при различных значениях параметра K (вертикальные линии показывают разброс значений показателя MSE при различных разбиениях на тестовое и обучающее множества)

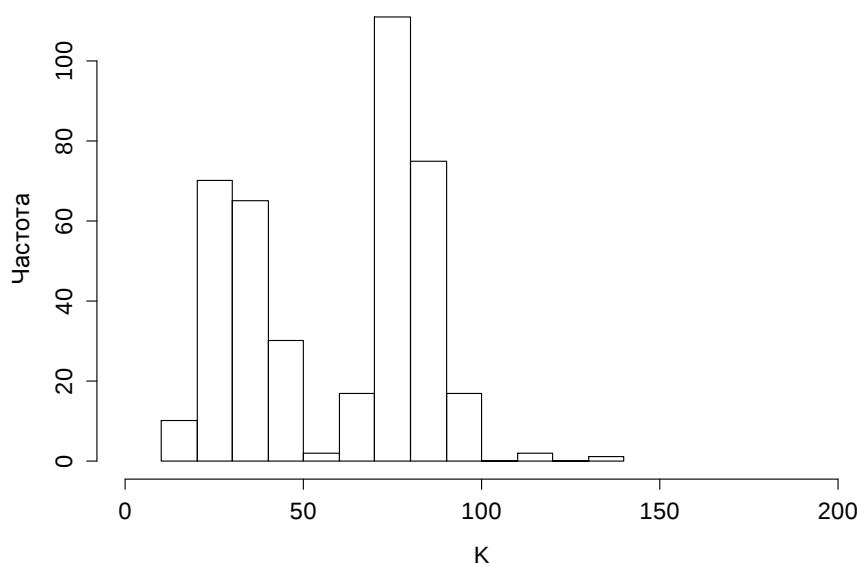


Рис. 2. Распределение оптимальных значений параметра K

Результаты оценки параметров моделей и значения показателя качества представлены в табл. 1 и 2 соответственно. Для каждого значения приведено его среднее значение и стандартное отклонение в скобках, которые получены усреднением результатов по множеству экспериментов. На рис. 3 приведены гистограммы распределения значений показателя MSE для каждой из моделей. На рис. 4 показано распределение ошибок прогноза по каждой из моделей на карте города.

Таблица 1

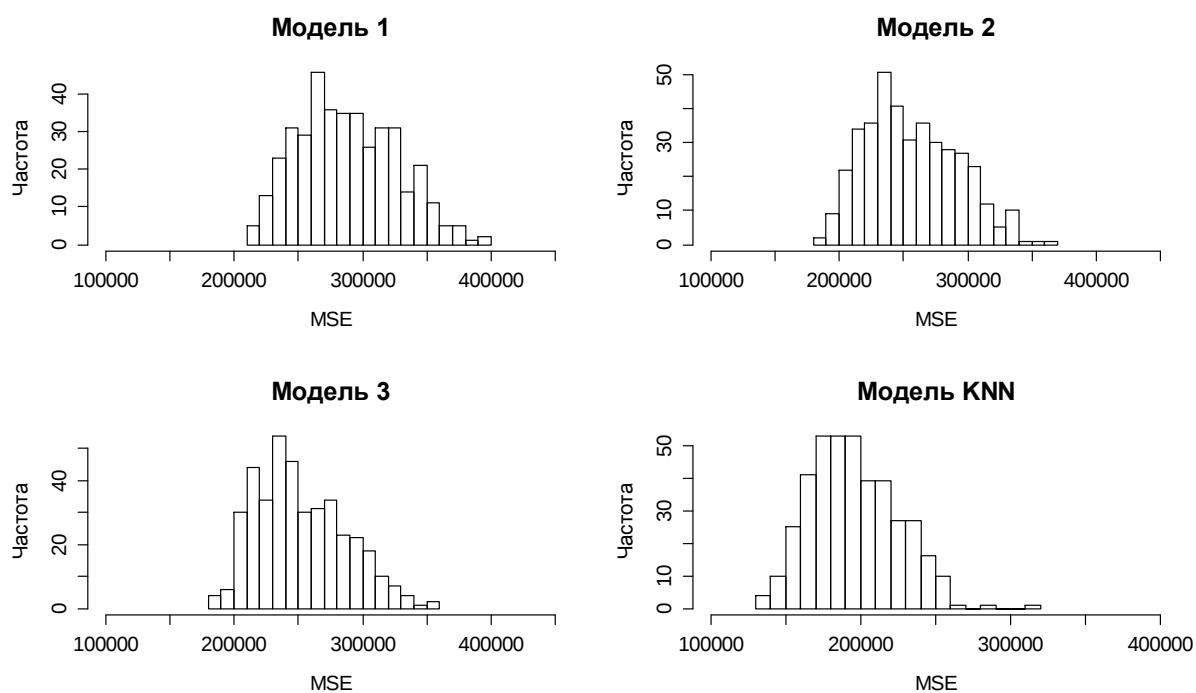
Средние значения оценок коэффициентов моделей 1–4

	$\hat{\beta}_{sq}$	$\hat{\beta}_{fl}$	$\hat{\beta}_{br}$	$\hat{\beta}_{do}$	$\hat{\beta}_{dk}$	$\hat{\beta}_{ds}$	$\hat{\beta}_{ce}$
Модель 1	69.6 (1.12)	–158 (13.7)	361.34 (14.0)				
Модель 2	70.11 (1.06)	–168.42 (12.76)	255.02 (13.63)	–185.18 (17.34)	220.41 (23.86)	252.82 (21.69)	
Модель 3	67.55 (1.05)	–160.77 (13.03)	242.65 (13.72)				–111.6 (4.67)
Модель 4 (KNN)	63.57 (2.09)	–123.68 (16.64)	199.6 (26.08)				

Таблица 2

Средние значения показателя качества моделей 1–4

	MSE	ΔMSE	$\Delta MSE(\%)$
Модель 1	288117.1 (38389.75)		
Модель 2	256285.8 (35959.66)	–31831.32 (8139.03)	–11.1 (2.6)
Модель 3	251666.7 (34638.11)	–36450.43 (9629.86)	–12.6 (2.8)
Модель 4 (KNN)	195320.7 (28691.52)	–92796.43 (21480.08)	–32.1 (5.6)

Рис. 3. Гистограммы распределения показателя MSE для каждой из моделей

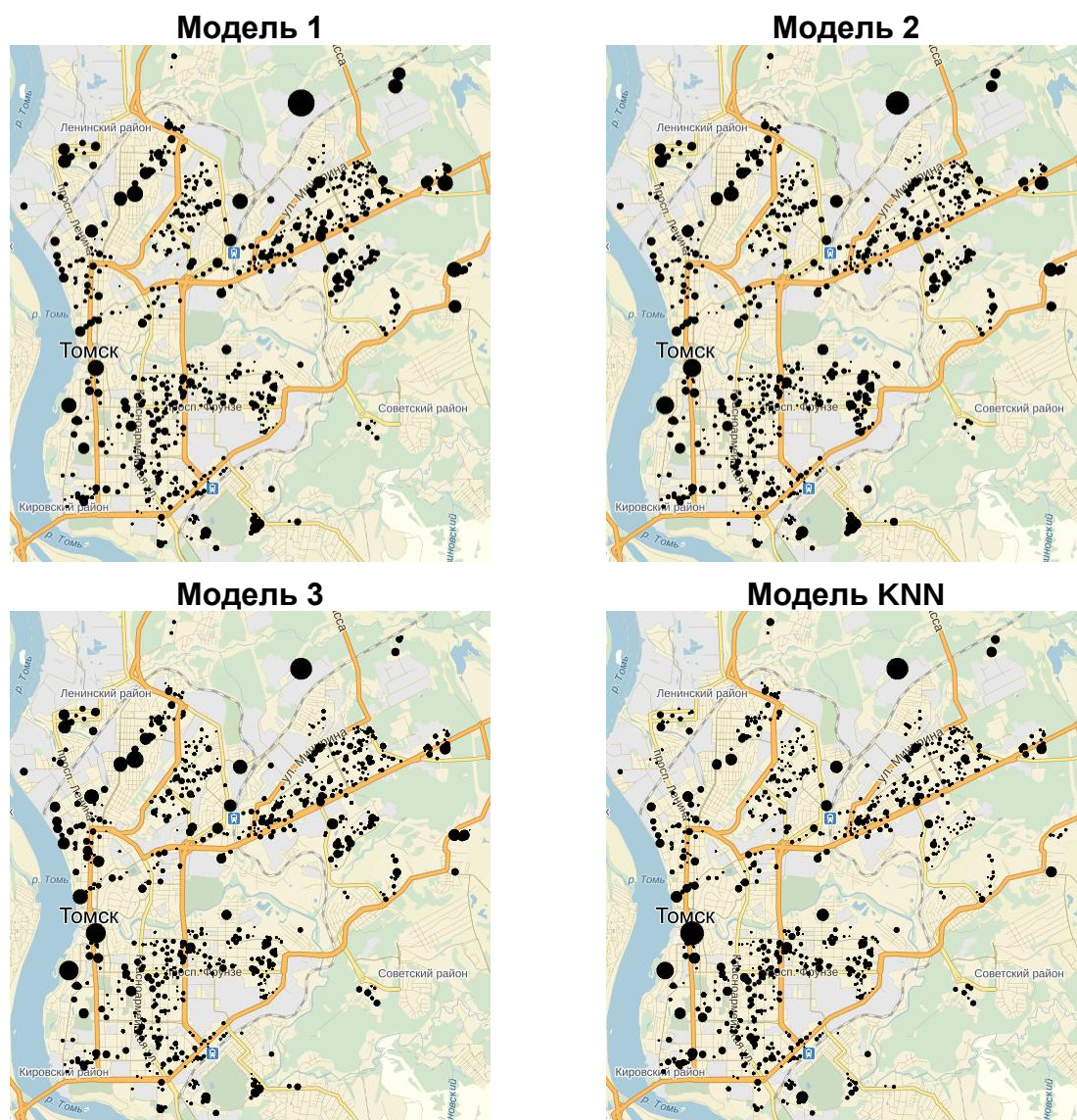


Рис. 4. Диаграммы распределения средней ошибки прогноза на карте г. Томска (радиус точки пропорционален размеру ошибки)

Как видно из табл. 2, добавление в базовую модель информации о районе расположения квартиры и информации о расстоянии до центра города оказывает примерно одинаковое влияние на точность прогноза: уменьшение показателя MSE составило 11.1 и 12.6 % соответственно. Модель, построенная по методу K ближайших соседей, оказалась наиболее точной: уменьшение показателя MSE по сравнению с базовой моделью составило 32.1 %.

Гистограммы на рис. 3 позволяют оценить форму распределения и диапазон изменений показателя MSE для каждой из моделей. Диаграммы распределения ошибок на рис. 4 позволяют определить, какие квартиры оказались наиболее «сложными» в смысле точности прогноза, например, из диаграмм видно, что северная часть города оказалась наиболее сложным участком для моделей 1–3.

Литература

1. Devijver, Pierre A.; Kittler, Josef. Pattern Recognition: A Statistical Approach. London, GB: Prentice-Hall, 1982.

МОДЕЛЬ ИДЕНТИФИКАЦИИ ЗВУКОВОГО СИГНАЛА С ПОЗИЦИЙ ТЕОРИИ АКТИВНОГО ВОСПРИЯТИЯ

В. Е. Гай

*Нижегородский государственный технический
университет им. Р. Е. Алексеева, Нижний Новгород, Россия*

1. Обзор существующих систем и алгоритмов идентификации звуковых сигналов

1.1. Описание задачи идентификации

Количество музыкальных композиций, хранящихся в настоящее время в сети Интернет, велико. Например, сервис Яндекс.Музыка хранит около пяти миллионов записей, сервис Shazam – пять миллиардов. Очевидно, что в такой ситуации актуальна задача быстрого и точного поиска по имеющимся музыкальным записям. Причём по мере увеличения числа звуковых записей значимость возможности поиска будет возрастать.

Существующие алгоритмы поиска звуковой информации можно разделить на два направления:

- 1) поиск по содержанию (Content-Based Sound Retrieval, CBSR);
- 2) поиск по текстовым аннотациям (тегам, Description-Based Sound Retrieval, DBSR).

В отличие от систем DBSR, системы поиска звуковых файлов по содержанию не требуют наличия какой-либо дополнительной информации о файле. Поиск в таких системах производится на основе анализа и сравнения характеристик звуковых файлов. Настоящая работа посвящена применению теории активного восприятия (ТАВ) к решению задачи идентификации звукового сигнала [1].

1.2. Обзор алгоритмов идентификации

Простейшее решение задачи идентификации звукового сигнала по его фрагменту состоит в прямом сравнении амплитуд искомого и исходного звуковых сигналов в определенный момент времени. Недостатки такого алгоритма заключаются в низкой устойчивости к искажению искомого сигнала, а также в большом времени его работы.

Установлено, что для построения помехоустойчивой системы идентификации звукового сигнала необходимо создать описание сигнала (цифровой отпечаток сигнала). Цифровой отпечаток в идеале должен быть устойчив к различным искажениям сигнала.

Рассмотрим некоторые известные алгоритмы генерации цифровых отпечатков.