

## **Экологическое состояние регионов России: кластерный анализ**

**Куриленко Анастасия Юрьевна**

*г. Москва, НИУ ВШЭ*

*Email: leucingerfu@gmail.com*

**Саенко Анастасия Игоревна**

*г. Москва, НИУ ВШЭ*

*E-mail: nastiasaenko@gmail.com*

*Научный руководитель: к.э.н., доцент Звездина Н.В.*

Значения экологических показателей – один из критериев оценки уровня развития страны и ее благосостояния. По данным официальной статистики [3] и мнению экспертов [2] уровень выбросов вредных веществ в атмосферу на региональном аспекте существенно неоднороден. Для кластеризации субъектов РФ по уровню загрязненности воздуха авторами была создана система показателей, описывающая регионы, по официальным данным за 2016 год: выбросы загрязняющих веществ в атмосферный воздух, (тыс. тонн), реальные доходы (руб.), численность населения (тыс. чел.), ВРП (млн. руб.), затраты на охрану окружающей среды (млн. руб.), удельный вес городского населения/уровень урбанизации (в процентах) и уровень промышленности (млн. руб.). Перед анализом данные были нормированы.

Предварительно был проведен корреляционный анализ с целью первичной оценки модели и поиска самых значимых для зависимой переменной связей. При расчете и проверке на значимость парных и частных коэффициентов корреляции было выявлено, что самая сильная связь наблюдается между загрязненностью воздуха и государственными затратами в сфере экологии, а также затратами и уровнем промышленности в регионе (на уровне значимости 0,05). Далее был проведен регрессионный анализ, для которого в качестве зависимой переменной были взяты выбросы в атмосферный воздух. Модель линейной регрессии показала, что данные в достаточной степени характеризуют зависимую переменную ( $R^2=0.4829$ ).

Кластеризация проводилась на языке программирования Python в несколько этапов. Сначала тестировались различные варианты представления данных с помощью алгоритмов снижения размерности, а именно TSNE - алгоритма, основанного на t-распределении. Было получено два вида представления данных - трехмерное и двухмерное. Вторым этапом было проведение иерархической кластеризации, использовались разные ее методы (ближнего и дальнего соседа, средней связи и центра тяжести, метод Уорда). Эффективность иерархического разделения проверялась двумя способами - визуальное оценивание полученных дендрограмм и оценивание наполненности кластеров при разных порогах. В результате было установлено, что наилучшее разбиение проходит по методу “дальнего соседа” (рис.1). Было также выдвинуто предположение о том, что 4 является оптимальным числом кластеров.

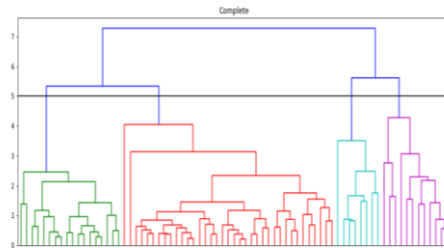


Рис. 1. Дендрограмма по методу дальнего соседа

Дальнейшая валидация проводилась с помощью метода силуэтов: рассчитывались silhouette-оценки для каждого варианта иерархической кластеризации от 2 до 20 кластеров включительно. Данная метрика установила, что при иерархической кластеризации оптимальным числом кластером является 3.

Для дальнейшей кластеризации методом k-средних была проведена дополнительная проверка методом силуэтов и методом локтя, которые установили, что для метода k-средних оптимальное число кластеров 4, что подтвердило разбиение, предложенное методом дальнего соседа. В итоге были получены 4 кластера. По их средним показателям были рассчитаны “места” кластеров относительно друг друга.

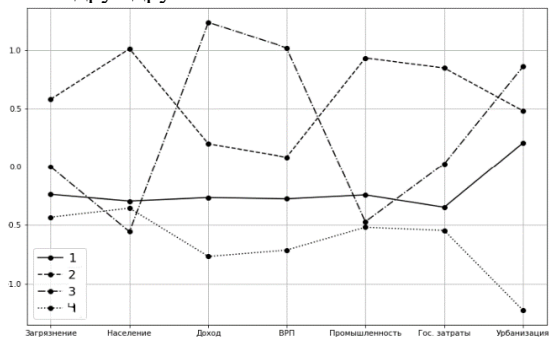


Рис. 2. График средних показателей по кластерам

Кластеры отчасти формировались и по географическому признаку - регионы, принадлежащие одному или близким по расположению федеральным округам, попадали в один кластер. Данные представлены в виде диаграмм рассеяния, предварительно размерность данных была снижена.

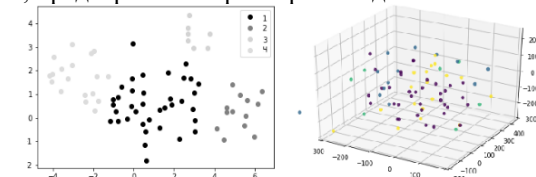


Рис. 3. Графическое представление кластеров. Цветом обозначена принадлежность к тому или иному кластеру. Графики получены путем снижения размерности, поэтому оси не относятся к какому-то конкретному показателю.

Таким образом, было получено 4 кластера, которые ранжируются в зависимости от общей экологической обстановки в регионах. В первый кластер (№1) вошли такие регионы, как Ивановская, Калужская и Брянская области. Эта группа регионов занимает второе место по числу населения и уровню промышленности и третье – остальным показателям, в том числе и по загрязнению воздуха. Во вторую группу (№2) вошли такие регионы, как Нижегородская область и Пермский край. Этот кластер характеризуется высоким уровнем загрязнения воздуха, производства и затрат на охрану природы и средним уровнем ВРП, доходов и урбанизации. В третий кластер (№3) попали Хабаровский край и Архангельская область, для него характерен высокий уровень доходов, ВРП и урбанизации, и низкое число населения. В четвертый кластер (№4) попали республики Северного Кавказа. Он характеризуется низкими значениями по всем показателям.

1. Федеральная служба государственной статистики. URL: <http://www.gks.ru> – (дата обращения – 20.11.2019).
2. Тагаева Татьяна Олеговна, Гильмундинов Вадим Манавирович, Казанцева Лидия Кузьминична Экологическая ситуация и природоохранная политика в регионах России // Экономика региона. 2016. №1. URL: <https://cyberleninka.ru/article/n/ekologicheskaya-situatsiya-i-prirodoohrannaya-politika-v-regionah-rossii> (дата обращения: 30.03.2020).
3. Национальный экологический рейтинг. Зима 2019-2020 // Зелёный патруль URL: <https://greenpatrol.ru/ru/stranica-dlya-obshchego-reytinga/ekologicheskii-reyting-subektov-rf?tid=388> (дата обращения: 30.03.2020).