Статистический анализ изменений частот сенсоров

Анализ матриц отдельных веществ

1. Тест распределения изменений частоты на нормальность

Большинство статистических методов анализа данных исходят из предположения о том, что данные имеют нормальное распределение, поэотму необходимо начать с проверки гипотезы о нормальном распределении отсчетов временных рядов. В пакете научных вычислений для Python Scipy [1] реализован [2] обобщённый тест Д'Агостино-Пирсона [3]. Графики построены с помощью графика квантилей, реализованного в виде функции qqplot [4] в пакете прикладной статистики statsmodels [5].

Беглый анализ графиков показывает: графики имеют S-образную форму, отсюда можно сделать вывод о том, что гипотеза о нормальном распределении отсчётов временных рядов отвергается для большинства веществ. Проверим это с помощью теста Харке-Бера. Сохраним результаты в Pandas DataFrame, а затем установим уровень значимости равным 5%: если значение столбца p-value превысит этот порог, то гипотеза о нормальности распределения отвергается для рассматриваемого временного ряда.

С помощью функции norm_analisys, которая выполняет анализ полученного dataframe, выясним, что в собранном массиве данных существует 48 рядов, для которых гипотеза о нормальном распределении отвергается с вероятностью 95%. Видно, что это верно не для целых матриц веществ, а для показаний отдельных сенсоров.

2. Тест стационарности

Выполним вычисление скользящих статистик (среднего и стандартного отклонения) и тест на единичные корни. В пакете statsmodels реализован расширенный тест Дики-Фуллера (ADF — Augmented Dickey-Fuller) [6]. Его преимущество перед обычным тестом Дики-Фуллера (DF) состоит в том, что, благодаря включению первых разностей, он даёт возможность работать с авторегрессиями не только первого, но и более высоких порядков, так как авторегрессия еще не была исследована. В результате обнаружено, что существует 228 нестационарных рядов из 288 — то есть можно сделать вывод, что большинство из них нестационарны, и, следовательно, большинство рядов

будут соответствовать моделям авторегрессий более высокого порядка.

3. Автокорреляционные и частичные автокорреляционные функции, радиус автокорреляции

По автокоррелограмме можно сделать вывод о периодичности рассматриваемых рядов и выявить наличие взаимосвязанных изменений ряда.

Коррелограмма негладкая для: ацетальдегида 0 (S7), ацетальдегида 3 (S2, S3, S7), ацетон 6 (S1, S7), бензин 7 (S2, S7), диоктилфталат 15 (S1), диоктилфталат 16 (S7), диоктилфталат 17 (S7), диоктилфталат 18 (S1, S7), диоктилфталат 20 (S7), диоктилфталат 21 (S2, S4, S7, S8), диоктилфталат 22 (S1, S4, S7, S8), диоктилфталат 23 (S7), этилацетат 32 (S1), этилацетат 33 (S2), этилацетат 35 (S2), гексан 14 (почти весь), пластизоль 26 (S2, S4, S5), пластизоль 27 (S3, S4, S7) — всего 32 вектора. Можно предположить наличие скрытых синусоидальных зависимостей.

Анализ радиусов автокорреляции показывает, что переход через 0 в большинстве графиков происходит в районе 40 отсчёта. Однако существуют такие графики, в которых поведение автокорреляционной функции резко отличается: ацетальдегид 0 (S7), бензин 7 (S2), гексан 14 (S2, S7), диоктилфталат 15 (S1, S7), диоктилфталат 17 (S7), диоктилфталат 18 (S7), диоктилфталат 22 (S1, S4, S7, S8), пластизоль 26 (S2, S4, S8), пластизоль 27 (S3, S5, S6, S8), этилацетат 32 (S1).

Наблюдаются пересечения между множеством векторов, которые имеют сильно отличающуюся от остальных коррелограмму, и множеством векторов с сильно отличающимся вектором автокорреляции.

4. Графы кросс-корреляции сенсоров

Для каждого вещества построены полносвязные взвешенные графы с 8 вершинами (S1, …, S8), где вершина sj (j=1…8) соответствует сенсору, а вес ребра ij присваиваются согласно значению корреляции между сенсорами si и sj.

Цвет ребра показывает, какому промежутку значений функции взаимной корреляции принадлежит данное ребро. Для наглядности рассматриваются следующие промежутки: вес ребра больше 0.9 – это сильно положительно

коррелированные сенсоры, вес в интервале (0.3, 0.9] – некое промежуточное, среднее значение корреляции, (0, 0.3] – слабая положительная корреляция. Аналогично для отрицательной корреляции. Положительная корреляция отражена оттенками красного, отрицательная – оттенками синего.

Вещество	**N**	**Тип**	**Характеристики**
Ацетальдегид	4	+	2 графа, содержащие исключительно сильную положительную корреляцию (№1 и №2); 2 графа, содержащие сильную положительную и преобладающую среднюю корреляцию (№0 и №3);
Ацетон	3	+	2 графа, содержащие исключительно сильную положительную корреляцию (№4 и №5); 1 граф, содержащий преобладающую сильную положительную и среднюю корреляцию (№6) от сенсора **S7** до каждого сенсора в виде **остовного дерева**;
Бензин	1	+	1 граф, преобладает средняя положительная корреляция (№7);
Бензол	4	+	2 графа, исключительно сильную положительную корреляцию (№9 и №11); 2 графа, содержащие преобладающую сильную положительную и среднюю корреляцию (№8 и №10), в обоих графах присутствуют ребра со слабой корреляцией (S3, S6) и (S3, S1);
Бутанол	1	+	1 граф, только сильная положительная корреляция (№12);
Бутилацетат	1	+	1 граф, только сильная положительная корреляция (№13);
ДОФ	9	+\-	Во всех рассматриваемых графах присутствуют ребра всех цветов, кроме темно-синего цвета, соответствующего максимальной отрицательной корреляции; В 7 графах из 9 (№№ 15, 16, 17, 19, 20, 21, 23) наблюдается явное **остовное дерево** из вершины **S7**, аналогичное ацетону, и представленное одним из трех видов корреляции (слабой положительной, средней отрицательной). В 2 оставшихся графах данное остовное дерево так же представлено, но, видимо, рассматриваются некие пограничные значения, поэтому отдельные ребра этого дерева ребра попадают в смежные интервалы (№18 – средняя и слабая положительная, №22 – слабая положительная и слабая отрицательная). Тем не менее, остовное дерево явно выделяется.
Этилацетат	4	+	2 графа, содержащий сильную положительную и преобладающую среднюю корреляцию (№32 и №35); 2 графа, содержащие исключительно сильную положительную корреляцию (№33, №34); Средняя положительная корреляция всегда представлена ребрами (S2, S5), (S2, S8). В графе №33 представлено почти полное **остовное дерево** из вершины **S2**, состоящее из ребер со слабой положительной корреляцией и одного ребра со средней (пограничные значения?)
Фенол	1	+	1 граф, только сильная положительная корреляция (№31);
Гексан	1	+\-	1 граф, содержащий сильную положительную, среднюю положительную, слабую положительную, слабую отрицательную (№14) корреляции;

			наблюдается явное **остовное дерево** из вершины **S7**, аналогичное ацетону, и представленное слабой отрицательной корреляцией; наблюдается неполное **остовное дерево** из вершины **S2**, представленное слабой и средней положительной корреляцией (вершины S3 и S4 имеют среднюю корреляцию, возможны пограничные значения); наблюдается клика (S4, S5, S6);
Изобутанол	1	+	1 граф, только сильная положительная корреляция (№24);
Изопропанол	1	+	1 граф, только сильная положительная корреляция (№25);
Пластизоль	2	+\-	2 графа, преобладает средняя положительная корреляция; Наблюдается **остовное дерево** из вершины **S7**, состоящее из ребер со слабой и средней отрицательной корреляцией (№26) слабой и средней отрицательной и слабой положительной корреляцией (№27)
Пропанол	1	+	1 граф, только сильная положительная корреляция (№28);
Стирол	1	+	1 граф, преобладает сильная положительная корреляция (№29);
Толуол	1	+	1 граф, только сильная положительная корреляция (№30);

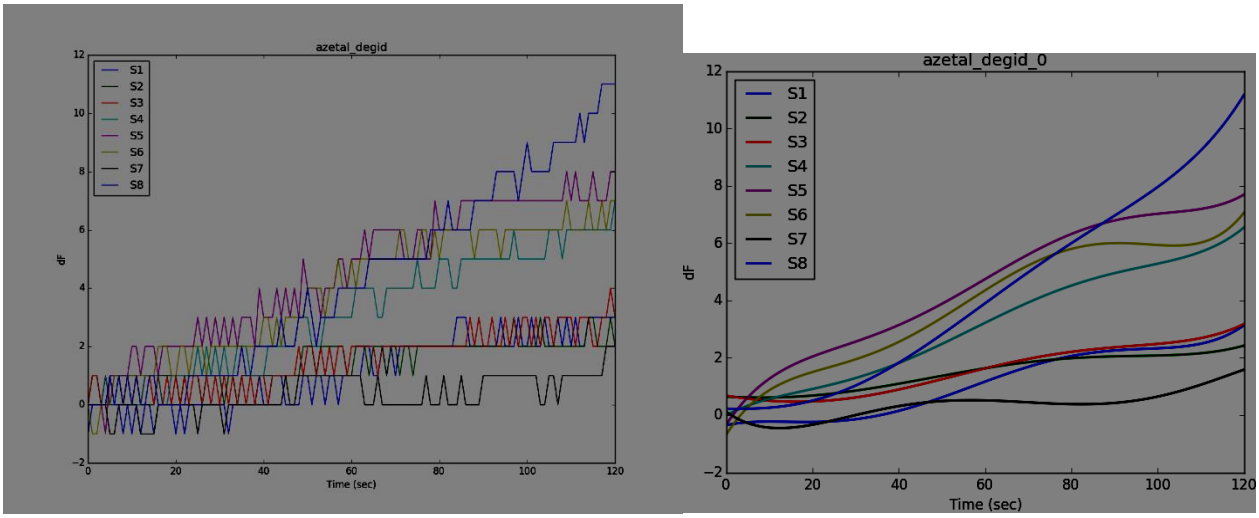
Таблица 1 – Сравнение графов кросс-корреляций сенсоров

Согласно таблице 1, для нас абсолютно неразличимы некоторые виды ацетальдегида, ацетона, бензола, бутанол, бутилацетат, фенол, изобутанол, изопропанол, пропанол, толуол, поскольку сенсоры у них имеют сильную взаимную корреляцию, а графы кросс корреляции идентичны.

Выявлены 2 остовных дерева – и вершины S2 и из вершины S7 (рис.1).

5. Сглаживание полиномом N-степени

Исходные матрицы визуализированы с помощью кода visualization.py для каждого рассматриваемого вещества. Анализ результатов показал, что из-за постоянных колебаний пьезосенсора в пределах 1 Гц, полученные графики непригодны для работы. Приближение полиномом N степени позволило



сгладить такой «шум» и оставить только общий вид графиков. Улучшение заметно на следующем примере. Справа исходный график, слева – сглаженный полиномом 5 степени.

Анализ полученных графиков приведен в таблице 2.

Вещество	**N**	**Тип**	**Характеристики**
Ацетальдегид	4	+	Визуально очень похожи графики №1 и №2; Между графиками №0 и №3 сходство проследить сложно; ы
Ацетон	3	+	Визуально похожи графики №4 и №5; График №6 значительно отличается от №4 и №5;
Бензин	1	+	**Сложно сказать (№7);**
Бензол	4	+	Визуально похожи графики (№8, №9 и №10); **Сложно сказать (№11);**
Бутанол	1	+	**Сложно сказать (№12);**
Бутилацетат	1	+	**Сложно сказать (№13);**
ДОФ	9	+\-	Среди рассматриваемых графиков нет сильно отличающихся от остальных. Есть сходство между ДОФ
Этилацетат	4	+	(№32 и №35); (№33, №34); В графе №33 представлено почти полное **остовное дерево** из вершины **S2**
Фенол	1	+	**Сложно сказать** (№31);
Гексан	1	+\-	**Сложно сказать** (№14) наблюдается явное **остовное дерево** из вершины **S7**, наблюдается неполное **остовное дерево** из вершины **S2**
Изобутанол	1	+	**Сложно сказать** (№24);
Изопропанол	1	+	**Сложно сказать** (№25);
Пластизоль	2	+\-	Наблюдается **остовное дерево** из вершины **S7**, (№26) (№27)
Пропанол	1	+	**Сложно сказать** (№28);
Стирол	1	+	**Сложно сказать** (№29);
Толуол	1	+	**Сложно сказать** (№30);

Таблица 2 – Сравнительный анализ кривых веществ

6. Авторегрессионные модели

Поскольку не все векторы стационарны, целесообразно остановить свой выбор на модели Бокса-Дженсинса [7] или ARIMA (Autoregressive Integrated Moving Average) [8]. Модель, реализованная в пакете statsmodels, предлагает

удобный интерфейс как для задания параметров модели [9], так и для применения этой модели к данным [10]. Функция fit позволяет настроить такие глобальные параметры: “на лету” приводить исследуемый ряд к стационарному (*transparams*), выбрать один из трех методов работы с максимальным правдоподобием (*method*), учет тренда (*trend*), солвер (*solver*), количество итераций (*maxiter*).

Текущие глобальные параметры: transparams=True (defalut), так как не все исследуемые временные ряды стационарны, method=css-mle (default) – условный метод максимального правдоподобия, trend=c (default) – учитывать константу, solver=newton.

Подбор параметров p, d, q будет осуществляться автоматически из следующих интервалов: $p \in [0,3], d \in [0,2], q \in [0,3]$, поскольку более интеллектуальный выбор параметров затрудняется большим количеством векторов (8 векторов из 36 матриц веществ – 288 векторов), поведение которых отличается. В качестве критерия отбора выбрана минимизация информационного критерия Акаике.

Дальнейшие эксперименты будут проводиться с увеличением порядка модели.

Анализ матриц смесей

1. Тест распределения на нормальность

Беглый анализ графиков показывает: графики имеют выпуклую форму, отсюда можно сделать вывод о том, что гипотеза о нормальном распределении отсчётов временных рядов отвергается для большинства веществ. Проверим это с помощью теста Харке-Бера. Сохраним результаты в Pandas DataFrame, а затем установим уровень значимости равным 5%: если значение столбца p-value превысит этот порог, то гипотеза о нормальности распределения отвергается для рассматриваемого временного ряда.

С помощью функции norm_analisys, которая выполняет анализ полученного dataframe, выясним, что в собранном массиве данных существует 12 рядов из 32, для которых гипотеза о нормальном распределении отвергается с вероятностью 95%. Видно, что это верно не для целых матриц веществ, а для

показаний отдельных сенсоров.

2. Тест стационарности

Выполним вычисление скользящих статистик (среднего и стандартного отклонения) и тест на единичные корни. В результате обнаружено, что существует 24 нестационарных ряда из 32 — то есть можно сделать вывод, что большинство из них нестационарны, и, следовательно, большинство рядов будут соответствовать моделям авторегрессий более высокого порядка. Смесь диоктилфталата и этилацетата полностью нестационарна.

3. Автокорреляционные и частично автокорреляционные функции

Коррелограмма негладкая для: ДОФ+ацетальдегида 3 (S1, S2, S3), ДОФ+ацетон 0 (S1, S2, S3, S4, S7, S8), ДОФ+бензол 2 (S1) — всего 10 векторов. Можно предположить наличие скрытых синусоидальных зависимостей. Все коррелограммы. ДОФ+этилацетат 1 гладкие.

Анализ радиусов автокорреляции показывает, что переход через 0 в большинстве графиков происходит в районе 43-44 отсчёта. Однако существуют такие графики, в которых поведение автокорреляционной функции резко отличается: ДОФ+ацетон 0 (S1), но отклонение здесь не значительно (32 отсчет).

4. Графы кросс-корреляций

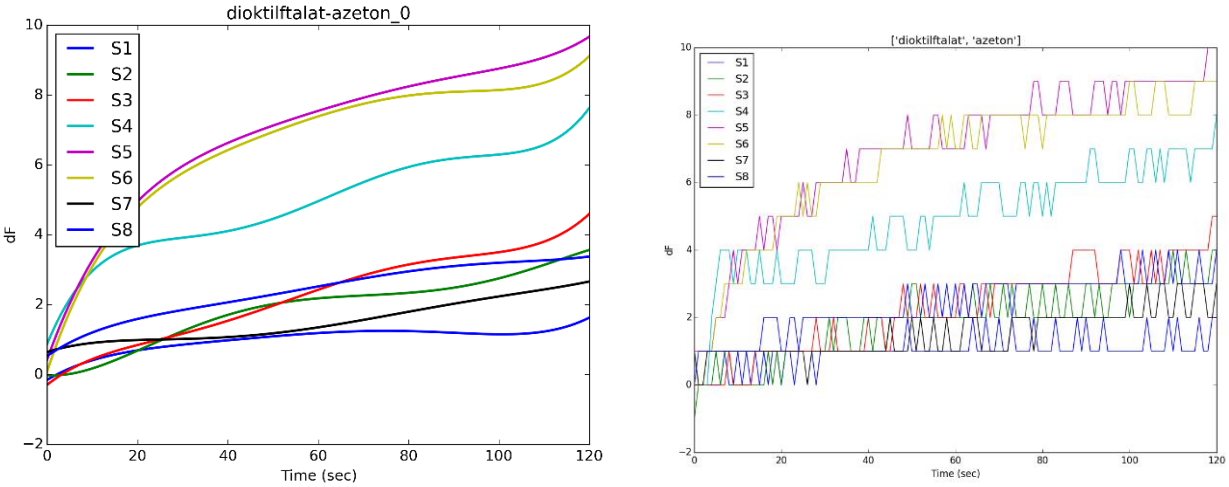
Для каждого вещества построены полносвязные взвешенные графы с 8 вершинами (S1, …, S8), где вершина sj (j=1…8) соответствует сенсору, а вес ребра ij присваиваются согласно значению корреляции между сенсорами si и sj.

Цвет ребра показывает, какому промежутку значений функции взаимной корреляции принадлежит данное ребро. Для наглядности рассматриваются следующие промежутки: вес ребра больше 0.9 – это сильно положительно коррелированные сенсоры, вес в интервале (0.3, 0.9] – некое промежуточное, среднее значение корреляции, (0, 0.3] – слабая положительная корреляция. Аналогично для отрицательной корреляции. Положительная корреляция отражена оттенками красного, отрицательная – оттенками синего.

Вещество	N	Тип	Характеристики
ДОФ+ацетон 0	1	+	В графе преобладает средняя положительная корреляция;
ДОФ+этилацетат 1	1	+	Сложно определить, какой вид ребер преобладает – сильная положительная или средняя положительна корреляция. Наблюдается остовное дерево из сенсора **S2** с одним ребром пограничной корреляции.
ДОФ+бензол 2	1	+	В графе преобладает средняя положительная корреляция;
ДОФ+ацетальдегида 3	1	+	В графе преобладает средняя положительная корреляция; Наблюдаются остовное дерево из сенсора **S2;**

5. Сглаживание полиномом N-степени

Исходные матрицы визуализированы с помощью кода visualization.py для каждого рассматриваемого вещества. Анализ результатов показал, что из-за постоянных колебаний пьезосенсора в пределах 1 Гц, полученные графики непригодны для работы. Приближение полиномом N степени позволило сгладить такой «шум» и оставить только общий вид графиков. Улучшение заметно на следующем примере. Справа исходный график, слева – сглаженный полиномом 5 степени.



Анализ полученных графиков приведен в таблице 2.

Вещество	N	Характеристики
ДОФ+ацетон 0	1	
ДОФ+этилацетат 1	1	
ДОФ+бензол 2	1	
ДОФ+ацетальдегида 3	1	

6. Авторегрессионные модели

Текущие глобальные параметры: transparams=True (defalut), так как не

все исследуемые временные ряды стационарны, method=css-mle (default) – условный метод максимального правдоподобия, trend=c (default) – учитывать константу, solver=newton.

Подбор параметров p, d, q будет осуществляться автоматически из следующих интервалов: $p \in [0,3], d \in [0,2], q \in [0,3]$, поскольку более интеллектуальный выбор параметров затрудняется большим количеством векторов (8 векторов из 36 матриц веществ – 288 векторов), поведение которых отличается. В качестве критерия отбора выбрана минимизация информационного критерия Акаике.

Дальнейшие эксперименты будут проводиться с увеличением порядка модели.

Обзор кода

1. Тест распределения на нормальность

   Код: `arima.py`, функция `norm_test`

   Выход: `graphs/norm` – содержит графики квантилей

   Код: `arima.py`, функция `jarque_bera_test`

   Выход: `jarque_bera.txt` – содержит таблицу с выводом результатов теста Харке-Бера

   Код: `arima.py`, функция `jarque_bera_test_analisys`

   Выход: в консоль – количество векторов, чье распределение не соответствует нормальному.

2. Тест стационарности

   Код: `arima.py`, функция `test_stationarity`

   Выход: `graphs/stat` – содержит графики скользящих статистик (математического ожидания и стандартного отклонения) и аутпут теста Дики-Фуллера

   Код: `arima.py`, функция `a_dickey_fully_test`

   Выход: adf_protocol.txt – содержит таблицу с результатами (единичный корень и стационарность)

Код: arima.py, функция a_dickey_fully_test_analisys

Выход: в консоль – количество нестационарных векторов

3. Автокорреляционные и частично автокорреляционные функции

   Код: arima.py, функция autocorr

   Выход: graphs/auto – содержит графики автокорреляции и частичной автокорреляции

4. Графы кросс-корреляций

   Код: arima.py, функция cross_corr

   Выход: графики в graphs/crosscorr содержат графы кросскорреляции сенсоров

5. Сглаживание полиномом N-степени

   Код: arima.py, функция fit_polynom

   Выход: графики в graphs/poly содержат графики, сглаженные полиномом 5 степени

6. Авторегрессионные модели

   Код: arima.py, функция arima_find_best

   Выход: arima_est.txt, содержит таблицу с наилучшими параметрами.

Ссылки

1. https://www.scipy.org/

2. https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.normaltest.html

3. D'Agostino, R. B. (1971), "An omnibus test of normality for moderate and large sample size," Biometrika, 58, 341-348

4. http://statsmodels.sourceforge.net/devel/generated/statsmodels.graphics.gofplots.qqplot.html

5. http://statsmodels.sourceforge.net/

6. http://statsmodels.sourceforge.net/devel/generated/statsmodels.tsa.stattools.adfuller.html

7. Box, G.E.P., and Jenkins, G., (1970) Time Series Analysis, Forecasting and Control, Holden-Day, San Francisco.

8. Box, G.E.P., and Pierce, D.A., (1970) "Distribution of the Residual Autocorrelations in Autoregressive-Integrated Moving-Average Time Series Models", Journal of the American Statistical Association, 65, 1509-1526.

9. http://statsmodels.sourceforge.net/devel/generated/statsmodels.tsa.arima_model.ARIMA.html

10. http://statsmodels.sourceforge.net/devel/generated/statsmodels.tsa.arima_model.ARIMA.fit.html