Руководство пользователя

Программа Schicksal предназначена для проведения статистического анализа данных, представленных в табличной форме. Таблицы статистических данных, с которыми работает программа, представляют собой файлы с расширением .sks, которые при установке программы ассоциируются с ней по расширению. Для работы с файлами, содержащими таблицы статистических данных, предназначено главное меню «Файл», показанное на рисунке 1.

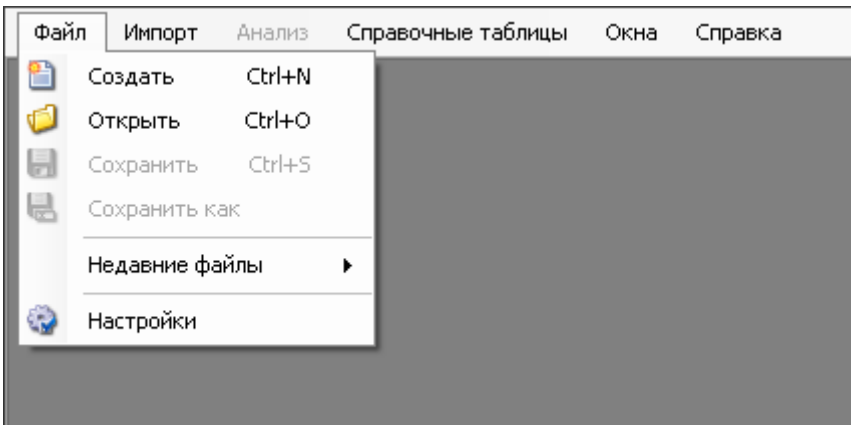


Рисунок 1. Главное меню «Файл».

Команда «Создать» открывает диалог создания нового файла, в котором предлагается создать колонки таблицы. Пример заполнения этого диалога показан на рисунке 2.

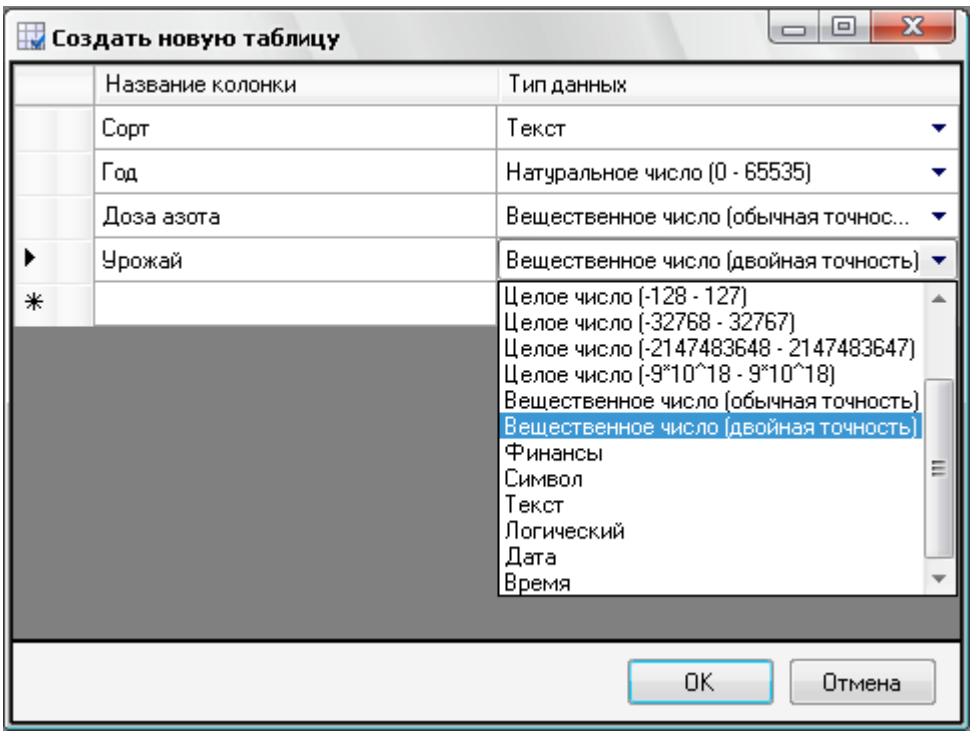


Рисунок 2. Диалог создания таблицы.

Рекомендуется для колонок, которые будут анализироваться, использовать тип данных, который выбирается по умолчанию – вещественное число с двойной точностью. Для колонок, содержащих описание, необходимо выбрать тип «Текст». Для остальных колонок тип выбирается исходя из того, что предполагается в них хранить. Таблица, которая получится при вводе таких данных в диалог создания таблицы, показана на рисунке 3.

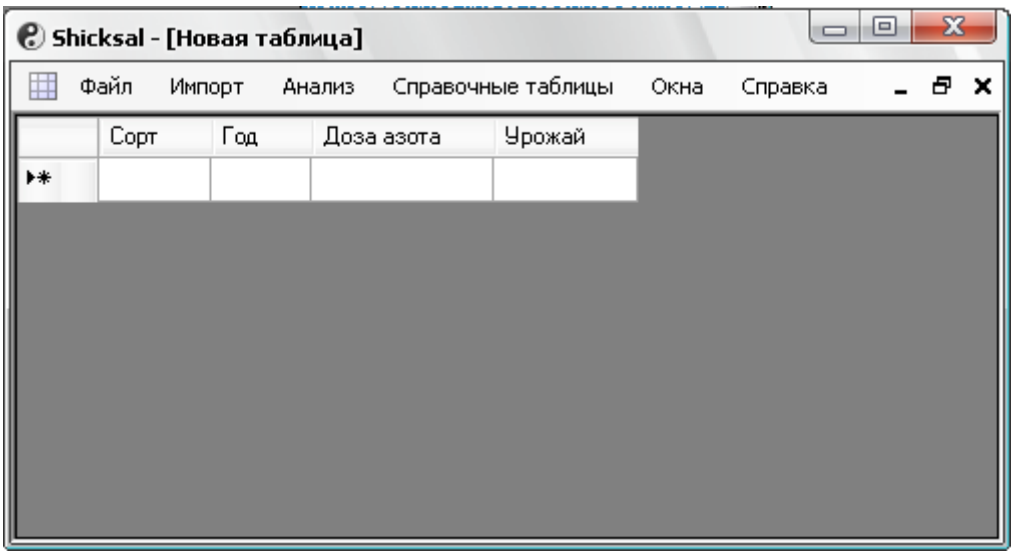


Рисунок 3. Созданная таблица.

Все ячейки созданной таблицы являются редактируемыми и проверяются в соответствии с теми типами данных, которые указаны при её создании. При вводе данных в строку, помеченную звёздочкой, новые строки создаются автоматически. После ввода данных таблицу можно сохранить в файл с помощью команд «Сохранить» или «Сохранить как». Созданный файл также можно открыть либо запустив программу и

используя пункт меню «Открыть», либо просто двойным кликом по файлу в файловом менеджере. При этом все файлы будут открываться в одной копии приложения, и переключаться между ними можно с помощью меню «Окна» (рисунок 4).

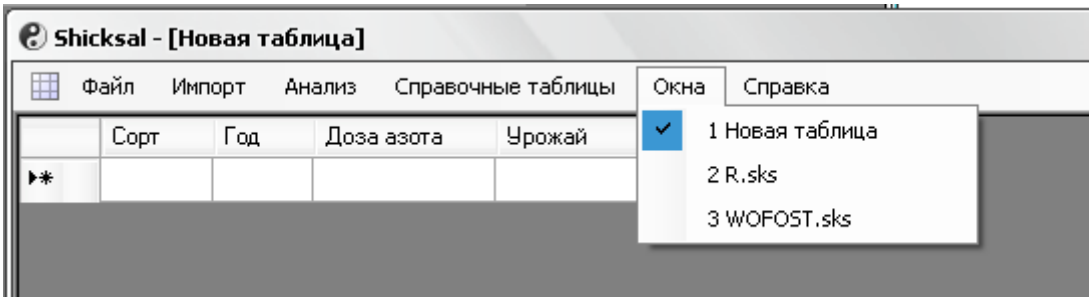


Рисунок 4. Переключение между открытыми файлами.

Кроме ручного ввода, данные можно импортировать. Для этого используется механизм плагинов, которые должны находиться в директории Plugins директории установки программы. В текущей версии единственный плагин для импорта импортирует выбранный диапазон из книги Excel. Рассмотрим, как создаётся новый плагин.

Структура директории, куда установлена программа, показана на рисунке 5. Видно, что среди установленных файлов есть файл Schicksal.Exchange.dll, предназначенный для поддержки плагинов.

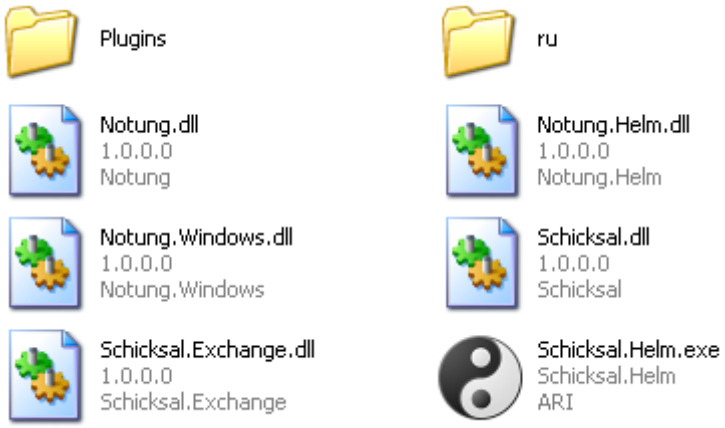


Рисунок 5. Содержимое директории установки программы.

В этом файле находится интерфейс **`ITableImport`**, содержащий единственный метод Import. Этот метод в качестве параметра принимает контекст (через него передаётся ссылка на главное окно программы – объект Windows Forms) и возвращает объект **`ImportResult`**, содержащий два свойства: текстовое описание результатов импорта и таблицу данных **`DataTable`**. Плагин импорта представляет собой класс, реализующий этот интерфейс, размещённый в dll, которую необходимо поместить в директорию Plugins. Кроме файла dll, необходимо поместить в эту директорию xml-файл с расширением .impot, содержащим описание плагина. Пример содержимого этого файла можно посмотреть в установленной программе:

```
<plugin assembly="ExcelJetImport.dll" name="EXCEL.JET.4.0.IMPORT" />
```

В атрибуте assembly необходимо указать имя файла dll, содержащего плагин. В атрибуте name можно написать произвольный текст, характеризующий сборку с

плагином. Любая сборка может содержать несколько плагинов. Библиотеки, от которых зависит сборка плагина, могут быть либо так же помещены в директорию Plugins, либо в основную директорию программы.

После того, как данные введены или импортированы, их можно проанализировать. Для этого предназначено меню «Анализ». В текущей версии имеется три варианта анализа данных: базовые статистики дисперсионный анализ и регрессионный анализ. При запуске любого из них вызывается диалог, показанный на рисунке 6. При анализе данных предполагается, что колонки таблицы можно подразделить на три категории: одна колонка должна быть выбрана как анализируемый показатель, результат воздействия факторов, в других колонках хранятся разные градации факторов, влияющих на результат, а остальные колонки не участвуют в анализе и создают стохастический шум, необходимый для оценки значимости влияния исследуемых факторов. Кроме того, диалог позволяет задать фильтр для того чтобы анализировать не все строки таблицы, а только выборку. Условия фильтрации можно объединять с использованием слов and и or, а также круглых скобок. Если имя колонки содержит пробелы, необходимо его взять в квадратные скобки. Ещё можно выбрать уровень значимости для исследований. При повторном запуске анализа одной и той же таблицы программа загружает ранее введённые данные, чтобы пользователю не приходилось многократно проделывать одну и ту же работу. Этого не происходит, если данные были введены некорректно, и расчёт статистики завершился ошибкой.

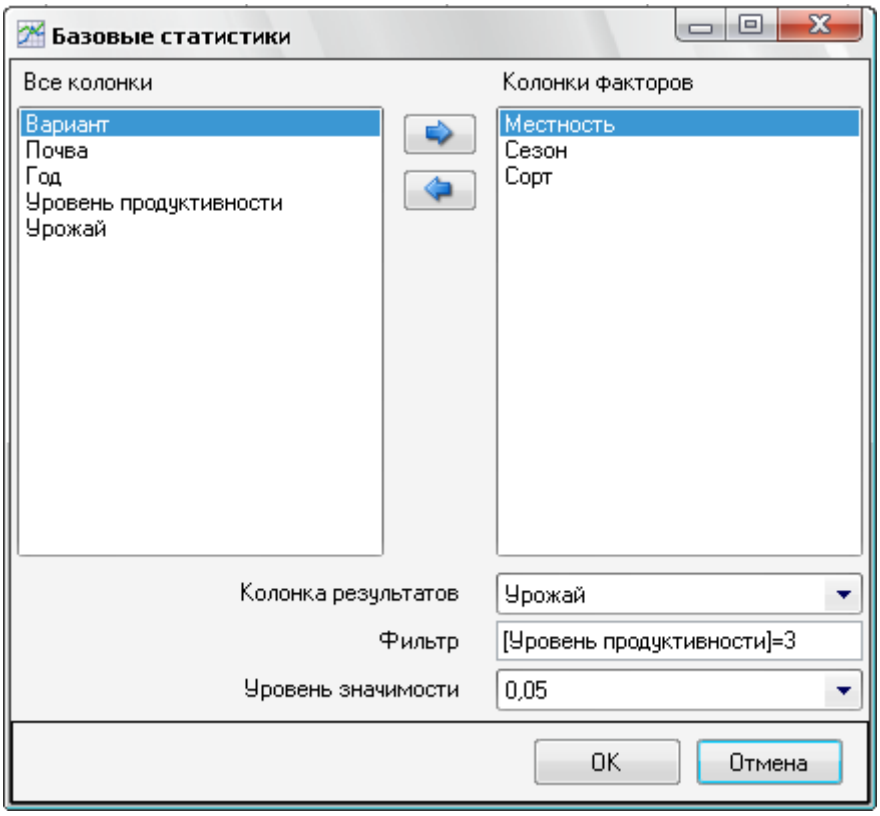


Рисунок 6. Диалог ввода параметров анализа.

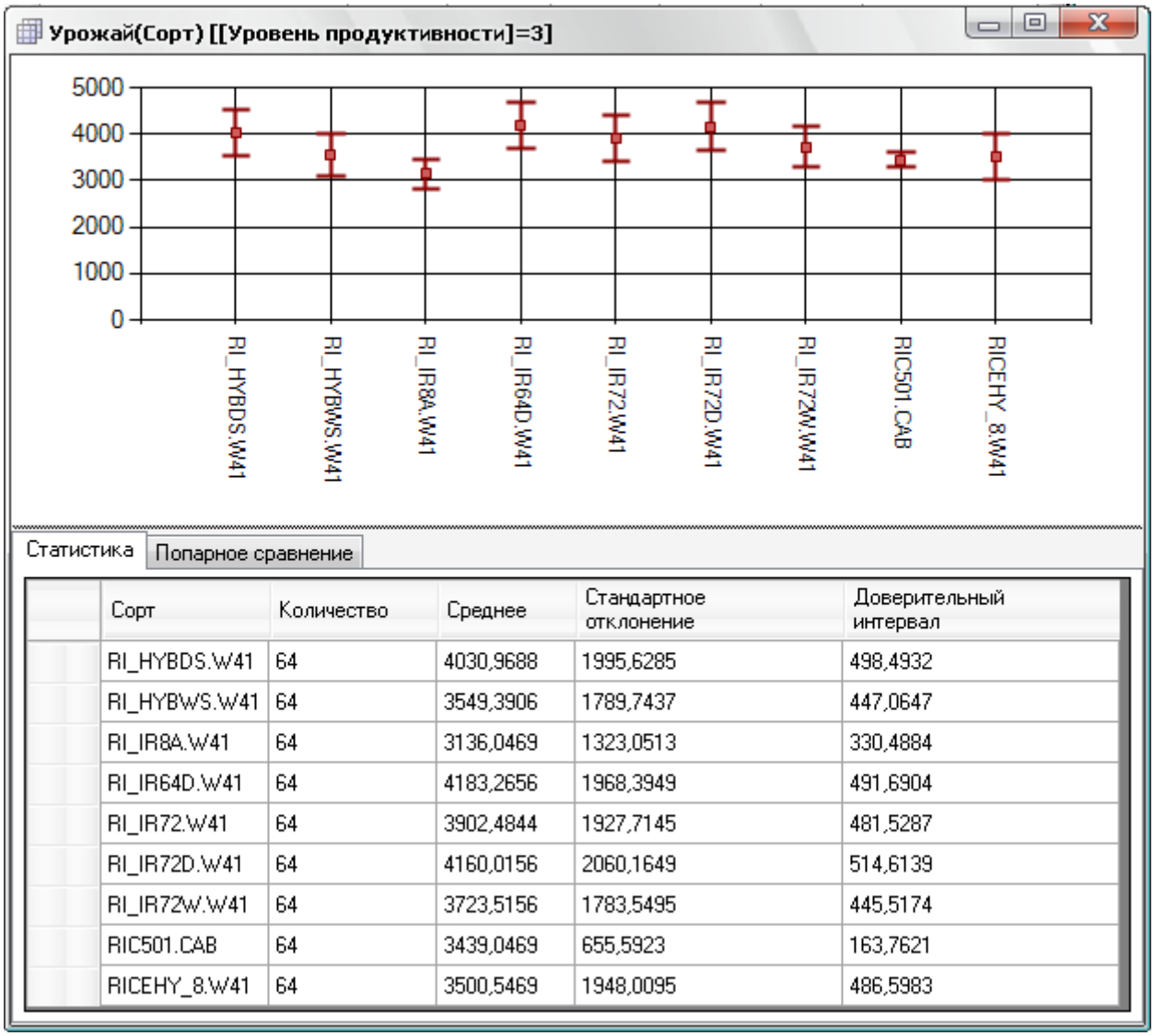
При расчёте базовых статистик получается плоская таблица, содержащая статистические показатели выбранной колонки результатов в разрезе выбранных факторов. Это такие показатели, как средние значения, минимумы, максимумы, медианы, стандартные отклонения и доверительные интервалы. Также в таблице показано количество строк анализируемой таблицы, попавших в каждую градацию факторов, в разрезе которых проводится анализ (Рисунок 7).

Описание	Среднее	Медиана	Минимум	Максимум	Количество	Стандартное отклонение	Доверительный интервал
Местность = 'Vara...	5337,5000	5519,0000	5156,0000	5519,0000	8	194,0317	162,2145
Местность = 'Vara...	4021,5000	4130,0000	3913,0000	4130,0000	8	115,9914	96,9712
Местность = 'Vara...	5408,0000	5608,0000	5208,0000	5608,0000	8	213,8090	178,7488
Местность = 'Pati...	3773,0000	5649,0000	1897,0000	5649,0000	8	2005,5284	1676,6637
Местность = 'Pati...	3325,5000	5005,0000	1646,0000	5005,0000	8	1795,4610	1501,0430
Местность = 'Pati...	2982,0000	4290,0000	1674,0000	4290,0000	8	1398,3108	1169,0171
Местность = 'Pati...	3949,0000	5822,0000	2076,0000	5822,0000	8	2002,3212	1673,9824
Местность = 'Pati...	3686,0000	5506,0000	1866,0000	5506,0000	8	1945,6618	1626,6140
Местность = 'Pati...	3875,5000	5791,0000	1960,0000	5791,0000	8	2047,7556	1711,9665
Местность = 'Pati...	3477,5000	5161,0000	1794,0000	5161,0000	8	1799,7372	1504,6179
Местность = 'Pati...	3290,0000	3986,0000	2594,0000	3986,0000	8	744,0553	622,0458
Местность = 'Pati...	3280,0000	5221,0000	1339,0000	5221,0000	8	2075,0163	1734,7570
Местность = 'New...	1815,5000	1831,0000	1800,0000	1831,0000	8	16,5702	13,8530

Рисунок 7. Результат анализа базовых статистик.

С помощью команды «Экспорт» в контекстном меню таблицы результаты можно выгрузить в html-файл с последующей возможностью его копирования и вставки.

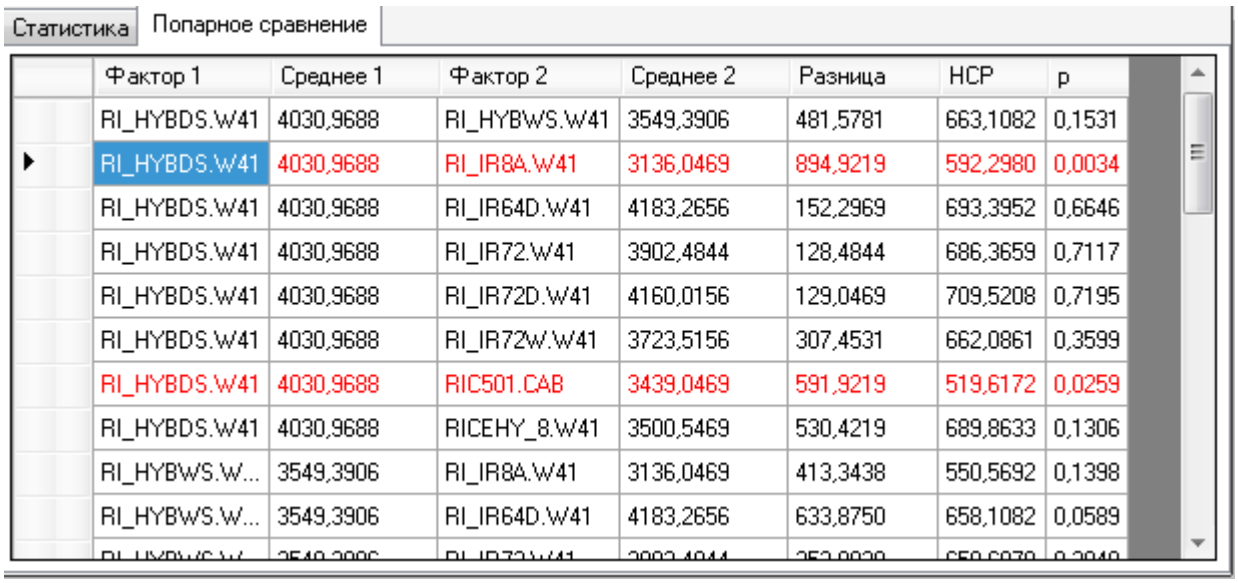
Дисперсионный анализ для запуска требует того же самого диалога, что и базовые статистики, но позволяет получить гораздо более детальные результаты. Таблица, отображаемая при завершении дисперсионного анализа, содержит оценку того, насколько значимо влияние каждого из выбранных факторов, а также их комбинаций. Те факторы и комбинации факторов, влияние которых значимо на выбранном уровне значимости, выделяются красным цветом. Цвет выделения, а также язык пользовательского интерфейса и генерируемых файлов, в которые экспортируются результаты анализа, можно задать в настройках, доступных через меню «Файл». Эту таблицу можно также выгрузить в файл через контекстное меню «Экспорт», однако при этом выгружается не только сама эта таблица, но и детальная информация по влиянию каждого фактора и комбинации факторов. Для доступа к этой информации в пользовательском интерфейсе программы необходимо два раза щёлкнуть мышью по выбранной строке, в которой показана значимость влияния фактора или комбинации фактора. Откроется окно, показанное, на рисунке 8.



Сорт	Количество	Среднее	Стандартное отклонение	Доверительный интервал
RI_HYBDS.W41	64	4030,9688	1995,6285	498,4932
RI_HYBWS.W41	64	3549,3906	1789,7437	447,0647
RI_IR8A.W41	64	3136,0469	1323,0513	330,4884
RI_IR64D.W41	64	4183,2656	1968,3949	491,6904
RI_IR72.W41	64	3902,4844	1927,7145	481,5287
RI_IR72D.W41	64	4160,0156	2060,1649	514,6139
RI_IR72W.W41	64	3723,5156	1783,5495	445,5174
RIC501.CAB	64	3439,0469	655,5923	163,7621
RICEHY_8.W41	64	3500,5469	1948,0095	486,5983

Рисунок 8. Окно детальных результатов анализа влияния фактора.

На примере окна, которое показано на рисунке, указано влияние урожая от сорта (это указывается, наряду с условием фильтрации, в заголовке окна). В верхней части окна на графике показаны средние значения каждой из градаций факторов, с доверительными интервалами, а в нижней части окна та же самая информация представлена в виде таблицы. Контекстное меню позволяет скопировать график и вставить его в любой документ, поддерживающий вставку изображений. На второй закладке «Попарное сравнение» показаны результаты попарного сравнения каждой градации этого фактора или комбинации факторов с каждой (рисунок 9).



	Фактор 1	Среднее 1	Фактор 2	Среднее 2	Разница	НСР	p
	RI_HYBDS.W41	4030,9688	RI_HYBWS.W41	3549,3906	481,5781	663,1082	0,1531
▶	RI_HYBDS.W41	4030,9688	RI_IR8A.W41	3136,0469	894,9219	592,2980	0,0034
	RI_HYBDS.W41	4030,9688	RI_IR64D.W41	4183,2656	152,2969	693,3952	0,6646
	RI_HYBDS.W41	4030,9688	RI_IR72.W41	3902,4844	128,4844	686,3659	0,7117
	RI_HYBDS.W41	4030,9688	RI_IR72D.W41	4160,0156	129,0469	709,5208	0,7195
	RI_HYBDS.W41	4030,9688	RI_IR72W.W41	3723,5156	307,4531	662,0861	0,3599
	RI_HYBDS.W41	4030,9688	RIC501.CAB	3439,0469	591,9219	519,6172	0,0259
	RI_HYBDS.W41	4030,9688	RICEHY_8.W41	3500,5469	530,4219	689,8633	0,1306
	RI_HYBWS.W...	3549,3906	RI_IR8A.W41	3136,0469	413,3438	550,5692	0,1398
	RI_HYBWS.W...	3549,3906	RI_IR64D.W41	4183,2656	633,8750	658,1082	0,0589

Рисунок 9. Закладка с попарным сравнением градаций факторов.

Для оценки статистической значимости различия каждой из пар используется критерий наименьшей существенной разницы (НСР). Статистически значимые различия (т.е., превышающие НСР) показаны красным цветом. Кроме того, контекстное меню содержит фильтр, позволяющий выбрать только значимые различия или только те пары, которые включают определённую градацию фактора или комбинации факторов. Если выбрано отображение только статистически значимых различий, таблица отображается не красным, а синим цветом (это можно поменять в настройках программы). Окно фильтрации показано на рисунке 10.

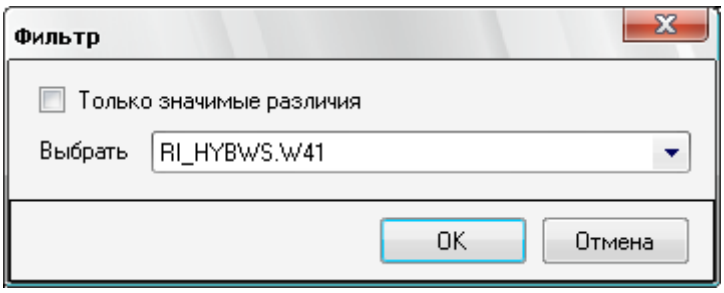


Рисунок 10. Окно фильтрации результатов попарного сравнения градаций.

Двойной клик по сравниваемой паре выводит информацию по строке в отдельном диалоге, откуда её можно скопировать.

Кроме этого, в программе присутствует меню «Справочные таблицы», в которых представлены значения статистичеких констант для различных уровней значимости и количествах степеней свободы.

Плагин для импорта данных, поставляемый вместе с программой, позволяет импортировать данные из Excel. Диалог импорта показан на рисунке 11. Он содержит всего два поля, которые требуется заполнить: путь к файлу Excel и лист или диапазон в этом файле, которую требуется импортировать. Для того чтобы импорт сработал, необходимо, чтобы в операционной системы был драйвер OleDb Jet 4.0, так как этот компонент использует его для чтения данных из Excel. Кроме того, сам файл Excel должен быть сформирован так, чтобы колонки соответствовали факторам и зависимым от них показателям, а строки – кортежам, содержащим эти данные для каждого набора измерений. Иначе говоря, структура таблицы Excel должна быть такой же, как структура

таблицы Schicksal, которую требуется в итоге получить. Первая строка импортируемых данных Excel должна содержать названия колонок. Эти названия не должны содержать символы квадратных скобок, сложения и запятых.

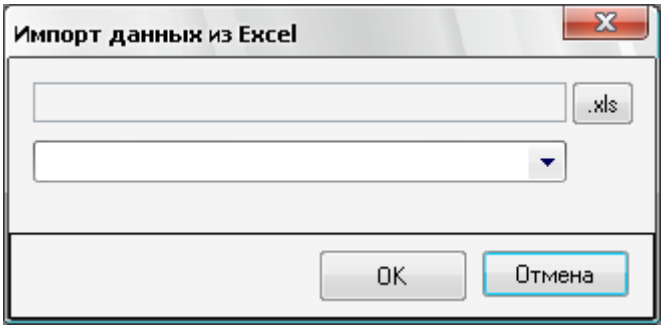


Рисунок 11. Окно простого импорта данных из Excel.

Кроме того, для лучшего качества импортированных данных рекомендуется импортируемые данные поместить в диапазон, а не просто на лист. Кроме того, если какие-то колонки предназначены для хранения чисел, следует для них явно задать числовой формат. Особенно это важно тогда, когда числа, подлежащие импорту, являются дробными: если разделитель целой и дробной части не соответствует тому, который установлен в настройках операционной системы, они импортируются как строки, и не смогут быть проанализированы статистически.

Второй вариант этого импорта позволяет импортировать вещественные числа, записанные в ячейках двумерной таблицы как в матрице. Этот импорт не просто импортирует данные из Excel в том виде, в котором они там есть, а формирует реляционную таблицу из матрицы. Для этого, помимо имени файла Excel и названия таблицы, плагин требует указать имена результирующих колонок, в которые превратятся значения из колонок, строк и ячеек импортируемой таблицы (рисунок 12).

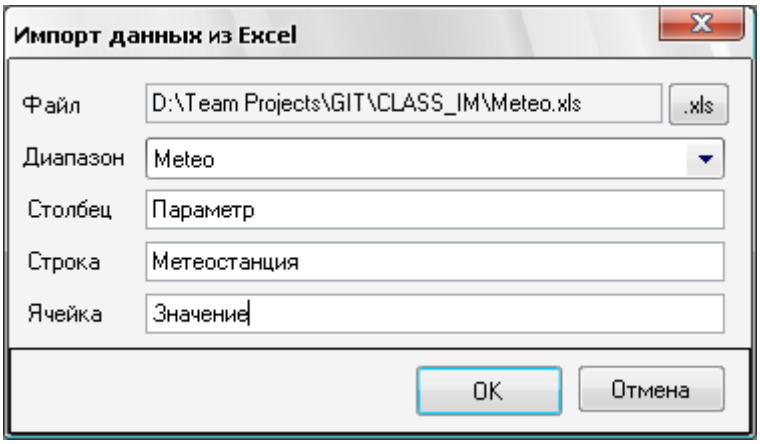
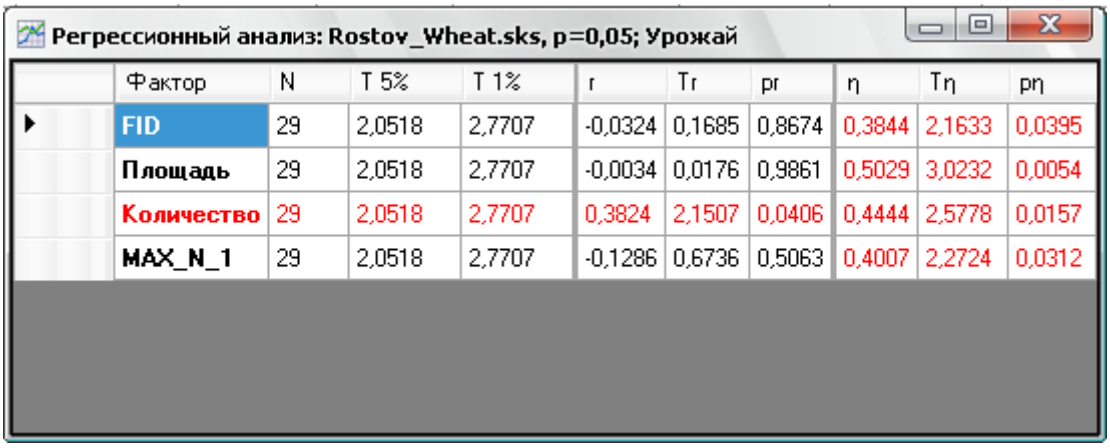


Рисунок 12. Матричный импорт данных из Excel.

Регрессионный анализ для ввода исходных данных требует того же самого диалога, что и базовые статистики, и дисперсионный анализ. После ввода данных в случае, если анализ прошёл успешно, показывается следующий диалог:



Фактор	N	T 5%	T 1%	r	Tr	pr	η	Tη	pη
FID	29	2,0518	2,7707	-0,0324	0,1685	0,8674	0,3844	2,1633	0,0395
Площадь	29	2,0518	2,7707	-0,0034	0,0176	0,9861	0,5029	3,0232	0,0054
Количество	29	2,0518	2,7707	0,3824	2,1507	0,0406	0,4444	2,5778	0,0157
MAX_N_1	29	2,0518	2,7707	-0,1286	0,6736	0,5063	0,4007	2,2724	0,0312

Рисунок 13. Результаты регрессионного анализа.

На этом диалоге каждая строка соответствует выбранному на предыдущем диалоге фактору, название которого отображается в первой колонке. Семантика последующих колонок такова:

- N – количество пар значений, которые программа нашла в таблице исходных данных. Это количество может быть меньше количества строк в исходной таблице по двум причинам:
  - К таблице на диалоге ввода данных для анализа применён фильтр
  - В некоторых строках не заполнены ячейки с одним из выбранных факторов или результатом
- T – значение критерия Стьюдента для фактического коэффициента корреляции и количества пар значений, обработанных в ходе анализа
- T 5%, T 1% - табличные значения критерия Стьюдента для уровней значимости 5 % и 1 %.
- r – коэффициент линейной корреляции Пирсена
- Tr – фактическое значение критерия Стьюдента для коэффициента линейной корреляции и количества пар значений.
- pr – уровень значимости, соответствующий фактическому значению критерия Стьюдента и количеству пар значений, обработанных в ходе анализа. Если он ниже критического значения, выбранного на диалоге ввода параметров статистического анализа, строка таблицы выделяется цветом.
- η – коэффициент криволинейной корреляции
- Tη – фактическое значение критерия Стьюдента для коэффициента криволинейной корреляции и количества пар значений.
- pη – уровень значимости, соответствующий фактическому значению критерия Стьюдента и количеству пар значений, обработанных в ходе анализа. Если он ниже критического значения, выбранного на диалоге ввода параметров статистического анализа, ячейки строки, начиная с коэффициента криволинейной корреляции, выделяются цветом.

Эта таблица имеет контекстное меню с пунктом «Экспорт», который позволяет экспортировать результаты регрессионного анализа в html-файл. При двойном щелчке мыши по любому фактору появляется окно с графиком (рисунок 14). Это окно позволяет выбрать вид зависимости из пяти вариантов (прямая линия, парабола, гипербола, кривая Михаэлиса и экспонента). При выборе рисуется график соответствующей зависимости.

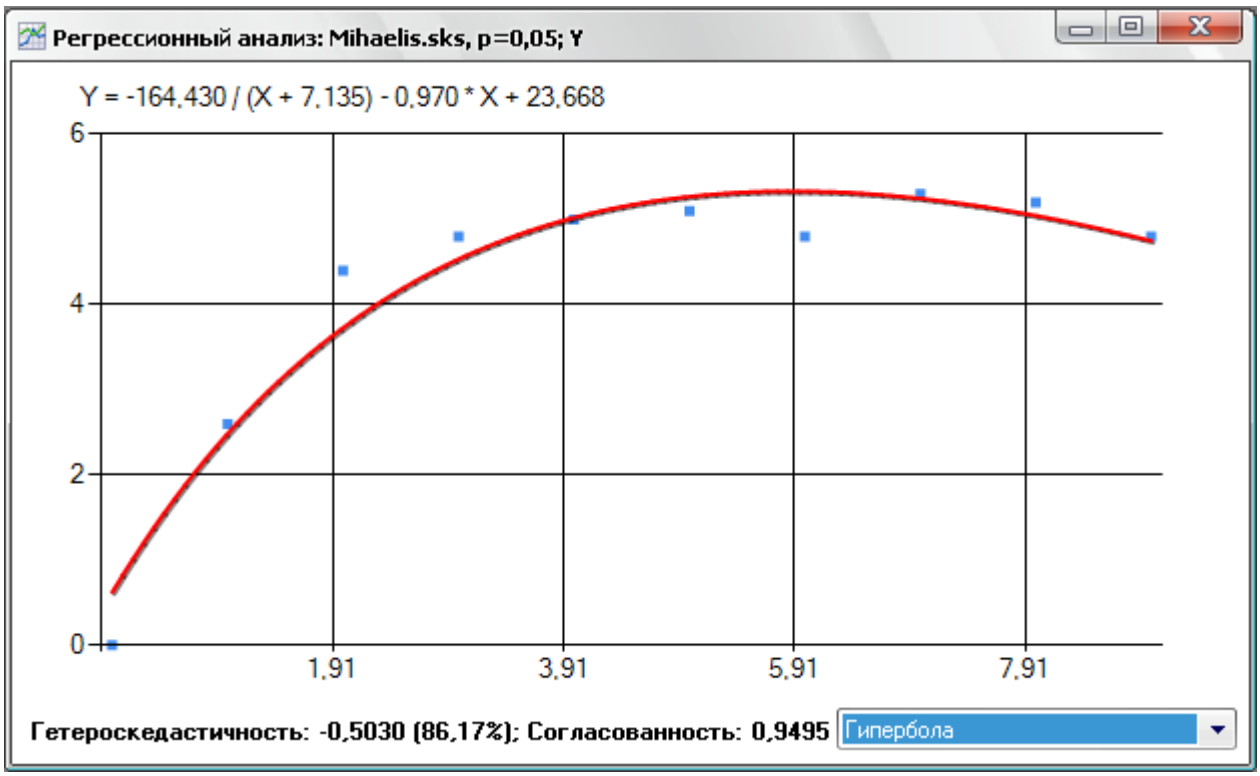


Рисунок 14. Окно детальных результатов регрессионного анализа.

Кроме графика регрессионной зависимости и точек, по которым построена регрессия, в левой нижней стороне показываются две характеристики, описывающие, насколько хорошо выбранный вид зависимости соответствует точкам: гетероскедастичность и согласованность. Гетероскедастичность рассчитывается как ранговый коэффициент корреляции Спирмана, показывающий зависимость между значением абсциссы и отклонением точек от графика. Кроме того, в скобках указана вероятность того, что этот коэффициент статистически значимо отличается от нуля. Она зависит от самого рангового коэффициента корреляции и количества точек. Согласованность – это то, насколько среднеквадратичное отклонение точек от графика меньше, чем среднеквадратичное отклонение точек от их среднего значения по ординате. Максимальное значение равно 1, минимальное значение при плохо подходящем виде зависимости может быть отрицательным (в этом случае график регрессионной зависимости не может пройти между точек и проходит рядом с ними). По умолчанию в выпадающем списке с видами зависимости выбирается тот, который даёт самую большую согласованность.

Метод наименьших квадратов, используемый при расчёте коэффициентов в формулах регрессионных зависимостей, достаточно чувствителен к выбросам – точкам, ордината которых выбивается из общей тенденции. Эти точки можно исключить с использованием фильтров, которые задаются при вводе параметров регрессионного анализа.