ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ»
(Н И У « Б е л Г У »)

ИНСТИТУТ ИНЖЕНЕРНЫХ ТЕХНОЛОГИЙ И ЕСТЕСТВЕННЫХ НАУК

КАФЕДРА ОБЩЕЙ МАТЕМАТИКИ

РАЗРАБОТКА МЕТОДА, АЛГОРИТМОВ И ПРОГРАММНЫХ КОМПОНЕНТОВ ДЛЯ ПРОГРАММЫ МНОГОМЕРНОГО АНАЛИЗА ВОЕННОЙ МОЩИ ВЕДУЩИХ СТРАН МИРА

Выпускная квалификационная работа
обучающегося по направлению подготовки
01.03.02, Прикладная математика и информатика
очной формы обучения,
группы 07001305
Ерёмина Вячеслава Васильевича

Научный руководитель
д.т.н., профессор
Аверин А.Г.

БЕЛГОРОД 2017

СОДЕРЖАНИЕ

СОДЕРЖАНИЕ

.

ВВЕДЕНИЕ

Исходная информация в социально-экономических исследованиях представляется чаще всего в виде набора объектов, каждый из которых характеризуется рядом признаков (показателей). Поскольку число таких объектов и признаков может достигать десятков и сотен, и визуальный анализ этих данных малоэффективен, то возникают задачи снижения количества исходных данных, выявления структуры и взаимосвязи между ними на основе построения обобщенных характеристик множества признаков и множества объектов. Такие задачи могут решаться методами многомерного статистического анализа. Многомерный статистический анализ - раздел математической статистики, посвященный математическим методам, направленным на выявление характера и структуры взаимосвязей между компонентами исследуемого многомерного признака и предназначенным для получения научных и практических выводов.

Основной целью работы является разработка метода, алгоритмов и программных компонентов для программы многомерного анализа военной мощи ведущих стран мира.

Актуальность данной работы обусловлена тем, что разрабатываемая система анализа позволит существенно упростить аналитическую работу, свести к минимуму скорость обработки больших объемов информации, а также получать актуализированные данные.

Задачи, решаемые в работе:

- анализ состояния вопроса;
- выбор методов многомерного анализа данных и разработка алгоритмов и программных компонентов;
- разработка, тестирование и реализация программного обеспечения.

Объектом исследования являются методы многомерного анализа данных. Предметом исследования исследования являются алгоритмы и программные компоненты для многомерного анализа данных.

Методы исследования используемые в работе:

- сбор данных и их систематизация, анализ состояния вопроса;
- методы многомерного анализа данных и способы визуализации информации;
- построение информационных моделей;
- объектно-ориентированное программирование и разработка тестовых примеров.

1 Методы многомерного статистического анализа

Многообразие свойств природных объектов и многофакторность природных процессов приводит исследователя к проблеме обработки огромной массы статистических наблюдений. В многомерных данных каждый объект наблюдений характеризуется множеством признаков (переменных). Многомерные методы позволяют одновременно изучать изменение набора характеристик. Конечной целью большинства многомерных статистических методов является предсказание (прогнозирование) тех или иных свойств изучаемых объектов, будь то гидрометеорологические, социально-экономические, экологические и т.д.

Можно привести много примеров гидрометеорологических, геохимических, социально-экономических и других данных, к которым применимы методы многомерного анализа. Среди них можно назвать химические анализы, в которых переменные представляют собой содержание микро или макро элементов в воде, почве, снеге. Примером многофакторного процесса может служить речной сток, являющийся результатом взаимодействия многих геофизических процессов (прямая и рассеянная радиация, осадки, температура воздуха и подстилающей поверхности,

давление и влажность воздуха, скорость и направление ветра и т.д.) физико-географических условий бассейна (ландшафт, почвы, геологическое строение, растительность) и т.д. Многомерные методы позволяют исследователю работать с большим числом переменных, объём которых невозможно обработать вручную без компьютера. Однако эти методы сложны как с методологической, так и с теоретической точки зрения. Статистические критерии и процедуры большей части этих методов разработаны лишь при очень сильных ограничениях, а поведение при решении реальных задач изучено слабо.

Некоторые процедуры многомерного анализа совсем не имеют теоретического обоснования, для них не созданы ещё критерии проверки соответствующих гипотез. Например, до сих пор не разработаны способы оценки адекватности результатов кластер-анализа. Тем не менее, эти методы ≪используют≫ и они дают неплохие результаты при условии сочетания их с профессиональным опытом и интуицией исследователя в конкретной предметной области, то есть реализуется принцип ≪*доказать нельзя, а использовать можно*≫.

Есть два пути решения проблемы обработки многомерных данных:

1) Исключить часть малоинформативных характеристик и возвратиться к мало размерным классическим задачам;

2) объединить характеристики в группы (в дальнейшем - факторы) для уменьшения признакового пространства.

Второй подход приводит к задаче обратного сведения множества характеристик к небольшому ряду обобщающих параметров, выражающих реально существующие закономерности в наборе данных, и соответственно сформировалось направление, получившее название ≪многомерный анализ≫ (Факторный, дискриминантный и кластерный анализ, 1989).

Развитие многомерного статистического анализа как науки началось с 1901-1904 гг. В это время появились статьи К. Пирсона и Ч. Спирмена, посвящённые теории факторного анализа. Методы многомерного

статистического анализа базируются на представлении исходной информации в многомерном признаковом пространстве и позволяют определять неявные, но объективно существующие закономерности в данных и тенденциях развития изучаемых явлений и процессов. Круг основных теоретических и практических задач, решаемых с помощью методов многомерной статистики, заключается в анализе и выявлении связей внутри комплекса исходных признаков, выделении групп случайных признаков, обладающих наиболее сильными связями, оценке вклада ведущих признаков и факторов (последние представляют комплекс генетически однородных характеристик) в общую дисперсию в типизации (группировке) объектов в многомерном пространстве.

Постепенно в многомерном анализе образовались разделы, взаимодополняющие друг друга - кластерный анализ, таксономия, распознавание образов, метод главных компонент, факторный анализ (Харман, 1972; Тыоки, 1981).

Особо стоит сказать о задаче классификации, одной из важнейших в обработке естественно-научных данных. Под решением *задачи классификации* понимается установление правил отнесения объекта к одной или нескольким группам (категориям, классам) на основании некоторого числа его характеристик (признаков) и построение описаний классов. же отнесение объекта к или иному классу с известным называется *идентификацией.* В работе по многих изданий классификация используется в смысле, включая . Если объектов разбивается на (классы) на основании признака, то классификация *монотетической.* для построения классификации и несколько признаков , то она называется *политетической.* Из известных ярким представителем метода является анализ (ДА). признаков, в ДА, как правило, невелико. того, задачу можно решать с методов *классификации,* которая у все признаки объектов. Это кластерного анализа, таксономии и т.д. Они к группе методов образов. Разновидностью классификации являются факторного (включая метод компонент). В них

классификация на основе нескольких показателей, факторами и компонентами.

1.1 корреляция и множественный анализ

Для успешного методов статистики необходимы

в таких областях математики как аналитическая , матричная , многомерный математический . Характерной особенностью методов является представление . Наблюдаемые объекты изобразить как точки в пространстве, соответствующем признаков, они характеризуются. Если разнородны, то их нормируют. методов многомерного анализа не является однозначным. , задачи группирования по принципу сходства можно и кластерным и факторным . У каждого метода свои сильные и стороны. корреляция используется для *степени тесноты* между признаками , а множественный анализ для определения *этой связи.* цель регрессионного - построить по матрице уравнение , по которому можно интерпретировать результаты и осуществлять . Одним словом, корреляция и множественный анализ применяются для и моделирования изучаемых признаков и . В статистике очень используется метод регрессии в направлениях - для восстановления по регрессии пропущенных и в целях прогноза. , по известным наблюдений за расходами на реках-аналогах и частично на реке, можно уравнение и по нему рассчитать () часть отсутствующих на контрольной реке при , естественно, требований, как к рекам-аналогам, так и к уравнению регрессии. многолетние данные по увлажнению , запасам снега, осадкам за период и т.д., можно найти между половодья и перечисленными формирования этого , а затем по ней спрогнозировать с заблаговременностью половодья в текущем .

1.2 Дискриминантный анализ

анализ является статистическим

решения классификационных , т.е. разделения *(дискриминации)* нормально распределённых на группы. На имеющихся формулируется правило, по новые единицы совокупности к одному из существующих , при этом новые не образуются. Таким , производится новых объектов по ≪эталонным группам≫. всего дискриминантный используется для разделения совокупностей па два , например, отделения территорий от незагрязнённых по ка химическому . В гидрометеорологии дискриминантный применяется чаще в улучшения качества прогнозов в с другими статистическими , например, с методом регрессии. В большинстве дискриминантному два класса объектом (, совокупность ситуаций нормы и ниже). В дискриминантном число классов () задаётся заранее.

1.3 К анализ

Кластерный это совокупность , предназначенных для разбиения объектов на однородные (кластеры). В большинстве кластерного заранее неизвестно, классов будет в данной совокупности . В отличие от анализа кластерный называется классификацией без ≪≫, потому что его методы не обучающую .

Задачи кластерного :

- классификация объектов с признаков, определяющих их ;
- проверка о структуре данных;
- новых классов.

, можно на основании геохимических в снежном покрове разбиение территории на классы по уровню нагрузки. таблицу с набором показателей с разных такого полезного как торф, разбить местороджения на классы по тому или направлению использования в или сельском .

1.4 Факторный анализ

анализ применяется в для ≪сжатия≫ данных, т.е. множества признаков к небольшому ≪обобщённых признаков≫, и латентных (скрытых,) факторов. Эта же может решаться не относительно признаков, но и объектов. Факторный после обобщённых показателей использовать для целей . Например, с его помощью решать районирования территории по формирования стока или по химических анализов лять группирование территории по превышения ПДК и т.д. Результаты, на основе статистических , зависят как от выбора самих м, так и точности исходных . К прикладной статистике применимо английского натуралиста : ≪*математику можно с мельницей превосходного , которая что угодно до любой . Тем не менее, то, что вы получите, от того, что вы засыпаете. И как великолепная в мире не доставит вам крупчатку из лебеды, так и формул не доставят вам результата из данных*≫.

1.5 Основные представления многомерных

Статистика имеет с совокупностями о, описываемых некоторыми (качественными или количественными). Если каждый имеет характеристику, то принято об *одномерных данных.* Если характеристик у каждого две и более, то рассматриваются как *многомерные.* природных (генеральных) пностей заключается в характеристик и получении выборочных или выборок. Генеральная характеристик природных может как случайная величина мерности, свойства оцениваются с помощью . Первичные *(выборку* многомерной величины) в науках о обычно формируют в таблицы *n*m,* где где *n* число строк, числу объектов, в выборку; *m* число , содержащих (измерения) каждого . Статистика же оперирует . *Матрицей* называется таблица из , содержащая некоторое строк (*n*) и некоторое столбцов *(m).* Если *m = n,* называется *,* а число *m* или *n* - её *порядком.* матрицы называется линейно независимых (или) матрицы.

Квадратными , например, ковариационная и матрицы, которые во всех многомерной прикладной . Матрицу будем заключённой в квадратные с, например [D] или [$D(d_{ij})$], где d_{ij} - матрицы; i - номер ; j- номер столбца. или строку матрицы рассматривать как . Будем для его обозначения использовать квадратные , например, [Xj] - j-й столбец [X]. Будем также полную записи для вектора, , вектор-строка [Xj] = {xi, x2, ... xn}, где x, и x2, ... xn - компоненты(, координаты) . Векторы называют да точками. Например, [Xj] можно геометрически как точку с Xj, x2j ... xnj в n-мерном пространстве. Это векторов используется в , дискриминантном и кластерном .

1.6 Основные представления многомерных

Многомерные данные быть отображены в виде на основе реляционных данных, а также и многомерными инструментальными .

Представление данных в рамках моделей может в виде трёх схем:

- «»;
- «снежинка»;
- «созвездие».

представление на плоскости на рисунке 1.

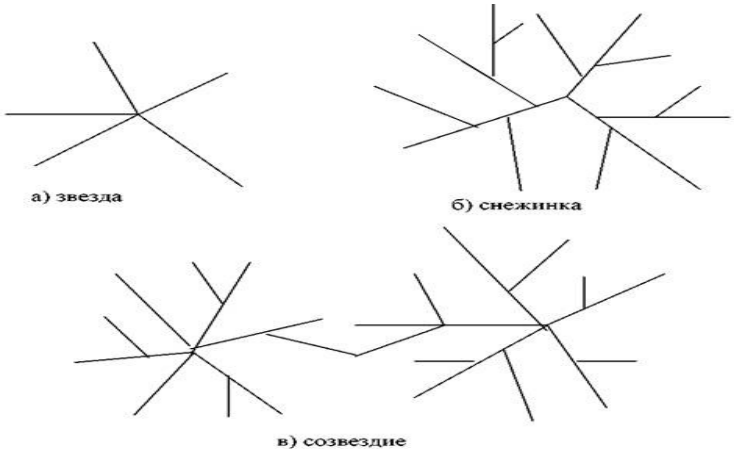


Рис. 1. Линейное реляционных многомерных данных

схемы являются таблиц реляционной . На рисунке 2 ены схема базы Northwind, входящей в поставки СУБД MS SQL и MS Access, а варианты схем на их основе кубов .

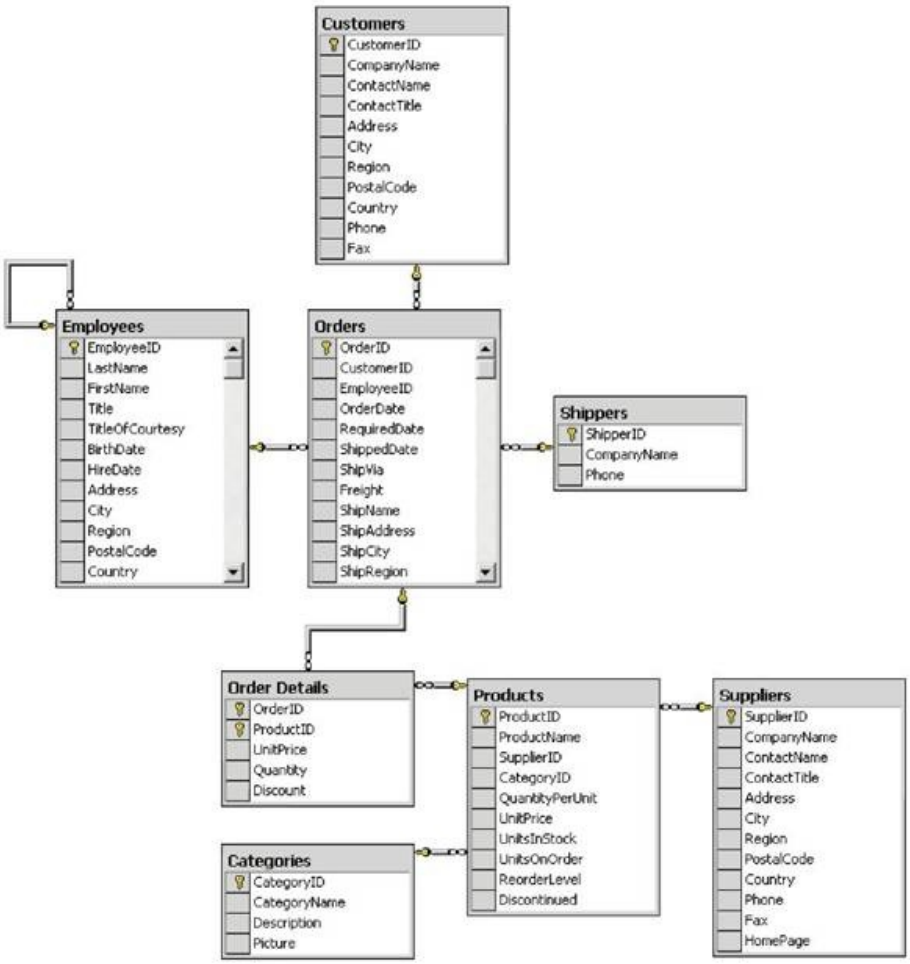


Рис. 2. Схема базы Northwind

В пуле информации большая центральная ица, называемая таблица (fact). В ней помещаются все данные интересующего пользователя показателя. Её окружают таблицы, данные по признакам, таблицы размерности или их называют измерений (table). размерности являются по отношению к таблице . Таблица факта дочерней. быть также таблицы (outrigger). Они присоединяются к таблицам и детализируют атрибуты. Консольные являются родительскими по к таблицам размерности. фактов числовые или качественные () значения.

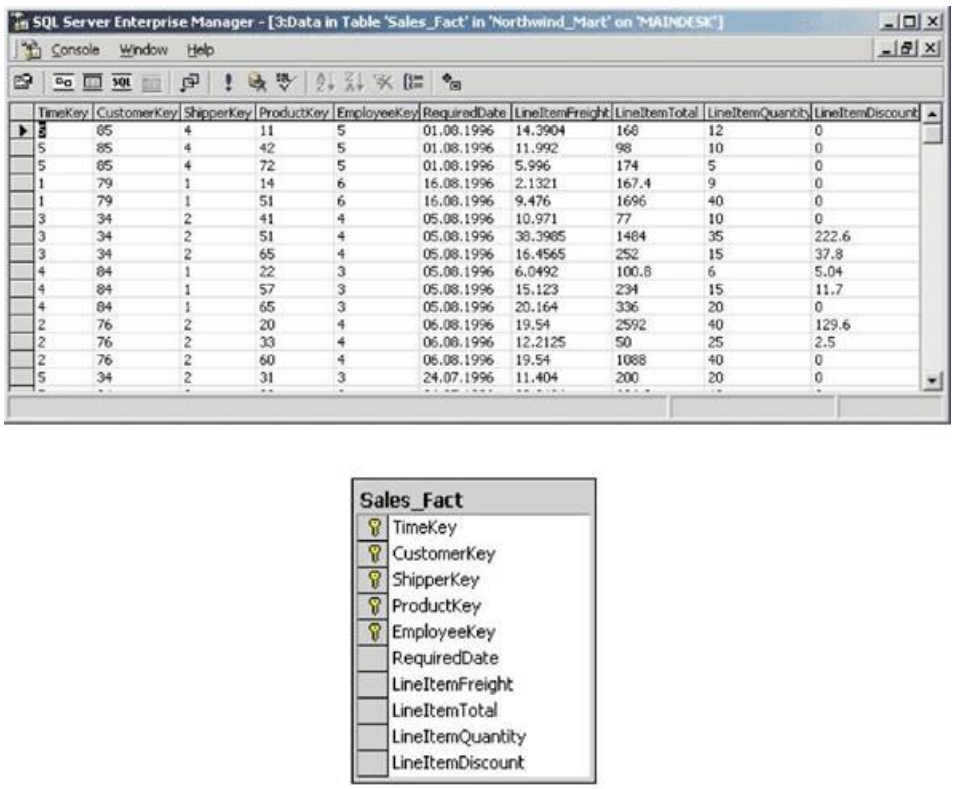
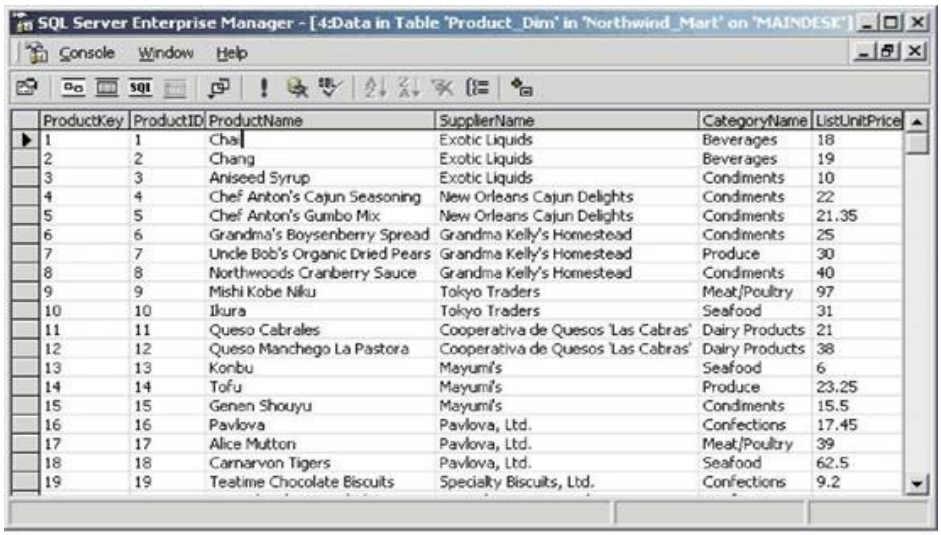


Рис. 3 Таблица table



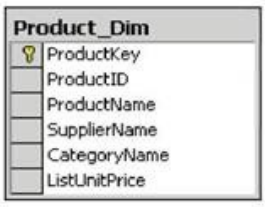


Рис. 4 Таблица table

При р базы данных по «звезда» или по другой схеме необходимо и тщательно предметную область; в центральную таблицу все характеризующие исследуемый данные, разработав систему . Консольные и

таблицы , а также таблица соединяются связями. Первичные родительских таблиц внешними ключами . Например, ключ таблицы является внешним таблицы факта. «звезда» только из таблиц ности и таблицы .

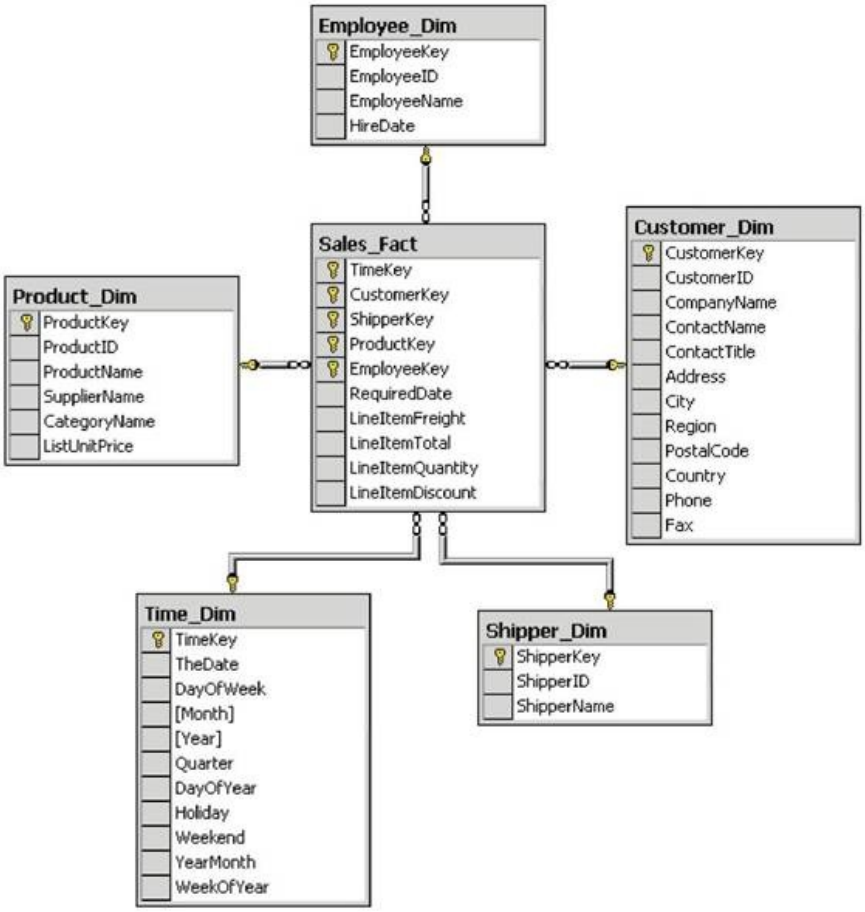


Рис. 5 Схема «звезда»

схемы «» является схема «» (snowflake schema). Её от первой схемы количество таблиц, они имеются на каждой таблице и могут иметь уровней и.

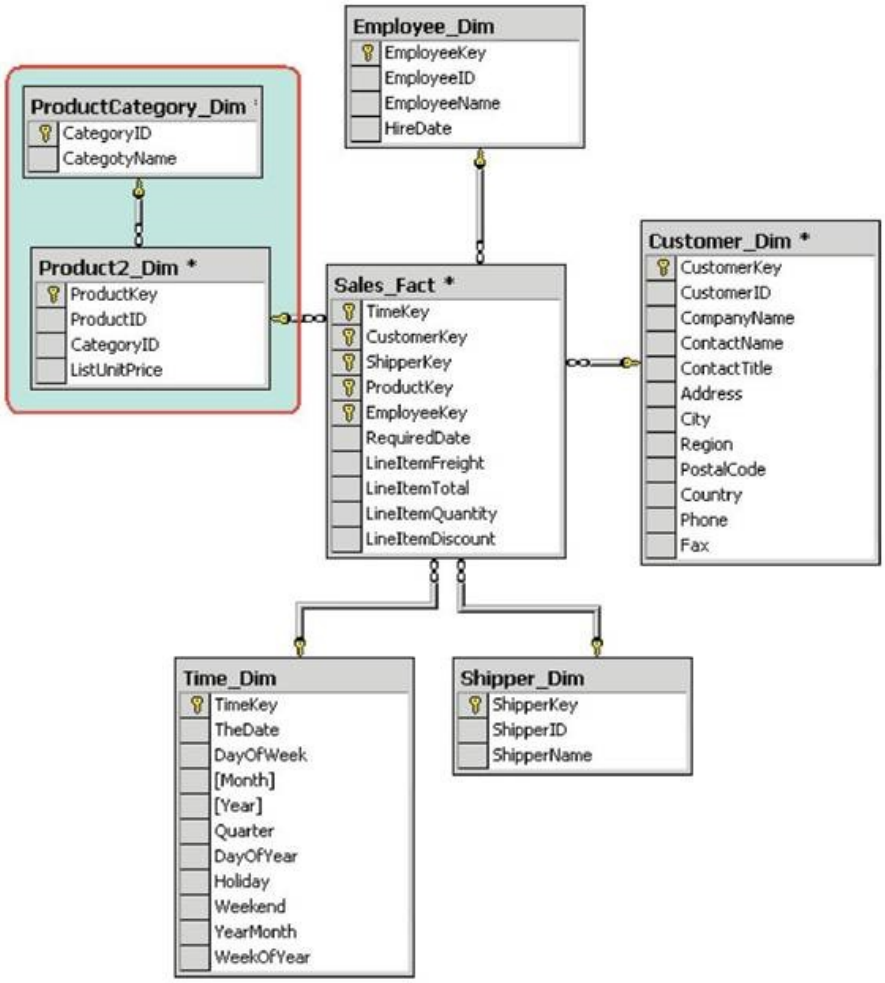


Рис. 6 Схема «снежинка»

«созвездие» (fact schema) получается из не таблиц . В этом варианте модели через или таблицы размерности несколько фактов, отображающих объектов с общими . В схемах «снежинка» и «ие» применение таблиц приводит к затратам времени на запроса. При проектировании фактор учитываться. При создании моделей на основе ионной базы рекомендуют длинные и узкие фактов и сравнительно и широкие таблицы (измерений). реализации многомерных баз на реляционной СУБД в ом виде приведены на 3 - 6. Многомерные данных на основе СУБД отличаются де, точнее отсутствием или нормализации. дублирование или избыточность . Ячейки гиперкубов, такими средствами, одинаковую , что также приводит к расходу ресурсов .

1.7 Математическое ожидание и многомерной величины

Многомерная величина. Так как строки и матрицы данных представить как векторы, то многомерная

величина может быть векторной. Двумерная величина в варианте может изображена с помощью гистограммы, имитирующей распре вероятностей (функцию). Гистограмму можно построить, используя ие процедуры Statistica: меню —≫ *3D Sequential Graphs* —> *Histograms*. Представить с обычных функцию распределения рной векторной величины затруднительно, можно функцию распределения из *m* одномерных случайных . Математическим ожиданием омерной величины [X], состоящей из *m* векторов [X1],[2], ... [Xm] является вектор [)] = {M(Xi), M(X2), ... M(Xm)}. О математического ожидания случайной величины *выборочный вектор* x= (xi, x2 , ... Xm), компонентами ого являются выборочные значения всех . Матрицу данных вычитания ора средних можно в матрицу вариаций [Y] *n*m* с элементами $y_{ij} = x_{ij} - x_i$ Изменчивость случайной ы характеризуется дисперсионной , называемой также цей ковариаций или вариационно- матрицей. Это матрица размером m*m. дисперсий и ковариаций представить как квадратную матрицу [V]. По этой матрицы суммы квадратов отк значений всех от своих :

$$v_n = \sum_{k=1}^{n} (¿¿¿ - x_i)^x$$

Недиагональные элементы [V] представлены суммами произведений:

$$v_{ij} = \sum_{k=1}^{n} (¿¿kj - x_j)^x (x_{ki} - x_i)$$

При делении элементов на объём выборки матрицу дисперсий и [D]. Если каждый матрицы [D] на корень квадратный из соответствующих

дисперсий, то корреляционную матрицу [R] с , называемыми коэффициентами корреляции и линейные зависимости свойствами объекта:

$$r_{ij}=\frac{d_{ij}}{\sqrt{d_{ij}}}$$

процедуры, со сравнением многомерных совокупностей, которое в дальнейшем в дискриминантном .

Сравнение двух многомерных совокупностей. Пусть две случайные векторные X(1) и X(2), о выборками объёма n_1 и n_2 случайная величина объект, описанный κ . Проверим о равенстве математических случайных величин H0: MX(1) = (2) Воспользуемся Хоттелинга (T^2), многомерным аналогом к Стьюдента. Статистика T^2 по достаточно громоздкой :

$$T^2=\frac{n_1 n_2}{n_1+n_2}\left(x^{-1}-x^{-2}\right)\left|D\right|^{-1}\left(x^{-1}-x^{-2}\right)$$

где D^{-1} - матрица, обобщённой эмпирической матрице системы, слева на вектор , а справа на столбец разности средних двух . Обобщённая ковариационная ма может получена путём выборок.

Таблицы рас Хоттелинга не всегда , поэтому F-критерий, связанный со T^2 , имеющей F-распределение:

$$F=\frac{T^2\left(n_1+n_2-k-l\right)}{\left(n_1+n_2-2\right)k}$$

$F_{рассч} > F_{крит}$ для значимости а и свободы κ и (n_1+n_2-k-1), то нулевая о равенстве векторов отвергается. Основное , на котором рассмотренный критерий, в том, что выборки взяты из нормально совокупностей, имеющих и ту же или одинаковые матрицы. Предположение о ности распределения и ковариационных матриц, как и при одномерных и дисперсий, в реальности нарушается.

1.8 Сравнение двух много выборочных

Сравнение ковариационных возможно с помощью *кри обобщённых дисперсий,* многомерным ом F-критерия. имеется две группы объёмом п1, и п2. Найдём для них матрицы [D1] и |2]. Сформулируем нулевую H0: |D 1|= | D 2| при альтернативе H1: |D1|не ровно |2|. Объединим выборки и обобщённую ковариационной [D], предполагаемую общей для генеральных совокупностей.

вычислим *M:*

$$M=(n_1+n_2-2)\ln|D|-\frac{1}{2}[(n_1-1)\ln|D1|+(n_2-1)\ln|D2|] \ ,$$

представляющую собой между логарифмом обобщённой ковариационной и средним зна логарифмов определителей ковариационных матриц. значение статистики M х-квадрат с числом степеней

равным 0,5$p(p$ + 1), где p — мерность величины:

$$X^2=M\,C^{-1}$$

Если значение M осходит критическое, то гипотеза о равенстве х матриц должна отвергнута в альтернативной.

1.9 Анализ за в многомерных данных

корреляционный и регрессионный относятся к немногих количественных , которые могут использованы для исследования природных . Основная задача анализа состоит в степени тесноты связи переменными, т.е. в расчёте матрицы по выборкам и частных и множественных корреляции и . Основное назначение анализа заключается в вида стохастических между . Он устанавливает форму между одной (Y), рассматриваемой в качестве , и значениями или нескольких переменных из этого же набора , рассматриваемых как *независимые* (X1 X2, ... X_n) и некоторые значения.

Зависимую называют также , а независимую *предиктором.* уравнение использовать для оценки влияния нескольких на данный процесс с его прогнозов и . Кроме того, метод позволяет

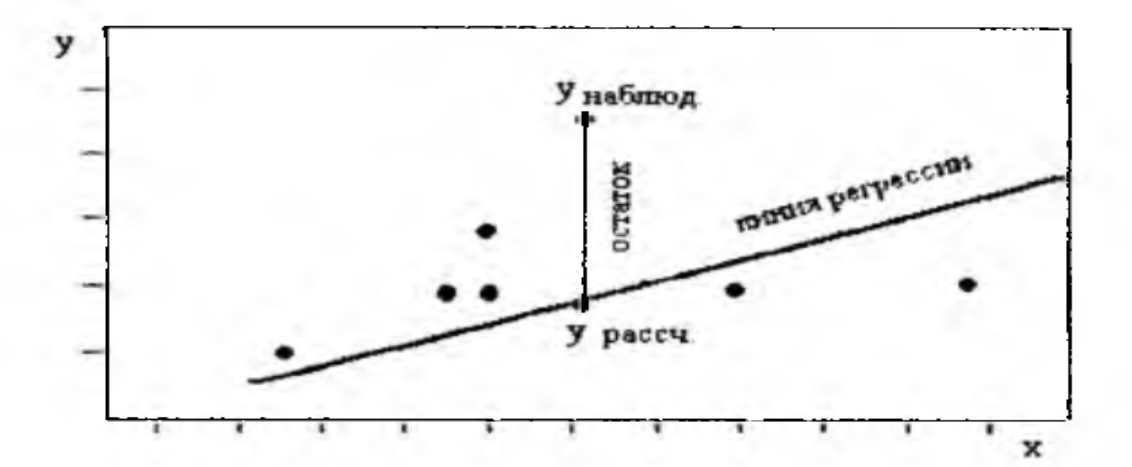
относительное влияние на каждого ф и измерять полный с помощью коэффициентов. также оценить связи зависимой и каждой переменной и получить ≪≫ расчётное уравнение.

1.10 остатков

называют разность наблюдения и предсказанного

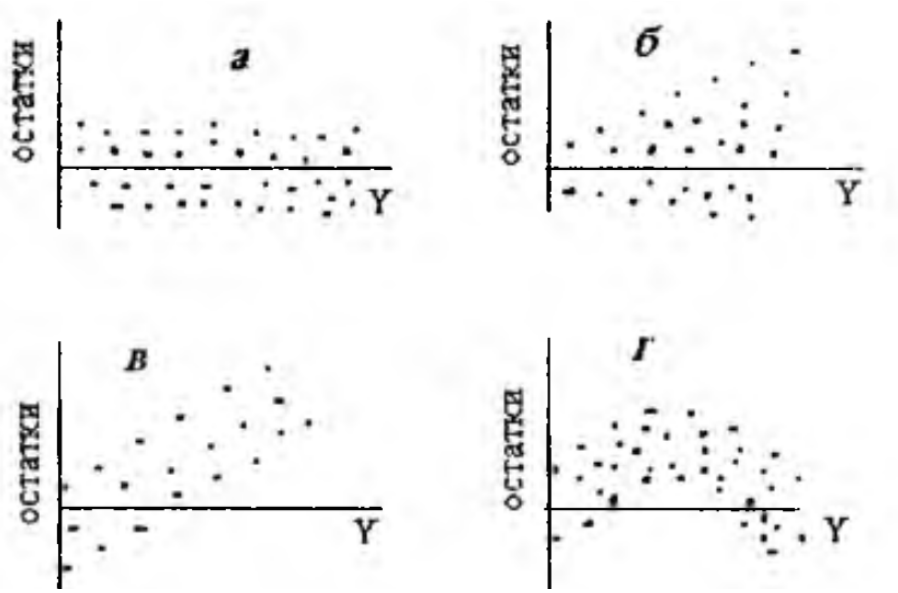
по регрессии. Анализ является из способов проверки модели или степени математической модели регрессии. остатки представляют временной ряд *случайных величин, распределенных по закону,* это служить обоснованием уравнения для прогноза. информативным в этих является представление зависимости от x или y. На графике остатки

вести себя хаотично, не быть резких , закономерностей в чередовании .



Если остатки в горизон полосу с центром по оси , то модель можно как адекватную. Если расширяется, то преобразование ряда Y.

, показывающий линейный даёт основание для в модель переменной. График , что в модель должен добавлен квадратичный .



1.11 Требования к данным

Для получения результатов при использовании множественной регрессии выполнение требований к исходной , соблюдение которых вообще не проверяется, в то как во многих они не выполняются или выполняются не .

Основные требования к наблюдений, следующие из математической , заключаются в следующем:

1) между всеми должны быть . Если связи очевидна, то рассмотреть или преобразование , или явно допустить нелинейных .

2) Исследуемые ряды подчиняться нормальному распределения. Близость распределения вы к нормальному является из главных показа надёжности математических основанных на метода наименьших .

3) Корреляция между должна отсутствовать или

незначительной. При тесной связи предикторами корреляционная становится вырождающейся, её стремится к , и возникают трудности в коэффициентов уравнения . Они становятся неустойчивыми. В случае *исключать дублирующие* .

4) Ряд-предиктант представлять собой значений величины, т.е. его значения быть не корелиованы собой. В применении ко рядам за

природными явлениями это не выполняется, так как для них характерно внутрирядной связности.

5) выборки в несколько раз превосходить независимых переменных (в 2-3 раза). Практика , что при использовании предиктора рядов *n* должна не менее 10, при двух предикторах длина должна составлять не 25-30, при четырёх - при пяти - 100-120 и т.д. в этом можно получить или менее надёжные параметров уравнения . Например, в уравнение регрессии м использоваться для практических при выполнении условий

$$R^2>0.5, \frac{R}{q_r}>2, \frac{b_i}{q_i}>2 \quad ,$$

где q_r - ошибка множественного коэффициента ; q_i — стандартная ошибка коэффициента уравнения.

1.12 множественной регрессии при решении задач

Рассмотрим , целью которой восстановление данных по уравнению ной линейной регрессии рек-аналогов. Имеется временных со среднегодовыми расходами рек, водосборные бассейны расположены на по физико- условиям территории. наблюдений составляет 24 . В одном из рядов пропуски. таблицу исходных в виде матрицы ≪объект - признак≫ 24*8. В принято наблюдения по конкретному (объекту) размещать в .

Для получения удовлетворительных при использовании множественной регрессии димо выполнение требований к исходной . Поэтому на ом шаге исследований проверить однородность и зентативность рядов , а также и тесноту связей ними. После корреляционной матрицы, легко в пакете Statistica, и на её основе дублирующих остались две независимые , удовлетворяющие ованиям, предъявляемым к как предикторов друг с , так и их связям с предиктантом. уравнение регрессии используется для пропущенных данных, то нт корреляции между и предиктантом быть не менее Независимые переменные Q2 и Q3 не *(r* = -0,08) и характеризуются связью с зависимой Q, (r = 0,76 и 0,84), поэтому с

зрения математических т их можно использовать в рек-аналогов. показывает, что при использовании предикторов минимальная рядов должна влять не 25-30. Только в случае можно по более или менее оценки уравнения регрессии. выборки в нашем (n = 24) намного превосходит независимых (m = 2), и можно рассчитывать на удовлетворительных результатов. Для адекватности модели ост проверить ряд на независимость и соответствие закону распределения. В для проверки независимости используется Дарбина - Уотсона, стандартным обнаружения их автокоррелир. Статистика - Уотсона d используется для гипотезы о том, что остатки п регрессионной модели не (корреляции нулю), против : остатки связаны зависимостью. Сравнение d = 2,35 с DW1 и DW2 из таблицы точек статистики - Уотсона при уровне p = 0,05 позволяет сделать об отсутствии внутрирядных связей в остатков. Стандартным проверяем соответствие остатков закону: визуально по , нормальному вероятностному и по теоретическим критериям - Смирнова, и Шапиро - Уилкса. Для прогноза введём обучающей и контрольной .

Обучающая - это просто матрица данных, на основе вычисляются коэффициенты регрессии;

К выборка - это совокупность , которые не использовались для регрессионных коэффициентов. Y по значениям из обучающей выборки прогнозом на зависимом але, а по данным из контрольной - на независимом.

2 имеющихся программных для многомерного анализа

2.1 Система Online Processing

(англ. Online Transaction), транзакционная система - транзакций в реальном . Способ БД, при котором система с небольшими по размерам транзакциями, но большим потоком, и при клиенту требуется от минимальное время .

Термин OLTP также к системам (). OLTP-системы для ввода, структурированного и обработки информации (, документов) в режиме реального времени.

- приложениями широкий спектр во многих отраслях - автоматизированные банковские системы, ERP-системы (системы планирования ресурсов предприятия), и биржевые операции, в - регистрация детали на конвейере, фиксация в посещений очередного веб-сайта, автоматизация , складского учёта и документов и т.п. Приложения , как правило, автоматизируют , повторяющиеся задачи данных, как ввод заказов и транзакции. OLTP-системы , настраиваются и оптимизируются для максимального транзакций за короткие времени. Как правило, гибкости здесь не , и чаще используется фиксированный надёжных и безопасных ввода, модификации, данных и оперативной отчётности. эффективности является транзакций, выполняемых за . Обычно возможности OLTP-систем ограничены (либо отсутствуют).

Требования OLTP:

- нормализованные модели данных;

- При возникновении ошибки должна целиком и вернуть систему к , которое до начала транзакции;

- данных в реальном времени.

Преимущества :

Высокая надёжность и данных, как транзакционного подхода. либо совершается и успешно, либо не и система к предыдущему состоянию. При исходе выполнения целостность данных не .

Недостатки OLTP:

OLTP-системы для небольших дискретных а вот запросы на некую информацию (к поквартальная динамика продаж по определённой товара в определённом), характерные для приложений (OLAP), породят соединения таблиц и таблиц целиком. На такой уйдет масса и компьютерных ресурсов, что обработку текущих .

2.2 Программное Statistical Package for the Sciences

SPSS (аббревиатура англ. Statistical for the Social - статистический для социальных наук) - компьютерная программа для статистической обработки , один из лидеров в области статистических продуктов, для проведения прикладных в социальных науках. По мнению некоторых , SPSS « ведущее положение программ, предназначенных для тической обработки ».

Возможности :

- Ввод и хранение ;
- Возможность исполь переменных разных ;
- Частотность , таблицы, графики, т сопряжённости, диаграммы;
- Пе описательная статистика;
- исследования;
- данных маркетинговых .

Преимущества SPSS:

-У графический интерфейс;

- О на социальных .

Недостатки SPSS:

- Д лицензий;

- Отсутствие в расчетах.

2.3 Программа для анализа «»

Statistica - программный пакет для статистического , разработанный компанией , реализующий функции анализа данных, управления данных, добычи данных, визуализации данных с статистических методов. Пакет широкими графическими , позволяет выводить в виде различных графиков (научные, деловые, и двухмерные графики в системах координат, ванные графики - гистограммы, , категорированные графики и др.), все графиков настраиваются.

Statistica:

- В параллельной работы в модулях;

- Выпущено литературы по работе с ;

- Понятный ;

- Содержит набор для базовой эконометрики;

- П русифицированной справочной ;

- Наличие версии и возможность собственных макросов;

- Б;

- Легкий импорт/экспорт в электронные и процессоры.

Недостатки :

- Высокая цена;

- О вкладок и кнопок в окнах воспроизводимость моделей;

- В параллельной обработки подгрупп данных в последних ;

2.4 Утилита для статистического Eviews 8

В данный последними версиями являются 8. Пакет представляет возможности при анализе рядов и панельных , что позволяет его в эконометрических исследованиях. К данного программного можно отнести недорогой версии (одногодичная Eviews стоит Интерфейс программы , как правило, осваивается студентами. с изучением командного возникают у студентов редко. В с руководством по полному методов, программа модули «Моделирование и » и «Анализ рядов», на базе возможно пост моделей временных .

Приемущества 8:

- Возможность одновременной с несколькими файлами;

- С огромный набор методов для эконометрики;

- Подробная (но не) справочная система;

- Л в освоении командный и интерфейс;

- Б;

- Легкая воспроизводимость и получения графиков;

- В создания собственных ;

- Доступная студенческой версии.

Eviews 8:

- Отсутствие версии и русифицированной системы;

- Мало литературы по работе в .

3 Создание базы информации

3.1 Директивы для данных

При данных постоянно необходимость в преобразовании или промежуточных данных. С целью следующие наборы .

1) Директивы для работы с , позволяющие производить строк, со нечисловые элементы, элементов, транспонирование, /слияние таблиц, строк/столбцов, строк/столбцов по столбцу/строке, строк/столбцов и т. д.

2) для работы с , позволяющие производить их , нормирование, выполнять операции, квант выравнивание, Мантеля и другие .

3) Директивы для вычисления мер /различия метрики , евклидовой , расстояния Жаккара, Джукса - Кантора, ра Кимуры и т. д. Часть директив указания набора строк/. Для выполнения задачи разработан , в котором выделить следующие особенности:

- возможность среди неуникальных ;
- возможность как лексикографического диапазона, так и по абсолютным значениям;
- указывать в качестве номера как буквенный , используемый в MS , так и десятичное число.

3.2 для многомерного анализа

Функционально из для многомерного анализа выделить директивы для размерности данных с потерями информации:

- метод компонент;
- метод координат;
- неметрическое шкалирование.

для анализа взаимосвязи :

- дискриминантный анализ;

- линейная регрессия;

- сети с распространением ошибки.

для ПЛС-анализа:

- 2B-PLS-анализ;

- .

Директивы для кластеризации:

- объединения;

- ближайшего соседа.

В реализации программного в рамках данной квалификационной были рассмотрены базы данных следующими сервисами:

- Nations (hdr.undp.org);

- Федеральная государственной статистики (www.gks.ru);

- Fire Power ();

- Central agency (www.cia.gov).

3.3 База United Nations

United Nations (Программа ООН (ПРООН)) - организация при ООН по помощи странам-участницам в развития. ПРООН помощь в проведении изысканий и природных ресурсов, в создании учебных , в развитии энергетических , предоставляет и экспертные услуги, специалистов, поставляет и т. д. Помощь ПРООН .

3.4 Базы сервиса Global Power

Рейтинг Firepower является из самых и авторитетных исследований в . Авторы этого самым тщательным изучают аспекты армий и выносят свой . Рейтинг стран с по военной составляется с использованием « мощи» (Power или PwrIndex). При анализе потенциала страны учитываются

различных параметров, в одну формулу. подсчетов число, достаточно отражающее потенциал той или страны. По мере военной страны ее PwrIndex и стремится к нулю. образом, чем меньше индекс государства, тем большей мощью оно располагает.
В подсчета индекса мощи 50 различных параметров, состояние экономики, и непосредственно вооруженных сил. того, при индекса применяется бонусных и штрафных . Также авторы Global учитывают некоторые государств, которые серьезно повлиять на .

Так, при подсчетах следующие правила:

- в страны не учитываются вооружения;

- при подсчетах во внимание особенности государств;

- не только количественные вооруженных сил;

- учитываются и потребление основных ресурсов;

- не выхода к морю не штрафуются за отсутствие ских сил;

- возможности ВМС являются для штрафа;

- не принимаются во политический курс и иные факторы.

База содержит полную по каждому вооруженному каждой , данных о количестве техники и прочего любого вида для все определённого .

3.5 Базы данных но-технического управления intelligence agency

управление ЦРУ одним из четырех структурных подразделений ЦРУ, и которых выполняют ЦРУ. Они решают путем эффективной целей, применения технологий и профессионального . Они создают, , разрабатывают и эксплуатируют технического сбора и высокоэффективные технологии для , обработки и информации.

База Central intelligence содержит максимально информацию социально-экономических аспектов государства. Информацию о руководстве всех структур.

3.6 данных федеральной государственной статистики

служба государственной (Росстат) - федеральный орган власти, осуществляющий по формированию официальной статистической информации о , экономическом, демографическом и экологическом положении страны, а функции по контролю и в области государственной деятельности на территории Федерации.

федеральной службы статистики предоставляет базы данных как :

- Центральная база статистических данных (ЦБСД);
- Единая межведомственная информационно – статистическая система (ЕМИСС);
- Показателей муниципальных образований;
- Список витрин данных.

4 Описание программного продукта

Идея программного заключается в том, что пользователь на скриптовом составляет программу, производит анализ . Скриптовый язык не наличия у ателя навыков , поэтому реализованы основные конструкции, как циклы, и вызов функций с .

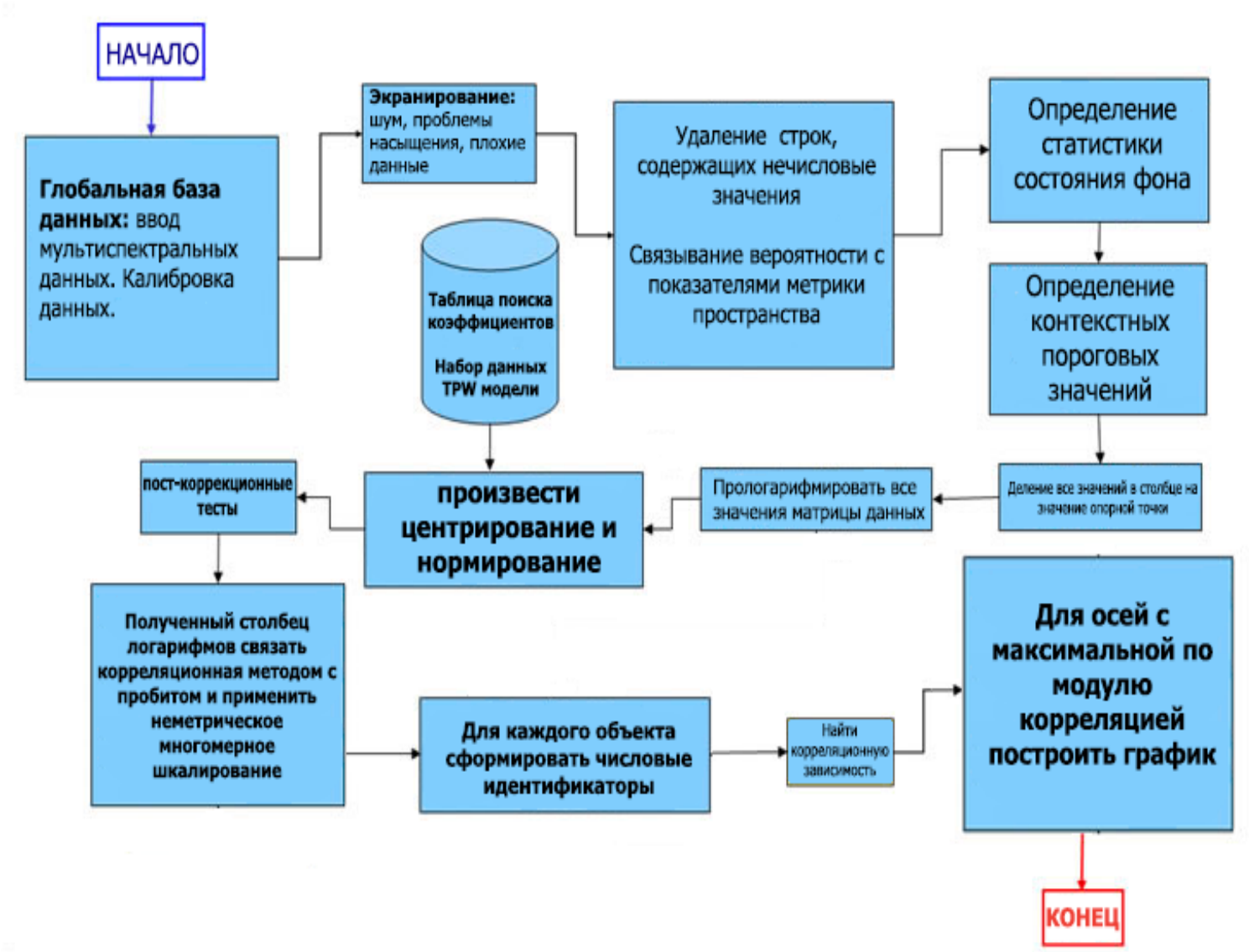


Рис. 7. Схема работы программного продукта

, что свой пользователь сможет в программе, аналогичной . При этом каждая записывается в ой ячейке. Предусмотрены следующих видов: менная, присваивание, последовательность, директивы, переменная , границы цикла, множества для цикла по , конец , слова, обозначающие и конец скрипта.

могут начинаться в месте. атель должен скрипт в формате csv(– точка с запятой). формат достаточно простым скрипта как в Excel, так и в текстовом редакторе, формат csv.

Для чтобы большой можно было в несколько этапов, понятие , который ограничивается словами «НАЧАЛО» и «». Если вскрипте место последовательность этих двух , то программа выдаст об ошибке.

Реализованный предусматривает определения переменной. может определяться один раз. Исключение переменные . Это сделано для удобства , потому что переменные, в , используются для сокращения строки.

, пользователь хочет файл D:\myDocuments\Folder\NextFolder\.csv, но не хочет много раз такое название. В этом он может до использования этого определить переменную: ; = ; D:\Bsu\Vkr\article\\GlobalFirePower.csv

После везде вместо «GlobFP» программа подставлять D:\Bsu\Vkr\\FirePower\GlobalFirePower.csv.

говоря, синтаксис описывается так: «Имя_ ; = ; значение».

В представлена возможность для с циклами. Одним из задания цикла цикл по , пробегающей целые в указанном пользователем , включая границы (табл. 1).

. 4.1 Пример цикла

LOOP BEGIN		1	7
log	B_<<index>>.csv	_<<index>>_	2
LOOP END			

Результатом такого цикла будет файлов B_1_LOG.csv, , в которых результаты работы для файлов B_1.csv, соответственно. На этом же можно особенность использования .

Индекс может и в качестве строки, в как назв файла, и в качестве . Однако чаще требуется его использование в качестве . В некоторых случаях потребность реализовать набор действий количество раз. Для предусмотрены циклы, в переменная не фигурирует, а лишь количество . Пример го цикла представлен на 2.

В результате работы цикла в файле sv будет 20 бутстрепов выборки, в input.csv.

Табл. 4.2 ер цикла по количеству

LOOP	20	
bootstrep	input.csv	
addtofile	output.csv	
LOOP END		

В реализованном продукте существует возможность я списка строковых для переменной цикла. циклы ваются циклами по (Таблица 3).

Табл. 4.3 цикла по множеству

OVER	index	file1.csv		f.csv	file3.csv	
log	index	LOG_<<>>	2			
LOOP END						

Рез работы этого (таблица 4.3) будут LOG_file1.csv, LOG_u.csv, , LOG_file3.csv, , в которых записаны по основанию 2 таблицы из file1.csv, u.csv, , file3.csv, соответственно.

Язык возможность задания любой глубины. При переменная цикла может оваться как граница .

4.1 Пользовательские скрипты

предусматривает создания пользовательских с параметрами, которые собой скрипты с входных . Глубина вложенности не ограничена. Пакет планируется поставлять с наборами скриптов-подпрограмм для включения в пользователей.

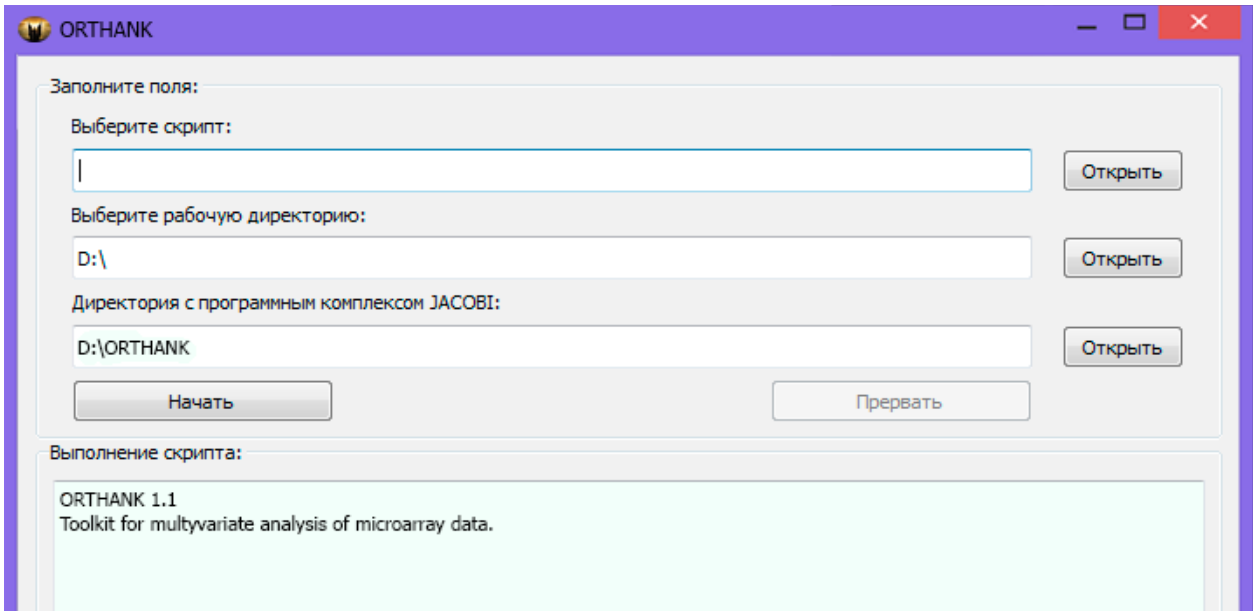


Рис. 4.2 Пользовательский

4.2 Пользовательский интерфейс

продукт имеет графический , ориентированный на пользователя, не большим опытом с ПК. Полноценный на данный момент на стадии разработки. бета-версия интерфейса доступна, с версии 1.3 (рис. 4.2).

4.3 возможности

Одним из работы утилиты является индивидуальных пользовательских имен директив. пользователи, работающие с пакетами для анализа, тратят времени на запоминание в случаев неочевидных операций. предоставляет пользователю самостоятельно изменять директив. Таким образом, будет с теми именами , к которым он привык, или с , которые ему кажутся для запоминания. принцип позволяет э много времени, обычно тратится на информации в и запоминание.

Кроме , пользователь может изменить набор в пакете. Как результаты тестирования, предпочитают иметь в только те директивы, им нужны. Это поиск, уменьшает на написание скриптов и воспринимается на подсознательном .

Частным изменения набора является добавление программ в пакет. Ни из распространенных не предоставляет возможности, однако в случаев ввиду задачи или ны метода функция быть не включена в программы. В Orthank изована так обертка - программа, позволяет работать со программами как с собственными пакета.

4.4 работы программного

В качестве примера задачу определения коэффициента мощи, взяв в такие страны как Федерация, США и Китайская Республика. Для целей возьмем предоставленные базами Global FirePower и службой статистики.

Опишем действий, которую выполнить для достижения цели. применением неметрического шкалирования необходимо подготовку данных, посчитать , таким образом, быть выполнены действия:

1) оставить , соответствующие и к общему количеству ;

2) удалить все строки, нечисловые значения;

3) вероятность с метрики пространства;

4) все значения в столбце на опорной точки;

5) все значения данных;

6) произвести и нормирование;

7) полученный логарифмов связать методом с ;

8) применить неметрическое шкалирование;

9) для каждого сформировать числовые ;

10) найти зависимость;

11) для осей с по модулю корреляцией график.

Пункты 1-10 в пакете, для этого следующий скрипт:

Копировать колонки	17.csv	1.CN.csv	ex.csv	[$1..
Удалить с нечисловыми значениями	1.	2.numbers.only.csv		
Логарифмировать	2. .csv	3.log2.csv	2	
Центрировать	3.	4.1.centre.csv		
	4.1.centre.csv	4.2.normalize.csv		
	4.2.normalize.csv	4.3.transpose.csv		
с пробитом	4.3. communicat .csv	5. c.csv	5	
Неметрическое	5.communicat.csv	6.	4	0.99

шкалирование				
Копировать колонки	.csv	7.1.2columns.csv	ex.csv	[$9;$27]
строки	7.1.2	7.4.CN.rows.csv	ex.csv	[$1..
Подпрограмма	заменить строк	7.4.CN.rows.csv	7.8.	[$2]
	6.nmds.csv	6.1.transposed.csv		
	7.8.grades.csv	7.9.transposed.csv		
	6.1.transposed.csv	7.9.transposed.csv	8.	
Конец				

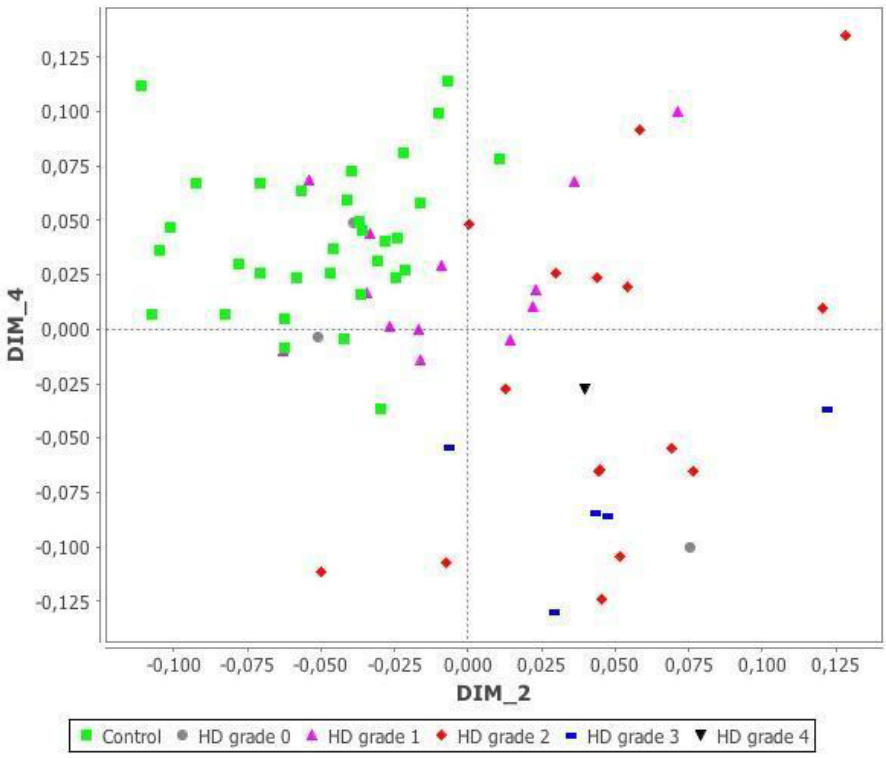


Рис. 4.3 корреляционной зависимости формирований (США, КНР, РФ)

В работы скрипта был файл, оси и файл, вклю значение корреляции к оси с логистическими и географическими , наличием ресурсов, наличием к морю и разнообразием . Для построения графика выбраны две оси с по модулю корреляцией: _2 с коэффициентом корреляции, 0,71573, и DIM_4 с корреляции, минус 0,54624.

4.5 дальнейшего развития продукта

К сентябрю г. планируется всех подсистем , механизма дельта- и создание базы для хранения . К 2019 г. - полный пакета на кластер, его функционала и дальнейшая пользовательского .

ЗАКЛЮЧЕНИЕ

ВКР была разработке метода, и программных компонентов для многомерного данных. Реализован продукт Orthank, быстро и эффективно однотипную для множества входных . Orthank протестирован пользователей с различным работы с ПК. В тестирования выявлен и ряд ошибок. Кроме , были учтены п и предложения, составили новые для комплекса.

Разработанный продукт позволит:

- данные распространенных баз;
- получать в табличном виде;
- данные в графическом ;
- обрабатывать данные различными анализа (такими как и кластерный);
- выводить в наиболее формате csv.

На основании можно сделать о том, что разработка метода, и программных для программы многомерного является целесообразной, и приносить реальную при его использовании алгоритма и программного .

В процессе выполнения получены следующие :

1. Разработан многомерного анализа ;
2. Разработан алгоритм данного метода;
3. программный реализующий разработанный многомерного анализа ;
4. Разработан способ полученных в графическом виде;

ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. « статистический анализ в задачах. моделирование в SPSS», учебник, 2009 г.

2. А.И. «Прикладная статистика» М.: «Экзамен»,

3. Фишер Р.А. «Статистические для исследователей», 1954 г.

4. В.Н., Соловьев В.И. «Введение в статистический » Учебное пособие , 2003;

5. Ахим Бююль, Цёфель, «SPSS: обработки » Изд-во DiaSoft, 2005.;

6. Итан Браун. с применением Node и . Полноценное и стека JavaScript. – Web De with Node and /Итан Браун.; - : Питер, – 336 с.

7. Айвазян С.А. Методы : учеб. – М. Магистр. , 2014.

8. Гафарова Е.А. прикладных при обучении эконометрическим // Современные проблемы и образования. – 2014. – № 6.

9. Д.Е. Программное эконометрического исследования // Нижегородского университета им. Н.И. , 2011, № 3 (2), с. 231–238.

10. А.И. Эконометрика: у для вузов / А.И. Орлов. – н/Д : Феникс, 2009. – 276 с.

11. А.Н., Орлова И.В., Математические в управлении: пособие - М.: Вузовский : ИНФРА-М, 2012. – 272 с.

ресурсы:

1. Айвязян С.А., Ц.С. Программное по статистическому анализу : Методология сравнительного и выборочный обзор .- Режим : http:// pub-health.spb.ru /SAS /.htm.
2. Data Analysis and Software/ StataCorp LP. 1996–2014. URL: http://www.stata.com.
3. EViews.com / IHS Global Inc. 2013. URL: http://www.eviews.com
4. Gnu Regression, Econometrics and Time-series Library/ Allin Cottrell, Wake Forest University. Riccardo "Jack" Lucchetti, Università Politecnica delle Marche. 2014. URL: http://gretl.sourceforge.net.
5. Predictive Solutions/ Predictive Solutions. 2012. URL: http://www.predictivesolutions.ru.
6. Prognoz/ JSC Prognoz. 2005–2014. URL: http://www.prognoz.ru
7. StatSoft Russia/ StatSoft Russia. 2014. URL: http://www.statsoft.ru