Научная статья

Original article

УДК 528:004.9

DOI 10.55186/25876740_2023_7_4_18

# ФОРМИРОВАНИЕ БАЗЫ ПРОСТРАНСТВЕННЫХ ДАННЫХ ДЛЯ ОБУЧЕНИЯ ПРЕДСКАЗАТЕЛЬНОЙ МОДЕЛИ АНАЛИЗА ИНВЕСТИЦИОННОЙ ПРИВЛЕКАТЕЛЬНОСТИ ЗЕМЕЛЬНЫХ УЧАСТКОВ

# FORMATION OF A SPATIAL DATA BASE FOR TRAINING A PREDICTIVE MODEL FOR ANALYZING THE INVESTMENT ATTRACTIVENESS OF LAND PLOTS

**Курлов Алексей Викторович**, заведующий научно-исследовательской лабораторией «Лаборатория городских технологий и пространственного развития» Московского государственного университета геодезии и картографии, (105064, Москва, Гороховский пер., 4), тел. +79215665459, Scopus ID: 57190620515, Researcher ID: CAF-6636-2022, ORCID: 0000-0003-3089-7288, kurlov-av@yandex.ru

**Alexey Viktorovich Kurlov**, head of the research laboratory "Laboratory of Urban Technologies and Spatial Development" of the Moscow State University of Geodesy and Cartography, (4 Gorokhovsky Lane, Moscow, 105064), tel. +79215665459, Scopus ID: 57190620515, Researcher ID: CAF-6636-2022, ORCID: 0000-0003-3089-7288, kurlov-av@yandex.ru

**Аннотация.** Задача предсказания инвестиционной привлекательности земельных участков не только является актуальной с экономической и финансовой точки зрения, но и в целом позволяет оценивать развитие населенных пунктов и их районов. В данной статье представлен опыт создания базы пространственных данных для обучения предсказательной модели анализа инвестиционной привлекательности земельных участков, а также процесс фильтрации аномальных значений. Данный подход стал многоэтапным процессом, который включал в себя сбор данных, первичный анализ и очистку данных. Благодаря этому была создана высококачественная база данных, которая может повысить точность и эффективность предсказательных моделей в задачах анализа инвестиционной привлекательности земельных участков. Кроме того, данный опыт позволяет брать его за основу для создания аналогичных баз данных в других областях, требующих высококачественных данных для обучения.

**Abstract.** The research problem of predicting the investment attractiveness of land plots is not only relevant from an economic and financial point of view, but in general allows assessing the development of settlements and their districts. This article presents the experience of creating a spatial database for training a predictive model for analyzing the investment attractiveness of land plots, as well as the process of filtering anomalous values. This approach became a multi-step process that included data collection, primary analysis, and data cleansing. Through this process, a high-quality database has been created that can improve the accuracy and efficiency of predictive models in the tasks of analyzing the investment attractiveness of land plots. In addition, this experience allows using it as a basis for creating similar databases in other areas that require high-quality data for training.

**Ключевые слова:** *инвестиционная привлекательность, база данных, кластеризация данных, пространственный фактор, кадастровая стоимость.*

**Keywords:** *investment attractiveness, database, data clustering, spatial factor, cadastral value.*

**Введение**

В ряде работ было выявлено, что на инвестиционную стоимость земельных участков в значимой степени влияет пространственный фактор [1-7].

Под пространственным фактором понимается ряд частных критериев, обязательных для учета при оценке инвестиционной стоимости земельных участков:

– транспортная доступность (подъезды к трассам, остановки общественного транспорт);

– административная принадлежность (развитость региона, локальные социально-экономические условия, местное законодательство);

– удаленность от объектов социальной инфраструктуры.

Для разработки модели, позволяющей предсказывать инвестиционную стоимость земельного участка, необходимо решить ряд задач, одной из которых является сбор и анализ данных для разработки корректной и наиболее эффективной модели.

Была выдвинута гипотеза о том, что если модель сможет предсказать кадастровую стоимость, то в дальнейшем ее можно будет обучить на данных о рыночной стоимости земельного участка. Это необходимо для разработки в дальнейшем прогнозной модели. В свою очередь, для оценки инвестиционной привлекательности необходимо будет ввести параметр динамики цены земельного участка, который является разностью между стоимостями участка в разные периоды времени. Предполагается, что параметр динамики цены земельного участка можно предсказать такой же разрабатываемой математической моделью, как и кадастровую, рыночную стоимости земельных участков. А значит, и определить инвестиционную привлекательность этого участка.

Для формирования базы данных требуется провести сбор данных об участках из открытых источников с использованием публичной кадастровой карты. Изначально была предпринята попытка извлекать данные по каждому отдельному участку, однако такой метод оказался неэффективным ввиду того,

что координатой участка из Росреестра является точка центра этого участка, в то время как нас интересует геометрия участка и данные, которые могут быть представлены в виде полигонов. Оцифровка плана участка с помощью языка программирования Python также заняла значительное количество времени для достижения поставленной задачи.

Для ускорения процесса сбора данных были предприняты следующие шаги:

– решено парсить[1] не отдельные участки, а кадастровые кварталы по всей Тверской области, они выгружались значительно быстрее;

– отказ от использования API Росреестра, так как это трудозатратно и долго, в пользу библиотеки языка программирования Python rosreestr2coord, ввиду большей возможности для автоматизации и более расширенного функционала - она позволяет выгружать атрибуты кадастровых участков или кварталов вместе с геометрией.

При парсинге кадастровых кварталов с кадастровой карты Росреестра были получены все данные по этим кварталам, включая их номера и координаты центров.

Получив данные о координатах кадастровых кварталов, было решено добавить в базу данные для оценки участков с точки зрения пространственного фактора. Так, была добавлена информация о расстоянии квартала до Твери, Москвы и социальных объектов, которое рассчитывалась по дорожному графу. Была добавлена информация о кратчайшем расстоянии до трассы М11 и Волги, которое считалось по прямой.

[1] Парсинг – автоматизированный сбор данных с вебсайтов.

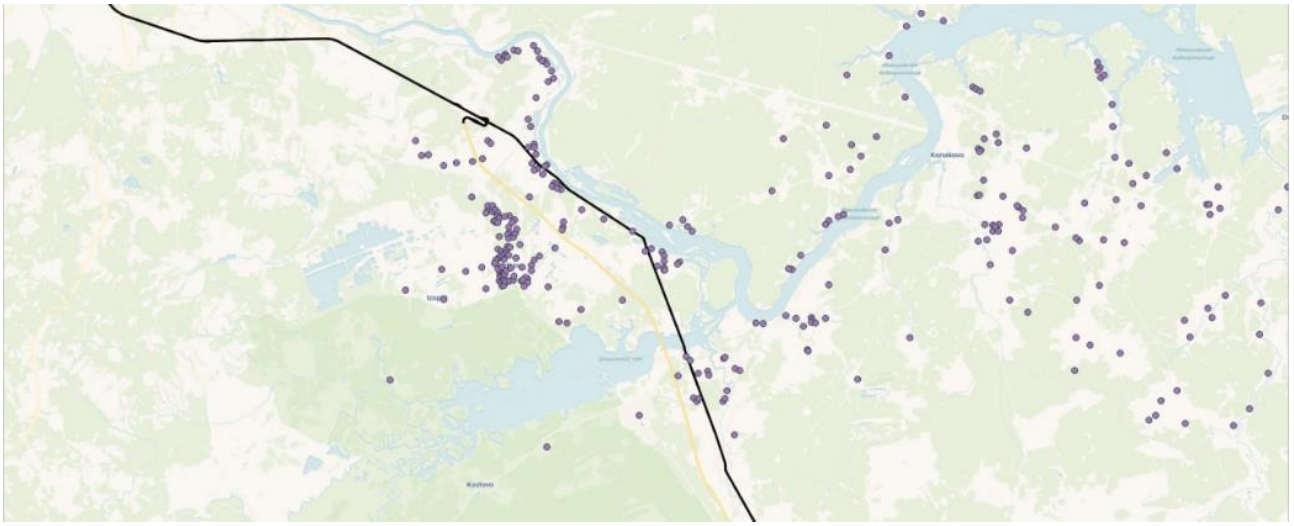

Рисунок 1. **Трасса М11 и кадастровые кварталы**

Figure 1. **M11 motorway and cadastral blocks**

В результате парсинга сайта Росреестра был получен датасет формата csv с атрибутами из Росреестра и пространственными атрибутами.

На момент написания статьи из открытых источников, описанных в п.1, была получена информация о 17 145 участках со следующими характеристиками:

- application_date – Дата подачи данных в реестр;
- cad_cost – Кадастровая стоимость;
- date_cost – Дата внесения данных в реестр;
- cc_date_entering – Дата внесения сведений;
- cc_date_approval – Дата утверждения;
- sale_doc_num – Номер документа продажи;
- parcel_tour – Объект недвижимости;
- sale_date – Дата продажи;
- parcel_build – Объект недвижимости;
- state_cd – Кадастровый номер области\муниципалитета;
- sale_dep – Департамент продажи;
- sale_doc_date – Дата документа продажи;
- cad_unit – Кадастровая единица;
- area_value – Площадь участка;

- children – Подконтрольные участки участка;
- sale_price – Цена продажи;
- util_by_doc – Тип;
- area_type – Тип площади;
- adress – Адрес;
- category_type – Тип категории земли;
- sale_doc_type – Тип документа продажи;
- parcel_build_attrs – Атрибуты объекта недвижимости;
- cn – Кадастровый номер;
- sale – Продажа;
- parcel_tour_attr – Атрибуты типа участка;
- parcel_type – Тип участка;
- area_unit – Единица площади;
- is_big – Большой участок (True\False);
- center_x – Координата X центра участка;
- center_y – Координата Y центра участка;
- avg_price_sotka – Средняя цена за сто квадратных метров по Тверской области;
- access_pt – Расстояние до остановки общественного транспорта;
- access_school – Расстояние до школ;
- access_polyclinic – Расстояние до поликлиник;
- access_kindergarten – Расстояние до детских садов;
- access_highway – Расстояние до трассы;
- hexpopulation – количество, проживающих людей в гексагоне размера 400 квадратных метров;
- hexosm_id – id гексагона;
- name – название муниципального района;
- avg_price_mun – средняя цена за сотку по муниципальному району.

Стоит отметить, что в первой итерации формирования датасета данные были неполными. Из-за этого было принято убрать из дальнейшей обработки ряд признаков, заполненных менее, чем на 70%, при детальном исследовании категориальных признаков было обнаружено, что некоторые признаки обладают слишком большим смещением распределения. Такие данные являются неприемлемыми для работы и также были убраны.

После фильтра данных, непригодных для дальнейшей обработки, было решено ставить следующие переменные:

- кадастровая стоимость (cad_cost);
- площадь участка (Area value);
- координаты участка (center_x, center_y);
- тип земельного участка (utils_fixed);
- расстояние до остановки общественного транспорта (access_pt);
- расстояние до трассы/скоростной магистрали (access_highway);
- дата внесения в реестр (cc_date);
- количество, проживающих людей в гексагоне размера 400 квадратных метров (hexpopulation);
- расстояние до поликлиники (access_polyclinic);
- расстояние до ближайшей школы (access_school);
- расстояние до детского сада (access_kindergarten).

С помощью алгоритма кластеризации был оценен уровень шума и аномальных выбросов в данных. Такая зашумленность наблюдалась для всех атрибутов, поэтому размерность датасета была снижена до 15 переменных.

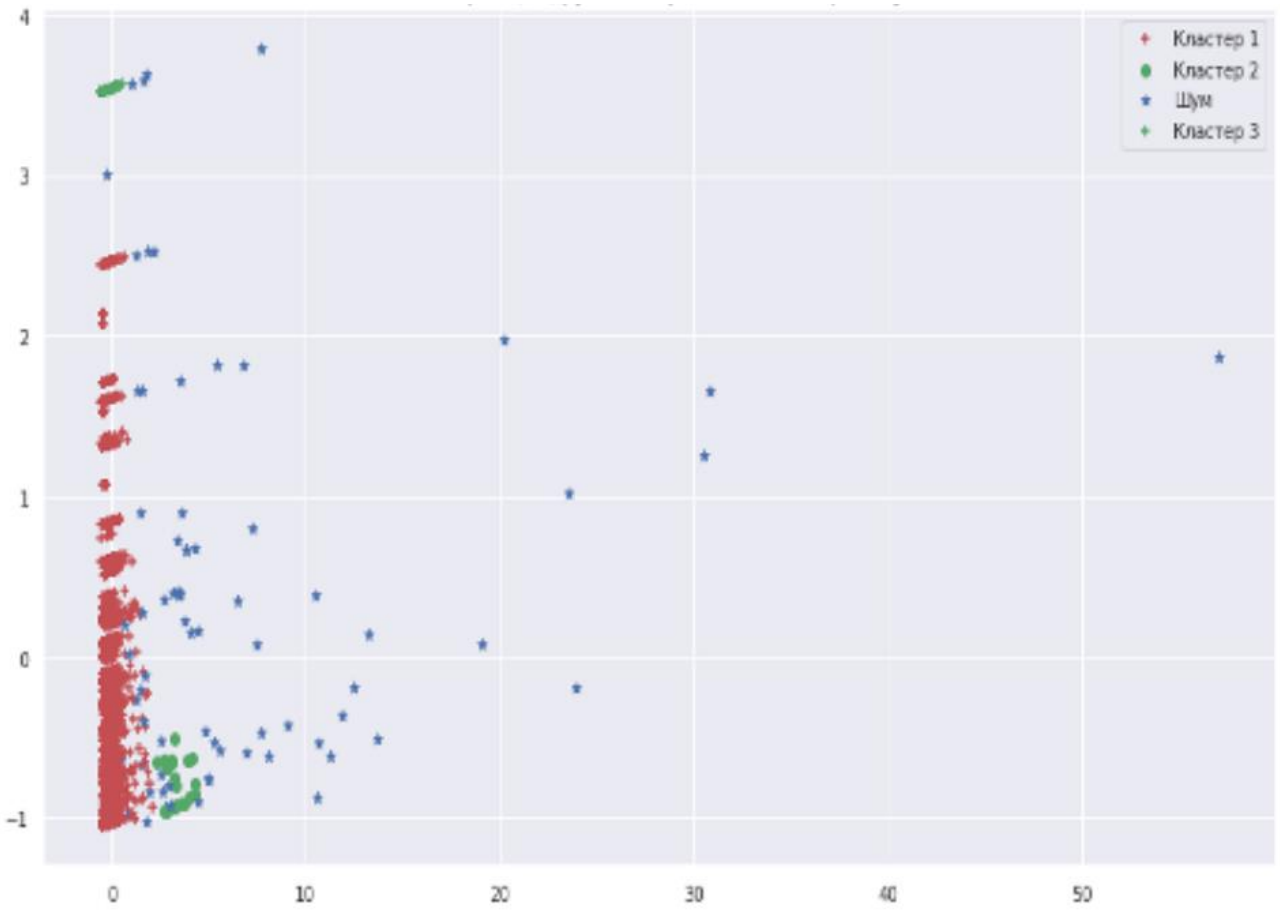


Рисунок 2. **Расстояние кластеров в зависимости от расстояния от трассы М11**

Figure 2. **The distance between clusters depends on the distance from the M11 motorway**

Было выделено 3 кластера и шумы, которые скорее всего являются грубыми промахами по параметрам. Например, кадастровой стоимости объекта (cad_cost).

Поскольку на первом этапе разработки модели были предприняты попытки предсказать кадастровую стоимость, начали с изучения именно этого параметра. Выявлен большой разброс в стоимости земельного участка. Для большинства объектов датасета кадастровая стоимость не превышает 25 000 000 рублей за земельный участок. Однако, существуют участки, которые на первый взгляд не сильно отличаются по площади от среднестатистического участка, но стоимость их доходит до 50 млн руб. Такие объекты были удалены из датасета, чтобы в дальнейшем «не путать» предсказательную модель.

Поскольку в целевой переменной возникли проблемы с большими значениями, рассмотрена также наименьшая кадастровая стоимость. Было обнаружено множество значений со стоимостью земельного участка в 1 руб. Такие участки появляются вследствие государственной кадастровой оценки земель. Согласно статье 2.5 Методических указаний № 39 [8] кадастровая стоимость земельных участков в составе земель населенных пунктов 16 вида разрешенного использования «Земельные участки улиц, проспектов, площадей, шоссе, аллей, бульваров, застав, переулков, проездов, тупиков; земельные участки земель резерва; земельные участки, занятые водными объектами, изъятыми из оборота или ограниченными в обороте в соответствии с законодательством Российской Федерации; земельные участки под полосами отвода водоемов, каналов и коллекторов, набережные» не рассчитывалась, а устанавливалась равной 1 (одному) рублю за земельный участок. Исходя из этого, было решено также отсечь участки, имеющие кадастровую стоимость меньше, чем 10 000 рублей. В результате этого распределение данных изменилось следующим образом.

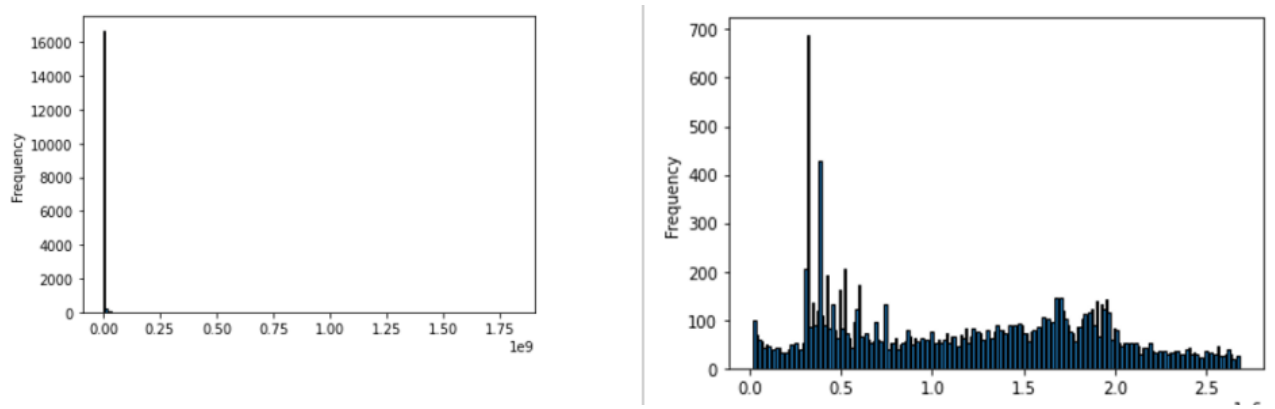


Рисунок 3. **Распределение данных после фильтрации**

Figure 3. **Distribution of data after filtering**

Был проведен анализ межквартильного диапазона - разницы между первым и третьим квартилями по значениям кадастровой стоимости cad_cost. Это позволило оценить разброс значений в наборе значений кадастровой стоимости. Преимуществом использования межквартильного диапазона для оценки

выбросов является возможность рассмотреть разброс средних значений в выборке.

В результате анализа были подобраны оптимальные квантили с учетом кадастровой стоимости ($Q_1$ = 70, $Q_2$= 25) и отброшены значения, которые выходят за пределы этого интервала, так называемые выбросы.

Как итог, было получено процентное соотношение выбросов в 17% от всех значений. Количество исследуемых объектов недвижимости составило 14 332 земельных участков.

Результатом предобработки данных является новый датасет, обладающий меньшей размерностью по количеству признаков, при этом, подходящий для обучения модели лучше, чем первоначальный вариант.

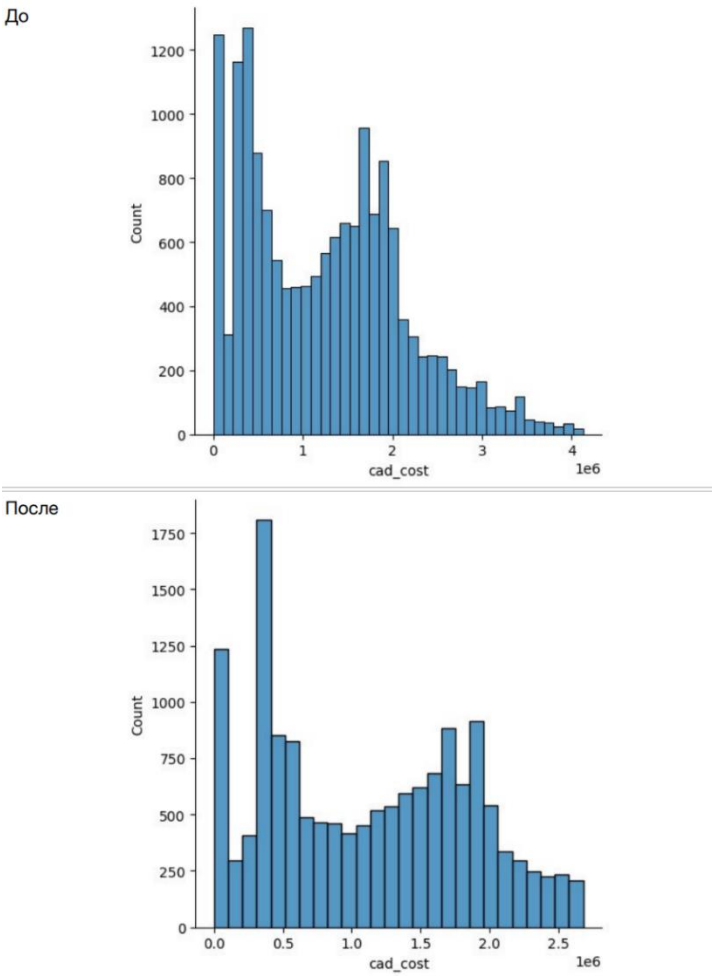


Рисунок 4. **Распределение данных после фильтрации**

Figure 4. **Distribution of data after filtering**

**Заключение**

В данной статье был представлен опыт создания базы пространственных данных для обучения предсказательной модели анализа инвестиционной привлекательности земельных участков, а также процесс фильтрации аномальных значений. Данный подход являлся многоэтапным процессом, который включал в себя сбор данных, первичный анализ и очистку данных. Благодаря проделанным шагам, была создана высококачественная база данных, которая может повысить точность и эффективность предсказательных моделей в задачах анализа инвестиционной привлекательности земельных участков. Кроме того, данный опыт позволяет брать его за основу для создания аналогичных баз данных в других областях, требующих высококачественных данных для обучения. Исследование было проведено в рамках исполнения государственного задания Министерства науки и высшего образования РФ, госбюджетная тема № FSFE-2022-0001.

**Литература**

1. Монин А.А., Плоткина А.Р. Местоположение как ключевой фактор формирования стоимости объекта недвижимости // Вестник Хабаровского государственного университета экономики и права. 2018. №6., с. 80-87. URL: https://cyberleninka.ru/article/n/mestopolozhenie-kak-klyuchevoy-faktor-formirovaniya-stoimosti-obekta-nedvizhimosti (дата обращения: 17.02.2023).

2. Плоткина А. Р. Региональные проблемы оценки и управления стоимостью собственности и реструктуризации промышленных предприятий в системе стратегического управления / А. Р. Плоткина // Экономика, статистика и информатика. Вестник УМО. 2015. №2 1. С. 94-97.

3. Монин А. А. Анализ и оценка факторов, влияющих на стоимость имущества предприятий, находящихся в процедуре банкротства / А. А. Монин // Финансовые аспекты структурных преобразований экономики : материалы Всероссийской науч.-практич. конференции. Иркутск : ИГУПС, 2011. С. 270-274.

4. Абакумов Роман Григорьевич, Моргунова Ольга Николаевна, Крылова Диана Дмитриевна Специфика ценообразования на рынке недвижимости и оценка влияния местоположения на стоимость недвижимости в городе Белгороде // Инновационная экономика: перспективы развития и совершенствования. 2018. №3 (29). URL: https://cyberleninka.ru/article/n/spetsifika-tsenoobrazovaniya-na-rynke-nedvizhimosti-i-otsenka-vliyaniya-mestopolozheniya-na-stoimost-nedvizhimosti-v-gorode-belgorode (дата обращения: 17.02.2023).

5. Овсянников А.С., Боева Т.А. Местоположение как один из факторов, определяющих инвестиционную привлекательность объектов d[fкоммерческой недвижимости // Экономика в инвестиционно-строительном комплексе и ЖКХ. 2019. №1. С. 78-83.

6. Беляева А.В. Учёт пространственных факторов в массовой оценке объектов недвижимости: сравнение эффективности различных методов // Управление большими системами. Выпуск 53, с. 6-26.

7. Ясницкий Л.Н. Ясницкий В.Л. Методика создания комплексной экономико-математической модели массовой оценки стоимости объектов недвижимости на примере квартирного рынка города Перми // Вестник Пермского университета. Сер. «Экономика» = Perm University Herald. Economy. 2016. № 2(29). С. 54–69. doi: 10.17072/1994–9960–2016–2–54–69

**References**

1. Monin A.A., Plotkina A.R. Mestopolozhenie kak klyuchevoi faktor formirovaniya stoimosti ob"ekta nedvizhimosti // Vestnik Khabarovskogo gosudarstvennogo universiteta ehkonomiki i prava. 2018. №6., s. 80-87. URL: https://cyberleninka.ru/article/n/mestopolozhenie-kak-klyuchevoy-faktor-formirovaniya-stoimosti-obekta-nedvizhimosti (data obrashcheniya: 17.02.2023).

2. Plotkina A. R. Regional'nye problemy otsenki i upravleniya stoimost'yu sobstvennosti i restrukturizatsii promyshlennykh predpriyatii v sisteme strategicheskogo upravleniya / A. R. Plotkina // Ehkonomika, statistika i informatika. Vestnik UMO. 2015. №2 1. S. 94-97.

3. Monin A. A. Analiz i otsenka faktorov, vliyayushchikh na stoimost' imushchestva predpriyatii, nakhodyashchikhsya v protsedure bankrotstva / A. A. Monin // Finansovye aspekty strukturnykh preobrazovanii ehkonomiki : materialy Vserossiiskoi nauch.-praktich. konferentsii. Irkutsk : IGUPS, 2011. S. 270-274.

4. Abakumov Roman Grigor'evich, Morgunova Ol'ga Nikolaevna, Krylova Diana Dmitrievna Spetsifika tsenoobrazovaniya na rynke nedvizhimosti i otsenka vliyaniya mestopolozheniya na stoimost' nedvizhimosti v gorode Belgorode // Innovatsionnaya ehkonomika: perspektivy razvitiya i sovershenstvovaniya. 2018. №3 (29). URL: https://cyberleninka.ru/article/n/spetsifika-tsenoobrazovaniya-na-rynke-nedvizhimosti-i-otsenka-vliyaniya-mestopolozheniya-na-stoimost-nedvizhimosti-v-gorode-belgorode (data obrashcheniya: 17.02.2023).

5. Ovsyannikov A.S., Boeva T.A. Mestopolozhenie kak odin iz faktorov, opredelyayushchikh investitsionnuyu privlekatel'nost' ob"ektov d[fkommercheskoi nedvizhimosti // Ehkonomika v investitsionno-stroitel'nom komplekse i ZHKKH. 2019. №1. S. 78-83.

6. Belyaeva A.V. Uchet prostranstvennykh faktorov v massovoi otsenke ob"ektov nedvizhimosti: sravnenie ehffektivnosti razlichnykh metodov // Upravlenie bol'shimi sistemami. Vypusk 53, s. 6-26.

7. Yasnitskii L.N. Yasnitskii V.L. Metodika sozdaniya kompleksnoi ehkonomiko-matematicheskoi modeli massovoi otsenki stoimosti ob"ektov nedvizhimosti na primere kvartirnogo rynka goroda Permi // Vestnik Permskogo universiteta. Ser. «EhkonomikA» = Perm University Herald. Economy. 2016. № 2(29). S. 54–69. doi: 10.17072/1994–9960–2016–2–54–69



**Для цитирования:** Курлов А.В. ФОРМИРОВАНИЕ БАЗЫ ПРОСТРАНСТВЕННЫХ ДАННЫХ ДЛЯ ОБУЧЕНИЯ ПРЕДСКАЗАТЕЛЬНОЙ МОДЕЛИ АНАЛИЗА ИНВЕСТИЦИОННОЙ ПРИВЛЕКАТЕЛЬНОСТИ ЗЕМЕЛЬНЫХ УЧАСТКОВ//International agricultural journal. 2023. №4, 1293-1305.