Казанский Федеральный Университет

Кафедра высоковязких нефтей и природных битумов

Kazan Federal University,

Department of high-viscosity oils and natural bitumen

Основы хемометрии. Методы редукции данных

Fundamentals of chemometric. Data reduction methods

Кемалов Руслан Алимович, Kemalov Ruslan Alimovich [a]

Кемалов Алим Фейзрахманович, Kemalov Alim Feizrahmanovich [b]

кандидат технических наук, доцент кафедры высоковязких нефтей и природных битумов, Член Экспертного совета РГО, и.о. руководителя группы «Водородная и альтернативная**,** [a]

доктор технических наук, профессор, заведующий кафедрой высоковязких нефтей и природных битумов [b]

Казань, Россия

E-mail: kemalov@mail.ru

Аннотация: в работе изучены области и методология редукции данных, факторный анализ, дано описание модуля Factor Analysis, приведена методология анализа главных компонент, дано описание модуля Principal Components&Classification Analysis.

Ключевые слова: области и методология редукции данных, факторный анализ, модуль Factor Analysis, приведена методология анализа главных компонент, модуль Principal Components&Classification Analysis

Abstract: the paper examines the areas and methodology of data reduction, factor analysis, describes the Factor Analysis module, describes the methodology of the analysis of the main components, describes the module Principal Components & Classification Analysis.

Keywords: areas and methodology of data reduction, factor analysis, the description of the Factor Analysis module, the methodology of the analysis of the main components, the description of the module Principal Components & Classification Analysis

Введение (Introduction). Методы редукции данных

1. Факторный анализ

Главными целями факторного анализа являются сокращение числа переменных (редукция данных) и определение структуры взаимосвязей между переменными, т.е. классификация переменных. Поэтому факторный анализ используется или как метод сокращения данных, или как метод классификации переменных.

Сокращение достигается путем выделения скрытых общих факторов, объясняющих связи между наблюдаемыми признаками (переменными) объекта, т.е. вместо исходного набора переменных появится возможность анализировать данные по выделенным факторам, число которых значительно меньше исходного числа взаимосвязанных переменных.

Взаимосвязи между переменными можно обнаружить с помощью диаграммы рассеяния. Полученная путем подгонки линия регрессии дает графическое представление зависимости. Если определить новую переменную на основе линии регрессии, изображенной на этой диаграмме, то такая переменная будет включать наиболее существенные черты обеих переменных. Итак, произошло сокращение числа переменных — две заменили одной. Причем новый фактор (переменная) является линейной комбинацией двух исходных. Приведенный пример, в котором две коррелированные переменные объединены в один фактор, показывает главную идею факторного анализа.

В основном процедура выделения факторов подобна вращению, максимизирующему дисперсию исходного пространства переменных. Например, на диаграмме рассеяния можно рассматривать линию регрессии как ось X, повернув ее так, что она совпадает с прямой регрессии. Этот тип вращения называется вращением, максимизирующим дисперсию (варимакс), так как цель вращения заключается в максимизации изменчивости новой переменной (фактора) и минимизации разброса исходных переменных. Если пример с двумя переменными распространить на большее число переменных, то вычисления становятся сложнее, однако основной принцип представления двух или более зависимых переменных одним фактором остается в силе.

Число наблюдаемых объектов может быть большим и взаимосвязи между ними чрезвычайно сложными. Однако наблюдая объект, выдвигаем гипотезу, что существует небольшое число факторов, которые влияют на измеряемые параметры. Естественно желание выделить как можно меньшее число скрытых общих факторов и чтобы выделенные факторы как можно точнее приближали наблюдаемые параметры, описывали связи между ними.

Выделяемые таким образом факторы называют общими, так как они воздействуют на все признаки (параметры) объекта, а не на какой-то один признак или группу признаков. Эти факторы являются гипотетическими, скрытыми, их нельзя измерить непосредственно, однако существуют статистические методы их выделения.

Описание модуля Factor Analysis

В меню Statistics щелкните по Multivariate Exploratory Techniques (многомерные исследовательские методы) и выберите команду Factor Analysis (анализ факторов). Откроется стартовая панель модуля. Рассмотрим все его компоненты и опишем некоторые из них. В поле Input File (файл входных данных) надо указать тип исходного файла, с которым предстоит работать. В модуле возможны следующие типы исходных данных:

- *Correlation Matrix* (корреляционная матрица);
- *Raw Data* (исходные данные).

Выберите, например, *Raw Data.* Это обычный файл данных, где по строкам записаны значения переменных. В правом нижнем углу окна, за всеми функциональными кнопками находится поле *MD deletion* (обработка пропущенных значений). В этом поле необходимо задать один из способов, которым будут обрабатываться при анализе пропущенные значения (незаполненные ячейки):

- *Casewise* (способ исключения пропущенных случаев);
- *Pairwise* (парный способ исключения пропущенных значений);
- *Mean Substitution* (подстановка среднего вместо пропущенных значений).

Способ *Casewise* состоит в том, что в электронной таблице, содержащей данные, игнорируются все строки (наблюдения), в которых имеется хотя бы одно пропущенное значение. Это относится ко всем переменным. Итак, в таблице остаются только те наблюдения, в которых нет ни одного пропуска.

В способе *Pairwise* игнорируются пропущенные наблюдения не для всех переменных, а лишь для выбранной пары. Все наблюдения, в которых нет пропусков, используются в обработке, например, при поэлементном вычислении корреляционной матрицы, когда последовательно рассматриваются все пары переменных.

Способ *Mean Substitution* предполагает при выполнении анализа заполнение пустых клеток средними значениями.

Очевидно, в способе *Pairwise* остается больше наблюдений для обработки, чем в способе *Casewise.* Тонкость, однако, состоит в том, что в способе *Pairwise* оценки различных коэффициентов корреляции строятся по различному числу наблюдений. Выберите, например, способ *Casewise.*

Дальнейшее рассмотрение требует работы уже с конкретными данными, поэтому следующим действием откройте файл, содержащий исходные данные для анализа (если он еще не открыт).

В качестве примера рассмотрите имеющийся в программе *STATISTICA* файл Factor.sta из библиотеки Examples. Об этом файле шла речь при изучении модуля Canonical Analysis. Теперь, когда есть данные для анализа, выбран способ обработки пропущенных значений, перейдем к выбору переменных, для которых будем проводить факторный анализ.

Для того чтобы сделать это, задействуйте кнопку Variables. Появится окно выбора переменных Select the variables for the factor analysis (выбрать переменные для факторного анализа). Кнопка Select All (выбрать все) позволяет выбрать все переменные сразу.

Щелкните в стартовом окне модуля кнопкой ОК. Программа начнет анализ выбранных неременных, появится окно Define Method of Factor Extraction (определить метод выделения факторов). В информационной части окна (рис.1) сообщается, что пропущенные значения обработаны методом *Casewise.*

Обработано 100 случаев и 100 случаев принято для дальнейших вычислений. Корреляционная матрица вычислена для 10 переменных. Нижняя часть текущего диалогового окна состоит из трех вкладок. Выделите вкладку Descriptives, так как факторный анализ надо начинать с вычисления корреляционной матрицы. Ее анализ позволит оценить степень коррелированности переменных между собой. И если эта степень окажется высокой, то данные переменные можно объединять в один фактор. А процедура вычисления корреляционной матрицы доступна именно из этого окна.

Кнопка Review corelations, means, standard deviations предназначена для построения корреляционной матрицы, вычисления средних, стандартных отклонений.

Кнопка Compute multiple regression analyses осуществляет запуск процедуры множественного регрессионного анализа.

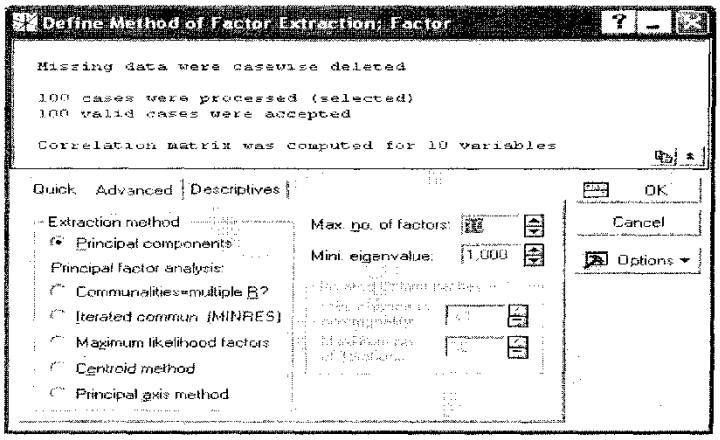

Рисунок 1 - Информационная часть окна

Нажмите кнопку Review corelations, means, standard deviations. Откроется окно Review Descriptive Statistics (обзор описательных статистик), на вкладке Quick (Advanced) нажмите кнопку Correlations. На рис. 2 изображен фрагмент корреляционной матрицы, из которого видно, что коэффициенты корреляции переменных *WORK* с переменными *HOME* имеют малые значения, в то время как с другими группами переменных принимают большие значения. Этот факт отразится на результатах последующих этапов факторного анализа.

Нажмите кнопку Cancel и вернитесь в исходное окно Define Method of Factor Extraction. Выделите вкладку Advanced, на этой вкладке имеются следующие поля:

- Maximum no. of factors (максимальное число факторов);
- Minimum eigenvalue (минимальное собственное значение).

В поле Minimum eigenvalue устанавливается минимальное собственное значение, т.е. если собственные значения окажутся меньше, чем установленный здесь минимум, то они игнорируются.

В поле Maximum no. of factors пользователь устанавливает количество факторов, которые необходимо выделить для анализируемых данных. Можно установить любое значение, не превышающее количество переменных, но не любой полученный таким образом результат окажется правильным. Для того чтобы получить интерпретируемый результат, на практике используют несколько полезных критериев.

Variable	Correlations (Factor) Casewise deletion of MD N=100 WORK1	WORK2	WORK3	HOBBY1	HOBBY2	HOME1	HOME2	HOME3
WORK1	1,00	0,65	0,65	0,60	0,52	0,14	0,15	0,14
WORK2	0,65	1,00	0,73	0,69	0,70	0,14	0,18	0,24
WORK3	0,65	0,73	1,00	0,64	0,63	0,16	0,24	0,25
HOBBY1	0,60	0,69	0,64	1,00	0,80	0,54	0,63	0,58
HOBBY_2	0,52	0,70	0,63	0,80	1,00	0,51	0,50	0,48
HOME1	0,14	0,14	0,16	0,54	0,51	1,00	0,66	0,59
HOME2	0,15	0,18	0,24	0,63	0,50	0,66	1,00	0,73
HOME3	0,14	0,24	0,25	0,58	0,48	0,59	0,73	1,00
MISCEL1	0,61	0,71	0,70	0,90	0,81	0,50	0,64	0,59
MISCEL2	0,55	0,68	0,67	0,84	0,76	0,42	0,59	0,52

Рисунок 2 - Фрагмент корреляционной матрицы

В методе главных компонент по умолчанию предполагается, что дисперсии всех переменных равны 1. Тогда общая дисперсия равна общему числу переменных (для нашего примера — 10). Это означает, что наибольшая изменчивость, которая потенциально может быть выделена, равна 10. Максимально возможное число выделяемых факторов равно числу переменных. Каждому фактору соответствует дисперсия, объясненная этим фактором. Дисперсии, соответствующие факторам, называются собственными значениями.

Для просмотра собственных значений факторов в окне Define Method of Factor Extraction произведите следующие установки параметров: *Maximum no. of factors = 10* и *Minimum eigenvalue* = 0. Далее нажмите ОК. В открывшемся окне Factor Analysis Results нажмите кнопку Eigenvalues, появится таблица с собственными числами (рис. 3).

Во втором столбце таблицы приведены дисперсии выделенных факторов — собственные числа. В третьем столбце для каждого фактора приводится процент от общей дисперсии (в данном примере она равна 10). Как видно, первый фактор объясняет 61% общей дисперсии, второй фактор — 18% и т.д. Четвертый столбец содержит накопленную или кумулятивную дисперсию. Как только получена информация о том, сколько дисперсии выделил каждый фактор, можно перейти к вопросу, сколько факторов следует оставить.

Критерий Кайзера. Сначала можете отобрать только факторы с собственными значениями, большими 1. По существу это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. Этот критерий предложен Кайзером и является, вероятно, наиболее широко используемым. В приведенном примере на основе данного критерия выделяются только два фактора, так как остальные не подходят под условие, наложенное на собственные значения.

Value	Eigenvalues (Factor) Extraction: Principal components			
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative%
1	6,118369	61,18369	6,11837	61,1837
2	1,800682	18,00682	7,91905	79,1905
3	0,472888	4,72888	8,39194	83,9194
4	0,407996	4,07996	8,79993	87,9993
5	0,317222	3,17222	9,11716	91,1716
6	0,293300	2,93300	9,41046	94,1046
7	0,195808	1,95808	9,60626	96,0626
8	0,170431	1,70431	9,77670	97,7670
9	0,137970	1,37970	9,91467	99,1467
10	0,085334	0,85334	10,00000	100,0000

Рисунок 3 - Таблица с собственными числами

Критерий каменистой осыпи. Критерий является графическим методом, впервые предложенным Кэттелем.

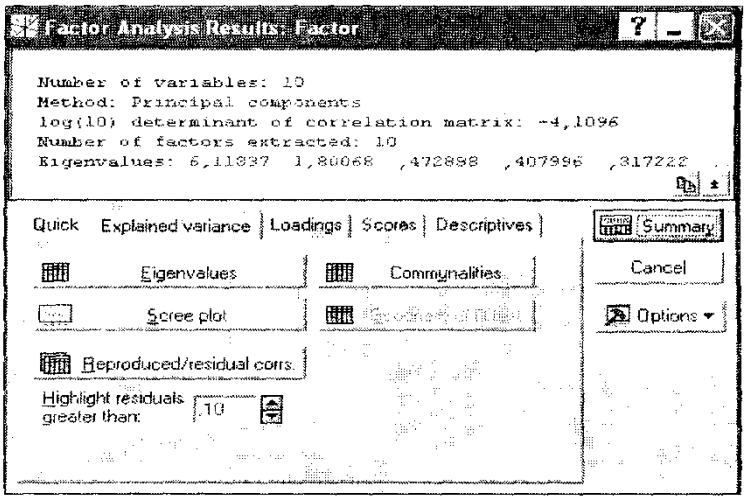

Рисунок 4 - Scree plot

Надо изобразить собственные значения, представленные в таблице в виде графика. Кэттель предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется, на вкладке Explained variance нажмите кнопку Scree plot (рис. 4).

Из построенного графика (рис. 5) видно, что в соответствии с этим критерием можно пытаться выделить 2 или 3 фактора.

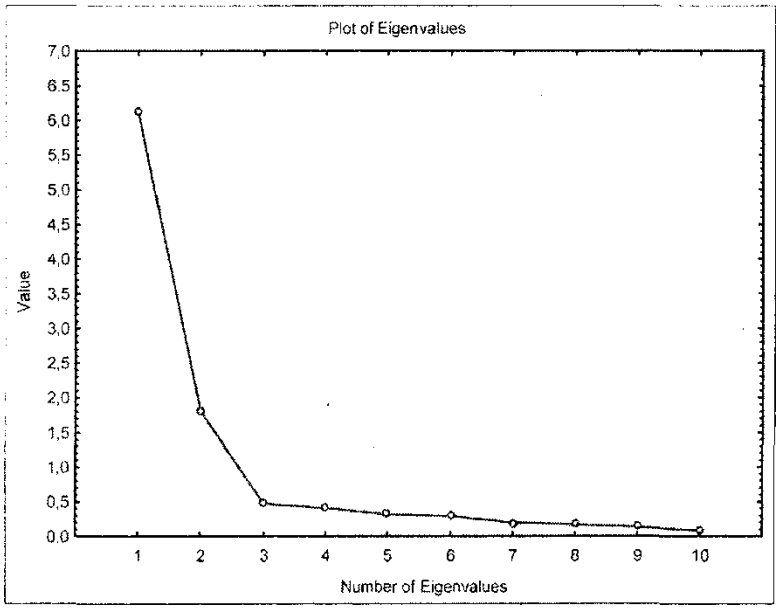

Рисунок 5 – Построение графика

Различные методы выделения факторов расположены на вкладке Advanced окна Define Method of Factor Extraction и объединены в группу опций под заголовком Extraction method (метод выделения).

В зависимости от критерия оптимальности возможен анализ либо методом *Principal components* (методом главных компонент), либо одним из методов, объединенных в группу *Principal factor analysis* (анализ главных факторов).

В группе *Principal factor analysis* предусмотрены следующие методы:

- *Communalities = multiple R**2* (общности равны квадрату коэффициента множественной корреляции);
- *Iterated Communalities (MINRES)* (итеративные общности или минимальные остатки);
- *Maximum likelihood factors* (максимальное правдоподобие);
- *Centroid method* (центроидный метод);
- *Principal axis method* (метод главных осей).

Выберите опцию *Principal components*. Чтобы лучше понять основные моменты факторного анализа, предположите, что неизвестны критерии определения числа факторов, и поэтому начните анализ с максимального числа факторов. Сохраните значения максимального числа факторов — 10 и минимального собственного значения — 0 (если собственное значение не будет установлено в 0, то количество выделенных факторов не будет равняться 10).

Щелкните кнопкой ОК, и на экране появится уже знакомое окно Factor Analysis Results. В верхней информационной части окна указаны:

- *Number of variables* (число анализируемых переменных);
- *Method* (метод анализа);
- *log(10) determination of correlation matrix* (десятичный логарифм детерминанта корреляционной матрицы);
- *Number of factor extraction* (число выделенных факторов);
- *Eigenvalues* (собственные значения). В нижней части окна находятся функциональные кнопки, позволяющие всесторонне численно и графически просмотреть результаты анализа.

Нажмите кнопку Summary. Factor loadings (итоги, факторные нагрузки), на рис. 6 приведен фрагмент таблицы с факторными нагрузками — корреляциями между переменными и выделенными факторами.

Из таблицы видно, что первому и второму факторам (*Factor 1, Factor 2)* соответствуют большие значения коэффициентов корреляции, чем остальным факторам. Причем с увеличением номера фактора значения коэффициентов корреляции стремительно уменьшаются. При правильно выбранном количестве факторов таблицы факторных нагрузок должны выявлять закономерности, проявляющиеся в следующем. Факторные нагрузки должны объединять переменные в группы, для которых коэффициенты корреляции с факторами принимают большие значения по одной группе и меньшие значения по другой.

	Factor Loadings (Unrotated) (Factor) Extraction: Principal components (Marked loadings are > ,700000)				
Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
WORK_1	-0,652601	0,514217	0,301687	0,439108	-0,0137
WORK_2	-0,756976	0,494770	-0,078826	-0,211795	-0,0908
WORK_3	-0,745706	0,456680	-0,104749	0,030826	-0,2049
HOBBY_1	-0,941630	-0,021835	0,012653	0,001861	0,1206
HOBBY_2	-0,875615	0,051643	0,099675	-0,324541	-0,0158
HOME_1	-0,576062	-0,604977	0,490999	-0,114927	-0,1125
HOME_2	-0,671289	-0,617962	-0,125776	0,159963	0,2250
HOME_3	-0,641532	-0,573925	-0,268572	0,152709	-0,3625
MISCEL_1	-0,951516	0,013513	-0,050164	0,026706	0,0767

Рисунок 6 - Factor loadings (итоги, факторные нагрузки)

Из сказанного следует нецелесообразность рассмотрения всех десяти факторов. Воспользуйтесь результатами этой таблицы, критерием Кэттеля, критерием Кайзера и назначьте число факторов — 2.

Из фрагмента таблицы результатов, приведенного на рис. 6, видно, что есть некоторая закономерность в значении факторных нагрузок, а именно группе переменных *WORK* соответствуют большие значения коэффициентов корреляции с фактором 1, чем с фактором 2. Аналогичные данные получим для групп переменных *HOBBY* и *MISCEL*. Но в такой форме выявленные закономерности трудно проинтерпретировать.

Чтобы получить интерпретируемое решение, надо применить повороты осей, которые достигаются вращением факторов. Как уже говорилось, бели пространство общих факторов найдено, то с помощью поворота системы координат в принципе можно получить бесчисленное множество решений. Конечно, такое количество решений — абсурд. Важно найти

интерпретируемое решение. Программа предлагает несколько способов вращения:

- *Varimax row* (варимакс исходных);
- *Varimax normalized* (варимакс нормализованных);
- *Biquartimax raw* (биквартимакс исходных);
- *Biquaitimax normalized* (биквартимакс нормализованных);
- *Quartimax raw* (квартимакс исходных);
- *Quartimax normalized* (квартимакс нормализованных);
- *Equamax raw* (эквимакс исходных);
- *Equamax normalized* (эквимакс нормализованных).

Метод варимакс предназначен для максимизации дисперсий квадратов исходных факторных нагрузок по переменным для каждого фактора, что эквивалентно максимизации дисперсий в столбцах матрицы квадратов исходных факторных нагрузок.

Целью метода биквартимакс является одновременная максимизация суммы дисперсий квадратов исходных факторных нагрузок по факторам и максимизация суммы дисперсий квадратов исходных фактоных нагрузок по переменным. Это эквивалентно одновременной максимизации дисперсий в строках и столбцах матрицы квадратов исходных факторных нагрузок.

Метод квартимакс означает максимизацию дисперсий квадратов факторных нагрузок по факторам для каждой переменной, что эквивалентно максимизации дисперсий в строках матрицы квадратов исходных факторных нагрузок.

Метод эквимакс можно рассматривать как взвешенную смесь вращения по методам варимакс и квартимакс, что эквивалентно одновременной максимизации дисперсий в строках и столбцах матрицы квадратов исходных факторных нагрузок. Однако в отличие от вращения по методу биквартимакс относительный вес, назначенный критерию варимакс при вращении, равен количеству факторов, деленному на *2*.

Дополнительный термин *normalized* (нормализованные) в названии методов указывает на то, что факторные нагрузки в процедуре нормализуются,

т.е. делятся на корень квадратный из соответствующей общности. Термин *raw* (исходные) показывает, что вращаемые нагрузки не нормализованы.

В поле Factor rotation окна Factor Analysis Results на вкладке Quick выберите метод поворота осей, например *Varimax raw,* и щелкните по Summary.

Из фрагмента таблицы факторных нагрузок (рис. 8) следует, что *Factor 1* имеет высокие факторные нагрузки по переменным *WORK* и низкие по переменным *HOME,* а *Factor 2* — наоборот: низкие по переменным *WORK* и высокие по переменным *HOME.* При этом факторные нагрузки, соответствующие переменным групп *HOBBY* и *MISCEL*, принимают промежуточные значения. Это и означает, что выделенные два фактора наилучшим образом характеризуют данные.

Variable	Factor Loadings (Unrotate Extraction: Principal comp (Marked loadings are > ,7 Factor 1	Factor 2
WORK_1	-0,652601	0,514217
WORK_2	-0,756976	0,494770
WORK_3	-0,745706	0,456680
HOBBY_1	0,941630	-0,021835
HOBBY_2	0,875015	0,051643
HOME_1	-0,576062	-0,604977
HOME_2	-0,671289	-0,617962
HOME_3	-0,641532	-0,573925
MISCEL_1	0,951516	0,013513
MISCEL_2	0,900333	0,048154
Expl.Var	6,118369	1,800682
Prp.Totl	0,611837	0,180068

Рисунок 7 - Таблица факторных нагрузок

Variable	Factor Loadings (Varima Extraction: Principal com (Marked loadings are > , Factor 1	Factor 2
WORK_1	0,830623	-0,019320
WORK_2	0,902408	0,058905
WORK_3	0,870524	0,082595
HOBBY_1	0,739857	0,582885
HOBBY_2	0,731191	0,484489
HOME_1	0,097371	0,829676
HOME_2	0,165722	0,897242
HOME_3	0,168370	0,844159
MISCEL_1	0,768988	0,560555
MISCEL_2	0,748801	0,502121
Expl.Var	4,561544	3,357507
Prp.Totl	0,456154	0,335751

Рисунок 8 - Таблица факторных нагрузок

Выявление и интерпретация закономерностей в таблицах факторных нагрузок — достаточно трудоемкий процесс. Процедура значительно упрощается, если использовать графическое представление факторных нагрузок. Нажмите кнопку Plot of factor loadings (двумерный график нагрузок). График, представленный на рис. 9, иллюстрирует соотношение между факторами и группами переменных. Видно, что группа переменных *WORK* занимает на плоскости крайнее левое верхнее положение, а группа переменных *HOME* — крайнее правое нижнее положение. Следовательно, *Factor 1* отвечает за удовлетворение, получаемое на работе, а *Factor 2* измеряет

удовлетворенность домашней жизнью. Поэтому можно сделать вывод, что общая удовлетворенность исследуемой группы людей, в основном, определяется двумя факторами — удовлетворенностью работой и удовлетворенностью домом.

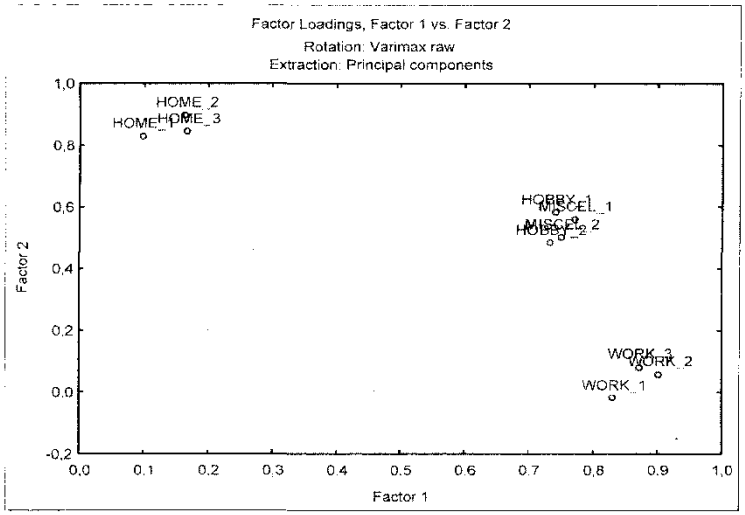

Рисунок 9 - Соотношение между факторами и группами переменных

В диалоге Factor Analysis Results перейдите на вкладку Scores (рис. 10). Нажмите кнопку Factor Score coefficients, откроется таблица с коэффициентами линейных уравнений регрессий (рис. 11), по которым программа посчитает значения факторов для каждого наблюдения (респондентов).

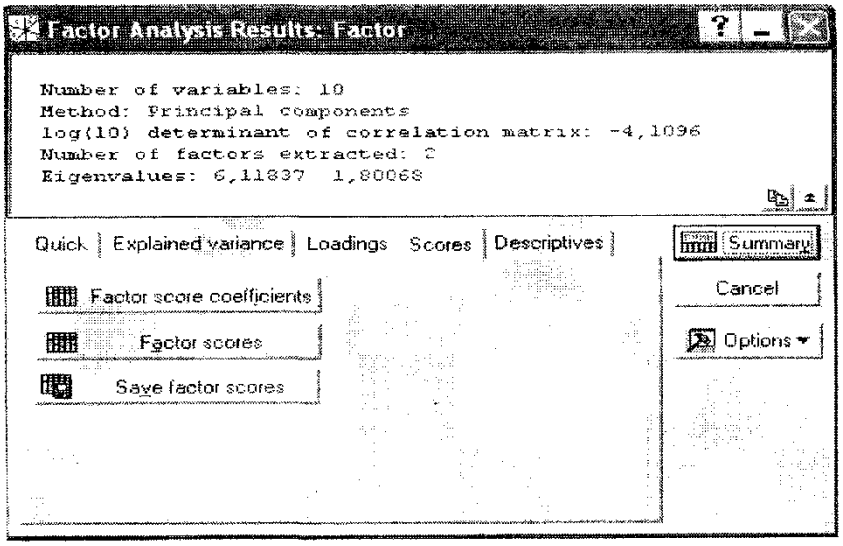

Рисунок 10 - Диалог Factor Analysis Results

Нажмите кнопку Factor Scores, появится таблица (рис. 12), в которой отображены значения факторов для каждого респондента. По этим значениям можно судить об отношении респондентов к *Factor 1* и *Factor2.* Положительное значение фактора соответствует позитивному отношению

респондента, а отрицательное — негативному.

Величина положительного фактора соответствует силе предпочтения данного фактора (для отрицательного — наоборот). Таким образом, процедура редукции данных позволила выделить два значимых фактора — *Factor 1* и *Factor 2* и сократить число переменных с 10 до 2.

Variable	Factor Score Coefficients (F Rotation: Varimax raw Extraction: Principal compor Factor 1	Factor 2
WORK_1	0,256768	-0,164304
WORK_2	0,263925	-0,145425
WORK_3	0,249750	-0,129616
HOBBY_1	0,115785	0,102111
HOBBY_2	0,131660	0,063002
HOME_1	-0,126453	0,325194
HOME_2	-0,118337	0,340306
HOME_3	-0,107542	0,317830
MISCEL_1	0,128865	0,087384
MISCEL_2	0,133727	0,066977

Рисунок 11 - коэффициенты линейных уравнений регрессий

Case	Factor Scores (Factor) Rotation: Varimax raw Extraction: Principal comp Factor 1	Factor 2
1	0,77326	-0,59909
2	-1,95924	-0,42839
3	-1,31803	-0,13560
4	0,17915	-0,70837
5	0,08277	-1,64135
6	-1,42460	0,42254
7	-0,19411	-0,39425
8	0,95212	-1,13020
9	0,03346	-0,20582
10	-0,70690	-0,41079
11	-0,18579	-1,75809
12	0,23559	1,19109

Рисунок 12 - Значения факторов для каждого респондента

Метод анализ главных компонент

На практике часто возникает задача анализа данных большой размерности. Метод анализ главных компонент и классификация позволяет решить эту задачу и служит для достижения двух целей:

- уменьшение общего числа переменных (редукция данных) для того, чтобы получить *«главные»* и *«некоррелирующие»* переменные;
- классификация переменных и наблюдений, при помощи строящегося факторного пространства.

Данный метод имеет сходство с факторным анализом в постановочной части решаемых задач, но имеет ряд существенных отличий:

- при анализе главных компонент не используются итеративные методы для извлечения факторов;
- наряду с активными переменными и наблюдениями, используемыми для извлечения главных компонент, можно задать вспомогательные переменные и/или наблюдения; затем вспомогательные

переменные и наблюдения проектируются на факторное пространство, вычисленное на основе активных переменных и наблюдений;

- перечисленные возможности позволяют использовать метод как мощное средство для классификации одновременно переменных и наблюдений.

Решение основной задачи метода анализ главных компонент и классификация достигается созданием векторного пространства латентных (скрытых) переменных (факторов) с размерностью меньше исходной (исходная размерность определяется числом переменных для анализа в исходных данных).

Предположим, необходимо выбрать объект (например, автомобиль) по двум критериями например, мощность двигателя и стоимость. Если значения этих двух критериев принять за координаты точек на плоскости, соответствующих различным автомобилям, то получим диаграмму рассеяния, которая покажет, что можно построить линию, проходящую через большинство точек и, в частности, через центр облака точек. В этом случае линия регрессии будет представлять два свойства автомобилей и, следовательно, может использоваться для выбора автомобиля. Тем не менее если принять во внимание и другие технические параметры автомобиля, например время разгона до 100 км/ч, то обычная парная регрессия переменных не поможет в принятии решения, так как она уже не будет представлять все три свойства автомобиля. Таким образом, становится ясным, что раз число переменных больше двух, то регрессия двух переменных уже не подходит для нашей задачи. Для случая с несколькими переменными требуется что-то, что является общим для всех переменных и может быть использовано как «значение» вида объектов. Если выразить геометрически, то это должна быть линия или линии (оси факторов), которые проходят через центр облака точек многомерного пространства.

Анализ главных компонент является тем методом, который может сделать это.

Новые факторные оси построены в пространстве меньшей размерности, на них можно спроектировать пространство переменных анализа.

Математически вычисление факторов в основном состоит в диагонализации симметричной матрицы: матрицы корреляций или ковариаций в зависимости от того, нужно ли данные стандартизировать или центрировать относительно средних значений. В обоих случаях результатом будет новый набор некоррелированных переменных (главных компонент), которые являются линейными комбинациями первоначальных переменных. Число переменных становится меньше,

и внутренняя дисперсия данных стремится к максимально возможному значению. Фактически в этом случае создается новое пространство — факторное, на которое можно спроектировать переменные и наблюдения, затем можно классифицировать на категории.

Главные компоненты — это прямые линии, которые наилучшим образом соответствуют облакам точек в векторных пространствах переменных и наблюдений, согласно критерию наименьших квадратов. По критерию наименьших квадратов главные компоненты (факторы) получаются как результат максимальной суммы квадратов ортогональных проекций. Следовательно, строится векторное подпространство меньшей размерности, которое заменяет первоначальное векторное пространство. Хотя фактор извлекается так, чтобы максимально объяснить разброс данных, редко удается сделать это полностью. Поэтому извлекается еще один фактор и т.д. По крайней мере число факторов, извлекаемых таким образом, никогда не превысит число переменных анализа.

В программе *STATISTICA* метод анализ главных компонент реализован для векторных пространств переменных и наблюдений. В соответствии с идеологией метода главных компонент можно разделить переменные на две группы: переменные анализа или активные переменные и вспомогательные переменные. Оба набора переменных относятся к одним и тем же данным и, следовательно, коррелируют между собой. Главные компоненты (факторы) будут вычислены только по переменным анализа (активным переменным).

Вспомогательные переменные можно затем спроектировать на подпространство факторов, чтобы сделать выводы об этих переменных, даже если они не участвовали непосредственно в вычислениях. То есть вспомогательные переменные используются только для интерпретации результатов. Заметим, что такое разделение переменных необязательно и должно исходить из существа задачи.

Аналогично наблюдения можно разделить на вспомогательные и активные наблюдения для анализа. Это может быть сделано с помощью группирующей переменной, с использованием одного из ее значений в качестве кода для задания наблюдений анализа.

Остальные наблюдения будут считаться вспомогательными наблюдениями. При этом только основные наблюдения будут участвовать в вычислениях главных компонент. Вспомогательные наблюдения позже проектируются на векторное подпространство, образованное факторами, которые были вычислены на основе переменных анализа и основных наблюдений. Выводы на основе вычисленных факторов применимы и к вспомогательным наблюдениям, даже если они не участвовали в наблюдениях.

Как было замечено, метод главных компонент позволяет вычислять главные компоненты с помощью матрицы корреляций или матрицы ковариаций. При реализации метода на вычисляемые факторы будут влиять различия вариабельности (изменчивости) активных переменных. Следовательно, анализ будет успешным, только если такие различия представляют интерес для проводимых исследований. В большинстве случаев эти различия несущественны просто потому, что они связаны с измерениями в различных шкалах. Например, два различных типа измерений температуры по Цельсию и Фаренгейту могут использоваться в двух переменных. Очевидно, что учет этих различий в анализе приведет к отрицательным результатам. В этом случае рекомендуется преобразовать данные, чтобы исключить различие в масштабах. Так как эти измерения произведены в шкале интервалов (связь между температурой по Фаренгейту и Цельсию имеет вид линейной зависимости — $F = (5/9)C + 42$) и измерения отличаются точкой начала отсчета

(62) и масштабом (5/9), данные надо преобразовать, а именно: центрировать относительно средних и масштабировать стандартными отклонениями, т.е. надо выбрать матрицу корреляций для вычисления главных компонент. Если измерения отличаются только точкой начала отсчета, данные нужно центрировать только относительно их средних, по этой причине главные компоненты необходимо вычислять через матрицу ковариаций. Очевидно, если в таблице исходных данных присутствуют разнотипные переменные (например, вес, длина, температура) или дисперсии однотипных переменных существенно отличаются, то для вычисления главных компонент надо выбрать корреляционную матрицу.

Собственные значения матрицы ковариаций или корреляций переменных анализа играют важную роль в вычислении главных компонент. Дополнительно к определению факторных координат переменных и наблюдений они предоставляют информацию о дисперсии, которую можно проанализировать по числу факторов. Эта информация может быть в дальнейшем использована для определения порядка, на который вы можете уменьшить размеры пространства первоначальных переменных и наблюдений без потери данных. На основе собственных значений построены различные критерии для определения оптимального числа факторов. Так как сумма собственных значений равна числу «активных» переменных и среднее собственных значений равно 1, то общий критерий состоит в том, чтобы начать с собственных значений, которые больше 1.

Один из важных вопросов, на который дастся ответ в методе главных компонент, является ли число главных компонент оптимальным, т.е. могли бы они (главные компоненты) идеально представить весь набор точек (переменных и наблюдений). Так как каждое собственное значение матрицы корреляций или ковариаций является показателем объясненной дисперсии каждой главной компоненты, то процент общей дисперсии (объясненной) можно приписать к данному числу факторов. Этот процент называют «качеством отображения», и он *является* важной мерой дисперсии, вычисляемой по данному набору главных компонент.

Описание модуля Principal Components & Classification Analysis

Из библиотеки Example --> Datasets откройте файл данных Activities, в котором приведены различные характеристики образа жизни для 28 групп людей. В качестве активных переменных используем 7 видов социальной активности: *WORK* (работа), *TRANSPORT* (транспорт), *CHILDREN* (дети), *HOUSEHOLD* (домашний быт), *SHOPPING* (покупки), *PERSONAL CARE* (личное время), *MEAL* (еда), которым посвящают время представители каждой из 28 групп. Показателем является общее время, посвященное данному виду активности представителями группы в часах. При анализе пропущенные данные замените на соответствующие средние. В качестве вспомогательных переменных укажите 3 переменные: *SLEEP (*сон), TV (телевизор) и *LEISURE* (досуг). Для того чтобы проиллюстрировать способ задания основных и вспомогательных наблюдений, в файл данных добавлена дополнительная группирующая переменная *GENDER* (пол), принимающая значения *MALE* (мужчины), *FEMALE* (женщины). Это означает, что одна часть групп состоит из женщин, другая — из мужчин. Для присвоения меток точкам на графиках добавлена переменная *GEO.REGION* (регион).

Цель данного анализа — изучение взаимосвязей между различными показателями социальной активности, чтобы выявить скрытые факторы, которые упростили бы процесс классификации изучаемых групп населения. Для достижения этой цели необходимо определить факторные оси в пространстве меньшей размерности, на которые можно спроектировать пространство переменных анализа, а также сделать возможной визуализацию этих групп, т.е. нанести результаты на карту полученного пространства.

В верхнем меню Statistics щелкните по Multivariate Exploratory Techniques и выберите команду Principal Components & Classification Analysis. Откроется стартовая панель модуля (рис. 13), в котором на вкладке Advanced нажмите кнопку Variables. В открывшемся окне Select the variables, в поле Variables for analysis (переменные для анализа) высветите переменные *WORK* (работа) — *MEAL* (еда), в поле Supplimentary variables (вспомогательные переменные) высветите переменные *SLEEP* — *LEASURE*, в поле Active cases variable

(переменные с основными наблюдениями) — *GENDER*, в поле Grouping variable — *GEO.REGION (*рис. 14).

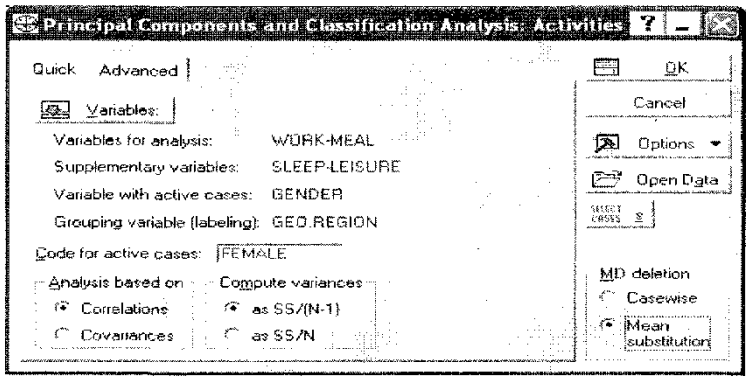

Рисунок 13 - Стартовая панель модуля

Нажмите кнопку OK. В открывшемся окне Principal Components and Classification Analysis в поле Code for active cases выберите значение группирующей переменной *FEMALE* в качестве кода для основных наблюдений.

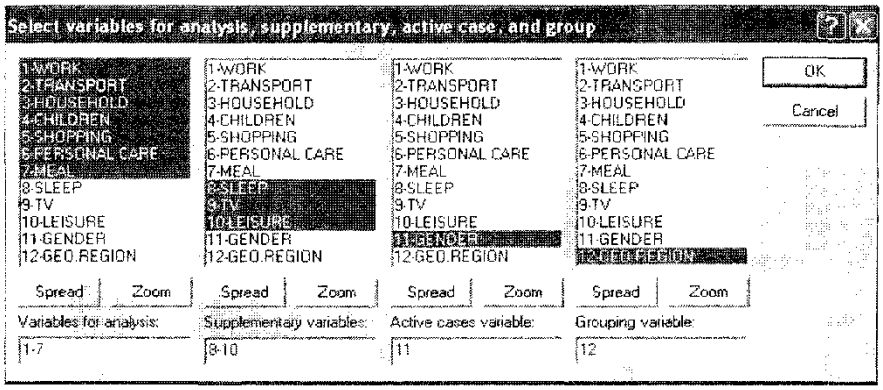

Рисунок 14 - Поле Grouping variable — *GEO.REGION*

При помощи контекстного меню, щелкнув правой кнопкой мыши последовательно на именах переменных и просмотрев в *Statistics of Block Data - * Block columns -* Means, SD's,* легко проверить, что дисперсии и средние переменных в файле исходных данных значительно отличаются. Поэтому в рамке Analysis based on (анализ основан на) выберите опцию *Correlations* для проведения анализа па основе корреляционной матрицы. Далее в рамке Compute variances (вычисление дисперсии) выберите опцию *as SS/N-1,* в рамке MD deletion (удаление пропущенных данных) — опцию *Mean Substitution* (замена средним) и нажмите кнопку OK.

В появившемся окне результатов анализа Principal Components and Classification Analysis Results в информационной части указано количество основных и вспомогательных переменных и наблюдений (рис. 15).

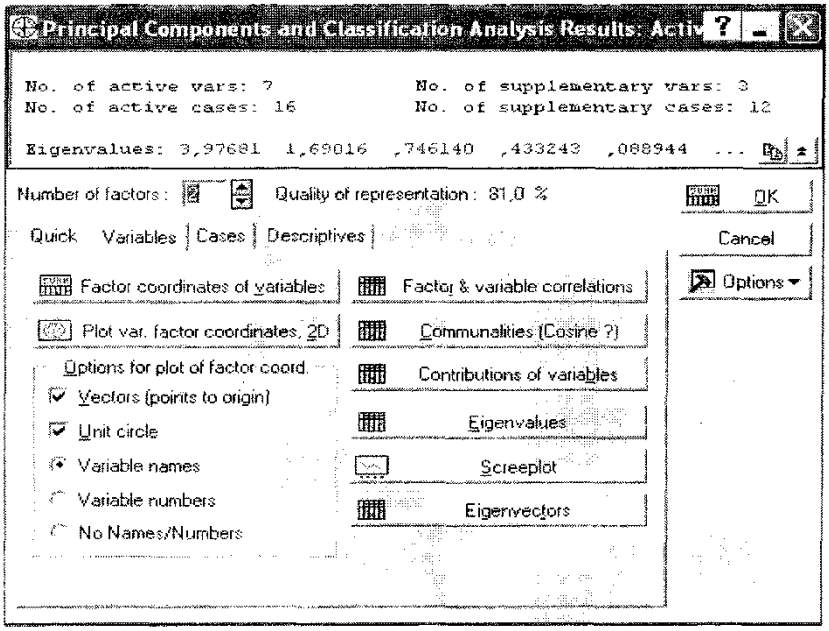

Рисунок 15 - Количество основных и вспомогательных переменных и наблюдений

Нажмите кнопку Screeplot, программа построит график каменистой осыпи, на котором в виде кусочно-линейной функции изображены собственные значения (рис. 16). По критерию Кэттеля надо определить собственное значение, начиная с которого «горка» теряет свою кривизну, т.е. убывание собственных значений максимально замедляется. Число выделяемых факторов должно быть равно номеру этого собственного значения. Из графика видно, что такими собственными значениями являются значения 2 или 3. Поэтому число выделяемых факторов может быть равно 2 или 3. В поле Number of factors установите число факторов, равным 2. При этом качество представления (*Quality of representation)* поменяет свое значение со 100% на 81%.

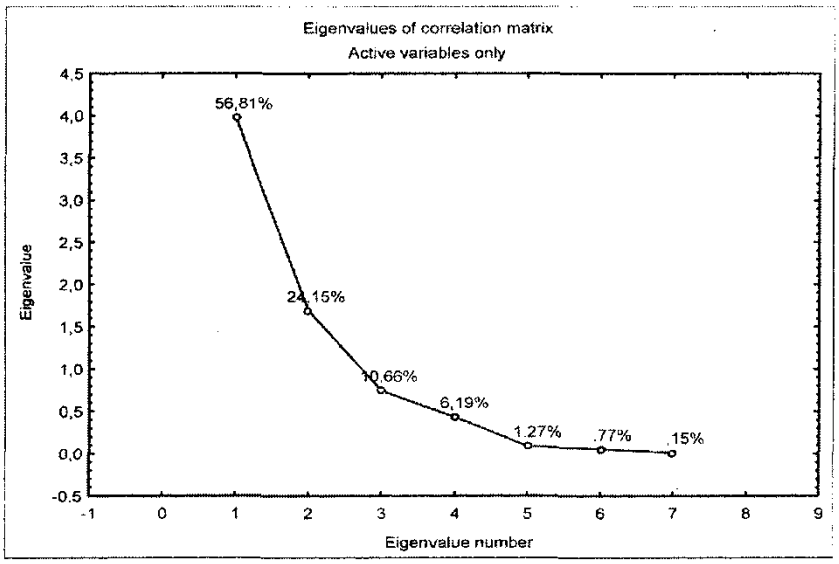

Рисунок 16 - Кусочно-линейные функции

Нажмите кнопку Eigenvalues (собственные значения), чтобы построить таблицу собственных значений (рис. 17). В этой таблице для каждого собственного значения также приведён процент объяснённой дисперсии (*Total variance*), кумулятивное собственное значение (*Cumulative Eigenvalue)* и кумулятивный процент (*Cumulative %)* объясненной дисперсии. Собственные значения представлены в порядке убывания, отражая тем самым степень важности соответствующих выделенных факторов для объяснения вариации исходных данных. Так, фактор, соответствующий максимальному собственному значению (3,976814), описывает приблизительно 56,8% общей вариации. Второй фактор для значения (1,690162) описывает 25,77% общей вариации и т.д.

Value number	Eigenvalues of correlation matrix, and related statist Active variables only			
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	3,976814	56,81163	3,976814	56,8116
2	1,690162	24,14518	5,666976	80,9568
3	0,746140	10,65914	6,413116	91,6159
4	0,433243	6,18918	6,846359	97,8051
5	0,088944	1,27063	6,935303	99,0758
6	0,054063	0,77233	6,989366	99,8481
7	0,010634	0,15191	7,000000	100,0000

Рисунок 17 - Таблица собственных значений

Когда анализируются корреляционные матрицы, сумма собственных значений равна числу (активных) переменных, для которых выделены (рассчитаны) факторы, при этом «среднее ожидаемое» собственное значение равно 1. На практике применяют различные критерии для правильного выбора размерности факторного пространства. Наиболее простой из них — оставить только те факторы, собственные значения которых больше 1. В данном примере только два первых собственных значения больше 1 и они объясняют приблизительно 82% общей вариации. Таким образом, значения собственных чисел подтвердили правильность выбора числа выделяемых факторов — 2.

Нажмите кнопку Factor coordinates of variables (факторные координаты переменных), чтобы получить таблицу координат исходных факторов в пространстве новых выделенных факторов (рис. 18).

Variable	Factor coordinates of th Active and Supplement *Supplementary variabl Factor 1	Factor 2
WORK	-0,941018	0,275054
TRANSPORT	-0,851971	-0,185457
HOUSEHOLD	0,912134	0,036525
CHILDREN	0,779245	-0,354216
SHOPPING	0,326204	-0,917236
PERSONAL CARE	-0,536329	-0,685359
MEAL	0,729504	0,377189
*SLEEP	0,590196	0,318393
*TV	0,280880	-0,568769
*LEISURE	0,476076	-0,318265

Рисунок 18 - Таблица координат исходных факторов в пространстве новых выделенных факторов

В терминологии факторного анализа факторные координаты в методе главных компонент также называют «факторными нагрузками». С точки зрения математики, главный компонент — это линейная комбинация переменных, которые сильно коррелируют с ним. В дальнейшем подразумевается, что факторные координаты переменной — это корреляции между переменной и факторными осями.

Следовательно, интерпретация главных компонент должна быть сделана в терминах корреляции, т.е. нужно выделить те переменные (наблюдения),

которые имеют наибольшие (абсолютные) значения факторных координат для данных факторов. Большее абсолютное значение факторной нагрузки переменной с каким-либо фактором говорит о том, что переменная сильнее связана с этим фактором. Другими словами, чем больше величина факторной координаты переменной, тем лучше переменные показывают структуру, представленную этим фактором. Например, фактор с высокими факторными нагрузками для трех измерений размеров человека, таких, как вес, рост и окружность грудной клетки, может рассматриваться как представляющий «размер» (т.е. абстракция трех переменных) человека.

Координаты отображаются как для активных переменных, так и для вспомогательных. Как видно из таблицы, первая факторная ось, соответствующая собственному значению 3,976, наиболее сильно коррелирует с переменными *WORK, TRANSPORT* (сильные отрицательные корреляции), *PERSONAL CARE* (умеренные отрицательные корреляции), *MEAL, SLEEP* (умеренная положительная корреляция), *HOUSEHOLD* и *CHILDREN* (сильные положительные корреляции). Поэтому можно субъективно обозначить первую выделенную факторную ось как социальную активность, связанную с работой, домом и детьми. Вторую же ось, соответствующую собственному значению 1,69, можно обозначить как социальную активность, связанную с такими видами деятельности, как покупки, личное время, телевидение (сильные и умеренные корреляции с *SHOPPING, TV, PERSONAL CARE).*

Аналогичная кнопка находится на вкладке Cases. Факторные координаты наблюдений (Factor coordinates of cases) — это не корреляции, как в случае с переменными. Наблюдения с большими координатами лучше показывают структуру, представленную фактором.

Процессу интерпретации факторов помогают графики факторных координат переменных и наблюдений. Нажмите на вкладке Variables кнопку Plot var. factor coordinates. 2D, чтобы построить соответствующий график для выделенных факторов (рис. 19). Как видно из рисунка, все переменные изображены в виде точек на единичном круге, так как корреляции (координаты точек) наблюдений с факторными осями принимают значения из интервала [0, 1].

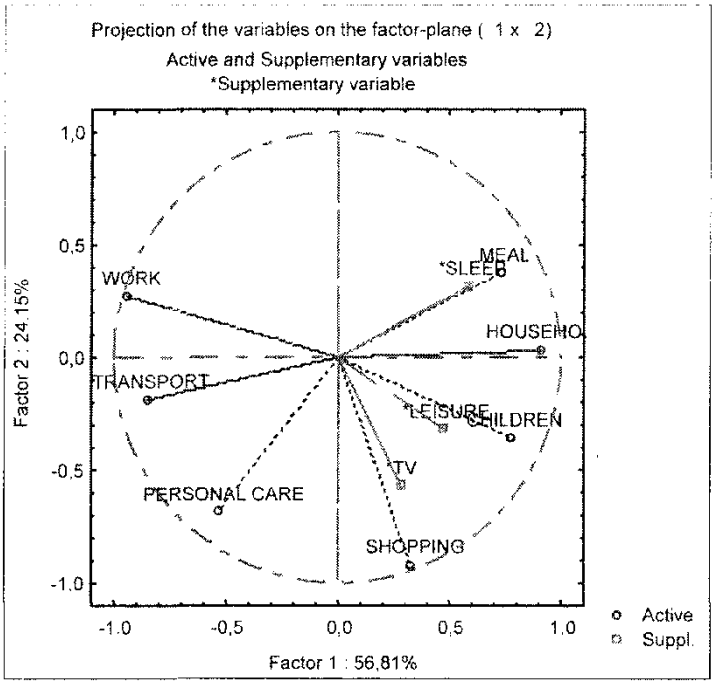

Рисунок 19 - График для выделенных факторов

Горизонтальная ось системы координат соответствует фактору 1 *(Factor 1*), а вертикальная — фактору 2 *(Factor 2).* В зависимости от знаков координат точки расположены в соответствующих квадрантах плоскости. Основные и вспомогательные переменные изображены (на мониторе) соответственно кружочком синего цвета и прямоугольником красного цвета. Этот круг является визуальным индикатором того, насколько хорошо каждая переменная воспроизводится текущим набором выделенных факторов — чем ближе переменная к единичной окружности, тем лучше она воспроизведена в найденной системе координат.

Variable	Variable contributions	
	Factor 1	Factor 2
WORK	0,222669	0,044762
TRANSPORT	0,182522	0,020350
HOUSEHOLD	0,209210	0,000789
CHILDREN	0,152691	0,074235
SHOPPING	0,026757	0,497776
PERSONAL CARE	0,072331	0,277913
MEAL	0,133820	0,084176

Рисунок 20 - Таблица вкладок основных переменных

Нажмите кнопку Contributions of variables (вклад переменных), появится таблица (рис. 20) с вкладами основных переменных.

Вклад переменной — это относительный вклад переменной в дисперсию факторной оси.

Значения этой статистики используются для отсеивания переменных, перед тем как они рассматриваются на основе факторных координат, т.е. корреляций для интерпретации факторных осей. Естественно, те переменные должны быть кандидатами для дальнейшей проверки, вклад (относительный) которых в дисперсию оси фактора больше. Обратите внимание, что значения вкладов «пропорциональны» факторным нагрузкам.

Аналогичная кнопка находится на вкладке Cases. Как и в случае переменных, вклады основных наблюдений *Contributions of cases* также являются их относительными вкладами в дисперсию факторной оси. Следовательно, вклад наблюдения — это мера важности наблюдения в качестве определителя факторной оси. Больший вклад наблюдения «утяжеляет» его в факторе. Следовательно, при переходе к интерпретации главных компонент сначала рассматриваются наблюдения с большим вкладом.

Нажмите кнопку Communalities [Cosine 2]. Программа построит таблицу общностей переменных (рис. 21).

Variable	Communalities, based Active and Supplemen *Supplementary variab	
	From 1 factor	From 2 factors
WORK	0,885515	0,961170
TRANSPORT	0,725854	0,760248
HOUSEHOLD	0,831988	0,833322
CHILDREN	0,607222	0,732691
SHOPPING	0,106409	0,947731
PERSONAL CARE	0,287648	0,757366
MEAL	0,532177	0,674448
*SLEEP	0,348331	0,449705
*TV	0,078893	0,402391
*LEISURE	0,226649	0,327941

Рисунок 21 - Таблица общностей переменных

Общность — это доля объясненной дисперсии, которая характеризует степень общности переменной (наблюдения) с другими переменными (наблюдениями) по заданному числу факторов. Геометрически это квадрат косинуса угла, образованного радиус вектором переменной (наблюдением) и факторной осью.

На вкладке Cases этой кнопке соответствует кнопка с названием Cosine 2. В таблице, которая откроется после нажатия этой кнопки (рис. 22), также представлена дополнительная информация о принадлежности наблюдения к основным или вспомогательным наблюдениям. Каждому наблюдению также будет поставлено в соответствие значение группирующей переменной *GEO REGION.*

На вкладке Cases щелкните по кнопке Plot cases factor coordinates,2D (график наблюдений в факторном пространстве). Появится график (рис. 23), на котором изображаются как основные (*FEMALES)* наблюдения, которые использовались при расчете факторов (кружочки синего цвета), так и вспомогательные *(MALES)* наблюдения (квадратики красного цвета).

Case	Cosine squares, based on correlations (Activities) Active cases variable: GENDER Labelling variable: C Code for active cases: FEMALE Suppl. case values Factor 1	Factor 2	GENDER	GEO.REGION
EMU	0,752438	0,012088	MALE	WEST
EWU	0,683246	0,250961	FEMALE	WEST
UWU	0,212589	0,598020	FEMALE	WEST
MMU	0,709125	0,036624	MALE	WEST
MWU	0,200835	0,638409	FEMALE	WEST
SMU	0,657691	0,100238	MALE	WEST
SWU	0,650255	0,229026	FEMALE	WEST
EMW	0,365504	0,494439	MALE	WEST
EWW	0,249937	0,609442	FEMALE	WEST
UWW	0,821704	0,053460	FEMALE	WEST
MMW	0,335141	0,523670	MALE	WEST
MWW	0,623147	0,291490	FEMALE	WEST
SMW	0,289582	0,445688	MALE	WEST
SWW	0,088667	0,456411	FEMALE	WEST
EME	0,765868	0,117273	MALE	EAST
EWE	0,723131	0,036003	FEMALE	EAST
UWE	0,812282	0,029792	FEMALE	EAST
MME	0,273680	0,390355	MALE	EAST
MWE	0,096166	0,066855	FEMALE	EAST
SME	0,736173	0,035253	MALE	EAST
SWE	0,850015	0,024420	FEMALE	EAST
EMY	0,747432	0,130184	MALE	EAST
EWY	0,507356	0,173083	FEMALE	EAST
UWY	0,772474	0,001093	FEMALE	EAST
MMY	0,744642	0,120001	MALE	EAST
MWY	0,396336	0,132181	FEMALE	EAST
SMY	0,731283	0,003974	MALE	EAST
SWY	0,835235	0,052683	FEMALE	EAST

Рисунок 22 - Таблица дополнительной информация о принадлежности наблюдений к основным или вспомогательным наблюдениям

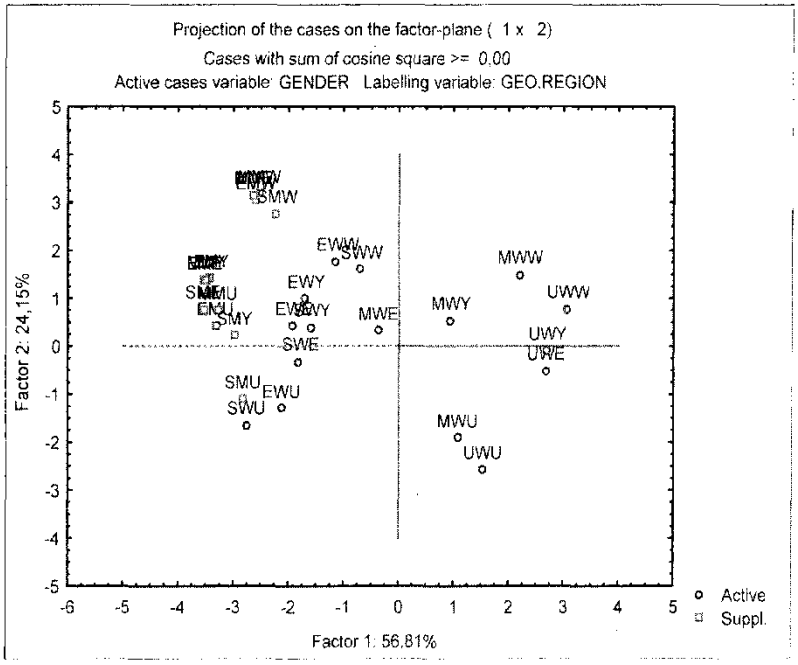

Рисунок 23 - График основных (*FEMALES)* наблюдений

Основные и вспомогательные наблюдения сгруппированы в разных областях плоскости, т.е. они объединены в группы однородности — кластеры. При этом кластер с вспомогательными мужскими группами расположен в центральной и нижней части второго квадранта, т.е. имеет отрицательные значения координат по первой, горизонтальной оси и положительные значения координат по второй, вертикальной оси (кроме одного наблюдения). Напомним, что горизонтальная факторная ось была нами интерпретирована как социальная активность, связанная с работой, домом и детьми. При этом отрицательную часть оси определяют переменные *WORK, TRANSPORT.*

Из графика видно, что мужские группы объединены в области переменной *WORK.* Это означает, что социальная активность мужских групп в основном сконцентрирована в работе. Проявления социальной активности женских групп носят более разносторонний характер — они сгруппированы более или менее равномерно на всей плоскости: в области переменных *PERSONAL CARE, CHILDREN, HOUSEHOLD, SHOPPING, MEAL, SLEEP, LEISURE.*

Список литературы (References):

1. Халафян А.А. STATISTICA 6. Статистический анализ данных. ООО "Бином-Пресс", 2007. 512 с.

2. Метод главных компонент / А. Померанцев http://www.chemometrics.ru/materials/textbooks/pca.htm

3. MatLab. Руководство для начинающих / Е. Михайлов, А. Померанцев http://www.chemometrics.ru/materials/textbooks/matlab.htm

4. Хемометрика в аналитической химии / О.Е. Родионова, А.Л. Померанцев. http://www.chemometrics.ru/materials/articles/

5. Российское Хемометрическое Общество. Доступно на http://rcs.chph.ras.ru/

6. Хемометрика в России. Доступно на http://chemometrics.ru/