Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Владимирский государственный университет имени Александра
Григорьевича и Николая Григорьевича Столетовых»
(ВлГУ)
Кафедра ИЗИ

ОТЧЕТ
по лабораторной работе №3
по дисциплине «МАД»
ТЕМА: «ПРЕДОБРАБОТКА ДАННЫХ»

Студент гр. ИСБ-121 __ Ходыкин Л.Ю.

Преподаватель __ Артюшина Л.А.

Владимир
2024

Оглавление

Задание.

Задание:

1.1 Используя текстовый редактор Word создать файл
1.2 Импортируйте данные из созданного файла в Deductor.
1.3 Восстановите пропущенные данные в столбце «Синус»
1.4 Удалите аномалии в данных
1.5 Сгладьте данные методом спектральной обработки
1.6 Удалите шумы из данных

2.1 импортировать данные из файла в Deductor;

2.2 используя инструмент Фильт, выделите из получившейся таблицы страны, согласно индивидуальному варианту (по 2 столбца таблицы);

2.3 визуализируйте данные;

2.4 при наличии заполните пропуски в данных. Обоснуйте выбранный способ заполнения;

2.5 визуализируйте данные после заполнения пропусков;

2.6 удалите аномалии в данных;

2.7 сгладьте данные одним из методов спектральной обработки. Обоснуйте выбор метода;

2.8 визуализируйте данные после удаления аномалий;

2.9 обоснуйте наличие шумов в данных. Удалите шумы из данных (при наличии);

2.10 Создайте многомерный отчет по количеству заболевших в виде OLAP-куба, показатели выберите самостоятельно (например, общее число заболевших по странам и т.п.).

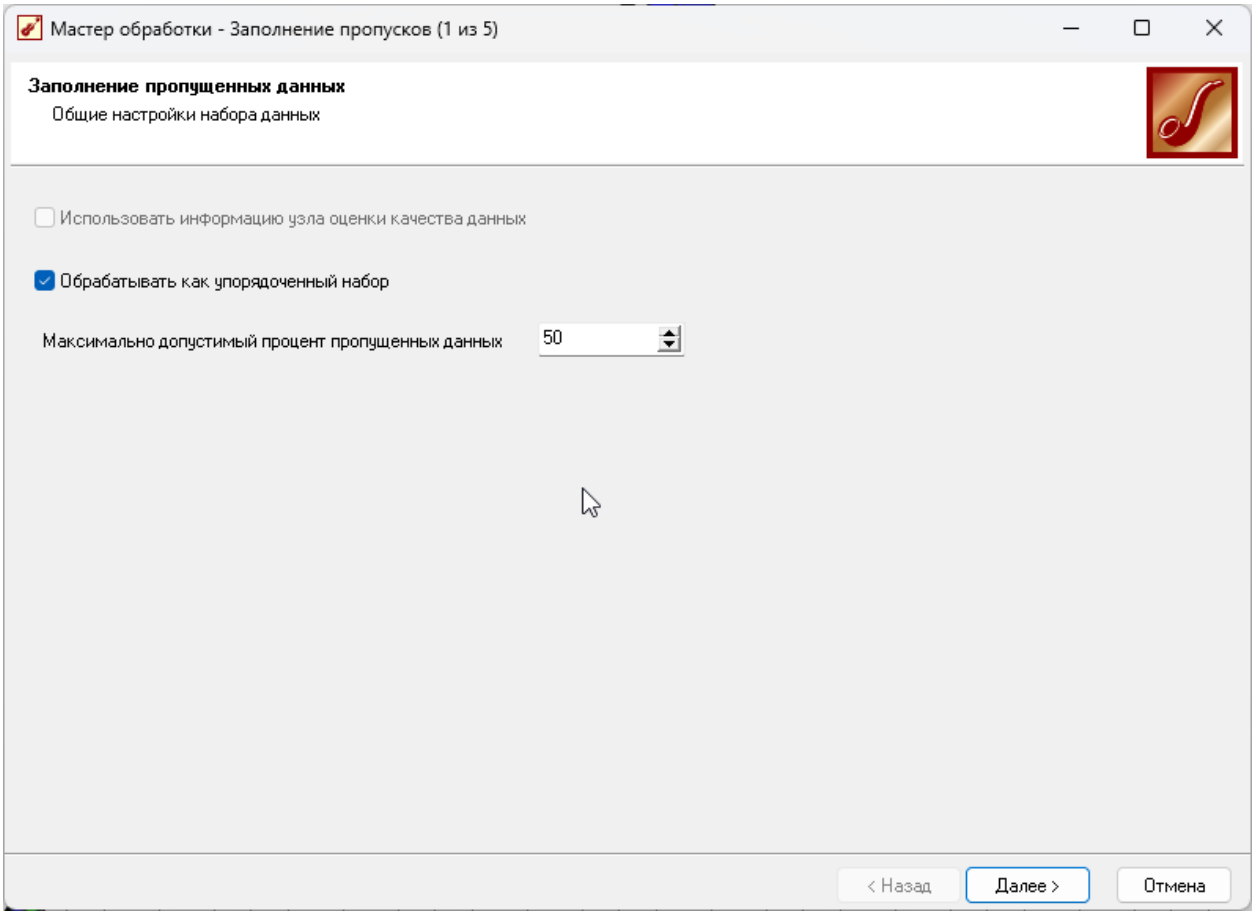
Ход работы 1 задание.

Создаем файл со столбцами – Аргумент, Синус, Аномалии, Большие Шумы, Средние Шумы, Малые Шумы.

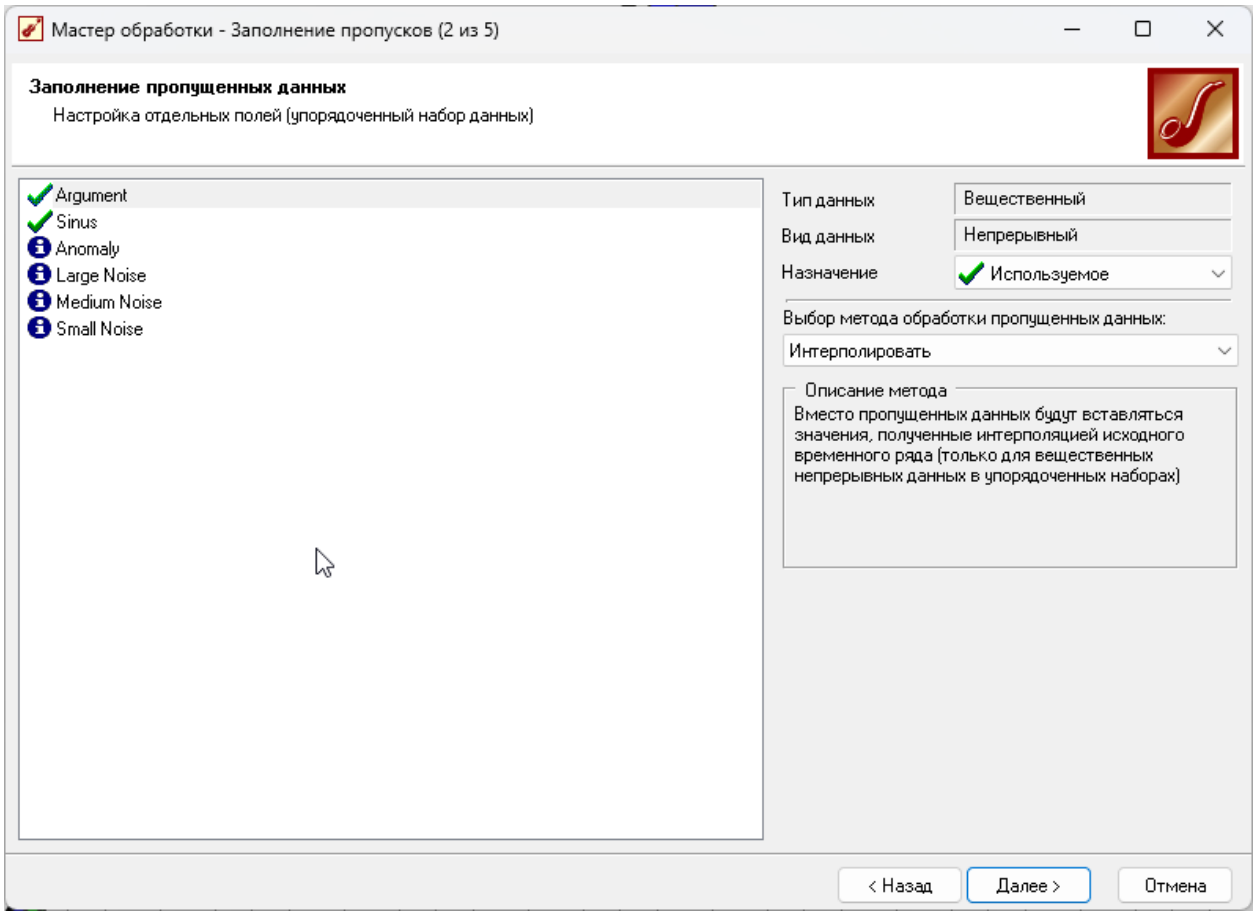
Argument	Sinus	Anomaly	Large Noise	Medium Noise	Small Noise
0.000000000	0.000000000	0.079914694	0.023834934	-0.049147772	-0.008032414
0.020000000	0.019998667	0.019998667	0.035346857	-0.023802290	0.010271700
0.040000000	0.039989334	0.059964006	0.134215134	0.101732203	0.056760332
0.060000000	0.059964006	0.059964006	0.064336934	0.036772284	0.040865458
0.080000000	0.079914694	0.079914694	-0.064865690	0.080033216	0.071900237
0.100000000	0.099833417	0.099833417	0.102934196	0.030902416	0.106550496
0.120000000	0.119712207	0.119712207	0.076163088	0.179586020	0.119247206
0.140000000	0.139543115	0.139543115	0.146402760	0.151334605	0.132139887
0.160000000	0.159318207	0.159318207	0.038418294	0.213413657	0.170795312
0.180000000	0.179029573	0.179029573	0.292217395	0.138959784	0.172577694
0.200000000	0.198669331	0.198669331	0.104018445	0.127732180	0.207973849
0.220000000	0.218229623	0.198669331	0.282018452	0.262107705	0.217532104
0.240000000	0.237702626	0.237702626	0.414769226	0.227303021	0.241287057
0.260000000	0.257080552	0.257080552	0.285008637	0.226875528	0.274392074
0.280000000	0.276355649	0.276355649	0.232482808	0.247364849	0.272814433
0.300000000	0.295520207	0.295520207	0.200311542	0.221370401	0.295651797
0.320000000	0.314566561	0.314566561	0.298577344	0.304446289	0.318777560
0.340000000	0.333487092	0.333487092	0.457663415	0.357547962	0.329305261
0.360000000	0.352274233	0.352274233	0.248535588	0.379479480	0.353890904
0.380000000	0.370920469	0.370920469	0.371209442	0.298687001	0.374706208
0.400000000	0.389418342	0.352274233	0.482408129	0.403232394	0.379481870
0.420000000	0.407760453	0.407760453	0.426426145	0.420541009	0.408829429
0.440000000	0.425939465	0.425939465	0.357989956	0.451032228	0.436068216
0.460000000	0.443948107	0.461779176	0.571466869	0.565075088	0.430901859
0.480000000	0.461779176	0.461779176	0.407071183	0.496868321	0.476819297
0.500000000	0.479425539	0.479425539	0.551837898	0.490657268	0.486050577
0.520000000	0.496880138	0.496880138	0.465588559	0.485270633	0.498147726
0.540000000	0.514135992	0.514135992	0.278569566	0.478830734	0.517984159
0.560000000	0.531186198	0.531186198	0.555910496	0.573211965	0.527753621
0.580000000	0.548023937	0.548023937	0.500921765	0.590447079	0.538078755
0.600000000	0.564642473	0.564642473	0.514768949	0.554359703	0.560216576
0.620000000	0.581035161	0.581035161	0.608518813	0.635938262	0.575216029
0.640000000	0.597195441	0.597195441	0.381441989	0.615273238	0.593728733
0.660000000	0.613116852	0.613116852	0.810053369	0.644584710	0.611191840
0.680000000	0.628793024	0.564642473	0.607654046	0.572206817	0.643697066
0.700000000	0.644217687	0.644217687	0.719469146	0.624530389	0.637676888
0.720000000	0.659384672	0.659384672	0.514273191	0.604166317	0.664693432
0.740000000	0.674287912	0.674287912	0.674762174	0.654361139	0.685770217
0.760000000	0.688921445	0.688921445	0.719196803	0.676934425	0.706751512
0.780000000	0.703279419	0.703279419	0.633862940	0.736795667	0.691378332
0.800000000	0.717356091	0.717356091	0.701852718	0.694907810	0.703719691
0.820000000	0.731145830	0.731145830	0.693102708	0.708464384	0.730554219
0.840000000	0.744643120	0.744643120	0.753082864	0.786260804	0.737966239
0.860000000	0.757842563	0.757842563	0.795911547	0.864086706	0.747049789
0.880000000	0.770738879	0.770738879	0.753916689	0.727164734	0.764832918

Далее загружаем данный файл и визуализируем по синусу и аргументу

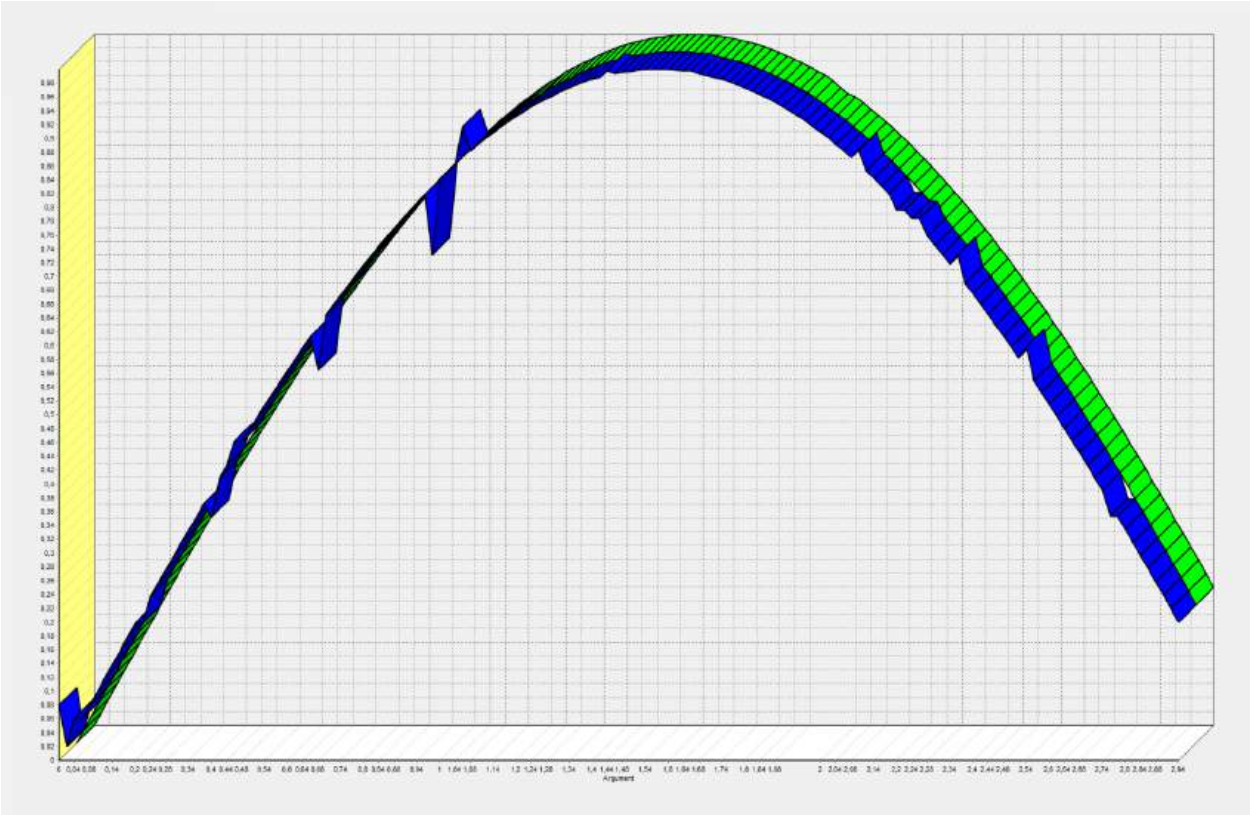
Дальше заполняем пропуски с помощью заполнения пропусков в мастере обработки



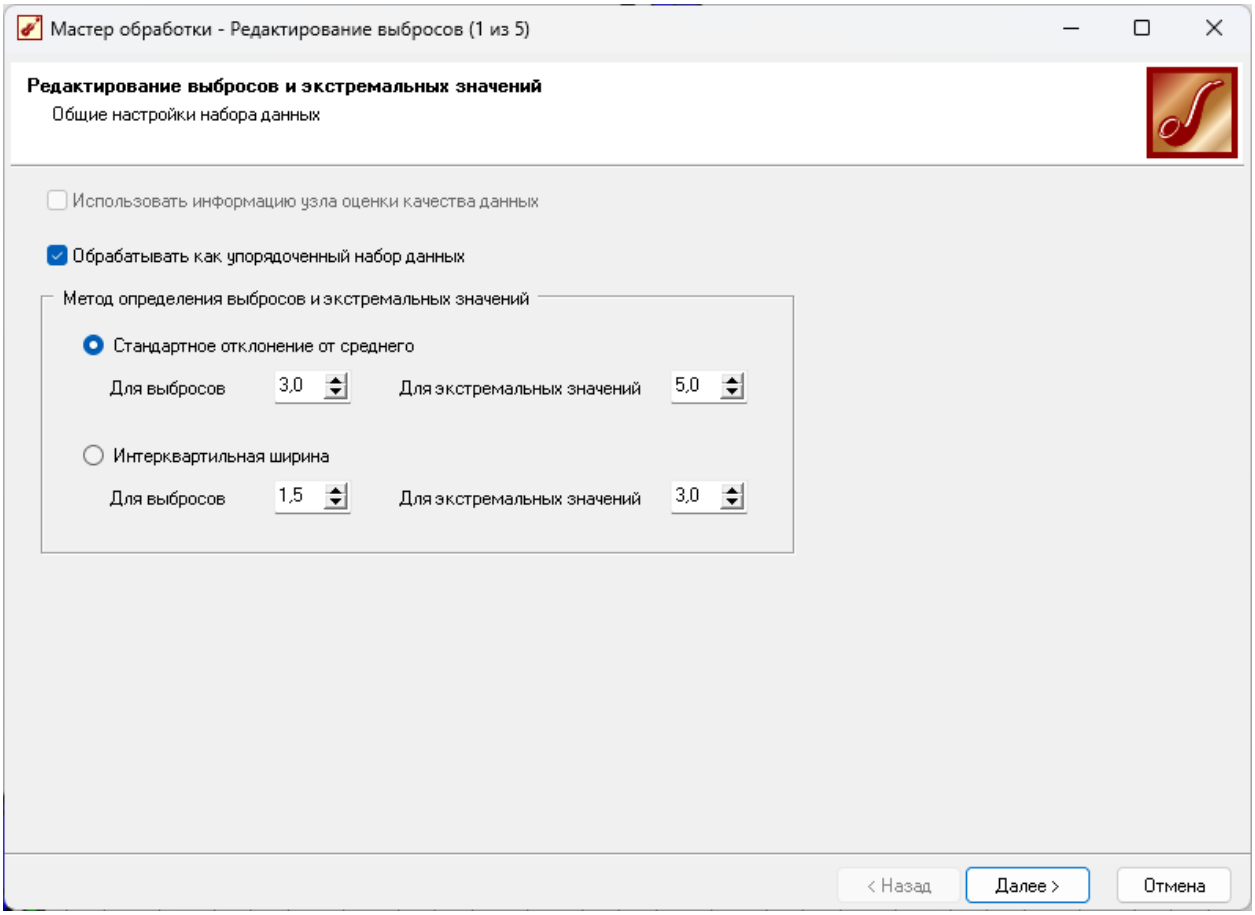
Нам необходимо только синус и аргумент интерполировать. Используется интерполяция потому что данные упорядочены.



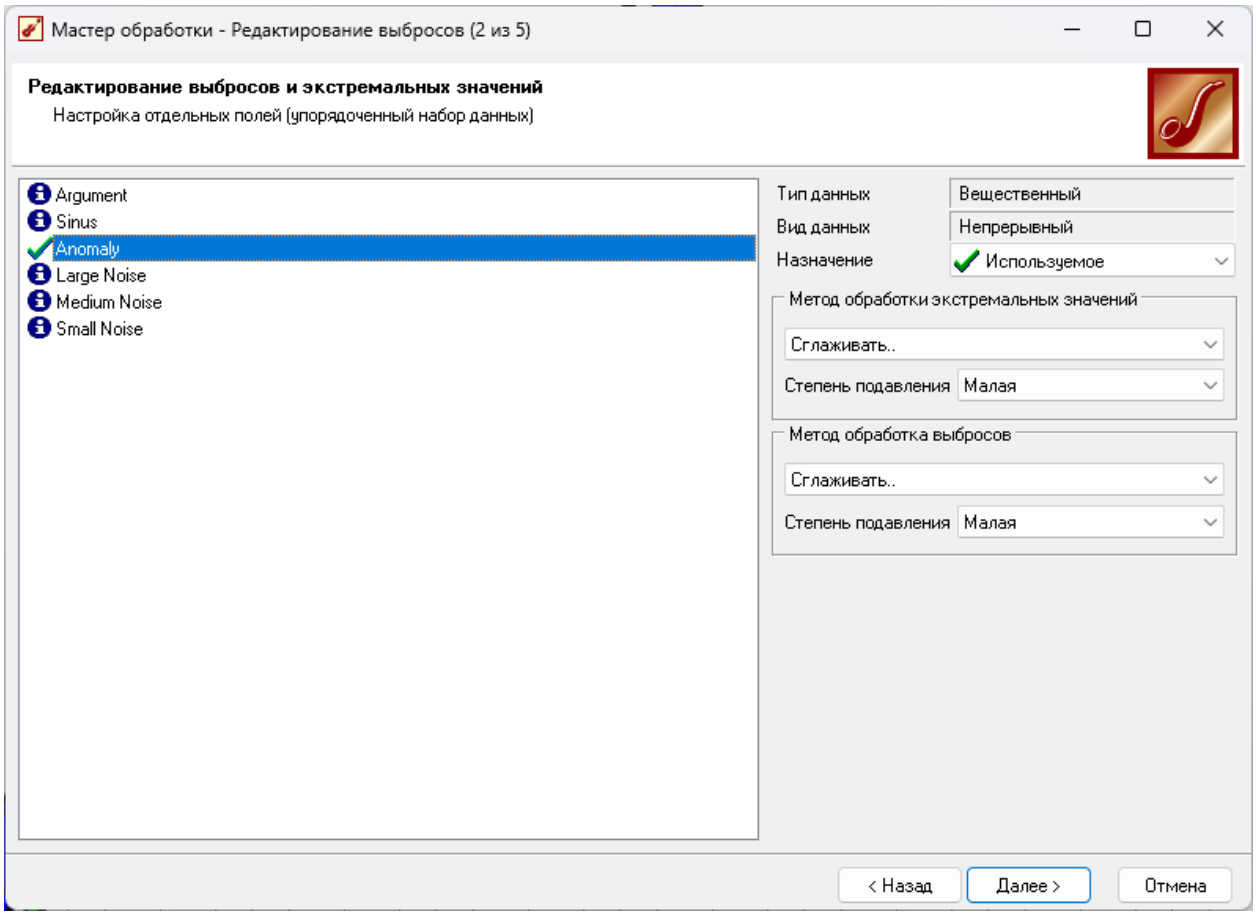
Получаем вот такую обновленную диаграмму для синуса (зеленая) и диаграмму для аномалий(синяя)



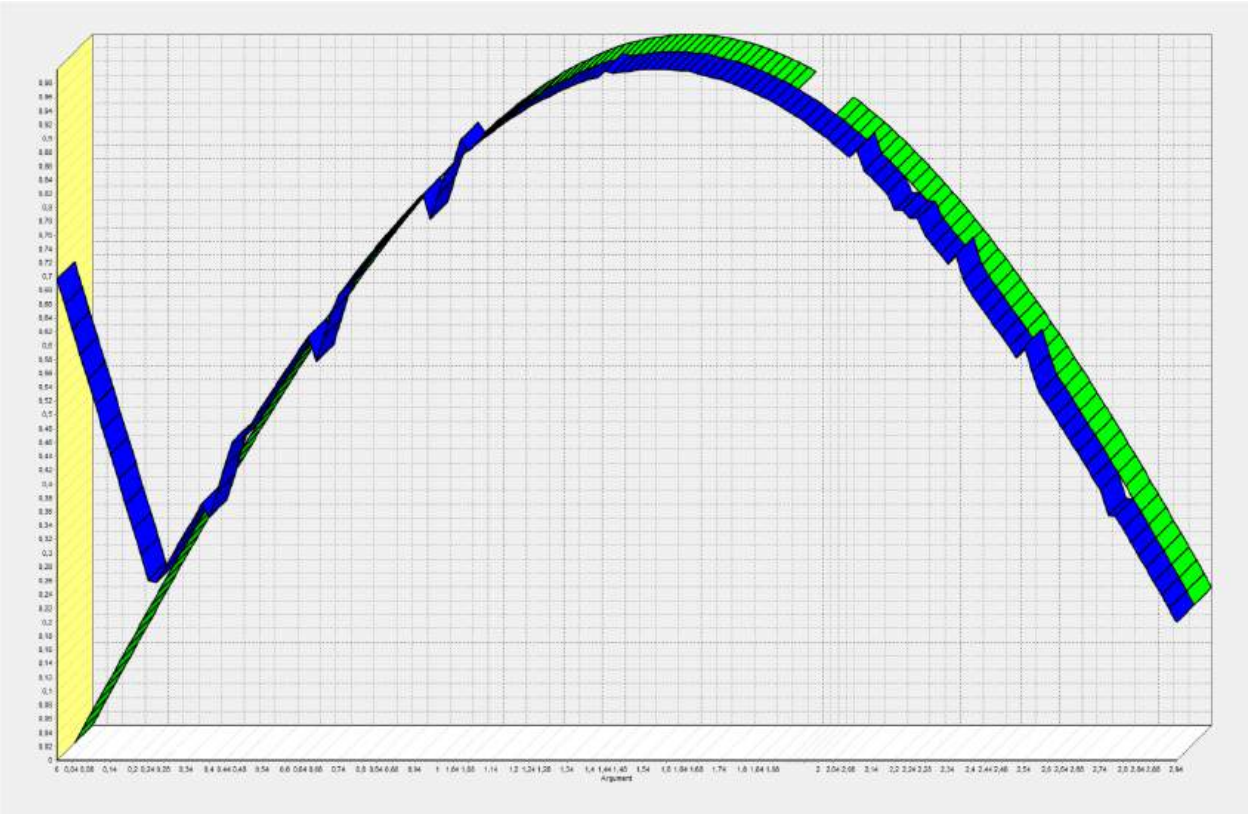
Далее удаляем аномалии. В данной версии дедуктора это делается через редактирование выбросов и экстремальных значений в мастере обработки



Далее выбираем нужный столбец в нашем случаее аномалии выбираем метод обработки – сглаживать и степень малая

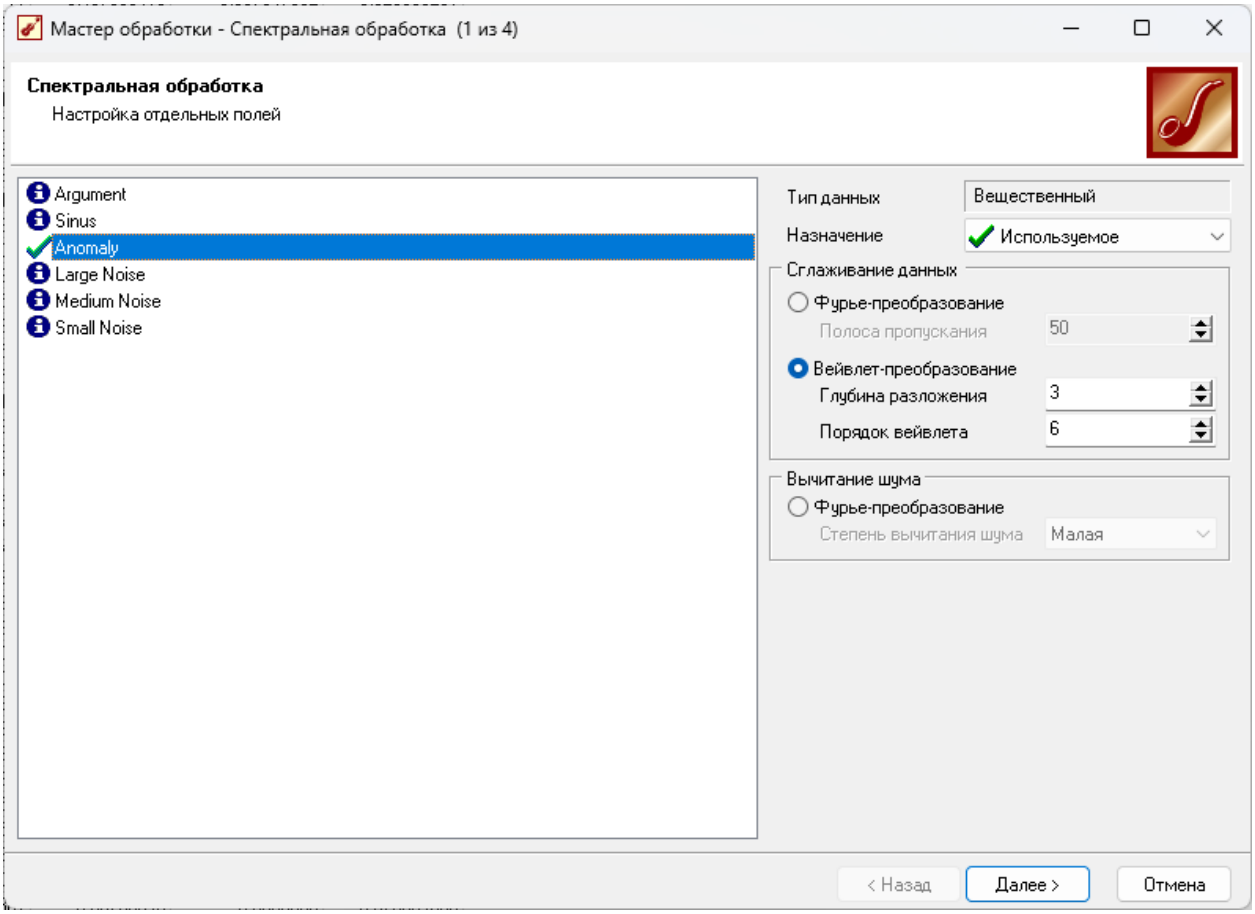


После чего получаем вот такую диаграмму со сглаженными экстремальными значениями по столбцу аномалии(синий)

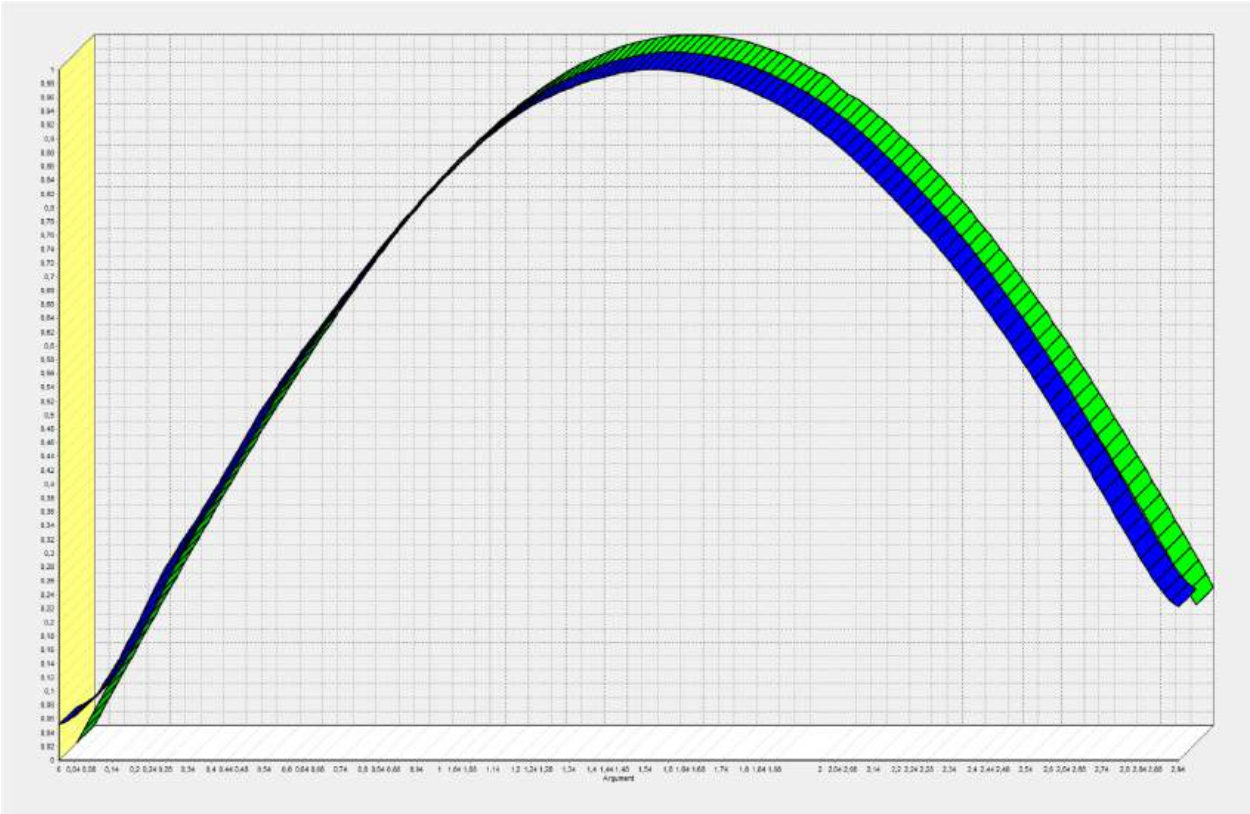


Далее необходимо удалить сгладить данные для аномалий. В данной версии дедуктора используется метод спектральной обработки в мастере обработки.

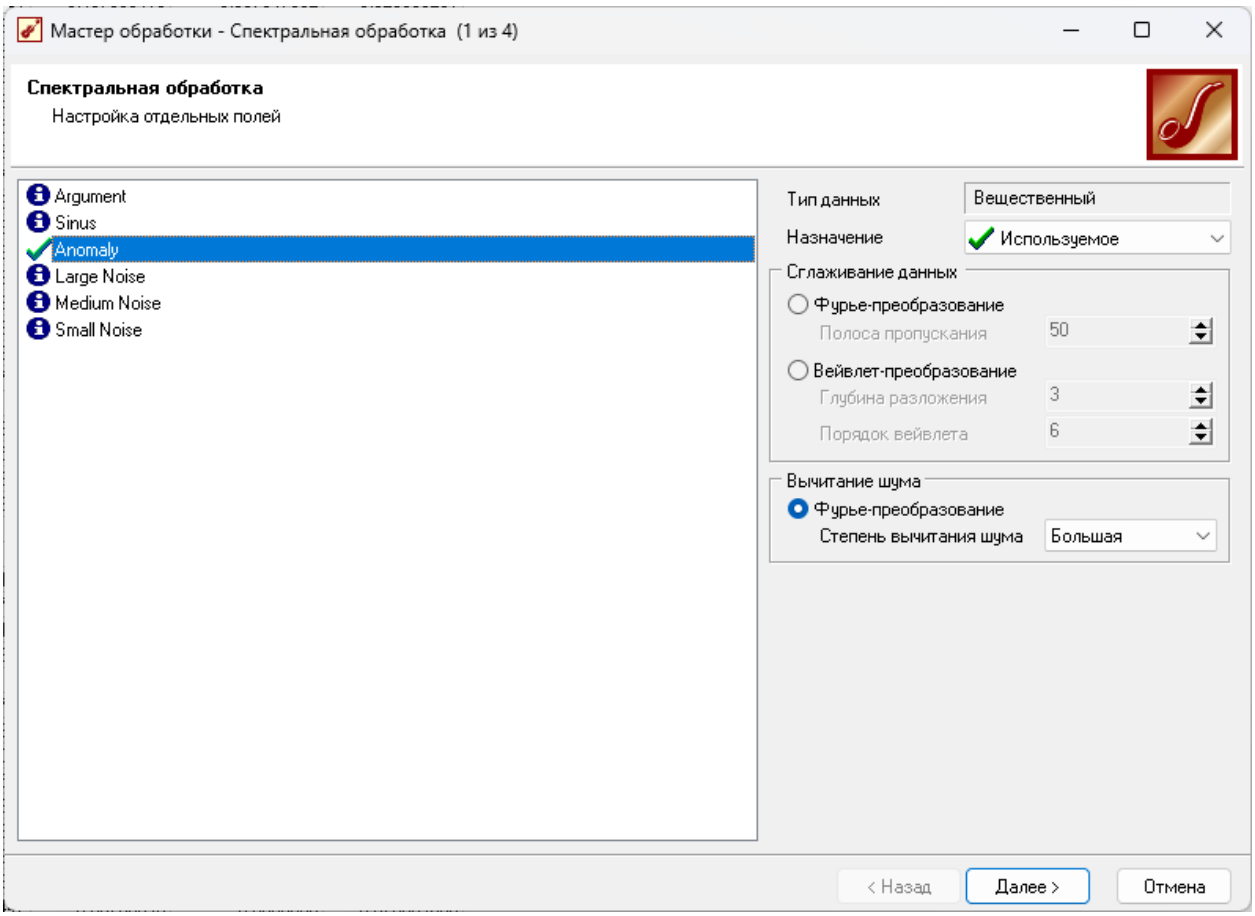
Выбираем нужный нам столбец и выбираем вейвлет-преобразование.



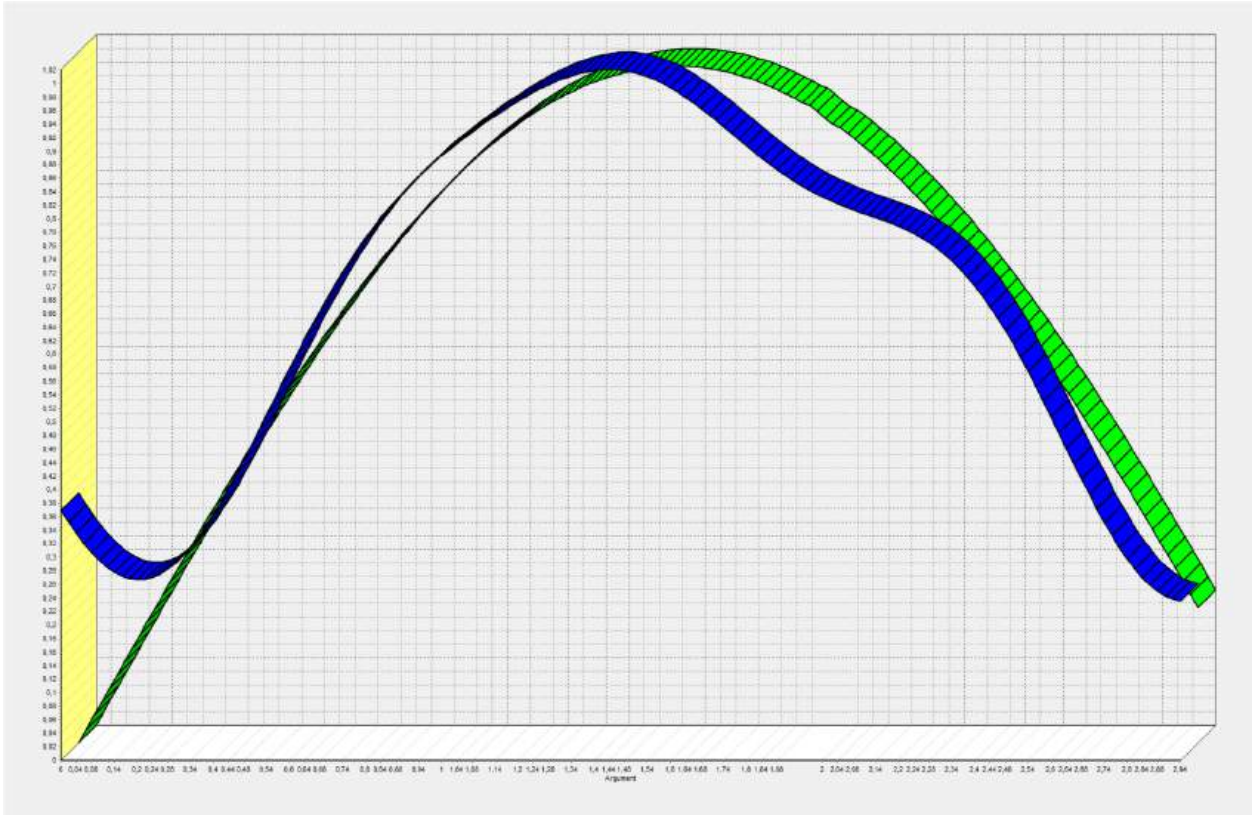
После чего получаем вот такую диаграмму



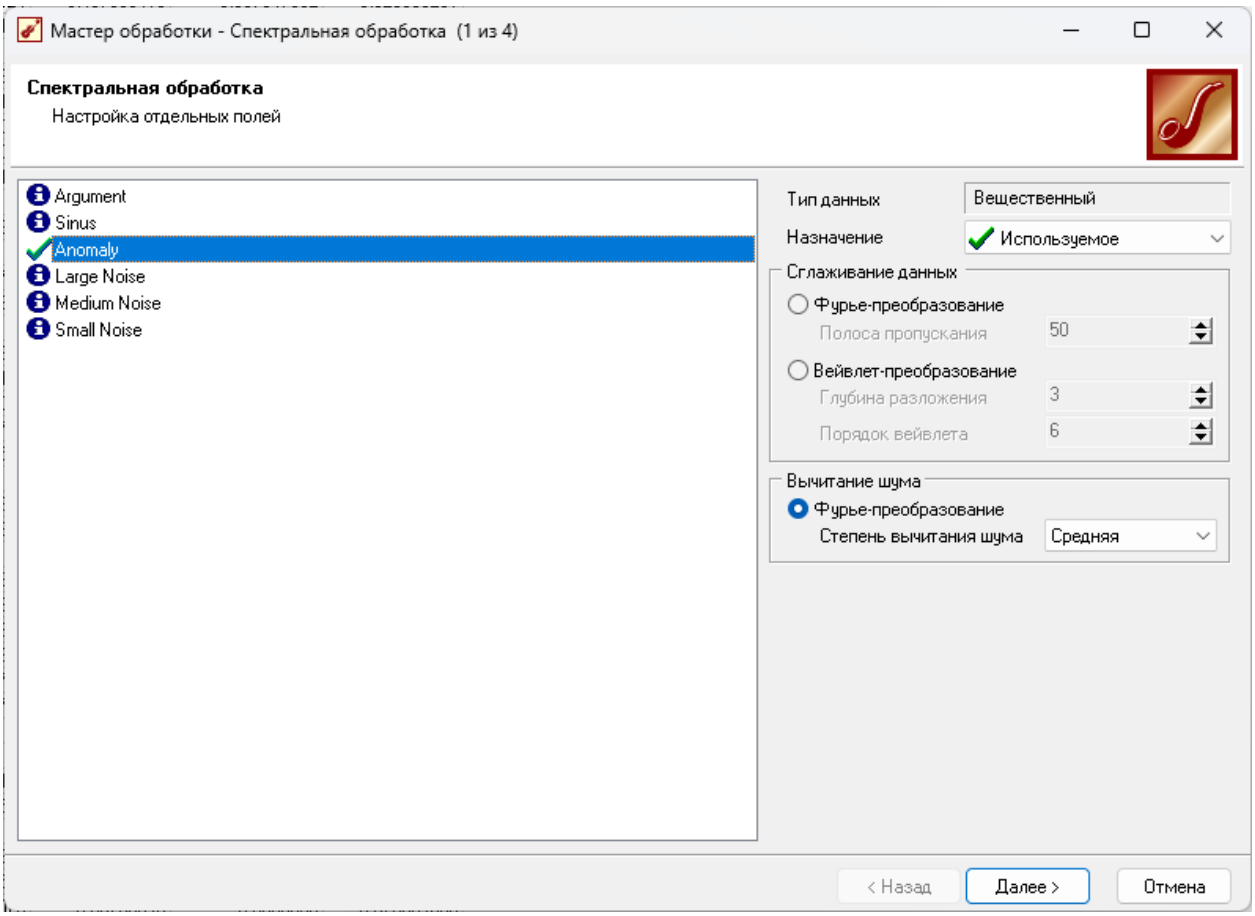
Далее необходимо удалить шумы. Шумы начинаем удалять от больших к меньшим. В данной версии дедуктора используется метод спектральной обработки в мастере обработки. Выбираем нужную колонку и метод вычитание шума, фурье-преобразование со степенью Большая.



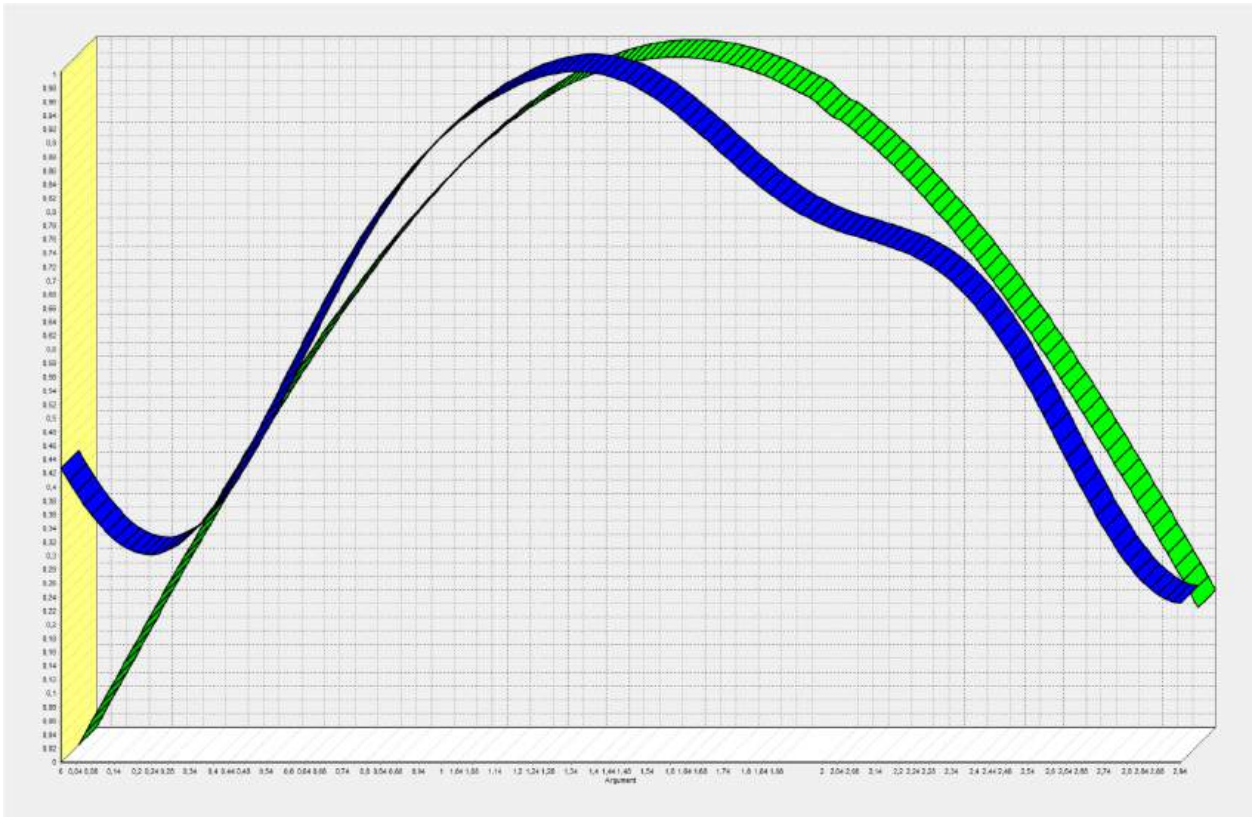
Получаем вот такую диаграмму



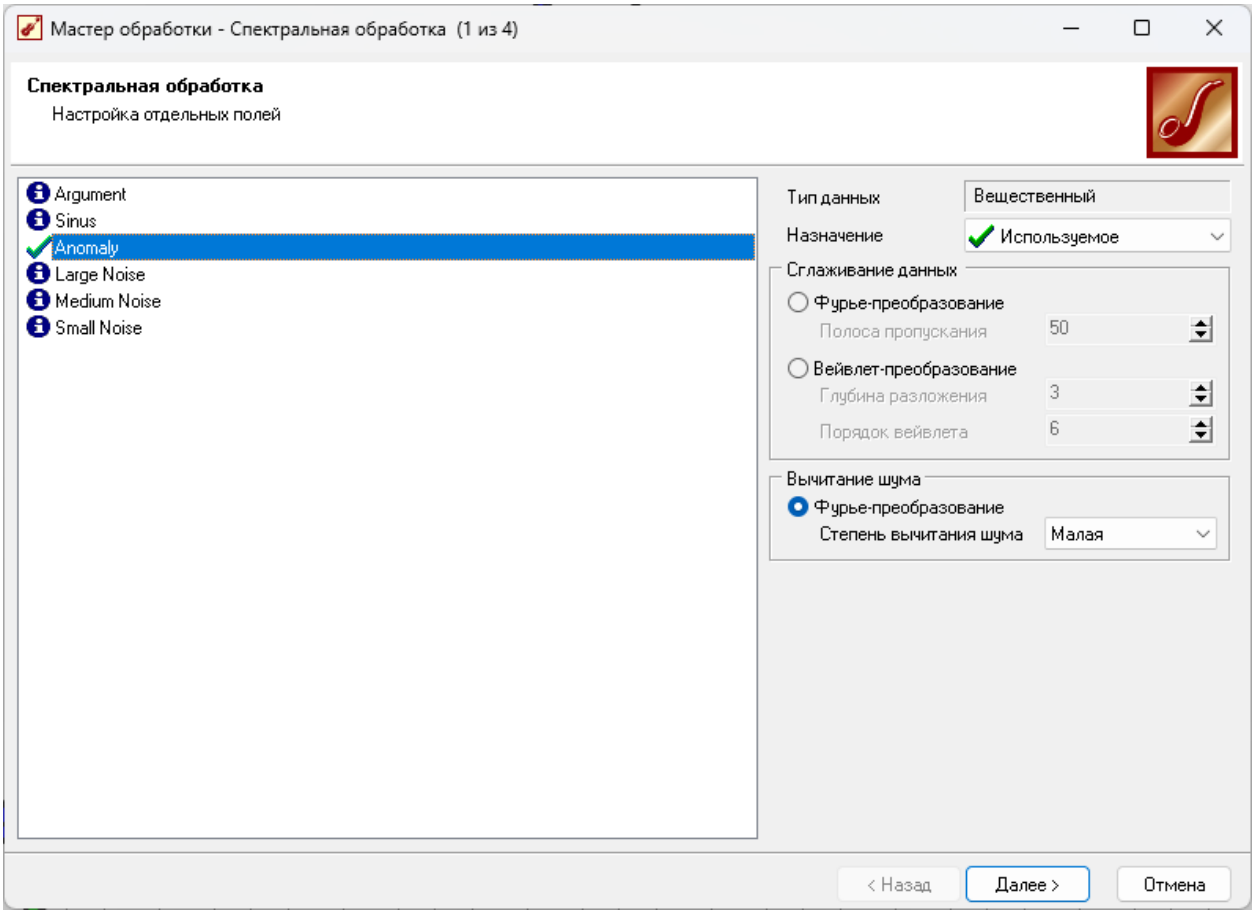
Дальше проделываем те же самые действия только ставим степень средняя.



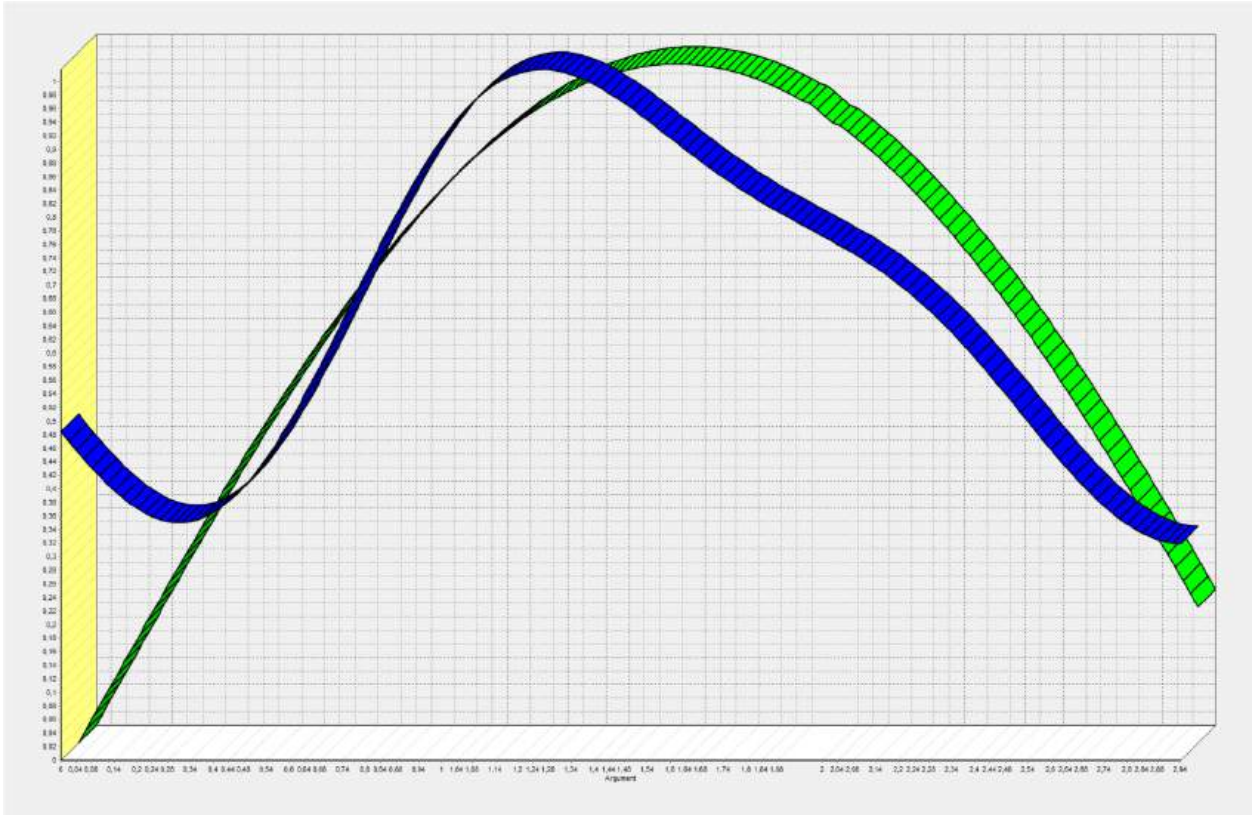
Получаем вот такую диаграмму



Дальше проделываем те же самые действия только ставим степень малая.



Получаем вот такую диаграмму

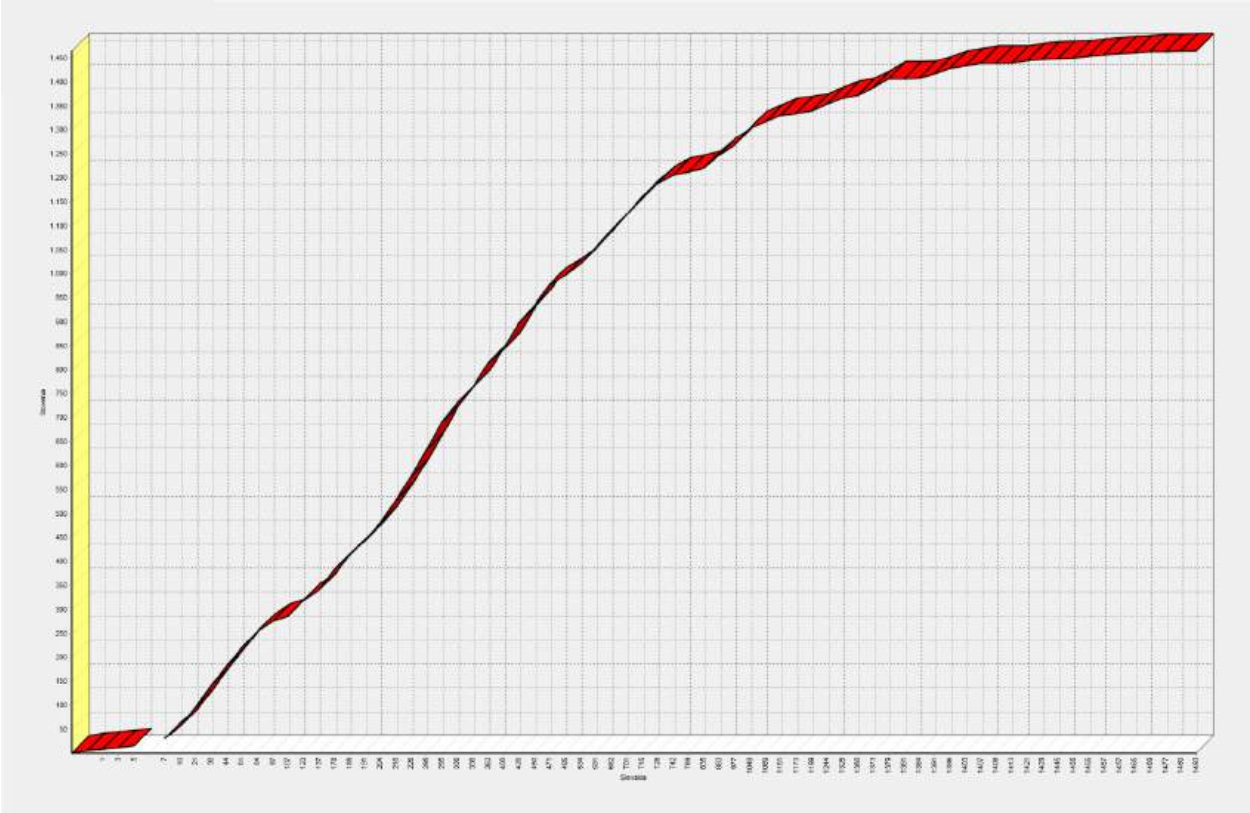


Ход работы 2 задание.

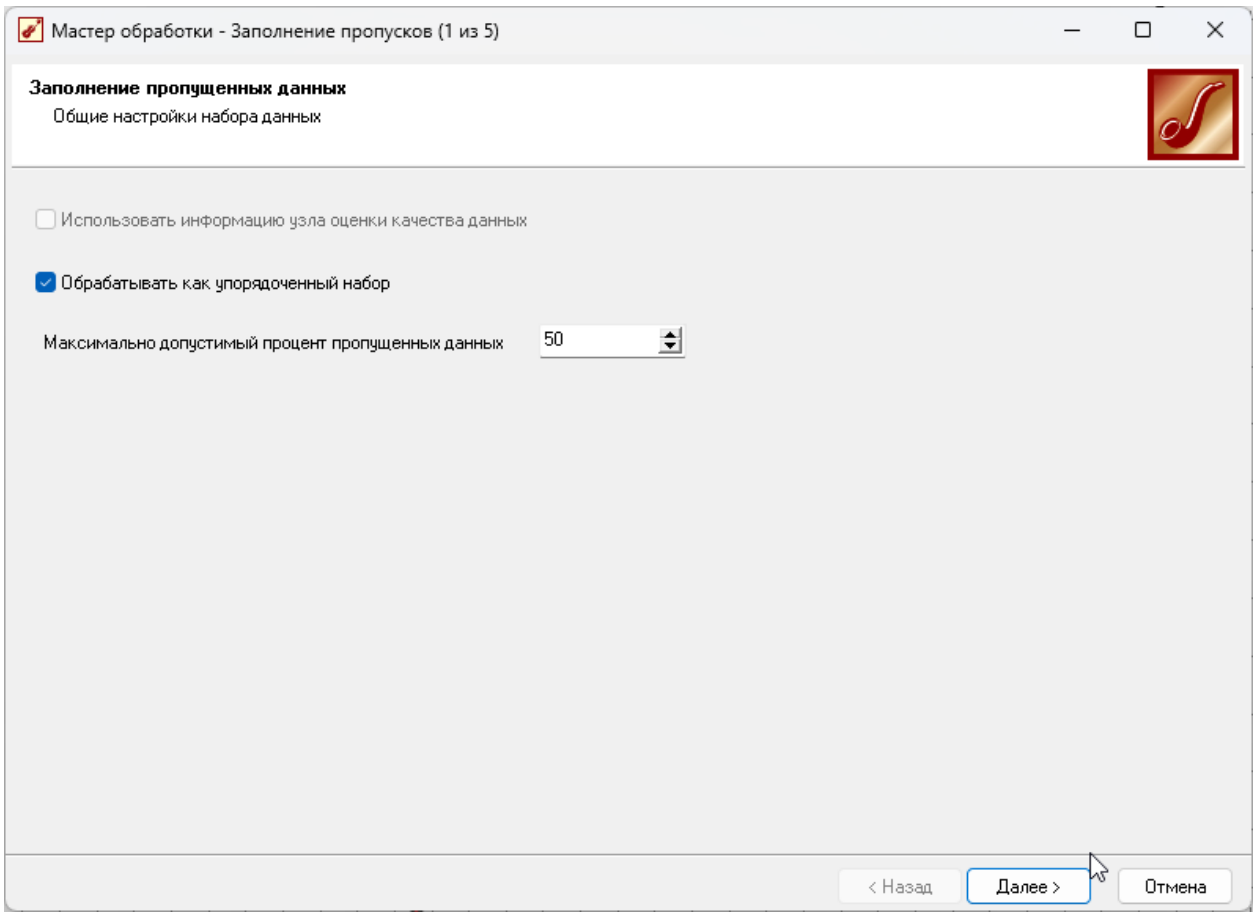
Для начала необходимо выбрать данные из excel таблицы по своему варианту и перенести их в txt файл.

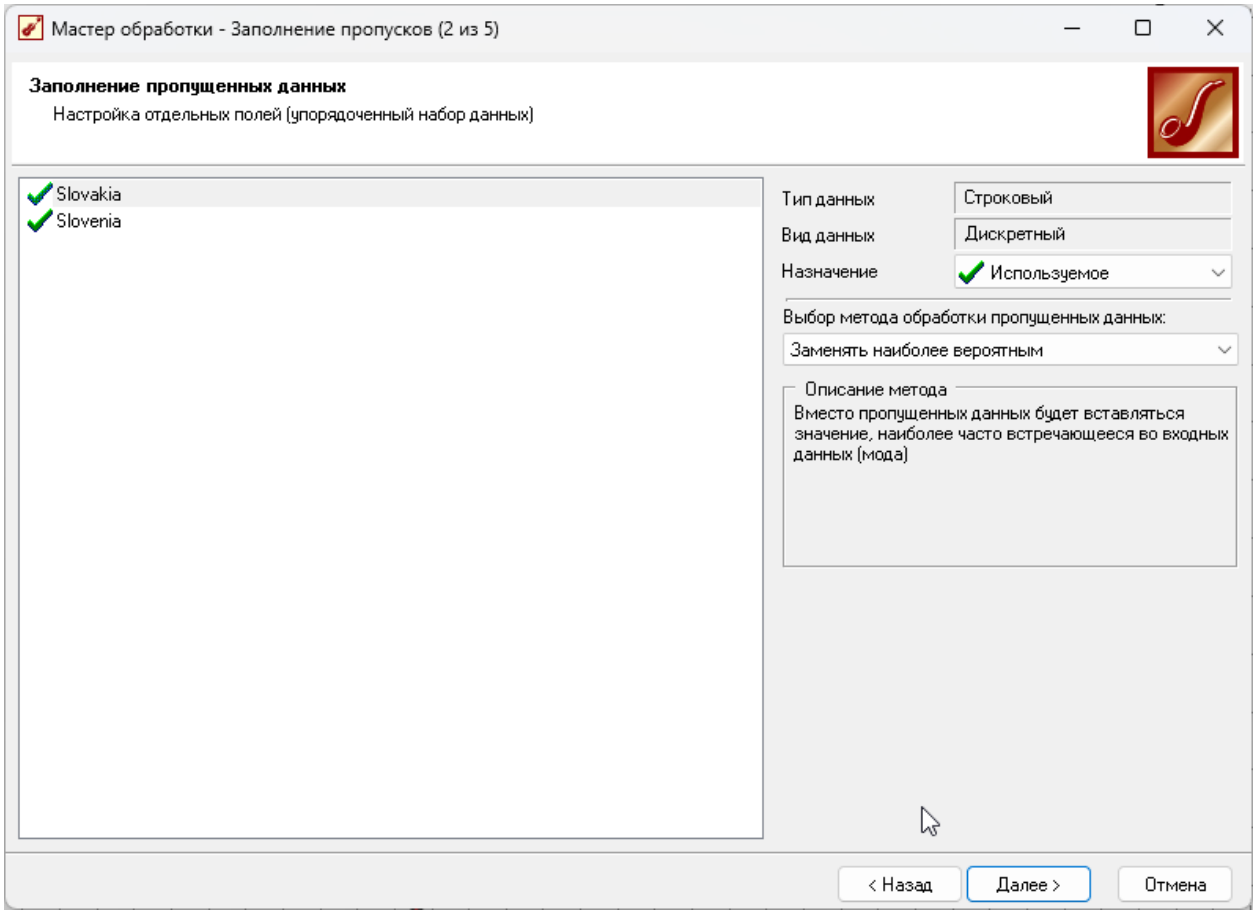
Slovakia	Slovenia
nan	1
nan	6
1	9
3	12
5	16
nan	nan
7	31
10	57
21	96
30	141
44	181
61	219
84	253
97	275
107	286
123	319
137	341
178	383
185	414
191	442
204	480
216	528
226	577
295	632
295	691
336	730
336	763
363	814
400	841
426	897
450	934
471	977
485	997
534	1021
581	1055
682	1091
701	1124
715	1160
728	1188
742	1205
769	1212
835	1220
863	1248
977	1268
1049	1304
1089	1317
1161	1330
1173	1335
1199	1340
1244	1353
1325	1366
1360	1373
1373	1388

После чего загружаем эти данные в deductor, строим диаграмму и смотрим на пропуски.

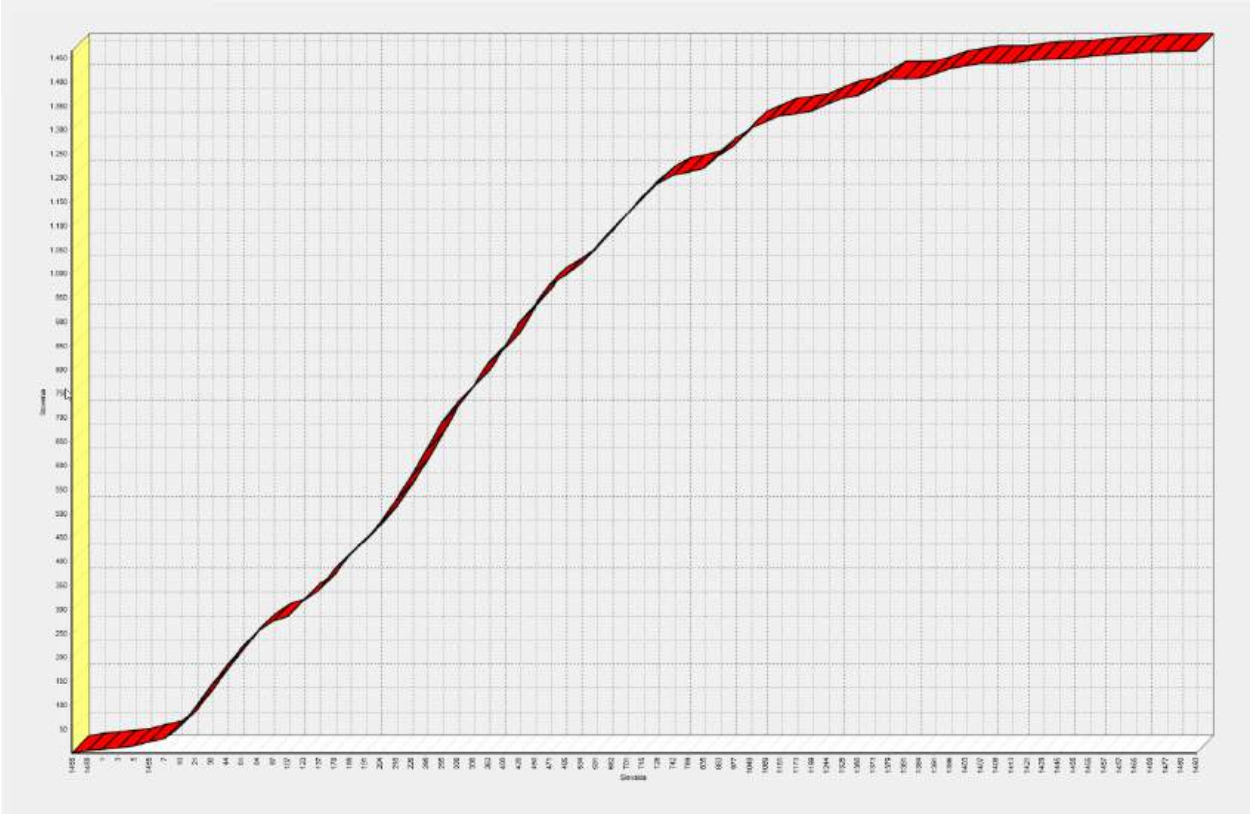


Для заполнения пропусков буду использовать метод заполнения пропущенных данных в мастере обработки. Данный метод используется, потому что по графику видно, что у данных есть некоторый порядок и мы можем заменить на наиболее вероятные значения.

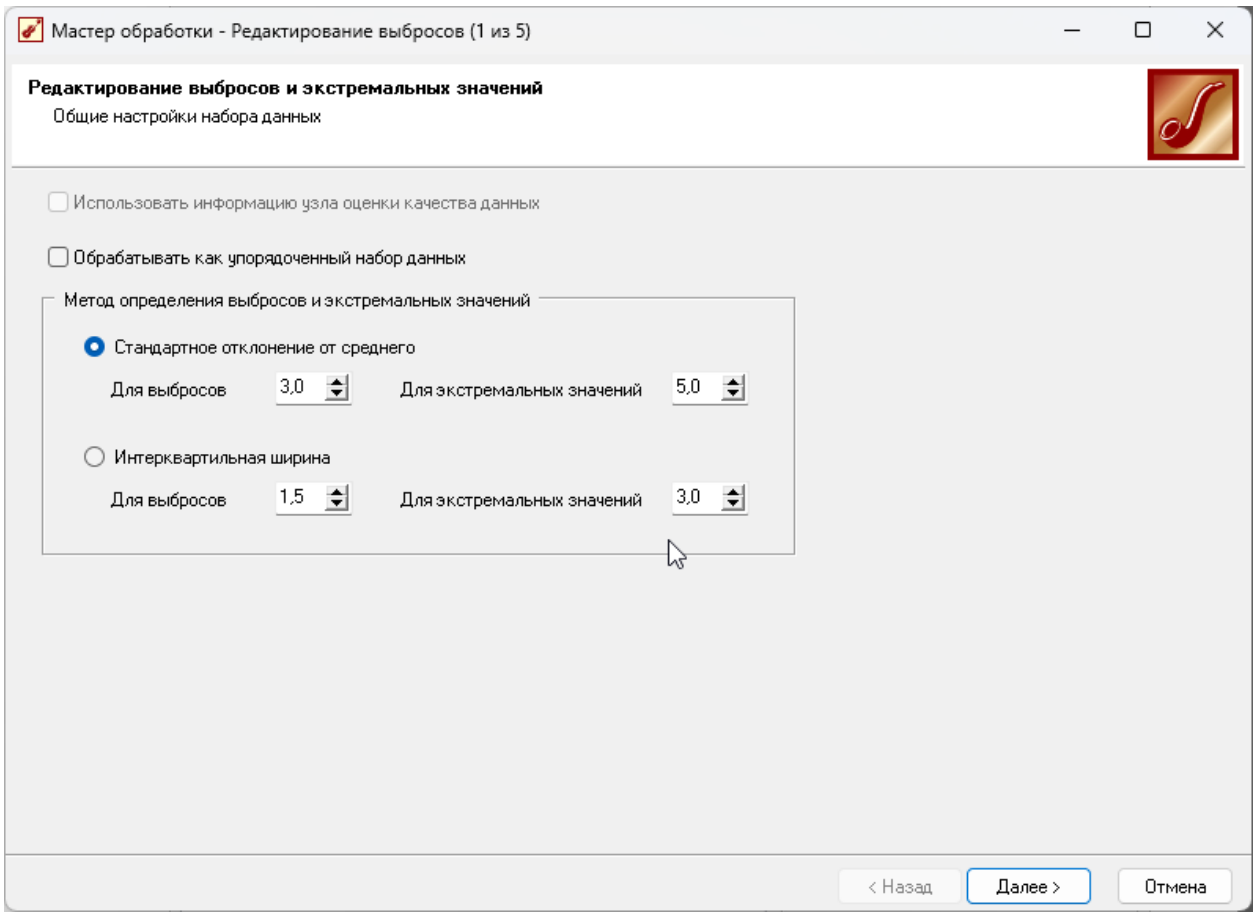

Мастер обработки - Заполнение пропусков (1 из 5)
Заполнение пропущенных данных
Общие настройки набора данных
Использовать информацию узла оценки качества данных
Обрабатывать как упорядоченный набор
Максимально допустимый процент пропущенных данных
50
< Назад
Далее >
Отмена

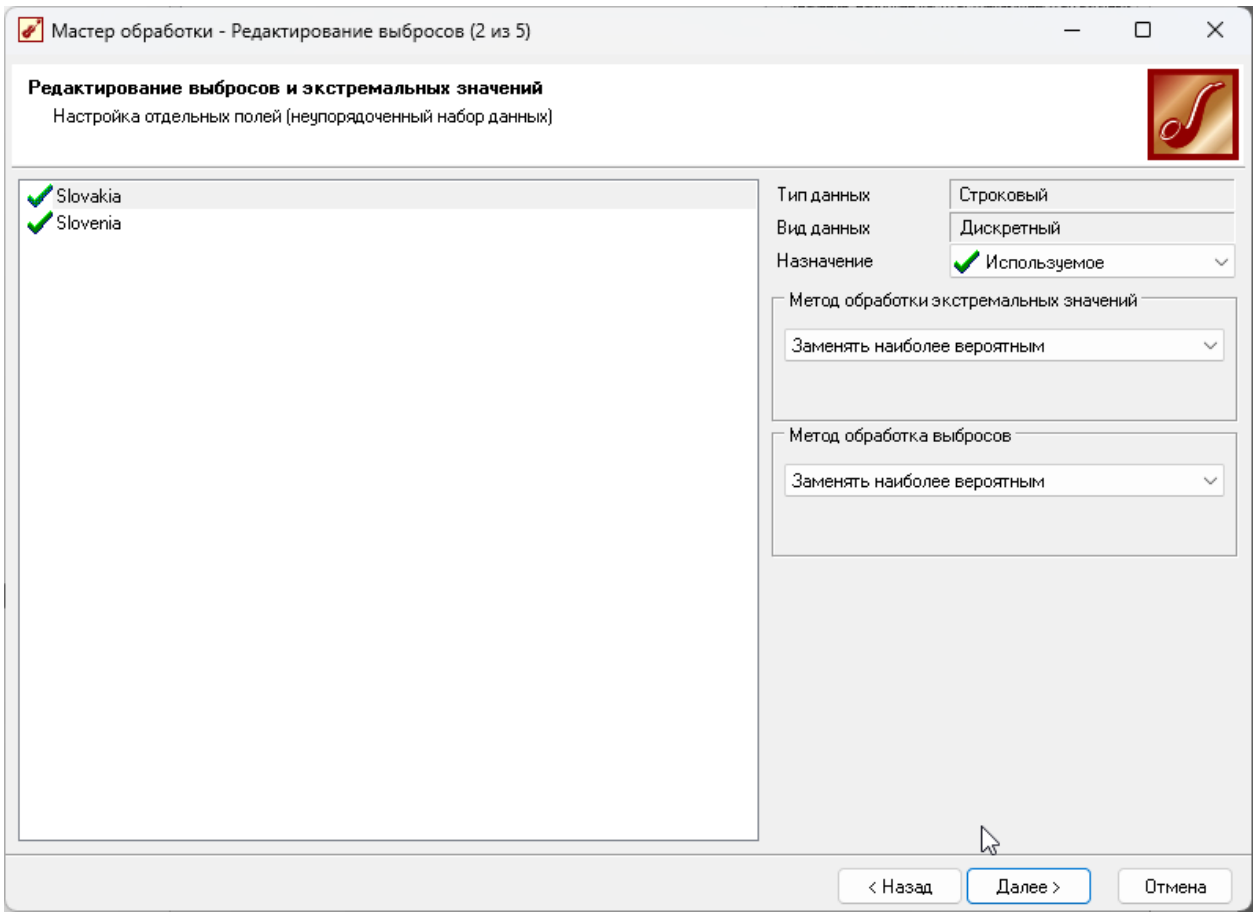

Мастер обработки - Заполнение пропусков (2 из 5)
Заполнение пропущенных данных
Настройка отдельных полей (упорядоченный набор данных)
Slovakia
Slovenia
Тип данных
Строковый
Вид данных
Дискретный
Назначение
Используемое
Выбор метода обработки пропущенных данных:
Заменять наиболее вероятным
Описание метода
Вместо пропущенных данных будет вставляться значение, наиболее часто встречающееся во входных данных (мода)
< Назад
Далее >
Отмена

Получаем обновленную диаграмму, в которой уже нету пропусков

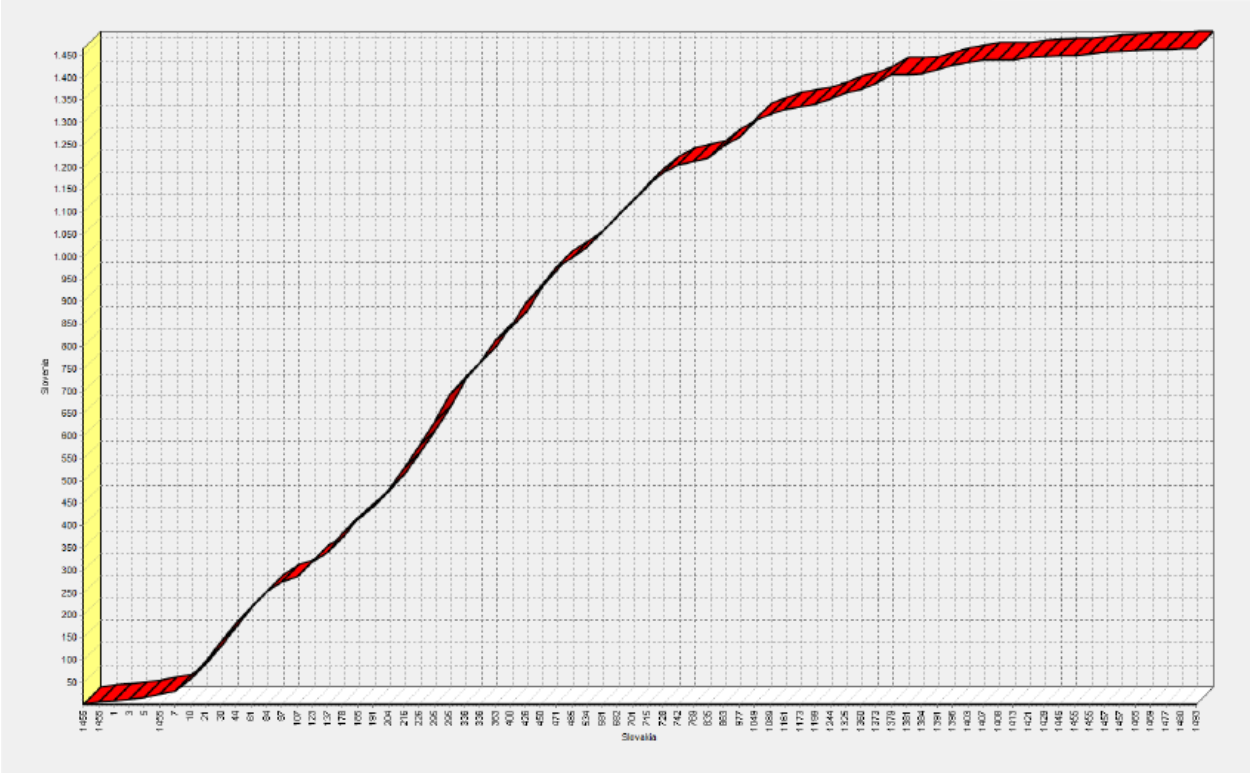


Т.к. значения являются строго упорядоченными можно не редактировать аномальные выбросы. Чтобы подтвердить данную теорию в мастере обработки выберем редактирование выбросов и экстремальных значений в мастере обработке и зададим стандартные параметры с заменой на наиболее вероятное значение.

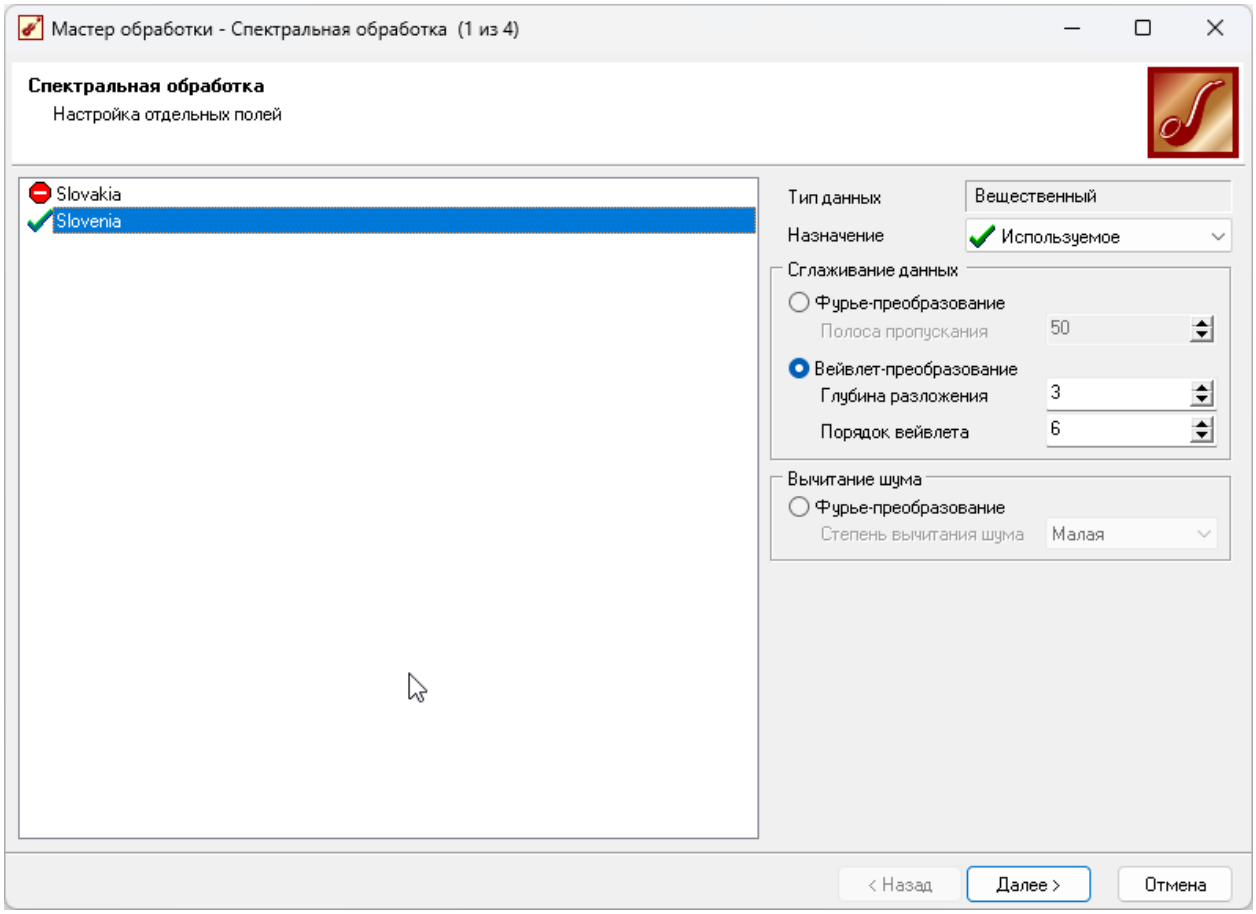

Мастер обработки - Редактирование выбросов (1 из 5)
Редактирование выбросов и экстремальных значений
Общие настройки набора данных
Использовать информацию узла оценки качества данных
Обрабатывать как упорядоченный набор данных
Метод определения выбросов и экстремальных значений
Стандартное отклонение от среднего
Для выбросов
3,0
Для экстремальных значений
5,0
Интерквартильная ширина
Для выбросов
1,5
Для экстремальных значений
3,0
< Назад
Далее >
Отмена


Мастер обработки - Редактирование выбросов (2 из 5)
Редактирование выбросов и экстремальных значений
Настройка отдельных полей (неупорядоченный набор данных)
Slovakia
Slovenia
Тип данных
Строковый
Вид данных
Дискретный
Назначение
Используемое
Метод обработки экстремальных значений
Заменять наиболее вероятным
Метод обработка выбросов
Заменять наиболее вероятным
< Назад
Далее >
Отмена

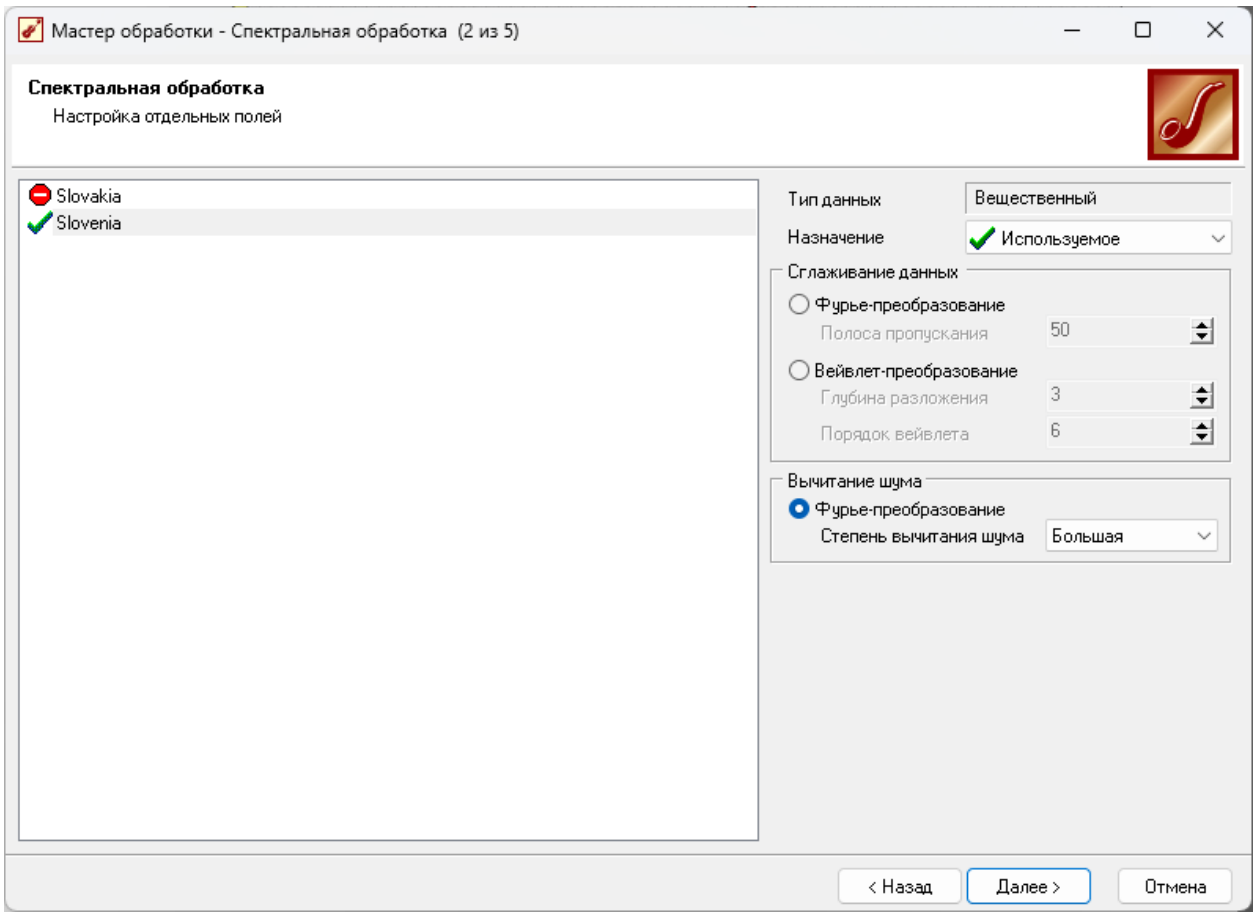
После построения диаграммы мы видим, что она никак не изменилась, т.е. у нас нету выбросов аномалий.



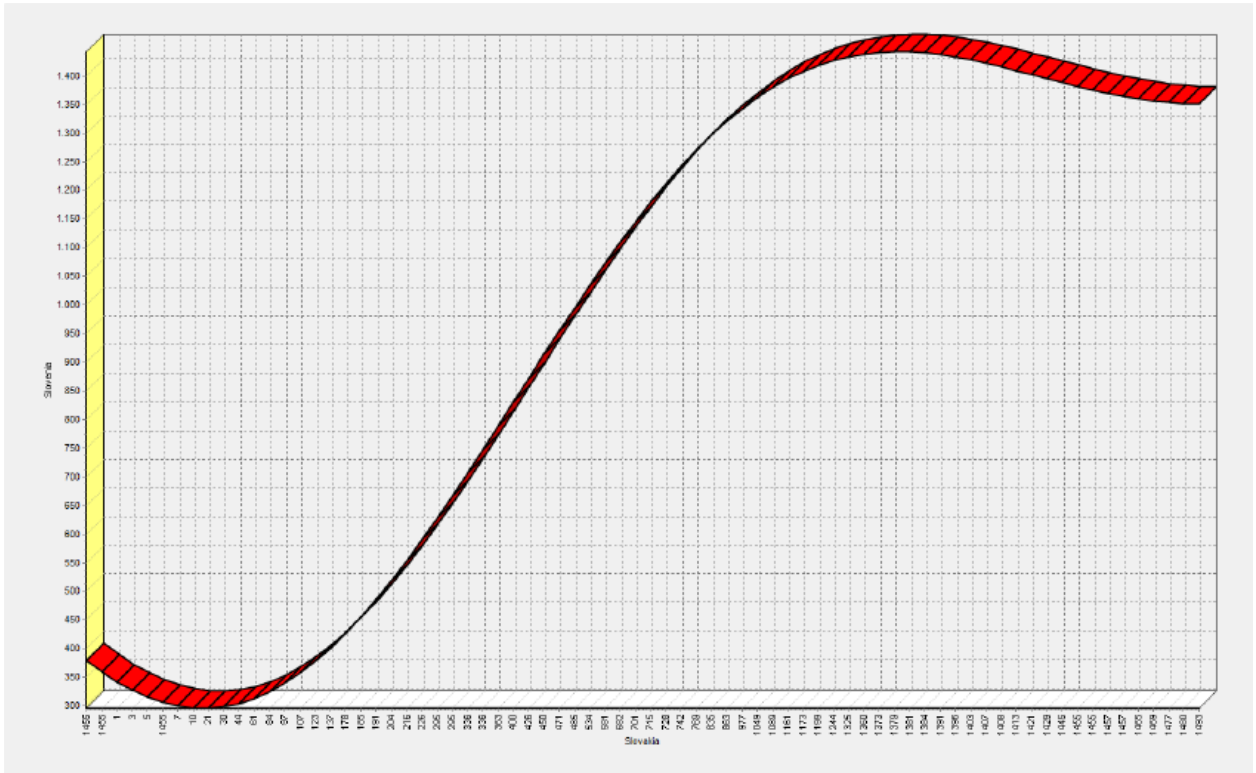
Далее сглаживаем данные с метода спектральной обработки в мастере обработки, с помощью вейвлет преобразования. Данный метод позволяет эффективно отделить сигнал от шума и так же он сохраняет важные составляющие – резкие изменения, выбросы и пр.



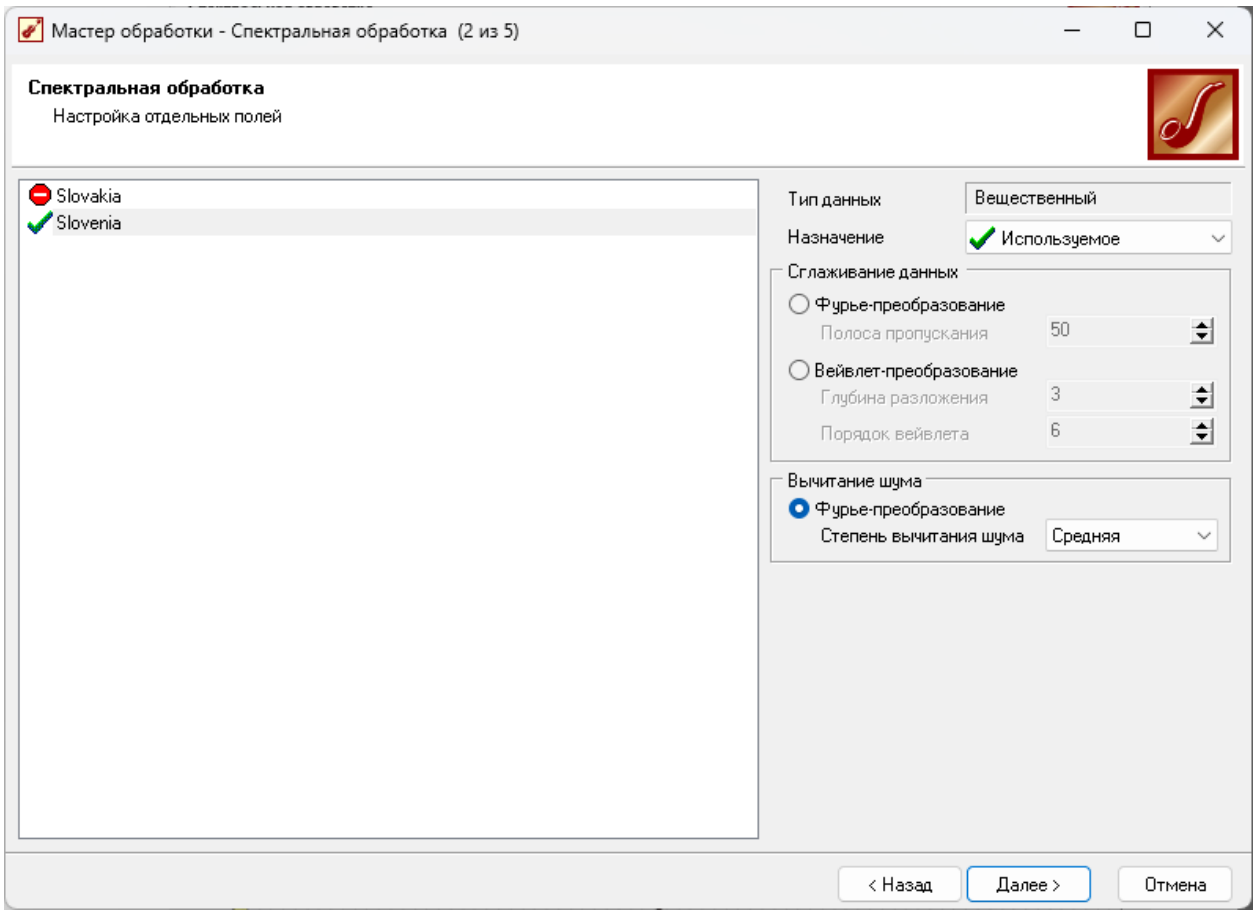
Далее удаляем шумы с больших до меньших. С помощью метода спектральная обработка в мастере обработке – вычитание шума. Для начала ставим степень Большая.



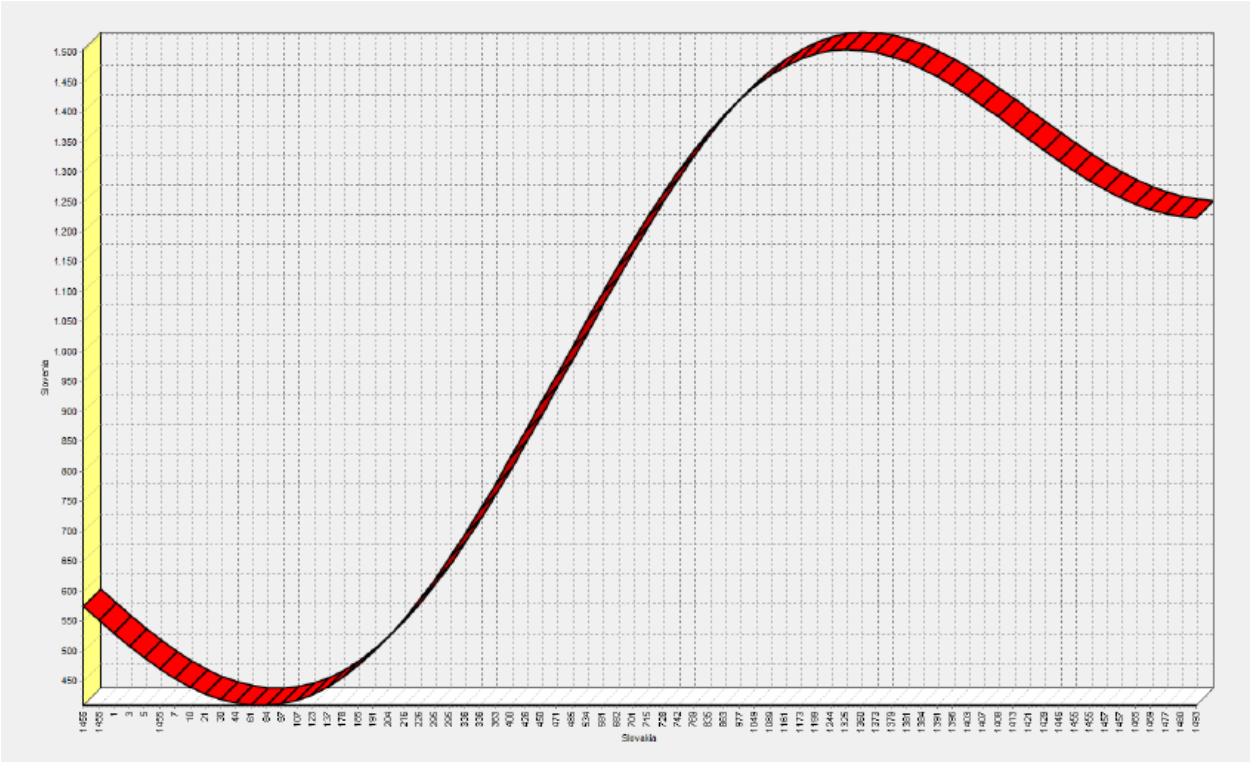
Получаем вот такую диаграмму



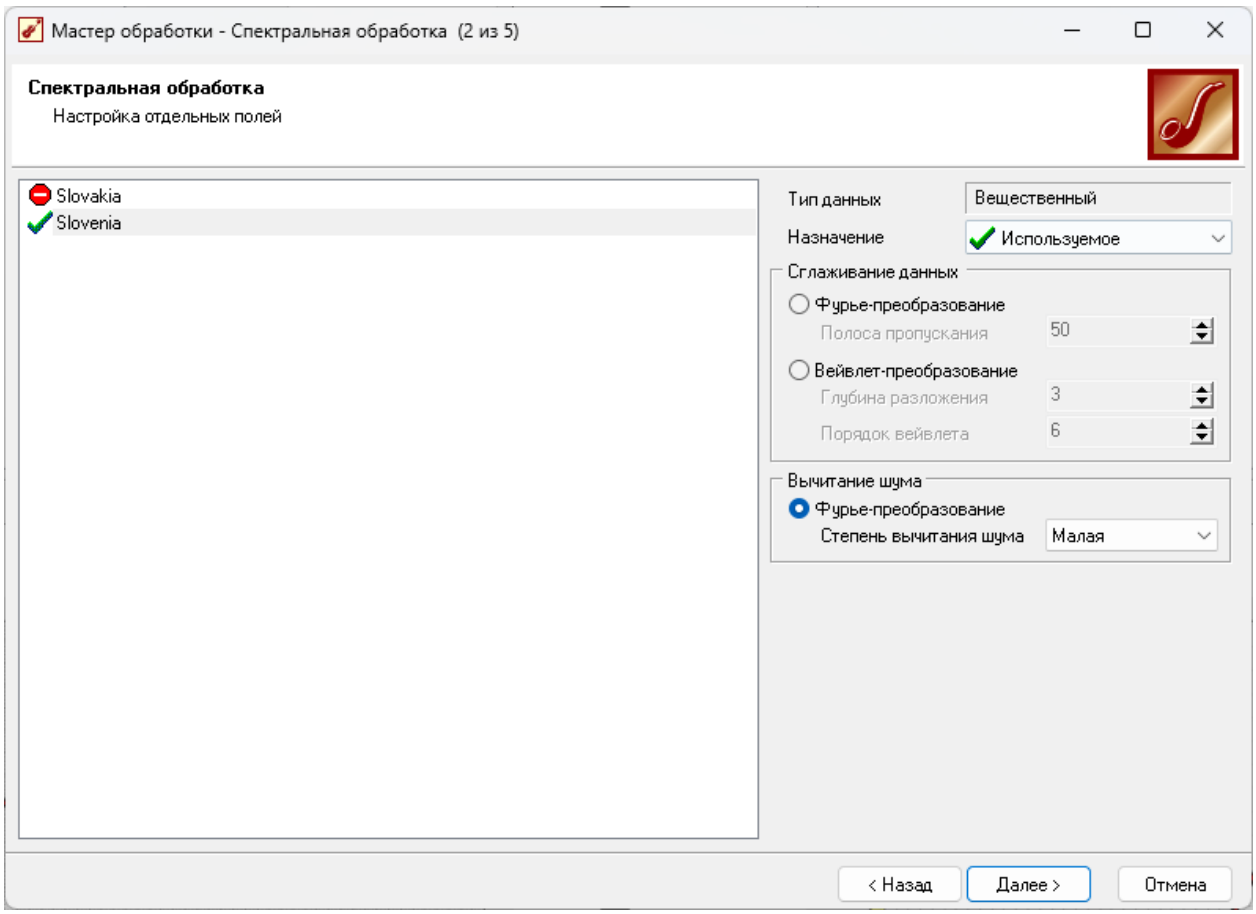
Теперь ставим степень Средняя.



Получаем вот такую диаграмму.



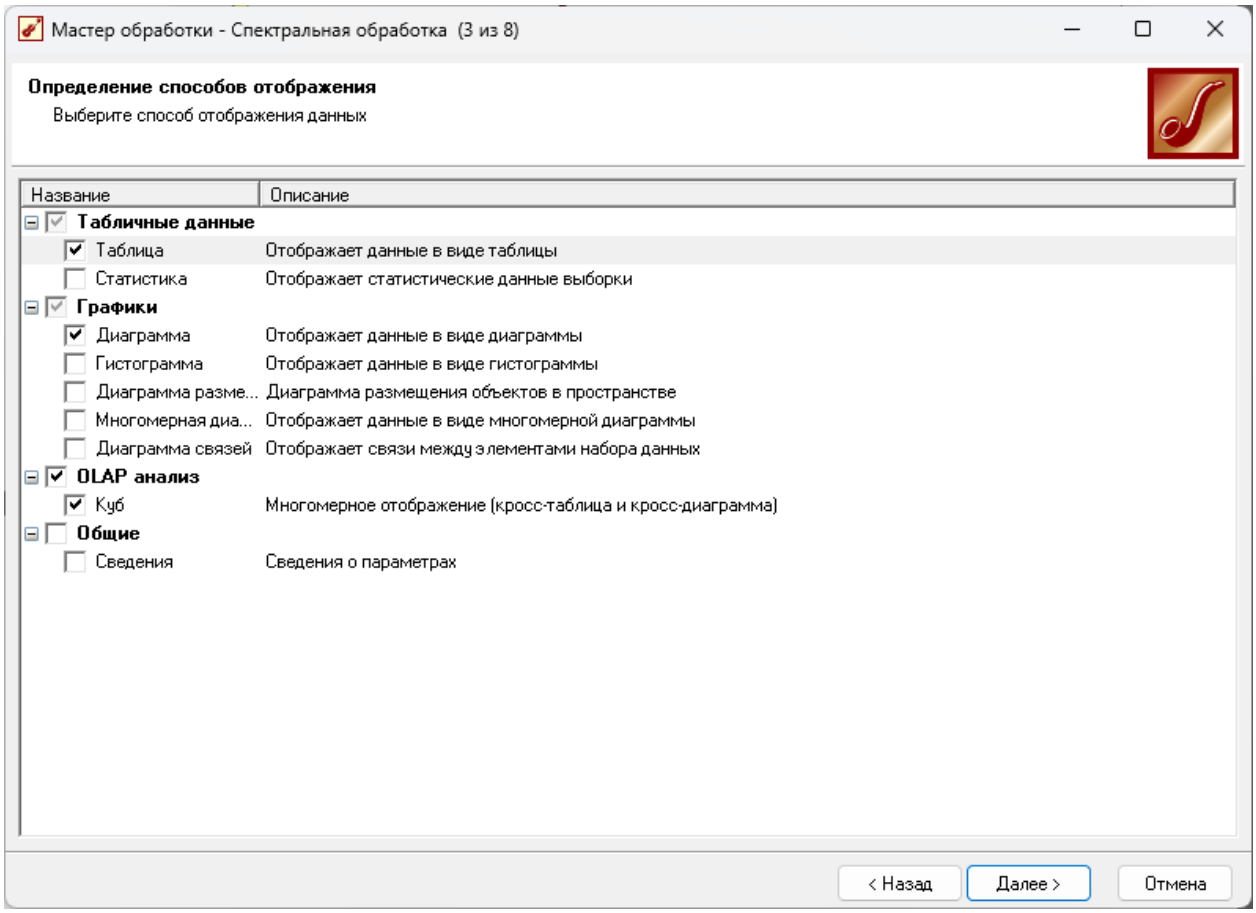
Теперь ставим степень Малая.



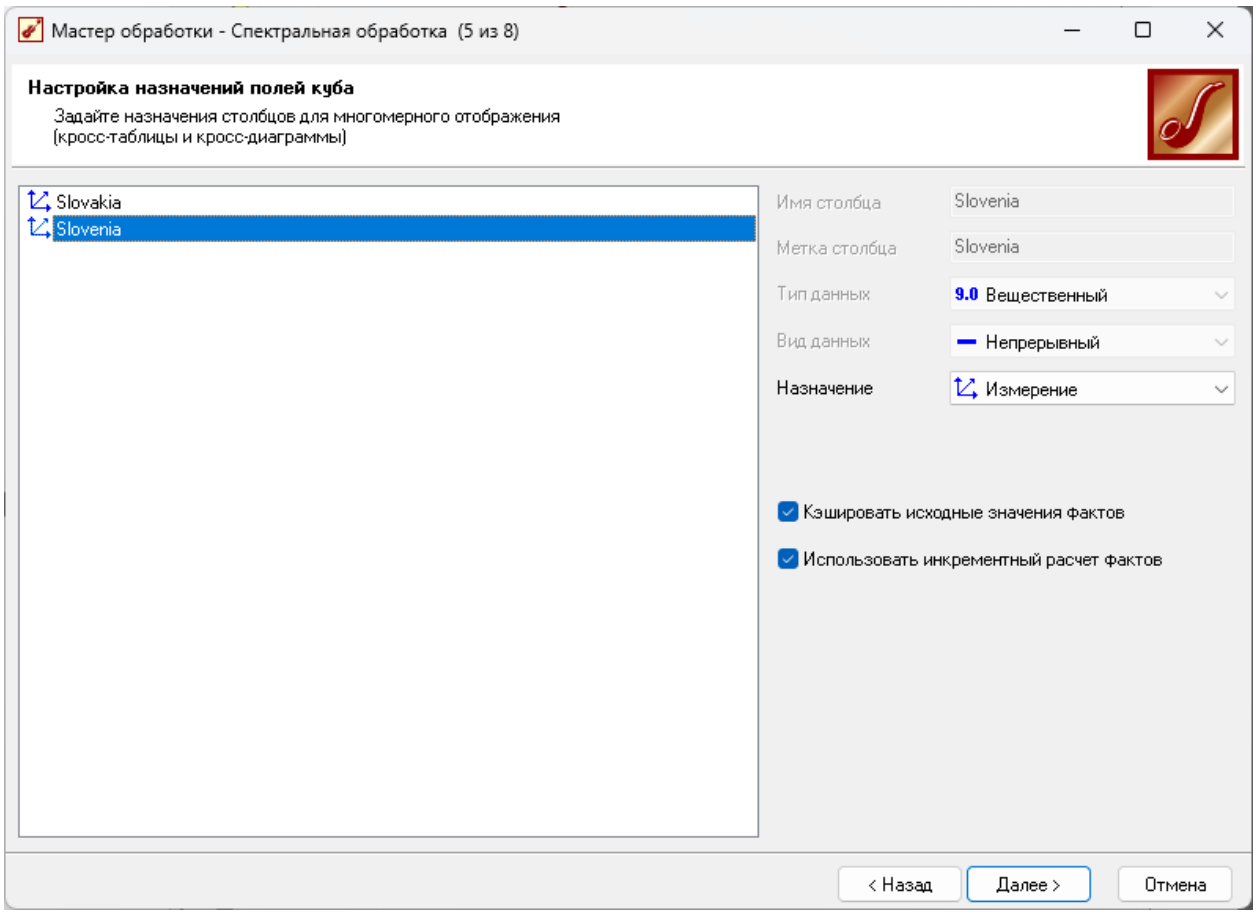
Получаем вот такую диаграмму.

В результате проделанной работы мы получаем качественные отфильтрованные данные.

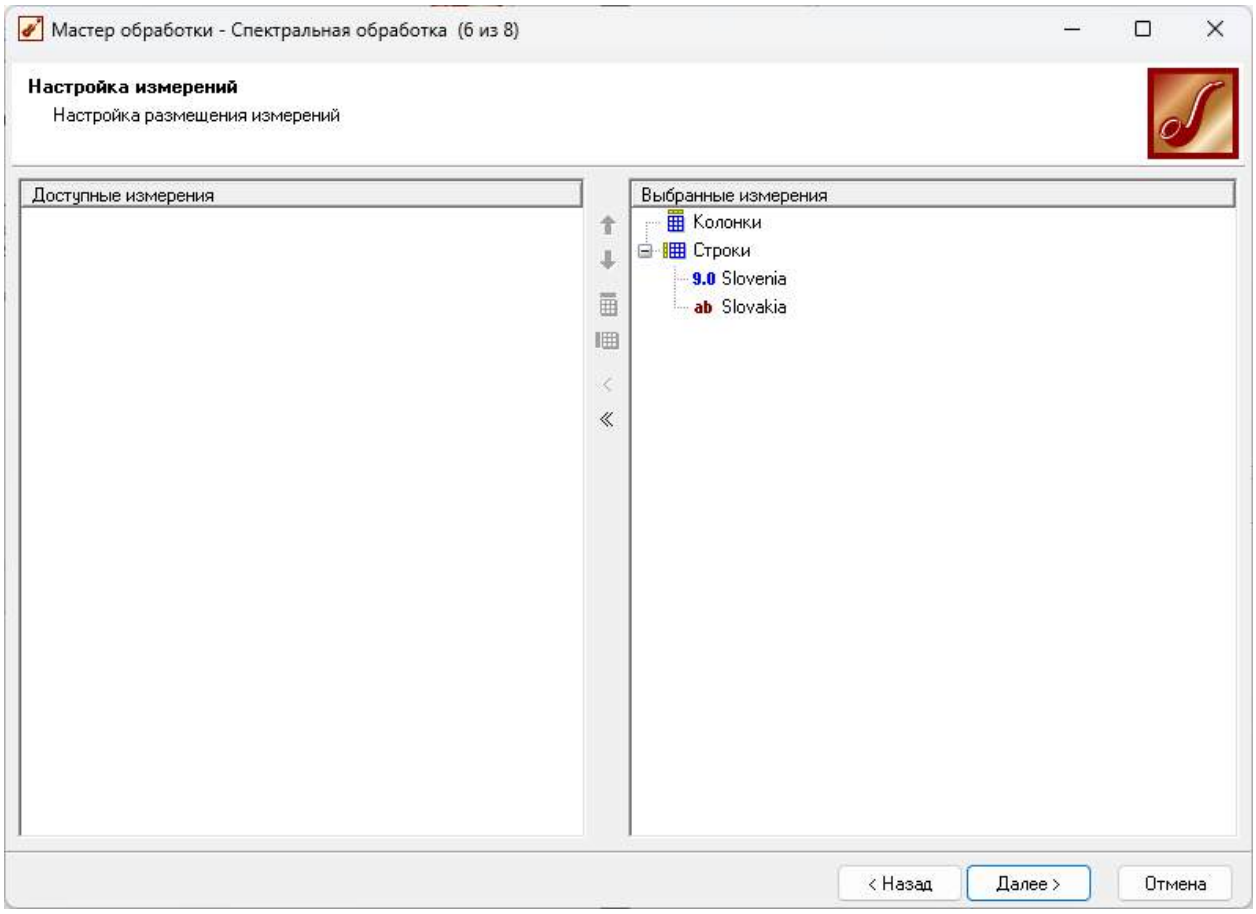
Во время построения последней диаграммы, нужно указать OLAP анализ – куб.



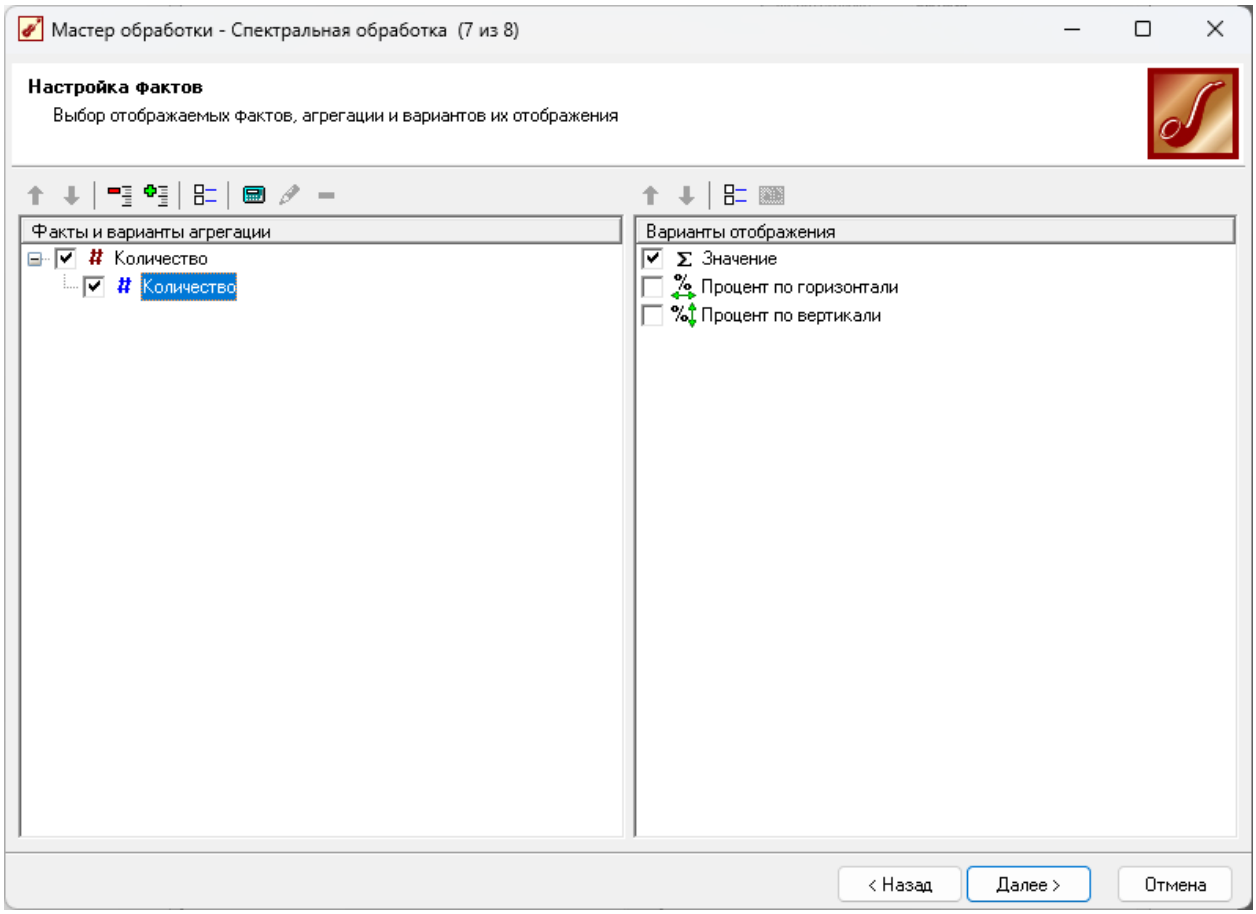
Указываем 2 измерения



Ставим эти измерения в строки

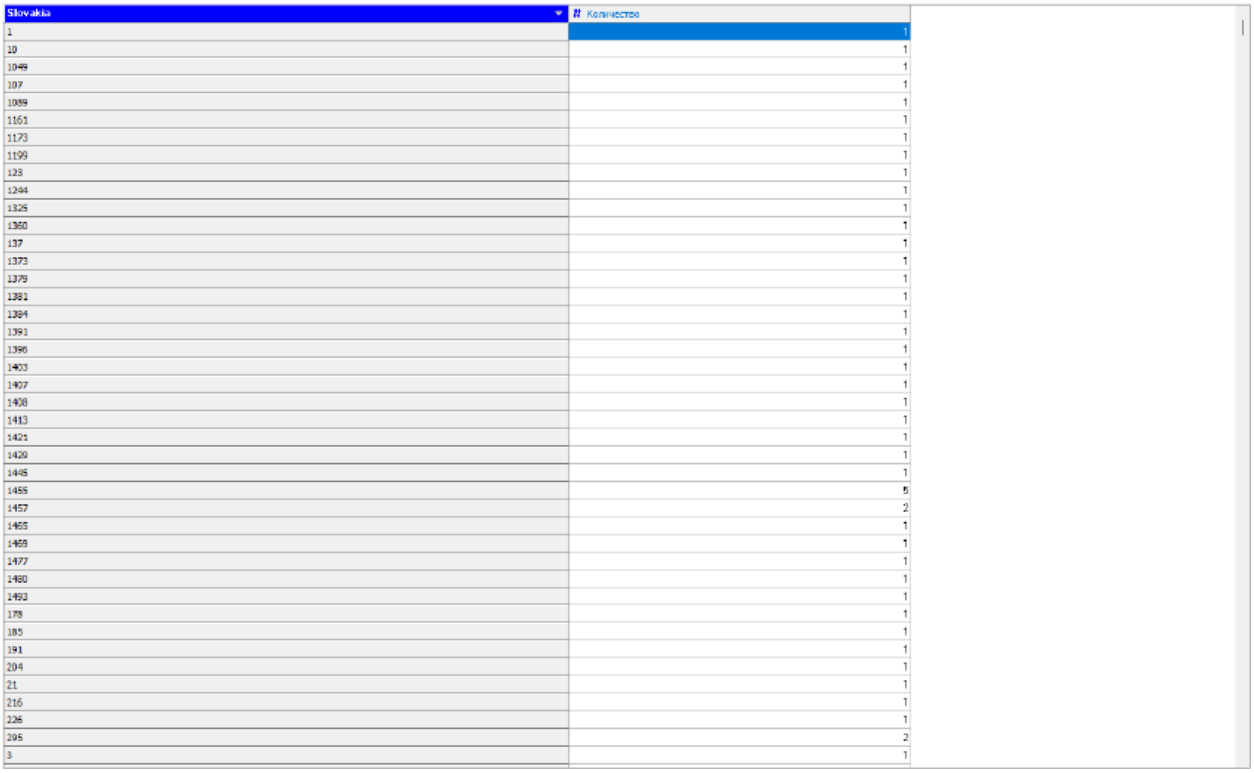


В фактах указываем количество

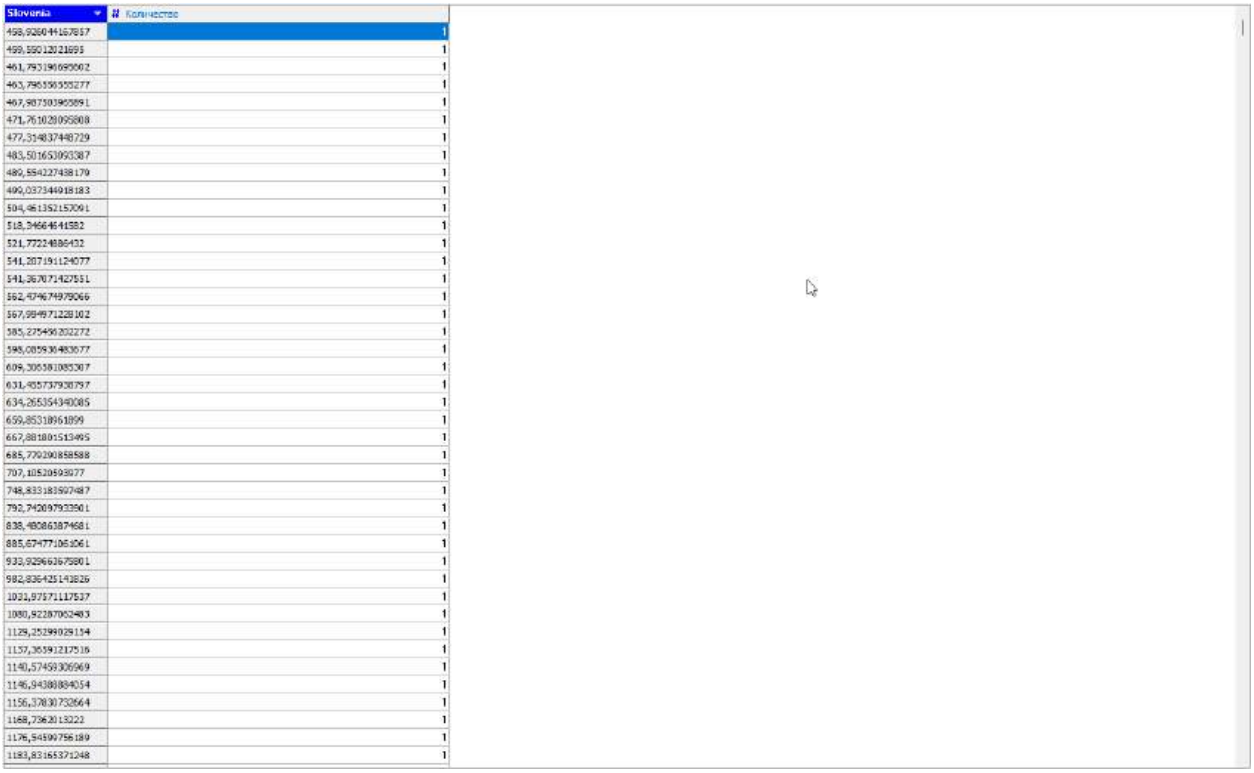


Получаем вот такой куб

Только словакия

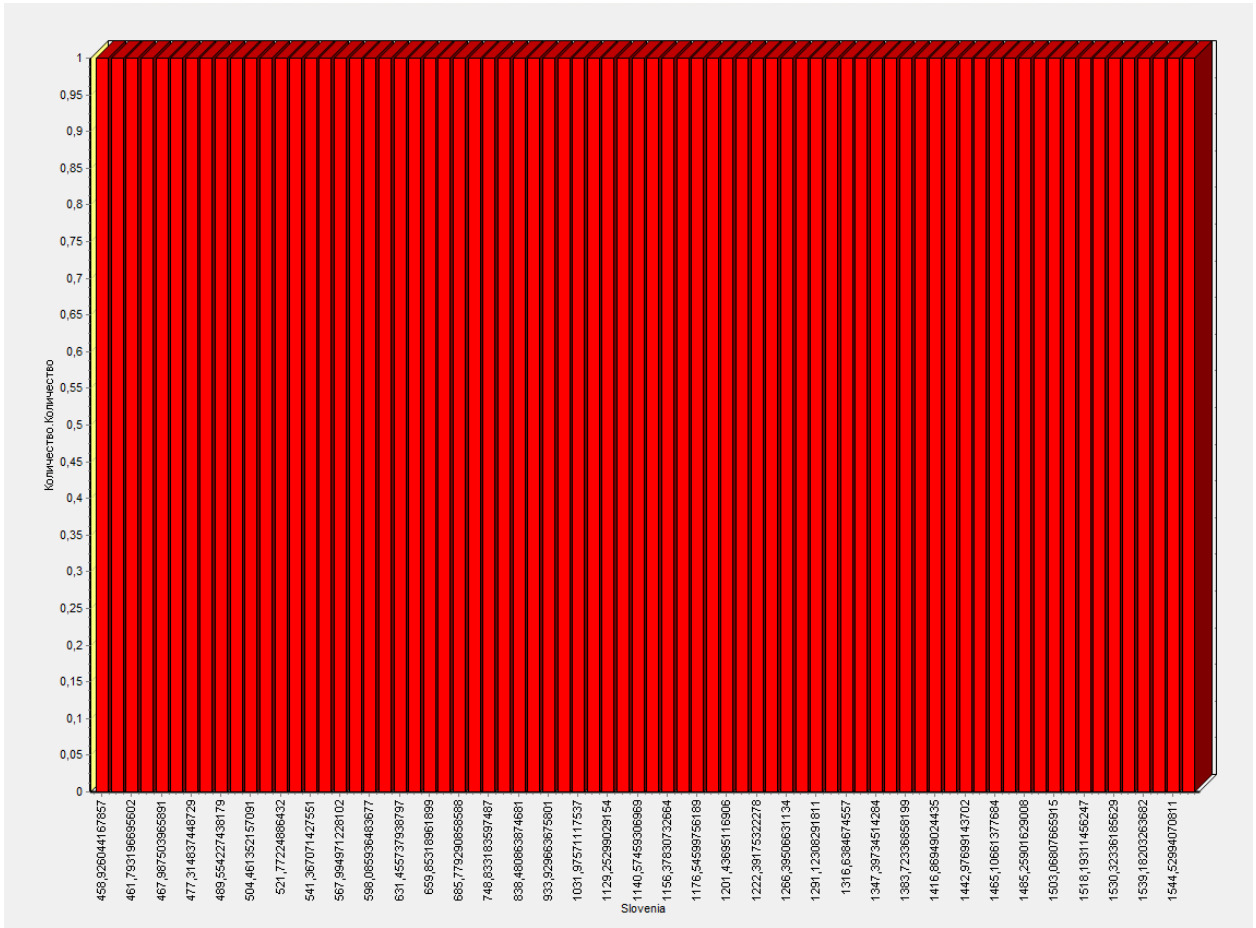


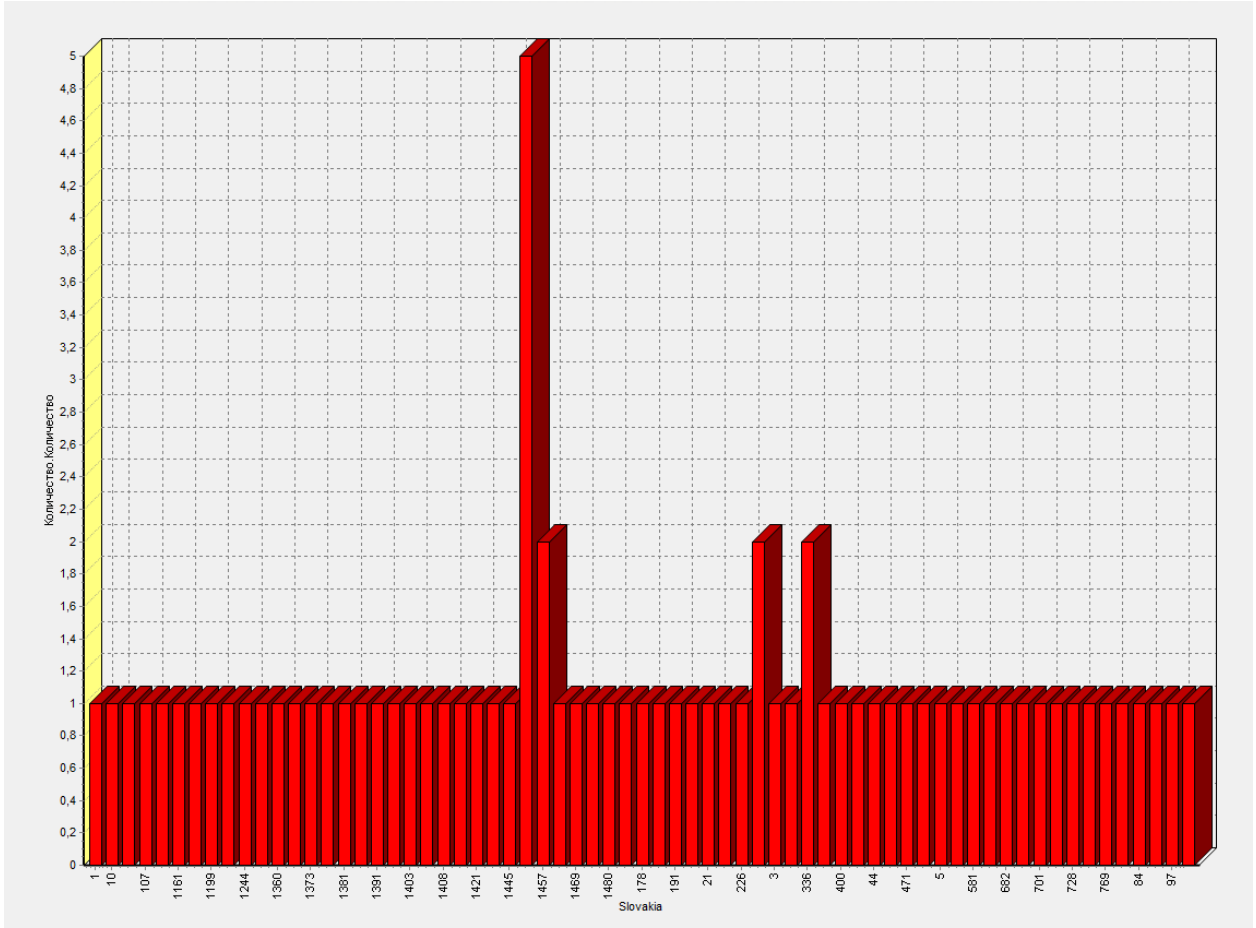
Только Словения



Slovenia	# Количество
458,926044167857	1
459,550120218991	1
461,795196099002	1
463,796556555277	1
467,987503965891	1
471,761028099808	1
477,514837448729	1
483,501653093387	1
489,554227438179	1
490,037344918183	1
504,461352153091	1
518,346646441582	1
521,77224886432	1
541,207191124077	1
541,367071427511	1
562,474674979066	1
567,984971228102	1
585,275496202272	1
598,085930483677	1
609,306581085307	1
631,455737930797	1
634,265354340085	1
659,85318961899	1
667,881801513495	1
685,770290858588	1
707,10520593077	1
748,833181997487	1
792,742097933901	1
838,480863874681	1
885,674771063061	1
933,929661675801	1
982,836425143826	1
1031,97571117537	1
1080,92287052403	1
1129,25299029154	1
1157,363912175116	1
1140,57459306969	1
1146,94388884054	1
1156,37830732664	1
1168,7962013222	1
1176,54599756189	1
1183,83165371248	1

Теперь построим диаграммы по кубу.


Количество.Количество
1
0,95
0,9
0,85
0,8
0,75
0,7
0,65
0,6
0,55
0,5
0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
458,926044167857
461,793196695602
467,987503965891
477,314837448729
489,554227438179
504,461352157091
521,77224886432
541,367071427551
567,994971226102
598,065936463677
631,455737938797
659,853189861899
685,779290858588
748,833183597487
838,480863874681
933,929663675801
1031,975711117537
1129,2529902954
1140,57459306969
1156,37830732664
1176,54599756189
1201,43695116906
1222,39175322278
1266,39506631134
1291,12306291811
1316,6384674557
1347,3973451428
1383,7233658199
1416,8694902435
1442,9769914370
1465,1066137768
1485,2590162900
1503,0668766591
1518,1931145624
1530,3233618562
1539,1820326368
1544,5299407081
Slovenia


Количество.Количество
5
4,8
4,6
4,4
4,2
4
3,8
3,6
3,4
3,2
3
2,8
2,6
2,4
2,2
2
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
1
10
107
1161
1199
1244
1360
1373
1381
1391
1403
1408
1421
1445
1457
1469
1480
178
191
21
226
3
336
400
44
471
5
581
682
701
728
769
84
97
Slovakia

Ответы на вопросы

1. Для чего следует проводить подготовку данных для анализа?

Подготовка данных обеспечивает качество и точность анализа, устраняя ошибки, пропуски и несоответствия, а также преобразуя данные в подходящий формат для эффективного использования аналитическими инструментами.

2. Что такое шумы и аномалии в данных?

- **Шумы** — случайные и нерегулярные колебания или ошибки в данных, не несущие полезной информации.
- **Аномалии** — значительные отклонения от нормального поведения данных, которые могут указывать на ошибки, необычные события или важные изменения.

3. Какими методами можно убрать шумы в системе Deductor?

- **Скользящее среднее**: сглаживание данных путем усреднения соседних точек.
- **Вейвлет-преобразование**: разделение сигнала и удаление высокочастотных шумов.
- **Фильтры низких частот**: устранение высокочастотных компонентов шума.

4. Какими методами можно убрать аномалии данных в системе Deductor?

- **Статистические методы**: использование Z-оценки или межквартильного размаха для выявления выбросов.
- **Методы кластеризации**: обнаружение точек, не принадлежащих ни одному кластеру.
- **Визуальный анализ**: применение графических методов, таких как ящики с усами или диаграммы рассеяния.

5. Для чего используется парциальная предобработка?

- Парциальная предобработка применяется для частичной очистки и трансформации данных на отдельных этапах обработки, улучшая качество данных перед основным анализом и позволяя более эффективно управлять большими объемами информации.

6. Для чего используется спектральная обработка?

Спектральная обработка анализирует частотные компоненты данных, позволяя выявлять скрытые паттерны, удалять шумы и аномалии, а также

сглаживать данные для улучшения их качества и подготовки к дальнейшему анализу.

7. Какие виды спектральной обработки имеются в системе Deductor?

- **Быстрое преобразование Фурье (БПФ)**: анализ частотных компонентов сигнала.
- **Вейвлет-преобразование**: многошкальный анализ данных.
- **Фильтрация частот**: удаление нежелательных частотных составляющих.

8. Что такое аппроксимация данных?

Аппроксимация данных — процесс приближения сложных данных более простыми моделями или функциями, сохраняя основные тенденции и структуры, что облегчает анализ и визуализацию.