# СИСТЕМНЫЙ АНАЛИЗ И УПРАВЛЕНИЕ
# SYSTEM ANALYSIS AND PROCESSING OF KNOWLEDGE

УДК 303.732.4
DOI 10.52575/2687-0932-2023-50-3-655-668

## Иерархическая кластеризация на языке R для производственно-экономических показателей пенитенциарной системы

**[1,2]Пономарев Д.С.**
[1]Филиал (г. Ижевск) федерального казенного учреждения
«Научно-исследовательский институт Федеральной службы исполнения наказаний»
Россия, 426004, Удмуртская Республика, г. Ижевск, ул. Коммунаров, 216
[2]Федеральное государственное бюджетное образовательное учреждение высшего образования
«Ижевский государственный технический университет имени М.Т. Калашникова»
Россия, 426069, Удмуртская Республика, г. Ижевск, ул. Студенческая, д. 7

**Аннотация.** Согласно официальным данным из открытых источников, структура производственного сектора уголовно-исполнительной системы Российской Федерации включает в себя 652 учреждения. В 2021 году общий объем производства товаров, выполненных работ и оказанных услуг составил 36,8 млрд рублей. На сегодняшний день в учреждениях пенитенциарной системы трудоустроено более 131 тысячи осужденных. Подразделениями ФСИН России ведется активная организационная работа по получению заказов на изготовление продукции. Таким образом, проведение исследований для производственного сектора с использованием современных научных методов является актуальным не только для уголовно-исполнительной системы, но и в целом для Российской Федерации. В свете трендов современной науки, одним из таких методов является машинное обучение без учителя, в частности – иерархический кластерный анализ. Его преимущества для поставленного вопроса являются очевидными: возможность вне зависимости от территориального уровня (что очень часто допускается в исследованиях) рассмотреть интересующие производственно-экономические показатели и возможность провести сегментацию объемов продукции с построением иерархий. Целью работы явилось проведение исследований в области машинного обучения без учителя (иерархической кластеризации) для сегментации производственно-экономических показателей пенитенциарной системы. Основным инструментом для реализации иерархической кластеризации явился язык программирования и статистической обработки — R (обработка данных проводилась в интегрированной среде разработки R-Studio). Новизной работы является: во-первых, исследование производственно-экономических показателей пенитенциарной системы с отрывом от территориального уровня (другими словами – производственно-экономические показатели были рассмотрены как часть глобальной системы, а не часть федеральных округов или территориальных органов уголовно-исполнительной системы), во-вторых, применение актуальных методов машинного обучения для сегментации и разделения на группы значений объема производства товаров, выполненных работ и оказанных услуг, связанный с привлечением осужденных к труду. Основными научными результатами в ходе проведенной работы явились: разработанный алгоритм для проведения иерархической кластеризации именно для пенитенциарной системы; сформированный ряд правил и норм по: выбору параметров, обработке данных, выборе гиперпараметров иерархической кластеризации. Кроме того, были выявлены новые зависимости для более глобального рассмотрения производственно-экономических показателей.

**Ключевые слова:** иерархическая кластеризация, производственно-экономические показатели, пенитенциарная система, язык *R*, системный анализ, разведочный анализ данных, машинное обучение

**Для цитирования:** Пономарев Д.С. 2023. Иерархическая кластеризация на языке *R* для производственно-экономических показателей пенитенциарной системы. Экономика. Информатика, 50(3): 655–668. DOI: 10.52575/2687-0932-2023-50-3-655-668

# Hierarchical Cluster Analysis in R for Production and Economic Indicators of the Penitentiary System

**[1,2] Dmitry S. Ponomarev**
[1] Branch (Izhevsk) Federal State Institution Research Institute of the Federal Penitentiary Service
216 st. Kommunarov, Izhevsk, Udmurt Republic, 426004, Russian Federation
[2] Kalashnikov Izhevsk State Technical University
7 Studencheskaya St, Izhevsk, Udmurt Republic, 426069, Russian Federation

**Abstract.** According to official information from the penitentiary system of the Russian Federation, the structure of the production sector of the penitentiary system of the Russian Federation includes 652 institutions. In 2021, the volume of production of goods and services amounted to 36.8 billion rubles. More than 131 thousand convicts are employed in the institutions of the penitentiary system. The divisions of the Federal Penitentiary Service of Russia are actively organizing work to receive orders for the manufacture of products. Thus, conducting research for the manufacturing sector using modern scientific methods is relevant not only for the penitentiary system, but also for the Russian Federation as a whole. One of these methods is machine learning hierarchical cluster analysis. Its advantages: the ability, regardless of the territories, to consider production and economic indicators of interest and the ability to segment the market with the construction of hierarchies. The purpose of this scientific study is to conduct research in the field of machine learning (hierarchical clustering) for segmenting the production and economic indicators of the penitentiary system. The main tool for implementing hierarchical clustering is the programming language and statistical processing ‑ R (data processing was carried out in the R-Studio environment). The novelty of the work is: the study of production and economic indicators of the penitentiary system, regardless of the territories and the use of relevant machine learning methods for segmentation and division into groups of values of the volume of production of goods. The main scientific results were: the developed algorithm for carrying out hierarchical clustering for the penitentiary system; formed a number of rules and norms for the choice of parameters, data processing, the choice of hyperparameters for hierarchical clustering. In addition, new dependencies were identified for a more global consideration of production and economic indicators.

**Keywords:** hierarchical cluster analysis, production and economic indicators, penitentiary system, R, systems analysis, exploratory data analysis, machine learning

**For citation:** Ponomarev D.S. 2023. Hierarchical Cluster Analysis in R for Production and Economic Indicators of the Penitentiary System. Economics. Information technologies. 50(3): 655–668 (in Russian). DOI: 10.52575/2687-0932-2023-50-3-655-668

## Введение

Иерархическая кластеризация — популярный алгоритм группировки данных, который часто может давать очень разные значения. Данный метод позволяет пользователю визуализировать эффект от разбиения разного количества кластеров. Он более чувствителен при

обнаружении отдаленных или аберрантных групп или записей. Иерархическая кластеризация также обеспечивает интуитивно понятное графическое отображение, что облегчает интерпретацию результатов [Stekh, 2006]. На сегодняшний день в исследованиях существует два основных алгоритма иерархической кластеризации: агломеративный алгоритм – кластеры образуются при помощи объединения данных т.е. «дерево» иерархической кластеризации создается от «листьев» к основанию; дивизивный алгоритм (в некоторой литературе дивизионный) – здесь кластеры образуются при помощи разделения данных (т.е. крупные кластеры разделяются на более малые), соответственно – «дерево» создается от основания к «листьям». В проведенной работе для поставленной задачи был использован агломеративный алгоритм – то есть рассматривались возможности и особенности укрупнения исследуемых массива данных [Everitt et al., 2011].

Эффективность иерархической кластеризации связана с мощностью вычислительных ресурсов, которые необходимы для нее. Поэтому, большинство методов иерархической кластеризации сосредоточены на относительно небольших наборах данных [Bruce et al., 2020]. При этом эффективность кластеризации зависит от качества данных, которые используются в ней.

Для исследования производственно-экономических показателей применение кластерного анализа поможет лучше оценить распределение ресурсов, объемы производства, сформировать правила и нормы для минимальных и максимальных показателей производственных и экономических процессов [Murtagh, Contreras, 2017]. Рассмотрение временных рядов для значений параметров позволит лучше оценить динамику изменения объемов производства, выполненных работ и оказанных услуг; оценить улучшение или ухудшение того или иного показателя. Если рассматривать кластеризацию для регионов, то построение иерархий позволит определить более доминирующие регионы по рассматриваемым показателям, оценить приоритетность. Построение иерархических структур также позволит разработать ряд правил и норм для экономических и производственных показателей.

Сбор данных по вышеуказанным вопросам в некоторых случаях может представлять определенные проблемы: учет показателей на различных друг от друга производств может отличаться; могут отличаться интервалы времени учета параметров; кроме того, следует учитывать различный род деятельности разных производств и объемы выпускаемой продукции.

Поэтому, в целях сбора данных для проведения иерархического кластерного анализа следует рассматривать ведение производственно-экономической деятельности как глобальный процесс, который подвержен единому глобальному плану и единым стандартам отчетности (как в плане учета параметров, так и в плане интервалов их мониторинга). Здесь, производства уголовно-исполнительной системы могут явиться достаточно хорошим примером для рассмотрения применения иерархического кластерного анализа. Целостность и глобальный охват отчетных форм является одним из ключевых моментов успешного сбора данных для исследования иерархического кластерного анализа, как раз эти условия и присутствуют для деятельности производств, которые находятся в ведении уголовно-исполнительной системы.

В первую очередь был разработан алгоритм для проведения иерархического кластерного анализа (рисунок 1).


Рис. 1. Алгоритм проведения иерархического кластерного анализа
Fig. 1. Algorithm for implementing hierarchical cluster analysis

В работе были рассмотрены параметры, которые связаны с одной стороны с объемом производства товаров, выполненных работ и оказанных услуг, а с другой стороны рассмотрены параметры, которые связаны с численностью осужденных, привлеченных к труду. Выборка была сформирована на основе данных отчетных форм по федеральным округам в период с 2019 по 2021 годы.

## Нормализация данных

При проведении кластерного анализа следует учитывать различность масштабов значения данных у каждого из них. Параметры с большим масштабом данных практически всегда будут доминировать не только при кластеризации данных, но и в применении методов машинного обучения без учителя в целом. Именно поэтому применение стандартизации и нормализации данных является практически всегда важным шагом. Рассмотрим процедуру более подробно (1) [Bruce et al., 2020].

$$z = \frac{x - x_{\text{ср}}}{s}, \quad (1)$$

где $s$ – стандартное отклонение, $x_{\text{ср}}$ – среднее значение, $x$ – фактическое значение. Как видно из представленной формулы, провести нормализацию значений не составляет большого труда: из фактического значения нужно вычесть среднее арифметическое и разделить на дисперсию. В результате проведения данной процедуры для каждого значения выборки будет получен новый массив данных со стандартизированными значениями. Именно данный «стандартизированный» массив и будет использоваться в дальнейшем для проведения иерархического кластерного анализа.

## Агломеративный алгоритм. Выбор расстояний между координатами

Основной алгоритм иерархической кластеризации – это агломеративный алгоритм, который итеративно объединяет похожие кластеры. Алгоритм агломерации начинается с того, что каждая запись создает свой собственный кластер из одной записи, а затем растет все

больше и больше кластеров. Первым шагом является вычисление расстояний между всеми парами записей. В работе были рассмотрены наиболее популярные и применяемые на сегодняшний день расстояния между координатами: Евклидово расстояние (2), квадрат Евклидова расстояния (3), Манхэттенское расстояние (4) [Everitt et al., 2011].

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \quad (2)$$

$$d(x,y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2 \quad (3)$$

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_k - y_k| \quad (4)$$

Проблема применения Эвклидова расстояния в том, что если есть большая разница в координатах, то при возведении в квадрат эта разница будет увеличиваться. Таким образом, следует для начала оценить распределение исследуемых точек и при большом разбросе координат рациональнее будет использовать Манхэттенское расстояние. В нашем же случае наблюдается равномерное распределение точек, а значит применение Эвклидова расстояния будет корректным.

**Выбор расстояния между кластерами**

Выбор расстояния между кластерами является достаточно важным параметром. В зависимости от выбранного межкластерного расстояния может зависеть конечный исход кластеризации. В некоторых случаях неправильно выбранное расстояние может привести к неправильно интерпретируемым выводам. Пожалуй, одним из самых «беспроигрышных» тактик в проведении кластерного анализа является попробовать все наиболее известные и используемые межкластерные расстояния на исследуемом наборе данных, а затем выбрать наиболее подходящее решение.

При использовании команды *hclust* в *R* [Kabacoff, 2011] можно выбирать и расстояние между кластерами (используется выбор параметров при помощи «*method*»). Наиболее популярные и применяемые расстояния между кластерами, которые рассматривались при проведении расчетов, обозначены далее.

Среднее невзвешенное расстояние (в некоторой литературе метод средней связи). Рассматривался только невзвешенный метод, где расстояние между кластерами приравнивается среднему расстоянию между их элементами (5):

$$\frac{1}{|X||Y|} \sum_{x \in X} \sum_{y \in Y} d(x,y) \quad (5)$$

где: *d(x, y)* – расстояние между элементами x и y, которые принадлежат кластерам *X* и *Y*; |*X*| и |*Y*| – мощности кластеров. На языке *R* реализуется следующим образом (где *data* – исследуемое нормализованное поле значений): *Clust <- hclust(d=data, method = "average")*.

Центроидный метод, расстояние между кластерами приравнивается расстоянию между их центрами масс (6):

$$\|C_X - C_Y\|, \quad (6)$$

где $C_x$ и $C_y$ – центроиды кластеров *X* и *Y*. Невзвешенный центроидный метод в *R* обозначен как «*centroid*», реализуется следующей командой: *Clust <- hclust(d=data, method = "centroid")*.

Метод «дальнего соседа» или метод полной связи, расстояние между кластерами приравнивается максимальному расстоянию между двумя точками этих кластеров (7):

$$max\{d(x,y): x \in X, y \in Y\}, \quad (7)$$

в *R* обозначен как «*complete*», реализация: *Clust <- hclust(d=data, method = "complete")*.

Метод «ближайшего соседа» или метод одиночной связи, в котором расстояние между кластерами приравнивается к минимальному расстоянию между двумя элементами из данных кластеров (8):

$$min\{d(x,y): x \in X, y \in Y\}, \quad (8)$$

в *R* реализуется как «single»: *Clust <- hclust(d=data, method = "single")*.

Ward's method (в некоторой литературе «Метод Варда», «Метод Уорда», далее – метод Уорда), суть метода заключается в объединении данных, при котором различие между двумя группами (кластерами) измеряется значением, на которое слияние этих двух кластеров увеличило бы среднее квадратическое расстояние от точки до ее центроида (9) [Ward, 1969].

$$d(X,Y) = \sum_{x \in X, y \in Y} \frac{d^2(x,y)}{|X|+|Y|} - \sum_{x,y \in X} \frac{d^2(x,y)}{|X|} - \sum_{x,y \in Y} \frac{d^2(x,y)}{|Y|}, \quad (9)$$

Остановимся на данном методе более подробно. На сегодняшний день существует два основных алгоритма реализации данного метода. В первом случае на *R* метод реализуется при помощи команды «*ward.D*» и соответствует алгоритму, представленному в работах Wishart D. [Wishart, 1969] и Murtagh F. [Murtag, 1983], во втором случае метод осуществляется при помощи команды «*ward.D2*» и реализует алгоритм, который представлен в более новых работах, а именно: Kaufman L., Rousseeuw P. [Kaufman, Rousseeuw, 1990] и Legendre P. [Legendre, 2012].

На сегодняшний день не существует оптимального метода для проведения кластерного анализа и даже результативность вышеуказанных методов может очень сильно зависеть от исследуемых данных, поэтому в работе были рассмотрены и использованы все из перечисленных методов (5-9) и расстояний (2-4), проведена различная группировка из пар «метод-расстояние», при этом были получены достаточно разные результаты. Количество кластеров также рассматривалось разное: от двух до восьми. Восемь (максимальное число кластеров) было выбрано по количеству рассмотренных федеральных округов, а один кластер не рассматривался из-за бессмысленности в этом случае проведения кластерного анализа).

## Построение дендрограмм

Дендрограмма при проведении кластерного анализа обычно используется для представления структуры дерева. Она представляет иерархическую структуру данных, используя вертикальную или горизонтальную диаграмму [Everitt et al., 2011]. В *R* построение дендрограммы возможно при помощи команды *plot*, например: *plot*(*hcl*) [Kabacoff, 2011]. При этом так называемые «листья» будут соответствовать исследуемым параметрам, а по длине «ветви» можно визуально определить степень различия между кластерами. Другими словами, чем длиннее линии иерархий в дендрограмме, тем большее различие между кластеризованными параметрами. Основное преимущество данного подхода заключается в простой интерпретируемости полученных результатов, возможности быстрого определения корректировки параметров иерархической кластеризации.

**Проведение кластеризации**

Реализация иерархической кластеризации проводилась при помощи языка *R* [Bruce et al., 2020; Metloff, 2019], для исследования данных, как упоминалось ранее, были применены все рассмотренные в данной статье гиперпараметры: расстояние между координатами (2,4) и расстояние между кластерами (5-9); была рассмотрена возможность разбиения данных от 2 до 8 кластеров. Наиболее удачным вариантом оказалась реализация применения Манхэттенского расстояния и метода Уорда с разбиением на 4 кластера (при этом стоит отметить, что разбиение на 2 кластера также является актуальным решением).

Пример кода с применением нормализации, разбиением на кластеры и построением дендрограммы представлен на рисунке 2 (разделение на четыре кластера, применение Манхэттенского расстояния и метода Уорда (причем метод Уорда рассматривался согласно алгоритмам, представленным в работах Wishart D. [Wishart, 1969] и Murtagh F. [Murtagh, 1983], а не Kaufman L., Rousseeuw P. [Kaufman, Rousseeuw, 1990] и Legendre P. [Legendre, 2012])).

```
df.scaled <- scale(df[,c(3,5)])
rownames(df.scaled) <- c(data$X1)
res.dist <- dist(x=df.scaled, method = "manhattan")
x <- as.matrix(res.dist)
round(x,digits = 3)
res.hc <-hclust(d=res.dist, method = "ward.D")
fviz_dend(res.hc, cex=0.8, lwd=0.8, k=4, rect = TRUE,
          k_colors = c("red", "blue", "grey"),
          rect_border = "jco",
          type = "rectangle",
          repel=TRUE,
          rect_fill = TRUE, horiz=TRUE)
```

Рис. 2. Реализация иерархической кластеризации и разработка дендрограммы при помощи языка R (выбрано: Манхэттенское расстояние между координатами и метод Уорда для межкластерного расстояния)
Fig. 2. Implementing hierarchical clustering and designing a dendrogram with R (used: Manhattan distance and Ward's method)

Исходя из правил иерархической кластеризации, которые были упомянуты ранее, было рассмотрено разбиение на 2 и 4 кластера. В рамках проведенного исследования более подробно рассмотрено распределение данных на 4 кластера (построение диаграмм методом «boxplot», оценка средних значений и их отклонения от медианы, оценка размаха основного массива данных, который был принят соответственно 25-му и 75-му процентилю). При этом одно из основных правил при проведении кластеризации было сформулировано так: федеральные округа при первоначальном объединении не должны перемешиваться друг с другом, т.е. первые итерации объединения в кластеры проходили бы только в рамках временного ряда только одного федерального округа. Таким образом, можно было бы достичь более целостной и интерпретируемой структуры иерархического «дерева». В результате была разработана дендрограмма (рисунок 3), которая отражает иерархическое объединение в кластеры (была рассмотрена кластеризация по всем исследуемым параметрам одновременно с учетом проведенной нормализации каждого из них).

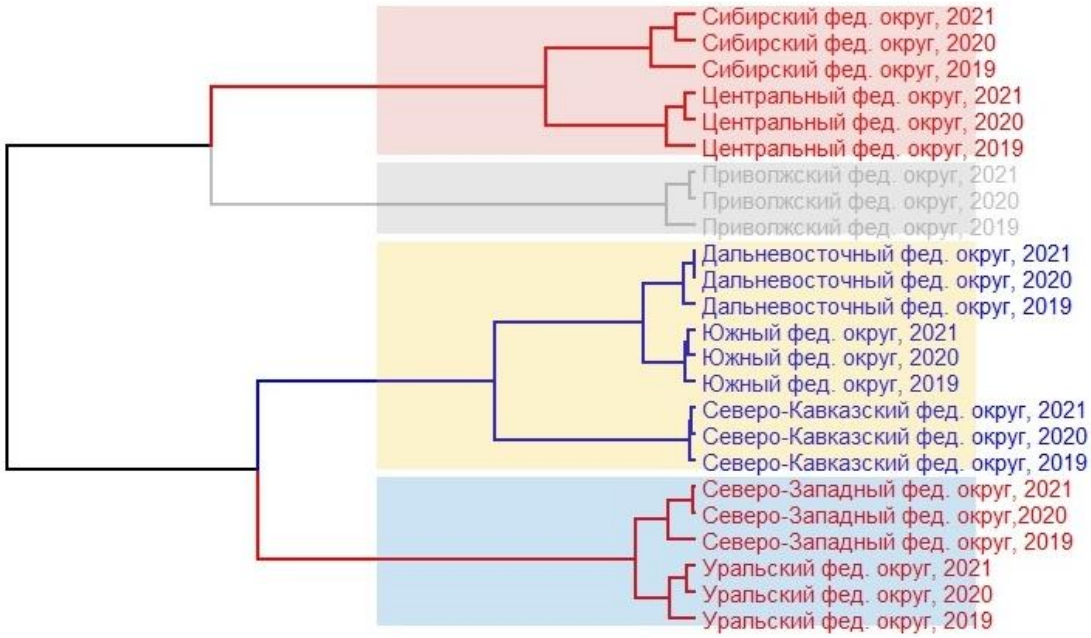


Рис. 3. Результат проведения иерархического кластерного анализа
для совокупности исследуемых параметров в виде дендрограммы
Fig. 3. The result of the hierarchical cluster analysis, the development of the dendrogram

Собранные данные в период с 2019 по 2021 годы в первой итерации были кластеризованы по федеральным округам. Причем была выделена интересная закономерность: поначалу происходила кластеризация данных за 2020 и 2021 годы и только затем «присоединялись» данные за 2019 год. Данная закономерность была выделена для всех федеральных округов. Здесь можно сделать вывод, что значения данных, которые отражают производственно-экономические показатели за 2019 год, могли достаточно заметно отличаться от данных в 2020 и 2021 году. Рассмотренная закономерность представляет достаточно большой интерес и является хорошим вопросом для проведения дальнейших исследований.

В последующих итерациях происходило объединение федеральных округов в все большие и большие кластеры. Наиболее значимыми при этом явилось разбиение на 2 и 4 кластера. Установленное ранее правило о том, что на начальных стадиях кластерного анализа федеральные округа не должны были «перемешиваться» между собой также было соблюдено.

Для интерпретации результатов проведенного кластерного анализа хорошим инструментом могут также явиться методы разведочного анализа данных. Проведение разведочного анализа данных (*exploratory data analysis*, далее – *EDA*) [Tukey, 1962] с учетом полученных кластеров поможет лучше сформировать выводы проведенной работы, определить ряд возможных правил и зависимостей между параметрами.

## Интерпретация результатов кластерного анализа при помощи методов разведочного анализа данных

В проводимой работе были применены следующие методы *EDA* [Tukey, 1977]: оценка центральных положений, распределение данных по процентилям, определение основного массива данных. Причем *EDA* проводился исходя из учета сформированных четырех кластеров. Одним из наиболее подходящих графических представлений получен-

ных результатов может явиться диаграмма, построенная методом «*boxplot*» (или коробчатая диаграмма) [Hintze, 1998]. На рисунке 4 представлена разработка графика методом «*boxplot*» на языке *R*.

```
BOXPLOT_X3 <- ggplot(data=data_boxplot)+
              geom_boxplot(aes(x=X1, y=X3),
                           fill="grey",
                           colour="black",
                           alpha=0.3,
                           notch = TRUE,
                           outlier.color = "red",
                           outlier.size = "10")+
              stat_summary(aes(x=X1, y=X3),
                           fun=mean,
                           geom="point",
                           shape=3,
                           size=8,
                           color="black",
                           alpha=1)
```

Рис. 4. Пример разработки графика методом «boxplot» на языке R
Fig. 4. An example of developing a boxplot in R

На примере параметра, который отражает объемы производства товаров, выполненных работ и оказанных услуг, была получена диаграмма для исследования распределений и интерпретации иерархического кластерного анализа при помощи метода «*boxplot*» (рисунок 5). Цветом представлено разделение на кластеры.

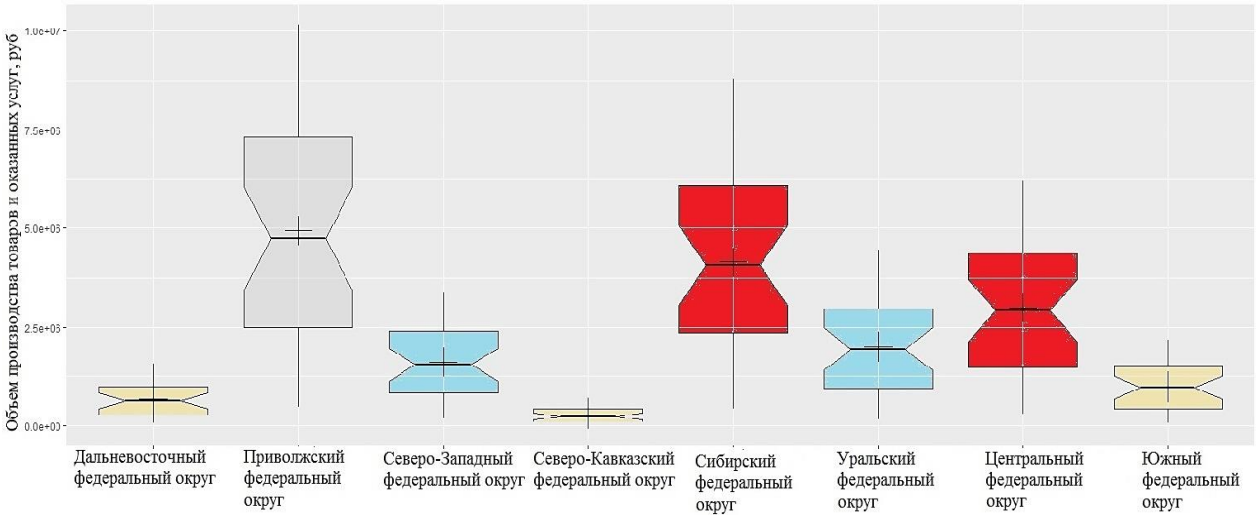


Рис. 5. Исследование распределений для интерпретации результатов иерархического кластерного анализа при помощи метода «boxplot»
Fig. 5. Interpretation of the results of hierarchical cluster analysis using a boxplot

Кроме того, исследуемые метрики рассматриваемых параметров из рисунка 5 можно представить и в виде таблицы (где: *Min* – минимальное значение, $Q_{25}$ и $Q_{75}$ – соответственно 25-й и 75-й процентиль исследуемых данных (Hyndman R.J., Yanan Fan, 1966), *Median* – медиана, *Mean* – среднее арифметическое, *Max* – максимальное значение).

Таблица
Table

Основные метрики распределения данных и центральных положений
с учетом проведенного кластерного анализа
The main metrics of the distribution of data and central positions,
taking into account the performed cluster analysis

| № кластера | Фед. округ | Min | $Q_{25}$ | Median | Mean | $Q_{75}$ | Max |
|---|---|---|---|---|---|---|---|
| 1 | Сибирский фед. округ | 7187493 | 7628825 | 8070157 | 8017197 | 8432048 | 8793940 |
| | Центральный фед. округ | 5353162 | 5671065 | 5988969 | 5844630 | 6090365 | 6191761 |
| 2 | Приволжский фед. округ | 9302259 | 9711546 | 10120832 | 9858558 | 10136708 | 10152584 |
| 3 | Дальневосточный фед. округ | 1308492 | 1438144 | 1567796 | 1485952 | 1574683 | 1581569 |
| | Южный фед. округ | 1944345 | 2008005 | 2071664 | 2071664 | 2135324 | 2198984 |
| | Северо-Кавказский фед. округ | 617579 | 632378 | 647178 | 663845 | 686979 | 726779 |
| 4 | Северо-Западный фед. округ | 2799142 | 3069550 | 3339958 | 3174473 | 3362139 | 3384320 |
| | Уральский фед. округ | 3739593 | 3973469 | 4207345 | 4128008 | 4322216 | 4437087 |

Исходя из представленной таблицы, можно оценить границы распределения для каждого кластера как по минимальным и максимальным значениям, так и по процентилям (что более предпочтительно). Также можно отметить незначительную (в рамках исследуемого масштаба данных) разницу между средним арифметически и медианным значением, что говорит о равномерном распределении данных.

Исследуя график на рисунке 5 и представленную соответствующую ему таблицу, можно сказать, что Дальневосточный, Южный и Северо-Кавказский федеральные округа, объединились в один кластер, который соответствует наименьшим порогам для параметра, который отражает объемы производства товаров, выполненных работ и оказанных услуг. Приволжский федеральный округ, наоборот, занимает в данном случае доминирующую позицию, настолько сильную, что был выделен в целый отдельный кластер.

Таким образом, результаты проведения кластерного анализа были получены достаточно адекватные и интерпретируемые. А применение *EDA* помогло еще более глубоко понять причины распределения федеральных округов в иерархии. Далее рассмотрим возможность применения кластеризации для каждого отдельно взятого параметра.

## Приоритет группового исследования параметров

Рассмотрим применение иерархического кластерного анализа для отдельно взятых параметров. В первую очередь была проведена кластеризация для параметра, который отражает значения объема производства товаров, выполненных работ и оказанных услуг, связанный с привлечением осужденных к труду (рисунок 6).

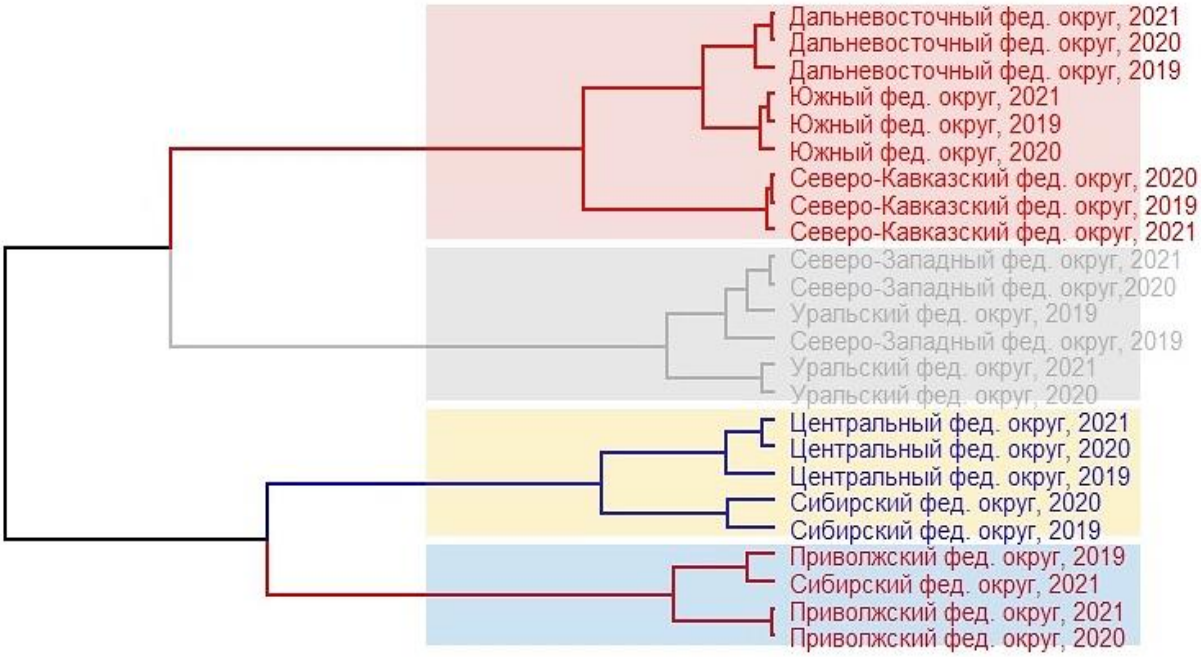


Рис. 6. Результат проведения иерархического кластерного анализа
в виде дендрограммы для параметра, который отражает значения объема производства товаров, выполненных работ и оказанных услуг, связанный с привлечением осужденных к труду
Fig. 6. The result of the hierarchical cluster analysis for the volume of production of goods, works and services, associated with the involvement of convicts in labor

Как видно из рисунка 6, при проведении кластерного анализа наблюдается «перемешивание» между регионами. Результат был получен следующий: Уральский федеральный округ на первых же этапах кластеризации стал объединятсья с Северо-Западным федеральным округом, Сибирский федеральный округ попал сразу в два кластера.

При разделении на 2 кластера для исследуемого параметра было установлено, что объединились округа: первый кластер – Дальневосточный федеральный округ, Южный федеральный округ, Северо-Кавказский федеральный округ, Северо-Западный федеральный округ, Уральский федеральный округ; второй кластер – Центральный федеральный округ, Сибирский федеральный округ, Приволжский федеральный округ.

При разделении на 4 кластера было установлено, что объединились округа: первый кластер – Дальневосточный федеральный округ, Южный федеральный округ, Северо-Кавказский федеральный округ; второй кластер – Северо-Западный федеральный округ, Уральский федеральный округ; третий кластер – Центральный федеральный округ, Сибирский федеральный округ; четвертый кластер – Сибирский федеральный округ, Приволжский федеральный округ.

Далее аналогичным образом была проведена иерархическая кластеризация для параметра, который отражает значения среднесписочной численности осужденных, привлеченных к труду. Результат представлен на рисунке 7.

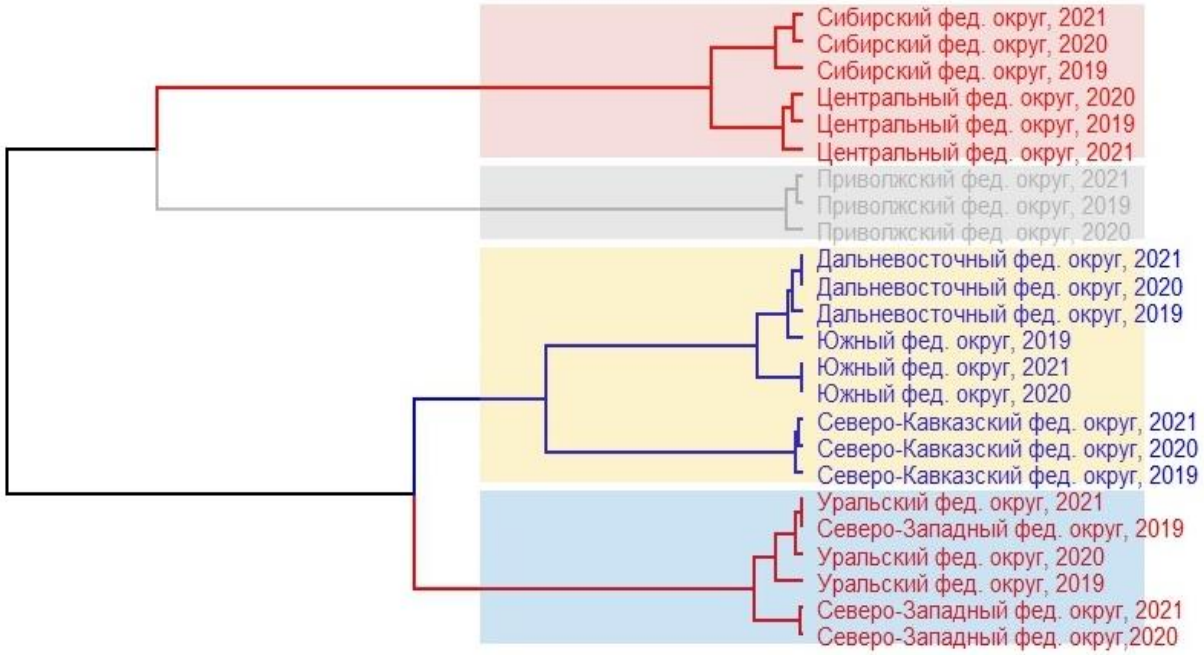


Рис. 7. Результат проведения иерархического кластерного анализа в виде дендрограммы для параметра, который отражает значения среднесписочной численности осужденных, привлеченных к труду
Fig. 7. The result of a hierarchical cluster analysis to monitor the average number of convicts involved in labor

Здесь, при проведении кластерного анализа, были получены следующие результаты: Уральский федеральный округ «перемешался» с Северо-западным федеральным округом; Южный федеральный округ объединился с Дальневосточнм федеральным округом.

Если рассматривать разделение на 2 кластера: первый кластер – Сибирский федеральный округ, Центральный федеральный округ, Приволжский федеральный округ; второй кластер – Дальневосточный федеральный округ, Южный федеральный округ, Северо-Кавказский федеральный округ, Уральский федеральный округ, Северо-Западный федеральный округ.

Если же рассматривать разделение на 4 кластера, то результаты были получены следующие: первый кластер – Сибирский федеральный округ, Центральный федеральный округ; второй кластер – здесь был только выделен один федеральный округ – Приволжский; третий кластер – Дальневосточный федеральный округ, Южный федеральный округ, Северо-Кавказский федеральный округ; четвертый кластер – Уральский федеральный округ, Северо-Западный федеральный округ.

Таким образом, при проведении кластерного анализа только для одного исследуемого параметра являются достаточно проблематичными сразу несколько позиций: первое – при первых же итерациях кластеризации федеральные округа начинают «перемешиваться» между собой; второе – некоторые федеральные округа попали сразу в несколько кластеров (например, при разбиении на четыре кластера). Все это приводит к тому, что полученные результаты практически невозможно интерпретировать, невозможно охарактеризовать каждый федеральный округ в рамках исследуемых параметров. Представляется также затруднительным проведение разведывательного анализа данных и определение доминирующих федеральных округов в рамках исследуемых параметров. Резюмируя сказанное в данном разделе, наилучшим решением в плане интерпретации и получении результатов при проведении иерархического кластерного анализа является выбор и исследование сразу нескольких интересующих параметров (значения которых следует предварительно нормализировать).

## Заключение

В работе были рассмотрены возможности применения иерархического кластерного анализа для изучения производственно-экономических параметров на примере данных уголовно-исполнительной системы. Были рассмотрены данные федеральных округов с 2019 по 2021 годы и установлено, что рассмотренный в статье метод машинного обучения является достаточно эффективным инструментом. Одними из значительных преимуществ иерархического кластерного анализа являются: прекрасная информативная визуализация при помощи построения дендрограмм; возможность оценить «доминирующие» федеральные округа в рамках исследуемых показателей.

Одними из открытых вопросов кластерного анализа считаются: выбор расстояний между координатами; выбор наиболее подходящего расстояния между кластерами; определение числа кластеров. На сегодняшний день единственным достоверным способом получить ответы на данные вопросы является применить данные гиперпараметры на практике. Кроме того, была также установлена важность применения предобработки данных перед проведением кластерного анализа – решение проблемы пропущенных данных и их нормализация могут сыграть в дальнейшем важную роль в интерпретации результатов.

Особым пунктом следует отметить эффективность применения *EDA* вместе с иерархической кластеризацией: в некоторых случаях методы *EDA* могут явиться наилучшим решением для интерпретации и объяснения результатов кластерного анализа.

В проведенной работе было установлено, что для интерпретируемых результатов иерархического кластерного анализа предпочтительнее рассматривать исследуемые параметры комплексно, а не по отдельности. В работе было отмечено, что данные за 2019 год для всех федеральных округов выделяются в отдельную «ветку» иерархии, что дает повод для проведения дополнительных исследований (здесь перспективным может явиться исследование временных рядов и разработка прогностических численных моделей).

## References

Brian S.E., Sabine Landau, Morven Leese, Daniel Stah. 2011. Cluster Analysis. Wiley, 5th Edition. 71-110.

Bruce P., Bruce A., Gedeck P. 2020. Practical statistics for Data Scientists. O'Reilly. 363 p.

Hintze J.L. 1998. Violin Plots: A Box Plot – Density Trace Synergism. The American Statistician. 2(52): 181–84.

Hyndman R.J., Yanan Fan. 1966. Sample Quantiles in Statistical Packages. American Statistician. 4(50): 361–65.

Kabacoff R.I. 2011. R in action. Manning Publications Co. 451 p.

Kaufman L., Rousseeuw P. 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley. 335 p.

Legendre P. 2012. Numerical ecology. 3rd English ed. - Amsterdam: Elsevier. 990 p.

Metloff N. 2019. The art of R pogramming. Starch Press. 416 p.

Murtagh F. 1983. A survey of recent advances in hierarchical clustering algorithms. The Computer Journal. №26. 354–359.

Murtagh F., Contreras P. 2017. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 7(6): 1219.

Stekh Y., Kernytskyy A., Lobur M. 2006. Hierarchical clustering algorithms for large datasets. Modern Problems of Radio Engineering, Telecommunications and Computer Science Proceedings of International Conference, TCSET 2006. 388-390.

Tukey J.W. 1962. The Future of Data Analysis. The Annals of Mathematical Statistics. № 1. 1–67.

Tukey J.W. 1977. Exploratory Data Analysis. Reading, Mass.: Addison Wesley. 688 p.

Ward J.H. 1963. Hierarchical grouping to optimize an objective function. J. of the American Statistical Association. 236 p.

Wishart D. 1969. An algorithm for hierachical classifications, Biometrics 25, 165–170.

**Конфликт интересов:** о потенциальном конфликте интересов не сообщалось.
**Conflict of interest:** o potential conflict of interest related to this article was reported.

## ИНФОРМАЦИЯ ОБ АВТОРЕ

**Пономарев Дмитрий Сергеевич**, кандидат технических наук, ведущий научный сотрудник филиала (г. Ижевск) федерального казенного учреждения «Научно-исследовательский институт Федеральной службы исполнения наказаний»; доцент кафедры «Водоснабжение и водоподготовка», Ижевский государственный технический университет имени М.Т. Калашникова, г. Ижевск, Россия

## INFORMATION ABOUT THE AUTHOR

**Dmitry S. Ponomarev**, Candidate of Technical Sciences leading researcher of the branch (Izhevsk) of the Federal State Institution Research Institute of the Federal Penitentiary Service of Russian Federation; Associate Prof. of Kalashnikov Izhevsk State Technical University, Izhevsk, Russian Federation