О возможности формализации интуитивных суждений экспертов посредством привлечения характеристик степенных распределений транзакционных данных

А.П. Никитин, О.Д. Чернавская, Д.С. Чернавский

Аннотация

Исследование степенных распределений остается актуальной научной проблемой как в теоретическом, так и в прикладном смысле. В работе продолжено рассмотрение вопроса о возникновении Парето-распределения в динамических системах, находящихся во внешнем шумовом поле. В качестве предметной области исследуется поведение экономических агентов, а именно, динамика клиентских покупок в розничных сетях. Обсуждается вопрос о том, какие прагматически ценные выводы могут быть сделаны из установления того факта, что наблюдаемые экономические показатели обнаруживают степенное поведение, а также из получения достоверных оценок параметров таких распределений. Рассмотрены примеры, когда подобный анализ позволяет формализовать применяемые экспертами термины и понятия, в частности, введя количественные меры лояльности клиентов.

On formalization of intuitive expert notions by applying characteristic parameters of power distributions in transaction data

A.P. Nikitin, O.D. Chernavskaya, D.S. Chernavskii

Abstract

Investigation of power distributions remains an actual scientific problem and is of theoretical as well as practical importance. In the present work the consideration of the Pareto distribution origin in dynamical systems subjected to external noise was continued. Behavior of economic agents, namely dynamics of customer purchases at retail networks is studied as a subject area of the current research. It is discussed which pragmatically valuable conclusions could be drawn from the fact that observed economic indicators revealed the power-law distribution and from the calculation of reliable estimates of this distribution parameters. Several examples are used to show how such analysis made it possible to formalize expert terms and concepts in particular by providing quantitative measures of client’s loyalty.

Введение

В ряде предыдущих публикаций [1, 2] нами рассматривался вопрос о том, при каких условиях в динамических системах возникают степенные распределения (распределения Парето). Была выдвинута задача исследовать класс подобных систем, где такие распределения проявляются, пусть и приблизительно, в широком диапазоне значений наблюдаемых переменных. Для аналитического исследования и компьютерного моделирования было выбрано несколько вариантов, описываемых двухъямными потенциалами специального вида. Изучалось влияние характеристик внешнего шума на размеры диапазона динамической переменной, в котором отклик системы может быть с хорошей точностью аппроксимирован степенным распределением.

В настоящей работе будет показано, как определенные характеристики наблюдаемых степенных распределений могут быть использованы для определения количественных значений ряда важных экономических показателей, формализации экспертных знаний и, тем самым, могут обеспечить информационную базу для прагматически значимых управленческих выводов.

Работа структурирована следующим образом. В первой части будет представлен краткий обзор экономической проблематики, связанной с анализом больших массивов накапливаемых транзакционных данных. Во второй части мы кратко повторим общетеоретическую постановку задачи, изложим методики количественного анализа выборочных степенных распределений и перечислим основные достигнутые в предыдущих работах результаты, формирующие базис текущего исследования. В третьей части на конкретных примерах будет продемонстрировано, как анализ степенных распределений в транзакционных данных позволяет получить количественные оценки ряда актуальных с маркетинговой точки зрения понятий.

1. Анализ транзакционных данных

Компании, специализирующиеся в секторах розничной и оптово-розничной торговли, стараются с помощью компьютеризированных систем учета собирать максимально полную и подробную информацию о каждой отдельной покупке и о каждом клиенте, в частности, выпуская для этой цели фирменные дисконтные карты. Среди первоочередных задач для аналитических подразделений компаний — содержательный анализ всего комплекса взаимоотношений с клиентами. Ключевым с точки зрения методологии является переход от сосредоточенности на одиночных покупках к понятию *lifetime value* [3] — суммарной прибыли от потребителя за период его активности (за весь период сотрудничества с ним).

Для любого бизнеса актуальна задача обеспечения притока новых клиен-

тов, в том числе и ранее приобретавших продукцию/услуги у компаний-конкурентов. Поэтому вторым важным показателем становится стоимость привлечения нового клиента (*CAC — cost of attracting a new customer*). Вместе с тем, удержание уже имеющихся клиентов во многих ситуациях оказывается экономически более оправданным, чем привлечение новых. При этом следует учитывать немаловажный аспект, касающийся анализа структуры клиентской базы, а именно то, что отследить привлечение новых клиентов гораздо проще, чем доказать эффективность мер, направленных на удержание имеющихся.

С позиции интенсивности взаимодействия потребителя и компании Ф. Котлер [4] выделяет 7 категорий покупателей (новый покупатель, повторный покупатель, клиент, адвокат, участник, партнер, сособственник), среди которых наибольший интерес представляет третья — постоянный лояльный клиент. От предшествующих категорий сегмент клиентов отличается также наличием более подробной персональной информации, накапливаемой за период взаимодействия с клиентом по удовлетворению компанией его потребностей. Взаимоотношения с клиентом существенно более продолжительны, чем с простым покупателем, и могут продолжаться не только годы, но и десятилетия.

Применительно к розничной торговле выделение 4 последних категорий не представляется насущно необходимым. В каждый последующий сегмент будет попадать всё более убывающее количество контрагентов компании, соответственно, будет уменьшаться и приходящаяся на сегмент доля выручки. С другой стороны, необходимо вычленить сегмент случайных покупателей, которые совершают покупки без заведения и без предъявления дисконтных карт.

Итак, базовый уровень анализа транзакционных клиентских баз данных подразумевает выделение сегментов случайных, новых, повторных покупателей, а также постоянных лояльных клиентов. Результатом такого анализа должен быть профиль клиента, т.е. статистическая модель клиента, базируясь на которой маркетолог разрабатывает и предлагает решения для удовлетворения потребностей клиента и, следовательно, повышения доходности бизнеса. При построении профиля используются различные имеющиеся сведения о клиенте, но наиболее информативна статистика его покупок.

Популярным методическим подходом к сегментированию клиентских баз данных является *RFM*-модель (сокращение от *Recency, Frequency, Monetary*), которая оперирует показателями, отражающими время, прошедшее от последней покупки клиента, частоту покупок и суммарную стоимость покупок [5]. Сегментация клиентской базы также может производиться по прибыльности, стереотипным действиям, демографическим и иным характеристикам [6].

Стандартная, описываемая в литературе схема использования *RFM*-модели подразумевает разбиение всего массива контрагентов на 5 квинтилей (по 20%) по каждому показателю. Отметим, что при этом порождается 125 сегментов, что ведет к избыточной детализации, даже учитывая тот факт, что

не все эти сегменты оказываются в равной степени «населенными».

Важной стороной аналитической работы является способность предупреждать о потере клиентом лояльности и его возможном уходе (*churn prediction*). Поэтому специальный анализ проводится для клиентов, переставших делать покупки или не совершающих их сравнительно долго.

Попутно отметим, что понятие лояльности в практике розничных компаний рассматривается на нескольких уровнях: это лояльность клиента ко всей торговой сети, лояльность к конкретным торговым маркам, лояльность к конкретному магазину и т.д. Однако менеджеры и маркетологи, как правило, располагают лишь общими представлениями о том, где «проходят границы» сегмента лояльных клиентов. Обычно они оперируют либо простыми, неверифицированными критериями («более двух покупок в год»), либо автоматически включают в списки лояльных покупателей заранее предопределенное число клиентов или фиксированную долю от общего размера клиентской базы.

Итак, одной из важных маркетинговых проблем является формализация критериев, служащих для выделения целевых сегментов клиентской базы, в частности, сегмента лояльных клиентов. В настоящей работе мы продемонстрируем применимость анализа степенных распределений для решения сформулированной подобным образом задачи.

2. Степенные распределения

2.1. Свойства распределения Парето.

Интенсивное научное исследование проблематики обнаружения зависимостей степенного вида в выборочных данных различной природы традиционно ведет свой отсчет от работ социолога В. Парето. Именно ему принадлежит наблюдение, что распределение людей по доходам и/или накоплениям подчиняется степенному закону: доля тех, чьи доходы выше порога x, описывается зависимостью $N(x) = (m/x)^{\alpha_{par}}$, где α_{par} — показатель Парето. Схожие закономерности были найдены лингвистом Дж. Ципфом при изучении частоты встречаемости слов естественного языка. Подобные же степенные зависимости проявляются в целом ряде физических, биологических, социально-экономических систем.

В общем случае распределение исследуемой переменной x будет иметь степенной характер (нами будет использоваться также термин «распределение Парето»), если ее плотность вероятности описывается как:

$$\rho(x) \sim \rho_0 x^{-\alpha} \qquad (1)$$

В (1) $\alpha = \alpha_{par} + 1$ и ниже по тексту такая форма показателя α будет интерпретироваться как характеристика изучаемых степенных распределений.

К настоящему времени распределению Парето посвящена обширная ли-

тература, прежде всего, социально-экономического направления. Предлагаются различные динамические и статистические модели (см., например, [7–8]), объясняющие появление подобных распределений, часто именуемых распределениями с «тяжелым хвостом». Обсуждается более сложное распределение Леви [9], которое в частных случаях переходит в распределение Парето.

Распределения с «тяжелыми хвостами» в реальных ситуациях играют существенную роль при анализе экономических показателей различного рода [10], при оценке вероятности катастроф и иных экстраординарных событий [11]. Легко показать, что для больших $\boldsymbol{x}$ формула (1) дает результаты, на много порядков отличающиеся от аналогичной оценки, полученной из предположения о том, что $\boldsymbol{x}$ подчиняется нормальному, гауссовому распределению.

Перечислим важнейшие свойства степенных распределений. При малых $\boldsymbol{x}$ и любых α выражение (1) неограниченно возрастает. Поэтому под «распределением Парето» обычно понимается распределение, для которого отсутствуют значения $\boldsymbol{x}$ меньше некоторого порога $\boldsymbol{x_{min}}$. При этом условии выражения для плотности вероятности $\rho(x)$ и функции распределения $F(x)$ будут иметь вид

$$\rho(x) = \begin{cases} \dfrac{(\alpha-1)x_{\min}^{\alpha-1}}{x^{\alpha}}, x \geq x_{\min} \\ 0, x < x_{\min} \end{cases} \qquad (2)$$

$$F(x) = P(X \leq x) = 1 - \left(\frac{x_{\min}}{x}\right)^{\alpha-1} \qquad (3)$$

На практике логично следовать соображению, что при $\boldsymbol{x} < \boldsymbol{x}_{\min}$ имеет место какое-либо другое распределение, а при $\boldsymbol{x} = \boldsymbol{x}_{\min}$ оно без разрыва переходит в распределение Парето.

Уместно задаться вопросом, не будет ли более конструктивно подобрать такой класс распределений, которые могут быть пригодны для аппроксимации выборочных характеристик транзакционных данных во всем диапазоне возможных значений. Например, для гамма-распределения с помощью метода максимального правдоподобия могут быть рассчитаны коэффициент формы и коэффициент масштаба, удовлетворяющие наилучшему приближению выборочного распределения теоретическим. Не отрицая возможности такого решения, отметим тот факт, что Парето-распределение зависит от одного параметра α, величина которого может быть сравнительно просто и с достаточной достоверностью оценена из имеющихся в распоряжении исследователя массивов данных, причем и массивов весьма скромной размерности.

Очевидно, что попытка оценить несколько параметров при недостатке данных может привести к тому, что корректно рассчитанные доверительные интервалы для оценок будут крайне широки. Так, в работе [12] для распределений по частоте повторных покупок предлагалась статистическая модель из 6 параметров, которые потом оценивались по методикам максимального прав-

доподобия. Однако дальнейший анализ показал, что удовлетворительную прогностическую точность дают модели, включающие лишь 2-3 параметра.

В целом, следует признать рациональным подход, заключающийся в том, что дополнительные параметры добавляются в модель, если исследуемый эффект не получает адекватного описания в рамках исходной модели. Как будет показано ниже, для решаемых нами задач модель распределения Парето позволяет получить актуальную информацию и сделать на ее основе выводы, важные с практической точки зрения.

Вернемся к свойствам распределения Парето, которые существенно зависят от показателя α. Так, интегралы от моментов распределения, т.е. величин $M_m = x^m \rho(x)$ расходятся на бесконечности при $m \geq \alpha - 1$. Например, при $\alpha \leq 2$ формально можно получить, что среднее значение $\boldsymbol{x}$ бесконечно. Очевидно, что для корректного описания реальных ситуаций выражение (2) должно быть ограничено также и в области больших $\boldsymbol{x}$.

Действительно, в реальных процессах неизбежно возникают факторы, ограничивающие $\boldsymbol{x}$ некоторым предельным $\boldsymbol{x}_{\max}$ таким образом, что $\rho(\boldsymbol{x})$ резко падает при $\boldsymbol{x} > \boldsymbol{x}_{\max}$, а значение интеграла $\int_{x_{\max}}^{\infty} \rho(x)dx$ становится пренебрежимо мало. По сути, $\boldsymbol{x}_{\max}$ также является параметром степенного распределения, однако его определение, как правило, не вызывает трудностей и может базироваться на анализе визуализированных выборочных распределений.

Подчеркнем еще раз, что при таком рассмотрении распределение Парето не является общим законом, а приближенно, хотя и с хорошей точностью выполняется в достаточно широком диапазоне значений $\boldsymbol{x}$. Иногда выдвигаются дополнительные условия, требующие, чтобы этот диапазон простирался в пределах не менее чем нескольких порядков по $\boldsymbol{x}$ ($\boldsymbol{x}_{\max} / \boldsymbol{x}_{\min} \sim 10^2 \div 10^4$). Однако на практике представляют интерес и случаи, когда степенное распределение проявляется в более коротких интервалах изучаемых величин.

2.2. Методики анализа степенных распределений.

На рис. 1 в качестве иллюстрации показан экономический пример степенного распределения: зависимость в двойном логарифмическом масштабе размера подгрупп покупателей, формируемых по признаку одинакового количества купленных ими товаров. Конкретная постановка задачи описана в следующем разделе статьи, пока же обратим внимание на форму визуализации, а также на то, что область степенного распределения простирается приблизительно от $\lg \boldsymbol{x}_{\min} = 0.6$ до $\lg \boldsymbol{x}_{\max} = 2.1$.

Итак, базовая методика анализа степенных распределений подразумевает их визуализацию посредством графического представления в двойном логарифмическом масштабе.

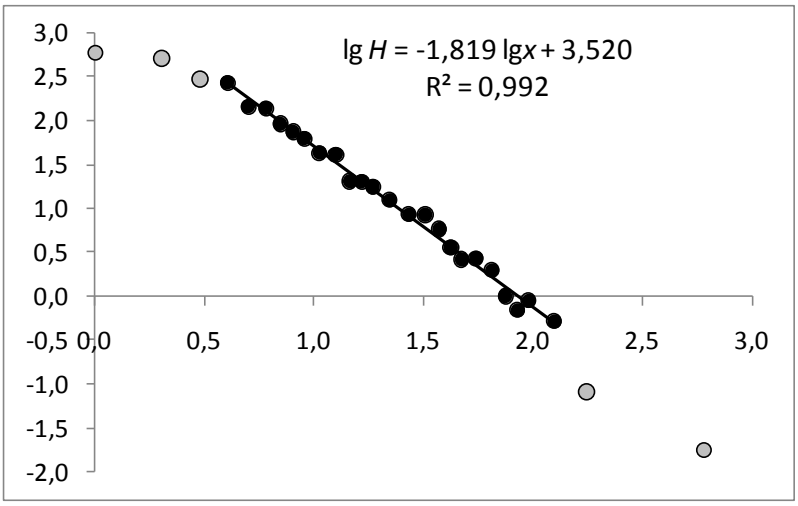

Рис. 1. Зависимость lg***H***(lg***x***) количества покупателей в подгруппах, отобранных по числу купленных товаров ***x*** (сеть **A**, регион **C**, все клиенты, сделавшие первую покупку в 2005 г. — см. в раздел 3)

Рассчитывается выборочная гистограмма плотности распределения ***H***, для чего диапазон значений ***x*** разбивается на выбираемое аналитиком число ***R*** интервалов — «бинов» $\boldsymbol{r_j}$ одинаковой ширины $\Delta R=(x_{max}-x_{min})/R$. Затем подсчитывается и (необязательно) нормируется количество элементов $\boldsymbol{H}(\boldsymbol{s_j})$, попадающих в $\boldsymbol{r_j}$ ($\boldsymbol{s_j}$ —середина j-го бина). В интересующем диапазоне по ***x*** ожидается $\boldsymbol{H} \sim \boldsymbol{x}^{-\alpha}$. Строится график в логарифмической шкале как по ***H***, так и по ***x***, на котором выделяется область линейной зависимости:

$$\lg(H) = -\boldsymbol{k} \lg(x) + \boldsymbol{c} \qquad (4)$$

Коэффициент ***c*** определяется объемом выборки и потому несущественен. Для вычисления коэффициента наклона ***k*** (как аппроксимации показателя α) используются средства регрессионного анализа. Элементарная проверка качества линейной модели проводится по близости коэффициента детерминированности R^2 к 1. Например, адекватные результаты (при достаточном объеме выборочных данных) позволяет получить условие:

$$R^2 > 0{,}99 \qquad (5)$$

Отличие такого подхода от обычной линейной регрессионной модели заключается в том, что остатки $\boldsymbol{\varepsilon_i}$ входят в нее не аддитивно, а мультипликативно. Это сказывается на значениях стандартных ошибок коэффициентов, ширине доверительных интервалов и накладывает серьезные ограничения на применимость статистических критериев проверки состоятельности модели.

Главный недостаток методики, связанной с построением гистограммы выборки, состоит в том, что она хорошо работает лишь при достаточно представительном массиве имеющихся данных. В противном случае приходится расширять интервал ΔR, что ведет к огрублению картины распределения.

Альтернативный способ визуализации можно реализовать, если сначала прологарифмировать диапазон по ***x***, выбрать шаг уже в логарифмическом

масштабе и затем просуммировать количество элементов, попадающих в интервал $\Delta R_l = (\lg x_{max} - \lg x_{min})/R$. Получившаяся гистограмма H_l также может быть аппроксимирована линейным трендом $\lg(\boldsymbol{H_l}) = -\boldsymbol{k_l} \lg(\boldsymbol{x}) + \boldsymbol{c_l}$. При этом следует ожидать $\boldsymbol{k_l} \approx \boldsymbol{k} - 1$. За счет иной схемы суммирования можно надеяться получить более гладкую гистограмму в области больших $\boldsymbol{x}$.

Однако во многих практических задачах накопленных данных недостаточно для построения сколько-нибудь информативной гистограммы выборки. Тогда исследуют ранговые распределения, для чего объекты упорядочиваются в порядке убывания и перенумеровываются, начиная с ранга 1, который получает объект с максимальным $\boldsymbol{x}$.

Привлекается выражение для выборочной функции распределения: для объекта $\boldsymbol{x_r}$ с рангом $\boldsymbol{r}$ она равна $F(x_r) = 1 - \frac{r-1}{N}$. Для степенных распределений согласно (3): $1 - F(x) \sim x^{-(\alpha-1)}$, откуда $r \sim x^{-(\alpha-1)}$ и $x \sim r^{-\frac{1}{(\alpha-1)}}$. В этом контексте эмпирическую степенную зависимость значения $\boldsymbol{x}$ от ранга объекта $\boldsymbol{r}$: $x \sim r^{-\beta}$ можно объяснить присутствием степенного распределения с показателем $\alpha = 1 + 1/\beta$. Понятно, что такое подтверждение существования степенного распределения гораздо более слабое, чем аргументация, базирующаяся на форме гистограмм плотности распределения объектов.

Если число объектов мало, применяются и более изощренные приемы аппроксимации, когда нумерация рангов объектов начинается со специально подбираемого $\boldsymbol{r}_0$ и $x \sim (r + r_0)^{-\beta}$.

Очевидно, что ранговые аппроксимации не могут претендовать на строгое обоснование степенного характера распределения объектов, а, следовательно, подобная интерпретация должна находить независимое дополнительное подтверждение. Вместе с тем в рассматриваемых нами задачах не требуется доказывать, что наблюдаемое в некотором диапазоне по $\boldsymbol{x}$ распределение является степенным и никаким иным. Достаточно выполнения существенно более слабого условия: что это распределение «похоже» на распределение Парето, может быть аппроксимировано им с удовлетворительной точностью.

2.3. Аналитическая модель возникновения степенных распределений.

В работах [1,2] был рассмотрен класс динамических систем, которые, будучи помещены в шумовое поле, в широком диапазоне $\boldsymbol{x}$ порождают распределение Парето. Отклик динамической системы на внешний шум представляет собой нерегулярный хаотический процесс, при этом существенно отличающийся от внешнего шума и характеризующийся иным распределением по величинам флуктуаций. Было показано, что основные характеристики отклика определяются свойствами самой динамической системы.

Для исследования подобных систем используется уравнение Ланжевена:

$$\frac{\partial x}{\partial t} = f(x) + G(t), \qquad (6)$$

где $\boldsymbol{x}$ — наблюдаемая величина, $f(x)$ — функция, описывающая нелинейную динамику системы, $G(t) = g(t) \cdot \xi(t)$ — случайная величина с заданным распределением.

Отметим, что модели, аналогичные (6), активно исследуются с теоретических позиций и применяются для описания социально-экономических процессов: динамики курсов акций, доходности финансовых инструментов и др.

Для решения (6) используется уравнение Фоккера-Планка (см. подробнее [1,2]), позволяющее получить вид плотности распределения $\rho(x,t)$. Стационарное решение этого уравнения имеет вид:

$$\rho(x) = \rho_0 e^{-\frac{2U(x)}{g}}, \qquad (7)$$

где $U(x) = -\int_0^x f(x')dx'$ — потенциал силового поля $f(x)$, g — амплитуда шума.

Из (7) следует, что распределение Парето имеет место, если потенциал $U(x)$ и функция $f(x)$ асимптотически ведут себя как:

$$U(x) \sim \ln x;\ f(x) \sim -1/x \qquad (8)$$

Показатель Парето при этом равен $\alpha = 2 / g$.

В [2] рассмотрено несколько моделей, самая простая из которых:

$$\frac{dx}{dt} = -\frac{a}{x+b} + \xi(t) \qquad (9)$$

Были проведены вычислительные эксперименты по моделированию в зависимости от параметров $\boldsymbol{a}$, $\boldsymbol{b}$, дисперсии $\boldsymbol{D}$ и типа распределения $\xi(t)$, которое представляет собой случайный дельта-коррелированный шум. Шумовая компонента $\xi(t)$ отлична от нуля только в моменты времени $t = \tau \cdot i$, где i — целое число. Таким образом, случайные импульсы с частотой $1/\tau$ вызывают мгновенное смещение координаты $\boldsymbol{x}$ на величину ξ_A. Плотность вероятности реализации конкретного значения ξ_A описывается стандартным некоторым распределением (например, гауссовым) с нулевым средним и дисперсией $D = \sigma^2$.

Нижняя граница распределения Парето $\boldsymbol{x}_{min}$ в модели (9) определяется нелинейной частью, а именно потенциалом $U(x) = a\ln(x+b)$.

Стационарное распределение имеет вид:

$$\rho(x) = \rho_0 \left(\frac{1}{x+b}\right)^{2a}. \qquad (10)$$

Форма (10) соответствует степенному распределению при $x >> b$. так что можно получить выражение для коэффициента Парето: $\alpha = 2a$. Отметим, что в

вычислительных экспериментах хорошее соответствие достигается уже при $x = x_{\min} \sim 3$.

После построения эмпирической гистограммы $\lg(H(x, a))$ для определения неизвестного пока значения параметра $\boldsymbol{a}$ следует исключить из рассмотрения элементы с $x < x_{\min}$, когда характер кривой $\lg(H)$ заметно отклоняется от линейности по $\lg(x)$. Аналогично требуется установить и $x_{\max}$ — верхнюю границу распределения Парето. В вычислительном эксперименте она определялась не моделью, а значением $\boldsymbol{N}$ — длиной временного ряда или, что то же самое, числом объектов в рассматриваемой выборке. Отклонения от Парето наступают, когда число объектов в j-м бине гистограммы H оказывается малым.

2.4. Экономическая интепретация.

Обсудим, какова может быть экономическая интерпретация $U(x)$ и $f(x)$ в выражениях (8). Интересно рассмотреть ситуацию (вернемся к рис.1), когда x — количество единиц товара, приобретенных контрагентом фирмы. В таком случае логарифмический потенциал $U(x)$ позволяет соотнести его с видом кривой общей полезности (*total utility*). Соответственно $f(x)$ можно интерпретировать как показатель, отражающий убывающую с ростом x предельную полезность (*margin utility*). Случайную компоненту $\xi(t)$ можно ассоциировать с отклонениями покупательского поведения от рациональности, когда принятие или отклонение решения о покупке не связано с объективной полезностью данного товара.

Возвращаясь к исходной задаче о шаблонах покупательского поведения, следует отметить, что в такой модели не учитывается уход покупателей, их перетекание к фирмам-конкурентам, предлагающим товары аналогичного назначения. Феномен ухода становится незначительным для сегмента лояльных покупателей, т.е. покупателей, входящих в категорию постоянных клиентов. При малых x доля ушедших клиентов, переориентировавшихся на другие предложения, заведомо значительна, что явно сказывается на форме наблюдаемых выборочных зависимостей.

Таким образом, можно использовать левую границу $\boldsymbol{x}_{\min}$ распределения Парето для количественного определения феномена лояльности. Покупателя следует отнести к сегменту постоянных клиентов, если приобретенное им количество товаров попадает в зону «линейности» $\lg(H)$ от $\lg(x)$.

3. Практический анализ транзакционных данных

3.1. Предварительная подготовка и разведочный анализ.

Для экспериментальной проверки теоретических положений нами были проанализированы транзакционные данные 4 торгово-розничных сетей в двух регионах Российской Федерации. По соображениям сохранения коммерческой тайны конкретные названия сетей, их географическое расположение, деталь-

ные характеристики деятельности остаются нераскрытыми. Мы будем использовать условные наименования сетей в виде латинских букв **A**, **B**, **D**, **E**, а для обозначения регионов — литеры **P** и **C**. Доступный для обработки временной диапазон транзакционных данных различен: более продолжительная «история» имеется для сетей **A** и **B**, более короткая — для сетей **D** и **E**.

Ведущим индикатором различия целевых аудиторий торговых сетей может служить соотношение средней цены (с учетом скидок) предлагаемой товарной единицы, которое составляет по сетям **A**, **B**, **D**, **E** приблизительно 10 – 5 – 1 – 3. Можно высказать гипотезу, что клиенты более «дорогих» сетей должны демонстрировать более высокие показатели лояльности.

Проведена фильтрация выборок по 4 критериям:

1) исключены выявленные злоупотребления;

2) исключены покупки, совершенные сотрудниками;

3) исключены покупки без предъявления клиентской карты;

4) исключены клиенты, делавшие покупки и в регионе **P**, и в регионе **C**.

Общая доля исключенных записей (не учитывая критерий №3) не выходит за пределы 5%.

Как уже говорилось, основным показателем, связанным с кривой полезности, будем считать количество купленных клиентом товаров $\boldsymbol{x}$. Для обеспечения сопоставимости результатов целесообразно выделить подгруппы клиентов по критерию продолжительности периода взаимодействия с компанией $\boldsymbol{T}$. Для определенности скомпонуем эти выборки по календарному году, когда была совершена первая покупка в данной сети. В англоязычной литературе для результата такого отбора применяется термин *cohort* (когорта). Таким образом для каждого фиксированного $\boldsymbol{T}$ будет строиться зависимость количества покупателей от количества купленных ими товарных единиц.

Еще одним критерием при формировании когорт можно взять признак «активности» клиента — потребовать, чтобы клиент совершал хотя бы одну покупку в торговой сети за последний год. При альтернативном варианте анализа такое условие не выдвигается и отбираются все покупатели, первая транзакция которых была зафиксирована в определенном календарном году без учета, продолжается ли их взаимодействие с компанией в настоящее время.

3.2. Получение эмпирических оценок коэффициента Парето.

На рис. 2 приведен пример, на котором видно, что происходит, когда заявленные выше требования к качеству аппроксимации линейным трендом (см. (5)) не выполняются, т.е. уже начинает сказываться случайный разброс количества элементов, попадающих в бины гистограммы.

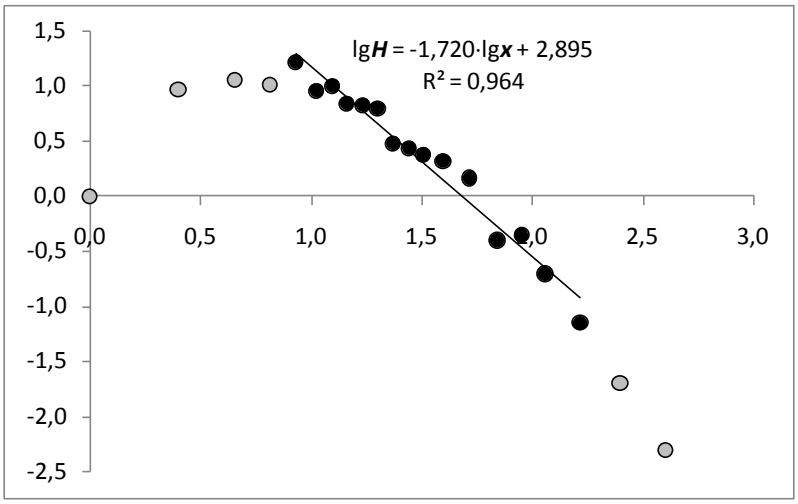

Рис. 2. Зависимость lg***H***(lg***x***) количества клиентов ***H*** в подгруппах, по равному числу купленных товаров ***x*** (Сеть **A**, регион **C**, клиенты, сделавшие первую покупку в 2007 г. и остающиеся активными)

Получив оценки ***k*** коэффициента Парето α для каждой когорты клиентов, вполне естественно построить зависимость α от продолжительности взаимодействия ***T*** с торговой сетью (см. рис. 3). На рис. 3 показаны интервалы «ошибок», соответствующие включению в линейный участок анализируемой лог-лог гистограммы соседних бинов.

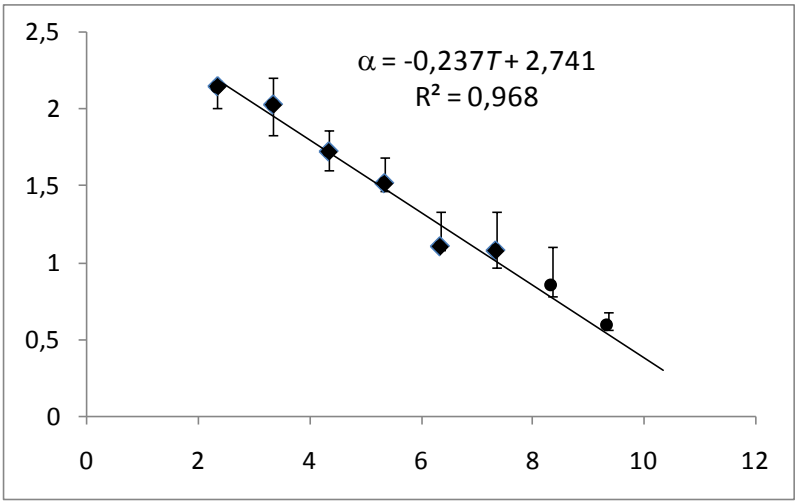

Рис. 3. Зависимость коэффициента Парето α от интервала лет ***T*** с момента первой покупки (активные клиенты сети **A** в регионе **C**)

Рис.3 позволяет заметить, что коэффициент Парето α зависит от ***T***, причем эта зависимость удовлетворительно описывается линейным трендом $\alpha = \alpha_0 + \nu\, \boldsymbol{T} = 2{,}762 - 0{,}239\, \boldsymbol{T}$. Таким образом, фигурирующий в выражении для потенциала $U(x) = a\ln(x+b)$ коэффициент a аналогичным образом зависит от времени $a = a_0 + w\, \boldsymbol{T}$. Исходя из $a = \alpha/2$ и оценки $\alpha_0 = 2{,}762 \pm 0{,}085$ ($p = 0{,}9$) можно получить «мгновенное значение» a_0 ($T = 0$) =1,38±0,04.

Располагая оценкой параметра α для сети **A** в регионе **C**, проверим, наблюдается ли подобная зависимость и в регионе **P**. Так как предлагаемый товарный ассортимент идентичен, то различия в шаблонах поведения покупателей, если таковые будут найдены, можно объяснить региональными (географическими) особенностями. Действуя по той же процедуре, получаем α = 2,616 – 0,216 $\boldsymbol{T}$, что дает в оценку для a_0 = 1,32±0,05. Так как доверительные интервалы оценок a_0 для двух исследованных регионов пересекаются, то можно говорить об отсутствии значимых различий между сегментами клиентов в регионе **C** и в регионе **P** с позиции сравнения кривых полезности.

3.3. Определение формы критерия лояльности клиентов.

Использованный алгоритм позволяет проследить (хотя бы качественно и со всеми необходимыми оговорками) зависимость левой границы $\boldsymbol{x}_{\min}$ распределения Парето от $\boldsymbol{T}$. Оказывается, что на выборке активных клиентов сети **A** $\boldsymbol{x}_{\min}$ можно считать постоянным в диапазоне 2002-09 гг. с весьма близкими средними значениями для регионов: 7,8 для **P** и 8,5 для **C**. Таким образом для активных клиентов сети **A** справедлив следующий критерий их отнесения к категории лояльных: минимум z_{act}=8 купленных товаров вне зависимости от года, когда была совершена первая покупка.

Небезынтересно рассмотреть те же зависимости, по-другому сформировав выборки клиентов. Откажемся от требования «активного» статуса клиента, отбирая всех, кто совершил первую покупку в торговой сети за конкретный календарный год, включая и «ушедших» клиентов, переставших взаимодействовать с сетью. Можно ожидать, что за счет расширения выборки будут легче выделяться участки линейности на лог-лог графиках, а также лучше выполняться условие близости R^2 к 1 (см. рис. 1).

Для сформированных по таким правилам выборок клиентов для сети **A** в регионе **C** получаем α = 2,722 – 0,141 $\boldsymbol{T}$, что дает в оценку для a_0 =1,36±0,05 (p = 0,9). Для той же сети в регионе **P**: a_0 =1,42±0,08.

Таким образом, можно заключить, что характеристики степенного распределения для сети **A** не обнаруживают значимой зависимости от способа формирования когорты-выборки клиентов. Из этого следует правдоподобность вывода, что и опосредованно оцениваемый параметр кривой полезности также не модифицируется при изменении алгоритма формирования выборки.

Однако критерий лояльности, формулируемый по величине $\boldsymbol{x}_{\min}$, видоизменяется — для отнесения к сегменту лояльных клиентов теперь требуется совершить лишь z_{any}=4 покупки (как для региона **C**, так и для региона **P**). Налицо противоречие — активные клиенты, не проходящие по первому варианту критерия лояльности, оказываются лояльными по второму критерию. Ведущим фактором, вызывающим такое расхождение в критериях, является существенная доля неактивных контрагентов, ограничивших своё взаимодействие с ком-

панией одной единственной покупкой, в результате которой было приобретено некоторое количество товаров.

Для прояснения вопроса следует перекомпоновать когорты, проследив зависимость $\boldsymbol{x}_{\min}$ от года совершения последней покупки в сети (т.е. от временной привязки факта прекращения взаимодействия с компанией). Оказывается, что для недавних клиентов (2010-11 г.) соответствующий показатель преобразуется в коэффициент 4 товара/ год. Т.е. для покупателя, чья первая транзакция совершена полтора года назад, критерий его отнесения к постоянным будет состоять в достижении порогового значения в 6 купленных товаров.

Итак, если считать активными тех клиентов, которые совершили последнюю покупку в произвольный момент из интервала последних 12 календарных месяцев, то для них критерий лояльности к сети **A** можно сформулировать как:

$L_1\,(t_{now}) = \boldsymbol{x}(t_{first}, t_{now}) \geq \mathrm{Min}\,(z_{act}, z_{any} \cdot \mathrm{datediff}\,(year, t_{first}, t_{now}))$

где t_{now} — дата проведения анализа, t_{now} — дата первой покупки клиента, $x\,(t_1,t_2)$ — количество купленных товаров в интервале $[t_1,t_2]$, $z_{act} = 8$, $z_{any} = 4$ — эмпирически найденные для изучаемой сети **A** параметры, а функция datediff вычисляет неокругленную величину временного интервала в единице измерения, тип которой задается первым аргументом, между датами, заданными вторым и третьим аргументами. В этом понимании лояльность определяется как текущая лояльность активных клиентов.

Однако лояльность клиента можно оценивать и ретроспективно, относя ее к периоду совершения покупок. Тогда клиента можно считать лояльным торговой сети на момент времени $\boldsymbol{t}$, т.е. по крайней мере, в момент совершения им некоторых из своих покупок, если им было куплено не менее $z_{any} = 4$ товаров:

$L_2\,(t) = \boldsymbol{x}(t_{first}, t) \geq z_{any}$.

Понятно, что критерий лояльности L_2 оказывается более слабым, чем L_1. Однако апелляция к некоторому неопределенному моменту предыстории клиента позволяет снять явное противоречие между этими критериями.

3.4. Анализ межрегиональных и межсетевых различий.

Аналогичные вышеприведенным оценки были получены для сети **B**: в регионе **C** для активных клиентов $a_0 = 1{,}83 \pm 0{,}08$ (см. рис. 4), для всех клиентов $a_0 = 1{,}84 \pm 0{,}15$, а в регионе **P** $a_0 = 2{,}15 \pm 0{,}11$ и $a_0 = 2{,}02 \pm 0{,}10$, соответственно. В данном случае на уровне $p = 0{,}9$ можно говорить о статистическом различии сегментов активных клиентов в исследованных регионах.

При определении порога лояльности для активных клиентов сети **B** в регионе **C** отмечаем отсутствие стационарного уровня, обнаруженного для сети **A** в том же регионе: устанавливается линейная зависимость от $\boldsymbol{T}$ с коэффициентов наклона $z_{act} \approx 4$ купленных товара в год (см. рис. 5). Аналогичный линейный тренд (с несколько большим значением z_{any}) проявляется и на выборке всех клиентов этой сети в регионе **C**. В регионе **P** активные клиенты сети **B**

характеризуются более «крутой» кривой лояльности с $z_{act} \approx 6$.

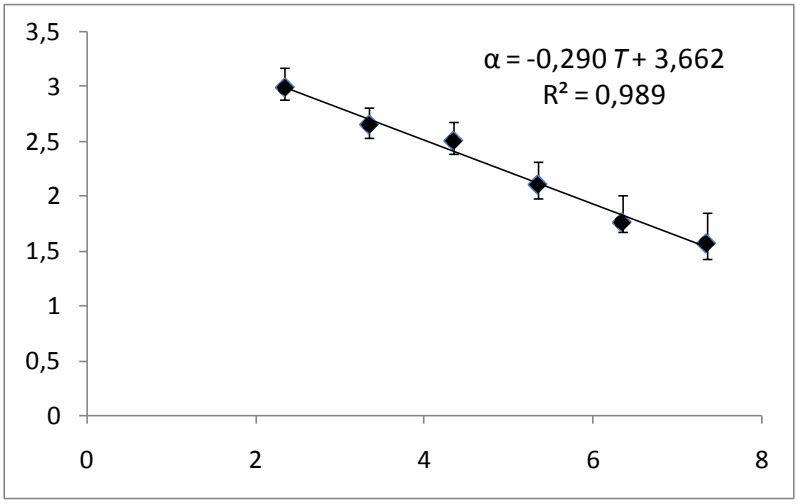

Рис. 4. Зависимость коэффициента Парето α от интервала лет ***T*** с момента первой покупки (активные клиенты сети **B** в регионе **C**)

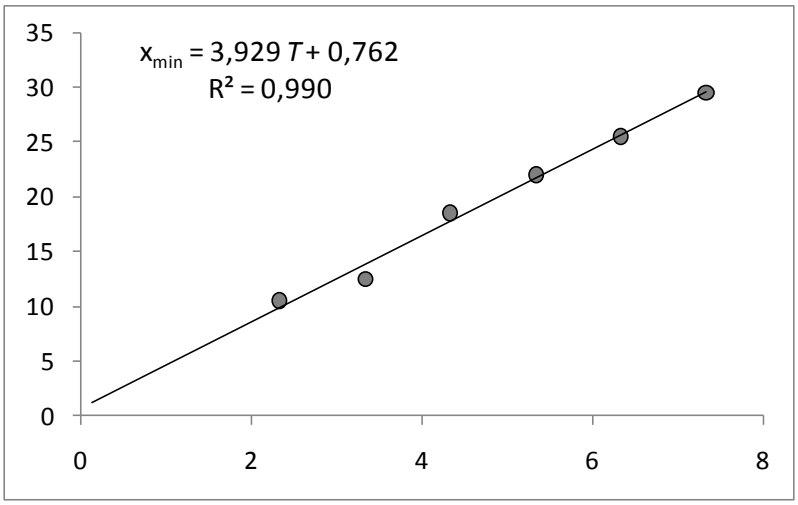

Рис. 5. Зависимость левой границы распределения Парето x_{min} от количества лет ***T*** с момента первой покупки (активные клиенты, сеть **B**, регион **C**)

Для сети **D** в регионе **C** по активным клиентам получаем $\alpha = 4{,}734 - 0{,}606\,T$, что дает оценку для параметра кривой полезности $a_0 = 2{,}37 \pm 0{,}03$. В регионе **P** тренд имеет существенно отличающийся коэффициент наклона: $\alpha = 4{,}200 - 0{,}258\,T$ и $a_0 = 2{,}10 \pm 0{,}25$ ($p = 0{,}9$). Из-за ограниченности периода времени доверительный интервал a_0 весьма широк. Уровень лояльности для обоих регионов можно рассматривать как постоянный (оценки динамики приходится делать с осторожностью из-за недолгой истории функционирования торговых предприятий этой сети) с $z_{act} \approx 15$.

Сеть **E** характеризуется сравнительно короткой историей функциониро-

вания и гораздо меньшим объемом клиентской базы по сравнению с другими сетями. Заранее понятно, что при количестве клиентов в пределах не более нескольких сотен гистограмма плотности выборочного распределения будет настолько «изрезана», что сколько-нибудь объективизированное выделение участков линейности обеспечить крайне затруднительно. Воспользуемся этим поводом для иллюстрации ранговых распределений (см. рис. 6).

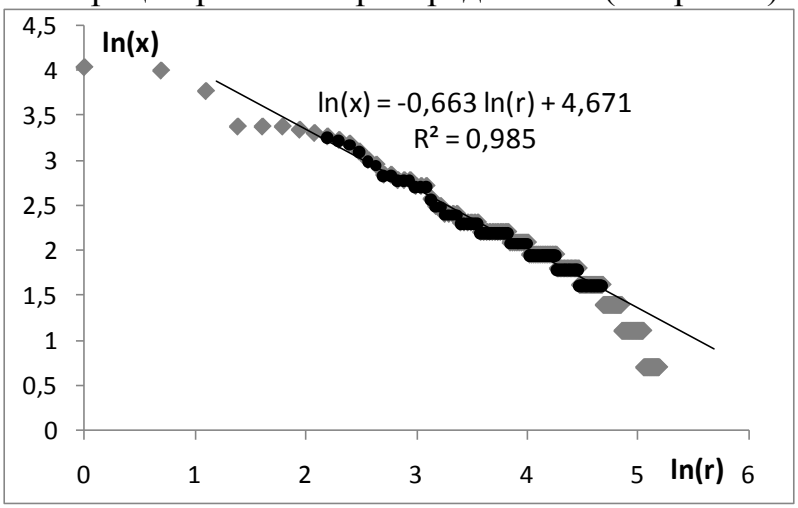

Рис. 6. Ранговая зависимость для количества купленных товаров (активные клиенты, сеть **E**, регион **P**, первая покупка в 2010 г.)

Из-за того, что ***x*** является целочисленной величиной, ранговая зависимость на рис. 6 имеет ступенчатый вид, что тем не менее не мешает выбрать участок линейности. Оценка коэффициента $\beta = 0{,}663 \pm 0{,}013$ ($p = 0{,}9$) позволяет рассчитать коэффициент степенной зависимости $\alpha = 1 / \beta + 1 = 2{,}51 \pm 0{,}03$, что хорошо согласуется с оценкой α из гистограммы плотности $2{,}56 \pm 0{,}66$ (p=0,9), которая, как можно заметить, оценивается с точностью более чем в 20 раз хуже. Порог лояльности активных клиентов для сети **E** линейно растет с темпом 4,5 купленных товара в год.

Заключение

Подводя итог, можно отметить, что степенные распределения с различными показателями Парето α весьма часто (хотя отнюдь и не повсеместно) проявляются при изучении количественных показателей, характеризующих структуру клиентской базы. Детальное изучение этих распределений, как, например, проанализированные выше выборочные гистограммы распределения клиентов в зависимости от количества купленных ими товаров, позволяет получить дополнительную информацию о механизмах формирования клиентской базы, о ее динамике (как о факторах роста, так и о факторах сужения базы). Кроме того, удается формализовать ряд критериев, которые исходно были выражены только в вербальной, словесной форме. Одним из таких критерием является продемонстрированный в настоящей работе порог для определения

сегмента лояльных клиентов.

Рассчитанные параметры степенных распределений позволяют делать выводы о схожести или несхожести характеристик сегментов клиентской базы, формируемых по определенным правилам (например, в региональном разрезе, в разрезе текущей активности или неактивности покупателей). Эти же методики позволяют судить о наличии значимых различий между сегментами клиентов, взаимодействующих с разными торговыми сетями. Вполне очевидно, что такой анализ должен сопровождаться содержательной интерпретацией выявленных особенностей в терминах рассматриваемой предметной области.

Список литературы.

1. *Чернавский Д.С., Никитин А.П., Чернавская О.Д.* О возникновении распределения Парето в нелинейных динамических системах // Биофизика. 2008. –т.53. –№2. –с. 351-358.
2. *Nikitin A.P., Chernavskaya O.D., Chernavskii D.S.* Pareto distribution in dynamical systems subjected to noise perturbation // Physics of Wave Phenomena. – 2009. – v.17. – No.3. – c.207-217
3. *Hughes A.M.* Strategic database marketing – 3rd Ed. – McGraw-Hill, 2006. – 438 p.
4. *Kotler Ph.* Kotler on marketing: how to create, win, and dominate markets. – NY, FreePress, 1999. – 256 p.
5. *Drozdenko R.G., Drake P.D.* Optimal database marketing: strategy, development, and data mining –Sage Publications, Th. Oaks, 2002. – 298 p.
6. *Меркулина И.А., Никитин А.П.* Экономические приложения интеллектуального анализа данных. – М.: ВГНА, 2007. – 370 с.
7. *Kechedzhi K.E., Usatenko O.V., Yampol'skii V.A.* Rank distribution of words on correlated symbolic systems and the Zipf law // Physical Review E. – 2005. – v.72. – No.4. – 046138.
8. *Solomon S., Richmond P.* Stable power laws in variable economies; Lotka-Volterra implies Pareto-Zipf // Eur. Phys. J. B. – 2002.– v.27. – p.257-261
9. *Романовский М.Ю., Романовский Ю.М.* Введение в эконофизику. Статистические и динамические модели. – М.: ИКИ, 2007. – 280 с.
10. *Малинецкий Г.Г., Подлазов А.В.* Парадигма самоорганизованной критичности. Иерархия моделей и пределы предсказуемости // Известия ВУЗов. Прикладная нелинейная динамика. – 1997. – т. 5. – №5. – с. 89–106.
11. *Малинецкий Г.Г.* Управление риском. Риск, устойчивое развитие, синергетика. – М.: Наука, 2000. – 432 с.
12. *Fader S.P., Hardie B., Huang C.-Y.* A dynamic changepoint model for new product sales forecasting // Marketing Science. – 2004. – No.23. – p.50-65.