Санкт-Петербургский политехнический университет имени Петра Великого

Институт прикладной математики и механики

Высшая школа прикладной математики и физики

# Анализ данных с интервальной неопределённостью

## Отчет по курсовой работе

Выполнил:
Студент гр. 5040102/10201
Митенев А.В.

Принял:
к. ф.-м. н., доцент
Баженов А.Н.

Санкт-Петербург

2022

# Оглавление

Список иллюстраций ..... 3
Список таблиц ..... 3
Постановка задачи ..... 4
Теория ..... 4
Линейная регрессия ..... 4
МНК ..... 5
Препроцессинг ..... 5
Коэффициент Жаккара ..... 5
Оптимизация ..... 6
Статус измерений ..... 6
Классы наблюдений ..... 6
Взаимные отношения интервалов анализируемого наблюдения и прогнозного интервала рассматриваемой модели. ..... 6
Набор неравенств для классификации измерений. ..... 6
Мода интервальной выборки ..... 7
Оценка влияния радиуса ..... 7
Реализация ..... 7
Результаты ..... 7
Обработка результатов ..... 20
Обсуждение ..... 22
Литература ..... 22

# Список иллюстраций

Figure 1. Исходные данные ..... 8
Figure 2. Обинтерваленные данные ..... 8
Figure 3. Обинтерваленные данные с МНК ..... 9
Figure 4. Обинтерваленные данные с добавкой ..... 9
Figure 5. Информационное множество внутренней части для снятого тока ..... 10
Figure 6. Диаграмма рассеяния и коридор совместности внутренней части для снятого тока ..... 10
Figure 7. Диаграмма рассеяния и коридор совместности внутри и вне внутренней части для снятого тока ..... 11
Figure 8. Информационное множество внутренней части для эталона ..... 11
Figure 9. Диаграмма рассеяния и коридор совместности внутренней части для эталона ..... 12
Figure 10. Диаграмма рассеяния и коридор совместности внутри и вне внутренней части для снятого тока ..... 12
Figure 11. Спрямленные данные ..... 13
Figure 12. Информационное множество внутренней части для снятого тока после устранения дрейфовой составляющей ..... 13
Figure 13. Диаграмма рассеяния и коридор совместимости внутренней части для снятого тока после устранения дрейфовой составляющей ..... 14
Figure 14. Диаграмма рассеяния и коридор совместимости внутри и вне внутренней части для эталона после устранения дрейфовой составляющей ..... 16
Figure 15. Диаграмма статусов наблюдений снятого тока ..... 16
Figure 16. Диаграмма статусов наблюдений эталона ..... 17
Figure 17. Диаграмма статусов для радиусов $10-4$ и $3*10-4$ ..... 18
Figure 18. Диаграмма статусов для радиусов $5*10-4$ и $7*10-4$ ..... 19
Figure 19. Диаграмма статусов для радиуса $9*10-4$ ..... 19
Figure 20. График ширины моды от ширины интервалов ..... 20
Figure 21. График к-та Жаккара от ширины интервалов ..... 21

# Список таблиц

Table 1. Численные характеристики при изменении радиуса интервалов ..... 20

# Постановка задачи

Калибровка датчика производится по эталону. Зависимость между квантовыми эффективностями и датчиков предполагается постоянной для каждой пары наборов измерений $QE_2 = \frac{I_2}{I_1} * QE_1$, где $QE_2, QE_1$ – эталонная эффективность эталонного и исследуемого датчика, $I_2, I_1$ – измеренные токи. Требуется определить коэффициент калибровки $R_{21} = \frac{I_2}{I_1}$ при помощи линейной регрессии на множестве интервальных данных и коэффициента Жаккара. Построить информационное множество параметров и совместимый коридор. Построить предсказание вне интервала имеющихся данных.

# Теория

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 $\varepsilon$.

$$x = \dot{x} + \varepsilon$$

Где $\varepsilon = [-\epsilon, \epsilon]$

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

## Линейная регрессия

Регрессионную модель описания данных называют простой линейной, если заданный набор данных аппроксимируется прямой с внесённой добавкой в виде некоторой нормально распределённой ошибки: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i \in \overline{1,n}$, где $\{x_n\}_{n\in N}$ – заданные значения, $\{y_n\}_{n\in N}$ – параметры отклика, $\{\varepsilon_n\}_{n\in N}$ – независимые, центрированные, нормально распределjнные случайные величины с неизвестной дисперсией $\delta$, суть предполагаемые погрешности, $\beta_0, \beta_1$ – параметры, подлежащие оцениванию. В данной модели мы считаем, что у заданных значений нет погрешности (пренебрегаем ей). Полагаем, что основная погрешность получается при измерении$\{y_n\}_{n\in N}$.

## МНК

Данный метод основан на минимизации $l^2$-нормы разности последовательностей полученных экспериментальных данных $\{y_n\}$ и значений аппроксимирующей функции $f(\{x_n\})$.

$$(f(\{x_n\}) - \{y_n\})^2 \to min$$

В данном случае мы ставим задачу линейного программирования таким образом, чтобы найти не только коэффициенты $\beta_0$ и $\beta_1$, но и вектор $w$ на который стоит домножить погрешности наших интервальных данных. Тогда задача ставится так: $\sum|w_i| \to min, i \in \overline{1,n}$

При ограничениях $\beta_0 + \beta_1 * x_i - w_i * \varepsilon \leq y_i, i \in \overline{1,n}$

## Препроцессинг

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

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

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

$$I_i^f(n) = \dot{x}(n) + \max_i(\varepsilon * w_i(n)), \quad n \in \overline{1,N}$$

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

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

## 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:

$$JK(x) = \frac{wid(x \wedge y)}{wid(x \vee y)}$$

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]

### Оптимизация

Для поиска оптимального параметра калибровки поставим следующую задачу максимизации:

$$JK\left(x_{all}(R)\right) \to max$$

Где $JK$ — это коэффциент Жаккара, $x_{all}$ — это выборка полученная как

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

Где $\frown$ обозначена операция конкатенации двух выборок. Поиск будем проводить методом дихотомии, а поиск оптимального $R$ будем проводить в отрезке [1, 1.5]. Тогда оптимальное $R$ это и будет $R_{21}$.

## Статус измерений

### Классы наблюдений

- Внутренние – полностью содержат прогнозируемый интервал
- Граничные – прогнозируемый интервал внутри, но выходит на границу
- Внешние – прогнозируемый интервал выходит за границы
- Выбросы – не пересекаются с прогнозируемым интервалом

### Взаимные отношения интервалов анализируемого наблюдения и прогнозного интервала рассматриваемой модели.

Взаимные отношения интервалов анализируемого наблюдения $(x, y)$ и прогнозного интервала рассматриваемой модели $Y(x)$ удобно характеризовать в специальных терминах. Введём понятия размаха (плечо, англ. — high leverage)

$$l(x, y) = \frac{rad\ Y(x)}{rad\ y}$$

и относительного остатка (относительное остаточное отклонение, относительное смещение, англ. — relative residual)

$$r(x, y) = \frac{mid\ y - mid\ Y(x)}{rad\ y}$$

### Набор неравенств для классификации измерений.

Размах и остаток позволяют установить статус наблюдения, проверив некоторые простые неравенства.

Так для внутренних наблюдений, содержащих в себе прогнозный интервал модели, выполняется нестрогое неравенство

$$|r(x,y)| \leq 1 - l(x,y)$$

а точное равенство в нём является характеристическим условием для граничных наблюдений.

Выбросы — наблюдения, не пересекающиеся с коридором совместных зависимостей, а потому они удовлетворяют неравенству

$$|r(x,y)| > 1 + l(x,y)$$

Интервальные измерения, у которых величина неопределённости меньше, чем ширина прогнозного интервала, то есть

$$l(x,y) > 1$$

могут оказывать очень сильное влияние на модель и потому называются строго внешними.

### Мода интервальной выборки

Мода интервальной выборки - интервал пересечения ее наибольшей совместной подвыборки.

Для вычисления моды интервальной выборки использовалась следующая идея. Для каждой верхней границы интервала вычислялось количество интервалов, которые требуется удалить, чтобы эта верхняя граница стала верхней границей моды. То есть количество объединения интервалов, у которых верхняя граница меньше, чем рассматриваемая и интервалов, у которых нижняя граница больше или равна рассматриваемому значению. Ширина интервала в таком случае будет равна разности между рассматриваемым значением и максимальным значением нижних границ не удаленных интервалов. Шириной моды в таком случае будет являться ширина объединения интервалов, для  которых требуется минимальное количество удалений других интервалов.

### Оценка влияния радиуса

Для того, чтобы оценить влияние радиуса был взят ряд значений и для каждого выполнены описанные шаги алгоритма для классификации наблюдений.

## Реализация

Работа реализована на языке программирования Python в среде Jupyetr Notebook с использованием библиотек numpy, pandas, scipy, matplotlib, seaborn.

Реализация работы: https://github.com/mitenevav/intervals/blob/main/lab3/lab3.ipynb

## Результаты

Результаты получены на данных Канал 1_800nm_0.2.csv и Канал 2_800nm_0.2.csv

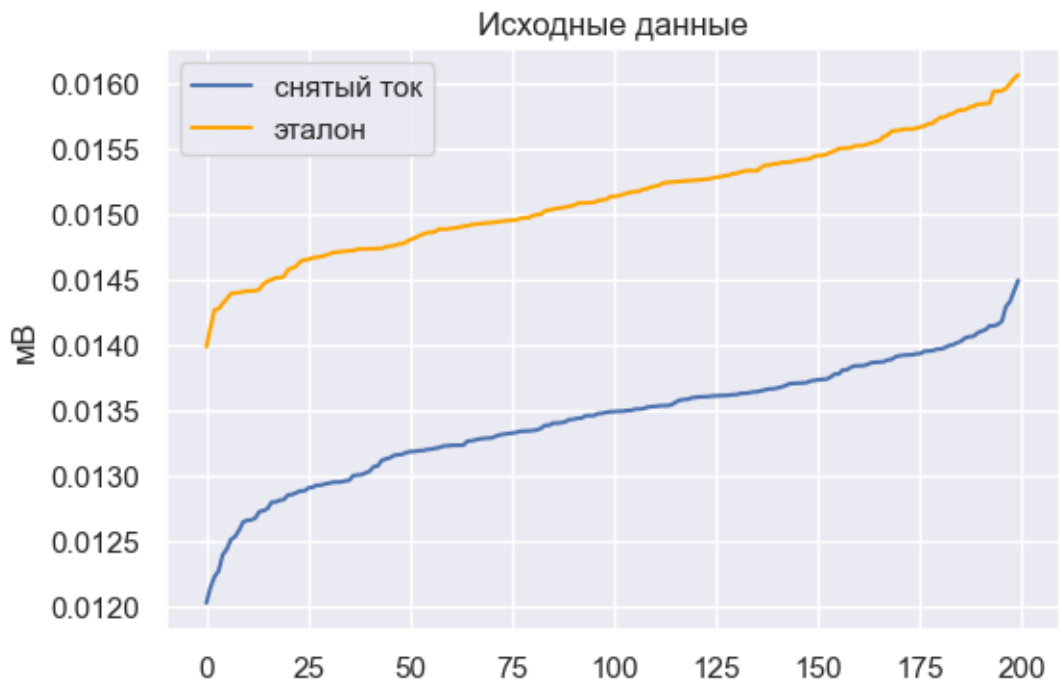


*Figure 1. Исходные данные*

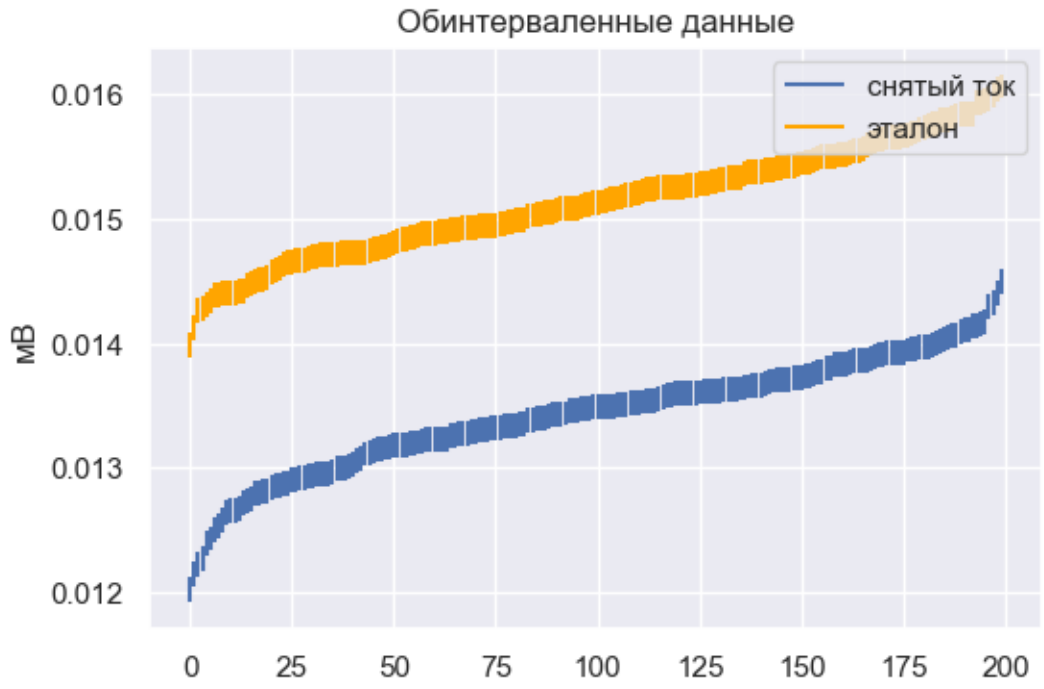


*Figure 2. Обинтерваленные данные*

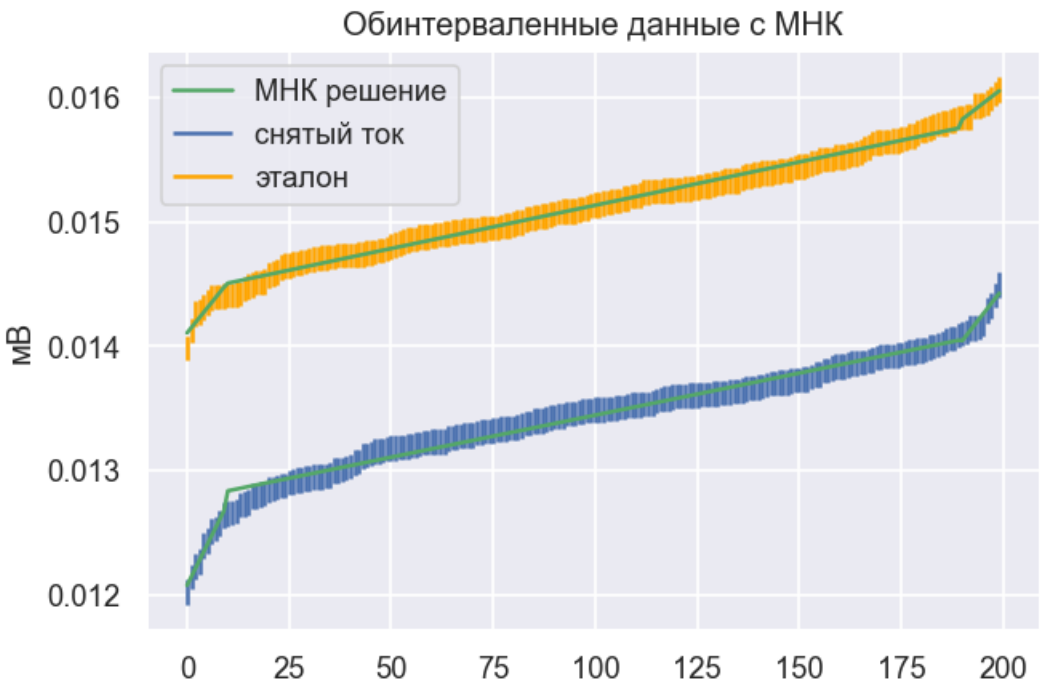


*Figure 3. Обинтерваленные данные с МНК*

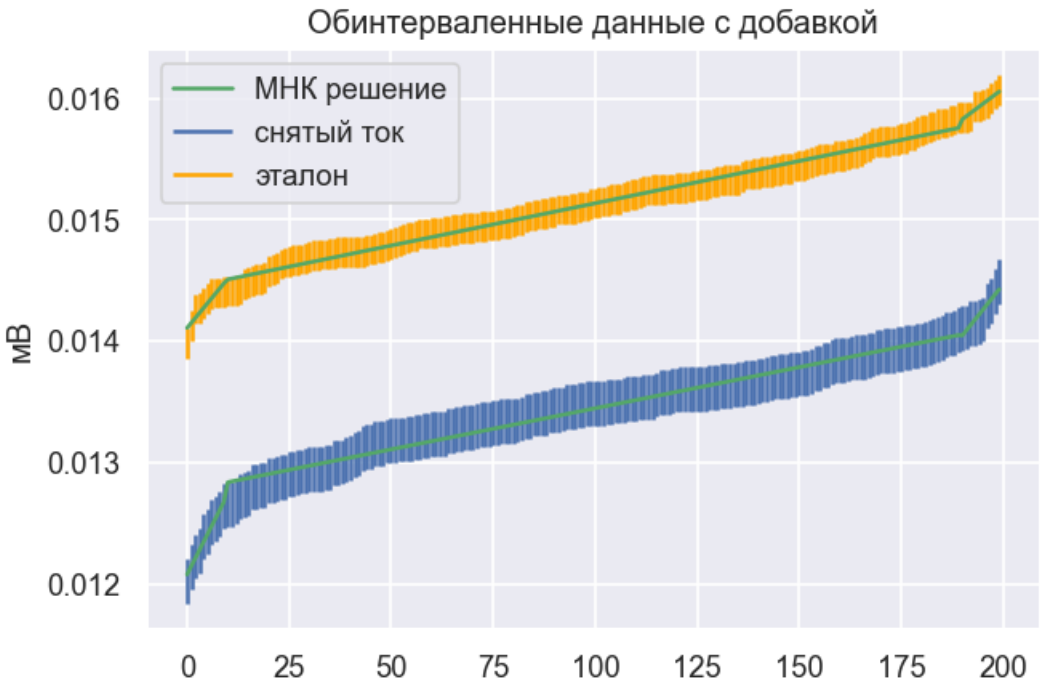


*Figure 4. Обинтерваленные данные с добавкой*

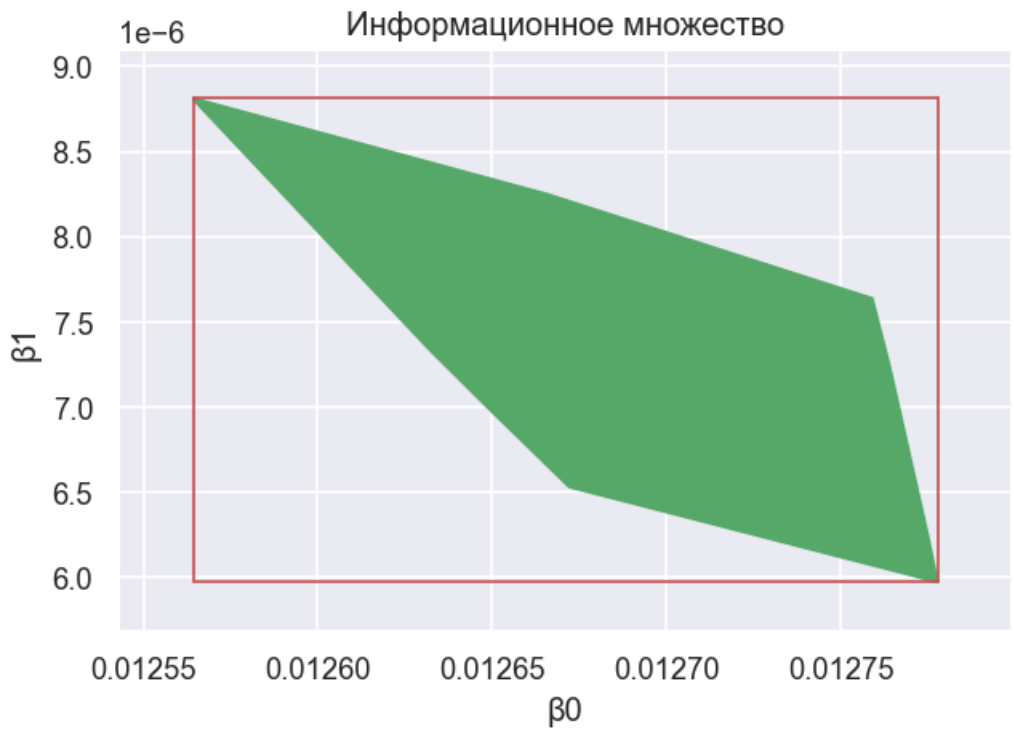


*Figure 5. Информационное множество внутренней части для снятого тока*

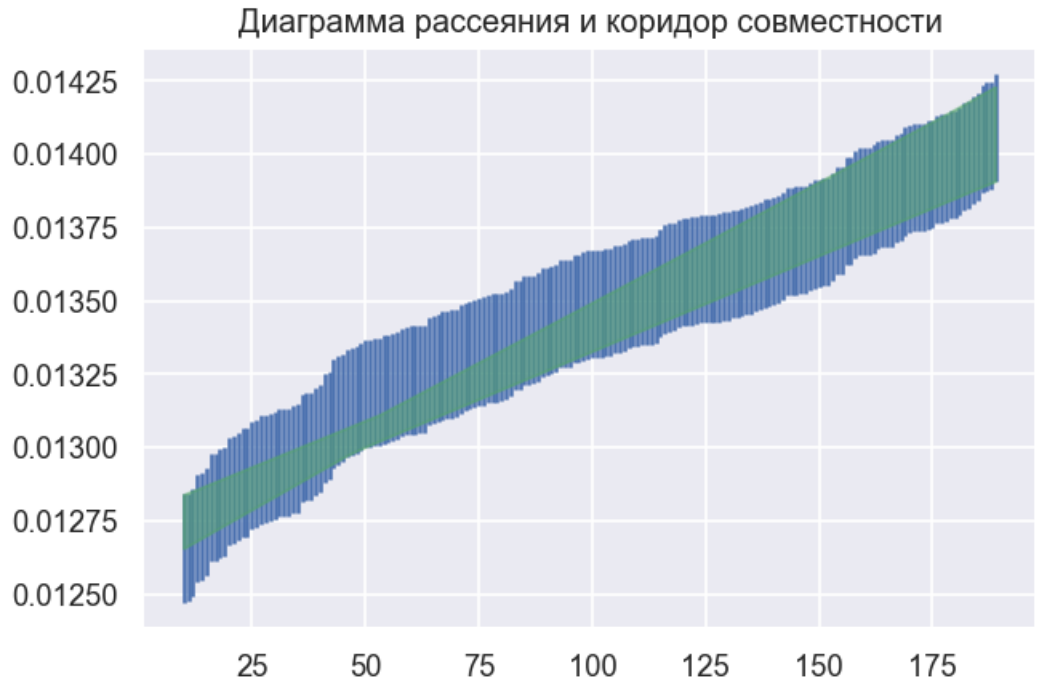


*Figure 6. Диаграмма рассеяния и коридор совместности внутренней части для снятого тока*

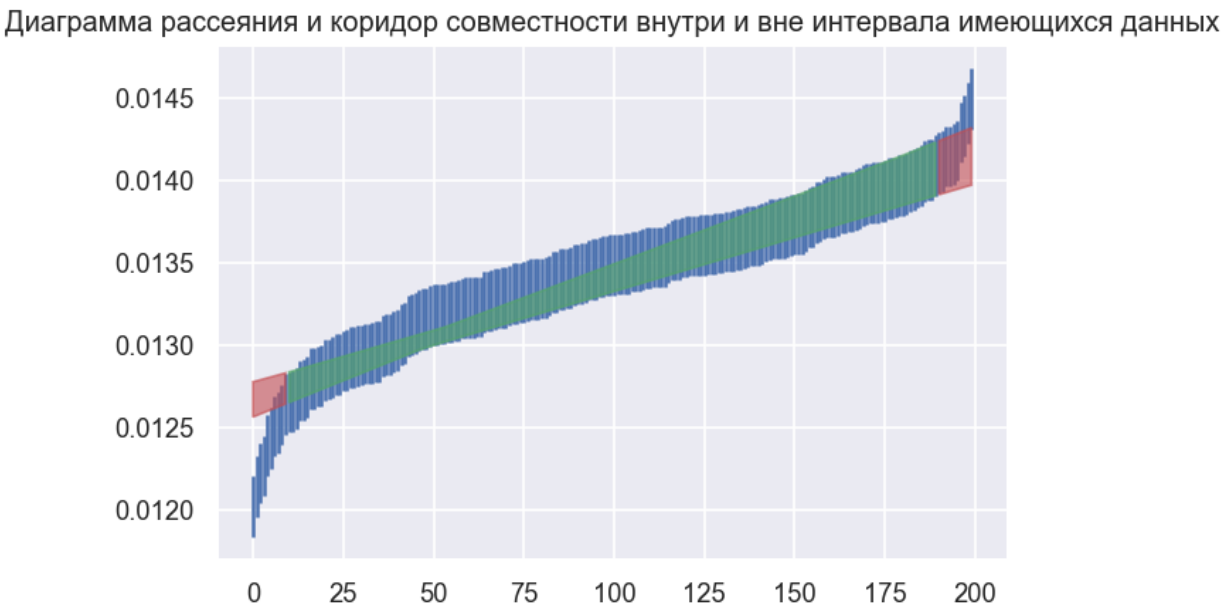


*Figure 7. Диаграмма рассеяния и коридор совместности внутри и вне внутренней части для снятого тока*

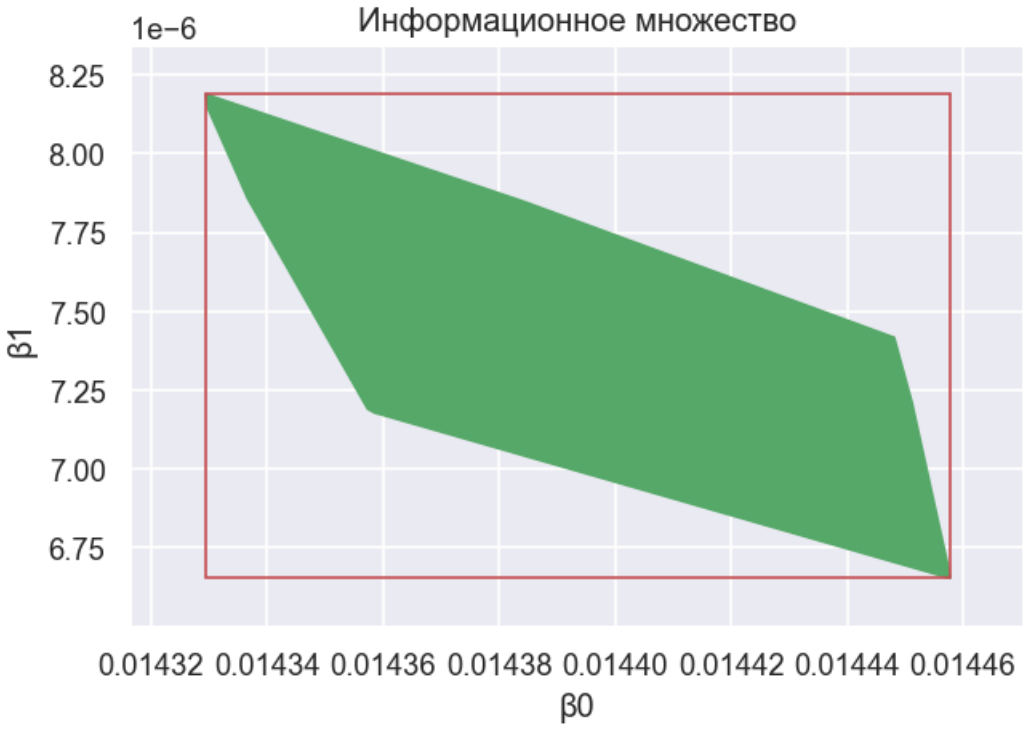


*Figure 8. Информационное множество внутренней части для эталона*

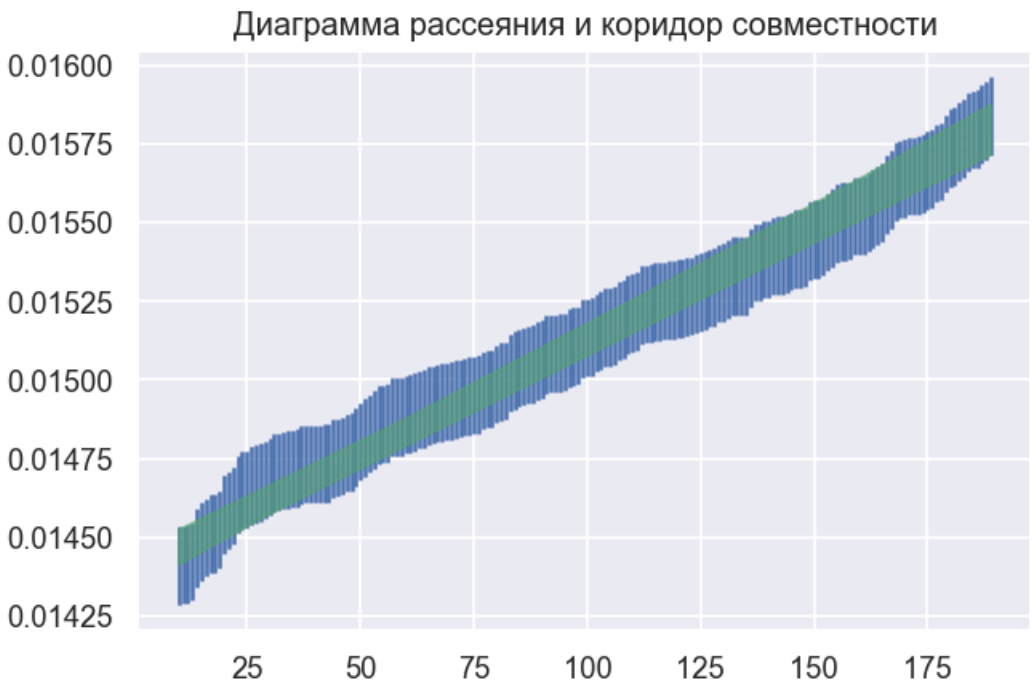


*Figure 9. Диаграмма рассеяния и коридор совместности внутренней части для эталона*

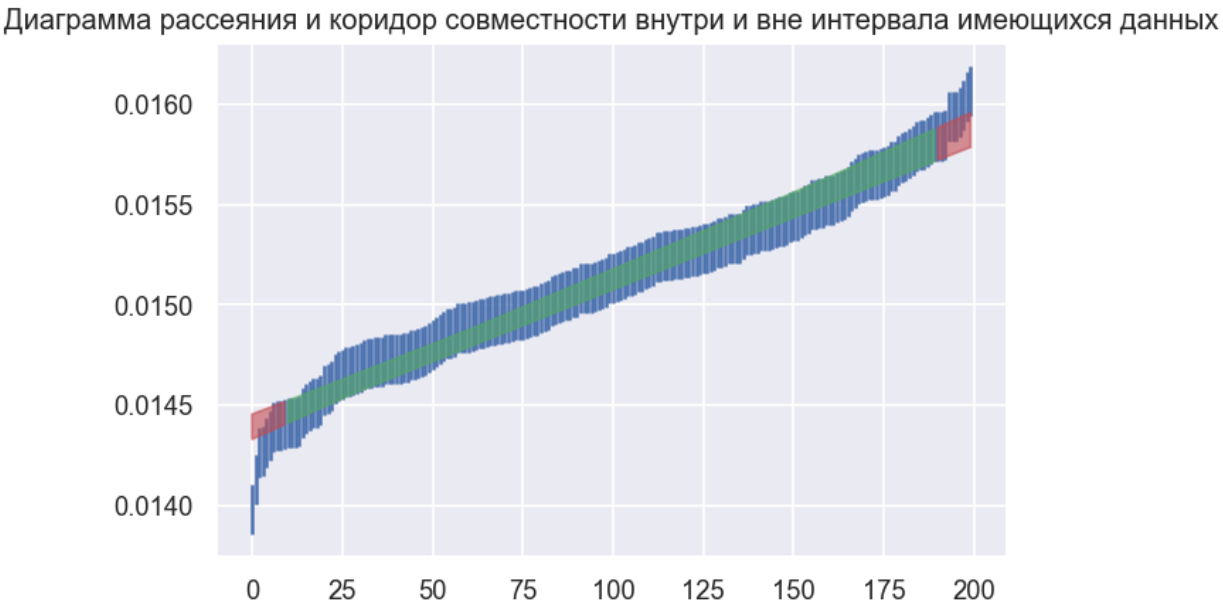


*Figure 10. Диаграмма рассеяния и коридор совместности внутри и вне внутренней части для снятого тока*

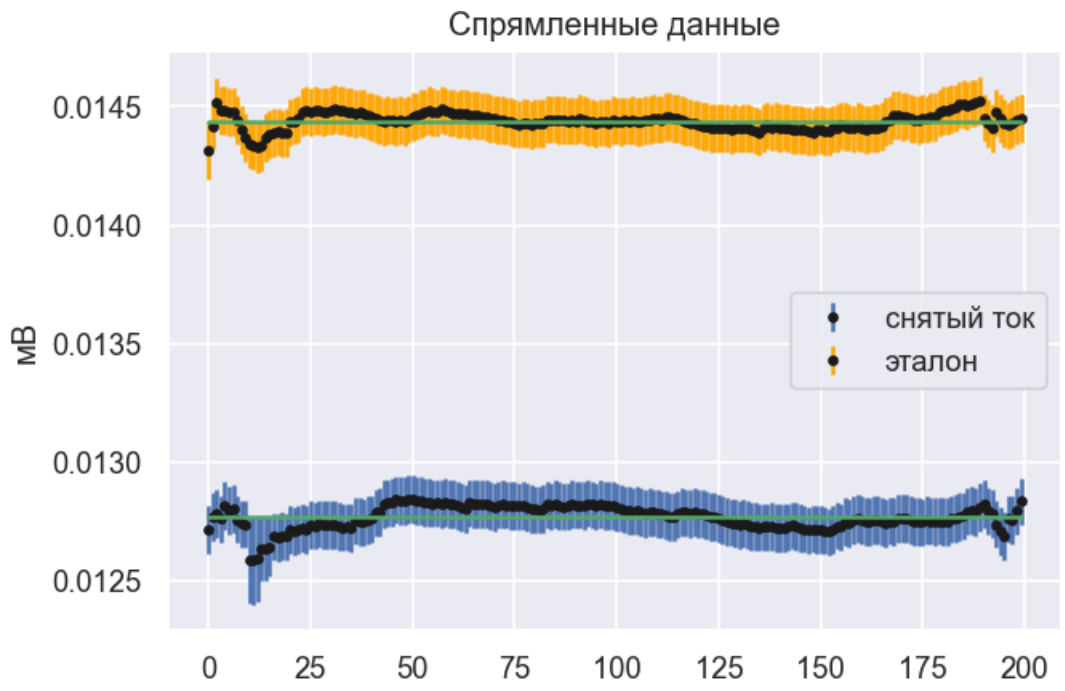


*Figure 11. Спрямленные данные*

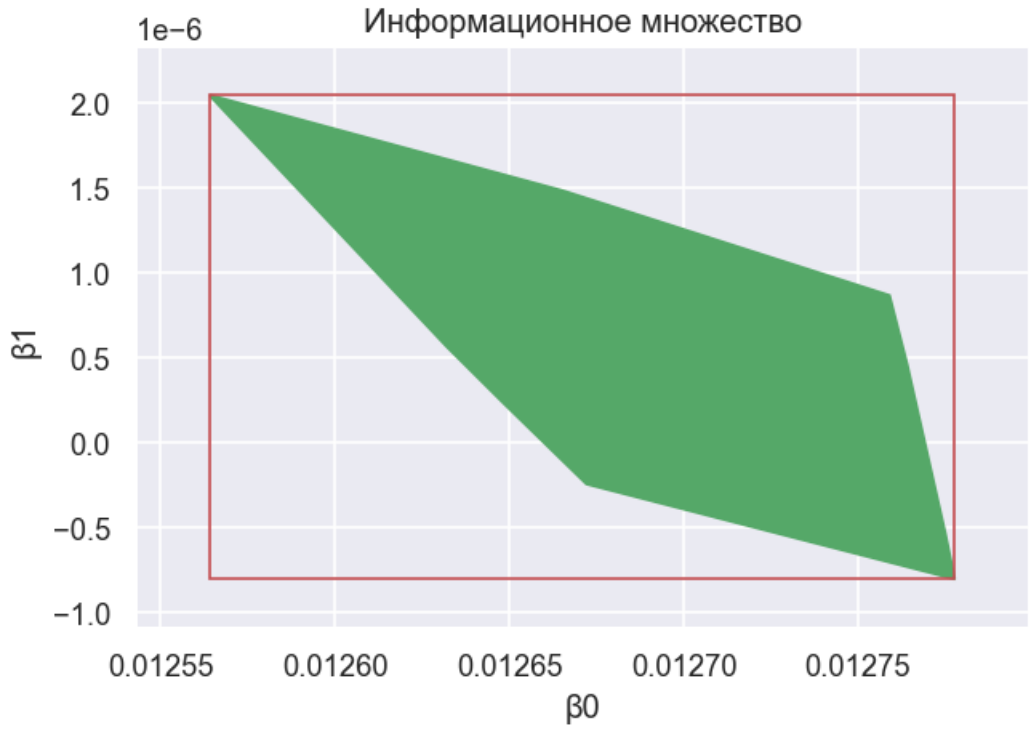


*Figure 12. Информационное множество внутренней части для снятого тока после устранения дрейфовой составляющей*

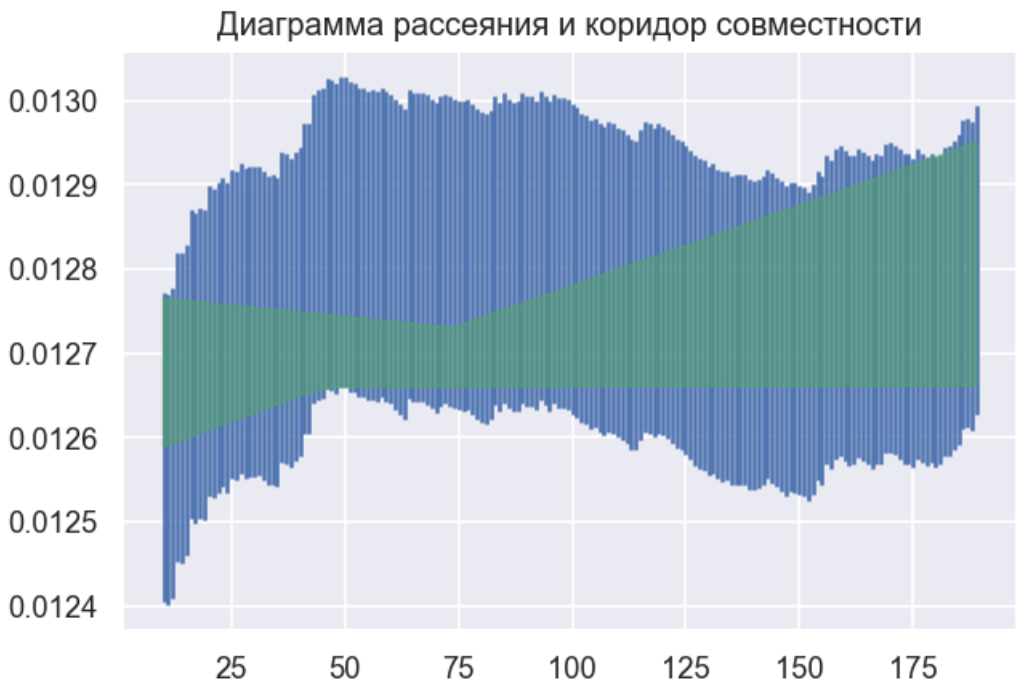


*Figure 13. Диаграмма рассеяния и коридор совместимости внутренней части для снятого тока после устранения дрейфовой составляющей*

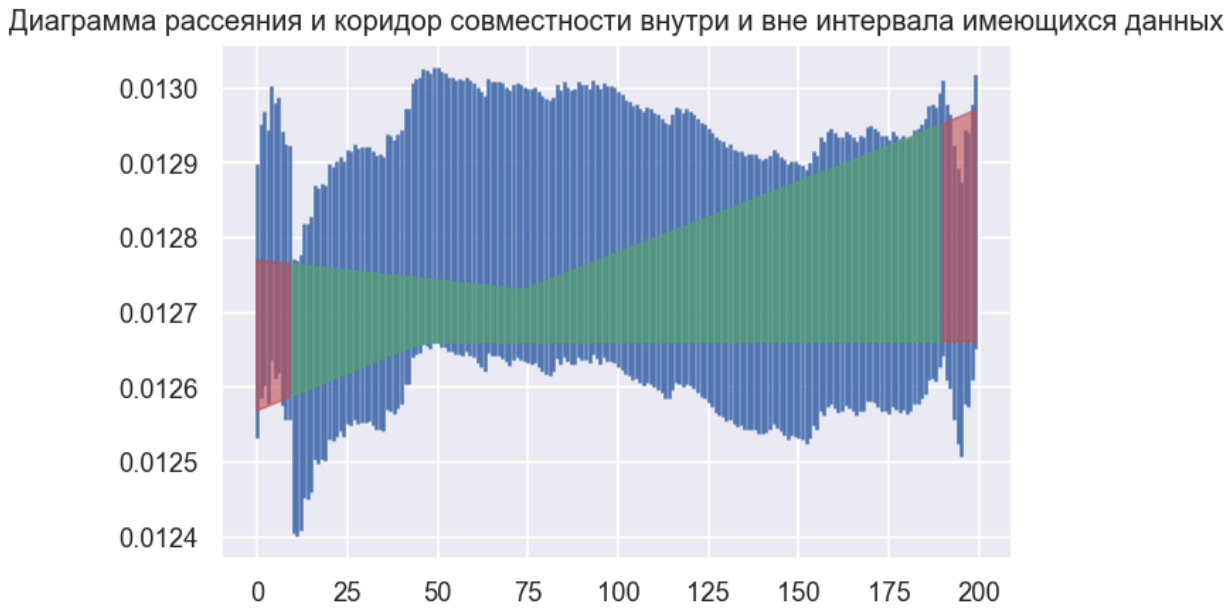


*Figure 14. Диаграмма рассеяния и коридор совместимости внутри и вне внутренней части для снятого тока после устранения дрейфовой составляющей*

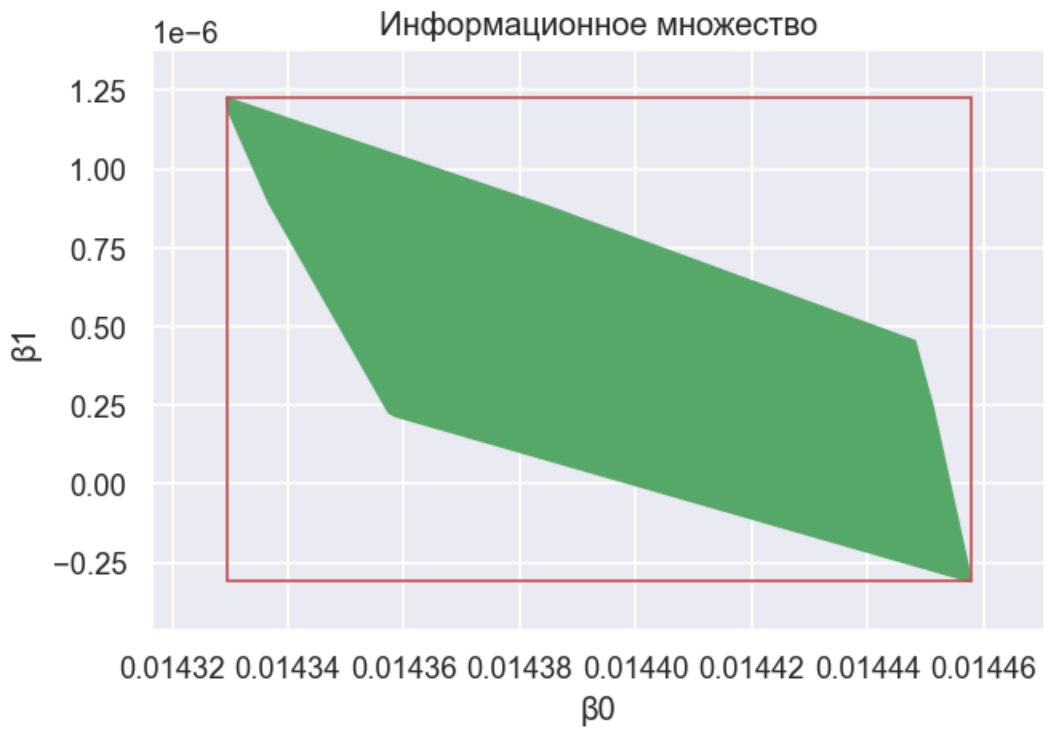


*Figure 15. Информационное множество параметров внутренней части для эталона после устранения дрейфовой составляющей*

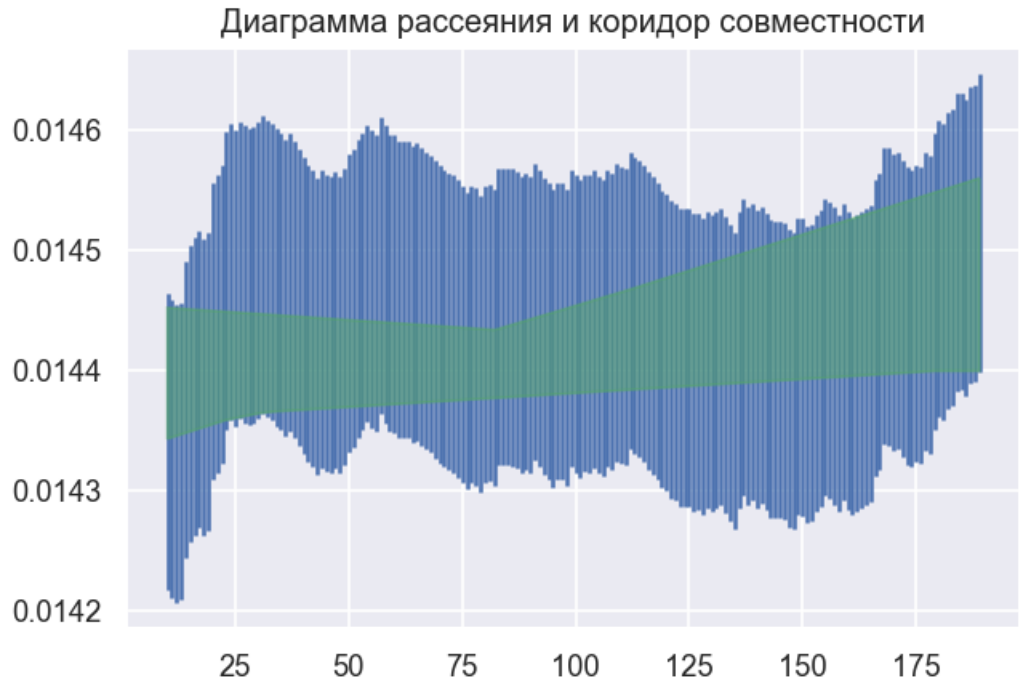


*Figure 16. Диаграмма рассеяния и коридор совместимости внутренней части для эталона после устранения дрейфовой составляющей*

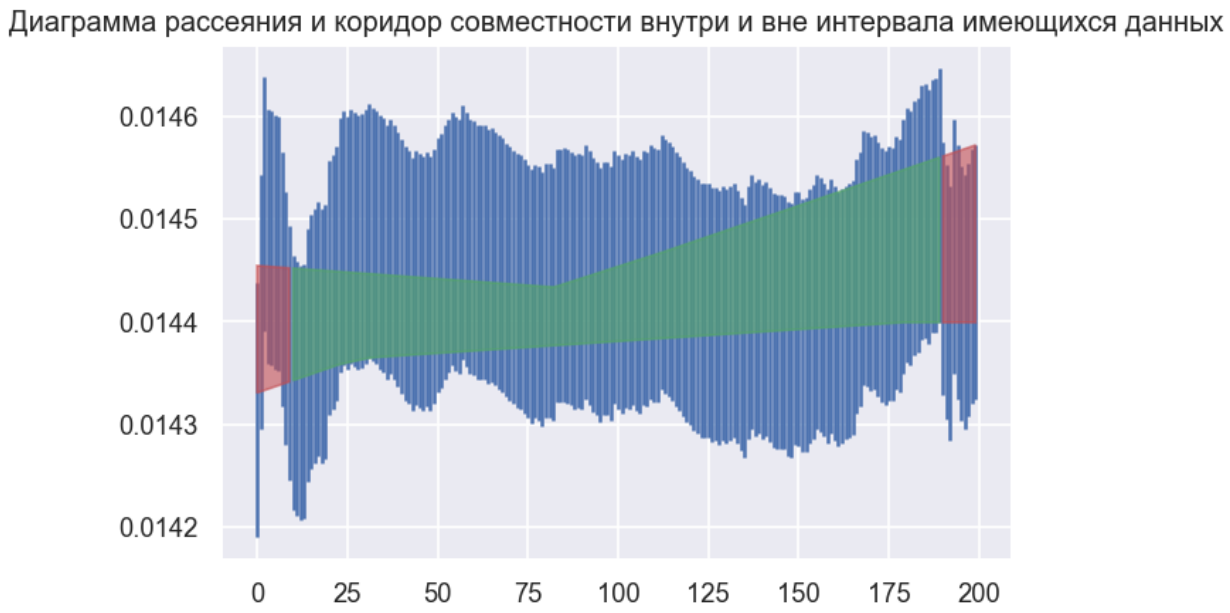


*Figure 14. Диаграмма рассеяния и коридор совместимости внутри и вне внутренней части для эталона после устранения дрейфовой составляющей*

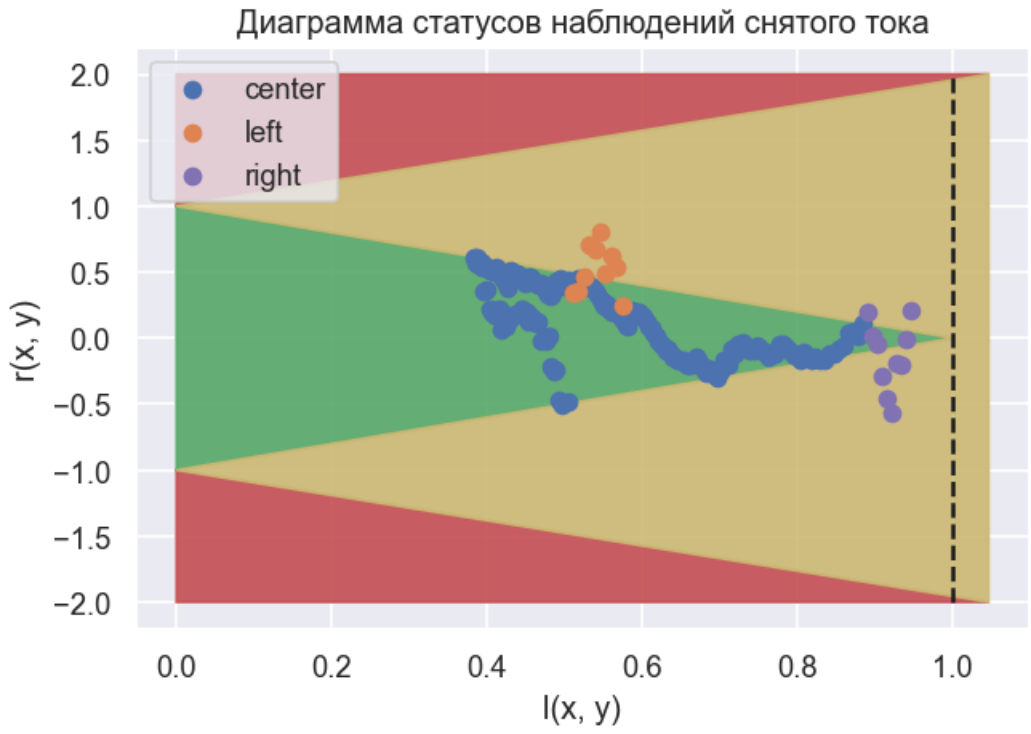


*Figure 15. Диаграмма статусов наблюдений снятого тока*

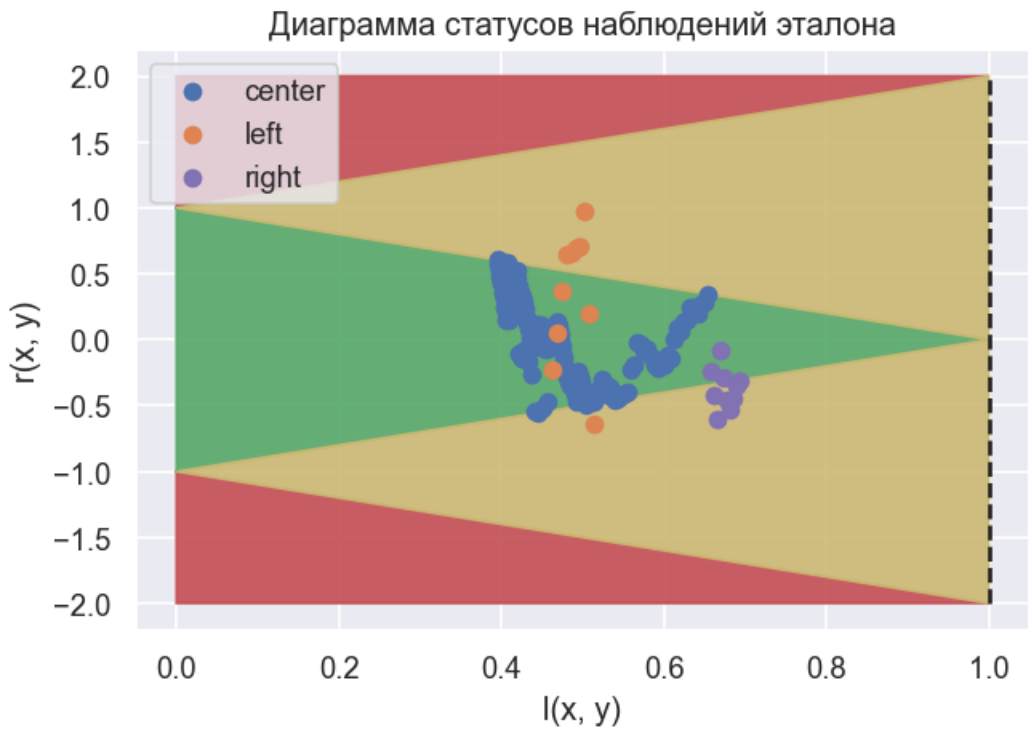


*Figure 16. Диаграмма статусов наблюдений эталона*

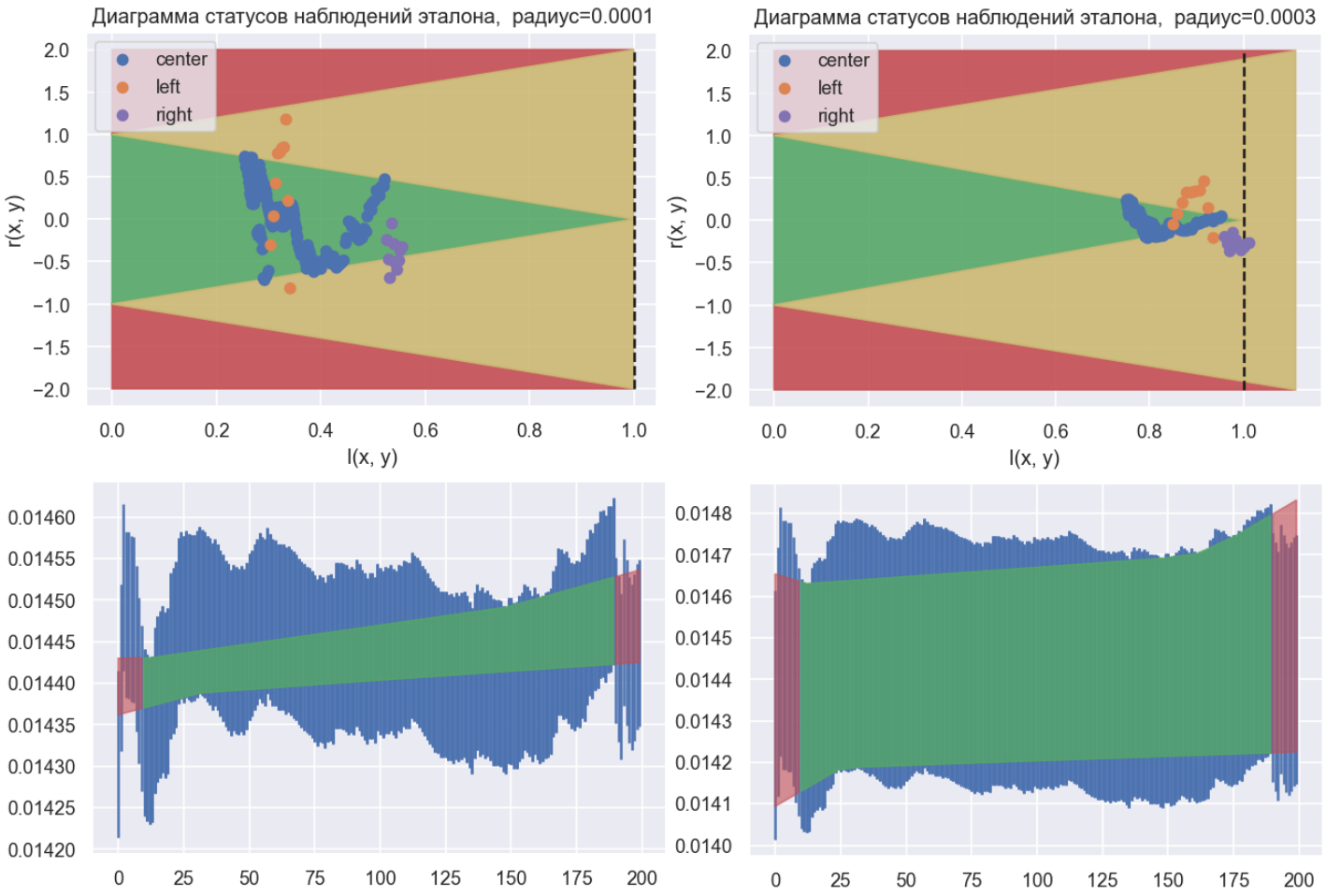


*Figure 17. Диаграмма статусов для радиусов* $10^{-4}$ *и* $3 * 10^{-4}$

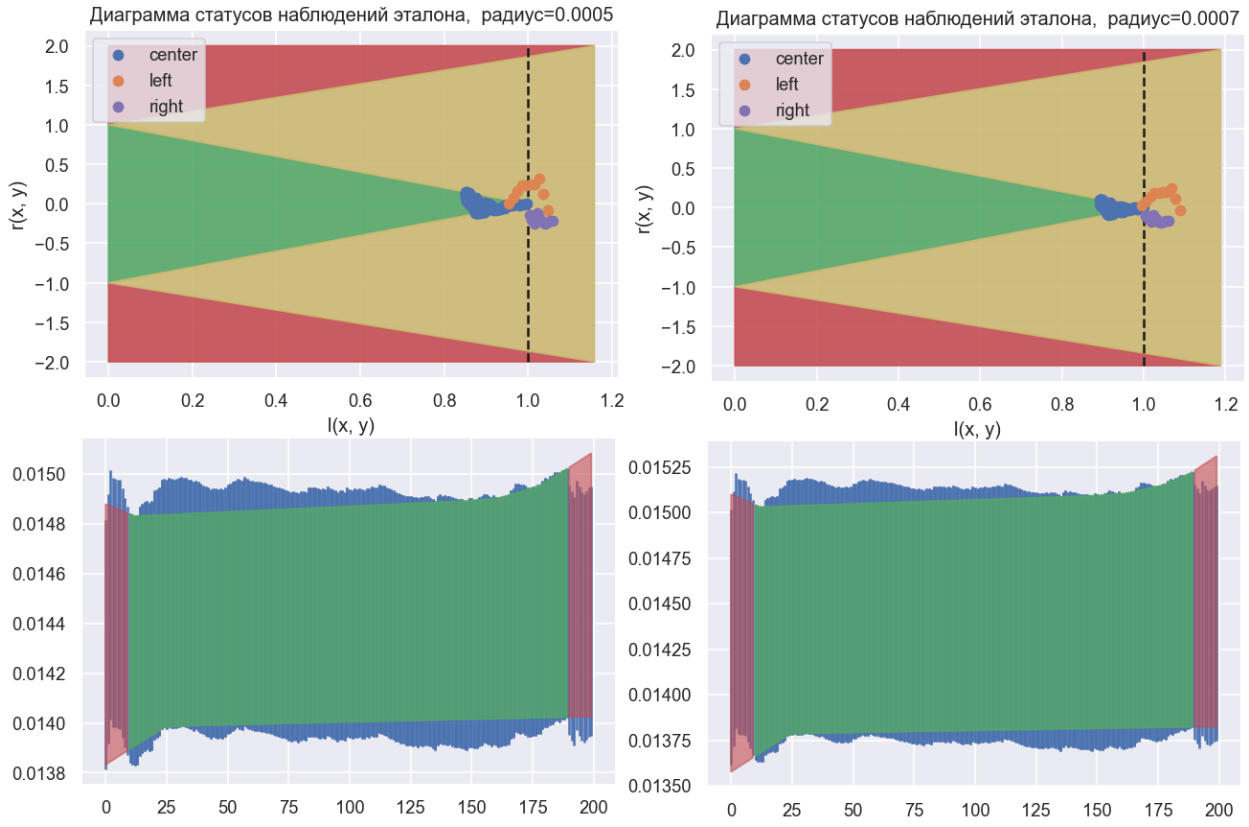


*Figure 18. Диаграмма статусов для радиусов* $5 * 10^{-4}$ *и* $7 * 10^{-4}$

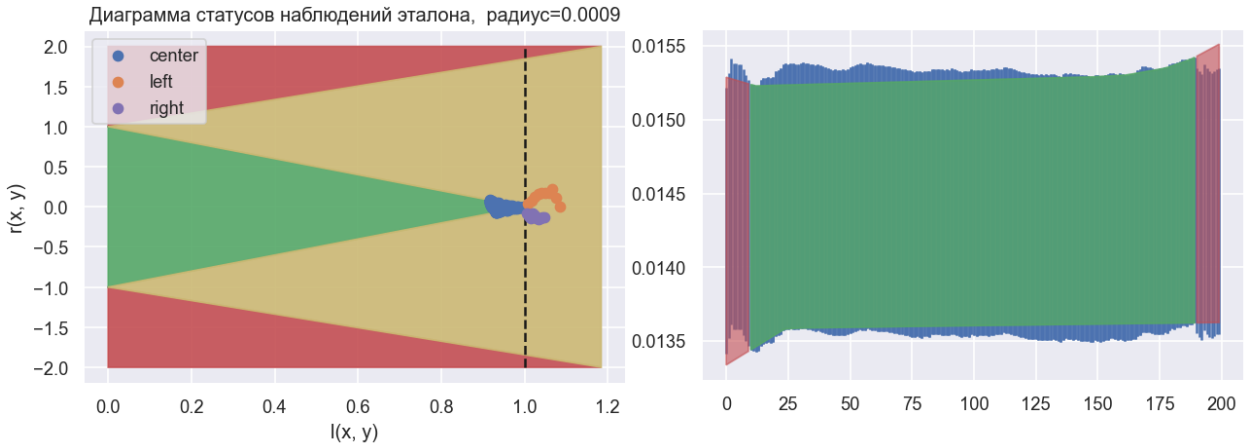


*Figure 19. Диаграмма статусов для радиуса* $9 * 10^{-4}$

## Обработка результатов

| Радиус | Мера Жаккара | Мода | Количество наблюдений в зеленой зоне |
|---|---|---|---|
| 1e-4 | -0.0207 | 0.000007 | 189 |
| 3e-4 | 0.4843 | 0.000392 | 177 |
| 5e-4 | 0.6550 | 0.000792 | 172 |
| 7e-4 | 0.7408 | 0.001192 | 177 |
| 9e-4 | 0.7924 | 0.001592 | 171 |

*Table 1. Численные характеристики при изменении радиуса интервалов*


*Figure 20. График ширины моды от ширины интервалов*


*Figure 21. График к-та Жаккара от ширины интервалов*

## Обсуждение

Из рисунков 15-16 видно, что в зеленую зону попадают в основном наблюдения из центральной части выборки. Предсказания же по большей части находиться в жёлтой зоне.

Из рисунков 17-19 можно заметить, что с увеличением радиуса интервалов количество наблюдений в зеленой зоне может уменьшаться из-за неточностей на промежутках вне рассматриваемого интервала. При этом значение $l(x, y)$ стягивается к единице, так как интервалы становятся больше. А значение $r(x, y)$ стягивается к нулю, так как знаменатель растет.

## Литература

Диаграмма статусов измерений выборки интервальных данных: https://github.com/AlexanderBazhenov/Students/blob/master/DataStatus.pdf