ЛАБОРАТОРНАЯ РАБОТА 5 ОЧИСТКА ДАННЫХ

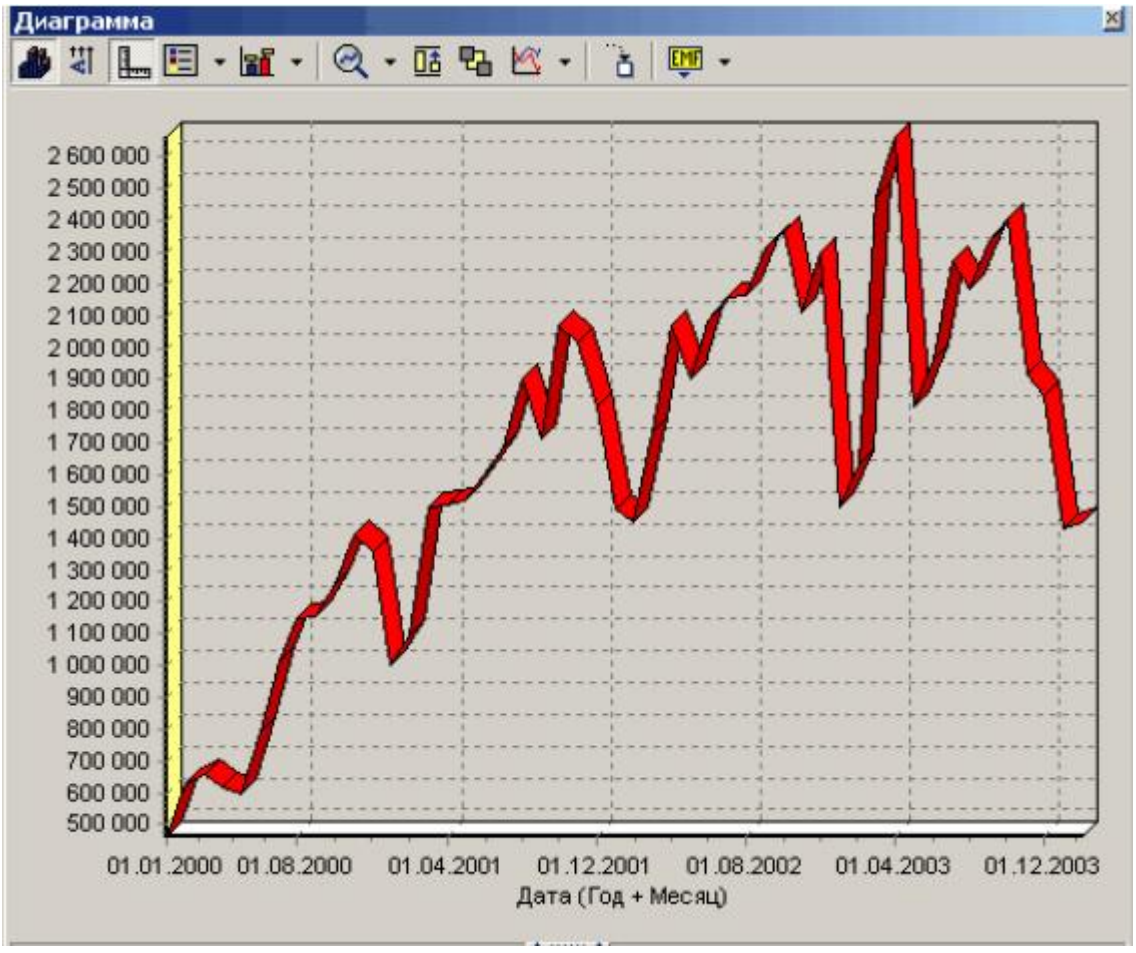
1. Парциальная предобработка

Парциальная предобработка служит для восстановления пропущенных данных, редактирования аномальных значений и спектральной обработки данных (например, сглаживания данных). Именно эти операции часто проводятся в первую очередь над данными.

Присутствие аномалий при построении моделей оказывает на них большое влияние, ухудшая качество результата.

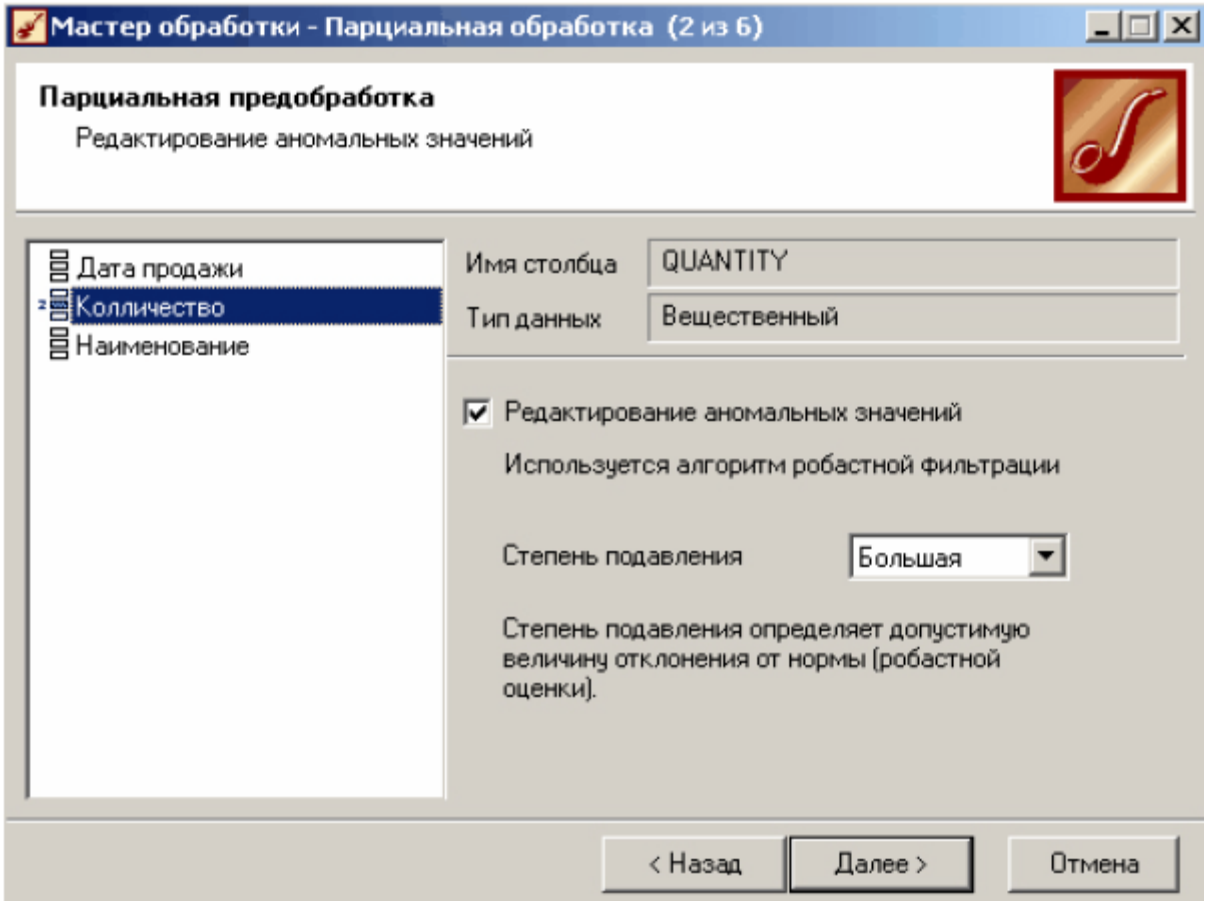
Исходные данные

В качестве примера возьмем данные из файла "Trade.txt". В данном файле находятся данные о продажах за некоторый период.



Обработка данных

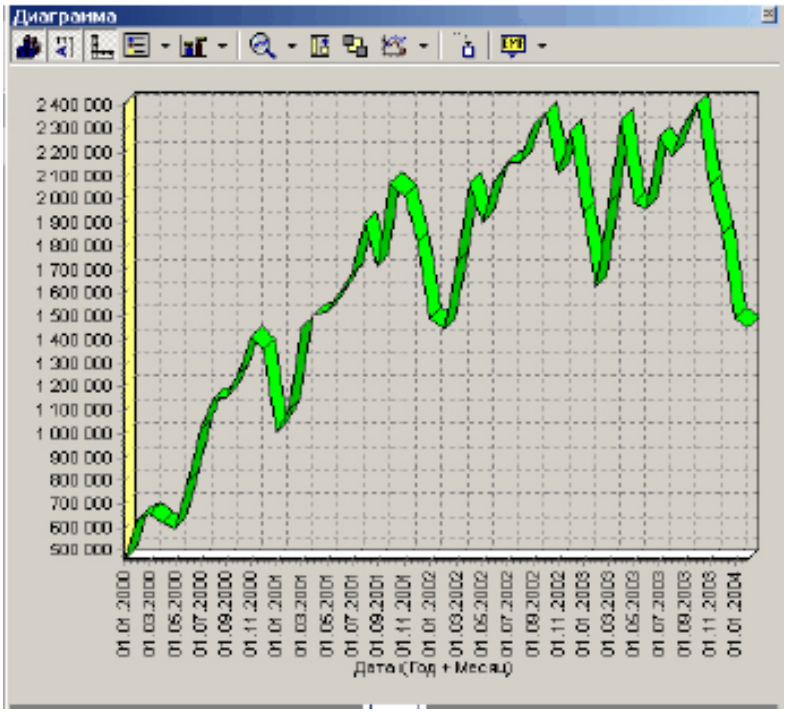
Как видно из диаграммы, выбросы ухудшают статистическую картину распределения данных. Воспользуемся Мастером обработки и выберем парциальную обработку.



В Мастере парциальной предобработки на втором шаге выбираем поле "Количество" и указываем ему тип обработки "Редактирование аномальных значений", степень подавления "Большая". Так как больше никаких действий над данными не планировалось, то переходим на шаг запуска процесса обработки и нажимаем "Пуск".

Результат

После выполнения процесса обработки на диаграмме видно, что выбросы уменьшились стала проясняться реальная картина продаж.



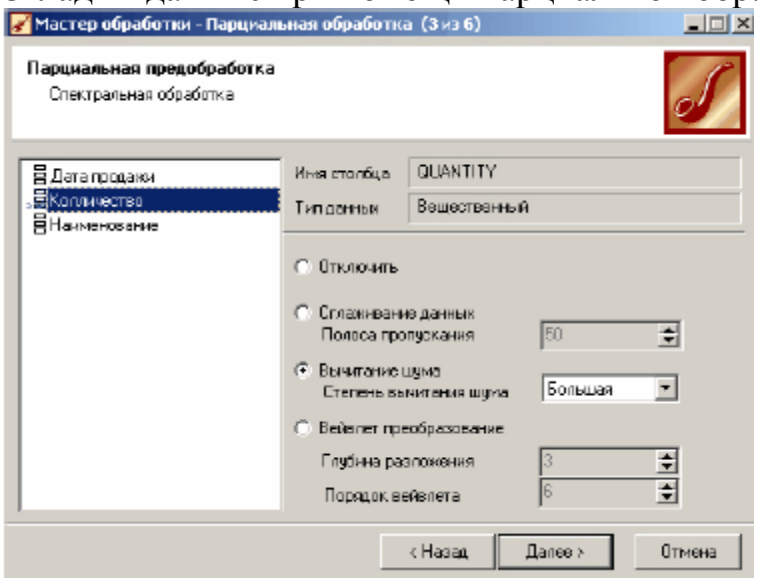
2. Спектральная обработка

Исходные данные

Продемонстрируем такой метод спектральной обработки, как вычитание шума. Для этого продолжим работу с данными файла "Trade.txt".

Обработка данных

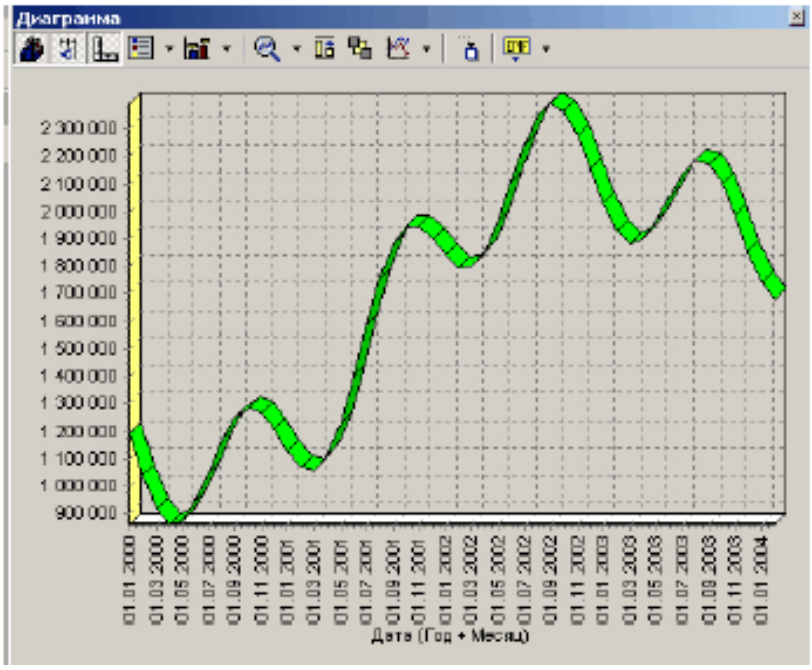
Сгладим данные при помощи парциальной обработки.



В Мастере парциальной предобработки на третьем шаге выбираем поле "Количество" и указываем ему тип обработки "Вычитание шума", степень подавления "Большая". Так как больше никаких обработок не планировалось, то переходим на шаг запуска процесса обработки и нажимаем "Пуск".

Результат

После выполнения процесса обработки выберем в качестве визуализации диаграмму.



Как видно из примера данные стали более сглаженными и могут служить для дальнейшей обработки. Взглянув на данные легко понять общую тенденцию.

2. Факторный анализ

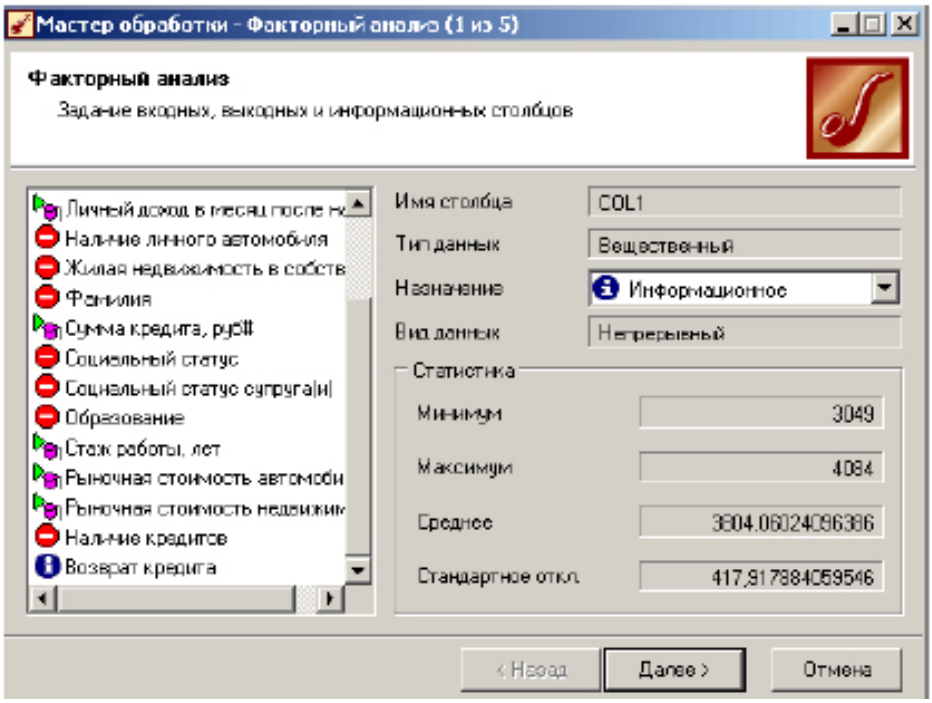
Факторный анализ служит для понижения размерности пространства входных факторов. Обработку можно выполнять как в автоматическом режиме (с указанием порога значимости), так и вручную (основываясь на значениях матрицы значимости).

Исходные данные

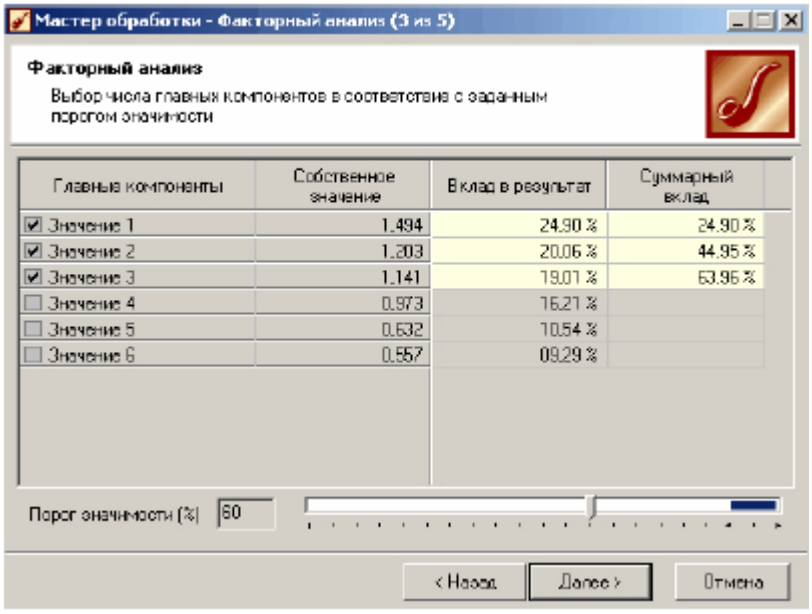
Рассмотрим применение обработчика на примере данных из файла "Anketa1.txt". Он содержит таблицу с информацией о кредитах граждан. Попробуем выявить значение факторов, влияющих на возврат кредита.

Понижение размерности пространства входных факторов

В Мастере обработки выберем факторный анализ и зададим входные поля - "Личный доход в месяц", "Сумма кредита", "Стаж работы", "Рыночная стоимость автомобиля", "Рыночная стоимость недвижимости".

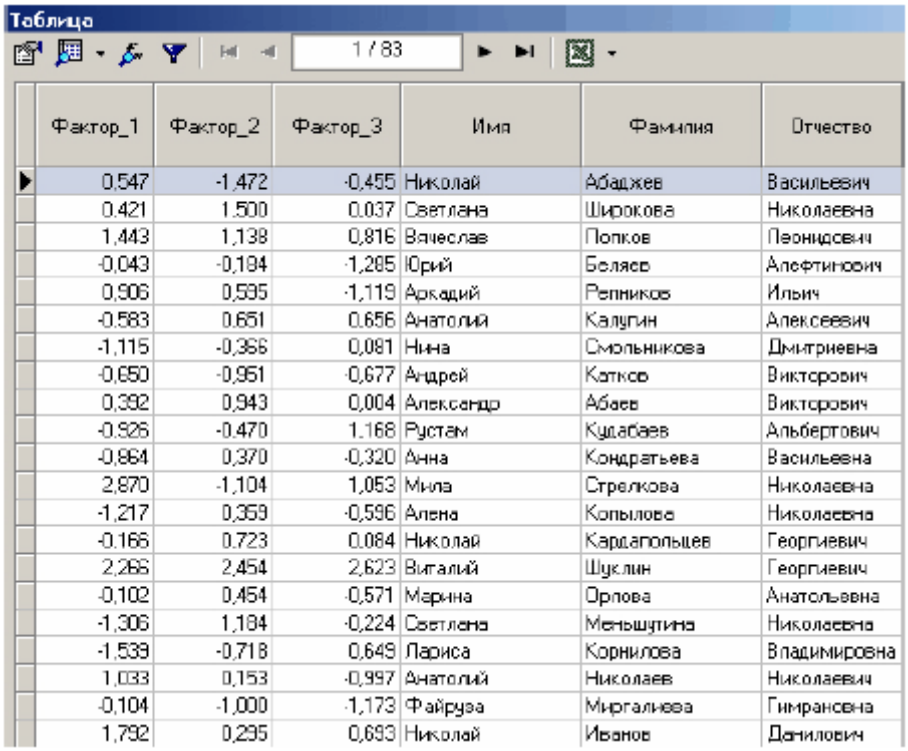


На следующем шаге предлагается запустить процесс понижения размерности пространства входных факторов. После завершения процесса можно выбрать, какие из полученных в результате обработки факторы оставить для дальнейшей работы. Это делается путем указания необходимого порога значимости.



Результат

Теперь необходимо перейти на следующий шаг и выбрать способ визуализации; результаты просмотрим в таблице.



Таблица

1 / 83

Фактор_1	Фактор_2	Фактор_3	Имя	Фамилия	Отчество
0,547	-1,472	-0,455	Николай	Абаджев	Васильевич
0,421	1,500	0,037	Светлана	Широкова	Николаевна
1,443	1,138	0,816	Вячеслав	Попков	Леонидович
-0,043	-0,184	-1,285	Юрий	Беляев	Алефтинович
0,906	0,595	-1,119	Аркадий	Репников	Ильич
-0,583	0,651	0,656	Анатолий	Калугин	Алексеевич
-1,115	-0,366	0,081	Нина	Смольникова	Дмитриевна
-0,650	-0,951	-0,677	Андрей	Катков	Викторович
0,392	0,943	0,004	Александр	Абаев	Викторович
-0,926	-0,470	1,168	Рустам	Кудабаев	Альбертович
-0,864	0,370	-0,320	Анна	Кондратьева	Васильевна
2,870	-1,104	1,053	Мила	Стрелкова	Николаевна
-1,217	0,359	-0,596	Алена	Копылова	Николаевна
-0,166	0,723	0,084	Николай	Кардапольцев	Георгиевич
2,266	2,454	2,623	Виталий	Шуклин	Георгиевич
-0,102	0,454	-0,571	Марина	Орлова	Анатольевна
-1,306	1,184	-0,224	Светлана	Меньшутина	Николаевна
-1,539	-0,718	0,649	Лариса	Корнилова	Владимировна
1,033	0,153	-0,997	Анатолий	Николаев	Николаевич
-0,104	-1,000	-1,173	Файруза	Миргалиева	Гимрановна
1,792	0,295	0,693	Николай	Иванов	Данилович

3. Корреляционный анализ

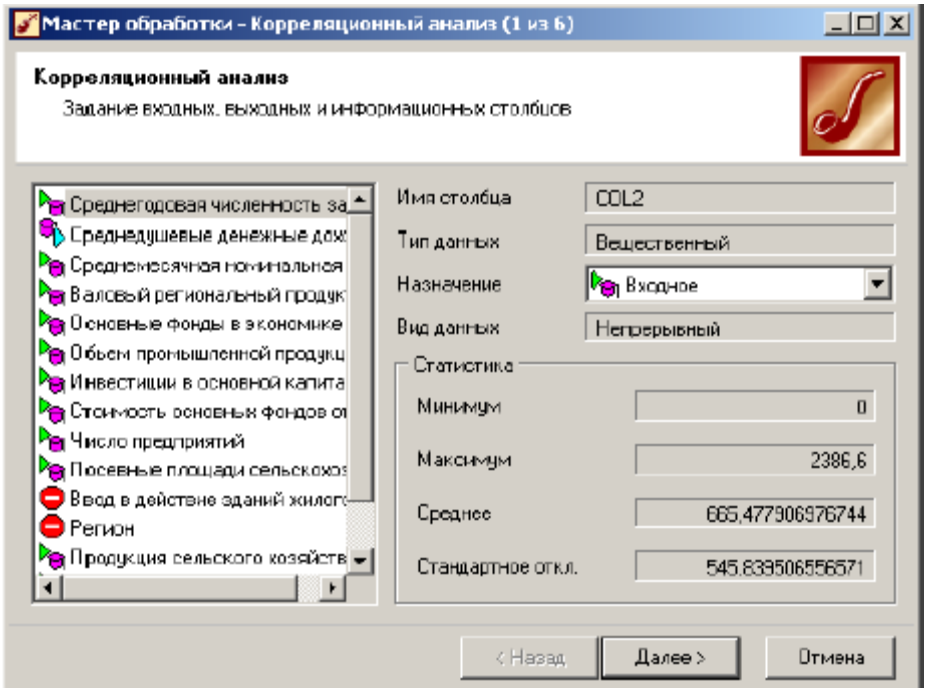
Корреляционный анализ применяется для оценки зависимости выходных полей данных от входных факторов и устранения незначащих факторов. Принцип корреляционного анализа состоит в поиске таких значений, которые в наименьшей степени коррелированны (взаимосвязаны) с выходным результатом. Такие факторы могут быть исключены из результирующего набора данных практически без потери полезной информации. Критерием принятия решения об исключении является порог значимости. Если корреляция (степень взаимозависимости) между входным и выходным факторами меньше порога значимости, то соответствующий фактор отбрасывается как незначащий.

Исходные данные

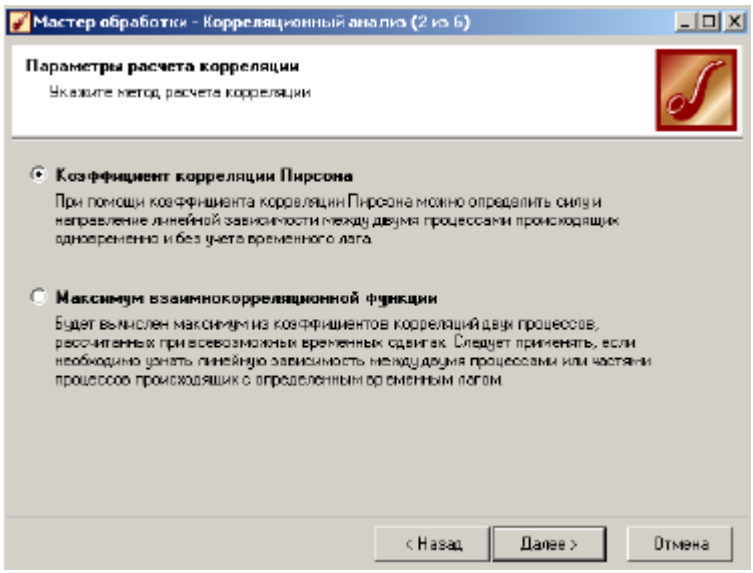
Рассмотрим применение обработки на примере данных из файла "region.txt". В данном примере определим степень влияния экономических показателей региона на среднедушевой доход жителей.

Устранение незначащих входных факторов

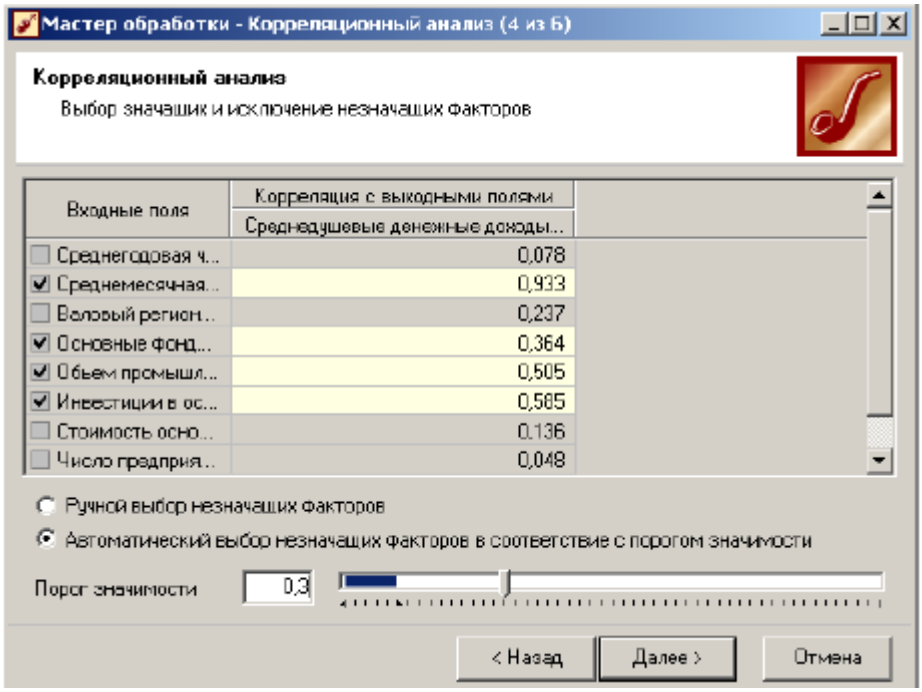
В Мастере обработки выбираем корреляционный анализ и задаем входные и выходные поля. Входными факторами будут являться все экономические показатели региона, а выходным будет "Среднедушевой денежный доход".



На следующем шаге необходимо выбрать метод на основе которого будет происходить расчет коэффициентов корреляции, выберем метод коэффициент корреляции Пирсона.

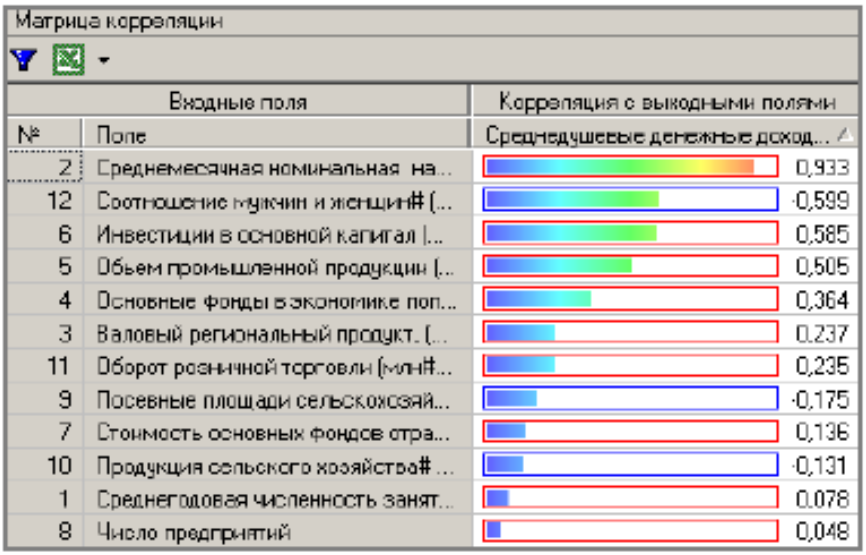


После выполнения предварительных настроек запускаем процесс корреляционного анализа, по результатам которого предлагается выбрать, какие факторы оставить для дальнейшей работы. Это делается либо вручную, основываясь на значениях матрицы ковариации, либо путем указания порога значимости (по умолчанию порог значимости равен 0.05).



Результат

По полученной матрице корреляции видно, какие факторы влияют сильнее, чем другие, и какие можно не учитывать при построении всевозможных моделей.



4. Дубликаты и противоречия

Одна из серьезных проблем, часто встречающаяся на практике, - наличие в данных дубликатов и противоречий.

Противоречивыми являются группы записей, в которых содержатся строки с одинаковыми входными факторами, но разными выходными. В такой ситуации непонятно, какое результирующее значение верное. Если противоречивые данные использовать для построения модели, то она окажется неадекватной. Поэтому противоречивые данные чаще всего лучше вообще исключить из исходной выборки.

Также в данных могут встречаться записи с одинаковыми входными факторами и одинаковыми выходными, т.е. дубликаты. Таким образом, данные несут избыточность. Присутствие дубликатов в анализируемых данных можно рассматривать как способ повышения "значимости" дублирующейся информации. Иногда они даже необходимы, например, если при построении модели нужно особо выделить некоторые наборы значений. Но все равно включение в выборку дублирующей информации должно происходить осознанно: в большинстве случаев дубликаты в данных являются следствием ошибок при подготовке данных.

Так или иначе возникает задача выявления дубликатов и противоречий. В **Deductor Studio** для автоматизации этого процесса есть соответствующий инструмент – обработка "Дубликаты и противоречия".

Суть обработки состоит в том, что определяются входные (факторы) и выходные (результаты) поля. Алгоритм ищет во всем наборе записи, для которых одинаковым входным полям соответствуют одинаковые (дубликаты) или разные (противоречия) выходные поля. На основании этой информации создаются два дополнительных логических поля – "Дубликат" и "Противоречие", принимающие значения "правда" или "ложь". В дополнительные числовые поля "Группа дубликатов" и "Группа противоречий" записываются номер группы дубликатов и группы противоречий, в которые попадает данная запись. Если запись не является дубликатом или противоречием, то соответствующее поле будет пустым.

Исходные данные

Рассмотрим механизм выявления дубликатов на примере данных файла "Anketa.txt". В этом файле находится информация об анкетных данных граждан, участвующих в кредитовании. Попробуем вычислить присутствие дубликатов.

Импортируем данные из текстового файла и посмотрим их в виде таблицы.

Так или иначе возникает задача выявления дубликатов и противоречий. В **Deductor Studio** для автоматизации этого процесса есть соответствующий инструмент – обработка "Дубликаты и противоречия".

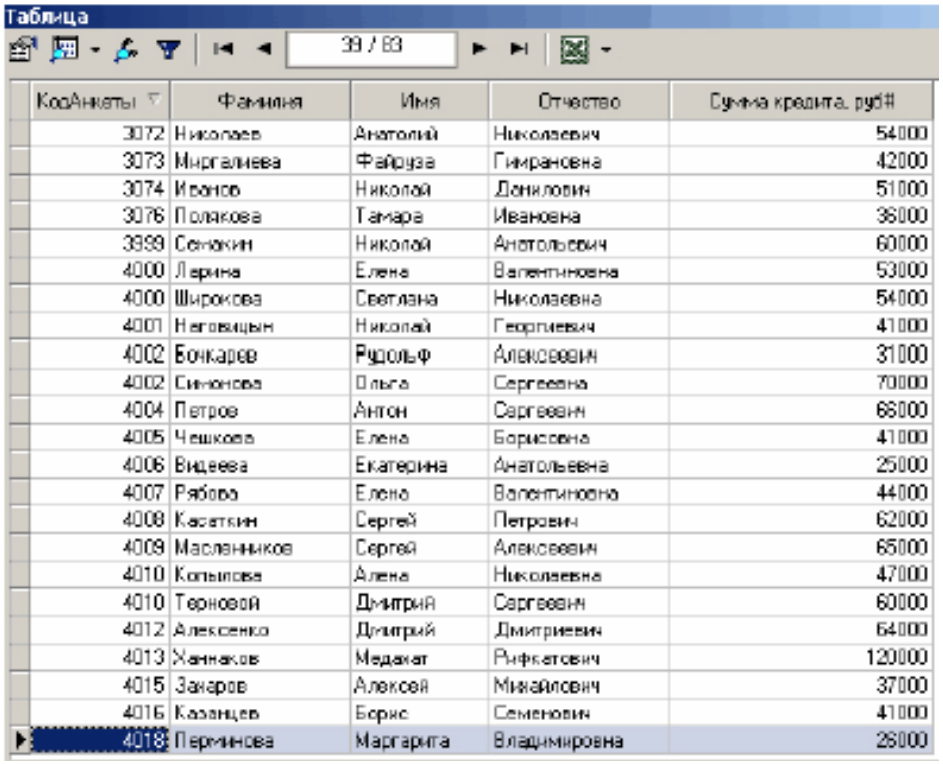
Суть обработки состоит в том, что определяются входные (факторы) и выходные (результаты) поля. Алгоритм ищет во всем наборе записи, для которых одинаковым входным полям соответствуют одинаковые (дубликаты) или разные (противоречия) выходные поля. На основании этой информации создаются два дополнительных логических поля – "Дубликат" и "Противоречие", принимающие значения "правда" или "ложь". В дополнительные числовые поля "Группа дубликатов" и "Группа противоречий" записываются номер группы дубликатов и группы противоречий, в которые попадает данная запись. Если запись не является дубликатом или противоречием, то соответствующее поле будет пустым.

Исходные данные

Рассмотрим механизм выявления дубликатов на примере данных файла "Anketa.txt". В этом файле находится информация об анкетных данных

граждан, участвующих в кредитовании. Попробуем вычислить присутствие дубликатов.

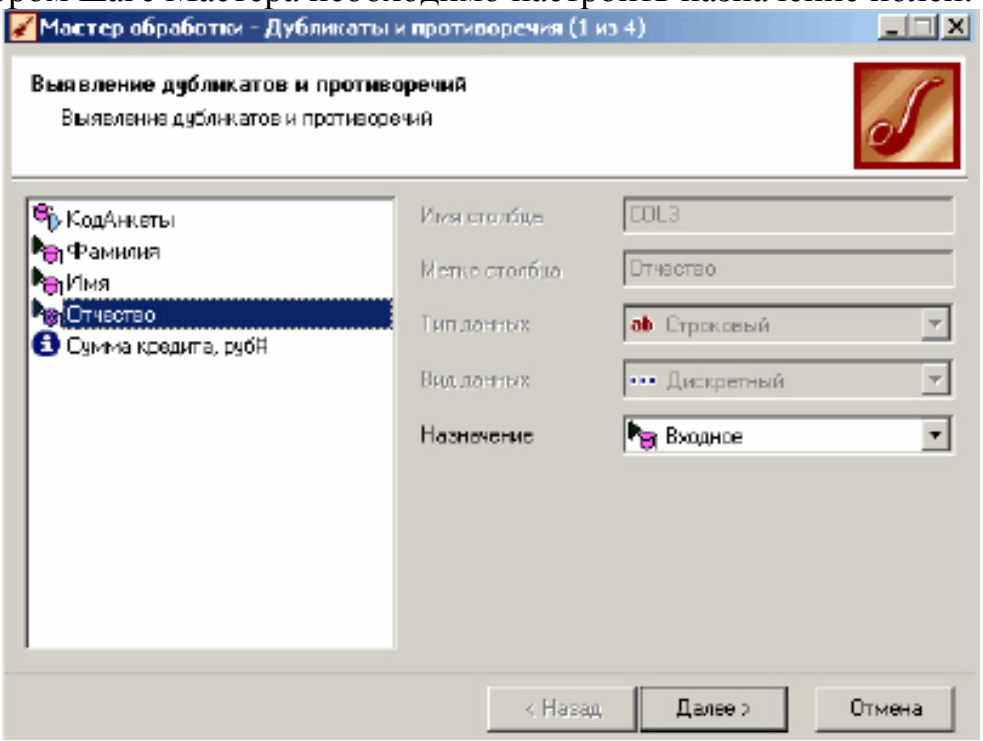
Импортируем данные из текстового файла и посмотрим их в виде таблицы.



Поиск дубликатов и противоречий

Для выявления дубликатов запустим Мастер обработки. В нем выберем тип обработки "Дубликаты и противоречия".

На втором шаге Мастера необходимо настроить назначение полей.

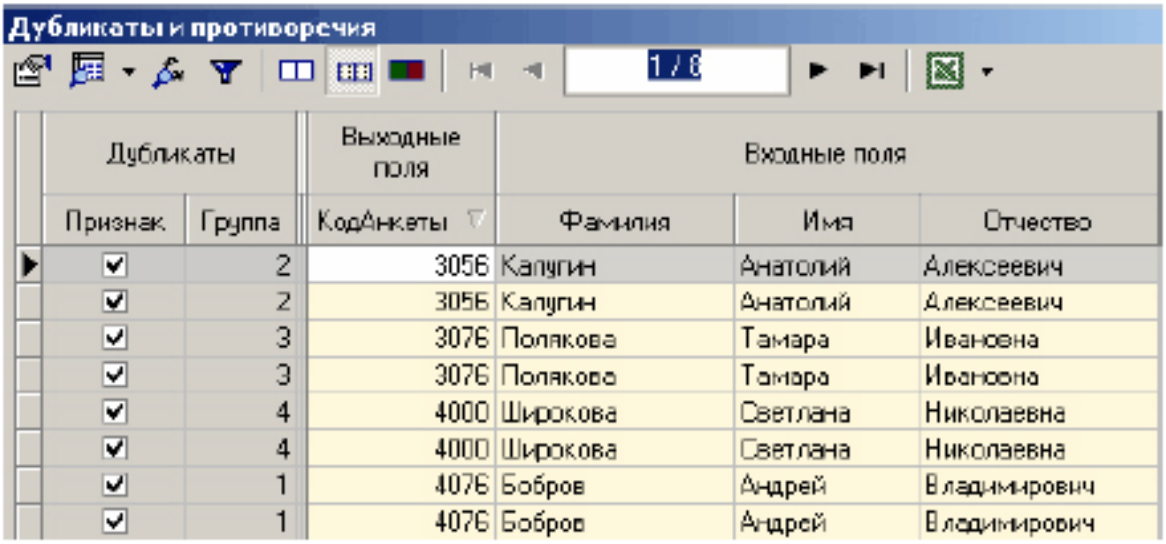


На следующем шаге необходимо запустить процесс обработки.

Результат

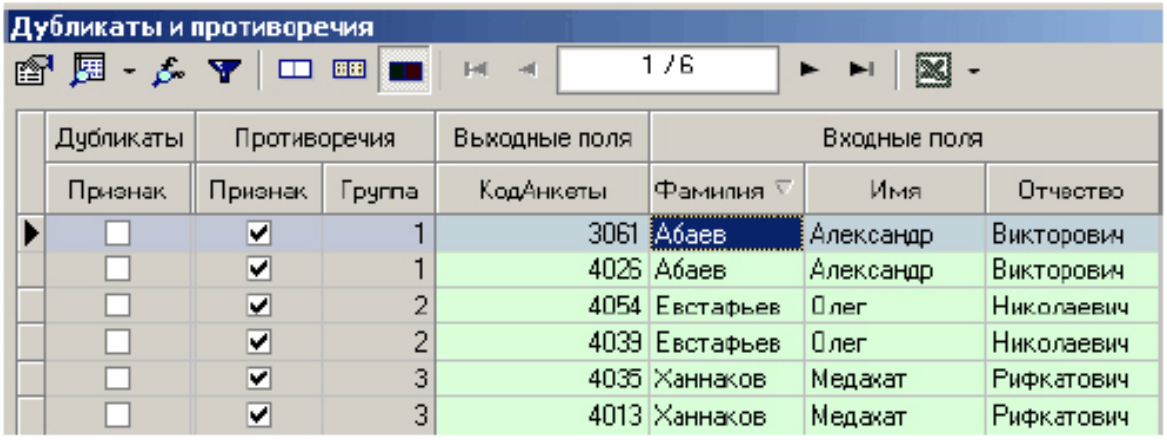
После завершения выявления дубликатов просмотрим результат в виде таблицы дубликатов и противоречий.

В первом случае видно, что существуют одинаковые строки, являющиеся дубликатами. Данный обработчик показывает дубликаты и их принадлежность к группам дубликатов.



Дубликаты		Выходные поля	Входные поля		
Признак	Группа	КодАнкеты	Фамилия	Имя	Отчество
☑	2	3056	Калугин	Анатолий	Алексеевич
☑	2	3056	Калугин	Анатолий	Алексеевич
☑	3	3076	Полякова	Тамара	Ивановна
☑	3	3076	Полякова	Тамара	Ивановна
☑	4	4000	Широкова	Светлана	Николаевна
☑	4	4000	Широкова	Светлана	Николаевна
☑	1	4076	Бобров	Андрей	Владимирович
☑	1	4076	Бобров	Андрей	Владимирович

Во втором случае видно, что при одинаковых "Фамилия", "Имя", "Отчество" оказываются различные Коды Анкет. В данном обработчике видно, у каких строк существуют противоречия и к какой группе они относятся.



Дубликаты	Противоречия		Выходные поля	Входные поля		
Признак	Признак	Группа	КодАнкеты	Фамилия	Имя	Отчество
☐	☑	1	3061	Абаев	Александр	Викторович
☐	☑	1	4026	Абаев	Александр	Викторович
☐	☑	2	4054	Евстафьев	Олег	Николаевич
☐	☑	2	4039	Евстафьев	Олег	Николаевич
☐	☑	3	4035	Ханнаков	Медахат	Рифкатович
☐	☑	3	4013	Ханнаков	Медахат	Рифкатович

Лабораторная работа:

1. Выполните **Парциальную предобработку** данных из файла "Trade.txt"
2. Выполните **Спектральную обработку** с данными файла "Trade.txt"
3. Выполните **Корреляционный анализ** данных из файла "region.txt"
4. Выполните **Выявления дубликатов** на примере данных файла "Anketa.txt".

Вопросы для проверки:

1. Назначения и ход выполнения **Парциальной обработки**
2. Назначение и ход выполнения **Спектральной обработки**
3. Назначение и ход выполнения **Корреляционного анализа**
4. Назначение и ход выполнения **Выявления дубликатов и противоречий**

Поиск дубликатов и противоречий

Для выявления дубликатов запустим Мастер обработки. В нем выберем тип обработки "Дубликаты и противоречия".

На втором шаге Мастера необходимо настроить назначение полей.

Поиск дубликатов и противоречий

Для выявления дубликатов запустим Мастер обработки. В не