Санкт-Петербургский политехнический университет имени
Петра Великого

Институт прикладной математики и механики
Высшая школа прикладной математики и физики

Математическая статистика
Отчёт по лабораторной работе №9

Выполнил:

Студент: Парусов Владимир
Группа: 5030102/90201

Принял:

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

2022 г.

Содержание

Список иллюстраций

1. Постановка задачи

Исследование из области солнечной энергетики. На Рис. 1 показана схема установки для исследования фотоэлектрических характеристик.

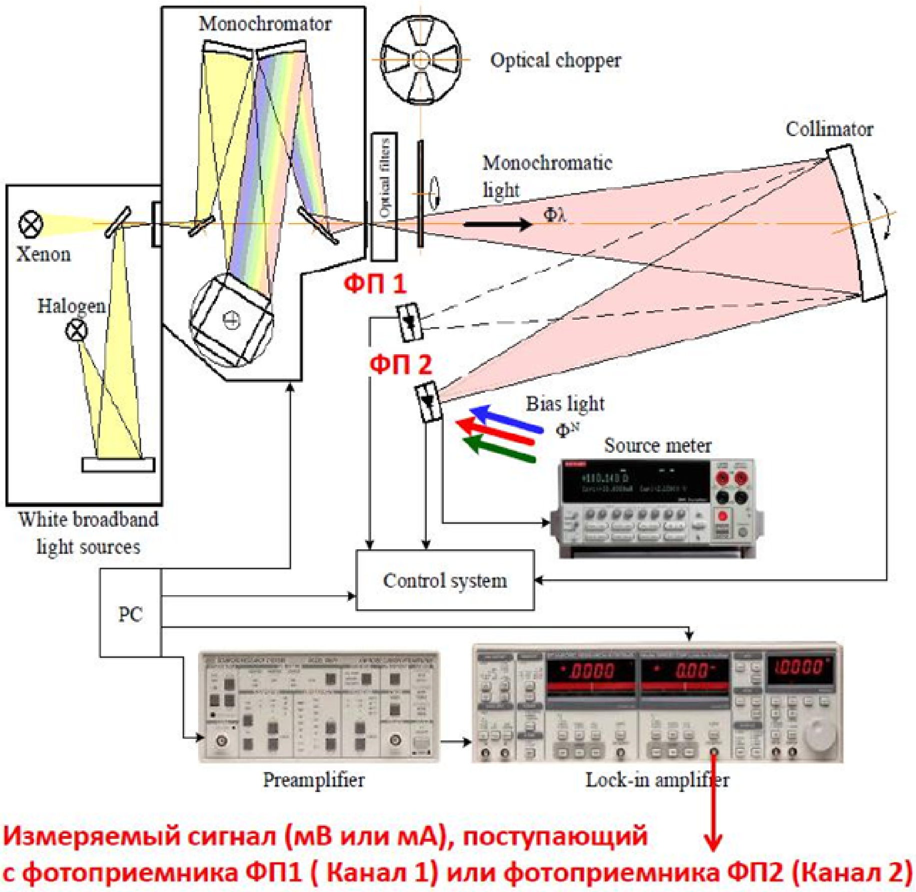

Рис. 1. Схема установки

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

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

QE_2, QE_1 – эталонная эффективность эталонного и исследуемого датчика, I_2, I_1 – измеренные токи. Данные с датчиков находятся в файлах Ch2_800nm_0.03.csv и Ch1_800nm_0.03.csv.

Требуется определить коэффициент калибровки

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

при помощи линейной регрессии на множестве интервальных данных и коэффициента Жаккара.

2. Теория

2.1. Представление данных

В первую очередь представим данные таким образом, чтобы применить понятия статистики данных с интервальной неопределённостью. Один из распространённых способов получения интервальных результатов в первичных измерениях — это «обинтерваливание» точечных значений, когда к точечному базовому значению $\dot{x}$, которое считывается по показаниям измерительного прибора прибавляется интервал погрешности ϵ.

$$x = \dot{x} + \epsilon \tag{3}$$

Интервал погрешности зададим как

$$\epsilon = [-\xi, \xi] \tag{4}$$

В конкретных измерениях примем $\xi = 10^{-4}$ мВ.

Согласно терминологии интервального анализа, рассматриваемая выборка - это вектор интервалов. или интервальный вектор $x = (x_1, x_2, x_3, x_4, ...)$.

Информационным множеством в случае оценивания единичной физической величины по выборке интервальных данных будет также интервал, который называют информационным интервалом. Неформально говоря, это интервал, содержащий значения оцениваемой величины, которые «совместны» с измерениями выборки («согласуются» с данными этих измерений).

2.2. Простая линейная регрессия

2.2.1. Описание модели

Регрессионную модель описания данных называют простой линейной, если заданный набор данных аппроксимируется прямой с внесённой добавкой в виде некоторой нормально распределённой ошибки:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i \in \overline{1, n} \tag{5}$$

где

$\{x_n\}_{n\in\mathbb{N}}$ – заданные значения,

$\{y_n\}_{n\in\mathbb{N}}$ – параметры отклика,

$\{\varepsilon_n\}_{n\in\mathbb{N}}$ – независимые, центрированные, нормально распределённые случайные величины с неизвестной дисперсией δ, суть предполагаемые погрешности,

β_0, β_1 – параметры, подлежащие оцениванию.

В данной модели мы считаем, что у заданных значений нет погрешности (пренебрегаем ей). Полагаем, что основная погрешность получается при измерении $\{y_n\}_{n\in\mathbb{N}}$.

2.2.2. Метод наименьших модулей

Данный метод основан на минимизации l^1-нормы разности последовательностей полученных экспериментальных данных $\{y_n\}$ и значений аппроксимирующей функции $f(\{x_n\})$.

$$\|f(\{x_n\}) - \{y_n\}\|_{l^1} \xrightarrow[\{\lambda_i\}]{} min \tag{6}$$

В данном случае мы ставим задачу линейного программирования таким образом, чтобы найти не только коэффициенты β_0 и β_1, но и вектор w на который стоит домножить погрешности наших интервальных данных. Тогда задача ставится так:

$$\sum |w_i| \to \min, i \in \overline{1,n} \tag{7}$$

При ограничениях

$$\beta_0 + \beta_1 * x_i - w_i * \xi \leq y_i, i \in \overline{1,n} \tag{8}$$

$$\beta_0 + \beta_1 * x_i + w_i * \xi \leq y_i, i \in \overline{1,n} \tag{9}$$

2.3. Предварительная обработка данных

Из последующих результатов ясно, что для оценки коэффициента калибровки необходима предварительная обработка данных. Для этого можем задаться линейной моделью дрейфа.

$$Lin_i(n) = A_i + B_i * n, n \in \overline{1,N} \tag{10}$$

Поставив задачу линейного программирования воспользуемся Методом наименьших модулей (7) и найдём коэффициенты A_i, B_i и вектор w_i множителей коррекции данных(где i = 1 соответствует данным с ФП1, а i = 2 соответственно ФП2). Множитель коррекции данных необходимо применить к погрешностям выборки, чтобы получить данные согласующиеся с нашей линейной моделью дрейфа.

$$I_i^f(n) = \dot{x}(n) + \epsilon * w_i(n), n \in \overline{1,N} \tag{11}$$

После построения линейной модели дрейфа необходимо построить «спрямлённые» данные выборки, вычтя из исходных данных (с применённым множителем коррекции данных) «дрейфовую» компоненту.

$$I_i^c(n) = I_i^f(n) - B_i * n, n \in \overline{1,N} \tag{12}$$

2.4. Коэффициент Жаккара

В различных областях анализа данных в науках о Земле, биологии, информатике используют множество мер сходства множеств. Иначе их называют коэффициентами сходства. Нами рассматривается модификация индекса Жаккара для интервальных данных:

$$JK(x) = \frac{wid(\bigwedge x_i)}{wid(\bigvee x_i)} \tag{13}$$

В качестве меры рассматривается ширина интервала, а вместо операций пересечения и объединения – операции взятия минимума и максимума по включению двух величин в интервальной арифметике (Каухера). Заметим что минимум по включению может быть неправильным интервалом, а значит данный коэффициент будет нормирован в отрезке [-1,1]

2.5. Процедура оптимизации

Для поиска оптимального параметра калибровки поставим следующую задачу максимизации:

$$JK(x_{all}(R)) \to \max \tag{14}$$

Где JK это коэффциент Жаккара((13)) x_{all} это выборка полученная как

$$x_{all} = I_1^f * R \frown I_2^f \tag{15}$$

где $\frown$ обозначена операция конкатенации двух выборок, а I_1^f и I_2^f посчитаны по формуле (12). Поиск будем проводить методом дихотомии, а поиск оптимального R Будем проводить в отрезке [1, 3]. Тогда оптимальное R это и будет $R21$((2)).

3. Реализация

Данная работа реализована на языке программирования Python с использованием редактора VIM и библиотек NumPy, MatPlotLib, Statsmodels, Scipy в ОС Ubuntu 19.04.

Отчёт подготовлен с помощью компилятора pdflatex и среды разработки TeXStudio.

4. Результаты

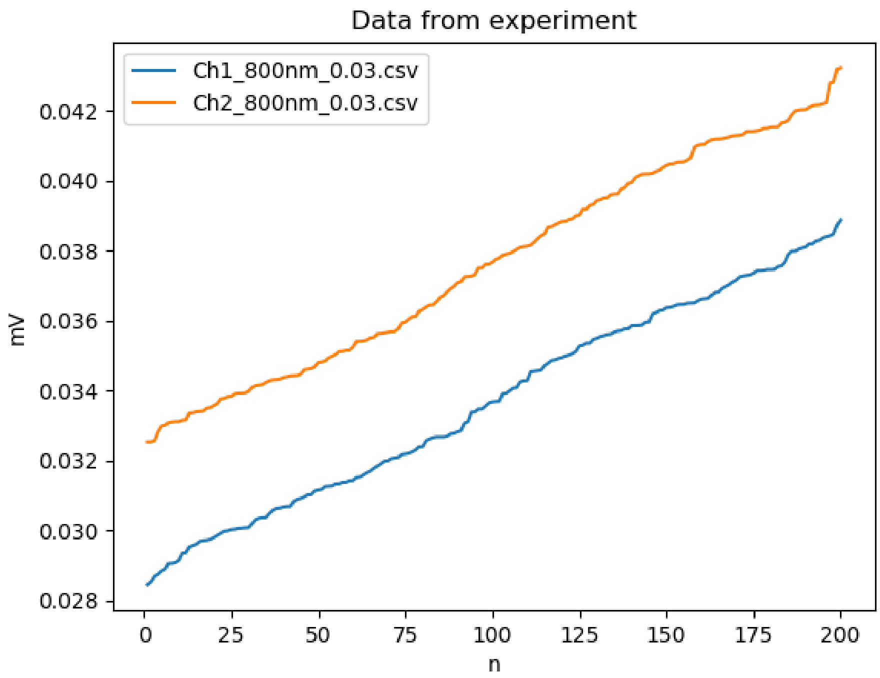

Рис. 2. Выборки полученные в ходе эксперимента

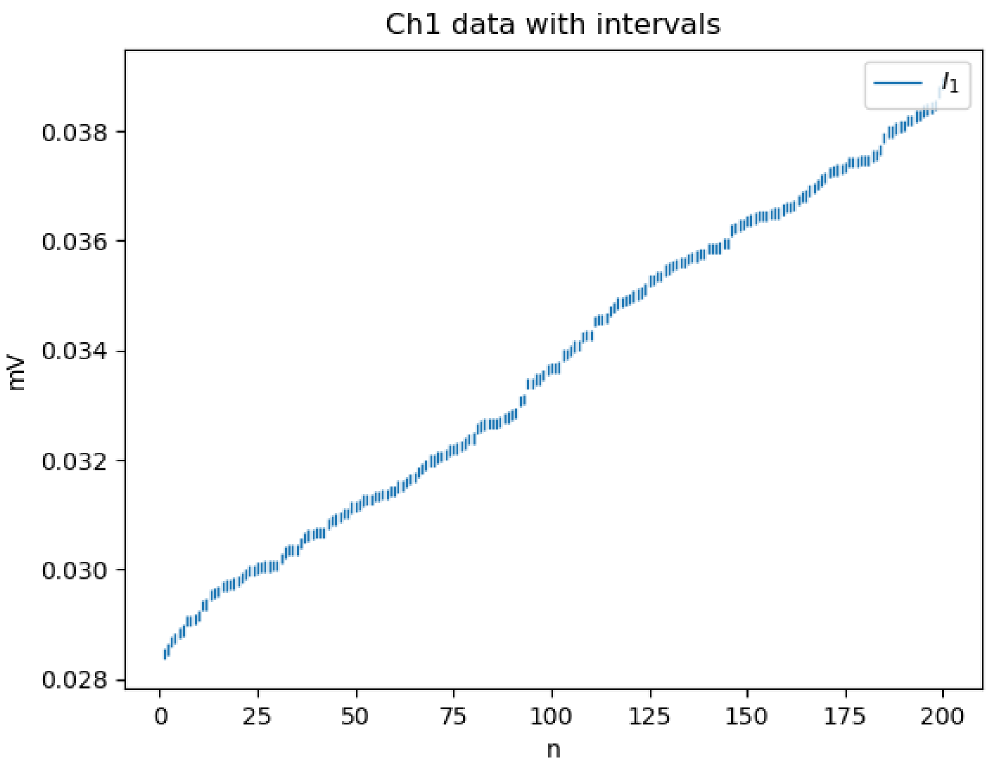

Рис. 3. Интервальное представление данных с первой выборки

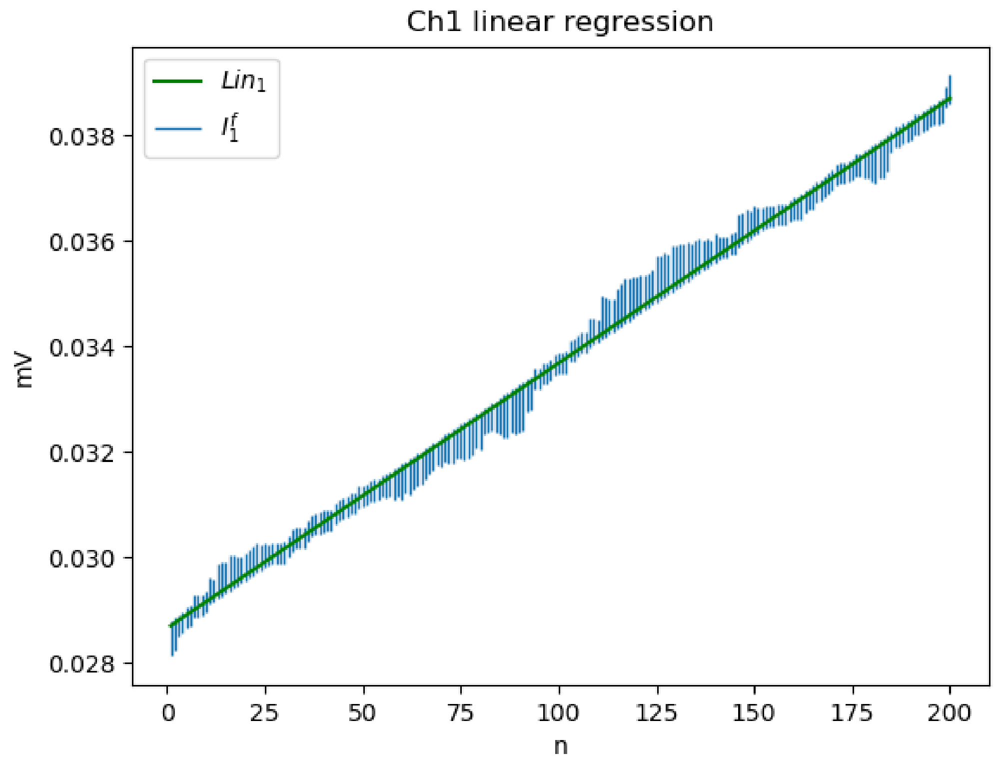

Рис. 4. I_1^f и Lin_1

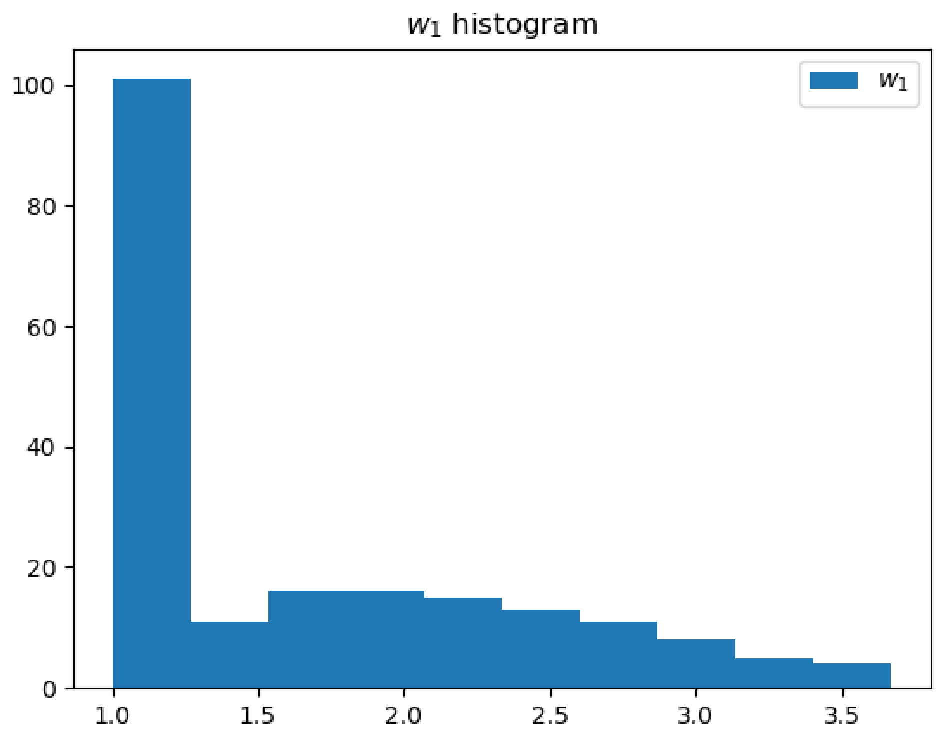

Рис. 5. Гистограмма значений w_1

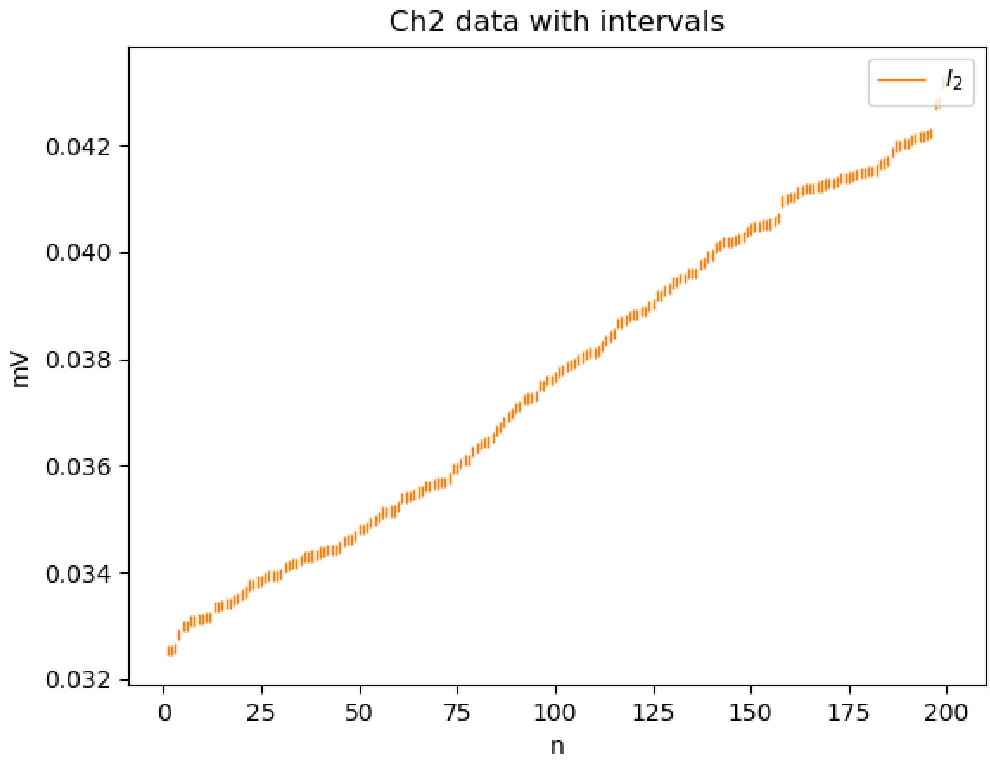

Рис. 6. Интервальное представление данных со второй выборки

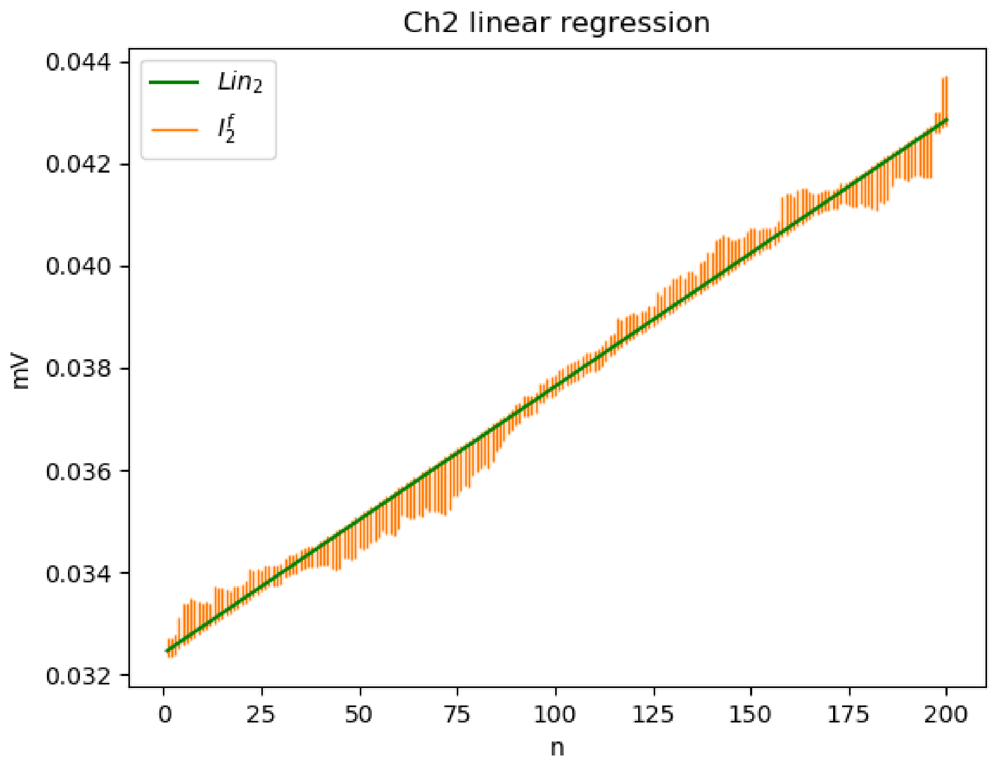

Рис. 7. I_2^f и Lin_2

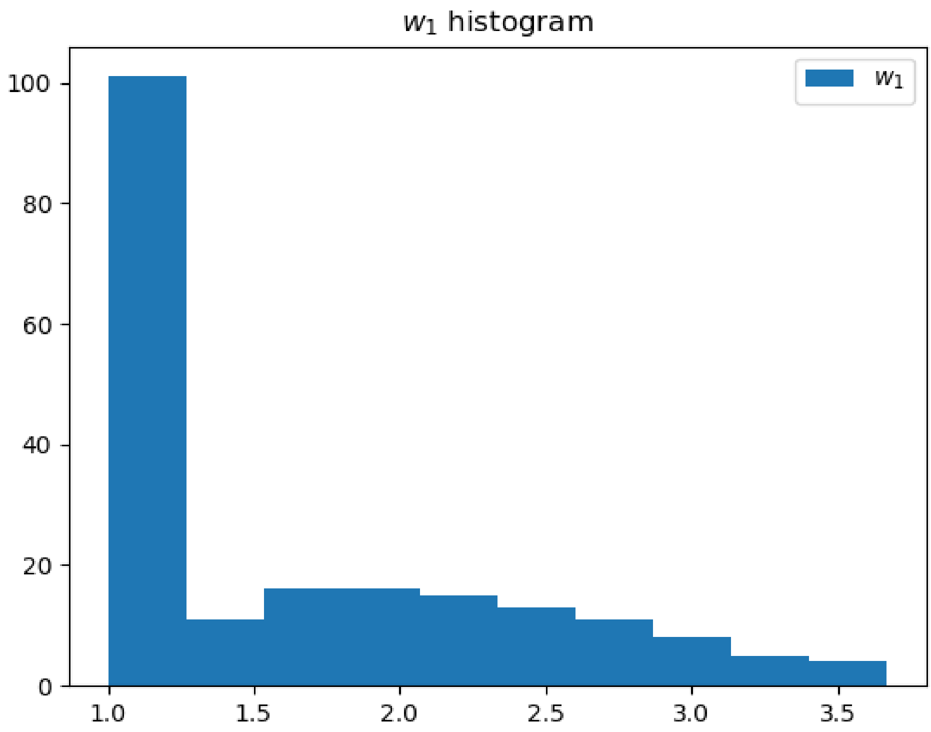

Рис. 8. Гистограмма значений w_2

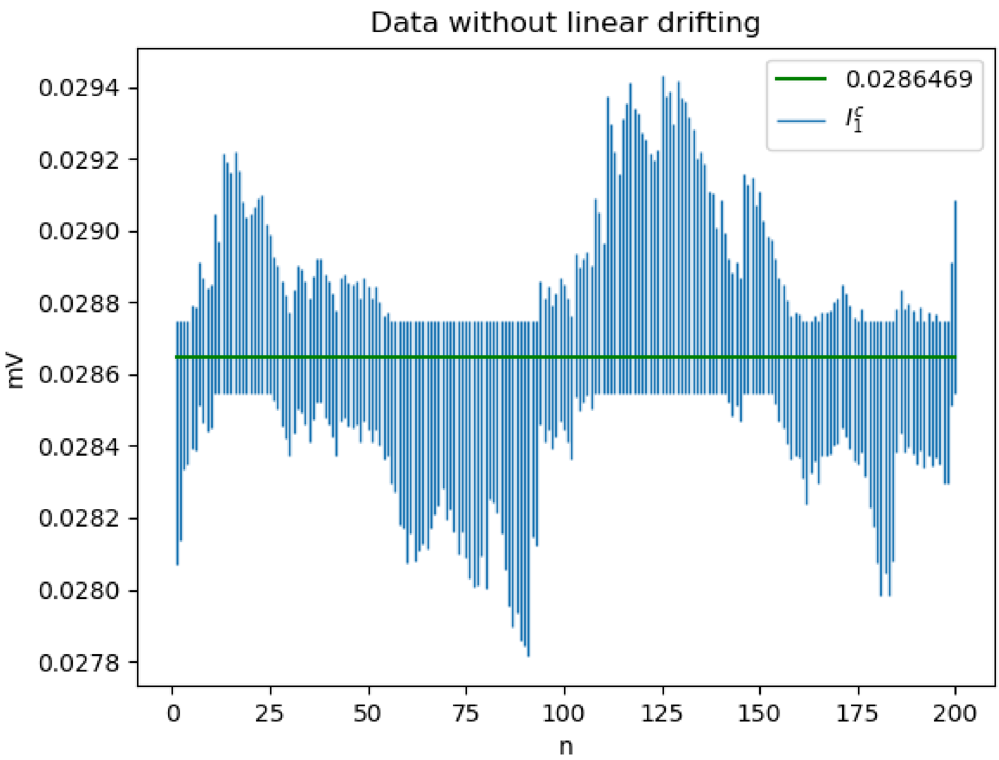

Рис. 9. I_1^c

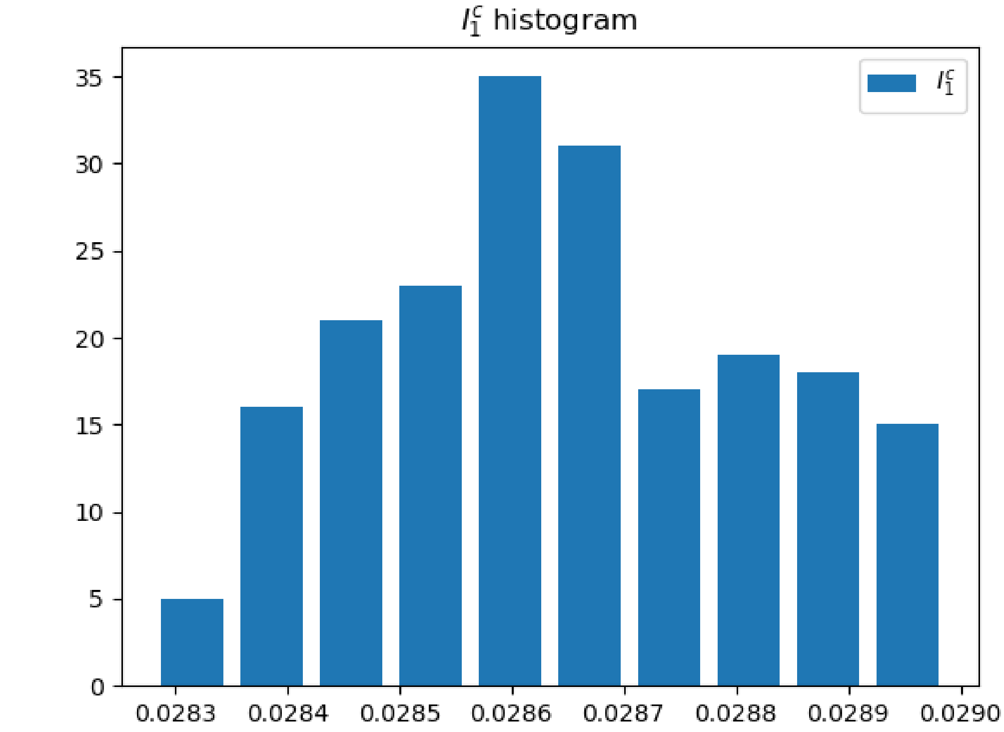

Рис. 10. Гистограмма I_1^c

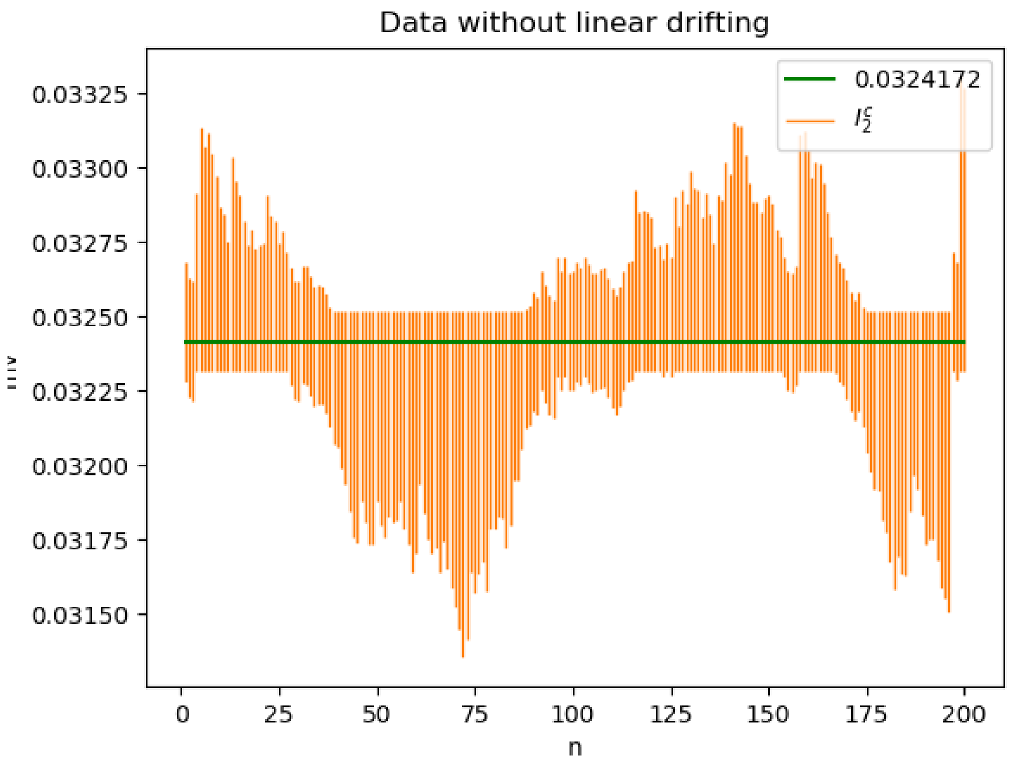

Рис. 11. I_2^c

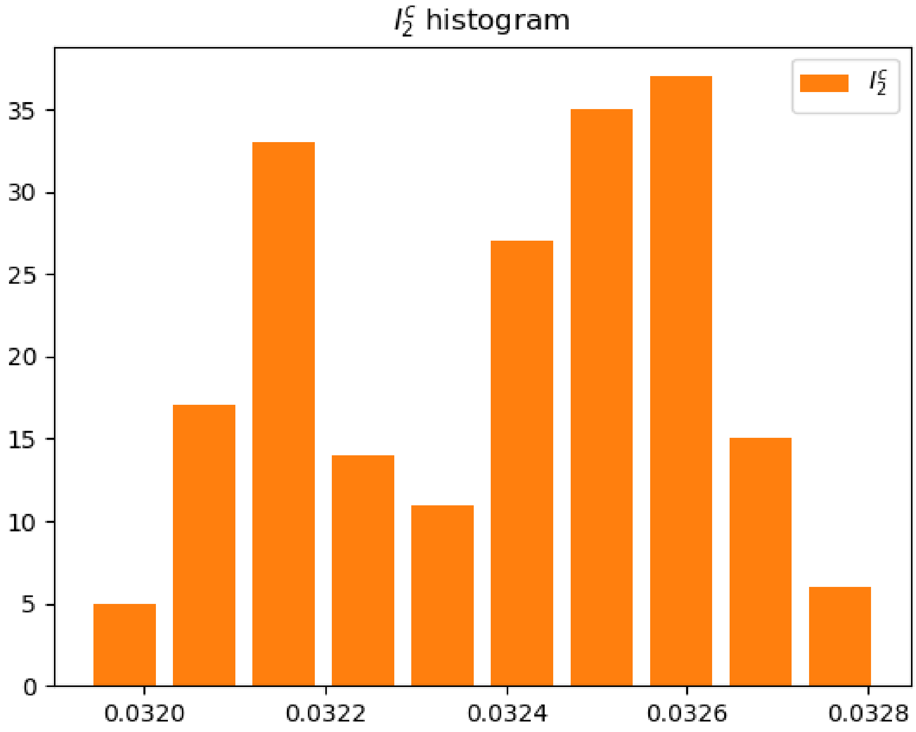

Рис. 12. Гистограмма I_2^c

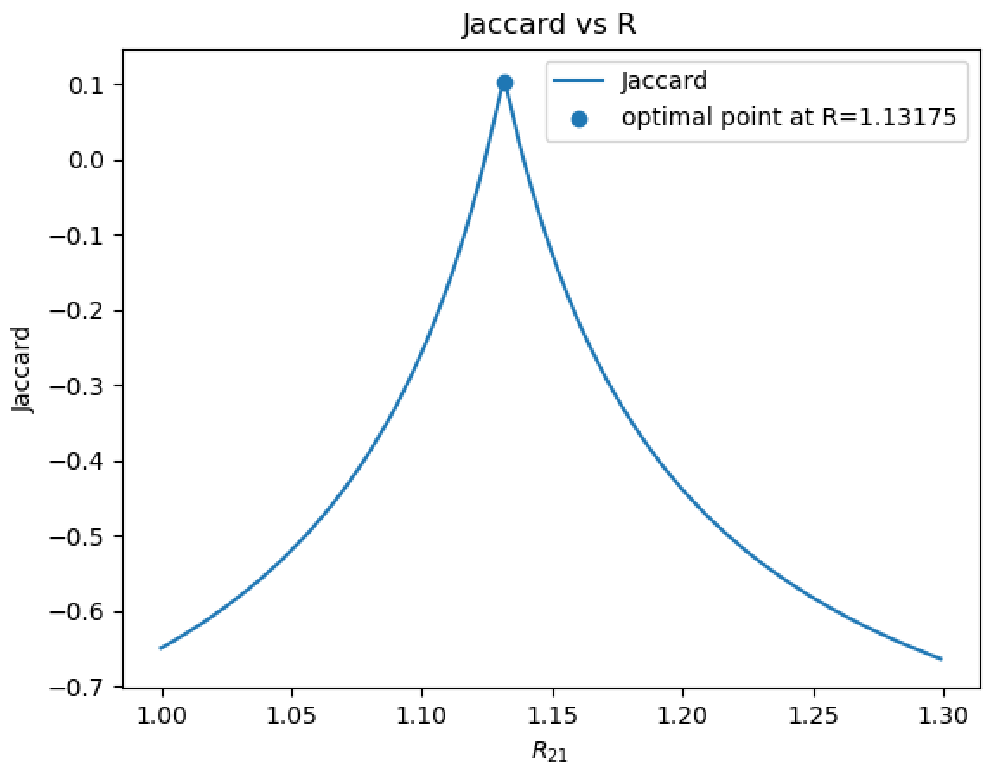

Рис. 13. Значение коэффициента Жаккара от калибровочного множителя

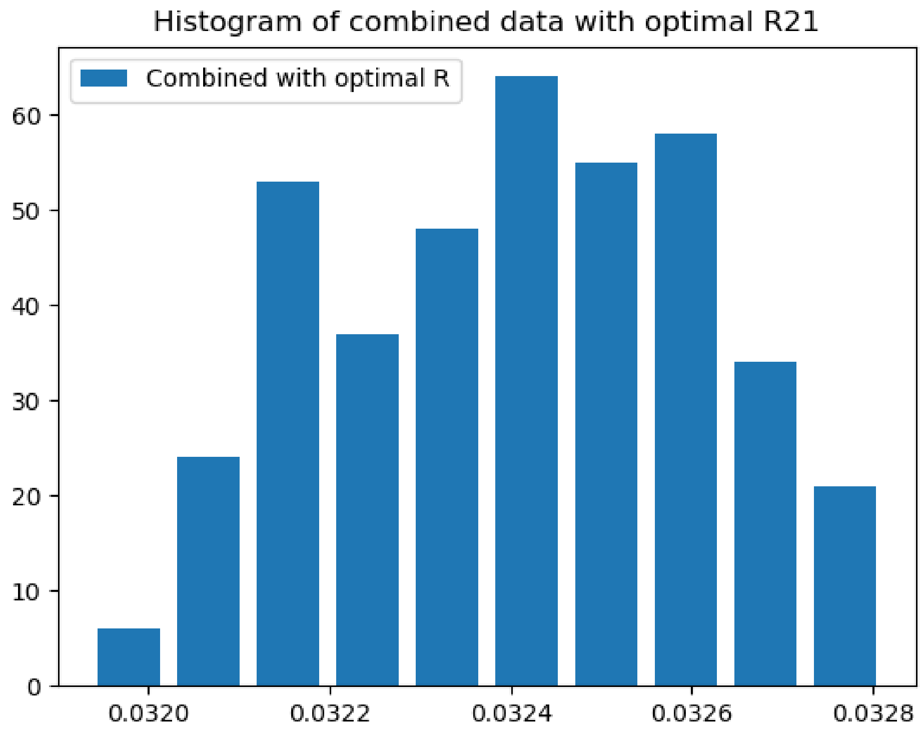

Рис. 14. Гистограмма объединённой выборки при оптимальном значении R_{21}

5. Обсуждение

5.1. Гистограммы w_1 и w_2

Рассмотрим Рис.5 и Рис.8. По преобладанию множителя 1, можно сказать что примерно половина данных не требует коррекции. Этот факт свидетельствует о том, что линейная модель дрейфа данных является разумным приближением.

5.2. Гистограммы I_1^f, I_2^f и Совмещённой выборки с оптимальным коэффициентом калибровки

Рассмотрим Рис.10 и Рис.12. Заметим что выборка I_1^f имеет характерную особенность в виде "пика"по центру, а I_2^f имеет 2 "пика"вокруг центра. В совмещенной выборке на Рис. 14 мы можем заметить что характерные особенности обеих гистограмм перенеслись на данную гистограмму, и можно наблюдать 3 "пика". При этом границы гистограммы совпадают с границами I_2^f.

5.3. Коэффициент Жаккара

Рассмотрим Рис.13. Оптимальное значение параметра калибровки R_{21} можно принять равным 1.13175. Помимо этого можно сказать, что поведение коэффициента Жаккара как функции от параметров несёт в себе гораздо больше информации, чем просто значение этого коэффициента. Например, в нашем эксперименте, максимум индекса Жаккара имеет значение чуть большее чем 0.1, но совершенно не близкое к 1. Это связано с наличием различных погрешностей, которые на практике невозможно устранить, но несмотря на их наличие, поведение функции Жаккара позволило найти оптимальный калибровочный коэффициент. Однако знак коэффициента Жаккара может свидетельствовать о том, является ли минимум по включению правильным интервалом, что в свою очередь говорит о совместности двух выборок. Таким образом можно сказать что область где $JK(R_{21}) >= 0$ является оценкой искомой величины R_{21}

6. Литература

[1] А.Н. Баженов, С.И. Жилин, С.И.Кумков, С.П.Шарый. Обработка и анализ данных с интервальной неопределенностью 2022.

[2] Коэффициент Жаккара https://en.wikipedia.org/wiki/Jaccard_index

[3] С.И. Жилин. Примеры анализа интервальных данных в Octave. https://github. interval-examples

7. Приложения

1. Репозиторий с кодом программы:

   https://github.com/sairsey/MathStats