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

Курсовая работа

по дисциплине «Анализ данных с интервальной неопределенностью»

на тему **«Обработка постоянной. Восстановление зависимостей. Статус данных измерений»**

Выполнил

студент гр. 5040102/10201 Гусаров Д.А. /_____/

Руководитель

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

Санкт-Петербург

2023

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 ε

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

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

Самый простой заключается в следующем. 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 ε_{1n} стоит растянуть, домножить на величину w_n, минимально возможную, для того, чтобы интервал коснулся линии регрессии.

Окончательно, интервальные данные представимы 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, $\tilde{\varepsilon}_n = w_n \cdot \varepsilon$, $w_n \geq 1$.

Другой способ состоит в построении интервальной регрессии. Пытаемся строить модель в классе линейных функций.

$$y_i = f(x_i, \beta_j) \quad (6)$$

Задача восстановления зависмостей заключается в том, чтобы имея набор значений переменных x_i и y_i, найти β_j, которые соответсвуют некоторой

функции f из параметрического семейства. Величины y_i являются интервальными и образуют систему, множество решений является информационным множеством. Иногда информационное множество задачи может оказаться пустым, это происходит при неправильно выбранном значении погрешности, которое не совпадает с реальным значением. В таком случае для уточнения погрешности решают задачу оптимизации:

$$\begin{cases} midy_i - w_i \cdot rady_i \le X\beta \le midy_i + w_i \cdot rady_i, i = \overline{1,n} \\ \sum_{i=1}^{n} w_i \to min \\ w_i \ge 0, i = \overline{1,n} \\ w, \beta = ? \end{cases} \quad (7)$$

где X – матрица $n \times 2$, в первом столбце которой единичные значения, во втором – значения x_i.

В наших обозначениях роль x_i играют значения n, y_i – значения x_{1i} или x_{2i}. Соответственно за $midy_i$ можем принять x_{1i} или x_{2i}, а $rady_i = \varepsilon$.

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 ε_{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, $\tilde{\varepsilon}_n = w_n \cdot \varepsilon$, $w_n \ge 1$.

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:

$$X_1' \leftarrow X_1 - A_1 \cdot n \tag{6}$$

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 Жаккакра, то есть ее модификация для интервальных данных:

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

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:

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

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}.

Определив параметры функциональной зависимости, мы можем предсказать значения в других точках области определения, хотя такое предсказание будет осуществляться с некоторой погрешностью, обусловленной неопределенностью самих данных, неоднозначностью процедуры восстановления зависимостей и другими факторами. Если информационное множество задачи восстановления зависимотей непусто, то обычно оно задает целое семейство зависимостей, совместных с данными задачи, которое имеет смысл рассматривать как единое целое в вопросах, касающихся оценивания неопределенности предсказания, учета возможных сценариев развития и так далее. Следовательно, возникает необходимость рассматривать как единое целое множество всех функций, совместных с интервальными данными задачи восстановления зависимостей. Такое множество называется коридором совместных зависмостей.

Граничными называются измерения, определяющие какой-либо фрагмент границы множества. Это свойство имеет смысл рассматривать для наблюдений, принадлежащих выборке, по которой строилась модель. Граничные измерения задают минимальную подвыборку, определяющую модель.

Для оценки качества моделе можно использовать функционал качества

$$T_w = \frac{\sum_{i=1}^{n} w_i}{n} \geq 1 \qquad (10)$$

Взаимные отношения интервалов анализируемого наблюдения (x, y) и прогнозного интервала рассматриваемой модели $\gamma(x)$ удобно характризовать в специальных терминах.

Введем понятие размаха (плечо):

$$l(x, y) = \frac{rad\ \gamma(x)}{rad\ y} \qquad (11)$$

и относительного остатка (относиткльное остаточное отклонение, относительное смещение):

$$r(x,y) = \frac{mid\ y - mid\ \gamma(x)}{rad\ y} \quad (12)$$

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

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

$$|r(x,y)| \leq 1 - l(x,y) \quad (13)$$

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

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

$$|r(x,y)| > 1 + l(x,y) \quad (14)$$

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

$$l(x,y) > 1 \quad (15)$$

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

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

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

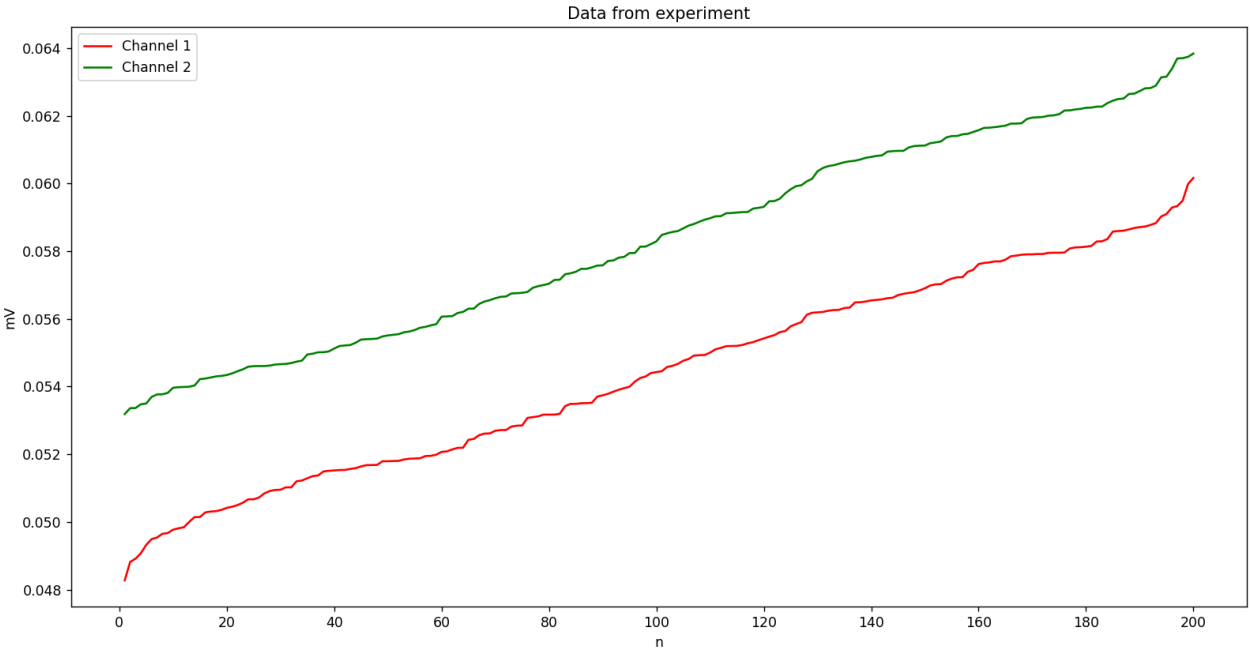

Рис. 1. Две выборки экспериментальных данных, измеренным датчиками

1. Измеренные данные:

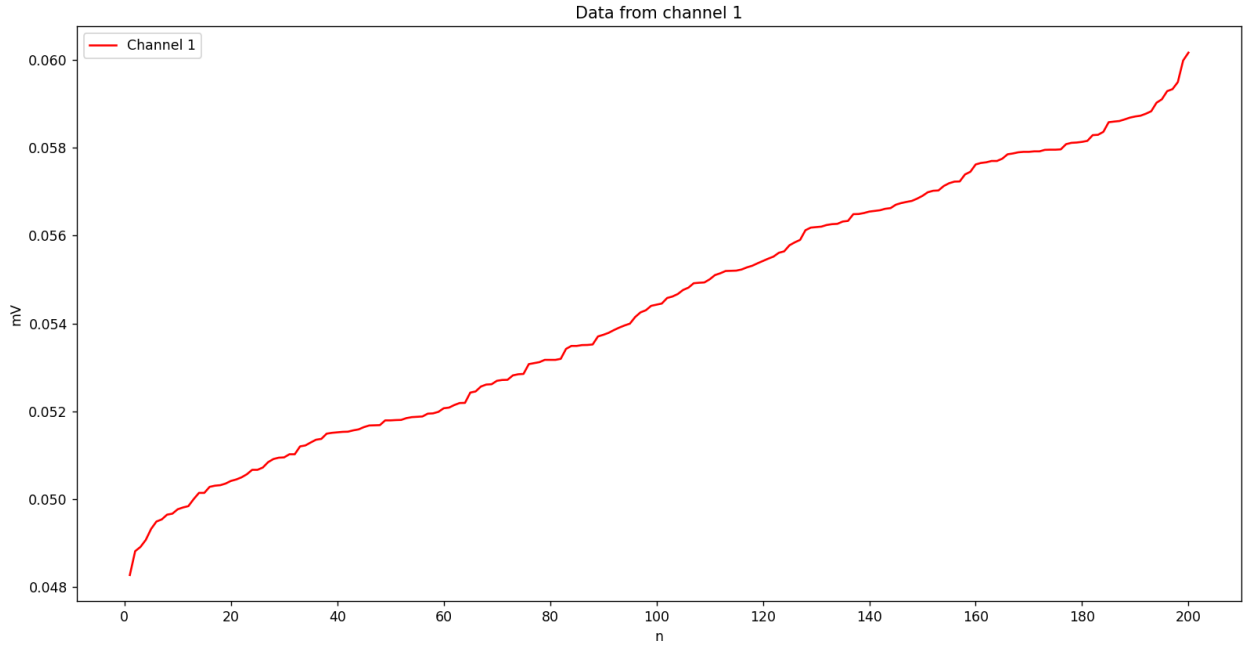

Рис. 2. Данные, измеренные датчиком ФП1

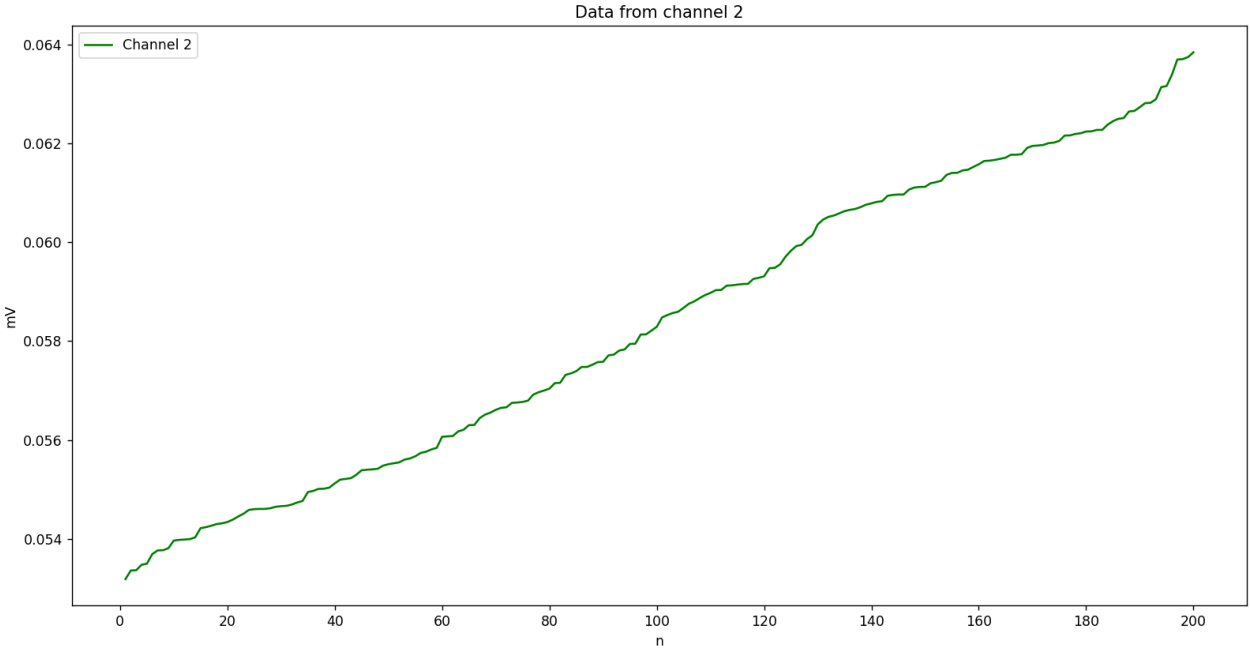

Рис. 3. Данные, измеренные датчиком ФП2

2. Интервальные данные:

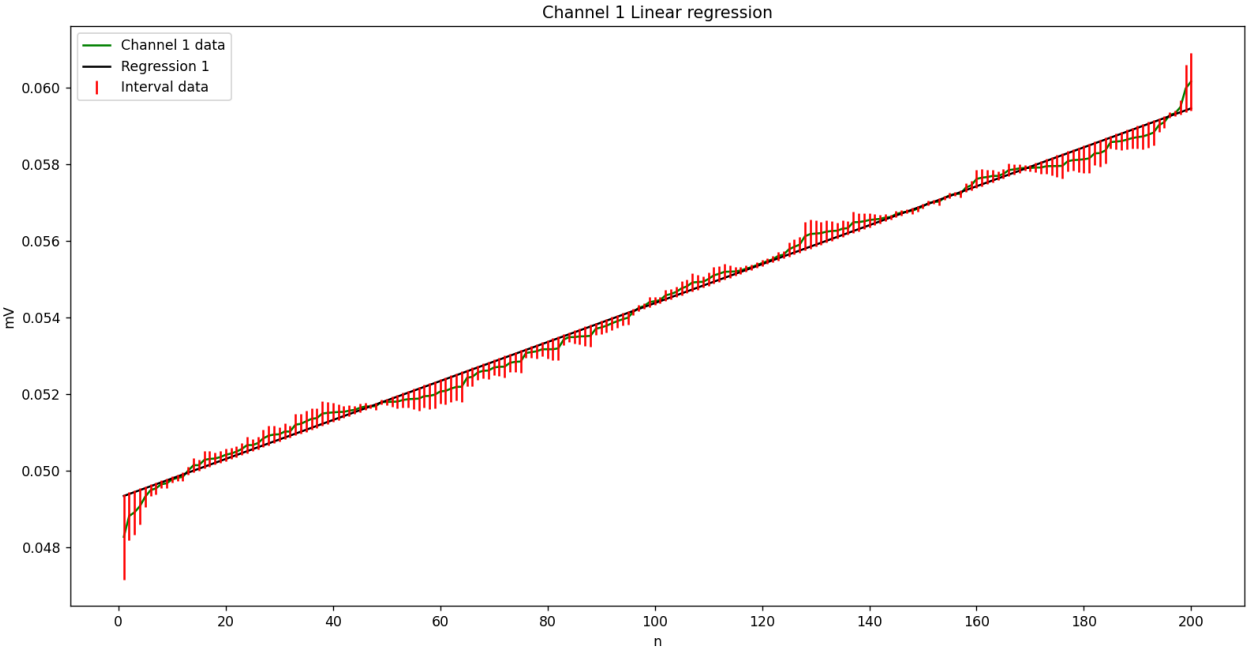

Рис. 4. Интервальные данные первой выборки и линейная регрессия

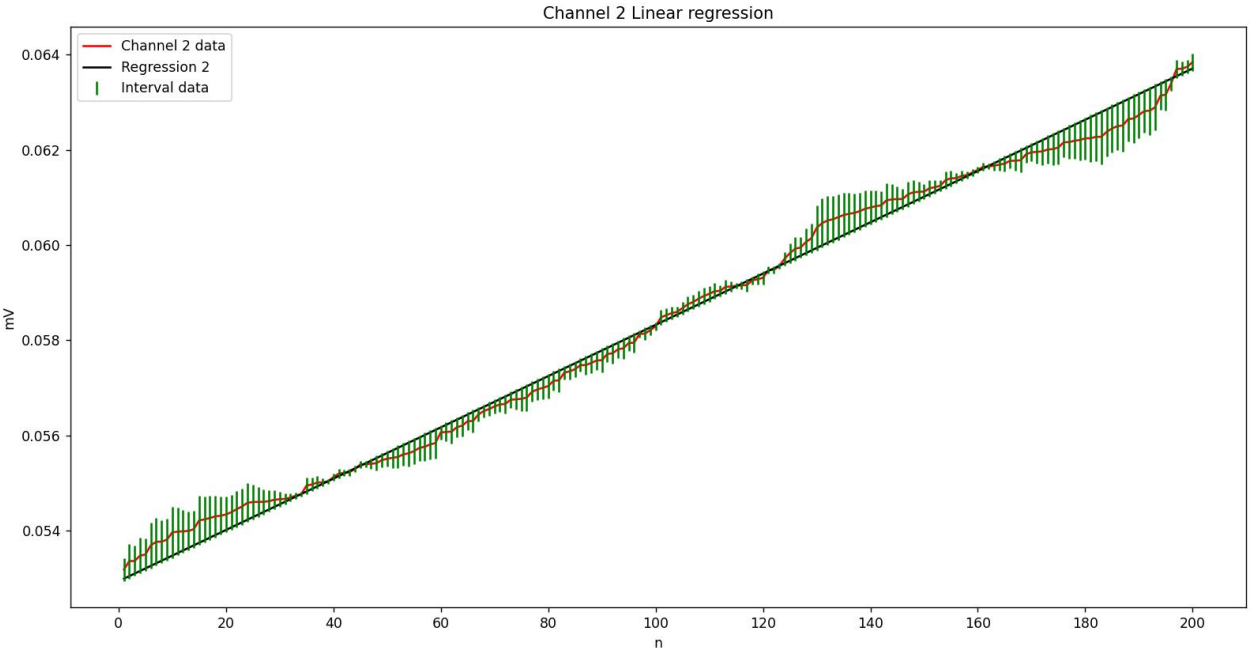

Рис. 5. Интервальные данные второй выборки и линейная регрессия

На рис. 6 визуализирован пример совместных выборок X_1', RX_2', что выполняется при R, обеспечивающим $JK > 0$.

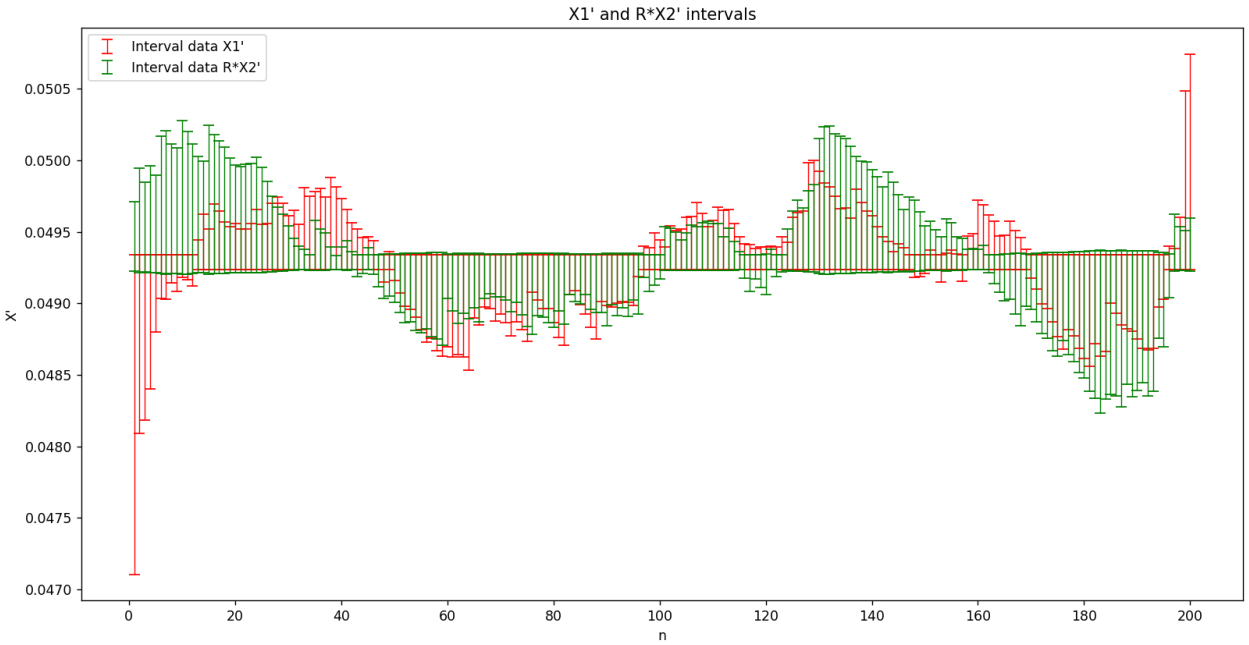

Рис. 6. Обработанные интервальные данные совместной выборки при R, обеспечивающем совместность выборок

3. Мультимера Жаккара.

На рис. 7 показана зависимость коэффицциента Жаккара от коэффициента калибровки R. Согласно внешней оценке оптимальное значение R_{opt} осуществлялся в диапазоне $\left[\underline{R}, \overline{R}\right] \approx [0.92457, 0.95941]$. 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 калибробочный коэффициент $R_{opt} \approx 0.93101$.

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

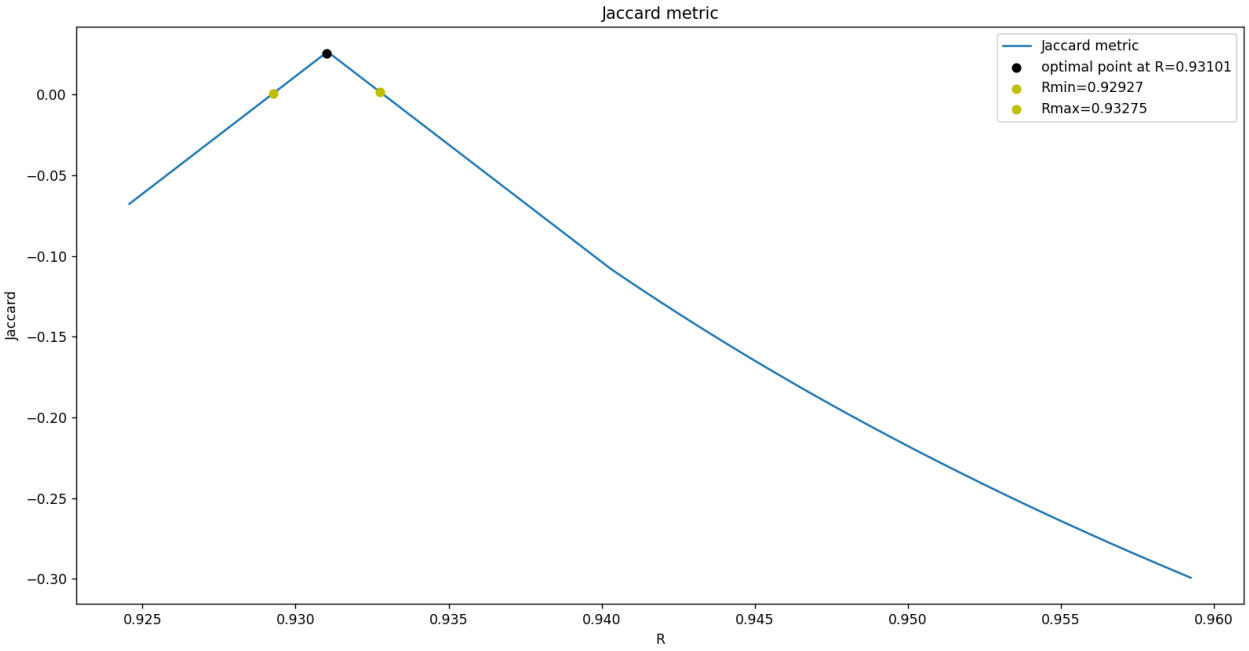

Рис. 7. Значения коэффициента Жаккара от коэффициента калибровки

4. Информационные множества.

При попытке построить информационное множество задачи оказалось, что оно пустое. Чтобы это исправить, была решена задача оптимизации.

Полученные информационные множества параметров модели изображены на рис. 8, где также обозначены центр наибольшей диагонали множества и его центр тяжести. Параметры интервальной регресии $\beta_1 \approx [4.8230 \cdot 10^{-2}, 4.9305 \cdot 10^{-2}], \beta_2 \approx [4.8886 \cdot 10^{-5}, 6.0523 \cdot 10^{-5}]$ для первой выборки и $\beta_1 \approx [5.2889 \cdot 10^{-2}, 5.3464 \cdot 10^{-2}], \beta_2 \approx [4.9362 \cdot 10^{-5}, 5.4194 \cdot 10^{-5}]$.

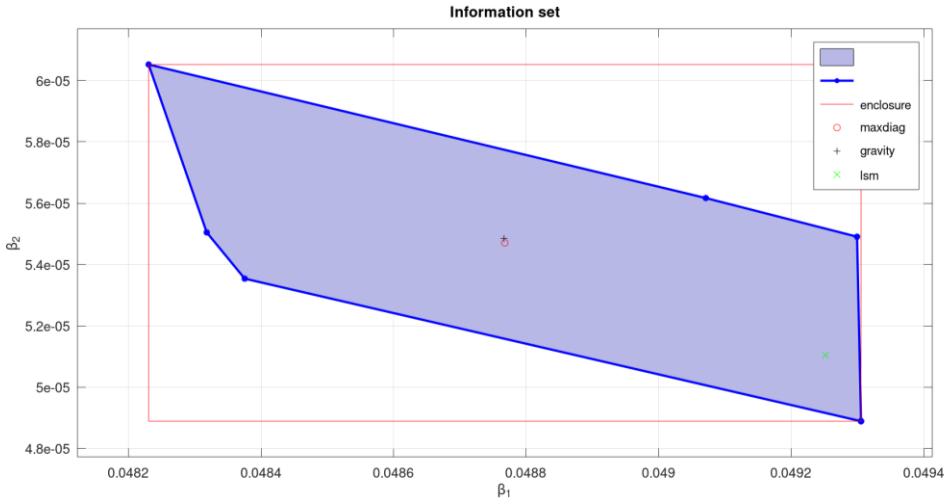

Рис. 8 (а). Информационное множество параметров для выборки 1

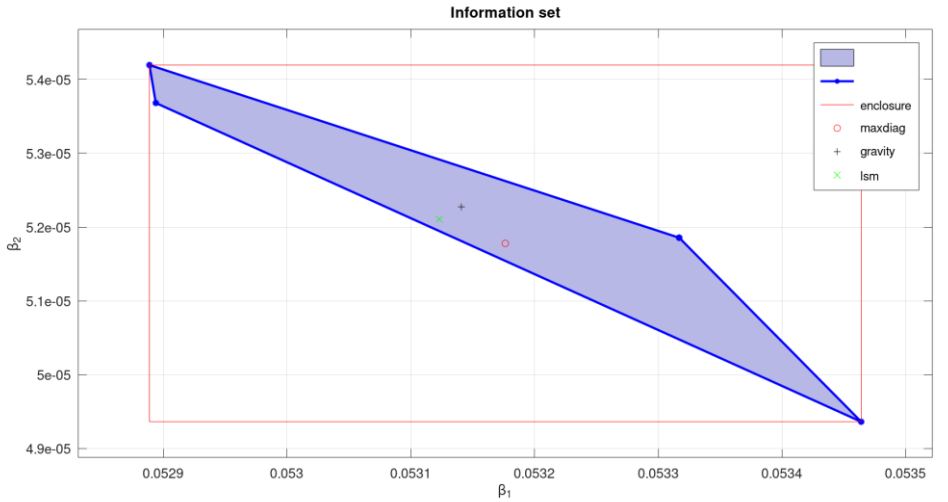

Рис. 8 (б). Информационное множество параметров для выборки 2

5. Коридор совместных зависимостей.

Графическое представление коридора совместных зависимостей модели интервальной регрессии представлено на рис. 9 и 10. Зависимость с параметрами, оцененными как центр наибольшей диагонали информационного множества обозначена красной сплошной линией, а с параметрами, оцененными как центр наибольшей диагонали информационного множества – синей пунктирной линией.

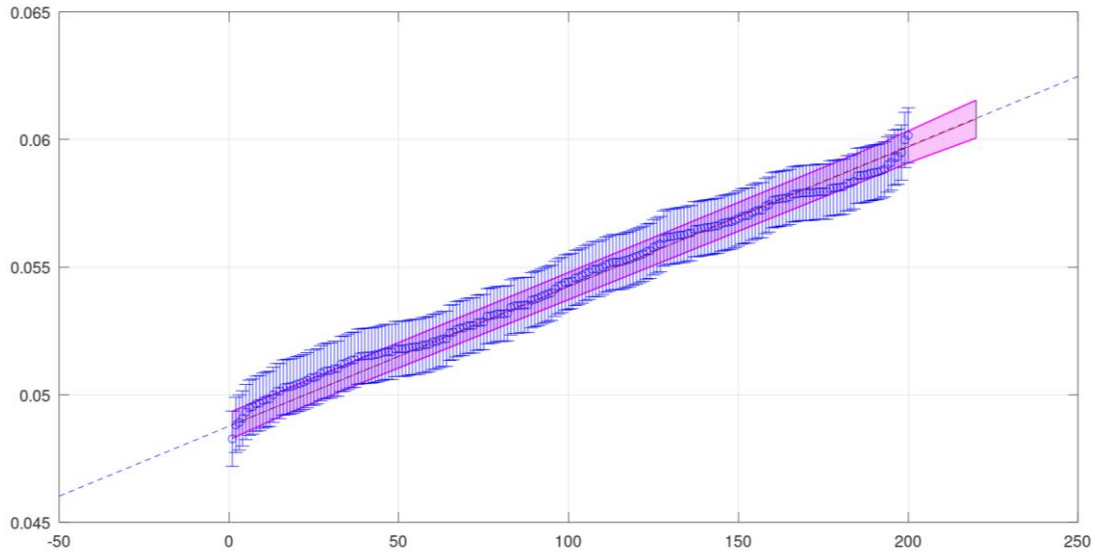

Рис. 9. Коридор совместных зависимостей (выборка 1)

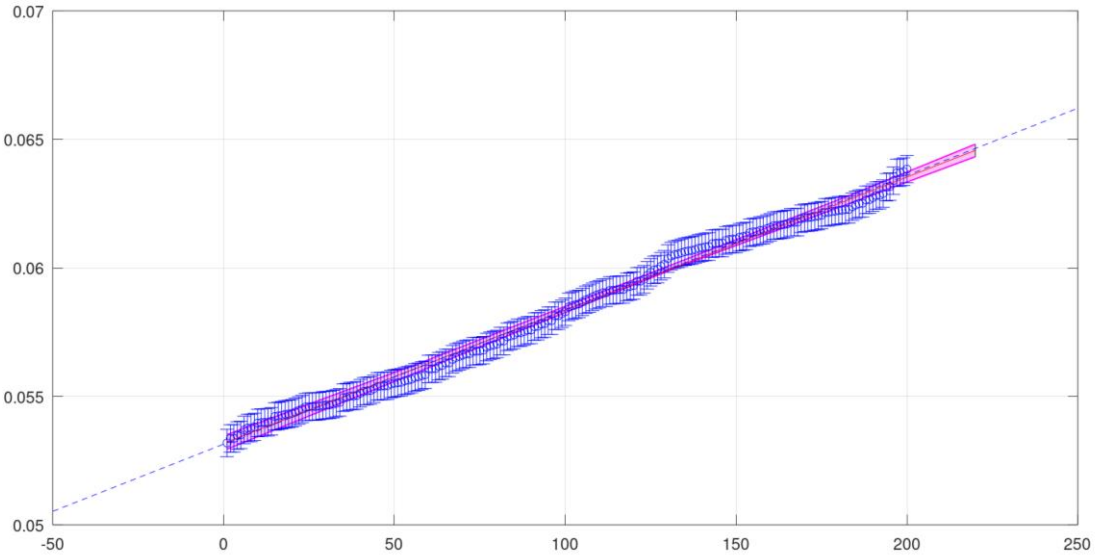

Рис. 10. Коридор совместных зависимостей (выборка 2)

6. Граничные точки.

Граничными точками оказались точки с номерами 1, 16, 38, 181, 193 в первой выборке и точки с номерами 10, 59, 132, 183 во второй выборке. В этом можно убедиться, посмотрев на рис. 13.1 – 13.5 и 14.1 – 14.4. Несмотря на имеющийся дефект визуализации, а именно наличие на рисунках «лишних» граничных синих линий интервалов от соседних точек, довольно четко прослеживается касание одного из концов интервала граничной точки границы коридора совместных зависимостей.

Рис. 13.1. Граничные точки в выборке 1

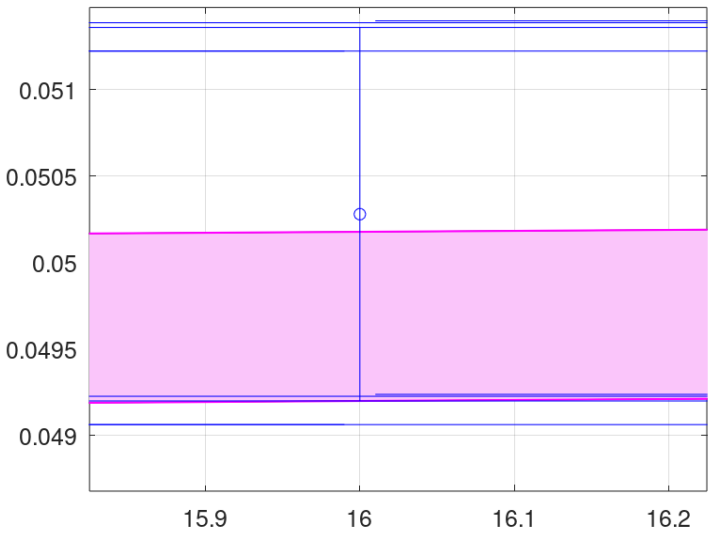

Рис. 13.2. Граничные точки в выборке 1

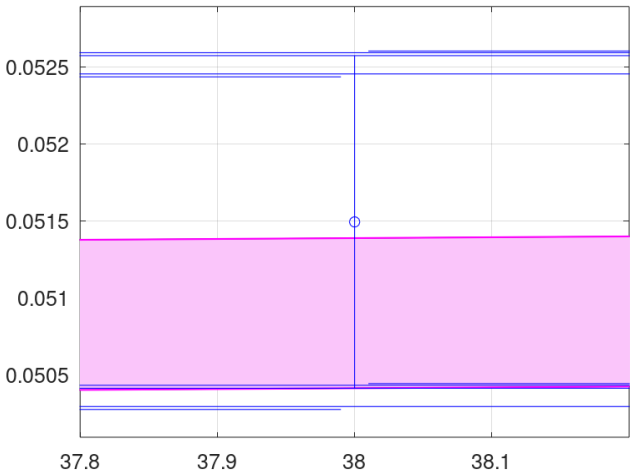

Рис. 13.3. Граничные точки в выборке 1

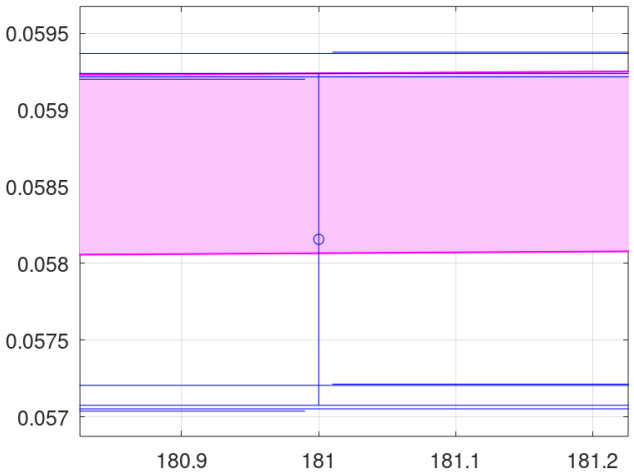

Рис. 13.4. Граничные точки в выборке 1

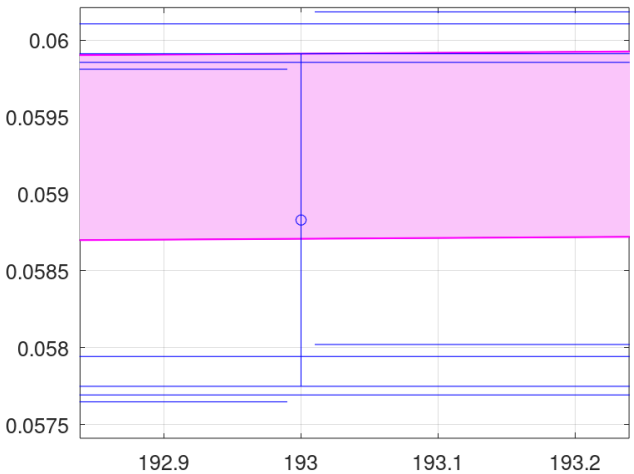

Рис. 13.5. Граничные точки в выборке 1

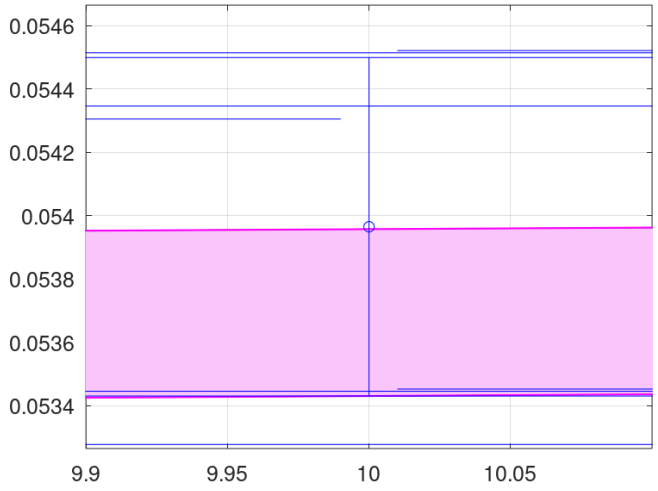

Рис. 14.1. Граничные точки в выборке 2

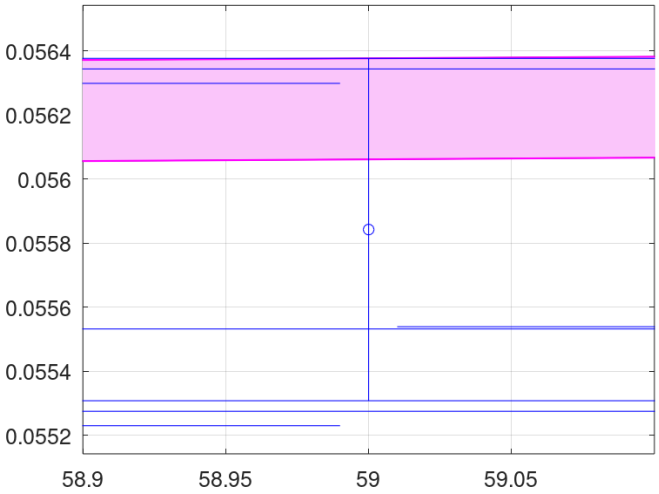

Рис. 14.2. Граничные точки в выборке 2

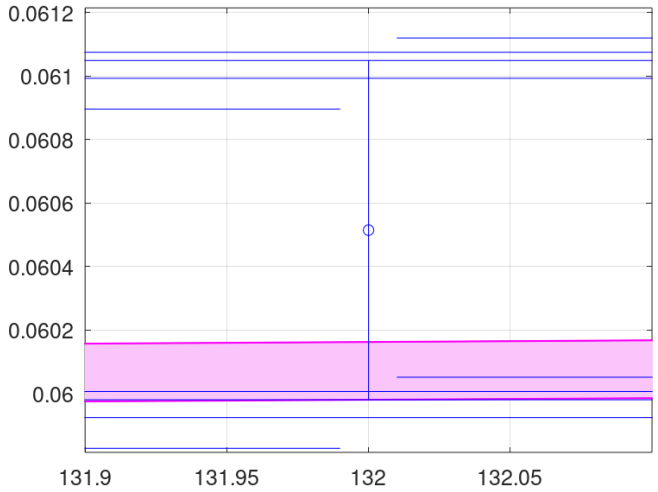

Рис. 14.3. Граничные точки в выборке 2

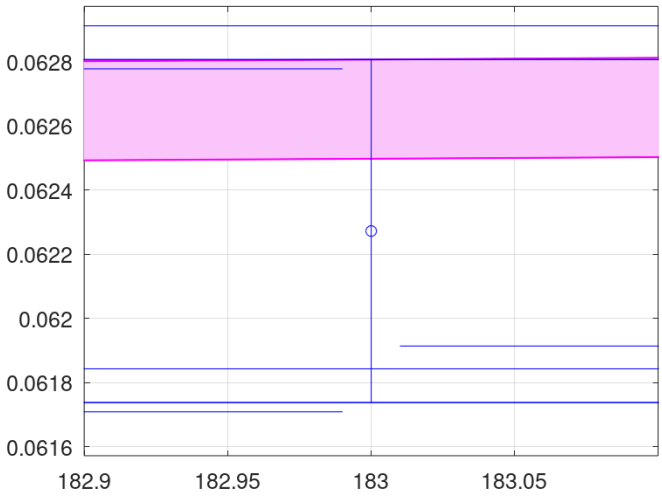

Рис. 14.4. Граничные точки в выборке 2

Теперь можно сравнить между собой построенные модели данных. На рис. 15 визуализирован пример совместных выборок X_1', RX_2', что выполняется при R, обеспечивающим меру Жаккара $JK > 0$ для выборки, приведенной к совместности с помощью вычитания дрейфовой компоненты с параметрами интервальной регрессии.

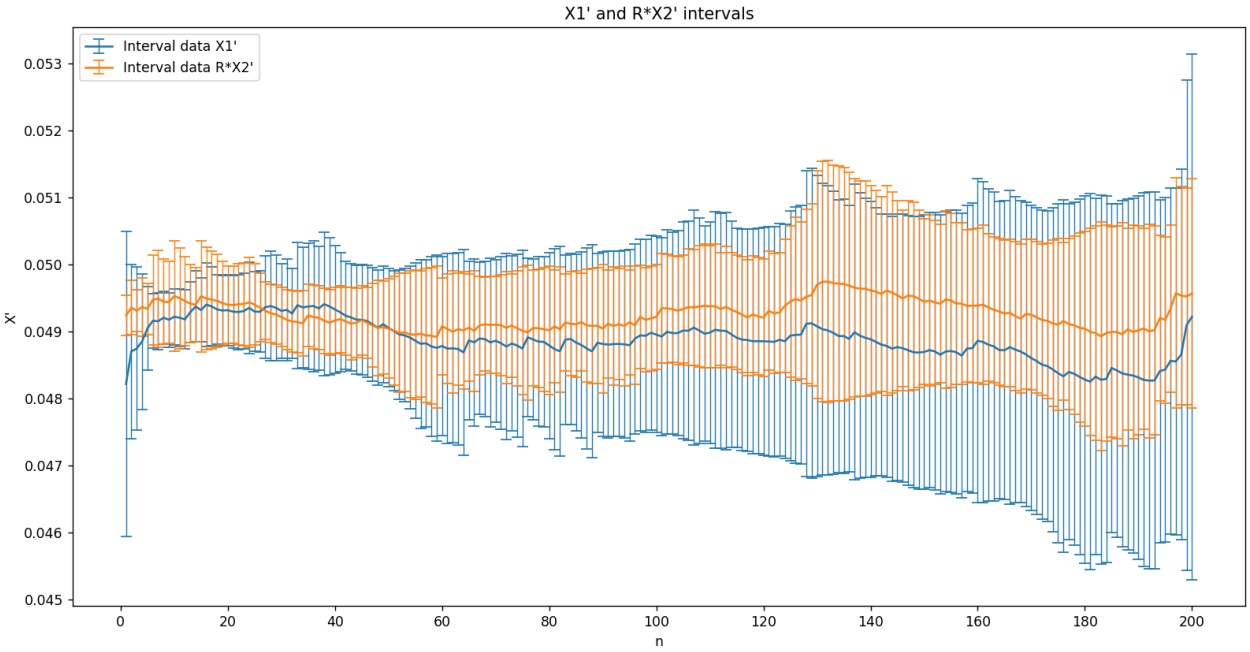

Рис. 15. Обработанные интервальные данные совместной выборки при R, обеспечивающем совместность выборок

Как видно, на рис. 15 обработанная выборка представляет собой более ровную полосу, чем и выборка, обработанная с точечными параметрами линейной регрессии, представленная на рис. 6.

На рис. 16 показана зависимость коэффициента Жаккара от коэффициента калибровки R. Согласно внешней оценке оптимальное значение R_{opt} осуществлялся в диапазоне $[\underline{R}, \overline{R}] \approx [0.88354, 0.93155]$. Как интервал можно представить $R_{12} \approx [0.92240, 0.93025]$. Максимум коэффициента Жаккара имеет значение 0.042.

При построении интервальной регресии удалось добиться максимума коэффициента Жаккара равного 0.042, что в 2 раза больше, чем в случае линейной регрессии. Это связано с тем, что модель интервальной регрессии описывает экспериментальные данные более точно и ошибки определения интервальных данных уменьшаются. Поведение коэффициента Жаккара позволило найти оптимальный калибровочный коэффициент $R_{opt} \approx 0.92677$.

Можно сказать, что мера совместности двух выборок увеличилась в условиях применения интервальной регрессии.

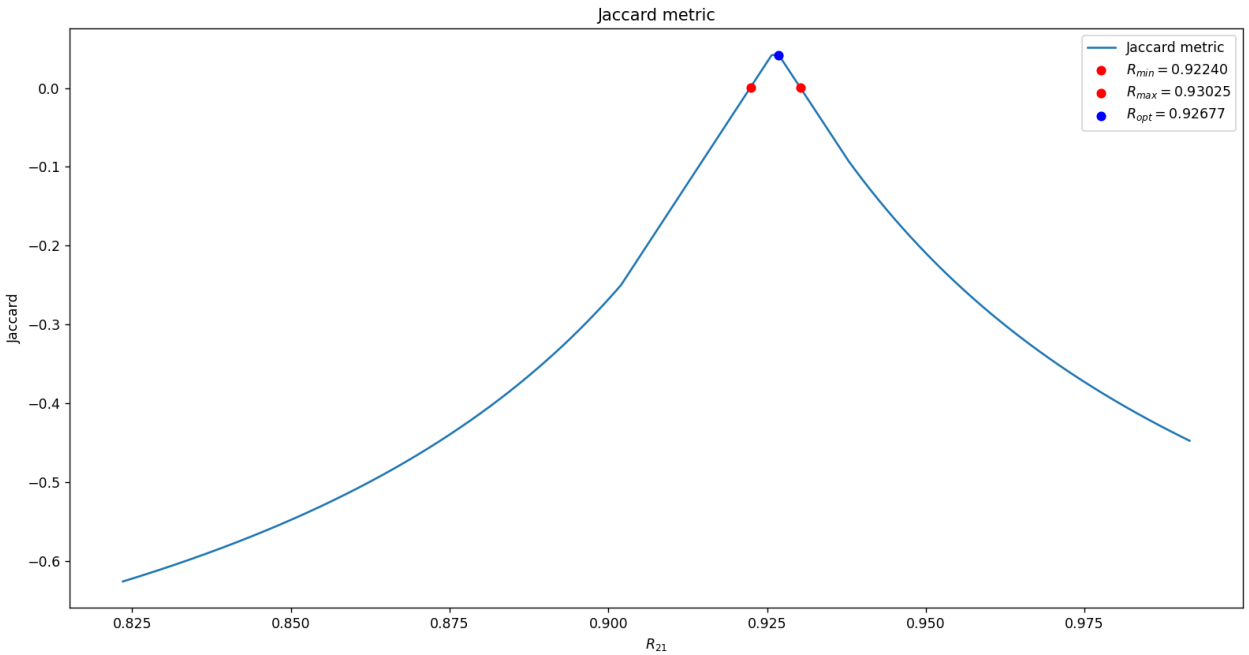

Рис. 16. Значения коэффициента Жаккара от коэффициента калибровки

Чтобы сравнить качество каждой из моделей, посчитаем функционал качества T_w по формуле (10). В случае интервальной регрессии его значение равно или практически равно 1 (с точностью до 3 знака), в случае линейной регрессии оно также близко к 1, но при этом больше по величине. Это также говорит о том, что качество модели интервальной регрессии выше.

Теперь рассмотрим интервальные регрессии по частям отрезков ломаных, определяемых угловыми точками. Визуально по рис. 4 и 5, либо рис. 9 и 10 можно выделить 5 характерных участков зависимости. Границам таких отрезков вполне соответствуют граничные точки, найденные выше. Это точки с номерами 1, 16, 38, 181, 193 в первой выборке и точки с номерами 10, 59, 132, 183 во второй выборке.

Соответствующие отрезки ломаных (участки) выделены цветом на рис. 17 и 18.

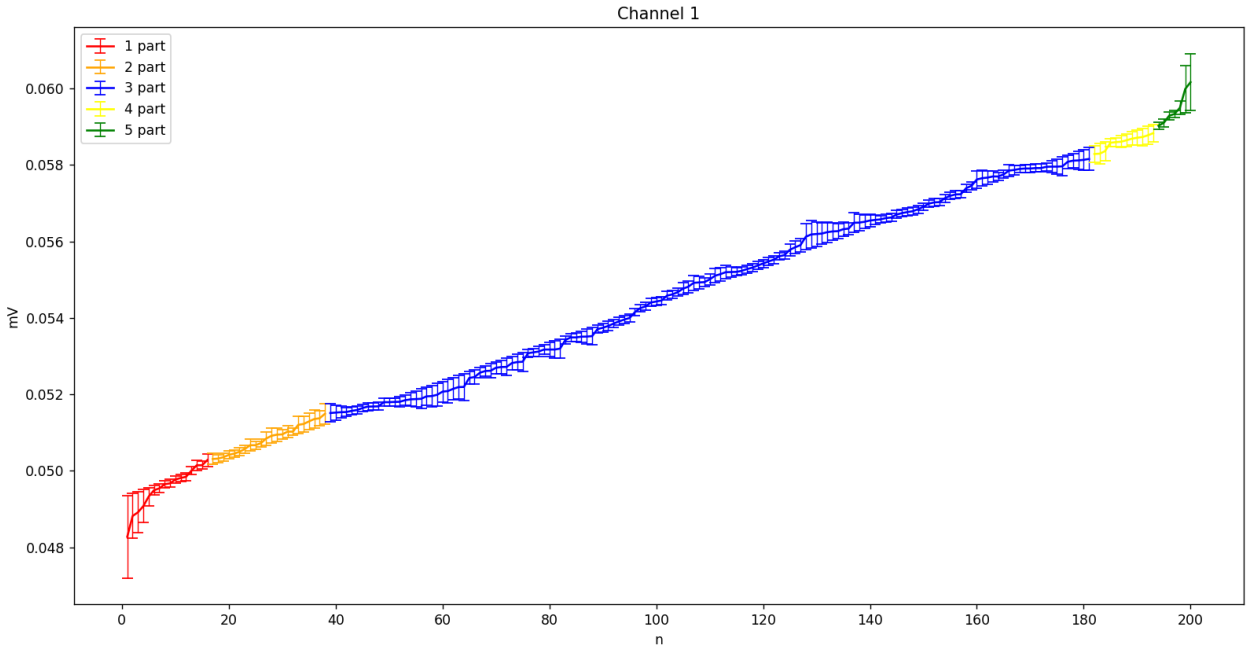

Рис. 17. Участки кусочной интервальной регрессии в 1 выборке

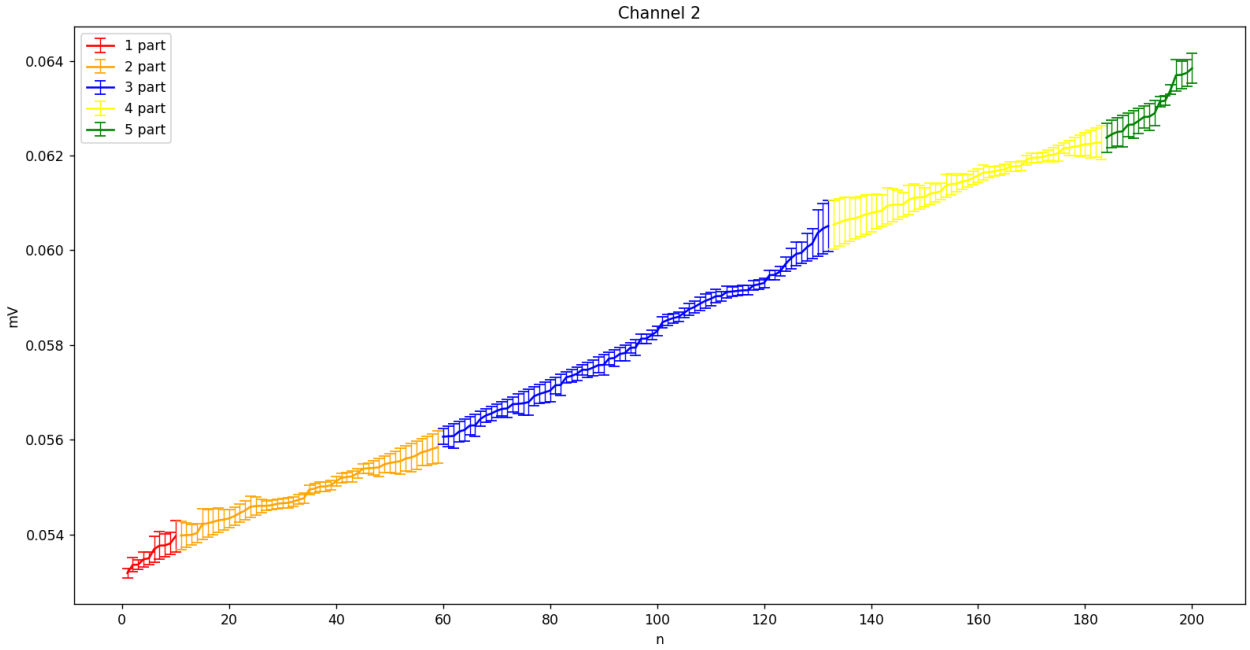

Рис. 18. Участки кусочной интервальной регрессии во 2 выборке

Коэффициенты интервальной регрессии по всей выборке:

№ выборки	β_2	β_1
1	[4.8886e-05, 6.0523e-05]	[4.8230e-02, 4.9305e-02]
2	[4.9362e-05, 5.4194e-05]	[5.2889e-02, 5.3464e-02]

Коэффициенты интервальной регрессии по отрезкам ломаных между угловыми точками следующие:

№ выборки	Граничные точки	β_2	β_1
1	1, 16	[6.1838e-05, 1.8761e-04]	[4.7913e-02, 4.8751e-02]
	16, 38	[4.9625e-05, 5.8875e-05]	[4.9238e-02, 4.9505e-02]
	38, 181	[4.8173e-05, 5.2607e-05]	[4.9063e-02, 4.9538e-02]
	181, 193	[5.6000e-05, 5.6250e-05]	[4.8074e-02, 4.8120e-02]
	193, 200	[1.4967e-04, 2.0900e-04]	[1.8302e-02, 3.0040e-02]
2	1, 10	[6.4333e-05, 9.3000e-05]	[5.3075e-02, 5.3222e-02]
	10, 59	[3.7922e-05, 3.8386e-05]	[5.3563e-02, 5.3579e-02]
	59, 132	[5.4972e-05, 6.5068e-05]	[5.1832e-02, 5.2928e-02]
	132, 183	[3.3118e-05, 3.7682e-05]	[5.5476e-02, 5.6228e-02]
	183, 200	[8.6000e-05, 1.0790e-04]	[4.2296e-02, 4.6523e-02]

7. Кусочная интервальная регрессия.

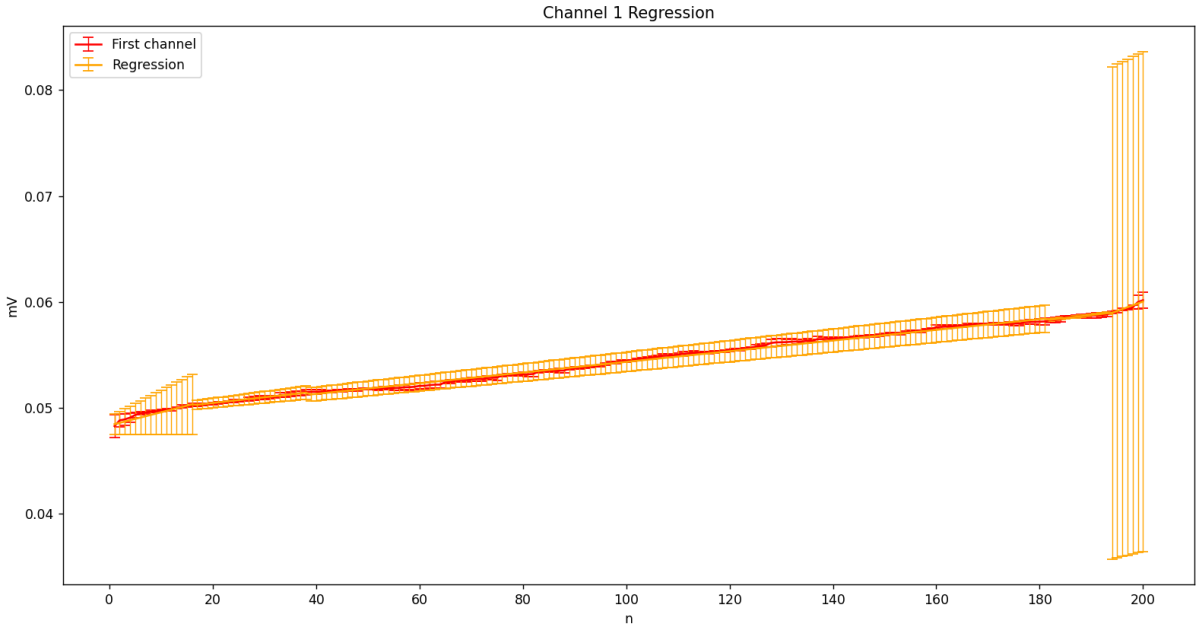

Рис. 19. Интервальные данные и кусочная интервальная регрессия – выб. 1

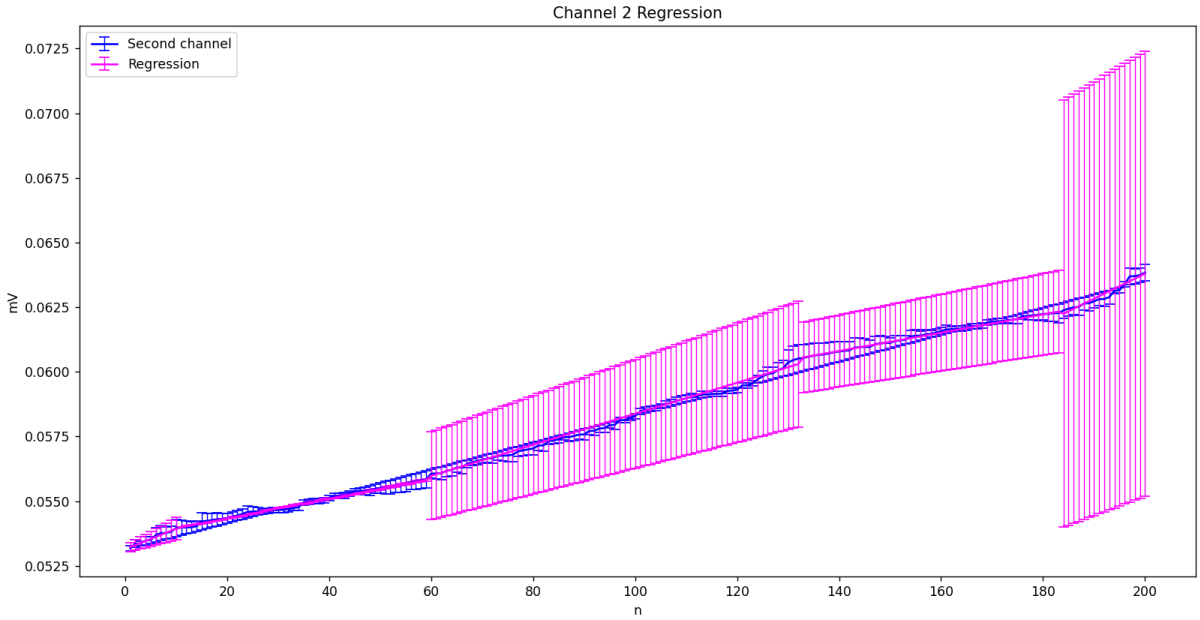

Рис. 20. Интервальные данные и кусочная интервальная регрессия – выб. 2

8. Остатки.

Вычтем из всех интервальных данных, построенных по этой модели, регрессию центральной выборки (точки 38 – 181 и 59 – 132 для 1 и 2 канала соответственно) – она хорошо описывает большую часть данных, поэтому ее можно принять за основную. Полученные остатки показаны на рис. 24, где также обозначен прогнозный интервал $\gamma(x_i) = \bigwedge_i x_i = [\max\min(x_i), \min\max(x_i)]$.

Диаграмма статусов (рис. 25): зеленая область – надежные данные, имеющие малое плечо и малую погрешность, желтая область – менее надежные данные, красная область – выбросы, или данные, образующие другую группу связности.

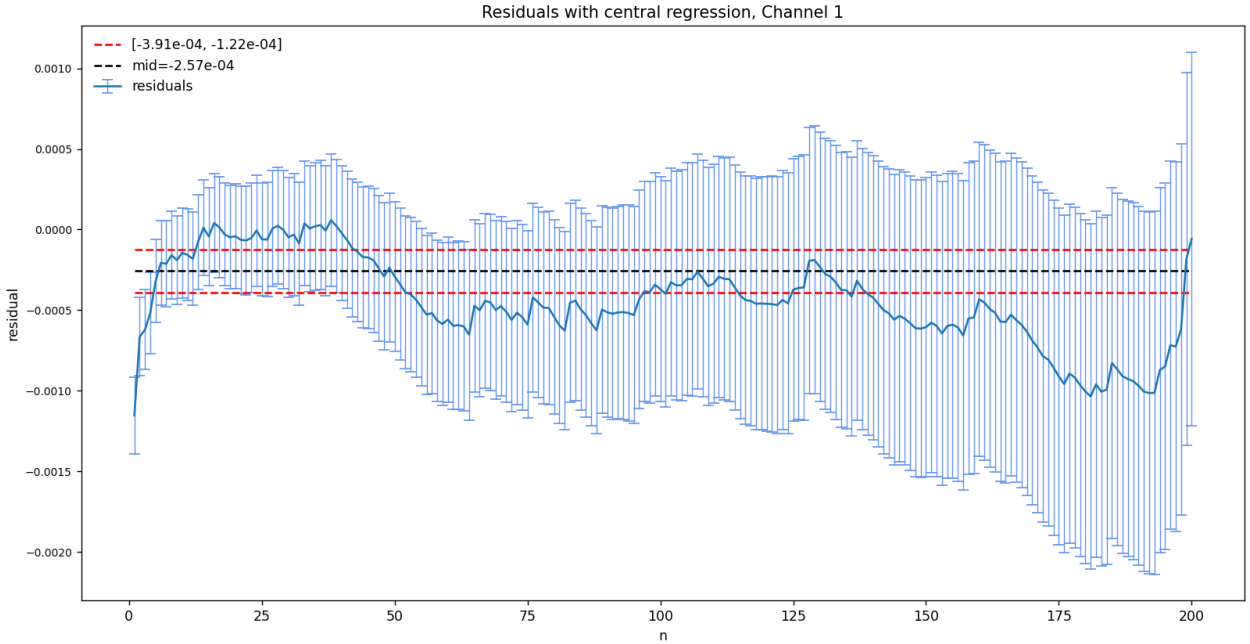

Рис. 24. Остатки при вычитании центральной регресии и прогнозный интервал для модели кусочной интервальной регресии (выборка 1)

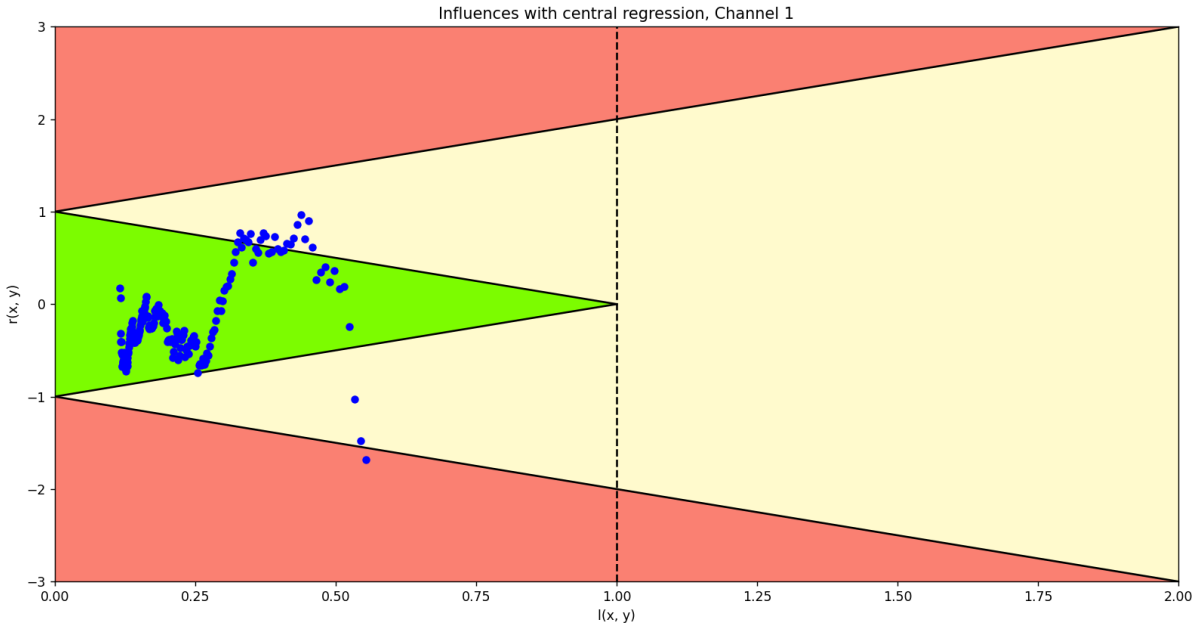

Рис. 26. Диаграммы статусов исходных данных (выборка 1)

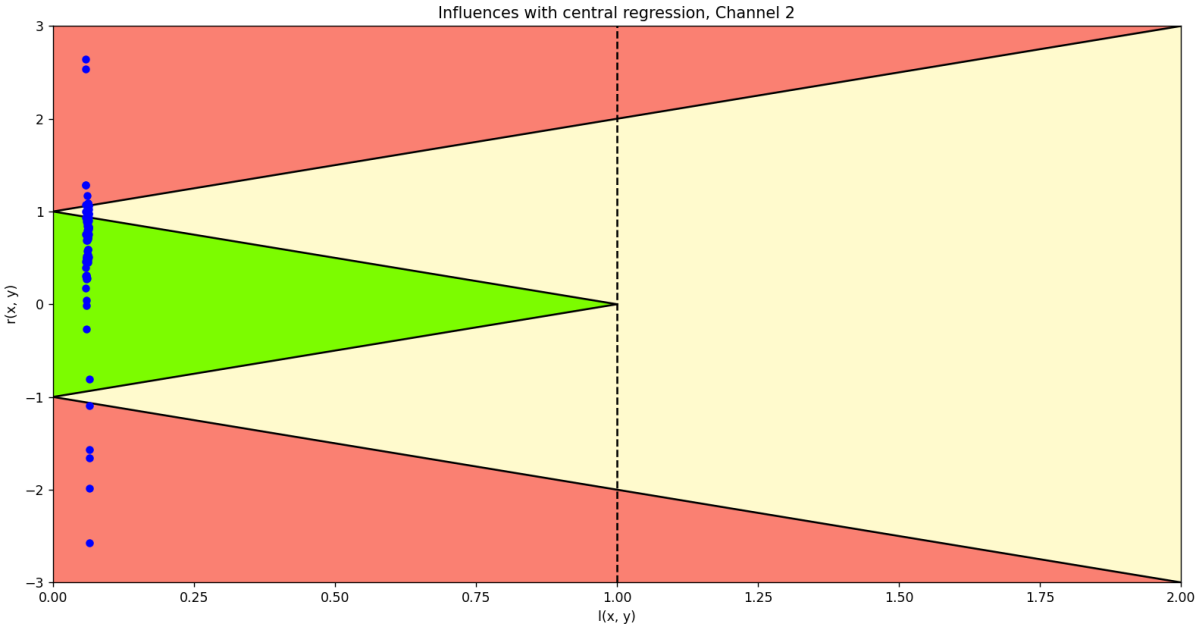

Рис. 27. Диаграммы статусов исходных данных (выборка 2)

Регуляризация остатков осуществляется следующим образом: следует растягивать интервалы остатков до границы пересечения большинства отатков, то есть до границы прогнозного интервала (рис. 26), после чего построим диаграмму статусов (рис. 27). Видим, что теперь все данные являются надежными и попадают в зеленую область.

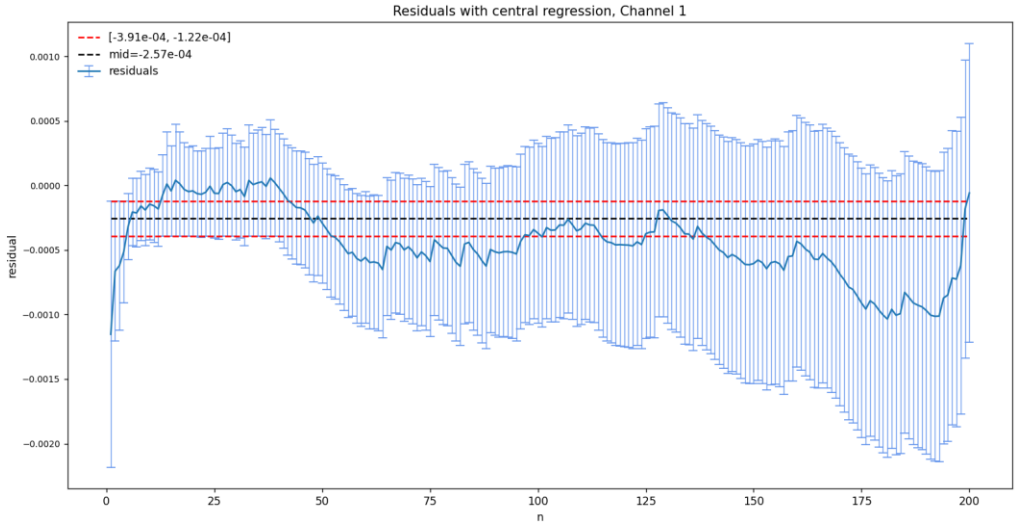

Рис. 28. Остатки при вычитании центральной регрессии и прогнозный интервал для модели кусочной интервальной регрессии после регуляризации (выборка 1)

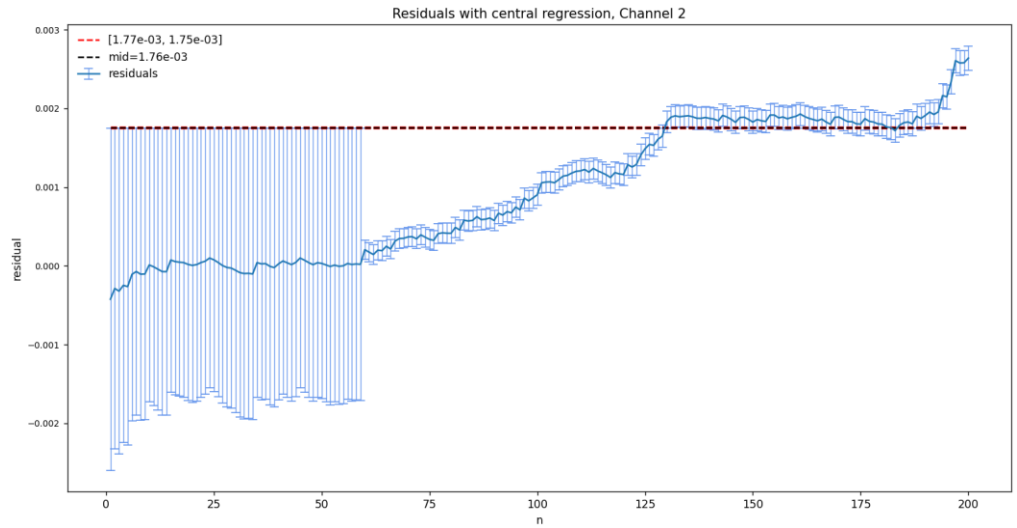

Рис. 29. Остатки при вычитании центральной регрессии и прогнозный интервал для модели кусочной интервальной регрессии после регуляризации (выборка 2)

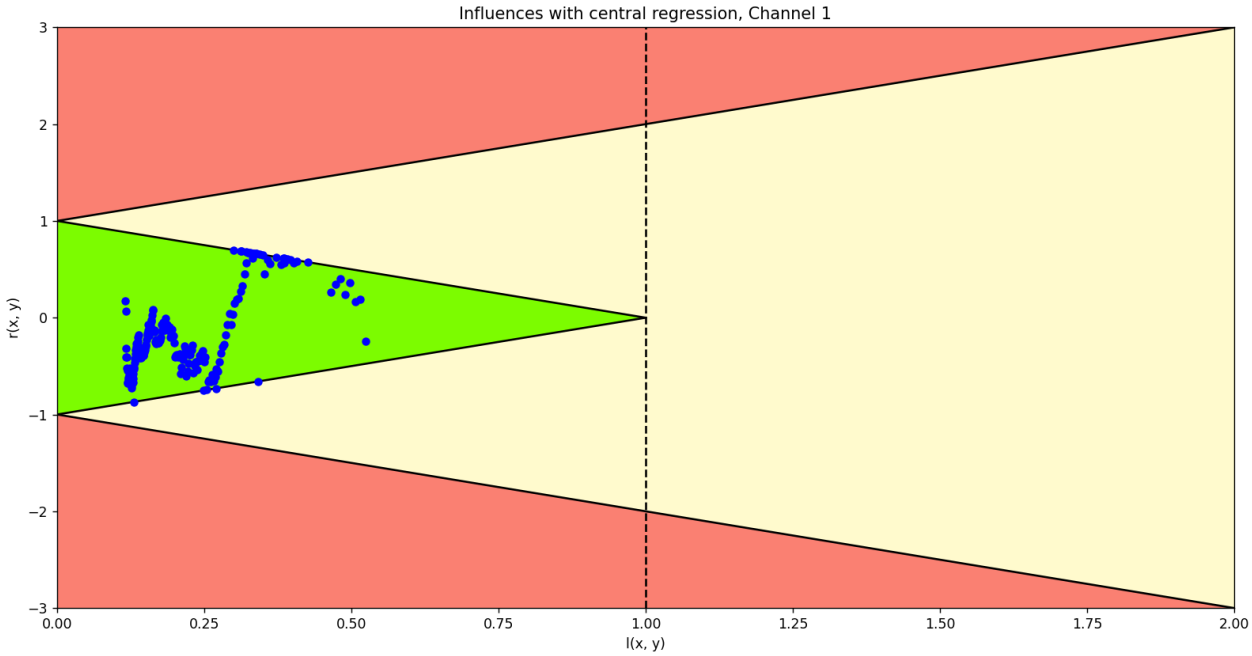

Рис. 30. Диаграммы статусов исходных данных после регуляризации (выборка 1)

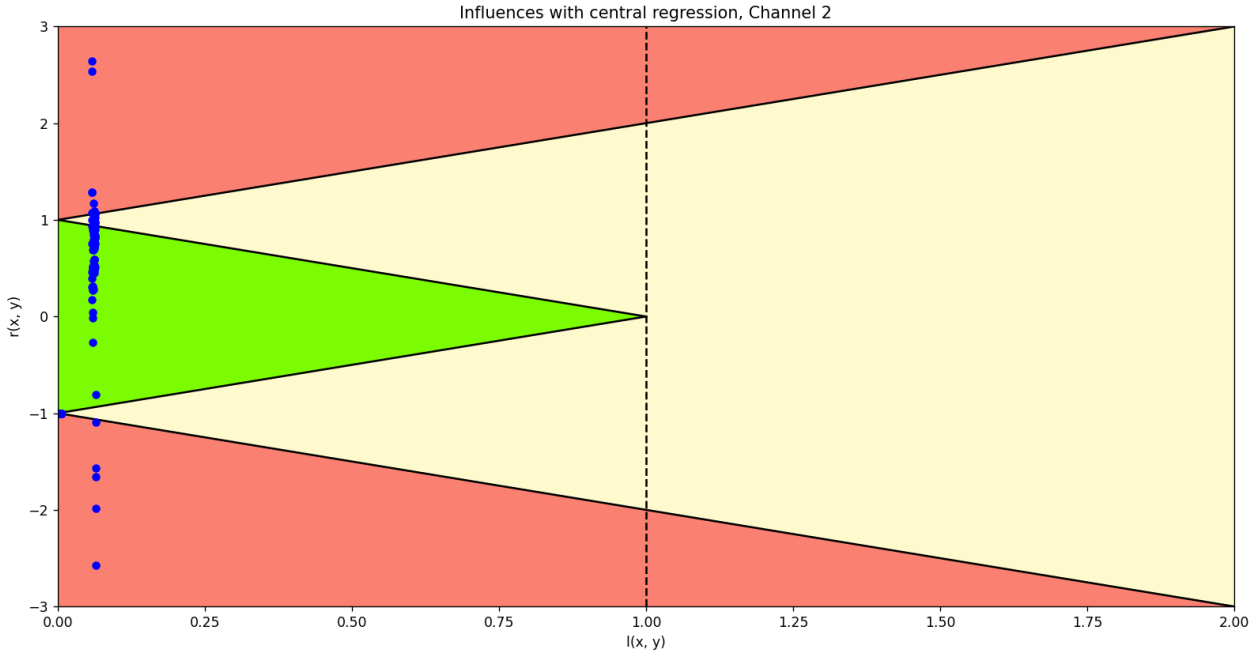

Рис. 31. Диаграммы статусов исходных данных после регуляризации (выборка 2)

Коэффициенты линейной регрессии

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

Коэффициенты интервальной регрессии

№ выборки	β_2	β_1
1	[4.8886e-05, 6.0523e-05]	[4.8230e-02, 4.9305e-02]
2	[4.9362e-05, 5.4194e-05]	[5.2889e-02, 5.3464e-02]

Коэффициенты интервальной регрессии по отрезкам ломаных, определяемых угловыми точками

№ выборки	Граничные точки	β_2	β_1
1	1, 16	[6.1838e-05, 1.8761e-04]	[4.7913e-02, 4.8751e-02]
	16, 38	[4.9625e-05, 5.8875e-05]	[4.9238e-02, 4.9505e-02]
	38, 181	[4.8173e-05, 5.2607e-05]	[4.9063e-02, 4.9538e-02]
	181, 193	[5.6000e-05, 5.6250e-05]	[4.8074e-02, 4.8120e-02]
	193, 200	[1.4967e-04, 2.0900e-04]	[1.8302e-02, 3.0040e-02]
2	1, 10	[6.4333e-05, 9.3000e-05]	[5.3075e-02, 5.3222e-02]
	10, 59	[3.7922e-05, 3.8386e-05]	[5.3563e-02, 5.3579e-02]
	59, 132	[5.4972e-05, 6.5068e-05]	[5.1832e-02, 5.2928e-02]
	132, 183	[3.3118e-05, 3.7682e-05]	[5.5476e-02, 5.6228e-02]
	183, 200	[8.6000e-05, 1.0790e-04]	[4.2296e-02, 4.6523e-02]

Ссылки на GitHub с реализацией всех работ

1. Ссылка на GitHub с реализацией работы 1 – «Линейная регрессия, МНК»:
https://github.com/dimerf99/interval_analysis/tree/main/lab_1

2. Ссылка на GitHub с реализацией работы 2 – «Интервальная регрессия»:
https://github.com/dimerf99/interval_analysis/tree/main/lab_2

3. Ссылка на GitHub с реализацией работы 3 – «Анализ остатков»:
https://github.com/dimerf99/interval_analysis/tree/main/lab_3

Файлы данных

Канал 1_700nm_0.03.csv

Канал 2_700nm_0.03.csv