А. В. Кудров

Оценка функции распределения максимумов выборок стационарных последовательностей с псевдостационарным трендом

Предлагается метод оценки функции распределения максимума выборки из случайной последовательности с псевдостационарным трендом. С помощью методов стохастического моделирования проведено сравнение данного подхода с классическим (без учета тренда). Проведена обработка данных о потреблении электроэнергии в России и температуре воздуха в Центральной Англии.

Задача оценки функции распределения максимумов выборок случайных последовательностей с псевдостационарным трендом играет важную роль при определении резервов, а также при прогнозировании пиков потребления (например, электроэнергии), экстремальных погодных явлений (например, высоких температур) и др. К решению этих задач можно подходить как с позиции результатов классической теории экстремумов[1], так и с позиции результатов, являющихся расширением классической теории экстремумов, с учетом сезонности данных. В настоящей работе применяются оба подхода.

В статье представлены результаты обработки данных о ежедневных максимумах температур воздуха в Центральной Англии за период с 1 января 1878 года по 31 декабря 1998 года, взятых с сайта Британского метеорологического центра[2], и почасовом потреблении электроэнергии в России за период с 1 июля по 10 сентября 2005 года, взятых с сайта системного оператора Единой энергетической системы России[3]. В первом случае решается задача оценки функции распределения годовых максимумов, во втором случае — ежедневных максимумов.

1. Теоретические положения

Классическая теория экстремумов изучает асимптотическое распределение максимумов

$$M_n = max(\xi_1,\ldots,\xi_n)$$

n независимых одинаково распределенных случайных величин с функцией распределения $F(x)$. В основе этой теории лежит теорема Фишера–Типпета–Гнеденко (теорема об экстремальных типах, см. [De Haan, Ferreira (2006)], [Fisher, Tippet (1928)], [Gnedenko (1943)], [Leadbetter, Lingren et al. (1983)]):

[1] С некоторыми понятиями теории экстремальных значений можно ознакомиться в разделе 1 настоящей статьи (об этом подробнее см. [De Haan (2006)], [Leadbetter (1983)]).

[2] http://www.metoffice.gov.uk

[3] http://www.so-cdu.ru

Теорема 1 (Фишера–Типпета–Гнеденко). *Если для функций распределения $F(x)$ и невырожденной $H(x)$ найдутся последовательности $a_n > 0$ и b_n, $n = 1, 2, \ldots$, такие, что*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = H(x) \qquad (1)$$

в каждой точке непрерывности $H(x)$, то $H(x)$ совпадает с точностью до линейного преобразования аргумента x с положительным коэффициентом масштаба с одной из трех функций распределения:

$H_1(x) = \exp\{-e^{-x}\}$ (*распределение Гумбеля*);

$H_2(x) = \exp\{-x^{\beta}\}, x > 0$ (*распределение Фреше с параметром* $\beta < 0$);

$H_3(x) = \exp\{-(-x)^{\beta}\}, x \le 0$ (*распределение Вейбулла с параметром* $\beta > 0$).

Если выполнено (1), то функцию $F(x)$ называют *максимально устойчивой*. Пусть $H(x)$ с точностью до линейного преобразования совпадает с $H_\nu(x)$, $\nu = 1, 2, 3$, тогда говорят, что $F(x)$ *принадлежит области притяжения* D_ν, и обозначают $F(x) \in D_\nu$. Параметр β называют *экстремальным индексом*.

Назовем (a_n, b_n), $n \ge 1$, из теоремы 1 *нормировочной последовательностью*. Конкретный вид одной из возможных нормировочных последовательностей (a_n, b_n), $n \ge 1$, в теореме 1 см., например, в [Leadbetter, Lingren et al. (1983)], [Кудров (2008)].

Позднее Лидбеттером (см. [Leadbetter (1974)], [Leadbetter, Lingren et al. (1983)]) классическая теория была расширена на случай стационарных случайных последовательностей, для которых выполняются условия слабой зависимости далеко отстоящих друг от друга элементов последовательности, принимающих большие значения. При этом получаемые предельные законы для нормированных максимумов таких последовательностей те же, что и в теореме 1.

В настоящей работе рассматривается модель

$$Y_i^{(n)} = X_i + a_n m_i, \qquad (2)$$

где $\{X_i,\ i = 1, 2, \ldots\}$ — строго стационарная случайная последовательность с функцией распределения $F(x)$[4];

$\{a_n,\ n = 1, 2, \ldots\}$ — нормировочная последовательность из теоремы Фишера–Типпета–Гнеденко, соответствующая функции распределения $F(x)$;

$\{m_i,\ i = 1, 2, \ldots\}$ — тренд, ведущий себя стационарным образом в определенном ниже смысле (см. условие 5).

Пусть

$$M_n = \max\{Y_i^{(n)},\ i = 1, \ldots, n\} = \max\{(X_i + a_n m_i), i = 1, \ldots, n\},\ n = 1, 2, \ldots .$$

Последовательность (M_n) нам потребуется при формулировке теоремы 2.

[4] Напомним, что последовательность $\{X_i,\ i = 1, 2, \ldots\}$ называют *строго стационарной с функцией распределения $F(x)$*, если для любого набора индексов $i_1, \ldots, i_k$ и любого натурального τ совместная функция распределения случайных величин $X_{i_1}, \ldots, X_{i_k}$ совпадает с совместной функцией распределения случайных величин $X_{i_1+\tau}, \ldots, X_{i_k+\tau}$ и $P(X_1 \le x) = F(x)$.

Введем случайную последовательность из нулей и единиц: $\{\delta_i, i = 1, 2, \ldots\}$; событие $(\delta_i = 0)$ символизирует пропуск наблюдения X_i.

Условие 1. *Последовательности* X_i и δ_i независимы.

Условие 2. *Последовательность* m_i ограничена сверху:

$$m = \sup_{i=1,2,\ldots} m_i < \infty.$$

Обозначим $u_n^1 = a_n x + b_n,\ u_n^2 = a_n y + b_n,\ u_n = max(u_n^1, u_n^2)$.

Введем условие типа Лидбеттера на перемешивание больших значений в модели (2), аналогичное введенным в работах [Mladenović, Piterbarg (2006)], [Ольшанский (2004)].

Условие 3 ($D^2(u_n^1, u_n^2, a_n, \{m_i\}_{i=1,\ldots,n})$). *Найдется семейство чисел* $\{\alpha_{n,l}\}$, $n, l = 1, 2, \ldots$, *и последовательность натуральных чисел* $\{l_n\}$ *такие, что* $l_n = o(n)$, $\alpha_{n,l_n} \to 0$, *и для любых x, y и произвольных множеств натуральных чисел* $I = \{i_1, \ldots, i_p\}, J = \{j_1, \ldots, j_q\}$ *таких, что*

$$1 \le i_1 < \ldots < i_p < j_1 < \ldots < j_q \le n,\ \ j_1 - i_p \ge l_n,$$

выполняется неравенство

$$\sup_{\sigma} \left| P\left(\bigcap_{r \in I \cup J} \{X_j \le u_n^{\sigma(r)} - a_n m_r\} \right) - P\left(\bigcap_{r \in I} \{X_r \le u_n^{\sigma(r)} - a_n m_r\} \right) P\left(\bigcap_{r \in J} \{X_r \le u_n^{\sigma(r)} - a_n m_r\} \right) \right| \le \alpha_{n,l_n},$$

где супремум берется по всем отображениям σ *из множества натуральных чисел в* $\{1, 2\}$.

Условие 3 означает *перемешивание* (*слабую зависимость*) *далеко отстоящих больших значений временного ряда* (2). В случае если $m_i = 0,\ \ i = 1, 2, \ldots$, это условие совпадает с условием перемешивания Лидбеттера (см. [Leadbetter (1974)], [Leadbetter, Lingren et al. (1983)]).

Условие 4 ($D'(u_n - ma_n)$). *Выполняется равенство*

$$\lim_{k \to \infty} \left\{ \limsup_{n \to \infty} n \sum_{2 \le j \le n/k} P\{X_1 > u_n - a_n m;\ \ X_j > u_n - a_n m\} \right\} = 0.$$

Условие 4 гарантирует *отсутствие кластеризации больших значений временного ряда*.

Введем эмпирические функции распределения значений тренда отдельно для пропущенных и наблюдаемых X_i:

$$G_\eta^{(n)}(x) = \frac{\#\{i: m_i \le x,\ \delta_i = \eta,\ 1 \le i \le n\}}{n},$$

$\eta = 0, 1$; знак # обозначает число элементов множества.

Пусть G — неубывающая неотрицательная непрерывная справа ограниченная функция. Обозначим через $a_+ = max(a, 0)$, где a — произвольное действительное число, и определим функции

$$L_1(z, G) = e^{-z} \int_{-\infty}^{+\infty} e^t\, dG(t);$$

$$L_2(z, G) = \int_{-\infty}^{+\infty} (z - t)_+^\beta\, dG(t), \quad \beta < 0;$$

$$L_3(z, G) = \int_{-\infty}^{+\infty} (t - z)_+^\beta\, dG(t), \quad \beta > 0.$$

Сформулируем *условие псевдостационарности* последовательности $\{m_k, k = 1, 2, \ldots\}$ относительно последовательности $\{\delta_k, k = 1, 2, \ldots\}$.

Условие 5. *Найдутся функции* $G_0(x)$ *и* $G_1(x)$ *такие, что для* $\eta = 0{,}1$ *имеет место сходимость по вероятности*

$$G_\eta^{(n)}(x) \xrightarrow{P} G_\eta(x) \text{ при } n \to \infty \tag{3}$$

во всех точках непрерывности x функции $G_\eta(x)$. Кроме того, если $F \in D_\nu$, $\nu = 1, 2, 3$, то для всех x и $\eta = 0{,}1$ существуют конечные пределы

$$\lim_{n\to\infty} E\big(L_\nu(x, G_\eta^{(n)})\big) = L_\nu(x, G_\eta) < \infty,$$

где $E\big(L_\nu(x, G_\eta^{(n)})\big)$ — математическое ожидание случайной величины $L_\nu(x, G_\eta^{(n)})$.

Пусть

$$\overline{M}_n = \max\{(X_i + a_n m_i),\ \delta_i = 1,\ i = 1, \ldots, n\}, \quad n = 1, 2, \ldots .$$

Примечание. В случае когда все значения X_i, $i = 1, 2, \ldots, n$, пропущены, по определению $\overline{M}_n = -\infty$.

Теорема 2. *Пусть в модели* (2) $F \in D_\nu$, *где* $\nu = 1, 2, 3$. *Предположим, что выполнены условия* 1–5. *Тогда:*

если $\nu = 1$ или $\nu = 3$, то для всех x, y

$$\lim_{n\to\infty} P\{M_n \le u_n^1;\ \overline{M}_n \le u_n^2\} = e^{-L_\nu(x, G_0) - L_\nu(\min(x,y), G_1)}; \tag{4}$$

если $\nu = 2$, то для всех x, $y > m$

$$\lim_{n\to\infty} P\{M_n \le u_n^1;\ \overline{M}_n \le u_n^2\} = e^{-L_2(x, G_0) - L_2(\min(x,y), G_1)}. \tag{5}$$

Доказательство теоремы приведено в работе [Кудров (2008)].

2. Моделирование

Продемонстрируем результаты теоремы 2 на примере смоделированной прореженной (с пропусками) выборки из некоторой случайной последовательности, каждый элемент которой представим в виде суммы стационарной последовательности и добавочной псевдостационарной составляющей. Кроме того, проведем сравнение данного подхода с классическим (без учета тренда).

Пусть (X_i) — последовательность независимых случайных величин, равномерно распределенных на отрезке [0; 1]. Как известно (см. [De Haan, Ferreira (2006)], [Leadbetter, Lingren et al. (1983)]), для $a_n = 1/n$, $b_n = 1$, $n \ge 1$, имеет место следующее предельное соотношение:

$$\lim_{n\to\infty} P\{\max(X_1, \ldots, X_n) \le a_n x + b_n\} = \begin{cases} e^x & \text{при } x \le 0; \\ 0, & \text{иначе,} \end{cases} \tag{6}$$

т. е. равномерное распределение на отрезке [0; 1] максимально устойчиво и принадлежит области притяжения D_3, а экстремальный индекс в этом случае равен $\beta = 1$.

Обозначим через $\boldsymbol{N}$ множество натуральных чисел, $\boldsymbol{N}_n = \{1,\dots,n\}$.

В качестве (δ_i) возьмем последовательность Бернулли, не зависящую от (X_i):

$$\delta_i = \begin{cases} 1 & \text{с вероятностью } \dfrac{9}{10}; \\ 0 & \text{с вероятностью } \dfrac{1}{10} \end{cases}$$

для каждого $i \in \boldsymbol{N}$.

Возьмем $m_k = \sin\left(\dfrac{2i\pi}{3}\right)$ для $k = 1, 2, \dots$. Последовательность (m_k) псевдостационарна с функциями

$$G_0(x) = \begin{cases} 0, & \text{если } x < -\dfrac{\sqrt{3}}{2}; \\ \dfrac{1}{30}, & \text{если } -\dfrac{\sqrt{3}}{2} \le x < 0; \\ \dfrac{2}{30}, & \text{если } 0 \le x < \dfrac{\sqrt{3}}{2}; \\ \dfrac{1}{10}, & \text{иначе} \end{cases} \quad \text{и} \quad G_1(x) = \begin{cases} 0, & \text{если } x < -\dfrac{\sqrt{3}}{2}; \\ \dfrac{9}{30}, & \text{если } -\dfrac{\sqrt{3}}{2} \le x < 0; \\ \dfrac{18}{30}, & \text{если } 0 \le x < \dfrac{\sqrt{3}}{2}; \\ \dfrac{9}{10}, & \text{иначе.} \end{cases}$$

Заметив, что для случайной последовательности выполнены условия теоремы 2, получим

$$\lim_{n\to\infty} P\{\overline{M}_n \le a_n x + b_n\} = \exp\left\{-\frac{9}{30}\left[\left(-\frac{\sqrt{3}}{2} - x\right)_+ + (-x)_+ + \left(\frac{\sqrt{3}}{2} - x\right)_+\right]\right\} = K_1(x) \tag{7}$$

и

$$\lim_{n\to\infty} P\{\max\{X_i, \delta_i = 1, i \in \boldsymbol{N}_n\} \le a_n x + b_n\} = \exp\left\{-\frac{9(-x)_+}{10}\right\} = K_2(x) \tag{8}$$

для любого действительного x, где $(-x)_+ = \max(-x, 0)$.

Функция $K_1(x)$ строго монотонна, непрерывна и принимает все значения из (0; 1]. Значит, для любого $p \in (0; 1]$ решение уравнения $K_1(x) = p$ существует и единственно. Обозначим это решение через $r(p)$.

Смоделируем выборки последовательностей (X_i) и (δ_i) объемом по 250 000 элементов каждая:

$$\hat{X}_1,\dots,\hat{X}_{250\,000};$$
$$\hat{\delta}_1,\dots,\hat{\delta}_{250\,000}.$$

Рассмотрим последовательность выборочных максимумов

$$\hat{M}_1 = \max\left\{\left(\hat{X}_i + a_{500}\sin\frac{2i\pi}{3}\right),\ \ \hat{\delta}_i = 1,\ \ i = 1,\dots,500\right\},$$

$$\dots$$

$$\hat{M}_{500} = \max\left\{\left(\hat{X}_{499\times500+i} + a_{500}\sin\frac{2i\pi}{3}\right),\ \ \hat{\delta}_{499\times500+i} = 1,\ \ i = 1,\dots,500\right\}.$$

Как и в предыдущем разделе, предполагаем, что максимум по пустому множеству равен $-\infty$.

Возьмем порядковую статистику для $\hat{M}_1, \ldots, \hat{M}_{500}$:

$$\hat{M}_{1,500} \le \ldots \le \hat{M}_{500,500}.$$

Определим *QQ*-график квантилей функции распределения $K_1(x)$ против квантилей эмпирической функции распределения нормированных максимумов $(\hat{M}_i)$ как:

$$A(\hat{X}) = \left\{ \left(r\left(\frac{i}{501}\right);\ \frac{\hat{M}_{i,500} - b_{500}}{a_{500}} \right),\ i = 1, \ldots, 500 \right\}. \quad (9)$$

Аналогично *QQ*-график квантилей функции распределения $K_2(x)$ против квантилей эмпирической функции распределения нормированных максимумов $(\hat{M}_i)$ примет вид

$$B(\hat{X}) = \left\{ \left(\frac{10}{9}\log\frac{i}{501};\ \frac{\hat{M}_{i,500} - b_{500}}{a_{500}} \right),\ i = 1, \ldots, 500 \right\}. \quad (10)$$

На рис. 1 по оси *X* откладываются квантили эмпирической функции распределения нормированных максимумов, по оси *Y* — квантили предельной функции распределения

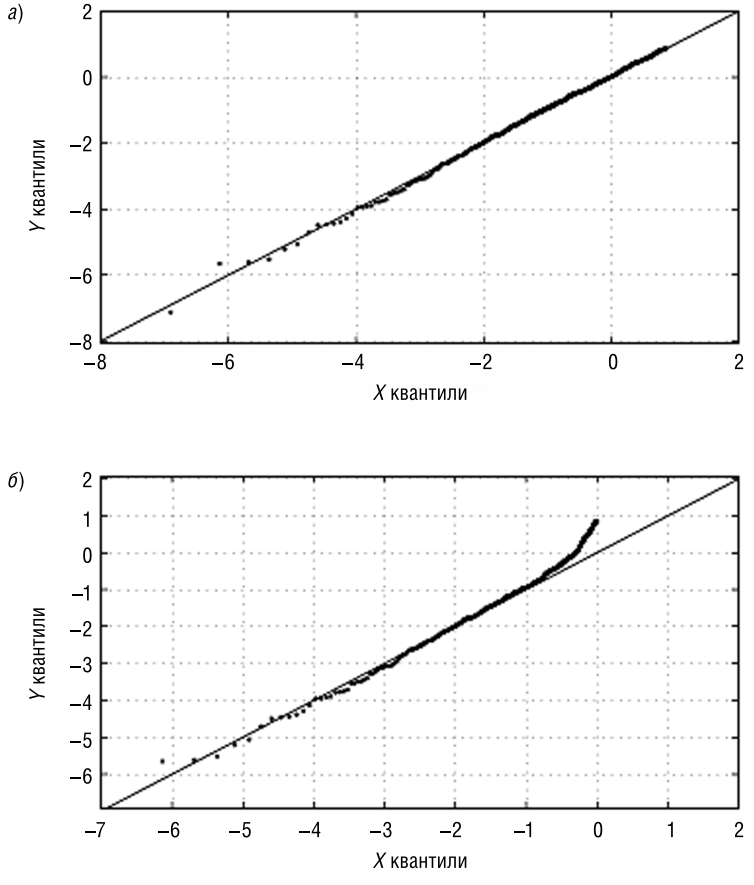

Рис. 1. *QQ*-графики $A(\hat{X})$ (*а*) и $B(\hat{X})$ (*б*)

для нормированных максимумов с учетом прореживания и детерминированного тренда $K_1(x)$ на рис. 1, *а* и с учетом прореживания и без учета детерминированного тренда $K_2(x)$ на рис. 1, *б*.

Интуитивно ясно, что приближение будет тем лучше, чем ближе будет расположен его *QQ*-график к прямой с коэффициентом наклона 1, проходящей через начало координат. Хотя в рассмотренном примере преимущество одного приближения над другим заметно визуально, измерим это преимущество численно, используя следующую меру точности:

$$Measure(t,c)=\sum_{i=c}^{500}\left(\frac{\hat{M}_{i,500}-b_{500}}{a_{500}}-K_t^{\leftarrow}\left(\frac{i}{501}\right)\right)^2, \qquad (11)$$

где $t=1,2$; c — пороговый индекс (будем брать $c=1,\ldots,470$), а $K_t^{\leftarrow}(\alpha)=\inf\{x: K_t(x)\geq\alpha\}$, $\alpha\in[0;1]$.

Здесь $Measure(1,c)$ — сумма квадратов отклонений квантилей эмпирической функции распределения нормированных максимумов от квантилей соответствующей предельной функции распределения нормированных максимумов, учитывающей прореживание и детерминированный тренд; $Measure(2,c)$ — сумма квадратов отклонений квантилей эмпирической функции распределения нормированных максимумов от квантилей соответствующей предельной функции распределения нормированных максимумов, учитывающей прореживание и не учитывающей детерминированный тренд.

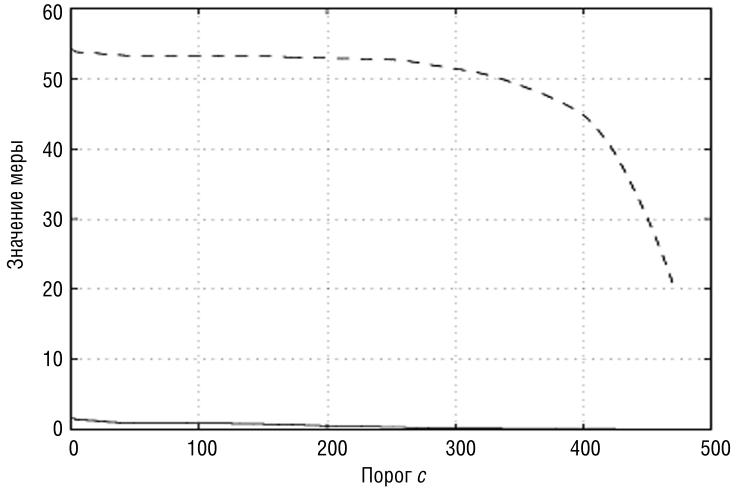

Рис. 2. Графики *Measure*(1, *c*) (сплошная линия) и *Measure*(2, *c*) (прерывистая линия)

Графики $Measure(1,c)$, $Measure(2,c)$ в зависимости от порогового индекса c представлены на рис. 2.

3. Обработка реальных данных

На конкретном примере оценим параметры предельных функций распределения максимумов из теоремы 1 и теоремы 2. Затем сравним их с эмпирической функцией распределения максимумов. Рассмотрим следующие реальные данные:

• ежедневные максимумы температур воздуха в Центральной Англии, взятые за период с 1 января 1878 года по 31 декабря 1998 года. Для этих данных оценим функцию распределения ежегодных максимумов;

• почасовое потребление электроэнергии в России за период с 7 июня по 22 июля 2005 года. Для этих данных оценим функцию распределения суточных максимумов.

3.1. Температура в Центральной Англии

Исследуем данные, представляющие собой выборку, состоящую из ежедневных максимумов температур воздуха в Центральной Англии, взятых за период с 1 января 1878 года по 31 декабря 1998 года. Опишем для них **две процедуры построения функции распределения годовых максимумов температур**. Первая основана на результатах теоремы 2, вторая — на результатах классической теории экстремумов (см. теорему Фишера–Типпета–Гнеденко). Далее сравним каждую из полученных функций распределения с эмпирической функцией распределения.

Для удобства анализа будем рассматривать выборку ежедневных максимумов температур (в градусах по Цельсию), умноженных на 10, и каждый элемент этой выборки будем называть *температурой* за соответствующий день.

Предположим, что элементы выборки температур $(\hat{T}_i)$ — это реализация значений случайного ряда (T_i), представимого в виде суммы некоторой детерминированной периодической составляющей (p_i) и стационарного временного ряда (X_i), в котором каждое X_i имеет нулевое среднее (в противном случае это среднее можно вычесть из стационарного ряда и добавить к детерминированной составляющей):

$$T_i = X_i + p_i.$$

Далее, положим, что детерминированная периодическая составляющая имеет период, равный 365, что соответствует представлениям о годовом температурном цикле.

Оценим (p_i). В силу периодичности достаточно оценить элементы $p_1, \ldots, p_{365}$. Будем использовать следующую оценку:

$$\hat{p}_i = \frac{\hat{T}_i + \hat{T}_{i+365} + \ldots + \hat{T}_{i+365(K-1)}}{K}, \tag{12}$$

где $1 \le i \le 365$, а K — количество лет, охваченных выборкой (в нашем случае $K = 120$). Таким образом, в качестве оценки для $\hat{p}_i$, $1 \le i \le 365$, берем обычное эмпирическое среднее температур i-го дня в каждый из последующих K лет. Заметим, что в условиях модели оценка $\hat{p}_i$ является несмещенной, т. е. $E\hat{p}_i = p_i$. На рис. 3 представлены значения $\hat{p}_i$, $1 \le i \le 365$.

Положим

$$\hat{X}_i = \hat{T}_i - \hat{p}_i, \ 1 \le i \le 365K.$$

Изменение температур в течение года подчинено сезонным закономерностям, так что можно выделить месяцы с самой высокой или самой низкой температурой. Поскольку нас интересуют годовые максимумы температур, выделим период года, в котором достигаются максимальные температуры.

Заметим, что если взять индекс i_0 максимального элемента последовательности $\hat{p}_1, \ldots, \hat{p}_{365}$ и рассмотреть отрезок времени $[i_0 - 50; i_0 + 50]$, то обнаруживается, что все годовые максимумы температур выборки за K лет попадают в этот промежуток, т. е. данные, входящие в этот

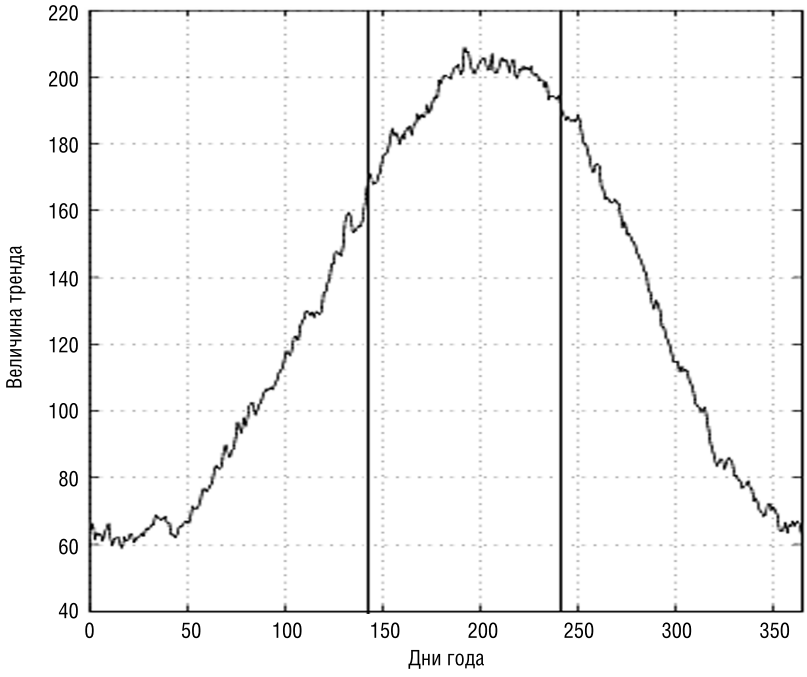

Рис. 3. График оценки периодического тренда для ежедневных максимумов температур воздуха

промежуток, определяют значения годовых максимумов температур. Поэтому будем рассматривать только данные из таких промежутков (сезонов) в каждый из годов, охваченных нашей выборкой. Заметим, что в таких промежутках тренд достаточно мал и можно применить теорему 2. На рис. 3 вертикальными линиями выделен отрезок, соответствующий этому интервалу времени.

Рассмотрим временные ряды $(\hat{T}_i^*)$, $(\hat{X}_i^*)$, $i = 1,\ldots,101K$:

$$\hat{T}_{j+365(m-1)}, \qquad j \in [i_0 - 50; i_0 + 50],\ m = 1, \ldots, K, \tag{13}$$

$$\hat{X}_{j+365(m-1)}, \qquad j \in [i_0 - 50; i_0 + 50],\ m = 1, \ldots, K, \tag{14}$$

и временной ряд $(\hat{p}_i^*)$, $i = 1,\ldots,101$:

$$\hat{p}_j, \qquad j \in [i_0 - 50; i_0 + 50]. \tag{15}$$

На каждом интервале индексов $[101(m-1)+1;\ 101m]$, где $m = 1, \ldots, K$, возьмем максимум элементов ряда $(\hat{T}_i^*)$:

$$\hat{M}_1, \ldots, \hat{M}_K$$

и ряда $(\hat{X}_i^*)$:

$$\hat{M}_1', \ldots, \hat{M}_K'.$$

Мы предполагаем, что последовательность центрированных данных $(\hat{X}_i^*)$ хорошо описывается стационарной случайной последовательностью (X_i^*), для которой выполняется условие Лидбеттера, а функция распределения с.в. X_1^* является максимально устойчивой. Тогда, применив теорему 2 (для случая, когда периодический тренд равен нулю), получим пре-

дельную функцию распределения для нормированных максимумов случайного ряда (X_i^*) — *функцию распределения экстремальных типов*, которая определяется одним параметром — *экстремальным индексом* β. Таким образом, для того чтобы оценить предельную функцию распределения, необходимо оценить экстремальный индекс функции распределения экстремальных типов.

Будем использовать оценку Пикандса для экстремального индекса β. Возьмем вариационный ряд для последовательности $(\hat{X}_i^*)$:

$$\hat{X}^*_{101K,101K} \le \hat{X}^*_{101K-1,101K} \le \ldots \le \hat{X}^*_{1,101K},$$

тогда **оценка Пикандса для экстремального индекса** имеет вид

$$\hat{\beta}_{i,101K} = -\left(\frac{1}{ln2}\ln\frac{\hat{X}^*_{i,101K} - \hat{X}^*_{2i,101K}}{\hat{X}^*_{2i,101K} - \hat{X}^*_{4i,101K}}\right)^{-1}. \qquad (16)$$

Кратко опишем **статистические свойства этой оценки** (см. [De Haan, Ferreira (2006)]):

- Если $i(K)$ такое, что $1 \le i(K) \le K$ и $i(K)/K \to 0$ при $K \to \infty$, то $\hat{\beta}_{i(K),101K}$ по вероятности стремится к β.
- При некоторых дополнительных условиях $\sqrt{i}\left(\frac{1}{\hat{\beta}_{i(K),101K}} - \frac{1}{\beta}\right)$ имеет асимптотическое нормальное распределение с нулевым средним и дисперсией

$$v(\beta) = \frac{\beta^{-2}\left(2^{-\frac{2}{\beta}+1}+1\right)}{4\left[\left(2^{-\frac{1}{\beta}}-1\right)\ln 2\right]^2}.$$

Для того чтобы выбрать *оптимальное значение оценки* $\hat{\beta}_{i,101K}$, прибегнем к часто используемой процедуре (см. [Embrechts, Kluppelberg et al. (1997)]):

1. Изобразим график множества

$$\{(i; -\hat{\beta}^{-1}_{i,101K}),\ i = 1, \ldots, 101K/4\}$$

(рис. 4).

2. Выберем наибольшую область (на рис. 4 она находится между двумя вертикальными линиями), где график приблизительно горизонтален, и возьмем в качестве оценки экстремального индекса β отвечающее этой области значение $\tilde{\beta} = 4{,}8054$. Для этого значения 95%-й асимптотический доверительный интервал (на основе асимптотической нормальности) равен [4,6642; 4,9529].

Обозначим эмпирическую функцию распределения величин $\hat{M}_1, \ldots, \hat{M}_K$ через $U(x)$, а величин $\hat{M}'_1, \ldots, \hat{M}'_K$ через $G(x)$.

Перейдем теперь к сравнению эмпирической функции распределения $G(x)$ и теоретической функции распределения экстремального типа с индексом $\tilde{\beta}$.

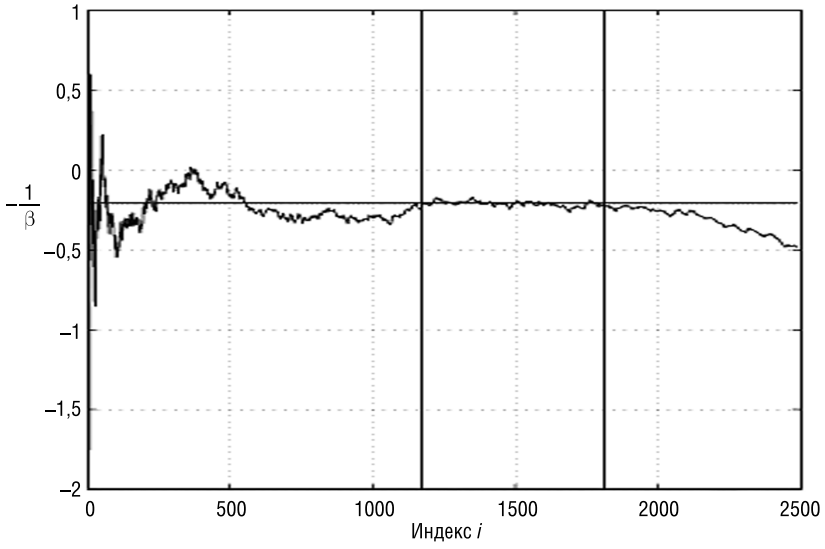

Рис. 4. График оценки Пикандса по данным за вычетом периодической составляющей

Рассмотрим *QQ*-график *A*, элементы которого составлены из пар, имеющих вид (квантиль уровня $i/(K+1)$ для эмпирической функции распределения $G(x)$; квантиль уровня $i/(K+1)$ для функции распределения экстремальных типов с экстремальным индексом $\widetilde{\beta}$):

$$A=\left\{\left(G^{\leftarrow}\left(\frac{i}{K+1}\right);-\left(-\ln\left(\frac{i}{K+1}\right)^{1/\widetilde{\beta}}\right)\right),\ i=0,\ldots,K\right\}. \tag{17}$$

На рис. 5 по оси *X* откладываются квантили эмпирической функции распределения нормированных максимумов данных за вычетом периодического тренда, а по оси *Y* — квантили стандартной функции распределения экстремальных типов, соответствующей оцененному экстремальному параметру $\widetilde{\beta}$. Как видно, точки множества *A* очень близко расположены к прямой $y=ax+b$, построенной по методу наименьших взвешенных квадратов.

Преобразуем линейно вторую координату $((y-b)/a)$ так, чтобы точки множества *A* располагались вдоль прямой $y=x$ (рис. 6).

На рис. 6 по оси *X* откладываются квантили эмпирической функции распределения нормированных максимумов данных за вычетом периодического тренда, а по оси *Y* — линейно преобразованные квантили стандартной функции распределения экстремальных типов, соответствующей оцененному экстремальному параметру $\widetilde{\beta}$.

Это линейное преобразование определяет нормировку для максимумов:

$$\frac{\hat{M}_1-b}{a},\ldots,\frac{\hat{M}_K-b}{a}. \tag{18}$$

Далее воспользуемся теоремой 2, которая утверждает, что функция распределения

$$P(x)=\exp\left\{-\frac{1}{101}\sum_{\left\{i:\ \frac{\hat{p}_i^*}{a}>x\right\}}\left(\frac{\hat{p}_i^*}{a}-x\right)^{\widetilde{\beta}}\right\} \tag{19}$$

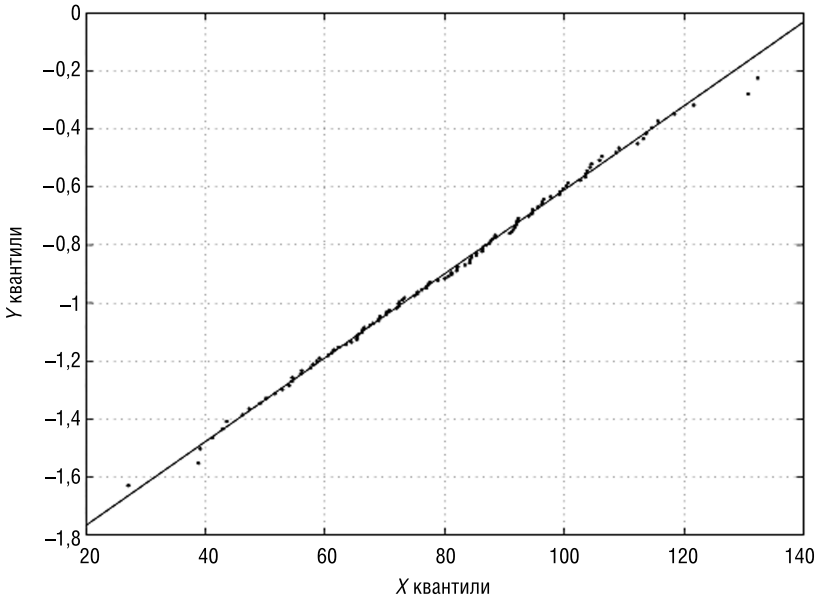

Рис. 5. *QQ*-график *A*

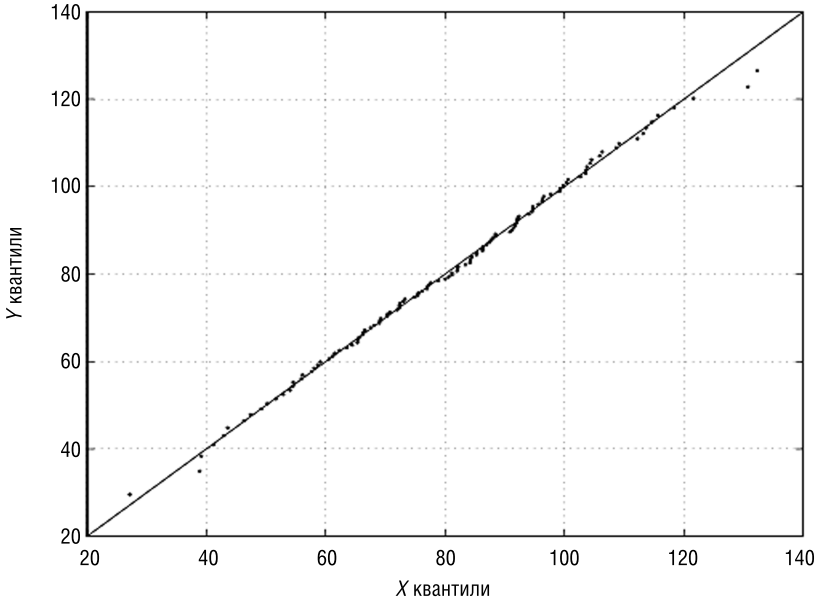

Рис. 6. *QQ*-график *A*, полученный после линейного преобразования второй координаты

должна приближать эмпирическую функцию распределения выборки нормированных максимумов (18). Для того чтобы увидеть, насколько хорошо одна функция распределения приближается другой функцией распределения, обратимся к множеству

$$B = \left\{ \left(U^{\leftarrow}\left(\frac{i}{K+1}\right); at\left(\frac{i}{K+1}\right) + b \right),\ i = 0,\ldots,K \right\}, \tag{20}$$

где $t(i/(K+1))$ — решение уравнения

$$\exp\left\{-\frac{1}{101}\sum_{\left\{i:\ \frac{\hat{p}_i^*}{a}>t\left(\frac{i}{K+1}\right)\right\}}\left(\frac{\hat{p}_i^*}{a}-t\left(\frac{i}{K+1}\right)\right)^{\tilde{\beta}}\right\}=\frac{i}{K+1}. \quad (21)$$

Заметим, что это уравнение всегда имеет решение, так как функция, стоящая слева, монотонна по $t(i/(K+1))$.

Построим на плоскости (x, y) QQ-график множества B (рис. 7).

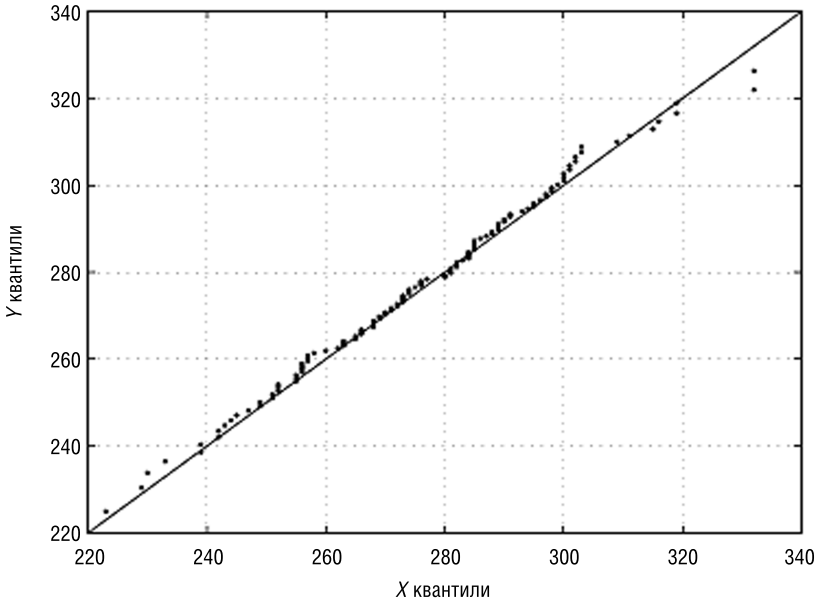

Рис. 7. *QQ*-график *B*

На рис. 7 по оси X откладываются квантили эмпирической функции распределения нормированных максимумов, а по оси Y — квантили теоретической функции распределения из теоремы 2, учитывающей периодическую составляющую.

Как видно, точки множества B расположены близко к прямой $y = x$, а это означает, что функция распределения $P(ax+b)$ достаточно точно приближает эмпирическую функцию распределения нормированных максимумов $U(x)$.

Сравним полученные результаты с результатами классической процедуры оценки функции распределения нормированных максимумов, когда эта функция приближается распределениями экстремальных типов.

Возьмем вариационный ряд последовательности $(\hat{M}_i)_{i=1}^K$:

$$\hat{M}_{K,K} \le \hat{M}_{K-1,K} \le \ldots \le \hat{M}_{1,K}.$$

Для оценки экстремального индекса применим оценку Пикандса:

$$\hat{\beta}'_{i,K} = -\left(\frac{1}{\ln 2}\ln\frac{\hat{M}_{i,K}-\hat{M}_{2i,K}}{\hat{M}_{2i,K}-\hat{M}_{4i,K}}\right)^{-1}. \quad (22)$$

Для того чтобы выбрать оптимальное значение оценки $\hat{\beta}'_{i,K}$, прибегнем к уже описанной процедуре:

1. Построим график функции $-(\hat{\beta}'_{i,K})^{-1}$, $i = 1, \ldots, K/4$:

$$\{(i; -(\hat{\beta}'_{i,K})^{-1}),\ i = 1, \ldots, K/4\}$$

(рис. 8).

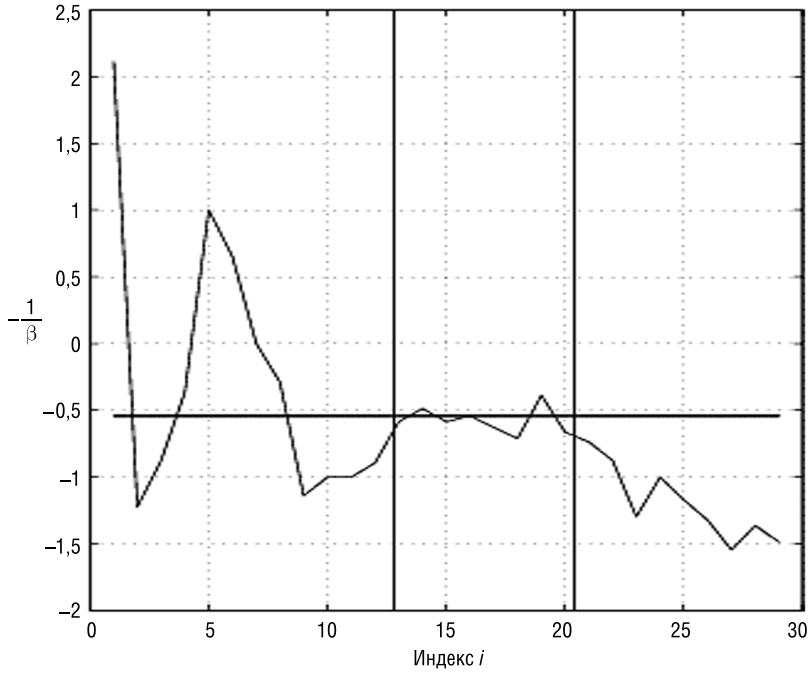

Рис. 8. График оценки Пикандса в зависимости от индекса максимальной порядковой статистики для ежегодных максимумов

2. В соответствии с ранее описанной процедурой выбора оценки Пикандса по графику статистики $-(\hat{\beta}'_{i,K})^{-1}$ выбираем значение, равное $\tilde{\beta}' = 1{,}8498$. Для этого значения 95%-й асимптотический доверительный интервал будет равен [1,4316; 2,6137].

Построим прямую $y = ax + b$ по методу наименьших взвешенных квадратов, приближающую *QQ*-график

$$C = \left\{ \left(-\left(-\ln\left(\frac{i}{K+1} \right)^{1/\tilde{\beta}} \right);\ G^{\leftarrow}\left(\frac{i}{K+1} \right) \right),\ i = 0, \ldots, K \right\}, \tag{23}$$

на плоскости (x, y). На рис. 9 в соответствии с этой линейной нормировкой изобразим график

$$D = \left\{ \left(G^{\leftarrow}\left(\frac{i}{K+1} \right);\ -a\left(-\ln\left(\frac{i}{K+1} \right)^{1/\tilde{\beta}} \right) + b \right),\ i = 0, \ldots, K \right\}. \tag{24}$$

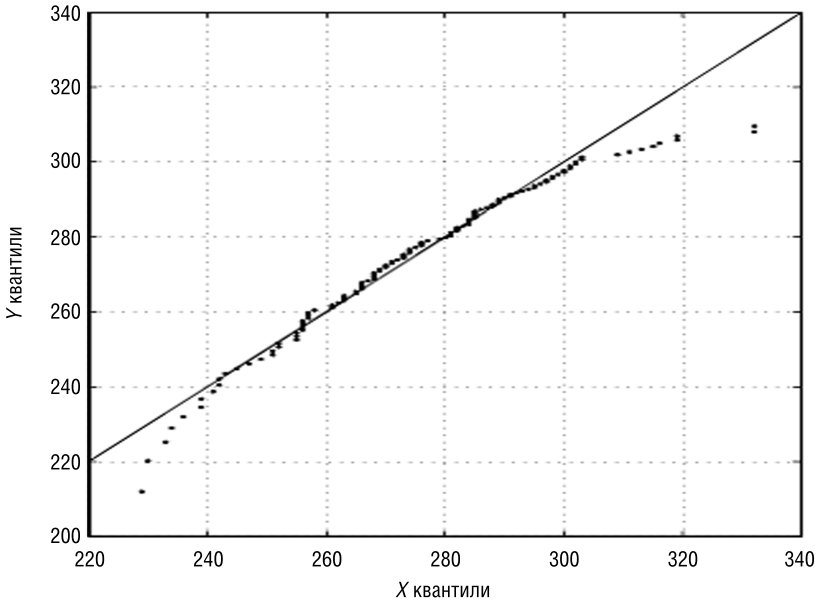

Рис. 9. *QQ*-график *D*

На рис. 9 по оси *X* откладываются квантили эмпирической функции распределения нормированных максимумов, а по оси *Y* — квантили теоретической функции распределения из теоремы 2, построенной по ежегодным максимумам.

Сравнивая графики *B* и *D* (см. рис. 7 и 9), приходим к выводу, что учет периодического тренда позволяет получить более хорошие оценки для описания эмпирической функции распределения максимумов, чем оценки, построенные на основании выборки, состоящей из ежегодных максимумов. Такие результаты можно объяснить тем, что учет нестационарности типа периодического тренда (даже такого малого, как в этом примере) позволяет оценивать параметры соответствующей функции распределения по большему числу данных по сравнению с числом данных, по которым оценивается функция распределения экстремального типа, а значит, позволяет получить более устойчивые оценки.

3.2. Потребление электроэнергии в России

Исследуем теперь данные о почасовом потреблении электроэнергии в России за период с 7 июня по 22 июля 2005 года. Визуальный анализ изменения потребления электроэнергии позволяет сделать вывод о периодичности потребления за сутки. Более того, можно увидеть, что имеется периодичность, связанная с днями недели, и годичная периодичность (однородность по сезонам). При полном исследовании экстремальных значений потребления необходимо учитывать и годичный тренд. Задача полного исследования данных, однако, в работе не ставится.

В нашем случае исследуемый промежуток взят как пример однородности по сезону. Мы будем рассматривать только данные со вторника по четверг каждой недели, так как максимумы потребления в течение недели достигаются только в эти дни. Кроме того, для этих дней наблюдается похожая структура потребления.

Обозначим через C_i потребление за i-й час рассматриваемого временного интервала. Пусть $(\hat{C}_i)$ — это реализация случайной последовательности (C_i). Предположим, что элементы этой случайной последовательности представимы в виде суммы детерминированной периодической составляющей (p_i) и стационарного временного ряда (X_i), имеющего нулевое среднее:

$$C_i = X_i + p_i.$$

Далее положим, что детерминированная периодическая составляющая имеет период, равный 24, что соответствует суткам.

Оценим (p_i):

$$\hat{p}_i = \frac{\hat{C}_i + \hat{C}_{i+24} + \ldots + \hat{C}_{i+24(K-1)}}{K}, \tag{25}$$

где $1 \le i \le 24$, а K — количество дней, охваченное выборкой (в нашем случае, $K = 28$). На рис. 10 представлены значения $\hat{p}_i$, $1 \le i \le 24$.

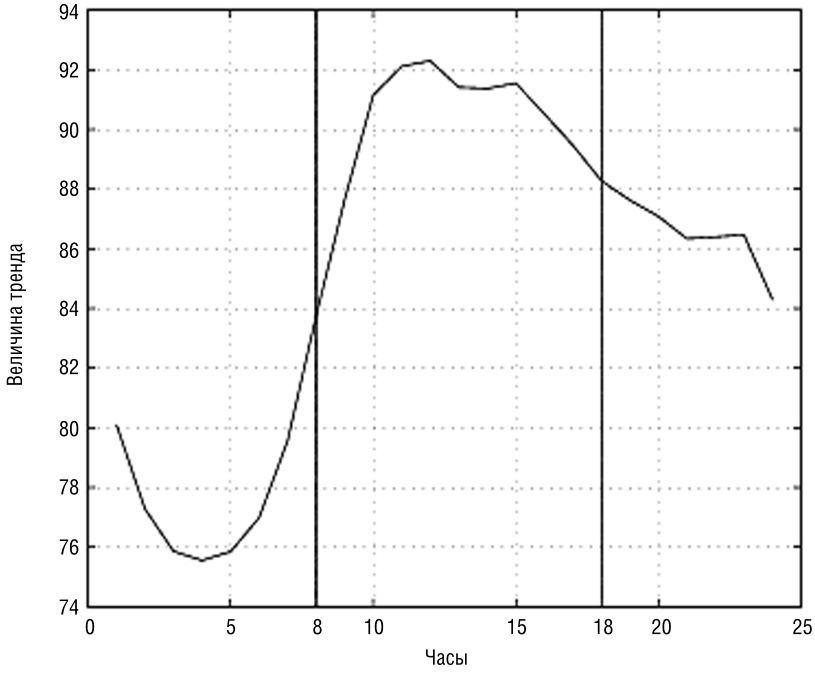

Рис. 10. График оценки периодического тренда для суточного потребления электроэнергии в России

Пусть

$$\hat{X}_i = \hat{C}_i - \hat{p}_i, \quad 1 \le i \le 24K.$$

Поскольку максимумы потребления в течение суток возникают в промежутке времени между 8:00 и 18:00, рассмотрим временные ряды $(\hat{C}_i^*)$, $(\hat{X}_i^*)$, $i = 1,\ldots,11K$, и $(\hat{p}_i^*)$, $i = 1,\ldots,11$, которые соответствуют этому промежутку, а именно:

$$\hat{C}_{j+24(m-1)}, \qquad j \in [8;18],\ m = 1,...,K, \tag{26}$$

$$\hat{X}_{j+24(m-1)}, \qquad j \in [8;18],\ m = 1,...,K, \tag{27}$$

$$\hat{p}_j, \qquad j \in [8;18]. \tag{28}$$

На рис. 10 выделенный вертикальными линиями отрезок соответствует рассматриваемому времени суток.

На каждом интервале индексов вида $[11(m-1)+1; 11m]$, где $m = 1,\ldots,K$, возьмем максимум временного ряда $(\hat{C}_i^*)$:

$$\hat{M}_1, ..., \hat{M}_K$$

и ряда $(\hat{X}_i^*)$:

$$\hat{M}_1', ..., \hat{M}_K'.$$

Предположим, что $(\hat{X}_i^*)$ представляет собой выборку из случайной стационарной последовательности (X_i^*), для которой выполняется условие Лидбеттера. Функция распределения с.в. X_1^* предполагается максимально устойчивой. Тогда, применив теорему 2 (случай, когда периодический тренд равен нулю), получим предельную функцию распределения для нормированных максимумов случайного ряда (X_i^*) — *функцию распределения экстремальных типов*. Для того чтобы оценить эту предельную теоретическую функцию распределения, необходимо оценить экстремальный индекс β функции распределения экстремальных типов. Для этого снова воспользуемся оценкой Пикандса:

$$\hat{\beta}_{i,11K} = -\left(\frac{1}{\ln 2}\ln\frac{\hat{X}^*_{i,11K} - \hat{X}^*_{2i,11K}}{\hat{X}^*_{2i,11K} - \hat{X}^*_{4i,11K}}\right)^{-1}, \tag{29}$$

где $\hat{X}^*_{11K,11K} \le \hat{X}^*_{11K-1,11K} \le \ldots \le \hat{X}^*_{1,11K}$ — вариационный ряд для последовательности $(\hat{X}_i^*)$.

Для выбора оптимального значения оценки Пикандса $\hat{\beta}_{i,11K}$ воспользуемся визуальным методом (см. раздел 3.1), для чего изобразим график множества

$$\{(i; -\hat{\beta}^{-1}_{i,11K}),\ i = 1,...,11K/4\}$$

и выберем наибольшую область, где график приблизительно горизонтален (на рис. 11 эта область находится между двумя выделенными вертикальными линиями). Таким образом, возьмем в качестве оценки экстремального индекса отвечающее этой области значение $\tilde{\beta} = 1{,}4267$. Для этого значения 95%-й асимптотический доверительный интервал будет равен [0,6339; 1,6759].

Обозначим эмпирическую функцию распределения величин $\hat{M}_1, \ldots, \hat{M}_K$ через $U(x)$, а величин $\hat{M}_1', \ldots, \hat{M}_K'$ через $G(x)$.

Перейдем теперь к сравнению эмпирической функции распределения $G(x)$ и теоретической функции распределения экстремального типа с индексом $\tilde{\beta}$. Для этого воспользуемся QQ-графиком множества

$$A = \left\{\left(G^{\leftarrow}\left(\frac{i}{K+1}\right); -\left(-\ln\left(\frac{i}{K+1}\right)^{1/\tilde{\beta}}\right)\right),\ i = 0,...,K\right\}. \tag{30}$$

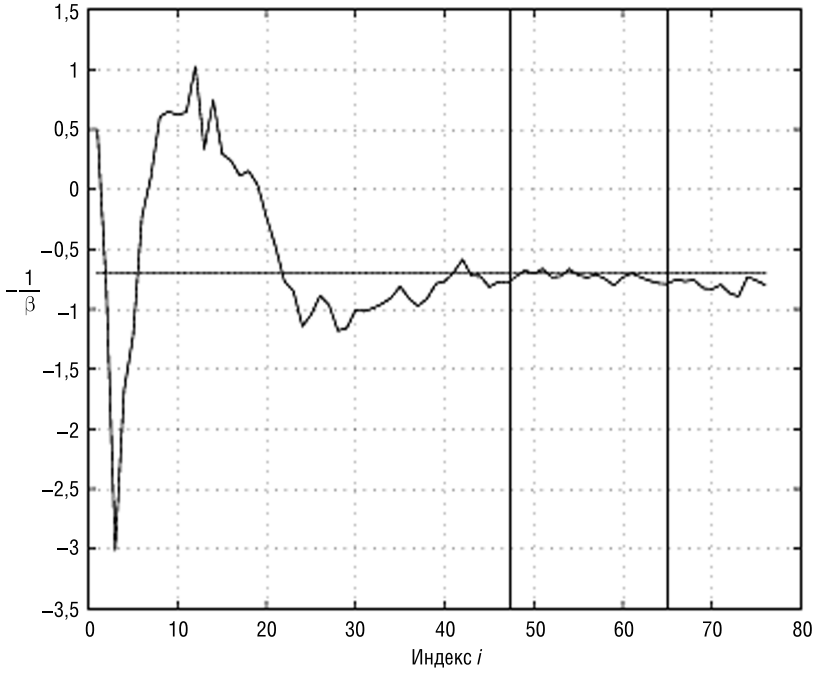

Рис. 11. График оценки Пикандса по данным за вычетом периодической составляющей

На рис. 12 по оси *X* откладываются квантили эмпирической функции распределения нормированных максимумов данных за вычетом периодического тренда, а по оси *Y* — квантили стандартной функции распределения экстремальных типов, соответствующей оцененному экстремальному параметру $\tilde{\beta}$.

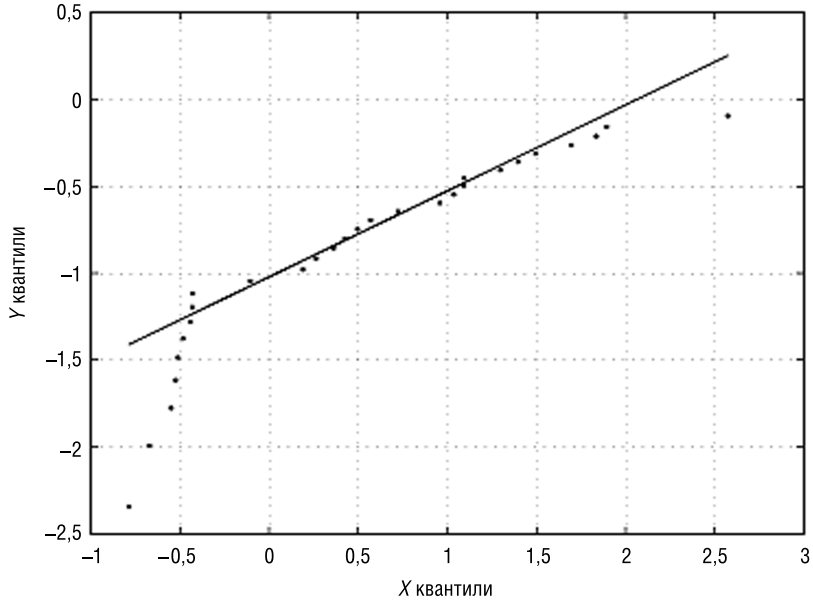

Рис. 12. *QQ*-график *A*

Как видно, элементы множества A очень близко расположены к прямой $y = ax + b$, построенной по методу наименьших взвешенных квадратов.

Преобразуем линейно вторую координату $((y - b)/a)$ так, чтобы точки множества A располагались вдоль прямой $y = x$.

Это линейное преобразование определяет нормировку для максимумов:

$$\frac{\hat{M}_1 - b}{a}, \ldots, \frac{\hat{M}_K - b}{a}. \tag{31}$$

На рис. 13 по оси X откладываются квантили эмпирической функции распределения нормированных максимумов данных за вычетом периодического тренда, а по оси Y — линейно преобразованные квантили стандартной функции распределения экстремальных типов, соответствующей оцененному экстремальному параметру $\tilde{\beta}$.

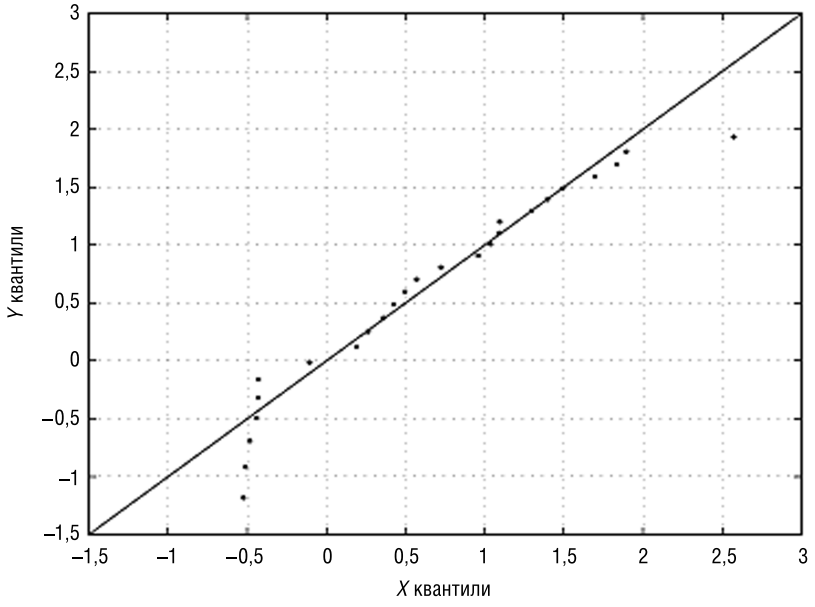

Рис. 13. *QQ*-график A, полученный после линейного преобразования второй координаты

Возьмем функцию распределения

$$P(x) = \exp\left\{-\frac{1}{11}\sum_{\left\{i;\ \frac{\hat{p}_i^*}{a} > x\right\}}\left(\frac{\hat{p}_i^*}{a} - x\right)^{\tilde{\beta}}\right\} \tag{32}$$

и посмотрим, насколько хорошо она приближает эмпирическую функцию распределения выборки нормированных максимумов (31). Для этого обратимся к множеству

$$B = \left\{\left(U^{\leftarrow}\left(\frac{i}{K+1}\right);\ at\left(\frac{i}{K+1}\right) + b\right),\ \ i = 0, \ldots, K\right\}, \tag{33}$$

где $t(i/(K+1))$ — решение уравнения

$$\exp\left\{-\frac{1}{11}\sum_{\left\{i:\ \frac{\hat{p}_i^*}{a}>t\left(\frac{i}{K+1}\right)\right\}}\left(\frac{\hat{p}_i^*}{a}-t\left(\frac{i}{K+1}\right)\right)^{\tilde{\beta}}\right\}=\frac{i}{K+1}. \quad (34)$$

Заметим, что это уравнение всегда имеет решение, так как функция, стоящая слева, монотонна по $t(i/(K+1))$.

Построим на плоскости (x, y) QQ-график B (рис. 14).

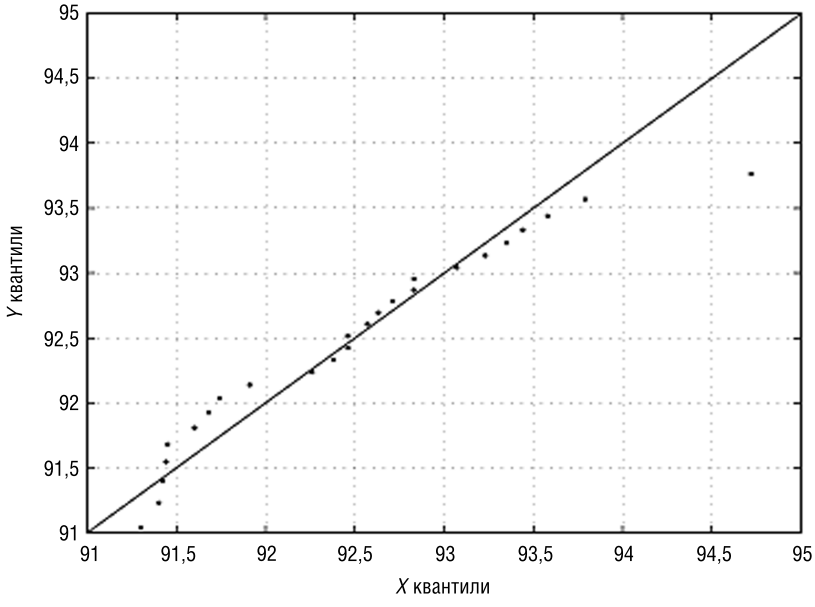

Рис. 14. QQ-график B

На рис. 14 по оси X откладываются квантили эмпирической функции распределения нормированных максимумов, а по оси Y — квантили теоретической функции распределения из теоремы 2, учитывающей периодическую составляющую.

Как видно, точки множества B расположены достаточно близко к прямой $y = x$, а это означает, что распределение $P(ax + b)$ достаточно точно приближает эмпирическую функцию распределения $U(x)$.

Поскольку периодическая составляющая в рассматриваемый интервал времени (с 8:00 до 18:00) достаточно плоская, то представляется разумным применить классическую теорию экстремумов, без учета влияния тренда, и сравнить результаты с результатами, полученными ранее.

Для оценки экстремального индекса в этом случае (периодическая составляющая не вычитается) снова воспользуемся оценкой Пикандса:

$$\hat{\beta}'_{i,11K}=-\left(\frac{1}{\ln 2}\ln\frac{\hat{C}_{i,11K}-\hat{C}_{2i,11K}}{\hat{C}_{2i,11K}-\hat{C}_{4i,11K}}\right)^{-1},\ 1\le i\le 11K/4, \quad (35)$$

где $\hat{C}_{11K,11K}\le\hat{C}_{11K-1,11K}\le\ldots\le\hat{C}_{1,11K}$ — вариационный ряд последовательности $(\hat{C}_i)$.

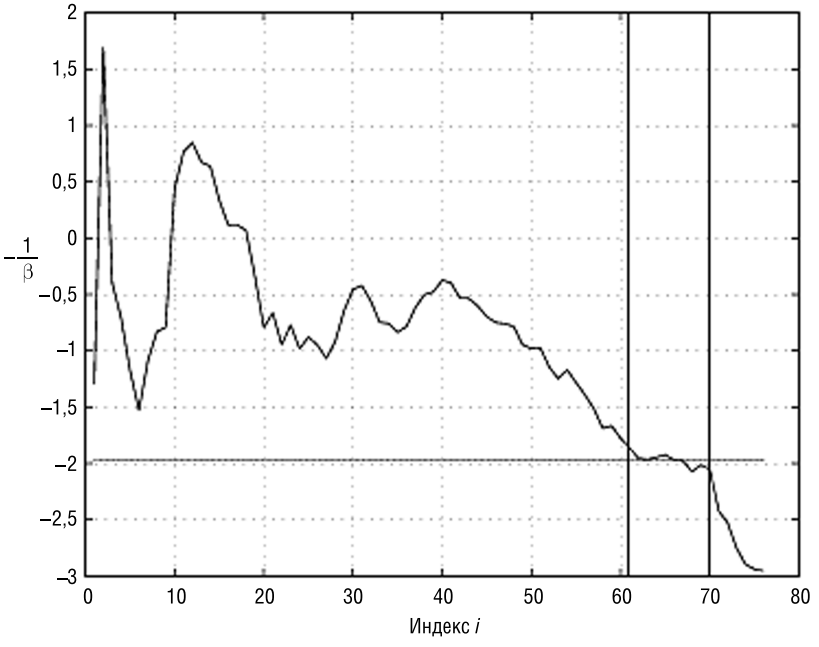

Рис. 15. График оценки Пикандса в зависимости от индекса максимальной порядковой статистики (по ежечасным данным)

На рис. 15 построен график функции $-(\hat{\beta}'_{i,11K})^{-1}$:

$$\{(i; -(\hat{\beta}'_{i,11K})^{-1}),\ \ i = 1,...,11K/4\}.$$

В соответствии с указанной процедурой визуально выбираем оценку Пикандса для экстремального индекса β: $\widetilde{\beta}' = 0{,}5079$, для которой асимптотический 95%-й доверительный интервал равен [0,4426; 1,4732].

Заметим, что эта оценка экстремального индекса $\widetilde{\beta}'$ (0,5079) по величине существенно отличается от оценки экстремального $\widetilde{\beta}$ для данных за вычетом периодической составляющей (1,4267).

Построим прямую $y = ax + b$ по методу наименьших взвешенных квадратов для *QQ*-графика

$$C = \left\{\left(-\left(-\ln\left(\frac{i}{K+1}\right)^{1/\widetilde{\beta}}\right); G^{\leftarrow}\left(\frac{i}{K+1}\right)\right),\ \ i = 0,...,K\right\}. \tag{36}$$

На рис. 16 изображено множество точек, соответствующее этой линейной нормировке:

$$D = \left\{\left(G^{\leftarrow}\left(\frac{i}{K+1}\right); -a\left(-\ln\left(\frac{i}{K+1}\right)^{1/\widetilde{\beta}}\right) + b\right),\ \ i = 0,...,K\right\}. \tag{37}$$

На рис. 16 по оси *X* откладываются квантили эмпирической функции распределения нормированных максимумов, а по оси *Y* — квантили теоретической функции распределения из теоремы 2, построенной по ежедневным максимумам.

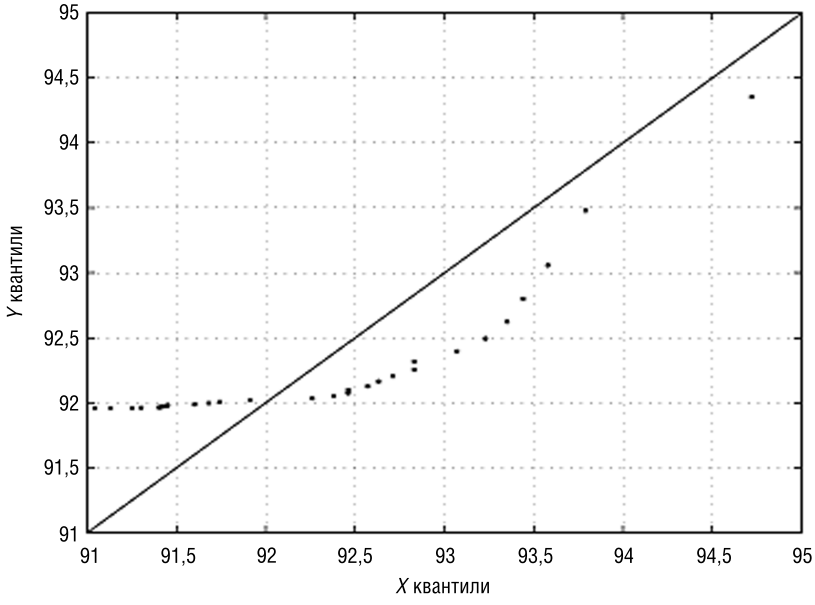

Рис. 16. *QQ*-график *D*

Сравнивая графики *B* и *D* (см. рис. 14 и 16), приходим к выводу, что и в этом случае учет периодического тренда позволяет получить более хорошие оценки для описания эмпирической функции распределения максимумов, чем оценки, построенные без учета периодичности.

4. Выводы

Решая задачу оценивания функции распределения максимумов данных с периодическим трендом, можно использовать два подхода: первый основан на результатах классической теории экстремумов, второй — на предельной теореме для нормированных максимумов выборок с псевдостационарным трендом, доказанной автором.

Существенным ограничением первого подхода является небольшое количество данных (максимумов), подлежащих обработке. Второй подход позволяет преодолеть это ограничение, так как учитывается наличие периодического тренда. Таким образом, применяя второй подход, можно рассматривать большее количество данных, а значит, получать более устойчивые оценки.

Для данных с периодическим трендом учет периодической составляющей позволяет получить более точные оценки функции распределения максимумов. В настоящей работе это показано как на примере смоделированных данных, так и на примере температур воздуха в Центральной Англии и потребления электроэнергии в России.

Список литературы

Кудров А. В. О максимумах частичных выборок случайных последовательностей с псевдостационарным трендом // Стат. методы оценивания и проверки гипотез. Пермский государственный университет, 2008.

Кузнецов Д. С. Предельные теоремы для максимума случайных величин // *Вестник МГУ*, Сер. Матем. механ. 2005. № 3. P. 6–9.

Ольшанский К. А. Об экстремальном индексе прореженного процесса авторегрессии // *Вестник МГУ*, Сер. Матем. механ. 2004. № 3. P. 17–23.

De Haan L., Ferreira A. Extreme value theory. An introduction // *Springer Series in Operations Research and Financial Engineering*. Springer, 2006.

Embrechts P., Kluppelberg C., Mikosch T. Modelling Extremal Events for Insurance and Finance. Berlin, Heidelberg: Springer-Verlag, 1997.

Fisher R. A., Tippet L. H. C. Limiting forms of the frequency distribution of the largest or smallest member of a sample // *Proc. Cambridge Phil. Soc*. 1928. № 24. P. 180–190.

Gnedenko B. V. Sur la distribution limite du terme maximum d'une série aléatoire // *Ann. Math*. 1943. № 44. P. 423–453.

Kudrov A. V., Piterbarg V. I. On maxima of partial samples in gaussian sequences with pseudo-stationary trends // *Liet. matem. rink*. 2007. № 47. No 1. P. 1–10.

Leadbetter M. R. On extreme values in stationary sequences // Z. Wahrscheinlichkeits-theorie verw. Gebiete, 1974. № 28. P. 289–303.

Leadbetter M. R., Lingren G., Rootzén H. Extreme and related properties of random sequences and precesses // *Springer Statistics Series*. Berlin-Heidelberg-New York: Springer, 1983.

Mladenović P., Piterbarg V. I. On asymptotic distribution of maxima of complete and incomplete samples from stationary sequences // *Stochastic Processes and their Applications*. 2006. № 116. P. 1977–1991.