Московский экономический журнал. № 5. 2023
Moscow economic journal. № 5. 2023
Научная статья
Original article
УДК 311.213
doi: 10.55186/2413046X_2023_8_5_205

ЭКОСИСТЕМА ДЛЯ АНАЛИЗА БОЛЬШИХ ДАННЫХ В СЕЛЬСКОМ ХОЗЯЙСТВЕ

ECOSYSTEM FOR BIG DATA ANALYSIS IN AGRICULTURE

Невзоров Александр Сергеевич, ассистент кафедры статистики и кибернетики, ФГБОУ ВО Российский государственный аграрный университет – МСХА им. К.А. Тимирязева, г. Москва, E-mail: a.nevzorov@rgau-msha.ru

Демичев Вадим Владимирович, канд. экон, наук, доцент, доцент кафедры статистики и кибернетики, ФГБОУ ВО Российский государственный аграрный университет – МСХА им. К.А. Тимирязева, г. Москва, E-mail: demichev_v@rgau-msha.ru

Nevzorov Alexander Sergeevich, assistant of the Department of Statistics and Cybernetics, Russian State Agrarian University – MTAA, Moscow, E-mail: a.nevzorov@rgau-msha.ru

Demichev Vadim Vladimirovich,Candidate of Economics, Associate Professor of the Department of Statistics and Cybernetics, Russian State Agrarian University – MTAA, Moscow, E-mail: demichev_v@rgau-msha.ru

Аннотация. Эффективность деятельности производителей аграрного сектора на современном этапе развития во многом зависит от возможности перехода на инновационные технологии, к числу которых относятся цифровые технологии, включая технологии Big Data (большие данные). В результате использования больших данных может быть достигнуто существенное

повышение производительности труда, качества использования сельскохозяйственных земель, снижение издержек производства. Важным моментом в использовании результатов анализа больших данных является развертывание инфраструктуры или экосистемы больших данных. Настоящее исследование направлено на описание основных компонентов экосистемы, начиная от языков программирования и их библиотек и заканчивая конкретными методами машинного обучения как важного инструмента анализа больших данных. В работе рассматриваются вопросы предобработки и исследования данных, а также их визуализация.

Abstract. In the modern world the efficiency of agricultural manufacturers mostly depends on innovative technologies involved, such as Information Technology and Big Data. As a rule, usage of Big Data leads to significant increase in labor performance as well as quality of agricultural land use and production costs reduction. An important part in Big Data analysis is infrastructure deployment or ecosystem deployment. The given research is to describe the basic components of the ecosystem starting from programming languages and their libraries and concluding with specific methods of machine learning as a crucial instrument of Big Data analysis. The article deals with the issues of preprocessing and studying data, as well as their visualization.

Ключевые слова: большие данные, сельское хозяйство, ридж-регрессия, экосистема

Keywords: big data, agriculture, ridge regression, ecosystem

Введение. Существует множество определений термина большие данные. Наиболее точная дефиниция приведена в тексте государственного стандарта «Информационные технологии. Большие данные. Обзор и словарь» (ГОСТ Р ИСО/МЭК 20546-2021). Согласно ГОСТу большие данные (big data) – это большие массивы данных, отличающиеся главным образом такими характеристиками, как объем, разнообразие, скорость

обработки и/или вариативность, которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа [5].

В свою очередь массивы данных (dataset) определены как идентифицируемая совокупность данных, к которой можно получить доступ или скачать в одном или нескольких форматах.

Помимо указанных в государственном стандарте характеристик в литературе также можно встретить такие характеристики больших данных как достоверность, визуализация, ценность [1].

Важным элементом экосистемы анализа больших данных является язык программирования и его пакеты (библиотеки), позволяющие существенно оптимизировать процесс анализа больших данных. Широкое применение для анализа больших данных получил высокоуровневый язык программирования Python. Данный язык программирования общего назначения применяется не только в анализе данных и Data Science, но и в разработке приложений, в том числе веб-приложений, задачах автоматизации и других сферах программирования [13].

Python – эффективный инструмент решения задач анализа больших данных и достигается это во многом благодаря наличию различных специализированных библиотек [4], в число которых входят такие библиотеки как Seaborn, Statmodels, Keras, PyMC3, Plotly, Altair, Geoplotlib, Gensin, Natasha, BeautifulSoup, Feather, Ibis, ParaText, Bcolz, Blaze, Xarray, Dask и др.

Говоря об анализе больших данных необходимо сформировать перечень источников генерации больших данных. В сельском хозяйстве генерировать большие данные могут всевозможные датчики в полях и фермах, а также других производственных площадках, отслеживающие экономические, организационные, производственные и технологические процессы. Сельское хозяйство становится одним из основных потребителей

новых технологий. А это означает, что цифровизация, связанная с индустрией, касается сельского хозяйства в первую очередь. Построение экосистемы анализа больших данных, особенно на первых этапах ее развития, требует развертывания всевозможных датчиков, приборов или отладки существующих источников генерации больших данных для их последующей записи и хранения в базах данных [9].

Целью представленного исследования является описание варианта экосистемы для анализа больших данных в сельском хозяйстве РФ, этапов анализа и более детальное рассмотрение этапа моделирования на примере прогнозирования урожайности зерновых культур.

Методы или методология проведения исследования, материалы

Для решения поставленных задач и реализации цели исследования применялись следующие общенаучные методы познания: анализ и синтез, сравнение, абстракция; а также специальные статистические и технические методы, направленные на выявление закономерности развития сельского хозяйства.

Для прогнозирования урожайности использовался метод ридж-регрессии, который можно использовать для подбора модели регрессии, когда в данных присутствует мультиколлинеарность и высокий уровень вариации признаков. То есть регрессия методом наименьших квадратов пытается найти оценки коэффициентов, которые минимизируют сумму квадратов остатков (RSS):

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (1)$$

где:

y_i – фактическое значение отклика для i -го наблюдения;

$\hat{y}_i$ – прогнозируемое значение отклика на основе модели множественной линейной регрессии.

И наоборот, ридж или гребневая регрессия стремится минимизировать следующее:

$$RSS + \lambda \sum_{i=1}^{n} \beta(j)^2 \qquad (2)$$

где j находится в диапазоне от 1 до p переменных-предикторов и $\lambda \geq 0$. Этот второй член уравнения известен как штраф за усадку. В гребневой регрессии мы выбираем значение λ, которое дает наименьшую возможную тестовую MSE (среднеквадратическую ошибку).

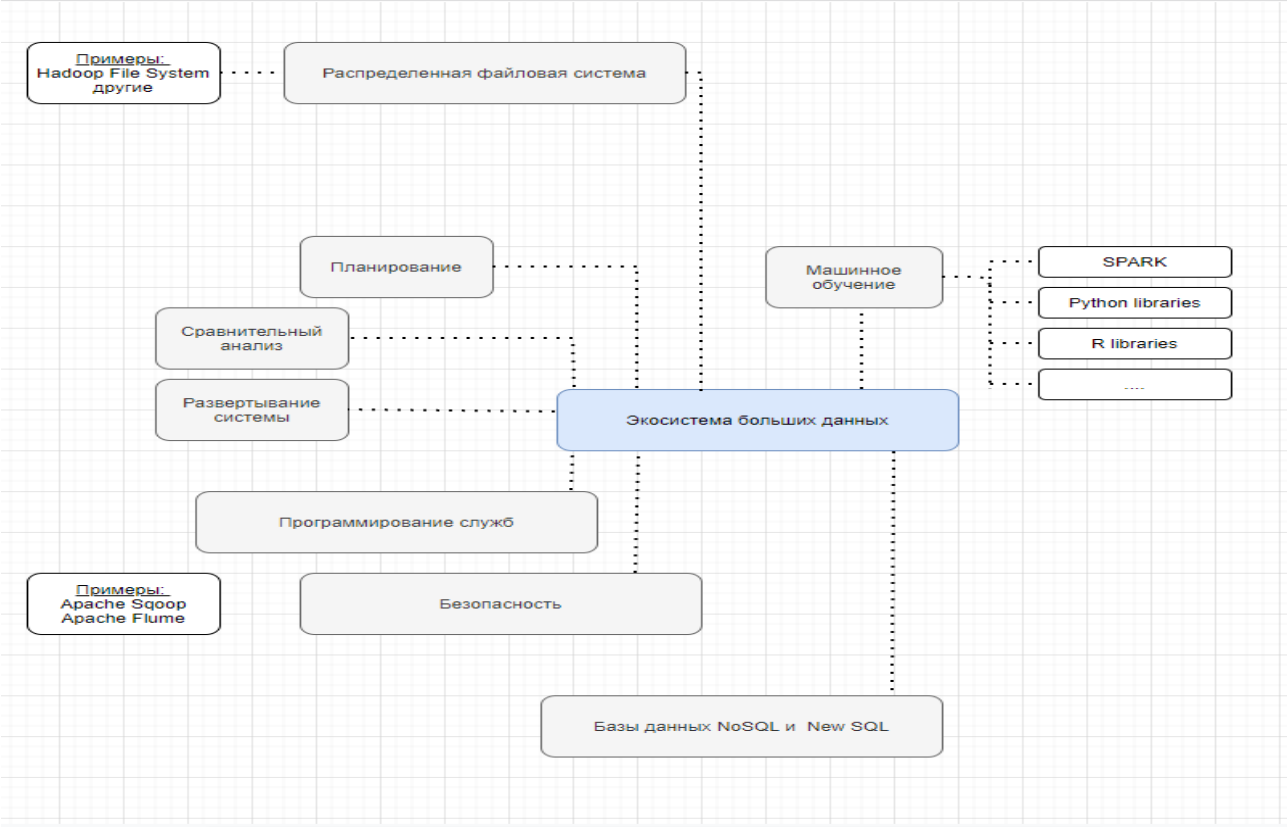
Построение ридж-регрессии проведено на языке python в среде разработки spyder.

Информационной базой проведенного анализа послужили статистические сборники Росстата («Регионы России», «Сельское хозяйство в России»), статистические издания Минсельхоза («АПК России»)[8].

Результаты и обсуждение

1. Обработка больших данных возможна и на одном компьютере, однако, на практике, мы имеем дело с целой экосистемой больших данных, реализуемой не на одном компьютере, а на целых кластерах компьютеров.

В настоящее время существует много разных инструментариев и инфраструктур больших данных, которые постоянно обновляются. Экосистема больших данных может быть разбита на группы по технологиям с похожими целями и функциональностью.



Источник: составлено автором с использованием источника [1]

Рисунок 1. **Экосистема больших данных**

Распределенная файловая система похожа на обычную файловую систему, но, в отличие от последней, она работает на нескольких серверах сразу. В основе любой файловой системы лежат такие действия, как хранение, чтение и удаление данных, а также реализация средств безопасности файлов.

Инфраструктура распределенного программирования. После того как данные будут сохранены в распределенной файловой системе, их необходимо использовать. Важный аспект работы с распределенным жестким диском состоит в том, что вы не перемещаете данные к программе, а скорее перемещаете программу к данным [2]. Сложности распределенного программирования: перезапуск сбойных заданий, отслеживание результатов из других субпроцессов и другие.

Инфраструктура машинного обучения. Когда данные оказываются на своем месте, наступает время их извлечения и анализа, с целью получению той или иной информации. С этой целью применяются методы машинного обучения, статистики и прикладной математики. Важным также является умение применять библиотеки машинного обучения, о которых мы с вами уже говорили: Scikit-learn, PyBrain, NLTK, TensorFlow и другие.

Для хранения огромных объемов данных требуется программное обеспечение, специализирующееся на управлении этими данными и формировании запросов к ним. В данном случае применяются нереляционные базы данных. Все остальные элементы являются основополагающими для развертывания системы больших данных, но не входят в компетенцию аналитика: инструменты планирования, сравнительного анализа, развертывания системы, программирование служб, безопасность [3].

2. Процесс работы с большими данными может быть описан следующим образом. На первом этапе осуществляется назначение *цели исследования*. Четко установленная цель и подробно описанные задачи позволяют эффективно спланировать весь процесс анализа больших данных. *На втором этапе* осуществляется сбор данных. Проведение качественного исследования требует сбора данных из всех доступных для авторов исследования источников. На этом этапе данные формируются в таблицах Excel, баз данных и т.д [7]. *Следующим этапом* является группировка и подготовка данных. На этом этапе данные из низкоуровневой формы преобразуются в данные, которые могут напрямую использоваться в ваших моделях. Выявление и исправление всевозможных ошибок, объединение и преобразование позволяет использовать необработанные данные в дальнейшем анализе.

Подготовка данных состоит из множества аспектов, работа над которыми существенно облегчит этап моделирования.

Очистка, интеграция и преобразование данных. Основная задача данного шага – это убрать дефекты и подготовить данные для использования в фазах моделирования и представления результатов. Это очень важный момент, потому что модели будут работать лучше и будет потрачено меньше времени на исправление аномальных результатов. Модель должна получать данные в конкретном формате, так что преобразование данных всегда будет играть важную роль[12].

Очистка данных представляет собой подпроцесс направленный на устранение ошибок в данных с тем, чтобы эти данные адекватно и последовательно представляли процесс, в результате которого они были получены. «Адекватное и последовательное представление» означает, что существует как минимум два типа ошибок. К первому типу относятся ошибки интерпретации, когда вы принимаете на веру значение в данных (пример: из данных следует, что возраст человека превышает 300 лет). Ошибки второго типа связаны с расхождениями между источниками данных или стандартизированными значениями. Например: в одной таблице денежные суммы хранятся в рублях, в другой - в долларах.

Ошибки ввода данных. Процессы сбора и ввода данных подвержены ошибкам. Они часто требуют человеческого участия, а поскольку люди не идеальны, они допускают опечатки или отвлекаются и вносят ошибки в технологическую цепочку. Впрочем, данные, собранные машинами или компьютерами, тоже не застрахованы от ошибок. Одни ошибки появляются из-за человеческого несовершенства, другие обусловлены сбоями машин или оборудования. В частности, ко второй категории относятся ошибки, происходящие из ошибок передачи данных или ошибок в фазах извлечения, преобразования и загрузки. Также могут встречаться такого рода ошибки как избыточные пробелы, расхождение в регистре символов, невозможные значения, отсутствующие значения, разные единицы измерения, разные уровни агрегирования, выбросы.

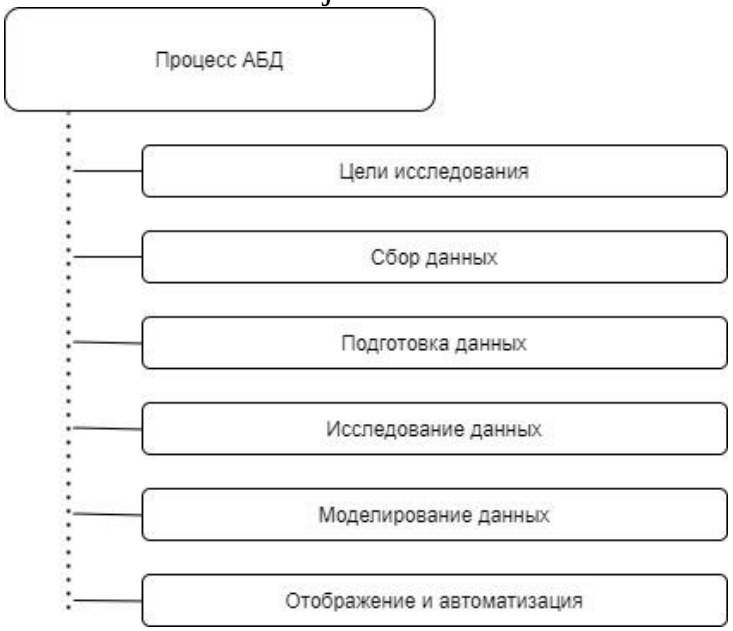
Выбросы (outliers). Выбросом называется результат наблюдений, заметно отклоняющийся от других результатов, или более конкретно – результат наблюдений, который обусловлен иной логикой или иным порождающим процессом, чем другие результаты. Простейший способ поиска выбросов основан на использовании диаграмм или таблиц с минимумами и максимумами

На четвертом этапе выполняется исследование данных. Конечной целью этого этапа является глубокое понимание данных. Осуществляется поиск закономерностей, корреляций и отклонений, основанных на визуальных и описательных методах.

Построение модели. На данном этапе осуществляется построение моделей реализации поставленных в исследовании целей – прогнозирования, классификации, кластеризации, регрессии и других. Модели могут быть достаточно сложными, например, модели машинного обучения.

Следующим этапом осуществляется отображение и автоматизация полученных результатов. Очень часто на данном этапе разрабатывается веб-приложение для дальнейшей автоматизации и отображения результатов статистического анализа. Также на данном этапе демонстрируются и интерпретируются полученные результаты, в том числе касательно наиболее важных выводов относительно предметной области исследования.

В таком виде результаты могут быть представлены заказчику или пользователю. Конечно, результаты могут быть представлены как в виде презентации, так и в виде научно-исследовательского отчета. Однако, создание веб-приложения позволяет автоматизировать производимый анализ, с возможностью его дальнейшего развития и углубления. Таким образом, результаты анализа, модели могут быть применены в другом проекте или задействованы в рабочем процессе при изменении или обновлении набора данных [10].



Источник: составлено автором

Рисунок 2. **Процесс анализа больших данных**

Представленные этапы не являются строго линейными в своем исполнении и зачастую носят итеративный характер, что означает возможность возвращения и корректировки каждого из этапов.

3. В качестве возможного варианта реализации этапа моделирования, рассмотрим построение модели прогнозирования урожайности зерновых культур на основе метода ридж-регрессии. Построение прогноза урожайности зерновых и зернобобовых на 2023, 2024, 2025 года на языке python позволяет наглядно изобразить корреляцию между зависимой и независимыми переменными.

Обозначение переменных представлено далее:

Y – Урожайность зерновых и зернобобовых (в весе после доработки), ц/га убранной площади;

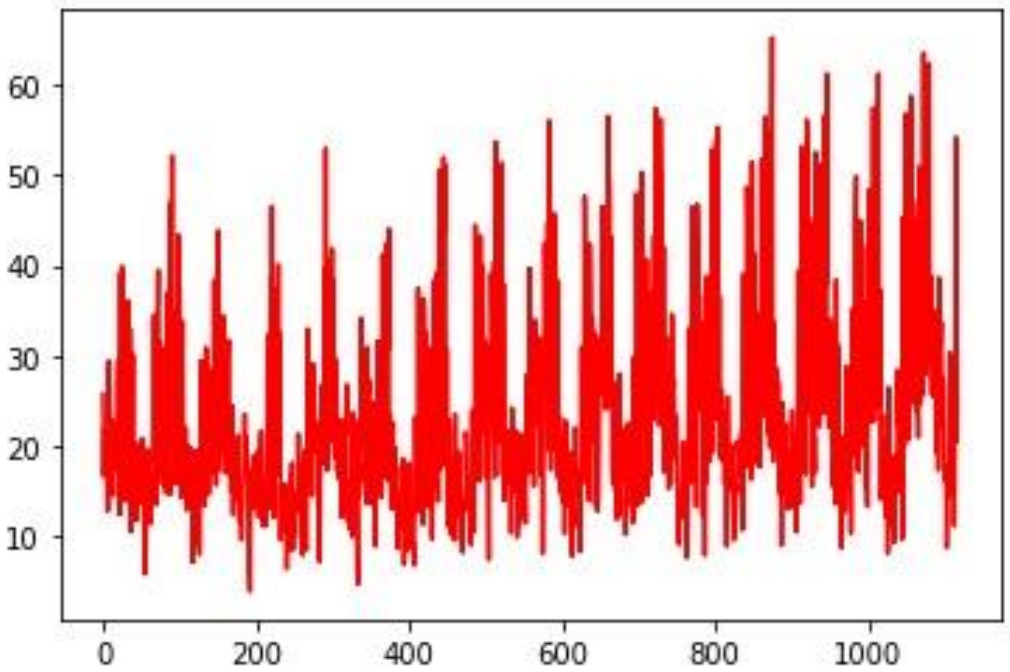
X1 – Субсидии на 1га пашни;

X2 – Внесение минеральных удобрений на один гектар посева сельскохозяйственных культур в сельскохозяйственных организациях (кг);

X3 – Удельный вес продукции растениеводства;

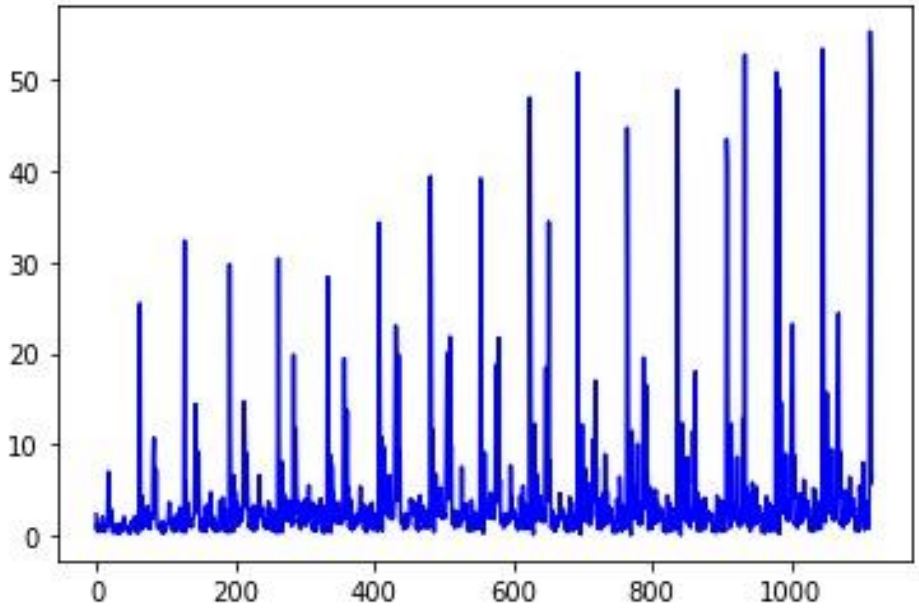
T – временная компонента (год).

В качестве набора данных были использованы панельные данные, то есть выборочная совокупность регионов России (77 регионов) за 15 лет [6]. Приведем графическое изображение ряда показателей исследования.



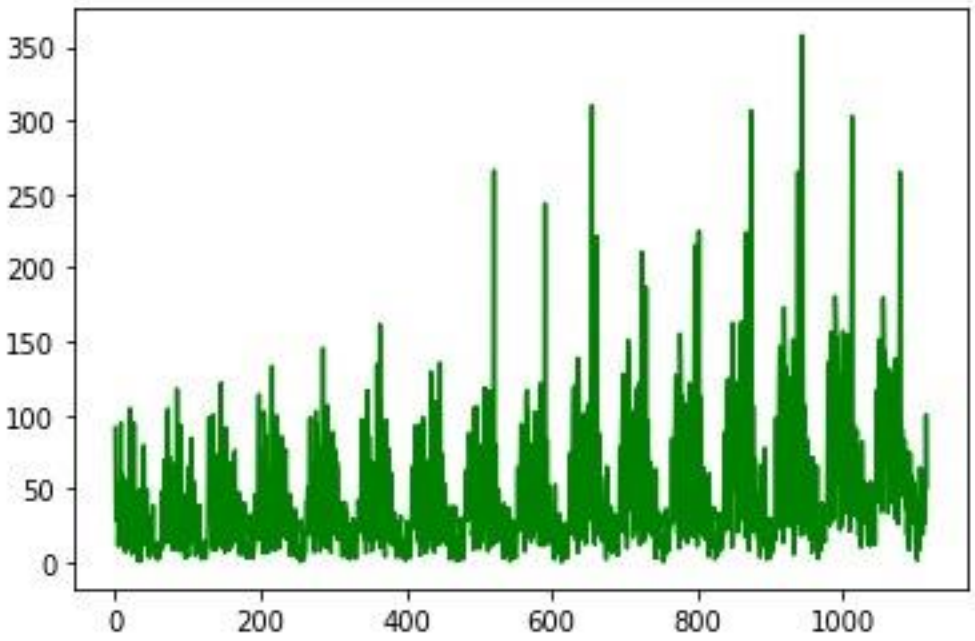
Источник: составлено автором

Рисунок 3. **Ряд панельных данных урожайности зерновых и зернобобовых культур по регионам РФ с 2007 – 2022 гг**



Источник: составлено автором

Рисунок 4. **Ряд панельных данных субсидий на 1га пашни по регионам РФ с 2007 – 2022гг**



Источник: составлено автором

Рисунок 5. **Ряд панельных данных внесения минеральных удобрений на один гектар посева сельскохозяйственных культур по регионам РФ с 2007 – 2022гг**

Данные представленные на рисунках 1-4 показывают высокий уровень вариации признаков, что обуславливает актуальность применения ридж-регрессии [11].

Построим корреляцию между переменными позволит увидеть тесноту связи (табл. 1).

Таблица 1. **Матрица парных коэффициентов корреляции**

	T	Y	X1	X2	X3
T	1.000	0.306	0.134	0.284	0.114
Y	0.306	1.000	0.057	0.309	-0.009
X1	0.134	0.057	1.000	0.003	-0.184
X2	0.284	0.309	0.003	1.000	0.345
X3	0.114	-0.009	-0.184	0.345	1.000

Величина и знак коэффициента корреляции указывают, что сила связи между переменными низкая. Коэффициент детерминации r^2 0,16 говорит о том, что только 16% вариации урожайности зерновых и зернобобовых за данный период объяснялось изменением субсидий на 1га пашни, внесением минеральных удобрений на один гектар посева сельскохозяйственных культур в сельскохозяйственных организациях и удельным весом продукции растениеводства, а 84% - другими факторами.

Низкие значения линейных коэффициентов корреляции объясняются высокой вариативностью показателей (рис. 3-5).

При построении ридж-регрессии были получены следующие результаты. Лучшее полученное значение alpha составило 0.99 для построения ридж-регрессии, которое дает наименьшую возможную тестовую MSE (среднеквадратическую ошибку). Предсказанные средние значения Y (урожайность зерновых по совокупности регионов) на 2023-2025: 28.4, 29,0, 29.6 ц/га соответственно, с уровнем доверия 95%.

Выводы. Для повышения эффективности сельскохозяйственного производства может быть развернута экосистема анализа больших данных, включающая инфраструктуру больших данных, процесс планирования и непосредственно этапы анализа больших данных. Это позволяет получать эффективные и точные оценки прогнозных значений, на основе которых могут быть приняты взвешенные управленческие решения. Основные этапы анализа больших данных – постановка цели, сбор данных, подготовка данных, исследование данных, моделирование данных, отображение и автоматизация. В качестве примера этапа моделирования рассмотрен пример обработки данных с применением ридж-регрессии с прогнозом урожайности зерновых и зернобобовых культур на 2023, 2024, 2025 годы с применением языка программирования Python.

Список источников

1. Демичев В.В. Влияние больших данных на развитие сельского хозяйства России // Российский экономический интернет-журнал. – 2020. – № 3. – С. 10.
2. Дэви, С. Основы Data Science и Big Data. Python и наука о данных / С. Дэви, М. Арно, А. Мохамед. – СПб.: Питер, 2018. – 336 с.
3. Миркин, Б. Г. Введение в анализ данных : учебник и практикум / Б. Г. Миркин. – Москва : Издательство Юрайт, 2023. – 174 с. – (Высшее образование). – ISBN 978-5-9916-5009-0. – Текст : электронный // Образовательная платформа Юрайт [сайт]. – URL: https://urait.ru/bcode/511121(Дата обращения: 5.04.2023).
4. Златопольский Д.М. Основы программирования на языке Python. – М.: ДМК Пресс, 2017. – 284 с.
5. Национальный стандарт РФ ГОСТ Р ИСО/МЭК 20546-2021 "Информационные технологии. Большие данные. Обзор и словарь" (утв. и введен в действие приказом Федерального агентства по техническому регулированию и метрологии от 13 июля 2021 г. N 632-ст).
6. Панельные данные Электронный ресурс URL: https://ru.wikipedia.org/wiki/%D0%9F%D0%B0%D0%BD%D0%B5%D0%BB%D1%8C%D0%BD%D1%8B%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5 (Дата обращения: 5.04.2023).
7. Парфенов, Ю. П. Постреляционные хранилища данных : учебное пособие для вузов / Ю. П. Парфенов ; под научной редакцией Н. В. Папуловской. – Москва : Издательство Юрайт, 2023. – 121 с. – (Высшее образование). – ISBN 978-5-534-09837-2. – Текст : электронный // Образовательная платформа Юрайт [сайт]. – URL: https://urait.ru/bcode/514724 (Дата обращения: 5.04.2023).
8. Федеральная служба государственной статистики. Электронный ресурс URL: https://rosstat.gov.ru/.
9. Цифровая трансформация сельского хозяйства России: офиц. изд. – М.: ФГБНУ «Росинформагротех», 2019 – 80 с.

10. 180 Data Science and Machine Learning Projects with Python. URL: https://medium.com/coders-camp/180-data-science-and-machine-learning-projects-with-python-6191bc7b9db9 (Дата обращения: 5.04.2023).

11. Data Visualization with Python. Электронный ресурс URL: https://www.geeksforgeeks.org/data-visualization-with-python/ /(Дата обращения: 5.04.2023)

12. Machine Learning Repository. Электронный ресурс URL: https://archive.ics.uci.edu/ml/datasets.php (Дата обращения: 5.04.2023).

13. Python Machine Learning Tutorials. Электронный ресурс URL: https://realpython.com/tutorials/machine-learning/ (Дата обращения: 5.04.2023).

References

1. Demichev V.V. The impact of big data on the development of Russian agriculture // Russian Economic Internet Journal. - 2020. - No. 3. - P. 10.

2. Davy, S. Fundamentals of Data Science and Big Data. Python and Data Science / S. Davy, M. Arno, A. Mohamed. - St. Petersburg: Peter, 2018. - 336 p.

3. Mirkin, B. G. Introduction to data analysis: textbook and workshop / B. G. Mirkin. - Moscow: Yurayt Publishing House, 2023. - 174 p. - (Higher education). – ISBN 978-5-9916-5009-0. – Text: electronic // Educational platform Urayt [website]. – URL: https://urait.ru/bcode/511121 (Date of access: 04/05/2023).

4. Zlatopolsky D.M. Fundamentals of programming in Python. – M.: DMK Press, 2017. – 284 p.

5. National standard of the Russian Federation GOST R ISO / IEC 20546-2021 "Information technology. Big data. Overview and dictionary" (approved and put into effect by order of the Federal Agency for Technical Regulation and Metrology dated July 13, 2021 N 632-st).

6. Panel data Electronic resource URL: https://ru.wikipedia.org/wiki/%D0%9F%D0%B0%D0%BD%D0%B5%D0%BB%D1%8C%D0%BD%D1

%8B%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5
(Date of access: 04/05/2023).

7. Parfenov, Yu. P. Post-relational data warehouses: a textbook for universities / Yu. P. Parfenov; under the scientific editorship of N. V. Papulovskaya. - Moscow: Yurayt Publishing House, 2023. - 121 p. - (Higher education). – ISBN 978-5-534-09837-2. – Text: electronic // Educational platform Urayt [website]. – URL: https://urait.ru/bcode/514724 (Date of access: 04/05/2023).

8. Federal State Statistics Service. Electronic resource URL: https://rosstat.gov.ru/.

9. Digital transformation of Russian agriculture: official. ed. - M.: FGBNU "Rosinformagrotech", 2019 - 80 p.

10. 180 Data Science and Machine Learning Projects with Python. URL: https://medium.com/coders-camp/180-data-science-and-machine-learning-projects-with-python-6191bc7b9db9 (Date of access: 04/5/2023).

11. Data Visualization with Python. Electronic resource URL: https://www.geeksforgeeks.org/data-visualization-with-python/ /(Date of access: 04/5/2023)

12. Machine Learning Repository. Electronic resource URL: https://archive.ics.uci.edu/ml/datasets.php (Date of access: 04/05/2023).

13. Python Machine Learning Tutorials. Electronic resource URL: https://realpython.com/tutorials/machine-learning/ (Date of access: 04/5/2023).

Для цитирования: Невзоров А.С., Демичев В.В. Экосистема для анализа больших данных в сельском хозяйстве // Московский экономический журнал. 2023. № 5. URL: https://qje.su/selskohozyajstvennye-nauki/moskovskij-ekonomicheskij-zhurnal-5-2023-13/