МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ. ЧИСЛЕННЫЕ МЕТОДЫ

УДК 519.87

МОДЕЛИ ОЦЕНКИ ПАРАМЕТРОВ ТЕЛЕКОММУНИКАЦИОННОГО ТРАФИКА В АВТОМАТИЗИРОВАННЫХ ИНФОРМАЦИОННО-УПРАВЛЯЮЩИХ СИСТЕМАХ

А.Н. Гуда, М.А. Бутакова, Н.А. Москат

ГОУ ВПО «Ростовский государственный университет путей сообщения», г. Ростов-на-Дону

Рецензент д-р физ.-мат. наук, профессор И.В. Павлов

Ключевые слова и фразы: математические модели телетрафика; оценка параметров; показатель Харста; *R*/*S*-статистики; самоподобные процессы.

Аннотация: Представлены модели анализа телекоммуникационного трафика в информационно-управляющих системах. Модели учитывают свойства самоподобия трафика, долговременной зависимости и тяжелого хвоста в распределении. Приведены численные и графические методы анализа.

Определение самоподобного процесса

При построении моделей, использующих математический аппарат теории очередей, как правило, потоки заявок в автоматизированную информационно-управляющую систему являются простейшими, а время обслуживания заявок – случайной экспоненциально распределенной величиной. Как известно, не во всех моделях потоков событий вероятность появления следующего события зависит только от времени, прошедшего с момента совершения предыдущего события, и не зависит от всей предыстории появления событий ранее [16]. Существуют потоки, в которых вероятность появления следующего события зависит от наступления событий в предыдущих интервалах времени. Типичным примером таких потоков являются потоки с ограниченным последействием. Для них задается конечный набор функций распределения для соседних интервалов τ_k между поступлением k событий.

Гуда А.Н. – доктор технических наук, профессор, заведующий кафедрой «Информатика», проректор по научной работе и информатизации; Бутакова М.А. – доктор технических наук, профессор кафедры «Информатика», e-mail: butakova@rgups.ru; Москат Н.А. – ассистент кафедры «Вычислительная техника и автоматизированные системы управления»; РГУПС, г. Ростов.

Однако в случае с телетрафиком такие модели не могут адекватно отразить реальный поток событий, так как в нем обнаруживается долговременная зависимость, то есть число событий на заданном временном интервале зависит от числа событий, поступивших в весьма отдаленные от него интервалы времени. Типичным способом измерения такой зависимости для случайных процессов является определение функции корреляции.

Как определил в своей работе [19] В.И. Нейман, «три источника и три составные части теории самоподобных процессов» выражены в медленном убывании дисперсии, долговременной зависимости и флуктуационном характере спектра мощности таких процессов. Самоподобные процессы в литературе также часто называют автомодельными.

В качестве значения случайного процесса будем рассматривать число событий, поступающих в систему в единицу времени. Понятно, что это неотрицательная случайная величина. Случайный процесс рассмотрим как дискретную последовательность таких величин, то есть аргументом будем считать порядковый номер такой единицы времени: $X = \{X_1 : t = 0, 1, 2, ...\}$.

Положим, что все рассматриваемые далее случайные процессы относятся к стационарным случайным процессам с ограниченной ковариацией, то есть $\mathrm{Cov}(X_1, X_{i+k}) = (X_1 - \overline{X})(X_{i+k} - \overline{X}) < +\infty$ с дисперсией $D(X) = \overline{(X_1 - \overline{X})^2} = \sigma^2$ и автокорреляционной функцией $r(k) = \dfrac{\mathrm{Cov}\,(X_t, X_{t+k})}{D(X)}$.

Для того чтобы охарактеризовать принадлежность процесса к классу процессов, имеющих долговременную зависимость или самоподобие, необходимо рассмотреть агрегированные процессы, построенные с помощью усреднения значений исходного процесса на пересекающихся временных интервалах:

$$X^{(m)} = \{X_k^{(m)} : k = 1, 2, 3, ...\}; \;\; X_k^{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{(k-1)m+i} \, .$$

Очевидно, что агрегированные процессы также будут стационарны и иметь ограниченную ковариацию.

Убывание дисперсии асимптотически описывается соотношением

$$D\left(X^{(m)}\right) \approx am^{-\beta}, \quad 0 < \beta < 1, \quad m \to \infty \, ,$$

то есть вариация агрегированных процессов – средних выборок – уменьшается медленнее, чем величина, обратная размеру выборки.

Долговременной зависимостью в самоподобных процессах называют наличие расходимости автокорреляционной функции процесса:

$$\sum_k r(k) = \infty, \quad r(k) \approx k^{-\beta} \, .$$

Это означает, что убывание автокорреляционной функции происходит гиперболически медленно.

Наконец, говоря о флуктуационном характере спектра мощности, понимают под этим аналогию со спектром мощности флуктуаций электронного потока:

$$f(\omega) \approx c\omega^{-\gamma}, \;\; \omega \to 0, \;\; \gamma = 1 - \beta,$$

где $f(\omega) = \sum_k r(k) \mathrm{e}^{-i\omega k}$.

Наличие перечисленных выше свойств у случайного процесса означает, что его автокорреляционная функция совпадает с автокорреляционными функциями агрегированных процессов точно $r(k) = r^m(k)$ или асимптотически $r(k) \to r^m k,\ m \to \infty$.

Собственно эти соотношения и определяют название самоподобного процесса: корреляционные свойства такого процесса, усредненного на различных временных интервалах, остаются неизменными.

Важным параметром, характеризующим «степень» самоподобности случайного процесса, является оценка показателя Харста [10]. Параметр Харста находится в интервале $0{,}5 < \mathrm{H} < 1$. Для процессов, не обладающих свойством самоподобия, параметр Харста равен величине в 0,5, а для фрактальных процессов с долговременной зависимостью данный параметр изменяется в пределах 0,7…0,9.

Самая известная из таких оценок основывается на следующих рассуждениях. Рассматривается выборка величиной n анализируемых данных $x_1,...,x_n$, для которой определяется выборочное среднее $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$ и выборочное стандартное отклонение $S_n = \sqrt{\frac{1}{n}\sum_{k=1}^{n} x_k^2 - \left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)^2}$.

Обозначив $R_n = \max_{k \le n}\left(\sum_{i=1}^{k} x_i - k\overline{X}_n\right) - \min_{k \le n}\left(\sum_{i=1}^{k} x_i - k\overline{X}_n\right)$, получаем нормированные порядковые статистики R_n / S_n, которые в случае наличия в данных свойства автомодельности будут эквивалентны $R_n / S_n \overset{d}{=} cn^{\mathrm{H}}$, $c = \mathrm{const},\ n \to \infty$. Рассматривая выборку из n значений и разбивая ее на k неперекрывающихся блоков, можно построить график зависимости $\ln(R(t_i, j)/S(t_i, j))$ от $\ln(j)$, где $t_1 = 1,\ t_2 = n/k + 1,\ t_3 = 2n/k + 1$, …, $(t_i - 1) + j \le n.$ По углу наклона прямой относительно оси абсцисс можно оценить величину показателя Харста.

Анализ трафика при наличии свойства долговременной зависимости

Самоподобные модели могут проявлять свойство долговременной зависимости, что означает проявление зависимости между событиями через достаточно большие промежутки времени. В задачах исследования функционирования компьютерных работ с сетями, присутствие или отсутствие долговременной зависимости играет определяющую роль в поведении, предсказанном аналитическими моделями.

Стационарный дискретный процесс обладает долговременной зависимостью, если автокорреляционная функция $r(k)$ не суммируема $\sum_k r(k) = \infty.$

Самые простые модели с долговременной зависимостью – самоподобные процессы, которые характеризуются гиперболически убывающими автокорреляционными функциями. Самоподобный и асимптотически самоподобный процессы являются особенно интересными, потому что долговременная зависимость может характеризоваться одним параметром H, который может быть оценен. Наиболее известными методами оценки параметра Харста H являются оценка Виттла [7] и оценка Хилла.

Рассмотрим некоторые свойства процессов с долговременной зависимостью. Возьмем $X = (X_n, n = 1, 2, ...)$ – стационарный дискретный процесс с функцией автокорреляции *r*. Допустим, что процесс обладает долговременной зависимостью, то есть $\sum_{n \geq 1} r(n) = \infty$. В этом случае автокорреляционную функцию можно представить при помощи асимптотического приближения $r(n) \sim n^{-\beta} L(n)$, где $0 < \beta < 1$, $\left(\mathsf{H} = 1 - \frac{\beta}{2}\right)$ и L – медленно меняющаяся функция. Тогда имеет место

$$\sum_{n \geq 1} n^{-\beta} L(n) = \infty \qquad (1)$$

и суммируемость/несуммируемость в данном случае не зависит от функции *L*. Это показывает, что соотношение $r(n) \sim n^{-\beta} L(n)$ может быть использовано в качестве определения процесса с долговременной зависимостью, который характеризуется гиперболически убывающей автокорреляционной функцией. Заметим также, что расходимость в формуле (1) гарантирует невырожденную корреляционную структуру процесса $X^{(m)} = \left\{X_j^{(m)}, j = 1, 2, ...\right\}$.

Долговременная зависимость стационарного процесса в непрерывном времени $X = \{X_t, t \in \mathbb{R}\}$ может также быть определена через автокорреляционную функцию $r(t) = \frac{\mathrm{Cov}(X(0), X(t))}{\mathrm{Var}(X(0))}$, $t \in \mathbb{R}$ условием $\int_0^\infty r(t)dt = \infty$.

Итак, процесс *X* называется долговременно зависимым, если автокорреляционная функция не суммируема $\sum_k r(k) = \infty$.

Анализ трафика при наличии хвоста распределения

В данном разделе рассматриваются модели анализа трафика при наличии хвоста распределения [4].

Функция распределения имеет тяжелый хвост, если

$$M(\lambda) = \{E(\exp(\lambda X)\} = \infty, \ \forall \lambda > 0.$$

Распределение с тяжелым хвостом имеет следующее свойство:

$$\lim_{x \to \infty} \mathrm{e}^{\lambda x}(1 - F(x)) = 0, \ \forall \lambda > 0.$$

Рассмотрим подкласс субэкспоненциальных распределений Sub в качестве примера распределения с тяжелым хвостом. Пусть $X_1, X_2, ..., X_n$ неотрицательные, независимые, одинаково распределенные случайные величины с функцией распределения *F*.

F называется субэкспоненциальным распределением $F \in \mathrm{Sub}$, если выполняется равенство $\lim\limits_{x\to\infty} \dfrac{P(X_1+...X_n>x)}{P(X_1<x)} = \dfrac{1-F^{(n)}(x)}{1-F(x)} = n,\ \forall n$, где $F^{(n)}(x)$ – n-мерная свертка функции распределения F.

К субэкспоненциальным распределениям относятся логнормальное распределение, распределения Вейбулла и Парето.

Если

$$f_X(x) = \begin{cases} \dfrac{1}{x\sigma\sqrt{2\pi}} \mathrm{e}^{-(\ln x-\mu)^2/2\sigma^2}, & x>0; \\ 0, & x\le 0, \end{cases}$$

где $\sigma>0,\ \mu\in\mathbb{R}$, то говорят, что X имеет логнормальное распределение с параметрами μ и σ, или $X \sim \ln n(\mu,\sigma^2)$.

Функция распределения F с хвостом $1-F(x)=\mathrm{e}^{-x\beta}$, где $\beta>0$, называется распределением Вейбулла.

Функция распределения F с хвостом $1-F(x)=\left(\dfrac{x_0}{x}\right)^{\alpha}$, где $\alpha>0,\ x_0>0,\ x\ge x_0,$ называется функцией распределения Парето.

При построении модели распределения с «тяжелым хвостом» необходимо воспользоваться надежным методом оценки параметра, определяющего степень тяжести хвоста. Одной из таких оценок является оценка Хилла [2, 11].

Отметим, что под распределением с тяжелым хвостом понимается распределение F, удовлетворяющее следующему условию: $\overline{F}(x) \sim x^{-\alpha}L(x),$ $x\to\infty$, $\alpha>0$, где L – медленно меняющаяся функция, такая что $\lim\limits_{t\to\infty} \dfrac{L(tx)}{L(x)} = 1,\ \forall x>0$.

Пусть $(X_n)_{n\in\mathbb{N}}$ – стационарная последовательность с одномерным распределением F, таким что $P\{X_n>x\}=1-F(x)$. Основной задачей является оценка необходимого параметра α, которую можно выполнить при помощи названной оценки Хилла следующим образом.

Пусть $X_1,...,X_n$ – независимые одинаково распределенные случайные величины с функцией распределения F. Пусть $X_{(1)}>X_{(2)}>...>X_{(n)}$ – порядковые статистики. Предположим, что F имеет тяжелый хвост:

$$1-F(x)=x^{-\alpha}L(x),\ x\to\infty.$$

Функция $\mathsf{H}_{k,n} = \dfrac{1}{k}\sum\limits_{i=1}^{k} \ln\dfrac{X_{(i)}}{X_{(k+1)}},\ k<n$, называется оценкой Хилла.

В данном случае используется только k порядковых статистик, так как лучше производить выборку из той части распределения, которая приблизительно выглядит как Парето. Объясним подробнее: набор $\dfrac{X_{(1)}}{X_{(k+1)}},...,\dfrac{X_{(i)}}{X_{(k+1)}}$

распределен как порядковые статистики из выборки размера k, и хвост распределения имеет вид $\frac{1-F(xX_{(k+1)})}{1-F(X_{(k+1)})}$, $x \geq 1$. Если $X_{(k+1)}$ большое, то $\frac{1-F(xX_{(k+1)})}{1-F(X_{(k+1)})} \sim x^{-a}$.

$$\text{Если } n \to \infty, k \to \infty, \text{ а } \frac{k}{n} \to 0, \text{ то } \mathsf{H}_{k,n} \to \alpha^{-1}. \quad (2)$$

Из условия, накладываемого на номера порядковых статистик, не ясно как получить результат (2). В данном случае можно выбрать оптимальное k, минимизирующее среднеквадратическую ошибку.

В практических задачах обычно строят оценку Хилла следующим образом: по оси x откладывается k, по оси y – $\mathsf{H}_{k,n}^{-1}$, $1 \leq k \leq n$, откуда определяют искомый параметр из устойчивой части графика. Иногда это затруднено тем, что выборка может быть не достаточно большого размера, и график слишком короткий. Фактически, традиционная оценка Хилла более эффективна, если выборка имеет распределение Парето или близкое к Парето.

Для распределения Парето, $1-F(x)=\left(\frac{x}{\sigma}\right)^{-\alpha}$, $x > \sigma$, $\sigma > 0$, ожидается, что график Хилла в правой части будет близок к γ, где $\gamma = \alpha^{-1}$, начиная с оценки $\mathsf{H}_{n-1,n}^{-1}$. Это следует из практики.

Если оценка Хилла недостаточно эффективна, то возможно построение альтернативной оценки Хилла. В этом случае, вместо построения $\left\{\left(k, \mathsf{H}_{n-1,n}^{-1}\right), 1 \leq k \leq n-1\right\}$, строится искомая альтернативная оценка $\left\{\left(k, \mathsf{H}_{n-1,n}^{-1}\right), 1 \leq k \leq 1\right\}$, где для k используется логарифмическая шкала. Здесь наблюдается эффект растяжения левой части графика, что дает более подробную кривую в сторону меньших значений k.

Формирование и агрегирование исходных данных

В качестве исследуемых исходных данных выступают информационные потоки автоматизированной системы оперативного управления перевозками (**АСОУП**), являющейся важнейшей информационной системой на железнодорожном транспорте. В основе экспериментальной проверки свойств информационных потоков лежат параметрические методы статистической оценки выборочных характеристик, дополненные методами, предназначенными для работы с телекоммуникационным трафиком. В качестве выборочных значений случайного процесса измерялись следующие характеристики сетевого трафика исследуемой системы:

1) межпакетное время прибытия дейтаграмм на сетевой интерфейс сервера системы;

2) размер файлов, передаваемых сервером в ответ на запросы пользователей.

Значения величин временных интервалов первой измеряемой характеристики выполнялись с помощью программы TCPdump [3]. Эта программа позволяет регистрировать прохождение пакета через сетевой интерфейс с погрешностью не более 1 мкс. Измерения производились блоками по 50000 дейтаграмм. Статистические характеристики для указанной выборки следующие: минимальное время – 0,000032 с, максимальное – 0,116172 с, выборочное среднее значение – 0,002689 с, выборочная дисперсия 0,000024 с, среднеквадратическое отклонение 0,004868 с.

Рассмотрим искомую выборку на предмет самоподобия. Для этого произведем следующий агрегационный процесс. Рассмотрим выборку из межпакетного времени прибытия 50000 дейтаграмм, которая затем агрегировалась (группировалась и усреднялась по 50 значений (рис. 1, *а*), по 10 значений (рис. 1, *б*), по 5 значений (рис. 1, *в*)). На рис. 1, *г* представлена исходная выборка.

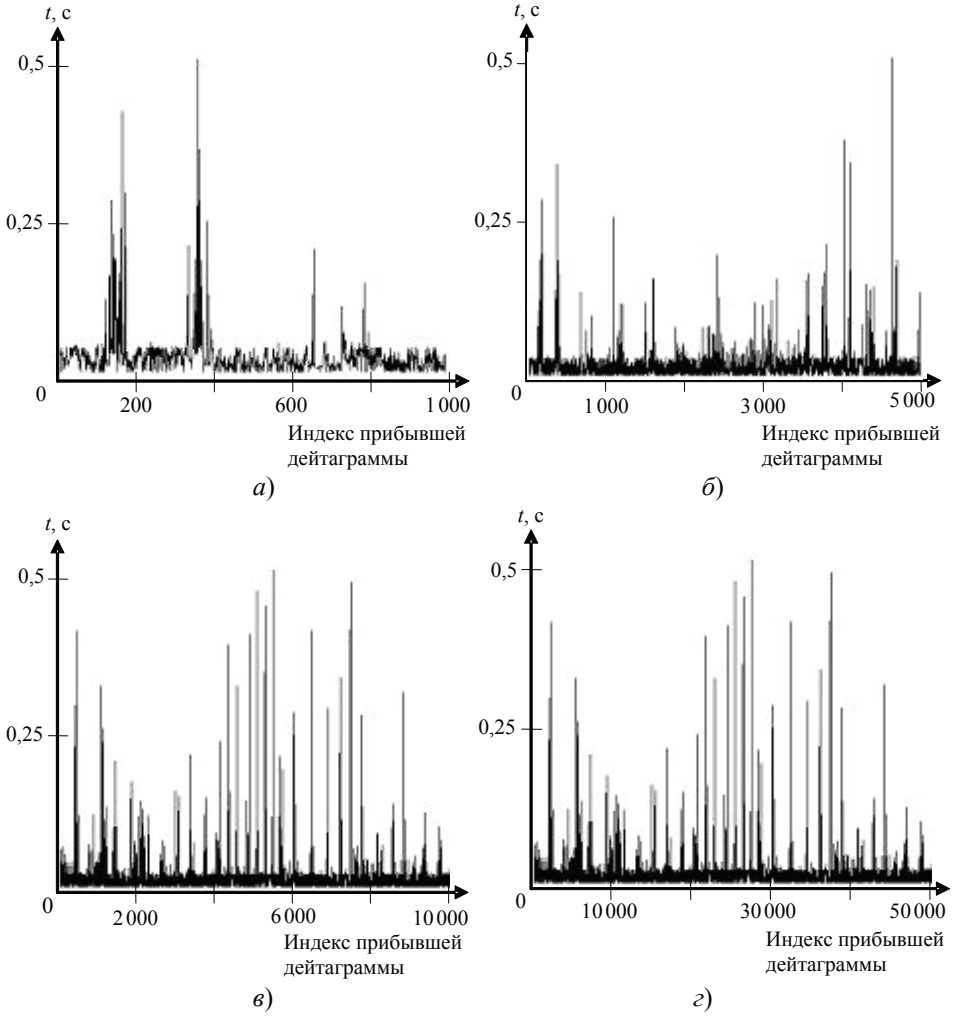


Рис. 1. Графики агрегирования выборки данных межпакетного времени прибытия дейтаграмм:
а – 1 000 дейтаграмм; *б* – 5 000 дейтаграмм;
в – 10 000 дейтаграмм; *г* – 50 000 дейтаграмм

Такие подготовительные действия предназначены для работы с выборкой в различных масштабах представления данных. Заметим, что на графиках (см. рис. 1) значения масштабы значений по оси ординат не меняются, что может являться одним из оснований для дальнейшего проверки данных на наличие свойства автомодельности. Проанализировав полученные результаты и основываясь на работах [9, 13], можно сделать предположение о том, что рассматриваемый трафик имеет самоподобную структуру. Также данный вывод можно сделать и основываясь на определении самоподобия, в котором сказано, что структура ряда, полученного усреднением групп элементов, остается такой же, как и структура исходного.

Для формирования выборок экспериментальных данных с целью исследования второй характеристики рассматривалось содержимое журналов протоколирования запросов к серверу баз данных MySQL.

Протоколирование работы сервера ведется непрерывно, поэтому возможна выборка значений за произвольные периоды времени с указанием времени предоставления ответа на запрос, размера файла и сетевого идентификатора запросившего клиента. Выполнение процесса агрегации для данных второй выборки показаны на рис. 2, *а – г* для 2000, 10000, 50000, 100000 измерений соответственно. На графиках (см. рис. 2) масштаб значений по оси ординат изменяется.

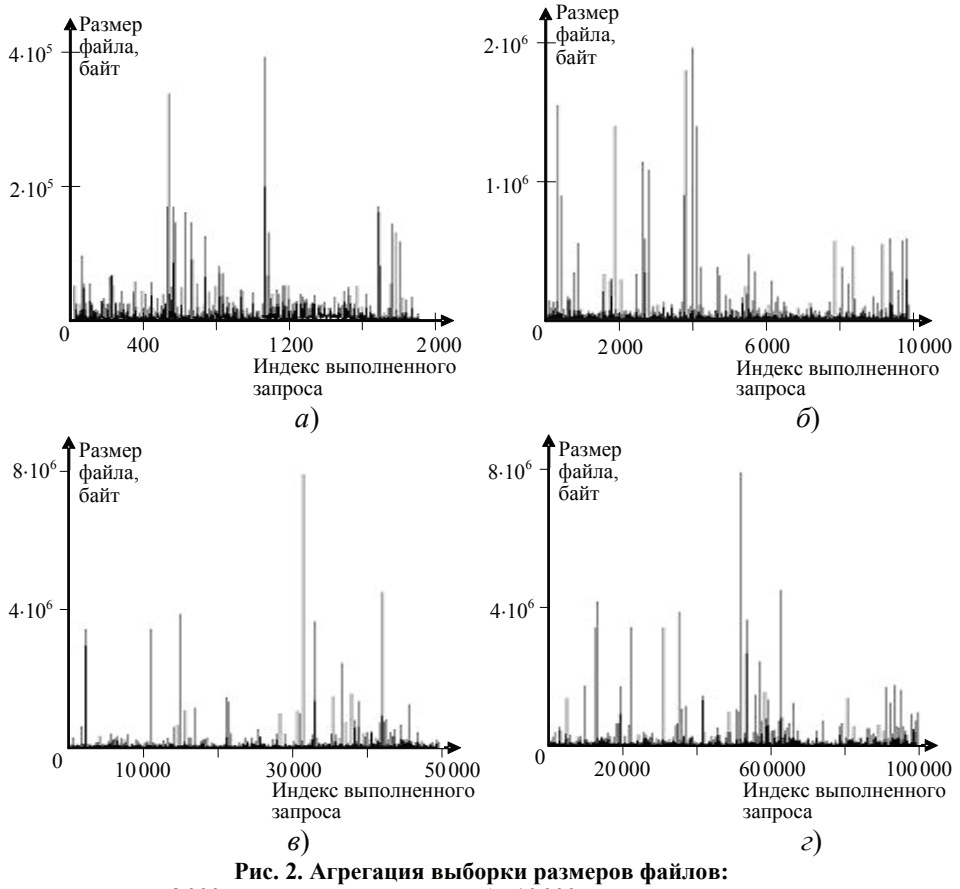
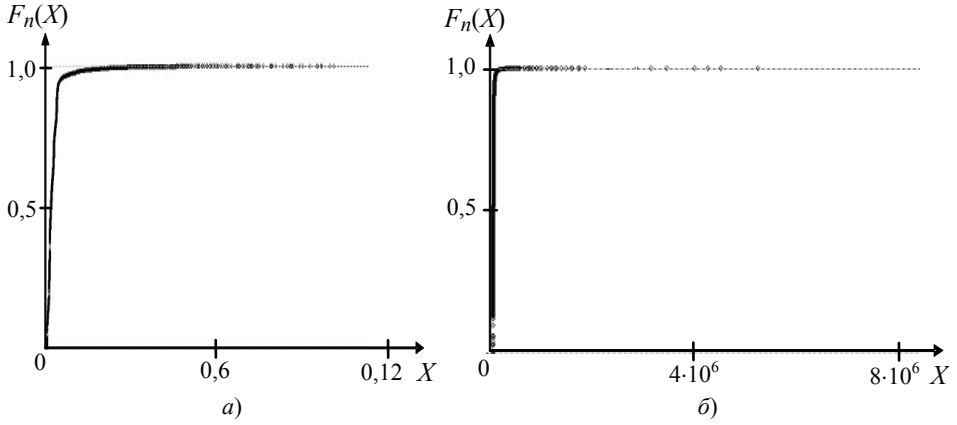


Рис. 2. Агрегация выборки размеров файлов:
а – 2 000 выполненных запросов; *б* – 10 000 выполненных запросов;
в – 50 000 выполненных запросов; *г* – 100 000 выполненных запросов

Анализ исходных выборок

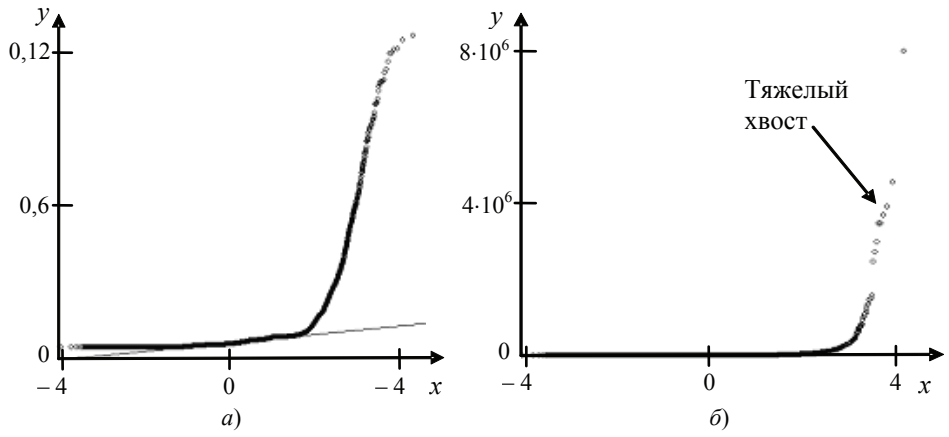
Выполним построение кумулятивной функции распределения. Упорядоченные по возрастанию данные из экспериментальной выборки $x_{(1)} \le ... \le x_{(n)}$ наносятся на ось абсцисс, и строится ступенчатая функция, возрастающая скачками величины $1/n$ в каждой точке, так что $F_n(x) = \frac{1}{n}\sum_{k=1}^{n} I(X_k < x)$. Для выборок первого и второго видов, $F_n(x)$ показаны на рис. 3, *а*, *б* соответственно. Сравнивая их с кумулятивными функциями распределений для выборок нормальных и экспоненциальных случайных величин, можно заметить существенные отличия. Виды функций на рис. 3, *а*, *б* показывают, что эмпирические распределения времени прибытия дейтаграмм и размера файлов имеют вид, отличный от нормального и экспоненциального распределений. Обратим внимание, что функция (см. рис. 3, *б*) имеет вид функции распределения с тяжелым хвостом, что отслеживается по оси абсцисс (величина масштаба представляемых значений по оси абсцисс имеет порядок 10^6, и такие значения встречаются достаточно редко, в хвосте распределения). Однако и на рис. 3, *а* хвост распределения убывает медленнее, чем экспоненциальный.

Проверим предположения о виде эмпирического распределения также с помощью квантильно-квантильных графиков (*Q*-*Q* plot). Такая методика часто используется для проверки согласия эмпирических данных семейству некоторых теоретических распределений, в том числе и распределений с тяжелыми хвостами [1]. Квантили для выборок значений обычно проверяются на соответствие нормальному распределению и строятся относительно теоретических квантилей нормального распределения. Для выборок нормальной плотности график представляет собой прямую линию, а для выборок экспоненциальной плотности – график вогнутый, имеет экспоненциально убывающий хвост.



а) *б*)

Рис. 3. Кумулятивная эмпирическая функция распределения:
а – время прибытия дейтаграмм; *б* – размер файлов



а) *б*)

Рис. 4. Квантильно-квантильные графики

У исследуемых данных, на рис. 4, *а* – по оси *y* – квантили для выборки времени поступления дейтаграмм и на рис. 4, *б* – по оси *y* – квантили для выборки размеров файлов, график еще более вогнутый, тем самым имеются хвосты распределений, убывающие медленнее, чем для экспоненциального распределения. По оси *x* отложены квантили нормального распределения.

Кроме графических методов требуется выполнить также проверки по критериям согласия изучаемых выборок на соответствие нормальному закону распределения. Наиболее известными и подходящими в нашем случае непараметрическими методами являются: метод Колмогорова–Смирнова [12], метод Шапиро–Уилка [5, 15], методы проверки на симметричность и значение эксцесса Пирсона–D'Agostino [14], Лиллиефорса [17], χ^2 Пирсона [18]. Работы, затрагивающие вопросы применения перечисленных методов к задачам проверки гипотез о распределениях типа Парето, и особенно с наличием тяжелых хвостов можно встретить не часто [8, 11].

Авторами выполнена проверка по пяти названным критериям согласия, результаты представлены в табл. 1. Значения тестовых статистик из таблицы свидетельствуют о том, что гипотезу о нормальности выборок следует отвергнуть, причем, как видно, по всем критериям согласия. Применение различных критериев согласия показывает, что исследуемые выборки не соответствует нормальному закону распределения.

Наличие тяжелого хвоста в выборке данных может быть выявлено при построении в логарифмическом масштабе графика функции в области хвоста распределения, то есть $\overline{F}_n(x) = 1 - F_n(x) = P[X > x]$. При больших значениях числа измерений в исследуемой выборке такой график будет представлять собой прямую линию, такую, что $\lim\limits_{x\to\infty} \dfrac{d \ln \overline{F}_n(x)}{d \ln x} = -\alpha$. По тангенсу угла наклона этого графика к оси абсцисс можно определить

Таблица 1

Результаты проверки выборочного распределения по критериям согласия с нормальным распределением

Наименование теста	Наименование статистики	Значение	
		Выборка 1 (объем)	Выборка 2 (объем)
Колмогорова–Смирнова для одной выборки	D	0,5 (5000)	1 (5000)
Шапиро–Уилка	W	0,4671 (1000)	0,3045 (2000)
Пирсона–D'Agostino	χ^2	15774,733 (1000)	24361,3128 (10000)
	S^3 (асимметрия)	107,04 (1000)	138,2978 (10000)
	S^4 (эксцесс)	64,4765 (1000)	72,3535 (10000)
Лиллиефорса	D	0,2948 (5000)	0,4455 (100000)
χ^2 Пирсона	P	155028,8414 (50000)	1778449, 9539 (100000)

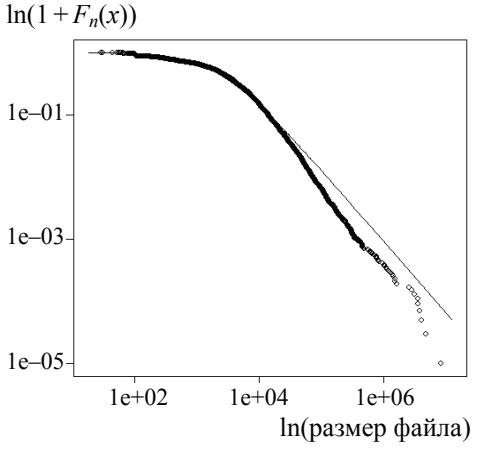


Рис. 5. График функции распределения в области хвоста

величину индекса тяжести хвоста, что, например, для исследуемой выборки размеров файлов показано на рис. 5. На приведенном графике угол наклона составляет 49,84°, таким образом, для исследуемой выборки $\alpha \approx 0{,}87$.

Анализ трафика при наличии свойств автомодельности и сильного последействия

Периодическая зависимость вариационного ряда в автомодельных процессах может быть формально определена как корреляционная зависимость некоторого k-го порядка между каждым i-м элементом ряда и $(i - k)$-м элементом. В таком случае величину k называют лагом (сдвигом, запаздыванием) и, обычно, он зависит от размера выборки n, числа выборок m, $k = 10 \lg(n / m)$. Автокоррелограмма (**АКФ**) для исследуемой выборки времени поступления дейтаграмм, приведенная на рис. 6, *а*, визуально показывает периодическую корреляционную связь, которая более заметна на частной АКФ, приведенной на рис. 6, *б*.

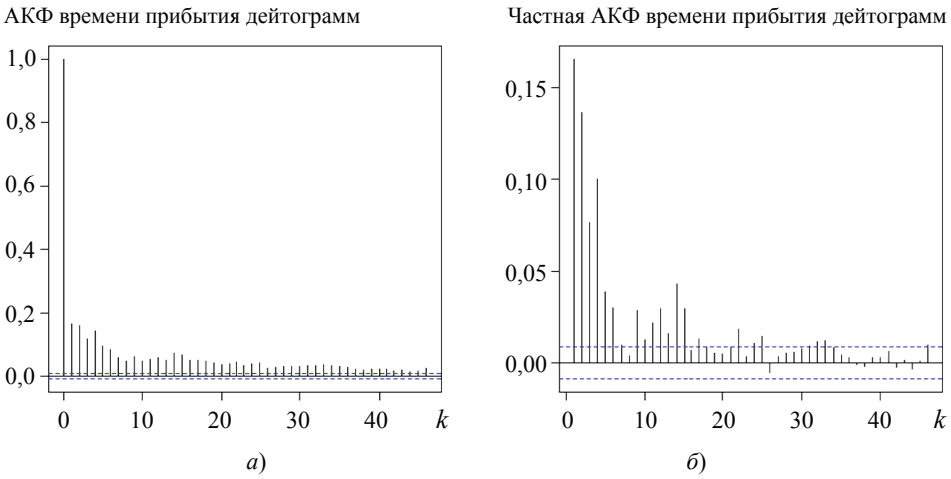


Рис. 6. Автокоррелограмма (*а*) и частная автокоррелограмма (*б*)

Далее показаны результаты проверки выборок на наличие долгой памяти потока (долговременной зависимости, сильной памяти процесса). Строится график автокорреляционной функции выборок в двойном логарифмическом масштабе, где по оси абсцисс откладывается величина логарифма лага $\ln k$, а по оси ординат откладывается величина $\ln$АКФ. О наличии долгой памяти можно судить по углу наклона линии регрессии относительно оси абсцисс.

Для выборки из межпакетного времени прибытия дейтаграмм угол наклона положительный, следовательно, корреляционная зависимость (долгая память потока) присутствует (рис. 7, *а*). Для выборки из размеров файлов свойства автомодельности в данных нет (рис. 7, *б*).

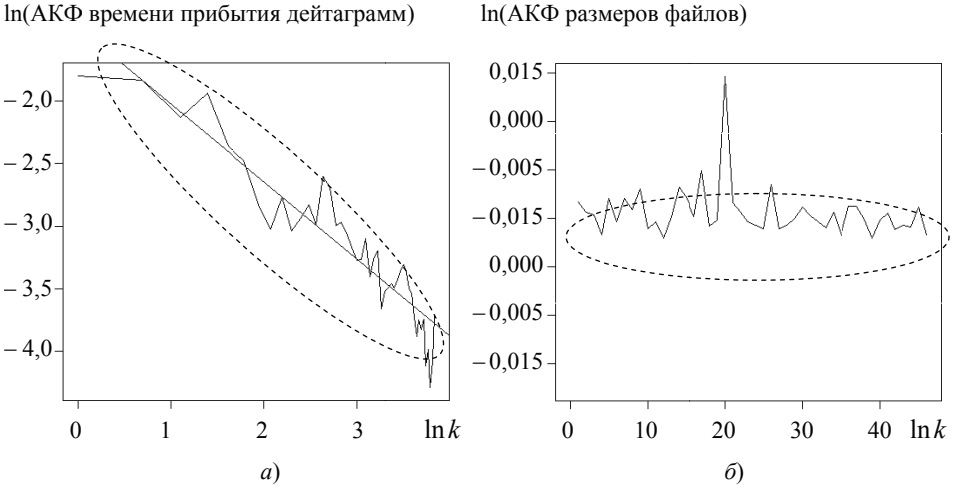


Рис. 7. Проверка выборок на наличие долговременной зависимости:
а – есть угол наклона к оси абсцисс; *б* – нет угла наклона к оси абсцисс

Перейдем к оценке показателя Харста. Для полноты анализа результатов выполнена оценка показателя Харста несколькими методами, обзор которых изложен в работе [6]. В представленных далее результатах исследовалась выборка значений времени межпакетного поступления дейтаграмм (см. рис. 1, *г*). Следует также отметить, что объединяет представленные ниже методы оценки показателя Харста, – все они основаны на нахождении углового коэффициента линии регрессии для различных вариантов агрегирования выборок данных.

Метод *R*/*S*-статистик изложен в [16], график исследуемой выборки представлен на рис. 8.

В методах агрегированных дисперсий вычисляются значения выборочной дисперсии $S^2_{(m)}$ в агрегированных блоках размера m из выборки n рассматриваемых значений: $S^2_{(m)} = \frac{1}{(n/m)-1}\sum_{k=1}^{n/m}\left(X_{(m)}(k)-\overline{X}\right)^2$, где $X_{(m)}$ – блок из m значений, взятый из выборки $X=\{X_1, X_2, ..., X_n\}$. Найденные значения откладываются в логарифмическом масштабе по оси ординат напротив соответствующих значений $\ln m$, пример графика для изучаемой выборки показан на рис. 9. Угловой коэффициент β линии регрессии связан в данном случае с показателем Харста соотношением $\beta = 2H - 2$.

Существует модификация этого метода, состоящая в том, что вместо выборочной агрегированной дисперсии используется разность ее соседних значений $\Delta S^2_{(t)} = S^2_{(m)} - S^2_{(m-1)}$, а график результатов вычислений представлен на рис. 10.

В методе периодограмм используется спектральное представление временного ряда, обычно в виде периодограммы Шустера

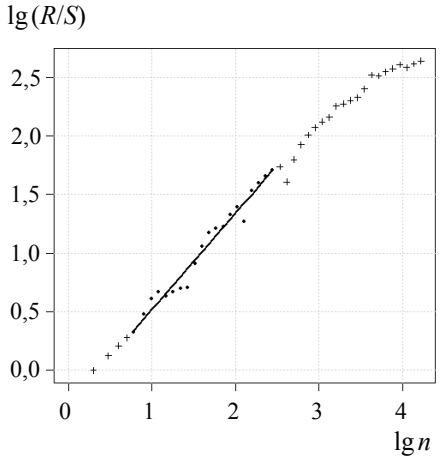


Рис. 8. Оценка показателя Харста методом *R*/*S*-статистик

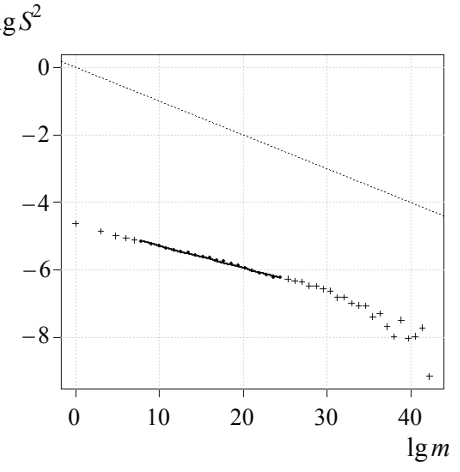


Рис. 9. Оценка показателя Харста методом агрегированной дисперсии

$$FT(\lambda) = \frac{1}{2\pi n}\left|\sum_{j=1}^{n} X_j \exp\{ij\lambda\}\right|^2,$$

где λ – система частот.

Для временных рядов, обладающих автомодельностью и сильным последействием, $FT(\lambda) \sim c_f |\lambda|^{-\beta}$, где c_f – некоторая положительная константа, $\beta \in (0,1)$. Как и прежде, параметр β и показатель Харста H связаны между собой соотношением $\mathsf{H} = (1+\beta)/2$, поэтому спектральная плотность такого временного ряда пропорциональна $|\lambda|^{1-2\mathsf{H}}$, а значит, зная угловой коэффициент линии регрессии, можно определить значение показателя Харста. Пример периодограммы для исследуемой выборки представлен на рис. 11.

Результаты расчетов показателя Харста для выборки межпакетного времени прибытия дейтаграмм, размером 50000 значений представлены в табл. 2.

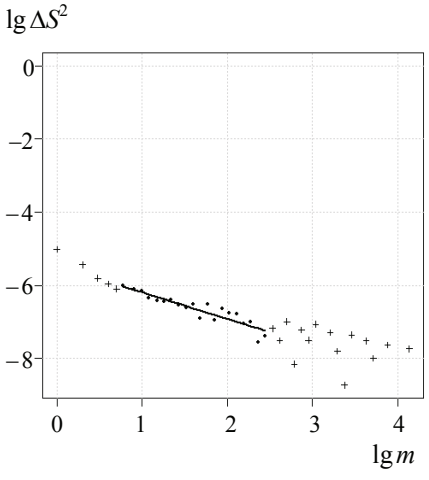


Рис. 10. Оценка показателя Харста методом разностей агрегированных дисперсий

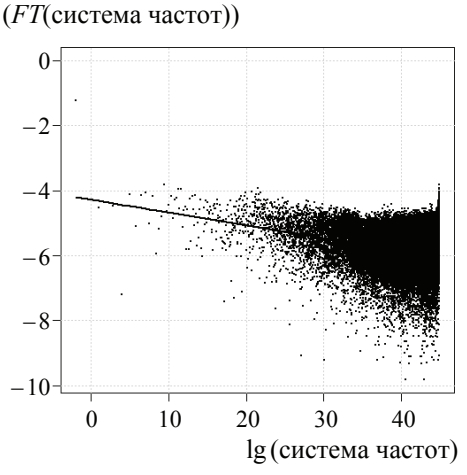


Рис. 11. Оценка показателя Харста методом периодограмм

Таблица 2

Результаты оценки показателя Харста для исследуемой выборки

Метод оценки показателя	Значение показателя
Метод *R*/*S*-статистик	0,8213
Метод агрегированных дисперсий	0,6723
Метод разностей агрегированных дисперсий	0,6362
Метод периодограмм	0,6978

Выводы

Выполнен анализ статистических данных по потокам информации в автоматизированной системе оперативного управления перевозками, и разработана комплексная методика исследования информационных потоков в автоматизированных информационно-управляющих системах. Методика позволяет выполнять экспериментальную оценку параметров потоков информации, а также проверять наличие в них свойств пиковых нагрузок, автомодельности и сильного последействия.

Работа выполнена при поддержке РФФИ (грант 09-08-00097-а).

Список литературы

1. Kratz, M.F. The Q-Q estimator and heavy tails / M.F. Kratz, S.I. Resnick // Stoch. Models. – 1996. – V. 12, No. 4. – P. 699–724.

2. Drees, H. How to make a Hill Plot / H. Drees, L. Haan, S. Resnick // Ann. Statistics. – 2000. – No. 28(1). – P. 254–274.

3. Tcpdump/libpcap. – URL : http://www.tcpdump.org/.

4. Crovella, M. Heavy Tailed-Probability distributions in the World Wide Web / M. Crovella, M. Taqqu, A. Bestavros // A Practical Guide To Heavy Tails / Robert J. Adler, Raisa E. Feldman, and Murad S. Taqqu, Eds. – New York, 1998. – Chapter 1. – P. 3–26.

5. Shapiro, S.S. An analysis of variance test for normality / S.S. Shapiro, M.B. Wilk // Biometrika. – 1965. – No. 3. – P. 591–611.

6. Taqqu, M. Estimators for long-range dependence: an empirical study / M. Taqqu, V. Teverovsky, W. Willinger // Fractals. – 1995. – No. 3(4). – P. 785–788.

7. Paxson, V. Fast, Approximate Synthesis of Fractional Gaussian Noise for Generating Self-Similar Network Traffic / V. Paxson // Computer Communications Review. – 1997. – October, V. 27, No. 5. – P. 5–18.

8. Беврани, Х. Оценка параметров распределений тяжелыми хвостами с помощью эмпирического распределения / Х. Беврани, К. Аничкин // Математика. Компьютер. Образование : сб. тр. XII междунар. конф. / под общ. ред. Г.Ю. Ризниченко. – Ижевск, 2005. – Т. 2. – С. 493–501.

9. Бондаренко, В.А. Фрактальное сжатие изображений по Барнсли-Слоану / В.А. Бондаренко, В.Л. Дольников // Автоматика и телемеханика. – 1994. – № 5. – С. 17–28.

10. Бутакова, М.А. Модели информационных потоков в системах массового обслуживания на транспорте : монография / М.А. Бутакова. – Ростов н/Д : Изд-во Рост. гос. ун-та, – 2006. – 200 с.

11. Бутакова, М.А. Особенности применения оценки Хилла для определения индекса устойчивости в распределениях с «медленно убывающим хвостом» / М.А. Бутакова // Обозрение прикладной и пром. математики. – М., 2005. – Т. 12, вып. 1. – С. 317–318.

12. Вентцель, Е.С. Марковские процессы. Потоки событий. Теория массового обслуживания / Е.С. Вентцель // Теория вероятностей / Е.С. Вентцель. – 4-е изд. – М., 1969. – Гл. 19. – С. 319–355.

13. Забарянский, С.Ф. Фрактальное сжатие изображений / С.Ф. Забарянский // Компьютеры + программы. – 1993. – № 6(39).

14. Кендалл, М. Статистические выводы и связи / М. Кендалл, А. Стьюарт. – М. : Наука, 1973.

15. Кобзарь, А.И. Прикладная математическая статистика / А.И. Кобзарь. – М. : Физматлит, 2006. – 238 с.

16. Крылов, В.В. Теория телетрафика и ее приложения / В.В. Крылов, С.С. Самохвалова. – СПб. : БХВ-Петербург, 2005. – 288 с.

17. Кулаичев, А.П. Методы и средства комплексного анализа данных. – М. : Форум–Инфра-М, 2006. – 162 с.

18. Лагутин М.Б. Наглядная математическая статистика. В 2 т. Т. 2. / М.Б. Лагутин. – М. : П-центр, 2003. – 174 с.

19. Нейман, В.И. Самоподобные процессы и их применение в теории телетрафика / В.И. Нейман // Тр. МАС. – 1999. – № 1 (9). – С. 1–15.

Models of Parameters Estimation of Telecommunication Traffic in Computer-Aided Management Information Systems

A.N. Guda, M.A. Butakova, N.A. Moskat

Rostov State University of Transport Communications,
Rostov-on-Don

Key words and phrases: Hurst metric; mathematical models of teletraffic; parameters estimation; R/S-statistics; self-similar processes.

Abstract: The paper presents the models of the analysis of the telecommunication traffic in management information systems. The models consider properties of self-similarity of the traffic, long-time dependence and a heavy tail in distribution. The numerical and graphic methods of analysis are presented.
