Санкт-Петербургский политехнический университет Петра Великого

Физико-механический институт

Высшая школа прикладной математики и вычислительной физики

Интервальный анализ
Отчёт по курсовой работе

Выполнил:

Студент: Дамаскинский Константин
Группа: 5040102/10201

Принял:

к. ф.-м. н., доцент
Баженов Александр Николаевич

2023 г.

Содержание

Список иллюстраций

Список таблиц

1. Постановка задачи

Проводится исследование из области солнечной энергетики. На рис. 1 показана схема установки для исследования фотоэлектрических характеристик.

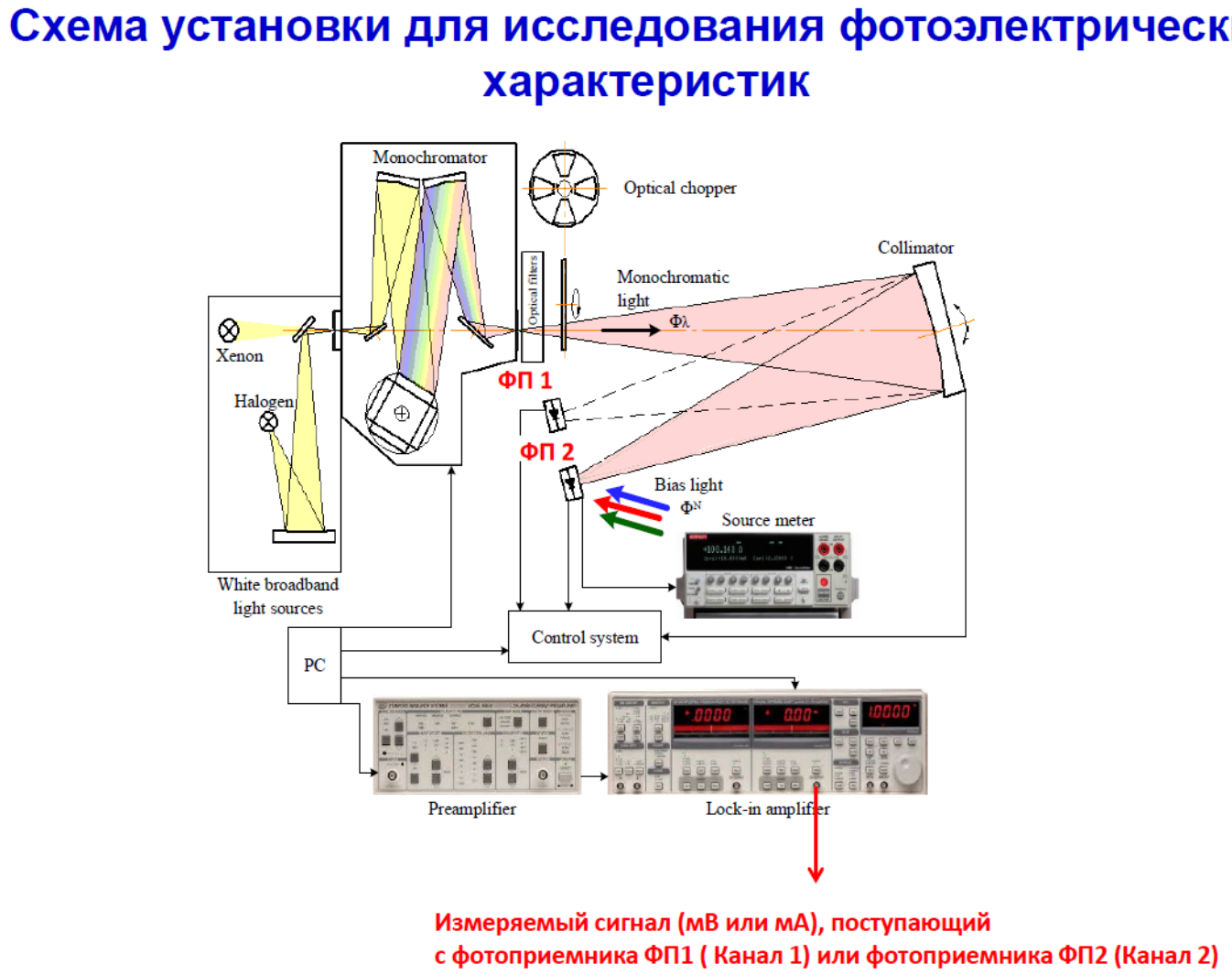

Рис. 1. Схема установки для исследования фотоэлектрических характеристик

Калибровка датчика ФП2 производится по эталону ФП1. Зависимость между квантовыми эффективностями датчиков предполагается постоянной для каждой пары наборов измерений:

$$QE_2 = \frac{I_2}{I_1} \cdot QE_1 \tag{1}$$

где QE_2, QE_1 – эталонная эффективность эталонного и исследуемого датчика, I_2, I_1 – измеренные токи. Данные с датчиков находятся в файлах **Канал2_800nm_0.2.csv**, **Канал1_800nm_0.2.csv** и полагаются линейными.

Требуется определить коэффициент калибровки:

$$R_{21} = \frac{I_2}{I_1} \tag{2}$$

- На основе линейной регрессии на множестве интервальных данных и коэффициента Жаккара
- На основе линейной регрессии на множестве интервальных данных

Также требуется построить информационное множество данной задачи и коридор совместных зависимостей для двух выборок, определить статус измерений.

2. Теория

2.1. Простая линейная регрессия для вещественных данных

Пусть заданы две последовательности $X = \{x_i\}_{i=1}^n, Y = \{y_i\}_{i=1}^n,\ x_i, y_i \in \mathbb{R}\ \forall i = \overline{1,n}$. **Простой линейной регрессией** для этих последовательностей называется функция:

$$f(x) = \beta_0 + \beta_1 \cdot x \tag{3}$$

подобранная так, чтобы вектор $F = \{f(x_i)\}_{i=1}^n$ был в каком-то смысле максимально близок к вектору Y.

Таким образом, для решения задачи простой линейной регрессии необходимо найти коэффициенты β_0, β_1. В зависимости от выбираемого метода поиска коэффициентов будет меняться и мера близости подобранной линейной функции к вектору Y.

В данной работе будет использоваться метод наименьших квадратов (МНК). Данный метод позволяет решить задачу простой линейной регрессии, поставив задачу минимизации второй (евклидовой) нормы разности векторов F и Y:

$$\sum_{i=1}^{n} \|\beta_0 + \beta_1 x_i - y_i\|_2 \underset{\beta_0,\beta_1}{\longrightarrow} \min \tag{4}$$

2.2. Обынтерваливание данных для интервальной регрессии

Поскольку показания датчиков обладают погрешностью, полученные данные на самом деле следует рассматривать как интервалы, центр которых совпадает со считанными показаниями, а радиус равен некоторой базовой погрешности ε, умноженной на вес w_i. ε является константой.

Для каждого из наборов данных $X^{(1)}$ и $X^{(2)}$, прочитанных из соответствующих файлов, построим простую линейную регрессию на вещественных числах в результате чего получим аппроксимацию:

$$Lin_k(i) = a_i^{(k)} \cdot i + b_i^{(k)},\ k \in \{1, 2\},\ i = \overline{1, n} \tag{5}$$

Определим для каждой из выборки вектор весов W_k простым способом: если значение аппроксимирующей прямой Lin_k в точке i не попадает в интервал $x_i^{(k)} \pm \varepsilon$, то увеличим радиус интервала в $w_i^{(k)}$ раз так, чтобы $Lin_k(i)$ оказалось на одной из границ интервала.

После того, как мы получили два интервальных вектора из $\mathbb{IR}^n$, вычтем из $x_i^{(k)}$ "наклонную" составляющую $a_i^{(k)} \cdot i$, получив таким образом "горизонтальные" векторы, для которых будем находить искомый коэффициент пропорциональности R_{21}.

2.3. Коэффициента Жаккара. Поиск R_{21}

Коэффициент Жаккара позволяет оценить, насколько хорошо совмещаются друг с другом заданные интервалы $x_1, \ldots, x_n$. Вычисляется путём деления длины интервала-пересечения на длину интервала объединения по формуле:

$$JK(x_1, \ldots, x_n) = \frac{wid\left(\bigcap\limits_{i=\overline{1,n}} x_i\right)}{wid\left(\bigcup\limits_{i=\overline{1,n}} x_i\right)} \tag{6}$$

Используя данный коэффициент, мы можем подобрать такой $R_{21} \in \mathbb{R}$, чтобы полученные интервалы X_2 и $R_{21} \cdot X_1$ были максимально совместны. Для этого необходимо вычислять коэффициент Жаккара для совокупности компонент этих векторов.

Таким образом, для того, чтобы найти R_{21}, необходимо задать нижнюю и верхнюю границы поиска $\underline{R}, \overline{R}$, а затем при помощи бинарного поиска найти точку максимума коэффициента Жакккара в зависимости от выбранного R_{21}.

Числа $\underline{R}, \overline{R}$ можно найти тривиально, поделив наименьшую верхнюю границу среди интервалов вектора $R_{21} \cdot X_1$ на наибольшую нижнюю границу среди интервалов вектора X_2 и, соответственно, наибольшую на наименбшую соответствующие границы.

2.4. Схема решения задачи интервальной регрессии

Будем, как и в прошлой работе, отдельно решать задачу интервальной регресии для двух наборов входных данных $(I, \mathbf{Y_1}), (I, \mathbf{Y_2})$. Здесь I – номера измерений, $\mathbf{Y_1}, \mathbf{Y_2}$ – обынтервaленные измеренные значения. В отличие от первой работы, будем решать эти задачи как задачи интервальной, а не вещественной регрессии, описанным ниже способом.

Далее, аналогично предыдущей работе, найдём оптимальный коэффициент R_{21}, максимизируя коэффициент Жаккара.

2.5. Интервальная регрессия как задача оптимизации

В данной работе для решения задачи интервальной регрессии будем использовать следующий подход.

Будем искать зависимость $y^{(k)} = \beta_0^{(k)} + \beta_1^{(k)} x$ таким образом, чтобы, минимально расширив интервалы исходного интервального вектора $\{\mathbf{y_i}\}_{i=1}^n$, получить набор интервалов, накрывающий аппроксимирующую прямую:

$$\begin{cases} \mathtt{midy}_\mathbf{i}^{(\mathbf{k})} - w_i^{(k)} \cdot \mathtt{rady}_\mathbf{i}^{(\mathbf{k})} \le \beta_0^{(k)} + \beta_1^{(k)} i \le \mathtt{midy}_\mathbf{i}^{(\mathbf{k})} + w_i^{(k)} \cdot \mathtt{rady}_\mathbf{i}^{(\mathbf{k})} & , \ i = \overline{1,n} \\ \sum\limits_{i=1}^{n} w_i^{(k)} \longrightarrow \min & \\ w_i^{(k)} \ge 0 & , \ i = \overline{1,n} \\ w^{(k)}, \beta_0^{(k)}, \beta_1^{(k)} =? & \end{cases} \tag{7}$$

Здесь $k \in \{1, 2\}$ – номер набора данных.

Данная задача является задачей линейного программирования. Как и в прошлой работе, примем $\varepsilon := \mathtt{rady}_\mathbf{i}^{(\mathbf{k})} = 10^{-4}$ для всех $i = \overline{1,n}$.

2.6. Информационное множество

Применительно к данной задаче, информационное множество – это все такие пары (β_0, β_1), при которых выполнено первое ограничение типа неравенства задачи оптимизации 7.

2.7. Коридор совместных зависимостей

В постановке задаче оптимизации 7 не ставится никаких ограничений и целей по минимизации для параметров β_0, β_1. Ясно, что параметры β_0, β_1, полученные в результате решения задачи оптимизации, будут не единственными допустимыми: информационное множество задает целое семейство допустимых β_0, β_1. Следовательно, имеет смысл рассматривать, как единое целое, множество всех функций, совместных с интервальными данными задачи восстановления зависимостей. Такое множество называется **коридором совместности**. **Граничными** называются измерения, определяющие какой-либо фрагмент границы множества. Это свойство имеет смысл рассматривать для наблюдений, принадлежащих выборке, по которой строилась модель. Граничные измерения задают минимальную подвыборку, определяющую модель.

2.8. Классификация измерений

В задаче интервальной регрессии важно классифицировать измерения по влиянию на итоговому модель. Мы будем разделять измерения следующим образом.

- Внутренние – это такие измерения, добавление которых в существующую модель не изменяет её (её информационное множество)
- Внешние – такие измерения, добавление которых в существующую модель изменяет её информационное множество

У внутренних и внешних измерений имеются важные частные случаи:

- Граничные – измерения, определяющие какой-либо фрагмент границы информационного множества задачи. Стоит заметить, что удаление из модели внутренних, но не граничных измерений, не изменит её
- Выбросы – такие измерения, которые делают информационное множество пустым

Для того, чтобы определить, к какому классу принадлежит очередное измерение, достаточно соотнести его с прогнозом существующей модели в данной точке.

- Внутреннее измерение полностью содержит в себе прогнозный интервал
- Граничное измерение имеет с ним общий конец
- Внешнее интервальное измерение не содержит в себе полностью прогнозный интервал
- Если пересечение внешнего интервального измерения с прогнозным интервалом пустое, то измерение – это выброс

2.9. Размах и относительный остаток

Для дальнейшего анализа измерений введём следующие понятия.

Определение 1 Размах (плечо). Размах – величина, показывающая, как соотносится ширина прогнозного коридора и полученного интервала в данной точке:

$$\ell(x, \mathbf{y}) = \frac{\mathtt{rad}\Upsilon(x)}{\mathtt{rady}} \tag{8}$$

Определение 2 Относительный остаток. Относительный остаток показывает, как соотносится расстояние между центром измерения и прогнозного коридора в данной точке и радиусом измерения:

$$r(x, \mathbf{y}) = \frac{\mathtt{midy} - \mathtt{mid}\Upsilon(x)}{\mathtt{rady}} \tag{9}$$

Для внутренних измерений, содержащих в себе прогнозный интервал, выполняется неравенство:

$$|r(x, \mathbf{y})| \leq 1 - \ell(x, \mathbf{y}) \tag{10}$$

Точное равенство будет выполнено исключительно для граничных наблюдений.

Выбросы удовлетворяют условию:

$$|r(x, \mathbf{y})| > 1 + \ell(x, \mathbf{y}) \tag{11}$$

Интервальные измерения, у которых величина неопределённости меньше, чем ширина прогнозного интервала, то есть плечо больше единицы, оказывают сильное влияние на модель. Их называют **строго внешними**.

2.10. Диаграмма статусов для интервальных измерений

На диаграмме статусов в зелёной области лежат внутренние измерения, в жёлтой – внешние, за вертикальной чертой $\ell = 1$ – строго внешние измерения. Наблюдения, расположенные на границе зеленой зоны, являются граничными.

Диаграмма статусов строится по каждому каналу. Для этого необходимо произвести следующие шаги:

1. Выполняется кусочно-линейная интервальная регрессия
2. Из обынтерваленных входных данных вычитается центральная часть полученной аппроксимации
3. Строится прогноз на всю выборку по центральной регрессии. Его используем для вычисления плеча и относительного остатка

3. Реализация

Данная работа реализована на языке программирования Python 3.10 с использованием пакетов `numpy` и `scikit`. Также использовался модуль `interval` вычислительного пакета Octave и библиотека программ С. Жилина. Код данного отчёта подготовлен с использованием редактора TeXstudio и компилятора pdflatex.

4. Результаты

Ниже приведены графики, полученные в результате работы реализованной программы.

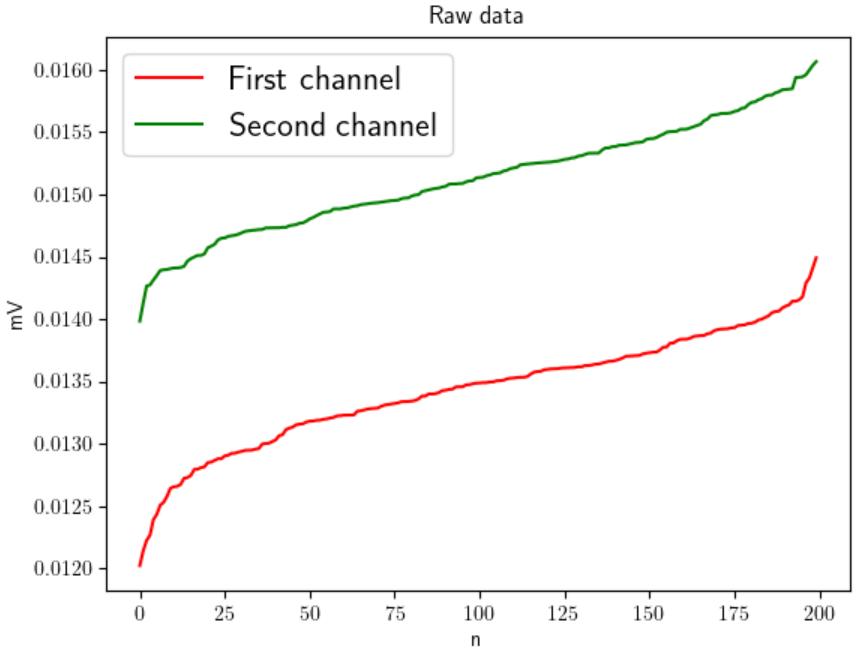

Рис. 2. Загруженные данные

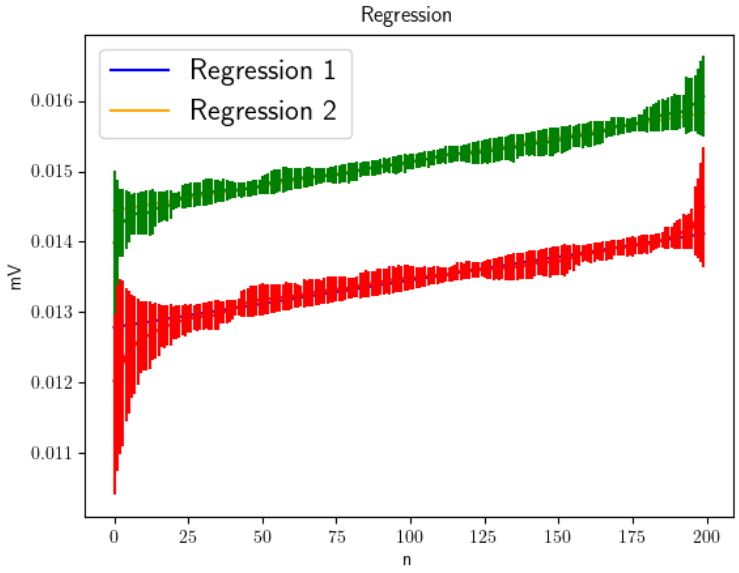

Рис. 3. Линейная регрессия для вещественных данных и результаты обынтерваливания

N	β_0	β_1
1	0.012	6.67e-6
2	0.014	6.96e-6

Таблица 1. Параметры линейной регрессии для двух входных наборов данных

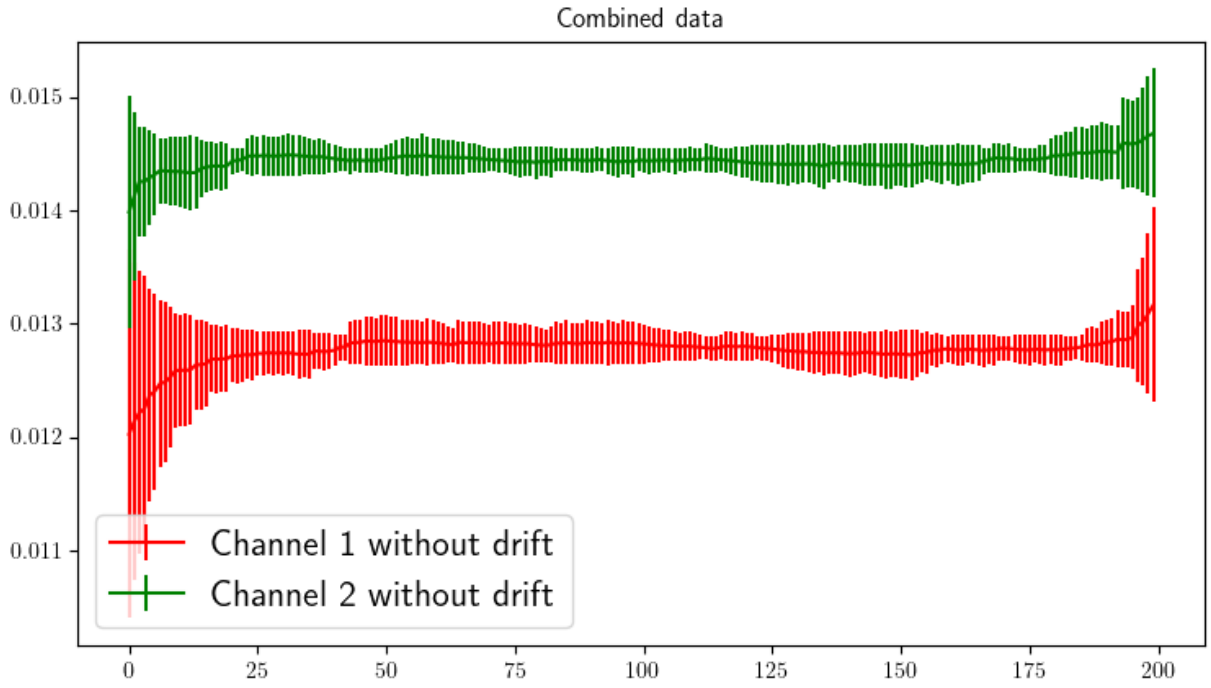

Рис. 4. Данные после вычитания “наклонной” составляющей

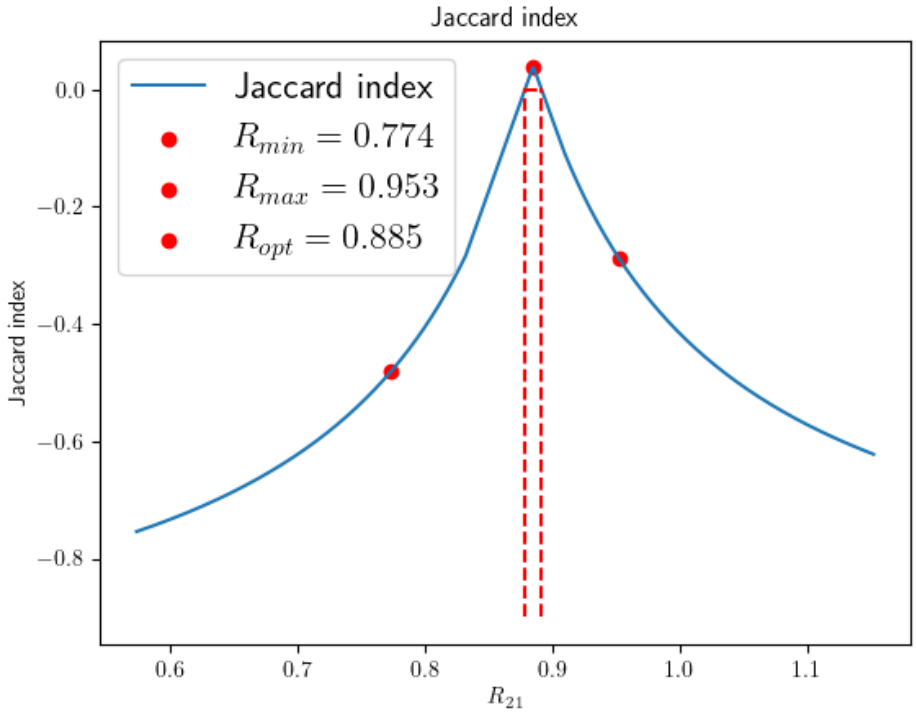

Рис. 5. Зависимость коэффициента Жаккара от R_{21}

Оптимальное соотношение $R_{opt} = 0.885$, было найдено в диапазоне $[0.774; 0.953]$.

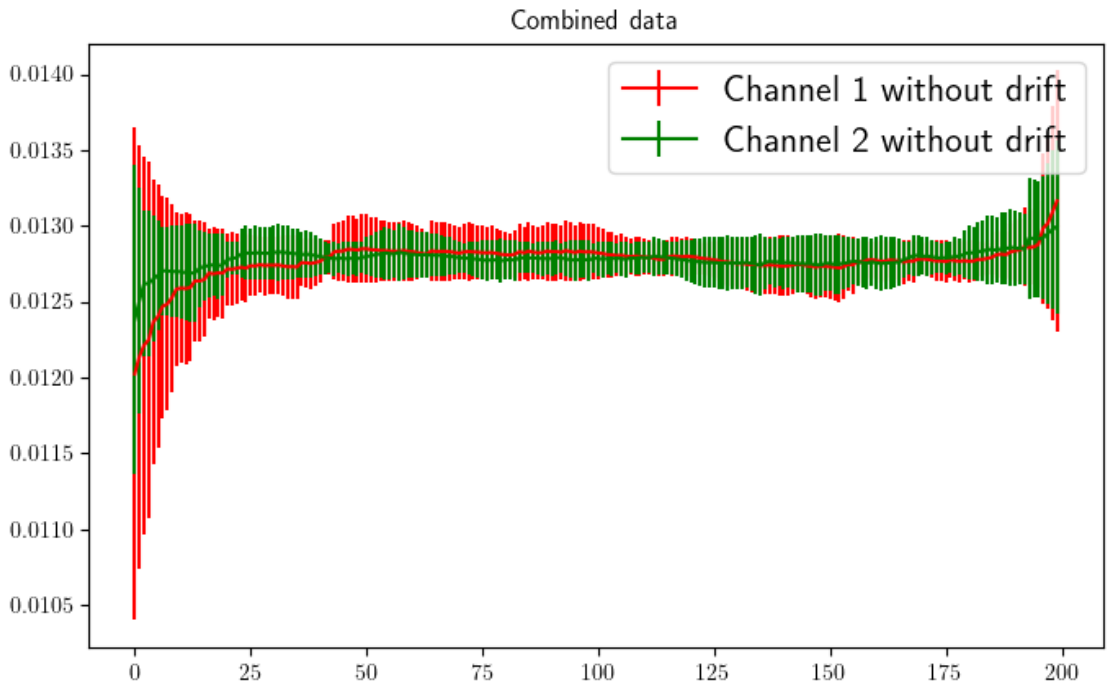

Рис. 6. Результат наложения данных при максимальном коэффициенте Жаккарда

4.1. Замечания относительно пакета glpk

Существенный объём времени был потрачен на выяснение причин, по которым пакет glpk не решал задачи, поставленные в модуле ir_outer.m (минимизация и максимизация β_0 и β_1 покомпонентно). Выяснилось, что солвер revised_simplex не в состоянии найти решения этих задач. Кроме того, солвер interior_point не мог найти граничные β_0 и β_1, когда на переменные не устанавливалось ограничений снизу. После того, как было установлено дефолтное ограничение снизу (т.е 0), interior_point со всеми задачами успешно справился. В то же время, солвер revised_simplex в пакете scipy успешно решал те же задачи без искусственных ограничений. Эти проблемы значительно усложнили реализацию данной лабораторной работы, и их, на мой взгляд, следует обнародовать среди студентов.

4.2. Линейная модель

Ниже приведены графики, полученные в результате работы реализованной программы.

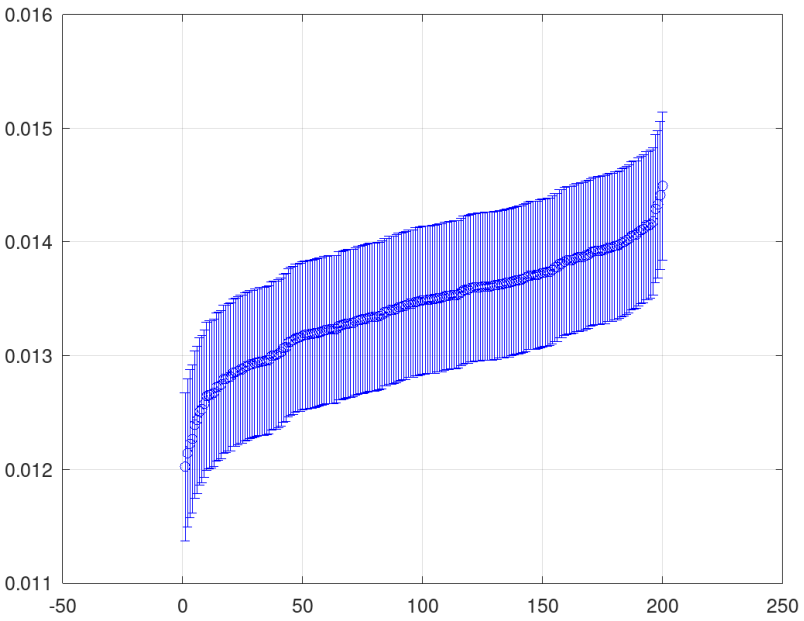

Рис. 7. Обынтерваленные данные. Модель 1

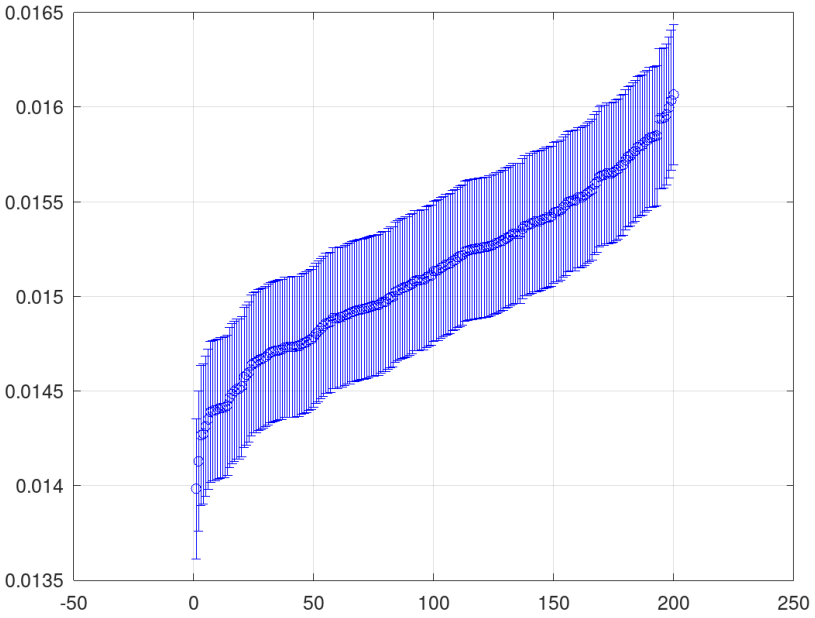

Рис. 8. Обынтерваленные данные. Модель 2

Для ускорения вычислительных процессов и более простого взаимодействия с данными, все интервалы были расширены в максимум из всех полученных весов раз: ширина интервалов составляети $\varepsilon \cdot \max_{i=\overline{1,n}}(w_i)$.

В следующей таблице приведены некоторые отличные от единицы веса:

Номер интервала	Вес (модель 1)	Вес (модель 2)
1	6.49	3.70
2	5.38	2.33
3	4.63	1.04
199	2.01	1.15
200	2.76	1.37

Таблица 2. Веса интервалов

В обоих случай максимальный вес пришёлся на первый интервал.

В следующей таблице указаны полученные параметры линейной интервальной регрессии (maxdiag).

Модель	β_0	β_1	$\max w$
1	0.012280	$1.0403 \cdot 10^{-5}$	6.49
2	0.014142	$9.3126 \cdot 10^{-6}$	3.70

Таблица 3. Параметры линейной интервальной регрессии

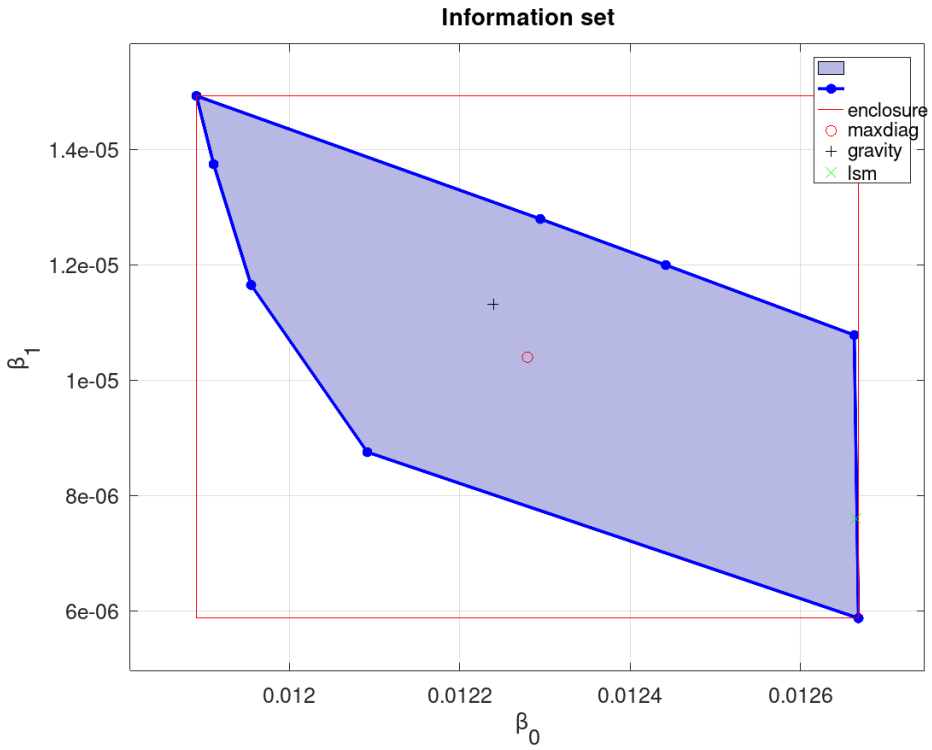

Рис. 9. Информационное множество. Модель 1

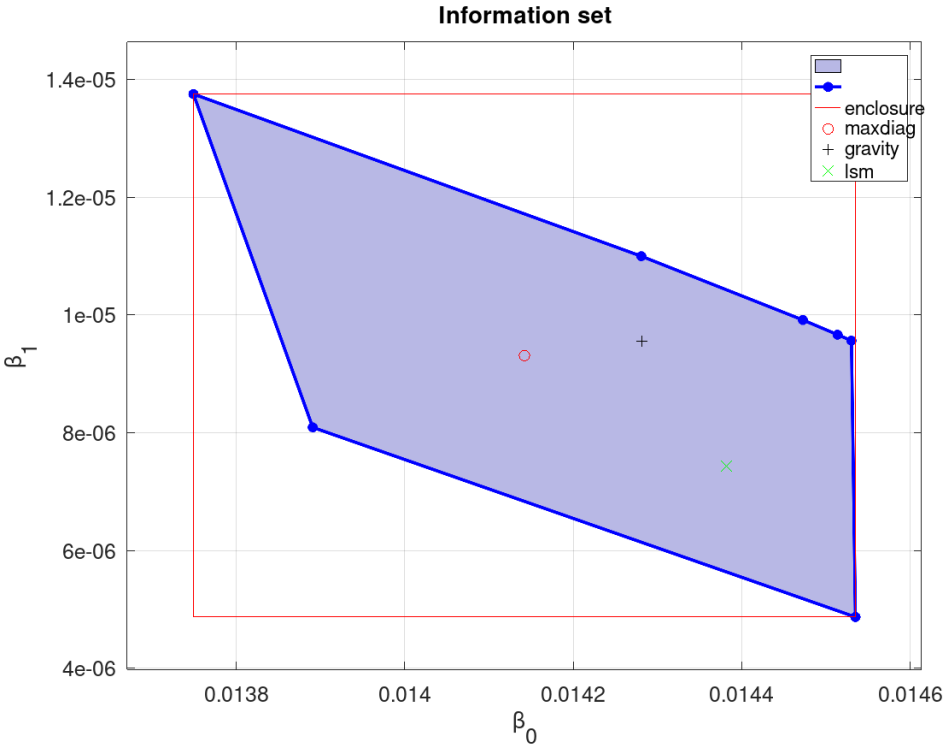

Рис. 10. Информационное множество. Модель 2

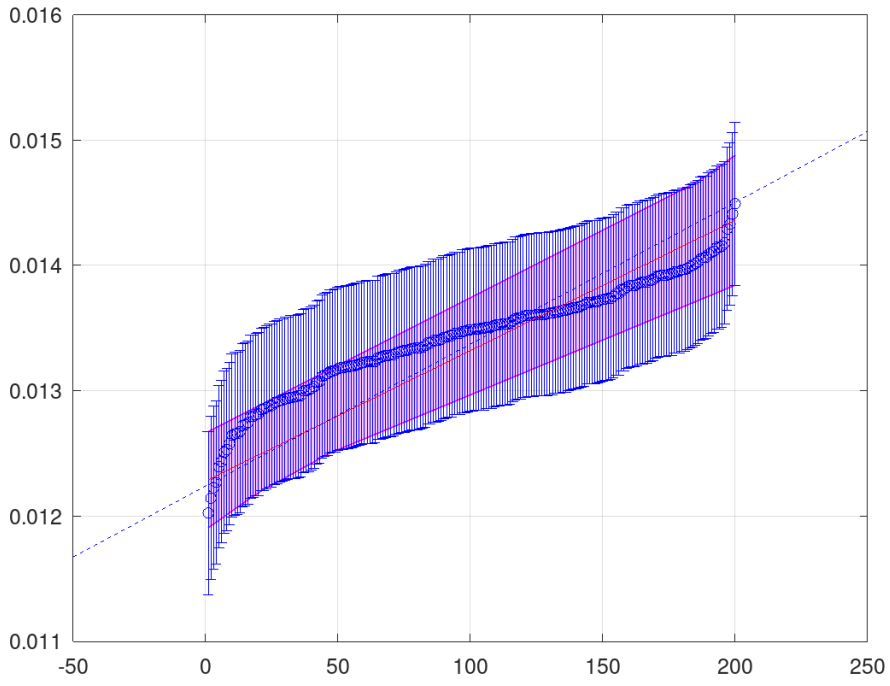

Рис. 11. Коридор совместных зависимостей. Модель 1

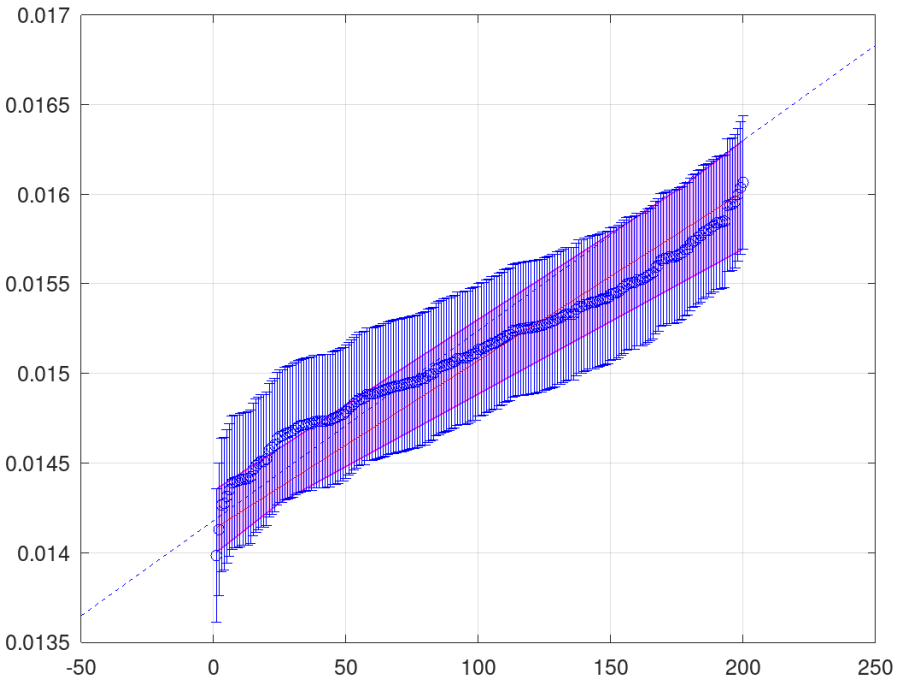

Рис. 12. Коридор совместных зависимостей. Модель 2

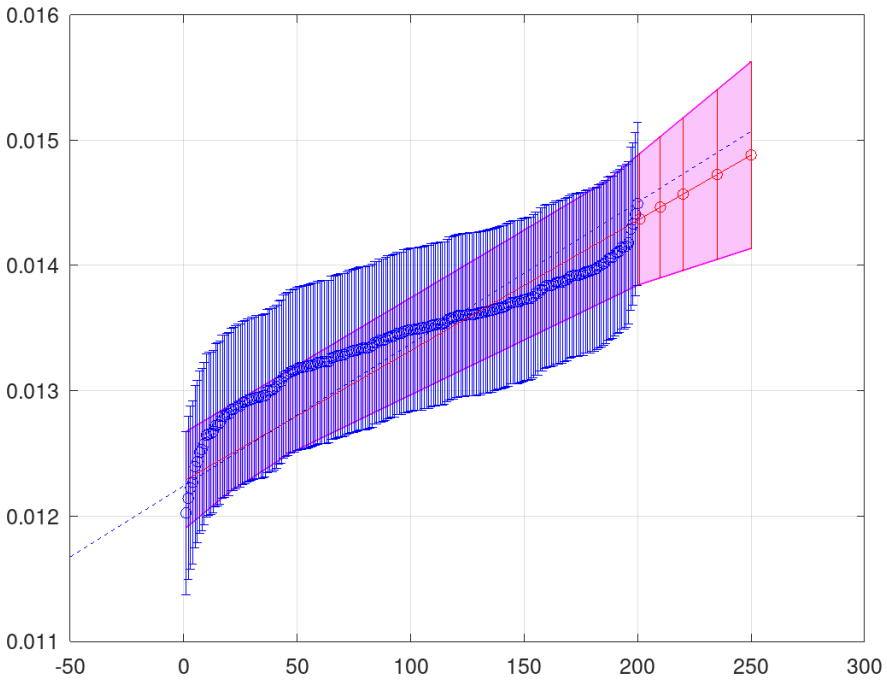

Рис. 13. Коридор совместных зависимостей. Предсказанные значения. Модель 1

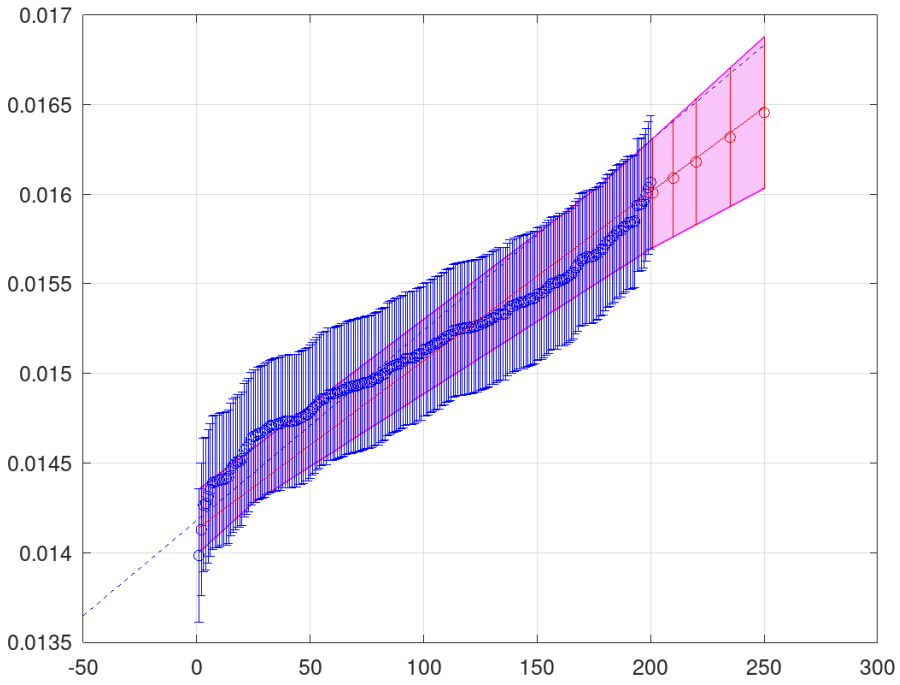

Рис. 14. Коридор совместных зависимостей. Предсказанные значения. Модель 2

Граничные точки в первой модели – точки под номерами 1, 17, 21, 47, 182, 184, 189, 200.

Граничные точки во второй модели – 1, 25, 162, 165, 177, 193, 200.

Максимальный коэффициент Жаккара, рассчитанный прежним методом при параметрах β_0, β_1, полученных как точка пересечения максимальных диагоналей (maxdiag) оказался равен 0.0615, в то время как в прошлой реализации он равен 0.037, что в 1.65 раз выше. Оптимальный коэффициент R_{21} в таком случае равен 0.882, что отличается от прошлого варианта на 0.003.

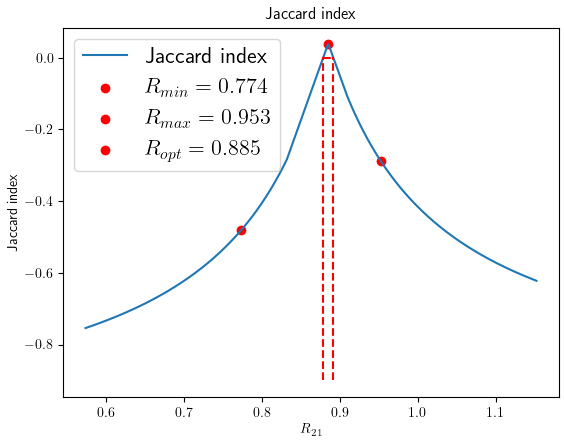

Рис. 15. Зависимость коэффициента Жаккара от множителя R_{21}

4.3. Кусочно-линейная модель

Далее была произведена процедура кусочно-линейной интервальной регрессии: выше описанная процедура была проделана для трёх отдельных участков данных: 1-50, 51-150, 151-200. В результате были получены следующие параметры регрессии:

Диапазон	β_0	β_1	$\max w$
1-50	0.01217	$2.0065 \cdot 10^{-5}$	6.54
51-150	0.01269	$7.4948 \cdot 10^{-6}$	1.00
151-200	0.01149	$1.471 \cdot 10^{-5}$	1.67

Таблица 4. Параметры кусочно-линейной интервальной регрессии. Модель 1

Диапазон	β_0	β_1	$\max w$
1-50	0.01420	$1.368 \cdot 10^{-5}$	2.31
51-150	0.01431	$8.109 \cdot 10^{-6}$	1.00
151-200	0.01318	$1.431 \cdot 10^{-5}$	1.00

Таблица 5. Параметры кусочно-линейной интервальной регрессии. Модель 2

Обынтерваленные данные выглядят следующим образом:

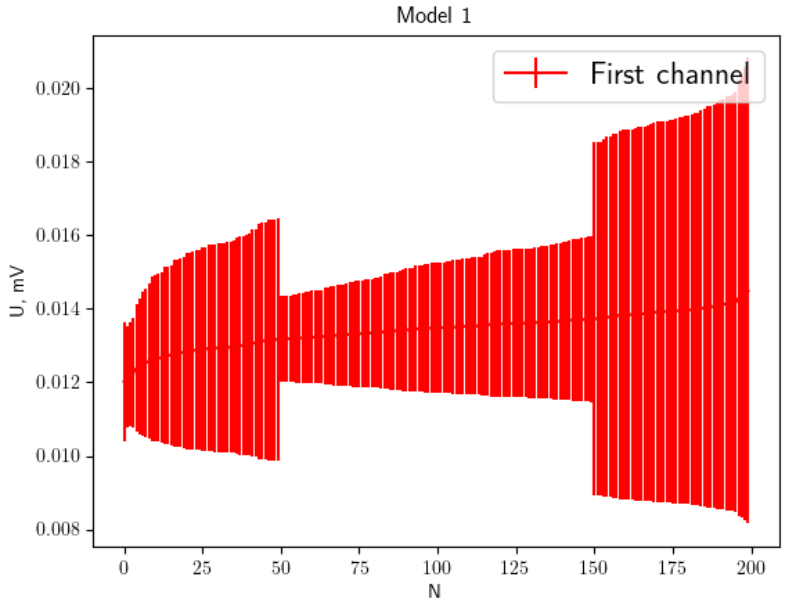

Рис. 16. Кусочно-линейная регрессия. Модель 1

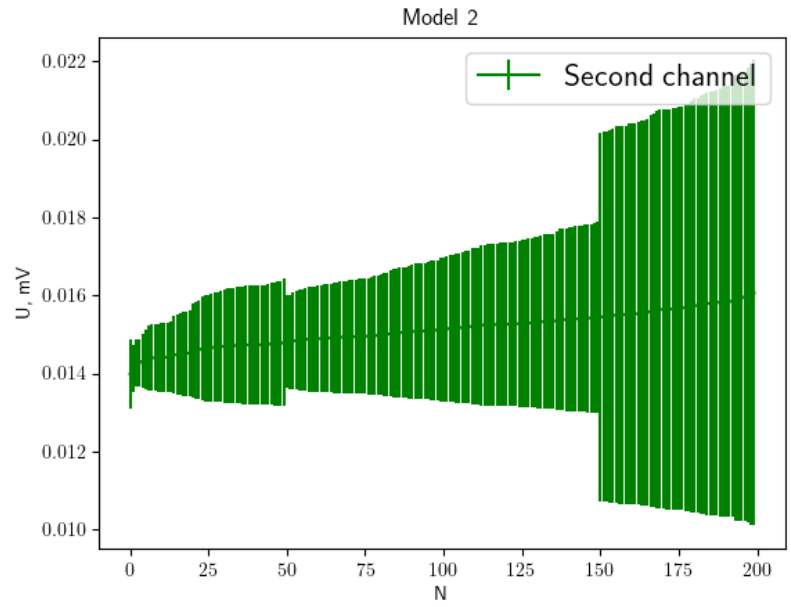

Рис. 17. Кусочно-линейная регрессия. Модель 2

При построении кусочно-линейной регрессии удалось добиться коэффициента Жаккара, равного 0.0667, что на 0.0052 больше, чем в случае линейной интервальной регрессии. Примечательно, что оптимальный множитель R_{21} в таком случае оказался равен 0.888: линейная регрессия дала отклонение коэффициента влево от точечной на 0.003, а кусочно-линейная – вправо на то же значение.

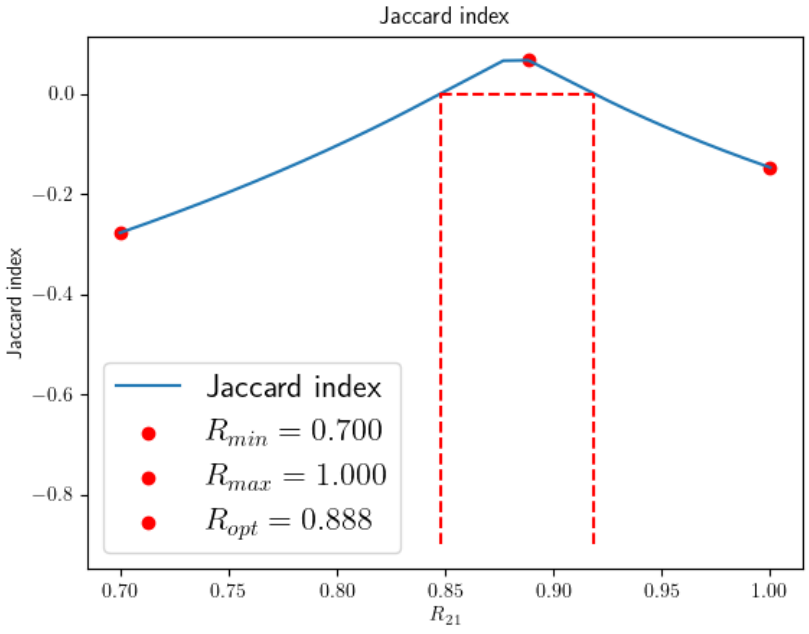

Рис. 18. Зависимость коэффициента Жаккара от множителя R_{21}

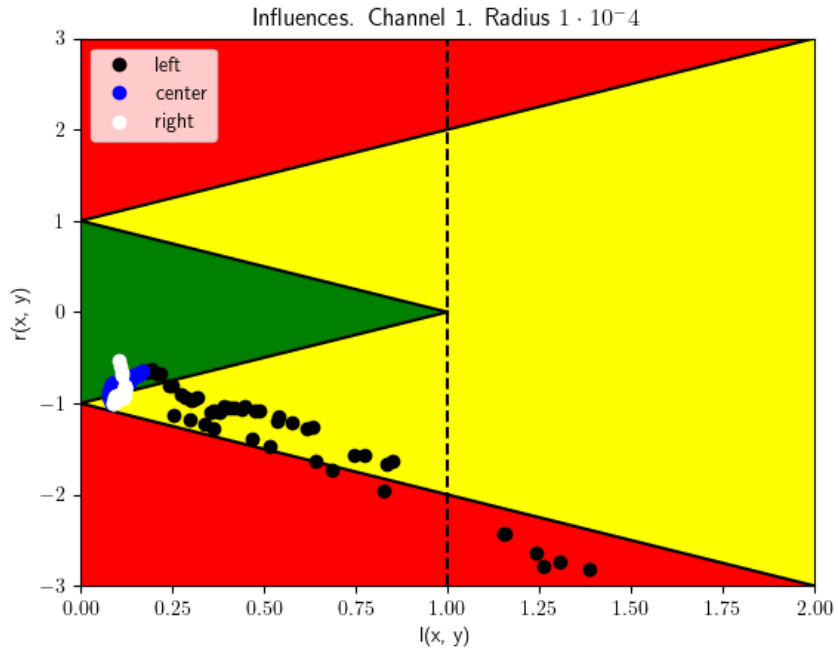

Рис. 19. Диаграмма статусов. Канал 1. Радиус интервала $1 \cdot 10^{-4}$

Для этого случая покажем диаграмму рассеяния, отметим на ней коридор совместных зависимостей и выбросы:

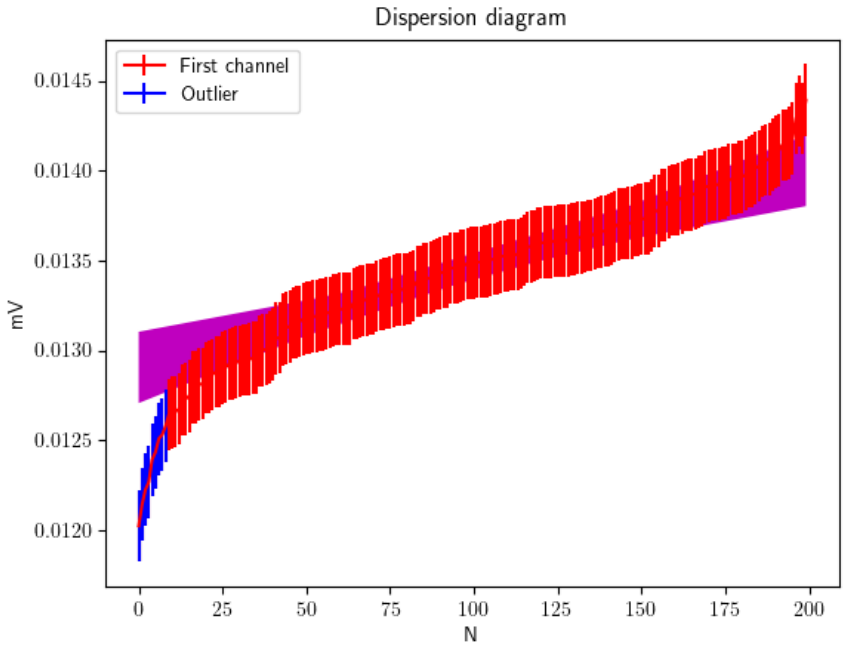

Рис. 20. Диаграмма рассеяния. Канал 1. Радиус интервала $1 \cdot 10^{-4}$

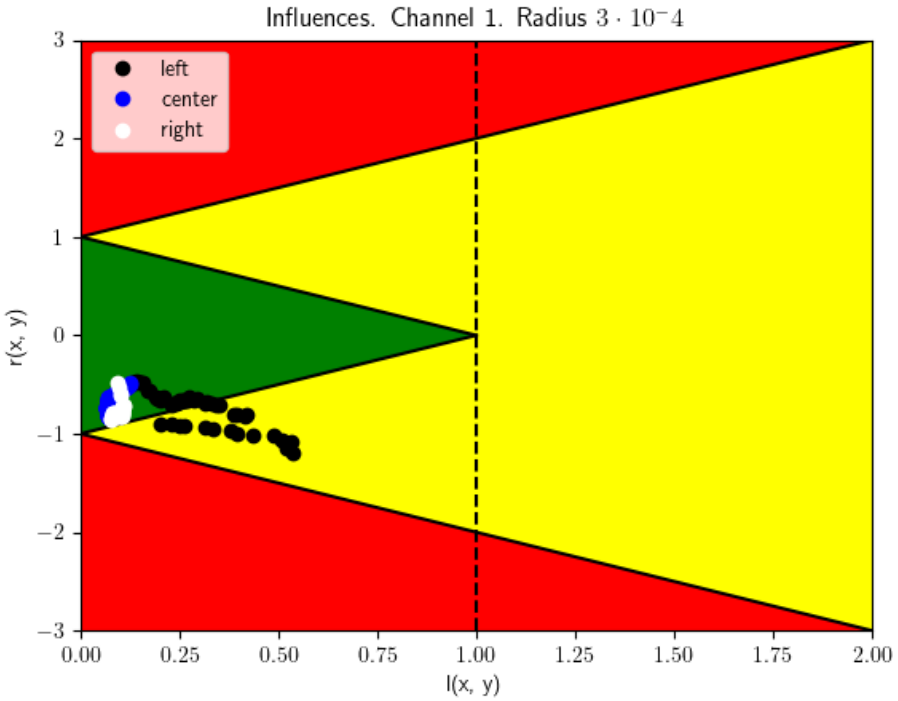

Рис. 21. Диаграмма статусов. Канал 1. Радиус интервала $3 \cdot 10^{-4}$

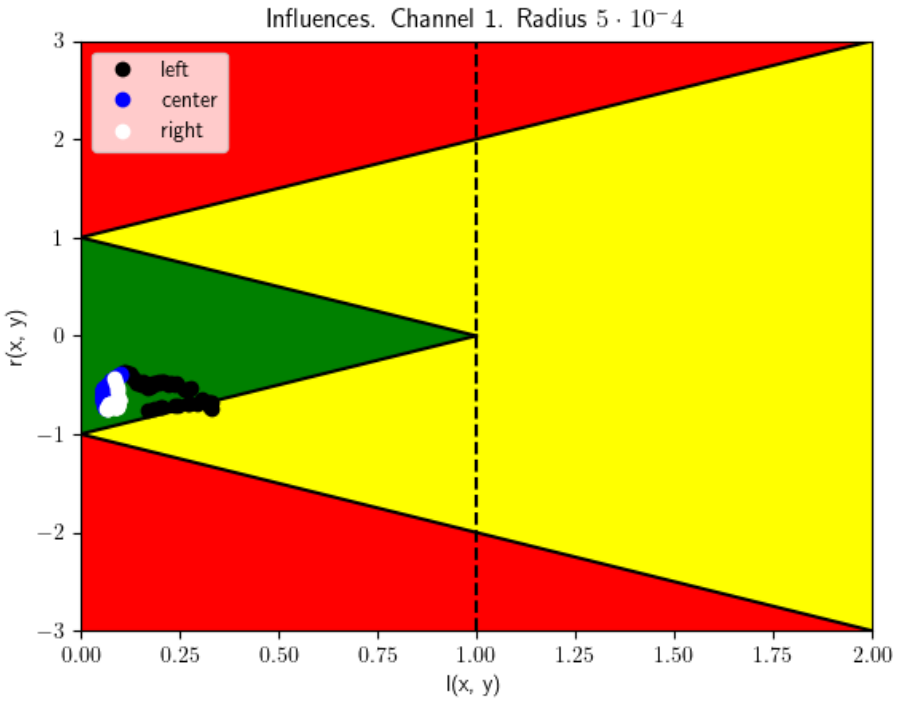

Рис. 22. Диаграмма статусов. Канал 1. Радиус интервала $5 \cdot 10^{-4}$

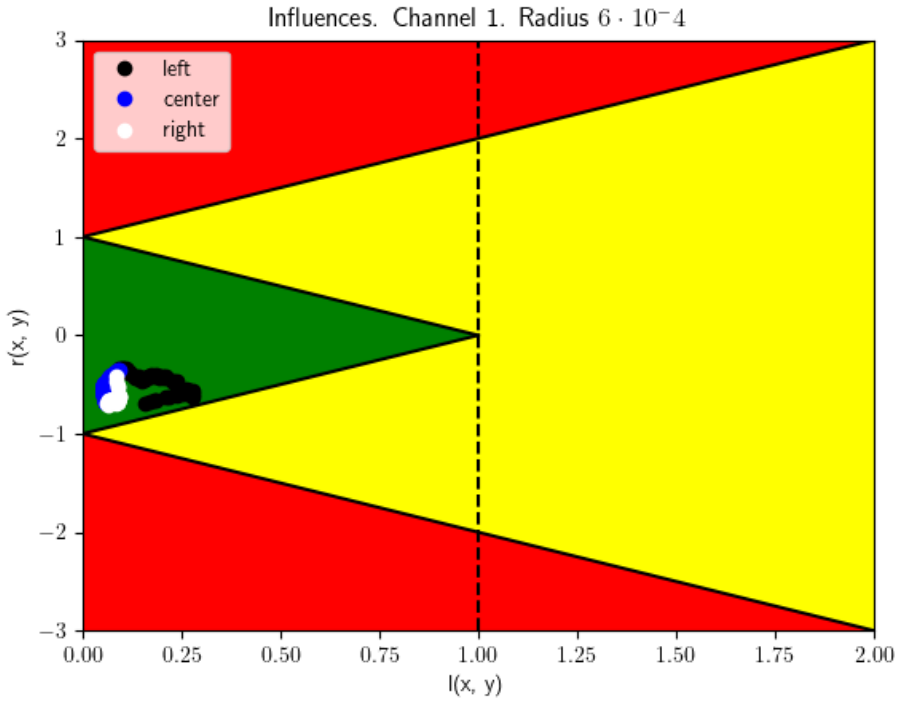

Рис. 23. Диаграмма статусов. Канал 1. Радиус интервала $6 \cdot 10^{-4}$

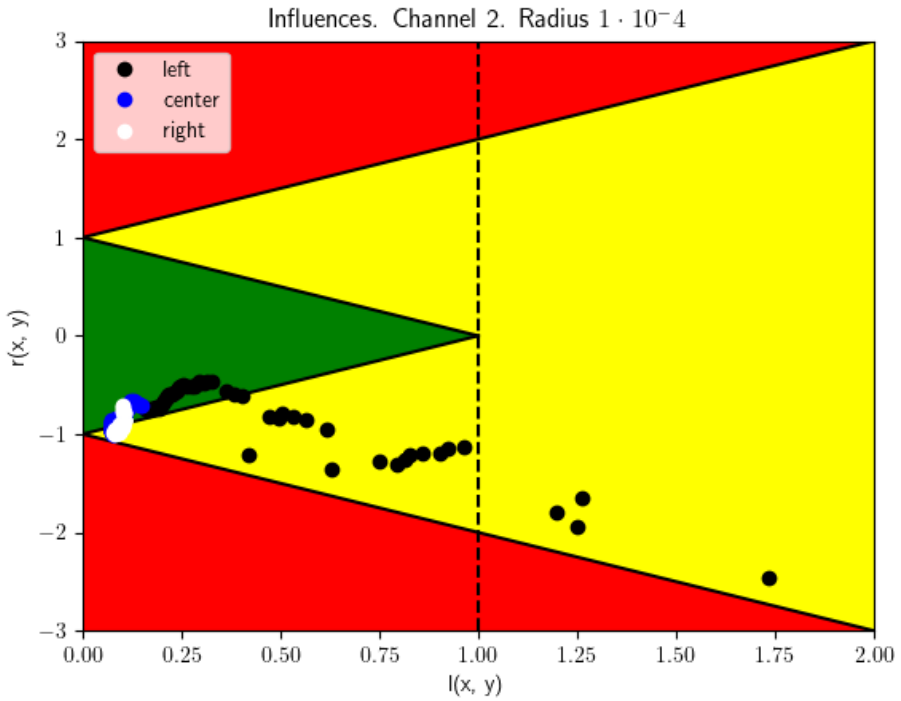

Рис. 24. Диаграмма статусов. Канал 2. Радиус интервала $1 \cdot 10^{-4}$

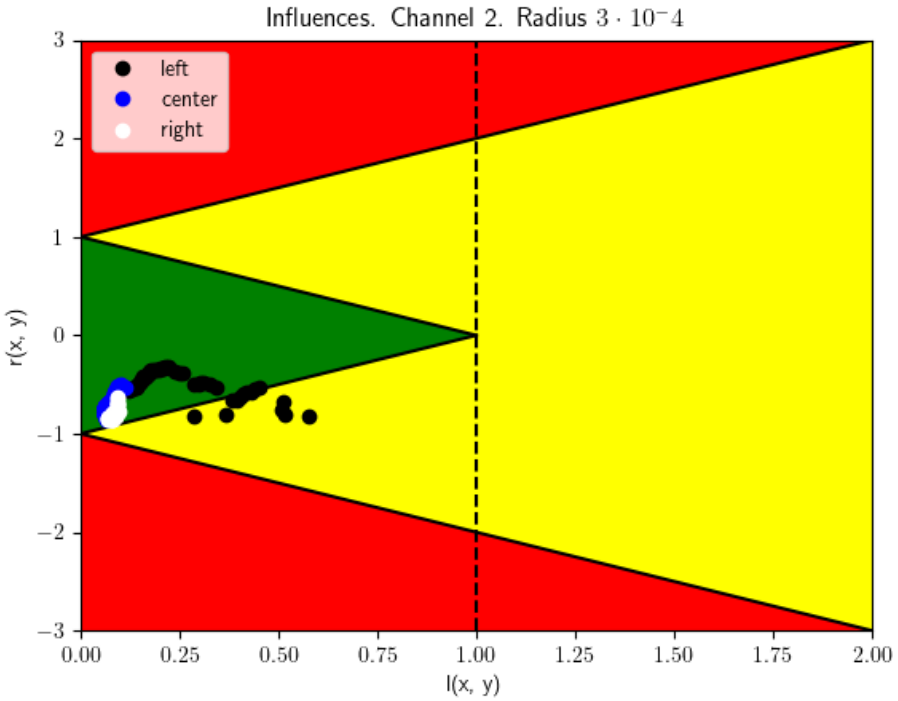

Рис. 25. Диаграмма статусов. Канал 2. Радиус интервала $3 \cdot 10^{-4}$

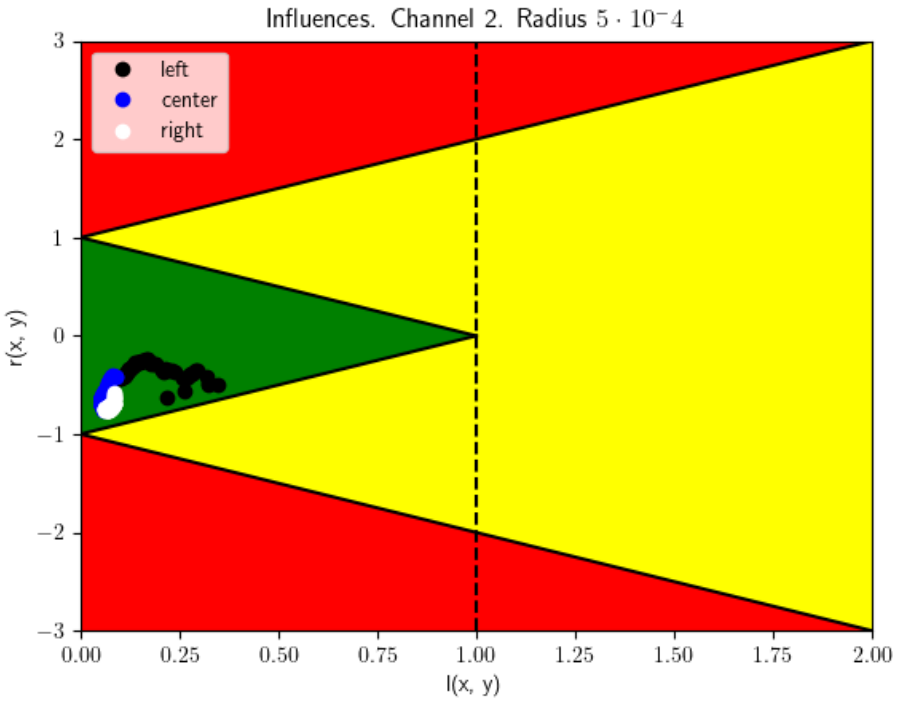

Рис. 26. Диаграмма статусов. Канал 2. Радиус интервала $5 \cdot 10^{-4}$

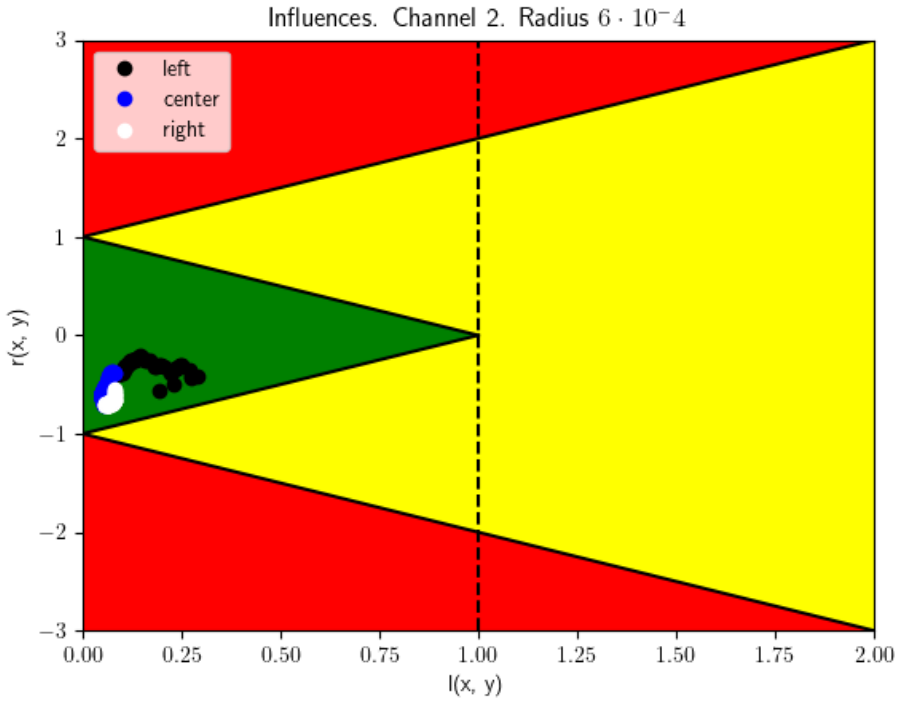

Рис. 27. Диаграмма статусов. Канал 2. Радиус интервала $6 \cdot 10^{-4}$

5. Обсуждение

Исходя из представленных графиков, можно судить о том, что все описанные в теории этапы выполнены правильно.

- Простая линейная регрессия и обынтерваливание проведены так, что каждый интервал содержит соответствующую точку аппроксимирующей прямой, при этом аппроксимирующая прямая лежит визуально близко к исходным данным.
- В результате отсечения наклонной части действительно получились визуально горизонтальные графики.
- График зависимости коэффициента Жаккара от искомого множителя ожидаемо имеет один локальный максимум. При этом видно, что оценка интервала R_{21} с точки зрения меры Жаккара действительно очень грубая: значение коэффициента Жаккара в нижней оценке приблизительно равно -0.5. Данное число привело бы к абсолютно неприемлемому результату интервальной регрессии: хоть точечные наборы и получились бы визуально похожими, этого нельзя было бы сказать про интервалы. Учитывая характер полученных данных, важно удостовериться именно в максимальном совпадении интервалов.
- На последнем рисунке видно, что значительная часть интервалов совпадает практически идеально, что также является показателем качественно

выполненной работы.

Представленные способы позволяют успешно решать задачу интервальной регрессии.

Модель линейной интервальной регрессии является более точной, нежели точечная, так как позволяет более корректно обынтервалить данные и найти такой коэффициент пропорциональности, при котором коэффициент Жаккара будет в 1.65 раз больше, чем в прошлом способе.

В то же время модель кусочно-интервальной регрессии позволяет получить ещё более качественную аппроксимацию: был получен коэффициент Жаккара, на 0.003 лучший, чем в прошлой модели.

При решении поставленной задачи были обнаружены баги в пакете glpk, с которыми удалось побороться путём изменения солвера. Видно, что точки из центральной части, как и ожидается, лежат в зелёной зоне. Чем больше увеличивается ширина интервалов, тем больше точек из левой и правой части попадают в зелёную зону. При радиусе в $6 \cdot 10^{-4}$ в обоих выборках точки из всех трёх частей попадают в зелёную зону. Также видно, что строго внешние измерения встречаются только во второй выборке при радиусе интервала 10^{-4}, а выбросы были обнаружены при том же радиусе в первой выборке.

Как можно заметить, выбросы действительно оказались вне прогнозного коридора, что верно в соответствие с приведённой теорией.

6. Приложения

1. Репозиторий с кодом программы и кодом отчёта:

   https://github.com/kystyn/interval2