Государственное и муниципальное управление. Ученые записки. 2023. № 3. С. 67–74
State and Municipal Management. Scholar Notes. 2023;(3):67–74
Проблемы экономики
Научная статья
УДК 303.722.4
https://doi.org/10.22394/2079-1690-2023-1-3-67-74 EDN NESYPG

Кластерная модель анализа больших данных в животноводческом производстве

Ольга Владимировна Галанина[1], Юлия Павловна Золотарёва[2]

[1, 2]Санкт-Петербургский государственный аграрный университет, Санкт-Петербург, Россия

[1]https://orcid.org/0000-0003-3156-2906 [2]https://orcid.org/0000-0003-1930-1519

Автор, ответственный за переписку: Ольга Владимировна Галанина, olga_galanina@inbox.ru

Аннотация. Интеллектуальные методы анализа, к которым относится задача кластеризации, все шире применяются в сфере экономики АПК. Задача кластеризации относится к классу задач обучения без учителя. Суть задачи – группировка объектов исследования по принципу схожести. Если рассматривать регионы РФ с точки зрения производства животноводческой продукции, их так же можно автоматически сгруппировать по принципу схожести. Метод k-средних на данный момент является основным методом решения задач кластеризации. Основным этапом задачи классификации является формирование набора данных, в который входят все основные характеристики объекта. Например, если рассматривать производство региона с точки зрения отрасли животноводства, то логичнее будет использовать x1 – производство мяса на душу населения и x2 – производство молока на душу населения. Критерием выбора количества кластеров является суммарная среднеквадратическая ошибка. Всего в анализе участвовало 79 регионов РФ. Оказалось, что рациональнее разбивать все регионы РФ на 7 кластеров схожести. Были выделены регионы с высоким производством молока и мяса (кластеры 4 и 6), регионы со средним производством молока и мяса (кластеры 2, 3, 5) и регионы с низким производством молока и мяса (кластеры 0, 1).

Ключевые слова: анализ данных, большие данные, животноводство, искусственный интеллект, кластерная модель, кластеризация, метод k-средних

Для цитирования: Галанина О. В., Золотарёва Ю. П. Кластерная модель анализа больших данных в животноводческом производстве // Государственное и муниципальное управление. Ученые записки. 2023. № 3. С. 67–74. https://doi.org/10.22394/2079-1690-2023-1-3-67-74. EDN NESYPG

Problems of Economics

Original article

Cluster model for big data analysis in livestock production

Olga V. Galanina[1], Julia P. Zolotaryova[2]

[1, 2]Saint-Petersburg State Agrarian University, St.-Petersburg, Russia

[1]https://orcid.org/0000-0003-3156-2906 [2]https://orcid.org/0000-0003-1930-1519

Corresponding author: Olga V. Galanina, olga_galanina@inbox.ru

Abstract. Intelligent methods of analysis, which include the problem of clustering, are widely used in the field of economics of the agro-industrial complex. The clustering problem belongs to the class of unsupervised learning problems. The essence of the problem is the grouping of research objects according to the use of similarity. If the regions of the Russian Federation are selected in terms of livestock production, they can also be automatically grouped according to the similarity recipe. The k-means method is currently a successful method for solving clustering problems. The main stage of solving the problem is the collection of data, which includes all the main characteristics of the object. For example, if you set up production in the region in terms of animal husbandry, then it would be more logical to x1 - meat production per capita and x2 – milk production per capita. The criterion for choosing the number of clusters is the root mean square error. In total, 79 regions of the Russian Federation participated in the analysis. It turned out that the potential to break all regions of the Russian Federation into 7 clusters of similarity. Regions with high milk and meat production (clusters 4 and 6), regions with high milk and meat production (clusters 2, 3, 5) and regions with low milk and meat production (clusters 0, 1) were identified.

Keywords: artificial intelligence, data analysis, big date, k-means method, cluster model, animal husbandry
For citation: Galanina O. V., Zolotaryova Ju. P. Cluster model for big data analysis in livestock production. *State and Municipal Management. Scholar Notes.* 2023;(3):67-74. (In Russ.). https://doi.org/10.22394/2079-1690-2023-1-3-67-74. EDN NESYPG

ВВЕДЕНИЕ. Успехи вычислительной техники, развитие теории искусственного интеллекта и возникновение феномена Big Data позволяют формулировать и эффективно решать задачи анализа данных и в сфере экономики сельского хозяйства. Кластеризация, или сегментация – одна из задач интеллектуального анализа данных, весьма трудоёмка без вычислительных мощностей ЭВМ. В процессе кластеризации (сегментации) объекты группируются в соответствии с их свойствами в один из кластеров (классов с заранее неизвестными параметрами). Количество кластеров может быть предварительно определено, но, каждый кластер не имеет четкого описания. Есть видимые различия между объектами из разных кластеров, в то время как объекты внутри одного кластера похожи по максимуму. Самый известный алгорит решения задачи кластеризации (сегментации) – это метод k-средних.

О возможностях применения кластерного интеллектуального анализа в эпоху цифровизации и трансформации экономики говорят многие исследователи. Например, в работах Погонышевой Д. А. [1, с. 61], Смелик Н. Л. [2, с. 98] указывается, что цифровизация, роботизация и интеллектуальные методы приобретают первоочередную значимость, особенно в отраслях АПК. Непосредственно в теории и практике кластерного анализа АПК и других отраслях производства посвящены работы Смагина Б. И. [3, с. 2], Меньшиковой М. А. [4, с. 457], Замбржицкой Е. С. [5, с. 110], Шамсутдинова Т.М. [6, с. 469]. Вся инновационная деятельность в сфере АПК должна быть нацелена на внедрение цифровых технологий и интеллектуальных методов [7, с. 29].

Интеллектуальные методы разделяются на методы машинного обучения и глубокого обучения. Глубокое обучение использует большие данные и большие вычислительные мощности, возможно без участия человека. Машинное же обучение может использоваться на небольших наборах данных, маломощных компьютерах и требует непосредственного контроля со стороны человека. Таким образом, пока у нас нет доступа к большим данным, можем методами машинного обучения решать задачи кластеризации на основе доступной статистической информации.

Представляет интерес деления регионов РФ по объемам производства животноводческой продукции [8, с. 190]. Алгоритмы кластеризации позволяют автоматически разделять регионы РФ по показателям производства животноводческой продукции на кластеры, и внутри каждого кластера будут определены регионы со схожими состояниями по показателям животноводческого производства. Конечно, можно выработать критерии отнесения регионов к тому или иному кластеру. Но гораздо рациональнее автоматизировать этот процесс.

В качестве исходных данных для подобного вида анализа можно выбирать информацию электронного документооборота [9, с. 215], внутреннюю отчетность с.-х. организаций [10, с. 53] и информацию нормативно-технических баз, информацию органов государственного управления [11, с. 70], данные, поступающие с «умных ферм» [12, с. 93].

ОБЪЕКТ И МЕТОДИКА. Исследуем отрасль производства животноводческой продукции по регионам. Поставим задачу сегментировать животноводческое производство на группы схожести, причем число сегментов определим экспериментально. Критерием выбора числа сегментов будем считать значение суммарной среднеквадратической ошибки . Для решения задачи потребуется:

1) выделить показатели, которые могут быть использованы для решения задачи кластеризации регионов по состоянию животноводства;

2) реализовать для выбранного набора данных метод k-средних для различного количества сегментов (кластеров);

3) рассчитать для каждого числа сегментов метрику (суммарную среднеквадратическую ошибку);

4) визуализировать результаты;

5) сделать выводы по результатам расчетов и визуализации.

Информационная модель задачи представлена на рис. 1. Некластеризованные регионы, проходя через метод k-средних, по принципу схожести группируются или сегментируются. Причем, исследователь задает число кластеров, получая на выходе значение суммарной среднеквадратической ошибки (метрики).

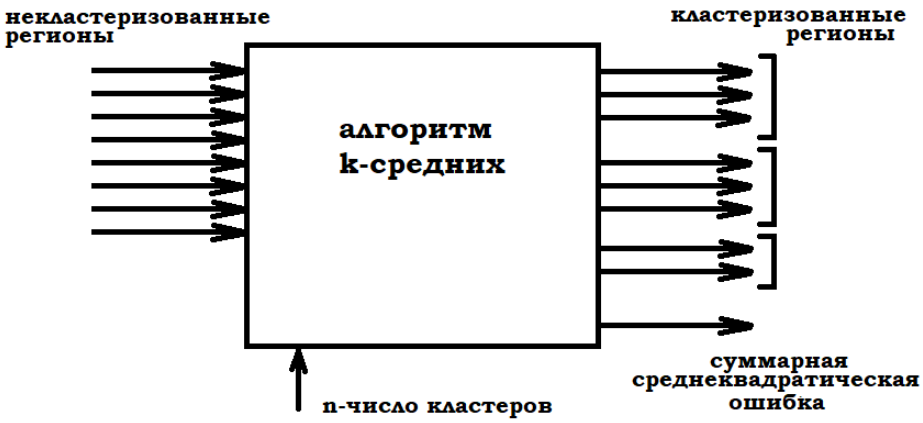

Рис. 1. Информационная модель исследования
Fig. 1. Research information model

В качестве оценки отрасли производства животноводческой продукции можно считать произведенное количество животноводческого продукта на душу населения, а именно:

X1 – производство в регионе скота и птицы на убой на человека в год, кг/чел/год;

X2 – производство в регионе молока на человека в год, кг/чел/год.

Эти значения могут быть рассчитаны из доступной статистической информации, а именно, в справочниках представлены следующие показатели за 2020г:

– среднегодовая численность населения в регионе РФ,

– производство скота и птицы на убой (в убойном весе) в регионе РФ,

– производство молока в регионе РФ.

Всего было вделено 79 регионов в составе РФ.

В качестве входных данных был сформирован датасет следующего содержания, представленный в табл. 1.

Таблица 1 – Общий вид набора данных для задачи кластеризации
Table 1 – General view of the data set for the clustering task

N	**Регион РФ**	**X1 – производство в регионе скота и птицы на убой на человека в год, кг/чел/год (2020г.)**	**X2 – производство в регионе молока на человека в год, кг/чел/год (2020г)**
1	Белгородская обл.	888,30	444,34
2	Брянская обл.	288,31	248,40
…	…	…	…
78	Еврейская АО	7,62	59,72
79	Чукотский АО	10,02	0,01

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ. Расчеты производились с использованием свободного программного обеспечения интеллектуального анализа данных WEKA для числа кластеров 3÷10. Суммарная среднеквадратическая ошибка (метрика) для каждого количества кластеров представлена на рис. 2. Выбор числа кластеров зависит от поведения суммарной среднеквадратической ошибки. До значений 7-8 кластеров ошибка уменьшается значительно. При разбиении более чем на 8 кластеров, ошибка уменьшается незначительно. Поэтому оптимальное число кластеров составляет 7-8.

Распределение регионов на 7 кластеров представлено в табл.2, а распределение регионов на 8 кластеров – в табл.3.

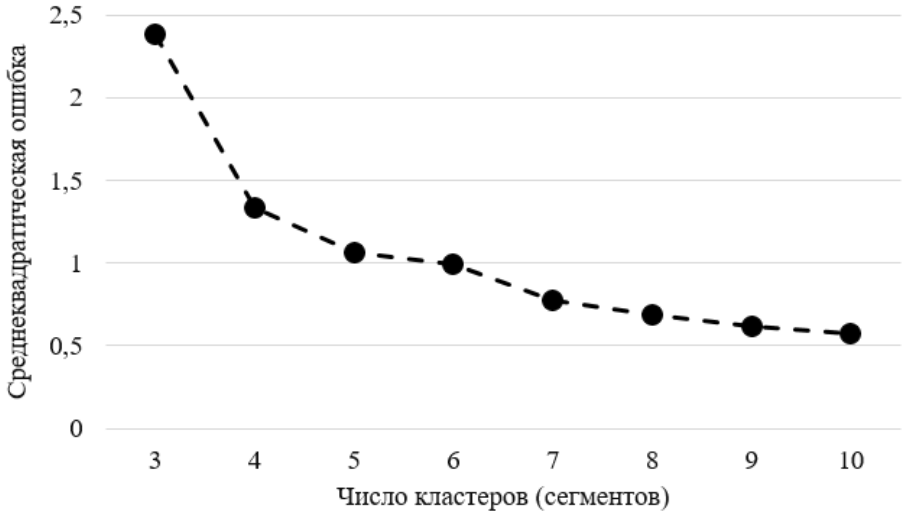

Рис. 2. Суммарная среднеквадратическая ошибка в зависимости от числа кластеров
Fig. 2. Total mean square error depending on the number of clusters

Таблица 2 – Распределение регионов РФ по 7 кластерам схожести
Table 2 – Distribution of regions of the Russian Federation by 7 clusters of similarity

Номер кластера	Доля	Регионы РФ
0	10 %	**Республика Коми, Мурманская обл., Камчатский край, Приморский край, Хабаровский край, Магаданская обл., Еврейская АО, Чукотский ОА**
1	18 %	Ивановская обл., **Московская обл., Тульская обл., Республика Карелия, Архангельская обл., Республика Крым, Самарская обл.,** Тюменская обл., **Челябинская обл., Кемеровская обл., Томская обл., Республика Бурятия,** Республика Саха (Якутия), **Сахалинская обл.**
2	30 %	**Костромская обл., Смоленская обл., Тверская обл.,** Ярославская обл., **Калининградская обл.**, Республика Адыгея, Краснодарский край, **Астраханская обл.,** Волгоградская обл., Ростовская обл, **Республика Дагестан, Республика Ингушетия,** Республика Северная Осетия – Алания, **Чеченская Республика, Пермский край, Нижегородская обл., Ульяновская обл.,** Курганская обл., **Свердловская обл., Республика Тыва,** Республика Хакасия, Красноярский край, **Иркутская обл., Амурская обл.**
3	15 %	**Владимирская обл., Калужская обл., Ленинградская обл., Карачаево-Черкесская Республика, Республика Башкортостан, Чувашская Республика, Оренбургская обл., Саратовская обл., Республика Алтай,** Новосибирская обл., **Омская обл, Забайкальский край**
4	11 %	**Воронежская обл., Рязанская обл., Вологодская обл., Кабардино-Балкарская Республика, Республика Мордовия, Республика Татарстан, Удмуртская Республика, Кировская обл., Алтайский край**
5	9 %	**Брянская обл., Липецкая обл., Орловская обл.,** Новгородская обл., **Республика Калмыкия**, Ставропольский край, **Пензенская обл.**
6	6 %	**Белгородская обл., Курская обл, Тамбовская обл, Псковская обл., Республика Марий Эл**

Таблица 3 – Распределение регионов РФ по 8 кластерам схожести

Table 3 – Distribution of regions of the Russian Federation by 8 clusters of similarity

Номер кластера	Доля	Регионы РФ
0	10 %	**Республика Коми**, **Мурманская обл., Камчатский край, Приморский край, Хабаровский край, Магаданская обл**., **Еврейская АО, Чукотский ОА**
1	15 %	**Московская обл., Тульская обл., Республика Карелия, Архангельская обл.**, Новгородская обл., **Республика Крым**, **Самарская обл., Челябинская обл., Кемеровская обл., Томская обл., Республика Бурятия, Сахалинская обл**.
2	24 %	Ивановская обл., **Костромская обл., Смоленская обл., Тверская обл., Калининградская обл., Астраханская обл., Республика Дагестан, Республика Ингушетия**, **Чеченская Республика**, Ставропольский край, **Пермский край, Нижегородская обл., Ульяновская обл., Свердловская обл.,** Тюменская обл., **Республика Тыва, Иркутская обл**., Республика Саха (Якутия), **Амурская обл.**
3	14 %	**Владимирская обл., Калужская обл., Ленинградская обл., Карачаево-Черкесская Республика, Республика Башкортостан, Чувашская Республика, Оренбургская обл., Саратовская обл., Республика Алтай**, **Омская обл**, **Забайкальский край**
4	11 %	**Воронежская обл., Рязанская обл., Вологодская обл., Кабардино-Балкарская Республика, Республика Мордовия, Республика Татарстан, Удмуртская Республика, Кировская обл., Алтайский край**
5	6 %	**Брянская обл., Липецкая обл., Орловская обл., Республика Калмыкия, Пензенская обл**.
6	6 %	**Белгородская обл., Курская обл, Тамбовская обл, Псковская обл., Республика Марий Эл**
7	13 %	Ярославская обл., Республика Адыгея, Краснодарский край, Волгоградская обл., Ростовская обл., Республика Северная Осетия – Алания, Курганская обл., Республика Хакасия, Красноярский край, Новосибирская обл.

Таким образом можно заметить, что существуют регионы, которые вне зависимости он числа разбиений на кластеры, группируются вместе. В табл. 2–3 эти регионы, стабильно схожие, выделены. На рис. 3 разбиение на кластеры представлено графически.

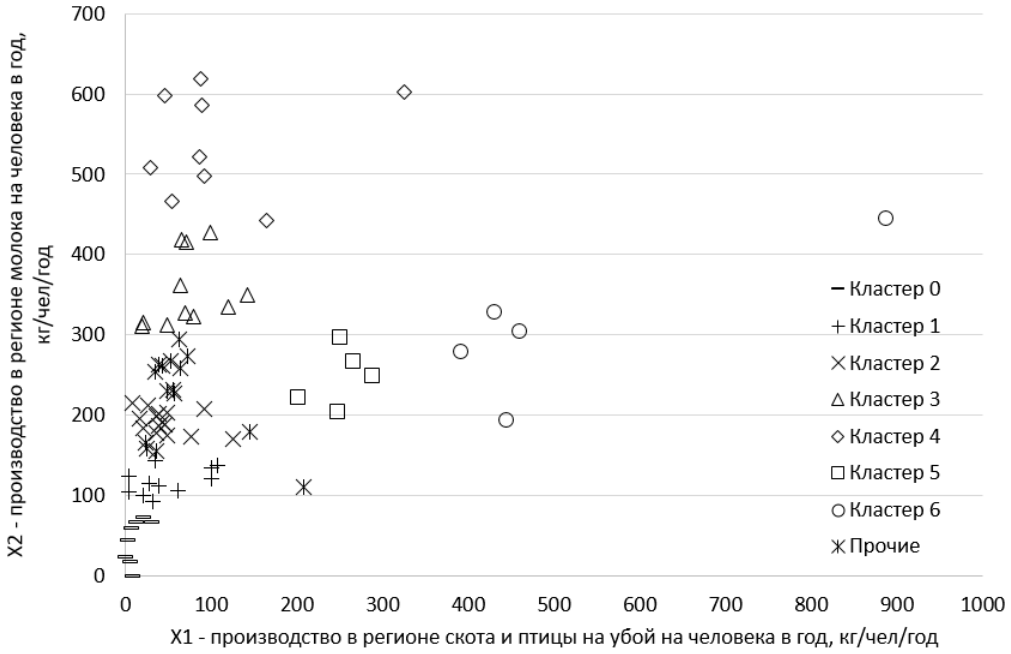

Рис. 3. Устойчивые кластеры схожести регионов РФ
Fig. 3. Stable clusters of similarity of regions of the Russian Federation

Поскольку разбиение на кластеры методом k-средних не учитывает принцип схожести, в нем нет четкого описания принципов схожести, исходя из визуализации рисунка 3 мы можем говорить, что можно установить принцип, по которому регионы разбились на группы схожести. А именно, кластер 0 – самые слабые с точки зрения производства продукции животноводства. В этих регионах минимальное или нулевое производство мяса и молока на человека. Кластер 1 – регионы с низким производством животноводческой продукции (малое производство молока и мяса). Кластеры 2, 3, 5 – средние по производству и мяса, и молока. Кластер 4 – регионы с высоким производством молока, но низким производством мяса, Кластер 6 – Регионы с высоким производством мяса и средним производством молока.

Так же разбиение на кластеры представлено на карте рисунка 4. В основном, кластеры 6 и 4 размещены в европейской части РФ.

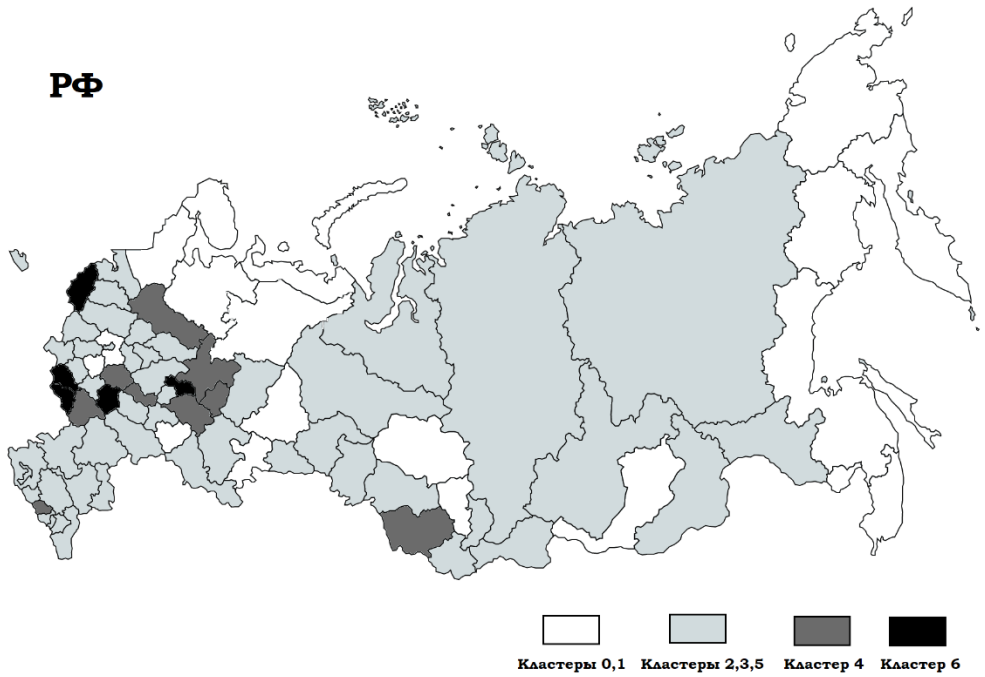

Рис. 4. Кластеры на карте
Fig. 4. Clusters on the map

ВЫВОДЫ.

1. Были выделены показатели для решения задачи кластеризации регионов по принципу производства животноводческой продукции, а именно, X1 – производство в регионе скота и птицы на убой на человека в год, кг/чел/год и X2 – производство в регионе молока на человека в год, кг/чел/год. Всего число записей составило 79 регионов.
2. Значение суммарной среднеквадратической ошибки показало, что разбиение на 7-8 кластеров схожести представляется наиболее приемлемым (рисунок 2).
3. Результаты кластеризации показали, что регионы, по состоянию производства животноводческой продукции группируются по следующему принципу схожести:

- регионы с высоким производством мяса и средним производство молока на душу населения (регионы кластера 6), находятся в европейской части РФ;
- регионы с высоким производство молока и низким производством мяса (регионы кластера 4), находятся преимущественно в европейской части РФ;
- регионы со средним производством молока и мяса (регионы кластеров 2, 3, 5);
- регионы с низким производством молока и мяса на душу населения (регионы кластеров 0 и 1), находятся преимущественно в северных и дальневосточных районах РФ.

Используя полученные результаты можно повысить качество и обоснованность решений в экономической сфере, принимая во внимание различия в уровне развития животноводства в разных регионах.

Список источников

1. Погонышева Д. А., Савин А. В., Серая Г. В., Тасоева Е. В. Цифровые технологии в кадровом менеджменте в сельском хозяйстве // Вестник Брянской государственной сельскохозяйственной академии. 2021. № 3(85). С. 60-66. DOI 10.52691/2500-2651-2021-85-3-60-66. EDN ZYGAOT.

2. Смелик Н. Л. Сущность трансформации экономической системы и ее механизма // Известия Санкт-Петербургского государственного аграрного университета. 2014. № 37. С. 97-103. EDN UXWKXL.

3. Смагин Б. И. Использование кластерного анализа в анализе экономических процессов сельскохозяйственного производства // Наука и Образование. 2021. Т. 4. № 2. EDN FPYMAK.

4. Меньшикова М. А., Ходыревская В. Н., Симахина О. Н., Шакирова Д. Ф. Территориальный анализ уровня развития строительства в России методом кластерного анализа // Экономика и предпринимательство. 2021. № 2(127). С. 456-460. DOI 10.34925/EIP.2021.127.2.086. EDN KSZPEO.

5. Замбржицкая Е. С. Кластерный анализ как предварительный этап анализа безубыточности // Приложение математики в экономических и технических исследованиях. 2020. № 1(10). С. 109-115. EDN WXKTIH.

6. Шамсутдинова Т. М. Технологии интеллектуального анализа статистических данных (на примере кластерного анализа показателей сельскохозяйственного производства субъектов РФ) // Современные научно-практические решения в АПК : Материалы международной научно-практической конференции, Воронеж, 06–07 июня 2017 года. Воронеж: Воронежский государственный аграрный университет им. Императора Петра I, 2017. С. 467–473. EDN ZWGBHD.

7. Kosyakova L. N., Popova A. L. Innovative policy in the agricultural sphere // British Journal for Social and Economic Research. 2016. Vol. 1, No. 2. P. 29-38. EDN XCPWLV.

8. Лаврова А. П. Роль личных подсобных хозяйств сельского населения в продовольственном обеспечении // Известия Санкт-Петербургского государственного аграрного университета. 2015. № 40. С. 186-191. EDN UXWNVZ.

9. Ульянова Н. Д. Искусственный интеллект в системах электронного документооборота // Проблемы энергообеспечения, автоматизации, информатизации и природопользования в АПК : Сборник материалов международной научно-технической конференции, Брянск, 30 апреля 2022 года. Брянск: Брянский государственный аграрный университет, 2022. С. 214-220. EDN DWLLIB.

10.Гринь М. Г., Гринь А. М. Внутренняя отчетность аграрных организаций и использование ее в системе управления // Вестник Брянской государственной сельскохозяйственной академии. 2011. № 4. С. 52-56. EDN THKTWH.

11.Золотарева Ю. П., Галанина О. В. Стратегическое планирование и программирование регионального и муниципального развития сельских территорий // Известия Международной академии аграрного образования. 2021. № 56. С. 69-72. EDN FBYYFL.

12.Галанина О. В. Big Data в планировании восстановления молочного стада КРС // Известия Международной академии аграрного образования. 2022. № 59. С. 92-95. EDN NAMWLU.

References

1. Pogonysheva D. A., Savin A.V., Seraya G. V., Tasoeva E. V. Digital technologies in personnel management in agriculture. *Bulletin of the Bryansk State Agricultural Academy.* 2021;3(85):60–66. DOI 10.52691/2500-2651-2021-85-3-60-66. EDN ZYGAOT. (In Russ.)

2. Smelik N. L. The essence of transformation of the economic system and its mechanism. Izvestiya of the St. Petersburg State Agrarian University. 2014;(37):97–103. EDN UXWKXL. (In Russ.)

3. Smagin B. I. The use of cluster analysis in the analysis of economic processes of agricultural production. *Science and Education.* 2021;4(2). EDN FPYMAK. (In Russ.)

4. Menshikova M. A., Khodyrevskaya V. N., Simakhina O. N., Shakirova D. F. Territorial analysis of the level of development of construction in Russia by the method of cluster analysis. *Economics and entrepreneurship.* 2021;2(127):456-460. DOI 10.34925/EIP.2021.127.2.086. EDN KSZPEO. (In Russ.)

5. Zambrzhitskaya E. S. Cluster analysis as a preliminary stage of break-even analysis. *Application of mathematics in economic and technical research.* 2020;1(10):109–115. EDN WXKTIH. (In Russ.)

6. Shamsutdinova T. M. Technologies of intellectual analysis of statistical data (on the example of cluster analysis of agricultural production indicators of the subjects of the Russian Federation). In: *Modern scientific and practical solutions in agriculture : Materials of the international scientific and practical conference*, Voronezh, June 06-07, 2017. Voronezh: Voronezh State Agrarian University named after Emperor Peter I; 2017:467–473. EDN ZWGBHD. (In Russ.)

7. Kosyakova L. N., Popova A. L. Innovative policy in the agricultural sphere. *British Journal for Social and Economic Research.* 2016;1(2):29–38. EDN XCPWLV.

8. Lavrova A. P. The role of personal subsidiary farms of rural population in food supply. *Izvestiya of St. Petersburg State Agrarian University.* 2015;(40):186–191. EDN UXWNVZ. (In Russ.)

9. Ulyanova N. D. Artificial intelligence in electronic document management systems. In: *Problems of energy supply, automation, informatization and environmental management in agriculture* : Collection of materials of the international scientific and technical conference, Bryansk, April 30, 2022. Bryansk: Bryansk State Agrarian University; 2022:214–220. EDN DWLLIB. (In Russ.)

10. Grin M. G., Grin A. M. Internal reporting of agricultural organizations and its use in the management system. *Bulletin of the Bryansk State Agricultural Academy.* 2011;(4):52–56. EDN THKTWH. (In Russ.)

11. Zolotareva Yu. P., Galanina O. V. Strategic planning and programming of regional and municipal development of rural territories. *Proceedings of the International Academy of Agrarian Education.* 2021;(56):69–72. EDN FBYYFL. (In Russ.)

12. Galanina O. V. Big Data in the planning of restoration of dairy cattle herd. *Proceedings of the International Academy of Agrarian Education.* 2022;(59):92–95. EDN NAMWLU. (In Russ.)

Информация об авторах

О. В. Галанина – кандидат экономических наук, доцент кафедры прикладной информатики, статистики и математики СПбГАУ.
Ю. П. Золотарёва – кандидат экономических наук, доцент кафедры земельных отношений и кадастра СПбГАУ.

Information about authors

O. V. Galanina – Cand. Sci. (Econ.), Associate Professor at the Department of Applied Informatics, Statistics and Mathematics of St. Petersburg State Agrarian University.
Ju. P. Zolotaryova – Cand. Sci. (Econ.), Associate Professor at the Department of Land Relations and Cadastre of St. Petersburg State Agrarian University.

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации.
Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Статья поступила в редакцию 05.06.2023; одобрена после рецензирования 26.06.2023; принята к публикации 27.06.2023.
The article was submitted 05.06.2023; approved after reviewing 26.06.2023; accepted for publication 27.06.2023.