9.3. ОБРАБОТКА И ИНТЕРПРЕТАЦИЯ РЕЗУЛЬТАТОВ, ПОЛУЧЕННЫХ ПУТЕМ АНАЛИЗА ДАННЫХ О ФИНАНСОВЫХ ПОТОКАХ КОМПАНИЙ МЕТОДОМ ГЛАВНЫХ КОМПОНЕНТ

Денисенко А.С., аспирант

Национальный исследовательский ядерный университет «МИФИ»

Перейти на Главное МЕНЮ
Вернуться к СОДЕРЖАНИЮ

В работе рассматривается обработка и интерпретация результатов, полученных путем анализа данных методом главных компонент. Опираясь на проведенные ранее исследования, автором разработана методика идентификации наиболее подозрительных компаний реального сектора экономики в Федеральном округе, исследован и интерпретирован закон распределения компаний в зависимости от их интегральных оценок, синтезированы интегральные оценки состояния отрасли экономики в части отмывания доходов в разрезе регионов федерального округа, показан эффект от внедрения полученных результатов в информационную систему для использования в процессе визуально-сетевого анализа.

ВВЕДЕНИЕ

Авторами методом главных компонент проведен анализ матрицы показателей, основанных на данных о финансовых потоках генеральной совокупности компаний отрасли реального сектора экономики отдельного региона. Основные результаты изложены в работе [2].

В работе представлены решения задач обработки и интерпретации результатов применения метода главных компонент. Данные области имеют весьма слабую научную проработку. Авторами разработана методика идентификации наиболее подозрительных компаний из всей анализируемой совокупности, исследован и интерпретирован закон распределения количества компаний в зависимости от их интегральных оценок, синтезированы интегральные оценки состояния отрасли в части отмывания доходов в целом в разрезе регионов, показан эффект от использования результатов в информационной системе для использования в процессе визуально-сетевого анализа.

Решение обратной факторной задачи

Каждая главная компонента является неким новым общим свойством всех объектов исследуемой выборки. Каждая компонента является функцией особенностей каждого из изучаемых объектов. Как правило, исследователь имеет дело с ситуацией, когда одна главная компонента связана с одним или несколькими признаками. Таким образом, особенно важно получение значений главной компоненты для каждого наблюдения. Это позволит ранжировать и классифицировать объекты по полученным рейтинговым оценкам.

Обратимся к модели метода главных компонент и развернем равенство:

$$y'_j = \sum_{r=1}^{n} a_{jr} f_r \qquad (1)$$

для ***j***-го признака:

$$y_j = a_{j1}f_1 + a_{j2}f_2 + \ldots + a_{jn}f_n; \qquad (2)$$

Выразим значения главных компонент через значения признаков. Для ***r***-й компоненты:

$$f_r = \frac{1}{\lambda}(a_{1r}y_1 + a_{2r}y_2 + \ldots + a_{nr}y_n). \qquad (3)$$

Ниже представлены синтезированные рейтинговые оценки причастности компаний (и аффилированных с ней лиц) отрасли в процесс легализации преступных доходов, полученные с использованием 1-й главной компоненты (табл. 1).

Таблица 1

РЕЙТИНГОВЫЕ ОЦЕНКИ

№	Компания	Рейтинг
1	ООО «Компания 1»	69,35441
2	ООО «Компания 2»	22,06919
3	ОАО «Компания 3»	15,75709
4	ООО «Компания 4»	11,84502
5	ОАО «Компания 5»	10,41410
6	ООО «Компания 6»	9,53931
7	ООО «Компания 7»	9,29094

Идентификация наиболее подозрительных компаний

Поскольку 1-я главная компонента характеризует уровень подозрительности компаний, увеличение рейтинговых оценок компаний связано с ростом степени их вовлеченности в противоправную деятельность. В этой связи в первую очередь интерес представляют компании, имеющие наибольшую вовлеченность в процесс легализации. Научный смысл задачи заключается в определении доверительного интервала для математического ожидания. Поскольку по определению метода главных компонент $M[f^{(v)}] = 0$, оценка доверительного интервала для среднего μ выборки может быть оценена по формуле:

$$\mu \pm t_{набл.}\frac{s}{\sqrt{n}} = 0,00 \pm 0,069. \qquad (4)$$

Диапазон между границами доверительного интервала содержит около 90% компаний. Их математическое ожидание равняется нулю. Эти компании имеют наименьшую вовлеченность либо осуществляют полностью законную деятельность. Компании, не попавшие в доверительный интервал (расположенные справа) от его большей границы, составляют 10% от исследуемой совокупности, и являются наиболее подозрительными.

Анализ природы распределения идентифицированной совокупности

За гипотезу H_0 о зависимости количества компаний от их рейтинговых оценок взято предположение, что выборка может быть асимптотически аппроксимирована логнормальной функцией:

$$y = \frac{1}{x\sigma\sqrt{2\pi}}e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}}. \qquad (5)$$

Расчеты показали, что экспериментальные гистограммы при уровне значимости α = 0,05 могут быть описаны логарифмически нормальным законом. Гипотеза подтверждается при проверке критерием хи-квадрат и критерием согласия Колмогорова. График

(рис. 3) демонстрирует, что кривая логнормальной теоретической функции асимптотически аппроксимирует дискретные эмпирические данные (рис. 1).

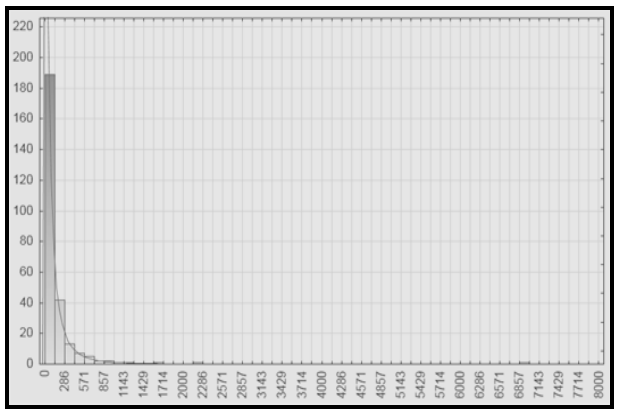

Рис. 1. Асимптотическая аппроксимация логнормальной функцией эмпирических данных о количестве компаний на интервалах рейтинговых оценок

Другим известным методом иллюстрации корректности гипотезы является график типа квантиль-квантиль (рис. 2). На графике отражены квантили эмпирических данных и теоретической логнормальной функции.

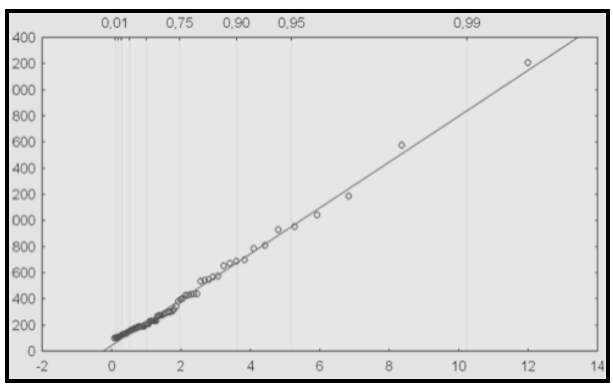

Рис. 2. Квантили эмпирических данных и теоретической логнормальной функции

Физический смысл логнормального распределения тесно связан с повторяющимся делением целого на части.

В качестве классической задачи, иллюстрирующей природу этого распределения, можно привести задачу о состоянии (наследстве). Предположим, землевладелец делит имущество среди наследников. Далее эти наследники делят то же имущество среди своих наследников, и так из поколения в поколение. Если T_j – случайная величина размера состояния в i-м поколении, mou_{j+1} – доля состояния отца, которая достается случайно выбранному наследнику следующего поколения. Если пропорции деления наследства в некоторой мере случайны и существует требуемая независимость между поколениями, то на основании изложенного можно ожидать, что распределение будет асимптотически логнормальным. Аналогичную аргументацию можно применить при рассмотрении повторяющегося деления любой количественно выраженной величины. Например, общество при изучении делится на все более мелкие группы, численность которых в пределе имеет приближенно логнормальное распределение. Как отмечено Айтчисоном и Брауном (1957), эти идеи приводят к разработке теории группирования, в соответствии с которой группы получают путем последовательного деления. Например, при изучении контингента рабочих они вначале подразделяются по уровню квалификации, полу, затем по уровню физического труда и т.д. Распределение численности рассматриваемых групп может быть асимптотически логнормальным [1].

Таким образом, чем больший уровень подозрительности у компаний группы, тем меньше компаний в этой группе. Закон убывания количества компаний в группах с ростом рейтинговых оценок может быть асимптотически аппроксимирован логнормальной функцией.

Использование результатов анализа данных методом главных компонент при визуально-сетевом анализе

Отдельные прикладные экономические задачи требуют визуализации отношений между компаниями анализируемой отрасли в виде неориентированного графа. Компании объединены в кластеры в зависимости от их специализации в отрасли. В результате препроцессорной обработки данных по финансовым потокам компаний методом главных компонент и последующей идентификации наиболее подозрительных, такие компании отмечены на схеме красным цветом (рис. 3, 4). Это позволило наглядно представить на схеме и, таким образом, упростить их выявление при проведении визуально-сетевого анализа.

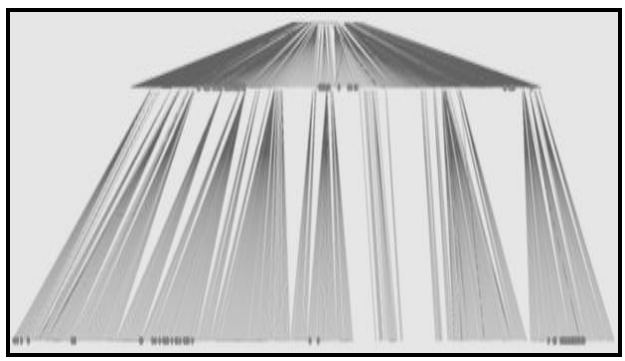

Рис. 3. Роли компаний на схеме цепочек собственников

Успешное применение метода главных компонент для синтеза рейтинговых оценок подозрительности компаний позволило получить оценки обстановки в сфере отмывания доходов в отрасли в целом на уровне регионов за определенный период. Данные о финансовых потоках по компаниям были агрегированы по регионам их регистрации. В результате анализа корреляционной матрицы выделены три главных компоненты. Для решения задачи синтеза интегральных оценок по регионам была выбрана 1-я главная компонента. Проекции показателей по каждому региону на 1-ю главную компоненту позволило синтезировать рей-

тинговые оценки отраслевой обстановки в регионах. Поскольку визуальное представление информации является одним из наиболее эффективных путей ее восприятия, проведено цветовое кодирование оценок и их представление на карте региона (рис. 5). Карта содержит снимок состояния отрасли региона за определенный период.

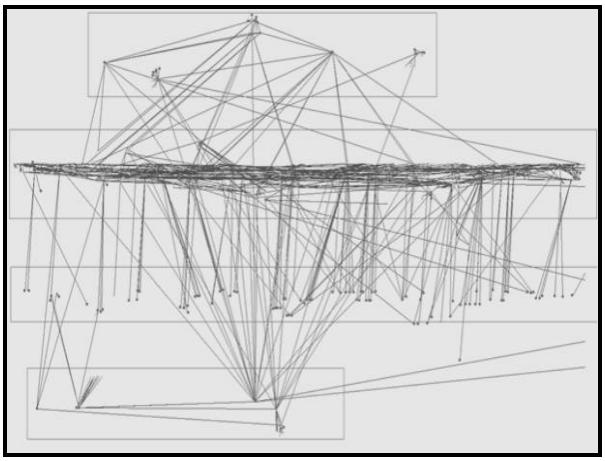

Рис. 4. Использование результатов анализа данных методом главных компонент при визуально-сетевом анализе отрасли

Регулярный мониторинг данных и их последующий анализ методом главных компонент позволит проводить анализ динамики интегральных показателей подозрительности, что в свою очередь способствует решению задачи анализа эффективности предпринимаемых экономических и законодательных мер.

Рис. 5. Цветовое кодирование оценок регионов

ЗАКЛЮЧЕНИЕ

Авторами исследованы способы обработки, интерпретации и прикладного использования результатов, полученных путем анализа данных о финансовых потоках методом главных компонент. Так, представлено решение обратной факторной задачи, в результате чего синтезированы рейтинговые оценки вовлеченности компаний в противоправную экономическую деятельность. Кроме того, идентифицированы наиболее подозрительные компании отрасли. Авторами проанализирована природа их распределения, в результате чего установлено, что зависимость между количеством компаний в группе и уровнем их подозрительности (т.е. значением их рейтинговых оценок) может быть асимптотически аппроксимировано логнормальной функцией.

Получены интегральные оценки состояния отрасли на уровне регионов в результате анализа методом главных компонент данных по региональным финансовым потокам. Кроме того, в результате проведенного исследования авторы пришли к выводу об эффективности и целесообразности применения метода главных компонент в качестве препроцессорного обработчика для повышения эффективности процесса визуально-сетевого анализа при решении различных прикладных экономических задач.

Литература

1. Бартоломью Д. Стохастические модели социальных процессов [Текст] / Д. Бартоломью. – М. : Финансы и статистика, 1985. – 296 с.
2. Денисенко А.С. Факторный анализ и интегральные оценки причастности промышленных предприятий к легализации преступных доходов [Текст] / А.С. Денисенко // Глобальный научный потенциал. – 2014. – №8.
3. Колмогоров А.Н. О логарифмически нормальном законе распределения размеров частиц при дроблении [Текст] / А.Н. Колмогоров // Докл. АН СССР. – 1941. – Т. 31 ; №2. – С. 99-101.
4. Andrukowich P.F. a. o. Abstract painting as a specific – Generale – Language. A Stat. Appr. To the problem // Metron XXIX. 1971. No. 1–2.
5. Dubrov A.M. Data processing with the principal components analysis. M. : Statistics, 1978. 130 p.
6. Dubrov A.M., Mhitarian V.S., Troshin L.I. Multidimensional statistic methods. M. : Finance and statistics, 1998.

Ключевые слова

Метод главных компонент; снижение размерности; факторный анализ; противодействие легализации (отмыванию) преступных доходов и финансированию терроризма; проверка статистических гипотез; визуально-сетевой анализ; доверительный интервал для математического ожидания.

Денисенко Андрей Сергеевич

РЕЦЕНЗИЯ

Статья Денисенко А.С. посвящена исследованию научных областей, имеющих весьма слабую проработку, таких как обработка и интерпретация результатов, полученных путем анализа данных методом главных компонент. В рамках исследуемой тематики, опираясь на проведенные ранее исследования, автором разработана методика идентификации наиболее подозрительных компаний реального сектора экономики в федеральном округе, исследован и интерпретирован закон распределения компаний в зависимости от их интегральных оценок, синтезированы интегральные оценки состояния отрасли экономики в части отмывания доходов в разрезе регионов федерального округа, показан эффект от внедрения полученных результатов в информационную систему для использования в процессе визуально-сетевого анализа.

Актуальность данной статьи обусловлена острой необходимостью применения наукоемких методов аналитической обработки данных в вопросах выявления и оценки рисков легализации преступных доходов в отраслях народного хозяйства Российской Федерации.

Автором проведено исследование в области анализа результатов применения математических методов факторного анализа: показано решение задачи идентификации наблюдений с существенным значением фактора, исследован закон распределения таких наблюдений и проведена его интерпретация в терминах исследуемой предметной области.

Малюк А.А., к.т.н., профессор Национальный исследовательский ядерный университет «МИФИ».