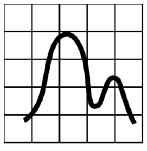
Краткое руководство пользователя ПО СтатСофт



СтатСофт®

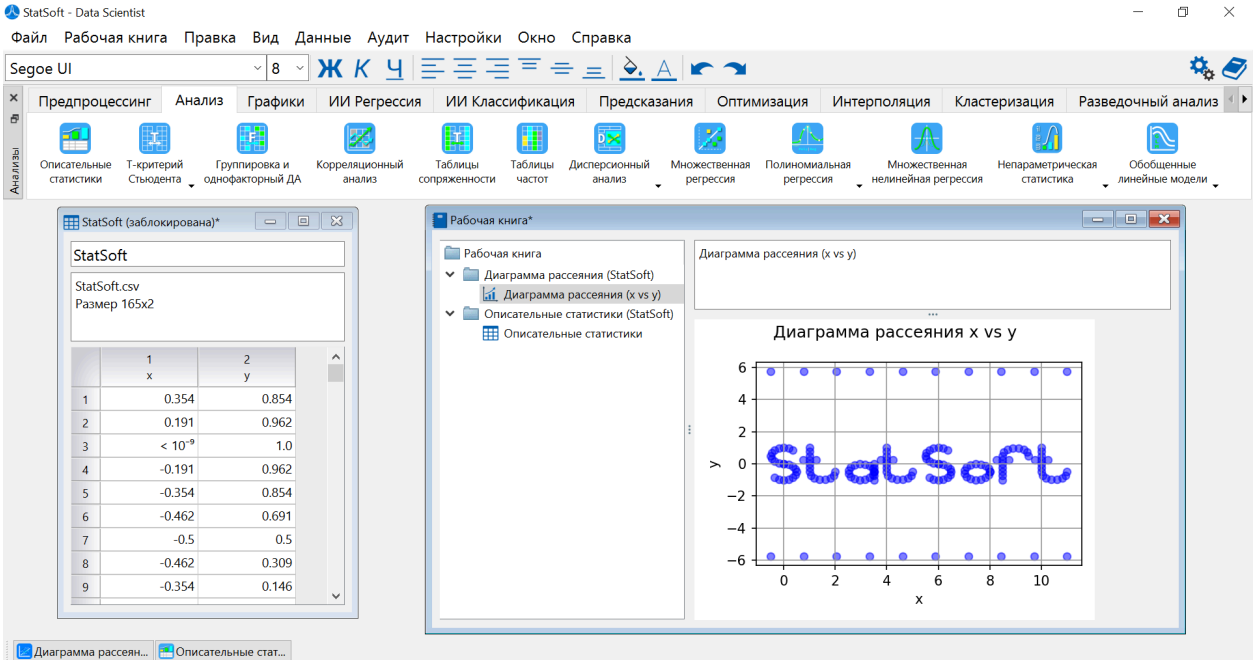
Сайт: https://statsoftai.ru

Оглавление

СтатСофт®

СтатСофт®

О программе



Предлагаемое пользователям ПО СтатСофт синтезирует лучшие современные достижения и мировые практики в области анализа данных, машинного обучения и искусственного интеллекта и отражает 20-летний опыт компании СтатСофт и тесное сотрудничество с ведущими мировыми IT компаниями.

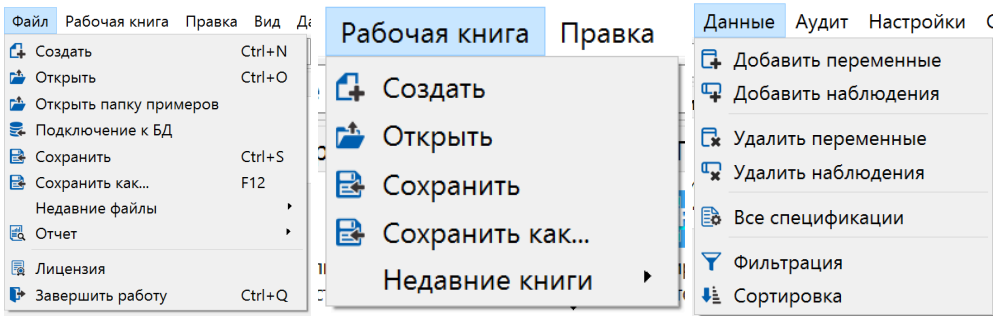
В ПО СтатСофт тесно слиты классические статистические методы и методы искусственного интеллекта, включая нейронные сети. Эти методы подкреплены качественными реальными примерами из области экономики, управления, бизнеса, финансов, промышленности. Необходимость совместного использования классических статистических методов и искусственного интеллекта диктуется тем, что данные методы взаимно дополняют и не противоречат друг другу.

Важным достоинством ПО СтатСофт является выбор лучших моделей на основе различных метрик и оптимизация на основе построенных моделей. Пользователь имеет возможность выбрать лучшую модель на основе заданных метрик, включая коэффициент детерминации, критерий Акаике и др.

Пользователи могут использовать данное ПО в различных комплектациях и версиях, включая десктоповские решения, сетевые (конкурентные), серверные решения.

СтатСофт®

Дейта менеджмент

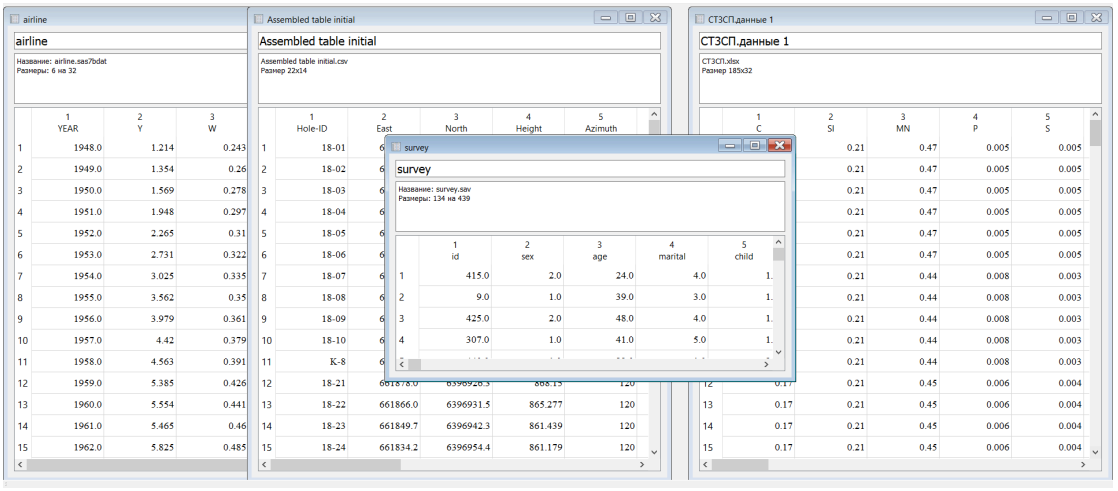


Таблицы данных

Таблицы данных ПО СтатСофт основаны на технологии мультимедийных таблиц, разработанной компанией СтатСофт. Система работает как с исходными данными, так и с численными и текстовыми результатами, поддерживаются технологии drag and drop, вставка объектов OLE.

Таблица данных ПО СтатСофт является двумерной таблицей, которая может содержать практически неограниченное число наблюдений (строк) и переменных (столбцов), при этом каждая ячейка может содержать неограниченное количество символов.

Отметим, что в ячейки можно помещать числа, текст и изображения.

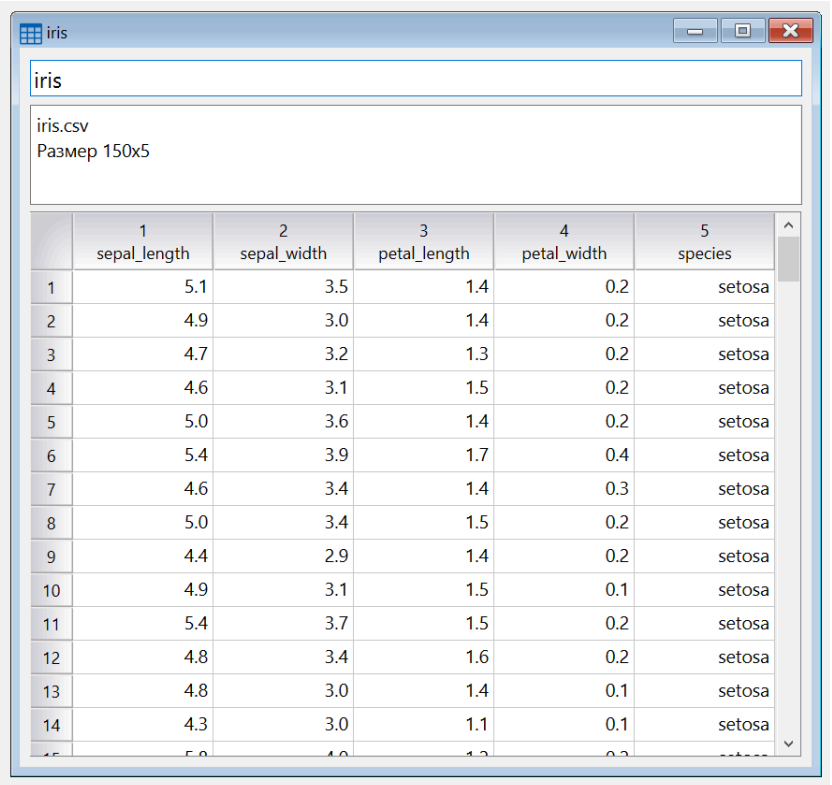


Данные в системе ПО СтатСофт организованы в виде наблюдений и переменных. Если вы не знакомы с этими понятиями, то можете представить себе, что наблюдения соответствуют записям таблицы базы данных (строкам), а переменные представляют

собой поля (столбцы). Каждое наблюдение состоит из набора значений переменных, в первом столбце файла содержатся номера наблюдений.

Окно таблицы данных

Окно Таблицы данных состоит из нескольких основных частей.



	1 sepal_length	2 sepal_width	3 petal_length	4 petal_width	5 species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa

Заголовок окна. В области Заголовок окна отображается название Таблицы. На рисунке выше в области Заголовок окна отображается текст iris

Заголовок. Дважды кликните мышью в области Заголовок, сверху окна над именами переменных для того, чтобы ввести или изменить текстовую информацию. На рисунке выше в области Заголовок содержится текст iris.

Информационное поле. Дважды нажмите в области Информационное поле для ввода или редактирования текста в этой области (например, для ввода дополнительных комментариев к Таблице данных).

Для увеличения размера области Информационное поле необходимо расположить указатель мыши в нижней части Информационное поле (указатель мыши принимает при этом знак ) и изменить размер.

На рисунке выше в области Информационное поле содержится текст Информационное поле.

Выделение таблицы. Для того, чтобы выделить всю Таблицу, нажмите на область, которая располагается в верхнем левом углу.

Номера наблюдений. Эти ячейки, расположенные в левой части окна Таблицы данных, содержат порядковые номера для каждого наблюдения.

Для того чтобы выделить всю строку наблюдения, нажмите один раз по данному наблюдению.

Для подгонки высоты строки Номера наблюдений необходимо расположить указатель мыши на нижней границе поля (указатель мыши принимает при этом знак )).

На рисунке выше ячейки Номера наблюдений содержат цифры 1, 2, ... n.

Имена переменных. Эти ячейки, расположенные в верхней части каждого столбца, содержат имена переменных.

Для просмотра и изменения спецификаций переменной столбца дважды нажмите в поле Имя переменной.

Для выделения всего столбца нажмите один раз в области Имя переменной.

Для подгонки ширины столбца необходимо расположить указатель мыши на правой границе поля Имя переменной (указатель мыши при этом принимает вид ).

На рисунке выше первые две ячейки Имя переменной содержат текст sepal_length и sepal_width.

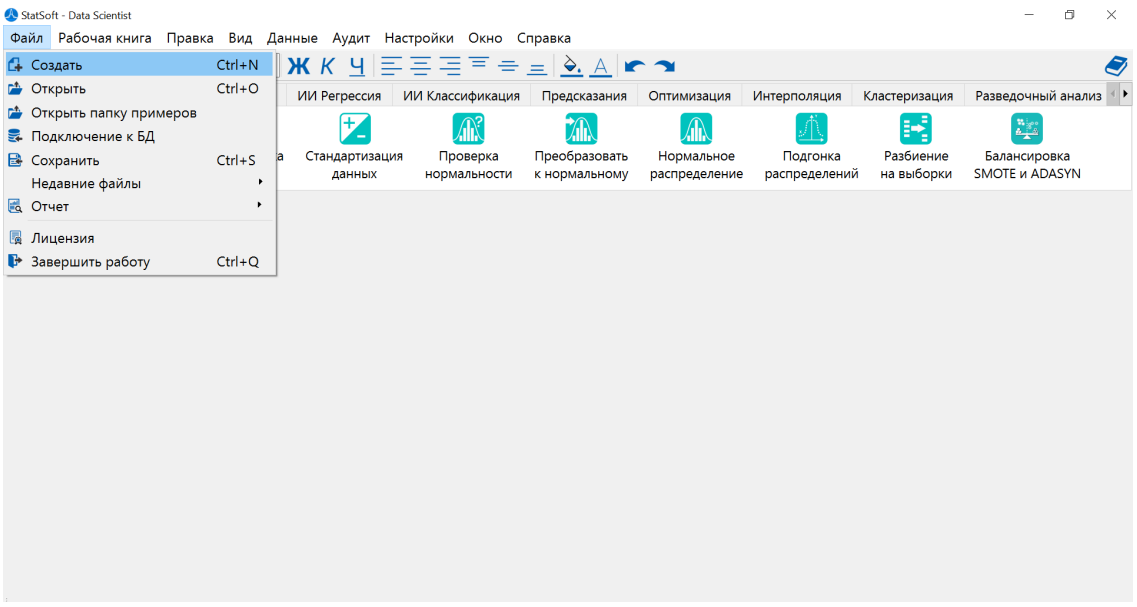
Данные (и редактирование внутри ячеек). Оставшаяся область Таблицы данных содержит сами данные, представленные в виде наблюдений и переменных. Текст в ячейках может иметь практически неограниченную длину.

Имеются широкие возможности для форматирования текста в ячейках.

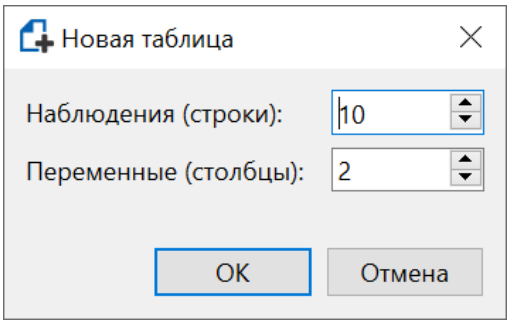
Создание таблиц

Чтобы создать новую Таблицу данных выполните следующие действия:

1. Выберите команду Создать в меню Файл.



2. В диалоге Новая таблица задайте Число переменных и Число наблюдений.

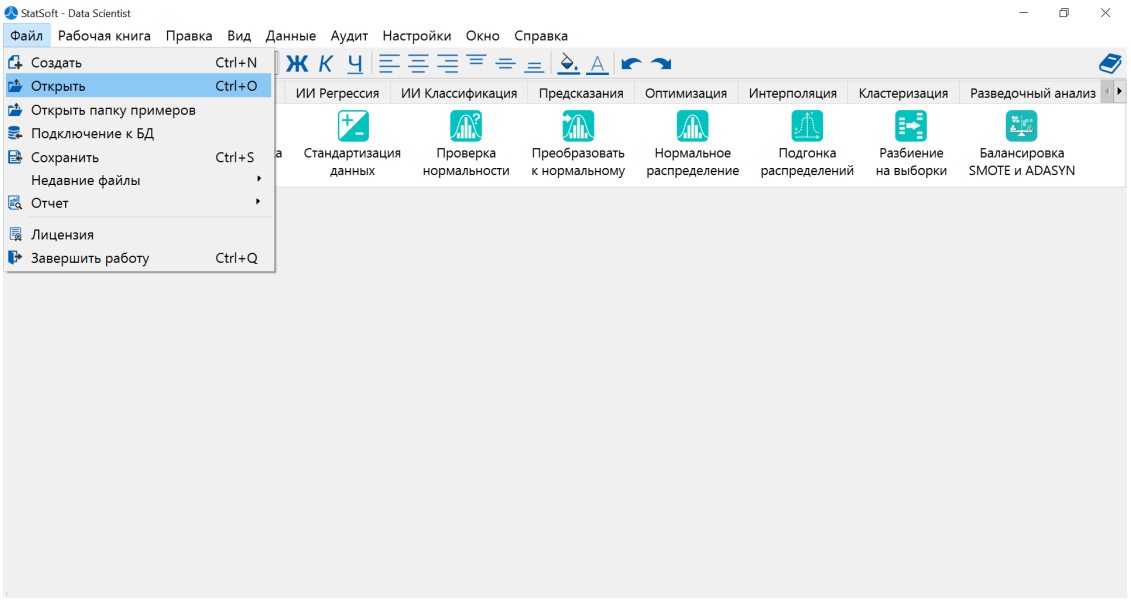


3. Нажмите кнопку OK.

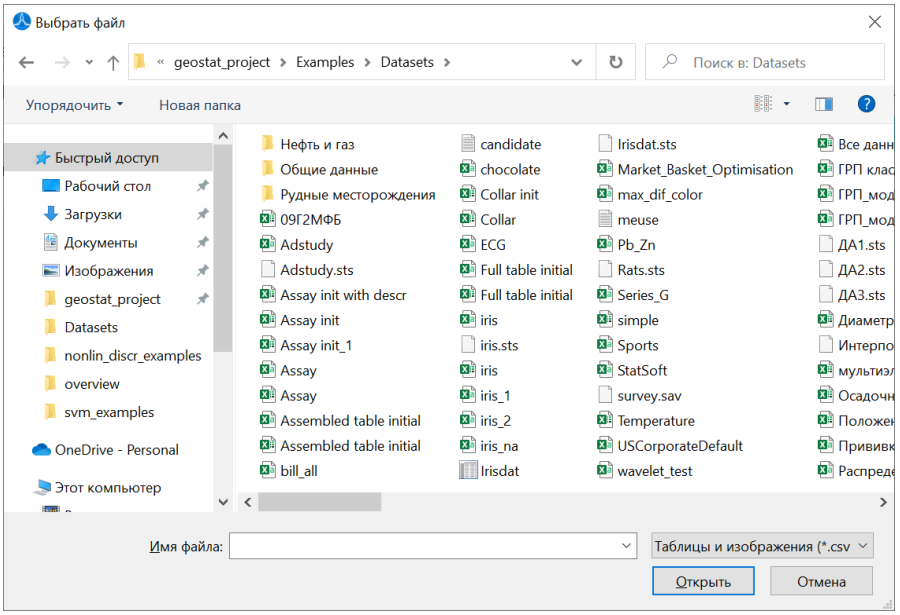
Открытие таблиц

Для того, чтобы открыть Таблицу данных (или несколько Таблиц данных), необходимо выполнить следующие действия:

1. Выберите команду Открыть в меню Файл в левом верхнем углу.



2. В диалоге Открыть укажите в поле *Имя файла* папку, в которой содержится требуемая Таблица данных.



3. Если потребуется, измените значение в поле Тип файлов, используя выпадающий список.

4. Выберите нужный файл и нажмите кнопку Открыть.

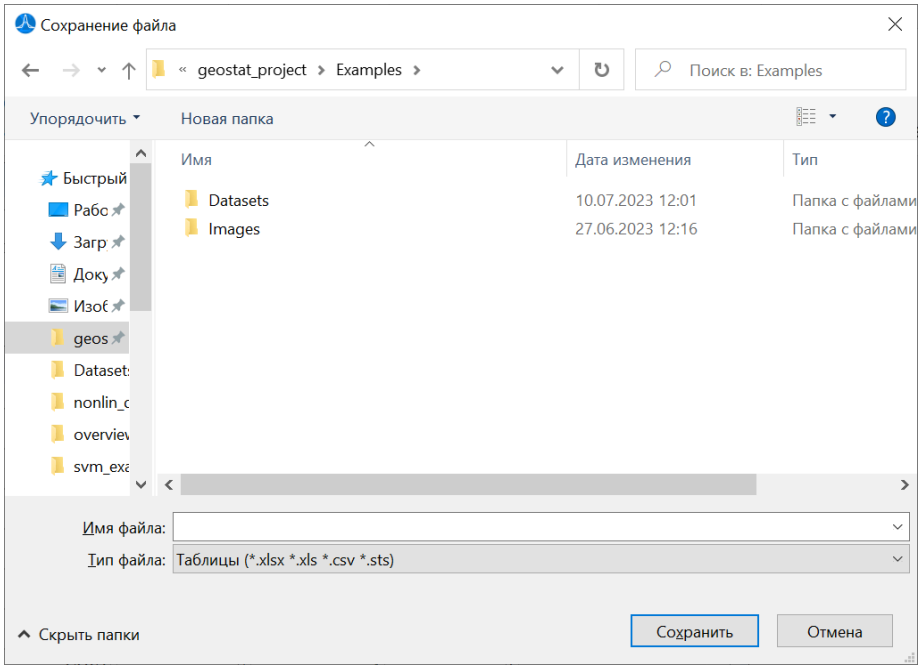
5. В зависимости от типа файла, в открывшимся диалоге выберите параметры импорта и нажмите ОК.

Сохранение таблиц

Для того, чтобы сохранить активную Таблицу данных, выполните последовательно следующие действия:

1. Выберите команду Сохранить в меню Файл

2. В появившемся диалоге выберите тип сохраняемого файла, укажите имя и нажмите Сохранить



3. В зависимости от типа сохраняемого файла откроется диалог с параметрами сохранения.

4. Укажите необходимые параметры и нажмите ОК.

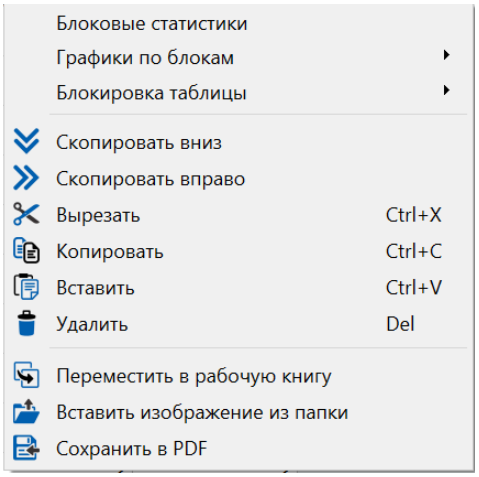
Блокировка и разблокировка таблиц

ПО СтатСофт позволяет защитить ваши данные от внесения каких-либо изменений. Для этого предусмотрены опции блокировки и разблокировки таблиц.

Блокировка таблиц. Чтобы заблокировать таблицу, сделайте следующее:

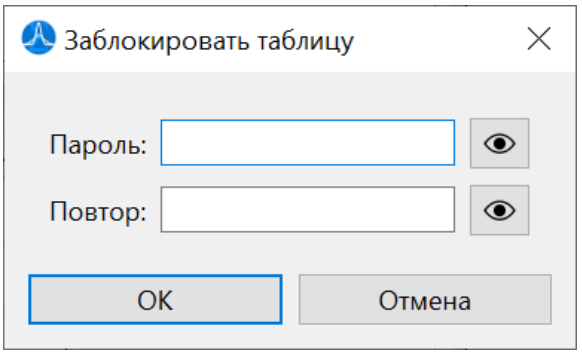
1. Откройте таблицу данных

2. Откройте контекстное меню, кликнув правой кнопкой мыши по таблице



3. В контекстном меню выберите Блокировка таблицы – Заблокировать

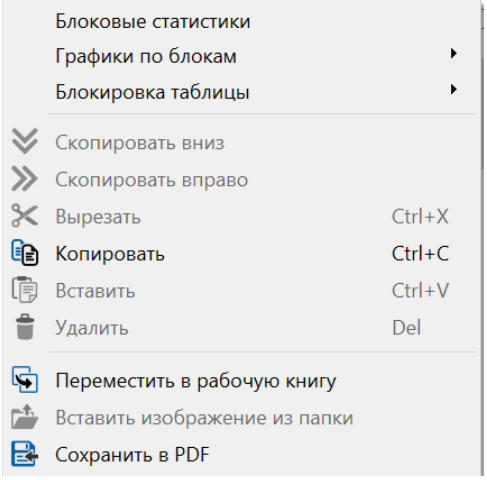
4. В открывшимся диалоге введите и подтвердите пароль, который будет использоваться при изменении блокировки таблицы данных



5. Нажмите OK для блокировки таблицы данных

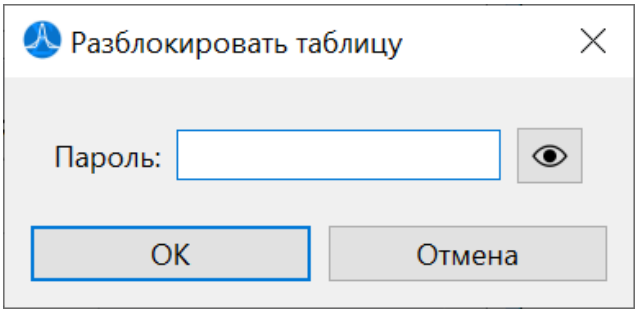
Разблокировка таблиц. Чтобы разблокировать выбранную таблицу, сделайте следующее:

1. Откройте контекстное меню, кликнув правой кнопкой мыши по таблице



2. В контекстном меню выберите Блокировка таблицы – Разблокировать

3. В открывшемся диалоге введите пароль

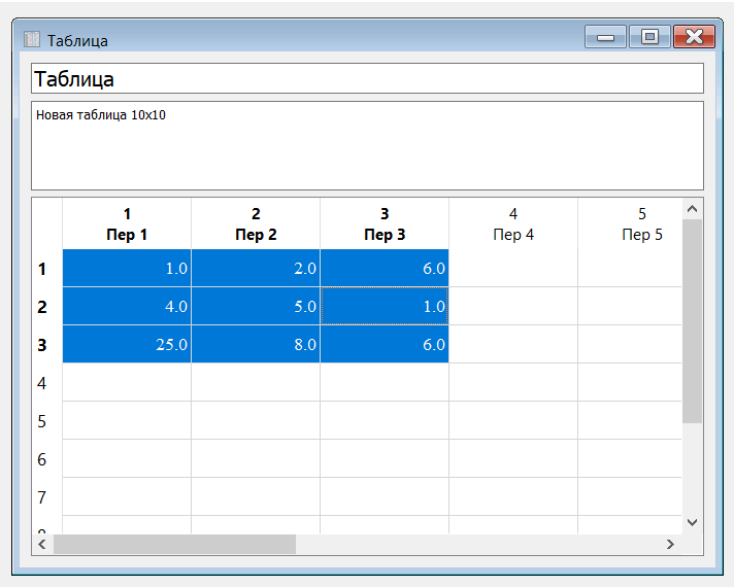


4. Нажмите OK, чтобы разблокировать таблицу данных

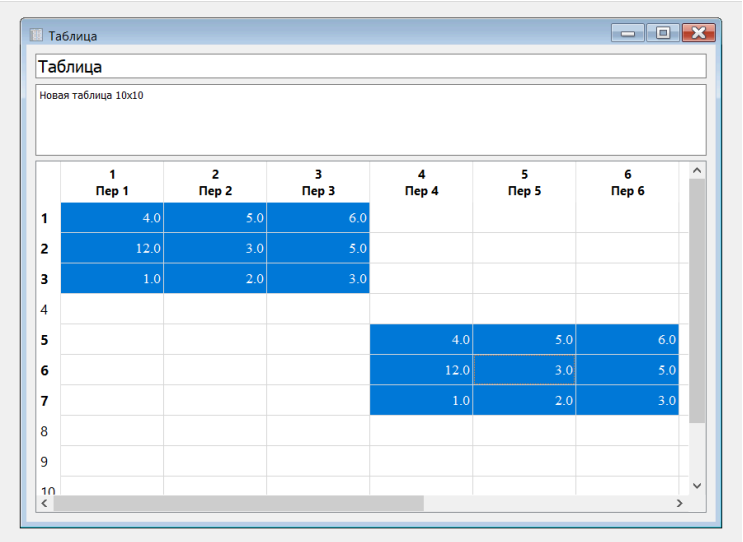
Метод перетаскивания

Таблицы данных ПО СтатСофт поддерживают все стандартные операции электронных таблиц типа "перетащить и отпустить" (копировать, переместить, вставить и т. д.).

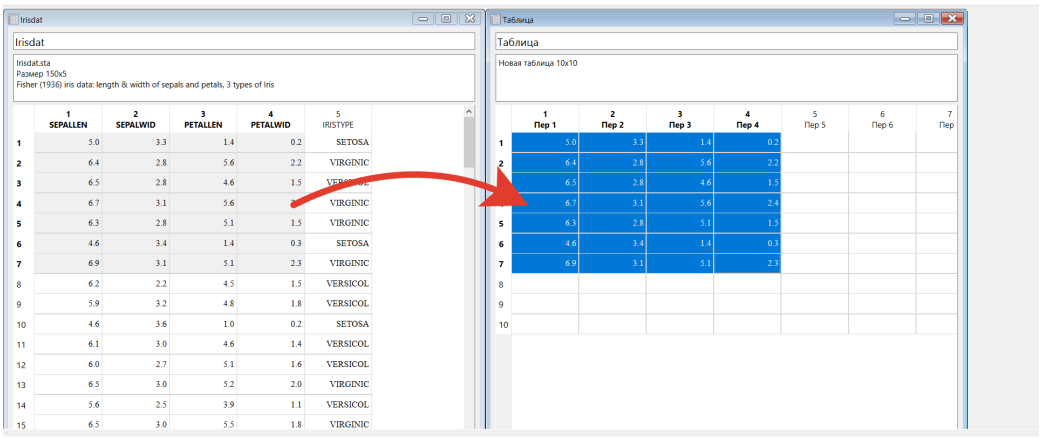
Перемещение блока внутри таблицы. Вы можете переместить блок данных в электронной таблице, указав границу выделения, и перетащить его на новое место.



Копирование блока внутри таблицы. Для копирования блока зажмите Alt, укажите границу выделения, затем перетащите выбранный фрагмент на нужное место.



Копирование и перемещение блока между таблицами. Вы можете переместить или скопировать блок данных в другую открытую таблицу аналогично копированию и перемещению блока внутри таблиц.

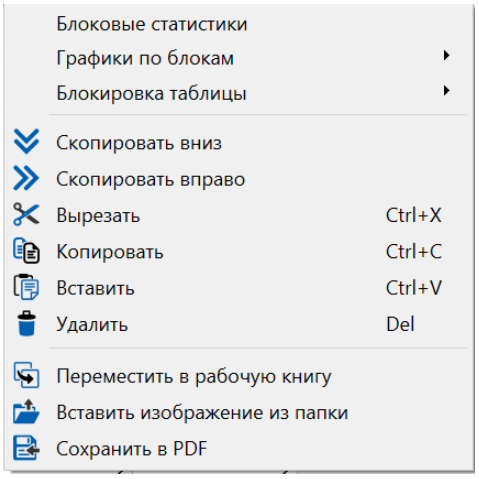


Добавление изображений

Копирование из папки. Для того, чтобы вставить изображение из папки в таблицу данных необходимо:

С помощью контекстного меню:

1. Открыть контекстное меню (нажатием правой кнопки мыши по **выделенной** ячейке).



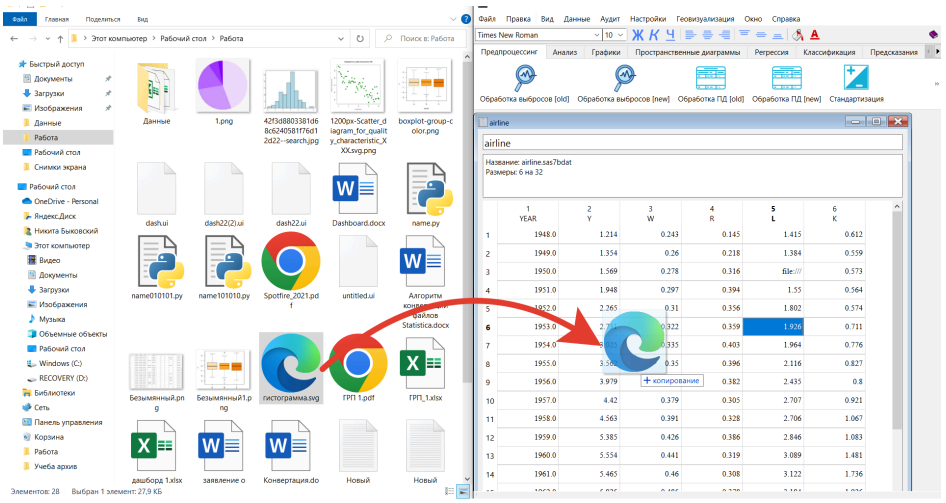
2. Нажать Вставить изображение из папки.

3. Выбрать необходимый файл. Нажать кнопку Открыть.

С помощью операции Перетащить и отпустить:

1. Выберите необходимый файл.

2. С помощью мышки перетащите его на нужную ячейку в ПО СтатСофт.



Копирование по ссылке. Для того, чтобы вставить изображение по ссылке необходимо:

С помощью копирования:

1. Скопировать URL изображения

2. Вставить URL в нужную ячейку

С помощью операции Перетащить и отпустить:

1. Выбрать необходимое изображение

2. С помощью мышки перетащите его на нужную ячейку в ПО СтатСофт.

Формат данных (.sts)

Формат ПО СтатСофт (*.sts) – это бинарный формат, предназначенный для представления табличных данных. Данный формат позволяет сохранять таблицу данных ПО СтатСофт, включая шрифты, размеры столбцов и строк, выравнивание, формат данных, а также изображения.

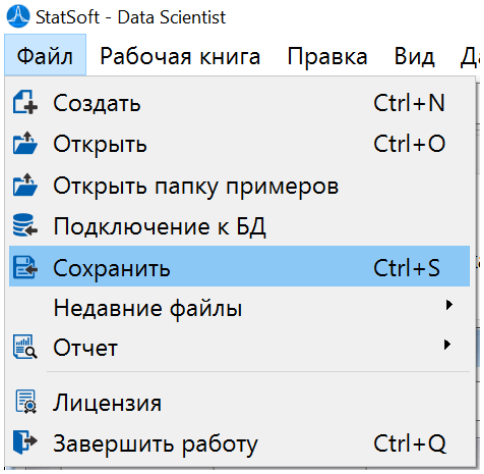
Открытие файлов sts возможно только в ПО СтатСофт, поэтому можно быть уверенным, что данные заблокированных таблиц не будут изменены.

Шифрование данных формата .sts

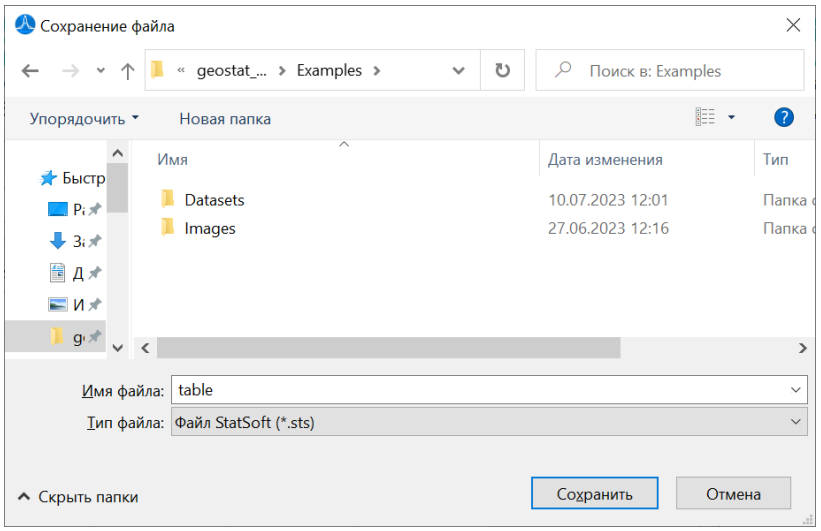
При сохранении файлов данных в формате ПО СтатСофт (.sts) программа предоставляет пользователю возможность зашифровать свои данные. Зашифрованные таблицы может открыть только пользователь, которому известен пароль, заданный при шифровании таблицы.

Процесс шифрования файла данных выглядит следующим образом:

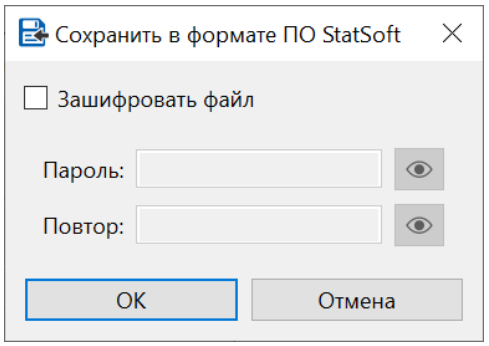
1. Выберите команду Сохранить в меню Файл



2. При сохранении задайте имя файла и выберите для него формат .sts

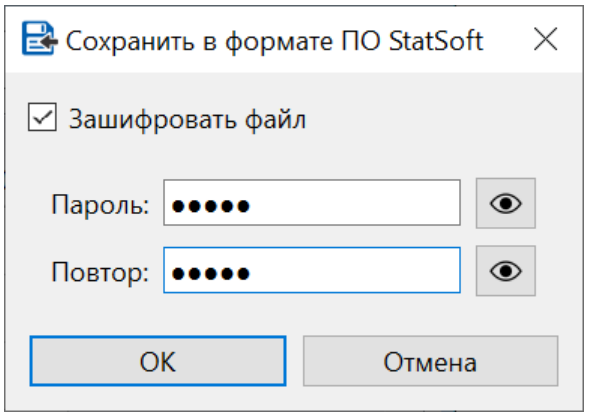


3. Откроется окно с предложением зашифровать ваш файл:

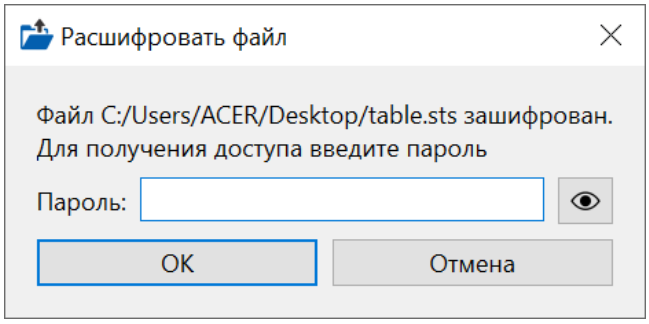


По умолчанию галочка в графе Зашифровать файл не стоит, поэтому, если шифрование вам не нужно, просто нажмите кнопку ОК.

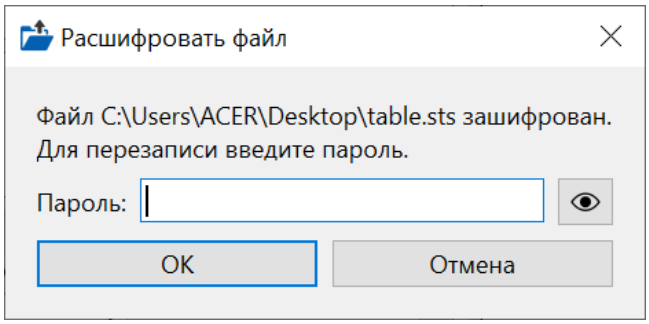
4. Если же вы хотите зашифровать ваш файл, поставьте галочку в графе Зашифровать файл и дважды введите секретный пароль. Вы можете нажать на символ  , если хотите сделать пароль видимым. После этого нажмите кнопку ОК



5. Теперь ваш файл зашифрован и сохранен. В дальнейшем, если кто-нибудь попытается открыть данный файл в ПО СтатСофт, ему необходимо будет ввести пароль для разблокировки и открытия файла.

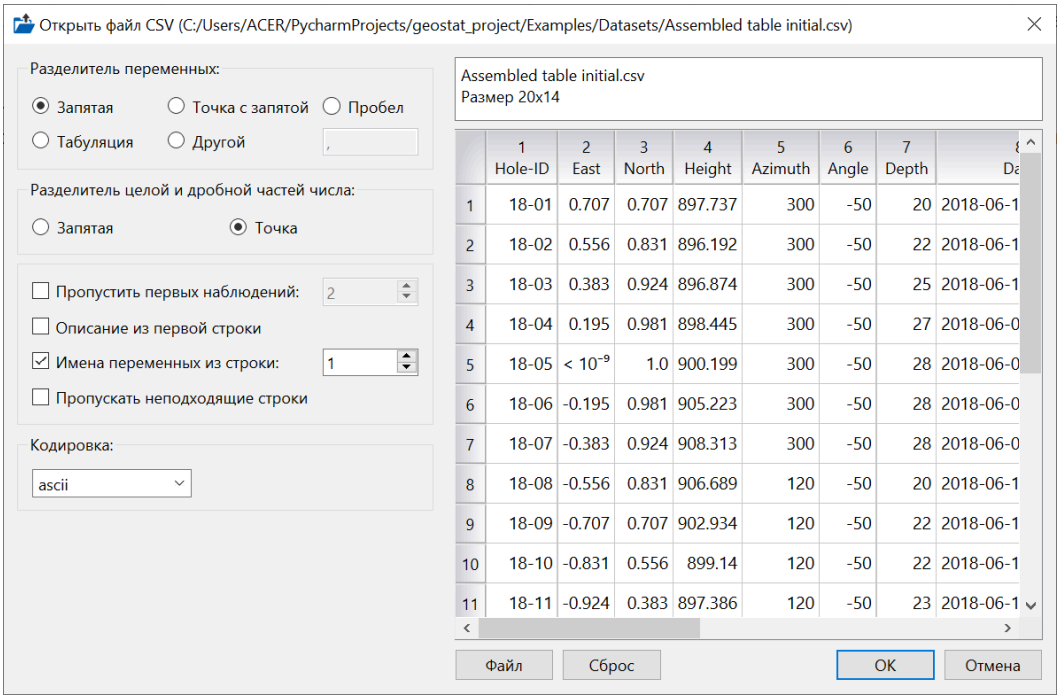


6. Также при попытке перезаписать уже существующий зашифрованный файл необходимо будет сначала ввести пароль:



Импорт текстовых и CSV файлов

Выберите пункт Открыть в меню Файл. В появившемся диалоге выберите текстовый и CSV файл (.txt, .csv) и нажмите кнопку Открыть. Перед вами откроется диалог Импорт файла.



Разделитель переменных. В блоке Разделитель переменных укажите специальный символ (или последовательность символов), который будет интерпретирован как разделительный. Вы можете выбрать один из стандартных четырех типов разделителей - Запятая, Точка с запятой, Пробел, Табуляция - или указать другой символ в текстовом поле, при выборе опции Другой.

Разделитель целой и дробной частей числа. Укажите в этом поле специальный символ, который должен быть интерпретирован как разделитель целой и дробной частей числа (Запятая или точка).

Пропустить первых наблюдений. Укажите в этом поле количество первых строк, которые должны быть пропущены при импорте файла.

Описание из первой строки. Установите этот флажок для считывания названий переменных из первой строки текстового файла в описание таблицы.

Имена переменных из строки. Укажите в этом поле номер строки для считывания названий переменных из строки текстового файла.

Пропускать неподходящие строки. Установите этот флажок для пропуска неподходящих строк.

Кодировка. Кодировка открываемого файла определяется автоматически. При необходимости изменить кодировку это можно сделать в блоке Кодировка: выберите одну из списка [utf-8, ansi, ascii, cp 1251, cp 1252, utf-16, koir-8] или укажите название вручную.

Файл. Нажмите кнопку Файл для обновления окна предварительного просмотра. В окне просмотра отобразятся последние параметры импорта.

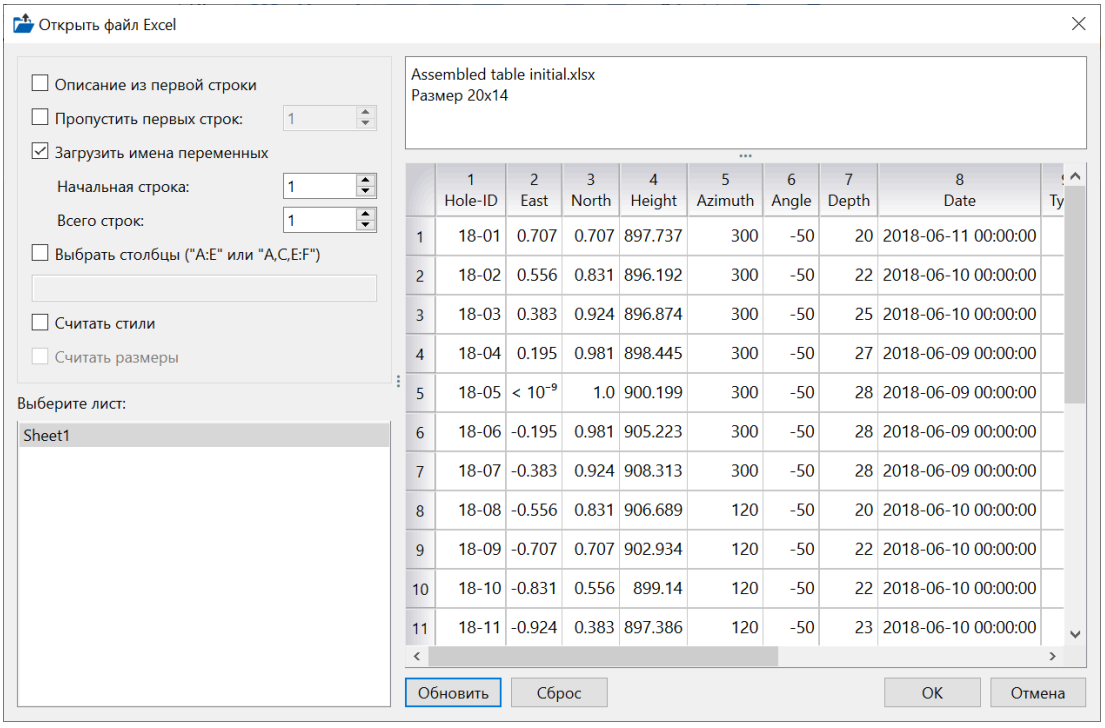
Сброс. Нажмите кнопку Сброс для восстановления настроек по умолчанию.

ОК. Нажмите кнопку ОК для импорта файла с текущими настройками.

Отмена. Нажмите кнопку Отмена для выхода из диалога без импорта файла.

Импорт файлов Excel

Выберите пункт Открыть в меню Файл. В появившемся диалоге выберите Книга Excel (.xlsx, .xls) и нажмите кнопку Открыть. Перед вами откроется диалог Импорт файла.



	1 Hole-ID	2 East	3 North	4 Height	5 Azimuth	6 Angle	7 Depth	8 Date
1	18-01	0.707	0.707	897.737	300	-50	20	2018-06-11 00:00:00
2	18-02	0.556	0.831	896.192	300	-50	22	2018-06-10 00:00:00
3	18-03	0.383	0.924	896.874	300	-50	25	2018-06-10 00:00:00
4	18-04	0.195	0.981	898.445	300	-50	27	2018-06-09 00:00:00
5	18-05	$< 10^{-9}$	1.0	900.199	300	-50	28	2018-06-09 00:00:00
6	18-06	-0.195	0.981	905.223	300	-50	28	2018-06-09 00:00:00
7	18-07	-0.383	0.924	908.313	300	-50	28	2018-06-09 00:00:00
8	18-08	-0.556	0.831	906.689	120	-50	20	2018-06-10 00:00:00
9	18-09	-0.707	0.707	902.934	120	-50	22	2018-06-10 00:00:00
10	18-10	-0.831	0.556	899.14	120	-50	22	2018-06-10 00:00:00
11	18-11	-0.924	0.383	897.386	120	-50	23	2018-06-10 00:00:00

Описание из первой строки. Установите этот флажок для считывания названий переменных из первой строки Excel листа в описание таблицы.

Пропустить первых строк. Укажите в этом поле количество первых строк, которые должны быть пропущены при импорте файла.

Загрузить имена переменных. Укажите в этих полях номера строк для считывания названий переменных из строк Excel листа.

Выбрать столбцы. Укажите в этом поле номера столбцов для считывания из Excel листа.

Выберите лист. В блоке Выберите лист укажите необходимый лист (листы) для импорта.

Обновить. Нажмите кнопку Обновить для обновления окна предварительного просмотра. В окне просмотра отобразятся последние параметры импорта.

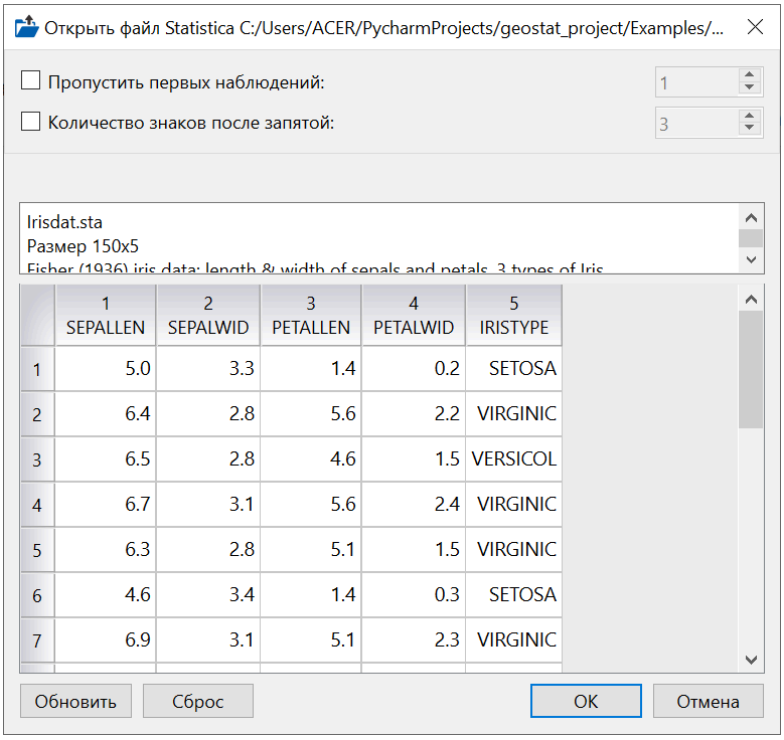
Сброс. Нажмите кнопку Сброс для восстановления настроек по умолчанию.

ОК. Нажмите кнопку ОК для импорта файла с текущими настройками.

Отмена. Нажмите кнопку Отмена для выхода из диалога без импорта файла.

Импорт SAS, SPSS, Statistica файлов

Выберите пункт Открыть в меню Файл. В появившемся диалоге выберите SAS (.sas7bdat, .xpt), SPSS (.sav) или Statistica (.sta) и нажмите кнопку Открыть. Перед вами откроется диалог Импорт файла.



Пропустить первых наблюдений. Укажите в этом поле количество первых строк, которые должны быть пропущены при импорте файла.

Количество знаков после запятой. Укажите количество знаков после запятой для считывания из исходного файла.

Обновить. Нажмите кнопку Обновить для обновления окна предварительного просмотра. В окне просмотра отобразятся последние параметры импорта.

Сброс. Нажмите кнопку Сброс для восстановления настроек по умолчанию.

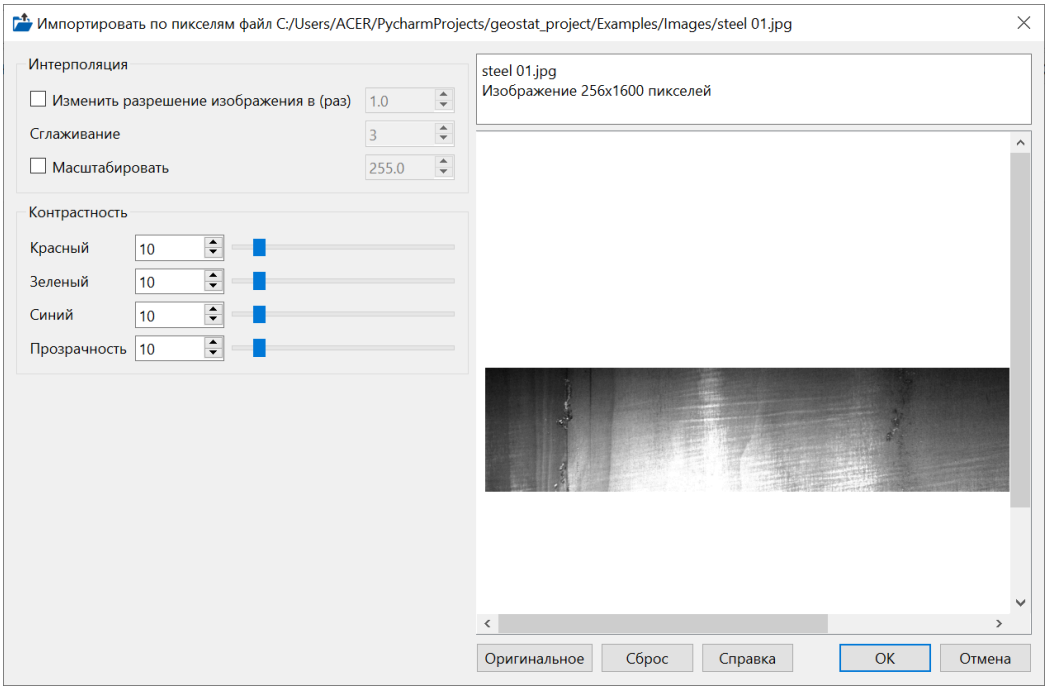
OK. Нажмите кнопку OK для импорта файла с текущими настройками.

Отмена. Нажмите кнопку Отмена для выхода из диалога без импорта файла.

Импорт изображений

Импорт изображений позволяет импортировать изображения по пикселям. Предварительно изображение преобразуется к монохромному и далее импортируется в ПО СтатСофт.

Чтобы выполнить импорт, выберите пункт Открыть в меню Файл. В появившемся диалоге выберите Изображение и нажмите кнопку Открыть. Перед вами откроется диалог Импортировать по пикселям. Выберите необходимые настройки и нажмите ОК.



Интерполяция:

1. *Изменить разрешение изображения в (раз).* В блоке Изменить разрешение изображения в (раз) укажите коэффициент с помощью которого разрешение импортируемого изображения увеличится или уменьшится в n раз.
2. *Сглаживание.* В блоке Сглаживание укажите степень полиномиальной поверхности для интерполяции (0 - *Ближайший сосед*, 1 - *Линейная интерполяция*, 2–5 - *Нелинейная интерполяция*).
3. *Масштабировать.* В блоке Масштабировать укажите коэффициент, на который будет умножено значение каждого пикселя.

Контрастность. В блоке Контрастность укажите необходимые веса цветов (красный, зеленый, синий) и прозрачность для оптимальной настройки импортируемого изображения.

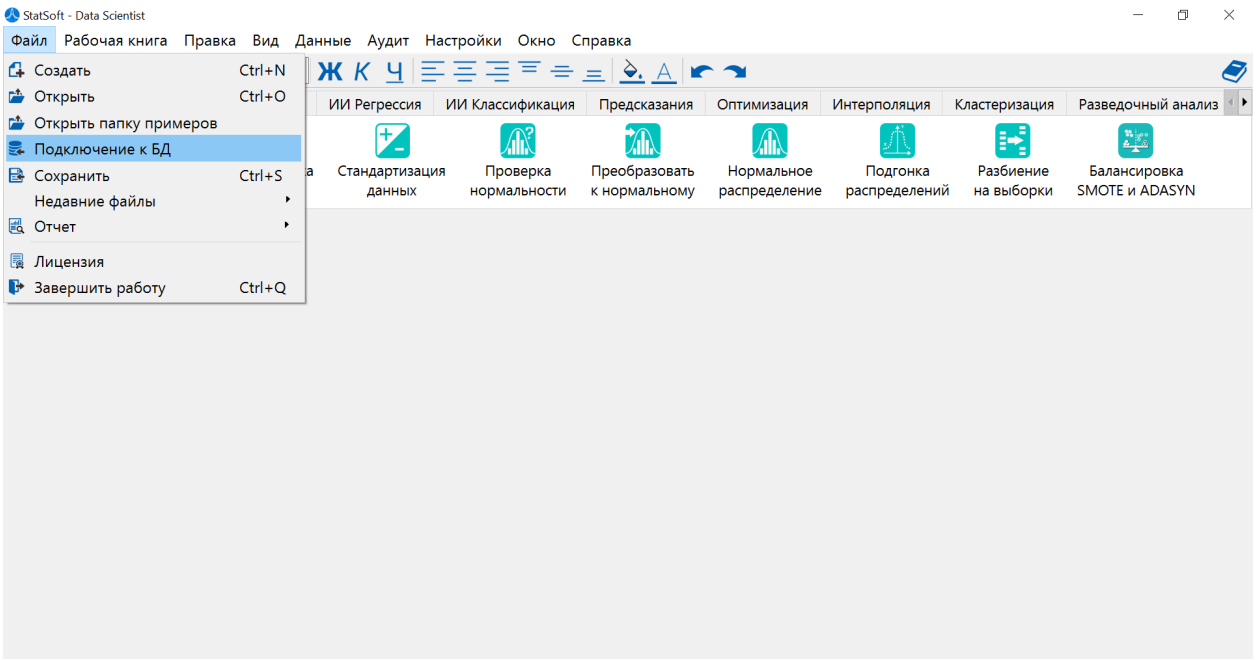
Оригинальное <-> Импортируемое. Кнопка Оригинальное <-> Импортируемое позволяет переключаться между оригинальным изображением и импортируемым (к которому были применены настройки)

Сброс. Кнопка Сброс задает все настройки по умолчанию.

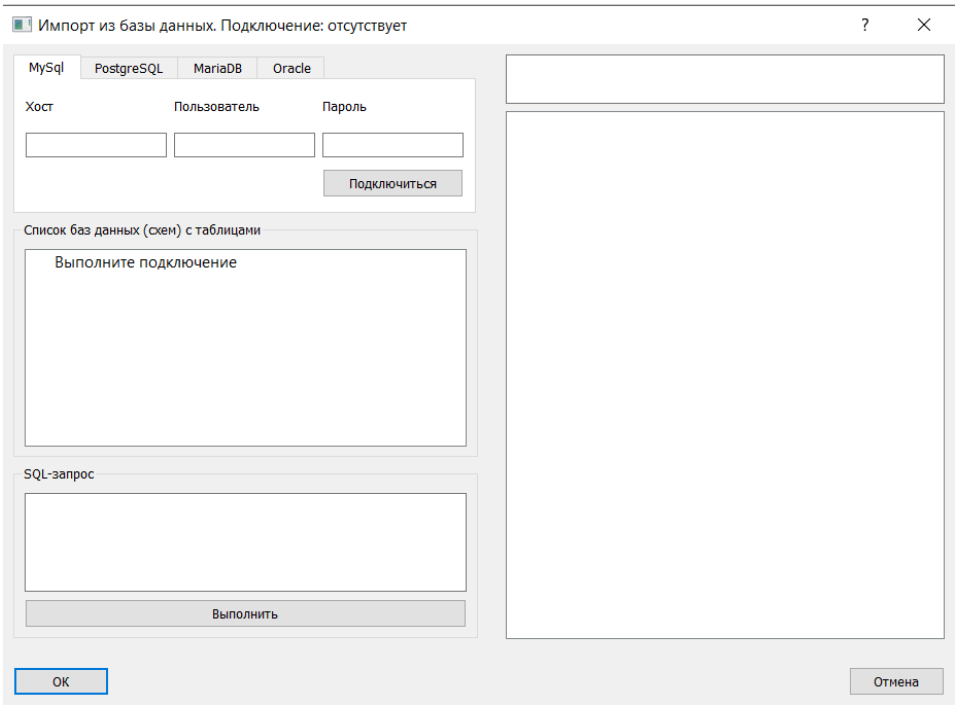
Подключение к базам данных

Возможности ПО СтатСофт позволяют подключаться к 4 видам баз данных, а именно MariaDB, MySql, PostgreSQL и Oracle. Подключившись к базе данных, можно делать импорт таблиц и схем, а также выполнять SQL-запросы, результат которых также можно импортировать в ПО СтатСофт.

Алгоритм подключения к БД. Рассмотрим алгоритм подключения к БД на примере БД MySql. Для подключения к серверу MySql необходимо нажать Файл -> Импорт из MySql

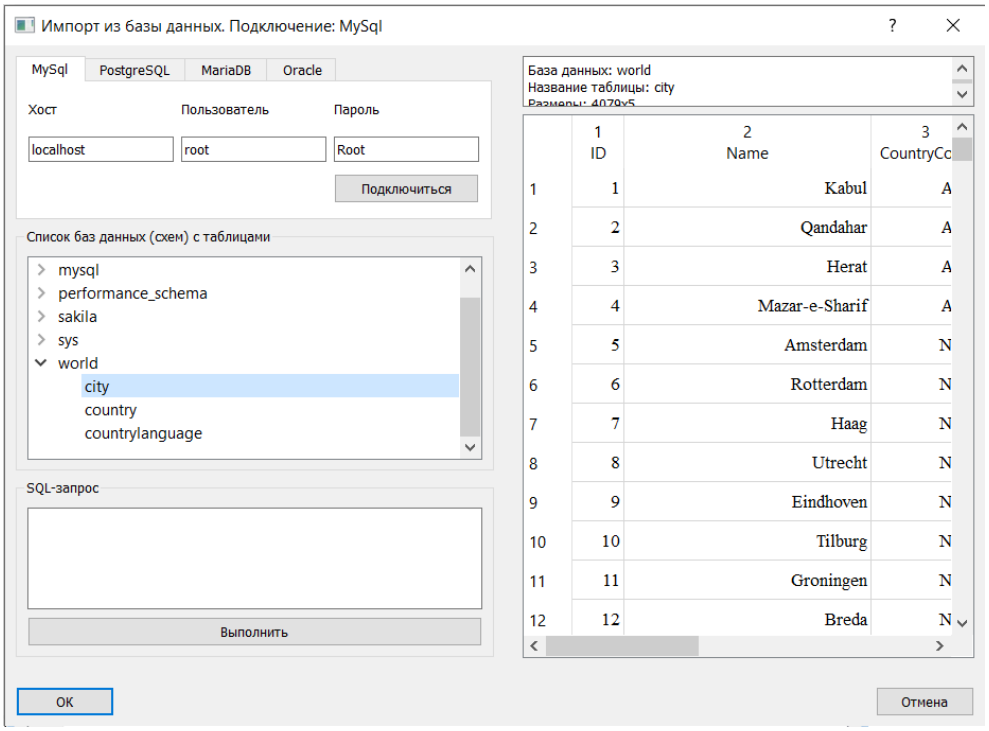


Откроется следующий диалог:

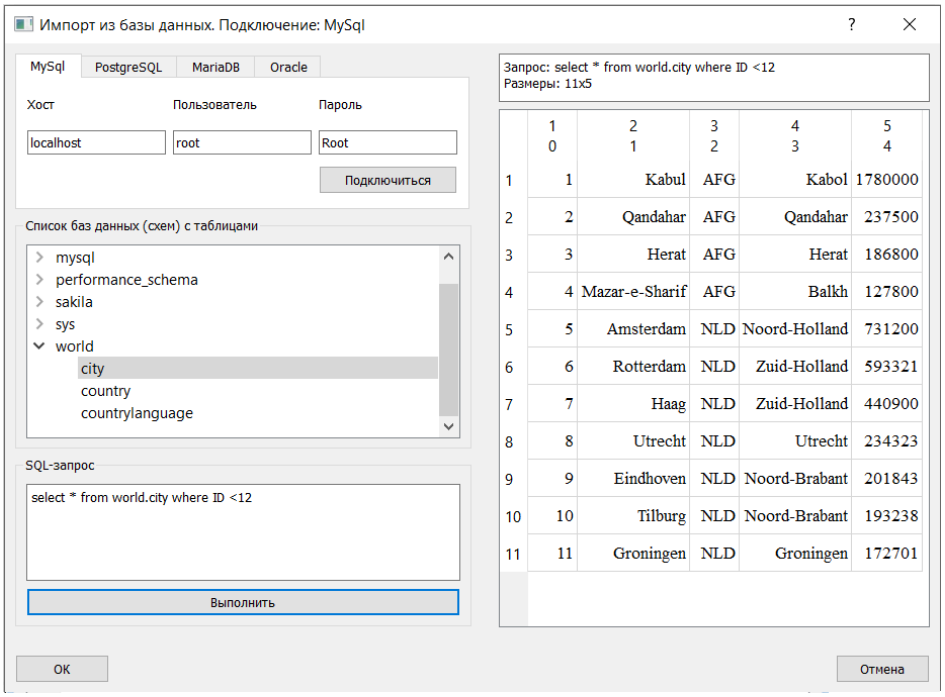


В открывшимся окне необходимо выбрать MySql, ввести IP адрес сервера MySql, имя пользователя и пароль пользователя, после чего нажать Подключиться.

После подключения, в списке слева появятся все базы данных и их таблицы доступные данному пользователю. Для просмотра необходимой таблицы нужно выбрать и нажать на нее.



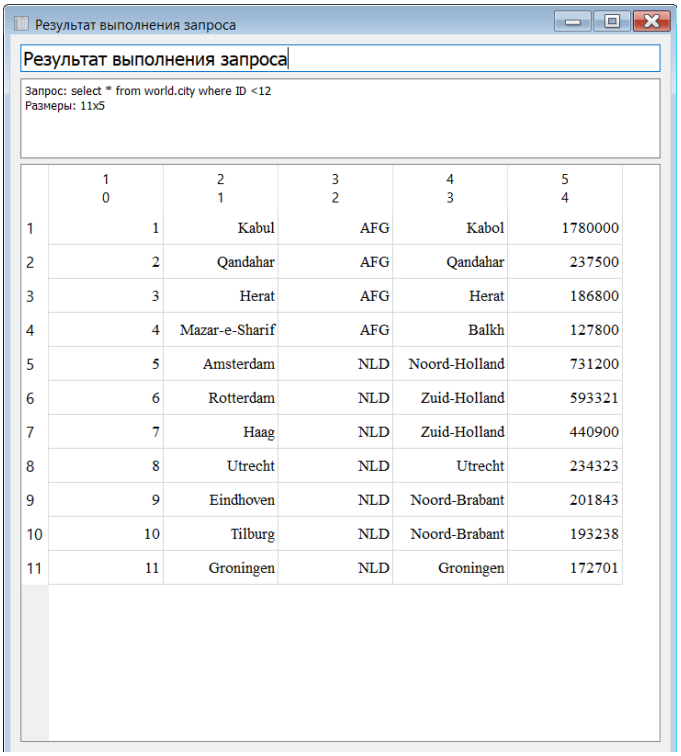
Чтобы выполнить SQL-запрос необходимо в поле для ввода ниже, написать SQL-запрос и нажать кнопку Выполнить.



СтатСофт®

Для импорта получившийся таблицы в ПО СтатСофт необходимо нажать кнопку ОК.

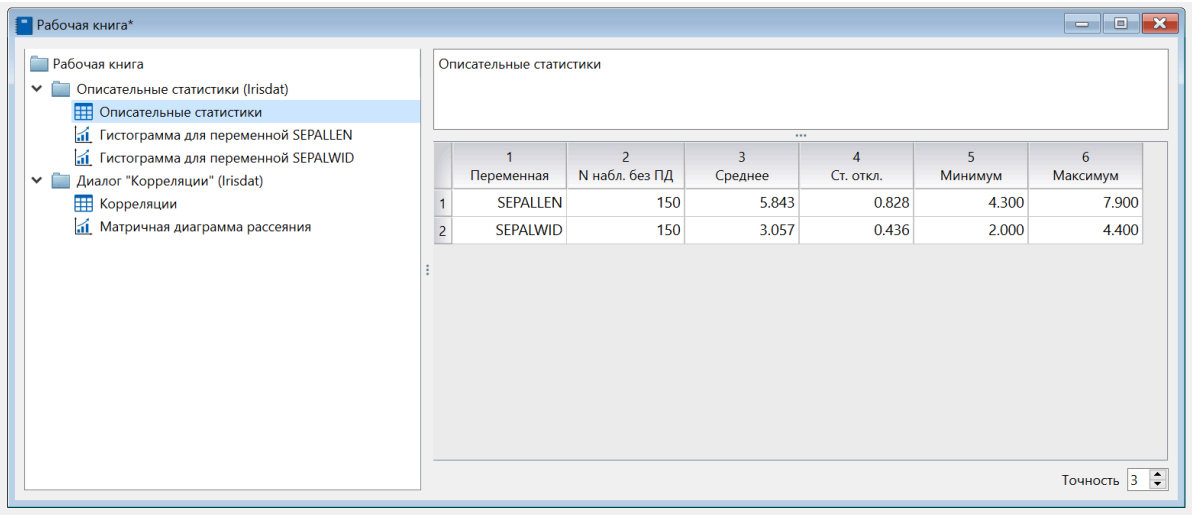
Таблица успешно импортировалась, теперь ей можно пользоваться в ПО СтатСофт.



	1 0	2 1	3 2	4 3	5 4
1	1	Kabul	AFG	Kabol	1780000
2	2	Qandahar	AFG	Qandahar	237500
3	3	Herat	AFG	Herat	186800
4	4	Mazar-e-Sharif	AFG	Balkh	127800
5	5	Amsterdam	NLD	Noord-Holland	731200
6	6	Rotterdam	NLD	Zuid-Holland	593321
7	7	Haag	NLD	Zuid-Holland	440900
8	8	Utrecht	NLD	Utrecht	234323
9	9	Eindhoven	NLD	Noord-Brabant	201843
10	10	Tilburg	NLD	Noord-Brabant	193238
11	11	Groningen	NLD	Groningen	172701

Обзор рабочих книг

Рабочие книги являются стандартным способом управления выводом ПО СтатСофт. В Рабочей книге каждый документ (например, Таблица данных или График) представлен в виде отдельной вкладки.

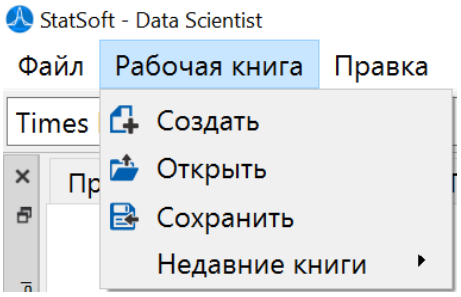


Говоря техническим языком, Рабочие книги являются оптимизированными контейнерами, которые позволяют эффективно обрабатывать большое количество документов. Все документы упорядочены в иерархическом виде с использованием отдельных папок или узлов документов (по умолчанию, для каждого нового Анализа создается отдельный узел).

Каждая Рабочая книга имеет две области: навигационное дерево, которое находится слева, и область просмотра документов справа. В дереве Рабочей книги (навигационное дерево) может находиться несколько различных узлов, которые используются для создания логической структуры (например, все результаты Анализа, созданные в текущем проекте, помещаются в отдельную папку).

Операции с Рабочими книгами

В ПО СтатСофт основные функции для взаимодействия с Рабочими книгами вынесены в отдельное меню Рабочая книга. Здесь можно открывать, сохранять и просматривать недавние Рабочие книги.



Рассмотрим подробнее, как происходит работа с Рабочими книгами.

Создание. Новая Рабочая книга создается в программе автоматически при получении любого результата анализа (таблицы или графика), при условии, что в программе не было открыто других Рабочих книг.

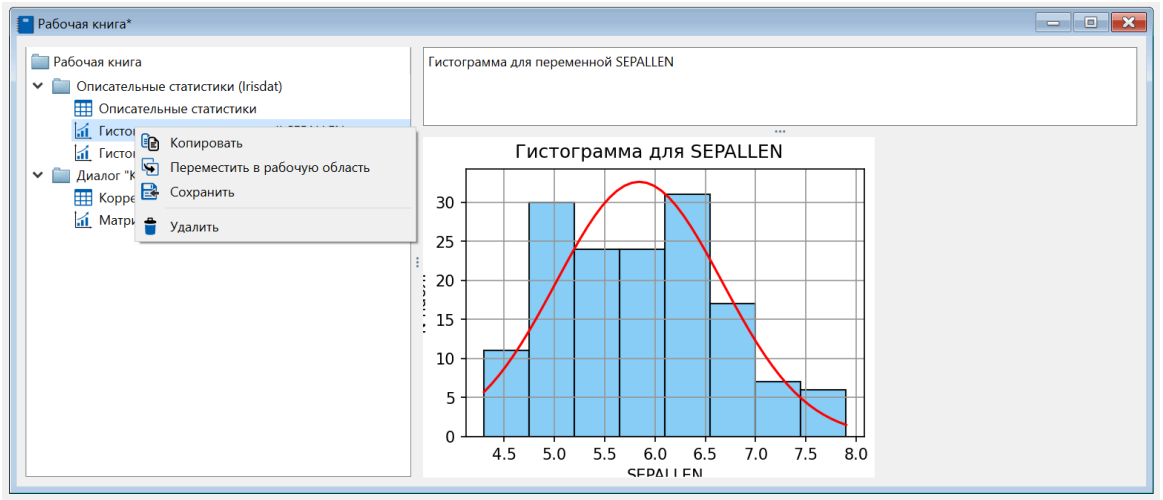
Сохранение. Выберите опцию Сохранить в меню Рабочая книга для сохранения Рабочей книги.

Загрузка. Выберите опцию Открыть в меню Рабочая книга для открытия существующей Рабочей книги с вашего компьютера. Обратите внимание, что одновременно в рабочей области может находиться только одна Рабочая книга, поэтому, если до этого была открыта другая, она будет удалена.

Недавние книги. Выберите опцию Недавние книги в меню Рабочая книга для просмотра списка недавно использовавшихся Рабочих книг. При желании можно также загрузить одну из них в рабочую область.

Операции с элементами Рабочих книг

Вы можете выполнять ряд операций с элементами Рабочей книги через контекстное меню, для вызова которого необходимо правой кнопкой мыши нажать на любой элемент Рабочей книги в навигационном дереве:



В программе доступны следующие опции взаимодействий:

Удалить. Выберите команду Удалить в контекстном меню, чтобы удалить выбранный элемент из Рабочей книги.

Сохранить. Выберите команду Сохранить в контекстном меню, чтобы сохранить выбранный элемент Рабочей книги как отдельную таблицу или график.

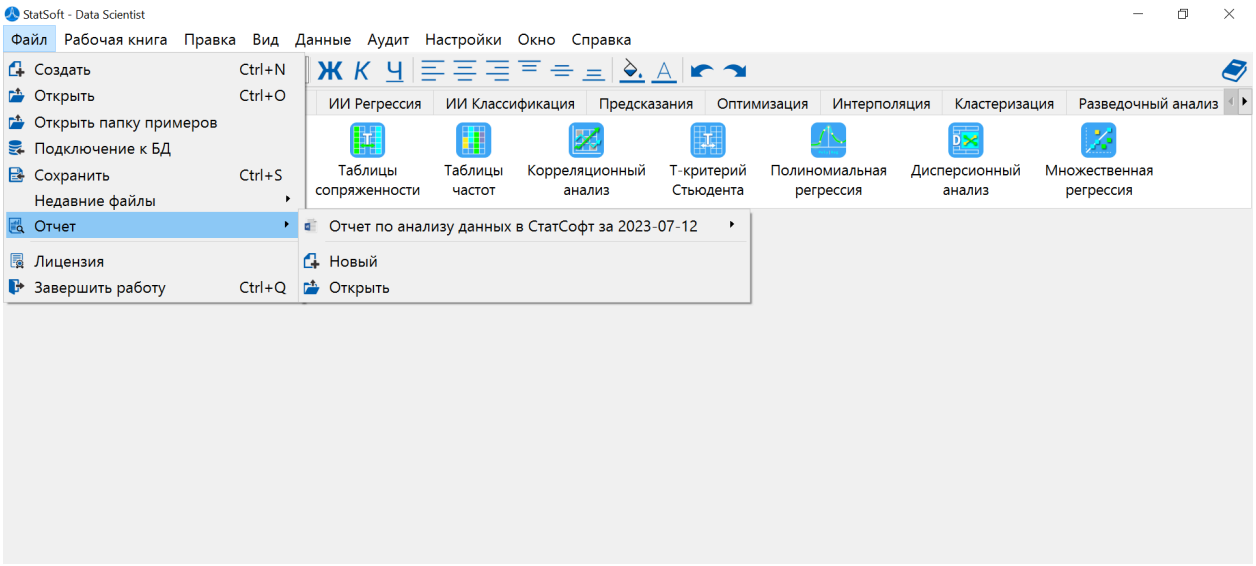
Переместить в рабочую область. Выберите команду Переместить в рабочую область. в контекстном меню, чтобы поместить копию таблицы в рабочую область программы для дальнейшей работы с ней.

Копировать (для графиков). Выберите команду Копировать в контекстном меню, чтобы скопировать выбранный график Рабочей книги в буфер обмена.

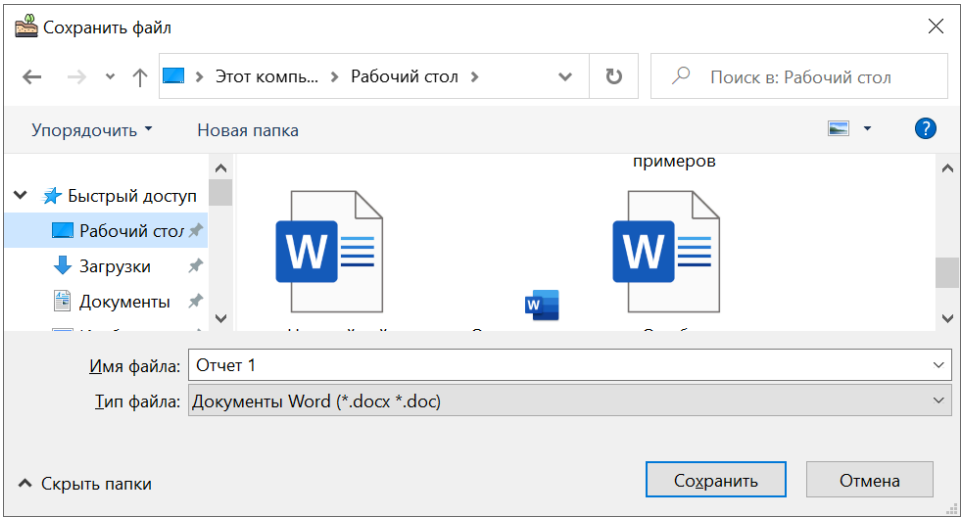
Отчеты

Отчет ПО СтатСофт — это тип документа в формате Word (.docx), который предоставляет простые, но, тем не менее, мощные инструменты для создания Отчетов. В Отчет можно добавлять все табличные и другие результаты, полученные в процессе проведения анализов.

Создание отчета. Чтобы создать новый отчет, необходимо в меню Файл – Отчет выбрать Новый.

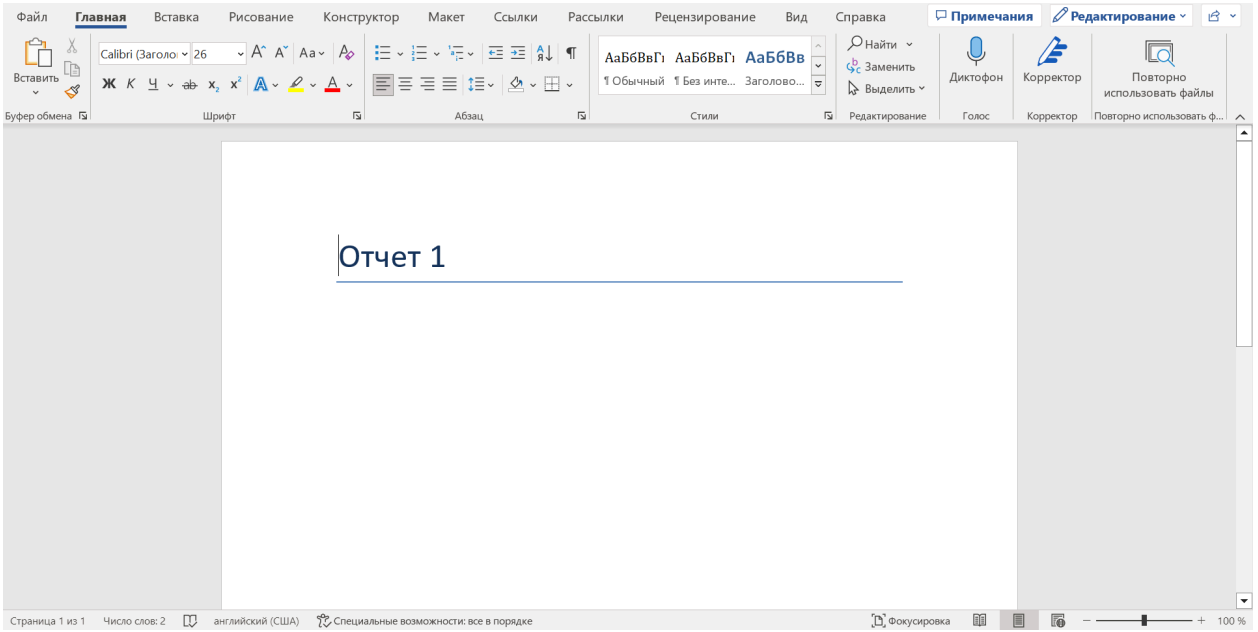


После этого откроется диалоговое окно, в котором необходимо выбрать место для сохранения нового отчета и при необходимости задать его название:

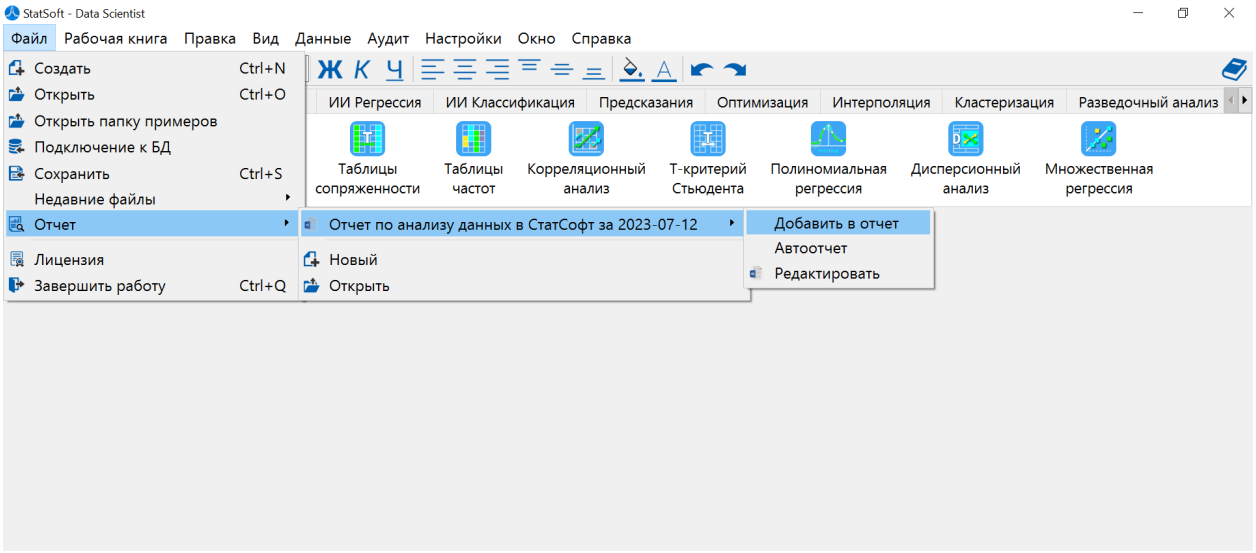


Нажмите кнопку Сохранить и ваш новый отчет будет создан.

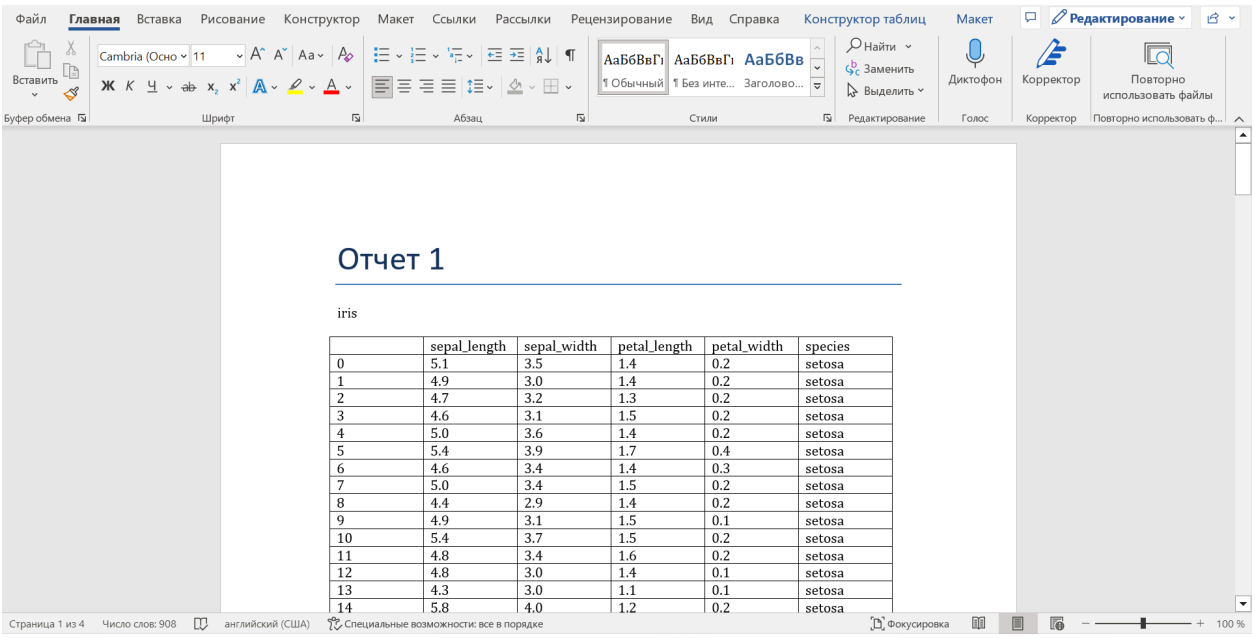
Добавление элемента в отчет. По умолчанию при создании отчета в него помещается активная таблица данных, если таковая имеется, либо отчет остается пустым:



При необходимости добавить в отчет новый элемент, его необходимо сделать активным, после чего в меню Файл – Отчет выбрать нужный отчет и нажать на опцию Добавить в отчет:



После этого отчет будет изменен:



	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
6	4.6	3.4	1.4	0.3	setosa
7	5.0	3.4	1.5	0.2	setosa
8	4.4	2.9	1.4	0.2	setosa
9	4.9	3.1	1.5	0.1	setosa
10	5.4	3.7	1.5	0.2	setosa
11	4.8	3.4	1.6	0.2	setosa
12	4.8	3.0	1.4	0.1	setosa
13	4.3	3.0	1.1	0.1	setosa
14	5.8	4.0	1.2	0.2	setosa

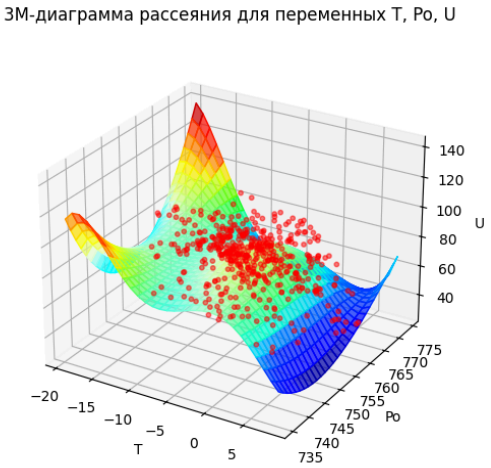
Автоотчет. При выборе этой опции в меню Файл – Отчет – (указать нужный отчет) все новые результаты анализов (такие как таблицы и графики будут заноситься в выбранный отчет автоматически.

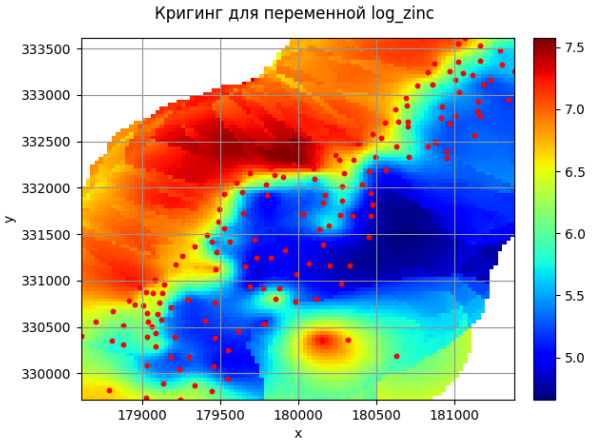
Редактирование отчета. При необходимости вручную внести любые изменения в отчет, в меню Файл – Отчет – (указать нужный отчет) выберите данную опцию. При этом выбранный

отчет будет открыт в отдельном окне в приложении Word, где вы сможете внести в него правки, используя любые инструменты программы Word по своему усмотрению.

Графика

Краткое руководство содержит разделы, посвященные описанию различных типов графиков, краткий обзор распространенных приложений, примеры типичного применения и описание различных свойств соответствующих типов графиков. Более подробное введение в науку (и искусство) графического представления данных можно найти в многочисленных работах, посвященных этому вопросу.



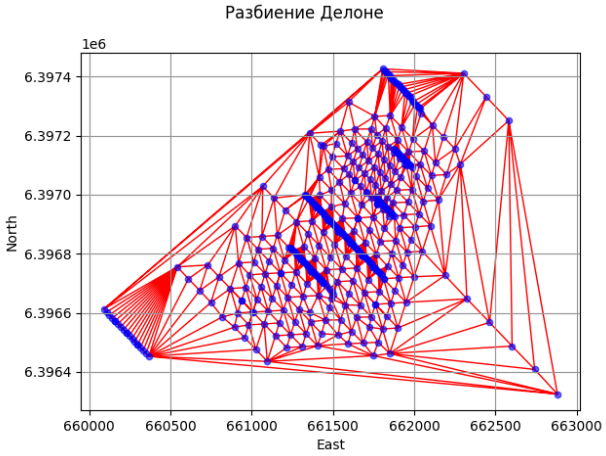


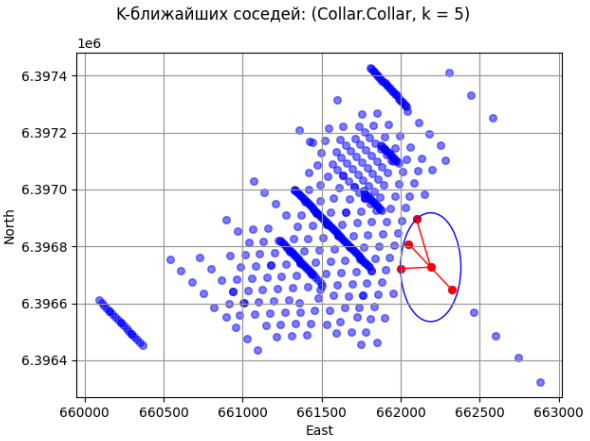
Среди источников, наиболее полно охватывающих спектр проблем, связанных с построением графиков, можно рекомендовать следующие: Buja и Tukey (1991), Chambers, Cleveland, Kleiner и Tukey (1983), Cleveland (1984, 1985), Kolata (1984), Tufte (1983, 1990), Tukey (1977), а также Tukey и Tukey (1981).

Большое число статей по специальным вопросам, связанным с выбором графического представления статистических данных, публикуется в сборнике докладов *Ежегодного семинара отделения статистической графики Американской статистической ассоциации*.

Большинство статистических графиков, предлагаемых по умолчанию, соответствуют установившимся соглашениям, которые описаны в литературе по использованию графиков в статистике и технике, или отвечают общепринятым стандартам, обычно используемым в большинстве научных журналов (например, в журнале SCIENCE).

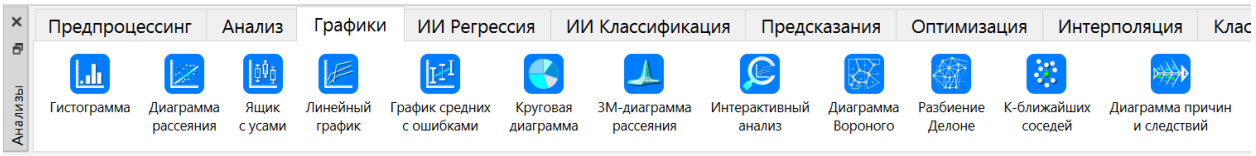
В то же время средства ПО СтатСофт позволяют изменять практически любые установленные по умолчанию параметры, если необходимо, чтобы они соответствовали каким-то специальным требованиям. Средства для построения графиков представляют собой гибкий инструмент, позволяющий выйти за рамки готовых шаблонов.





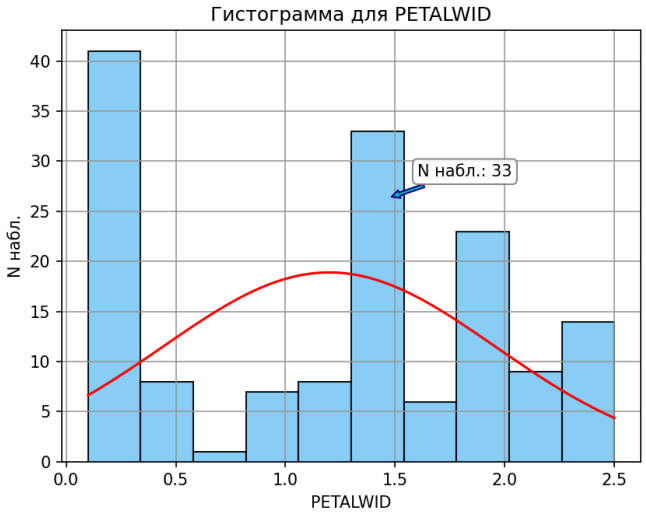
Графики

Графический анализ в ПО СтатСофт удобно проводить с использованием графических инструментов, расположенных на вкладке меню Графики.

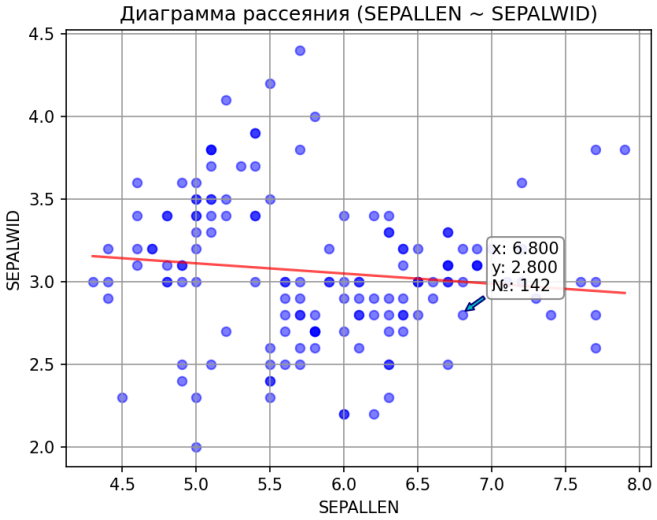


Графики в ПО СтатСофт обладают свойствами интерактивности и позволяют пользователю просто и наглядно оценить результаты анализа. При наведении курсора на тот или иной элемент графика пользователь увидит его выделение, содержащую информацию по данному элементу.

Например, при наведении курсора на столбец гистограммы, будет указано количество наблюдений, попавших в заданный интервал.



А при наведении на точку в диаграмме рассеяния, будут даны такие ее характеристики, как порядковый номер в таблице исходных данных, а также значения переменных, по которым строилась данная диаграмма рассеяния.



В ПО СтатСофт окно графика служит не только для отображения графика, но также представляет очень эффективное и сложное средство настройки графика, позволяющее изменять практически любые параметры графика (например, оси, шкалы, метки, цвета, шаблоны и т. д.) с помощью вызова окна настроек графика двойным щелчком по нему.

Виды графиков

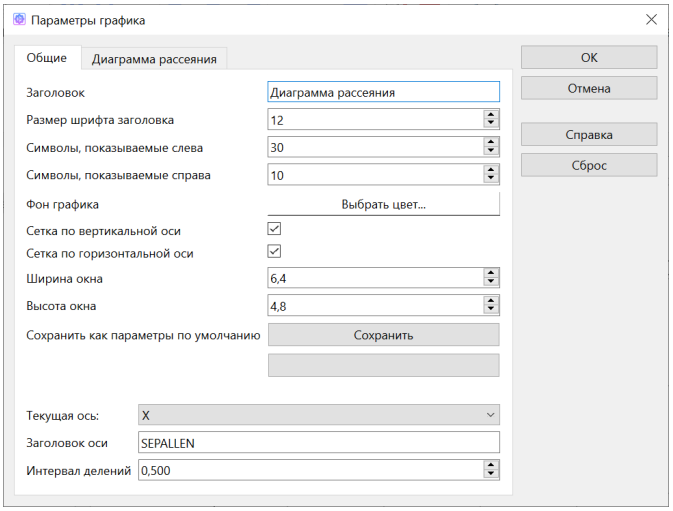
Графики в меню Графики предлагают самый большой выбор параметров среди графиков ПО СтатСофт. Заметьте, что в отличие от Графиков блоковых данных, все типы графиков из меню Графики не ограничены по значениям текущей таблицы результатов. Данные для построения этих графиков берутся непосредственно из текущей активной таблицы исходных данных.

Основные типы графиков в меню Графика включают:

- Гистограммы
- Диаграммы рассеяния
- Ящики с усами
- Линейные графики
- Графики средних с ошибками
- Круговые диаграммы
- 3M диаграммы рассеяния
- Диаграммы причин и следствий
  - Диаграммы Вороного
  - Разбиение Делоне
  - К ближайших соседей

Настройка графиков

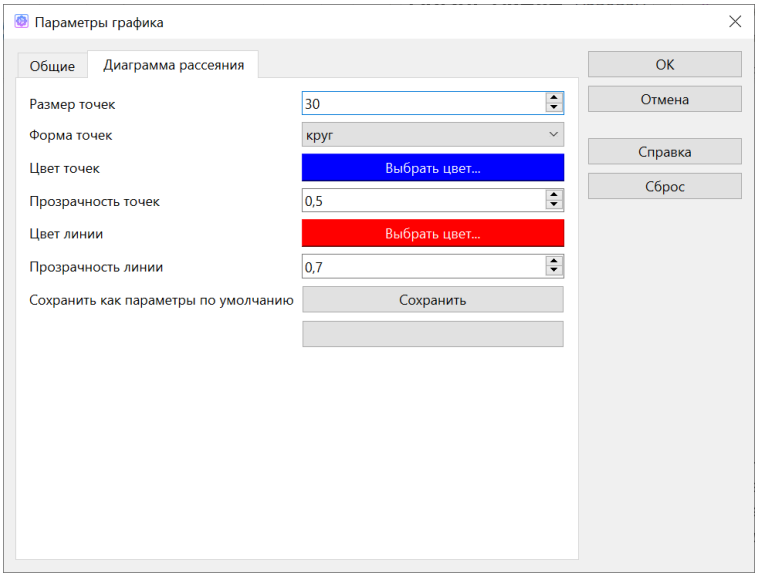
Интерактивная настройка графиков. Опции настройки графиков ПО СтатСофт охватывают множество свойств и инструментов для корректировки каждого элемента отображения графика и соответствующих данных. Эти опции располагаются в окне Параметры графика, вызвать которое можно двойным нажатием на любой элемент графика.



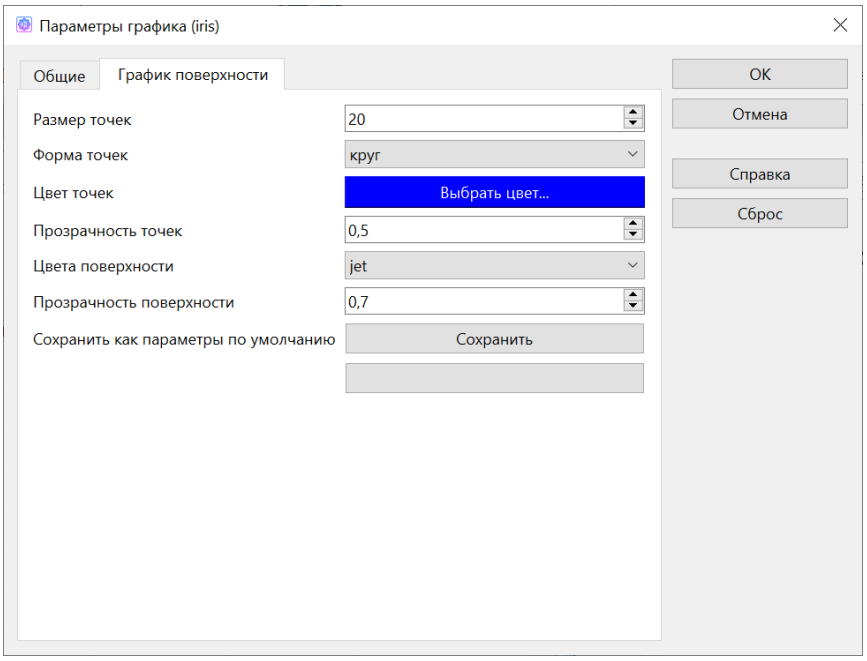
Общие и индивидуальные настройки. Окно настроек графиков содержит как опции для редактирования элементов, присущих всем графикам ПО СтатСофт (заголовок, размер заголовка, сетка и т.д.), так и опции, характерные для конкретного типа редактируемого графика.

Опции для типов графиков располагаются на отдельной вкладке в окне Параметры графика и могут включать настройки различных элементов графиков таких как размеры и цвета точек, прозрачность линий и другие.

Например, так выглядят расширенные опции для редактирования графиков типа диаграмма рассеяния:



А так для графиков типа 3D поверхность:



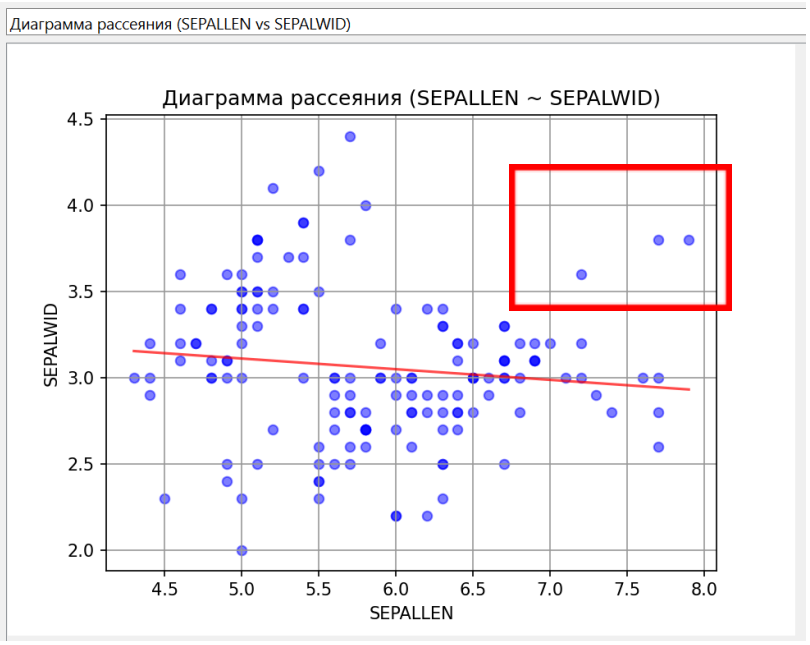
Постоянные установки и опции автоматизации. Начальные (по умолчанию) установки всех свойств можно легко изменить таким образом, чтобы даже вид и свойства графика по умолчанию будут удовлетворять вашим запросам и требовать минимального вмешательства с вашей стороны. Для этого в окне Параметры графика необходимо сохранить текущие настройки как параметры по умолчанию.

Интерактивный графический анализ

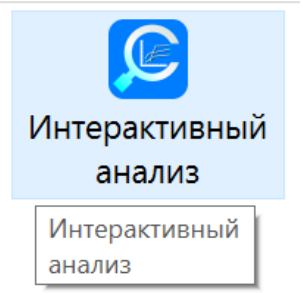
Интерактивный графический анализ является одним из методов разведочного анализа данных. Это интерактивный способ, который дает возможность выбрать на экране определенные точки данных или подмножества данных и идентифицировать их (например, пометить). Этот подход широко применяется для работы с выбросами.

Интерактивный анализ доступен для графиков типа диаграмма рассеяния и диаграмма размаха. Чтобы он стал доступен, сначала необходимо построить график одного из этих типов.

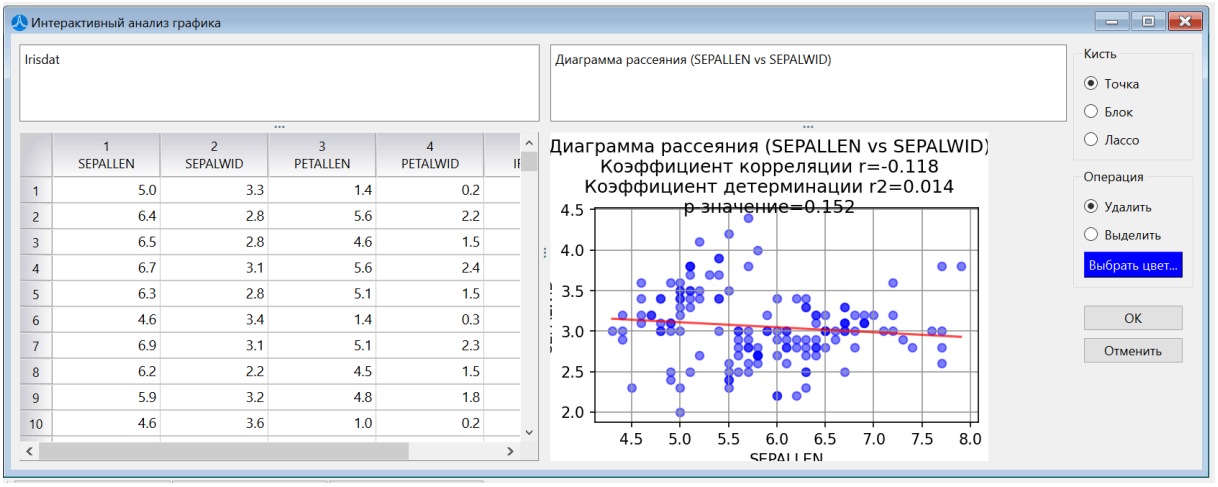
Исходный график (красным выделены точки, с которыми мы будем проводить изменения):



Для того, чтобы открыть панель Интерактивный анализ, необходимо на вкладке Графики выбрать Интерактивный анализ.



После этого откроется окно Интерактивного анализа:

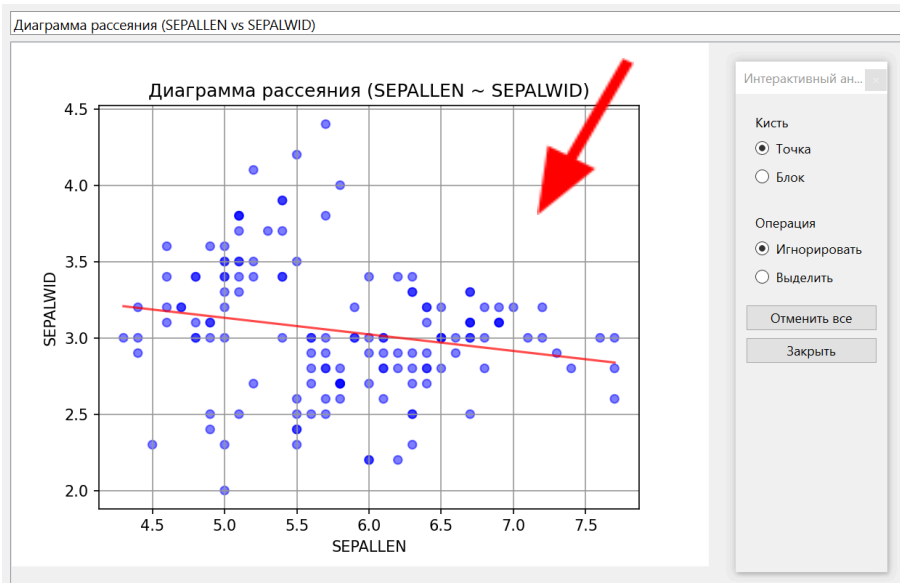


	1 SEPALLEN	2 SEPALWID	3 PETALLEN	4 PETALWID
1	5.0	3.3	1.4	0.2
2	6.4	2.8	5.6	2.2
3	6.5	2.8	4.6	1.5
4	6.7	3.1	5.6	2.4
5	6.3	2.8	5.1	1.5
6	4.6	3.4	1.4	0.3
7	6.9	3.1	5.1	2.3
8	6.2	2.2	4.5	1.5
9	5.9	3.2	4.8	1.8
10	4.6	3.6	1.0	0.2

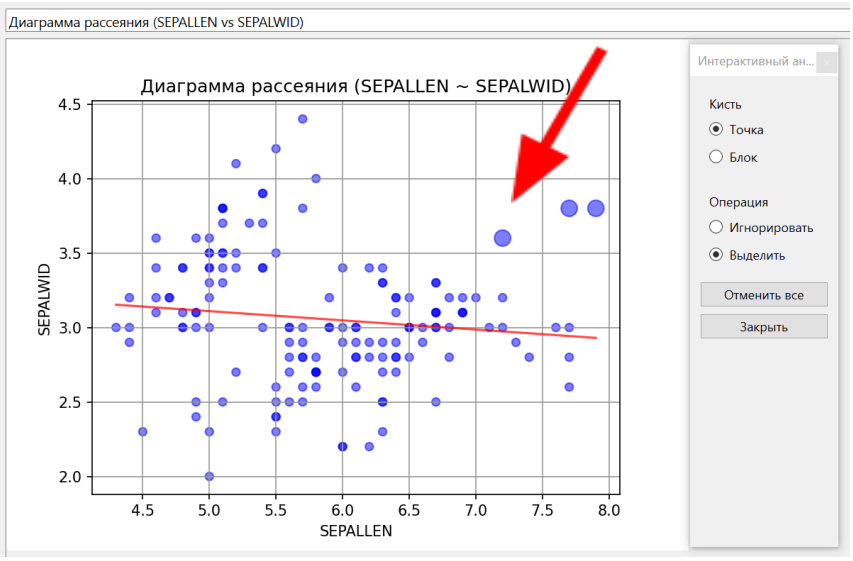
Отменить. С помощью этой кнопки можно удалить все выполненные преобразования.

Операция. В этом поле указывается операция, которая осуществляется над выделенными точками: Игнорировать или Выделить.

Игнорировать. Эта операция позволяет удалить с графика выбранные точки. При этом соответствующие этим точкам наблюдения также удаляются из таблицы исходных данных.



Выделить. Эта операция позволяет идентифицировать выбранные точки данных соответствующей меткой.



Кисть. Опции поля Кисть используются, чтобы выбрать тип кисти для выделения точек данных, которые вы хотите маркировать/удалить на графике.

Точка. В этом режиме программа позволяет выделять отдельные точки, щелкая на них курсором.

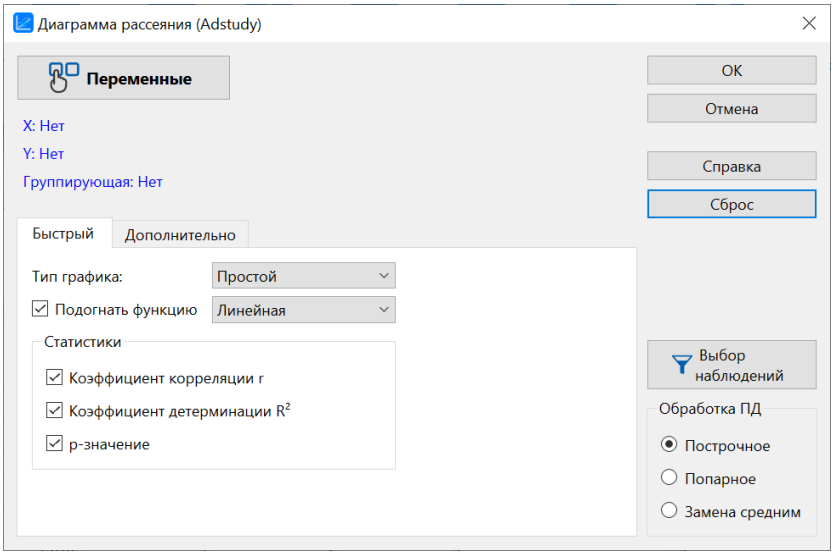
Блок. В этом режиме имеется возможность нарисовать прямоугольную область вокруг выбранного множества точек и выделить все точки внутри этой области.

Пример 1. Диаграмма рассеяния

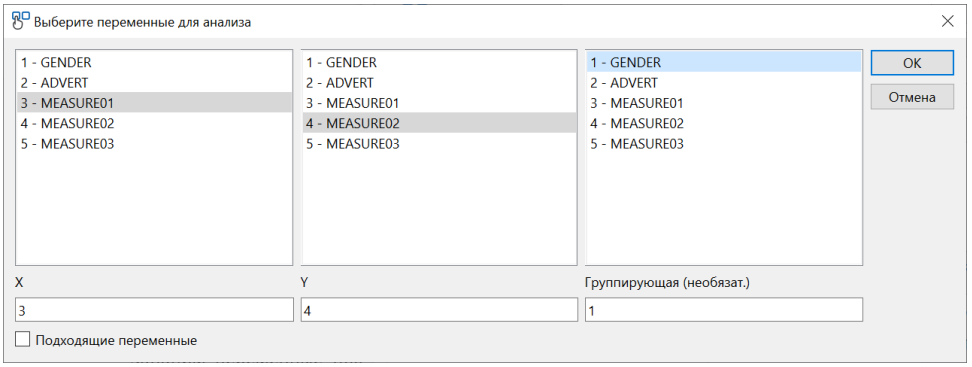
Для этого примера использовался файл Adstudy.sts, расположенный в Папке примеров. Чтобы открыть этот файл, выберите пункт меню Файл – Открыть папку примеров. Файл Adstudy.sts содержит предполагаемые результаты исследования, проведенного в гостинице для посетителей. Помимо анкетирования о предпочтительных кола-напитках, исследователи попросили посетителей оценить качество гостиницы по нескольким пунктам, таким как обслуживание комнат, чистота, работа портье и т. д.

Построение графика. Откройте файл Adstudy.sts и выберите опцию меню Графики – Диаграмма рассеяния.

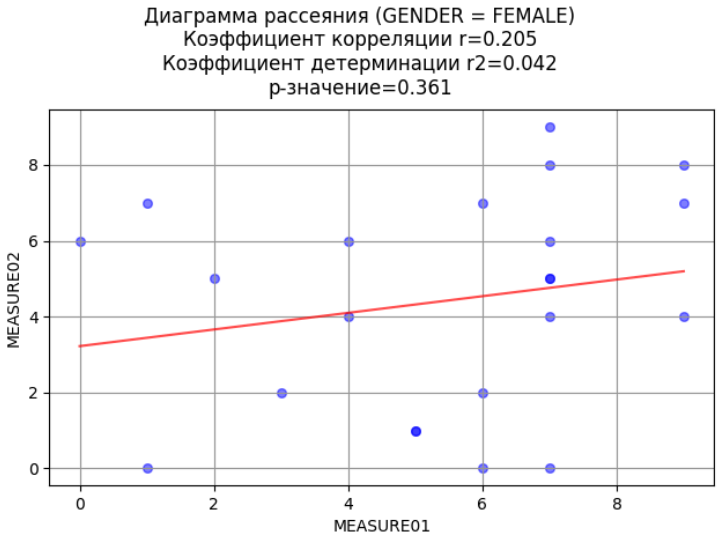
Появится диалоговое окно Диаграмма рассеяния. Здесь вы можете задать для вашего графика необходимые настройки, используя вкладки Быстрый и Дополнительно.

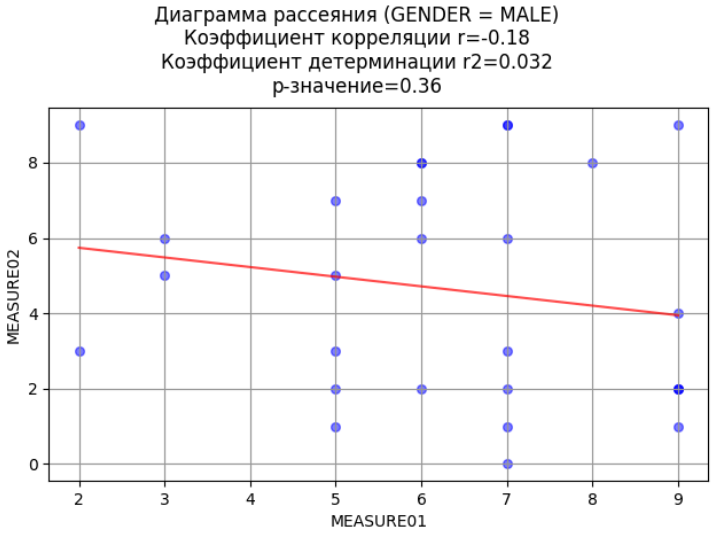


Зададим переменные для построения гистограммы. В качестве переменной по оси X выберем MEASURE01, по оси Y – MEASURE02, а в качестве необязательной группирующей переменной – GENDER. Нажмем ОК.



Далее в окне Диаграмма рассеяния также нажмем ОК. В рабочей книге появятся два графика – по одному на каждый из представленных в таблице данных гендеров.

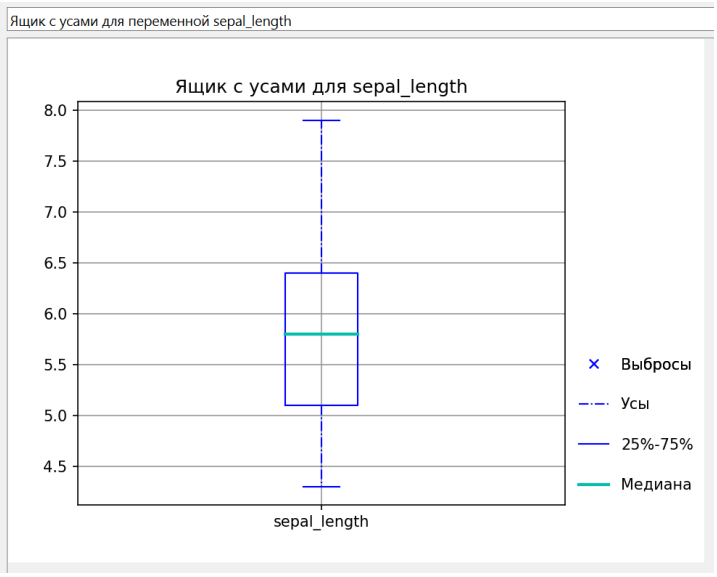




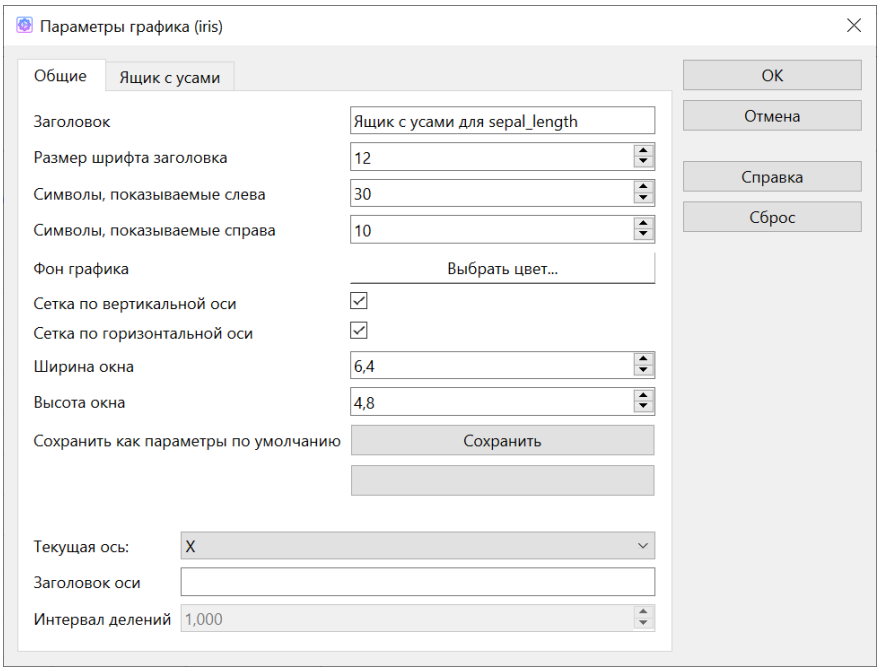
Из графиков видим, что для разных гендеров наблюдаются противоположные картины. Так, для графика GENDER = FEMALE наблюдается слабая положительная корреляция между рассматриваемыми переменными, так как для графика GENDER = MALE – слабая отрицательная.

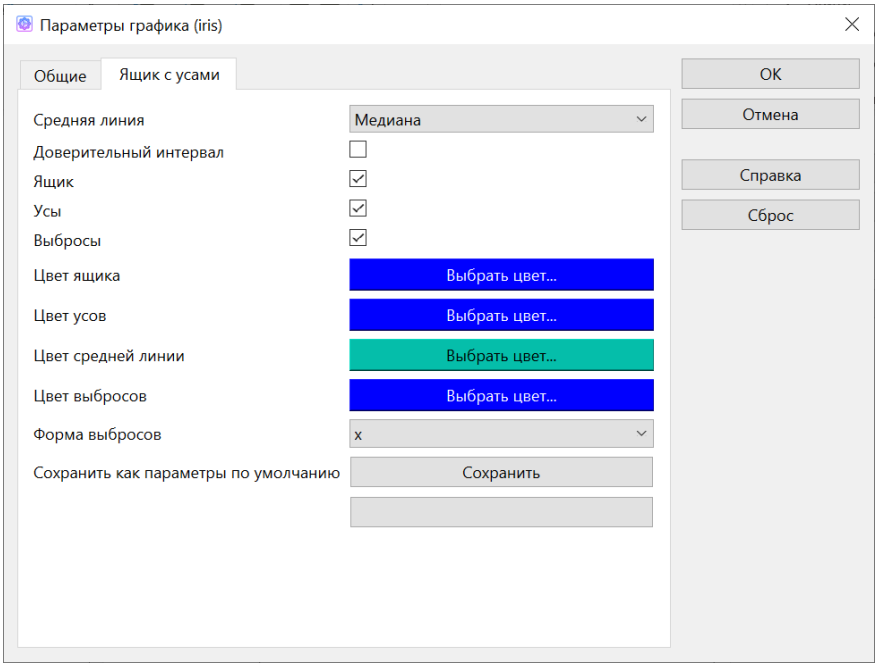
Пример 2. Настройка графика Ящик с усами

Любой график в ПО СтатСофт можно гибко настраивать под свои потребности. Рассмотрим пример настройки графика типа Ящик с усами.



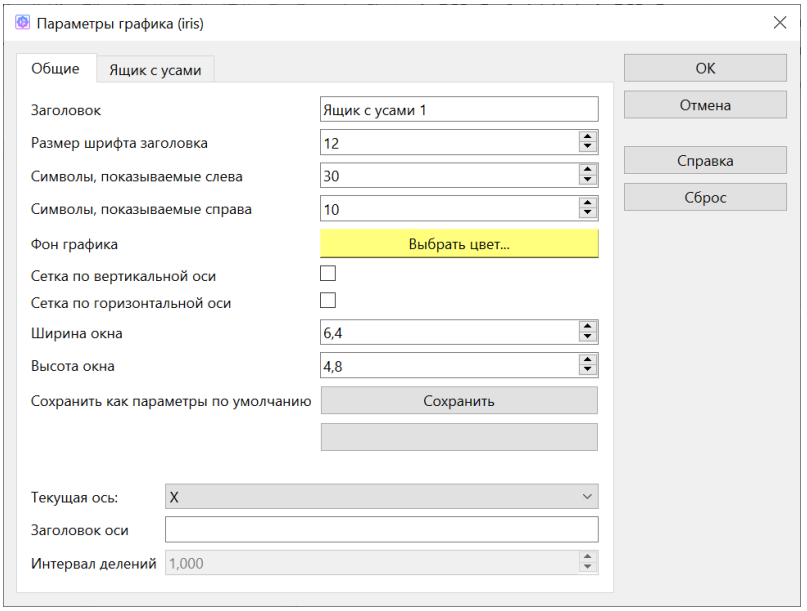
У нас есть некоторый график Ящик с усами, настройки которого мы хотим поменять. Для того, чтобы перейти в меню Параметры графика, дважды щелкнем мышью по любой области графика. Откроется следующее диалоговое окно:



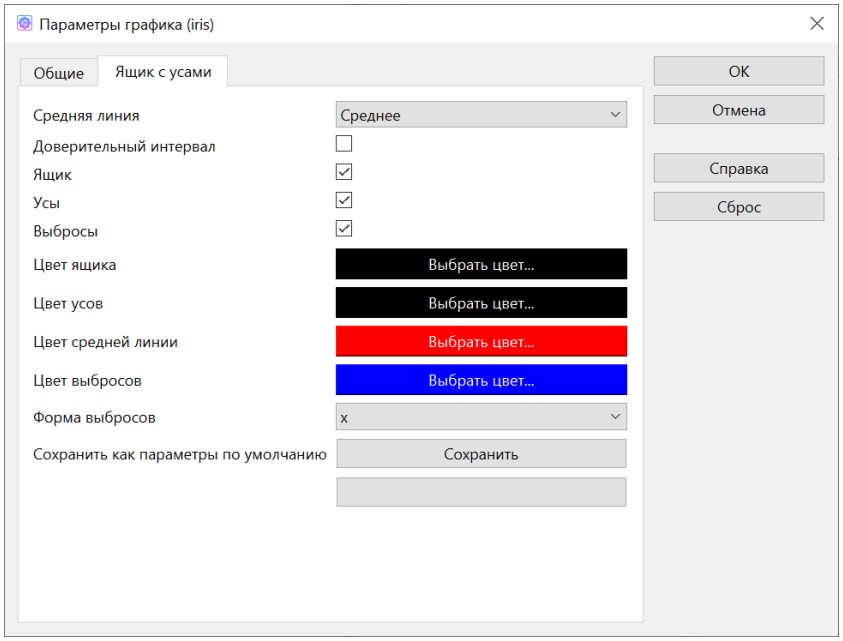


Окно Параметры графика содержит две вкладки для большинства типов графиков, которые можно построить в ПО СтатСофт: это вкладка с общими параметрами графика (присутствует для любых типов графиков) и вкладка, соответствующая типу настраиваемого графика (в данном случае Ящик с усами).

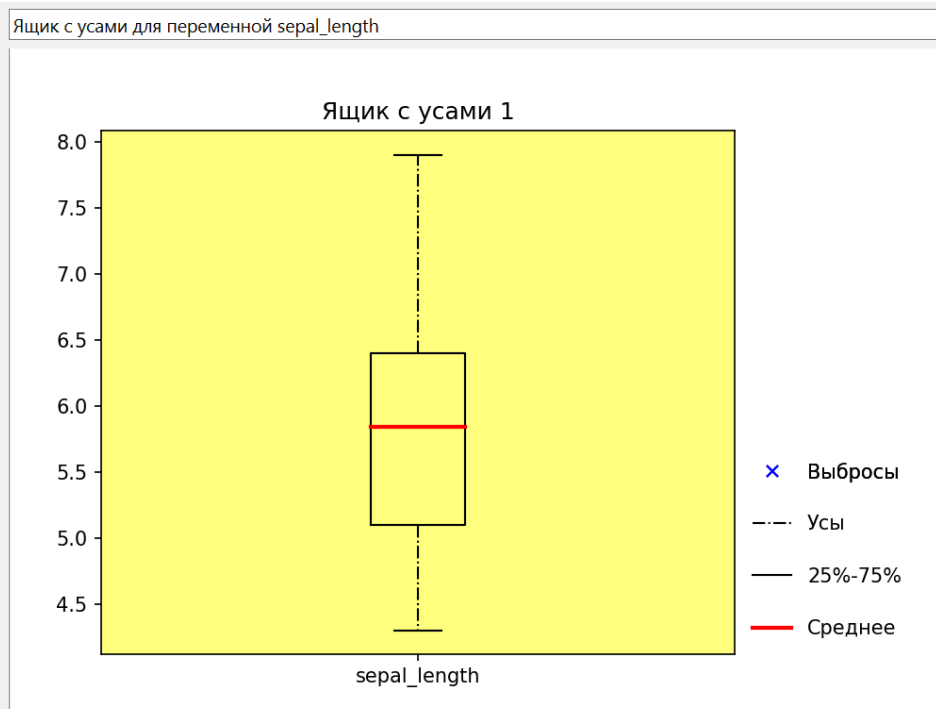
Попробуем внести изменения на обеих вкладках. На вкладке Общие поменяем название графика, цвет его фона и уберем сетку:



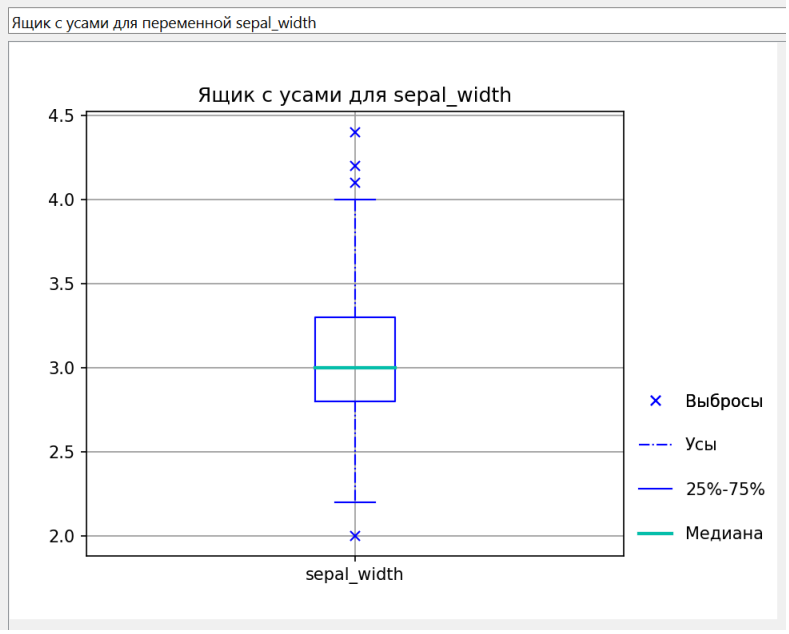
На вкладке Ящик с усами изменим тип средней линии графика с медианы на среднее, а также поменяем цвета ящика, усов и средней линии:



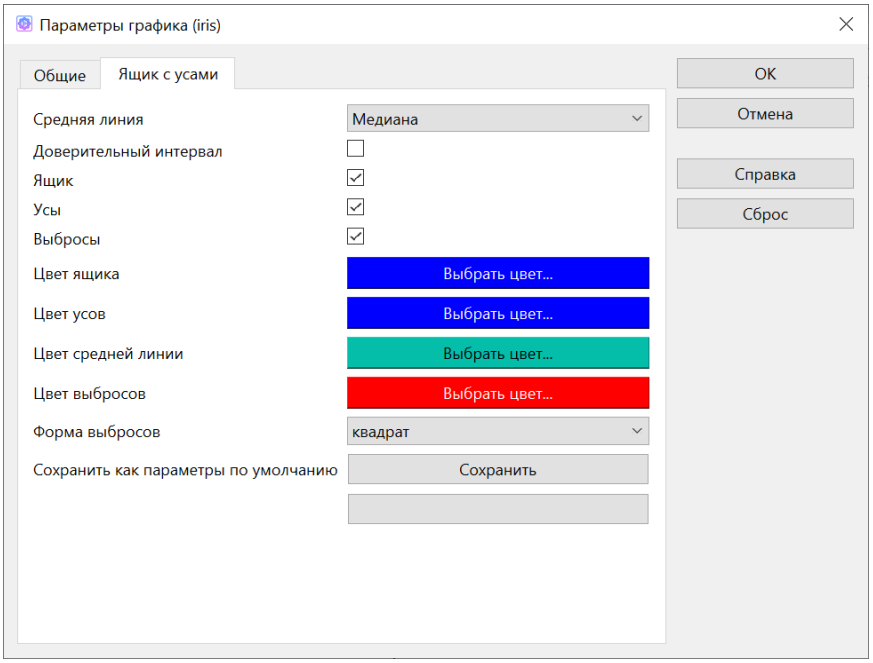
В итоге получим следующее изображение:



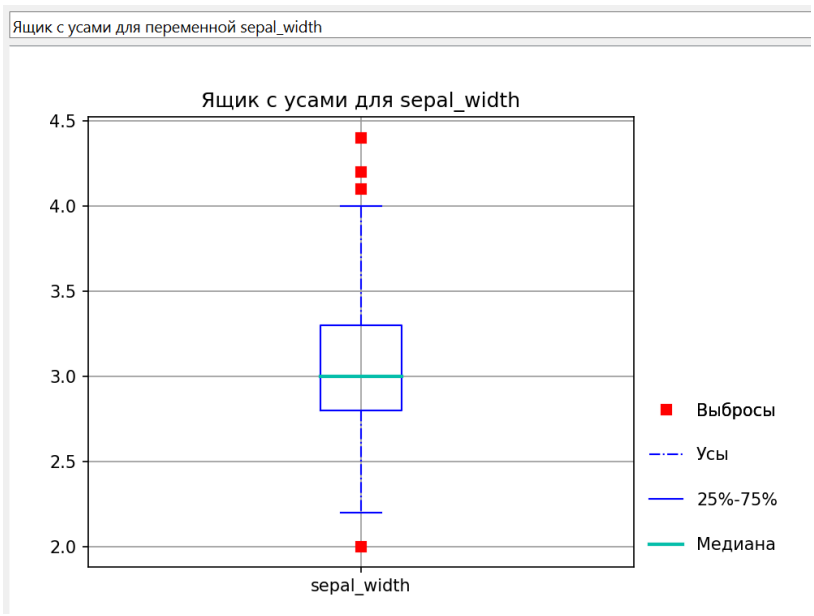
Если на графике ящик с усами есть выбросы, что бывает довольно часто, то для них тоже есть отдельные критерии настроек. Рассмотрим пример:



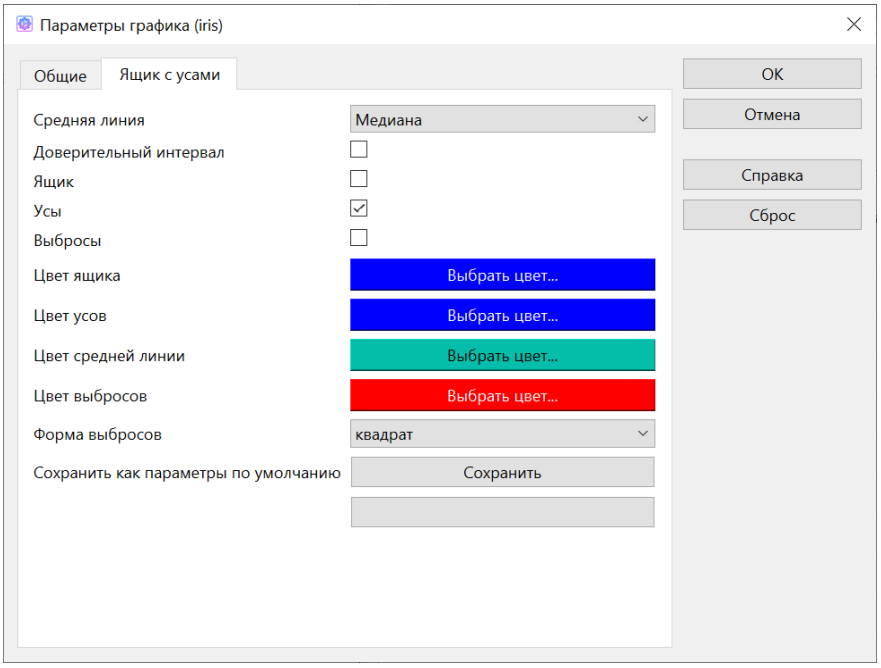
По умолчанию выбросы на данном типе графика обозначаются символом X синим цветом. Допустим, мы хотим сделать выбросы более явными. Для этого перейдем в окно Параметры графика и цвет выбросов сделаем красным, а в качестве формы укажем квадрат:



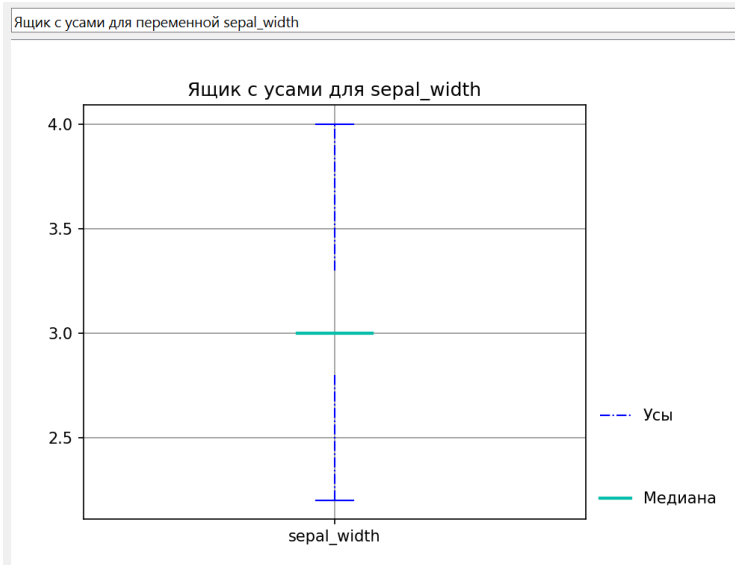
Вот что в итоге у нас получилось:



При желании мы также можем вообще скрыть выбросы на графике, как и некоторые другие его элементы. Попробуем, к примеру скрыть выбросы и “ящик”:



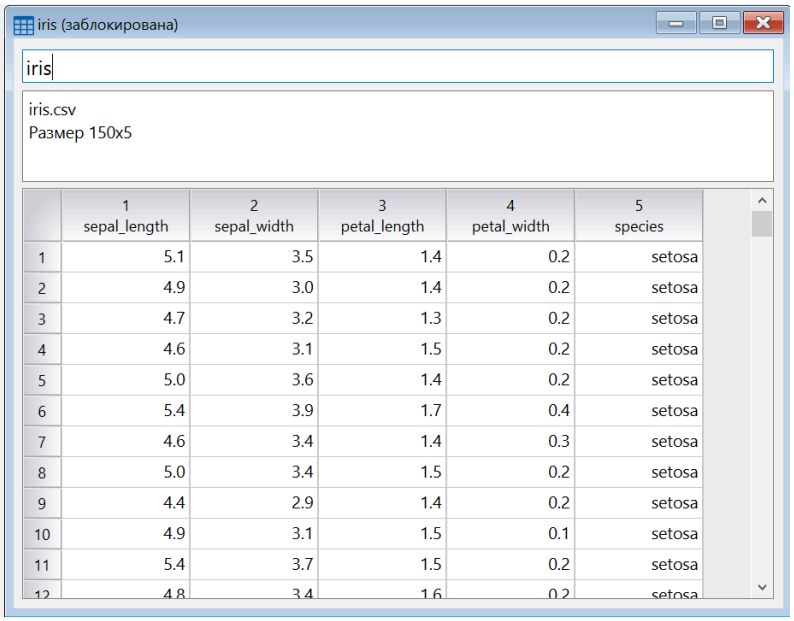
В результате получаем:



Вывод. ПО СтатСофт предоставляет инструменты для гибкой настройки всех элементов своих графиков. В зависимости от типа графика настройки для них могут меняться.

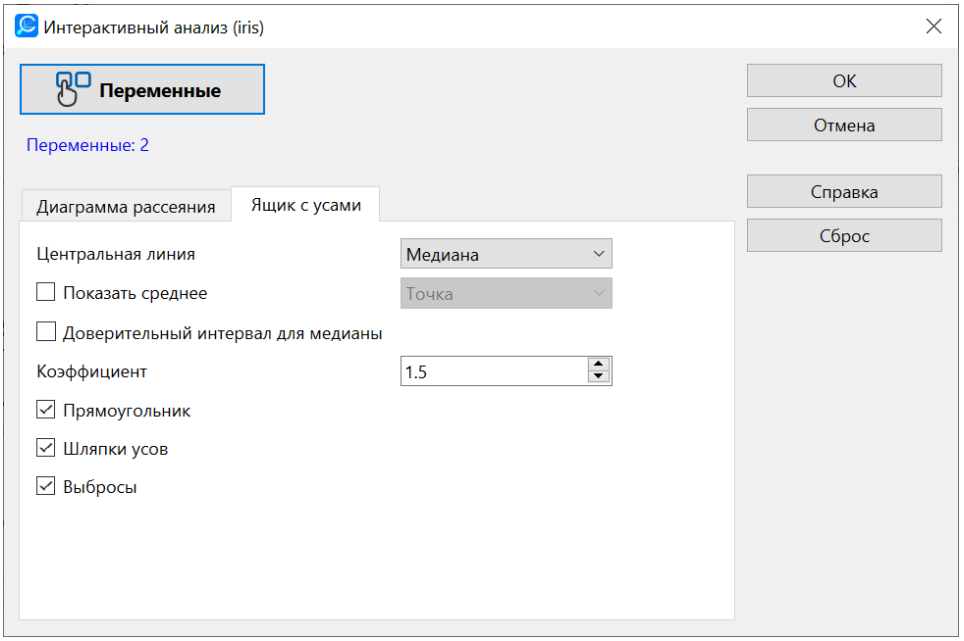
Пример 3. Интерактивное удаление выбросов на графике Ящик с усами

Рассмотрим пример удаления выбросов для диаграммы размаха. Откроем файл iris из папки примеров:

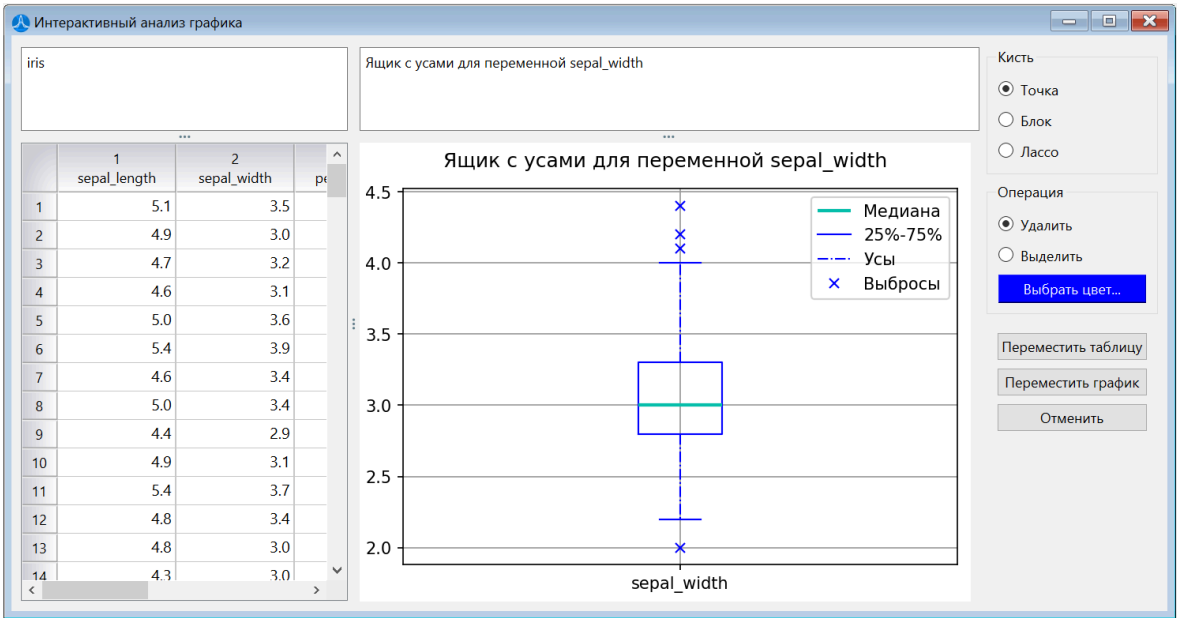


	1 sepal_length	2 sepal_width	3 petal_length	4 petal_width	5 species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa

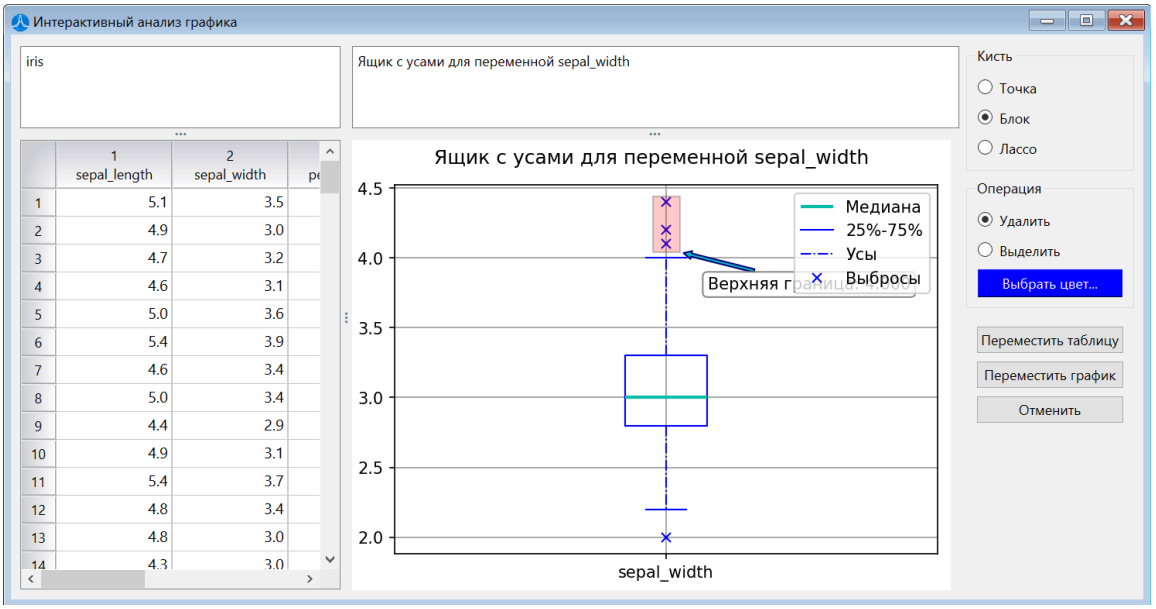
Откроем модуль Интерактивный анализ для этой таблицы и для переменной sepal_width построим график ящик с усами.



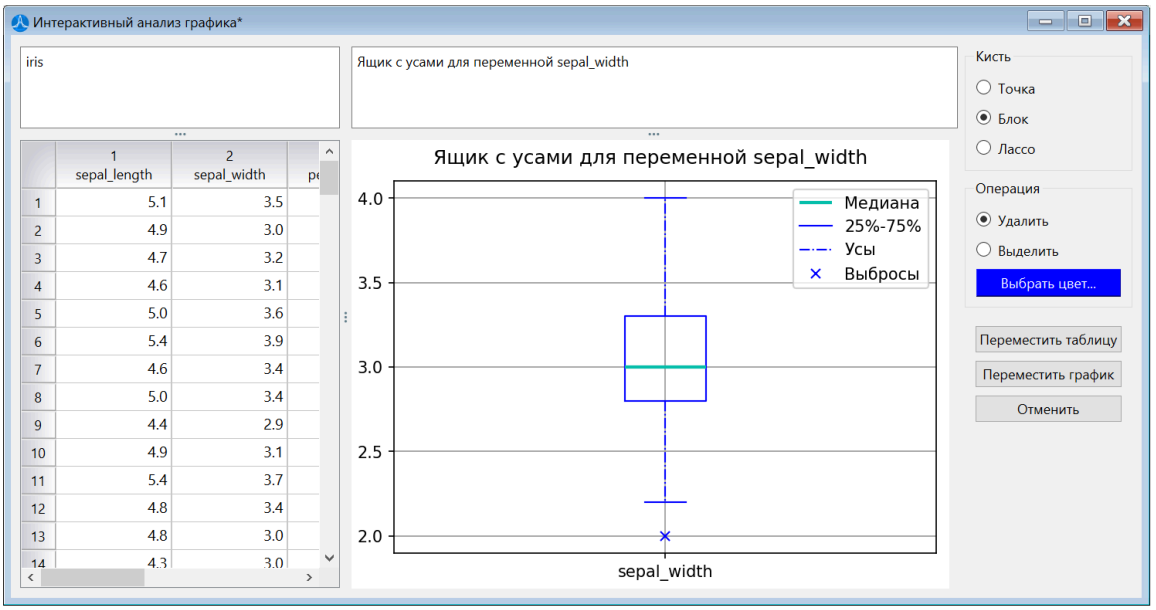
Видим, что на построенном графике имеются несколько выбросов:



С помощью кисти типа Блок и операции Удалить выделим три верхних выброса:



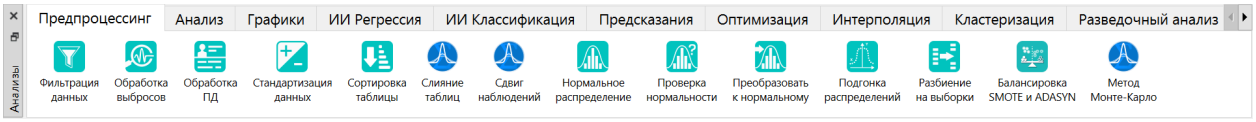
Сразу после этого выбросы на графике исчезают:



Если обратиться к таблице данных слева, мы заметим, что наблюдений в ней стало на три меньше (147 против исходных 150). Таким образом, удаленные нами точки также были удалены из таблицы. Для ее сохранения в качестве активной таблицы данных для дальнейшей работы, нажмем на кнопку Переместить таблицу.

Анализы

Предпроцессинг



Предпроцессинг данных является важным этапом перед основным анализом данных и позволяет обеспечить качественные и надежные результаты. Он включает в себя ряд методов и техник, которые помогают очистить данные от ошибок, шума, выбросов и пропущенных значений, а также преобразовать данные для их более удобного и понятного использования. Ниже перечислены наиболее распространенные методы предпроцессинга.

Фильтрация данных: это процесс удаления нежелательных или неинформативных данных. Например, можно убрать дубликаты строк, удалить столбцы, которые не несут значимой информации, или отфильтровать данные по заданным условиям.

Обработка выбросов: выбросы — это значения, которые значительно отличаются от большинства остальных данных. Такие аномальные значения могут исказить результаты анализа данных. При обработке выбросов можно удалить такие значения, заменить их на более типичные или использовать специальные методы статистической обработки выбросов.

Обработка пропущенных данных: пропущенные данные могут появляться по разным причинам, например, из-за ошибок сбора данных, повреждения файлов или отсутствия ответов на определенные вопросы в опросниках. Одной из широко используемых техник является заполнение пропущенных значений интерполяцией, средними или медианными значениями, восстановление данных с использованием методов машинного обучения или полное удаление строк или столбцов с пропущенными значениями.

Сортировка: сортировка данных может быть полезна для более эффективного доступа к информации или для проведения дальнейших анализов, основанных на упорядоченных данных.

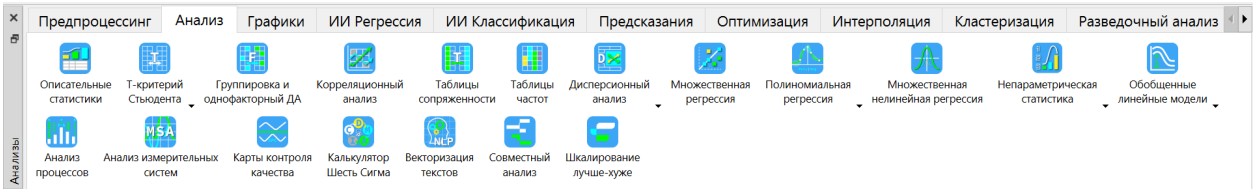
Стандартизация данных: стандартизация данных заключается в приведении значений к определенному диапазону или шкале. Это помогает устранить различия в масштабах и единицах измерения. Например, методы стандартизации могут быть использованы для

приведения всех значений к диапазону от 0 до 1 или к стандартному нормальному распределению.

Приведение данных к нормальному распределению: некоторые методы анализа данных, такие как статистические тесты или модели машинного обучения, предполагают нормальное распределение данных. В некоторых случаях данные могут быть далеки от нормального распределения, поэтому преобразование данных с использованием методов, таких как логарифмирование, квадратный корень или степенное преобразование, может помочь привести данные к более близкому к нормальному распределению виду.

Таким образом, предпроцессинг данных выполняет важную роль в обработке и подготовке данных перед основным анализом. Он позволяет улучшить качество данных, сделать их более доступными и пригодными для анализа, а также снизить искажения и ошибки, которые могут возникнуть при использовании "сырых данных".

Основные статистики и таблицы



Основные статистики и таблицы являются фундаментальными инструментами классического статистического анализа данных. Они используются для обработки и интерпретации информации, содержащейся в выборке или наборе данных.

Описательные статистики являются первым шагом в анализе данных. Они предоставляют сводную информацию о переменных, такую как среднее значение, медиана, мода, размах, стандартное отклонение и другие показатели. Описательные статистики позволяют оценить центральную тенденцию и разброс данных, что позволяет сделать выводы о характере выборки.

Т-критерии используются для проверки статистической значимости различий между средними значениями двух групп. Они позволяют определить, насколько вероятно получение таких различий в случае отсутствия истинного эффекта. Т-критерии широко используются в медицинских и социальных исследованиях, для сравнения эффективности различных лечебных методов или воздействий.

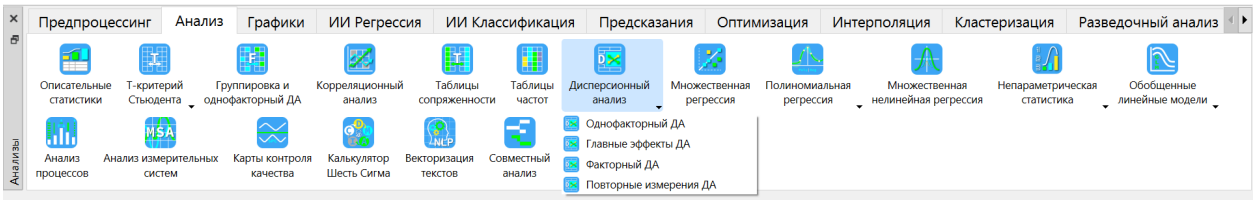
Корреляционный анализ позволяет определить степень взаимосвязи между двумя или более переменными. Наиболее распространенным показателем корреляции является коэффициент корреляции Пирсона, который может принимать значения от -1 до +1. Корреляционный анализ используется для исследования зависимостей и взаимосвязей между переменными, например, для оценки связи между давлением и пульсом у пациентов.

Таблицы частот представляют собой сводные таблицы, показывающие распределение переменных по категориям или группам. Они позволяют визуально оценить частотность различных значений или категорий и провести сравнение между группами. Такие таблицы особенно полезны при работе с категориальными переменными и позволяют исследовать различные параметры, например, процентное соотношение группы пациентов с разными диагнозами.

Таблицы сопряженности используются для исследования связей между двумя категориальными переменными. Они представляют данные в виде пересечений строк и столбцов, позволяя оценить статистическую значимость этих связей. Таблицы сопряженности часто используются в социальных исследованиях для изучения взаимосвязи между различными группами населения или для оценки связи между гендером и предпочтениями потребителей, например.

В целом, основные статистики и таблицы ПО СтатСофт предоставляют аналитикам и исследователям мощные инструменты для анализа данных. Они позволяют описать и интерпретировать информацию, выявить статистически значимые различия и взаимосвязи, а также делать выводы на основе полученных результатов.

Дисперсионный анализ



Основной целью дисперсионного анализа является исследование значимости различия между средними.

Может показаться странным, что процедура сравнения средних называется дисперсионным анализом. В действительности, это связано с тем, что при исследовании

статистической значимости различия между средними двух (или нескольких) групп на самом деле сравниваются (т.е. анализируются) выборочные дисперсии.

Перечислим основные термины, используемые в дисперсионном анализе.

Пусть имеется зависимая переменная Y и k независимых (X_1, X_2, …, X_k) переменных, задающих разбиение на группы. Независимые переменные называют факторами. Значения, которые может принимать фактор, – уровни фактора. Значение зависимой переменной – отклик.

Примеры факторов: пол, оператор оборудования, тип геологической скважины, наличие катализатора.

Примеры зависимых переменных: рост человека, прочность изделия, содержание золота.

Поскольку факторы задают разбиение на группы, для которых потом сравниваются средние, они должны являться категориальными переменными, то есть принимать какое-то конечное множество значений. В таком случае уровень фактора - это одно из уникальных значений, которое фактор может принять. Если же независимая переменная является непрерывной и должна участвовать в качестве фактора в дисперсионном анализе, нужно предварительно ее табулировать, то есть разбить возможный диапазон значений на промежутки и каждому из них сопоставить определенный уровень фактора.

Целевая характеристика (зависимая переменная) – характеристика процесса/изделия, влияние факторов на которую исследуется в ходе эксперимента. Например, временное сопротивление разрыву металла.

Отклик – значение целевой характеристики при определенной комбинации уровней факторов (то есть при определенных значениях). Например, при содержании C = 0,15% и Si = 0,21% временное сопротивление разрыву равно 425 Н/мм2.

Эффект фактора – изменение отклика при переходе фактора с одного уровня на другой. Например, сменили оператора станка, и качество продукции выросло на 10%.

Если изучаются несколько факторов, можно рассматривать их воздействие по-отдельности (тогда это называется главными эффектами). Но сложность заключается в том, что факторы могут взаимодействовать между собой, и эти взаимодействия тоже нужно учитывать.

Итак, в дисперсионном анализе набор категориальных факторов задает разбиение объектов на группы (по уровням этих факторов), и ставится задача о проверке гипотезы о равенстве средних во всех группах.

При этом процедура дисперсионного анализа основывается на следующих предположениях:

- Данные в каждой группе имеют нормальное распределение.
- Данные в каждой группе имеют одинаковую дисперсию.

Для реальных данных эти требования выполняются не всегда, но это можно компенсировать, если соответствующим образом составить выборку данных:

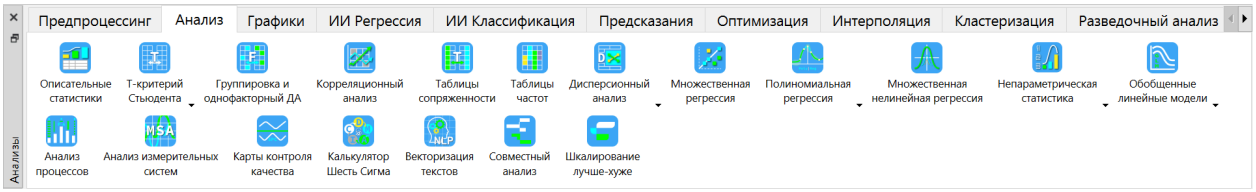
- Отсутствие нормальности можно компенсировать большим числом наблюдений.
- Различия в дисперсиях можно компенсировать одинаковым количеством наблюдений в каждой группе.

Суть процедуры дисперсионного анализа заключается в следующем. Дисперсия изучаемого признака делится на отдельные компоненты, обусловленные влиянием конкретных факторов. Сравнивая компоненты дисперсии друг с другом посредством F-критерия Фишера, можно определить статистическую значимость различия между средними. Если это различие значимо, гипотеза о равенстве средних отвергается и принимается альтернативная гипотеза о существенном различии между средними.

Существуют разные виды дисперсионного анализа.

- Однофакторный. Изучается воздействие только одного фактора.
- Главные эффекты. Изучаются воздействия нескольких факторов без учета их взаимодействия.
- Факторный. Изучаются воздействия нескольких факторов с учетом их взаимодействия.
- Повторные измерения. Изучаются наблюдения за одним и тем же объектом в разное время или при разных обстоятельствах.

Множественная регрессия



Модель множественной регрессии строит зависимость целевой переменной от предикторов в виде их линейной комбинации, то есть подбирает оптимальные веса (коэффициенты, параметры модели) для каждой независимой переменной. Регрессионные коэффициенты (или B-коэффициенты) представляют независимые вклады каждой независимой переменной (предикторов) в предсказание зависимой переменной (целевой). Если предикторы стандартизованы, то есть имеют одинаковое среднее и дисперсию, то чем больше по модулю значение коэффициента, тем сильнее вклад соответствующей переменной.

Пусть x_1, x_2, …, x_k – значения k независимых переменных какого-то наблюдения. Тогда множественная регрессия построит следующую зависимость:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

Здесь y – предсказанное моделью значение, а b_0, b_1, b_2, …, b_k – подобранные моделью веса, при которых предсказанные значения максимально близки к наблюдаемым для рассматриваемой выборки (b_0 также называется свободным членом или сдвигом).

Поскольку модель представляет собой линейную комбинацию предикторов (сумму с некоторыми весами), то ее также часто называют линейной регрессией.

Если рассматривается только один предиктор, то уравнение регрессии выглядит как уравнение прямой: $y = b_0 + b_1x_1$. Таким образом, если отмечать на графике по оси X значения предиктора, а по оси Y - предсказанные значения, то предсказания модели образуют прямую линию. Она называется линией регрессии.

Если предикторов больше одного, то уравнение регрессии задает не линию, а поверхность в соответствующем пространстве. Она называется поверхностью регрессии.

Линия регрессии (или поверхность регрессии) выражает наилучшее предсказание зависимой переменной (y) по независимым переменным (X). Однако, природа редко (если вообще когда-нибудь) бывает полностью предсказуемой и обычно имеется существенный разброс наблюдаемых точек относительно подогнанной прямой. Отклонение отдельной точки от линии регрессии (от предсказанного значения) называется остатком.

Соответственно, программа строит линию или поверхность регрессии таким образом, чтобы минимизировать сумму квадратов этих остатков. Поэтому на эту общую процедуру иногда ссылаются как на оценивание по методу наименьших квадратов.

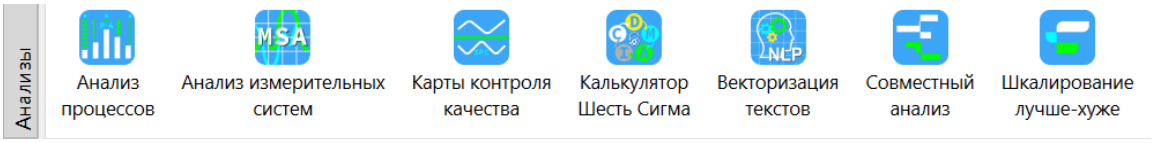
В стандартной регрессии используются все указанные исследователем переменные. Помимо этого, существует пошаговая регрессия, в которой автоматически отбираются наиболее статистически значимые переменные на основании некоторого критерия качества (p-значение должно быть менее 0,05).

Пошаговая с включением. При таком подходе независимые переменные будут по отдельности включаться или исключаться из модели на каждом шаге регрессии до тех пор, пока не будет получена "наилучшая" регрессионная модель.

Пошаговая с исключением. При таком подходе независимые переменные будут исключаться из модели по одной на каждом шаге до тех пор, пока не будет получена "наилучшая" регрессионная модель.

Преимуществами модели множественной регрессии является ее простая интерпретируемость и небольшое число параметров для обучения (примерно столько же, сколько предикторов), что позволяет применять эту модель к наборам данных небольшого объема.

Карты контроля качества



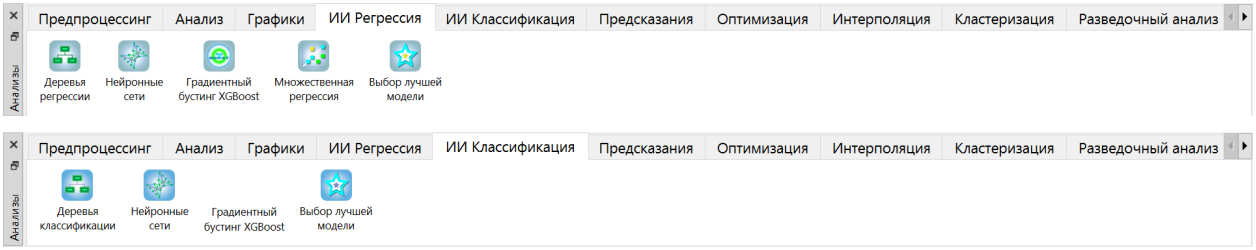
Многие, если не большинство, производственных процессов полностью автоматизированы и автоматически контролируются датчиками и измерительными устройствами, что дает большой объем информации, относящейся к качеству конечного продукта и общей производительности процесса.

Целью модуля Карты контроля качества является предоставление инструментов и методов для анализа информации, предоставляемой процессом производства или предоставления услуг, а также для выявления тенденций, закономерностей и причинно-следственных связей, которые можно использовать для дальнейшего повышения качества и производительности процессов.

Карты контроля качества (ККК) - это инструмент качественного контроля, который используется для оценки соответствия продукции или процессов определенным стандартам и требованиям. Они используются для повышения качества продукции, улучшения процессов и сокращения отходов благодаря идентификации и устранению проблем на ранних этапах производства. Они могут также использоваться для обучения персонала в области контроля качества и анализа данных.

Карты контроля качества являются важным инструментом для улучшения качества продукции и процессов, а также для увеличения эффективности и экономичности производства.

Нейронные сети



Если классические методы анализа не работают или точность результатов не удовлетворяет исследователя, встает задача выбора инструмента. Таким инструментом могут служить нейронные сети, позволяющие строить сложные нелинейные зависимости.

Простейший вариант нейронной сети состоит из трех слоев: входного, скрытого и выходного.

Исходные данные (набор значений переменных для какого-либо наблюдения) поступают на входной слой.

Скрытый слой состоит из нескольких нейронов, каждый из которых строит линейную комбинацию из исходных данных (как в модели линейной регрессии) и применяет функцию активации.

Вычисленные таким образом значения с каждого нейрона поступают на выходной слой, где из них снова строится линейная комбинация с применением функции активации. Полученное число и будет являться ответом модели на имеющиеся данные.

Нейронные сети позволяют решать широкий спектр задач, включающий задачи классификации и регрессии.

Примеры

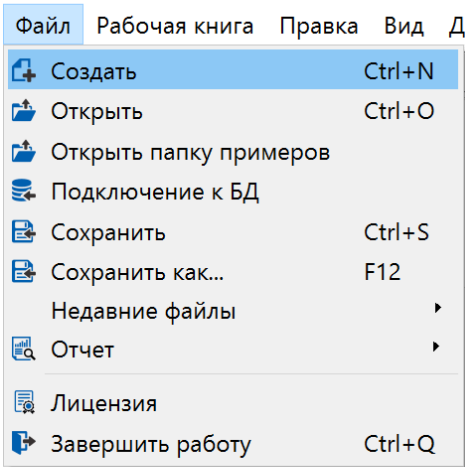
Пример 1. Корреляционный анализ

На металлообрабатывающем заводе у 60 марок стали проводят замеры предела текучести F (признак X, кг/мм²) и предела прочности σ_B (признак Y, кг/мм²). В итоге получают следующие пары значений:

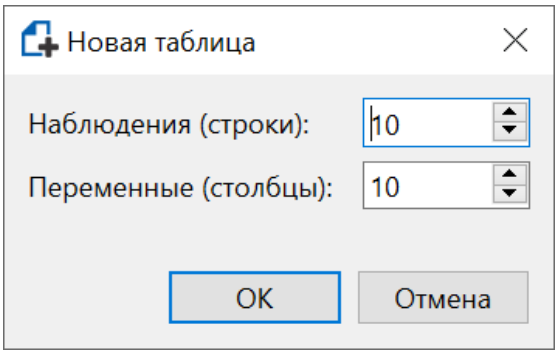
F (X)	σ_B (Y)	F (X)	σ_B (Y)	F (X)	σ_B (Y)
154	178	83	98	73	76
133	164	106	111	77	85
58	75	92	104	47	61
145	161	85	103	68	85
94	107	112	118	137	142
113	141	98	102	44	69
86	97	103	108	92	116
121	127	99	119	141	157
119	138	104	128	155	193
112	125	107	118	136	155
85	97	98	140	82	81
41	72	97	115	136	163
96	113	105	101	72	79
45	89	71	93	66	81
99	109	39	69	42	61
51	95	122	147	113	123
101	114	33	52	42	85
169	209	78	117	133	147
87	101	114	138	153	179
88	139	125	149	85	91

Эти 60 пар значений образуют реализации случайного вектора (X; Y) (предел текучести F и предел прочности σ_B). Спрашивается, существует ли какая-либо связь между пределом текучести и пределом прочности стали? Для решения этого вопроса воспользуемся модулем Корреляционный анализ.

Файл данных. Создадим в ПО СтатСофт новую таблицу и занесем в нее результаты эксперимента. Для этого в меню Файл нажмем Создать:

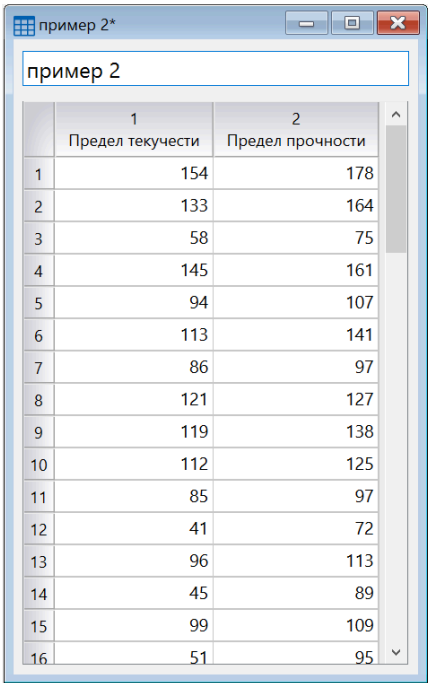


Откроется диалоговое окно Новая таблица, в котором необходимо задать количество переменных и наблюдений:



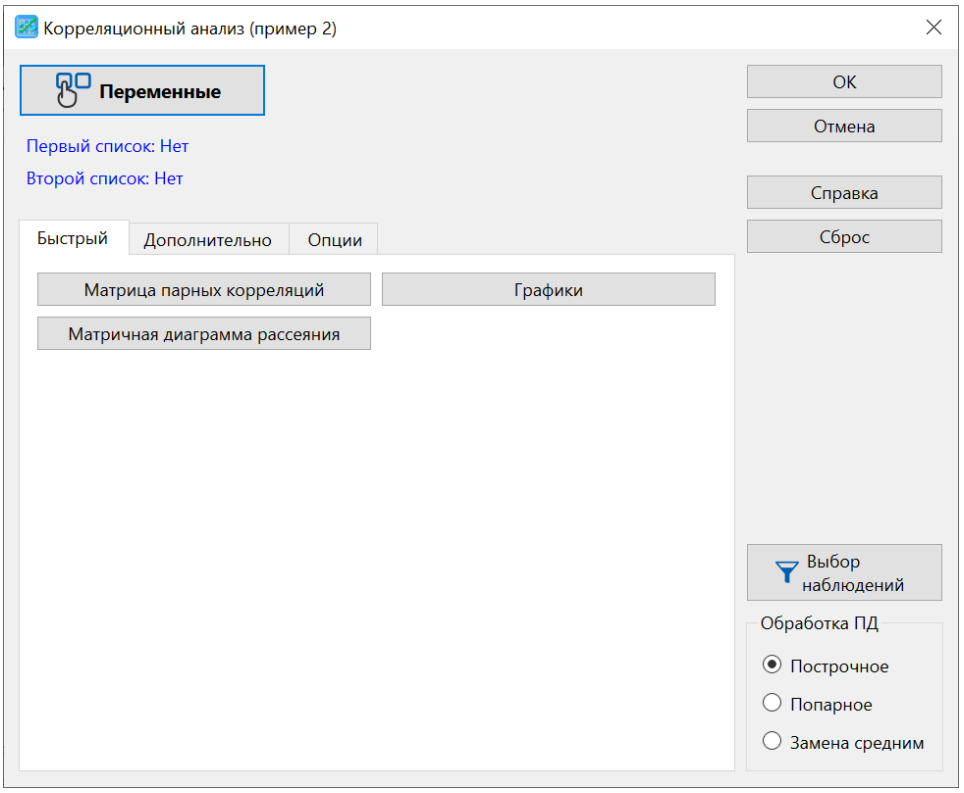
Зададим количество наблюдений 60 и количество переменных 2, после чего нажмем ОК. В рабочей области появится пустая таблица указанных размеров, занесем в нее наши данные и по желанию переименуем столбцы.

Итоговая таблица выглядит следующим образом:

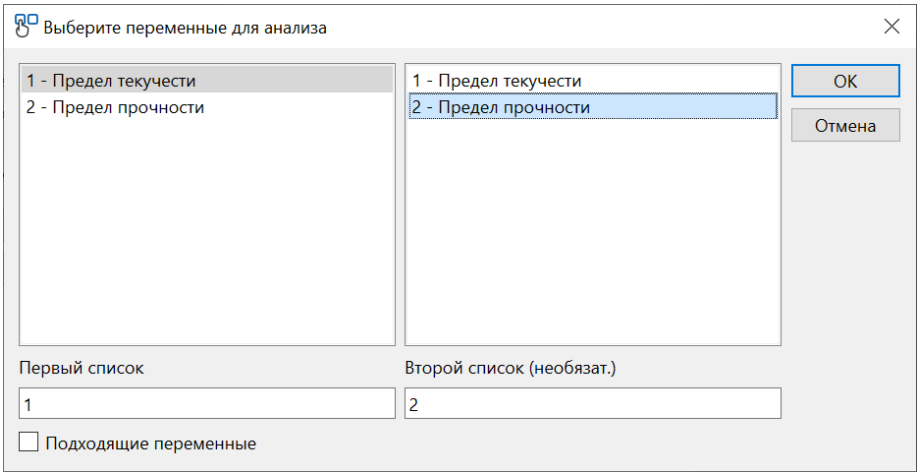


	1 Предел текучести	2 Предел прочности
1	154	178
2	133	164
3	58	75
4	145	161
5	94	107
6	113	141
7	86	97
8	121	127
9	119	138
10	112	125
11	85	97
12	41	72
13	96	113
14	45	89
15	99	109
16	51	95

Задание анализа. Выберите опцию Корреляционный анализ в меню Анализ для отображения диалогового окна Корреляционный анализ:

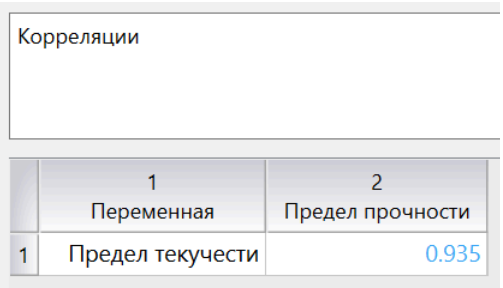


Нажмем на кнопку Переменные для открытия стандартного диалогового окна Выбор переменных. Отнесем переменную Предел текучести в первый список, а Предел прочности - во второй.

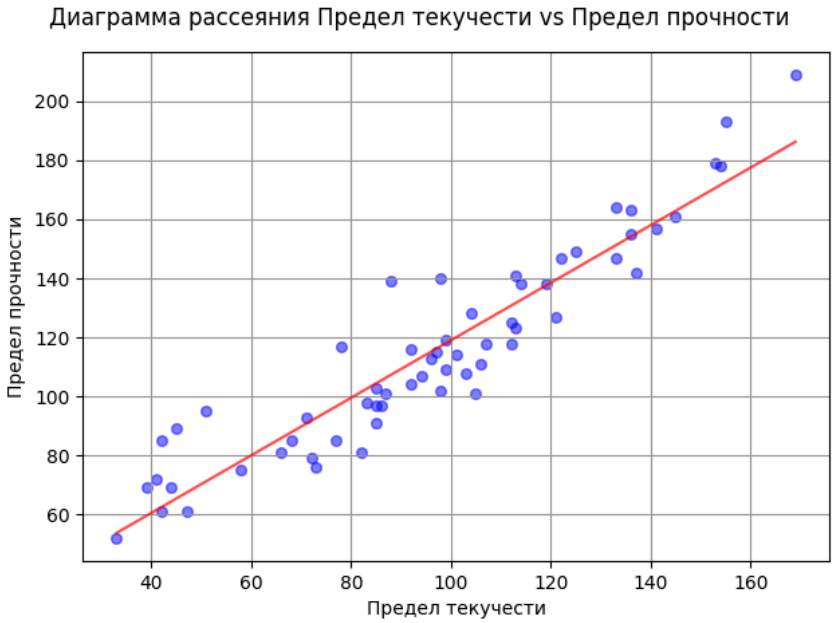


Нажмем кнопку ОК.

Просмотр результатов. В диалоговом окне Корреляционный анализ нажмем кнопку Матрица парных корреляций для получения таблицы коэффициентов корреляции:



Видим, что коэффициент корреляции между двумя исследуемыми величинами составляет примерно 0.94, что говорит об очень сильной положительной корреляции. Проиллюстрируем полученный результат графиком, выбрав на вкладке Дополнительно диалогового окна Корреляционный анализ опцию 2M Рассеяния:



На этом графике данные в виде точек отображаются в декартовой системе координат, причем по оси абсцисс откладываются значения Предела текучести, а по оси ординат - значения Предела прочности. Из графика видно, что между переменными присутствует ярко выраженная линейная зависимость, причем с увеличением значений одной переменной увеличиваются и значения второй.

Это графическое представление подтверждает полученный ранее большой положительный коэффициент корреляции.

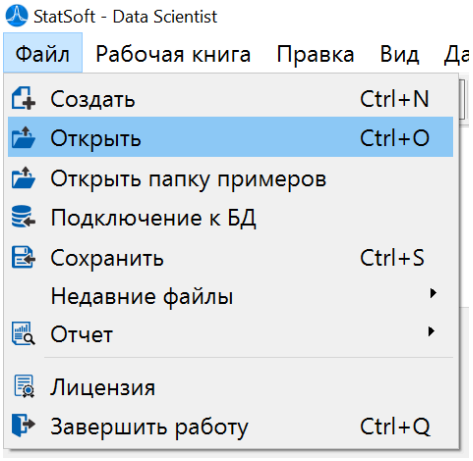
Выводы. С помощью использования Корреляционного анализа нам удалось найти ответ на вопрос о существовании зависимости между исследуемыми переменными и выяснить, что между ними существует сильная положительная корреляции.

Пример 2. Однофакторный дисперсионный анализ

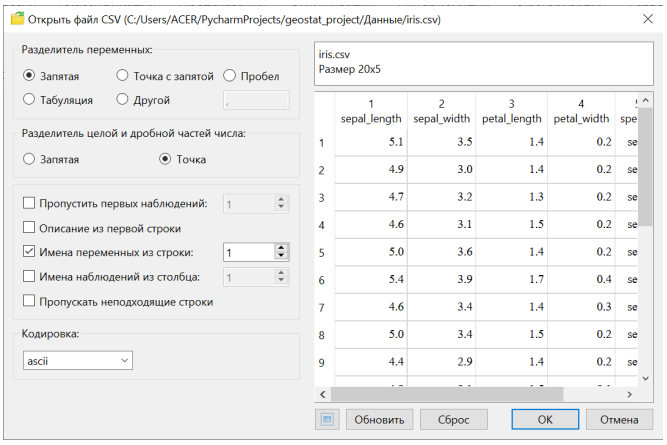
В данном примере рассматривается применение модуля "Однофакторный ДА" для проверки гипотезы о равенстве средних в трех группах.

Пусть данные разбиты на три группы (то есть фактор имеет три уровня), и для каждой группы было проведено 50 измерений (всего 150). Необходимо понять, является ли значимым фактор разбиения на эти группы.

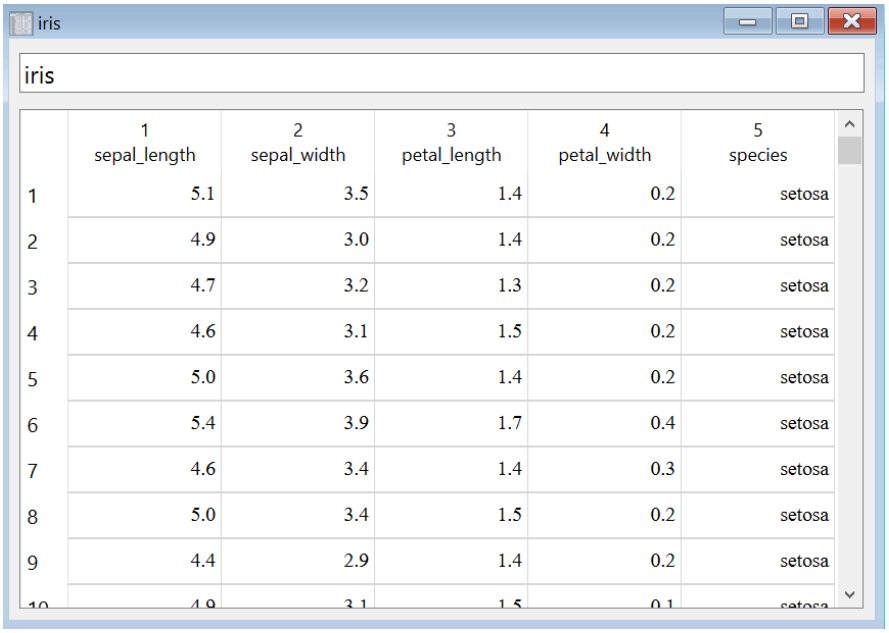
Загрузка и обзор данных. Загрузим новую таблицу с данными. В меню «Файл» выберем опцию «Открыть».



Выберем файл "iris.csv" и нажмем "Открыть", далее, если это необходимо, зададим параметры текстового файла с разделителями и нажмем кнопку «ОК».



В рабочей области появится сформированная таблица с данными.



	1 sepal_length	2 sepal_width	3 petal_length	4 petal_width	5 species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

Ирисы Фишера — это набор данных для задачи классификации, на примере которого Рональд Фишер в 1936 году продемонстрировал работу разработанного им метода дискриминантного анализа, задачей которого было установить связь между параметрами цветка ириса и классом, к которому он относился.

Этот набор данных стал классическим и часто используется в литературе для иллюстрации работы различных статистических алгоритмов.

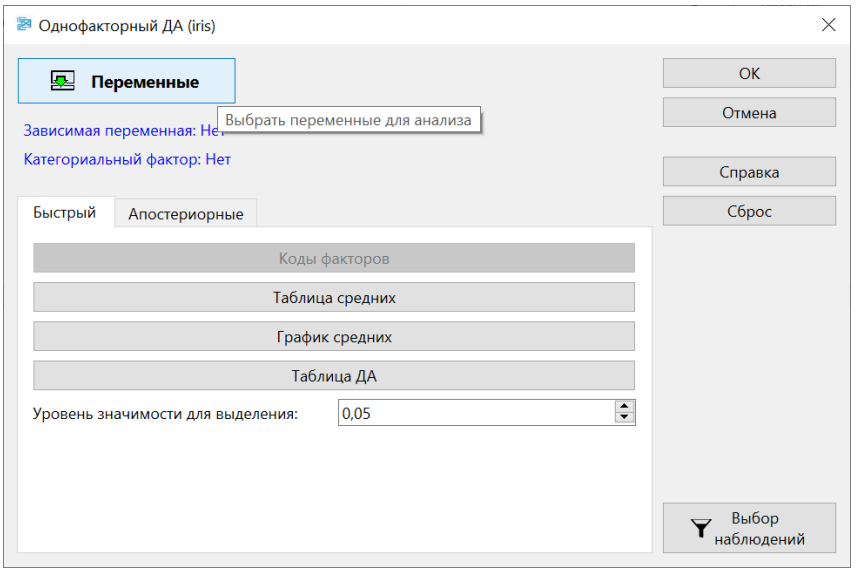
Структура данных:

- sepal_lenght – длина чашелистиков
- sepal_widht – ширина чашелистиков
- petal_lenght – длина лепестков
- petal_widht – ширина лепестков
- species – тип ириса

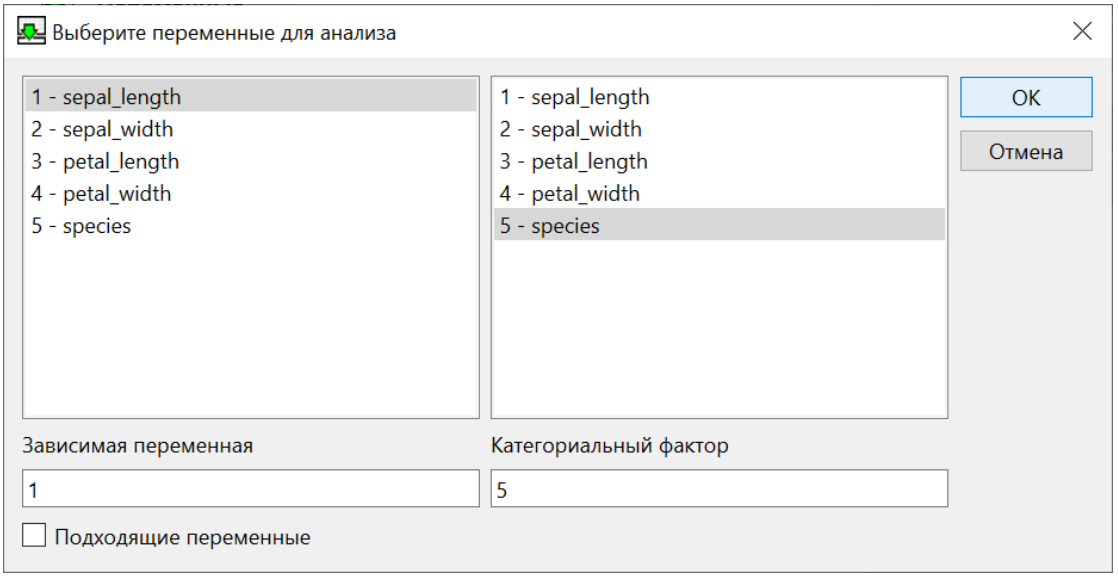
Столбец 5 показывает, к какой группе ирисов относится конкретное наблюдение, а столбцы 1–4 - результаты измерений параметров цветков. Мы будем исследовать влияние фактора на значение первой переменной – «sepal_lenght». Всего в таблице 150 строк, поскольку в исследовании было изучено 150 ирисов.

Проведение дисперсионного анализа. На вкладке "Анализ" выберем "Дисперсионный анализ", в открывшемся окне подменю выберем "Однофакторный ДА".

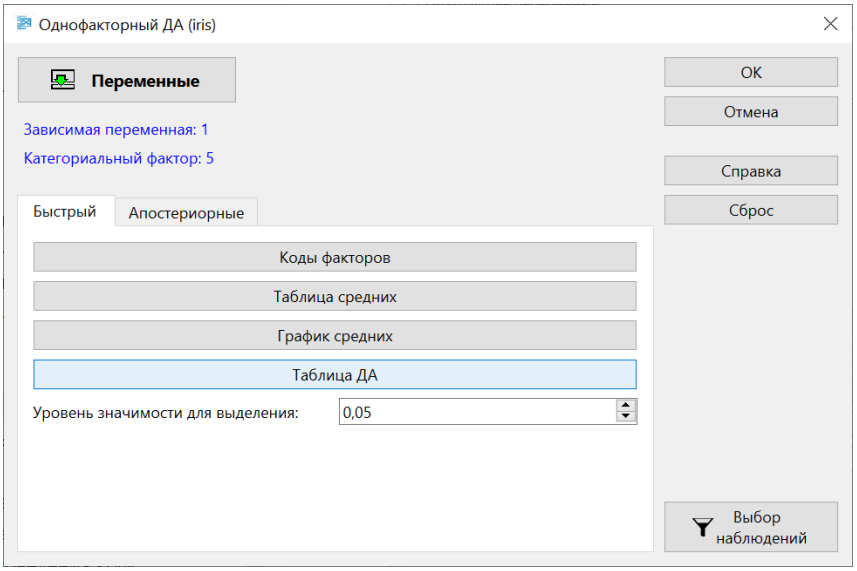
Откроется новое окно. Здесь нажмем кнопку "Переменные".



Установим переменную 1 в качестве зависимой и переменную 5 в качестве фактора, нажмем «ОК».



Затем в окне анализа нажмем кнопку "Таблица ДА".



В рабочей книге отобразится таблица с результатами дисперсионного анализа.

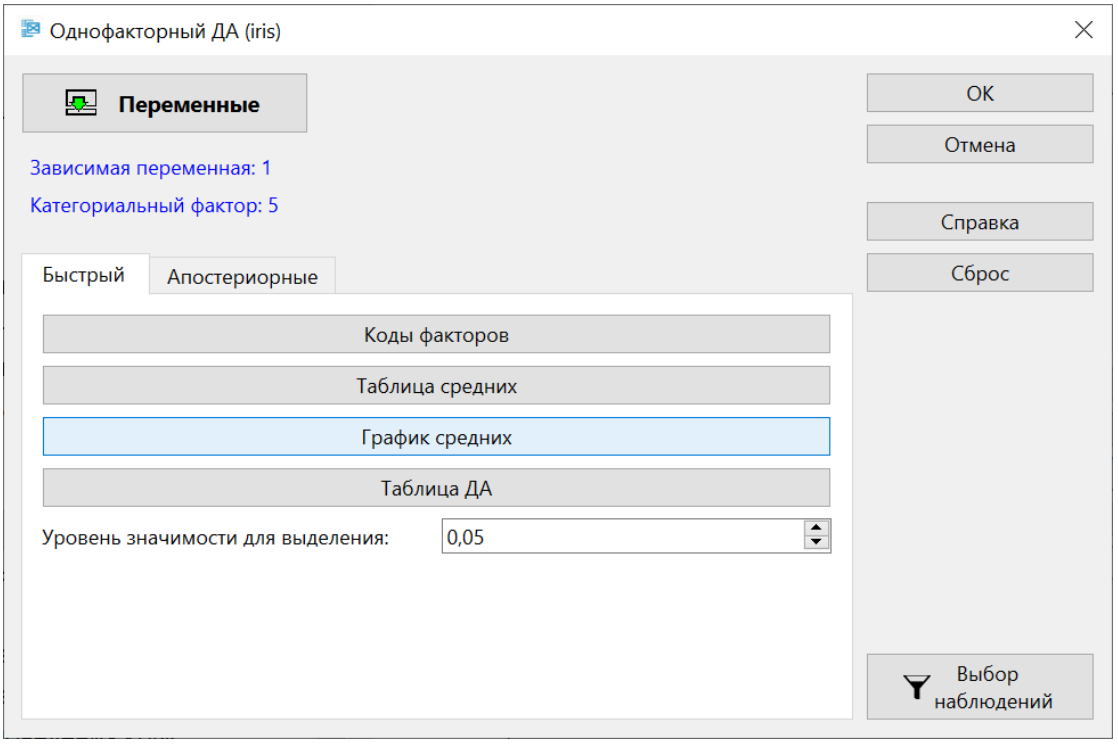
Однофакторный ДА для sepal_length

	1 Фактор	2 Сумма квадратов	3 Степени свободы	4 F	5 p
1	species	63.212	2.0	119.265	0.0
2	Ошибка (sepal_length)	38.956	147.0		

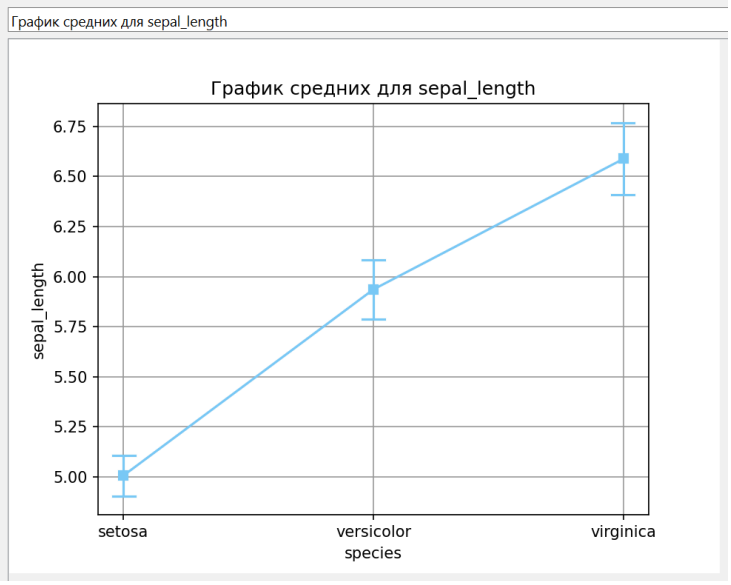
Данная таблица показывает сумму квадратов отклонений, объясняемую фактором "species". С помощью F-статистики определяется, насколько эта сумма отклонений отличается от оставшейся ошибки, и по p-уровню для F-статистики можно определить значимость фактора.

В данном примере p-уровень ниже 0,05, поэтому фактор является значимым и влияет на среднее значение наблюдений.

Нажмем на кнопку "График средних".



В рабочей книге появится график средних значений, по которому также можно увидеть, что средние значения в трех группах значимо различаются.



Выводы. Функционал приложения позволяет быстро проводить однофакторный дисперсионный анализ и исследовать его результаты как в виде таблицы, так и в виде графика.

СтатСофт®

В данном примере исследуемый фактор оказался значимым, следовательно, он существенно влияет на зависимую переменную. Поэтому средние значения зависимой переменной на разных уровнях этого фактора значимо отличаются друг от друга.

Пример 3. Логит модель

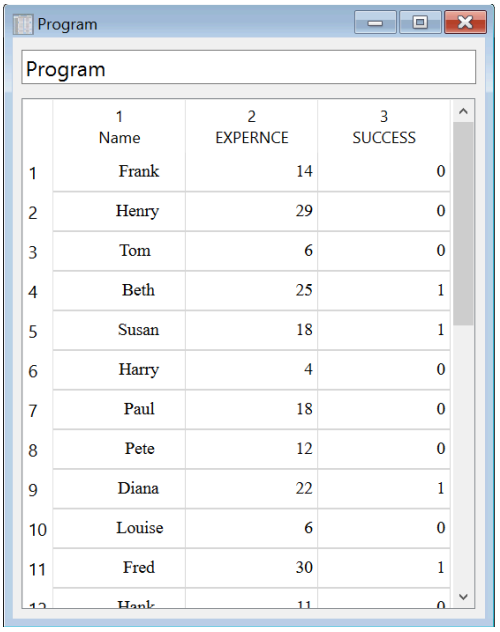
Вводный обзор. В этом примере мы рассмотрим логит модель для бинарных откликов. Эта модель встроена в модуль Обобщенные линейные модели (ОЛМ) и может быть выбрана в соответствующем диалоговом окне.

Модуль ОЛМ содержит опции построения биномиальных и многомерных моделей отклика и содержит методы пошагового выбора предикторов, а также способ выбора наилучшего подмножества.

Этот пример основан на данных из работы Neter, Wasserman and Kutner (1985, стр. 357. Однако отметим, что они использовали для подгонки линейную регрессионную модель). Предположим, что вы хотите проверить, правда ли, что стаж работы помогает программистам в написании сложных программ, если на написание отпущен ограниченный промежуток времени.

Для исследования были выбраны двадцать пять программистов с различным стажем работы (выраженным в месяцах). Их попросили написать сложную компьютерную программу за определенный промежуток времени. Бинарная переменная отклика принимала значение 1, если программист справился с поставленной задачей и 0, если нет.

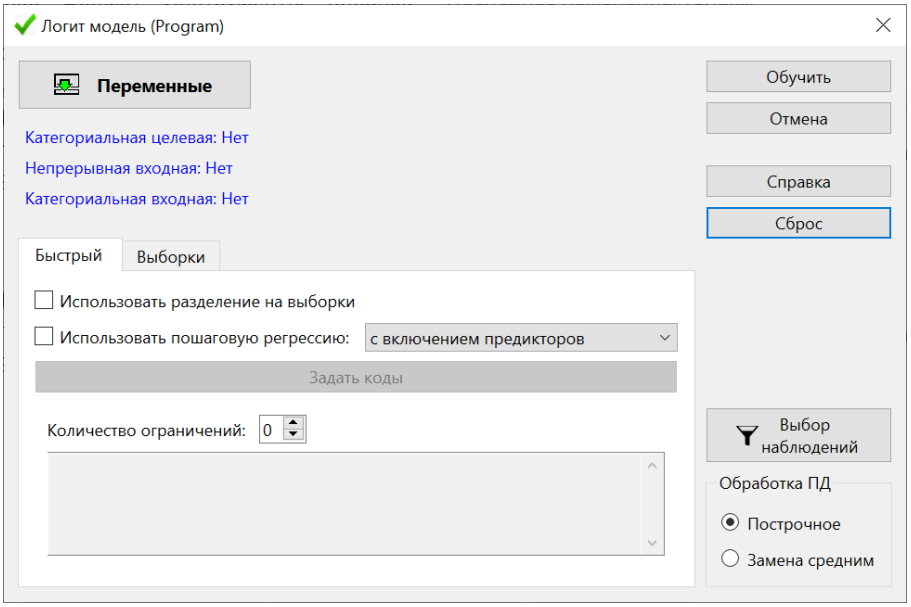
Эти данные были сохранены в файле Program.csv; ниже представлен фрагмент этого файла:



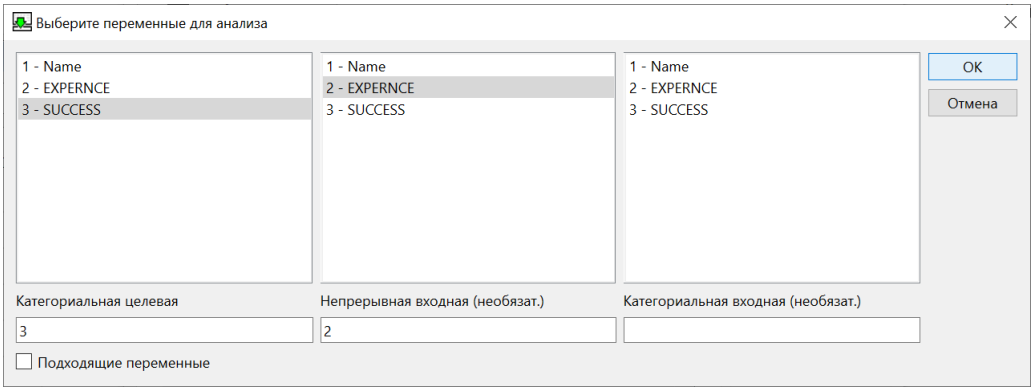
	1 Name	2 EXPERNCE	3 SUCCESS
1	Frank	14	0
2	Henry	29	0
3	Tom	6	0
4	Beth	25	1
5	Susan	18	1
6	Harry	4	0
7	Paul	18	0
8	Pete	12	0
9	Diana	22	1
10	Louise	6	0
11	Fred	30	1
12	Hank	11	0

Задание анализа. Откройте файл Program.csv с помощью меню Файл, выбрав команду Открыть папку примеров. Выберите опцию Обобщенные линейные модели в меню Анализ для отображения всех доступных в ПО видов обобщенные линейные модели:

В раскрывшемся меню выберите опцию Логит модель, после чего откроется диалоговое окно Логит модель:



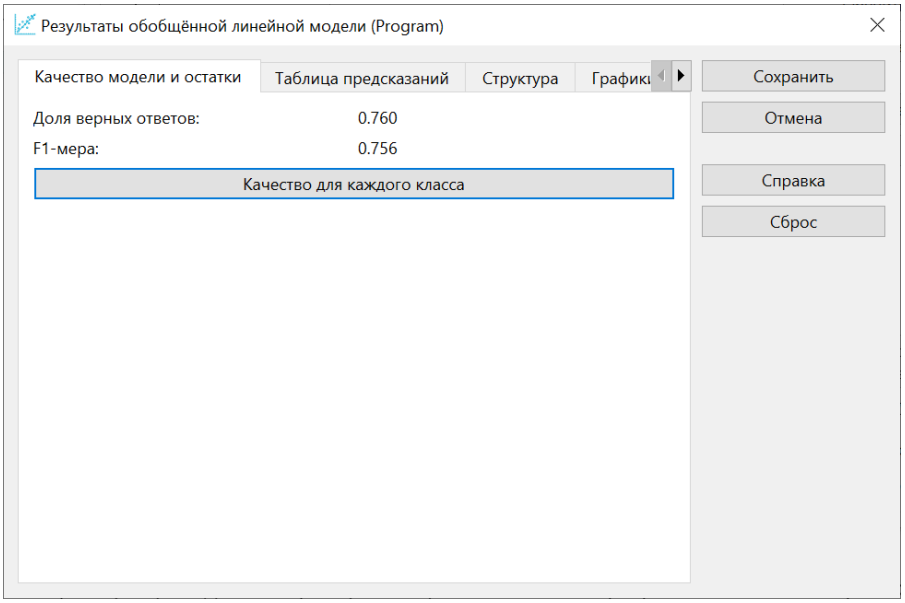
Далее нажмите кнопку Переменные для отображения на экране стандартного диалогового окна Выбор переменных. Выберите переменную Success как целевую и Expernce как непрерывную входную и нажмите ОК.



После этого программа автоматически определит коды для возможных значений зависимой переменной, однако их также можно задать вручную, нажав на кнопку Задать коды.

Отметим, что модуль Обобщенные линейные модели для оценки параметров логит и пробит моделей всегда использует метод максимума правдоподобия. Обычный метод наименьших квадратов основывается на предположении об одинаковой дисперсии ошибок (остатков) при различных значениях независимых переменных. В случае бинарной зависимой переменной это предположение явным образом нарушается, и поэтому для оценивания параметров логит и пробит регрессионных моделей следует использовать метод максимума правдоподобия.

После задания переменных нажмите кнопку ОК, перейдя тем самым в диалоговое окно Результаты обобщенной линейной модели.



Просмотр результатов. Сверху открытого окна отображается качество построенной модели: доля верных ответов и средняя f1-мера. Для получения более подробной информации о качестве для каждого класса нажмем на кнопку Качество каждого класса. Получим следующую таблицу:

Качество модели для каждого класса

	1 Класс	2 precision	3 recall	4 f1-score	5 support
1	0	0.786	0.786	0.786	14.0
2	1	0.727	0.727	0.727	11.0
3	accuracy	0.76	0.76	0.76	0.76
4	macro avg	0.756	0.756	0.756	25.0
5	weighted avg	0.76	0.76	0.76	25.0

По таблице видим, что полнота для класса 0 равна 0.786, то есть модель верно распознает большинство объектов класса 0. Аналогично, точность для класса 1 равна 0.727.

Теперь просмотрим полученные оценки параметров, нажав на кнопку Структура модели на вкладке Структура диалогового окна Результаты обобщенной линейной модели. В таблице видим B-коэффициенты для каждого из параметров модели (в данном случае входной непрерывной переменной EXPERNCE и константы const).

Коэффициенты множественной регрессии для переменной SUCCESS

	1 Переменная	2 B-коэффициент	3 Нижн. 95%	4 Верхн. 95%	5 Бета-коэффициент	6 p-значение
1	const	-3.06	-5.528	-0.591	-0.334	0.015
2	EXPERNCE	0.161	0.034	0.289	1.436	0.013

Интерпретация оценок параметров. В принципе, оценки параметров могут быть проинтерпретированы, как и в случае стандартной линейной регрессионной модели, т.е. в терминах свободного члена (const) и углового коэффициента (Expernce).

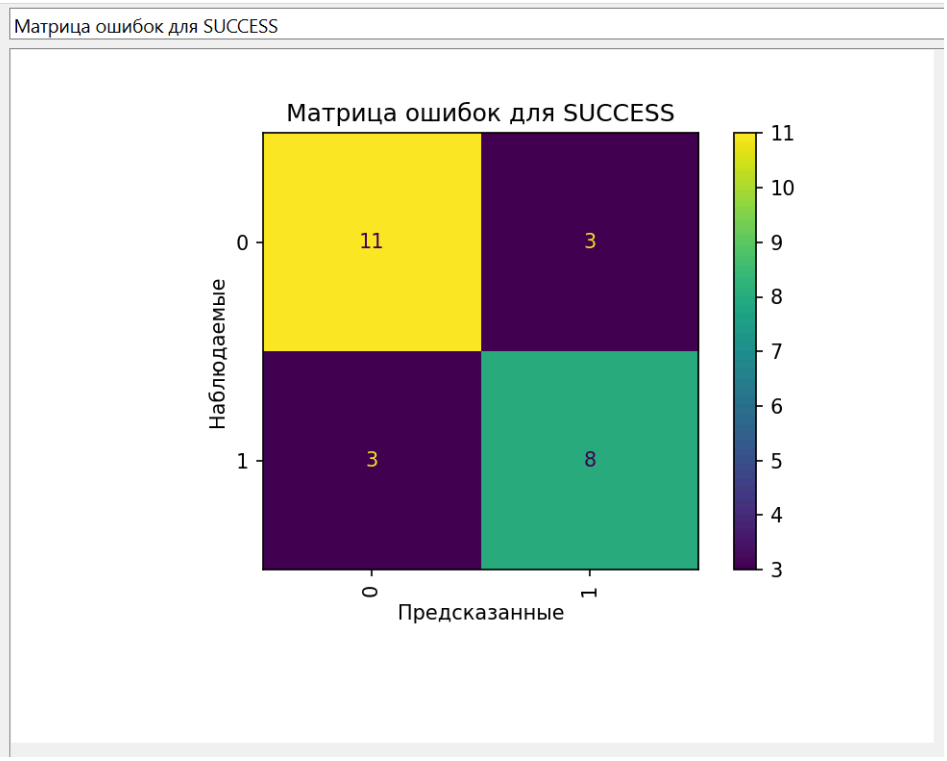
Предсказанные значения. Теперь рассмотрим предсказанные значения. Для этого нажмите кнопку Предсказания на вкладке Таблица предсказаний, поставив галочку напротив пункта Выводить вероятности.

Напомним, что регрессионная модель логит гарантирует, что предсказанные значения всегда будут находиться внутри отрезка [0,1]. Поэтому вы можете рассматривать полученные значения как вероятности (столбец 4). Например, предсказанная вероятность успеха для второго программиста равна 0.835.

Предсказания для переменной SUCCESS

	1 SUCCESS	2 SUCCESS предсказанный	3 Вероятность	4 Предсказанное числовое значение P(SUCCESS = 1)
1	0.0	0.0	0.69	0.31
2	0.0	1.0	0.835	0.835
3	0.0	0.0	0.89	0.11
4	1.0	1.0	0.727	0.727
5	1.0	0.0	0.538	0.462
6	0.0	0.0	0.918	0.082
7	0.0	0.0	0.538	0.462
8	0.0	0.0	0.754	0.246
9	1.0	1.0	0.621	0.621

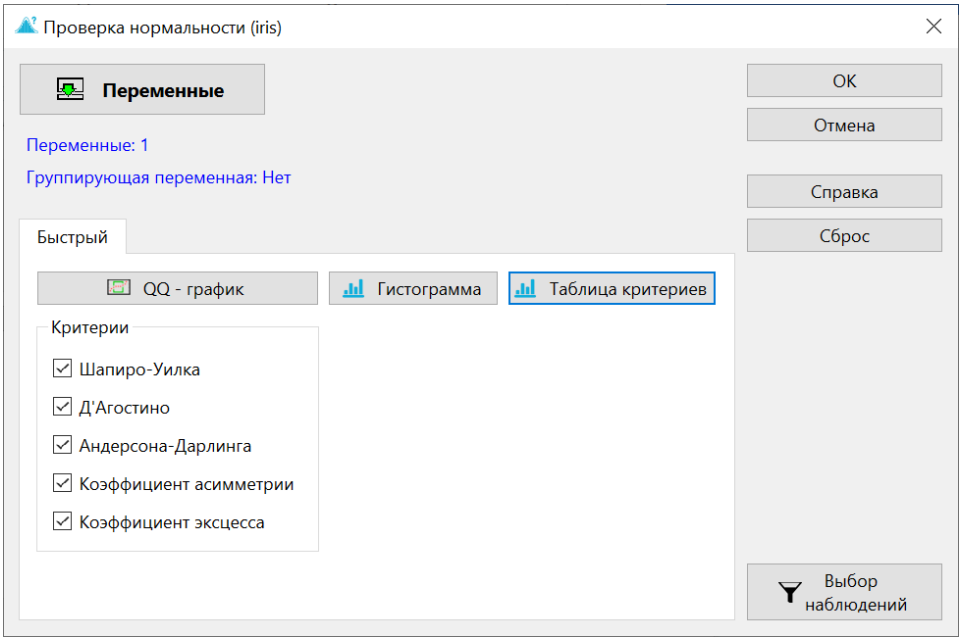
Классификация наблюдений. Выберите опцию Матрица ошибок на вкладке Графики, чтобы вывести на экран график с числом наблюдений, которые были правильно и неправильно классифицированы в соответствии с полученной моделью.



Из графика видим, что всего верно было классифицировано 19 наблюдений (сумма клеток на главной диагонали), а неверно - 6 (сумма всех остальных клеток).

Пример 4. Проверка нормальности

Рассмотрим пример проведения Проверки нормальности в программе ПО СтатСофт по шагам. Выбираем Проверка нормальности на вкладке «Предпроцессинг».



Открывается окно Проверка нормальности, где необходимо указать переменные и выбрать интересующие нас критерии для проверки нормальности.

Моделирование Монте-Карло показало, что тест Шапиро-Уилка имеет наилучшую мощность для заданного уровня статистической значимости, за ним следует тест Андерсона-Дарлинга при сравнении тестов Шапиро-Уилка, Колмогорова-Смирнова и Лиллиефорса.

Тест Шапиро-Уилка проверяет нулевую гипотезу о том, что исследуемая выборка происходит из нормально распределенной совокупности значений. Нулевая гипотеза этого теста состоит в том, что данные распределены нормально.

То есть, если значение p меньше выбранного альфа-уровня, то нулевая гипотеза отвергается и есть свидетельство того, что тестируемые данные не имеют нормального распределения. С другой стороны, если значение p больше выбранного альфа-уровня, то нулевая гипотеза (о том, что данные получены из нормально распределенной совокупности) не может быть отвергнута.

Как и большинство тестов статистической значимости, если размер выборки достаточно велик, этот тест может обнаружить даже тривиальные отклонения от нулевой гипотезы. Таким образом, обычно рекомендуется дополнительное исследование размера эффекта, например, в нашем случае используем для этого QQ-график.

После того как интересующие нас критерии и переменная заданы, выбираем QQ-график и нам выводится квантиль-квантиль график.

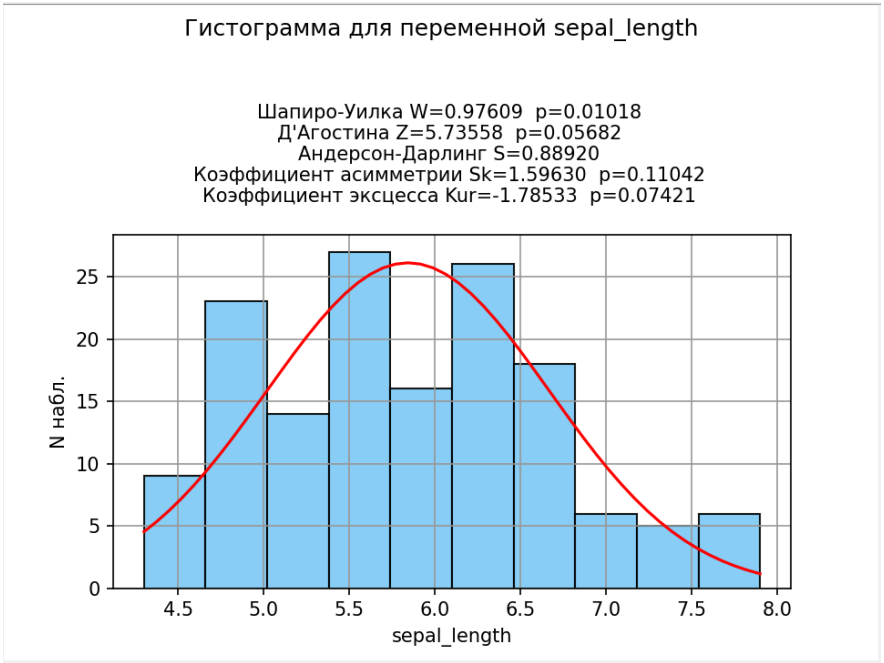
Данный график показывает, насколько хорошо наша переменная соответствует нормальному закону распределения, в идеальном случае все точки лежат на красной прямой. По полученному графику не совсем понятно как наши данные оцениваются нормальным распределением.

Поэтому посмотрим на таблицы результатов выбранных критериев, нажав на кнопку Таблица критериев. По ней видно, что значение p у Шапиро-Уилка меньше 0.05, а это значит, что он отвергает гипотезу на уровне значимости 0.05.

Статистики критериев для переменной sepal_length

	1 Критерий	2 Значение критерия	3 p-value
1	Шапиро-Уилка	0.976	0.01
2	Д'Агостина	5.736	0.057
3	Андерсон-Дарлинг	0.889	
4	Коэффициент асимметрии	1.596	0.11
5	Коэффициент эксцесса	-1.785	0.074

По гистограмме видим, что данные имеют не колоколообразное распределение, так как не все колонки гистограммы лежат на красной линии (плотности нормального распределения), а это значит, что данные с большей вероятностью имеют нормальное распределение.



Вывод. Посмотрев на результаты графиков и критериев, можно утверждать, что данные не имеют нормальное распределение на уровне значимости 0.05.

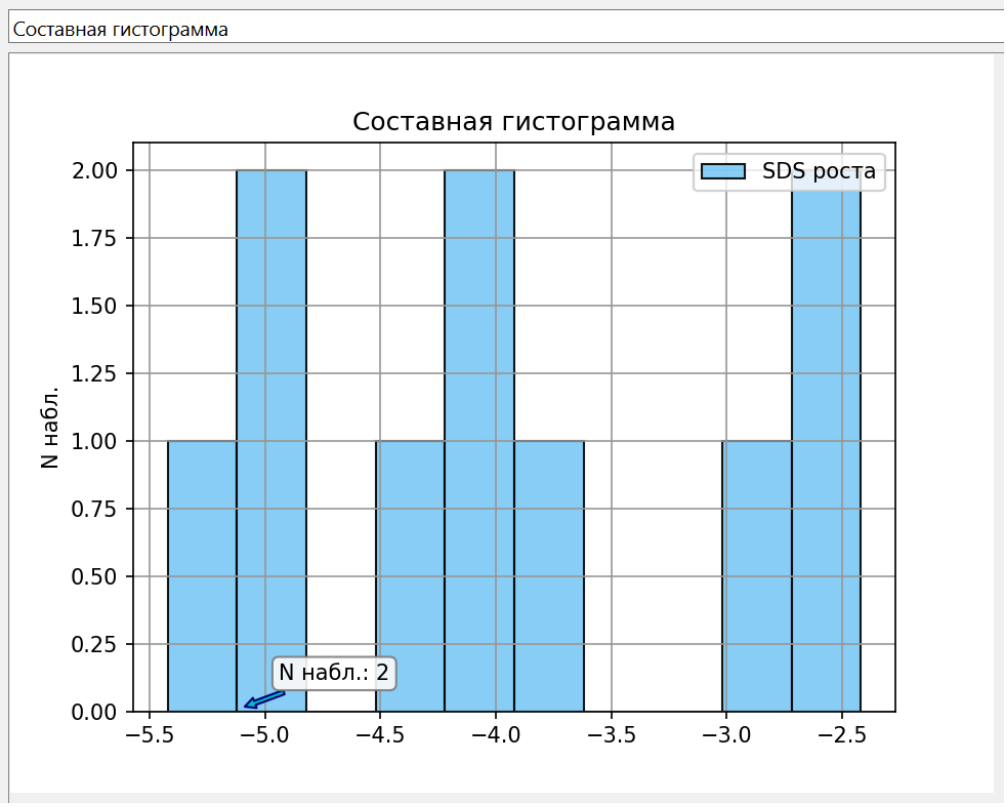
Пример 5. Непараметрическая статистика

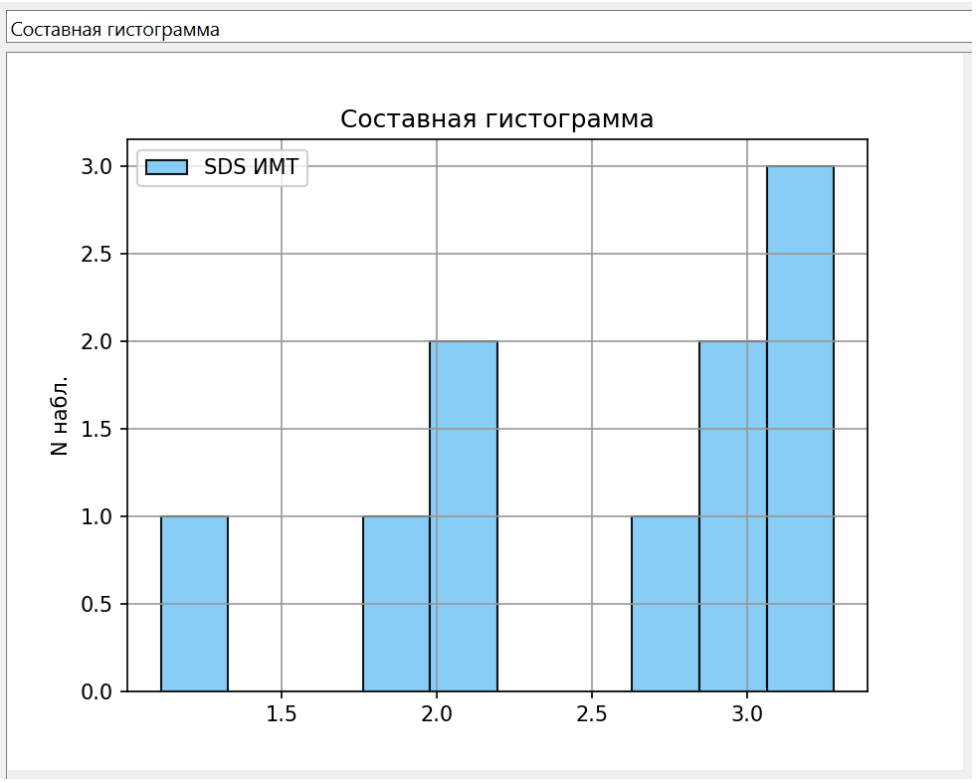
Имеется таблица, содержащая медицинские показатели для 10 пациентов разного пола и возраста (от 0 до 18 лет). Необходимо выяснить, имеется ли статистически значимая разница для переменных SDS роста и SDS ИМТ (см. замечание в конце статьи) в зависимости от пола и возраста пациента. Эти данные содержатся в файле Пациенты.csv, фрагмент которого представлен на изображении ниже. Откройте этот файл с помощью меню Файл - Открыть.

Пациенты

	1 Пол	2 Дата рождения	3 Рост (см)	4 SDS роста	5 ий сегме	6 ерхний се	7 ий сегме	8 жний се	9 Вес (кг)	10 ИМТ (г/м2)	11 SDS ИМТ	12 ость голс	13 ужность	14 КВ (г)	15 Рост отца	16 Рост матери	17 Возраст	18 SDSимт
1	Ж	17.09.2013	112.0	-2.5					23.3	18.57	1.11	54.0	1.75	9.0	170	165	5-9 лет	избыток
2	М	22.06.2018	78.0	-4.84	55.0	-1.1	23.0	-7.72	12.5	20.55	3.12	54.5	2.74	2.3	173	172	0-4 лет	ожирение 3 ст
3	Ж	22.10.2018	78.5	-3.93					13.0	21.1	2.98	54.7	4.22	1.6	172	160	0-4 лет	ожирение 2 ст
4	М	18.03.2018	80.0	-5.42					12.8	19.53	2.64			3.5	176	164	0-4 лет	ожирение 2 ст
5	Ж	05.04.2017	85.0	-4.35	58.2	-1.0	26.8	-6.47	14.0	19.38	2.16	54.0	2.65	3.6	183	170	5-9 лет	ожирение 1 ст
6	Ж	31.07.2020	73.0	-3.8					10.5	19.7	2.1	54.0	4.43	1.6	170	164	0-4 лет	ожирение 1 ст
7	М	21.06.2015	97.0	-3.97	68.0	0.9	28.0	-7.4	21.0	22.79	3.28	60.0	6.18	6.0	175	160	5-9 лет	ожирение 3 ст
8	Ж	28.07.2015	93.0	-4.86					17.0	19.66	1.78	54.5	2.39	6.8	148	160	5-9 лет	избыток
9	Ж	02.12.2010	121.5	-3.02	78.3	0.7		-5.7	46.9	31.77	3.22	55.0	1.72	14.0	170	165	10-14 …	ожирение 3 ст
10	Ж	28.06.2017	94.3	-2.42	63.5	1.1	30.8	-5.06	19.0	21.37	2.92	55.0	3.44	4.8	156	168	5-9 лет	ожирение 2 ст

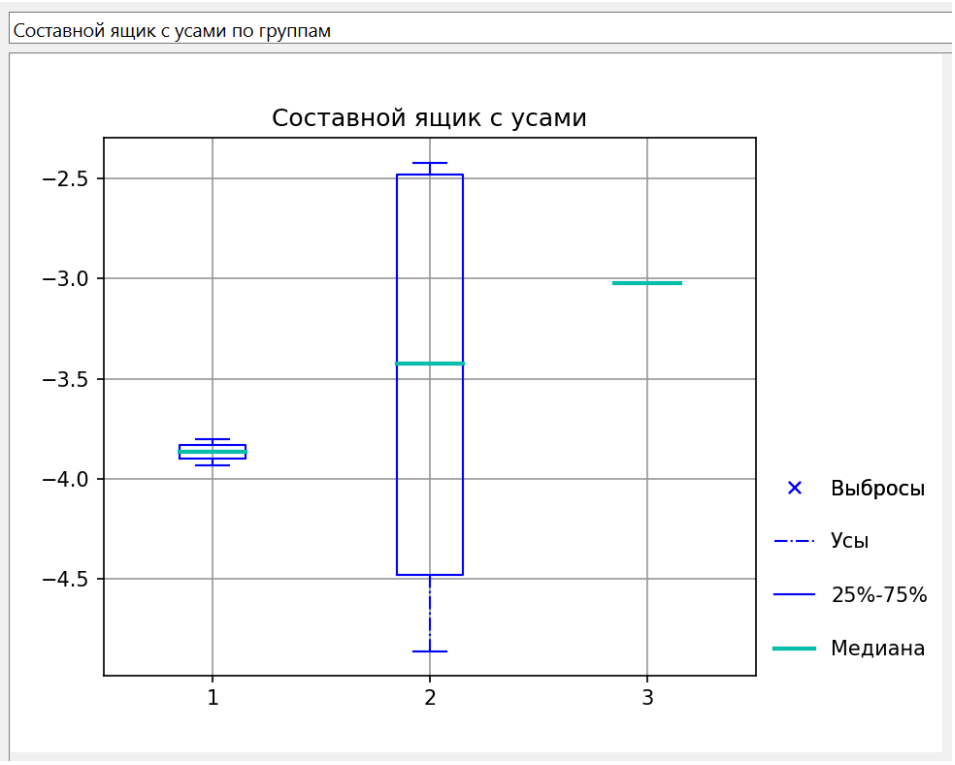
Предварительный анализ данных. В качестве первого этапа анализа построим гистограммы для изучаемых переменных.



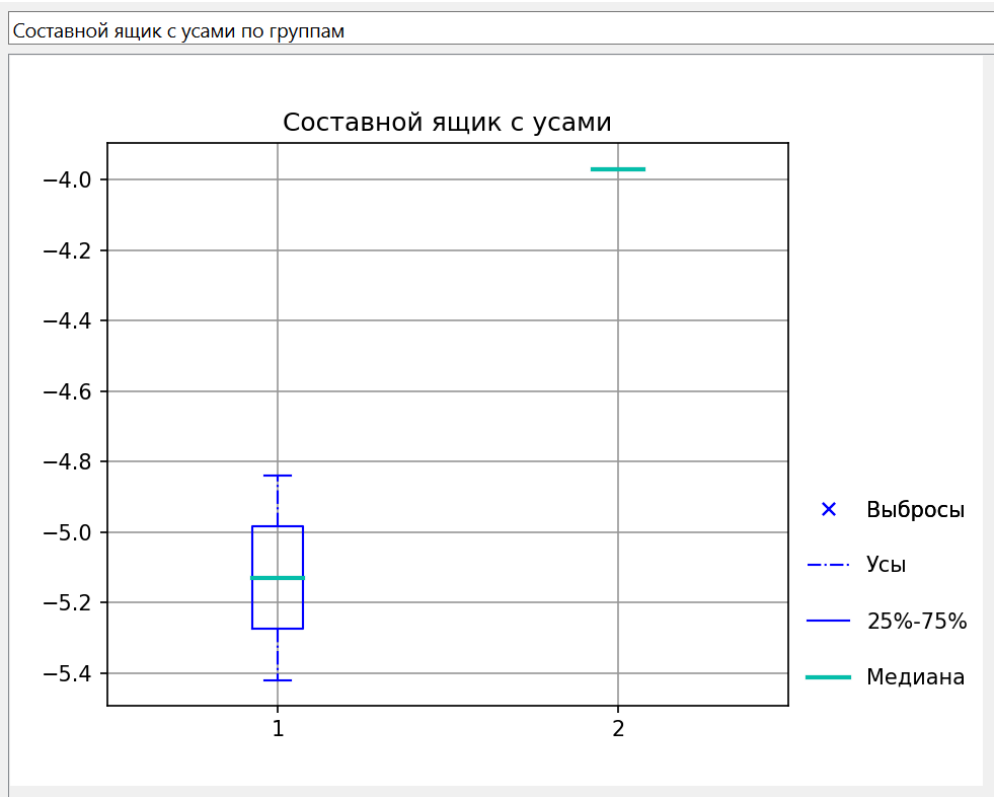


Из полученных графиков видим, что данные распределены ненормально, а значит для сравнения групп в данном случае необходимо использовать методы непараметрической статистики.

Теперь построим диаграммы размаха для переменной SDS роста отдельно для пациентов женского и мужского пола, сгруппировав данные по возрастам.



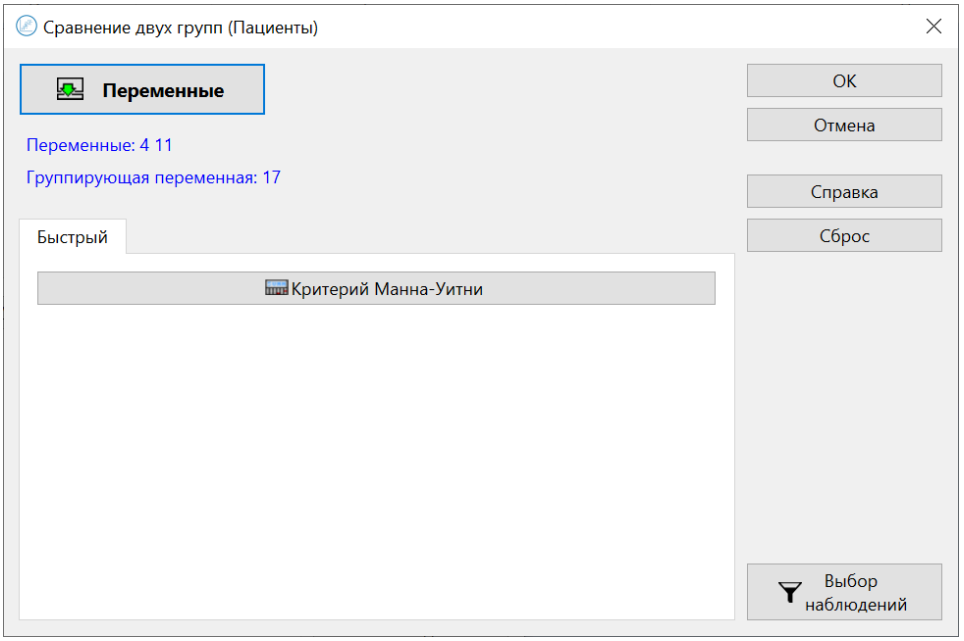
Видим, что группа Ж 10-14 лет слишком малочисленна, чтобы включать ее в анализ, а значит в дальнейшем будем сравнивать между собой только группы Ж 0-4 лет и Ж 5-9 лет.



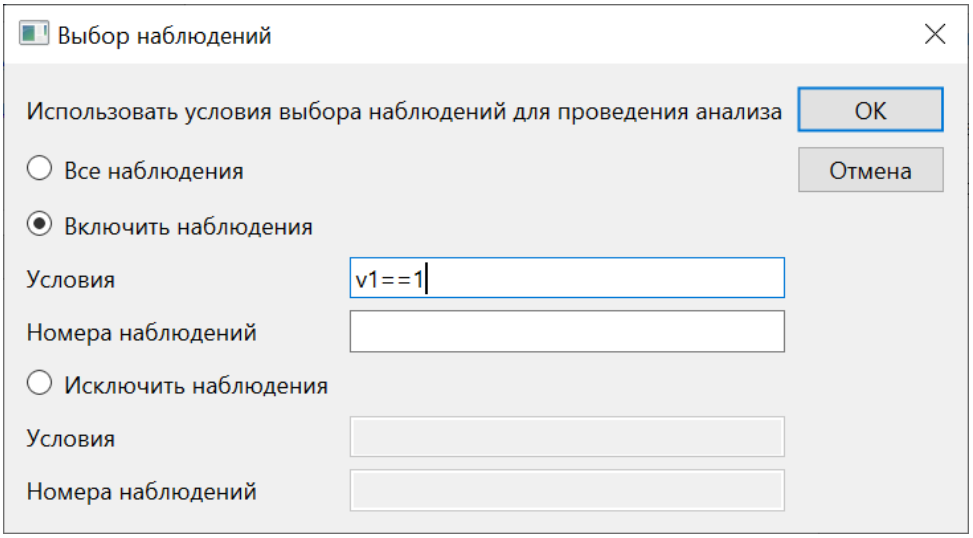
Аналогичную картину с недостаточным количеством наблюдений видим для группы М 5-9, но поскольку это одна из всего двух представленных для сравнения групп у пациентов мужского пола, то для мальчиков анализ проводить не имеет смысла.

После проведения предварительного анализа представленных данных, можем переходить к непосредственному сравнению оставшихся для рассмотрения групп пациентов.

Выбор анализа. Выберите команду Непараметрическая статистика в меню Анализ для отображения стартовой панели модуля Непараметрическая статистика. Затем выберите опцию Сравнение двух независимых групп и нажмите кнопку ОК для отображения диалогового окна Сравнение двух групп. Нажмите кнопку Переменные для отображения диалогового окна Выбор переменных. В списке выберите переменные SDS роста и SDS ИМТ в качестве зависимых переменных и переменную Возраст в качестве группирующей переменной.



Теперь нажмите кнопку Выбор наблюдений и в открывшемся диалоговом окне задайте ограничение на переменную пол так, чтобы в анализе учитывались только пациенты женского пола (см. изображение ниже). Далее нажмите OK.



Просмотр результатов. Теперь нажмите кнопку OK в окне Сравнение двух групп, чтобы выполнить анализ. Таблица с результатами появится на экране.

Критерий Манна-Уитни по группам 1 & 2

	1 Сум. ранг 1	2 Сум. ранг 2	3 U	4 p-уровень	5 N 1	6 N 2
SDS роста	1.0	9.0	0.0	0.5	1.0	3.0
SDS ИМТ	3.0	7.0	1.0	1.0	1.0	3.0

Как можно увидеть из таблицы, разница между возрастными группами 1 (0-4 лет) и 2 (5-9 лет) в этом исследовании в отношении переменных SDS роста и SDS ИМТ не имеет статистической значимости ($p > 0.05$).

Замечание. SDS (Standard Deviation Score) - коэффициент стандартного отклонения - интегральный показатель, применяемый для оценки соответствия индивидуального роста ребенка референсным для соответствующего возраста и пола данным. SDS показывает, сколько стандартных (сигмальных) отклонений составляет разница между средним арифметическим и измеренным значением.

Пример 6. Векторная авторегрессия

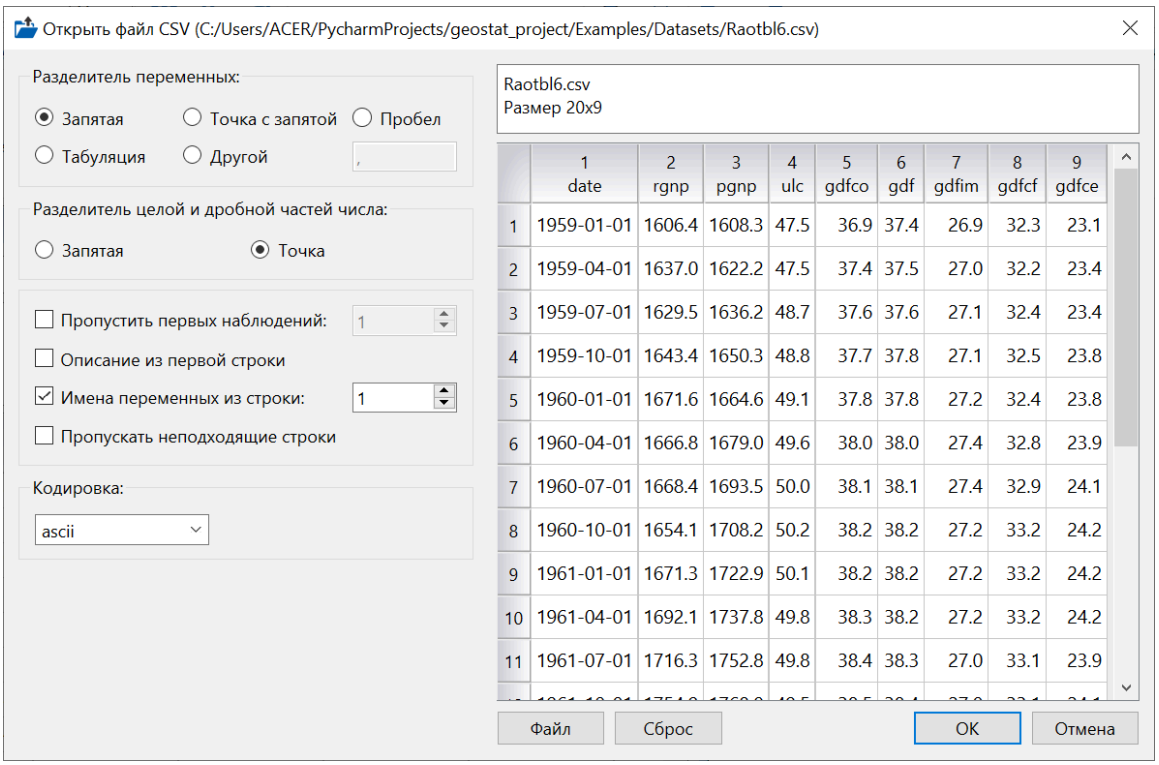
Рассмотрим пример построения модели векторной авторегрессии для прогнозирования значений нескольких временных рядов, взаимно влияющих друг на друга.

Структура данных. Для этого примера рассмотрим файл данных, использовавшийся в статье Яша П. Мехры 1994 года: «Рост заработной платы и инфляционный процесс: эмпирический подход».

Этот набор данных содержит следующие 8 квартальных временных рядов:

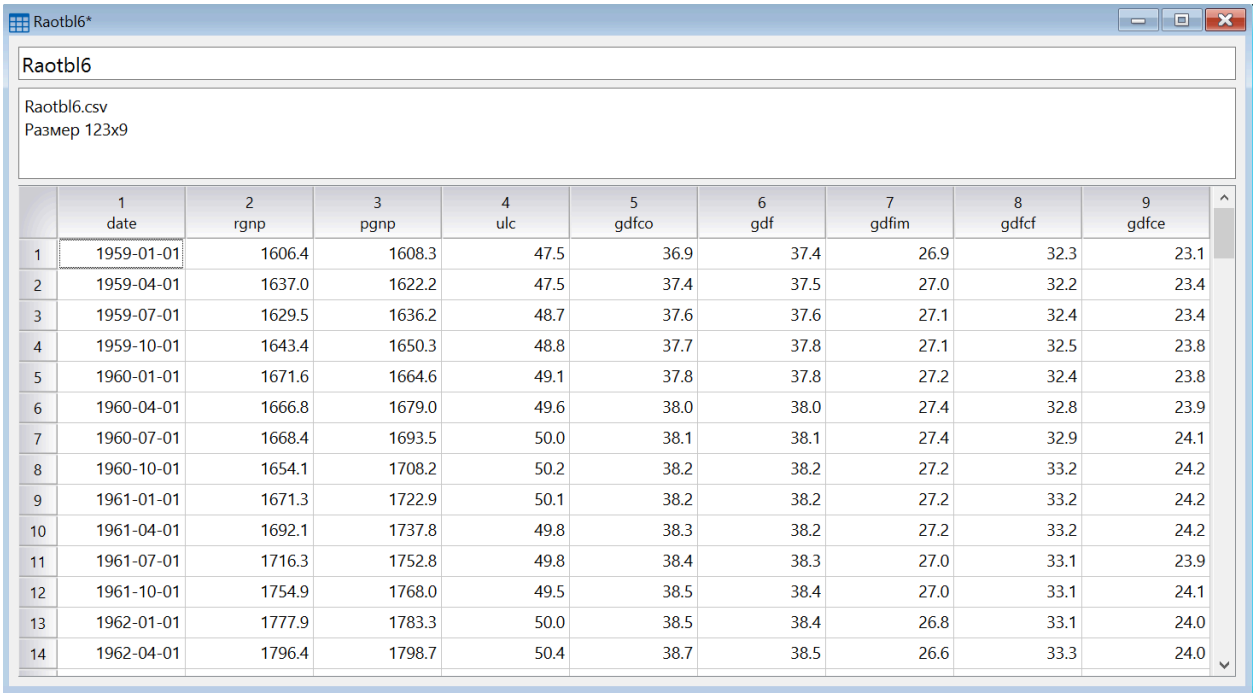
- rgnp – реальный ВНП (валовый национальный продукт)
- pgnp – потенциальный реальный ВНП
- ulc – стоимость рабочей силы
- gdfco – дефлятор фиксированного веса для расходов на личное потребление, исключая продукты питания и энергию
- gdf – дефлятор фиксированного веса для ВНП
- gdfim – дефлятор фиксированного веса для импорта
- gdfcf – дефлятор фиксированного веса для продуктов питания в расходах на личное потребление
- gdfce – дефлятор фиксированного веса для энергии в расходах на личное потребление

Загрузка данных. С помощью меню Файл -– Открыть папку примеров найдите и откройте файл Raotbl6.csv. В окне импорта задайте для файла необходимые настройки:



	1 date	2 rgnp	3 pgnp	4 ulc	5 gdfco	6 gdf	7 gdfim	8 gdfcf	9 gdfce
1	1959-01-01	1606.4	1608.3	47.5	36.9	37.4	26.9	32.3	23.1
2	1959-04-01	1637.0	1622.2	47.5	37.4	37.5	27.0	32.2	23.4
3	1959-07-01	1629.5	1636.2	48.7	37.6	37.6	27.1	32.4	23.4
4	1959-10-01	1643.4	1650.3	48.8	37.7	37.8	27.1	32.5	23.8
5	1960-01-01	1671.6	1664.6	49.1	37.8	37.8	27.2	32.4	23.8
6	1960-04-01	1666.8	1679.0	49.6	38.0	38.0	27.4	32.8	23.9
7	1960-07-01	1668.4	1693.5	50.0	38.1	38.1	27.4	32.9	24.1
8	1960-10-01	1654.1	1708.2	50.2	38.2	38.2	27.2	33.2	24.2
9	1961-01-01	1671.3	1722.9	50.1	38.2	38.2	27.2	33.2	24.2
10	1961-04-01	1692.1	1737.8	49.8	38.3	38.2	27.2	33.2	24.2
11	1961-07-01	1716.3	1752.8	49.8	38.4	38.3	27.0	33.1	23.9

После этого нажмите кнопку ОК, файл будет успешно импортирован в программу:



Raotbl6*

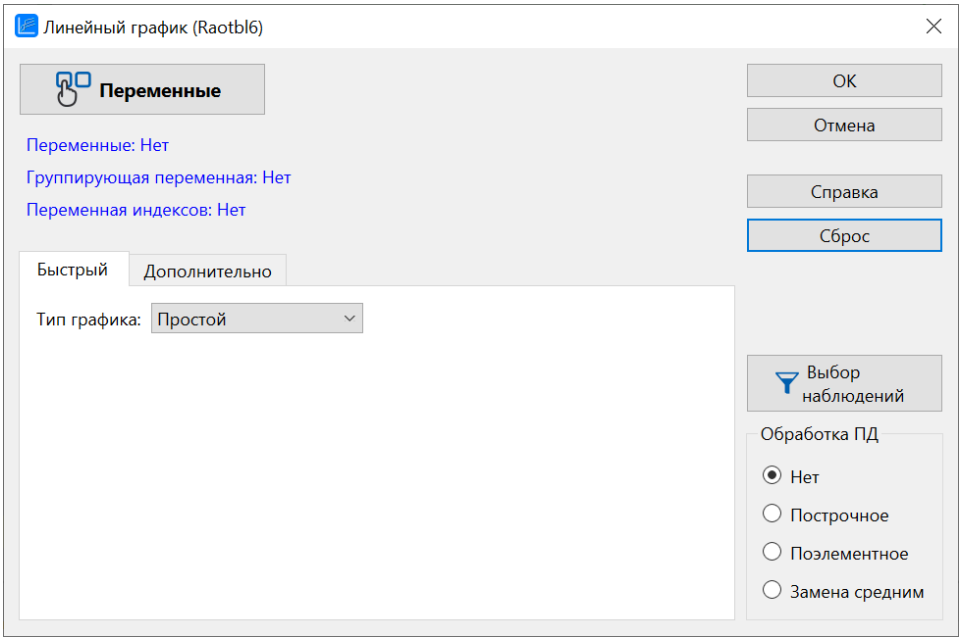
Raotbl6

Raotbl6.csv
Размер 123x9

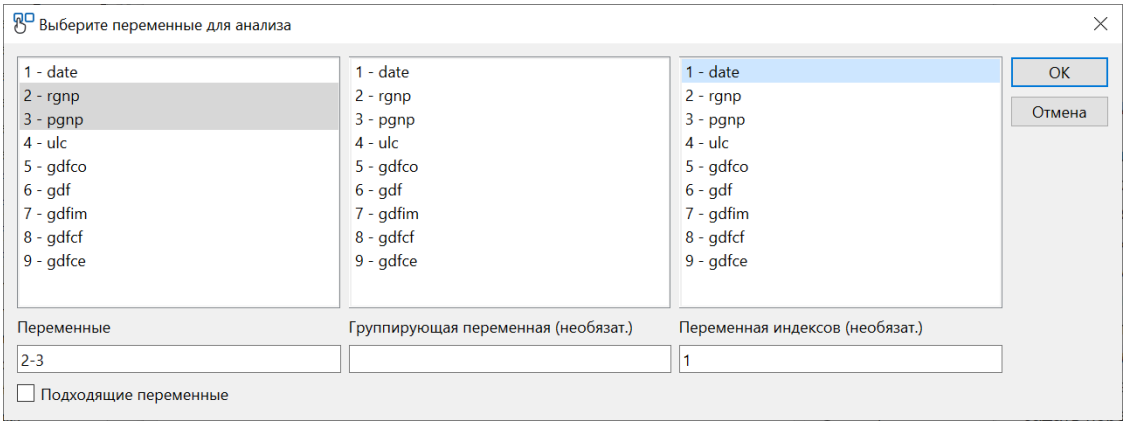
	1 date	2 rgnp	3 pgnp	4 ulc	5 gdfco	6 gdf	7 gdfim	8 gdfcf	9 gdfce
1	1959-01-01	1606.4	1608.3	47.5	36.9	37.4	26.9	32.3	23.1
2	1959-04-01	1637.0	1622.2	47.5	37.4	37.5	27.0	32.2	23.4
3	1959-07-01	1629.5	1636.2	48.7	37.6	37.6	27.1	32.4	23.4
4	1959-10-01	1643.4	1650.3	48.8	37.7	37.8	27.1	32.5	23.8
5	1960-01-01	1671.6	1664.6	49.1	37.8	37.8	27.2	32.4	23.8
6	1960-04-01	1666.8	1679.0	49.6	38.0	38.0	27.4	32.8	23.9
7	1960-07-01	1668.4	1693.5	50.0	38.1	38.1	27.4	32.9	24.1
8	1960-10-01	1654.1	1708.2	50.2	38.2	38.2	27.2	33.2	24.2
9	1961-01-01	1671.3	1722.9	50.1	38.2	38.2	27.2	33.2	24.2
10	1961-04-01	1692.1	1737.8	49.8	38.3	38.2	27.2	33.2	24.2
11	1961-07-01	1716.3	1752.8	49.8	38.4	38.3	27.0	33.1	23.9
12	1961-10-01	1754.9	1768.0	49.5	38.5	38.4	27.0	33.1	24.1
13	1962-01-01	1777.9	1783.3	50.0	38.5	38.4	26.8	33.1	24.0
14	1962-04-01	1796.4	1798.7	50.4	38.7	38.5	26.6	33.3	24.0

Всего файл включает 9 переменных и 123 наблюдения.

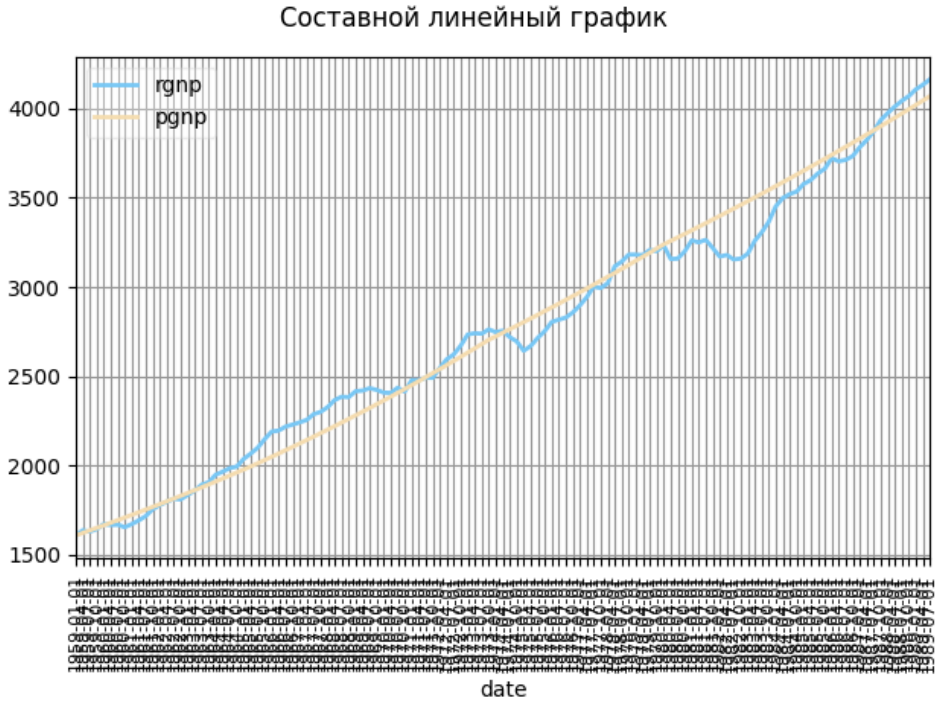
Визуализация. До проведения основного анализа можем построить линейные графики для некоторых или всех исследуемых переменных. К примеру, можем посмотреть, как отличаются реальные и потенциальные показатели ВНП. Для этого в меню Графики выберите Линейный график. Откроется окно анализа, в котором необходимо задать переменные и настройки для графиков:



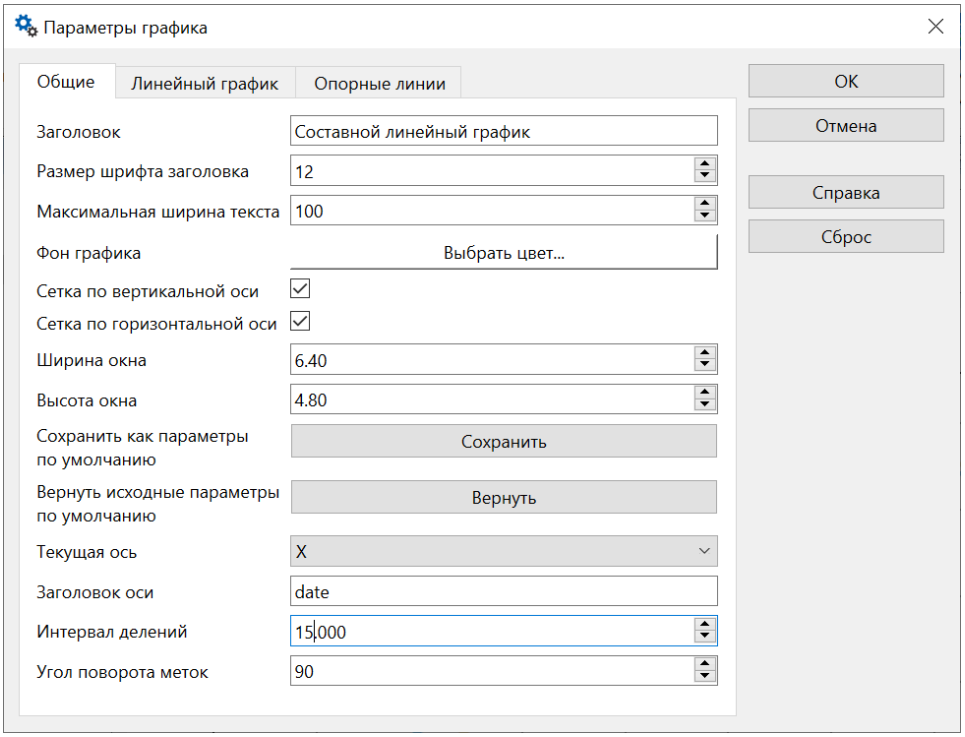
Нажмите на кнопку Переменные и в открывшемся окне выберите переменные 2-3 в качестве основных и переменную 1 в качестве переменной индексов:



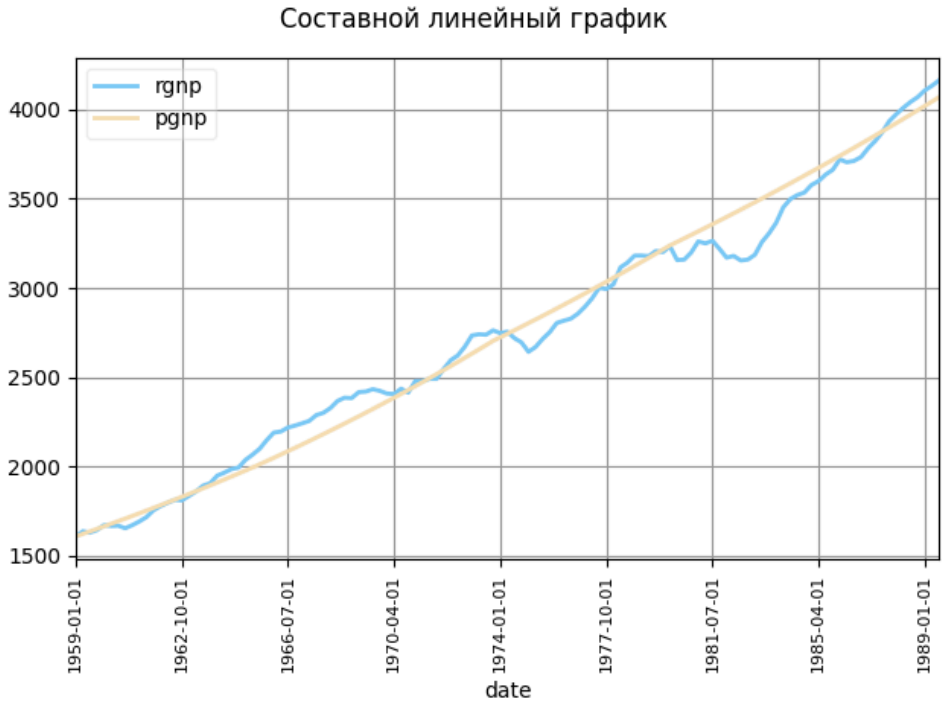
В графе Тип графика укажите Составной и нажмите кнопку OK. Построенный график отобразится в рабочей книге:



Так как наблюдений достаточно много, индексы накладываются друг на друга и создают слишком плотную сетку. Чтобы настроить внешний вид графика, дважды нажмите на него левой кнопкой мыши, будет открыто окно настроек линейного графика. В этом окне укажите более редкий интервал делений для оси X, например 15:



После этого график примет более красивый и лаконичный вид:

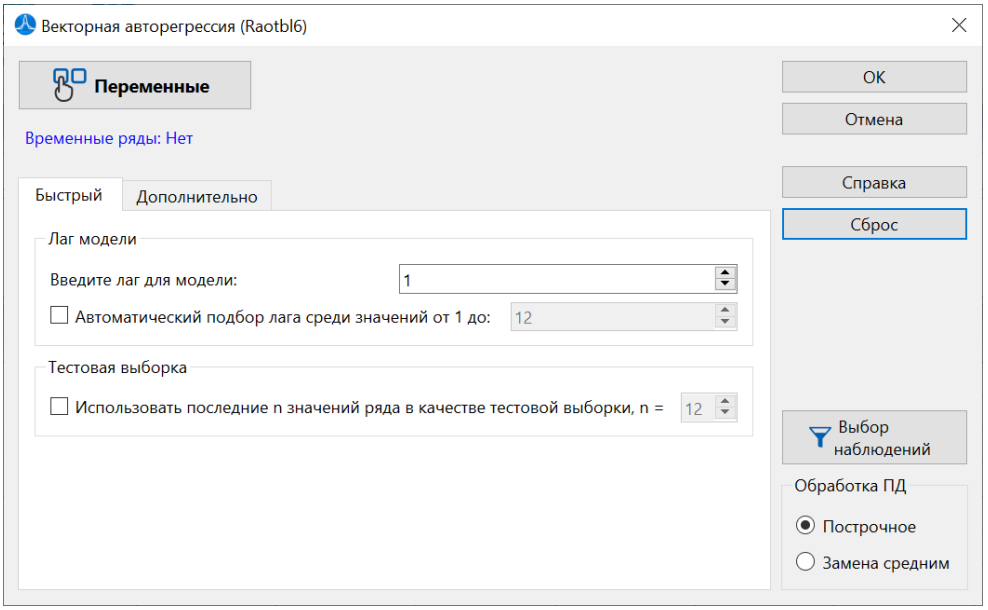


Таким образом, мы провели первичную визуализацию двух из восьми исследуемых временных рядов, аналогичные действия можно повторить и для остальных.

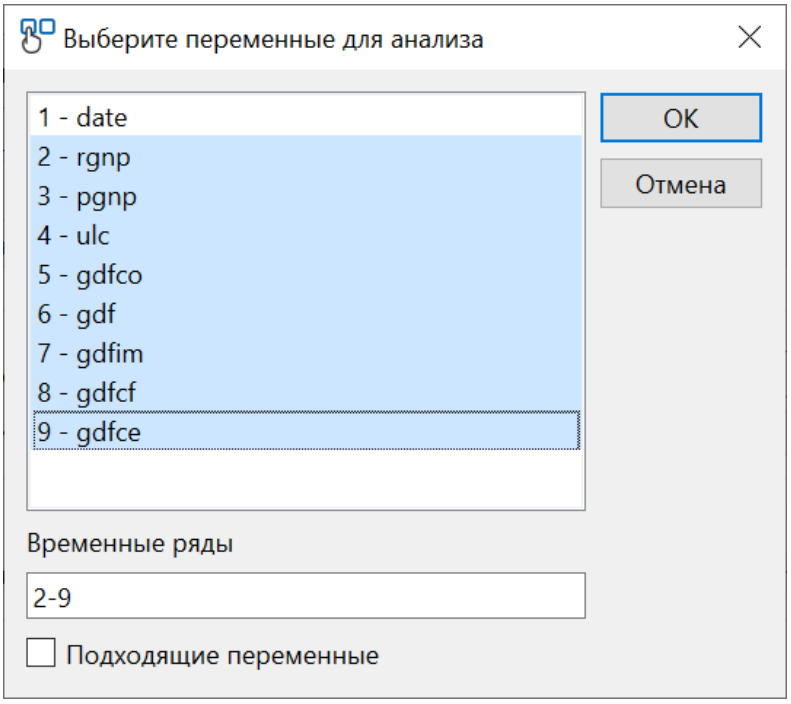
Теперь перейдем к построению модели векторной авторегрессии.

Векторная авторегрессия. Проведем анализ по шагам.

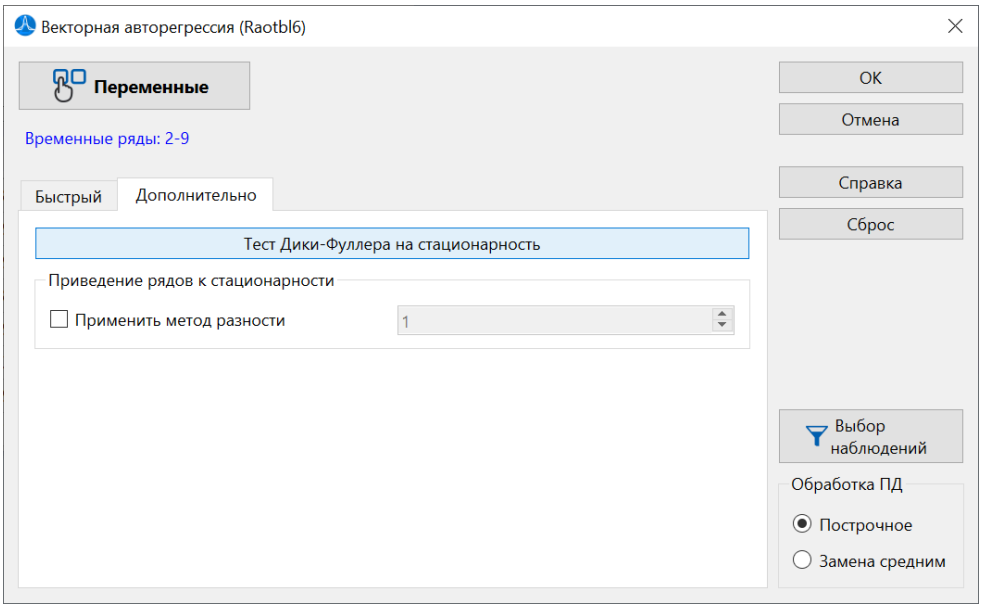
Шаг 1 – выбор модуля анализа. На вкладке меню Временные ряды выберите пункт Векторная авторегрессия. Откроется окно соответствующего модуля:



Шаг 2 – выбор переменных. Нажмите на кнопку Переменные для выбора временных рядов для анализа. Укажите в появившемся списке 8 интересующих нас рядов:



Шаг 3 – проверка рядов на стационарность. На вкладке Дополнительно выберите тест Дики-Фуллера на стационарность:

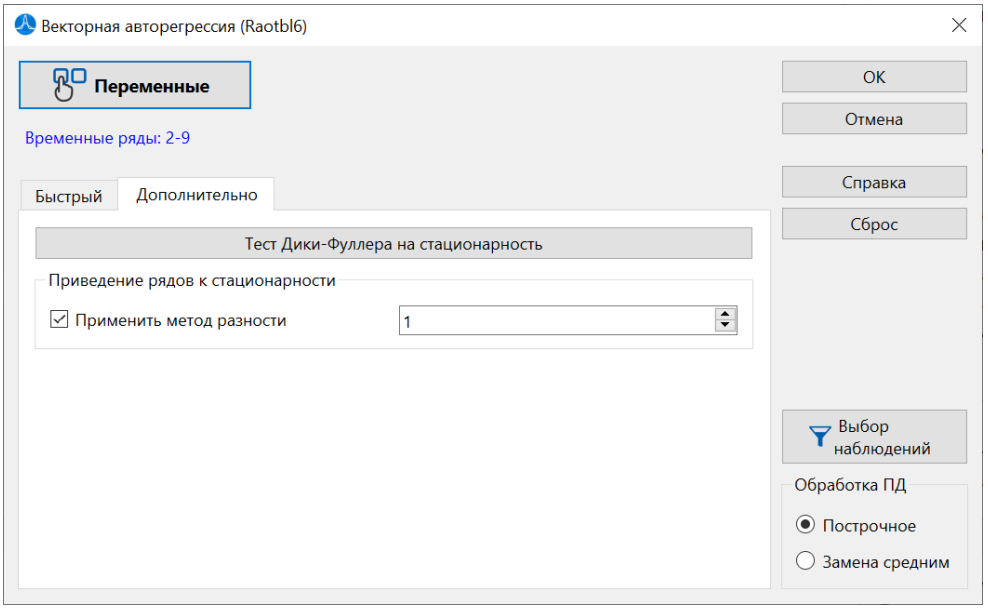


Откроется таблица статистик. По p-уровням значимости (>0.5) видно, что ряды не являются стационарными:

Тест Дики-Фуллера для таблицы Raotbl6

	1 Переменная	2 ADF статистика	3 p-значение	4 Количество лагов	5 Количество наблюдений
1	rgnp	0.642	0.989	2	120
2	pgnp	1.274	0.996	1	121
3	ulc	1.397	0.997	2	120
4	gdfco	0.576	0.987	5	117
5	gdf	1.113	0.995	7	115
6	gdfim	-0.199	0.939	1	121
7	gdfcf	1.669	0.998	9	113
8	gdfce	-0.816	0.814	13	109

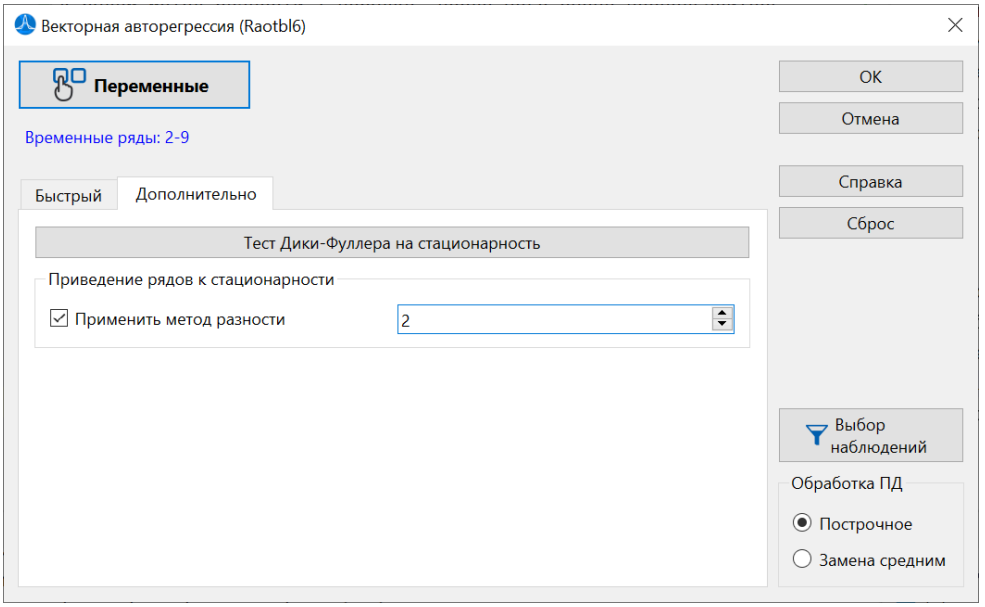
Так как модель векторной авторегрессии подходит для работы только со стационарными рядами, необходимо привести исходные ряды к стационарному состоянию. Для этого вернитесь на вкладку Дополнительно и примените к рядам метод разности 1 порядка, после чего снова воспользуйтесь тестом Дики-Фуллера:



Тест Дики-Фуллера для таблицы Raotbl6 с использованием метода разности

	1 Переменная	2 ADF статистика	3 p-значение	4 Количество лагов	5 Количество наблюдений
1	rgnp	-5.428	0.000	1	120
2	pgnp	-1.759	0.401	0	121
3	ulc	-3.576	0.006	1	120
4	gdfco	-1.093	0.718	4	117
5	gdf	-1.400	0.582	12	109
6	gdfim	-4.244	0.001	0	121
7	gdfcf	-1.824	0.369	6	115
8	gdfce	-2.045	0.267	12	109

По новой таблице видно, что часть рядов приобрела стационарность, а часть – нет. Снова вернитесь на вкладку Дополнительно и примените к рядам метод разности 2 порядка:

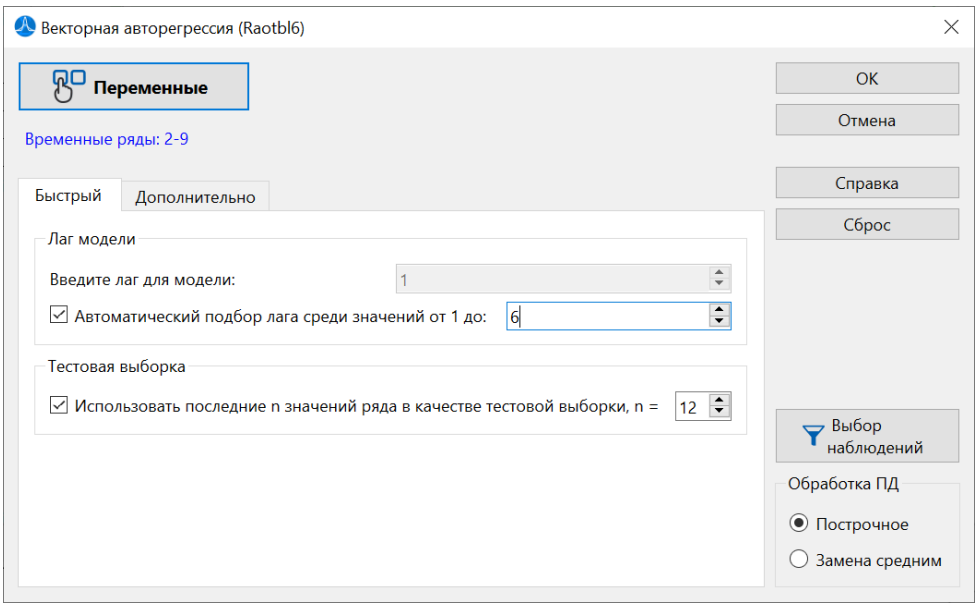


Тест Дики-Фуллера для таблицы Raotbl6 с использованием метода разности_(1)

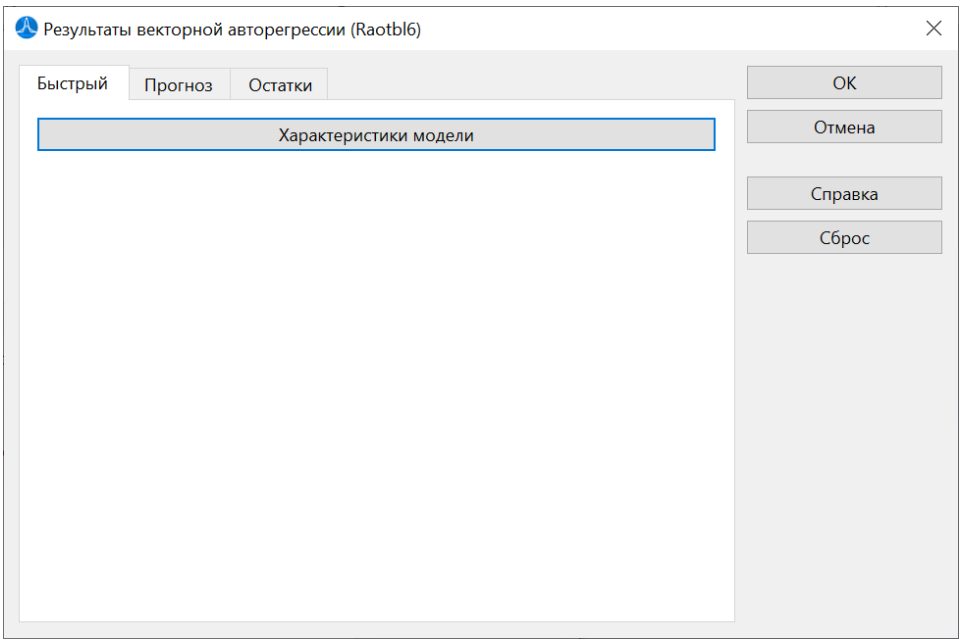
	1 Переменная	2 ADF статистика	3 p-значение	4 Количество лагов	5 Количество наблюдений
1	rgnp	-9.164	$< 10^{-9}$	2	118
2	pgnp	-11.172	$< 10^{-9}$	0	120
3	ulc	-8.838	$< 10^{-9}$	2	118
4	gdfco	-8.772	$< 10^{-9}$	3	117
5	gdf	-4.328	0.000	11	109
6	gdfim	-9.558	$< 10^{-9}$	1	119
7	gdfcf	-7.219	$< 10^{-9}$	5	115
8	gdfce	-4.375	0.000	13	107

После выполненных преобразований все ряды проходят проверку стационарности и могут быть использованы для построения модели.

Шаг 4 – задание параметров модели. Переключитесь на вкладку Быстрый. Здесь можно задать параметры для построения модели – ее лаг и наличие тестовой выборки. Выберите автоматический выбор наилучшего лага модели среди значений от 1 до 6 и тестовую выборку из 12 наблюдений:



Шаг 5 – построение модели и ее характеристики. После задания всех параметров нажмите на кнопку ОК, откроется окно результатов анализа:



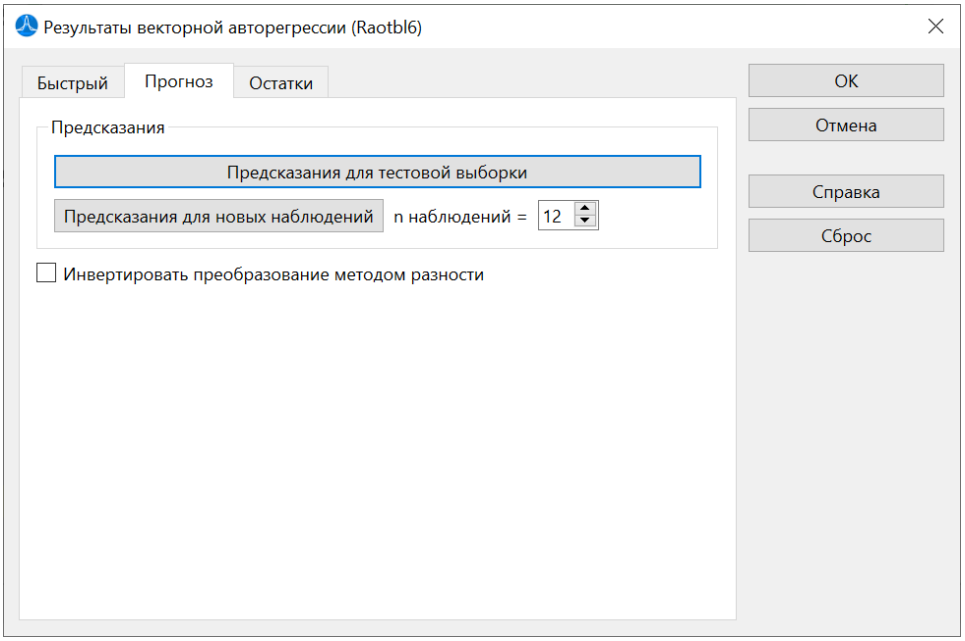
В этом окне нажмите на кнопку Характеристики модели для открытия соответствующей таблицы:

Характеристики векторной авторегрессии

	1 Характеристика	2 Значение
1	AIC	-3.177
2	BIC	6.850
3	FPE	0.081
4	HQIC	0.884
5	Lag	6.000

Из нее видим, что программой была выбрана модель с лагом = 6 в качестве оптимальной.

Шаг 6 – проверка на тестовой выборке. Так как была выбрана тестовая выборка для данных, мы можем проверить результаты построенной модели на ней. Для этого перейдите на вкладку Прогноз и нажмите на кнопку Предсказания для тестовой выборки:



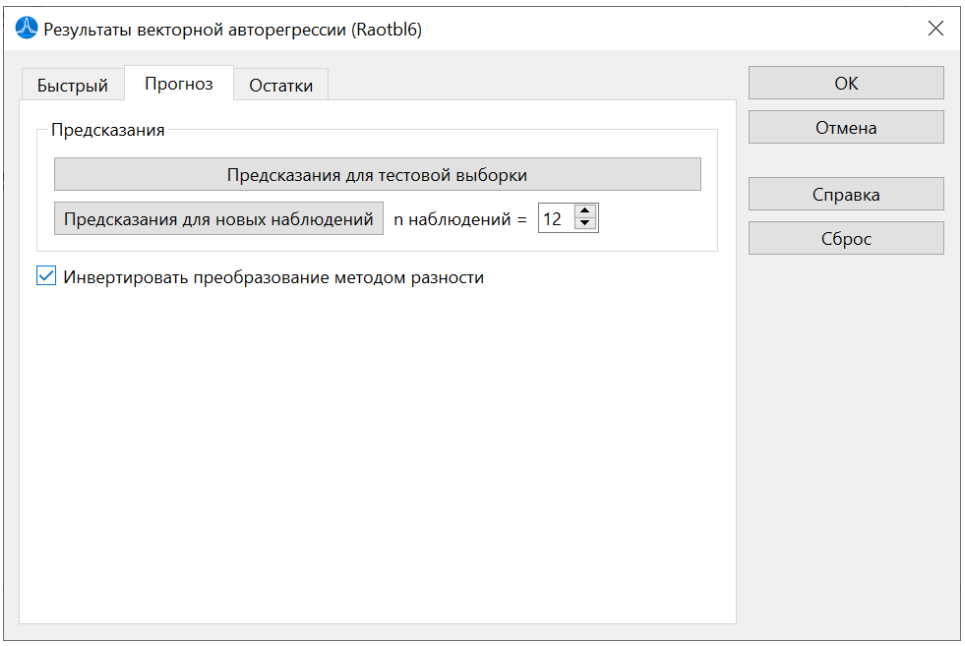
Откроется следующая таблица прогнозов и значений тестовой выборки, всего на 16 переменных:

Предсказанные и наблюдаемые значения для тестовой выборки

	8 gdfce	9 rgnp предсказанное	10 pgnp предсказанное	11 ulc предсказанное	12 gdfco предсказанное	13 gdf предсказанное	14 gdfim предсказанное	15 gdfcf предсказанное	16 gdfce предсказанное
1	1.400	67.130	0.459	-2.598	-1.410	-0.105	-3.067	-0.417	-0.334
2	5.800	-49.584	1.997	1.107	1.363	-0.397	0.567	-1.440	-1.487
3	-1.900	16.077	0.002	-0.249	-0.239	0.693	1.045	1.885	3.667
4	-0.100	-0.070	0.671	0.316	-0.639	-0.716	1.641	-2.220	-2.052
5	-2.000	-75.597	-0.568	0.223	0.957	-0.092	-0.283	0.520	-0.551
6	-1.200	96.078	-0.787	-0.508	-0.742	0.743	-1.506	2.369	-0.192
7	2.200	-61.009	0.514	1.874	0.377	-0.285	3.108	-1.763	1.160
8	-0.300	58.928	0.009	-2.443	0.229	0.106	-0.902	1.144	0.539
9	-0.400	-8.593	0.104	0.101	-0.355	0.056	-1.510	-0.571	-2.597
10	1.000	-46.993	0.329	1.567	0.436	0.049	0.567	0.364	2.512
11	5.500	34.111	-0.454	-0.545	-0.360	0.178	0.020	-0.220	-0.427
12	-9.000	-33.311	0.065	0.632	0.118	-0.131	0.022	-0.176	-1.415

Видим, что значения переменных и прогнозов не соответствуют тем, что были в исходной таблице данных, поскольку ранее к ним был применен метод разности. Чтобы получить

исходные значения, вернитесь на вкладку Прогноз и поставьте галочку в графе Инвертировать преобразование методом разности:



Снова постройте таблицу прогнозов:

Предсказанные и наблюдаемые значения для тестовой выборки_(2)

	8 gdfce	9 rgnp предсказанное	10 pgnp предсказанное	11 ulc предсказанное	12 gdfco предсказанное	13 gdf предсказанное	14 gdfim предсказанное	15 gdfcf предсказанное	16 gdfce предсказанное
1	85.600	3787.330	3808.759	171.302	119.890	116.095	90.233	114.783	83.866
2	89.400	3812.675	3833.815	172.511	121.342	116.493	88.133	115.026	78.646
3	91.300	3854.097	3858.874	173.471	122.555	117.584	87.077	117.154	77.093
4	93.100	3895.449	3884.603	174.747	123.128	117.959	87.662	117.062	73.488
5	92.900	3861.204	3909.765	176.246	124.659	118.242	87.964	117.490	69.332
6	91.500	3923.037	3934.140	177.237	125.447	119.268	86.761	120.288	64.984
7	92.300	3923.862	3959.029	180.103	126.612	120.009	88.665	121.322	61.796
8	92.800	3983.614	3983.926	180.525	128.006	120.857	89.666	123.500	59.148
9	92.900	4034.773	4008.928	181.049	129.045	121.760	89.158	125.108	53.903
10	94.000	4038.939	4034.258	183.141	130.520	122.713	89.216	127.079	51.169
11	100.600	4077.216	4059.134	184.687	131.634	123.843	89.295	128.831	48.008
12	98.200	4082.181	4084.075	186.865	132.867	124.842	89.395	130.407	43.433

Теперь можно оценить, насколько близки оказались полученные значения и хотим ли мы использовать данную модель для предсказания новых наблюдений.

Шаг 7 – анализ остатков модели. На вкладке Остатки можно построить таблицу остатков модели, а также гистограммы и нормальные вероятностные графики для них:

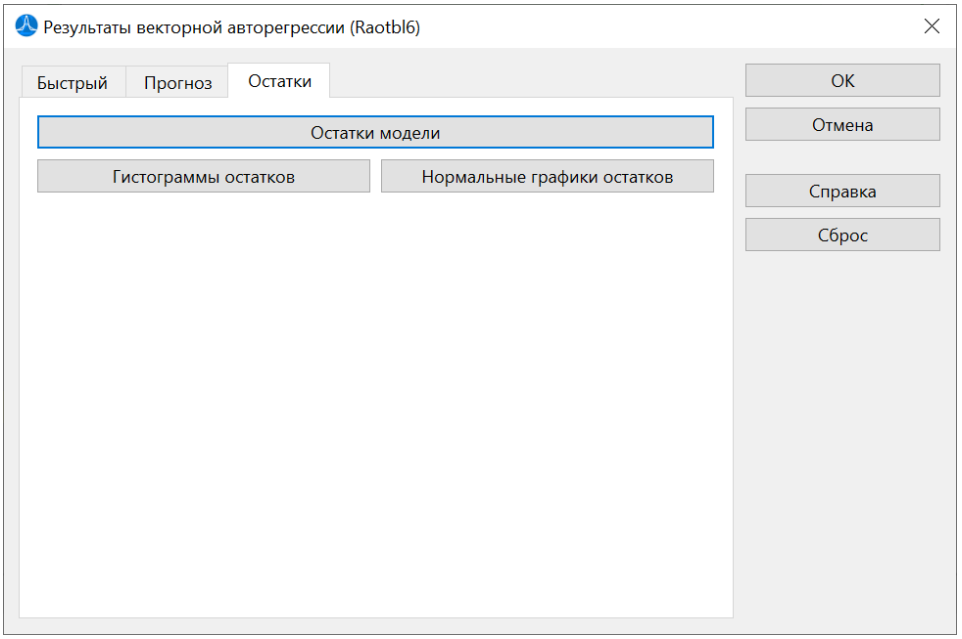
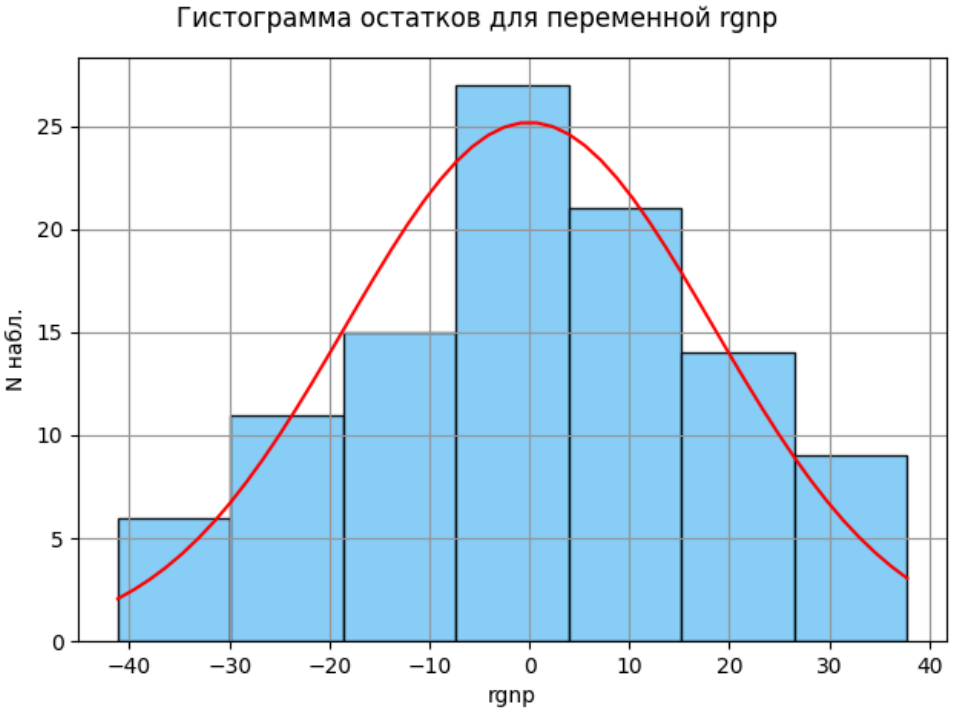
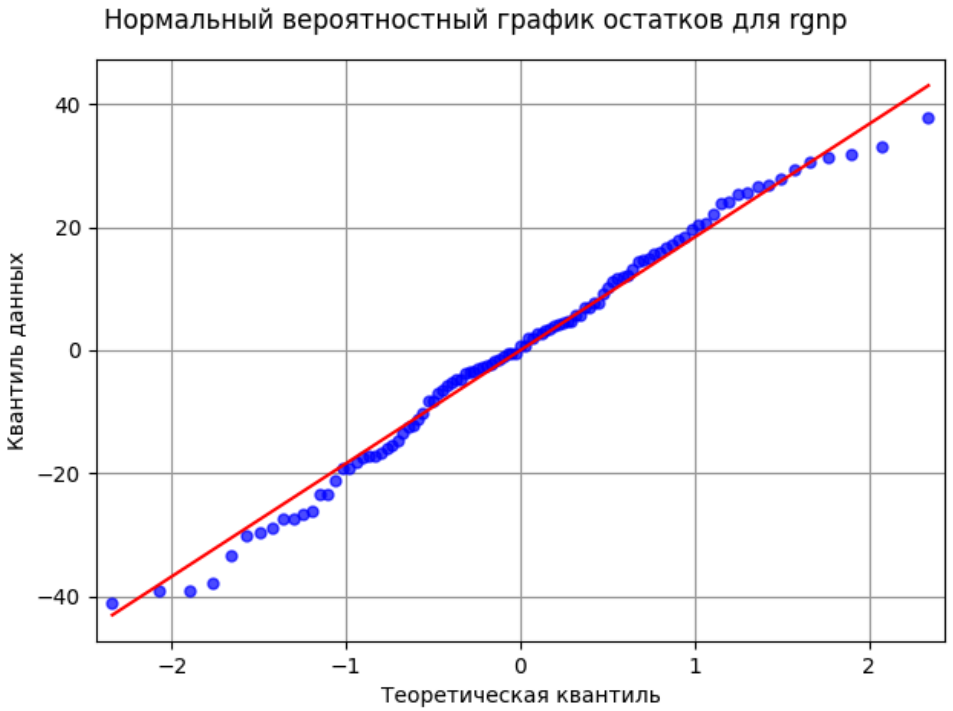


Таблица остатков модели

	1 rgnp	2 pgnp	3 ulc	4 gdfco	5 gdf	6 gdfim	7 gdfcf	8 gdfce
1	-0.422	0.088	-0.022	-0.328	-0.093	-0.463	-0.302	-0.426
2	4.048	0.099	0.002	0.235	0.002	-0.601	0.201	-1.165
3	7.687	-0.179	-0.101	-0.012	0.142	-0.368	-0.133	-0.397
4	30.567	0.098	-0.348	0.059	0.032	0.764	0.052	1.376
5	3.264	0.269	0.036	-0.225	-0.157	0.016	-0.413	-0.141
6	-3.801	-0.254	0.370	0.218	0.048	-0.358	0.370	-0.152
7	-6.543	-0.199	-0.874	0.104	0.055	0.235	-0.211	0.335
8	-18.252	-0.141	-0.240	0.602	-0.064	0.311	-0.185	0.975
9	10.255	0.817	0.478	-0.295	-0.003	-0.035	0.154	-0.197
10	-12.466	-0.063	-0.071	-0.728	-0.100	-0.448	-0.159	-0.361
11	19.633	-0.184	-0.288	-0.149	-0.056	0.133	0.001	-0.990
12	-16.641	0.280	0.140	-0.259	-0.023	-0.234	-0.119	-0.847
13	18.399	-0.262	-0.210	0.114	0.142	0.231	-0.012	0.977
14	12.142	0.253	0.223	-0.050	0.028	0.237	-0.272	0.392
15	-10.159	0.495	0.253	0.101	-0.021	-0.657	-0.039	-0.804
16	-4.780	-0.636	-0.354	0.093	-0.093	-0.196	-0.179	0.364

Для примера посмотрим, как выглядят графики остатков для переменной rgnp:





Выводы. В данном примере было рассмотрено пошаговое проведение анализа на нескольких временных рядах с помощью модели векторной авторегрессии. Была

проведена визуализация данных, выбраны параметры для модели и рассмотрен анализ полученных результатов.

Машинное обучение

Добыча данных

Data Mining в ПО СтатСофт представляют собой систематический метод построения расширенных аналитических моделей для связывания одной или нескольких целевых (зависимых) величин с рядом входных (независимых) переменных-предикторов.

Целевые переменные могут быть непрерывными или категориальными. Непрерывные целевые переменные обычно связаны с задачами регрессии, а категориальные переменные используются в задачах классификации. ПО СтатСофт может работать с обоими типами переменных и, таким образом, способна строить прогностические модели для решения задач регрессии и классификации.

Data Mining в ПО СтатСофт — это комплексное решение, которое превращает процесс построения прогностической модели и анализа данных в систематический и пошаговый процесс. Построение модели начинается с предварительного анализа данных, предварительной обработки переменных анализа, уменьшения размерности и устранения любой избыточности, которая может существовать в наборе данных.

После завершения определения и подготовки данных можно создавать различные прогностические модели (такие как нейронные сети, методы опорных векторов, деревья и т. д.) для моделирования значений целевых данных из входных переменных. За этим шагом следует оценка модели и, наконец, самое важное, развертывание модели, в котором прогностические модели могут использоваться для прогнозирования невидимых (новых) данных (например, для «оценки» баз данных).

Сохранение и внедрение моделей машинного обучения

ПО СтатСофт предоставляет возможность сохранять построенные модели в виде отдельных файлов для их последующих загрузки и использования в приложении.

Python-объекты, представляющие собой предсказательные модели внутри программы, конвертируются в поток байтов и сохраняются в бинарном файле с расширением «.pickle». При загрузке этого файла в программу он десериализуется и позволяет снова использовать сохраненные объекты Python. Такие файлы могут использоваться только в ПО СтатСофт.

Возможность сохранять файлы в таком формате имеется в следующих модулях:

- Вкладка Анализ
  - Множественная регрессия
  - Множественная нелинейная регрессия
  - Обобщенные линейные модели
  - Логит модель
  - Пробит модель
  - Полиномиальная логит модель
  - Логит для порядковой шкалы
  - Пробит для порядковой шкалы
  - Нормальная лог модель
- Вкладка ИИ Регрессия
  - Деревья регрессии
  - Нейронные сети
  - Градиентный бустинг XGBoost
  - Множественная регрессия
  - Выбор лучшей модели
- Вкладка ИИ Классификация
  - Деревья классификации
  - Нейронные сети
  - Градиентный бустинг XGBoost
  - Выбор лучшей модели
- Вкладка Разведочный анализ
  - Метод опорных векторов
  - Случайные леса

Модули множественной регрессии и множественной нелинейной регрессии со вкладки Анализ позволяют дополнительно сохранить модель в формате PMML («.pmml») для совместимости с другими приложениями, поддерживающими данный формат.

PMML – это язык разметки, созданный для описания предсказательных моделей, с определенной структурой. Файлы такого формата не могут быть повторно загружены в ПО СтатСофт, но могут быть использованы в приложениях, поддерживающих формат PMML, например в STATISTICA.

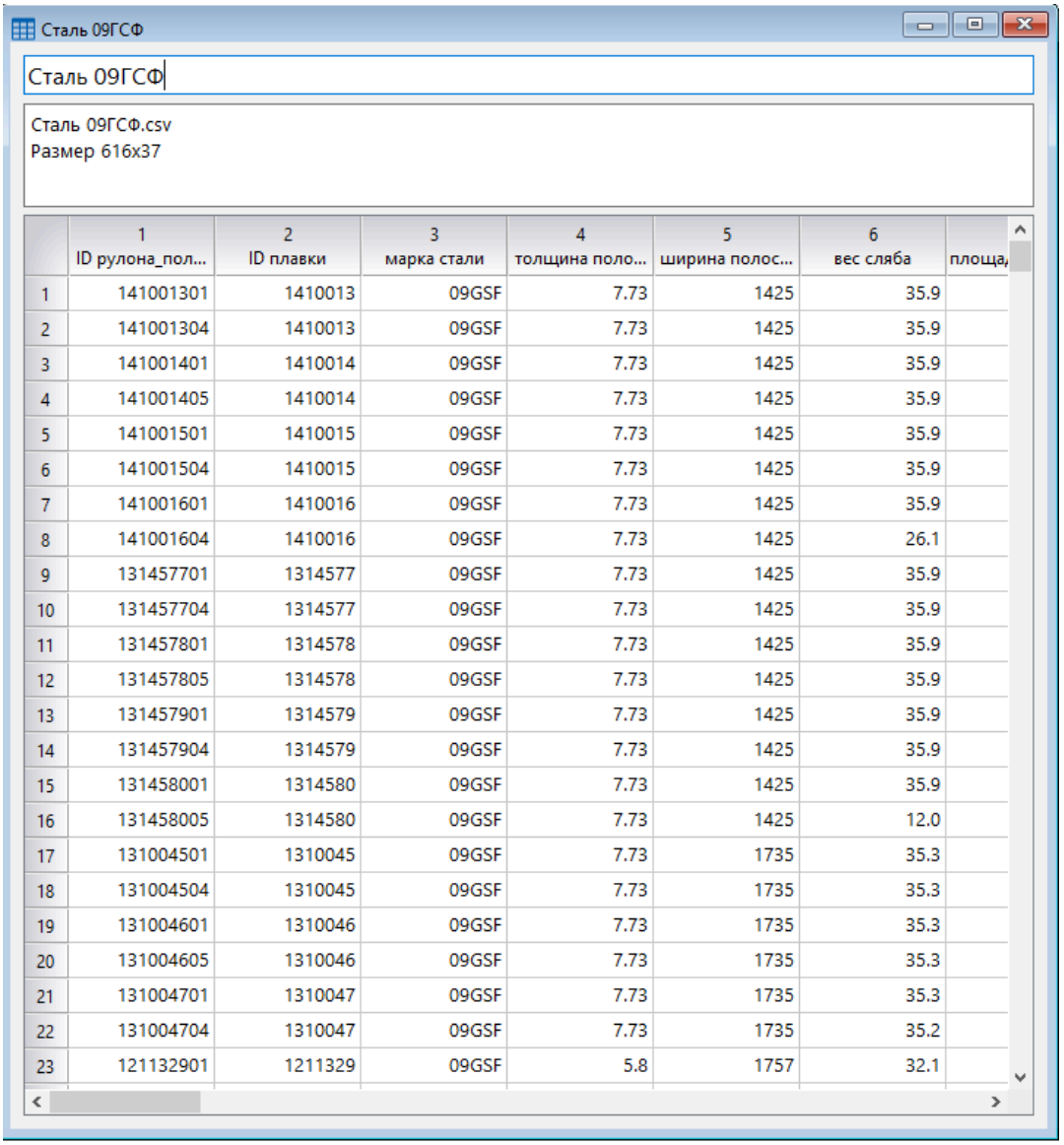
Пример построения предсказательной модели

Постановка задачи. Сталь 09ГСФ - Конструкционная легированная сталь повышенной коррозионной стойкости и хладостойкости.

Необходимо построить в ПО СтатСофт деревья регрессии и нейронные сети для предсказания механических свойств стали: предела текучести, предела прочности и относительного удлинения в зависимости от химического состава и параметров прокатки на данных стали марки 09ГСФ.

Обзор данных. Данные представлены в файле «Сталь 09ГСФ.csv». Каждая строка соответствует одной прокатке стали, каждый столбец – какой-либо фиксируемой характеристике. Всего в таблице приводятся 616 строк и 37 столбцов. Переменные передают информацию о химическом составе плавки, параметрах плавки и прокатки, а также об измеренных механических свойствах.

Открыть файл в приложении можно с помощью меню «Файл».

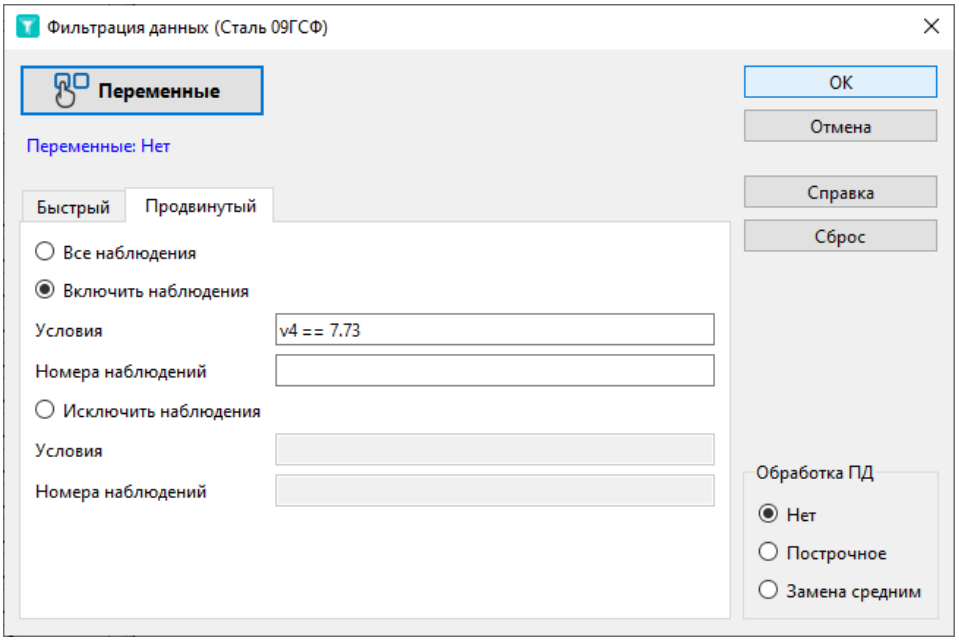


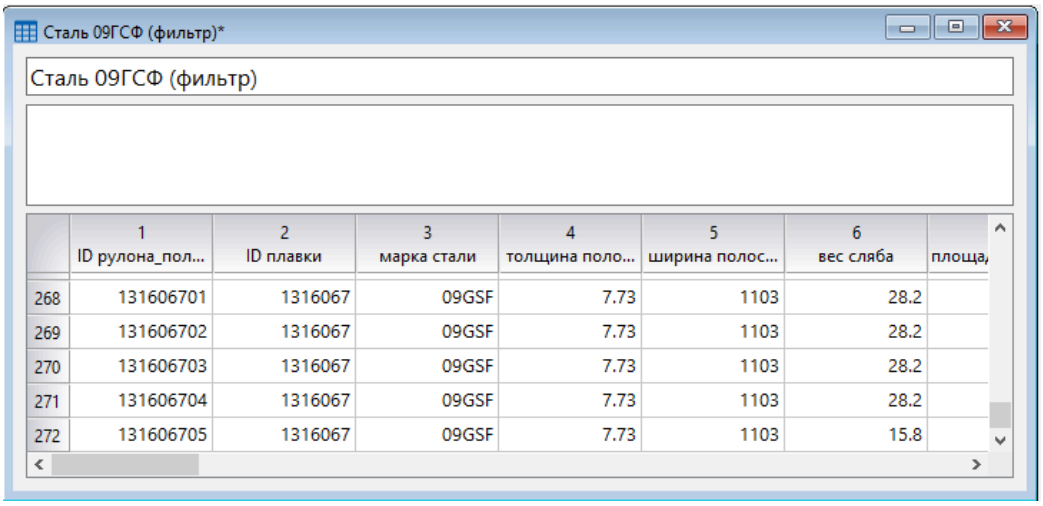
	1 ID рулона_пол...	2 ID плавки	3 марка стали	4 толщина поло...	5 ширина полос...	6 вес сляба	площа
1	141001301	1410013	09GSF	7.73	1425	35.9	
2	141001304	1410013	09GSF	7.73	1425	35.9	
3	141001401	1410014	09GSF	7.73	1425	35.9	
4	141001405	1410014	09GSF	7.73	1425	35.9	
5	141001501	1410015	09GSF	7.73	1425	35.9	
6	141001504	1410015	09GSF	7.73	1425	35.9	
7	141001601	1410016	09GSF	7.73	1425	35.9	
8	141001604	1410016	09GSF	7.73	1425	26.1	
9	131457701	1314577	09GSF	7.73	1425	35.9	
10	131457704	1314577	09GSF	7.73	1425	35.9	
11	131457801	1314578	09GSF	7.73	1425	35.9	
12	131457805	1314578	09GSF	7.73	1425	35.9	
13	131457901	1314579	09GSF	7.73	1425	35.9	
14	131457904	1314579	09GSF	7.73	1425	35.9	
15	131458001	1314580	09GSF	7.73	1425	35.9	
16	131458005	1314580	09GSF	7.73	1425	12.0	
17	131004501	1310045	09GSF	7.73	1735	35.3	
18	131004504	1310045	09GSF	7.73	1735	35.3	
19	131004601	1310046	09GSF	7.73	1735	35.3	
20	131004605	1310046	09GSF	7.73	1735	35.3	
21	131004701	1310047	09GSF	7.73	1735	35.3	
22	131004704	1310047	09GSF	7.73	1735	35.2	
23	121132901	1211329	09GSF	5.8	1757	32.1	

Предварительная обработка данных - фильтрация наблюдений. В таблице представлены данные для разных значений толщина полосы. Узнать распределение количества наблюдений по значениям данной переменной можно с помощью модуля «Таблицы частот».

	1 Значение	2 Частота
1	7.73	272.000
2	5.8	17.000
3	9.7	3.000
4	6.83	10.000
5	7.68	87.000
6	7.75	119.000
7	5.83	24.000
8	6.0	35.000
9	5.0	32.000
10	9.73	6.000
11	6.8	8.000
12	9.65	3.000

Как правило, построение отдельных моделей для каждой толщины приводит к повышению качества по сравнению с одной общей моделью. Построим модели для наиболее часто встречающейся толщины 7.73. Для этого отбора наблюдений воспользуйтесь модулем «Фильтрация данных». На вкладке «Продвинутый» задайте условие включения «v4 == 7.73» и нажмите «ОК». В рабочей области сформируется таблица, состоящая из 272 наблюдений.

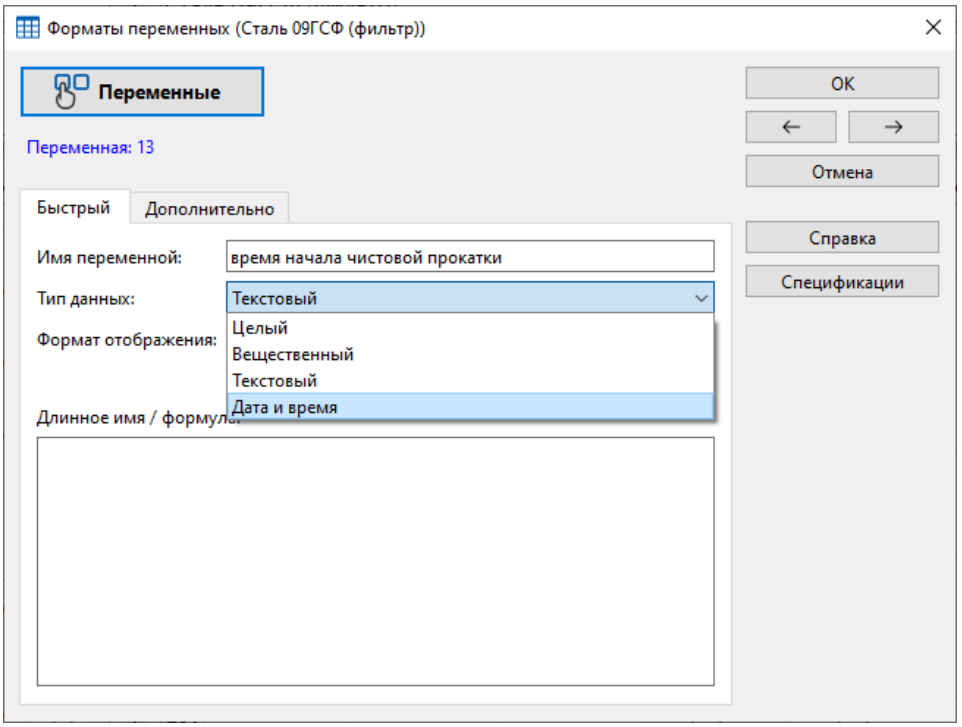




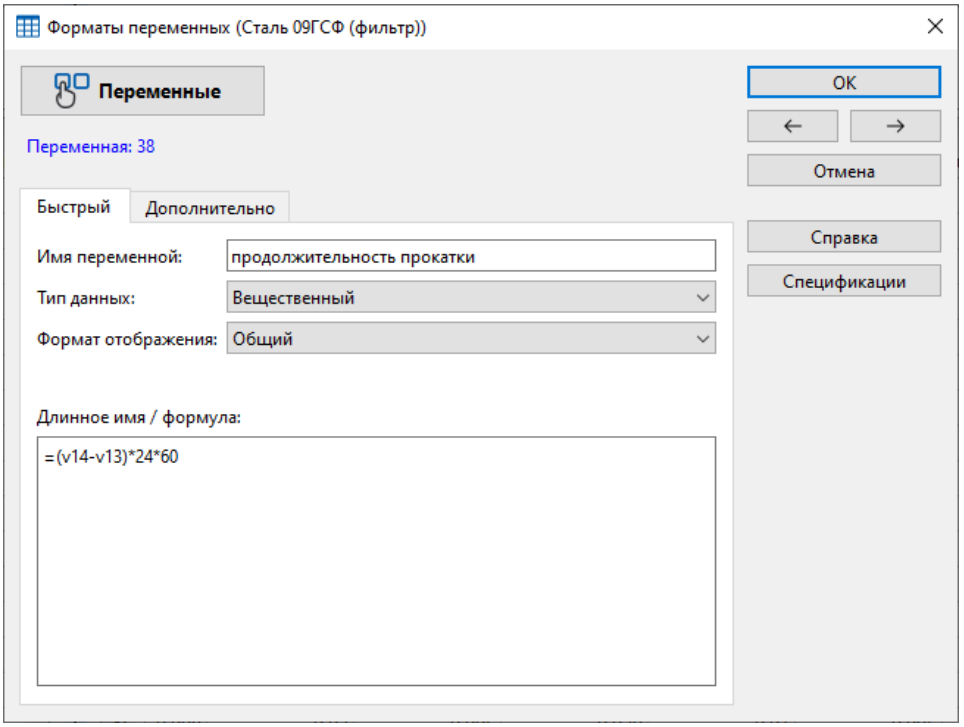
	1 ID рулона_пол...	2 ID плавки	3 марка стали	4 толщина поло...	5 ширина полос...	6 вес сляба	площа
268	131606701	1316067	09GSF	7.73	1103	28.2	
269	131606702	1316067	09GSF	7.73	1103	28.2	
270	131606703	1316067	09GSF	7.73	1103	28.2	
271	131606704	1316067	09GSF	7.73	1103	28.2	
272	131606705	1316067	09GSF	7.73	1103	15.8	

Создание переменной – продолжительность прокатки. В таблице присутствуют данные о времени начала и конца чистовой прокатки. Сами по себе эти значения ничего не говорят, но их разность представляет собой время прокатки.

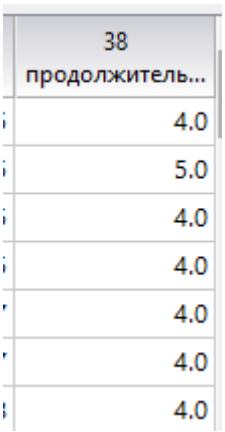
Поскольку заранее неизвестно, какие переменные окажут наибольшее влияние на целевую, добавьте новый столбец в таблицу, где вычислите время прокатки. Для этого двойным щелчком по имени переменной «время начала чистовой прокатки» откройте окно спецификаций.



Переведите переменную сначала в формат «Дата и время» и нажмите «ОК». После этого переведите переменную в «Вещественный» формат и снова нажмите «ОК». Аналогичным образом поступите с переменной «время конца чистовой прокатки». Данные действия закодируют значения столбцов некоторыми вещественными значениями. После этого создайте новую переменную, указав для нее формулу.

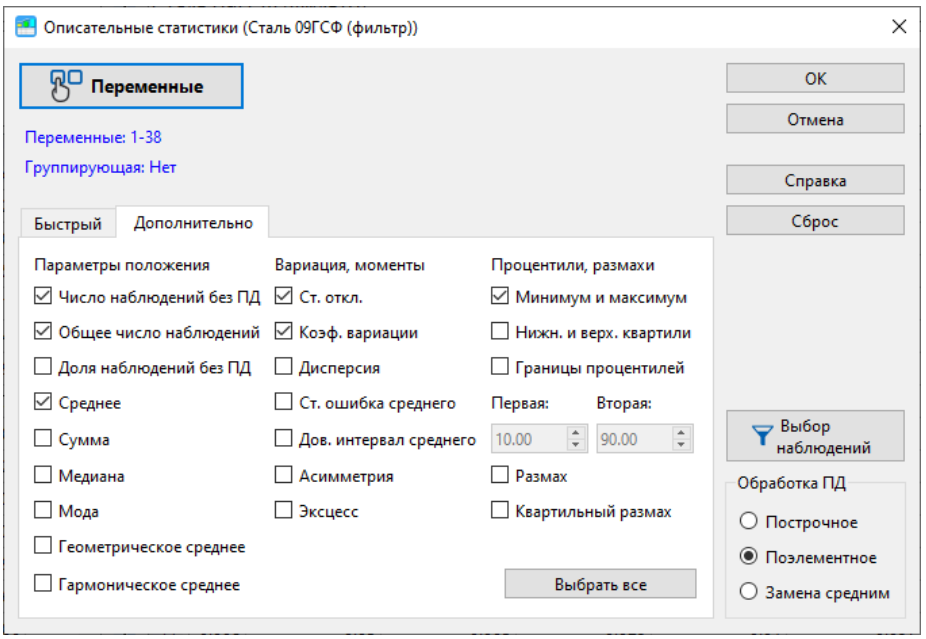


Разность закодированных в столбцах 13 и 14 чисел показывает, какую долю от суток занимает данный диапазон времени. Чтобы перевести эту величину в минуты, проводится умножение на 24 и на 60. Вы получите новый столбец в таблице.



	38 продолжитель...
	4.0
	5.0
	4.0
	4.0
	4.0
	4.0
	4.0

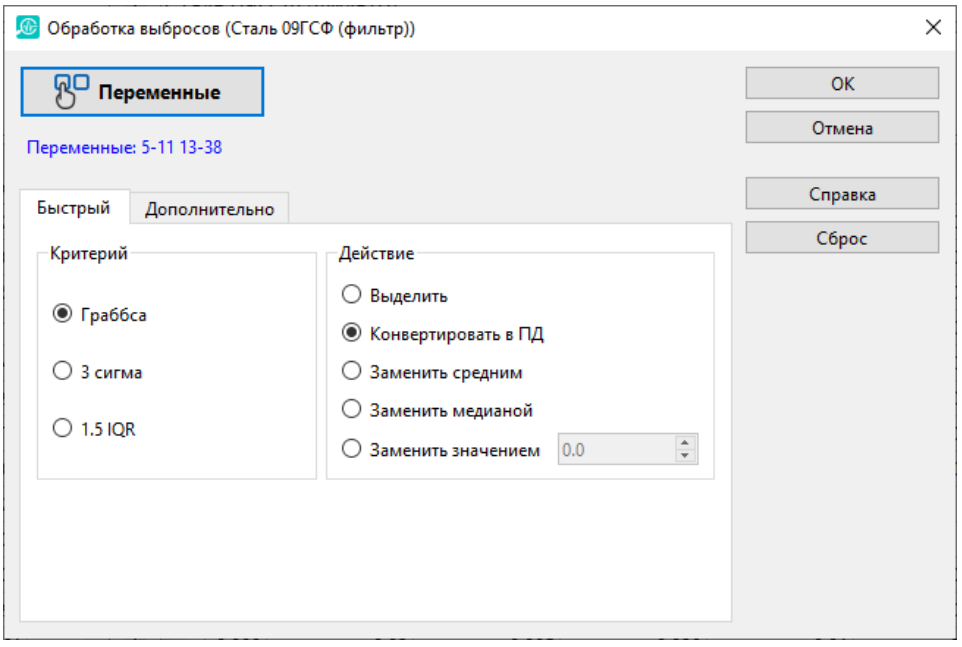
Обработка выбросов. При помощи анализа «Описательные статистики» исследуйте наличие пропущенных данных в таблице.

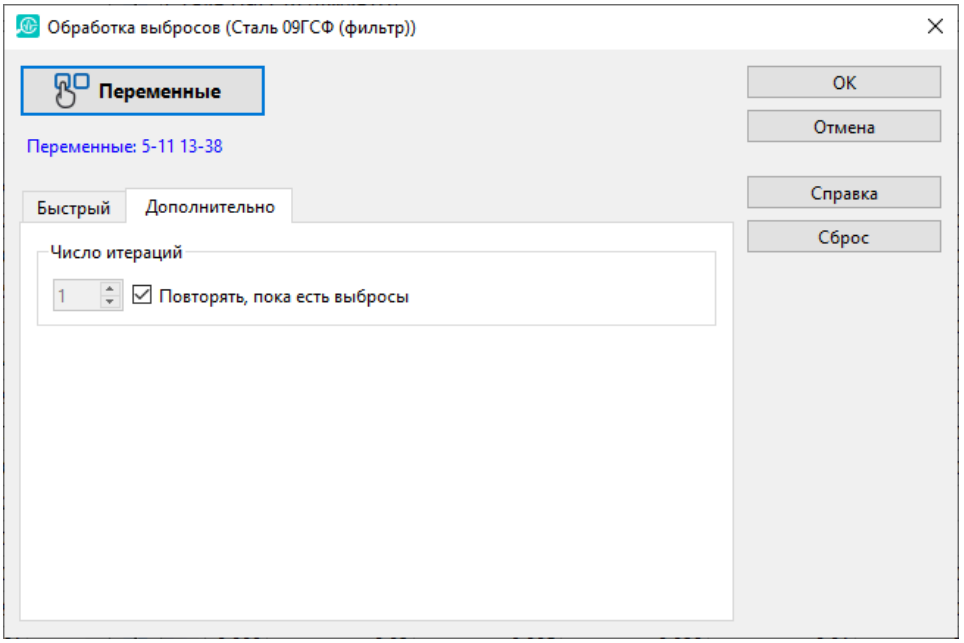


Вы увидите, что пропуски присутствуют только в нескольких наблюдениях целевых переменных.

	1 Переменная	2 Общее число наблюдений	3 Число наблюдений без ПД
14	SI	272	272
15	TI	272	272
16	V	272	272
17	вес сляба	272	272
18	время конца чистовой прокатки	272	272
19	время начала чистовой прокатки	272	272
20	время транспортировки на рольганге	272	272
21	марка стали	272	272
22	относительное удлинение %	272	268
23	площадь поверхности	272	272
24	площадь поверхности внешняя	272	272
25	предел прочности МПа	272	268
26	предел текучести МПа	272	268
27	продолжительность прокатки	272	272
28	скорость прокатки м\сек	272	272
29	стратегия охлаждения на отводящем ...	272	272
30	температура 1 черновой прокатки	272	272
31	температура конца прокатки	272	272
32	температура окончания черновой ...	272	272
33	температура смотки	272	272
34	температура смотки на центральной ...	272	272
35	толщина полосы	272	272

Далее с помощью критерия Граббса замените на пропуски все значения-выбросы, поставив на вкладке «Дополнительно» отметку «Повторять, пока есть выбросы».

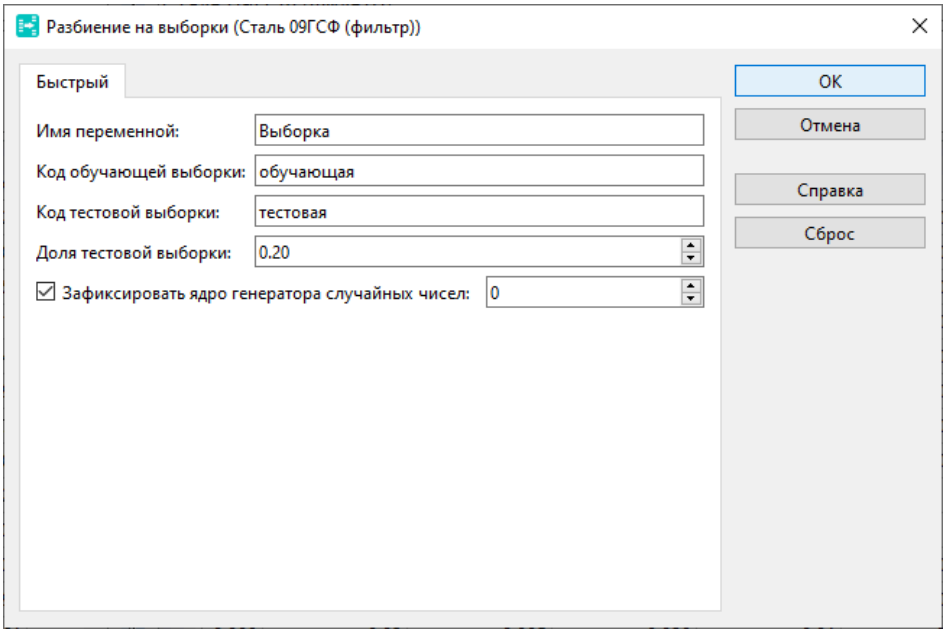




Постройте заново таблицу описательных статистик, чтобы выяснить, сколько выбросов было удалено.

	1 Переменная	2 Общее число наблюдений	3 Число наблюдений без ПД
14	SI	272	270
15	TI	272	272
16	V	272	270
17	вес сляба	272	272
18	время конца чистовой прокатки	272	272
19	время начала чистовой прокатки	272	272
20	время транспортировки на рольганге	272	272
21	марка стали	272	272
22	относительное удлинение %	272	268
23	площадь поверхности	272	254
24	площадь поверхности внешняя	272	251
25	предел прочности МПа	272	268
26	предел текучести МПа	272	268
27	продолжительность прокатки	272	272
28	скорость прокатки м\сек	272	272
29	стратегия охлаждения на отводящем ...	272	272
30	температура 1 черновой прокатки	272	271
31	температура конца прокатки	272	272
32	температура окончания черновой ...	272	268
33	температура смотки	272	272
34	температура смотки на центральной ...	272	255
35	толщина полосы	272	272

Видно, что из разных переменных были удалены выбросы, однако их общее количество не слишком велико. Теперь откройте модуль «Разбиение на выборки» и сформируйте обучающую и тестовую выборки в соотношении 8:2.

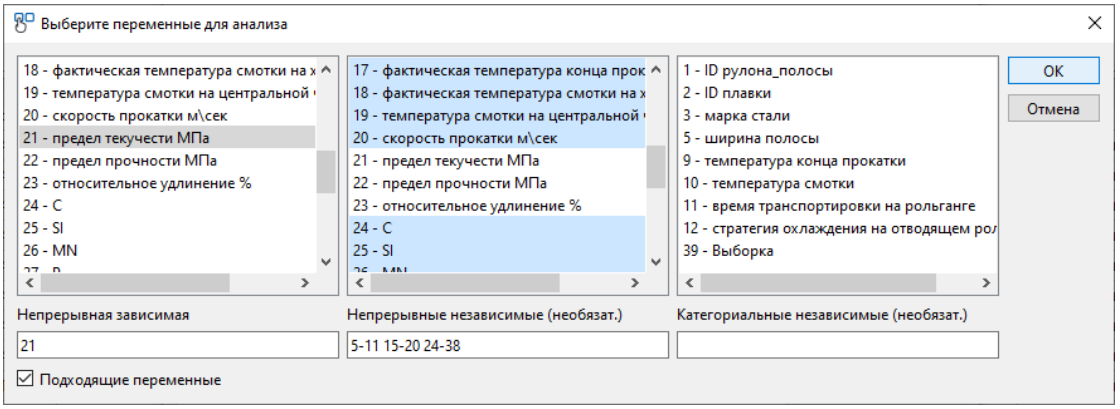


К данным добавится соответствующий столбец.

		34 N	35 NB	36 Tl	37 V	38 продолжитель...	39 Выборка
1	0.031	0.005	0.028	0.009	0.095	4.0	обучающая
2	0.031	0.005	0.028	0.009	0.095	5.0	обучающая
3	0.026	0.006	0.029	0.01	0.095	4.0	обучающая
4	0.026	0.006	0.029	0.01	0.095	4.0	обучающая
5	0.026	0.006	0.029	0.009	0.097	4.0	обучающая
6	0.026	0.006	0.029	0.009	0.097	4.0	обучающая
7	0.029	0.005	0.03	0.011	0.098	4.0	обучающая
8	0.029	0.005	0.03	0.011	0.098	4.0	тестовая
9	0.033	0.006	0.028	0.01	0.1	4.0	тестовая
10	0.033	0.006	0.028	0.01	0.1	5.0	обучающая

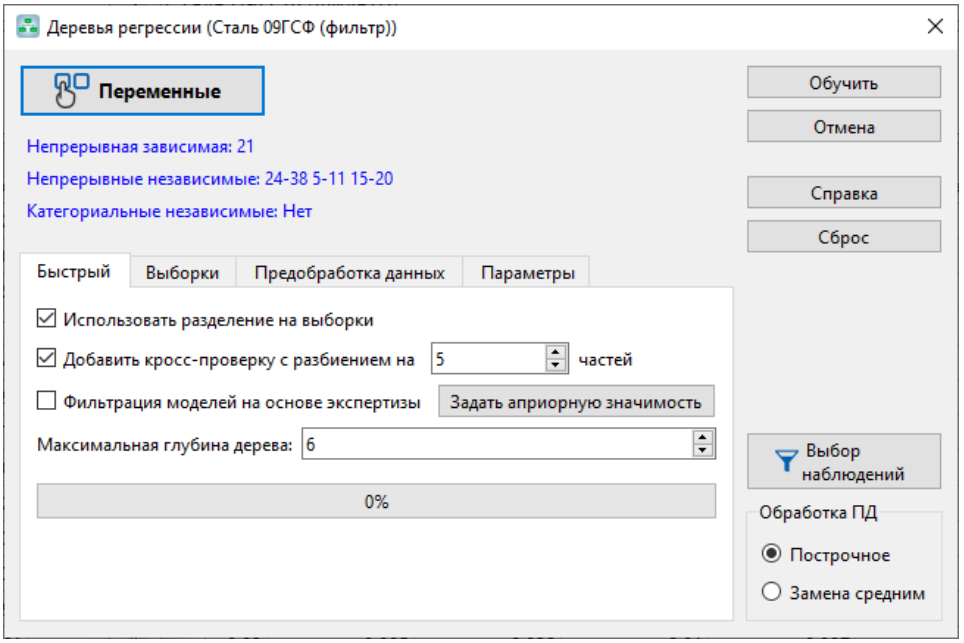
Обработанную таблицу рекомендуется сохранить в виде отдельного sts-файла.

Построение деревьев регрессии. Перейдите на вкладку «ИИ Регрессия» и откройте модуль «Деревья регрессии». Задайте переменные для анализа:

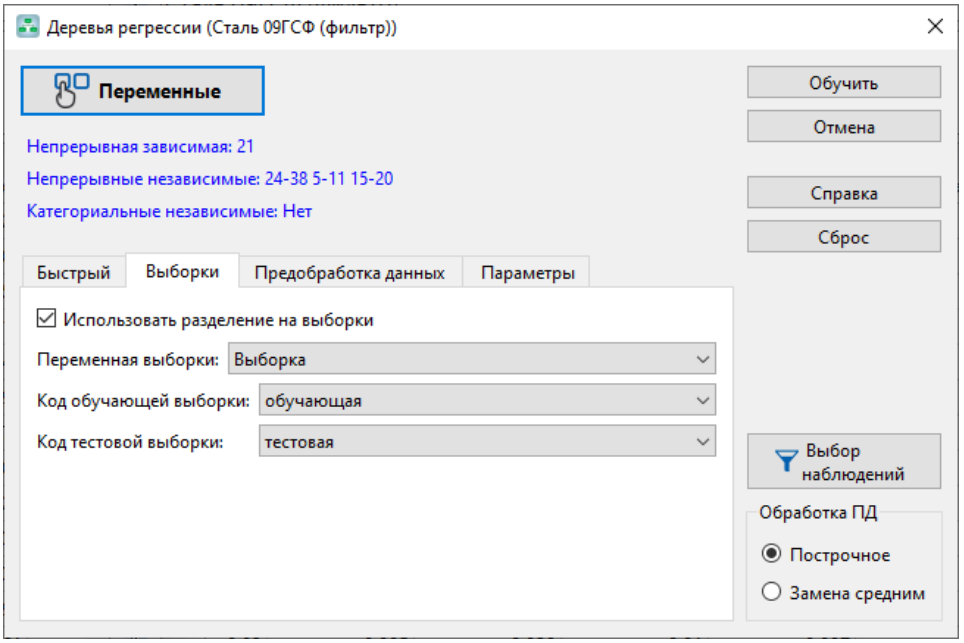


Время начала и время конца прокатки не включаются, потому что вместо них используется созданная переменная продолжительности прокатки.

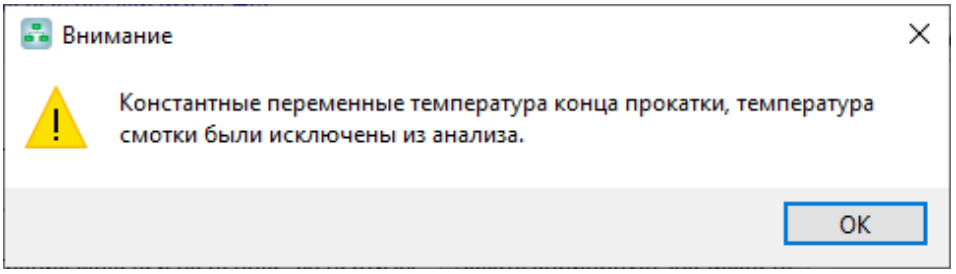
Включите опцию, позволяющую использовать разбиение на выборки. Также включите кросс-проверку с разбиением на 5 частей для подбора оптимальной комбинации гиперпараметров.



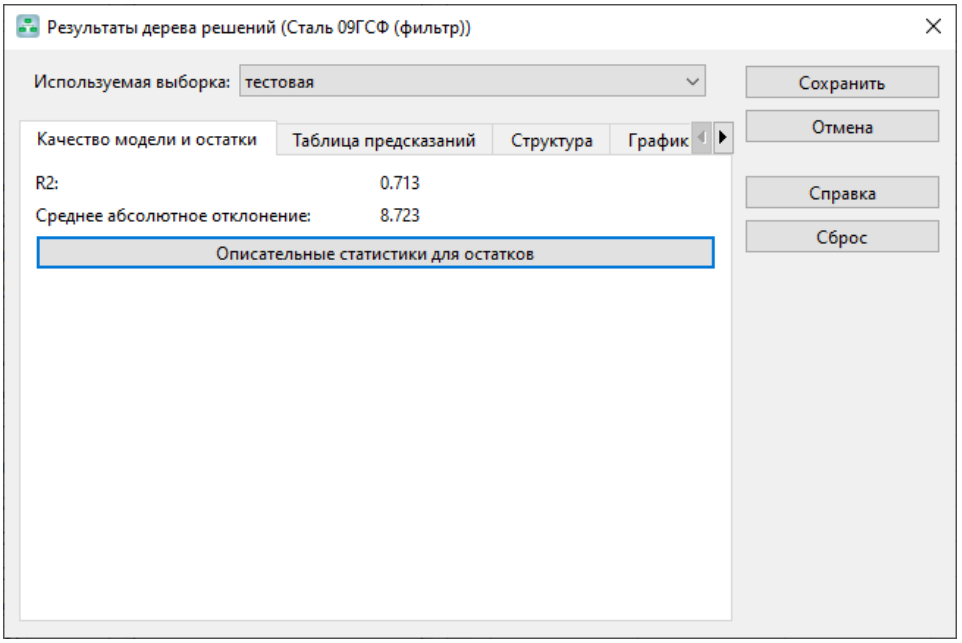
Задайте коды обучающей и тестовой выборок.



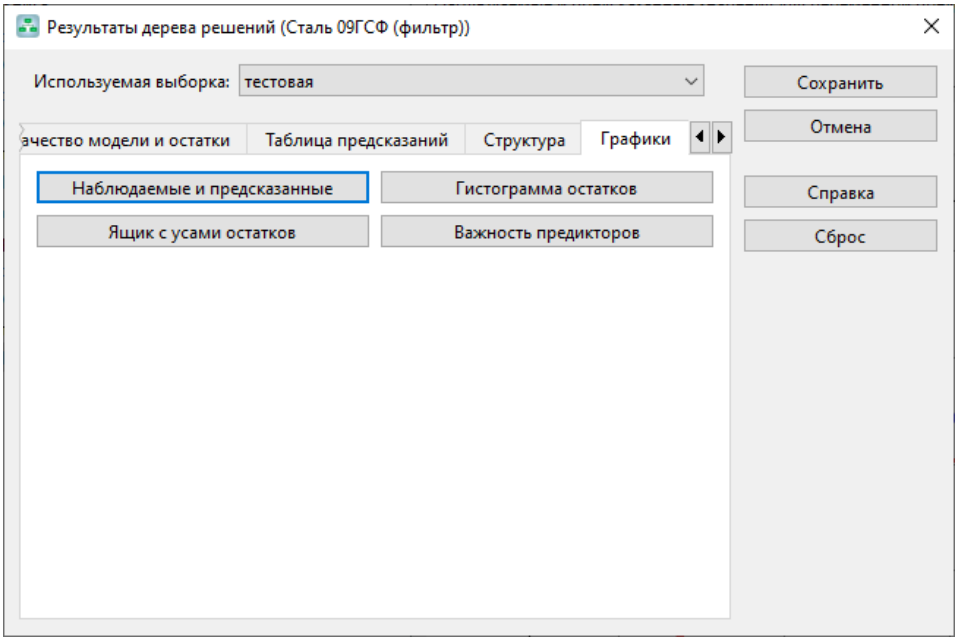
Отметьте построчное удаление ПД, чтобы исключить из анализа строки с ПД. Нажмите «Обучить». Появится предупреждение о наличии константных переменных в данных, согласитесь с ним.

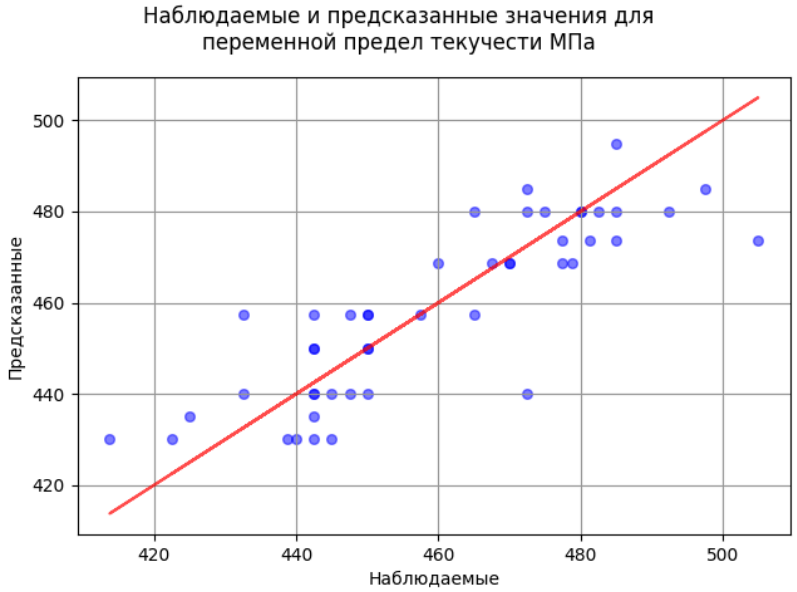


На вкладке «Быстрый» будет отображаться прогресс построения моделей. Когда процесс завершится, откроется окно результатов, где вы сразу увидите числовые характеристики качества модели: R2 и среднее абсолютное отклонение.

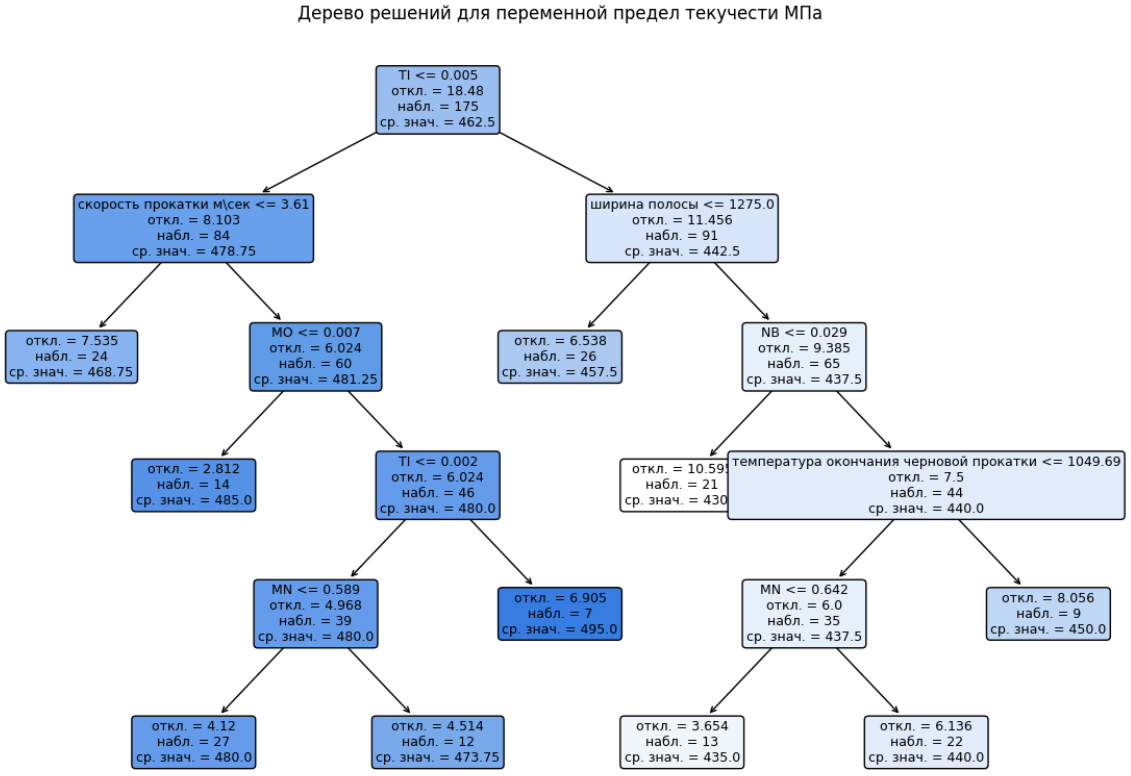


Перейдите на вкладку графиков и постройте диаграмму наблюдаемых и предсказанных значений.

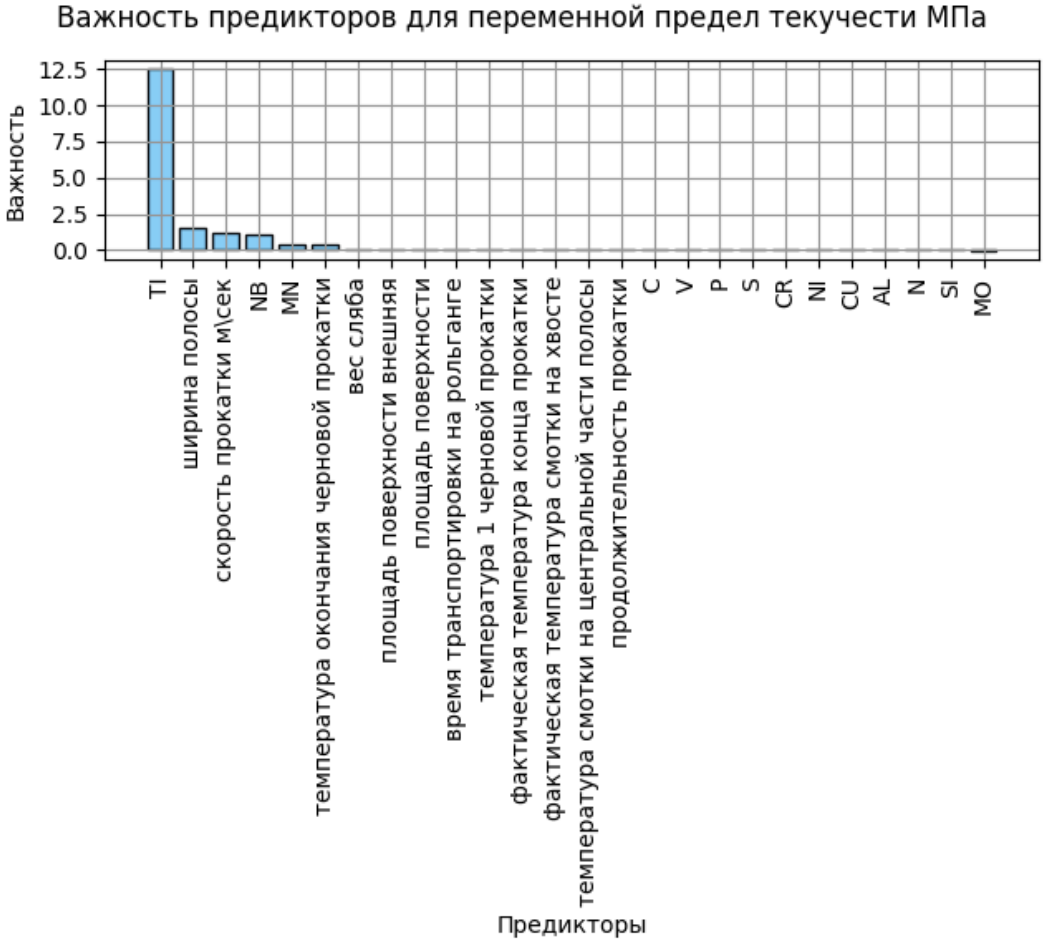




На вкладке «Структура» можно отобразить структуру построенной модели.



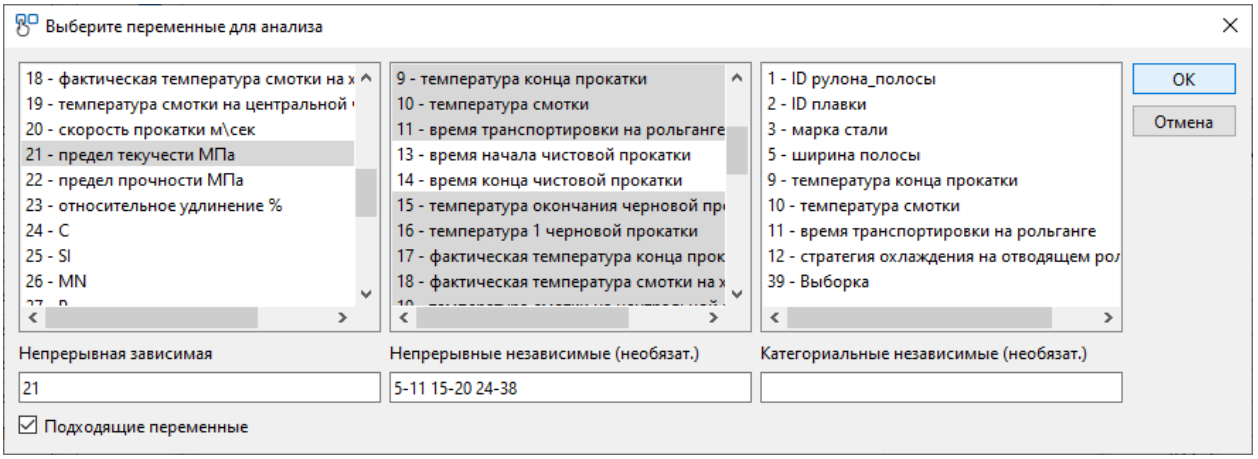
Важность предикторов:



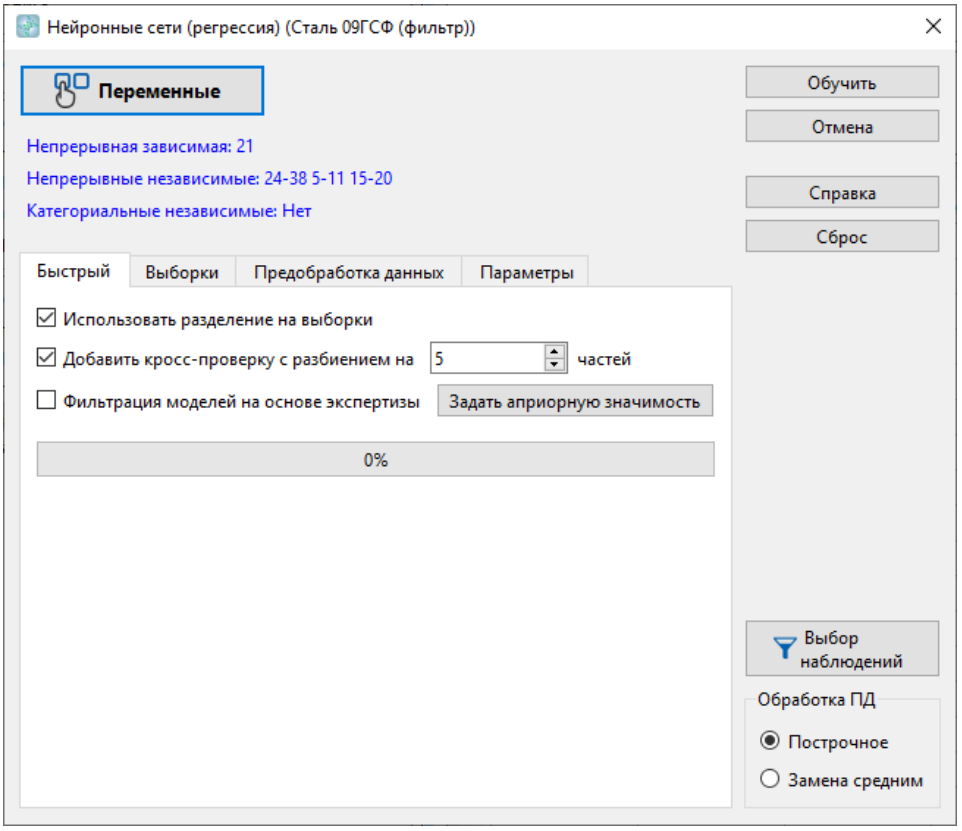
По графику важности можно понять, что наибольшее влияния на значение предела текучести с точки зрения модели оказывает содержание титана. Также определенный вклад вносят ширина полосы, скорость прокатки, содержание ниобия и марганца, а также температура окончания черновой прокатки.

Исходя из графика дерева можно сделать вывод, что наибольшее значение предела текучести достигается при содержании титана от 0,002 до 0,005.

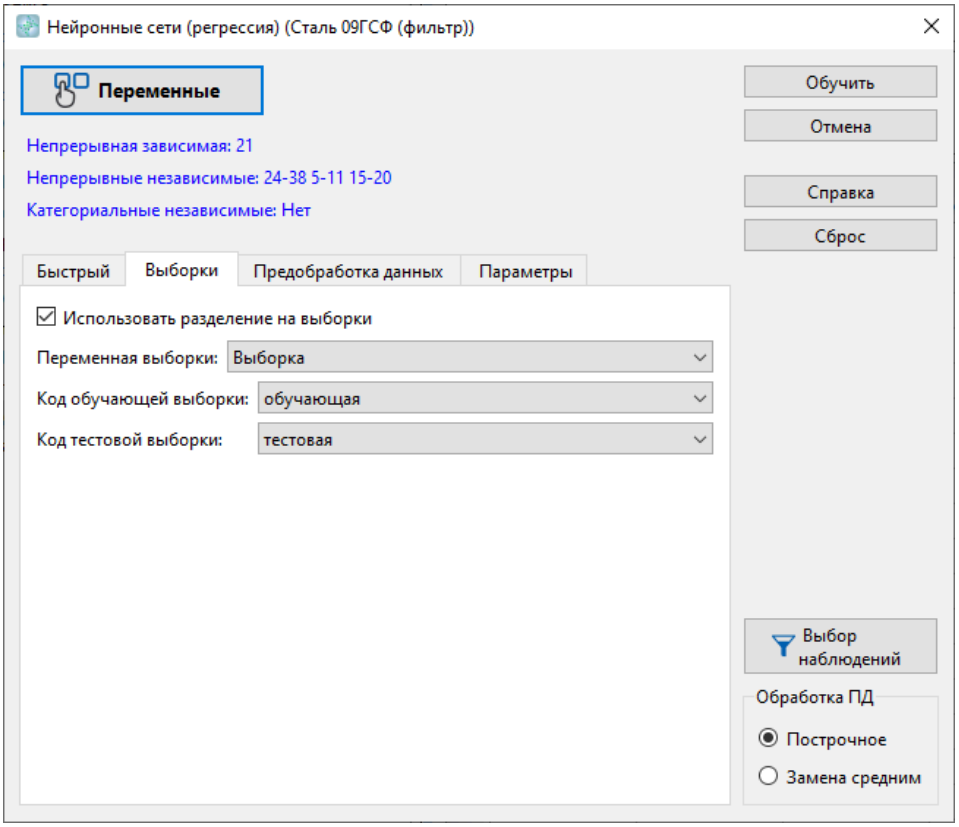
Построение нейронных сетей. Откройте модуль «Нейронные сети» на вкладке «ИИ Регрессия». Задайте переменные для анализа:



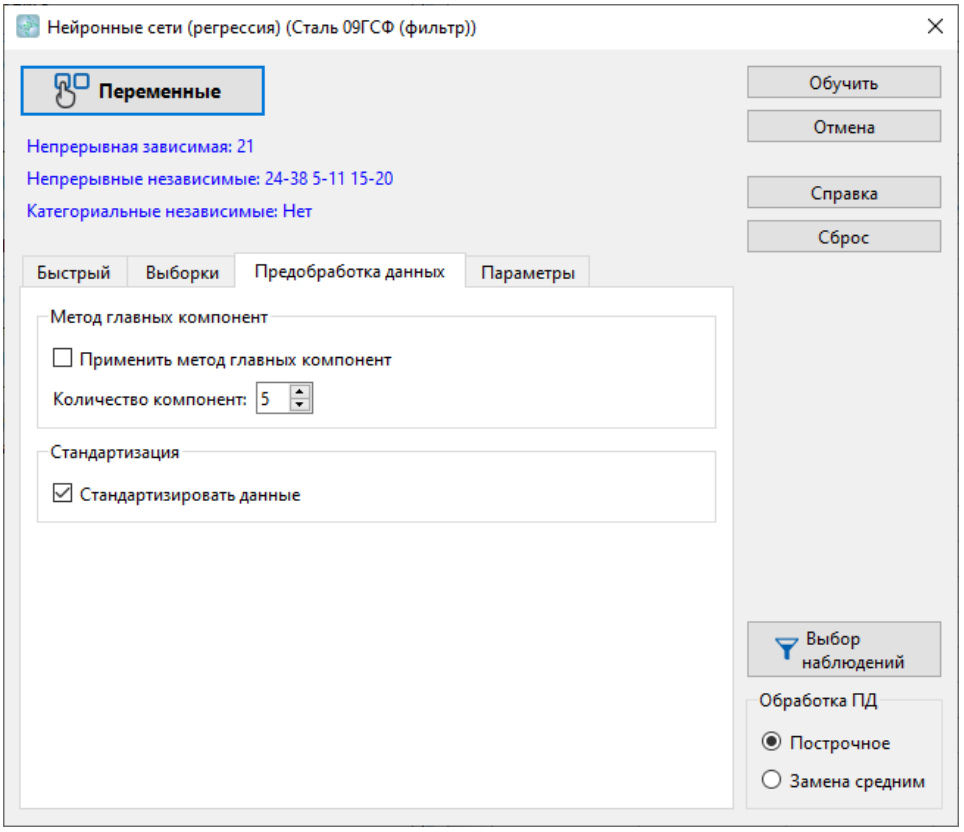
Включите опцию, позволяющую использовать разбиение на выборки. Также включите кросс-проверку с разбиением на 5 частей для подбора оптимальной комбинации гиперпараметров.



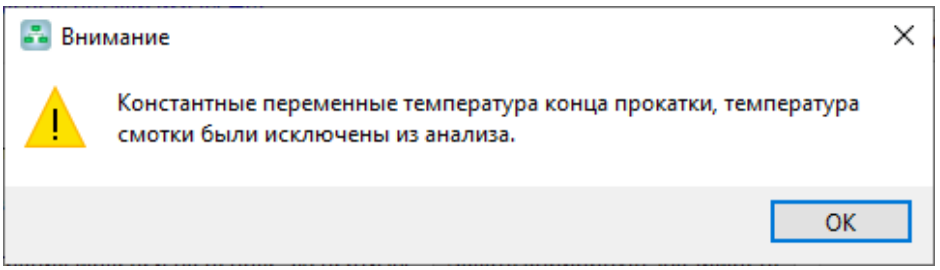
Задайте коды обучающей и тестовой выборок.



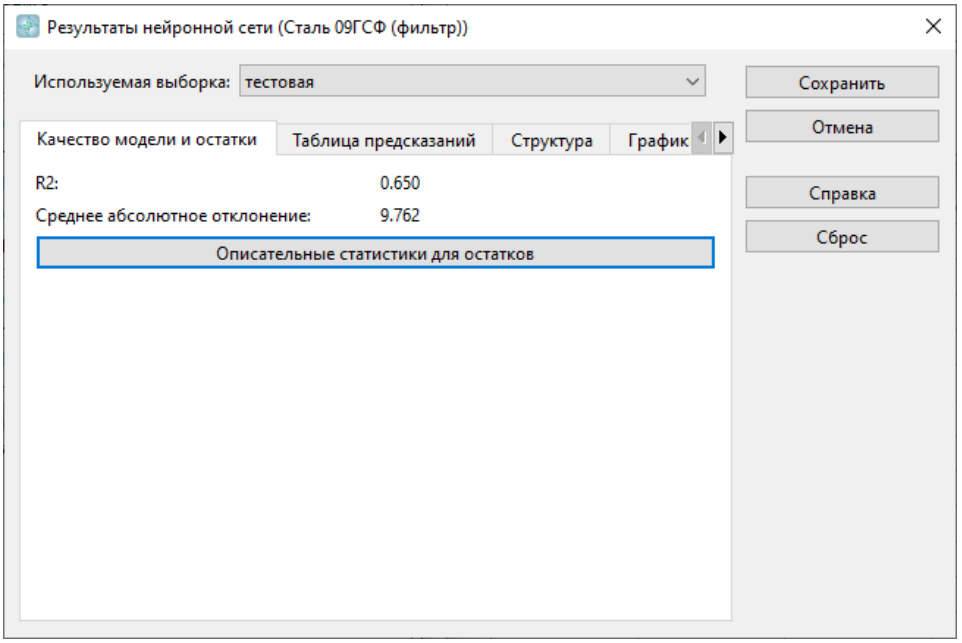
На вкладке «Предобработка данных» включите стандартизацию.



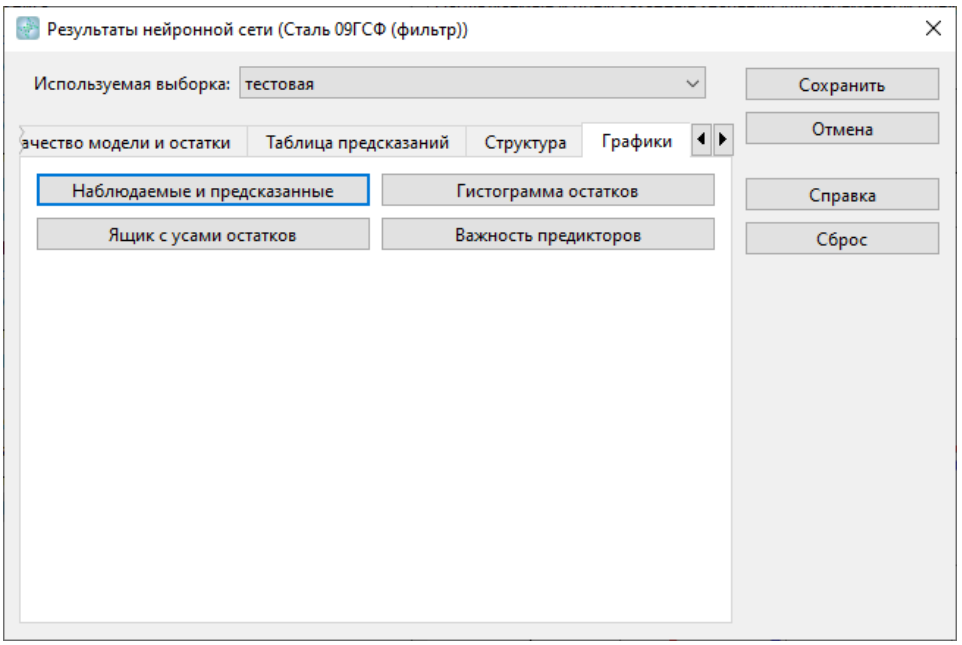
Отметьте построчное удаление ПД, чтобы исключить из анализа строки с ПД. Нажмите «Обучить». Появится предупреждение о наличии константных переменных в данных, согласитесь с ним.

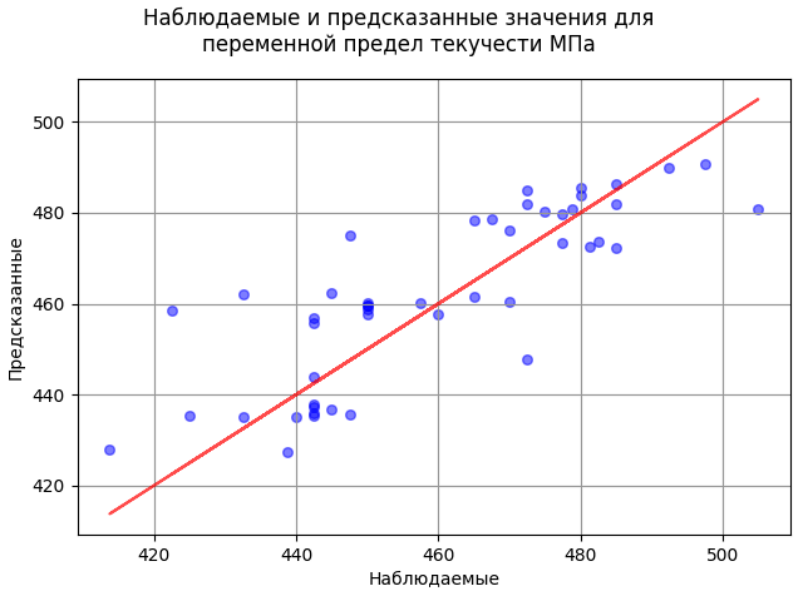


На вкладке «Быстрый» будет отображаться прогресс построения моделей. Когда процесс завершится, откроется окно результатов, где вы сразу увидите числовые характеристики качества модели: R2 и среднее абсолютное отклонение.



Перейдите на вкладку графиков и постройте диаграмму наблюдаемых и предсказанных значений.





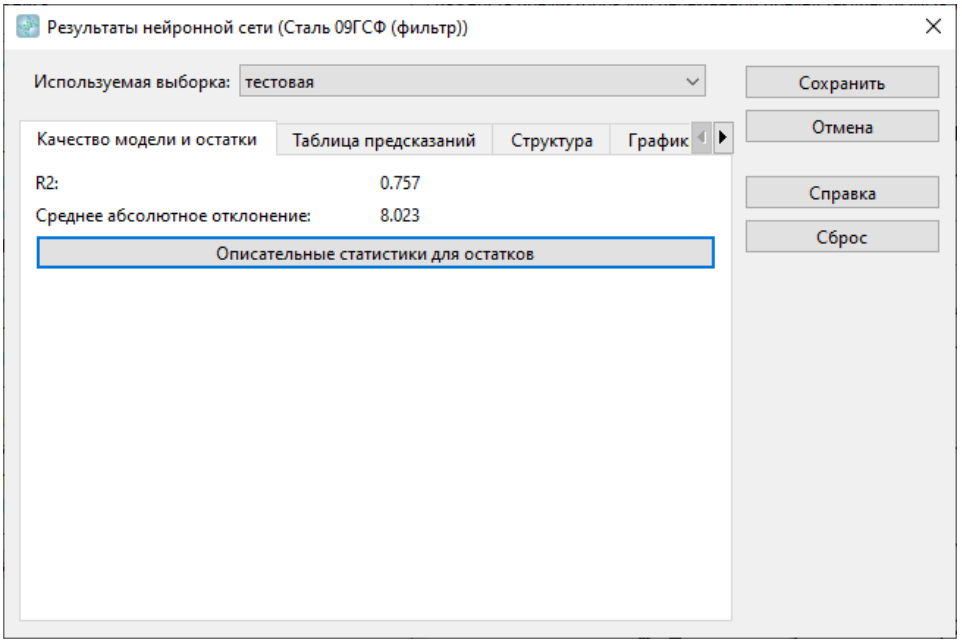
На вкладке «Структура» можно отобразить структуру построенной модели.

	1 Характеристика	2 Значение
1	Число скрытых слоёв	3
2	Число нейронов на скрытых слоях	30
3	Функция активации	Логистическая
4	Оптимизатор	lbfgs
5	L2 регуляризация	1e-05
6	Размер батча	auto
7	Шаг обучения	Адаптивный
8	Начальный шаг	0.001
9	Степень уменьшения шага	0.5
10	Максимальное число эпох	50
11	Толерантность	0.0001
12	Начальная установка случайных чисел	22

Важность предикторов:

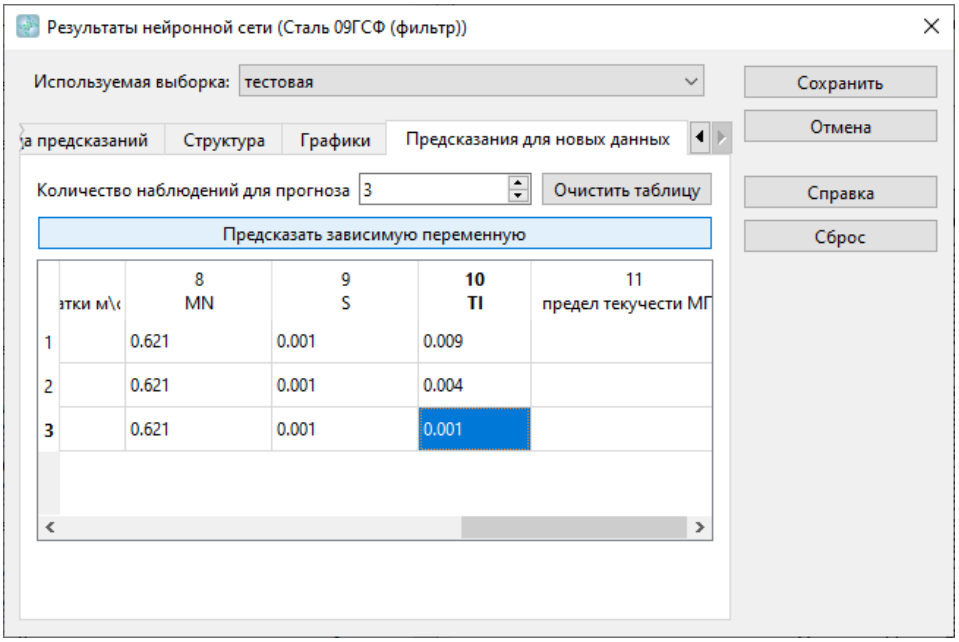
По графику важности можно понять, что наибольшее влияния на значение предела текучести с точки зрения модели оказывает содержание титана. Также определенный вклад вносят скорость прокатки, ширина полосы, площадь поверхности, время транспортировки, содержание марганца, вес сляба, а также температура окончания черновой прокатки. При этом большое некоторое количество переменных с отрицательной важностью, наличие которых, вероятно, запутывает модель.

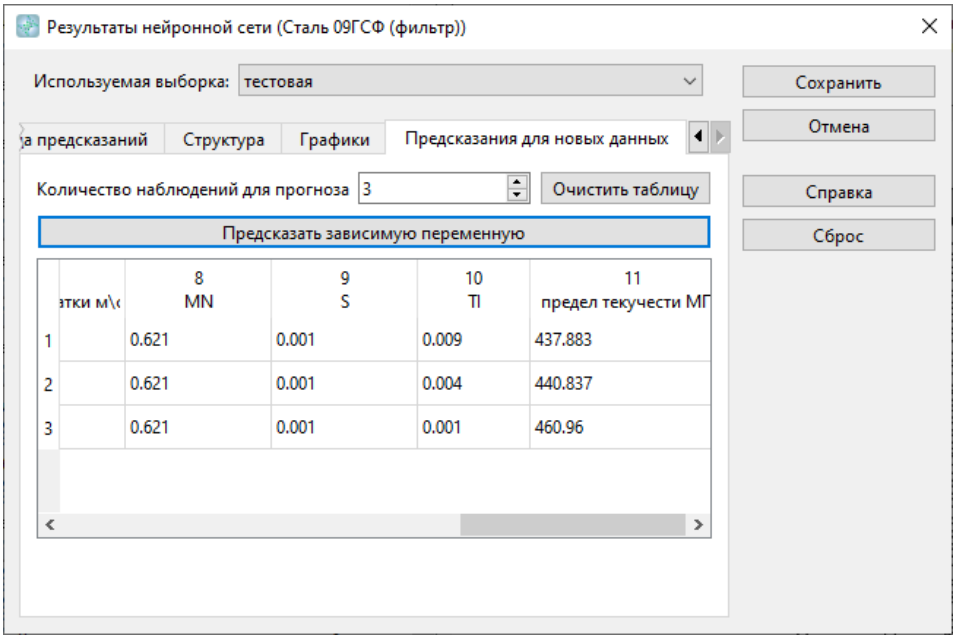
Скопируйте из описания графика номера переменных с положительной важностью и постройте модель только с ними.



В данном случае качество модели на тестовой выборке возрастает, потому что модель перестает учитывать закономерности, не относящиеся к тестовой выборке.

Чтобы лучше изучить зависимость целевой переменной от содержания титана (или любых других переменных) воспользуйтесь таблицей на вкладке «Предсказания для новых данных» окна результатов модели. Задайте там наблюдения с разным содержанием титана и нажмите «Предсказать зависимую переменную».





По трем рассмотренным точкам можно сказать, что зависимость обратная (предел текучести снижается при высоком содержании титана), однако, зависимость,

восстановленная моделью может быть и более сложной. Для более детального изучения лучше использовать больше разных точек (наблюдений).

Качество моделей. Приведем в сводных таблицах качество деревьев регрессии и нейронных сетей для всех двухцелевых переменных.

MAE – среднее абсолютное отклонение.

Если было обнаружено большое количество переменных с отрицательной важностью на тестовой выборке, то нейронные сети обучались повторно на переменных с положительной важностью.

Обучающая выборка:

	Дерево регрессии	**Нейронная сеть**
Предел текучести	R2 = 0.819 MAE = 6.180	R2 = 0.715 MAE = 8.454
Предел прочности	R2 = 0.840 MAE = 5.981	R2 = 0.904 MAE = 4.356

Тестовая выборка:

	Дерево регрессии	**Нейронная сеть**
Предел текучести	R2 = 0.713 MAE = 8.723	R2 = 0.757 MAE = 8.023
Предел прочности	R2 = 0.567 MAE = 9.109	R2 = 0.673 MAE = 8.282

Приведем аналогичные таблицы для группы с толщиной полосы 7,75.

Обучающая выборка:

	Дерево регрессии	**Нейронная сеть**
Предел текучести	R2 = 0.644 MAE = 10.974	R2 = 0.758 MAE = 9.821
Предел прочности	R2 = 0.621 MAE = 8.186	R2 = 0.940 MAE = 2.608

Тестовая выборка:

	Дерево регрессии	**Нейронная сеть**
Предел текучести	R2 = 0.714 MAE = 9.826	R2 = 0.720 MAE = 10.007
Предел прочности	R2 = 0.392 MAE = 9.213	R2 = 0.380 MAE = 10.167

Для остальных групп в таблице данных представлено слишком мало наблюдений.

Приведем для сравнения таблицы для моделей, построенных на всех данных без разделения на группы. (Нейронные сети переобучались на значимых переменных.)

Обучающая выборка:

	Дерево регрессии	Нейронная сеть
Предел текучести	R2 = 0.702 MAE = 9.987	R2 = 0.799 MAE = 7.811
Предел прочности	R2 = 0.691 MAE = 8.279	R2 = 0.787 MAE = 6.742

Тестовая выборка:

	Дерево регрессии	Нейронная сеть
Предел текучести	R2 = 0.601 MAE = 10.598	R2 = 0.630 MAE = 10.732
Предел прочности	R2 = 0.498 MAE = 9.153	R2 = 0.540 MAE = 8.413

Вывод. Для группы с наибольшим числом наблюдений были получены модели среднего качества для предела текучести и предела прочности. В группе с меньшим числом наблюдений работоспособная модель была получена только для предела текучести. В целом, это говорит о том, что информации, представленной в таблице данных, недостаточно для качественного восстановления зависимости целевых переменных (механические свойства) от предикторов (химический состав и параметры прокатки).

Тем не менее, проведение группировки по толщине полосы позволяет несколько повысить точность моделей, если при этом в группе остается достаточное число наблюдений.

Практически во всех моделях ключевой переменной, влияющей на итоговое механическое свойства, является содержание титана, причем прочность выше при небольших значениях этой переменной.

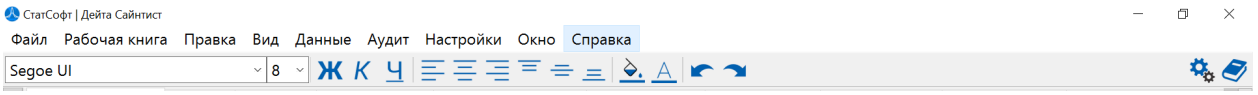
Электронная справочная система

Электронное руководство ПО СтатСофт предназначено для знакомства пользователя с функциями программы. Вы можете найти в нем общие сведения об интерфейсе программы, описания всех функций анализа, опции для работы с графиками и т.д.

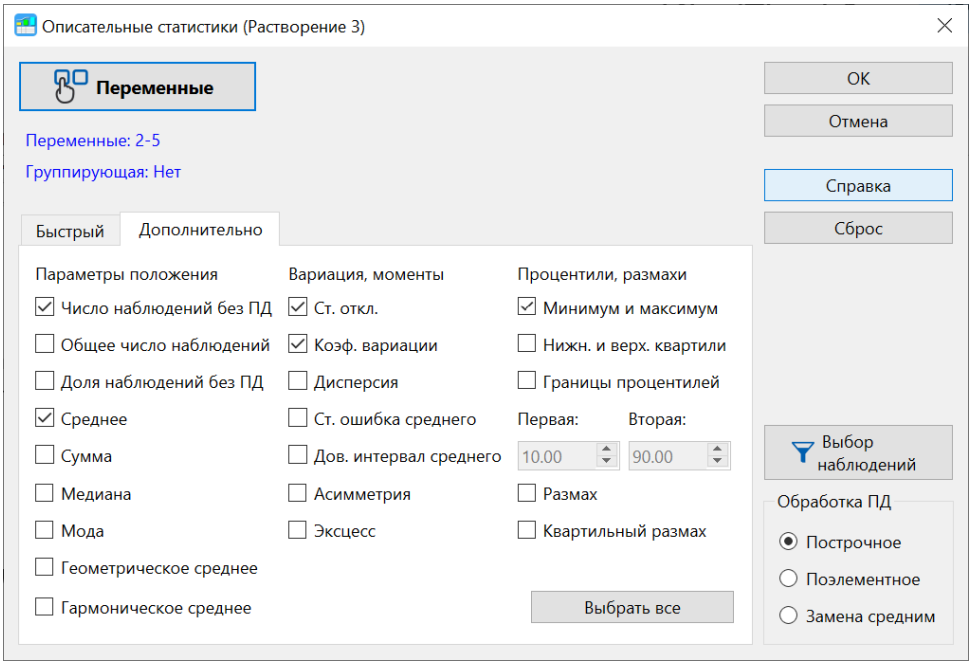
Помимо этого, в электронном руководстве представлены примеры проведения анализа с использованием различных методов и интерпретациями полученных результатов.

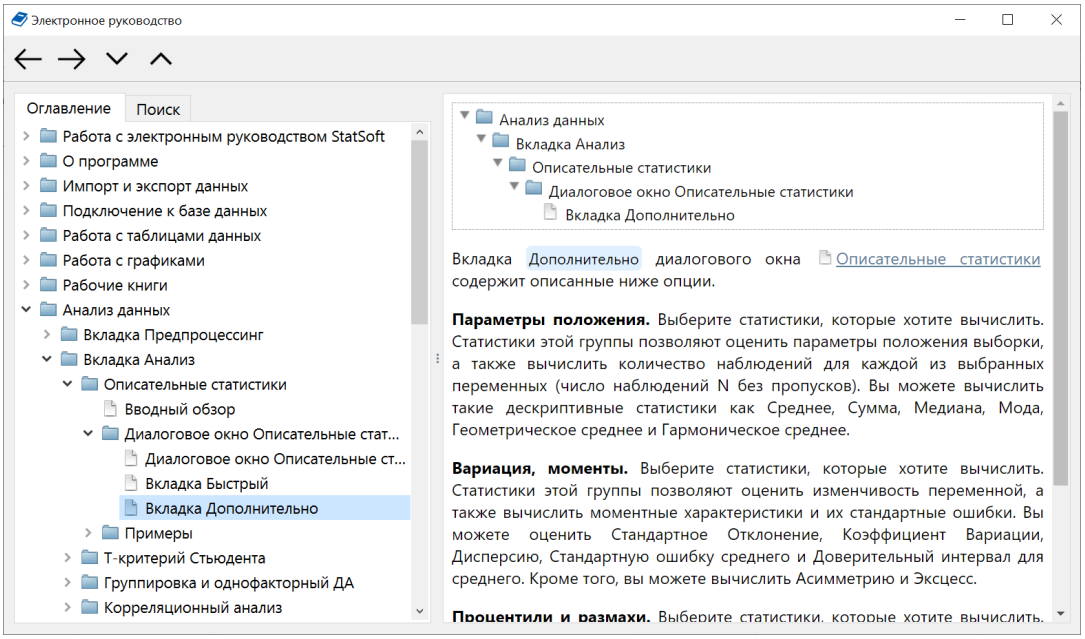
Использование справки во время работы

Во время работы в ПО СтатСофт вы можете получить доступ к Справке в меню Справка или нажав на значок книги в правом углу:



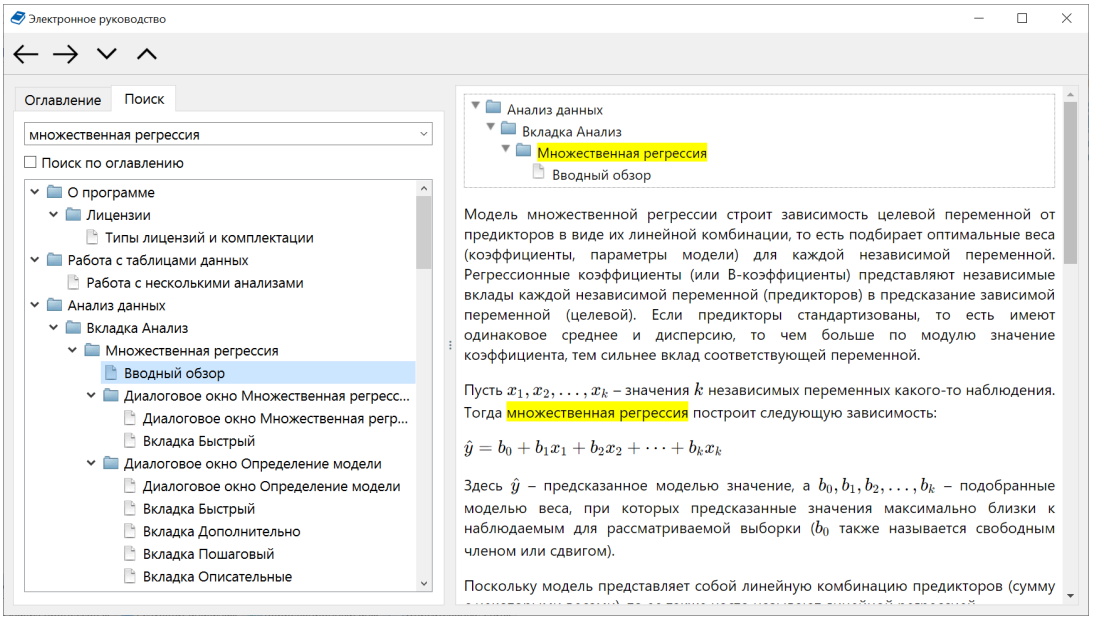
Помимо этого, каждый модуль анализа также имеет кнопку Справка, при нажатии на которую будет открыта статья Справки, соответствующая этому окну анализа и его вкладке:





Структура электронного руководства

Окно справки состоит из двух частей: панель навигации (слева) и область просмотра (справа). На панели навигации содержится многоуровневое оглавление, а отдельная вкладка "Поиск" в том же окне позволяет осуществлять поиск по словам по всем разделам справки.



Приложение. Комплектации программы и список модулей

Комплектации ПО

Для программных продуктов компании СтатСофт предусмотрены следующие виды комплектации, различающиеся по функциональным возможностям, предоставляемым пользователю:

- *Аналист - Analyst*, базовая комплектация
- *Моделер - Modeler*, продвинутая комплектация
- *Дейта Сайнтист - Data Sientist*, максимальная комплектация ПО, включающая классические статистические методы и методы искусственного интеллекта, нейронных сетей, оптимизацию, внедрение и другие аналитические технологии
- *Академик Аналист - Academic Analyst*, базовая комплектация для вузов и академических учреждений
- *Академик Дейта Сайнтист - Academic Data Sientist*, максимальная комплектация для вузов и академических учреждений
- *Триал - Trial*, пробная версия ПО

Комплектации *Analyst, Modeler, Data Scientist* отличаются составом и ценой. Длительность лицензии – 365 календарных дней.

Комплектация *Trial* предназначена для ознакомительных целей и предоставляется бесплатно. Важное отличие этой комплектации от остальных – сокращенный набор примеров. Длительность лицензии – 30 календарных дней.

В новых версиях ПО СтатСофт проводится обновление и дополнение функционала ПО СтатСофт.

Состав комплектаций

Краткая сравнительная таблица комплектаций

Раздел	Analyst	Modeler	Data Scientist	Academic	Trial
Предпроцессинг	10	10	10	10	10
Анализ	18	18	18	18	12

Графики	12	12	12	12	10
ИИ Регрессия	0	4	5	4	2
ИИ Классификация	0	2	3	2	2
Предсказания	0	5	5	5	5
Оптимизация	0	0	1	0	1
Интерполяция	4	4	4	4	2
Кластеризация	5	5	5	5	3
Разведочный анализ	9	9	9	9	3
Временные ряды	3	3	4	3	0
Обработка изображений	1	1	1	1	0

Полная сравнительная таблица комплектаций

	Раздел / анализ	**Analyst**	**Modeler**	**Data Scientist**	**Academic**	**Trial**
	Предпроцессинг					
0	Разбиение на выборки	✓	✓	✓	✓	✓
1	Фильтрация данных	✓	✓	✓	✓	✓
2	Нормальное распределение	✓	✓	✓	✓	✓
3	Проверка нормальности	✓	✓	✓	✓	✓
4	Обработка выбросов	✓	✓	✓	✓	✓

5	Балансировка SMOTE и ADASYN	✓	✓	✓	✓	✓
6	Обработка ПД	✓	✓	✓	✓	✓
7	Подгонка распределений	✓	✓	✓	✓	✓
8	Стандартизация данных	✓	✓	✓	✓	✓
9	Преобразовать к нормальному	✓	✓	✓	✓	✓
	Анализ					
10	Т-критерий Стьюдента	✓	✓	✓	✓	✓
11	Анализ процессов	✓	✓	✓	✓	✓
12	Полиномиальная регрессия	✓	✓	✓	✓	✓
13	Непараметрическая статистика	✓	✓	✓	✓	
14	Дисперсионный анализ	✓	✓	✓	✓	✓
15	Множественная нелинейная регрессия	✓	✓	✓	✓	
16	Анализ измерительных систем	✓	✓	✓	✓	✓
17	Шкалирование лучше-хуже	✓	✓	✓	✓	
18	Группировка и однофакторный ДА	✓	✓	✓	✓	
19	Обобщенные линейные модели	✓	✓	✓	✓	
20	Таблицы частот	✓	✓	✓	✓	✓

21	Описательные статистики	✓	✓	✓	✓	✓
22	Калькулятор Шесть Сигма	✓	✓	✓	✓	✓
23	Совместный анализ	✓	✓	✓	✓	
24	Таблицы сопряженности	✓	✓	✓	✓	✓
25	Карты контроля качества	✓	✓	✓	✓	✓
26	Корреляционный анализ	✓	✓	✓	✓	✓
27	Множественная регрессия	✓	✓	✓	✓	✓
	Графики					
28	Диаграмма Вороного	✓	✓	✓	✓	✓
29	3М Диаграмма рассеяния	✓	✓	✓	✓	✓
30	Диаграмма рассеяния	✓	✓	✓	✓	✓
31	К-ближайших соседей	✓	✓	✓	✓	
32	График средних с ошибками	✓	✓	✓	✓	✓
33	Линейный график	✓	✓	✓	✓	✓
34	Диаграмма причин и следствий	✓	✓	✓	✓	✓
35	Интерактивный анализ	✓	✓	✓	✓	
36	Гистограмма	✓	✓	✓	✓	✓
37	Разбиение Делоне	✓	✓	✓	✓	✓

38	Ящик с усами	✓	✓	✓	✓	✓
39	Круговая диаграмма	✓	✓	✓	✓	✓
	ИИ Регрессия					
40	XGBoost (регрессия)		✓	✓	✓	
41	Множественная регрессия		✓	✓	✓	
42	Выбор лучшей модели			✓		
43	Деревья регрессии		✓	✓	✓	✓
44	Нейронные сети (регрессия)		✓	✓	✓	✓
	ИИ Классификация					
45	Выбор лучшей модели (классификация)			✓		
46	Нейронные сети (классификация)		✓	✓	✓	✓
47	Деревья классификации		✓	✓	✓	✓
	Предсказания					
48	Предсказания деревьев решений		✓	✓	✓	✓
49	Предсказания нейронных сетей		✓	✓	✓	✓
50	Предсказания множественной регрессии		✓	✓	✓	✓
51	Предсказания обобщенных линейных моделей		✓	✓	✓	✓

52	Предсказания XGBoost		✓	✓	✓	✓
	Оптимизация					
53	Деревья решений			✓		✓
	Интерполяция					
54	Кригинг	✓	✓	✓	✓	
55	Вариограммы	✓	✓	✓	✓	
56	Сплайн	✓	✓	✓	✓	✓
57	РБФ	✓	✓	✓	✓	✓
	Кластеризация					
58	Карты Кохонена	✓	✓	✓	✓	
59	Кластеризация К-средних	✓	✓	✓	✓	✓
60	Иерархическая кластеризация	✓	✓	✓	✓	✓
61	EM алгоритм	✓	✓	✓	✓	
62	DBSCAN кластеризация	✓	✓	✓	✓	✓
	Разведочный анализ					
63	Пробит регрессия	✓	✓	✓	✓	
64	Логит регрессия	✓	✓	✓	✓	✓
65	Факторный анализ	✓	✓	✓	✓	
66	Дискриминантный анализ	✓	✓	✓	✓	✓
67	Нелинейный дискр. анализ	✓	✓	✓	✓	✓
68	Канонический анализ	✓	✓	✓	✓	

69	Метод опорных векторов	✓	✓	✓	✓	
70	Случайные леса	✓	✓	✓	✓	
71	Алгоритм Априори	✓	✓	✓	✓	
	Временные ряды					
72	Сезонная АРПСС	✓	✓	✓	✓	
73	Спектральный анализ	✓	✓	✓	✓	
74	Преобразование временных рядов	✓	✓	✓	✓	
75	Вейвлет анализ			✓		
	Обработка изображений					
76	Описательные статистики	✓	✓	✓	✓	