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 Пестряков Д.Д. /_____/

Руководитель

доцент, к.ф.-м.н. Баженов А.Н. /_____/

Санкт-Петербург

2022

Оглавление

Постановка задачи

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

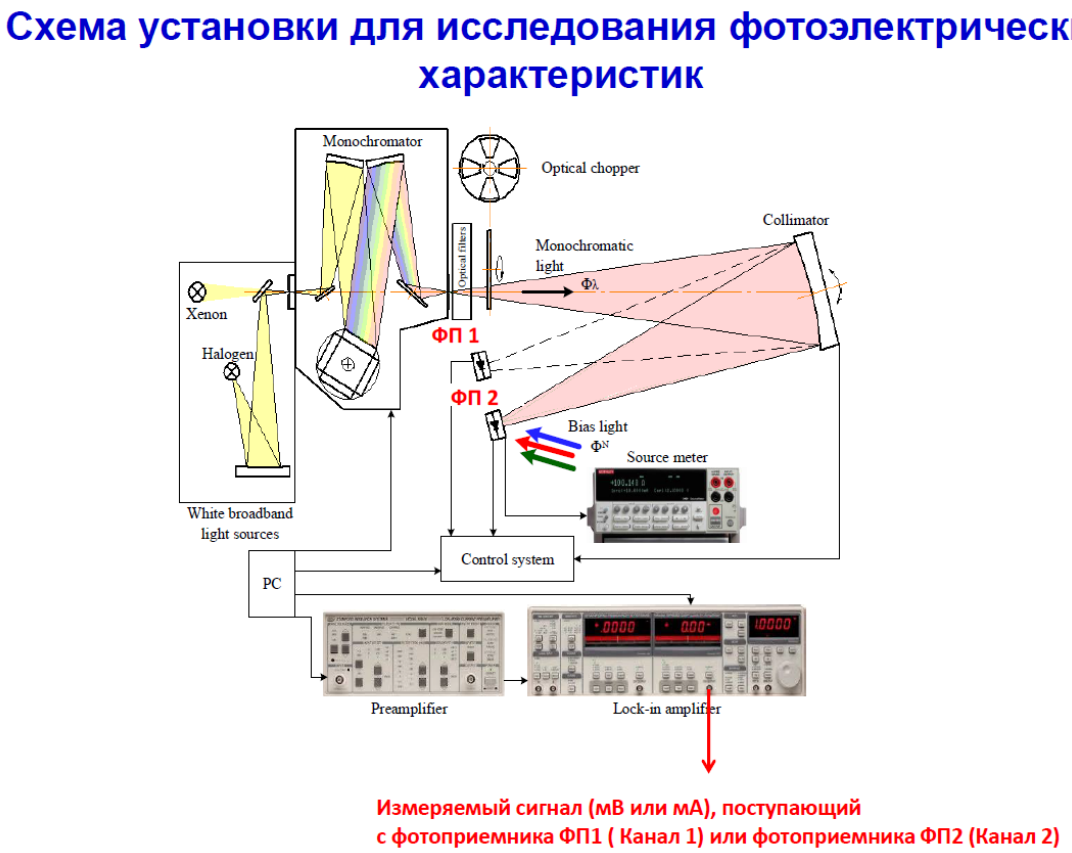

Рис. 1. Схема установки

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

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

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 Channel_1_800nm_0.23.csv и Channel_2_800nm_0.23.csv.

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

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

при помощи линейной регрессии, интервальных данных и коэффициента Жаккара.

Теория и методы

Представим данные таким образом, чтобы применить понятия статистики данных с интервальной неопределенностью. Один из распространенных способо получения 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 ε

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

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

Интервалы будем строить простым способом. 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

$$\varepsilon_{1n} = \varepsilon + L_1(n) - \overline{x_{1n}}, \text{если интервал ниже прямой}$$
$$\varepsilon_{1n} = \varepsilon + \underline{x_{1n}} - L_1(n), \text{если интервал выше прямой} \quad (4)$$

Окончательно, интервальные данные представимы в виде:

$$X_{1i} = x_{1i} + [-\varepsilon_{1i}, +\varepsilon_{1i}] \quad (5)$$

или кратко X_1 – множество всех интервальных данных, построенных по измерениям датчика ФП1.

Чтобы сделать интервальную величину более константной и в дальнейшем оценить совместность двух выборок экспериментальных измерений, вычтем 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} выпоним аналогичные вычисления. Найдем линейную 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. Будем использовать мультимеру Жаккакра, то есть ее модификацию для интервальных данных:

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

Мера Жаккара $-1 \le JK \le 1$ 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 $\left[\underline{R}, \overline{R}\right]$ – 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.2 представлены экспериментальные данные, измеренные двумя датчиками. На рис. 3 и 4 показаны построенные согласно описанной выше теории интервальные данные и линейная регрессия с коэффициентами $A_1 \approx 3.242 \cdot 10^{-6}, B_1 \approx 0.472, A_2 \approx 5.454 \cdot 10^{-6}, B_2 \approx 0.503$.

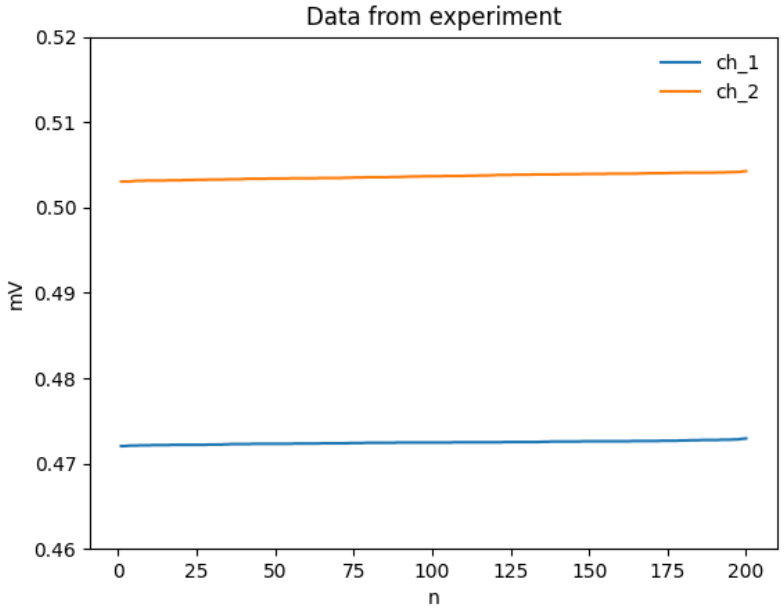

Рис. 2. Две выборки экспериментальных данных, измеренным датчиками

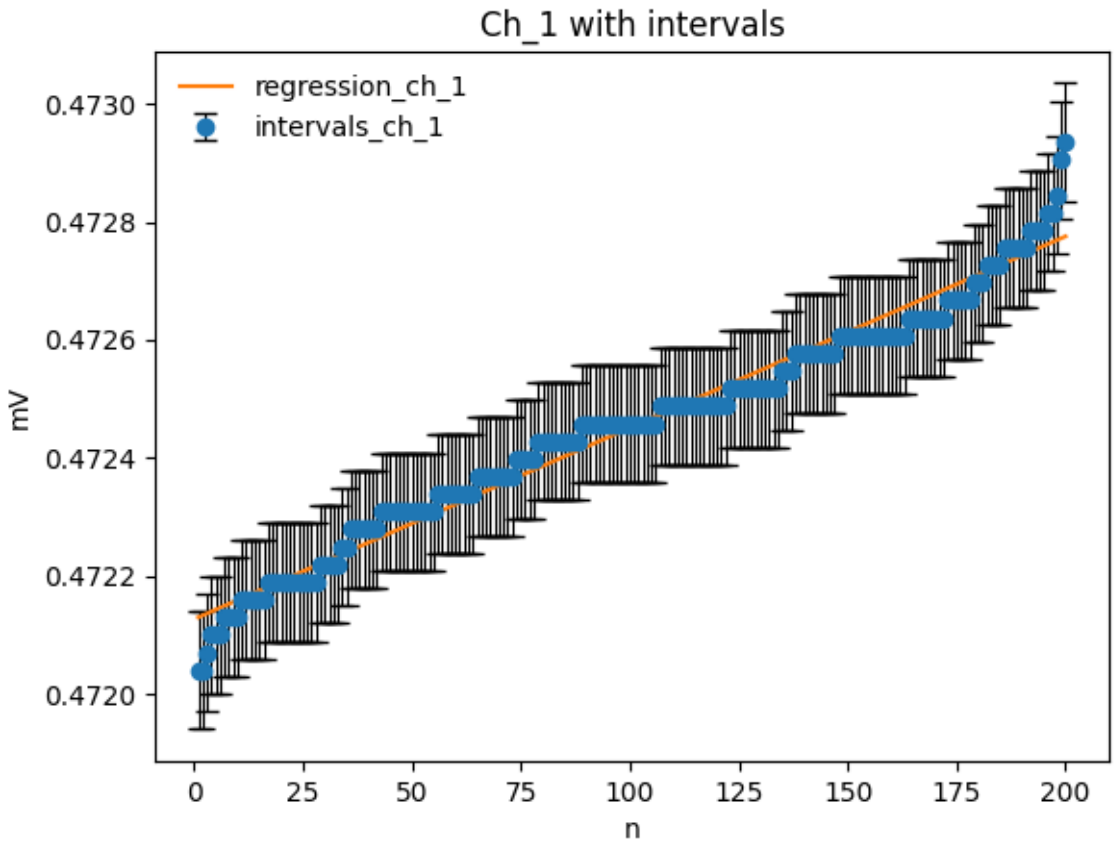

Рис. 3. Интервальные данные первой выборки и линейная регрессия

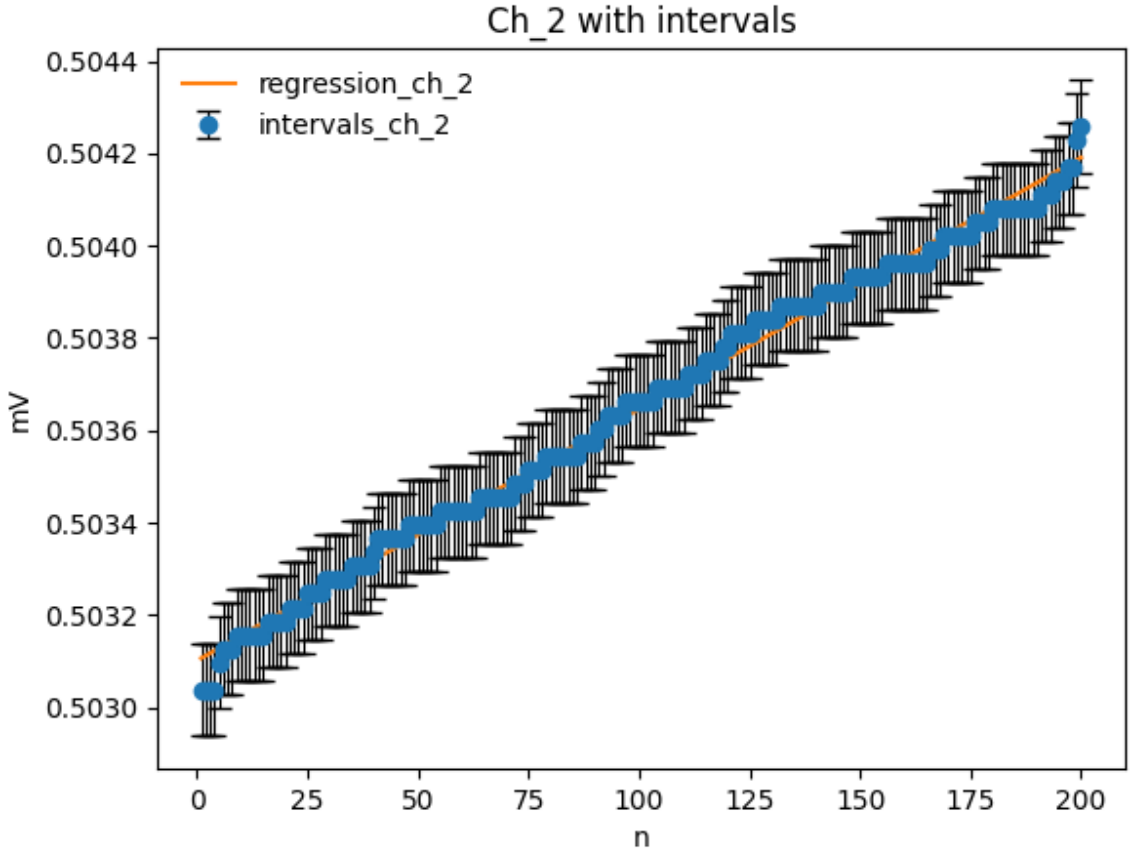

Рис. 4. Интервальные данные второй выборки и линейная регрессия

Исходя из Рис. 3 и Рис. 4 можно отметить следующее:

- Данные с канала 1 носят линейно-ступенчатый характер с изломами в первых и последних 20 точках
- Данные с канала 2 имеют кусочно-линейный характер, но с ярко выраженной регрессией, а именно точки примерно с 3 по 100 и с 127 по 175 (что составляет около 75% всех точек) имеют от линейной регрессии малое отклонение
- Для канала 2 прямая линейной регрессии пересекает все интервалы, для канала 1 не пересекает только 2 последних измерения, поэтому их пришлось растянуть

На Рис. 5 представлена гистограмма весов растяжения интервалов 1 канала. Для канала 2 такую не приводим, так как там растяжения не было

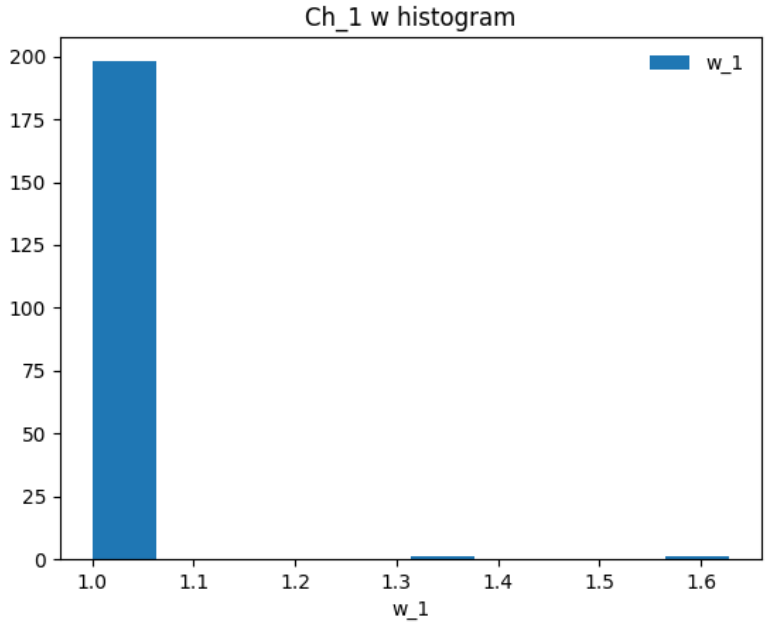

Рис. 5. Гистограмма весов растяжения интервалов для 1 канала

На Рис. 6 и Рис. 7 представлена константа и интервалы после вычитания дрейфовой компоненты из регрессии

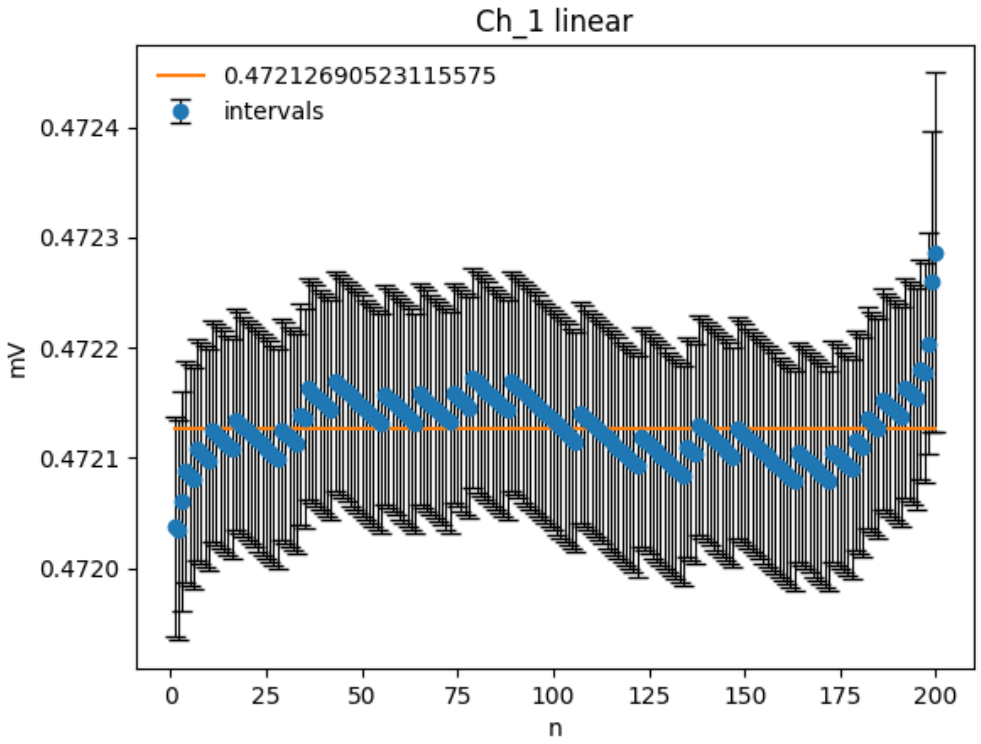

Рис. 6. Интервалы после вычитания дрейфовой компоненты для канала 1

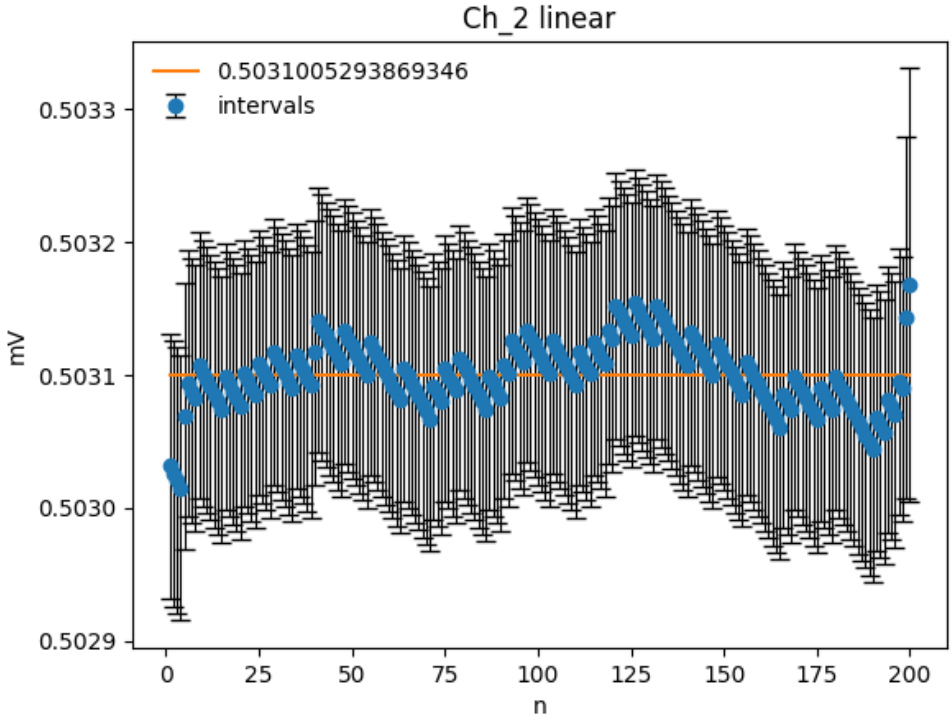

Рис. 7. Интервалы после вычитания дрейфовой компоненты для канала 2

Можно отметить, что измерения канала 2 визуально имеют более широкий канал совместности, чем измерения канала 1.

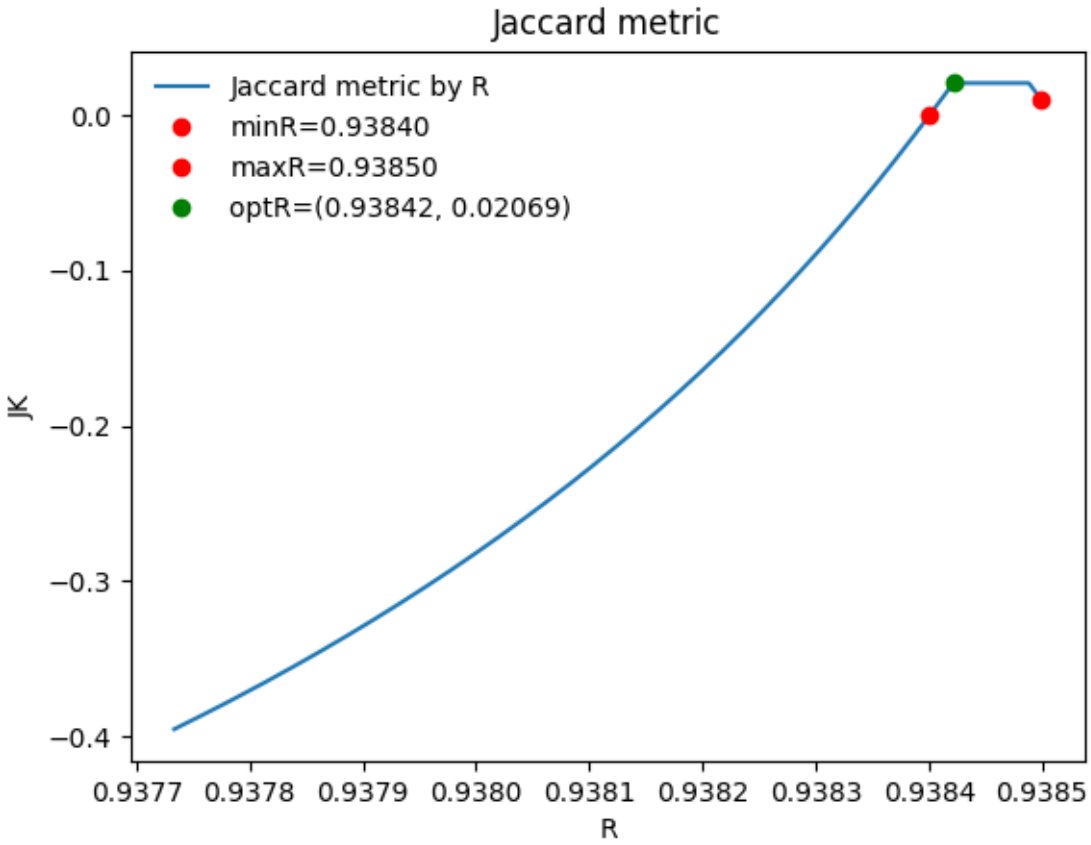

Рис. 8. Мера Жаккара для различных R

На Рис. 8 представлено изменение меры Жаккара в зависимости от калибровочного коэффициента R.
Внешняя оценка (округленная до 4 знаков) для R = [0.9377, 0.9385]
Размах внешней оценки равен 0.0008
Внутренняя оценка для R = [0.9384, 0.9385]
Размах внутренней оценки равен 0.0001
То есть, размах внутренней оценки стал в 10 раз меньше. С учетом того, что внутренняя оценка имеет достаточно узкий размах, то это можно считать приемлемым результатом.
Оптимальное R можно взять 0.93842. При этом R мера Жаккара принимает значение 0.02069, что говорит о наличии пересечения, представленного некоторым интервалом.

Приложение

Ссылка на GitHub с реализацией:
DanilPestryakov/intervals_first_lab (github.com)