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.

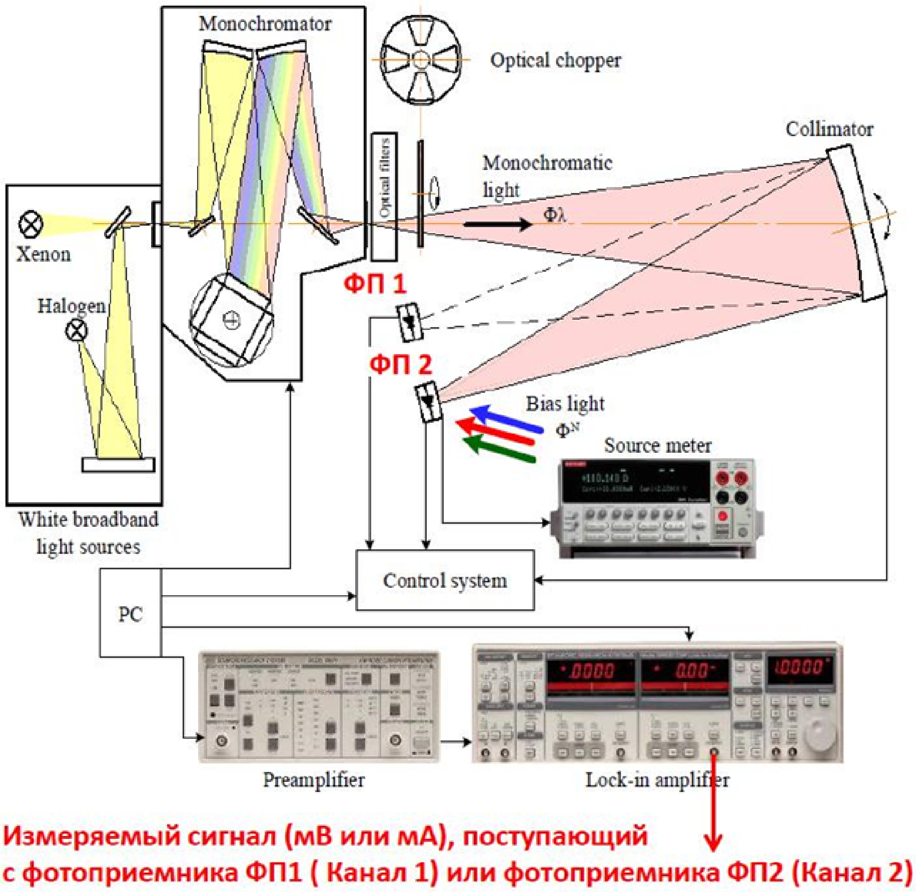

Рис. 1. Схема установки

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 $f(\{x_n\})$. Увы, автору данного отчёта неизвестно метода, позволяющего решить, как в случае МНК, данную задачу минимизации для линейной комбинации заданного количества базисных функций, действующих на $\mathbb{R}$, однако метод позволяет решать задачу для линейной функции любой размерности:

$$\|[\boldsymbol{a}, \{x_n\}] - \{y_n\}\|_{l^1} \xrightarrow[\{\lambda_i\}]{} min \tag{6}$$

Данную задачу минимизации можно решать точно, например, используя алгоритм спуска по узловым направлениям. Метод основан на теореме о том, что точка минимума искомой функции лежит в одной из точек нарушения дифференцируемости минимизируемой функции (в точке, где какой-либо модуль обращается в ноль), заданного данными и реализует направленный перебор всех таких точек [2].

Кроме того, можно решать численно, методом Вейсфельда [3]. Суть метода в том, что вместо решения негладкой задачи мы на каждой итерации минимизируем взвешенную l^2-норму разности, где вес равен величине, обратной невязке на предыдущем шаге (таким образом, мы делим квадрат невязки на текущем шаге на невязку на предыдущем, и получаем "почти невязку" в первой степени, что соответствует l^1-норме).

2.3. Простая линейная регрессия

2.3.1. Описание модели

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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i \in \overline{1, n} \tag{7}$$

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 $\{y_n\}_{n\in\mathbb{N}}$.

2.3.2. Метод наименьших модулей

Данный метод основан на минимизации l^1-нормы разности последовательностей полученных экспериментальных данных $\{y_n\}$ и значений аппроксимирующей функции $f(\{x_n\})$. Увы, автору данного отчёта неизвестно метода, позволяющего решить, как в случае МНК, данную задачу минимизации для линейной комбинации заданного количества базисных функций, действующих на $\mathbb{R}$, однако метод позволяет решать задачу для линейной функции любой размерности:

$$\|[\boldsymbol{a}, \{x_n\}] - \{y_n\}\|_{l^1} \xrightarrow[\{\lambda_i\}]{} min \tag{8}$$

Данную задачу минимизации можно решать точно, например, используя алгоритм спуска по узловым направлениям. Метод основан на теореме о том, что точка минимума искомой функции лежит в одной из точек нарушения дифференцируемости минимизируемой функции (в точке, где какой-либо модуль обращается в ноль), заданного данными и реализует направленный перебор всех таких точек [2].

Кроме того, можно решать численно, методом Вейсфельда [3]. Суть метода в том, что вместо решения негладкой задачи мы на каждой итерации минимизируем взвешенную l^2-норму разности, где вес равен величине, обратной невязке на предыдущем шаге (таким образом, мы делим квадрат невязки на текущем шаге на невязку на предыдущем, и получаем “почти невязку” в первой степени, что соответствует l^1-норме).

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

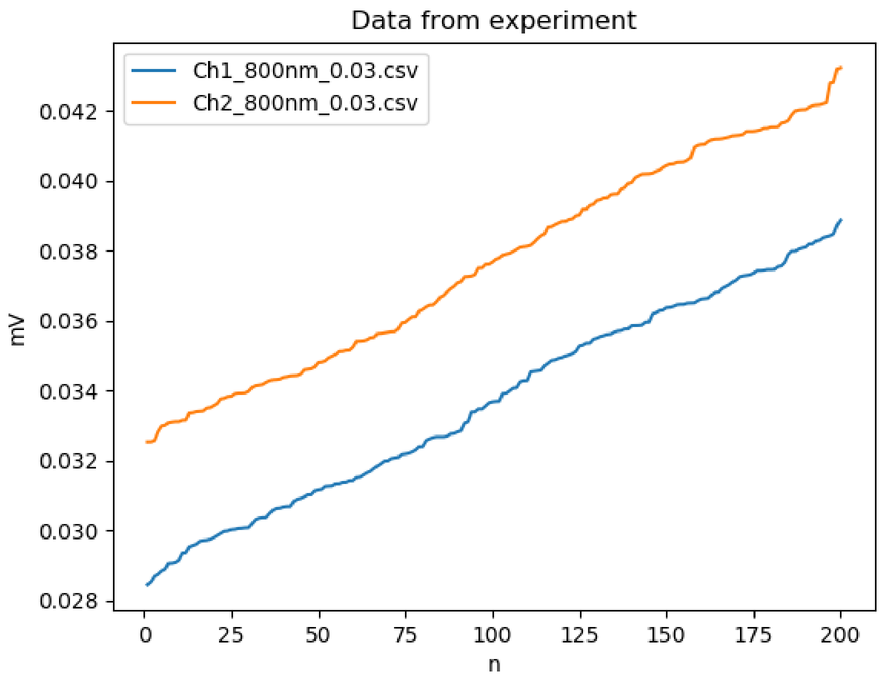

Рис. 2. Выборки полученные в ходе эксперимента

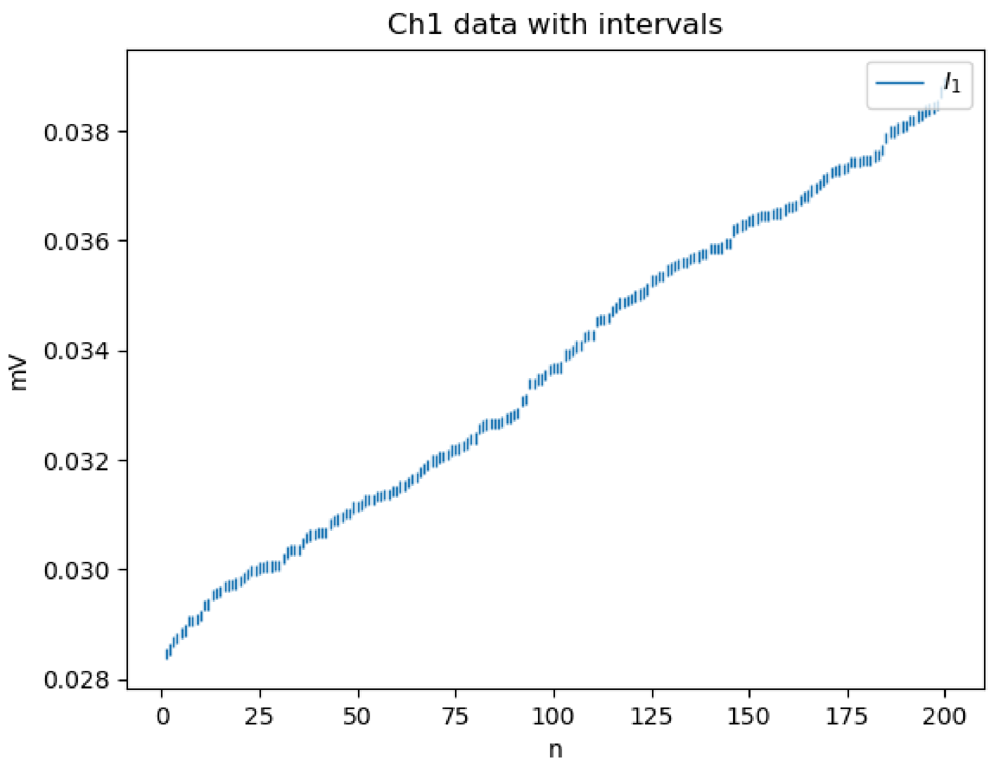

Рис. 3. Интервальное представление данных с первой выборки

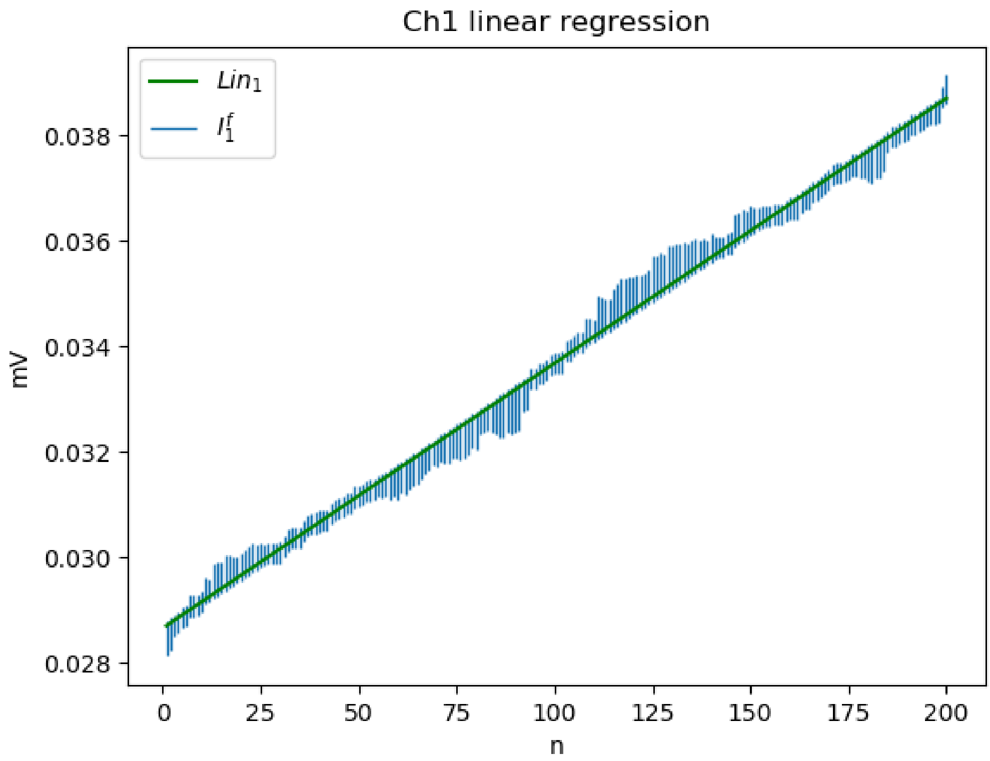

Рис. 4. I_1^f и Lin_1

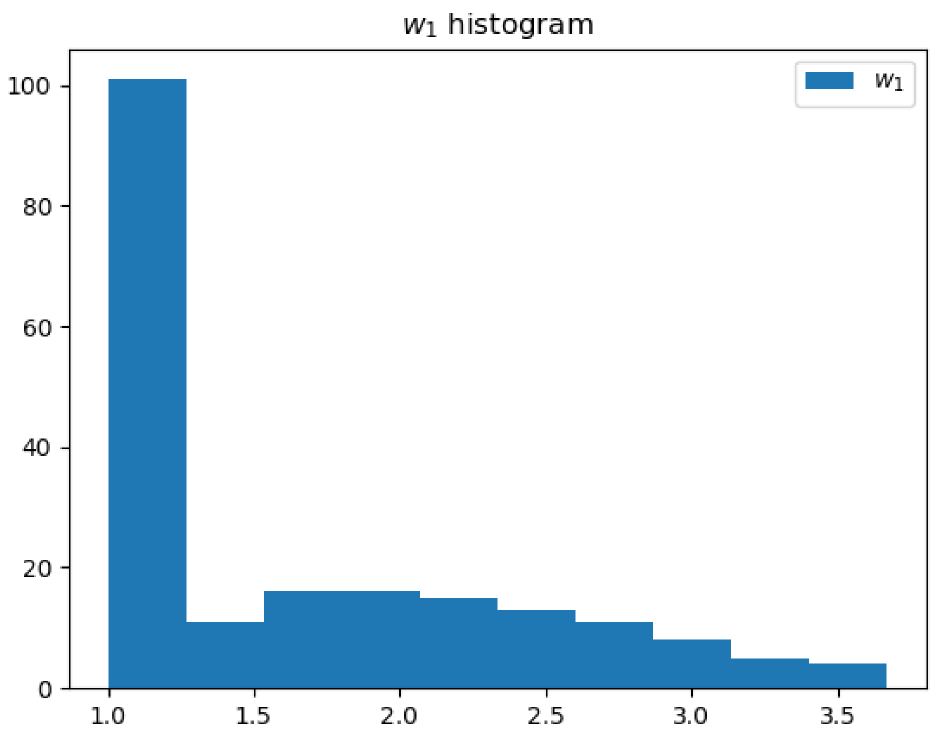

Рис. 5. Гистограмма значений w_1

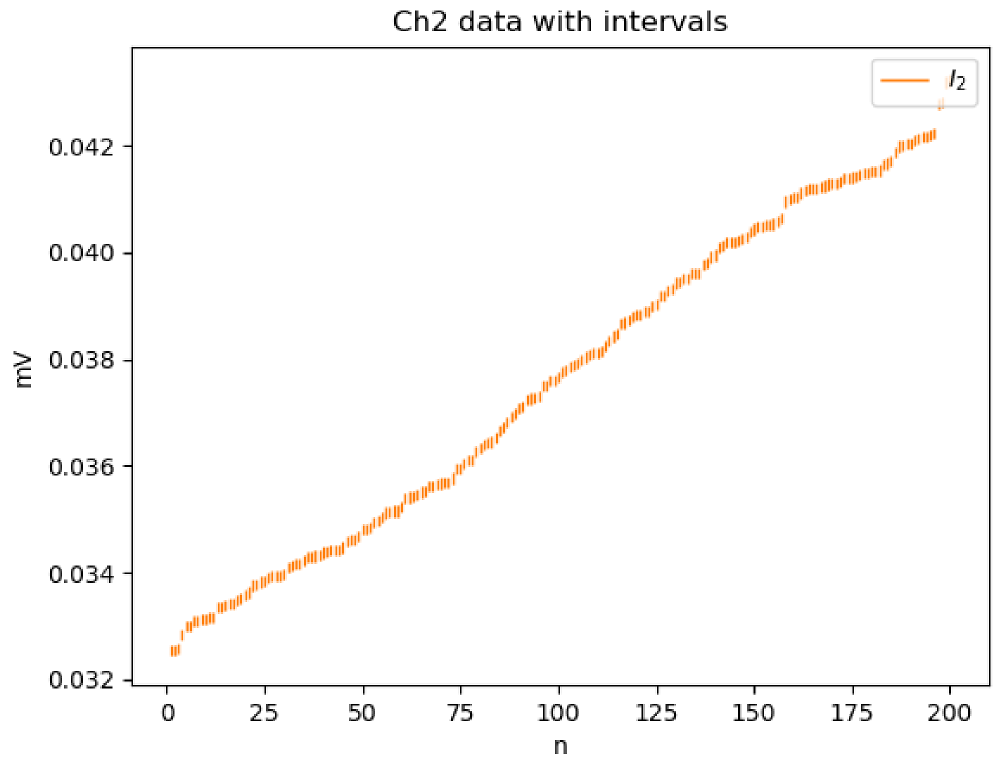

Рис. 6. Интервальное представление данных со второй выборки

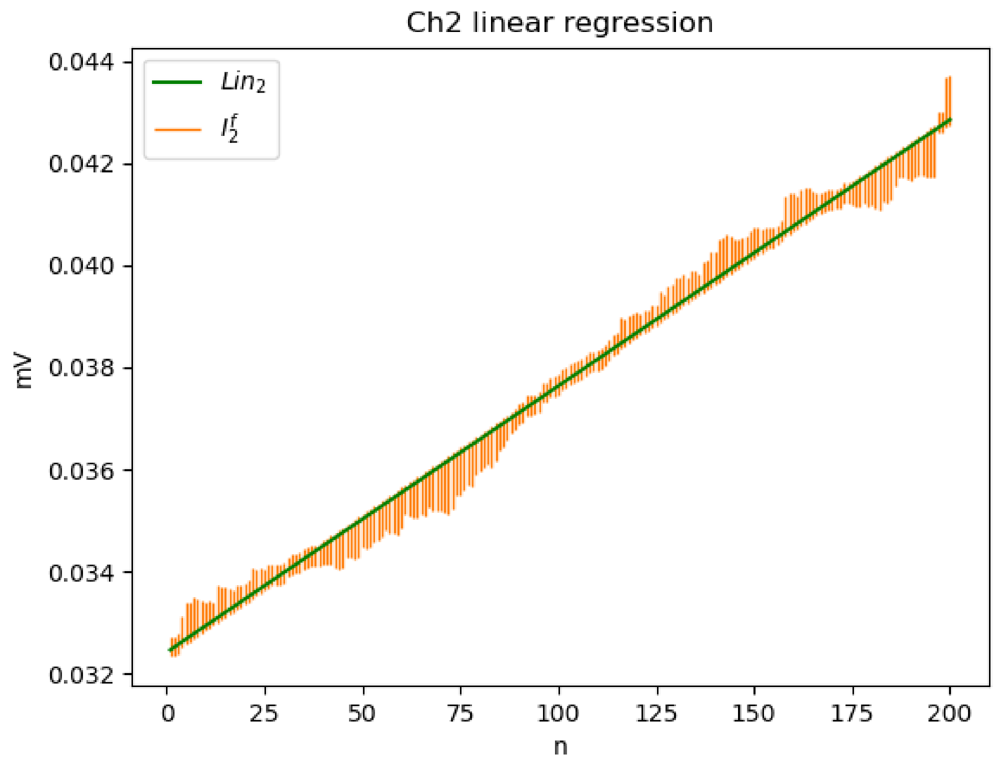

Рис. 7. I_2^f и Lin_2

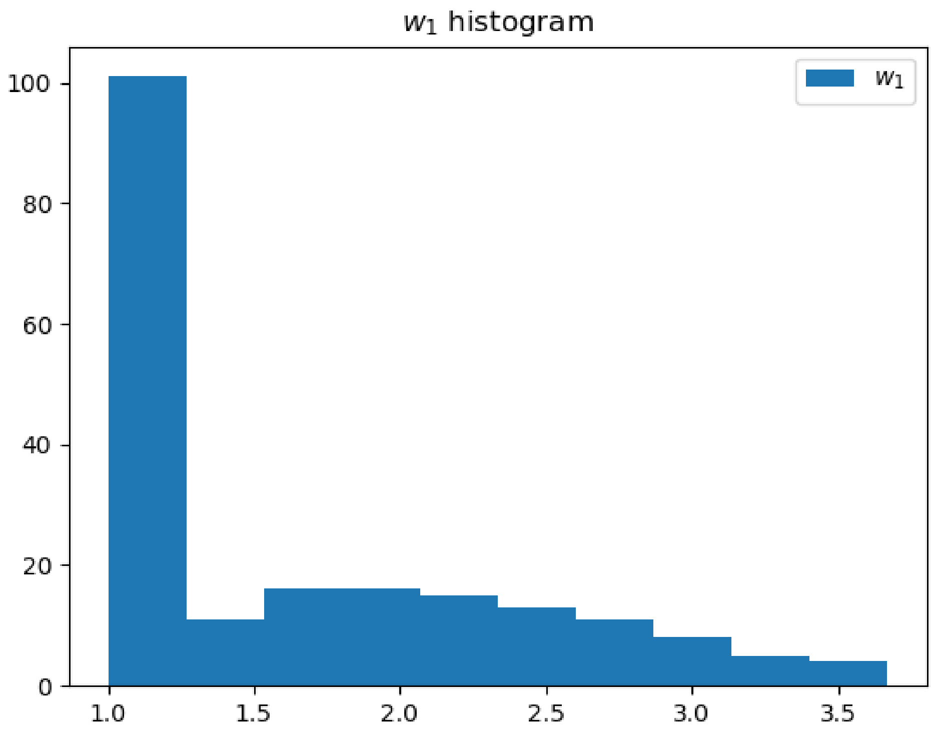

Рис. 8. Гистограмма значений w_2

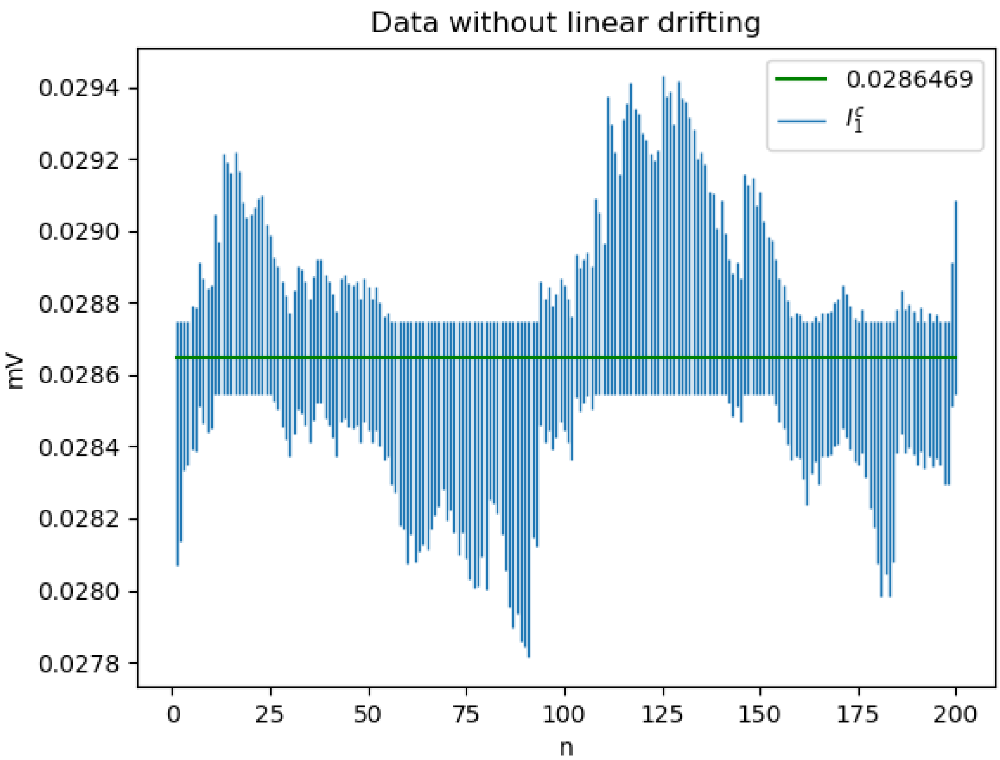

Рис. 9. I_1^c

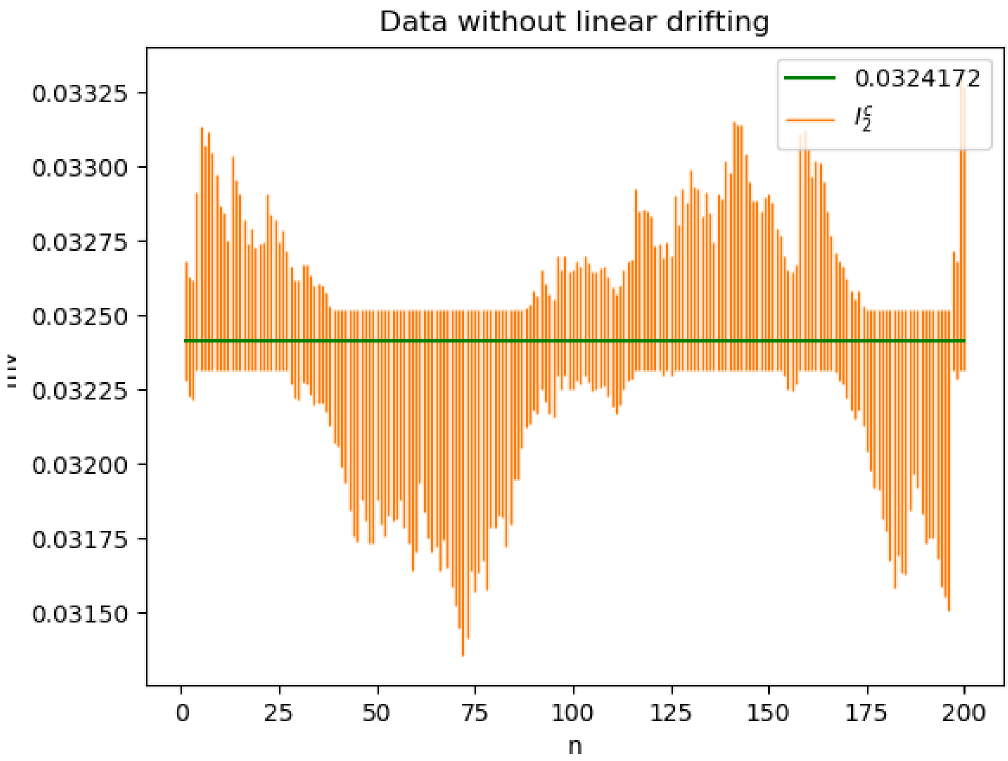

Рис. 10. I_2^c

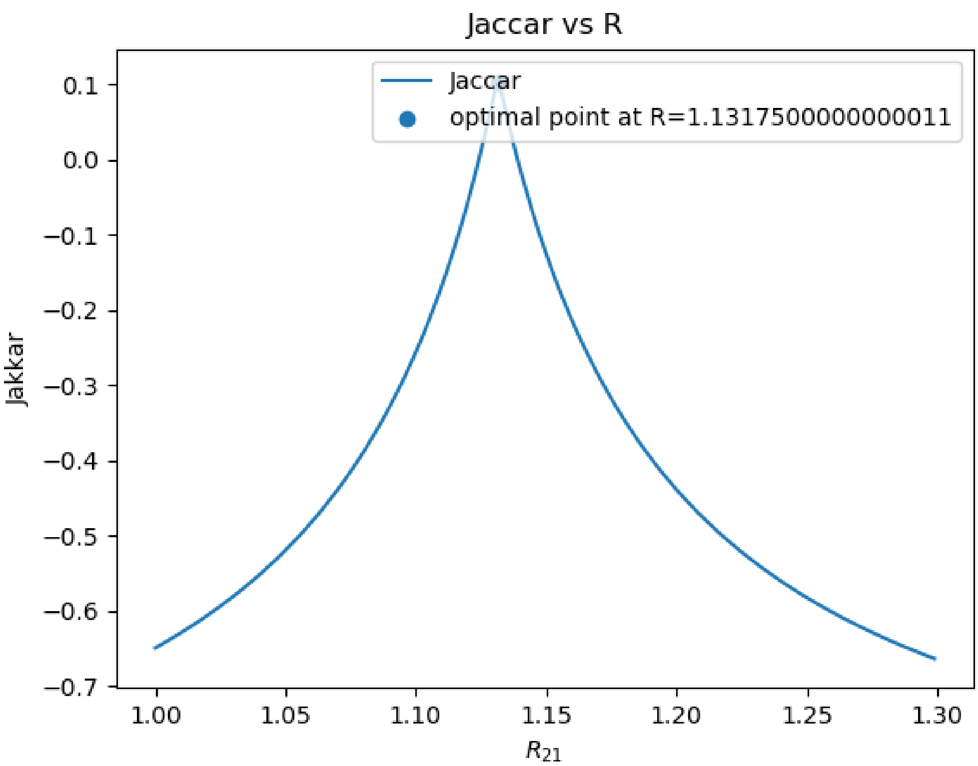

Рис. 11. Значение коэффициента жаккара от калибровочного множителя

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. Коэффициент Жаккара

Рассмотрим Рис.11. 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.

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