

ОСОБЕННОСТИ АНАЛИЗА СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ В СФЕРЕ ИНФОКОММУНИКАЦИЙ

Галина Гайворонская, Петр Яцук, Юлия Казак

Аннотация: Огромные объемы статистической информации, которая стала доступной при построении информационного общества, привели к увеличению случаев некорректного использования методов анализа этой информации, что зачастую влечет за собой ошибочные выводы, которые могут привести к негативным последствиям. Во избежание такой ситуации в работе даны некоторые рекомендации по условиям применения методов анализа данных и сформирована методика поэтапного анализа статистики в сфере электронных коммуникаций.

Ключевые слова: статистическая информация, анализ данных, информационно-коммуникационные технологии, инфокоммуникационные услуги.

ACM Classification Keywords: G.3 Mathematics Of Computing – Probability And Statistics – Statistical Computing, Time Series Analysis, Correlation and Regression Analysis; C.2 Computer-Communication Networks.

Введение

В настоящее время в процессе перехода от индустриального общества к информационному невозможно переоценить роль всеобъемлющей и достоверной информации о происходящих в обществе процессах и явлениях. Естественно такую информацию целесообразно получать путем анализа имеющейся статистической информации, однако в последние годы результаты статистических отчетов все чаще вызывают недоверие, связанное с неадекватным отображением ими действительности. Это связано не только с заведомой подтасовкой и искажением имеющейся статистической информации, но зачастую и с неумением правильно применять методы анализа данных.

Если в предыдущем столетии анализом статистики, связанной с передачей информации по информационным сетям, занимались в основном специалисты, осуществляющие весь комплекс мероприятий, связанных с получением, (чаще всего путем специально организованных измерений) и обработкой статистической информации, то теперь, когда информация и

программные средства ее обработки стали более доступными, ее анализом пытаются заниматься и дилетанты, плохо представляющие условия применения тех или иных методов анализа данных (АД). Такая ситуация приводит к неверным результатам на основании которых делаются ошибочные выводы, иногда приводящие к далеко идущим последствиям не лучшим образом, сказывающимся на развитии нашей страны.

Еще одной особенностью сложившейся ситуации стало то, что если раньше специалисты в основном сталкивались с нехваткой достоверной статистической информации, то сейчас для получения несмещенных оценок исследуемой случайной величины необходимо обрабатывать сверхбольшие объемы статистики. Эти объемы измеряются десятками, а то и сотнями гигабайт, что тоже вызывает определенные трудности, так как ни один программный продукт, предназначенный для обработки статистики не способен корректно обработать такие большие массивы данных.

Следующая сложность связана с наличием большого количества специализированных программных продуктов, предназначенных для анализа данных (ППАД). Их наличие иногда, вместо того, чтобы помочь специалистам, сокращая трудозатраты на выполнение трудоемких рутинных операций, приводит к некорректному использованию, при котором исследователь зачастую даже не понимает, что именно рассчитывается ППАД, по каким формулам и при каких условиях допустимо использование тех или иных методов АД, что априори приводит к заведомо неверным результатам, а соответственно и к ошибочным выводам, сделанным на основе этих результатов.

Постановка задачи

Все это, также и ряд других немаловажных аспектов привели к необходимости разработки методики корректного применения методов анализа данных в инфокоммуникациях. Одному из подходов решения такой задачи посвящена эта работа. Существует большое количество учебной литературы и научных исследований, в которых дается определение тех или иных методов анализа статистической информации. К наиболее известным из них можно отнести работы Андерсона Т. [Андерсон, 1963; Андерсон, 1976], Айвазяна С.А. [Айвазян, 1998; Айвазян, 1989], Боровкова А.А. [Боровков, 2010], Горяинова В.Б. [Горяинов, 2001], Иберлы К. [Иберла, 1980], Кобзаря А.И. [Кобзар, 2006], Орлова А.И. [Орлов, 2004], Мендаля И.Д. [Мандель, 1988], Ким Дж. О. [Ким, 1989], Лоули Д. [Лоули, 1967], Окунь Я. [Окунь, 1974], Хармана Д. [Харман, 1972], Ферстера Э. [Ферстер, 1983], Ренца Б. [Ферстер, 1983], Кендалла М. [Кендалл, 1976], Стьюарта А. [Кендалл, 1976], Лемана Е. [Леман, 1964], Ллойда Э. [Ллойд, 1989; Ллойд, 1990], Ледермана В.

[Ллойд, 1989; Ллойд, 1990], и т.д. Однако большинство из них позволяют хорошо изучить теорию, но не дают практических рекомендаций по получению навыков применению этих методов и использованию для этих целей, имеющихся ППАД.

Наиболее полно задача использования методов АД и существующих ППАД отражена в работах проф. Н.Д. Вайсфельд [Вайсфельд, 2006], которая на протяжении многих лет консультировала сотрудников кафедры информационно-коммуникационных технологий Одесской национальной академии пищевых технологий (ИКТ ОНАХТ) в процессе диссертационных исследований, связанных с применением методов АД, при решении разнообразных задач, связанных с анализом и синтезом информационных сетей Украины. Так в частности результаты этих исследований опубликованные в работах [Гайворонская, 2009а; Гайворонская, 2007а; Гайворонская, 2007b; Павлов, 2007; Котова, 2009а; Котова, 2010а; Котова, 2010b; Котова, 2008; Котова, 2009b; Сахарова, 2008; Бондаренко, 2010а; Бондаренко, 2010b; Бондаренко, 2011], дали возможность обобщить некоторые особенности применения методов АД при исследованиях в области инфокоммуникаций.

Так, например, в работах [Гайворонская, 2009b; Ганницкий, 2009; Гайворонская, 2009c; Ганницкий, 2011; Гайворонская, 2007c; Ганницкий, 2008а; Ганницкий, 2008b; Gannitskiy, 2010; Ганницкий, 2010а; Ганницкий, 2010b; Gannytskyi, 2012], опубликованных по результатам диссертационных исследований И.В. Ганницким на основе анализа методов обработки статистической информации показано, что применение существующих статистических программных продуктов не дает возможности обрабатывать сверхбольшие массивы данных объемом более 1 млн. записей. Поэтому им разработан метод аналитической обработки статистических данных, включающий кеширование, семплинг, иерархическое структурирование исходной информации, что позволило обработать сверхбольшие объемы данных и повысить точность и скорость обработки для исследования массивов статистических данных сверхбольшого объема (более 250 ГБ).

Обработка результатов измерений, выполненная с использованием авторского программного продукта, позволила проанализировать тенденции изменения параметров потоков вызовов и поступающей нагрузки. В результате показано, что существенным является явно выраженное снижение темпов роста нагрузки операторов мобильной связи, количество вызовов мобильной связи и их длительность имеют ярко выраженную тенденцию падения, а пики нагрузки в

праздничные дни становятся все более ярко выраженным при общей тенденции сглаживания нагрузки.

Разработанный автором метод повышения эффективности обработки сверхбольших объемов исходных данных с использованием системы управления базой данных (СУБД) *Oracle*, отличающийся использованием специфических внутренних функций СУБД, переносом методов статистической обработки на уровень работы СУБД позволяет:

- Сократить время, необходимое для обработки статистических данных;
- Повысить скорость доступа к обработанным данным в 6–7 раз;
- Автоматизировать создание отчетов в различных временных срезах и режимах;

Увеличить точность полученных результатов на 12-15% за счет подбора интервала усреднения и обеспечить взаимодействие с существующими программными продуктами, предназначенными для обработки статистики.

Можно утверждать, что в большинстве исследований, связанных с развитием сферы инфокоммуникаций в той или иной степени присутствуют результаты, полученные с применением методов АД.

Эта работа направлена на разработку пошагового алгоритма, определяющего условия корректного применения методов АД и рекомендаций по использованию ППАД, наиболее удобных для решения отдельных задач АД, относящихся к анализу функционирования и планированию развития информационных сетей и информационного общества в целом.

Основная часть

Прежде чем приступить к использованию методов математической статистики необходимо определить способ представления имеющейся статистической информации, которая может представлять собой выборку или временной ряд. Если имеющаяся статистическая информация упорядочена во времени - она представляет собой временной ряд, если же сформирована по каким-то другим признакам, ее называют выборкой. В литературе, посвященной использованию методов математической статистики [Кобзар, 2006] под выборкой понимают результат n независимых последовательных наблюдений случайной величины X из рассматриваемой генеральной совокупности.

Для того чтобы определить какие именно методы корректно использовать для анализа конкретной выборки, необходимо оценить ее вид, то есть определить является ли выборка стандартной, парной, комбинированной, стратифицированной, кластерной или систематической.

- Если используется парная статистическая выборка, то прежде, всего необходимо оценить является ли она репрезентативной, для этого она должна содержать не менее 30 объектов, в случае, если выборка содержит меньшее количество объектов, использование статистических методов для ее исследования считается некорректным.
- Если стандартная статистическая выборка репрезентативна, проводится анализ характера распределения исследуемой случайной величины.
- Если выборка подчиняется хотя бы одному из распределений известного вида, то ее можно исследовать методами параметрической статистики, при этом определяются основные центральные моменты, а именно математическое ожидание, дисперсия и среднеквадратическое отклонение.
- Если вид функции распределения исследуемой случайной величины не может быть описан известными законами распределения, то расчет перечисленных основных центральных моментов не имеет смысла. В этом случае необходимо использовать методы непараметрической статистики, к которым относится вычисление: медианы, то есть числа, которое делит вариационный ряд на две части, содержащие одинаковое число элементов, если объем выборки $n = 2l + 1$ (непарный), то $Me_x = x^{(l+1)}$, если же объем выборки парный, то $Me_x = \frac{(x^{(l+1)} + x^{(l+1)})}{2}$; моды – величины признака, чаще всего встречающегося в данной совокупности $M_o = x_0 + n \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})}$; размаха – разницы между максимальным и минимальным значением признака $R = x_{\max} - x_{\min}$ и квартиля – значения признака в ранжированном ряду распределения, выбранного таким образом, что 25% единиц совокупности будут меньше по размеру Q_1 ; 25% будут заключены между Q_1 и Q_2 ; 25 % – между Q_2 и Q_3 ; другие 25 % превосходят Q_3 :

$$Q_1 = x_{Q_1} + i_{Q_1} \frac{\frac{1}{4} \sum f - S_{Q_1-1}}{f_{Q_1}}, \quad Q_2 = x_{Q_2} + i_{Q_2} \frac{\frac{2}{4} \sum f - S_{Q_2-1}}{f_{Q_2}}, \quad Q_3 = x_{Q_3} + i_{Q_3} \frac{\frac{3}{4} \sum f - S_{Q_3-1}}{f_{Q_3}}.$$

На Рисунке 1 представлен общий подход к использованию методов математической статистики.

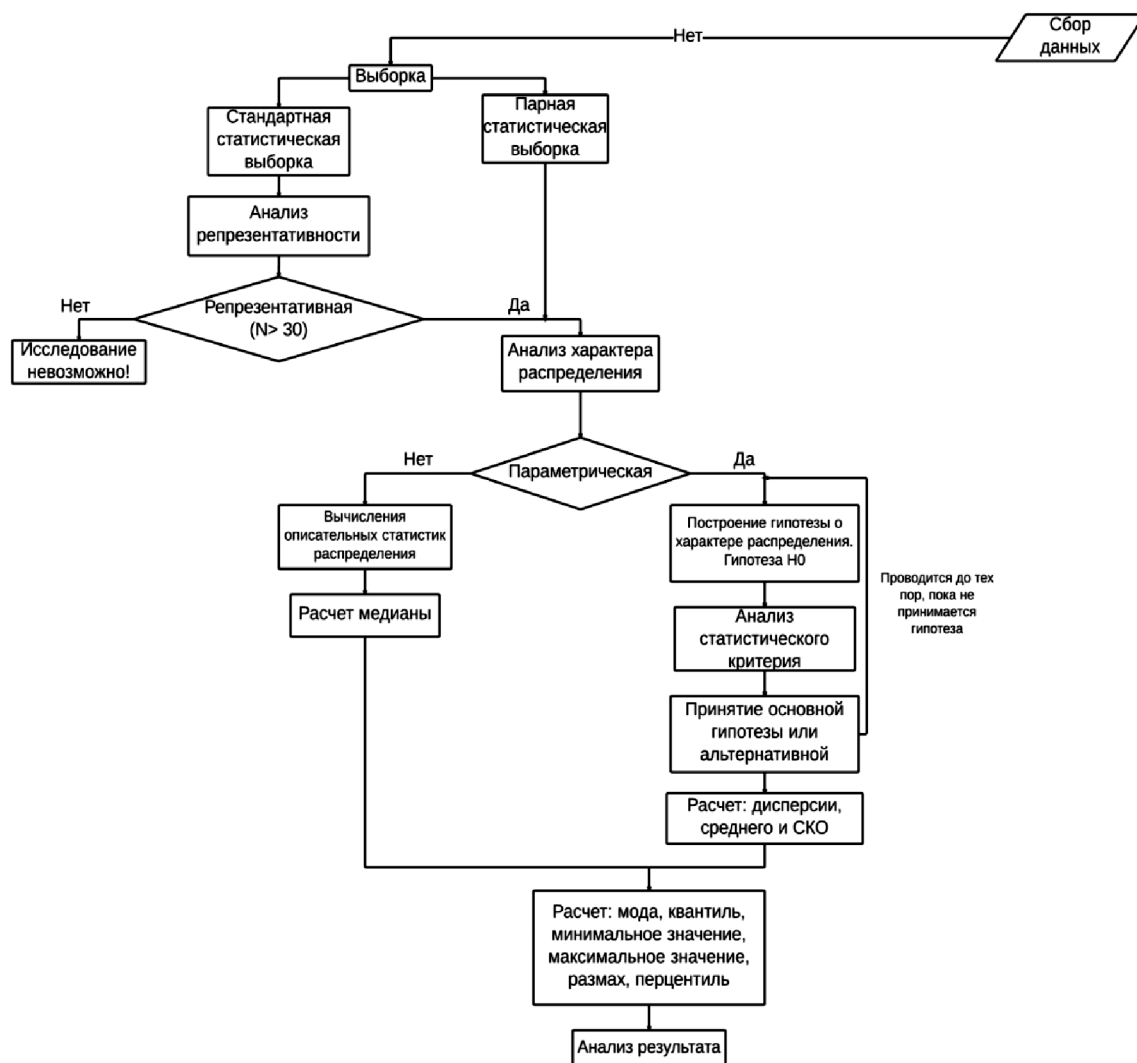


Рисунок 1. Методы анализа статистической информации для выборки

В качестве примера предварительного анализа статистической информации рассмотрим результаты, полученные при анализе состояния телекоммуникационной отрасли в различных странах мира, размещенной на сайтах Международного союза электросвязи (*International Telecommunication Union, ITU*) [ITU] и Европейской экономической комиссии ООН (*United Nations Economic Commission for Europe, UNECE*) [UNECE Statistic Database]. На основе этих данных выполнен анализ информации, характеризующей количество пользователей четырех видов связи: фиксированной проводной широкополосной связи (ФПШС), обозначенной T1, фиксированной телефонной связи (ФТС) (T2), мобильной связи (МС) и Интернет (T4). В

исследовании использованы показатели для 53 стран мира за период с 2000 по 2012 годы. Пример анализируемых исходных данных приведен в Табл. 1 - 4, отражающих использование ФПШС, ФТС, МС и процент пользователей Интернет.

Таблица 1. Количество точек подключения Т1

Страна	2000	2001	2002	...	2009	2010	2011	2012
Австрия	190500	320600	451000	...	1845600	2029000	2078000	2118000
Молдова		239	418	...	186973	269067	355099	417177
Россия	0	0	11000	...	12900000	15700000	17420161	20703653
США	7069874	12792812	19881549	...	78349000	82759000	86445000	90341000
Узбекистан		0	0	...	88735	118000	147760	212729
Украина			0	...	1906725	2954556	3169396	3643460
Швейцария	56416	140000	455220	...	2739149	2911504	3076384	3210631

Таблица 2. Количество точек подключения Т2

Страна	2000	2001	2002	...	2009	2010	2011	2012
Австрия	3997000	3997000	3883000	...	3253000	3398000	3388000	3380000
Молдова	583811	639165	719286	...	1138729	1161148	1179953	1205768
Россия	32070000	33278200	35500000	...	45379601	44915829	44151461	42168388
США	192513000	191570800	189250143	...	152873000	149652000	143319000	138595000
Узбекистан	1655044	1662963	1681127	...	1856592	1892164	1927735	1980038
Украина	10417000	10669600	10833300	...	13026293	12941346	12680881	12182142
Швейцария	5235733	5383483	5387568	...	5131810	4907773	4898770	4721981

Таблица 3. Количество точек подключения ТЗ

Страна	2000	2001	2002	...	2009	2010	2011	2012
Австрия	8562000	11132000	12670000	...	22200000	22500000	24490000	24338000
Молдова	139000	225000	338225	...	2784832	3165052	3587431	4080143
Россия	3263200	7750499	17608756	...	230050000	237689224	256116581	261886329
США	109478031	128500000	141800000	...	274283000	285118000	298293000	310000000
Узбекистан	53128	128012	186900	...	16417914	20952000	25441789	20274090
Украина	818524	2224600	3692700	...	54942815	53928830	55576481	59343693
Швейцария	4638519	5275791	5736303	...	9322580	9644157	10082636	10460000

Таблица 4. Количество точек подключения Т4

Страна	2000	2001	2002	...	2009	2010	2011	2012
Австрия	33,73	39,19	36,56	...	73,45	75,17	78,74	80,00
Молдова	1,28	1,49	3,79	...	27,50	32,30	38,00	43,37
Россия	1,98	2,94	4,13	...	29,00	43,00	49,00	63,80
США	43,08	49,08	58,79	...	71,00	71,69	69,73	79,30
Узбекистан	0,48	0,60	1,08	...	17,06	20,00	30,20	36,52
Украина	0,72	1,24	1,87	...	17,90	23,30	28,71	35,27
Швейцария	47,10	55,10	61,40	...	81,30	83,90	85,19	85,20

Из Таблиц 1 - 4 видно, что имеющаяся статистическая информация представляет собой стандартную статистическую выборку, состоящую из 53-х объектов, что подтверждает ее репрезентативность. При анализе характера распределения выборка сводится к однородному

показателю. Случайная величина, подчиняющаяся нормальному распределению на графике, имеет вид кривой „колоколообразной” формы, в пределах которой находятся все ее значения.

На Рис. 2 представлены гистограммы распределения случайной величины, представленной в анализируемой выборке с помощью критерия χ -квадрат.

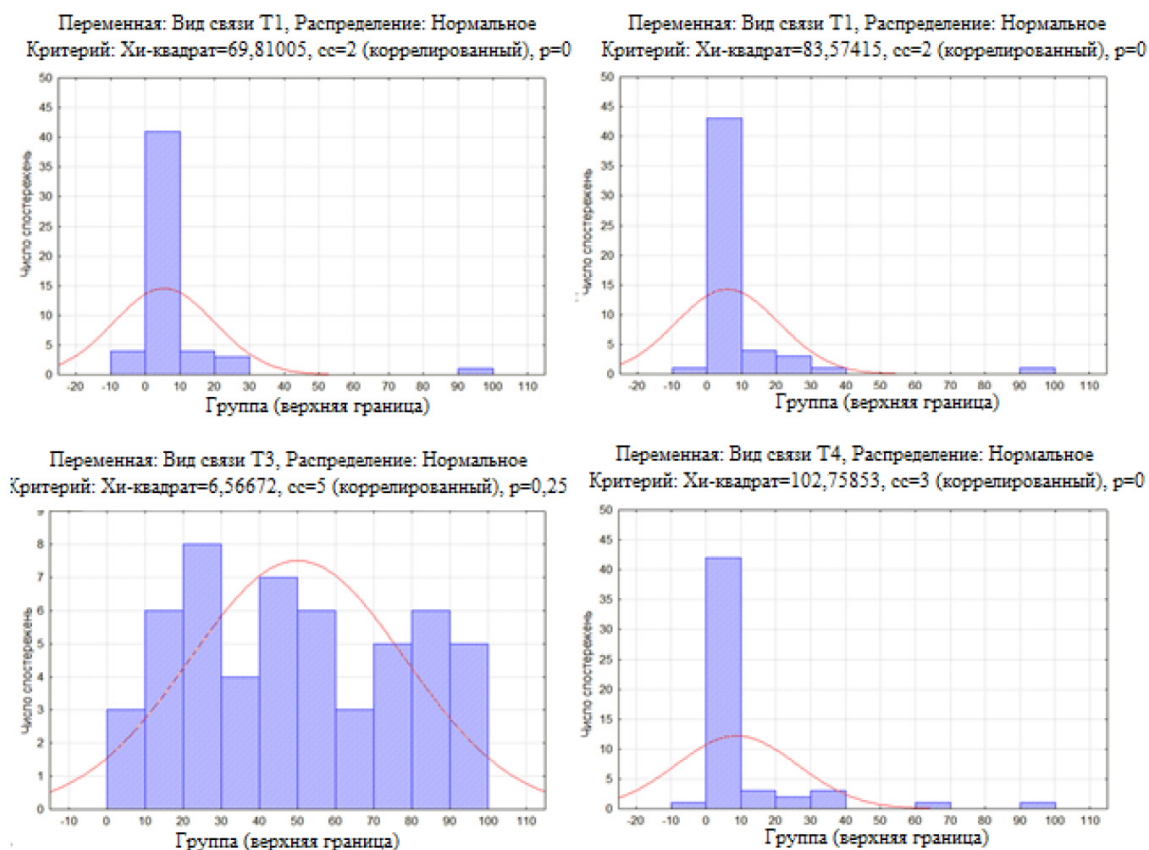


Рисунок 2. Гистограммы анализа нормального распределения для видов связи T1 - T4

Для достоверности результатов проверки нормального распределения выполнено исследование тех же выборок с использованием еще трех критериев:

- Колмогорова-Смирнова основанного на максимуме разности между кумулятивным распределением выборки и предполагаемым кумулятивным распределением;
- Лиллиефорса представляющего модификацию критерия Колмогорова-Смирнова;
- W -критерия Шапиро-Уилки являющегося наиболее эффективным, так как он обладает большей мощностью по сравнению с альтернативными критериями проверки нормальности.

Результаты продемонстрированы на Рис. 3.

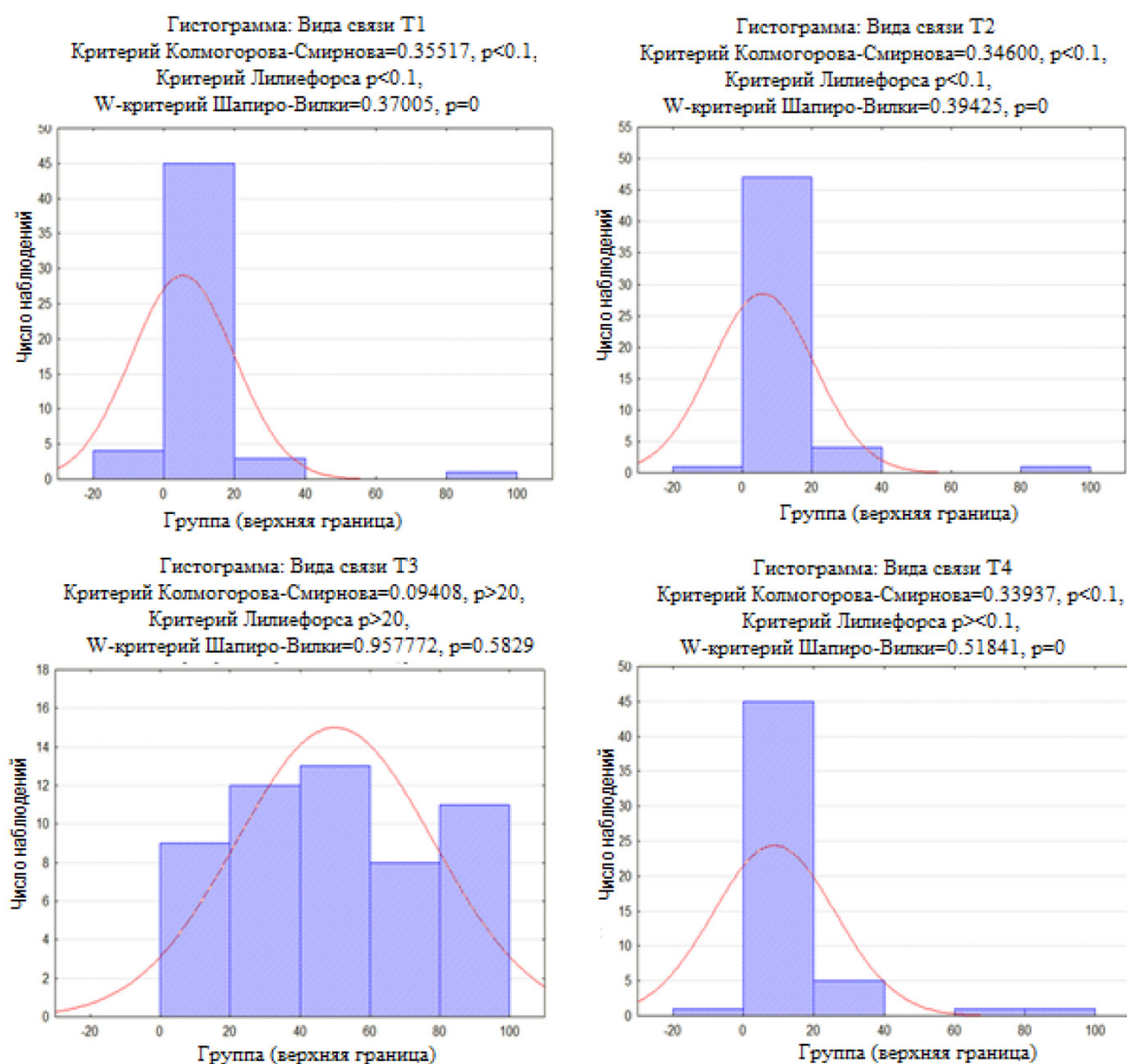


Рисунок 3. Гистограммы анализа нормального распределения для видов связи T1 - T4

Из Рис. 2 и Рис. 3 видно, что значения не одной из анализируемых выборок не находятся в пределах колоколообразной кривой, следовательно, анализируемые выборки не могут быть описаны нормальным распределением. Поэтому для исследования основных характеристик распределения анализируемой случайной величины адекватным является применение методов непараметрической статистики. Полученные результаты представлены в Табл. 5.

Таблица 5. Результат расчета основных статистических характеристик для T1 - T4

Описательные характеристики	T1	T2	T3	T4
Среднее значение выборки	2965924	9965559	19501476	42,14
Максимальное значение выборки	90341000	192513000	310000000	96,21
Минимальное значение выборки	0	18470	1160	0,05
Размах	90341000	192494530	309998840	96,16
Мода	–	–	–	–
Медиана	540850	2490022	5828157	40,41
Квартиль Q ₁	108804	587642	1817835	23,55
Квартиль Q ₂	557167	2303484	5952441	41,39
Квартиль Q ₃	1830508	5396746	15414791	60,71
Среднее линейное отклонение	2886107	9799232	19350479	41,55

Из Таблицы 5 видно, что среднее количество точек подключения (ТП) для ФПШС составляет 2 965 924, при нулевом минимуме, с максимумом в 90341000 ТП. Среднее значение ФТС – 9 965 559 ТП, МС – 19 501 476 ТП.

Следовательно, в настоящее время наиболее широко распространенной в анализируемых странах является мобильная связь, среднее количество ТП для которой практически в два раза больше, чем у ФТС и в 10 раз превышает среднее значение пользователей ФПШС. Это можно пояснить тем, что широкополосная связь пока еще не достигла максимального развития, платежеспособный спрос на эти услуги еще не успевает за предложением.

Из Таблиц 1 - 4 видно, что наиболее часто встречающегося значения в анализируемой выборке не существует, поэтому моду определить невозможно, однако серединой выборки для T1 является значение 540 850, для T2 – 2 490 022, для T3 – 5 828 157, отсюда видно, что использование фиксированной телефонной связи по-прежнему актуально и пока превосходит в 4 раза использование фиксированной проводной широкополосной связи.

Помимо этого в таблице представлены три значения, которые делят исследуемую выборку на четыре равные части (квартили Q_1 , Q_2 , Q_3) что, например, для ФТС Q_3 составляет 5 396 746 , а это означает, что значения количества ТП достаточно сконцентрированы относительно середины выборки.

Для того чтобы учесть различия значений исследуемой выборки рассчитано среднее линейное отклонение (СЛО), которое для ФПШС составило 2 886 107, для ФТС – 9 799 232, для МС – 19 350 479. Эта характеристика показывает величину разброса значений исследуемых случайных величин относительно среднего значения выборки, например, для Т1 СЛО составило 2 886 107 относительно середины выборки, это объясняется тем, что значения ТП для этого вида связи существенно отличаются друг от друга, поскольку этот вид связи еще не достиг максимального развития и его использование существенно отличается в различных странах, в соответствии с их экономическим развитием в целом.

На следующем этапе выполняется анализ признаков имеющейся выборки, для этого разработано множество методов АД, к которым относятся, в частности, факторный, корреляционный, кластерный и регрессионный анализы, которые можно использовать как независимо друг от друга, так и в совокупности.

Факторный анализ [Иберла, 1980] позволяет оценить влияние отдельных факторов на результирующие характеристики процесса и включает пять этапов. На первом этапе отбираются факторы, определяющие исследуемые показатели.

Следующим этапом является классификация и систематизация факторов с целью обеспечения комплексного системного подхода к исследованию их влияния на исследуемую случайную величину.

Далее определяется форма зависимости и моделируется взаимосвязь между факторами и результирующими показателями. На четвертом этапе рассчитывается влияние факторов и оценка роли каждого из них в изменении величины результирующего показателя. Заключительным этапом является практическая работа с факторной моделью и собственно факторный анализ.

Последовательность выполнения этих этапов отображает Рис. 4.

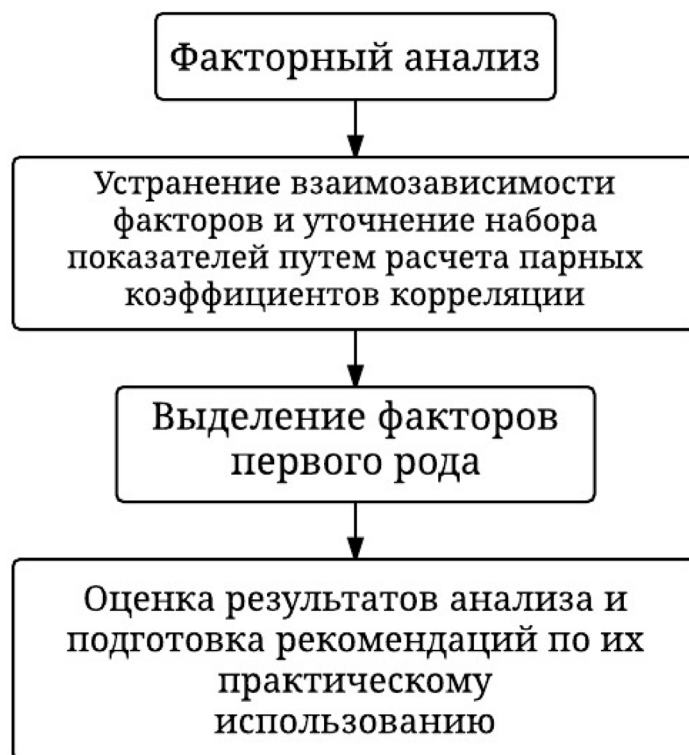


Рисунок 4. Последовательность выполнения факторного анализа

В качестве примера приведены результаты использования методов факторного анализа при выявлении степени влияния отдельных факторов на процесс развития инфокоммуникационных технологий.

В качестве исходных данных использована статистическая информация по макроэкономической и гендерной статистике для 53-х стран мира *UNECE* [*UNECE Statistic Database*]. Макроэкономическая информация – это совокупность экономических показателей для стран всего мира, а гендерная статистика представляет собой социальные данные, классифицированные по половому признаку. Информацию, полученную из баз данных *UNECE*, можно охарактеризовать параметрами, выбор которых обусловлен тем, что на сайте Европейская экономическая комиссия Организации Объединенных Наций (ЕЭК ООН) представлены их реальные значения, а также, тем что доступ к другим источникам аналогичной информации затруднителен.

Проанализированы следующие факторы:

1. Население страны: территория; общая численность населения; плотность населения, чел. на кв. км.
2. Численность населения страны в отдельных возрастных группах по полу: мужчины возрастных групп 0-14, 15-64, 64+ и женщины тех же возрастных групп, процентная составляющая мужчин и женщин в общем населении страны.
3. Доля населения страны, использующего персональный компьютер (ПК), разделенная по половой принадлежности и распределенная в отдельные группы по возрасту и полу: мужчины в возрасте 16-24, 25-54, 55-74 и женщины тех же возрастных групп.
4. Занятость населения, с точки зрения международной стандартной отраслевой классификации (МСОК) и безработица населения: уровень безработицы молодежи; занятость (МСОК версия 3.1 раздел А и В) *UNECE* [*UNECE Publications*]; занятость в промышленности и электроэнергии (МСОК версия 3.1 С-Е); занятость (МСОК версия 3.1 разделы G-I и J-K); занятость в других услугах (МСОК версия 3.1 L-P).
5. Обобщенные экономические показатели в международно-сопоставимых ценах, включающие:
 - Валовой внутренний продукт (ВВП) в текущих ценах и паритет покупательской способности (ППС) в млн. долл. США;
 - ВВП в текущих ценах, млн. национальной валюты;
 - ВВП на душу населения в ценах и ППС (США) текущего года, долларов США;
 - ВВП на душу населения в ценах текущего года, единиц национальной валюты.
6. Обобщенные экономические показатели, включающие: уровень безработицы; занятость и темпы ее роста; индекс потребительских цен и темпы его роста; паритет покупательной способности (ППС); расходы на конечное потребление инфокоммуникационных услуг (ИКУ) на душу населения.
7. Экономическая деятельность:
 - Доля экспорта и импорта товаров и услуг в процентах от ВВП;
 - Внешний баланс товаров и услуг, % от ВВП;
 - Внешний долг, млн. долл. США; ВВП: сельское хозяйство, промышленность (включая строительство), услуги: производственный метод, индекс 2005 = 100, ценах и ППС 2005;
 - ВВП: МСОК А-В, С-Е, F, G-I, J-K, L-P, производственный метод, доля в процентах от валовой добавленной стоимости (ВДС), ценах и ППС текущий год.

Факторный анализ, независимо от используемых методов, начинается с определения количества факторов, которые имеет смысл использовать для анализа. Этот этап является очень важным, так как исследователь зачастую имеет большой набор характеристик, но заранее неизвестно, сколько факторов необходимо и достаточно для предоставления данного набора характеристик.

Сама же процедура факторного анализа предполагает предварительное задание числа факторов. Исходя из этого, исследователь должен заранее определить или оценить возможное количество факторов. Для определения числа факторов удобно использовать программный пакет для статистического анализа информации „STATISTICA-8”, с помощью которого в нашем случае рассчитано два критерия: метод главных компонент и критерий каменной осыпи.

Одним из наиболее распространенных критериев поиска факторов является метод главных компонент. Оценка значимости главных компонент производится по их собственным значениям, представляющим собой дисперсию стандартизованных исходных данных. Основная идея метода основана на следующем предположении: чем выше % общей дисперсии, тем больше информации содержит значение. Поэтому метод главных компонент сводится к последовательной процедуре: вначале ищется первый фактор, объясняющий наибольшую часть дисперсии, затем независимый от него второй фактор, объясняющий наибольшую часть оставшейся дисперсии, и т.д.

Для работы с программным пакетом „STATISTICA-8” необходимо ввести все значения исследуемой случайной величины в главный модуль, после чего запускается модуль „Анализ главных компонент”, после чего выбираются переменные для анализа (в нашем случае это параметры представленные выше). После того как переменные заданы, необходимо принять решение как будет проводится анализ: на основе ковариаций, либо корреляций. Ковариация несет тот же смысл, что и коэффициент корреляции – она показывает, есть ли линейная взаимосвязь между двумя случайными величинами, и может рассматриваться как „двумерная дисперсия”. Знак ковариации указывает на вид линейной связи между рассматриваемыми величинами: если она больше 0 – это означает прямую связь (при росте одной величины растет и другая), ковариация <0 указывает на обратную связь. При ковариации равной 0 линейная связь между переменными отсутствует. В нашем случае анализ проведен на основе корреляционной матрицы, поэтому, выбрана опция „Анализ основан на корреляциях”, а для того, чтобы заменить пропущенные значения анализируемых величин на их средние значения, в группе опций

„Удаление пропущенных данных” установлено значение „Замена средним”. После этого нажав кнопку „ОК” получаем результаты, представленные на Рисунке 5.

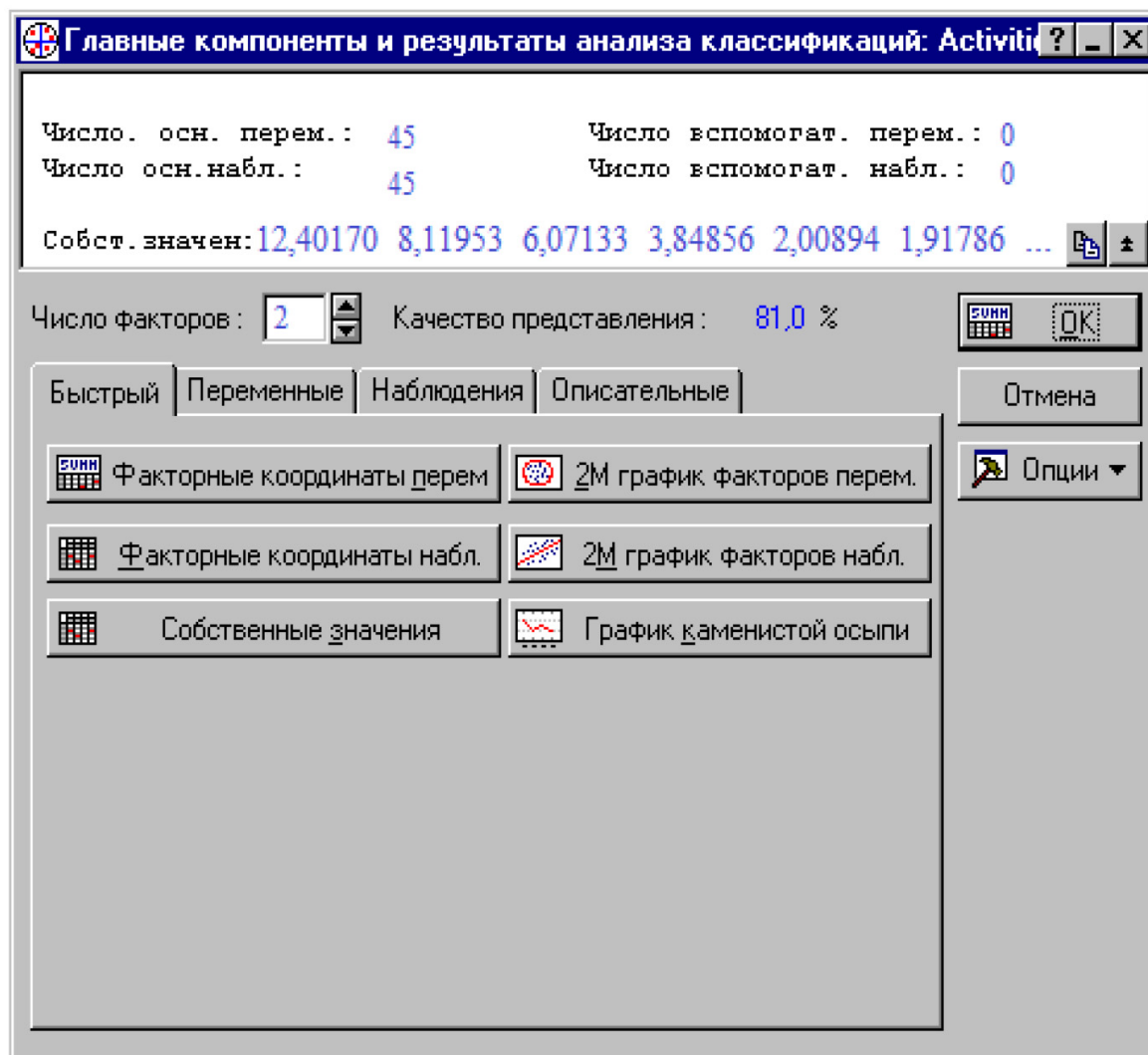


Рисунок 5. Форма модуля “Анализ главных компонент”

В информационном поле диалога „Главные компоненты и результаты анализа классификаций” представлена общая информация о результатах текущего анализа. Это число активных и вспомогательных переменных, наблюдений и собственные значения. При нажатии кнопки „Собственные значения”, получаем таблицу собственных значений (Табл. 6), в которой для каждого собственного значения представлен процент объясненной дисперсии, кумулятивное собственное значение и кумулятивный процент объясненной дисперсии.

Первое собственное значение всегда является наибольшим и превышает единицу; это определяется алгоритмом работы процедуры – факторы вычитаются в порядке их влияния на дисперсию переменных. затем вычисляется дисперсия, обуславливаемая определенным фактором. С изъятием каждого нового фактора собственные значения уменьшаются, а кумулятивный процент приближается к 100.

Таблица 6. Собственные значения матрицы корреляции

Фактор	Собственное значение параметров	% общей дисперсии	Кумулятивное собственное значение	Кумулятивный процент
1	12,40170	27,55932	12,40170	27,55932
2	8,11953	18,04340	20,52123	45,60273
3	6,07133	13,49185	26,59256	59,09457
4	3,84856	8,55236	30,44112	67,64694
5	2,00894	4,46431	32,45006	72,11125
6	1,91786	4,26191	34,36792	76,37316
7	1,61008	3,57796	35,97800	79,95112
8	1,46741	3,26090	37,44541	83,21202
9	1,17164	2,60364	38,61705	85,81566
10	1,10646	2,45880	39,72351	88,27446

Как видно из Таблицы 6, собственное значение для первого фактора равно 12,4, то есть этот фактор описывает примерно 27,6% общей выборки. Второй фактор, имеющий значение 8,12 отвечает за 18,04% общей выборки и т.д.

В рассматриваемом примере оставляем для дальнейшего исследования только первые пять факторов, которые имеют значение больше двух, в соответствии с критерием Кайзера [Kaiser, 1960].

Другим способом определения наиболее значимых факторов является построение и анализ графика каменной осыпи (Рис. 6), на котором отображается последовательность собственных значений факторов. Р. Б. Кеттел [Cattell, 1966] предложил определить на графике собственное значение, начиная с которого „горка” теряет свою кривизну и выходит на примерно постоянный уровень. Правая часть графика является лишь незначительными остатками „каменистой осыпи”. Таким образом, при использовании такого подхода нужно оставить только факторы, расположенные слева от „каменной осыпи”. На Рис. 6 показан график, на котором точка, где непрерывное падение собственных значений замедляется, может соответствовать пятому или sixthому фактору.

В результате сопоставления таблицы собственных значений и графика „каменной осыпи” определено, что в анализируемом случае целесообразно выделить пять факторов.

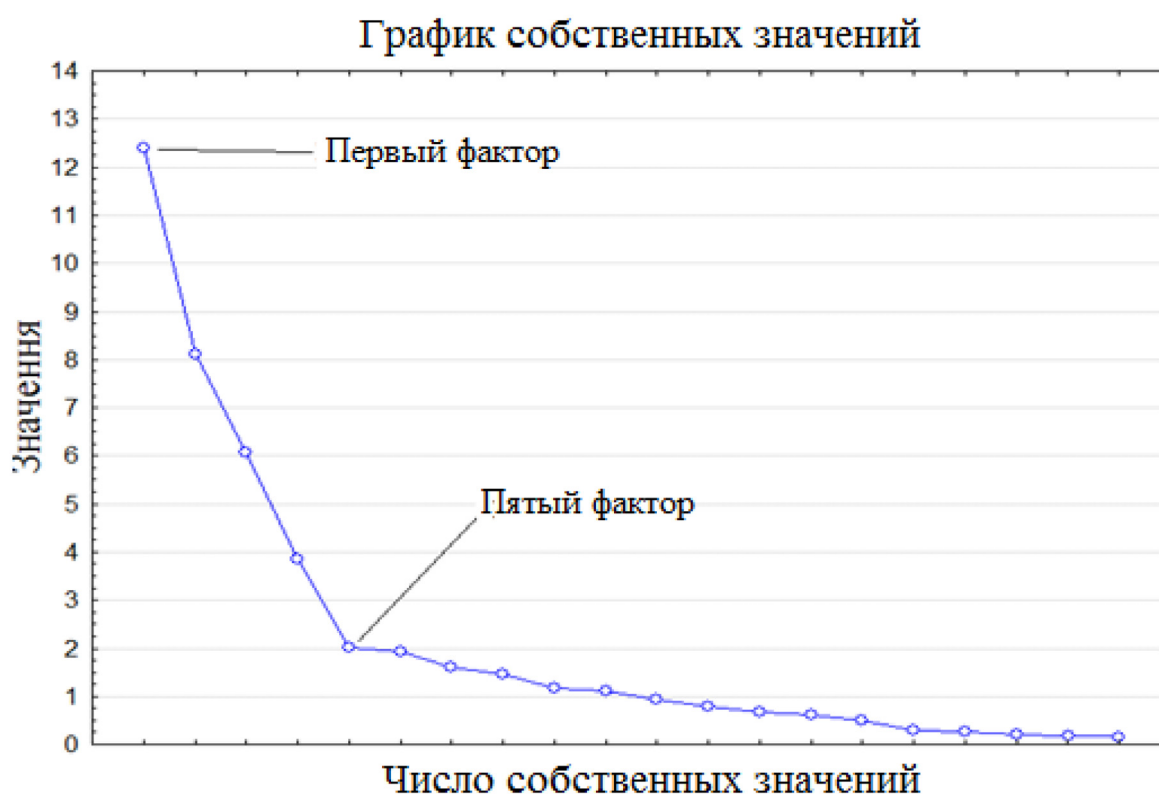


Рисунок 6. График собственных значений факторов

Основные результаты факторного анализа выражаются в наборах факторных нагрузок, то есть значений коэффициентов корреляции каждого из исходных признаков каждым из выявленных факторов и факторных рычагов. Факторными рычагами называют количественные значения выделенных факторов для каждого из n имеющихся объектов. Чем теснее связь конкретного признака с каким-то фактором, тем выше значение факторной нагрузки. Положительный знак факторной нагрузки указывает на прямую (а отрицательный знак – на обратную) связь данного признака с фактором.

В примере использовано несколько видов факторных нагрузок, касающиеся способов вращения, целью которых является получение интерпретируемой матрицы нагрузок, то есть факторов, отмеченных высокими нагрузками для некоторых переменных и низкими для других. При этом использованы методы без обращения, варимакс и квартимакс [Лоули, 1967]. Варимакс [Ким, 1989] – наиболее часто используемый на практике метод, цель которого минимизировать количество переменных, имеющих высокие нагрузки на данный фактор, что способствует упрощению описания фактора за счет группировки вокруг него только тех переменных, которые с ним связаны в большей степени, чем остальные.

Квартимакс [Ким, 1989] – в определенном смысле противоположен варимаксу, т.к. минимизирует количество факторов, необходимых для объяснения данной переменной. Поэтому он усиливает интероперабельность переменных. Квартимакс – вращение приводит к выделению одного из общих факторов с достаточно большими нагрузками на большинство переменных. Пример результата расчетов для одного из этих методов приведен в Таблице 7, где приняты следующие обозначения: М – мужчины, Ж – женщины, ВГ – возрастная группа.

При анализе Таблицы 7, можно заметить, что начальные параметры объединились, образуя факторы. Например, общая численность населения и возрастные группы образуют фактор два, а общая численность мужчин и женщин объединились в фактор три, и так далее. Исходя из результатов Таблицы 7, в качестве основных выделено следующие пять факторов: использование ПК в производственных целях, численность населения, производственная необходимость, экономический фактор и уровень безработицы.

Корреляционный анализ [Ферстер, 1983] двух случайных величин включает в себя: построение корреляционного поля и составление корреляционной таблицы, вычисление выборочных коэффициентов корреляции и корреляционных отношений, а также проверку статистической гипотезы значимости связи (Рис. 7).

Таблица 7. Факторы нагрузок без вращения

Переменные	Фактор 1	Фактор 2	Фактор 3	Фактор 4	Фактор 5
Общая численность населения	-0,60644	0,784758	-0,048107	0,028718	0,027946
ВГ 0-14, М	-0,56028	0,805252	-0,035146	0,023690	-0,008963
ВГ 0-14, Ж	-0,55984	0,805216	-0,034697	0,023430	-0,008984
ВГ 15-64, М	-0,60767	0,782602	-0,049392	0,028653	0,032558
ВГ 15-64, Ж	-0,60305	0,783924	-0,051367	0,024812	0,032395
ВГ 64+, М	-0,66333	0,718933	-0,046818	0,062626	0,061124
ВГ 64+, Ж	-0,64517	0,725005	-0,059302	0,033788	0,063937
Общее население, М (%)	-0,10863	-0,047701	-0,720109	0,002136	-0,107401
Общее население, Ж (%)	-0,08897	-0,006833	-0,762461	-0,033656	-0,016059
Использование ПК, 16-24, М	-0,71622	-0,432270	-0,212928	-0,051869	0,130498
Использование ПК, 16-24, Ж	-0,73283	-0,440734	-0,197314	-0,051307	0,141277
Использование ПК, 25-54, М	-0,79541	-0,431768	-0,174681	-0,113321	0,047776
Использование ПК, 25-54, Ж	-0,79837	-0,426325	-0,161282	-0,116773	0,061987

Переменные	Фактор 1	Фактор 2	Фактор 3	Фактор 4	Фактор 5
Использование ПК, 55-74, М	-0,80783	-0,358019	-0,093864	-0,195795	-0,094792
Использование ПК, 55-74, Ж	-0,78132	-0,322063	-0,082339	-0,190662	-0,077259
ВВП в текущих ценах, (млн. национальной валюты)	0,17189	0,309228	-0,089602	-0,834416	0,294933
ВВП на душу населения в ценах текущего года, (единиц нац. валюты)	0,18696	0,076216	-0,104545	-0,820006	0,308456
Расходы на конечное потребление ИКУ на душу населения, (дол. США)	-0,85094	-0,269061	-0,070655	-0,056333	-0,031010
ППС, единиц национальной валюты за долл. США	0,32237	0,217275	-0,085943	-0,804335	0,249007
Внешний долг, (млн. долл. США)	-0,71501	0,490130	-0,008225	0,064208	-0,003780
ВВП: сельское хозяйство	-0,11039	-0,080288	-0,802218	-0,117216	-0,164368
ВВП: промышленность	-0,02994	0,029975	-0,817944	-0,138809	-0,257610
ВВП услуги	-0,01826	0,039206	-0,840393	-0,105549	-0,267240
ВВП: МСОК А-В	0,70899	0,321233	-0,323788	0,058450	-0,191523
ВВП: МСОК J-K	-0,76247	-0,299351	-0,202627	0,030646	0,129242
Занятость (МСОК вер. 3.1 J-K)	-0,77092	-0,345382	0,331043	-0,011503	-0,049646



Рисунок 7. Последовательность корреляционного анализа

Основное назначение корреляционного анализа – выявление связи между двумя или более изучаемыми переменными, которая рассматривается как совместное согласование изменения двух исследуемых характеристик. Такая изменчивость описывается тремя основными характеристиками: формой, направлением и силой. По форме корреляционная связь может быть линейной или нелинейной.

Более удобной для выявления и интерпретации корреляционной связи является линейная форма. Для линейной корреляционной связи выделяют два основных направления: положительное („прямая связь”) и отрицательное („обратная связь”). Сила связи указывает, насколько ярко проявляется общая изменчивость исследуемых переменных. В качестве числовой характеристики вероятностной связи используют коэффициенты корреляции, значения которых изменяются в диапазоне от -1 до +1 [Ферстер, 1983]

$$r_{x,y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \quad (1)$$

где x_i – значение, принимаемое в выборке X ;

y_i – значение, принимаемое в выборке Y ;

$\bar{x}$ и $\bar{y}$ – средние значения выборок X и Y .

После проведения расчетов, как правило, отбираются только самые сильные корреляции, которые затем интерпретируются. Критерием для отбора „достаточно сильных” корреляций может быть, как абсолютное значение самого коэффициента корреляции (от 0,6 до 1), так и относительная величина этого коэффициента, определяемого по уровню статистической значимости (от 0,01 до 0,1), зависящая от размера выборки [Ферстер, 1983].

Таким образом, задача корреляционного анализа сводится к установлению направления и формы связи между варьируемыми признаками, измерению ее тесноты и к проверке уровня значимости полученных коэффициентов корреляции. Используется множество коэффициентов корреляции, чаще всего применяют коэффициенты r -Пирсона, r -Спирмена и τ -Кендалла. Выбор метода вычисления коэффициента корреляции зависит от типа шкалы, к которой относятся переменные (Табл. 8). Для переменных с интервальной и номинальной шкалой используется коэффициент корреляции Пирсона, а если, по меньшей мере, одна из двух переменных имеет порядковую шкалу или не может быть описана нормальным распределением, используется ранговая корреляция по Спирмену или Кендаллу.

Для условий примера, рассмотренного выше, проанализируем применение методов корреляционного анализа с целью выявления зависимостей между найденными ранее факторами и рассматриваемыми видами связи. В ходе подготовки данных к проведению расчетов на первом этапе выявлено отсутствие нормального распределения, в связи с этим используются методы непараметрической статистики, результаты расчета основных характеристик которых представлены в Табл. 9 в виде матрицы.

Таблица 8. Меры расстояний

Типы шкал		Мера связи
Переменная	Переменная	
Интервальная или отношений	Интервальная или отношений	Коэффициент Пирсона
Ранговая, интервальная или отношений	Ранговая, интервальная или отношений	Коэффициент Спирмена
Ранговая	Ранговая	Коэффициент Кендалла
Дихотомическая	Дихотомическая	Коэффициент ϕ , четырёхполевая корреляция
Дихотомическая	Ранговая	Рангово-бисериальный коэффициент
Дихотомическая	Интервальная или отношений	Бисериальный коэффициент
Интервальная	Ранговая	Не разработаны

Таблица 9. Таблица расчета непараметрической статистики

	AS	D	MSD	MaxS	MinS	R	M
Фактор 1	0,518	0,114	0,338	1	0	1	0,620
Фактор 2	0,013	0,001	0,026	0,167	0	0,167	0,004
Фактор 3	0,776	0,054	0,232	1	0	1	0,843
Фактор 4	0,042	0,022	0,147	1	0,00092	0,999	0,005
Фактор 5	0,315	0,044	0,209	1	0,02021	0,979	0,274
T1	0,044	0,015	0,121	0,831	0	0,831	0,008
T2	0,034	0,008	0,087	0,588	0,00007	0,588	0,007
T3	0,037	0,005	0,073	0,419	0,00005	0,419	0,012
T4	0,500	0,072	0,269	1,000	0,02270	0,977	0,492

В Табл. 9 использованы следующие показатели:

- *AS (average of sample)* – среднее значение выборки,
- *D (dispersion)* – дисперсия,
- *MSD (mean square deviation)* – среднеквадратическое отклонение,
- *MaxS (maximum sample)* – максимальное значение,
- *MinS (minimum sample)* – минимальное значение,
- *R (range)* – размах,
- *M (median)* – медиана.

Результирующие значения линейной корреляции представлены в виде Табл. 10, где каждая ячейка содержит значение коэффициента корреляции двух выборок.

При значении коэффициента корреляции больше или равно 0,6 [Ферстер, 1983], можно утверждать, что между двумя массивами, которые исследуются, существует зависимость. В Табл. 10 соответствующие значения, указывающие на наличие линейной корреляции, выделены жирным шрифтом.

По результатам рассчитанных значений корреляции, выделено следующие зависимости:

1. Очень высокое значение расчетного показателя корреляции, практически равное единице, отмечено между ФПШС (Т1) и ФТС (Т2).
2. Средний показатель корреляции присутствует между фактором первым и использованием Интернет. Это обусловлено тем, что большинство компьютеров, используемых в производственных целях, подключено к Интернет.
3. Достаточно высокий показатель корреляции имеют фиксированная проводная широкополосная связь, фиксированная телефонная связь и мобильная связь. Данный показатель объясняется тем, что количество пользователей любой технологии имеет прямую зависимость от количества населения.
4. Исходя из описанного выше и прямой зависимости количества пользователей ФПШС и ФТС высокое значение коэффициента корреляции между Т3 и Т1, Т2 обусловлено потребностью населения в связи независимо от их местоположения.

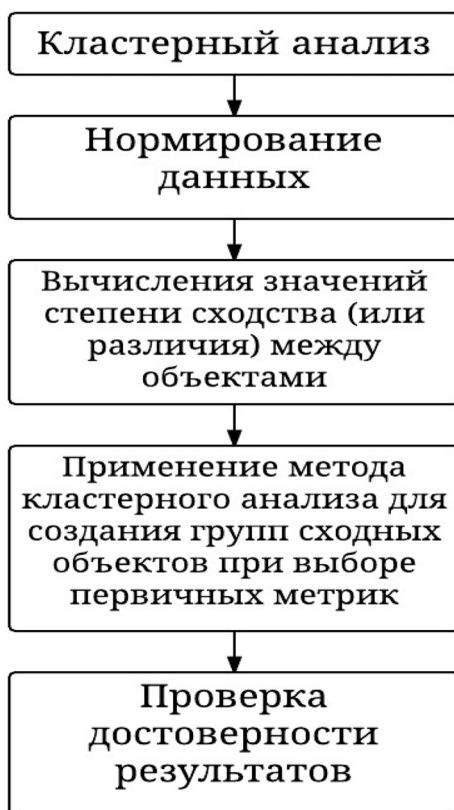


Рисунок 8. Последовательность кластерного анализа

Таблица 10. Таблица рассчитанных значений корреляции

	Фактор 1	Фактор 2	Фактор 3	Фактор 4	Фактор 5	T1	T2	T3	T4
Фактор 1	1,00	0,17	0,18	-0,28	-0,12	0,21	0,20	0,23	0,65
Фактор 2	0,17	1,00	0,06	-0,05	-0,11	0,93	0,97	0,98	0,11
Фактор 3	0,18	0,06	1,00	0,10	0,25	0,00	0,02	0,09	-0,14
Фактор 4	-0,28	-0,05	0,10	1,00	-0,20	-0,08	-0,07	-0,08	-0,22
Фактор 5	-0,12	-0,11	0,25	-0,20	1,00	-0,11	-0,10	-0,10	-0,18
T1	0,21	0,93	0,00	-0,08	-0,11	1,00	0,99	0,89	0,28
T2	0,20	0,97	0,02	-0,07	-0,10	0,99	1,00	0,94	0,22
T3	0,23	0,98	0,09	-0,08	-0,10	0,89	0,94	1,00	0,15
T4	0,65	0,11	-0,14	-0,22	-0,18	0,28	0,22	0,15	1,00

Кластерный анализ представляет собой совокупность математических методов, предназначенных для формирования отдельных групп объектов похожих по информации, свойствами или другим критериям [Дюран, 1977], такие группы объектов называют кластерами. Кластерный анализ широко используется в науке как средство типологического анализа, фактически кластерный анализ – это обобщенное название большого набора алгоритмов, используемых при классификации объектов, Рис. 8.

В любой научной деятельности классификация является одной из фундаментальных составляющих, без которой невозможны построение и проверка научных теорий и гипотез.

Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма [Тгуон, 1939].

Для применения кластерного анализа необходимо проверить удовлетворяют ли данные следующим требованиям:

1. Выборка должна быть однородной.
2. Показатели не должны быть связаны между собой и не должны быть безразмерными.
3. Распределение показателей должно быть близким к нормальному.
4. Отсутствие влияния на значение показателей случайных факторов.

После получения результатов и их последующего анализа возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата, определяемого высокой степенью сходства объектов внутри каждого кластера. Результатом применения процедуры кластеризации [Мандель, 1988] может быть формирование нескольких подгрупп кластеров объектов исследования, в каждом из которых содержится объекты наблюдения „похожи” по некоторым выбранным показателям или по интегрированному показателю.

Для определения схожести используют следующие метрики:

1. Эвклидово расстояние [Мандель, 1988] – наиболее общий тип расстояния, являющийся геометрическим расстоянием в многомерном пространстве

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{k=1}^n (p_k - q_k)^2} \quad (2)$$

2. Квадрат эвклидова расстояния [Мандель, 1988] – используется для придания большего веса более отдаленным друг от друга объектам

$$\rho(x, x') = \sum_i^n (x_i - x'_i)^2 \quad (3)$$

3. Расстояние городских кварталов (манхэттенское расстояние) [Мандель, 1988] – это просто среднее разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для расстояния Евклида. Однако для этой меры влияние отдельных выбросов уменьшается, так как они не возводятся в квадрат

$$\rho(x, x') = \sum_i^n |x_i - x'_i| \quad (4)$$

4. Расстояние Чебышева [Мандель, 1988] полезно использовать, когда необходимо определить два объекта как „различные”, если они отличаются по какой-либо одной координате

$$\rho(x, x') = \max(|x_i - x'_i|) \quad (5)$$

5. Степенное расстояние [Мандель, 1988] используют в том случае, когда необходимо прогрессивно увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются

$$\rho(x, x') = \sqrt[p]{\sum_i^n (x_i - x'_i)^p} \quad (6)$$

В примере применения кластерного анализа использован интегрированный показатель, в качестве которого выбран средний показатель ТП для каждого из видов связи. Для выявления групп стран, близких по уровню принятия интегрированного показателя, применен метод „ближнего соседа” при использовании трех видов метрик: Евклидовой, Чебышева и Манхэттенской.

В Таблице 11 рассмотрены начальные статистические данные для каждой группы показателей, и видно, что они имеют большой разброс, что негативно влияет на построение кластера. Для уменьшения разброса начальных данных применяется процедура их нормирования.

Таблица 11. Данные о количестве пользователей

Страны	Фиксированная проводная широкополосная связь	Мобильная телефонная связь	Фиксированная телефонная связь	Процент пользователей Интернет
1	2	3	4	5
Австрия	1262746	9509198	3614769	60
Азербайджан	267279	4476405	1208219	21
...	...	...	...	...
Молдова	119321	1718217	955713	19
Польша	2690346	31933336	10111327	41
США	53646550	218217541	167908456	67
1	2	3	4	5
Туркменистан	753	1132026	444938	2
Узбекистан	51350	8125834	1801042	11
Украина	1132626	33905591	12011422	11
Франция	12018732	49366166	36033248	53
Швейцария	1791004	7615364	4989000	72

Полученные значения расстояний между кластерами стран с использованием Евклидовой метрики представлены в Табл. 12, при этом учтена симметричность Евклидовой метрики, согласно которой расстояние, например, между Молдовой и Австрией равно расстоянию между Австрией и Молдовой, поэтому Евклидово расстояние рассчитано только для одной из каждой пары стран.

Таблица 12. Матрица расстояний между странами

Страны	Австрия	...	Молдова	Россия	США	Узбекистан	Украина	Франция	Швейцария
Австрия	0	...	0,50	0,74	1,66	0,60	0,60	0,33	0,15
...	...	0	...	...	...	...	...	...	...
Молдова	0,50	...	0	0,67	1,81	0,10	0,19	0,55	0,64
Россия	0,74	...	0,67	0	1,32	0,66	0,53	0,54	0,86
США	1,66	...	1,81	1,32	0	1,83	1,72	1,36	1,68
Узбекистан	0,60	...	0,10	0,66	1,83	0	0,14	0,62	0,74
Украина	0,60	...	0,19	0,53	1,72	0,14	0	0,56	0,75
Франция	0,33	...	0,55	0,54	1,36	0,62	0,56	0	
Швейцария	0,15	...	0,64	0,86	1,68	0,74	0,75	0,40	0

После получения значений Евклидовых расстояний между странами проведена кластеризация объектов с помощью программного пакета статической обработки исходных данных „STATISTICA-8”, в результате которой выявлены группы стран, обобщенный уровень которых, по признакам T1 - T4, достаточно близок. Для большей наглядности приведены дендриты, иллюстрирующие содержание кластеров (Рис. 9 и Табл. 11, 12), описывающие объединение стран в кластеры 1, 2 и 3 по различным показателям.

Таблица 13. Страны, составляющие кластер № 1

Страна	Среднее количество пользователей / коэффициент нормирования данных за 10 лет для:							
	T1	Коеф.	T2	Коеф.	T3	Коеф.	T4	Коеф.
Кластер 1								
Франция	12019886	0,22	36090709	0,22	49366166	0,23	52	0,63
Турция	3434316	0,06	17657764	0,11	46403659	0,21	24	0,29
Великобритания	11225663	0,21	34046706	0,20	66470794	0,30	68	0,82
Германия	14708381	0,27	52743846	0,31	84013461	0,39	65,83	0,79
США	53849396	1	167861456	1	218217541	1	66	0,79
Россия	6647832	0,12	40524691	0,24	135816791	0,60	23,06	0,28
Испания	6062649	0,11	19073325	0,11	42725493	0,20	47,28	0,57
Италия	7469673	0,14	24586898	0,15	75095222	0,34	39,36	0,48

Кластер 1 представляет собой группу стран, имеющих наибольшее значение показателей T1 - T4. Эти страны представлены в Табл. 13. Кластер 2 представляет собой группу стран со средним уровнем развития видов связи. В этом кластере, уровень развития примерно в 1, 2 раза ниже, чем в кластере 1. Данные предоставлены в Табл. 14.

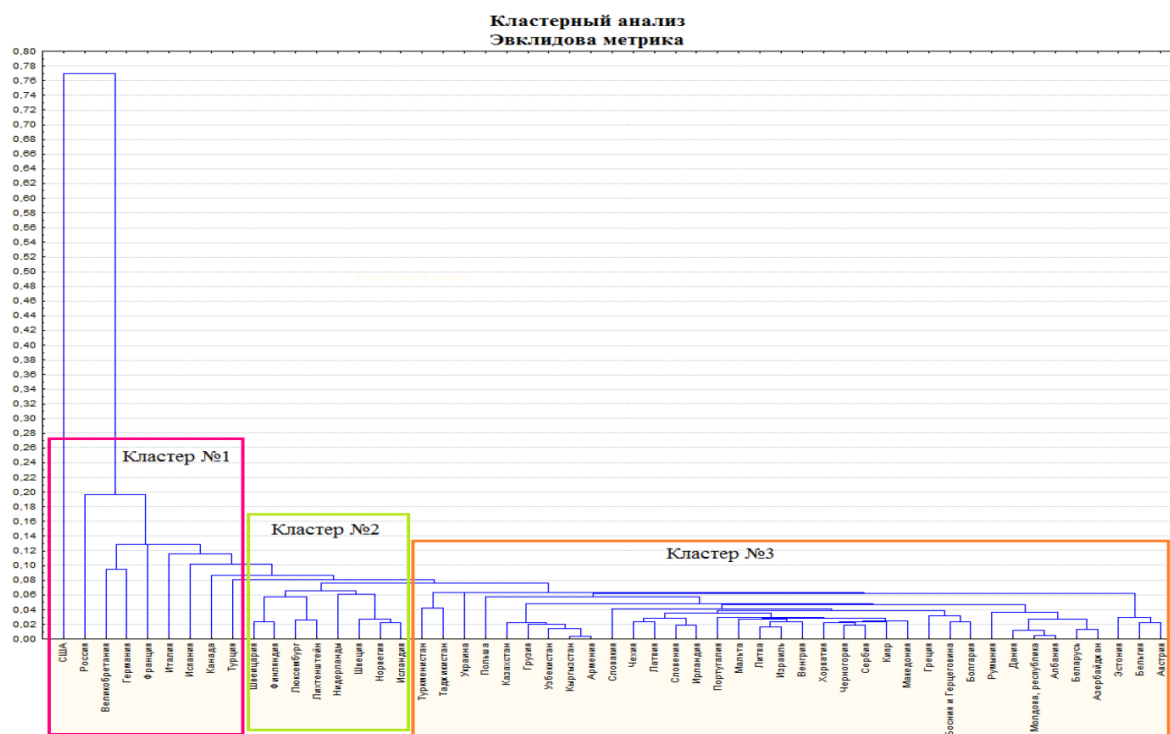


Рисунок 9. Кластер №1, 2, 3

В кластер 3 включены страны, в которых уровень развития ФПШС, МС, ФТС и доступа к Интернет в три раза ниже, чем в кластере 1. Список стран можно определить из общей выборки, исключив из нее страны, принадлежащие к кластерам 1 и 2.

Таблица 14. Страны, составляющие кластер № 2

Страна	Среднее количество пользователей / коэффициент нормирования данных за 10 лет для:							
	T1	Коеф.	T2	Коеф.	T3	Коеф.	T4	Коеф.
Кластер 2								
Исландия	71939	0,001	191338	0,001	300933	0,001	82,7	1,000
Лихтенштейн	10273	0,0002	19707	0,0001	26973	0,0001	65,9	0,796
Люксембург	89519	0,002	255968	0,0015	598625	0,003	67,9	0,821
Нидерланды	4100413	0,076	7719577	0,046	16530562	0,0758	76,2	0,921
Канада	7265525	0,11	18537085	0,04	19230156	0,07	71,4	0,86
Польша	2658205	0,049	10097934	0,06	31933336	0,15	40,5	0,489
Украина	1338558	0,025	1201422	0,07	33905591	0,16	10,9	0,131
Норвегия	1037630	0,019	1989249	0,012	4726491	0,022	76,2	0,921
Финляндия	1064900	0,020	195275	0,012	6181239	0,028	73,2	0,885
Швейцария	1788976	0,033	5090043	0,030	7615364	0,035	71,9	0,869
Швеция	2069195	0,038	5374828	0,320	9419770	0,043	80,2	0,969

Таким образом, по результатам кластерного анализа, можно сделать вывод о существовании трех основных групп стран с равным развитием четырех видов связи.

Первую группу составляют страны с высокоразвитым уровнем экономики. Это Великобритания, США, Италия, Турция, Россия, Франция, Канада, Испания, Германия.

Второй кластер объединяет высокоразвитые страны Центральной и Северной Европы: Исландию, Лихтенштейн, Люксембург, Нидерланды, Норвегию, Финляндию, Швейцарию, Швецию.

Наиболее многочисленным является кластер 3, объединяющий страны, которые быстро развиваются, это Австрия, Азербайджан, Албания, Армения, Беларусь, Бельгия, Болгария, Босния и Герцеговина, Македония, Венгрия, Греция, Грузия, Дания, Израиль Ирландия, Казахстан, Кипр, Кыргызстан, Латвия, Литва, Мальта, Молдова, Польша, Португалия, Румыния, Сербия, Словакия, Словения, Таджикистан, Туркменистан, Узбекистан, Украина, Хорватия, Черногория, Чехия, Эстония.

Результаты кластерного анализа позволяют также оценить степень различия между объектами и степень информационного неравенства между отдельными государствами. Государства, вошедшие в один кластер, имеют минимальный цифровой разрыв, под которым подразумевается разница между объемом использования сферы инфокоммуникаций в развитых и развивающихся странах.

Цифровой разрыв (*digital divide*) – понятие, получившее в последнее время широкое распространение в связи с возросшим значением новых информационно-коммуникационных технологий, усилением процессов глобализации, становлением информационного общества и переходом к глобальной экономике. Чем больше разница в номерах кластеров, в которые вошли отдельные государства, тем больше цифровой разрыв между ними.

Заключение

Ввиду интенсивного развития современного общества невозможно переоценить роль информации являющейся активным звеном всех сфер жизнедеятельности человека, которую целесообразно оценивать путем анализа имеющейся статистической информации. Однако в связи с большим объемом данных, большим количеством программ для обработки статистической информации и специалистов, плохо представляющих условия применения тех или иных методов анализа, оценку имеющейся информации выполнить не так просто.

Исследование существующих методов анализа статистических данных, учет условий применения каждого отдельного метода анализа, формирование методики и алгоритма анализа статистических данных позволяет корректно использовать методы анализа статистической информации для решения задач, относящихся к анализу функционирования и планирования развития информационных сетей.

Благодарности

Настоящая работа выполнена при поддержке интернационального проекта ITHEA XXI Института информационных теорий и их приложений FOI ITHEA и Ассоциации ADUIS Украина (Ассоциация разработчиков и пользователей интеллектуальных систем).

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Литература

- [Cattel, 1966] R.B. Cattel, "The scree test for the number of factors", Multivariate Behavioral Research, 1966, №1, pp. 245-276
- [Gannitskiy, 2010] Illiya Gannitskiy, "Adaptation time-series analysis to number of calls on the modern telecommunication network", Modern Problems of Radio Engineering, Telecommunications and Computer Science: X int. conf.: proceedings Lviv: Publishing House of Lviv Polytechnic, 2010, 197 p.
- [Gannytskyi, 2012] Illia Gannytskyi, "Short-term Forecasts Parameters the Stream of Calls on the Telecommunication Networks", Modern problems of radio engineering, telecommunications and computer science: XIth int. conf.: Lviv: Publishing House of Lviv Polytechnic, 2012, 261 p.
- [ITU] International Telecommunication Union ITU [Электронный ресурс], Режим доступа: <http://www.itu.int/>. Дата обращения: 15.03.2015.
- [Kaiser, 1960] H.F. Kaiser, "The application of electronic computers to factor analysis", Educational and Psychological Measurement, 1960, № 20, pp.141 - 151
- [Tryon, 1939] R.C. Tryon, "Cluster analysis", London: Ann Arbor Edwards Bros, 1939, 139 p.
- [UNECE Publications] United National Economic Commission for Europe UNECE Publications [Электронный ресурс], Режим доступа <http://www.unece.org/publications>, Дата обращения: 31.05.2015.

[UNECE Statistic Database] United National Economic Commission for Europe UNECE Statistic Database
[Електронний ресурс], Режим доступа <http://w3.unece.org/pxweb/>, Дата обращения: 15.03.2015.

[Кендалл, 1976] Кендалл М., Стьюарт А. "Многомерный статистический анализ и временные ряды", Москва: Наука, 1976, 375 с.

[Айвазян, 1989] Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д., "Прикладная статистика: Классификация и снижение размерности", Москва: Финансы и статистика, 1989, 607 с.

[Айвазян, 1998] Айвазян С.А., Мхитарян В.С., "Прикладная статистика и основы эконометрики", Москва: ЮНИТИ, 1998, 1004 с.

[Андерсон, 1963] Андерсон Т., "Введение в многомерный статистический анализ", М.: Государственное издательство физико-математической литературы, 1963, 499 с.

[Андерсон, 1976] Андерсон Т., "Статистический анализ временных рядов", М.: Мир, 1976, 757 с.

[Бондаренко, 2010a] Бондаренко А.А., "Оптимізація методів обробки результатів вимірювань", Збірник тез IV Міжнародної НТК „Проблеми телекомунікацій – 2010”, Київ: КПІ., 2010, с. 246

[Бондаренко, 2010b] Бондаренко А.А., "Методи зберігання і обробки великих обсягів результатів вимірювань", Збірник тез X Всеукраїнської НТК студентів і аспірантів „Інформаційні системи і технології”, Одеса, 2010, с. 59 - 60

[Бондаренко, 2011] Бондаренко А.А., "Некоторые аспекты обработки сверхбольших объемов статистических данных", XI Всеукраїнська НТК студентів і аспірантів „Стан, досягнення і перспективи інформаційних систем і технологій”, Одеса: ОДАХ, 2011, с. 81 - 82

[Боровков, 2010] Боровков А.А., "Математична статистика", М.: Лань, 2010, 704 с.

[Вайсфельд, 2006] Лободзинская И.Г., Вайсфельд Н.Д., Реут О.В., "Математична статистика", Основи аналізу даних Навчально-методичний посібник, Одеса, 2006, 124 с.

[Гайворонская, 2007a] Гайворонская Г.С., Котелевец А.И., "Анализ использования беспроводных технологий в сетях абонентского доступа", Труды III Международной научно-технической конференции „Сучасні інформаційно-комунікаційні технології” /COMINFO’ 2007, 24-28 вересня 2007, р.м. Ялта, смт. Лівадія

[Гайворонская, 2007b] Гайворонская Г.С., Ганницкий И.В., "Анализ параметров потоков вызовов на современной телекоммуникационной сети", Труды III Международной научно-технической конференции „Сучасні інформаційно-комунікаційні технології” /COMINFO’ 2007, 24-28 вересня 2007, р.м. Ялта, смт. Лівадія

- [Гайворонская, 2007с] Г. С. Гайворонская, И. В. Ганницкий, "Оценка влияния вероятностно-временной структуры потоков вызовов на объем оборудования телекоммуникационных сетей", Проблемы телекомунікацій: перша наук.-техн. конф. : зб. тез., К. : НТУУ „КПІ”, 2007, С. 170 – 171
- [Гайворонская, 2009а] Гайворонская Г. С., Котова А. И., "Исследование структурных характеристик абонентских линий непараметрическими методами обработки статистики", Сборник тезисов IX Международной НТК „Математическое моделирование и информационные технологии”, Одесса: ОГАХ, 2009, С.102
- [Гайворонская, 2009b] Г. С. Гайворонская, И. В. Ганницкий, "Анализ параметров потоков вызовов на телекоммуникационной сети", Зв'язок., 2009, № 3 (87), С. 14 – 17.
- [Гайворонская, 2009с] Г. С. Гайворонская, И. В. Ганницкий, "Обработка исходных данных при анализе параметров потоков вызовов на телекоммуникационной сети", Холодильна техніка і технологія, Одеса: ОДАХ, 2009, № 3 (119), С. 73 – 76
- [Ганницкий, 2008а] И. В. Ганницкий, "Особенности обработки статистических данных при анализе параметров потоков вызовов на современной телекоммуникационной сети", Сучасні інформаційно-комунікаційні технології: IV міжнар. наук.-техн. конф.: зб. тез., К.: ДУІКТ, 2008, С. 86 - 87
- [Ганницкий, 2008b] И.В. Ганницкий, "Влияние объема выборки на установление закона распределения длительности обслуживания потока вызовов на телекоммуникационной сети", Математическое моделирование и информационные технологии: IX НТК.: сбор. тез., Одесса: ОГАХ, 2009, С. 104
- [Ганницкий, 2009] И. В. Ганницкий, "Применение метода нормированного размаха для оценки характеристик потока вызовов на телекоммуникационной сети", Цифрові технології, Одеса: ОНАЗ, 2009, Вип. 6, С. 71 – 76
- [Ганницкий, 2010а] И. В. Ганницкий, "Обработка результатов измерений потоков вызовов в современной телекоммуникационной сети", Сучасні інформаційно-комунікаційні технології: VI міжнар. наук.-техн. конф.: зб. тез., К. : ДУІКТ, 2010, С. 64 – 65
- [Ганницкий, 2010b] І. В. Ганницький, А. А. Бондаренко, "Розробка програмного забезпечення для підвищення швидкості обробки даних надвеликого об'єму", Проблеми телекомунікацій: п'ята наук.-техн. конф.: зб. тез., К.: НТУУ „КПІ”, 2011, С. 107.
- [Ганницкий, 2011] І. В. Ганницький, А.А. Бондаренко, "Підвищення швидкості обробки результатів вимірювань параметрів потоку викликів", Холодильна техніка і технологія, Одеса: ОДАХ, 2011, № 1 (129), С. 69 – 72

- [Горяинов, 2001] В. Б. Горяинов, И. В. Павлов, Г. М. Цветкова, А. И. Тескин; Под ред. В.С. Зарубина, А.П. Крищенко, "Математическая статистика: Учебник для вузов", М.: Издательство МГТУ им. Н.Э. Баумана, 2001, 424 с.
- [Дюран, 1977] Дюран Б., Одделл П., "Кластерный анализ", М.: Статистика, 1977, 128 с.
- [Иберла, 1980] Иберла К., "Факторный анализ", М.: Статистика, 1980, 398 с.
- [Ким, 1989] Ким Дж.-О., "Факторный, дискриминантный и кластерный анализ", М.: Финансы и статистика, 1989, 215 с.
- [Кобзар, 2006] Кобзар А.И., "Прикладна математична статистика", М.: Физматлит, 2006, 816 с.
- [Котова, 2008] Г.С. Гайворонская, А. И. Котова, "Оценка параметров абонентских линий", Збірник тез науково-технічної конференції „Сучасні інформаційно-комунікаційні технології”, К.: ДУІКТ, 2008, С. 62.
- [Котова, 2009a] Г. С. Гайворонская, А. И. Котова, "Структурные характеристики абонентских линий", Холодильна техніка і технологія, Одеса: ВЦ ОДАХ, 2009, № 4 (120), С. 78 – 81
- [Котова, 2009b] Г.С. Гайворонская, А.И. Котова, "Исследование структурных характеристик абонентских линий непараметрическими методами обработки статистики", Збірник тез IX міжнародної НТК „Математичне моделювання і інформаційні технології”, Одеса: ОДАХ, 2009, С. 102.
- [Котова, 2010a] О.І. Котова, "Дослідження груп абонентських ліній сільських районів України", Збірник наукових праць, К.: ВІТІ НТУУ „КПІ”, 2010, № 1, С. 11 – 15.
- [Котова, 2010b] Г. С. Гайворонская, А. И. Котова, "Выбор технологии доступа на основании анализа структурных характеристик существующих абонентских сетей", Холодильна техніка і технологія, Одеса: ВЦ ОДАХ, 2010, № 2 (122), С. 63 – 67.
- [Леман, 1964] Леман Е., "Проверка статистических гипотез", М.: Наука, 1964, 501 с.
- [Ллойд, 1989] Ллойд Э., Ледерман В., "Справочник по прикладной статистике", Том 1, М: Финансы и статистика, 1989, 510 с.
- [Ллойд, 1990] Ллойд Э., Ледерман В., "Справочник по прикладной статистике", Том 2, М: Финансы и статистика, 1990, 526 с.
- [Лоули, 1967] Лоули Д., "Факторный анализ как статистический метод", М. Мир, 1967, 144 с.
- [Мандель, 1988] Мандель И.Д., "Кластерный анализ", М: Финансы и статистика, 1988, 176 с.
- [Окунь, 1974] Окунь Я., "Факторный анализ", М. Статистика, 1974, 250 с.

[Орлов, 2004] Орлов А.И., "Прикладная статистика", М.: Экзамен, 2004, 672 с.

[Павлов, 2007] Павлов С.В., "Снижение размерности параметров предоставления инокоммуникационных услуг методом факторного анализа", Труды III Международной научно-технической конференции „Сучасні інформаційно-комунікаційні технології” COMINFO' 2007, 24-28 вересня 2007, р.м. Ялта, смт. Лівадія

[Сахарова, 2008] С.В. Сахарова, "Исследование параметров сетей абонентского доступа", Материалы VIII МНТК „Математическое моделирование и информационные технологии” ММИТ-2008, Одесса: ОГАХ, 2008, с.29.

[Ферстер, 1983] Ферстер Э., Ренц Б., "Методы корреляционного и регрессионного анализа", Руководство для экономистов, М.: Финансы и статистика, 1983, 304 с.

[Харман, 1972] Харман Г., "Современный факторный анализ", М.: "Статистика", 1972, 486 с.

Информация об авторах



Галина Сергеевна Гайворонская – Институт холода, криотехнологий и экоэнергетики им. В. С. Мартыновского ОНАПТ, факультет информационных технологий и кибербезопасности, д.т.н., профессор, зав. кафедрой информационно-коммуникационных технологий, советник ректора по инфокоммуникациям; Украина, Одесса, 65026, ул. Дворянская, 1/3; тел. (048)-720-91-48;

e-mail: gsgayvoronska@gmail.com

Основные направления научных исследований: оптимизация переходных периодов при эволюции телекоммуникационных сетей. Потoki вызовов, нагрузка и межузловое тяготение в сетях. Проблемы создания сетей доступа. Проблема построения полностью оптических сетей и систем коммутации.



Петр Петрович Яцук – Национальная комиссия, осуществляющая государственное регулирование в сфере связи и информатизации, г. Киев, Украина;

e-mail: petr_yatsuk@icloud.com

Основные направления научных исследований: телекоммуникационные сети и технологии, процессы реализации технологий передачи данных в телекоммуникационных сетях.



Юлия Сергеевна Казак – Институт холода, криотехнологий и экоэнергетики им. В. С. Мартыновского ОНАПТ, факультет информационных технологий и кибербезопасности, аспирант кафедры информационно-коммуникационных технологий, ул. Дворянская, 1/3, г. Одесса, Украина, 65082, e-mail: flyger1@bigmir.net

Основные направления научных исследований: использование методов математической статистики в современных телекоммуникационных сетях.

Peculiarities Analysis of Statistical Information in ICT

Galina Gayvoronska, Petro Yatsuk, Yulya Kazak

Abstract: *Huge amounts of statistical information, which has become available at the information society creation, have led to the increase of number of this information analysis methods' incorrect usage. This often entails erroneous conclusions that may lead to negative consequences. Some guidance on the conditions of data analysis methods' usage are given in the work in order to avoid such a situation as well as electronic communications statistics' stepwise analysis method is formed.*

Keywords: *statistical information, information and communication technology, information and communication services.*