САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Направление: 01.03.02 «Прикладная математика и информатика»
ООП: Прикладная математика, фундаментальная информатика и программирование

ОТЧЕТ ПО УЧЕБНОЙ ПРАКТИКЕ (НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ)

Тема задания: Статистический анализ с применением методов кластерного анализа для исследования социально-экономических данных.

Выполнил: Дашкова Олеся Вячеславовна 20.Б02

Научный руководитель: Буре В.М., доктор технических наук

Руководитель практики: Раевская А.П., доцент

Санкт-Петербург

2023

Введение.

Задача кластеризации достаточно популярна в различных сферах статистического анализа и является одной из основных в сфере обучения без учителя в машинном обучении. Существует множество различных методов и алгоритмов. Некоторые из них подходят для реализации некоторого разбиения, но при этом могут быть непригодными для другого. Если обратиться к определению - задача кластерного анализа заключается в том, чтобы на основании данных, каких - либо наблюдаемых показателей или характеристик, разбить множество объектов на k кластеров (подмножеств) так, чтобы каждый объект принадлежал одному и только одному подмножеству разбиения и чтобы объекты принадлежащие одному и тому же кластеру, были сходными, в то время как объекты, принадлежащие разным кластерам были разнородными.[1] Соответственно, для основательного анализа методов кластеризации необходим подходящий для исследования набор данных и меры качества.

В настоящее время всё чаще поднимается вопрос о неравенстве в развитии регионов РФ. Об этом говорит приоритетное направление министерства экономического развития РФ, а именно: реализация различных проектов и задач в области социально-экономического развития субъектов Российской Федерации.[8] Таким образом, регионы России могут быть классифицированы в зависимости от уровня социально-экономического развития.

Объектами в исследовании выступили все субъекты РФ, за исключением тех, по которым отсутствуют значения некоторых признаков. В качестве вектора характеристик были рассмотрены экономические, инновационные и научные показатели. Основные данные были взяты мной с сайта Федеральной службы государственной статистики за 2020 год.[9] Реализация практической составляющей проведена с использованием языка R.

Обзор литературы/публикаций.

В качестве основной информационной базы для моего исследования выступила статья о результатах кластеризации регионов России.[6] Ориентируясь на данный источник был осуществлён выбор показателей для задачи кластеризации.

Основанием для определения экономических показателей послужила статья интернет-журнала “Науковедение” об оценке экономического потенциала региона.[10] Были рассмотрены производственный и ресурсный потенциал, эластичность спроса экономики, величина и динамика валового продукта на душу населения, трудовой потенциал, инвестиционная привлекательность и инвестиционная составляющая. Для исследования использовались три показателя: валовый региональный продукт на душу населения, инвестиции в основной капитал на душу населения, объем основных фондов на душу населения.

Для рассмотрения инновационных признаков развития региона была использована статья об основных факторах инновационного развития регионов.[11] В ней выделено множество различных признаков, но для исследования использовались только три, а именно: коэффициент изобретательской активности без учета полезных моделей, внутренние затраты на исследования и разработки в расчете на одного работника и объем отгруженных инновационных товаров и услуг на душу населения. Данные о коэффициенте изобретательской активности взяты со статистического сборника аналитического центра ФИПС.[12]

Реализация практической части основывалась на алгоритмах и методах, описываемых в электронной книге об алгоритмах Data Mining с использованием R.[13] Также была использована книга о методах прикладной статистики в R.[4]

Постановка задачи.

Первостепенной задачей является подбор вектора характеристик для рассматриваемых объектов (субъектов РФ), которые максимально точно описывают социально-экономическое развитие региона. Затем предобработка готового массива данных.

Следующий важный шаг - рассмотрение методов кластеризации для дальнейшего применения к выбранным объектам. Для оценки качества разбиений также следует рассмотреть возможные меры качества.

Основной задачей является применение рассмотренных методов кластеризации к полученному набору объектов и их признаков. Затем, вычисление меры качества для каждого из разбиений. На основе полученных данных сделать попытку выбора того метода, который даёт наиболее высокую оценку.

Модели, методы, алгоритмы.

Условно методы кластеризации можно разбить на две группы - иерархические и неиерархические. Первые заключаются в построении так называемого дерева вложенных кластеров, основываясь на различных способах вычисления расстояния между ними. Вторые же более удобны в случае большого количества объектов, подлежащих объединению, и строят одно разбиение на кластеры.

Метод k-средних(k-means). Самый популярный метод кластеризации, благодаря своей простоте в применении. Главной проблемой при его использовании является необходимость выбора гиперпараметра - числа кластеров.
Алгоритм k-means:

1. Случайным образом выбираются позиции центроидов (геометрический центр кластера). Для каждого объекта определяется, к какому из центроидов он ближе всего - в тот кластер и распределяем рассматриваемый объект.
2. Перемещаем центроиды в геометрический центр всех объектов для каждого кластера.
3. Снова для каждого объекта определим, к какому центроиду он ближе всего и обновим их принадлежность к кластерам.
4. Повторяем шаги 1-3 пока на очередном шаге не будет происходить обновление принадлежности объектов к некоторому кластеру.

Методы для определения оптимального числа кластеров:

1. Метод “локтя” (Elbow method).

Он рассматривает характер изменения общей внутригрупповой суммы квадратов (total within-cluster sum of squares) с увеличением числа кластеров k. На некотором шаге происходит замедление снижения дисперрсии - на графике это происходит в точке, называемой “локтем”.

2. Метод “силуэтов” (Silhouette method).

Коэффициент “силуэт” вычисляется с помощью среднего внутрикластерного расстояния (a) и среднего расстояния от центроида до объектов ближайшего кластера (b). Формула для вычисления имеет вид: $\frac{b-a}{\max(a,b)}$.

Методы иерархической кластеризации. Идея заключается в измерении расстояния между всеми объектами (построение матрицы расстояний), затем поочередно происходит группировка объектов в кластеры до определенного момента. Какие объекты группируются - определяется различными методами. Результатом является дендограмма - дерево кластеров.

1. Метод одиночной связи / ближайшего соседа (single linkage):
1) Количество кластеров приравнивается к количеству объектов.
2) Строится матрица расстояний для каждого объекта.
3) Две точки с минимальным расстоянием объединяются в один кластер - таким образом общее число кластеров уменьшилось на единицу.
4) Высчитывается новая матрица расстояний, но уже с центроидом образованного на третьем шаге кластера.
5) Процесс продолжается до тех пор, пока не будет получен один кластер.

В результате исследователь сам выбирает, как будут кластеризованы данные.

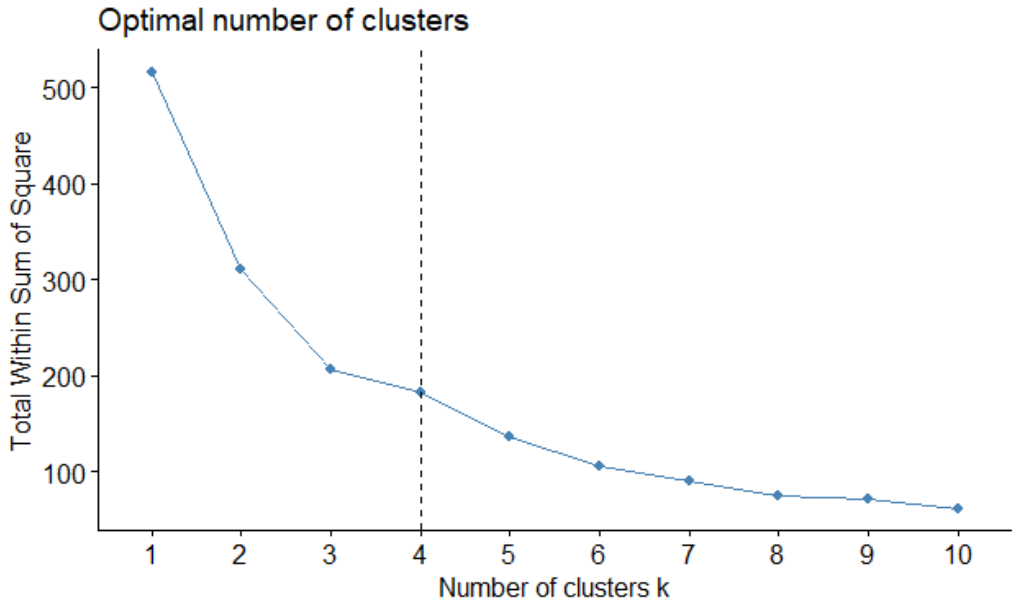
2. Метод полной связи / дальнего соседа (complete linkage).

Отличается от метода одиночной связи тем, что расстояние определяется как максимум из множества расстояний между элементом первого кластера и элементом второго кластера.

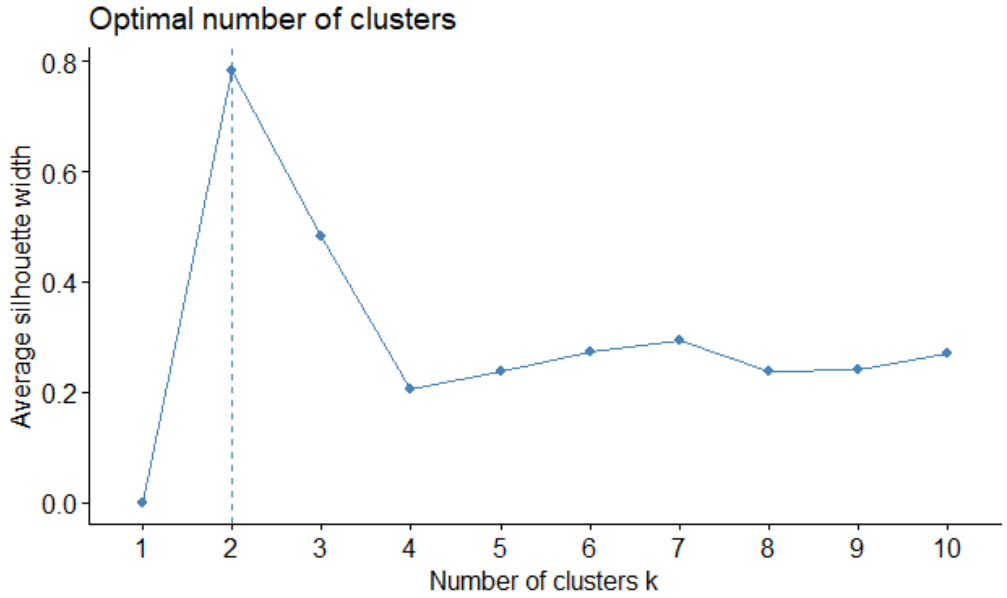
Предварительные результаты.

Проведена предобработка исходного массива данных. Пропущенные значения заменены средним значением по признаку с целью предотвратить наличие сильных выбросов. Все характеристики являются количественными, поэтому применяем нормализацию.

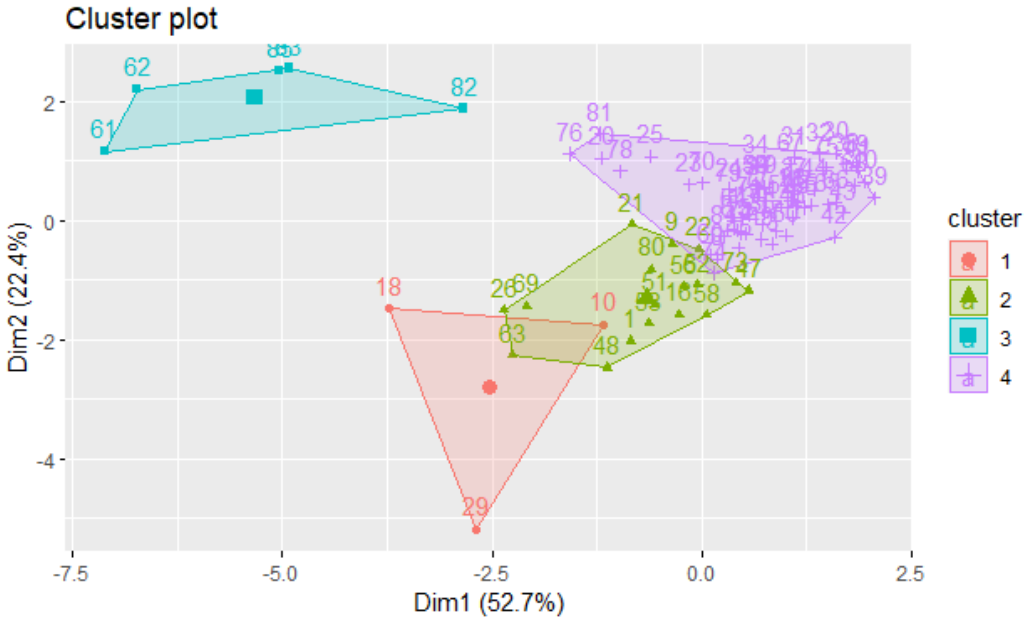
Метод “локтя” определения оптимального числа кластеров дает неоднозначный результат. Но предварительно количество кластеров равно четырём.



Поэтому решено так же реализовать метод силуэтов для точного определения числа кластеров. Метод силуэтов наилучшим выбрал разбиение на два кластера. Можем заметить, что в условиях неопределенности различные алгоритмы могут порождать конкурирующие решения.

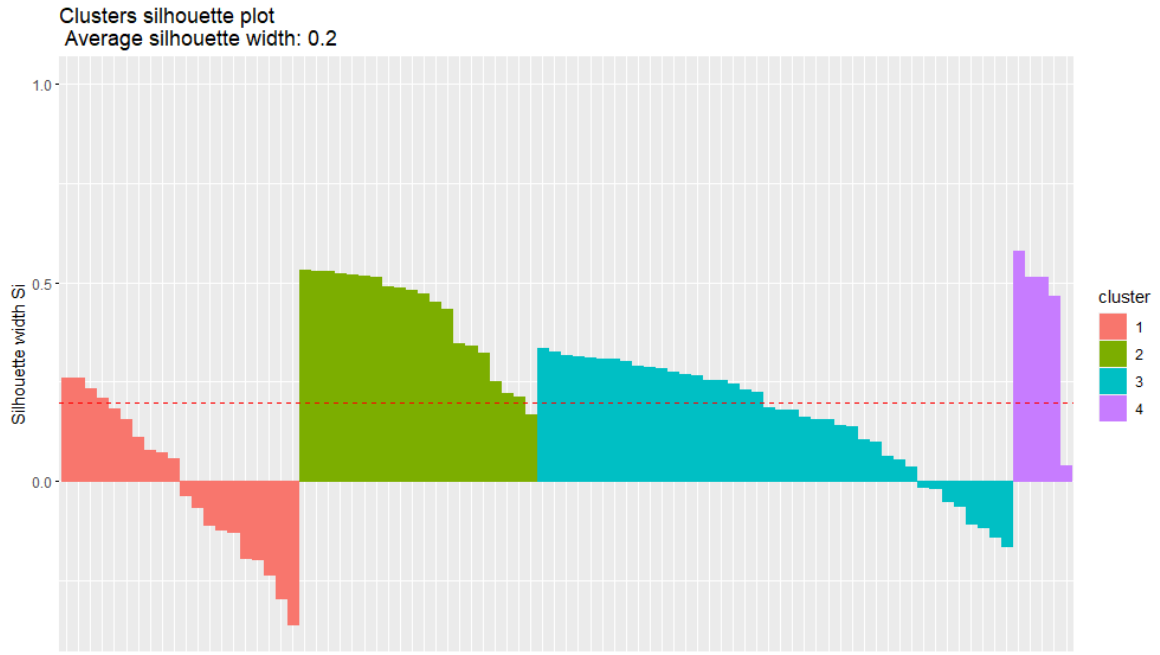


Итак, выберем четыре кластера в качестве оптимального количества. После многократной кластеризации методом k-means, в большинстве случаев получаем данный результат.



Первый кластер: Московская область, город Москва, Санкт-Петербург.
Второй кластер: Белгородская область, Липецкая область, Тульская область, Архангельская область включая Ненецкий автономный округ, Архангельская область без Ненецкого автономного округа, Мурманская область, Республика Татарстан, Республика Мордовия, Ульяновская область, Самарская область, Пермский край, Нижегородская область, Кировская область, Хабаровский край, Омская область, Тюменская область без автономных округов, Красноярский край.
Третий кластер: Ханты-Мансийский автономный округ, Тюменская область включая автономные округ, Магаданская область, Сахалинская область, Чукотский автономный округ.
Четвертый кластер: оставшиеся регионы вошли в данный кластер.

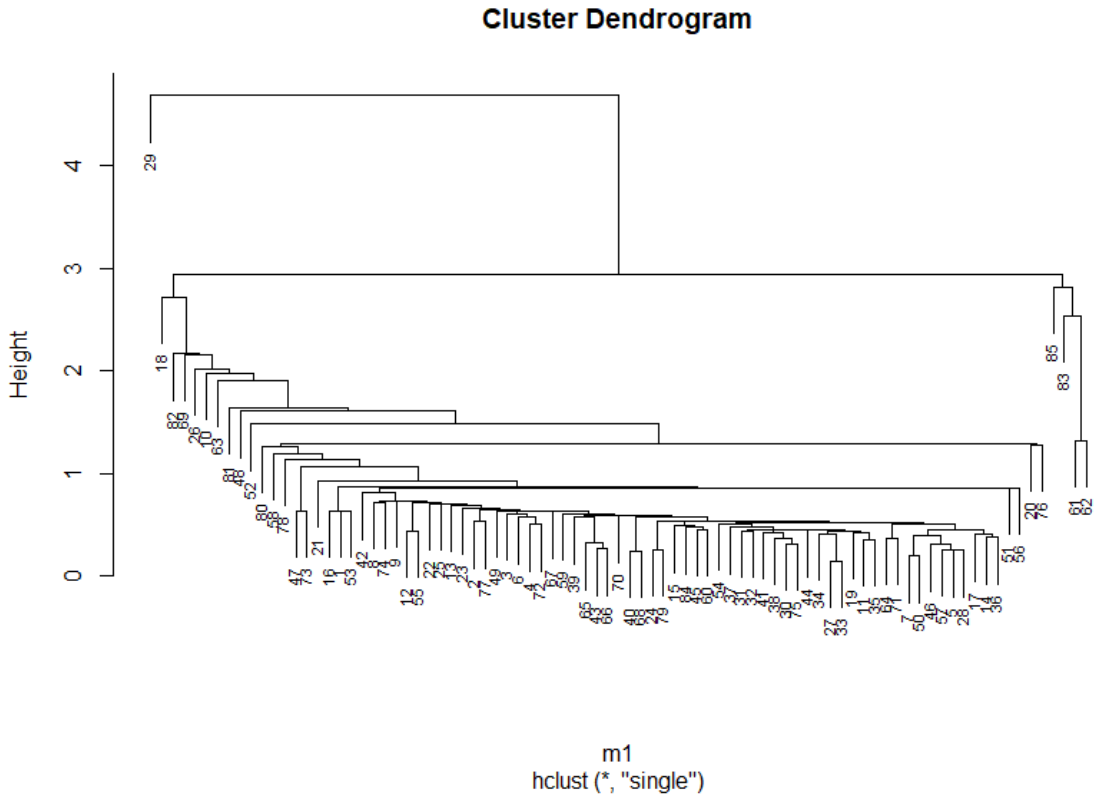
Для оценки качества разбиения с помощью метода k-means воспользуемся также диаграммой силуэтов. Она показывает, что для двух кластеров из четырёх объекты распределены не очень качественно (значение коэффициента силуэта приближается к -1) .



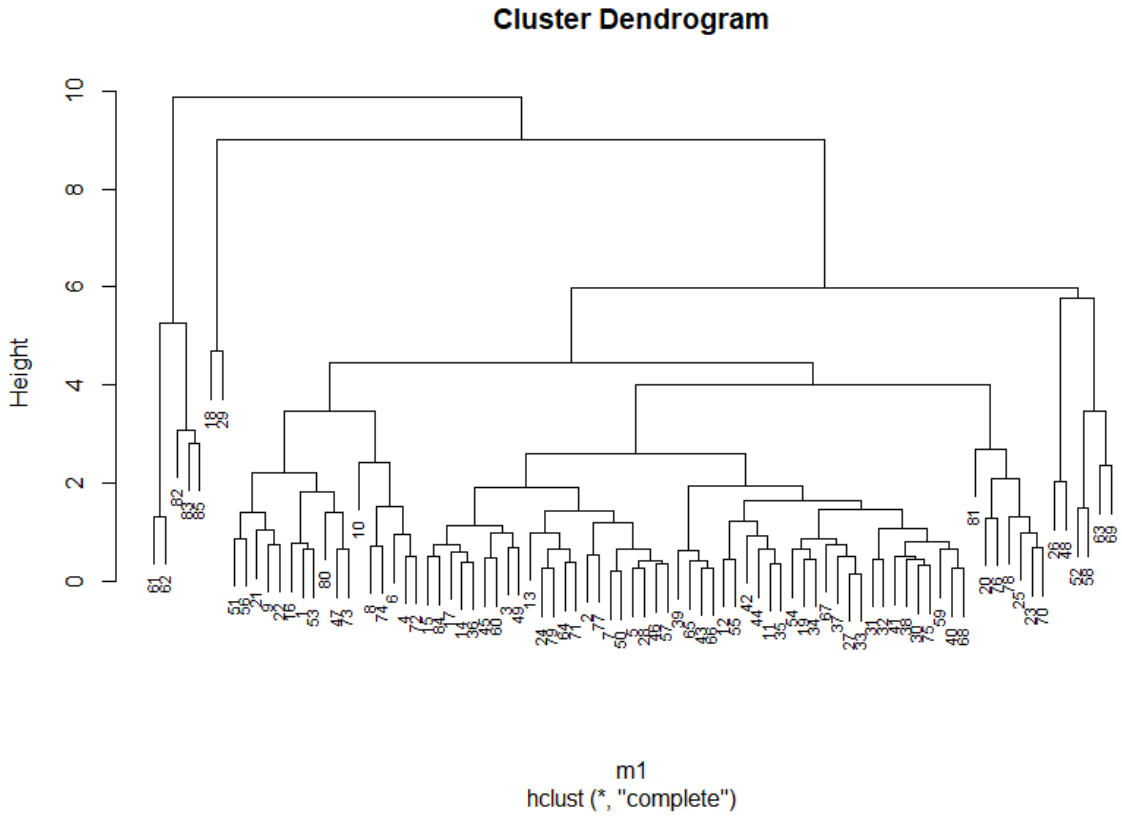
Возможно, следует рассмотреть еще какое-либо число кластеров, либо попробовать использовать другие методы разбиения с иным вычислением внутрикластерного расстояния.

Применение методов иерархической кластеризации.

Метод одиночной связи не дает нам похожего разбиения на четыре кластера.



В отличие от метода полной связи, который показывает неплохой результат, в котором можно обнаружить некоторые сходства с кластеризацией методом k-means. Видим две небольшие группы слева и справа и две большие посередине.



Заключение.

Таким образом, были осуществлены задачи, связанные с поиском необходимой информационной базы, произведен выбор характеристик для исследуемых объектов и реализована предобработка данных. Проведено исследование алгоритмов и методов кластеризации, которые планируется использовать в дальнейшей работе. Также реализован метод k-means, вместе с методами выбора оптимального числа кластеров. Проведена оценка данного разбиения. Были реализованы методы иерархической кластеризации, построены графики дендограмм и проведен эмпирический анализ полученных результатов.

В дальнейшем планируется реализовать оценку качества иерархической кластеризации. Более того, хотелось бы сравнить полученные результаты разбиений основываясь на мерах качества. Есть вариант рассмотрения каждого объекта (субъекта РФ) в качестве временного ряда (экономические и инновационные показатели в течение 10 лет). Так же хотелось бы рассмотреть методы для выбора оптимального числа кластеров для метода k-means, предложенные старшим преподавателем Ореховым А.В. в научных публикациях.

Литература.

1. B.S. Duran, P.L. Odell "Cluster Analysis: A Survey", 1974. Перевод с английского: Е.З. Демиденко, с. 15.
2. И.Д. Мандель "Кластерный анализ", 1988.
3. В.М. Буре, Е.М. Парилина "Теория вероятностей и математическая статистика".
4. В.М. Буре, Е.М. Парилина, А.А. Седаков "Методы прикладной статистики в R и Excel".
5. «Кластерный анализ: Википедия. Свободная энциклопедия. Режим доступа: https://ru.wikipedia.org/wiki/Кластерный анализ (дата обращения: 31.05.2023).
6. Заварухин В.П., Чинаева Т.И., Чурилова Э.Ю. Регионы России: результаты кластеризации на основе экономических и инновационных показателей. Статистика и Экономика. 2022;19(5):35-47. https://doi.org/10.21686/2500-3925-2022-5-35-47.
7. Оценка качества кластеризации. Режим доступа: https://ranalytics.github.io/data-mining/103-Clustering-Quality.html.
8. Приоритетные направления регионального развития https://www.economy.gov.ru/material/directions/regionalnoe_razvitie/.
9. Федеральная служба государственной статистики https://rosstat.gov.ru/.
10. Козина Е.В., Гостева С.В. Проблемы определения и оценки экономического потенциала региона // Интернетжурнал «НАУКОВЕДЕНИЕ» Том 8, №5 (2016) http://naukovedenie.ru/PDF/99EVN516.pdf (доступ свободный).
11. Петрухина, Е.В. Основные факторы инновационного развития регионов / Е.В. Петрухина ; Орловский государственный университет // Научные ведомости БелГУ. Сер. История. Политология. Экономика. Информатика. - 2012. - №7(126), вып.22/1.-С. 56-65.
12. Статистический сборник аналитического центра ФИПС. https://new.fips.ru/about/deyatelnost/sotrudnichestvo-s-regionami-rossii/a-iz-akt-2021.pdf
13. Шитиков В. К., Мастицкий С. Э. (2017) Классификация, регрессия, алгоритмы Data Mining с использованием R. - Электронная книга, адрес доступа: https://github.com/ranalytics/data-mining.