ЛАБОРАТОРНАЯ РАБОТА 8. АНАЛИТИЧЕСКАЯ ПЛАТФОРМА DEDUCTOR

ЦЕЛЬ И ЗАДАЧИ РАБОТЫ, ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ЕЕ ВЫПОЛНЕНИЯ

Целью выполнения лабораторной работы является ознакомление с архитектурой, основными частями и пользовательским интерфейсом Deductor Studio, освоение и закрепление навыков применения факторного и корреляционного анализа, ознакомление с методами трансформации данных в среде Deductor Studio.

Основными задачами выполнения лабораторной работы являются:

1. получение навыков импорта данных,
2. изучение парциальной обработки,
3. изучение восстановления данных,
4. применение методов удаления аномалий в данных,
5. изучение спектральной обработки данных,
6. получение навыков удаления шумов в данных,
7. изучить факторный и корреляционный анализ,
8. получить навыки по работе с ними,
9. научиться понимать разницу между ними,
10. осознать области применения факторного и корреляционного анализа в data mining,
11. изучение способов разбиения данных,
12. применение методов квантования данных,
13. получения навыков фильтрации данных.

Результатами работы являются:

- Файл сценария Deductor, содержащий решение задачи по варианту
- Подготовленный отчет

ХАРАКТЕРИСТИКА ОБЪЕКТА ИЗУЧЕНИЯ, ИССЛЕДОВАНИЯ

Deductor Studio

Deductor Studio – программа, реализующая функции импорта, обработки, визуализации и экспорта данных. Deductor Studio может функционировать и без хранилища данных, получая информацию из любых других источников, но наиболее оптимальным является их совместное использование. В Deductor Studio включен полный набор механизмов, позволяющий получить информацию из произвольного источника данных, провести весь цикл обработки (очистку, трансформацию данных, построение моделей), отобразить полученные результаты наиболее удобным образом (OLAP, диаграммы, деревья…) и экспортировать результаты на сторону. Это полностью соответствует концепции извлечения знаний из баз данных (KDD).

Интерфейс Deductor Studio (Рис. 8.1) состоит из главного окна, внутри которого располагаются панели сценариев, отчетов, источников данных и результаты моделирования (таблицы, графики, кросс-диаграммы, правила и т.д.).

Для автоматизации получения данных из любого источника, предусмотренного в системе, следует использовать мастер импорта. На первом шаге мастера импорта открывается список всех предусмотренных в системе типов источников данных. Число шагов мастера импорта, а также набор настраиваемых параметров отличается для разных типов источников.

Мастер обработки предназначен для настройки всех параметров выбранного алгоритма.

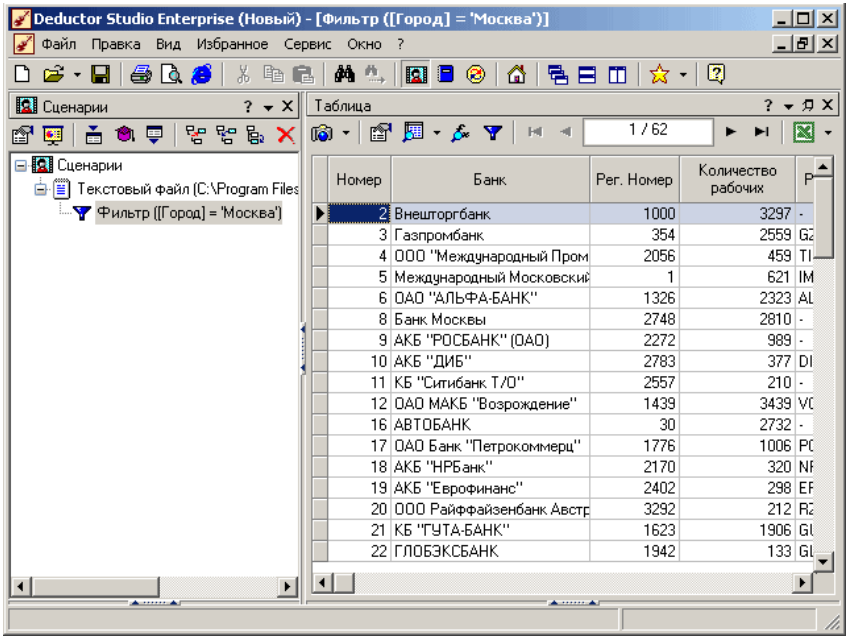


Рис. 8.1. Интерфейс Deductor Studio

Пошаговом режиме выбрать и настроить наиболее удобный способ представления данных можно с помощью мастера отображений. В зависимости от обработчика, в результате которого была получена ветвь сценария, список доступных для него видов отображений будет различным. Например, после построения деревьев решений их можно отобразить с помощью визуализаторов «Деревья решений» и «Правила». Эти способы отображения не доступны для других обработчиков.

Импорт данных является отправной точкой анализа данных. Импорт в Deductor может осуществляться из таких форматов хранения данных, как Excel, Access, MS SQL, Oracle, текстовый файл и прочих. Кроме того, имеется универсальный доступ к любому источнику данных посредством ADO или ODBC.

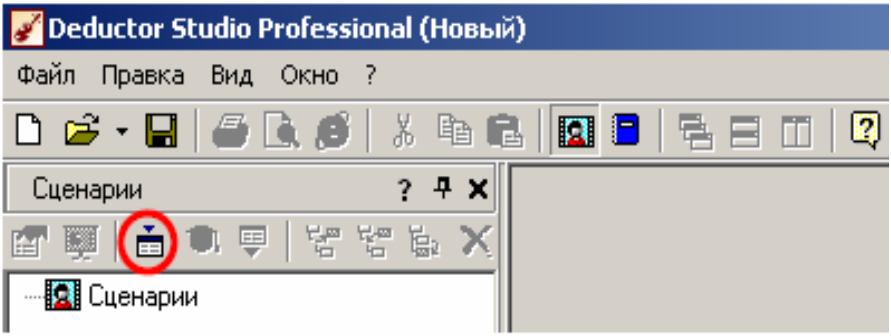


Рис.8.2 Импорт данных

Вид данных определяет – конечный ли это набор (дискретные) или бесконечный (непрерывные). Назначение столбцов определяет характер их использования в алгоритмах обработки (при импорте можно оставить значение по умолчанию).

Часто исходные данные являются недостаточно полными либо имею различные шумы и не годятся для анализа, а качество данных влияет на качество результатов. Так что вопрос подготовки данных для последующего анализа является очень важным. Обычно «сырые» данные содержат в себе различные шумы, за которыми трудно увидеть общую картину, а также аномалии – влияние случайно, либо редко происходивших событий. Очевидно, что влияние этих факторов на общую модель необходимо минимизировать, т.к. модель, учитывающая их, получится неадекватной.

Парциальная предобработка

Парциальная предобработка служит для восстановления пропущенных данных, редактирования аномальных значений и спектральной обработке данных (например, сглаживания данных). Именно этот шаг часто проводится в первую очередь.

Восстановление пропущенных данных

Часто бывает так, что в столбце некоторые данные отсутствуют в силу каких- либо причин (данные не известны, либо их забыли внести и т.п.). Обычно из–за этого пришлось бы убрать из обработки все строки, которые содержат пропущенные данные. Но механизмы Deductor Studio позволяют решить эту проблему. Один из шагов парциальной обработки как раз отвечает за восстановление пропущенных значений.

Если данные упорядочены (например, по времени), то рекомендуется в качестве восстановления пропущенных значений использовать аппроксимацию. Алгоритм сам подберет значение, которое должно стоять на месте пропущенного значения, основываясь на близлежащих данных. Если же данные не упорядочены, то следует использовать режим максимального правдоподобия, когда алгоритм подставляет вместо пропущенных данных наиболее вероятные значения, основываясь на всей выборке.

Удаление аномалий

Аномалии – это отклонения от нормального поведения чего-либо. Это может быть, например, резкое отклонение величины от ее ожидаемого значения.

Автоматическое редактирование аномальных значений осуществляется с использованием методов робастной фильтрации, в основе которых лежит использование робастных статистических оценок, таких, например, как медиана. При этом можно задать эмпирически подобранный критерий того, что считать аномалией. Например, задание в качестве степени подавления аномальных данных

значения «слабая» означает наиболее терпимое отношение к величине допустимых выбросов.

По существу аномалии вообще не должны оказывать никакого влияния на результат. Если же они присутствуют при построении модели, то оказывают на нее весьма большое влияние. Т.е. предварительно их необходимо устранить. Также они портят статистическую картину распределения данных. Данные с аномалиями, а также гистограмма их распределения представлены на рис.8.3:

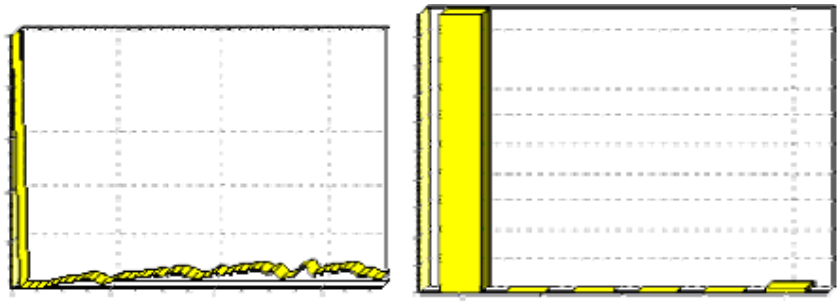


Рис.8.3 Данные с аномалиями и гистограмма распределения

Очевидно, что аномалии не позволяют определить как характер самих данных, так и статистическую картину. Данные после устранения аномалий представлены на рис.8.4.

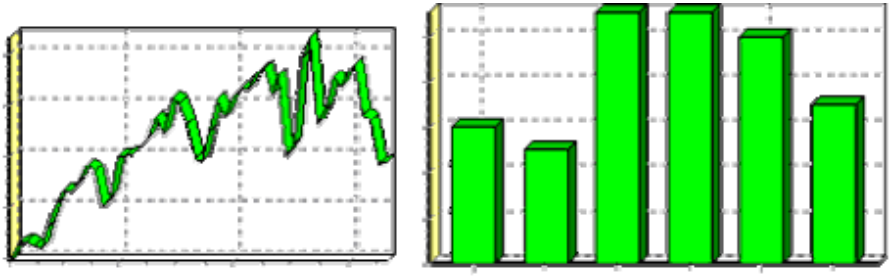


Рис.8.4 Результат удаления аномалий

Спектральная обработка.

Сглаживание данных применяется для удаления шумов из исходного набора. Платформа Deductor Studio предлагает несколько видов спектральной обработки: сглаживание данных путем указания полосы пропускания, вычитание шума путем указания степени вычитания шума и вейвлет преобразование путем указания глубины разложения и порядка вейвлета.

Удаление шумов

Шумы в данных не только скрывают общую тенденцию, но и проявляют себя при построении модели прогноза. Из-за них модель может получиться с плохими обобщающими качествами.

Спектральная обработка позволяет сделать это с помощью указания для этих полей в качестве типа обработки «Вычитание шума». Настройки обладают определенной гибкостью. Так, существует большая, средняя и малая степень вычитания шума. Аналитик может подобрать степень, устраивающую его.

В некоторых случаях неплохие результаты удаления шумов дает вейвлет преобразование.

ЗАДАЧИ И ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ

Содержание работы

Используя текстовый редактор "блокнот" создать файл «TestForPPP.txt» (рис.8.5), содержащий такие столбцы, как «Аргумент», «Синус», «Аномалии», «Больше шумы», «Средние шумы», «Малые шумы». Разделителем между столбцами является знак табуляции. Столбцу «Аргумент» присваиваются значения от 0 до 2,96 с шагом 0,02. В столбце «Синус» принимаются значения синуса (9 знаков после запятой). При любых двадцати значениях аргумента, ввод данных в значениях синуса пропустить. Значения столбца «Аномалии» равны значениям столбца «Синус», но не имеют пропущенных данных, однако 10 значений резкое отклоняются от истинного значения синуса аргумента.

Значения столбцам «Больше шумы», «Средние шумы», «Малые шумы» имеют значения близкие к значению синуса аргумента, но имеют некоторое отклонение (дисперсию и выбираются из промежутка -1,5 до 1,5)(рис.5).

Выполнить импорт данных, созданного файла, обработку данных, восстановить пропущенные значения синуса, выполнить парциальную обработку, удалить аномалии и шумы.

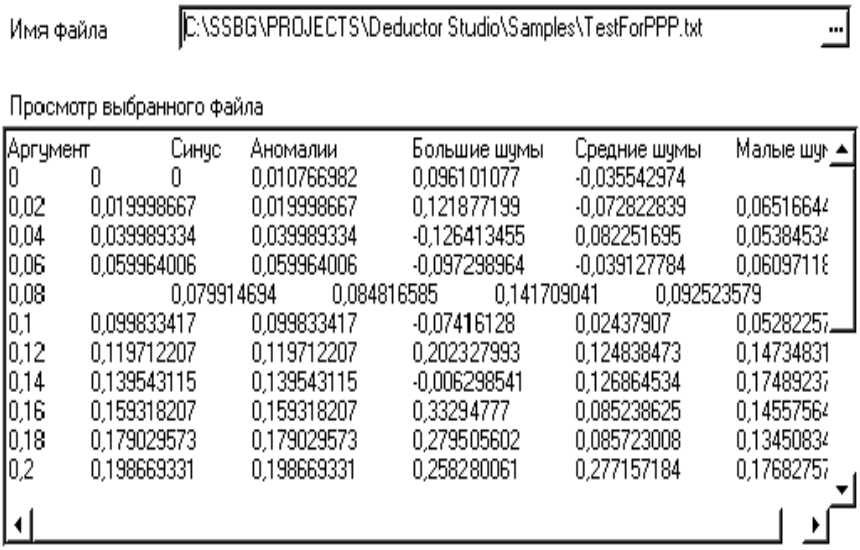
Аргумент	Синус	Аномалии	Большие шумы	Средние шумы	Малые шумы
. . . . .					
0,14	0,139543115	0,139543115	-0,006298541	0,126864534	0,174892379
0,16	0,159318207	0,159318207	0,33294777	0,085238625	0,145575649
0,18	0,179029573	0,179029573	0,279505602	0,085723008	0,134508346
0,2	0,198669331	0,198669331	0,258280061	0,277157184	0,176827573
0,22	0,218229623	0,5	0,139744277	0,28969863	0,231458805
0,24	0,237702626	0,237702626	0,317646146	0,177087062	0,282627785
0,26	0,257080552	0,257080552	0,266913669	0,19389017	0,288192043
0,28	0,276355649	0,276355649	0,131932825	0,325878695	0,243445006
0,3		0,295520207	0,403496144	0,30362014	0,30123014
0,32		0,314566561	0,430780976	0,306273759	0,312952364
. . . . .					

Рис. 8.5 Пример заполнения файла «TestForPPP.txt»

Порядок выполнения работы

Импорт осуществляется путем вызова мастера импорта на панели «Сценарии»

После запуска мастера импорта укажем тип импорта «Текстовый файл с разделителями» и перейдем к настройке импорта. Укажем имя файла, из которого необходимо получить данные (пример для парциальной обработки). В окне просмотра выбранного файла можно увидеть содержание данного файла(рис.8.6).



Аргумент		Синус	Аномалии	Большие шумы	Средние шумы	Малые шум
0	0	0	0,010766982	0,096101077	-0,035542974	
0,02	0,019998667		0,019998667	0,121877199	-0,072822839	0,0651664
0,04	0,039989334		0,039989334	-0,126413455	0,082251695	0,0538453
0,06	0,059964006		0,059964006	-0,097298964	-0,039127784	0,0609711
0,08	0,079914694		0,084816585	0,141709041	0,092523579	
0,1	0,099833417		0,099833417	-0,07416128	0,02437907	0,0528225
0,12	0,119712207		0,119712207	0,202327993	0,124838473	0,14734831
0,14	0,139543115		0,139543115	-0,006298541	0,126864534	0,1748923
0,16	0,159318207		0,159318207	0,33294777	0,085238625	0,1455756
0,18	0,179029573		0,179029573	0,279505602	0,085723008	0,1345083
0,2	0,198669331		0,198669331	0,258280061	0,277157184	0,1768275

Рис.8.6 Окно просмотра

Далее перейдем к настройке параметров импорта (рис.8.7). На этой странице мастера предоставляется возможность указать, с какой строки следует начать импорт, указать, то, что первая строка является заголовком, возможность добавить первичный ключ. Указать, что является символом-разделителем столбцов, а также указать ограничитель строк, разделитель целой и дробной части вещественного числа, разделитель компонентов даты и ее формат.

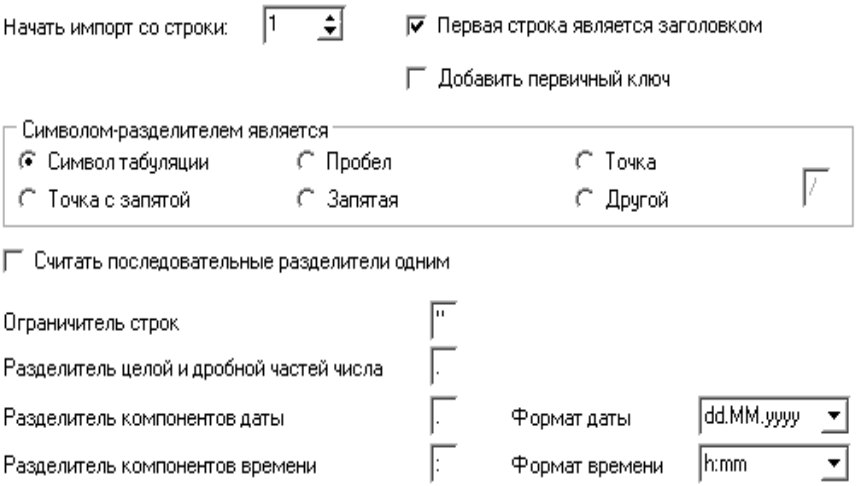


Рис.8.7 Настройка параметров импорта

В данном случае параметры по умолчанию на этой странице мастера установлены правильно, а именно: начать импорт с первой строки, первая строка является заголовком, разделителем между столбцами является знак табуляции, разделителем целой и дробной частей является запятая.

На следующем шаге мастера предоставляется возможность настроить имя, название (метку), размер, тип данных, вид данных и назначение. Некоторые свойства (например, тип данных) можно задавать для выделенного набора столбцов(рис.8.8).

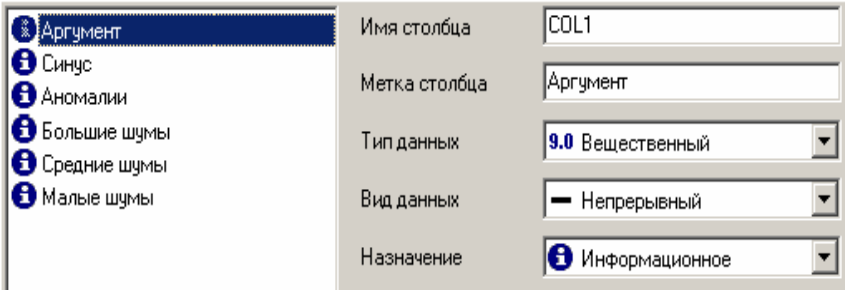


Рис.8.8 Настройка свойств столбцов

Для правильного импорта данных необходимо изменить тип данных у первых трех столбцов («АРГУМЕНТ», «СИНУС», «АНОМАЛИИ»).

Тип данных по умолчанию неверный, поскольку программа определяет его, основываясь на значениях первой строки данных. В данном случае там находятся нули – целые числа. Поэтому программа определила, что столбец содержит целочисленные значения. Выделим их с помощью мыши и укажем им тип данных – «Вещественный». Далее осталось только выполнить импорт данных, нажав на кнопку «Пуск» на следующем шаге мастера импорта.

После импорта данных на следующем шаге мастера необходимо выбрать способ отображения данных (рис.8.9). В данном случае самым информативным является диаграмма, выберем ее.

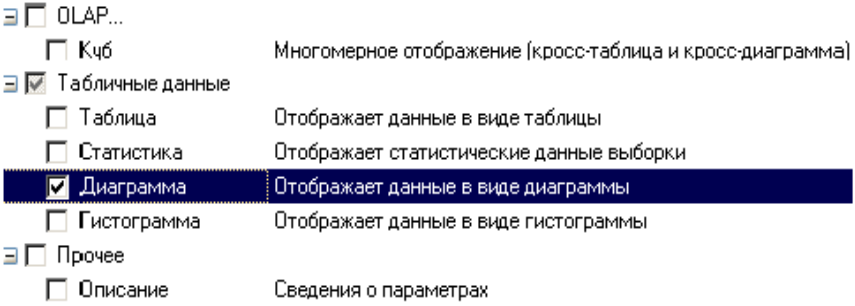


Рис.8.9 Способ отображения данных

От того, какие способы отображения будут выбраны на этом этапе, зависят последующие шаги мастера. В данном случае необходимо настроить, какие столбцы диаграммы следует отображать и как именно.

Выберем для отображения поле «СИНУС» (рис.8.10) и тип диаграммы «Линии».

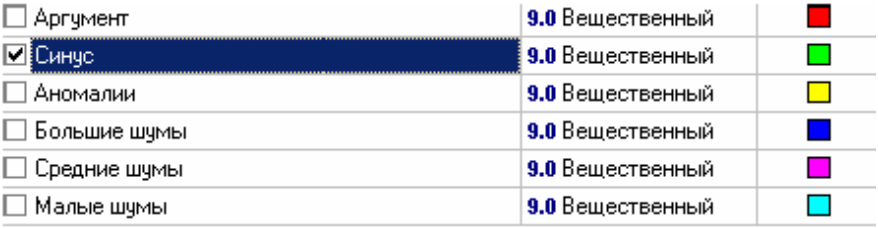


Рис.8.10 Настройка способа отображения

На последнем шаге мастера необходимо указать название ветки в дереве сценариев. На этом работа мастера импорта заканчивается.

Теперь в дереве сценариев появится новый узел с необходимыми данными. В главном окне программы представлены все выбранные отображения данных этого узла. В данном случае только диаграмма(рис.8.11)

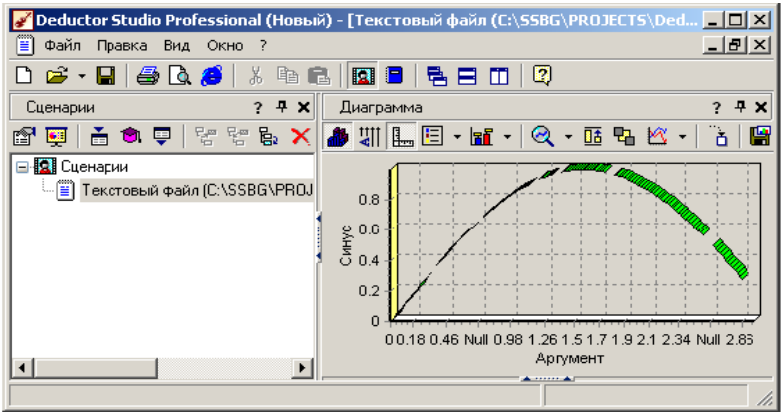


Рис.8.11 Главное окно программы

Далее сделаем обработку данных из файла «TestForPPP.txt». Он содержит таблицу со следующими полями: «АРГУМЕНТ» – аргумент, «СИНУС» – значения синуса аргумента (некоторые значения пустые), «АНОМАЛИИ» – синус с выбросами, «БОЛЬШИЕ ШУМЫ» – значения синуса с большими шумами, «СРЕДНИЕ ШУМЫ» – значения синуса со средними шумами, «МАЛЫЕ ШУМЫ» – значения синуса с малыми шумами. Все данные можно увидеть на диаграмме после импорта из текстового файла(рис.8.12).

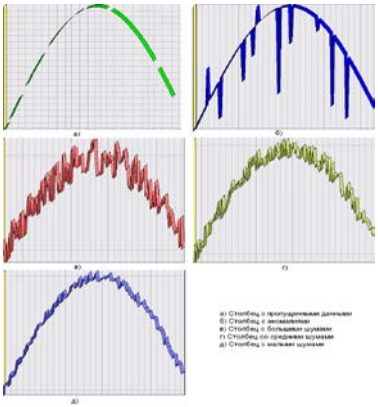


Рис.8.12 Диаграммы

Импортировав файл можно увидеть, что в столбце «СИНУС» содержатся пустые значения. На диаграмме выше видно, что некоторые значения синуса пропущены. Для дальнейшей обработки необходимо их восстановить. Для этого следует запустить мастер парциальной обработки.

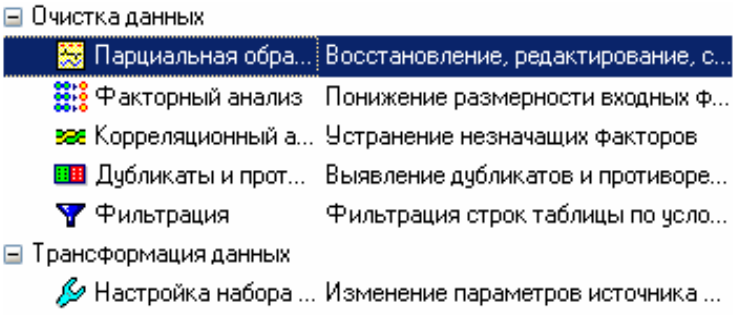


Рис.8.13 Запуск мастера парциальной обработки

Поскольку данные в исходном наборе упорядочены, на следующем шаге мастера обработки выделим поле «СИНУС» и укажем для него тип обработки «Аппроксимация» (рис.8.14). Так как в данном случае больше ничего не требуется, то остальные параметры обработки оставляем отключенными. Перейдя на страницу запуска процесса обработки, выполняем ее, нажав на пуск, и далее выбираем тип визуализации обработанных данных (как в примере импорта).

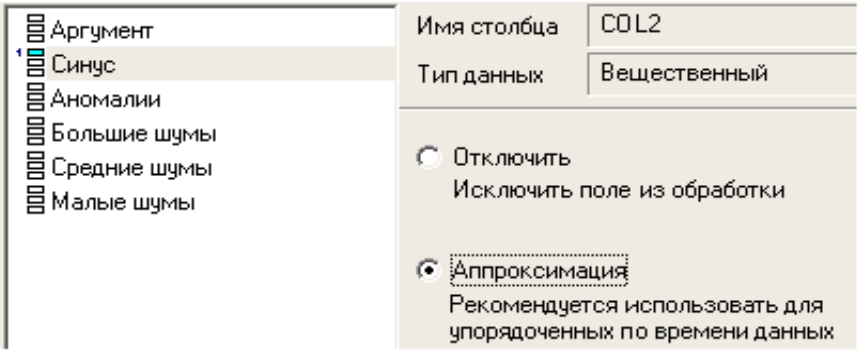


Рис.8.14 Окно мастера обработки

После выполнения процесса обработки, как видно из рисунка 8.15, на диаграмме пропуски в данных исчезли, что и было необходимо сделать.

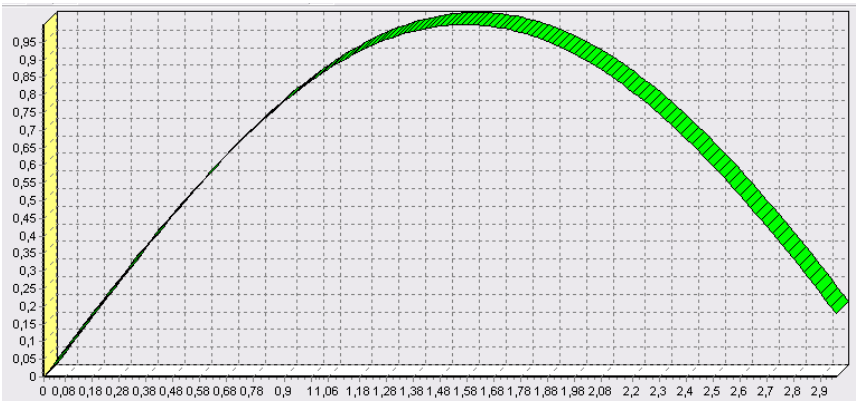


Рис.8.15 Диаграмма после процесса обработки

Далее удалим аномалий из поля «АНОМАЛИИ» импортированной таблицы.

В мастере парциальной предобработки на третьем шаге выбираем поле «АНОМАЛИИ» и указываем ему тип обработки «Удаления аномальных явлений», степень подавления «Большая» (рис.8.16).

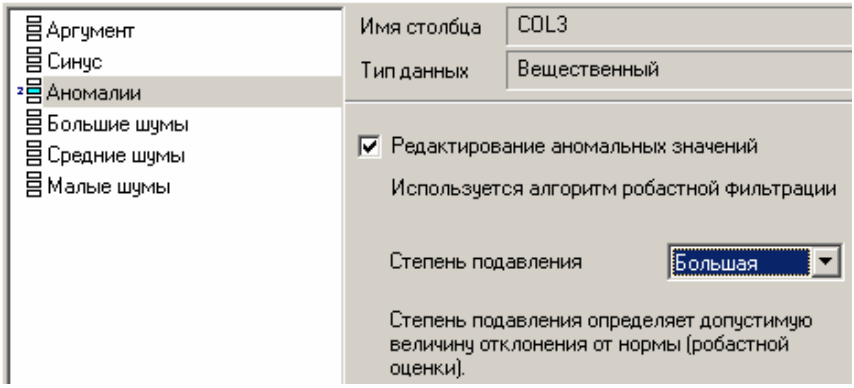


Рис.8.16 Окно мастера обработки

Так как больше никаких обработок не планировалось, то переходим на шаг запуска процесса обработки и нажимаем «Пуск».

После выполнения процесса обработки на диаграмме видно, что выбросы исчезли, остались лишь небольшие возмущения, которые легко сгладить при помощи спектральной обработки.

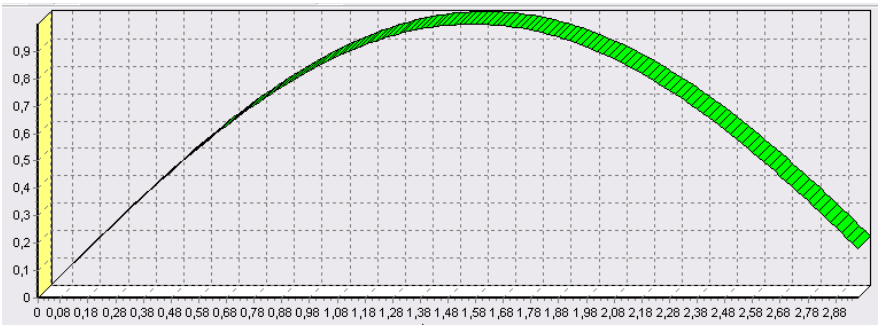


Рис.8.17 Итоговая диаграмма

Как видно на рисунке 8.17, аномалии были устранены, однако небольшие возмущения остались. Сгладим их при помощи парциальной обработки. Для этого после удаления аномалий вновь запустим мастер парциальной обработки. В нем на четвертом шаге выберем поле «АНОМАЛИИ» и укажем ему тип обработки «Вейвлет преобразование» с параметрами по умолчанию (глубина разложения 3, порядок вейвлета 6).

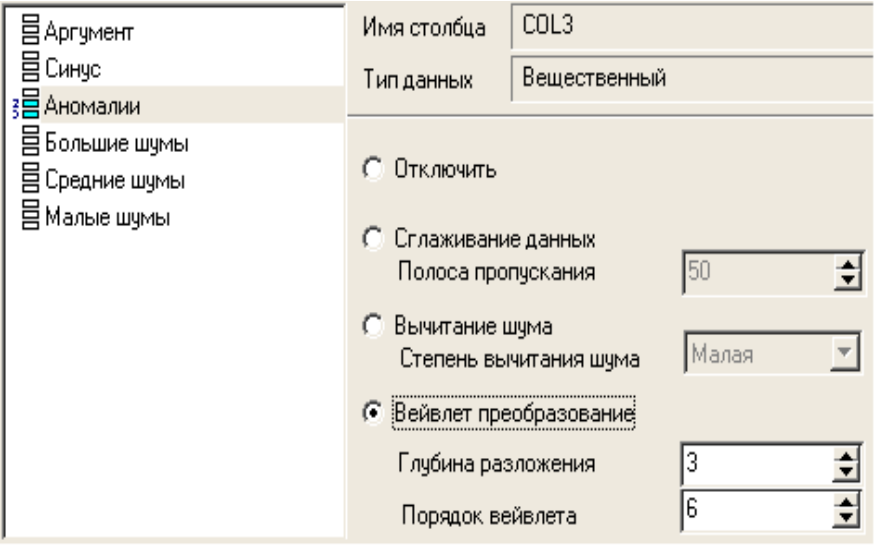


Рис.8.18 Окно мастера обработки

Так как больше ничего не планировалось, то перейдем с шагу запуска процесса обработки и выполним ее. В качестве визуализатора укажем диаграмму.

После обработки можно убедиться на диаграмме в отсутствии выбросов и сравнить результат с эталонным значением синуса (столбец «СИНУС»). На рисунке 8.19 зеленый (светлый) график – значения синуса, синий (темный) – значения сглаженного синуса после устранения аномалий.

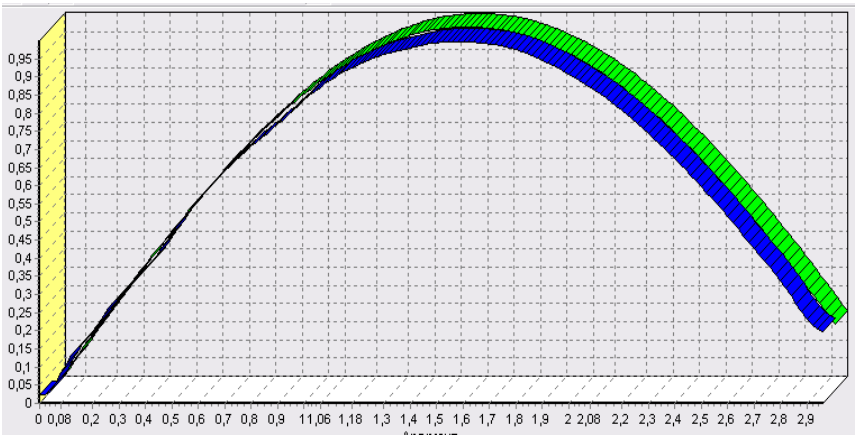


Рис.8.19 Диаграмма после обработки

В примере по парциальной обработке, как было показано ранее, есть 3 столбца с шумами: «БОЛЬШИЕ ШУМЫ», «СРЕДНИЕ ШУМЫ», и «МАЛЫЕ ШУМЫ» - соответственно синус с большими, средними и малыми шумами. Ясно, что для дальнейшей работы с данными эти шумы необходимо устранить.

Таким образом, в мастере парциальной обработки на четвертом шаге выберем по очереди поля «БОЛЬШИЕ ШУМЫ», «СРЕДНИЕ ШУМЫ» и «МАЛЫЕ ШУМЫ», зададим тип обработки «Вычитание шума» и укажем степень подавления – «большая», «средняя» и «малая» соответственно. После выполнения обработки на диаграмме можно просмотреть полученные результаты(рис.8.20).

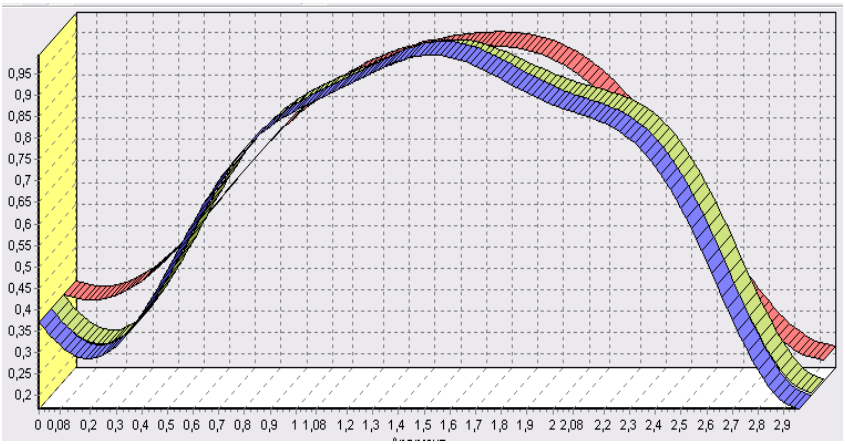


Рис.8.20 Результат парциальной обработки

Теперь удалим шумы с помощью вейвлет преобразования. В мастере парциальной обработки выберем поля «БОЛЬШИЕ ШУМЫ», «СРЕДНИЕ ШУМЫ» и «МАЛЫЕ ШУМЫ», укажем тип обработки «Вейвлет преобразование», оставив параметры обработки по умолчанию (глубина разложения – 3, порядок вейвлета – 6). На диаграмме можно убедиться в том, что данные сгладились(рис.8.21). Повысить качество сглаживания шумов таким способом можно, путем подбора удовлетворительных параметров обработки.

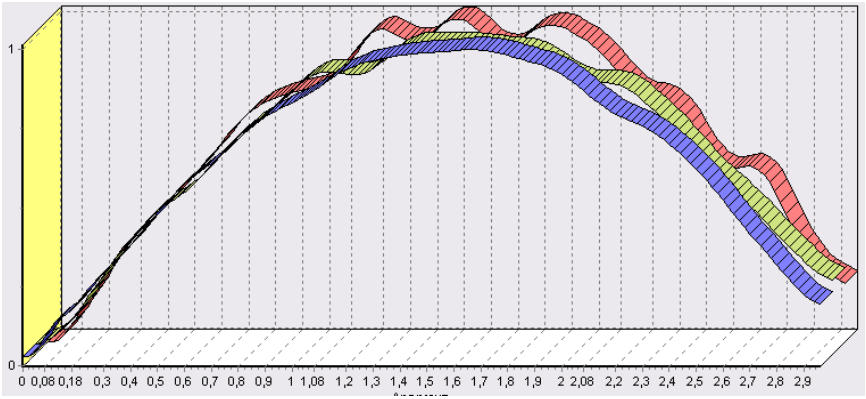


Рис.8.21 Результат удаления шума

ЗАДАНИЕ НА ЛАБОРАТОРНУЮ РАБОТУ

Создать сценарий в среде Deductor Studio согласно варианту, полученному у преподавателя.

ТРЕБОВАНИЯ К РЕАЛИЗАЦИИ

Задание выполняется согласно варианту в среде Deductor Studio. По завершении готовится отчет.

ВАРИАНТЫ ЗАДАНИЙ

1. Создать файл с данными аналогично файлу, приведенному в лабораторной работе. В качестве основной функции использовать *cos(x)*. Выполнить импорт данных, созданного файла, обработку данных, восстановить пропущенные значения синуса, выполнить парциальную обработку, удалить аномалии и шумы.

2. Создать файл с данными аналогично файлу, приведенному в лабораторной работе. В качестве основной функции использовать: *cos(x)*sin(x)*. Выполнить импорт данных, созданного файла, обработку данных, восстановить пропущенные значения функции, выполнить парциальную обработку, удалить аномалии и шумы.

3. Создать файл с данными аналогично файлу, приведенному в лабораторной работе. В качестве основной функции использовать: $cos^2(x)$. Выполнить импорт данных, созданного файла, обработку данных, восстановить пропущенные значения функции, выполнить парциальную обработку, удалить аномалии и шумы.\

4. Создать файл с данными аналогично файлу, приведенному в лабораторной работе. В качестве основной функции использовать: *sin(2x)*. Выполнить импорт данных, созданного файла, обработку данных, восстановить пропущенные значения функции, выполнить парциальную обработку, удалить аномалии и шумы.

5. Создать файл с данными аналогично файлу, приведенному в лабораторной работе. В качестве основной функции использовать: *cos(2x)*. Выполнить импорт данных, созданного файла, обработку данных, восстановить пропущенные значения функции, выполнить парциальную обработку, удалить аномалии и шумы.

6. Создать файл с данными аналогично файлу, приведенному в лабораторной работе. В качестве основной функции использовать: x^2. Выполнить импорт данных, созданного файла, обработку данных, восстановить пропущенные значения функции, выполнить парциальную обработку, удалить аномалии и шумы.

7. Создать файл с данными аналогично файлу, приведенному в лабораторной работе. В качестве основной функции использовать: *x*sin(x)*. Добавить аномальные выбросы в столбец с шумами. Выполнить импорт данных, созданного файла, обработку данных, восстановить пропущенные значения функции, выполнить парциальную обработку, удалить аномалии и шумы.

8. Создать файл с данными аналогично файлу, приведенному в лабораторной работе. В качестве основной функции использовать: *cos(x)*. Добавить пропуски данных в столбец с аномалиями. Выполнить импорт данных, созданного файла, обработку данных, восстановить пропущенные значения функции, выполнить парциальную обработку, удалить аномалии и шумы.

9. Создать файл с данными аналогично файлу, приведенному в лабораторной работе. В качестве основной функции использовать: *cos(x)*sin(x)*. Выполнить импорт данных, созданного файла, обработку данных, восстановить пропущенные значения функции, выполнить парциальную обработку, удалить аномалии и шумы. Сравнить результаты удаления шумов при задании различных типов обработки данных в мастере парциальной обработки.

10. Создать файл с данными аналогично файлу, приведенному в лабораторной работе. В качестве основной функции использовать: *cos(x)+sin(x)*. Выполнить импорт данных, созданного файла, обработку данных, восстановить пропущенные значения функции, выполнить парциальную обработку, удалить аномалии и шумы. Сравнить результаты удаления аномалий при задании различной степени подавления выбросов.

КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ

1. Опишите назначение мастера импорта программы Deductor Studio.
2. Опишите назначение мастера обработки программы Deductor Studio.
3. Опишите назначение мастер отображений программы Deductor Studio.
4. Раскройте основные цели проведения подготовки данных для анализа.
5. Дайте определение шумов и аномалий в данных.
6. Перечислите методы устранения шумов в системе Deductor.
7. Перечислите методы устранения аномалий данных в системе Deductor.
8. Опишите назначение парциальной предобработки.
9. Опишите назначение спектральной обработки.
10. Перечислите виды спектральной обработки в системе Deductor.

ХАРАКТЕРИСТИКА ОБЪЕКТА ИЗУЧЕНИЯ, ИССЛЕДОВАНИЯ

Факторный анализ

Факторный анализ - группа методов многомерного статистического анализа, которые позволяют представить в компактной форме обобщенную информацию о структуре связей между наблюдаемыми признаками изучаемого объекта на основе выделения некоторых непосредственно не наблюдаемых факторов. Факторный анализ служит для понижения размерности пространства входных факторов. Обработку можно выполнять как в автоматическом режиме (с указанием порога значимости), так и самостоятельно (основываясь на значениях матрицы значимости).

Первым этапом факторного анализа является выбор новых признаков, которые являются линейными комбинациями прежних и «вбирают» в себя большую часть общей изменчивости входных факторов. Поэтому они содержат большую часть информации, заключенной в первоначальных данных. В обработчике «Факторный анализ» это осуществляется с помощью метода главных компонент. Этот метод сводится к выбору новой ортогональной системы координат в пространстве наблюдений. В качестве первой главной компоненты избирают направление, вдоль которого массив данных имеет наибольший разброс. Выбор каждой последующей главной компоненты происходит так, чтобы разброс данных вдоль нее был максимальным и чтобы эта главная компонента была ортогональна другим главным компонентам, выбранным прежде.

Корреляционный анализ

Корреляционный анализ- совокупность основанных на математической теории корреляции методов обнаружения корреляционной зависимости между двумя случайными признаками или факторами. 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

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 степень взаимозависимости между входным и выходным факторами меньше порога значимости, то соответствующий фактор отбрасывается как незначащий.

ЗАДАЧИ И ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ

Содержание работы

В блокноте создать файл «TestForCPP.txt», содержащий такие столбцы, как «Аргумент», «Фактор1», «Фактор2», «Фактор3», «Результат1», «Результат2» (рис.8.22). Разделителем между столбцами является знак табуляции. Столбец «Аргумент» заполняется значениями в диапазоне от 0 до 6,25 с шагом 0,05. Столбцы «Фактор1», «Фактор2», «Фактор3» являются входными значениями и задаются в диапазоне от -1 до 1 (в любом порядке). Столбцы «Результат1», «Результат2» является выходными значениями и принимаются равными значениям столбцов «Фактор1», «Фактор2» соответственно.

Значения столбцам нужно задавать такие, чтобы получилась циклическая функция.

Выполнить обработку данных факторным и корреляционным анализом.

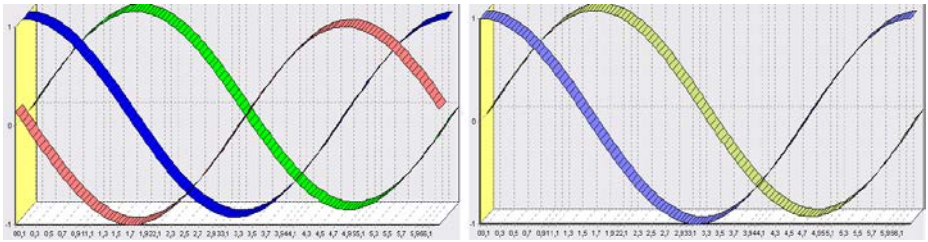
Аргумент	Фактор1	Фактор2	Фактор3	Результат1	Результат2
0,00	0,00	1,00	0,14	0,00	1,00
0,05	0,05	1,00	0,09	0,05	1,00
0,10	0,10	1,00	0,04	0,10	1,00
0,15	0,15	0,99	-0,01	0,15	0,99
0,20	0,20	0,98	-0,06	0,20	0,98
0,25	0,25	0,97	-0,11	0,25	0,97
0,30	0,30	0,96	-0,16	0,30	0,96
0,35	0,34	0,94	-0,21	0,34	0,94
0,40	0,39	0,92	-0,26	0,39	0,92
0,45	0,43	0,90	-0,30	0,43	0,90
0,50	0,48	0,88	-0,35	0,48	0,88
0,55	0,52	0,85	-0,40	0,52	0,85
0,60	0,56	0,83	-0,44	0,56	0,83
0,65	0,61	0,80	-0,49	0,61	0,80

Рис.8.22 Пример заполнения файла «TestForCPP.txt»

Порядок выполнения работы

Выполним обработку данных из файла «TestForCPP.txt» при помощи факторного анализа. Он содержит таблицу со следующими полями: «АРГУМЕНТ» – аргумент, «ФАКТОР1», «ФАКТОР2», «ФАКТОР3» –

входные значения, «РЕЗУЛЬТАТ1», «РЕЗУЛЬТАТ2» – выходные значения (рис.8.23).



а) Входные факторы б) Выходы

Рис.8.23 Данные для обработки

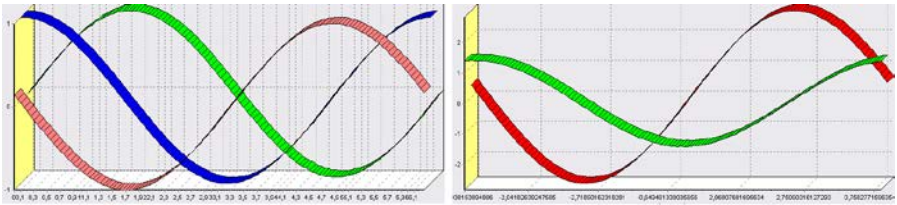
В мастере факторного анализа зададим «ФАКТОР1», «ФАКТОР2», «ФАКТОР3» входными полями, «РЕЗУЛЬТАТ1», «РЕЗУЛЬТАТ2» - выходными, а поле «АРГУМЕНТ» – непригодным.
Следующий шаг предлагает запустить процесс понижения размерности пространства входных факторов. После завершения процесса на следующем шаге выбираем, какие из полученных в результате обработки факторы оставить для дальнейшей работы (рис.3). Это делается путем указания необходимого порога значимости (по умолчанию порог значимости равен 90%, не будем его менять).

Главные компоненты	Собственное значение	Вклад в результат	Суммарный вклад
☑ Значение 1	2.000	66.66 %	66.66 %
☑ Значение 2	1.000	33.34 %	100.00 %
☐ Значение 3	0.000	00.00 %	

Порог значимости (%) 90

Рис.8.24 Указание порога значимости

Теперь необходимо перейти на следующий шаг и выбрать способ визуализации. Просмотрим результаты на диаграмме, изображенной на рисунке 8.25.



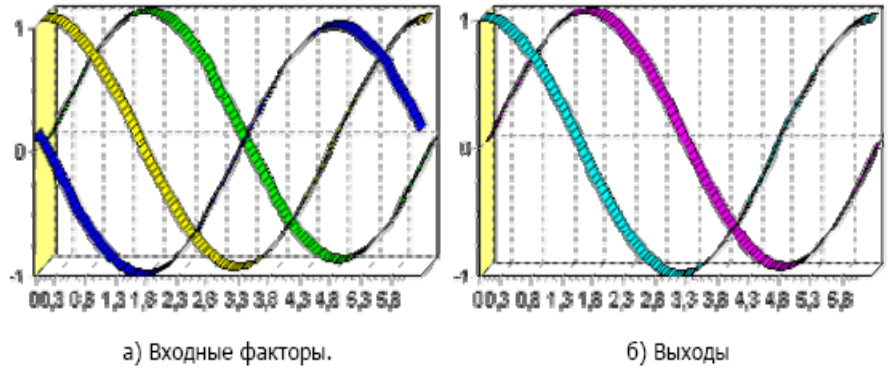
а) Исходные входные факторы б) Полученные входные факторы

Рис.8.25 Результаты факторного анализа

После обработки в наборе данных вместо трех исходных входных полей появились два новых поля – «Фактор1» и «Фактор2» – это результат понижения размерности (было 3 входных фактора, стало 2). На диаграмме видно, что «Фактор2» – близок к полю «ФАКТОР3», следовательно, «Фактор1» – это преобразованные факторы «ФАКТОР1» и «ФАКТОР2».

Далее выполним обработку данных из файла «TestForCPP.txt» при помощи корреляционного анализа. Он содержит таблицу со следующими полями: «АРГУМЕНТ» – аргумент, «ФАКТОР1», «ФАКТОР2», «ФАКТОР3» – входные значения, «РЕЗУЛЬТАТ1», «РЕЗУЛЬТАТ2» – выходные значения (рис.8.26).

Определим степень влияния входных факторов на один из выходов – «РЕЗУЛЬТАТ2» и оставим только значимые факторы.



а) Входные факторы. б) Выходы

Рис.8.26 Данные для обработки

В мастере корреляционного анализа зададим «ФАКТОР1», «ФАКТОР2», «ФАКТОР3» входными полями, «РЕЗУЛЬТАТ2» -

выходными, а поля «АРГУМЕНТ» и «РЕЗУЛЬТАТ1» – информационным.

На следующем шаге запускаем процесс корреляционного анализа. После завершения процесса выбираем, 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). Из рассчитанной матрицы ковариации видно, что выходное поле «РЕЗУЛЬТАТ2» напрямую зависит от поля «ФАКТОР2» (вообще, значение коэффициента, равное 1.000 говорит о том, что эти поля идентичны), и в меньшей степени от остальных факторов. В данном случае без потери полезной информации можно исключить из дальнейшего рассмотрения «Фактор1» и «Фактор3»

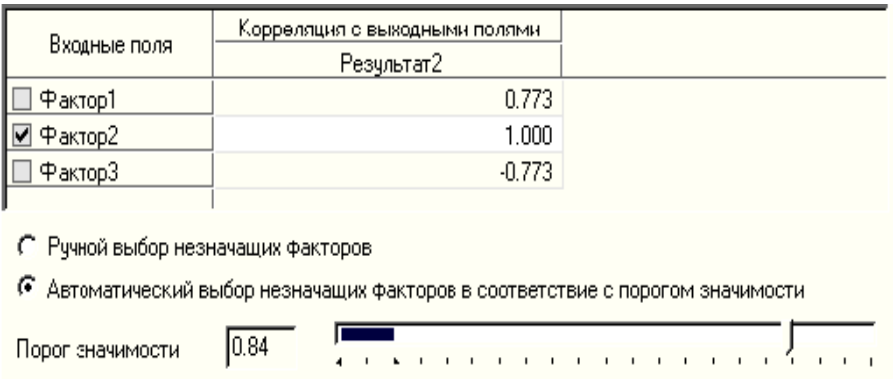


Рис.8.27 Указание порога значимости

Теперь необходимо перейти на следующий шаг и выбрать способ визуализации. Просмотрим результаты на диаграмме (например, можно убедиться в идентичности полей «Фактор2» и «Результат2»). Таким образом, корреляционный анализ позволил проанализировать влияние входных факторов на результат и исключить незначащие факторы из дальнейшего анализа.

ЗАДАНИЕ НА ЛАБОРАТОРНУЮ РАБОТУ

Выполнить обработку данных факторным и корреляционным анализом согласно варианту

ТРЕБОВАНИЯ К РЕАЛИЗАЦИИ

Задание выполняется согласно варианту в среде Deductor studio. По завершении готовится отчет.

ВАРИАНТЫ ЗАДАНИЙ

1. Исследовать зависимость прибыли от следующих показателей: цена, себестоимость, номенклатура, количество проданного товара

2. На n предприятиях была проанализирована среднемесячная заработная плата и количество уволившихся сотрудников. Необходимо определить зависимость числа уволившихся сотрудников от средней зарплаты.

3. Определить, есть ли взаимосвязь между временем работы станка и стоимостью его обслуживания

4. Определить зависимость объема продаж продукции от цены за единицу товара

5. Определить зависимости годового объема производства Y от основных фондов X по n предприятиям

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Дайте определение факторного анализа.
2. Опишите назначение факторного анализа при обработке и анализе данных.
3. Дайте определение корреляционного анализа.
4. Опишите назначение корреляционного анализа при обработке и анализе данных.
5. Раскройте сущность критерия принятия решения об исключении фактора при корреляционном анализе.
6. Приведите примеры других методов статистического анализа.
7. Дайте определение матрицы ковариации.

ХАРАКТЕРИСТИКА ОБЪЕКТА ИЗУЧЕНИЯ, ИССЛЕДОВАНИЯ

Трансформация данных

Трансформация данных – перенос и преобразование проводится на основе правил трансформации, сопоставляющих аналитику двух "учетных зон" и определяющих критерии передачи данных.

Разбиение данных на группы. Часто для проведения анализа или построения модели прогноза приходится разбивать данные на группы, исходя из определенных критериев. В первом случае такая необходимость возникает, если аналитик желает просмотреть, к примеру, информацию не по всей совокупности данных, а по определенным группам (например, какую сумму кредита берут на те или иные цели, либо кредиторы того или иного возраста). Во втором случае (прогнозирование) аналитику необходимо учитывать тот факт, что определенные группы (в данном случае группы кредиторов) ведут себя по разному, и что модель прогноза, построенная на всех данных не будет учитывать нюансов, возникающих в этих группах. Т.е. лучше построить несколько моделей прогноза, например, в зависимости от суммовой группы кредита и строить прогноз на них, нежели построить одну модель прогноза. Исходя из этого и не только, в Deductor Studio предоставляется широкий набор инструментов, тем или иным способом позволяющие разбивать исходные данные на группы, группировать любым способом всевозможные показатели и т.п.

Разбиение даты (по неделям). Разбиение даты служит для анализа всевозможных показателей за определенный период (день, неделя, месяц, квартал, год). Суть разбиения заключается в том, что на основе столбца с информацией о дате формируется другой столбец, в котором указывается, к какому заданному интервалу времени принадлежит строка данных. Тип интервала задается аналитиком, исходя из того, что он хочет получить – данные за год, квартал, месяц, неделю, день или сразу по всем интервалам.

Квантование. Часто аналитику необходимо отнести непрерывные данные (например, количество продаж) к какому-либо конечному набору (например, всю совокупность данных о количестве продаж необходимо разбить на 5 интервалов – от 0 до 100, от 100 до 200 и т.д., и

отнести каждую запись исходного набора к какому – то конкретному интервалу) для анализа или фильтрации исходя именно из этих интервалов. Для этого в Deductor Studio применяется инструмент квантования (или дискретизации). Квантование предназначено для преобразования непрерывных данных в дискретные. Преобразование может проходить как по интервалам (данные разбиваются на заданное количество интервалов одинаковой длины), так и по квантилям (данные разбиваются на интервалы разной длины так, чтобы в каждом интервале находилось одинаковое количество данных). В качестве значений результирующего набора данных могут выступать номер интервала, нижняя или верхняя граница интервала, середина интервала, либо метка интервала (значения определяемые аналитиком).

Фильтрация данных. Почти всегда исходный набор данных, или набор данных после обработки аналитику необходимо отфильтровать. Фильтрация бывает необходима для разбиения данных на какие либо группы (например, товарные группы) для последующей обработки или анализа данных уже отдельно по каждой группе. Также некоторые данные могу не подходить, или наоборот, подходить для дальнейшего анализа в силу накладываемых условий (например, если на каком – либо этапе обработки данных были выявлены противоречивые записи, то их необходимо исключить из последующей обработки). Здесь тоже возникает необходимость фильтрации. Фильтрация позволяет из базового набора данных получить набор данных, удовлетворяющий определенным аналитиком условиям. В Deductor Studio механизм построения условий фильтрации прост для понимания. В окне мастера можно определить несколько элементарных условий фильтрации (<ПОЛЕ> <ОТНОШЕНИЕ> <ЗНАЧЕНИЕ>), последовательно связанных логическими операциями (И, ИЛИ).

Группировка данных. Совокупные данные намного более информативны, тем более, если их можно получить в различных разрезах. В Deductor Studio предусмотрен инструмент, реализующий сбор сводной информации – «Группировка». Группировка позволяет объединять записи по полям - измерениям и агрегируя данные в полях-фактах для дальнейшего анализа.

ЗАДАЧИ И ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ

Содержание работы:

1. Создать в блокноте файл «Credit.txt», содержащий данные кредитования. В файле должны быть такие столбцы, как «Сумма кредита», «Дата кредитования» (в формате ДД.ММ.ГГ), «Цель кредитования», «Возраст» (рис.8.28).

Сумма кредита	Дата кредитования	Цель кредитования	Возраст
7000	01.01.03	Оплата услуг	49
14578	01.01.03	Покупка товара	30
34567	03.01.03	Ремонт недвижимости	23
23567	04.02.03	Турпоездка	22
7500	05.02.03	Оплата услуг	56
12345	05.02.03	Покупка недвижимости	78

Рис.8.28 Пример заполнения файла «Credit.txt»

2. Импортировать в систему Deductor Studio текстовый файл «Credit.txt».

2.1. Произвести разбиение данных по рискам кредитования физических лиц.

2.2. Получить данные по суммам взятых кредитов по неделям.

2.3. Разбить данные о возрасте кредиторов на 5 интервалов (до 30 лет, от 30 до 40, от 40 до 50, от 50 до 60, старше 60 лет). Причем представить данные в разрезе по неделям.

3. Создать текстовый файл «banks.txt» и импортировать в систему. Файл должен содержать статистику по банкам России за определенный период («Банк», «Филиал», «Город», «Прибыль») (рис.8.29).

Банк	Филиал	Город	Прибыль
Внешторгбанк	32	Москва	355197
Газпромбанк	1786	Казань	0
АВТОБАНК	100	Москва	4678389
Банк ЗЕНИТ	24	Челябинск	0
"ОАО ""АЛЬФА-БАНК"	ALFM	Москва	356564

Рис.8.29 Пример заполнения файла «banks.txt»

3.1. Выявить ряд городов, в которых прибыль банков самая большая.

Порядок выполнения.

Интересующие нас столбцы: «СУММА КРЕДИТА», «ДАТА КРЕДИТОВАНИЯ», «ЦЕЛЬ КРЕДИТОВАНИЯ» и «ВОЗРАСТ». После импорта данных из текстового файла наиболее информативно просмотреть данные можно с помощью визуализатора «Куб», выбрав в качестве измерений столбцы «ВОЗРАСТ» и «ЦЕЛЬ КРЕДИТОВАНИЯ», а в качестве факта – столбец «СУММА КРЕДИТА». Остальные столбцы установить как непригодные (рис.8.30).

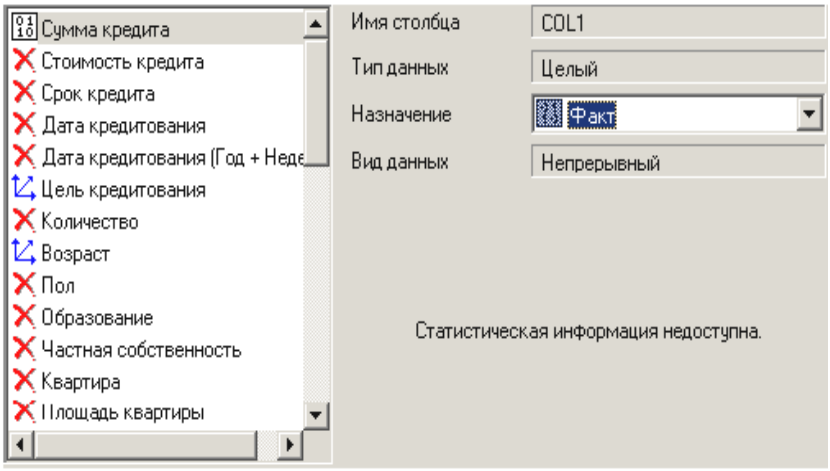


Рис.8.30 Настройка визуализатора «Куб»

На следующем шаге настройки куба следует указать измерение «ЦЕЛЬ КРЕДИТОВАНИЯ» как измерение в сроках, а измерение «ВОЗРАСТ» как измерение в столбцах, перетащив их с помощью мыши в соответствующие окна из области доступных измерений (рис.8.31).

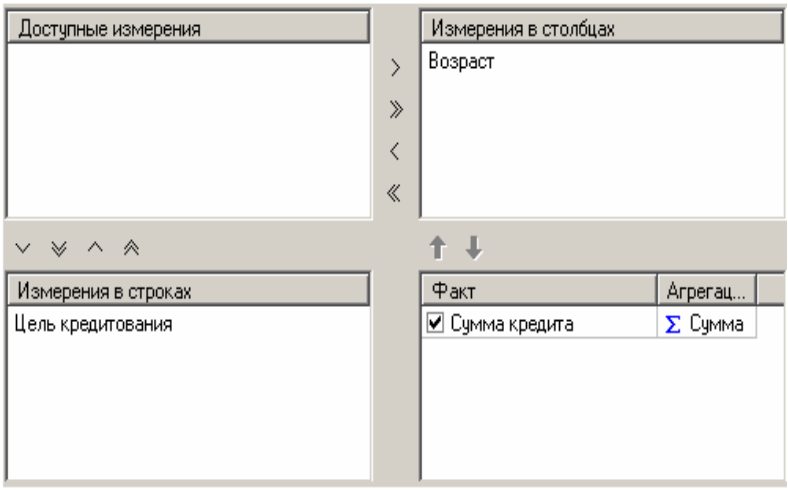


Рис.8.31 Настройка визуализатора «Куб»

В итоге, на кросс диаграмме (одна из закладок визуализатора куб) можно просмотреть исходные данные (рис.8.32).

	Возраст		
Цель кредитования	19	20	21
Иное	50 000,00	17 000,00	8 500,00
Оплата за образование		17 500,00	29 500,00
Оплата услуг (мед., юрид. и т.п.)			
Покупка и ремонт недвижимости	78 000,00		13 000,00
Покупка товара	46 500,00	73 500,00	76 500,00
Турпоездки, развлечения и т.п.		30 500,00	
Итого	174 500,00	138 500,00	127 500,00

Рис.8.33 Кросс-диаграмма

В мастере обработки «Дата и Время» на втором шаге выберем поле «ДАТА КРЕДИТОВАНИЯ» используемым, в появившейся после этого таблице настроек выберем назначение «Используемое» в столбце «Строка» напротив строки «Год + Неделя».

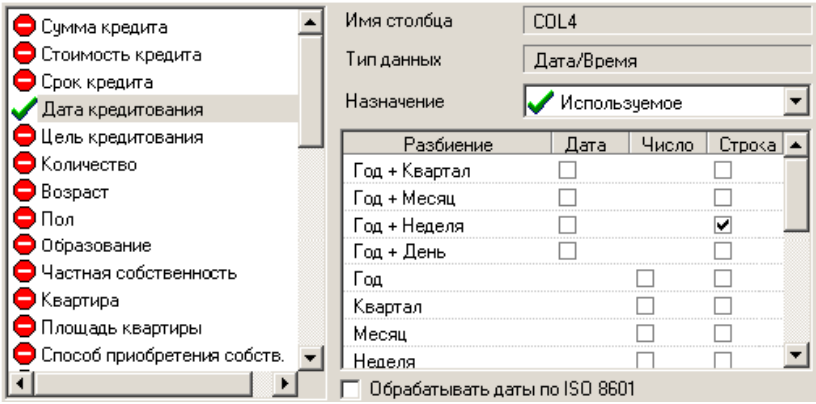


Рис.8.34 Окно мастера обработки «Дата и Время»

Больше никакие настройки не понадобятся, поэтому перейдем далее к выбору типа визуализации. Выберем в качестве визуализаторов «Таблицу» и «Куб», поставив галочки в соответствующих позициях. В мастере настройки полей куба выберем в качестве измерения появившийся после обработки столбец «ДАТА КРЕДИТОВАНИЯ_YWStr (Год + Неделя)» и столбец «ЦЕЛЬ КРЕДИТОВАНИЯ», а в качестве факта – «СУММА КРЕДИТА». Остальные поля сделаем неиспользуемыми. На следующем шаге перенесем одно измерение из области «доступных» в область «Измерения в строках», а другое – в область «Измерения в столбцах».

Таким образом, на кросс диаграмме имеем суммы взятых кредитов по неделям (за первые две недели года) в разрезе целей кредитования (рис.8.35).

	Дата кредитования (Год + Неделя)		
Цель кредитования	2003-W01	2003-W02	Итого
Иное	358 000,00	137 000,00	495 000,00
Оплата за образование	62 000,00	312 000,00	374 000,00
Оплата услуг (мед., юрид. и т.п.)	110 500,00	191 000,00	301 500,00
Покупка и ремонт недвижимости	404 000,00	538 000,00	942 000,00
Покупка товара	642 000,00	643 500,00	1 285 500,00
Турпоездки, развлечения и т.п.	35 500,00	113 500,00	149 000,00
Итого	1 612 000,00	1 935 000,00	3 547 000,00

Рис.8.35 Кросс диаграмма

В таблице с данными видно, что новое поле - «ДАТА КРЕДИТОВАНИЯ_YWStr (Год + Неделя)» содержит одинаковые значения (дата начала недели) для строк, которые попадают в одну и ту же неделю (дата начала недели или номер недели с начала года) (рис.8.36).

Срок кредита	Дата кредитов	Дата кред	Цель кредитования
24	05.01.03	2003-W01	Покупка товара
12	05.01.03	2003-W01	Покупка товара
30	05.01.03	2003-W01	Иное
36	06.01.03	2003-W02	Покупка и ремонт недвижимости
12	06.01.03	2003-W02	Оплата за образование
18	06.01.03	2003-W02	Иное
6	06.01.03	2003-W02	Покупка товара

Рис.8.36 Таблица данных

Далее используем инструмент квантования для разбиения данных о возрасте кредиторов на 5 интервалов (до 30 лет, от 30 до 40, от 40 до 50, от 50 до 60, старше 60 лет). Исходные данные распределятся по пяти интервалам именно так, поскольку, согласно статистике, минимальное значение возраста кредитора 19, а максимальное 69 лет. Это необходимо аналитику для оценки кредиторской активности разных возрастных групп, с целью принятия решения о стимулировании кредиторов в группах с низкой активностью (например, уменьшение стоимости кредита для этих групп) и, быть может, увеличение прибыли

в возрастных группах кредиторов с высоким риском (путем предложения дополнительных платных услуг). Причем аналитик желает видеть данные в разрезе по неделям. Воспользуемся мастером квантования (рис.8.37).

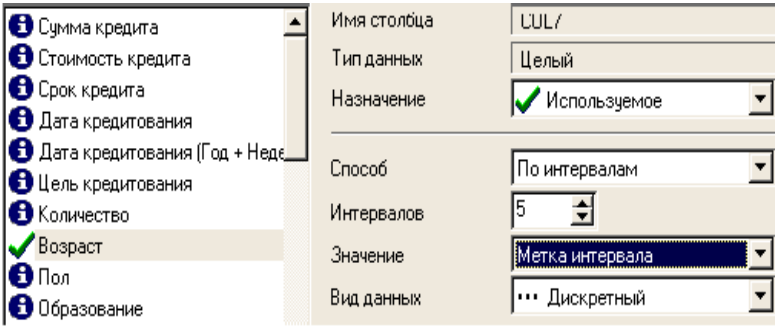


Рис.8.37 Мастер квантования

В нем выберем назначение поля «Возраст» используемым, укажем способ разбиения «По интервалам», зададим количество интервалов равное 5, в качестве значения выберем «Метку интервала».

На следующем шаге мастера определим сами метки соответственно возраста кредиторов: «до 30 лет», «от 30 до 40 лет» и т.д. (рис.8)

Столбцы	
Имя	Интервалов
12 Возраст	5

Интервалы (изменены)		
NN	Граница	Метка
	0	
0	29	До 30 лет
1	39	От 30 до 40 лет
2	49	От 40 до 50 лет
3	59	От 50 до 60 лет
4	1000	Старше 60 лет

Рис.8.38 Определение меток

После обработки выберем в качестве способа отображения «Куб». В мастере укажем «СУММА КРЕДИТА» в качестве факта, «ВОЗРАСТ» и поле «ДАТА КРЕДИТОВАНИЯ (Год + Неделя)» в качестве измерения, остальные поля укажем неиспользуемыми. Далее перенесем «ВОЗРАСТ» из доступных измерений в «Измерения в строках», а «ДАТА КРЕДИТОВАНИЯ (Год + Неделя)» в «Измерения в столбцах». На кросс диаграмме (рис.8.39) теперь видна информация о том, какие суммы кредитов берут кредиторы определенных возрастных групп в разрезе по неделям.

	Дата кредитования (Год + Неделя)		
Возраст	2003-W01	2003-W02	Итого
До 30 лет	798 500,00	808 500,00	1 607 000,00
От 30 до 40 лет	298 000,00	561 000,00	859 000,00
От 40 до 50 лет	195 000,00	295 500,00	490 500,00
От 50 до 60 лет	111 000,00	209 500,00	320 500,00
Старше 60 лет	209 500,00	60 500,00	270 000,00
Итого	1 612 000,00	1 935 000,00	3 547 000,00

Рис.8.39 Кросс диаграмма

Теперь аналитик, получив такие данные, может дать рекомендации о снижении стоимости кредита для лиц, старше 50 лет, либо о применении каких – нибудь других мер, способных привлечь большее количество кредиторов этих групп, либо мер, направленных на то, чтобы кредиторы брали кредит на большие суммы.

Рассмотрим ситуацию, когда аналитику необходимо спрогнозировать кредитоспособность потенциального кредитора. Предполагается, что кредиторы, берущие суммы разного диапазона ведут себя по-разному, следовательно, модели прогноза должны свои для каждой группы. Т.е. для дальнейшего построения моделей прогноза кредитоспособности определенных аналитиком категорий необходимо использовать фильтрацию.

Определим, для примера группу кредиторов, взявших кредит менее 10000 руб. Воспользуемся данными предыдущего примера. Для этого, находясь на узле импорта данных из текстового файла, запустим мастер обработки. В нем в качестве метода обработки выберем фильтрацию. На втором шаге мастера можно видеть одно неопределенное условие фильтрации (при необходимости их можно добавлять или удалять соответствующими кнопками на форме). Поскольку необходимо отфильтровать данные только по кредиторам, взявших кредит менее 10000, то в графе «Имя поля» выбираем поле «СУММА КРЕДИТА», в графе «Условие» выбираем знак меньше, в графе «Значение» пишем «10000», как изображено на рисунке 8.40:

Операция	Имя поля	Условие	Значение
	12 Сумма кредита	<	10000

Рис.8.40 Фильтр данных

Больше никаких условий не требуется, поэтому переходим на следующий шаг мастера и запускаем процесс фильтрации. После выполнения обработки можно манипулировать уже только с данными по кредиторам выбранного кредитного диапазона (рис.8.41). В правильности выполненной операции можно легко убедиться, выбрав в качестве визуализации данных статистику и просмотрев значения минимального и максимального значения поля «СУММА КРЕДИТА».

Метка столбца	Тип данных	Минимум	Максимум
Сумма кредита	12 Целый	2000	9500
Стоимость кредита	12 Целый	400	1900
Срок кредита	12 Целый	6	6
Дата кредитования	7 Дата/Время	01.01.03	11.01.03
Дата кредитован...	ab Строковый		
Цель кредитования	ab Строковый		
Количество	12 Целый	1	1

Рис.8.41 Процесс фильтрации

Теперь допустим, что у аналитика имеется статистика по банкам России за определенный период. Перед ним стоит задача выявления ряда городов, в которых прибыль банков самая большая для использования этих данных в дальнейшем. Для этого аналитик должен обратить внимание на следующие поля таблицы из файла: «БАНК», «ФИЛИАЛЫ», «ГОРОД», «ПРИБЫЛЬ». Т.е. информация о названии банка, городе, в котором он находится, (филиалы банка могут находиться в разных городах – следовательно, по одному и тому же банку может быть несколько записей с данными по разным городам) и прибыль банка.

Для решения поставленной задачи первым делом необходимо найти суммарную прибыль всех банков в каждом городе. Для этого и необходима группировка. Для начала следует импортировать данные по банкам из текстового файла. Просмотреть исходную информацию

можно в виде куба, где по строкам будут названия банков, а по столбцам – города. С помощью визуализатора «Куб» также можно получить требуемую информацию, выбрав в качестве измерения поле «ГОРОД», а в качестве факта «ПРИБЫЛЬ». Но нам необходимо получить эти данные для последующей обработки, следовательно, необходимо сделать аналогичную группировку.

Находясь в узле импорта, запустим мастер обработки. Выберем в качестве обработки группировку данных. На втором шаге мастера установим назначение поля «ГОРОД» как измерение (рис.8.42), а назначение поля «ПРИБЫЛЬ» как факт. В качестве функции агрегации у поля «ПРИБЫЛЬ» следует указать Сумму.

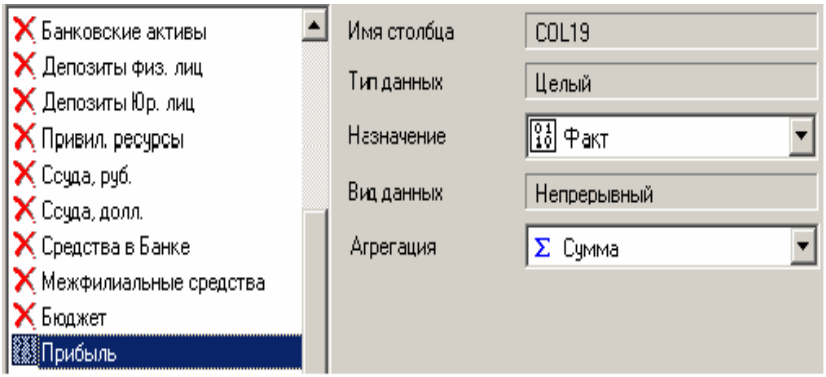


Рис.8.42 Окно мастера обработки

Таким образом, после обработки получим суммарные данные по прибыли всех банков по каждому городу. Их можно просмотреть, используя таблицу (рис.8.43).

Город	Прибыль
Москва	6076922
Санкт-Петербург	233620
Уфа	370468
Санкт-Петербург	128038
Ханты-Мансийск	30679
Казань	68576
Челябинск	63956

Рис.8.43 Таблица данных

ЗАДАНИЕ НА ЛАБОРАТОРНУЮ РАБОТУ

Создать сценарий в среде Deductor Studio согласно варианту, полученному у преподавателя.

ТРЕБОВАНИЯ К РЕАЛИЗАЦИИ

Задание выполняется согласно варианту в среде Deductor Studio. По завершении готовится отчет.

ВАРИАНТЫ ЗАДАНИЙ

1. Добавить в файл с информацией о банках столбец с количеством клиентов. Разбить данные на 3 интервала по количеству клиентов: малые, средние и крупные банки. Сравнить прибыль банков из разных групп.

2. Добавить в файл с информацией о банках столбец с количеством клиентов. Разбить данные на 3 интервала по количеству клиентов: малые, средние и крупные банки. Сгруппировать данные по названию банка. Определить количество малых, средних и крупных филиалов у всех банков

3. Добавить в файл с информацией о банках столбец с количеством клиентов. Определить банки, у которых прибыль ниже определенного значения и количество при этом больше 100000.

4. Добавить в файл с информацией о кредитах столбец с идентификатором клиента. С помощью OLAP-куба получить информацию по каждому клиенту: количество кредитов по каждой цели кредитования.

5. Добавить в файл с информацией о кредитах столбец с идентификатором клиента. С помощью OLAP-куба получить информацию по каждому клиенту: средняя, максимальная и минимальная сумма кредита.

6. Добавить в файл с информацией о кредитах столбец с идентификатором клиента. С помощью OLAP-куба получить информацию по каждому клиенту: средняя сумма кредита. В зависимости от средней суммы кредита разделить клиентов на 3 интервала: крупные кредиторы, средние кредиторы, мелкие кредиторы. Представить данную информацию в разрезе по возрасту клиентов.

7. Добавить в файл с информацией о кредитах столбец с идентификатором клиента. С помощью OLAP-куба получить информацию по каждому клиенту: количество кредитов. В зависимости от количества кредитов разделить клиентов на 3 интервала: частые кредиты, средние кредиты, редкие кредиты. Отобразить суммы кредитов, сгруппированные по частоте получения кредитов, в разрезе по неделям.

КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ

1. Раскройте значение понятия «разбиение данных на группы».
2. Опишите назначение разбиения данных на группы в ходе анализа данных.
3. Определите вид анализа, использующий разбиение по дате.
4. Дайте определение фильтрации данных.
5. Опишите назначение фильтрации данных в анализе решений.
6. Перечислите способы осуществления фильтрации данных в системе Deductor Studio.
7. Дайте определение группировки данных.
8. Опишите назначение группировки данных в анализе решений.
9. Перечислите способы осуществления группировки данных в системе Deductor Studio.
10. Опишите назначение квантования данных в анализе решений.
11. Перечислите способы осуществления квантования данных в системе Deductor Studio.
12. Дайте определение кросс-диаграммы.

ФОРМА ОТЧЕТА ПО ЛАБОРАТОРНОЙ РАБОТЕ

На выполнение лабораторной работы отводится 2 занятия (4 академических часа: 3 часа на выполнение и сдачу лабораторной работы и 1 час на подготовку отчета).

Отчет на защиту предоставляется в печатном виде.

Структура отчета (на отдельном листе(-ах)): титульный лист, формулировка задания, описание процесса выполнения лабораторной работы, результаты выполнения работы, выводы.