Санкт-Петербургский государственный университет
факультет прикладной математики – процессов управления

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Шелест Арина Александровна

О применении методов обнаружения выбросов к задаче исследования проб нефти

Направление 01.04.02. – Прикладная математика и информатика
Программа – Методы прикладной математики и информатики
в задачах управления

Научный руководитель:
профессор кафедры теории игр
и статистических решений,
доктор физико-математических наук
Е.В. Громова

Рецензент:
научный сотрудник
химико-биологического кластера,
Университет ИТМО,
кандидат химических наук
Е.В. Андрусенко

Санкт-Петербург
2020

Содержание

Введение

Методы многомерного анализа данных тесно связаны с математической статистикой, она широко используется в физическом и химическом анализе, в частности, для вычисления средних, отклонений, пределов обнаружения, проверки гипотез. Проблема обнаружения измерений, негативно влияющих на обеспечение чистоты анализа данных – одна из основных проблем анализа данных.

Данная работа посвящена анализу проб нефти с целью обнаружения выбросов – результатов измерения, выделяющихся из общей выборки [4] – и дальнейшей кластеризации полученных проб. По результатам лабораторных анализов судят о среднем составе и параметрах добытой нефти. Достоверность выявленных параметров нефти зависит не только от точности лабораторного оборудования и тщательности соблюдения методик измерений физико-химических свойств нефти, но и от правильности отбора проб. Если проба оказалась случайной и не характеризует средний состав прошедшей нефти, то и её анализ не может отразить действительного физико-химического состава. Под погрешностью отбора проб понимают отклонение значения величины, характеризующей состав, свойства пробы вещества от значения этой же величины характеризующей состав, свойства объекта аналитического контроля в целом.

В данной работе описан и модифицирован один из методов обнаружения выбросов в одномерных наборах данных, проведен корреляционный анализ, описан кластерный анализ, применимый к задаче исследования проб нефти, введена метрика для оценки качества кластеризации.

Постановка задачи

Получены данные газовой хроматографии по 23 пробам, взятые из 6 пластов одного месторождения, каждая проба характеризуется набором 330 параметров, представляющие собой индексы удерживания единичного межалканового пика (индексы Ковача).

Изначально данные были сгруппированы по скважинам. В таблице 1 приведены характеристики каждой из проб (привязка: месторождение отбора; скважина/куст; статиграфическая привязка (свита/горизонт, возраст); глубина, интервал отбора, место взятия от верха; пласт).

Таблица 1 – Соответствие проб и пластов

№ пробы	Параметры скважины	Пласт
1	П-10419ГС125-АС10.4(6)	АС10.4(6)
2	П-10420ГС127-АС10.4(6)	АС10.4(6)
3	П-1258А-АС10.1-3(1)	АС10.1-3(1)
4	П-15932БС8А-АС10.1-3(1)	АС10.1-3(1)
5	П-16499БС44-АС10.4(1)	АС10.4(1)
6	П-19659ГС124А-АС10.1-3(1)	АС10.1-3(1)
7	П-30111А-АС10.1-3(1)	АС10.1-3(1)
8	П-30611А-АС10.1-3(1)	АС10.1-3(1)
9	П-30711-АС10.1-3(1)	АС10.1-3(1)
10	П-30937ГС76А-АС10.0.1(1)	АС10.0.1(1)
11	П-30977ГС76А-АС10.0.1(1)	АС10.0.1(1)
12	П-35375ГС722-АС10.0.1(1)	АС10.0.1(1)
13	П-35376ГС721А-АС10.1-3(1)	АС10.1-3(1)
14	П-35422ГС721А-АС10.0.1(1)	АС10.0.1(1)
15	П-35423ГС721А-АС10.0.1(1)	АС10.0.1(1)
16	П-35424ГС721А-АС10.0.1(1)	АС10.0.1(1)
17	П-41932ГС21Б-АС12.3-5(4)	АС12.3-5(4)
18	П-42095ГС25А-АС12.3-5(4)	АС12.3-5(4)
19	П-42097ГС25А-АС12.3-5(4)	АС12.3-5(4)
20	П-42099ГС25А-АС12.3-5(4)	АС12.3-5(4)
21	П-42104ГС14А-АС12.3-5(4)	АС12.3-5(4)
22	П-42106ГС14А-АС12.3-5(4)	АС12.3-5(4)
23	П-42108ГС14А-АС12.1(2)	АС12.1(2)

Необходимо:

- обнаружить измерения, негативно влияющие на обеспечение чистоты анализа данных;
- определить взаимосвязь между полученными данными;
- осуществить разбиение проб на k кластеров.

Обзор литературы

Краткое изложение теории ошибок представлено в работе [4] Зайдель А.Н. Для данной выпускной работы важен III раздел «Приемы вычислений», где подробно описаны некоторые методы количественной оценки погрешностей измерений. Также в разделе I описаны типы ошибок, объяснена их природа происхождения и, что наиболее интересно, приведены поясняющие примеры.

Современные методы анализа распределений вероятностей, оценки параметров распределений, проверки статистических гипотез, оценки связей между случайными величинами в понятных терминах излагаются в работе [5] Кобзарь А.И. В главе 3 «Методы анализа законов распределения вероятностей случайных величин» разобраны на примерах, в том числе используемые в работе, критерий Колмогорова-Смирнова, критерий Шапиро-Уилка, критерий Андерсона-Дарлинга.

В работе [2] Буре В.М., Парилиной Е.М. изложены основные разделы современного курса математической статистики, включая принципы статистического оценивания, методы построения доверительных интервалов, методы проверки статистических гипотез. Коэффициент ранговой корреляции Спирмена, вычисляющийся для порядковых шкал, коэффициент корреляции Пирсона, применяемый к данным измеренным в шкале отношений, рассмотрены в главе 17.

Основные физические свойства и методика исследования нефтеносных пород, нефти и газа в пластовых условиях рассмотрены в работе [3] Гиматудинова Ш.К.

Вопросы организации высокоэффективных хроматографических процессов рассмотрены в работе [7] Руденко Б.А., Руденко Г.И. Здесь описаны системы, обеспечивающие высокую общую эффективность процесса и позволяющие осуществлять разделение очень близких по свойствам веществ, а также системы, обеспечивающие эффективность, близкую к предельно

достижимым теоретически параметрам для данной хроматографической системы. Государственный стандарт [1] устанавливает определения основных понятий в области газовой хроматографии. Работа [8] Шакировой Д.И., Рождественского Д.А. содержит определения и расчеты параметров, применимых ко всем хроматографическим методам в общем случае; описание приборов в [9].

Выбранный в работе метод для обнаружения выбросов описан в книге [17] Дж. Тьюки. Именно в ней описан принцип построения коробочной диаграммы, так называемого «ящика с усами».

Глава 1. Исследование проб нефти

1.1. Нефтеносный пласт

Нефть – маслянистая жидкость, обычно черного или красно-коричневого цвета со специфическим запахом и горючими свойствами. Сегодня из данного вещества получают топливо, поэтому можно смело говорить о том, что это наиболее ценное полезное ископаемое на планете Земля (наряду с природным газом).

Залежью называется естественное локальное скопление нефти в одном или нескольких гидродинамически связанных пластах. Месторождение – это совокупность залежей нефти, приуроченных к одной или нескольким ловушкам, расположенных на одной локальной площади. Месторождение может быть одно – или многопластовое. В зависимости от площади месторождения могут быть большими (30*60 км), средними (10*20 км) и малыми (до 10 км^2).

Пласт – это геологический слой, являющийся основной формой залегания осадочных горных пород и отражающий их последовательность отложения.

Состав пласта более или менее однороден и ограничивается двумя параллельными поверхностями. Верхняя поверхность называется кровлей, а нижняя – подошвой. Если залегание не нарушено, каждый вышележащий пласт считается более молодым, чем нижележащий.

Нефтяной пласт – это горная осадочная порода, имеющая скопления капиллярных каналов и трещин с большой поверхностью. Эта порода пропитана водой, нефтью и газом. Закономерности движения нефти и её вытеснение из пористой среды пласта зависит от свойств, которыми обладают пограничные слои фаз и процессов, происходящих на поверхности контактов воды, нефти и газа с породой.

Нефть представляет собой в основном смесь углеводородов различного состава, хотя в ней обычно преобладают углеводороды метанового

(парафинового) или нафтенового рядов. В меньших количествах встречаются углеводороды ароматического ряда и другие. Товарные качества нефти определяются содержанием легких и тяжелых углеводородов, составом жидких и твердых углеводородов и наличием примесей. В нефти в небольших количествах встречаются хлор, йод, фосфор, мышьяк, калий, натрий, кальций, магний. Из кислородных соединений наибольшее значение имеют нафтеновые и жирные кислоты, асфальтены и смолы.

Нефтеносные пласты – это несистемное чередование проницаемых нефтенасыщенных горных пород. На рисунке 1 представлен геологический разрез нефтегазоконденсатного месторождения, где 1 – газ; 2 – нефть; 3 – вода; 4 – глина.

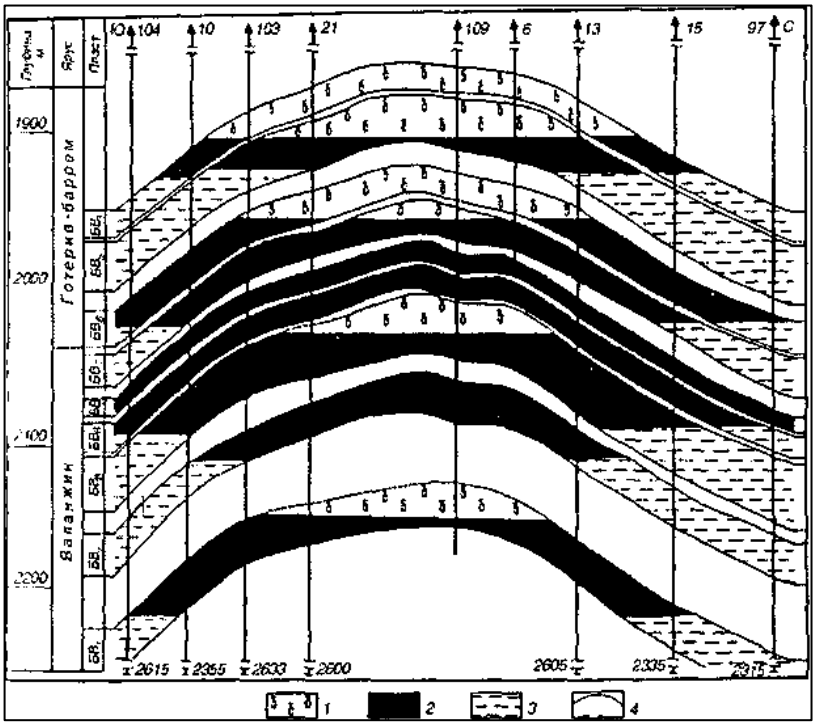

Рисунок 1 – Геологический разрез месторождения

Исследование свойств нефти начинают с отбора проб. Отобранные пробы нефти переводятся в специальные контейнеры или транспортируются в лаборатории в корпусе пробоотборника.

1.2. Метод газовой хроматографии при исследовании проб нефти

Газовая хроматография – широко используемый аналитический метод разделения, анализа и физико-химических исследований, основанный на распределении вещества между двумя несмешивающимися фазами: подвижной и неподвижной. Здесь газ – агрегатное состояние подвижной фазы.

Процесс хроматографирования начинается со стадии ввода разделяемой смеси веществ в хроматографическую колонку. При этом компоненты разделяемой смеси перемещаются по колонке с потоком газа-носителя. По мере движения разделяемая смесь многократно распределяется между подвижной и неподвижной фазами (сорбент), то есть компоненты смеси селективно задерживаются на неподвижной фазе. Сначала из колонки выходят те компоненты, которые хуже задерживаются на неподвижной фазе, а после них те, которые лучше. Таким образом, достигается полное разделение смеси на компоненты при достаточной длине колонки. После выхода из нее вещества регистрируются детектором. В результате получают хроматограмму, состоящую из ряда пиков, каждый из которых характеризует количество компонента или группы компонентов. Время выхода каждого компонента является строго постоянной величиной. Определяя время, прошедшее с момента подачи пробы до выхода компонента из колонки, можно качественно расшифровать хроматограмму. Для каждого пика находят площадь либо высоту с помощью программного обеспечения.

Пик – это некоторый участок хроматограммы с записанным сигналом детектора при элюировании из колонки одного или более неразделенных компонентов. Элюирование – извлечение вещества вымыванием его подходящим растворителем – элюентом. Пик может быть охарактеризован площадью пика, или высотой пика (h) и шириной пика на половине высоты (w_h), или высотой пика (h) и шириной пика между точками перегиба (w_i).

Хроматограммы представляют собой последовательность гауссовых пиков, расположенных на базовой линии [8].

Для гауссовых пиков выполняется соотношение:

$$w_h = 1{,}8w_i$$

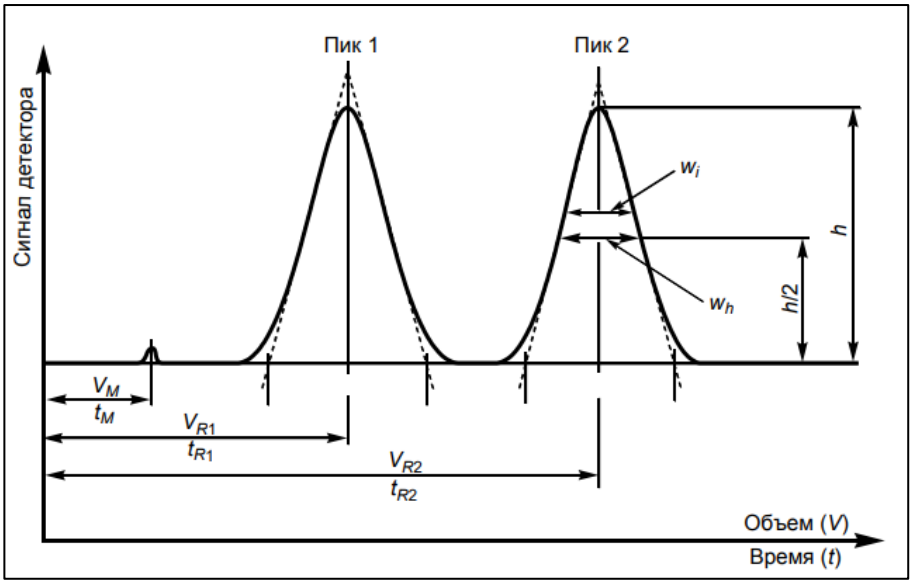

Рисунок 2 – Схематическое изображение хроматограммы

На рисунке 2:

t_R – время удерживания, необходимое для элюирования компонента.

V_R – объем удерживания, то есть объем подвижной фазы, необходимый для элюирования компонента. V_R рассчитывается по следующей формуле:

$$V_R = t_R * F,$$

где F – скорость подвижной фазы (в миллилитрах в минуту).

t_M – время, необходимое для элюирования неудерживаемого компонента, так называемое «мертвое» время.

V_M – объем подвижной фазы, необходимый для элюирования неудерживаемого компонента, так называемый «мертвый» объем. V_M рассчитывается по формуле:

$$V_M = t_M * F.$$

k – коэффициент удерживания – характеристика, определяемая по формуле:

$$k = \frac{\text{количество вещества в неподвижной фазе}}{\text{количество вещества в подвижной фазе}} = K_c \frac{V_s}{V_m},$$

где K_c – константа распределения (известная также как коэффициент равновесного распределения, V_s – объем неподвижной фазы, V_m – объем подвижной фазы.

Коэффициент удерживания компонента может быть определен из хроматограммы по формуле:

$$k = \frac{t_R - t_M}{t_M}.$$

Для сопоставления полученных значений с данными, полученными на других приборах, необходимо ввести ряд поправок на объем газа-носителя, не принимающего участия в вымывании компонентов пробы. Приведенное время удерживания пересчитывают с учетом поправки на время удерживания несорбирующегося компонента t_M:

$$t'_{Ri} = t_R - t_M$$

Для представления величин удерживания в газовой хроматографии используется индекс удерживания Ковача. По определению Ковача индекс удерживания – это мера относительного удерживания веществ, причем в качестве стандартного сравнения, как правило, используется нормальные углеводороды. Индекс удерживания задается следующей формулой:

$$l_i = 100\left(\frac{\ln t'_{Ri} - \ln t'_{Rm}}{\ln t'_{R(m+1)} - \ln t'_{Rm}} + m\right), \text{ где}$$

t'_{Ri} – приведенное время удерживания определяемого компонента;

$t'_{Rm}, t'_{R(m+1)}$ – приведенные времена удерживания н-алканов с числом атомов углерода m и (m+1), элюирующихся до и после определяемого вещества.

m- мера нормального алкана, содержащего m атомов углерода;

l_i – индекс удерживания Ковача рассматриваемого вещества i, выходящего между алканами m и m+1.

В процессе расшифровки хроматограммы также важны следующие характеристики:

A_s – характеристика симметричности пика. Коэффициент симметрии рассчитывается по формуле:

$$A_s = \frac{w_{0.05}}{2d},$$

где $w_{0.05}$ – ширина пика на одной двадцатой его высоты, d – расстояние между перпендикуляром, опущенным из максимума пика, и передней границей пика на одной двадцатой его высоты.

Если $A_s = 1.0$, то пик считается полностью симметричным. Если $A_s > 1.0$, пик имеет растянутый задний фронт («хвост»); если $A_s < 1.0$, пик имеет растянутый передний фронт как на рисунке 3.

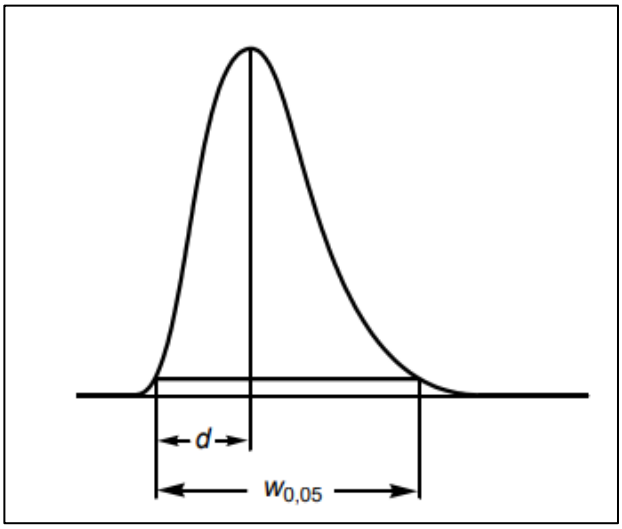

Рисунок 3 – Схематическое изображение пика с растянутым задним фронтом

R_s – характеристика степени разделения между пиками двух компонентов, которая может быть рассчитана по формуле:

$$R_s = \frac{1{,}18(t_{R2} - t_{R1})}{w_{h1} + w_{h2}},$$

где t_{R2}, t_{R1} – времена удерживания пиков, $t_{R2} > t_{R1}$ и w_{h1}, w_{h2} – ширина пиков на половине высоты.

Отношение пик/впадина ($^{p}/_{v}$) – характеристика, используемая в качестве критерия пригодности хроматографической системы в испытании на родственные примеси, когда разделение двух пиков до базовой линии не достигнуто (рисунок 4) и рассчитываемая по формуле:

$$^{p}/_{v} = \frac{H_p}{H_v},$$

где H_p – высота меньшего пика относительно экстраполированной базовой линии, H_v – высота над экстраполированной базовой линией наиболее низкой точки кривой, разделяющей меньший и больший пики.

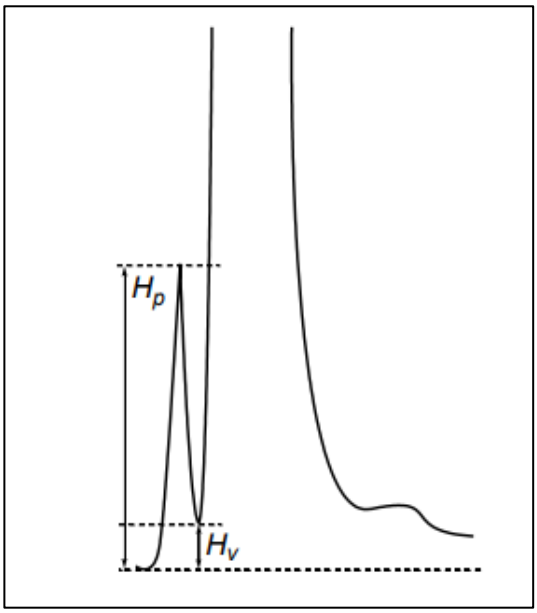

Рисунок 4 – Схематическое изображение неразделившихся пиков

r – относительное удерживание – характеристика, рассчитываемая по формуле:

$$r = \frac{t_{Ri} - t_M}{t_{Rst} - t_M},$$

где t_{Ri} – время удерживания пика определяемого компонента, t_{Rst} – время удерживания пика сравнения (обычно пик испытуемого вещества), t_M – «мертвое» время.

Определение индексов удерживания обычно лежит в основе качественного анализа. В основе количественного хроматографического анализа лежит измерение площади регистрируемого пика, которая пропорциональна концентрации вещества в пробе. На современных приборах площадь пика определяется интегратором. При отсутствии интегратора площадь может быть определена как произведение высоты пика на его полуширину (ширина пика на половине высоты). Таким образом, входными данными в работе служит матрица из 23 столбцов и 330 строк, представляющая собой индексы удерживания межалканового пика (индексы

удерживания Ковача), где столбцы – пробы, взятые из разных пластов одного месторождения, а строки – компоненты нефтяной пробы представлены на рисунке 5.

	prob1	prob2	prob3	prob4	prob5	prob6	prob7
comp1	625,9673259	626,0309278	626,0569456	625,9930915	625,989673	626,0794473	626,0757315
comp2	660,0171969	660,137457	660,13805	660,0172712	660,1549053	660,1036269	660,1549053
comp3	670,0773861	670,2749141	670,2329594	670,1208981	670,1376936	670,2936097	670,2237522
comp4	682,9750645	683,161512	683,175151	683,074266	683,0464716	683,2469775	683,2185886
comp5	690,1117799	690,2920962	690,336497	690,2417962	690,2753873	690,4145078	690,3614458
comp6	722,9408792	723,1158321	723,0496454	722,964087	722,9873418	723,1589639	723,0496454
comp7	725,9727135	726,1507334	726,1398176	726,0495701	726,1265823	726,2569832	726,1904762
comp8	728,2465892	726,1507334	728,4194529	728,2751644	728,4050633	728,5424073	728,4701114
comp9	732,2890349	732,5240263	732,4721378	732,3216995	732,4556962	732,6053834	732,5734549
comp10	737,0389085	737,2787051	737,2340426	737,12696	737,2151899	737,4301676	737,3353597
comp11	744,5679636	744,8153768	744,7821682	744,6130501	744,7594937	744,9466734	744,8834853
comp12	755,3815058	755,6904401	755,6231003	755,4881133	755,5949367	755,8151346	755,7244174
comp13	755,3815058	755,6904401	757,5987842	755,4881133	757,5696203	755,8151346	757,6494428
comp14	760,6366852	760,748609	760,6889564	760,6474456	760,7088608	760,8430675	760,739615

Рисунок 5 – Фрагмент исходных данных

Глава 2. Обработка данных

2.1. Первичный анализ результатов

Описательная (дескриптивная) статистика – один из разделов статистической науки, в рамках которого изучаются методы описания и представления основных свойств данных. Так, чтобы обобщить первичные полученные данные газовой хроматографии, были найдены следующие характеристики (величины) для всех 23 проб:

- mean – среднее арифметическое, рассчитываемое по формуле:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i;$$

- min – наблюдаемый минимум среди параметров пробы;
- max – наблюдаемый максимум среди параметров пробы;
- std – стандартное отклонение, рассчитываемое по формуле:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}};$$

- первый квартиль – значение параметра пробы, ниже которого в упорядоченном по возрастанию массиве находится четверть (25%) данных;
- второй квартиль – медиана – значение параметра пробы, расположенное в середине массива, упорядоченного по возрастанию;
- третий квартиль – значение параметра пробы, выше которого в упорядоченном по возрастанию массиве находится четверть данных.

На рисунке 6 представлены данные величины для проб № 1-7.

	prob1	prob2	prob3	prob4	prob5	prob6	prob7
mean	1560.524316	1560.233128	1560.838494	1560.567900	1560.862477	1552.048694	1560.778008
std	619.886396	619.831453	619.912510	619.999637	619.987900	615.488135	619.912005
min	625.967326	626.030928	626.056946	625.993091	625.989673	379.752475	626.075732
25%	1046.600064	1047.035128	1046.890106	1046.487271	1046.889104	1043.715759	1046.875000
50%	1414.945227	1415.231010	1415.119467	1414.806110	1415.242947	1411.276429	1415.158637
75%	1989.503042	1989.908722	1989.984787	1989.579107	1990.035497	1979.817444	1989.832657
max	3339.248641	3341.876543	3341.732673	3341.873142	3342.016807	3205.416192	3342.482690

Рисунок 6 – Величины проб № 1-7

Для наглядного представления рассогласования, наблюдаемого и теоретического (нормального) распределения были построены графики квантиль-квантиль (QQPlot) – это графики, на которых квантили из двух распределений расположены относительно друг друга.

На квантильном графике по оси X отложены квантили стандартного нормального распределения, а по оси Y – квантили наблюдаемых значений. Если распределение исходной величины нормальное, то точки лягут приблизительно по прямой линии.

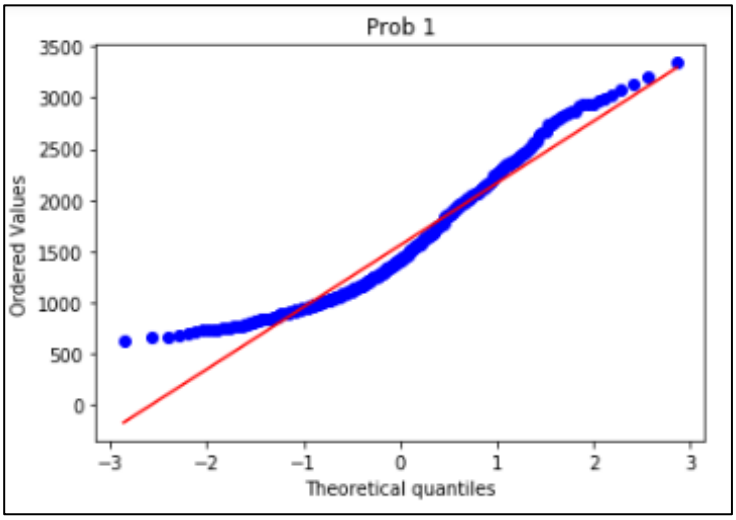

Рисунок 7 – QQPlot для пробы № 1

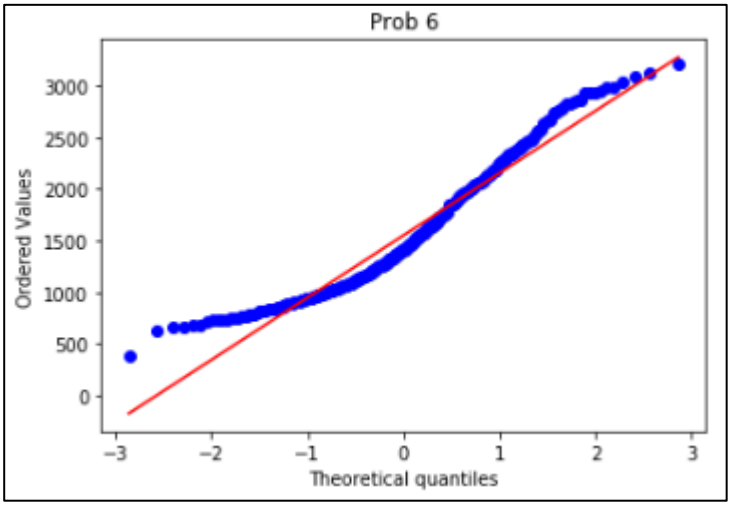

Рисунок 8 – QQPlot для пробы № 6

Среднее арифметическое является одной из наиболее распространённых мер центральной тенденции. Так для проб № 1-5, 7, 9, 11-12, 15-16, 18, 22, 23

центральная тенденция составляет ≈ 1560.8, для проб № 10, 13-14, 17, 20-21 ≈ 1551.8, для проб № 18, 19 ≈ 1556.1, для пробы № 6 ≈ 1552.

Стандартное отклонение является мерой разнообразия входящих в группу объектов и показывает, на сколько в среднем отклоняются варианты от средней арифметической изучаемой совокупности. Стандартное отклонение близкое к 0 говорит о маленькой вариабельности данных. Для проб № 1-5, 7-9, 11, 15-16, 18-19, 22, 23 стандартное отклонение составляет ≈ 619.9, для проб № 6, 10, 13-14, 17, 20-21 ≈ 615.4, для проб № 18, 19 ≈ 1556.1, для пробы № 12 ≈ 620. Полученные значения свидетельствуют о большой вариабельности данных.

При этом существенное различие между средним и медианой всех проб говорит об асимметричности распределений, а их визуальный анализ даёт сделать вывод о том, что распределение скошено вправо.

По построенным графикам квантиль-квантиль можно сделать предварительный вывод о том, что распределение данных проб № 1-23 не соответствует нормальному.

2.2. Проверка данных на нормальность распределения

Сначала необходимо проверить нормальность распределения данных каждой из проб. В данной работе был проведен тест на нормальность с помощью трех критериев: критерия Шапиро-Уилка, критерия Колмогорова-Смирнова, критерия Андерсона-Дарлинга.

При выборе критериев была использована таблица результатов исследования сравнительной мощности критериев нормальности распределения вероятностей случайных величин для различных альтернативных распределений, представленная в работе [5]. Критерии представлены в порядке предпочтения в зависимости от коэффициента эксцесса – числовой характеризующей степени остроты пика распределения случайной величины.

Коэффициент эксцесса распределения случайной величины x определяется формулой:

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3,$$

где $\mu_4 = \sum[(x - \sum x)^4]$ – четвёртый центральный момент случайной величины x, $\sigma^2 = D[x] = \sum[(x - \sum x)^2]$ – дисперсия или второй центральный момент случайной величины x.

Для полученных проб $\gamma_2 \in [2.52612156; 2.59823676]$. Таким образом, выбранные тесты обладают одними из наибольших мощностей.

Критерий Шапиро-Уилка используется для проверки гипотезы H_0 «случайная величина распределена нормально» и является наиболее эффективным по сравнению с альтернативными критериями проверки нормальности. Статистика критерия следующая:

$$W = \frac{[\sum_{i=1}^{n} a_{n-i+1}(x_{n-i+1} - x_i)]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $n = 330$.

Значения коэффициентов a_{n-i+1} заданы в таблице первоисточника [16]. Уровень значимости $W = 0.05$ Если $W < 0.05$, то нулевая гипотеза H_0

отклоняется. Таблицы коэффициентов a_{n-i+1} удобны лишь при малых значениях n. Существует аппроксимация, позволяющая применить данный критерий для больших значениях n (n>100).

Значение находится по формуле:

$$W_1 = \left(1 - \frac{0{,}6695}{n^{0{,}6518}}\right)\frac{s^2}{B},$$

где $B = \left\{\sum_{j=1}^{m} a_j * (x_{n-j} - x_j)\right\}^2$; $m = \left[\frac{n}{2}\right]$; $a_0 = \frac{0{,}899}{(n-2{,}4)^{0{,}4162}} - 0{,}02$;

$$a_j = a_0\left[z + \frac{1483}{(3-z)^{10{,}845}} + \frac{71{,}6*10^{-10}}{(1{,}1-z)^{8{,}26}}\right]; z = \frac{n-2j+1}{n-0{,}5}$$

Гипотеза отклонена для всех проб.

Критерий согласия Колмогорова-Смирнова – непараметрический критерий согласия, предназначенный для проверки простых гипотез о принадлежности анализируемой выборки некоторому известному закону распределения (нормальному).

В основе лежит статистика вида:

$$D = \sup_x |F_n(x) - F(x)|,$$

где sup S - точная верхняя грань множества S, F_n - функция распределения исследуемой совокупности, *F(x)* - функция нормального распределения.

Результаты критерия Колмогорова-Смирнова подтверждают факт о том, что распределение не является нормальным для всех проб.

Критерий согласия Андерсона-Дарлинга предназначен для проверки простых гипотез о принадлежности выборки некоторому закону распределения с известными параметрами, то есть для проверки гипотез вида $H_0: F_n(x) = \mathrm{F}(x)$.

Статистика критерия имеет следующий вид:

$$S = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln\left(\mathrm{F}(x_i)\right) + (1 - \frac{2i-1}{2n})\ln(1 - \mathrm{F}(x_i)\right\},$$

где n – объем выборки, $x_1, x_2, .., x_n$ – упорядоченные по возрастанию элементы выборки.

При уровне значимости = 0.05 гипотеза H_0 так же отклонена для всех проб.

Таким образом, все выбранные тесты на нормальность не подтвердили гипотезу о нормальном распределении каждой пробы.

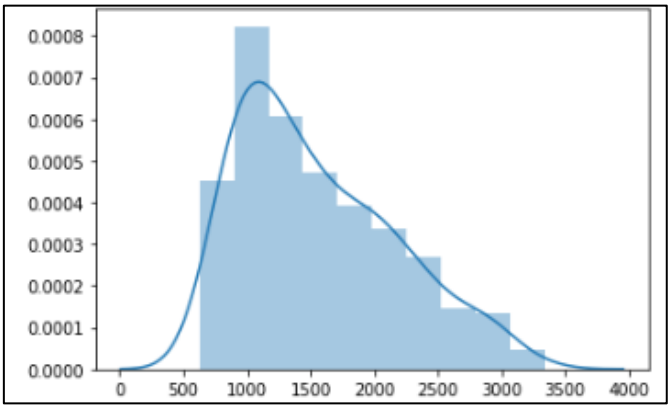

Рисунок 9 – Распределение данных пробы № 1

2.3. Приведение распределения отличного от нормального

Для приведения распределения отличного от нормального универсальным считается логарифмическое преобразование $\ln(x_i) = \widehat{x_i}$. Оно не дало результатов: распределение не стало нормальным ни для одной из проб.

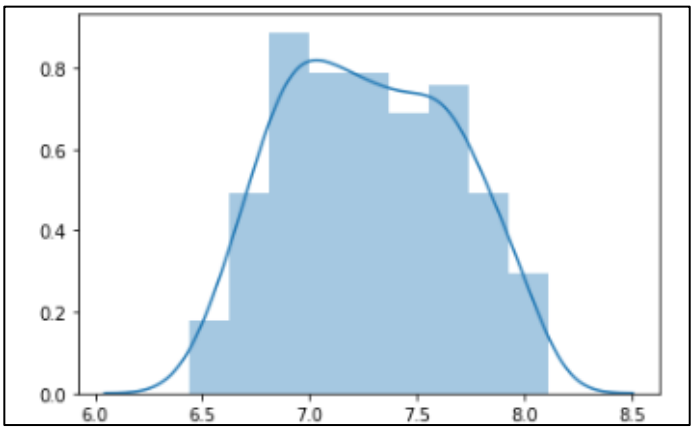

Рисунок 10 – Распределение данных пробы №1 после применения логарифмирования

Если истинное нормализующее преобразование неизвестно, наилучшим считается преобразование Бокса-Кокса (Box-Cox transformation - Box, Cox, 1964).

Пусть некоторая совокупность X представлена вектором непрерывных данных $x_i, i = 1,..n$. Преобразование определяется следующим образом:

$$x(\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \lambda \neq 1 \\ \ln(\lambda), & \lambda = 0 \end{cases} \quad (1)$$

Выражение (1) – это параметрическое семейство преобразований. Здесь используется значение λ, максимизирущее логарифм функции правдоподобия. Логарифм функции правдоподобия выгляди таким образом:

$$f(x,\lambda) = -\frac{n}{2} * \ln\left[\sum_{i=1}^{n} \frac{(x_i(\lambda)-\bar{x}(\lambda))}{n}\right] + (\lambda - 1) * \sum_{i=1}^{n} \ln(x_i),$$

где $\bar{x}(\lambda) = \frac{1}{N} * \sum_{i=1}^{n} x_i(\lambda)$ – среднеарифметическая Бокса-Кокса преобразованных данных.

В данном случае распределение так же не стало нормальным ни для одной из проб.

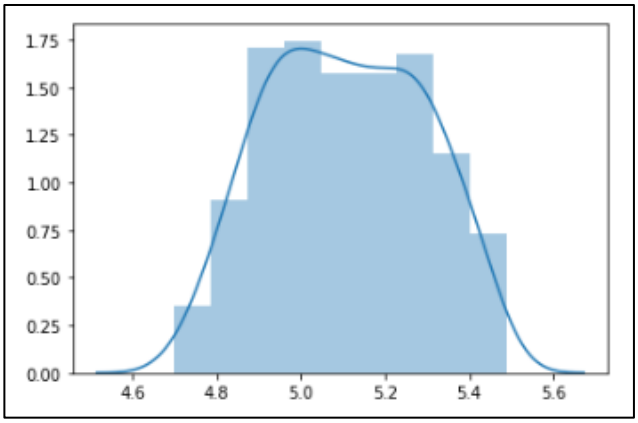

Рисунок 11 – Распределение данных пробы №1 после применения метода Бокса-Кокса

2.4. Определение взаимосвязи между пробами

Для определения статистической взаимосвязи между пробами была построенная корреляционная матрица – квадратная таблица, в которой на пересечении строки и столбца с номером пробы находится коэффициент корреляции между соответствующими пробами. Корреляционная матрица представляет собой симметричную квадратную матрицу размером 23*23, главная диагональ которой заполнена единицами, а недиагональные элементы представляют собой меру тесноты связи между парой проб.

В качестве показателя оценки тесноты связи был выбран коэффициент корреляции Спирмена, коэффициент корреляции Пирсона, коэффициент корреляции Кендалла.

Коэффициент корреляции r-Спирмена применяется для исследования корреляционной взаимосвязи между двумя ранговыми переменными. Расчет коэффициента состоит из следующих этапов:

1. Ранжирование признаков по возрастанию. Ранг – это порядковый номер. Если встречаются два одинаковых значения, им присваивают одинаковое значение ранга, равное среднему арифметическому рангов этих значений.
2. Вычисление разности рангов каждой пары рассматриваемых значений

$$d_i = d_x - d_y.$$

3. Возведение в квадрат найденного значения d_iи нахождение общей суммы $\sum {d_i}^2$.
4. Вычисление коэффициента корреляции рангов по следующей формуле:

$$r = 1 - 6\frac{\sum {d_i}^2}{n^3 - n},$$

где ${d_i}^2$ – квадрат разности между рангами, n – общее количество признаков.

Если связь между рассматриваемыми признаками имеет линейный характер, то коэффициент корреляции Пирсона точно устанавливает тесноту этой связи.

Выборочный коэффициент корреляции Пирсона определяется выражением:

$$r_{xy} = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 * \sum(y_i - \bar{y})^2}},$$

где $\bar{x}, \bar{y}$ – выборочные средние: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Еще одной мерой линейной связи между случайными величинами является коэффициент корреляции Кенделла является. Корреляция Кенделла так же является ранговой. Коэффициент инвариантен по отношению к любому монотонному преобразованию шкалы измерения.

Пусть заданы две выборки $x = (x_1, .. x_n), y = (y_1, .. y_n)$. Коэффициент корреляции Кенделла вычисляется по формуле:

$$\tau = 1 - \frac{4}{n(n-1)} R,$$

где $R = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left[[x_i < x_j] \neq [y_i < y_j] \right]$ – количество инверсий, образованных величинами y_i, расположенными в порядке возрастания соответствующих x_i. Коэффициент τ принимает значения из отрезка $[-1; 1]$. Равенство $\tau = 1$ указывает на строгую прямую линейную зависимость, $\tau = -1$ на обратную.

В таблице 2 представлен фрагмент попарной корреляции проб, вычисленной с помощью коэффициента корреляции Пирсона, в таблице 3 – с помощью коэффициента корреляции Спирмена, в таблице 4 – с помощью коэффициента корреляции Кендалла.

Таблица 2 – Значения коэффициента корреляции Пирсона для проб № 1-7

	Prob1	Prob2	Prob3	Prob4	Prob5	Prob6	Prob7
Prob1	1.000000	0.999998	1.000000	1.000000	1.000000	0.964909	1.000000
Prob2	0.999998	1.000000	0.999998	0.999998	0.999998	0.964834	0.999998
Prob3	1.000000	0.999998	1.000000	1.000000	1.000000	0.964860	1.000000
Prob4	1.000000	0.999998	1.000000	1.000000	0.999999	0.964856	1.000000

	Prob1	Prob2	Prob3	Prob4	Prob5	Prob6	Prob7
Prob5	1.000000	0.999998	1.000000	0.999999	1.000000	0.964858	1.000000
Prob6	0.964909	0.964834	0.964860	0.964856	0.964858	1.000000	0.964840
Prob7	1.000000	0.999998	1.000000	1.000000	1.000000	0.964840	1.000000

Таблица 3 – Значения коэффициента корреляции Спирмена для проб № 1-7

	Prob1	Prob2	Prob3	Prob4	Prob5	Prob6	Prob7
Prob1	1.000000	0.999996	0.999999	0.999999	0.999999	0.981707	0.999999
Prob2	0.999996	1.000000	0.999996	0.999996	0.999996	0.981703	0.999996
Prob3	0.999999	0.999996	1.000000	0.999999	1.000000	0.981707	0.999999
Prob4	0.999999	0.999996	0.999999	1.000000	0.999999	0.981707	0.999999
Prob5	0.999999	0.999996	1.000000	0.999999	1.000000	0.981707	0.999999
Prob6	0.981707	0.981703	0.981707	0.981707	0.981707	1.000000	0.981706
Prob7	0.999999	0.999996	0.999999	0.999999	0.999999	0.981706	1.000000

Таблица 4 – Значения коэффициента корреляции Кендалла для проб № 1-7

	Prob1	Prob2	Prob3	Prob4	Prob5	Prob6	Prob7
Prob1	1.000000	0.999568	0.999944	0.999944	0.999934	0.987710	0.999897
Prob2	0.999568	1.000000	0.999550	0.999550	0.999559	0.987330	0.999578
Prob3	0.999944	0.999550	1.000000	0.999925	0.999972	0.987729	0.999897
Prob4	0.999944	0.999550	0.999925	1.000000	0.999897	0.987691	0.999897
Prob5	0.999934	0.999559	0.999972	0.999897	1.000000	0.987701	0.999906
Prob6	0.987710	0.987330	0.987729	0.987691	0.987701	1.000000	0.986440
Prob7	0.999897	0.999578	0.999897	0.999897	0.999906	0.986440	1.000000

Коэффициент корреляции может принимать значения из интервала [-1;1], причём если значение находится ближе к 1, то это означает наличие сильной связи, а если ближе к 0, то слабой. Отрицательный коэффициент корреляции означает наличие противоположной связи: чем выше значение одной переменной, тем ниже значение другой.

Результаты по всем пробам следующие:

$r - Spearman's \in [0.981703; 1]$, $r - Pearson \in [0.964818; 1]$, $\tau - Kendall \in [0.987330; 1]$.

Для словесного описания величины коэффициента корреляции используются градации, представленные в таблице 4.

Таблица 4 – Словесная интерпретация коэффициентов корреляции

Значение	Интерпретация
до 0,2	Очень слабая корреляция
до 0,5	Слабая корреляция
до 0,7	Средняя корреляция
до 0,9	Высокая корреляция
свыше 0,9	Очень высокая корреляция

Из построенных корреляционных матриц можно сделать вывод о том, что пробы сильно коррелируют между собой.

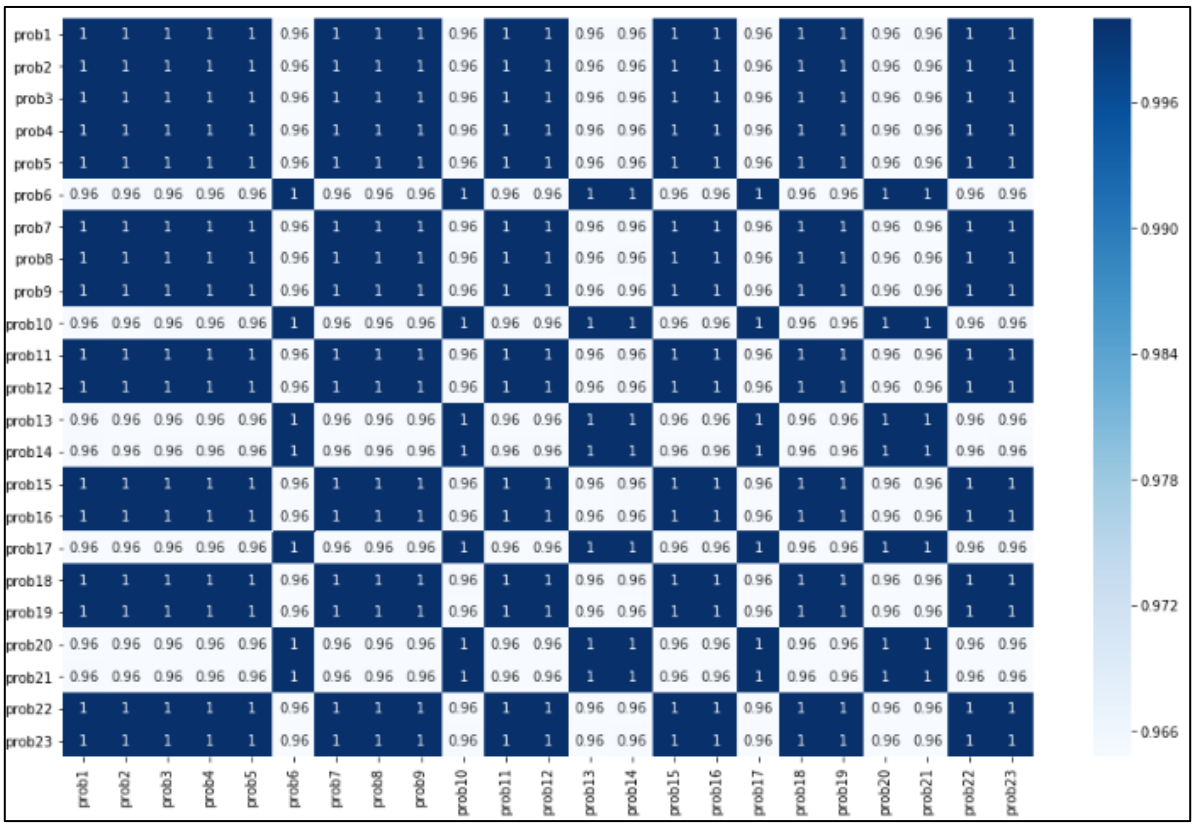

Рисунок 12 – Корреляционная матрица Пирсона

2.5. Обнаружение выбросов

Необходимым начальным этапом обработки данных перед дальнейшим использованием статистических методов является нормирование данных, то есть применение линейного преобразования всех значений признаков таким образом, чтобы значения признаков попадали в сопоставимые по величине интервалы по формуле:

$$z_i = \frac{x_i - \bar{x}}{s},$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – среднее арифметическое выборки, $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ – стандартное отклонение.

Для обнаружения выбросов был выбран метод Джона Тьюки, использующий квартили, которые являются стойкими к критическим величинам, подходящим в том случае, когда распределение не соответствует нормальному или неизвестное. Блочную диаграмму характеризуют пять величин:

- нижний квартиль (Q1) – величина, ниже которой лежит 25% значений из набора данных;
- Верхний квартиль (Q3) – величина, выше которой лежит 25% значений из набора данных;

Q1 и Q3 определяют границы ящика, внутри которого находятся 50% наблюдений.

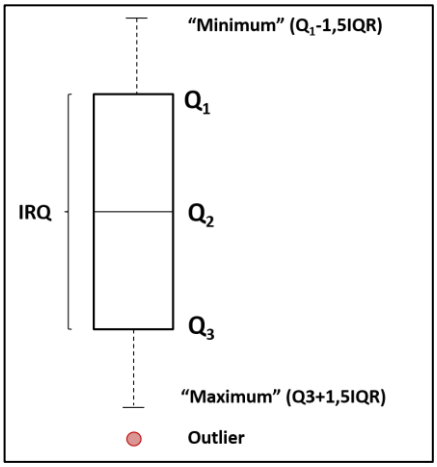

Рисунок 13 – Коробчатая диаграмма

Для выводов о наличии резко выделяющихся наблюдений в выборке находится интервальный размах: IQR=Q3-Q1.

Границы расположены на расстоянии 1.5 IQR ниже Q1 и выше Q3.

[Q1-1.5IQR; Q3+1.5IQR].

Концы усов – края статистически значимой выборки. Данные, выходящие за границы усов (выбросы), отображаются на графике в виде точек как представлено на рисунке 13.

Нет никакого статистического обоснования, почему Тьюки использует 1.5 как коэффициент перед IQR для построения границ [17] Было решено протестировать и другие коэффициенты перед IQR: 1, 1.2. Результаты приведены на рисунках 14-16, где так же отмечено выборочное среднее символом «●».

Рисунок 14 – Boxplot с границами [Q1-1IQR; Q3+1IQR]

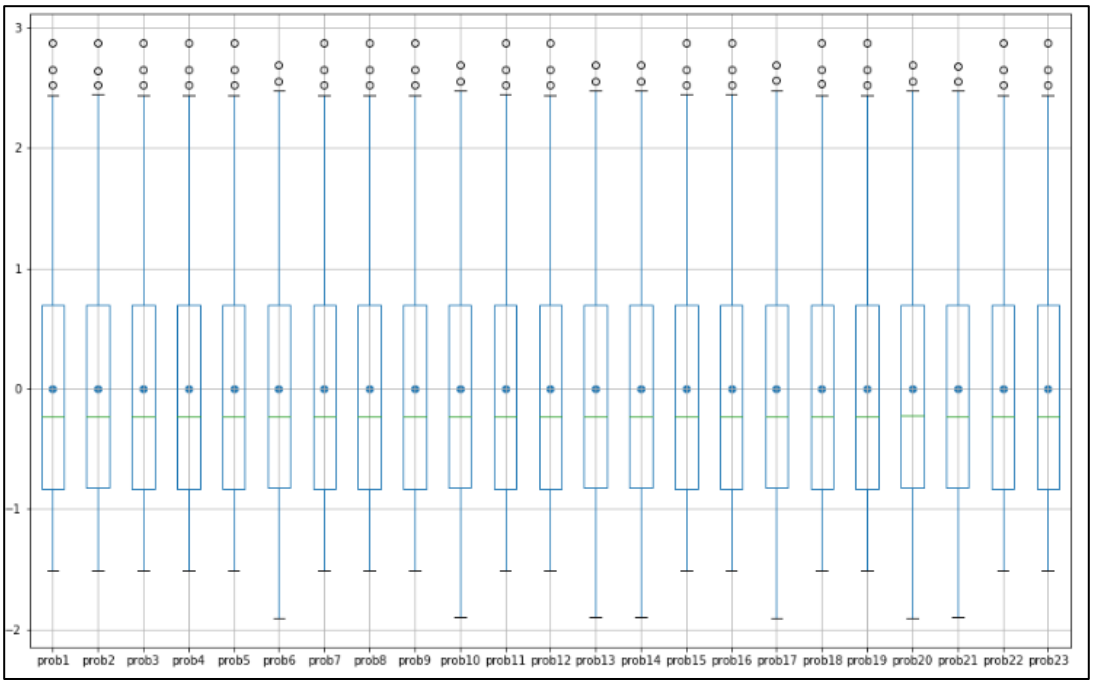

Рисунок 15 – Boxplot с границами [Q1-1.2IQR; Q3+1.2IQR]

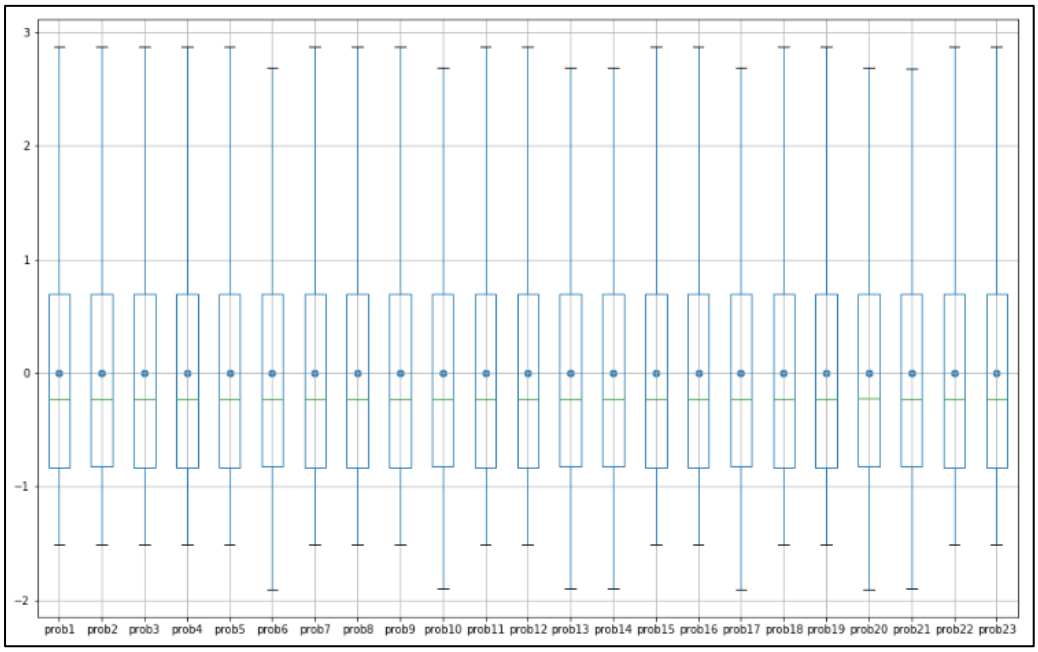

Рисунок 16 – Boxplot с границами [Q1-1.5IQR; Q3+1.5IQR]

Взяв границы [Q1-1.2IQR; Q3+1.2IQR] как оптимальные и очистив данные от выбросов, был проведен следующий этап работы – кластеризация.

Глава 3. Кластеризация

3.1. Постановка задачи и проблемы

Кластеризация – одна из важнейших задач, встречающихся при анализе данных. Кластерный анализ – это многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем разбивающая объекты на однородные в некотором смысле группы [6].

Формальная постановка задачи кластеризация следующая: пусть X – множество объектов, Y – множество номеров кластеров. Имеется функция расстояния между объектами $\rho(x, x')$. Существует некоторая конечная обучающая выборка объектов $X^m = \{x_1, .., x_m\} \subset X$. Необходимо разделить выборку на непересекающиеся подмножества (кластеры), таким образом, чтобы каждый кластер включал в себя объекты, близких по метрике ρ, и при этом объекты разных кластеров значительно отличались. Каждому объекту $x_i \in X^m$ приписывается номер кластера y_i.

Одной из основных проблем задачах кластеризации является проблема, связанная с экспоненциальным возрастанием количества данных из-за увеличения размерности пространства. В более широком смысле термин «Проклятие размерности», введенный Р. Беллманом в 1961 году [10], применяется по отношению ко всем «неудобным» или необычным свойствам многомерных пространств и к трудностям их исследования. «Проклятие размерности» особенно явно проявляется при работе со сложными системами, которые описываются большим числом параметров. В исследуемых данных после удаления выбросов осталось 324 параметра. Основная идея при решении данной проблемы – понизить размерность пространства, а именно спроецировать данные на подпространство меньшей размерности. Для решения данной проблемы был применен метод главных компонент (PCA) и агломеративный метод иерархической кластеризации (Feature Agglomeration).

3.2. Методы снижения размерности данных

Для снижения размерности данных был выбран и агломеративный метод иерархической кластеризации (Feature Agglomeration) и метод главных компонент (PCA).

Иерархическое кластеризация – это совокупность алгоритмов упорядочивания данных, направленных на создание иерархии (дерева) вложенных кластеров путем их последовательного слияния или разделения. В агломеративных методах иерархической кластеризации новые кластеры создаются путем объединения более мелких кластеров и, таким образом, дерево создается от листьев к стволу, то есть каждое наблюдение начинается в своем собственном кластере, затем кластеры последовательно объединяются.

Под дендрограммой обычно понимается дерево, построенное по матрице мер близости. Дендрограмма позволяет изобразить взаимные связи между объектами из заданного множества. Для создания дендрограммы требуется матрица сходства (или различия), которая определяет уровень сходства между парами кластеров. Для построения матрицы сходства (различия) необходимо задать меру расстояния между двумя кластерами [14].

Так как составленные корреляционные матрицы Спирмена, Пирсона, Кэндалла показали сильную корреляцию между пробами, для определения расстояния между кластерами был выбран метод Уорда, подходящий для подобных задач. Здесь в качестве расстояния между кластерами берётся прирост суммы квадратов расстояний объектов до центра кластера, получаемого в результате их объединения. Первоначально каждый объект формирует свой собственный кластер. Во всех кластерах для всех наблюдений вычисляются средние значения отдельных переменных. Далее рассчитываются квадраты евклидовых расстояний от отдельных наблюдений каждого кластера до кластерного среднего значения и

вычисленные дистанции суммируются. На каждом шаге алгоритма объединяются такие кластеры, которые дают наименьший прирост общей суммы дистанций.

Метод Feature Agglomeration основан на агломеративном методе иерархической кластеризации, но объединяет не сами объекты, а их параметры.

Принцип построения метода PCA заключается в переходе к новой системе координат с меньшим базисом. В новой системе координат главная ось задается линией, для которой сумма квадратов расстояний до всевозможных точек будет минимальна. Таким образом основная доля информации содержится именно в ней. Вторая ось ортогональна первой и задается таким же образом. Это позволяет выявить независимые друг от друга факторы. При наличии более двух факторов принцип построения главных компонент остается прежним. В результате, основная изменчивость исходного набора данных представлена несколькими первыми компонентами, и появляется возможность, отбросив менее существенные, перейти к пространству меньшей размерности.

3.3. Метрика качества кластеризации

Проблема оценки качества в задаче кластеризации трудноразрешима. Алгоритм кластеризации – это такая функция $\alpha: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие идентификатор кластера $y \in Y$. Иногда Множество Y известно заранее, но чаще ставится задача определить оптимальное число кластеров, используя тот или иной критерий качества кластеризации.

В 2015 году Д. Клейнберг вывел три аксиомы кластеризации и доказал теорему, связывающую эти свойства [13]:

- Аксиома № 1: алгоритм кластеризации α является масштабно инвариантным если для любой функции расстояния ρ и любой константы $\alpha>0$ результаты кластеризации с использованием расстояний ρ и $\rho*\alpha$ совпадают.
- Аксиома № 2: полнота. Множество результатов кластеризации алгоритма α в зависимости от изменения функции расстояния ρ должно совпадать со множеством всех возможных разбиений множества объектов X.
- Аксиома № 3: алгоритм кластеризации является согласованным, если результат кластеризации не изменяется после допустимого преобразования функции расстояния. Функция расстояния ρ' является допустимым преобразованием функции расстояния ρ, если:
  1. $\rho'(x_i, x_j) \leqslant \rho(x_i, x_j)$, если x_i и x_j лежат в одном кластере;
  2. $\rho'(x_i, x_j) \geqslant \rho(x_i, x_j)$, если x_i и x_j лежат в разных кластерах.

Аксиома № 1 говорит о том, что функция кластеризации должна быть нечувствительна к линейному преобразованию метрического пространства обучающей выборки и не должна зависеть от системы счисления функции расстояния. Аксиома № 2 постулирует о том, что алгоритм кластеризации должен разбить обучающую выборку на любое фиксированное количество кластером для функции расстояния ρ. Аксиома № 3 требует неизменного

количества кластеров как при уменьшении расстояния внутри кластера, так и при увеличении расстояния между кластерами.

Исходяи из этих свойств Клейнберг сформулировал и доказал теорему о невозможности: «Для множества объектов, состоящего из двух и более элементов, не существует алгоритма кластеризации, который был бы одновременно масштабно-инвариантным, согласованным и полным» [13].

Принято выделять две группы методов оценки качества кластеризации: Внешние меры основаны на сравнении результата кластеризации с априори известным разделением на класс, а внутренние меры отображают качество кластеризации только по информации в данных.

В задаче, поставленной в работе, количество кластеров известно, поэтому для выбора наилучшего метода кластеризации была введена внешняя метрика качества кластеризации.

Рассмотрим пары (x_i, x_j) из элементов кластеризуемого множества X. Подсчитаем количество пар, в которых:

- элементы принадлежат одному кластеру и одному классу – TP;
- элементы принадлежат одному кластеру, но разным классам – TN;
- элементы принадлежат разным кластерам, но одному классу – FP;
- элементы принадлежат разным кластерам и разным классам – FN.

Индекс Рэнда оценивает, насколько много из тех пар элементов, которые находились в одном классе, и тех пар элементов, которые находились в разных классах, сохранили это состояние после кластеризации алгоритмом [15].

$$Rand = \frac{TP + FN}{TP + TN + FP + FN}$$

Данная метрика имеет область определения от 0 до 1, где 1 – полное совпадение кластеров с заданными классами, а 0 – отсутствие совпадений.

3.4. Метод k-means

В данной работе была поставлена задача кластеризации, которая позволяет выделить однородные группы объектов. Для решения данной задачи был выбран метод k-means.

Метод k-средних – это один из методов кластерного анализа, целью которого является разбиение m наблюдений из пространства R^n на k кластеров таким образом, что каждое наблюдение относится к тому кластеру, к центроиду которого оно ближе всего.

В качестве меры близости используется Евклидово расстояние:

$$p(x, y) = \|x - y\| = \sqrt{\sum_{p=1}^{n} (x_p - y_p)^2},$$

где $x, y \in R^n$.

Рассмотрим ряд наблюдений $(x^{(1)}, x^{(2)}, .., x^{(m)}), x^{(j)} \in R^n$. Метод k-means разделяет m наблюдений на k групп (или кластеров) ($k \leq m$) $S = \{S_1, S_2, .. S_k\}$, чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центроидов этих кластеров:

$$\min \left[\sum_{i=1}^{k} \sum_{x^{(j)} \in S_i} \left\| x^{(j)} - \mu_i \right\|^2 \right],$$

где $x^{(j)} \in R^n, \mu_i \in R^n, \mu_i$ – центроид для кластера S_i.

Метод k-средних выполняет кластеризацию следующим образом:

1. Назначается число групп (k), на которые должны быть разбиты данные. Случайно выбирается k объектов исходного набора как первоначальные центры кластеров.

2. Каждому наблюдению присваивается номер группы по самому близкому центроиду, т.е. на основании наименьшего евклидова расстояния между объектом и точкой C_k.

3. Пересчитываются координаты центроидов μ_k всех k кластеров и вычисляются внутригрупповые разбросы (within-cluster variation) $W(C_k) = \sum x_i \in C_k (x_i - \mu_k)^2$. Если набор данных включает p переменных, то μ_k представляет собой вектор средних с p элементами.

4. Минимизируется общий внутригрупповой разброс $W_{total} = \sum_k W(C_k) \to min$, для чего шаги 2 и 3 повторяются многократно, пока назначения групп не прекращают изменяться или не достигнуто заданное число итераций.

3.5. Результаты

Для очищенных от выбросов данных была проведена кластеризация следующими способами:

- метод 1: k-means без уменьшения размерности данных;
- метод 2: k-means с уменьшением размерности данных методом PCA;
- метод 3: k-means с уменьшением размерности данных методом Feature Agglomeration.

Выбранная метрика качества кластеризации показала, что метод 3 является наиболее подходящим из выбранных: $Rand_3 = 0.6837944664031621$. При этом метод 1 оказался наименее подходящим. Метрика кластеризации методом 1 меньше метрики оценки алгоритма кластеризации методом 2: $Rand_1 = 0.5770750988142292 < Rand_2 = 0.6205533596837944$.

Результаты кластеризации методами 1-3 представлены в таблице 5. На рисунках 17-18 также отмечены центры кластеров.

Таблица 5 – Распределение проб по кластерам

№ пробы	Пласт	Кластеры для метода 1	Кластеры для метода 2	Кластеры для метода 3
1	АС10.4(6)	1	1	1
2	АС10.4(6)	2	3	1
3	АС10.1-3(1)	1	5	1
4	АС10.1-3(1)	1	5	1
5	АС10.4(1)	1	1	1
6	АС10.1-3(1)	3	4	2
7	АС10.1-3(1)	1	1	1
8	АС10.1-3(1)	5	5	1
9	АС10.1-3(1)	1	1	1
10	АС10.0.1(1)	3	4	2
11	АС10.0.1(1)	1	1	4
12	АС10.0.1(1)	5	5	1

№ пробы	Пласт	Кластеры для метода 1	Кластеры для метода 2	Кластеры для метода 3
13	AC10.1-3(1)	3	4	2
14	AC10.0.1(1)	0	0	0
15	AC10.0.1(1)	1	1	4
16	AC10.0.1(1)	1	1	4
17	AC12.3-5(4)	3	4	5
18	AC12.3-5(4)	1	1	1
19	AC12.3-5(4)	5	5	3
20	AC12.3-5(4)	3	0	0
21	AC12.3-5(4)	4	2	5
22	AC12.3-5(4)	1	1	4
23	AC12.1(2)	1	1	4

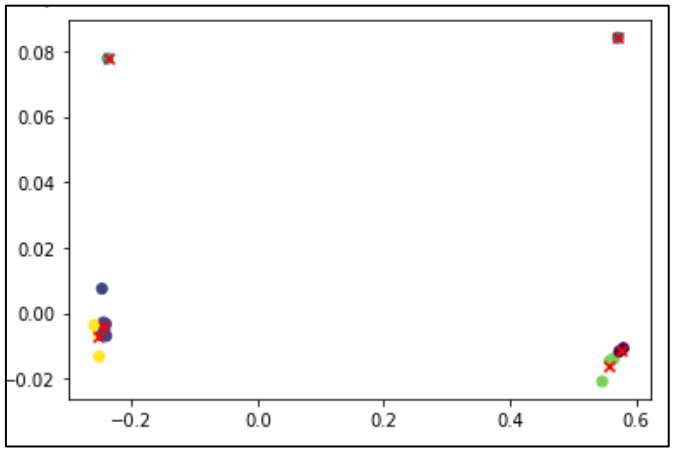

Рисунок 17 – Результаты k-means + PCA

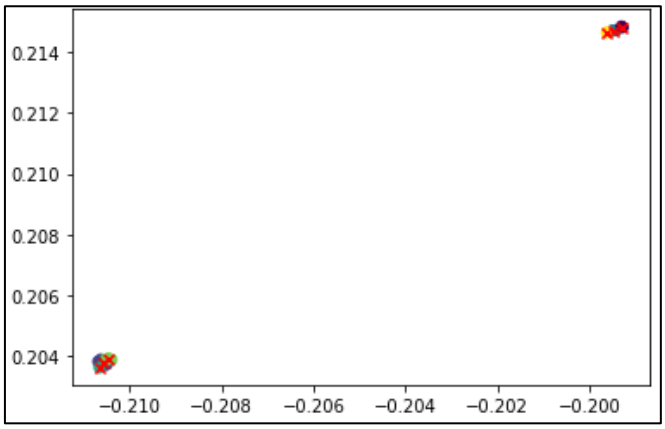

Рисунок 18 – Результаты k-means + Feature Agglomeration

Заключение

Целью данной работы был анализ проб нефти на наличие выбросов, а также кластеризация найденных проб.

При первичном анализе данных с помощью инструментов дескриптивной статистики был описан характер распределения проб, построенные графики квантиль-квантиль показали рассогласование исследуемого и нормального распределения в каждой пробе.

Построенные корреляционные матрицы Спирмена, Пирсона, Кендала показали, что полученные данные проб сильно коррелируют между собой.

На последующем этапе обработки полученных данных с помощью тестов Шапиро-Уилка, Колмогорова-Смирнова и Адерсона-Дарлинга нулевая гипотеза H_0 «случайная величина распределена нормально» была отклонена.

Универсальные преобразования логарифмирования и метод Бокса-Кокса не привели распределение данных к нормальному, поэтому, для следующего этапа работы – обнаружения выбросов, – был выбран метод Тьюки, использующийся в таких задачах, где распределение неизвестно или не соответствует нормальному. На основе экспериментов была произведена его модификация, определены подходящие границы для обнаружения и удаления выбросов.

Для снижения размерности данных был использован метод главных компонент (PCA) и агломеративный метод иерархической кластеризации (Feature Agglomeration).

Далее для «очищенных» данных была произведена кластеризация 3 способами: методом k-means без уменьшения размерности данных (метод 1), методом k-means с уменьшением размерности данных при помощи PCA (метод 2); методом k-means с уменьшением размерности данных при помощи Feature Agglomeration (метод 3).

Для оценки качества алгоритма была введена метрика оценки кластеризации Rand. На ее основе наиболее подходящий из рассмотренных в работе алгоритмов кластеризации оказалась кластеризация методом 3.

Тем не менее, рассмотренные алгоритмы не сгруппировали все пробы по кластерам, соответствующим пластам. Это может быть связано со специфичностью данных – малый объем выборки, зависимость результатов химического эксперимента от таких составляющих как точность оборудования, способ забора проб и другого. Альтернативными методами кластеризации являются метод нечеткой кластеризации c-means и Expectation-maximization (EM)-алгоритм. На большей выборке данных возможно использовать иерархические методы классификации.

Таким образом, пробы № 1-2 отнесены к пласту АС10.4(6) (кластер 1), пробы № 3-4, 6-9 отнесены к пласту АС10.1-3(1) (кластер 1) с вероятность 0.83. Проба № 5 так же отнесена к кластеру 1, но взята из пласта АС10.4(1). Проба № 23 отнесена к пласту АС12.1(2) (кластер 4). Пробы № 10-12, 14-16 отнесены к пласту АС10.0.1(1) (кластер 4) с вероятностью 0.5. С вероятностью 0.33 пробы № 17-22 отнесены к пласту АС12.3-5(4) (кластер 5).

Список литературы

1. ГОСТ 17567-81. Хроматография газовая. Термины и определения. – М.: Издательство стандартов, 1981. – с. 12.
2. Буре В.М., Парилина Е.М. Теория вероятностей и математическая статистика. – СПб.: Лань, 2013. – с. 334-338.
3. Гиматудинов Ш.К. Физика нефтяного и газового пласта. – М.: Недра, 1971. – 310 с.
4. Зейдель А.Н. Элементарные оценки ошибок измерений. – М.: Наука, 1965. – 96 с.
5. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ., 2006. – с. 220-221, 233, 238-241, 278.
6. Мандель И.Д. Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.
7. Руденко Б.А., Руденко Г.И. Высокоэффективные хроматографические процессы. – М.: Наука, 2003. – 425 с.
8. Шакирова Д.И., Рождественский Д.А. Газовая хроматография – Режим доступа: https://eurasiancommission.org/ru/act/texnreg/deptexreg/LS1/Documents/2.2.28%20Газовая%20хроматография.pdf (дата обращения 02.05.2020).
9. Шакирова Д.И., Рождественский Д.А. Хроматографические методы разделения. – Режим доступа: eurasiancommission.org/ru/act/texnreg/deptexreg/LS1/Documents/2.2.46%20Хроматографические%20методы%20разделения.pdf (дата обращения 02.05.2020).
10. Bellman R.E. Adaptive Control Processes. – Princeton University Press, Princeton, NJ, 1961. – 255 p.
11. Frigge M., Hoaglin, D., Iglewicz, B. Some Implementations of the Boxplot. – The American Statistician, 1989. – p. 120.

12. Iglewicz B., Hoaglin, D. How to detect and handle outliers. – ASQC Quality Press, 1993. – 458 p.

13. Kleinberg J. An impossibility theorem for clustering. – Режим доступа: https://www.cs.cornell.edu/home/kleinber/nips15.pdf (дата обращения 02.05.2020).

14. Lance G.N., Williams W.T. A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems. – The Computer Journal, V.9, I.4, 1967. – p. 373-380.

15. Rand W.M. Objective criteria for the evaluation of clustering methods. – Journal of the American Statistical Association. American Statistical Association, 1971. – p. 846-850.

16. Shapiro S.S., Wilk M.B. An analysis of variance test for normality (complete samples). – Biometrika, 1965. – 611 p.

17. Tukey J. Exploratory Data Analysis. – Addison Wesley Publishing Company, 1970. – 722 p.

Приложение 1 Фрагмент данных газовой хроматографии

	prob 1	prob 2	prob 3	prob 4	prob 5	prob 6	prob 7	prob 8	prob 9	prob 10	prob 11	prob 12	prob 13	prob 14	prob 15	prob 16	prob 17	prob 18	prob 19	prob 20	prob 21	prob 22	prob 23
comp 1	625, 9673 259	626, 0309 278	626, 0569 456	625, 9930 915	625, 9896 73	626, 0794 473	626, 0757 315	626, 1245 675	626, 1432 269	626, 0757 315	625, 9706 644	625, 9896 73	626, 0120 586	626, 0309 278	626, 0757 315	625, 9706 644	625, 9930 915	626, 1843 239	626, 0606 061	626, 0757 315	626, 1206 897	626, 0569 456	625, 9896 73
comp 2	660, 0171 969	660, 1374 57	660, 1380 5	660, 0172 712	660, 1549 053	660, 1036 269	660, 1549 053	660, 2076 125	660, 1380 5	660, 1549 053	660, 0517 688	659, 9827 883	660, 1205 857	660, 1374 57	660, 1549 053	660, 0517 688	660, 1036 269	660, 2067 183	660, 0865 801	660, 0688 468	660, 1724 138	660, 0517 688	659, 9827 883
comp 3	670, 0773 861	670, 2749 141	670, 2329 594	670, 1208 981	670, 1376 936	670, 2936 097	670, 2237 522	670, 3287 197	670, 2329 594	670, 2237 522	670, 1466 782	670, 0516 351	670, 1981 051	670, 1890 034	670, 2237 522	670, 1466 782	670, 2072 539	670, 3703 704	670, 2164 502	670, 2237 522	670, 2586 207	670, 2329 594	670, 1376 936
comp 4	682, 9750 645	683, 1615 12	683, 1751 51	683, 0742 66	683, 0464 716	683, 2469 775	683, 2185 886	683, 3044 983	683, 2614 323	683, 1325 301	683, 0888 697	682, 9604 131	683, 1180 017	683, 1615 12	683, 2185 886	683, 0888 697	683, 0742 66	683, 2902 67	683, 2034 632	683, 1325 301	683, 2758 621	683, 1751 51	683, 0464 716
comp 5	690, 1117 799	690, 2920 962	690, 3364 97	690, 2417 962	690, 2753 873	690, 4145 078	690, 3614 458	690, 3979 239	690, 4227 783	690, 2753 873	690, 2502 157	690, 0172 117	690, 2670 112	690, 2920 962	690, 3614 458	690, 2502 157	690, 2417 962	690, 4392 765	690, 3896 104	690, 2753 873	690, 4310 345	690, 3364 97	690, 1893 287
comp 6	722, 9408 792	723, 1158 321	723, 0496 454	722, 9640 87	722, 9873 418	723, 1589 639	723, 0496 454	723, 1199 187	723, 0730 223	723, 0496 454	722, 9716 024	722, 8744 939	723, 0886 076	723, 0886 076	723, 0886 076	723, 0106 437	723, 0339 929	723, 1237 323	723, 0573 895	723, 0263 158	723, 1744 422	723, 0730 223	722, 9640 87
comp 7	725, 9727 135	726, 1507 334	726, 1398 176	726, 0495 701	726, 1265 823	726, 2569 832	726, 1904 762	726, 2703 252	726, 2170 385	726, 1398 176	726, 1156 187	725, 9615 385	726, 1772 152	726, 1772 152	726, 1772 152	726, 1023 822	726, 1288 686	726, 3184 584	728, 4916 201	726, 1639 676	726, 2677 485	726, 1663 286	726, 0495 701

	prob 1	prob 2	prob 3	prob 4	prob 5	prob 6	prob 7	prob 8	prob 9	prob 10	prob 11	prob 12	prob 13	prob 14	prob 15	prob 16	prob 17	prob 18	prob 19	prob 20	prob 21	prob 22	prob 23
comp 8	728, 2465 892	726, 1507 334	728, 4194 529	728, 2751 644	728, 4050 633	728, 5424 073	728, 4701 114	728, 5060 976	728, 4989 858	728, 4194 529	728, 3975 659	728, 1882 591	728, 4050 633	726, 1772 152	728, 4556 962	728, 3324 886	728, 4119 736	728, 6004 057	728, 4916 201	728, 4412 955	726, 2677 485	728, 4482 759	728, 3257 461
comp 9	732, 2890 349	732, 5240 263	732, 4721 378	732, 3216 995	732, 4556 962	732, 6053 834	732, 5734 549	732, 5711 382	732, 5557 809	732, 5227 964	732, 4036 511	732, 2368 421	732, 5063 291	732, 5569 62	732, 5569 62	732, 3872 276	732, 4708 27	732, 6572 008	732, 5545 962	732, 4898 785	732, 6572 008	732, 5050 71	732, 3722 812
comp 10	737, 0389 085	737, 2787 051	737, 2340 426	737, 1269 6	737, 2151 899	737, 4301 676	737, 3353 597	737, 3475 61	737, 3225 152	737, 2847 011	737, 2210 953	736, 9939 271	737, 2658 228	737, 3164 557	737, 3164 557	737, 1515 459	737, 2399 797	737, 4239 351	737, 3285 932	737, 2469 636	737, 4746 45	737, 2718 053	737, 1269 6
comp 11	744, 5679 636	744, 8153 768	744, 7821 682	744, 6130 501	744, 7594 937	744, 9466 734	744, 8834 853	744, 9186 992	744, 8782 961	744, 8328 267	744, 7261 663	744, 5344 13	744, 8101 266	744, 8607 595	744, 8607 595	744, 7034 972	744, 7995 941	745, 0304 26	744, 8958 862	744, 7874 494	745, 0304 26	744, 8275 862	744, 6636 318
comp 12	755, 3815 058	755, 6904 401	755, 6231 003	755, 4881 133	755, 5949 367	755, 8151 346	755, 7244 174	755, 7926 829	755, 7302 231	755, 6737 589	755, 5780 933	755, 3643 725	755, 6455 696	755, 6962 025	755, 7468 354	757, 4252 408	757, 6357 179	755, 8316 43	755, 7135 602	755, 6174 089	755, 8316 43	755, 6795 132	755, 4881 133
comp 13	755, 3815 058	755, 6904 401	757, 5987 842	755, 4881 133	757, 5696 203	755, 8151 346	757, 6494 428	760, 7723 577	757, 7079 108	755, 6737 589	757, 5557 809	755, 3643 725	757, 6202 532	755, 6962 025	755, 7468 354	757, 4252 408	757, 6357 179	755, 8316 43	760, 7414 931	755, 6174 089	755, 8316 43	755, 6795 132	757, 4102 175
comp 14	760, 6366 852	760, 7486 09	760, 6889 564	760, 6474 456	760, 7088 608	760, 8430 675	760, 7396 15	760, 7723 577	760, 7505 071	760, 7396 15	760, 6997 972	760, 6275 304	760, 7594 937	760, 8101 266	760, 8101 266	760, 7197 162	760, 7305 936	760, 8012 17	760, 7414 931	760, 7287 449	760, 8012 17	760, 6997 972	760, 6474 456

Приложение 2 Код программы

Программа реализована на языке Python.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn import preprocessing
from sklearn.cluster import KMeans
from numpy.random import normal
from scipy.stats import boxcox, shapiro, kstest, pearsonr, normaltest,
kurtosis, anderson
import pylab as pb
from scipy import stats
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import normalize
from sklearn.decomposition import PCA
from sklearn.cluster import FeatureAgglomeration

data = pd.read_csv('RI_compare.csv')
data.head(10)

# Minuses remove
data = data.drop(axis = 0, index = [326, 328])

# NA remove
data.dropna(inplace = True)

#Descriptive statistics
data.describe()

#Exapmple of histogram prob1
sns.distplot(data['prob1'])

#Find kurt to choose normality test
kurtosis(data,  fisher=False)

#Check data for normality using Anderson-Darling criterion
for i in range(1, data.shape[1]):
    if anderson(data['prob' + str(i)],dist='norm')[1][2] < 0.05:
        print('normal')
    else:
        print('not normal')

#Check data for normality using Kolmogorov-Smirnov criterion
for i in range(1, data.shape[1]):
    if kstest(data['prob' + str(i)], 'norm')[1] < 0.05:
        print('not normal')
    else:
        print('normal')

#Check data for normality using Pearson criterion
for i in range(1, data.shape[1]):
    if normaltest(data['prob' + str(i)])[1] < 0.05:
        print('not normal')
    else:
```

```
        print('normal')

#Check data for normality using Shapiro-Wilk criterion
for i in range(1, data.shape[1]):
    if shapiro(data['prob' + str(i)])[1] < 0.05:
        print('not normal')
    else:
        print('normal')

#Normal QQ Plot and General QQ Plot
for i in range(1, data.shape[1]):
    stats.probplot(data['prob' + str(i)], dist="norm", plot=pb)
    plt.title('Prob ' + str(i))
    pb.show()

#Pairwise Pearson correlation coefficient
corel=data.corr(method ='pearson')
print(corel)

#Pairwise Spearman correlation coefficient
corel=data.corr(method ='kendall')
print(corel)

#Pairwise Spearman correlation coefficient
corel=data.corr(method ='spearman')
print(corel)

#Pairwise Spearman correlation coefficient
corel=data.corr(method ='kendall')
print(corel)

#Log transformation
for i in range(1, 24):
    if shapiro(np.log(data['prob' + str(i)]))[1] < 0.05:
        print('not normal')
    else:
        print('normal')

#Example of prob1 after log transformation
sns.distplot(np.log(data.prob1))

#Boxcox transformation
data_boxcox = data.copy()
for column in data_boxcox:
    data_boxcox[column] = boxcox(data_boxcox[column])[0]

for i in range(1, 24):
    if shapiro(data_boxcox['prob' + str(i)])[1] < 0.05:
        print('not normal')
    else:
        print('normal')

#Example of prob1 after Boxcox transformation
sns.distplot(data_boxcox.prob1)

#Normalize data
```

```
data = (data - data.mean(axis = 0)) / data.std(axis = 0)
data.head(10)

#Correlation matrix
plt.figure(figsize=(16,10))
cor=data.corr(method ='pearson')
sns.heatmap(cor, annot = True, cmap = plt.cm.Blues)
plt.show()

#Box-and-whiskers diagrams1
plt.figure(figsize=(16,10))
data.boxplot(whis = 1.5)
#Mark the mean
means = data.mean()
pb.scatter(list(range(1, 24)), means)

#Box-and-whiskers diagrams2
plt.figure(figsize=(16,10))
data.boxplot(whis = 1)
#Mark the mean
means = data.mean()
pb.scatter(list(range(1, 24)), means)

#Box-and-whiskers diagrams3
plt.figure(figsize=(16,10))
data.boxplot(whis = 1.2)
#Mark the mean
means = data.mean()
pb.scatter(list(range(1, 24)), means)


#boxplot vs histogram
f, (ax_box, ax_hist) = plt.subplots(2, sharex=True,
gridspec_kw={"height_ratios": (.15, .85)})
f, (ax_box, ax_hist) = plt.subplots(2, gridspec_kw={"height_ratios":
(.15, .85)})

sns.boxplot(data.prob1, whis = 1.5, ax=ax_box)
sns.distplot(data.prob1, ax=ax_hist)
plt.xlim()
plt.show()

#metric
def metric(y_true, y_pred):
  TP = 0
  TN = 0
  FP = 0
  FN = 0
  for i in range(len(y_true)):
    for j in range(i + 1, len(y_true)):
      if y_true[i] == y_true[j] and y_pred[i] == y_pred[j]:
        TP += 1
      if y_true[i] == y_true[j] and y_pred[i] != y_pred[j]:
        FP += 1
      if y_true[i] != y_true[j] and y_pred[i] == y_pred[j]:
        TN += 1
```

```
      if y_true[i] != y_true[j] and y_pred[i] != y_pred[j]:
        FN += 1
  return (TP + FN) / (TP + TN + FP + FN)


#Cut outliers
data_wo_outliers=data.copy()
min_max = 2.5
data = data[ data < min_max]

data.dropna(inplace = True)

#k-means w/o outliers
kmeans = KMeans(n_clusters = 6, random_state = 5)
data_tr = data_wo_outliers.transpose()
kmeans.fit(data_tr)
data.cluster = kmeans.labels_

kmeans.labels_

y_pred =  kmeans.labels_

#metric for k-means w/o outliers
metric(y_true, y_pred)

#PCA
d = pd.read_csv('RI_compare.csv')
# Minuses remove
d = d.drop(axis = 0, index = [326, 328])
# NA remove
d.dropna(inplace = True)

scaler = StandardScaler()
d_scaled = scaler.fit_transform(d)
d_normalized = normalize(d_scaled)
d_normalized = pd.DataFrame(d_normalized)

d_normalized = d_normalized.T
pca = PCA(n_components = 2)
d_principal = pca.fit_transform(d_normalized)
d_principal = pd.DataFrame(d_principal)
d_principal.columns = [1st principal component','2nd principal component']
print(d_principal.head(3))
print(len(d_principal))

kmeans = KMeans(n_clusters = 6, random_state = 5)
kmeans.fit(d_principal)
data.cluster = kmeans.labels_

plt.scatter(d_principal['1st principal component'], d_principal['2nd
principal component'], c=data.cluster, label='o')
plt.scatter(kmeans.cluster centers .T[0], kmeans.cluster centers .T[1]
,marker="x", color="red")
plt.title('Result of  k-means + PCA')

kmeans.labels_
```

```
y_pred =  kmeans.labels_

#metric for k-means with pca
metric(y_true, y_pred)

#FeatureAgglomeration
d = pd.read_csv('RI_compare.csv')
# Minuses remove
d = d.drop(axis = 0, index = [326, 328])
# NA remove
d.dropna(inplace = True)

scaler = StandardScaler()
d_scaled = scaler.fit_transform(d)
d_normalized = normalize(d_scaled)
d_normalized = pd.DataFrame(d_normalized)

d_normalized = d_normalized.T
fa = FeatureAgglomeration(2)
d_principal = fa.fit_transform(d_normalized)
d_principal = pd.DataFrame(d_principal)
d_principal.columns = ['P1', 'P2']
print(d_principal.head(3))
print(len(d_principal))

kmeans = KMeans(n_clusters = 6, random_state = 5)
kmeans.fit(d_principal)
data.cluster = kmeans.labels_

plt.scatter(d_principal['P1'], d_principal['P2'], c=data.cluster,
label='o')
plt.scatter(kmeans.cluster_centers_.T[0], kmeans.cluster_centers_.T[1]
,marker="x", color="red")
plt.title('Result of  k-means + Feature Agglomeration')

kmeans.labels_

y_pred = kmeans.labels_

#metric for k-means with FeatureAgglomeration
metric(y_true, y_pred)
```