Санкт-Петербургский государственный университет

Направление 02.03.03 «Математическое обеспечение и администрирование информационных систем»

Основная образовательная программа СВ.5006.2019 «Математическое обеспечение и администрирование информационных систем»

Профиль Информационно-аналитические системы

Говорова Диана Игоревна

Выпускная квалификационная работа

Система анализа двигательной активности по показаниям датчиков

Уровень образования: бакалавриат

Научный руководитель:

к. ф-м. н., доцент Графеева Н.Г.

Рецензент:

к. ф. н., доцент Егорова О.Б.

Санкт-Петербург

2023

Saint-Petersburg State University

Speciality 02.03.03 «Software and Administration of Information Systems»

Programme CB.5006.2019 «Software and Administration of Information Systems»

Profile Information and analytical systems

Govorova Diana

Bachelor's Thesis

Motion activity analysis system based on sensor readings

Scientific supervisor:

PhD, docent Grafeeva N.G.

Reviewer:

KfD, docent Egorova O.B.

Saint-Petersburg

2023

Оглавление

Введение

В современном мире ежедневно производятся, извлекаются и обрабатываются огромные объемы данных из самых разных областей человеческой жизни. При таком стремительном росте количества данных появляется естественная потребность в создании систем, помогающим людям с более быстрыми и качественными их сбором и обработкой. Ведь никак не обработанные данные не приносят пользы, их трудно анализировать и делать на основе анализа действительно значимые выводы. На сегодняшний день проблема грамотного, точного и быстрого анализа становится все актуальнее, и в этом специалистам помогают различные приложения для подготовки, обработки и анализа данных. Их неоспоримые плюсы заключаются в минимизации человеческих ресурсов, более быстрых и точных обработке и получения нужных результатов, а также сведении к минимуму ошибок, возникающих из-за человеческого фактора. Целью данной дипломной работы является создание приложения, позволяющего быстро и точно обрабатывать результаты экспериментов по данным, полученным от наших коллег с медицинского факультета.

1. Постановка задачи

Целью данной дипломной работы является разработка приложения для обработки данных по экспериментам, полученных от наших коллег с медицинского факультета.

В ходе исследования были поставлены следующие задачи:

1. Исследовать экспериментальные и контрольные данные, собранные в ходе исследования влияния препаратов на животных, полученные от наших коллег
2. Провести сбор и анализ требований к системе и составить техническое задание
3. Разработать прототип приложения
4. Провести тестирование

2. Исследование

2.1. Введение в предметную область

Поскольку основным объектом изучения является временной ряд, введем некоторые основные определения, связанные с ним[1][2].

Временной ряд – это упорядоченная последовательность значений какого-либо показателя за несколько периодов времени. Основная характеристика, которая отличает временной ряд от простой выборки данных, — указанное время измерения или номер изменения по порядку.

Временные ряды применяются во многих предметных областях, таких как медицина, экономика и т.д.

2.2. Данные

Все полученные данные делились на контрольные и экспериментальные и содержали следующую информацию:

1) OBJECT_TYPE (CONTR/EXP) - Тип объекта (контрольный/экспериментальный)
2) OBJECT – Номер животного
3) ID_VALUE – Номер измерения
4) VALUE – Значение, полученное в результате данного измерения
5) TYPE_VALUE – Тип измерения; участок животного, в котором находился датчик
6) TIME_VALUE – Время эксперимента (BEFORE/AFTER EXPERIMENT)
7) SUBSTANCE_VALUE – Тип введенной субстанции

Пример данных показан на Рис.1

	A	B	C	D	E	F	G
1	_TYPE (C	HOM	OMEF	VALUE (ИЗМ	TYPE_V	/ALUE(BEFOR	SUBSTANCI
2	EXP	1	1	153,01	DIST	BEFORE	ACH
3	EXP	1	2	153,01	DIST	BEFORE	ACH
4	EXP	1	3	153,01	DIST	BEFORE	ACH
5	EXP	1	4	153,01	DIST	BEFORE	ACH
6	EXP	1	5	153,01	DIST	BEFORE	ACH

Рис. 1

3. Обзор

3.1. Обзор существующего решения

В предыдущем семестре в силу необходимости обработки экспериментов в качестве временного решения все расчеты производились при помощи программы Excel. Ввиду большого количества экспериментов вычисления проводились при помощи макросов, созданных с помощью языка VBA, где были прописаны все формулы для расчета необходимых характеристик[3].

Пример расчета (для экспериментальных данных №1) показан на Рис. 2

	A	B	C	D	E	F	G	H	I
1			AMPLITUDE	COUNT_1000	COUNT_3000	TONUS	PERCENTILE		
2							0,995	0,05	0,85
3	1EXP	DIST BEFORE ACH	61,998	39,5	13,5	144,014	419,007	146,014	164,014
4	1EXP	DIST AFTER ACH	303,993	32	24,5	126,014	567,003	128,014	209,012
5	1EXP	DIST BEFORE BUT	53,438	31,5	16	313,067	503,382	303,692	348,692
6	1EXP	DIST AFTER BUT	95,626	27,5	17,5	243,691	410,568	244,628	288,691

Рис. 2

Также для каждого эксперимента были построены графики временных рядов по рассчитанным сглаженным значениям. Визуализация временных рядов была выполнена при помощи встроенных инструментов Excel для создания графиков и диаграмм.

Пример построенного графика по данным одного из экспериментов показан на Рис. 3

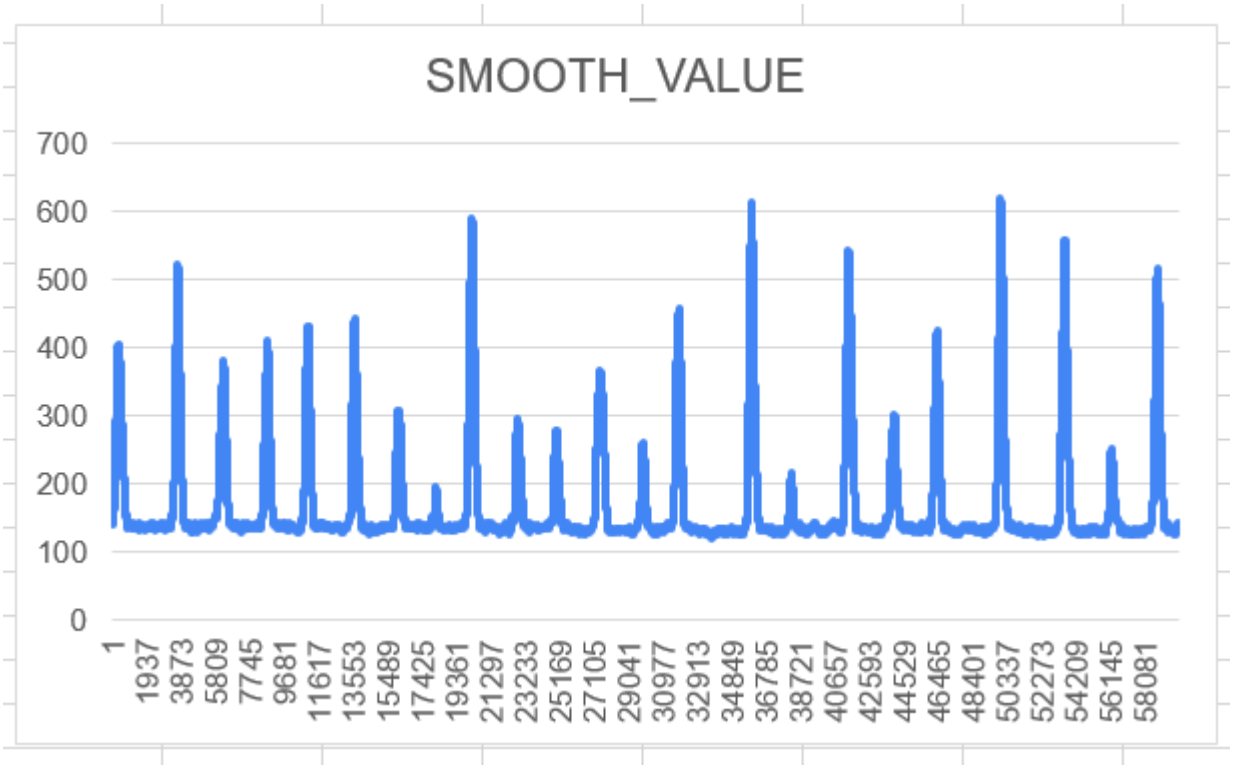

Рис. 3

Однако ввиду того, что расчет необходимых показателей временных рядов таким образом имеет ряд недостатков, таких как непрерывная вовлеченность человеческого ресурса в процесс подсчета, долгая обработка данных экспериментов ввиду медленной работы макросов, необходимость ручного сбора данных по каждому эксперименту в единый файл, появилась необходимость создания приложения, обладающего удобным и понятным интерфейсом, а также быстрой скоростью обработки необходимых файлов.

4. Реализация

4.1. Общие требования к системе

Основной идеей приложения является обеспечение быстрой, точной и качественной обработки результатов экспериментов.

Система должна поддерживать:

1) Загрузку с компьютера одного или нескольких файлов в формате xlsx из выбранной пользователем папки

2) Расчет следующих характеристик временных рядов:

- Амплитуда
- Частота на основе линейного тренда с шагом в 1000 значений
- Частота на основе линейного тренда с шагом в 3000 значений
- Тонус
- Значения k-ой персентили для трех различных порогов приемлемости (0,995; 0,05; 0,85)

3) Сохранение сводной таблицы по всем результатам в формате xlsx в выбранную пользователем папку

4) Визуализация графиков по каждому эксперименту и сохранение изображения в формате jpg в выбранную пользователем папку

5) Агрегирование данных по всем животным по параметрам экспериментов (контрольная/экспериментальная группа; участок животного, в который вводили препарат; до/после введения; тип препарата) и сохранение полученной таблицы в формате xlsx

6) Расчет коэффициента изменения каждой характеристики до/после эксперимента, сохранение результата в формате xlsx

7) Визуализация графиков изменения каждой характеристики в разрезе типа группы, участка введения препарата и типа препарата до и после эксперимента, сохранение результата в формате jpg

4.2. Описание сценария использования системы

1) При запуске приложения пользователю предлагается выбрать один или несколько файлов с данными в формате xlsx, которые необходимо будет обработать. Выбрать файлы можно в окне проводника Windows.

2) Выбор папки для сохранения сводной таблицы с результатами расчетов также происходит в начале работы программы. Это удобнее для пользователя, так как нет необходимости ждать окончания обработки файла. Выбрать папку для сохранения результатов также можно в окне проводника.

3) В выбранную ранее папку сохранятся также все графики с визуализациями временных рядов, построенных по данным экспериментов.

4) Кроме того, туда же сохранятся медианные значения характеристик, агрегированных по параметрам экспериментов, а также их коэффициенты изменения до/после введения препарата и столбчатые диаграммы изменения каждой из них.

4.3. Реализация основных функций

После сбора необходимых требований к приложению у заказчиков, началась непосредственно его разработка. В качестве основного инструмента был выбран язык Python, ввиду его неоспоримых плюсов: скорость разработки, простой и понятный синтаксис, гибкость, а также множество фреймворков и библиотек для работы с данными и их визуализацией (например: pandas, numpy и другие).

Прежде всего для устранения шумов и кратковременных флуктуаций к временному ряду был применен метод сглаживания, используя скользящее среднее с шагом в 9 позиций по снятым показаниям экспериментов.

После чего был рассчитан ряд значений, которые в дальнейшем будут необходимы для вычисления итоговых характеристик временных рядов:

1) Используя оконные функции, реализованных при помощи DataFrame.rolling() из библиотеки Pandas[4], были посчитаны скользящие минимальные и максимальные значения с шагом в 3000 по сглаженным значениям, а также разность между ними.
2) Вычислены точки, в которых прямые, построенные по скользящему окну в 1000 и 3000 точек, меняют знак углового коэффициента прямой[5].

После формирования датасета с предварительными расчетами, было необходимо вычислить итоговые характеристики временных рядов, построенных по данным экспериментов, такие как:

1) Амплитуда

   Амплитуда – пиковое значение в положительном или отрицательном направлении ряда[6].

Характеристика была рассчитана как медиана всех амплитуд точек, рассчитанных ранее.

2) Частота на основе линейного тренда с шагом в 1000 и 3000 значений

Частота - это количество наблюдений до повторения сезонной картины[7]. Данная характеристика была вычислена как половина от количества точек, в которых угловой коэффициент функции меняет знак.

3) Тонус

Тонус ряда рассчитывался как медиана скользящих минимумов.

4) Персентиль

Персентиль – методика измерения в статистике, которая показывает процент значений измеряемой метрики, который находится ниже значения персентиля[8].

В качестве порогов приемлемости были выбраны значения 0,995, 0,05 и 0,85. Для вычисления была использована встроенная функция языка Python *numpy.percentile()*

Пример итоговой сводной таблицы после расчета всех характеристик показан на Рис. 4.

EXP/CONTR	NUM_OF_RAT	AMPLITUDE	COUNT_1000	COUNT_3000	TONUS	PERCENTILE_0.995	PERCENTILE_0.05	PERCENTILE_0.85	PARAMETR_1	PARAMETR_2	PARAMETR_3
EXP	2	181,88	30	15,5	277,44	724,64	280,26	402,13	DIST	BEFORE	BUT
EXP	2	451,88	28,5	11,5	225,02	1494,41	225,02	745,33	DIST	BEFORE	ACH
EXP	2	937,51	27	20	218,38	1205,62	215,57	550,26	DIST	AFTER	BUT
EXP	2	607,51	30,5	15	229,7	1306,9	212,83	759,4	DIST	AFTER	ACH
EXP	2	54,3	27,5	8,5	296,19	423,49	255,25	376,31	PROX	BEFORE	BUT
EXP	2	97,98	32	20	242,75	372,04	241,74	327,6	PROX	BEFORE	ACH
EXP	2	67,65	38	20,5	288,18	381,65	247,24	354,06	PROX	AFTER	BUT
EXP	2	26,26	38	10,5	162,95	220,52	158,91	202,34	PROX	AFTER	ACH
EXP	1	53,44	31,5	16	313,07	503,38	303,69	348,69	DIST	BEFORE	BUT
EXP	1	62	39,5	13,5	144,01	419,01	146,01	164,01	DIST	BEFORE	ACH
EXP	1	95,63	27,5	17,5	243,69	410,57	244,63	288,69	DIST	AFTER	BUT
EXP	1	303,99	32	23,5	126,01	567	128,01	209,01	DIST	AFTER	ACH

Рис. 4

Для реализации диалогового окна общения с пользователем использовалась библиотека Tkinter. Вызов окон загрузки и сохранения файла выполнялся с помощью функций filedialog.askopenfilenames и filedialog.askdirectory[9].

Для визуализации графиков временных рядов, построенных по сглаженным значениям экспериментов, использовалась библиотека matplotlib.pyplot[10].

Пример графика представлен на Рис. 5 (параметры: участок введения препарата: dist; препарат: but; время замера: после введения препарата)

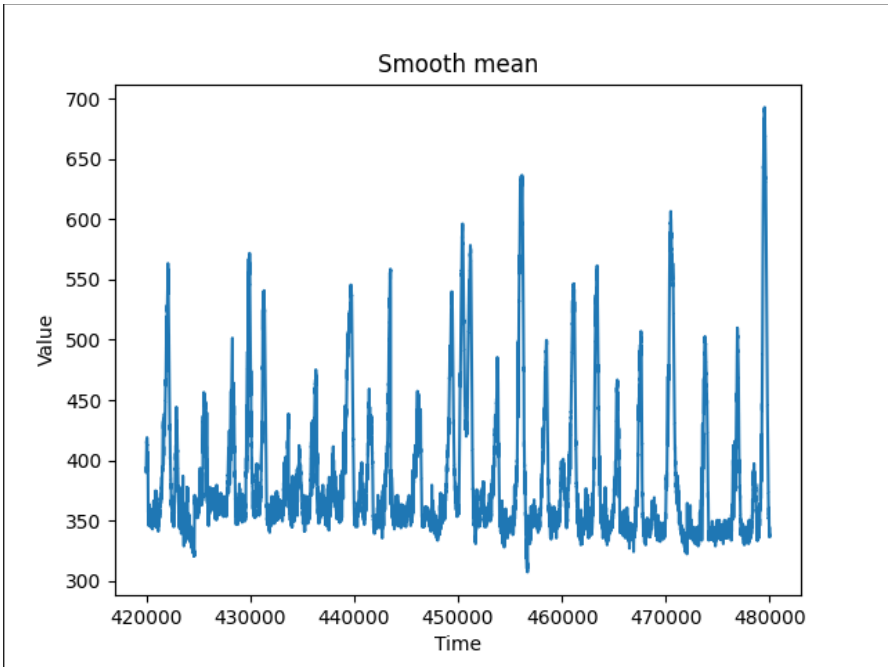

Рис. 5

При создании таблицы по всем экспериментам, агрегированной по параметрам (типу данных; участку введения препарата; типу препарата; времени снятия данных(до/после введения препарата)), была рассчитана медиана, после чего для каждой характеристики был рассчитан коэффициент изменения показателя до и после проведения эксперимента.

Пример таблицы, агрегированной по параметрам экспериментов и таблицы индексов представлен на Рис. 6

EXP/CONTR	TYPE_VALUE	TIME_VALUE	SUBST_VALUE	AMPLITUDE	COUNT_1000	COUNT_3000	TONUS	PERCENTILE_0.995	PERCENTILE_0.05	PERCENTILE_0.85
			Таблица медианных значений параметров экспериментов							
EXP	DIST	BEFORE	BUT	202,03	42,75	20	332,29	642,61	314	467,76
EXP	DIST	AFTER	BUT	202,5	44,75	23,25	297,6	565,73	299,47	427,44
EXP	DIST	BEFORE	ACH	33,04	35	7,5	337,34	441,91	293,72	412,87
EXP	DIST	AFTER	ACH	35,95	44	20	338,96	415,85	333,76	363,34
EXP	PROX	BEFORE	BUT	202,72	31	17	1030,83	1274,46	1018,11	1199,01
EXP	PROX	AFTER	BUT	292,72	46	21,5	745,38	1119,01	751,74	993,56
EXP	PROX	BEFORE	ACH	233,78	27,5	20	614,48	892,74	605,17	823,43
EXP	PROX	AFTER	ACH	56,89	27,5	7,5	811,02	917,57	736,54	887,57
			Таблица индексов изменение характеристик до/после введения препарата							
EXP	DIST		BUT	1	1,05	1,16	0,9	0,88	0,95	0,91
EXP	DIST		ACH	1,09	1,26	2,67	1	0,94	1,14	0,88
EXP	PROX		BUT	1,44	1,48	1,26	0,72	0,88	0,74	0,83
EXP	PROX		ACH	0,24	1	0,38	1,32	1,03	1,22	1,08

Рис. 6

Для визуализации графиков изменения каждой характеристики в разрезе группы, агрегированной по параметрам была использована библиотека matplotlib.pyplot[10].

Пример графика изменения амплитуды представлен на Рис. 7

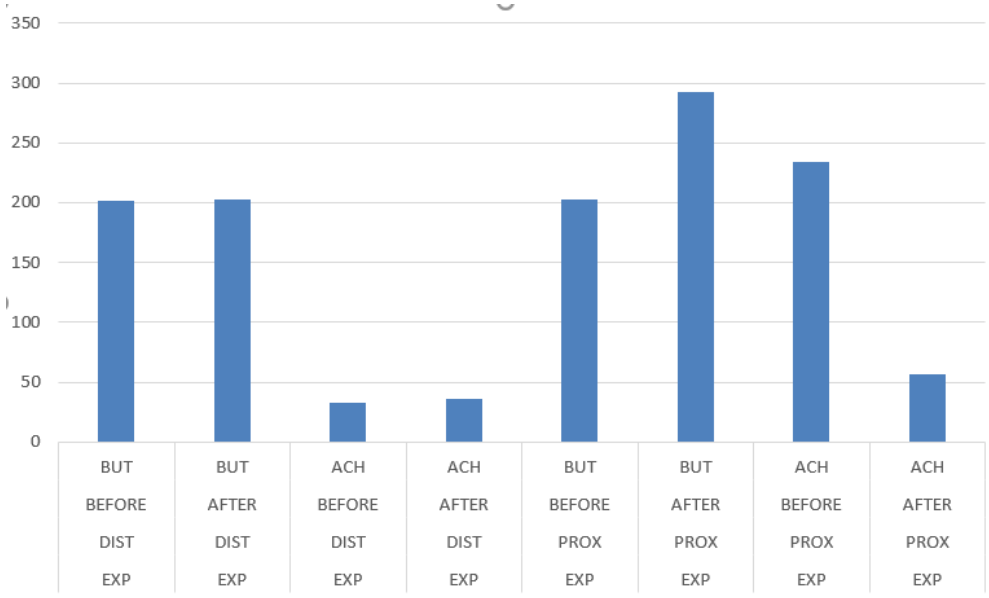

Рис. 7

4.4. Оптимизация решения

Ввиду того, что данные по каждому эксперименту независимы друг от друга и могут обрабатываться раздельно, было решено попробовать сделать параллельную обработку файлов с экспериментами для большего сокращения времени обработки.

Распараллеливание проводилось при помощи следующей библиотеки языка Python:

1) *multiprocessing*[11] - это пакет, поддерживающий порождение процессов с использованием API, позволяющий программисту полностью использовать несколько процессоров на компьютере.

Для параллельной обработки данных использовался *multiprocessing.Pool*. Класс Pool() создает объект, который управляет пулом рабочих процессов, в который отправляются задания. Пул рабочих процессов поддерживает асинхронное выполнение задач с тайм-аутами, обратными вызовами, а также имеет параллельную реализацию[12]. При распараллеливании выполнялось разделение файла на отдельные эксперименты по типу объекта, номеру животного, типу введенной субстанции, времени замера эксперимента, а также участка, в который был введен препарат, после чего эксперименты обрабатывались параллельно. Такая оптимизация не уменьшила скорость обработки файла, в котором содержится один эксперимент, однако показала очень хорошее уменьшение скорости на файлах с данными, содержащими несколько экспериментов.

4.5. Инструкция по использованию

Перед запуском системы необходимо выполнит следующие шаги:

1) Установить Python с официального сайта : https://www.python.org/downloads/
2) Установить следующие пакеты:

   -pandas

   -numpy

   -matplotlib

   Для установки необходимо открыть командную строку и прописать команду: pip install *название_пакета*

   Например: pip install pandas
3) Скачать код приложения по ссылке ниже: https://github.com/Diana-Govorova/Diploma/blob/main/Diplom.py

   Поместить его в выбранную папку.
4) Для запуска приложения необходимо написать в командной строке: python *путь_до_файла*.py

   Например: python C:\Users\diana \Diplom\Diplom.py

5) После запуска откроется диалоговое окно, в котором необходимо выбрать один или несколько файлов для обработки (выбрать несколько файлов можно зажав Ctrl):

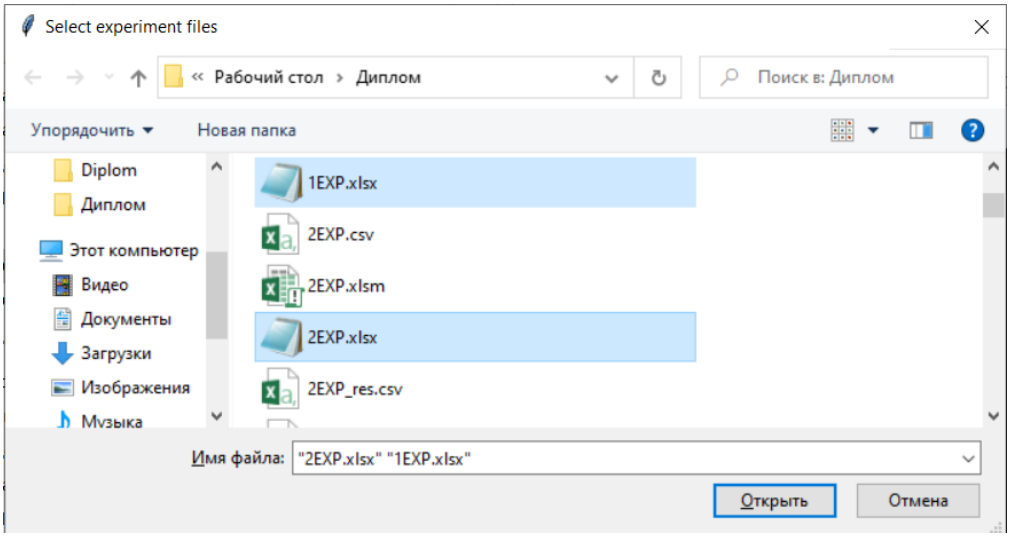

Рис. 4

6) После нажатия кнопки открыть (см Рис.4), появится еще одно диалоговое окно, в котором необходимо выбрать папку для сохранения результатов расчетов и графиков:

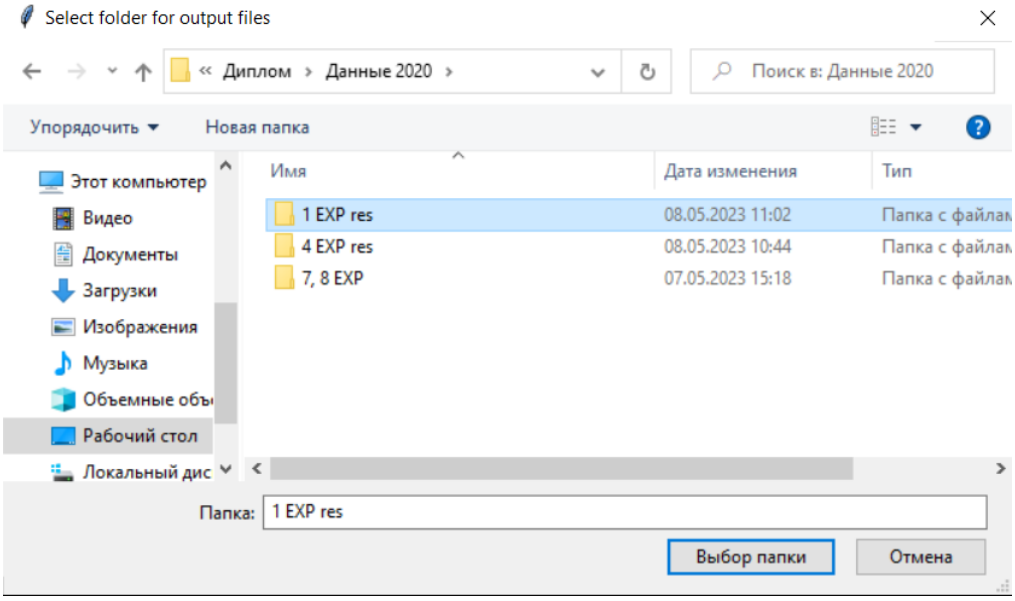

Рис. 5

5. Тестирование

После создания финальной версии приложения необходимо было протестировать и сравнить его с уже существующей версией обработки файлов при помощи макросов, а также с первоначальной версией приложения, в которой не использовалось распараллеливание.
Описание данных и среды, в которой проводилось тестирование:

-операционная система: Windows

-модель ноутбука, его характеристики (процессор, оперативная память): *Acer Aspire 5 A515-54G-511G*, процессор *Intel Core i5 8265U*, оперативная память (тип, частота, размер) *DDR4, 2133 МГц, 8 ГБ.*

-версия Excel*: Microsoft Office 2016 Professional Plus*

-количество данных в одном эксперименте: *60 000 строк*

-количество экспериментов в одном файле: *8 экспериментов*

- количество экспериментов в двух файлах: *16 экспериментов*

Для сравнения были выделены следующие характеристики:
1) Скорость обработки данных по 1 эксперименту
2) Скорость обработки одного файла, состоящего из 8 экспериментов
3) Скорость обработки 2х файлов, состоящих в общей сложности из 16 экспериментов
4) Человеческий ресурс: количество кликов, чтобы обработать данные по одному эксперименту

5) Риск ошибки из-за человеческого фактора. Под риском понимается возможное возникновение ошибки при объединении данных по экспериментам в итоговую таблицу

Для каждого параметра проводилось по 10 запусков эксперимента, после чего считались среднее значение и стандартное отклонение выборки (мера разброса данных относительно среднего значения) [13].

	Макрос	Приложение до распараллеливания	Приложение после распараллеливания
Скорость обработки данных по 1 экс.	μ: 155,8 с σ: 9,2 с	μ: 82, 2 с σ: 6,9 с	μ: 79,1 с σ: 5,6 с
Скорость обработки 1 файла с 8 экс.	μ: 1310,1 с σ: 40,5 с	μ: 1011,9 с σ: 16,6 с	μ: 379,0 с σ: 6,8 с
Скорость обработки 2х файлов (16 экс.)	μ: 2720,7 с σ: 33,8 с	μ: 2035,2 с σ: 19,2 с	μ: 543,4 с σ: 11,0 с
Количество кликов	30	5	5
Риск ошибки	да	нет	нет

Таблица 1 Сравнение характеристик

(μ – среднее значение, σ – стандартное отклонение)

Основываясь на данных в таблице, может сделать вывод, что финальная версия приложения дала не только большой выигрыш в скорости обработки файлов:

-в 1,9 раз по скорости обработки одного эксперимента по сравнению со скоростью обработки при помощи макросов

-в 3,5 раза по скорости обработки одного файла, состоящего из 8 экспериментов по сравнению со скоростью обработки при помощи макросов, и в 2,7 раз по сравнению со скоростью обработки эксперимента при помощи первоначальной версии приложения

-в 5 раза по скорости обработки 2х файлов по сравнению со скоростью обработки при помощи макросов, и в 3,7 раз по сравнению со скоростью обработки эксперимента при помощи первоначальной версии приложения

, а также сократило затраты человеческого ресурса для обработки файлов, измеряемого в количестве кликов, в 6 раз для одного эксперимента и свело на нет риск ошибки, вызванной человеческим фактором.

7. Заключение

В результате выполнения дипломной работы было разработано приложение для обработки данных по экспериментам, которое будет в дальнейшем использоваться нашими коллегами с медицинского факультета.

- В ходе работы был проведен анализ предметной области, выполнен обзор существующего решения
- Проведен сбор и анализ требований к системе, составлено техническое задание
- Разработан прототип приложения для решения поставленной задачи
- Проведено тестирование и сравнение по некоторым характеристикам с уже существующим решением

Созданное приложение является прототипом и в будущем будет иметь возможность загружать данные по экспериментам из базы данных SQL, так как файл Excel может содержать только ограниченное количество строк.

Список литературы:

1) Статья «Временной ряд.»
   URL: https://blog.skillfactory.ru/glossary/vremennoj-ryad-2/ (дата обращения: 2022-10-14)
2) Статья «What Is a Time Series and How Is It Used to Analyze Data?»
   URL: https://www.investopedia.com/terms/t/timeseries.asp (дата обращения: 2022-10-28)
3) Статья «Введение в VBA: Макросы.»
   URL: https://excelpedia.ru/makrosi-v-excel/vvedenie-v-vba-makrosy-chast-1-iz-3 (дата обращения: 2022-10-28)
4) Документация «pandas.DataFrame.rolling»
   URL: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rolling.html (дата обращения: 2023-01-25)
5) Статья «Slope.»
   URL: https://www.cuemath.com/geometry/slope/ (дата обращения: 2023-01-25)
6) Лекция «Time Series Analysis.»
   URL: http://people.rses.anu.edu.au/heslop_d/TSA2011%20-%20Lectures.pdf (дата обращения: 2023-02-04)
7) Электронная книга «Forecasting: Principles and Practice.»
   URL: https://otexts.com/fpp2/ts-objects.html (дата обращения: 2023-02-04)
8) Статья «Percentiles, Percentile Rank & Percentile Range: Definition & Examples»
   URL: https://www.statisticshowto.com/probability-and-statistics/percentiles-rank-range/ (дата обращения: 2023-02-15)

9) Электронный учебник «Руководство по Tkinter»
URL: https://metanit.com/python/tkinter/1.1.php (дата обращения: 2023-02-18)
10) Документация «matplotlib.pyplot»
URL: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html (дата обращения: 2023-02-23)
11) Статья «Python Multiprocessing: The Complete Guide.»
URL: https://superfastpython.com/multiprocessing-in-python/ (дата обращения: 2023-02-25)
12) Электронный учебник «Python Multiprocessing Pool: The Complete Guide. »
URL: https://superfastpython.com/multiprocessing-pool-python/ (дата обращения: 2023-02-25)
13) Статья «Standard Deviation Formula and Uses vs. Variance. »
URL: https://www.investopedia.com/terms/s/standarddeviation.asp (дата обращения: 2023-03-01)