

Санкт-Петербургский политехнический университет Петра Великого  
Физико-механический институт  
**Высшая школа прикладной математики и вычислительной физики**

Лабораторная работа  
по дисциплине «Анализ данных с интервальной неопределенностью»  
на тему **«Обработка постоянной. Применение меры совместности  
к анализу данных»**

Выполнил

студент гр. 5040102/10201

Гусаров Д.А.

/\_\_\_\_\_/

Руководитель

доцент, к.ф.-м.н.

Баженов А.Н.

/\_\_\_\_\_/

Санкт-Петербург

2022

## Постановка задачи

Проводится исследование из области солнечной энергетики.

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

$$QE_1 = \frac{X_1}{X_2} \cdot QE_2 \quad (1)$$

$QE_1$ ,  $QE_2$  – эталонная эффективность эталонного и исследуемого датчика,  $X_1$ ,  $X_2$ , или  $\{x_{1i}\}_{i=1}^{200}$ ,  $\{x_{2i}\}_{i=1}^{200}$  – измеренные мощности. Данные датчиков находятся в файлах “Канал 1\_700nm\_0.03.csv” и “Канал 2\_700nm\_0.03.csv”.

Требуется определить параметры постоянной величины на основе двух выборок  $\{x_{1i}\}_{i=1}^{200}$ ,  $\{x_{2i}\}_{i=1}^{200}$ , в частности коэффициент калибровки

$$R_{12} = \frac{X_1}{X_2} \quad (2)$$

при помощи линейной регрессии, интервальных данных и коэффициента Жаккара.

## Теория

Один из распространенных способов получения интервальных результатов в первичных измерениях – это «обинтерваливание» точечных значений, когда к точечному базовому значению  $x_{1i}$ , которое считывается по показаниям измерительного прибора прибавляется интервал погрешности  $\varepsilon$

$$X_{1i} = x_{1i} + [-\varepsilon, +\varepsilon] \quad (3)$$

В конкретных измерениях  $\varepsilon = 10^{-4}$  мВ. Согласно терминологии интервального анализа, рассматриваемая выборка – это вектор интервалов, или интервальный вектор  $X_1 = \{X_{1i}\}_{i=1}^{200}$

Построение интервалов будет происходить следующим образом:

Вначале построим линейную регрессию по известному методу наименьших квадратов в виде  $L_1(n) = A_1 \cdot n + B_1$ , где  $n$  – номер измерения;  $L_1(n)$  – прямая, аппроксимирующая экспериментальные измерения  $\{x_{1i}\}_{i=1}^{200}$ . Отклонение можно вычислить как

$$\varepsilon_{1n} = |x_{1n} - L_1(n)| \quad (4)$$

Если отдельные интервалы не заключают в себе линейную регрессию, к отклонение  $\varepsilon_{1n}$  стоит растянуть, домножить на величину  $w_n$ , минимально возможную, для того, чтобы интервал коснулся линии регрессии.

Интервальные данные представляются в виде:

$$X_{1n} = x_{1n} + [-\tilde{\varepsilon}_n, +\tilde{\varepsilon}_n] \quad (5)$$

или кратко  $X_1$  – множество всех интервальных данных, построенных по измерениям датчика ФП1,  $\tilde{\varepsilon}_n = w_n \cdot \varepsilon$ ,  $w_n \geq 1$ .

Чтобы сделать интервальную величину более константной и в дальнейшем оценить совместность двух выборок экспериментальных измерений, следует вычесть из интервальных данных линейную зависимость (фактически из концов интервала), получим:

$$X'_1 \leftarrow X_1 - A_1 \cdot n \quad (6)$$

Для базовых значений  $x_{2i}$  выполняются аналогичные вычисления. Находится линейная зависимость  $L_2(n) = A_2 \cdot n + B_2$ , интервалы  $X_{2i}$  по формуле (5) и обработанные интервалы  $X'_2$  по формуле (6) с соответствующими индексами.

В различных областях анализа данных используют различные меры сходства множеств, иными словами, коэффициенты сходства. В данной работе используется мультимера Жаккара, то есть ее модификация для интервальных данных:

$$JK = \frac{wid(\cap y_i)}{wid(\cup y_i)} \quad (7)$$

Мера Жаккара  $-1 \leq JK \leq 1$  численно характеризует меру совместности интервальных данных. В качестве  $y_i$  рассматриваются интервальные данные объединенной выборки  $X' = \{X'_1, RX'_2\}$ .  $JK$  – число, получаемое в результате деления пересечения интервалов на их объединение. Заметим, что если при подборе калибровочного множителя  $R$  получается  $JK > 0$ , то выборка совместна (имеет положительную меру совместности). Поиск оптимального  $R_{opt}$  можно представить так:

$$R_{opt} = arg \left\{ \max_R JK(X') \right\} \quad (8)$$

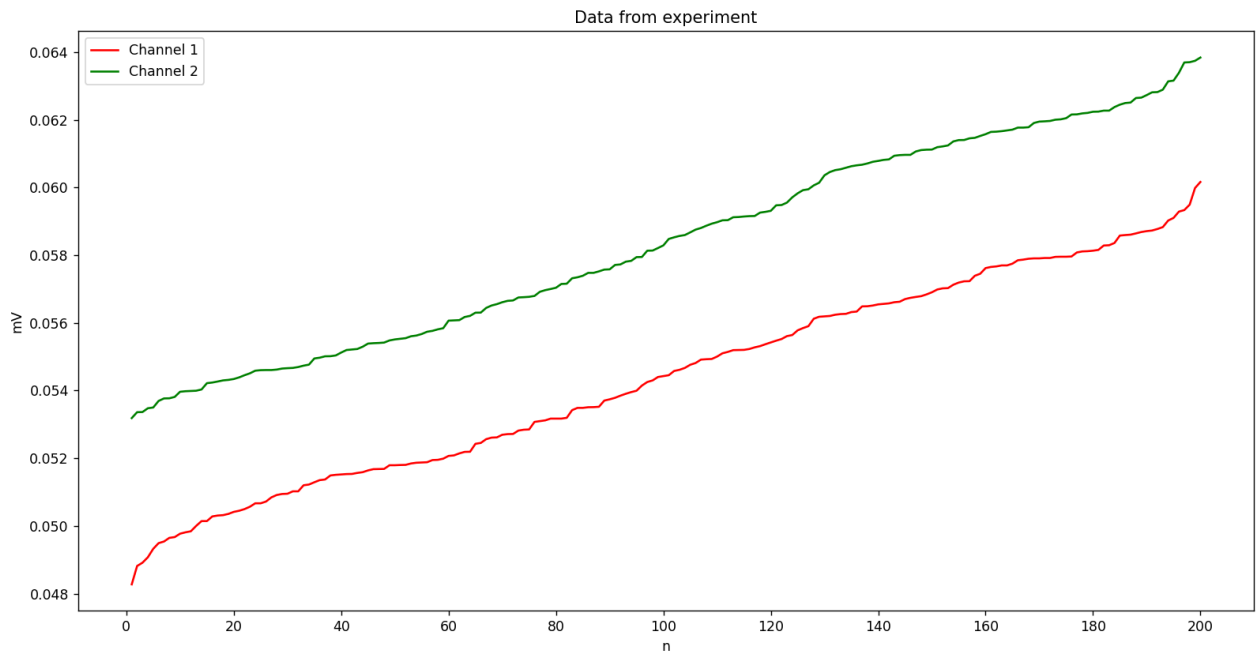
$R_{opt}$  – это аргумент, у которого реализуется данный функционал, максимальная оценка коэффициента калибровки  $R_{12}$  из формулы (2). Внешнюю оценку для  $R_{opt}$  можно найти разными способами, проще всего путем деления интервалов двух выборок  $R = \frac{x_1}{x_2}$ , в результате чего получим интервал внешней оценки  $[R, \overline{R}]$  – такой интервал, в котором можно найти  $R_{opt}$ , перебирая  $R$  с некоторым шагом и вычисляя функционал (8). Интервал, в пределах которого наблюдается  $JK > 0$  является внутренней оценкой коэффициента  $R_{opt}$ .

## Результаты

Программный код написан на языке программирования Python с использованием библиотек Matplotlib, NumPy и Sklearn.

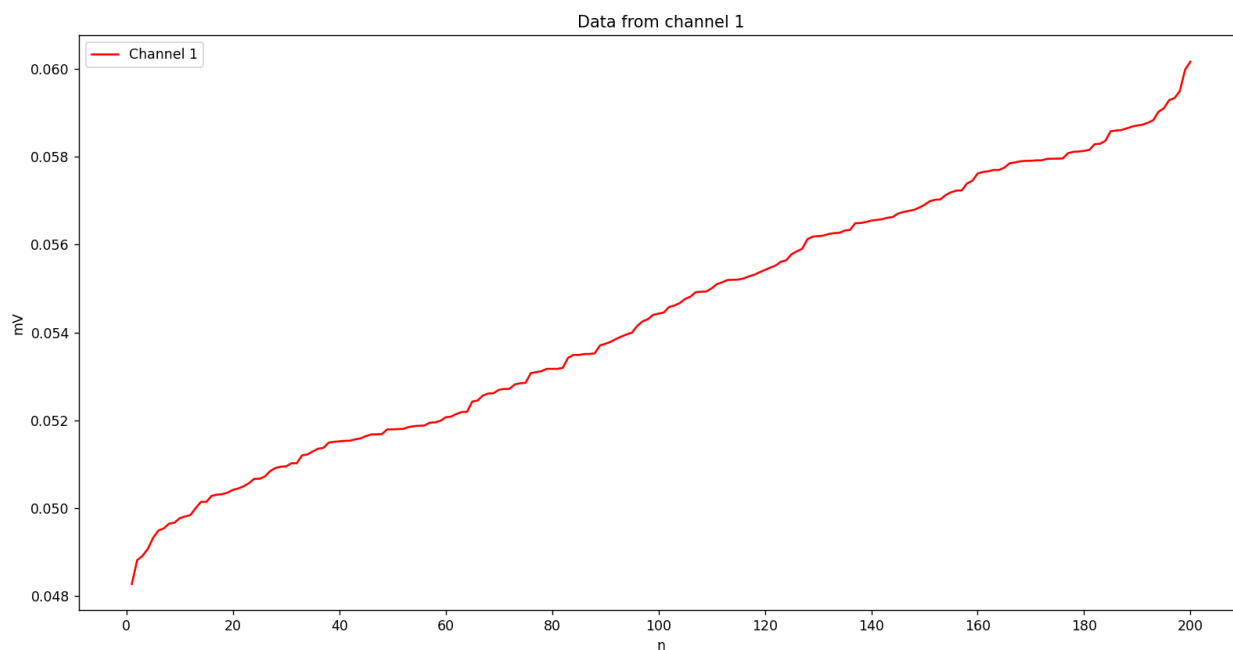
На рис.1 представлены экспериментальные данные, измеренные двумя датчиками, на рис.2 и рис.3 – те же данные, но в другом масштабе. На рис. 4 и 5 показаны построенные согласно описанной выше теории интервальные данные и линейная регрессия с коэффициентами

$A_1 \approx 5.0867 \cdot 10^{-5}$ ,  $B_1 \approx 0.04928$ ,  $A_2 \approx 5.3844 \cdot 10^{-5}$ ,  $B_2 \approx 0.0529$ .

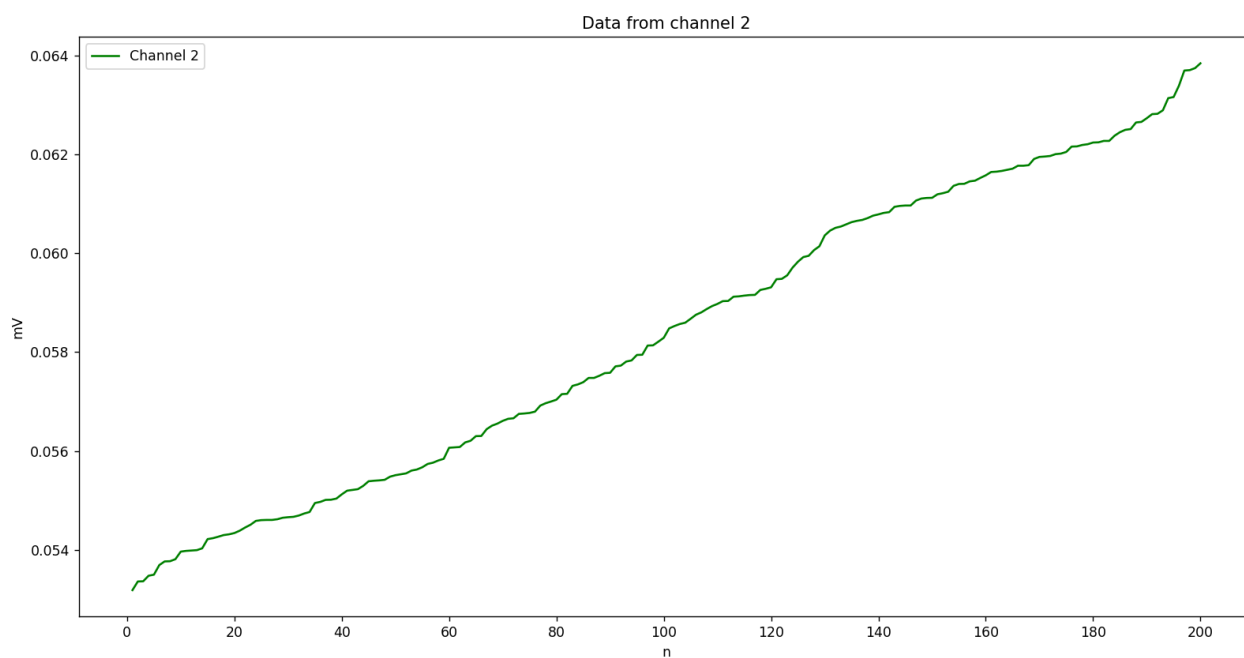


*Рис. 1. Две выборки экспериментальных данных, измеренным датчиками*

## 1. Измеренные данные:



*Рис. 2. Данные, измеренные датчиком ФП1*



*Рис. 3. Данные, измеренные датчиком ФП2*

## 2. Интервальные данные:

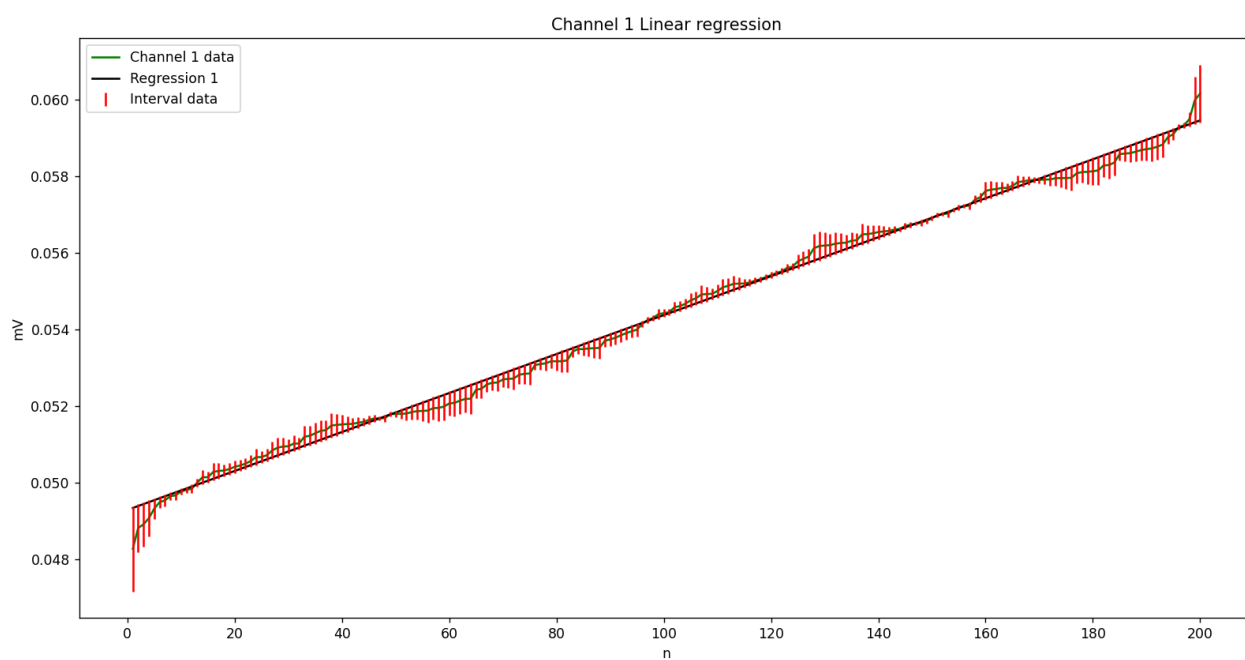


Рис. 4. Интервальные данные первой выборки и линейная регрессия

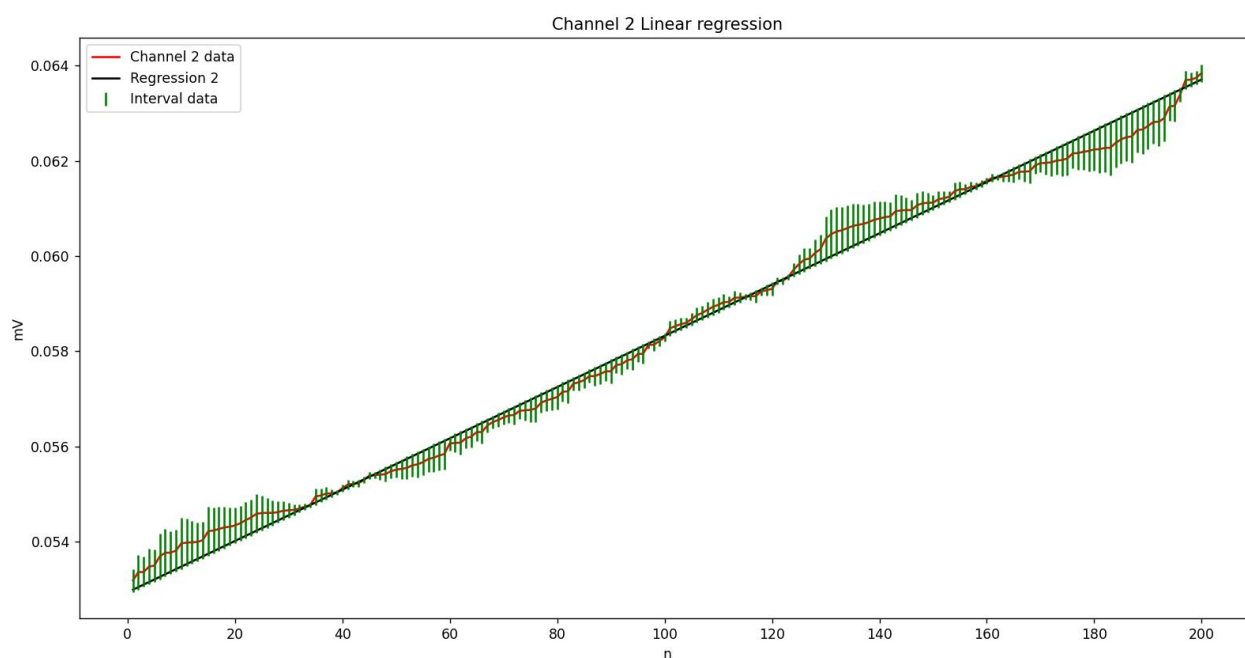
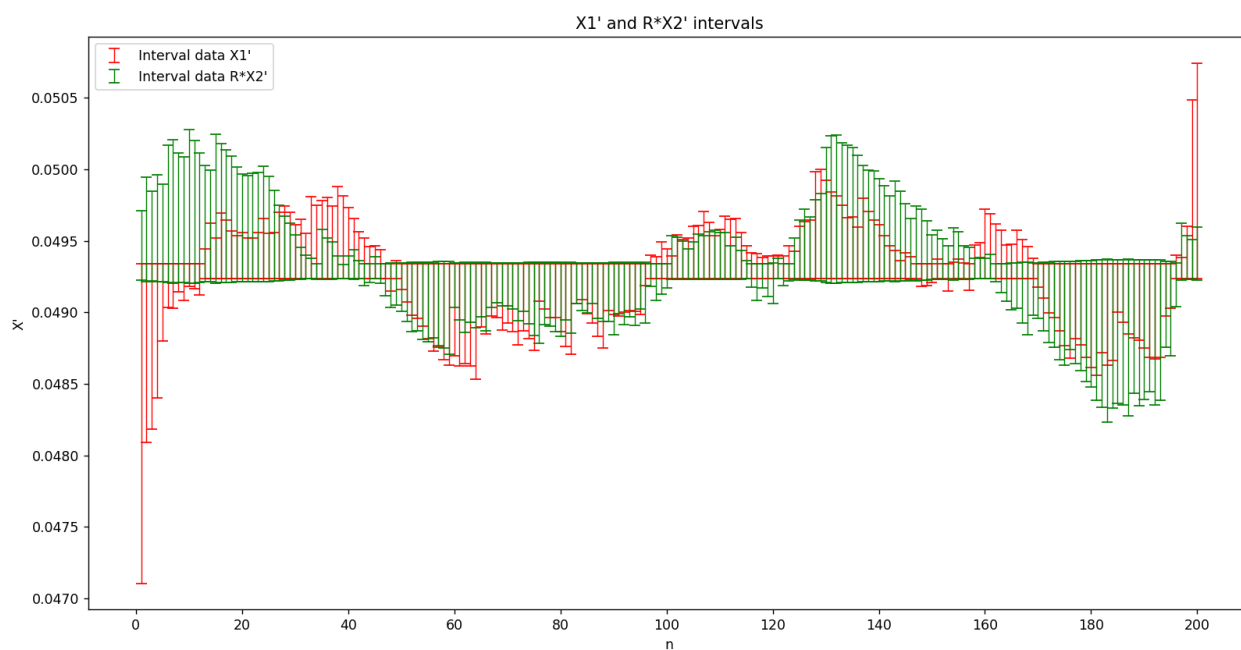


Рис. 5. Интервальные данные второй выборки и линейная регрессия

На рис. 6 визуализирован пример совместных выборок  $X'_1, RX'_2$ , что выполняется при  $R$ , обеспечивающим  $JK > 0$ .



*Рис. 6. Обработанные интервальные данные совместной выборки при  $R$ , обеспечивающем совместность выборок*



### 3. Мультимера Жаккара:

На рис. 7 показана зависимость коэффициента Жаккара от коэффициента калибровки  $R$ . Согласно внешней оценке оптимальное значение  $R_{opt}$  осуществлялся в диапазоне  $[R, \bar{R}] \approx [0.92457, 0.95941]$ . Как интервал можно представить  $R_{12} \approx [0.92927, 0.93275]$ . В нашем эксперименте, максимум коэффициента Жаккара имеет значение 0.026.

Это связано с наличием различных погрешностей, которые на практике невозможно устранить, но несмотря на их присутствие, поведение коэффициента Жаккара позволило найти оптимальный калибровочный коэффициент  $R_{opt} \approx 0.93101$ .

Таким образом, можно сказать, что область, где  $JK(R_{12}) \geq 0$  является оценкой искомой величины  $R_{12}$ .

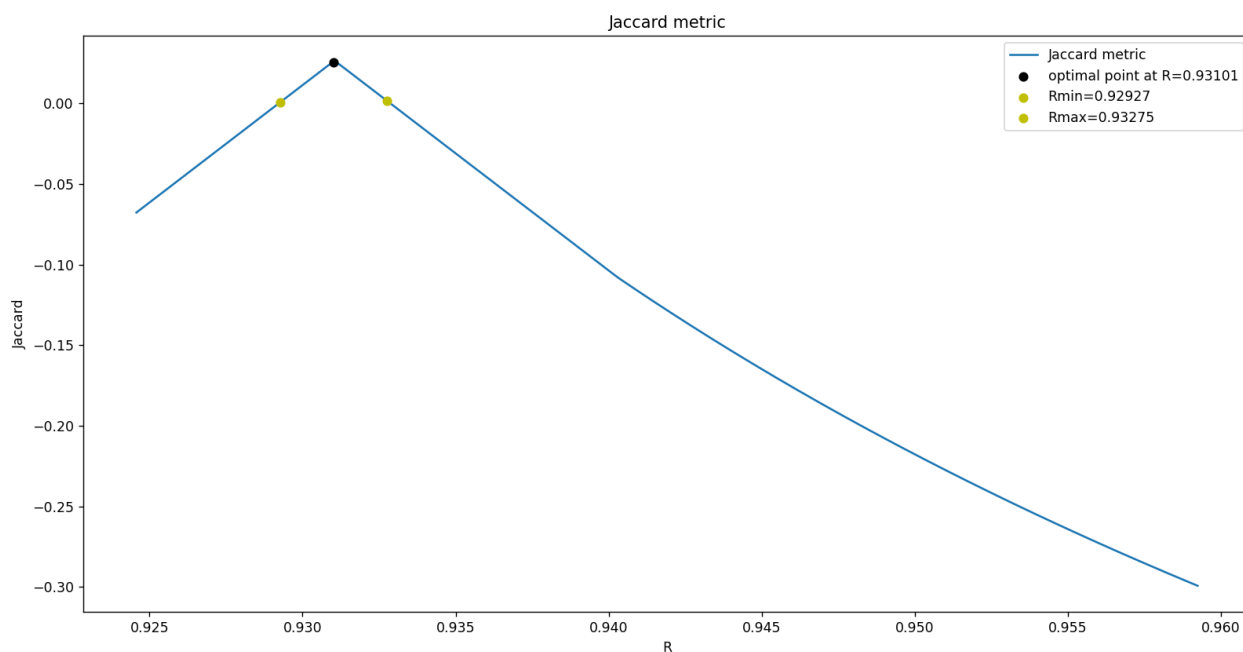
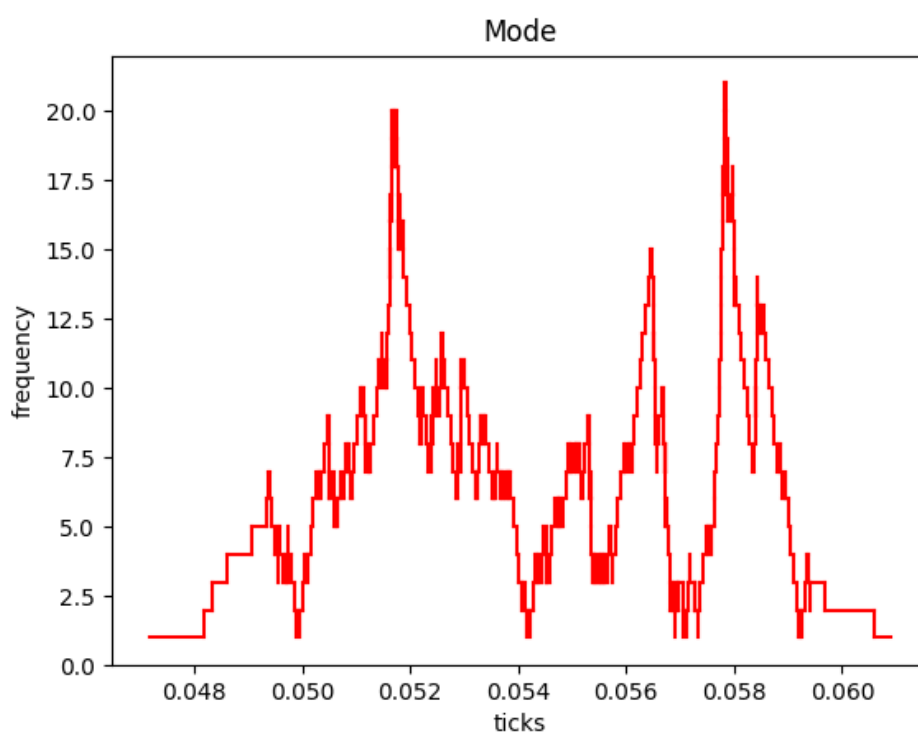


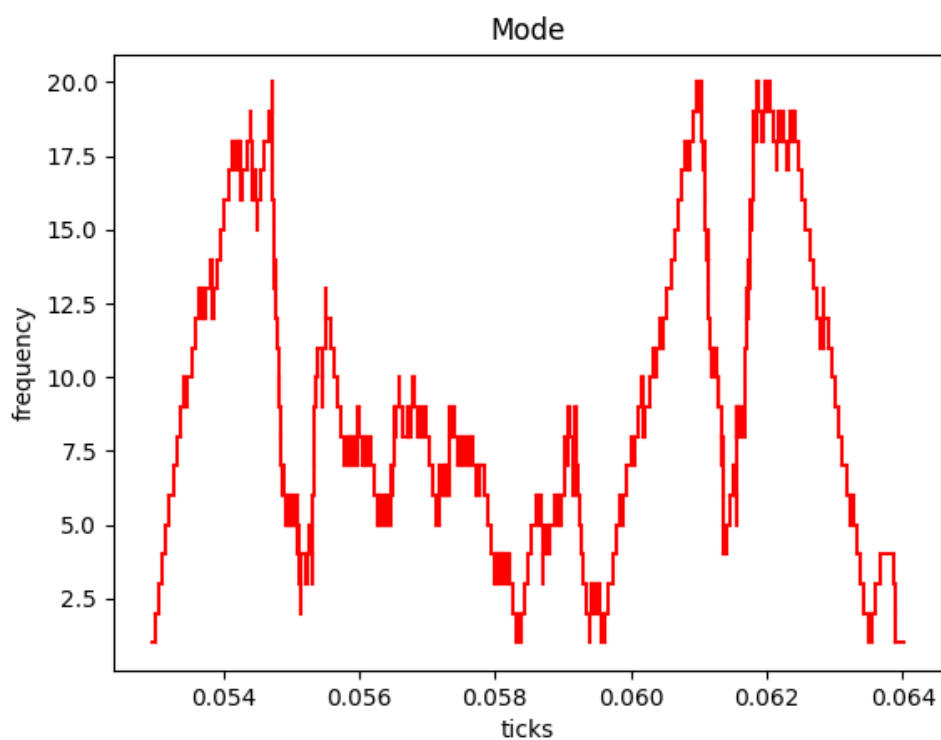
Рис. 7. Значения коэффициента Жаккара от коэффициента калибровки

#### 4. Нахождение мод выборок.

Для исходных интервальных данных (до вычитания тренда):

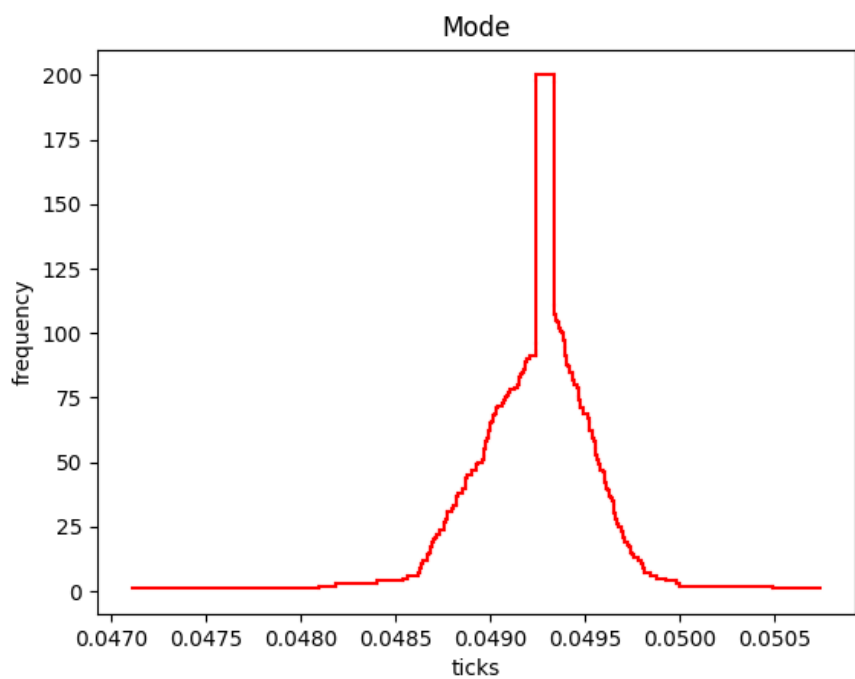


*Рис. 8. Мода выборки 1*

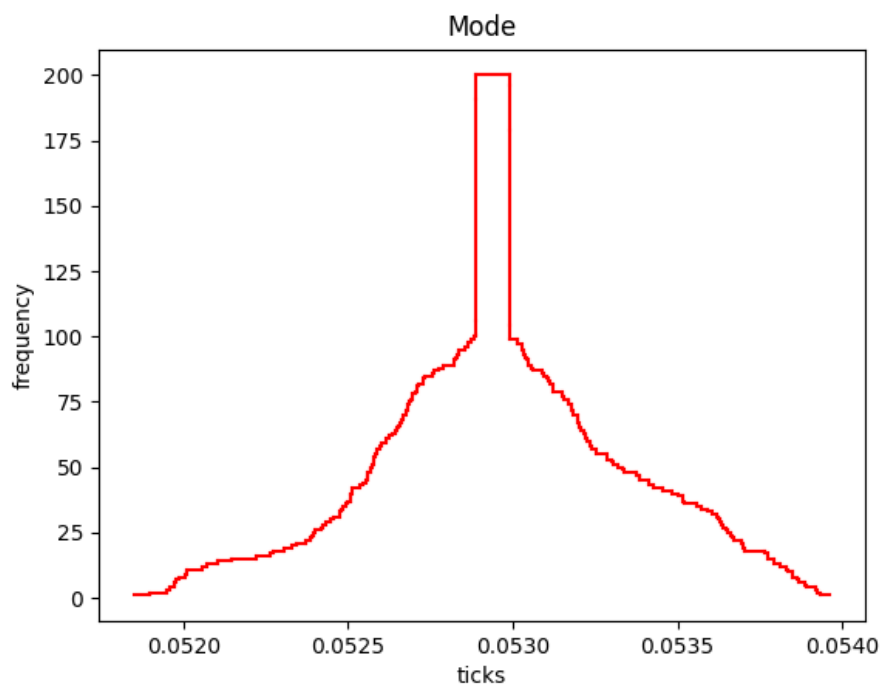


*Рис. 9. Мода выборки 2*

Для данных, полученных после вычитания дрейфовой компоненты:



*Рис. 10. Мода выборки 1*



*Рис. 11. Мода выборки 2*

### Ссылка на GitHub с реализацией

[https://github.com/dimerf99/interval\\_analysis/tree/main/lab\\_1](https://github.com/dimerf99/interval_analysis/tree/main/lab_1)

Файлы данных:

Канал 1\_700nm\_0.03.csv

Канал 2\_700nm\_0.03.csv

### Коэффициенты линейной регрессии

№ выборки	$A_i$	$B_i$
1	5.0867e-05	0.0492885
2	5.3843e-05	0.0529391