Введение

В текущей работе исследуется эффективность использования основных фондов в отраслях экономики РФ с помощью регрессионного и кластерного анализа.

TODO

1. Провести кластеризацию по исходным признакам. Лучше ли результат? Добавить раздел вопросов

2. Регрессия по фиктивным переменным

3. Исправить цель работы

1. Постановка задачи

Одним из наиболее важных показателей эффективности использования основных фондов (далее – ОФ), является фондоотдача RoFA (return on fixed assets):

|  |  |  |
| --- | --- | --- |
|  |  | () |

где *Q* – стоимость произведенной продукции,

*К* – стоимость ОФ.

Из экономической теории известно, что стоимость продукции связана с затратами на труд и затратами на основные фонды производственной функцией Кобба-Дугласа. С учетом технического прогресса функция имеет вид:

|  |  |  |
| --- | --- | --- |
|  |  | () |

где – время,

– затраты на трудовые ресурсы,

– темп прироста объема производства благодаря техническому прогрессу.

Разделив обе части уравнения (2) на и учитывая (1), получим производственную функцию для фондоотдачи:

|  |  |  |
| --- | --- | --- |
|  |  | () |

Производственная функция Кобба-Дугласа и выведенные из нее (такие, как функция фондоотдачи) справедливы для предприятий только внутри одной отрасли. Однако на межотраслевом уровне, они до сих пор не использовались из-за значительных различий отраслей. Одно из наиболее существенных различий отраслей заключается в структуре основных фондов.

В текущей работе выдвигается гипотеза о том, что группировка отраслей по структуре их основных фондов позволяет получить приемлемое качество регрессионной модели фондоотдачи. Исследование строится в два этапа: сначала производится кластеризация, а затем регрессионный анализ.

1. Использованные данные и методология

В исследовании использовались данные Росстата о 18-ти видах экономической деятельности (далее – ВЭД) согласно общероссийскому классификатору ОКВЭД2 за 6 лет – с 2017 по 2022. На основе этих данных было сформировано две различных таблицы, одна из которых использовалась в кластеризации, а другая – в регрессии.

3.1 Первая таблица представляет собой данные о долях видов ОФ в общем объеме ОФ для каждого ВЭД. Всего было использовано четыре вида ОФ: «Сооружения», «Машины и оборудование», «Жилые здания», «Здания».

Однако в качестве признаков кластеризации использовались доли ОФ только четырех видов(вместо пяти) – для увеличения точности алгоритма за счет уменьшения размерности. В частности, признаки «доля Жилых зданий» и доля «Нежилых зданий» были объединены в один признак «доля Зданий» суммированием.

На основании данных о среднегодовом наличии ОФ для каждого ВЭД сначала рассчитывались значения долей видов ОФ за каждый год с 2017 по 2022. Затем для каждого ВЭД рассчитывались средние доли видов ОФ за 6 лет, которые в итоге использовались как факторы для кластеризации ВЭД.

В эконометрической модели (4) использовались годовые показатели ВЭД, описание которых приведено в таблице 1. При этом показатели затраты на рабочую силу и фондоотдача были рассчитаны на основе промежуточных показателей.

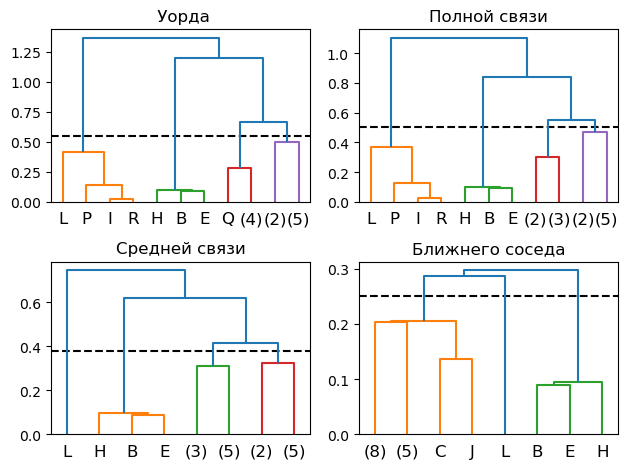
Таблица 1 – Переменные (показатели) эконометрической модели

|  |  |  |  |
| --- | --- | --- | --- |
| Тип | Обозначение | Наименование показателя | Формула расчета |
| Промежуточные |  | Среднегодовая численность занятых, чел. | - |
|  | Среднемесячная номинальная начисленная заработная плата, руб./чел. | - |
|  | Валовая добавленная стоимость, млн. руб. | - |
| Основные |  | Среднегодовое наличие основных фондов, млн. руб. | - |
|  | Суммарные годовые затраты на рабочую силу, млн. руб. |  |
|  | Год | - |
|  | Фондоотдача, руб./руб. |  |

1. Кластеризация

Кластеризация производилась по сжатой таблице (отраслям)

Были использованы четыре вида иерархической агломеративной кластеризации, а именно методы Уорда, полной связи, средней связи и ближнего соседа с помощью библиотеки Python «scipy.cluster». В результате работы программы были получены так называемые дендрограммы – диаграммы последовательности объединения кластеров в иерархию (Рис. 1).



Латинскими буквами обозначены отрасли, числами в скобках – количество отраслей, входящих в кластер

Рисунок 1 – Дендрограммы методов кластеризации.

По полученным дендрограммам производился выбор разбиения на кластеры так, чтобы получить как можно более однородные кластеры, и таким образом добиться наилучшего качества регрессии на них. При этом было учтено основное свойство иерархической кластеризации – чем больше кластеров (и меньше элементов в них), тем больше их однородность. Таким образом выбиралось как можно более детализированное разбиение, при этом следя за тем, чтобы в каждом кластере было не слишком маленькое количество элементов, достаточное для статистической значимости соответствующих регрессий.

Минимальное количество элементов в одном кластере для -го признака рассчитывалось по формуле:

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

где – значение критической точки двустороннего распределения Стьюдента с уровнем значимости и степенями свободы ,

– стандартное отклонение -го признака общей выборки,

– количество элементов в общей выборке,

– предельная ошибка (точность), принятая равной .

Можно показать, что если рассчитывать предельную ошибку для каждого признака одинаковым образом, то рассчитанное количество элементов не будет зависеть от конкретных стандартных отклонений различных признаков. Таким образом, минимальное количество элементов в одном кластере будет единственно: . В нашем случае при размере исходной выборки минимальное количество наблюдений в одном кластере оказалось равно 16. Учитывая, что по каждой отрасли было собрано по 6 наблюдений (по одному за каждый год с 2017 по 2022), минимальное количество отраслей в одном кластере составило 3.

Итак, для каждого метода кластеризации было выбрано такое разбиение на кластеры, чтобы каждый кластер содержал как можно меньше отраслей, но не менее трех. На дендрограммах (Рис. 1) показаны полученные разбиения: цветом выделены минимальные кластеры, а пунктирной линией – место разреза дерева для получения этих кластеров.

Далее, для каждого метода была посчитана суммарная внутригрупповая дисперсия разбиения в зависимости от количества выбранных кластеров. На соответствующих графиках (Рис. 2) можно увидеть так называемые точки перегиба, соответствующие количеству кластеров, при увеличении дальше которого внутригрупповая средняя уменьшается несущественно. Оказалось, что выбранное нами количество кластеров достаточно близко к оптимальному для всех методов кластеризации, кроме метода ближнего соседа.

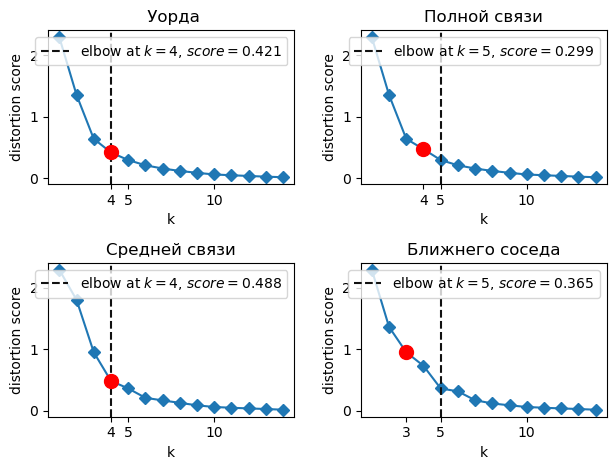


Рисунок 2 – Графики зависимости суммарной внутригрупповой дисперсии от количества кластеров

В итоге лучшим оказалось разбиение на кластеры методом Уорда, так как значение внутригрупповой дисперсии для выбранного количества кластеров было наименьшим (Таблица 2).

Таблица 2 – Зависимость средней внутригрупповой дисперсии от количества кластеров в разбиении для каждого метода

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Количество кластеров | Уорда | Полной связи | Средней связи | Ближнего соседа |
| 1 | 2.29 | 2.29 | 2.29 | 2.29 |
| 2 | 1.36 | 1.36 | 1.81 | 1.37 |
| 3 | 0.64 | 0.64 | 0.95 | 0.95 |
| 4 | **0.42** | 0.48 | 0.49 | 0.73 |

Анализируя центроиды кластеров, полученных методом Уорда, (Таблица 3) были сделаны следующие выводы об их профилях.

Во-первых, выделились два кластера с доминирущим видом основных фондов в структуре основных фондов. Так, в первый кластер попали отрасли, наиболее активно задействующие здания, такие как «Образование» и «Деятельность гостиниц и предприятий общественного питания». Во втором кластере доминирующим видом основных фондов оказались сооружения. В него вошли такие отрасли, как «Добыча полезных ископаемых» и «Транспортировка и хранение». Показать параллельное улучшение качества

Остальные два кластера не имеют такой узкой направленности. В третий кластер попали отрасли, в равной степени задействующие здания и оборудование, такие как «Здравоохранение» и «Торговля». В четвертый кластер были объединены отрасли, в равной степени задействующие каждый из видов основных фондов, кроме транспортных средств, например, «Строительство» и «Административная деятельность».

Таблица 3 – Центры полученных кластеров

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Номер кластера | Здания | Машины и оборудование | Сооружения | Транспортные средства |
| 1 | **0.71** | 0.12 | 0.10 | 0.02 |
| 2 | 0.09 | 0.14 | **0.70** | 0.06 |
| 3 | **0.41** | **0.31** | 0.12 | 0.09 |
| 4 | **0.21** | **0.34** | **0.31** | 0.08 |

1. Регрессионный анализ
   1. Построение моделей на кластерах

На каждом из четырех полученных кластеров была оценена линеаризованная функции фондоотдачи:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

Во всех кластерах уравнения регрессии оказались значимыми, причем коэффициенты оказались значимыми во всех кластерах, кроме одного – на одном кластере с номером 3 коэффициенты (константа) и незначимы. Проверим данные кластера номера 3 на аномальность с помощью расстояния Махаланобиса [?], которое отражает расстояние от вектора до центра многомерного нормального распределения: It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D.

где – наблюдение, проверяемое на аномальность,

– матожидание выборочного распределения,

– симметричная положительно определенная матрица ковариации выборки.

Квадрат расстояния Махаланобиса – случайная величина, распределенная по закону с количеством степеней свободы, равным количеству признаков в выборке (в нашем случае ). Гипотеза об аномальности одного наблюдения формулируется следующим образом:

На уровне значимости проверим гипотезу для всех наблюдений рассматриваемого, сравнивая соответствующие рассчитанные расстояния с критическим значением .

Было выбрано минимальное расстояние, при котором

С помощью дисперсионного анализа убедиться, что однородность данных улучшилась? Возможно, это слишком и можно объяснить просто. Как? Качество регресии улучшается при улучшении качества данных, например, при исключении аномалий. Можно ли исключать не совсем аномальные данные, а только отклоняющиеся от среднего? Посчитать distortion на кластерах до и после перенесения элементов

Сравнить гетероскедастичности до и после шафла!!

1. Как связана -статистика с характеристиками выборки?

2. Как расслоить данные?

3. Построить 3д

Многомерный тест на нормальность каждой подвыборки – Тест на равенство вариаций – Дисперсионный анализ проверяет только стат. Значимость различий, но не их величину

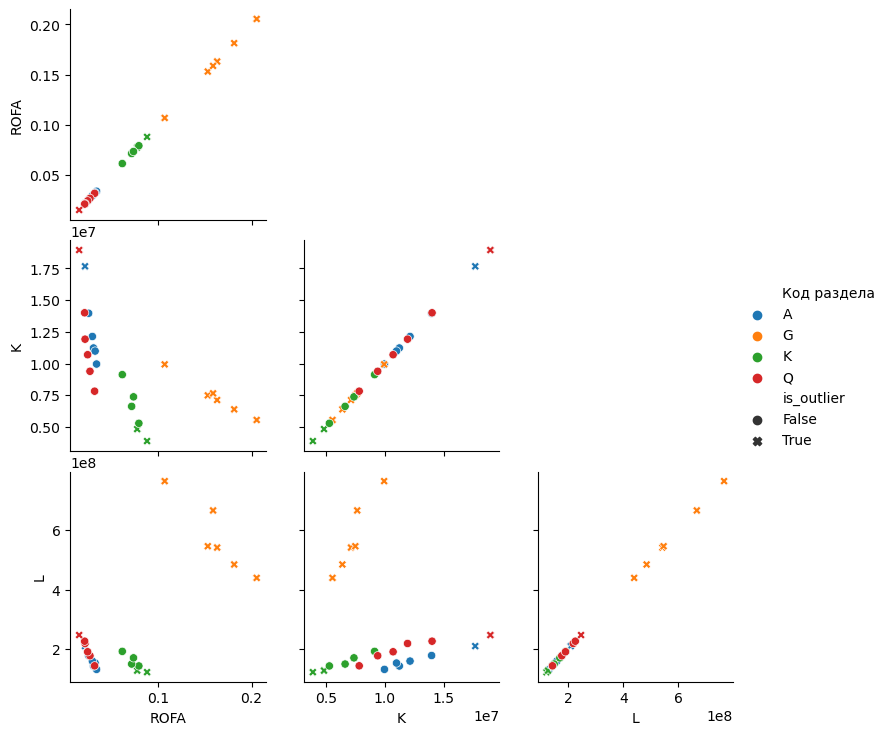


Рисунок 3 – Во втором кластере наблюдения отрасли G являются аномальными 5 по критерию Махаланобиса

* 1. Анализ свойств регрессий

После оценки коэффициентов моделей на кластерах они были проверены на наличие мультиколлинеарности, гетероскедастичности и авторегрессии. (а на нормальность?) Результаты сопутствующих тестов приведены в Таблице

Таблица 3 – Результаты тестов свойств моделей

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Номер кластера | Тест Уайта, p-знач. | для | для | для | Тест Джарка-Бера,  -знач. |  |
| 0 | 3 | 24 | 0.0664 | 1.6002 | 1.1659 | 1.7519 | 0.3318 |
| 1 | 2 | 18 | 0.1053 | 28.8948 | 28.8565 | 1.0253 | 0.2886 |
| 2 | 4 | 42 | 0.0009 | 1.1772 | 1.1957 | 1.0996 | 0.2682 |
| 3 | 1 | 24 | 0.0328 | 1.1123 | 1.1415 | 1.0875 | 0.6764 |

4.2.2 Мультиколлинеарность выявлялась с помощью коэффициента инфляции дисперсии :

где – коэффициент детерминации уравнения регрессии на все остальные экзогенные переменные модели.

Значения -коэффициента лежат в интервале от 1 до бесконечности, причем значения, лежащие близко к 1 свидетельствуют об отсутствии мультиколлинеарности, а значения, большие 3 – о её наличии.

Для каждой переменной оцениваемой модели были рассчитаны -коэффициенты на данных каждого кластера. Только в регрессии, построенной по второму кластеру, была обнаружена мультиколлинеарность между переменными и .

2. Скопировать statsmodels.OLS изменить эстиматор на гребневую

4.2.3 Гетероскедастичность построенных моделей проверялась с помощью теста Уайта, в котором предполагается, что дисперсии ошибок регрессии представляют собой одну и ту же функцию от наблюдаемых значений регрессоров:

Если такая функциональная зависимость существует, то существует статистическая зависимость между ошибками регрессии и значениями регрессоров:

Тогда проверяемая гипотеза об отсутствии гетероскедастичности есть то же самое что и гипотеза о значимости этого уравнения регрессии и проверяется по -статистике. При этом включает все регрессоры, их квадраты и все их попарные произведения.

По полученным -значениям -статистики теста Уайта был сделан вывод о наличии гетероскедастичности в кластерах с номерами 1 и 4 на уровне 5%.

Автокорреляция

Список источников

1. https://stackoverflow.com/questions/46827580/multivariate-outlier-removal-with-mahalanobis-distance