ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ
ГЕОЛОГОРАЗВЕДОЧНЫЙ УНИВЕРСИТЕТ ИМЕНИ
СЕРГО ОРЖОНИКИДЗЕ

КАФЕДРА ВЫСШЕЙ МАТЕМАТИКИ И
МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

А.А. Любушин

АНАЛИЗ ПЕРИОДИЧЕСКИХ КОМПОНЕНТ ИНТЕНСИВНОСТИ ТОЧЕЧНЫХ ПРОЦЕССОВ.

УЧЕБНОЕ ПОСОБИЕ ДЛЯ СТАРШИХ КУРСОВ ГЕОФИЗИЧЕСКОГО ФАКУЛЬТЕТА

МОСКВА
2006

Введение. В пособии приводится основные сведения по теории точечных процессов и рассматривается один метод обнаружения периодических компонент в интенсивности потока событий. Метод реализован программно и может быть использован как в учебных, так и в исследовательских целях. Приведены примеры анализа реальных геофизических данных.

Всюду ниже будут использоваться обозначения: $\Pr\{...\}$ - вероятность события, условие которого записано в фигурных скобках, $M\{\xi\} = \int x\varphi_\xi(x)dx$ - математическое ожидание или среднее значение случайной величины ξ, где $\varphi_\xi(x)$ - плотность вероятности распределения случайной величины ξ, $D\{\xi\} = M\{(\xi - M\{\xi\})^2\}$ - дисперсия случайной величины, которая также может быть записана в виде $D\{\xi\} = M\{\xi^2\} - (M\{\xi\})^2$.

Про случайную величину ξ будем говорить, что она распределена по нормальному закону с параметрами a и σ^2 и записывать это как $\xi \sim \mathbb{N}(a, \sigma^2)$, если плотность вероятности случайной величины ξ равна $\varphi_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-(x-a)^2/2\sigma^2)$. Известно, что в этом случае $M\{\xi\} = a$, $D\{\xi\} = \sigma^2$. Вообще, знак «~» означает «распределено как».

1. *Точечным процессом* называется последовательность случайных моментов времени t_j, $j \in Z$. Ее *считающей функцией* называется величина:

$$N(t) = \sum_{t_j < t} r(t - t_j), \quad r(t) = \begin{cases} 1, t \geq 0 \\ 0, t < 0 \end{cases} \tag{1}$$

Таким образом, $N(t)$ – монотонно невозрастающая кусочно-постоянная функция, скачкообразно растущая на 1 в случайные моменты времени t_k. Если существует функция $\lambda(t) \geq 0$, такая что вероятности возникновения и отсутствия события на малом интервале времени $(t, t+\varepsilon]$ выражаются формулами:

$$\begin{aligned} &\Pr\{N(t+\varepsilon) - N(t) = 1\} = \lambda(t)\cdot\varepsilon + o(\varepsilon) \\ &\Pr\{N(t+\varepsilon) - N(t) = 0\} = 1 - \lambda(t)\cdot\varepsilon + o(\varepsilon), \quad \varepsilon \to 0 \end{aligned} \tag{2}$$

то $\lambda(t)$ называется *интенсивностью точечного процесса*. Напомним, что $o(\varepsilon) = \varepsilon \cdot \nu(\varepsilon)$, где $\nu(\varepsilon)$ – произвольная величина, такая что $\nu(\varepsilon) \to 0$ при $\varepsilon \to 0$, причем, использование одного и того же символа $o(\varepsilon)$ не означает, что соответствующие функции $\nu(\varepsilon)$ будут

одинаковы. Интенсивность имеет смысл числа событий в единицу времени и размерность, обратную к размерности времени. Кроме того, потребуем выполнение условия *ординарности* процесса, то есть того, что вероятность возникновения 2-х и более событий на малом интервале времени длиной ε есть величина $o(\varepsilon)$:

$$\Pr\{N(t+\varepsilon)-N(t)\geq 2\}=o(\varepsilon),\quad \varepsilon\to 0 \tag{3}$$

2. *Пуассоновский процесс* есть точечный процесс с постоянной интенсивностью $\lambda(t)=\mu=const>0$ (условие стационарности), удовлетворяющий условиям (2) и (3) и для которого вероятность возникновения события на будущем интервале времени любой длины $(\tau,\tau+h]$ не зависит от того, сколько и в какие моменты времени событий произошло в прошлом, до момента времени τ (отсутствие последействия или памяти о прошлых событиях). Из условий стационарности и независимости возникновения событий следует, что любой момент времени наблюдения пуассоновского процесса можно считать начальным или нулевым.

Выведем форуму для вероятности того, что на интервале времени $(0,t]$ произойдет ровно k событий, то есть для величины $P_k(t)=\Pr\{N(t)=k\,|\,k>0\}$ при условии, что $N(0)=0$. Вероятность $P_k(t+\varepsilon)$ того, что на интервале времени $(0,t+\varepsilon]$ произошло ровно $k\geq 1$ событий состоит из суммы 2-х вероятностей: того что на интервале $(0,t]$ произошло k событий, а на интервале $(t,t+\varepsilon]$ – ни одного и вероятности того, что на интервале $(0,t]$ произошло $k-1$ событие, а на интервале $(t,t+\varepsilon]$ – одно. При $\varepsilon\to 0$ всеми прочими вариантами можно пренебречь в силу условия ординарности (3). В силу независимости событий, стационарности процесса и из условия (2) получаем

$$\begin{aligned}&P_k(t+\varepsilon)=\Pr\{N(t)=k\}\cdot\Pr\{N(t+\varepsilon)-N(t)=0\}+\\&+\Pr\{N(t)=k-1\}\cdot\Pr\{N(t+\varepsilon)-N(t)=1\}=\\&=P_k(t)\cdot(1-\mu\varepsilon+o(\varepsilon))+P_{k-1}(t)\cdot(\mu\varepsilon+o(\varepsilon))\end{aligned} \tag{4}$$

откуда следует дифференциальное уравнение для $P_k(t)$:

$$\frac{dP_k(t)}{dt}=-\mu\cdot(P_k(t)-P_{k-1}(t)),\quad P_k(0)=0,\quad k\geq 1 \tag{5}$$

Что же касается вероятности $P_0(t) = \Pr\{N(t) = 0\}$, то для нее

$$P_0(t+\varepsilon) = \Pr\{N(t) = 0\} \cdot \Pr\{N(t+\varepsilon) - N(t) = 0\} = P_0(t) \cdot (1 - \mu\varepsilon + o(\varepsilon)) \tag{6}$$

откуда сразу следует, что

$$\frac{dP_0(t)}{dt} = -\mu \cdot P_0(t), \quad P_0(t) = e^{-\mu t} \tag{7}$$

поскольку очевидно, что $P_0(0) = 1$ (за «нулевой» промежуток времени от начала наблюдений не произойдет ни одного события). Рассмотрим производящую функцию

$$f(t,z) = \sum_{k=0}^{\infty} P_k(t) \cdot z^k \tag{8}$$

от комплексного аргумента z, которая является аналитической при $|z| < 1$ в силу условия $P_k(t) < 1$ и нормировки $\sum_{k=0}^{\infty} P_k(t) = 1$. Если удастся найти явно функцию (8), то искомые вероятности определяются дифференцированием:

$$P_k(t) = \frac{1}{k!} \cdot \frac{d^k f(t,z)}{dz^k}\bigg|_{Z=0} \tag{9}$$

Имеем, учитывая (5) и (7):

$$\begin{aligned} &\frac{\partial f(t,z)}{\partial t} = \sum_{k=0}^{\infty} \frac{dP_k(t)}{dt} \cdot z^k = -\mu P_0(t) - \mu \sum_{k=1}^{\infty} (P_k(t) - P_{k-1}(t)) \cdot z^k = \\ &= -\mu \cdot f(t,z) + \mu z \cdot f(t,z) = \mu \cdot (z-1) \cdot f(t,z) \end{aligned} \tag{10}$$

откуда получаем:

$$f(t,z) = \exp(\mu (z-1)t) \tag{11}$$

в силу начального условия $f(0,z)=\sum_{k=0}^{\infty}P_k(0)\cdot z^k=P_0(0)\cdot z^0=1$. Подставляя (11) в (9) получим формулу для искомой вероятности, которая годится и для $k=0$:

$$P_k(t)=\Pr\{N(t)=k\}=\frac{e^{-\mu t}(\mu t)^k}{k!},\quad k\geq 0 \tag{12}$$

так как $0!=1$. Нетрудно получить среднее значение случайной величины $N(t)$:

$$M\{N(t)\}=\sum_{k=0}^{\infty}k\cdot P_k(t)=\left.\frac{df(t,z)}{dz}\right|_{z=1}=\mu t \tag{13}$$

Таким образом, интенсивность μ пуассоновского процесса есть ничто иное, как скорость роста среднего числа событий с увеличением длительности интервала наблюдения. Для вычисления дисперсии $N(t)$ используем формулу:

$$D\{N(t)\}=M\{(N(t)-M\{N(t)\})^2\}=M\{N^2(t)\}-(M\{N(t)\})^2=M\{N^2(t)\}-(\mu t)^2 \tag{14}$$

Вычислим $M\{N^2(t)\}$:

$$M\{N^2(t)\}=\sum_{k=0}^{\infty}k^2P_k(t)=\left.\frac{d^2f(t,z)}{dz^2}\right|_{z=1}+\left.\frac{df(t,z)}{dz}\right|_{z=1}=(\mu t)^2+(\mu t) \tag{15}$$

откуда, с учетом (14):

$$D\{N(t)\}=\mu t \tag{16}$$

Итак, дисперсия пуассоновского процесса растет линейно со временем с той же скоростью, что его математическое ожидание, равной интенсивности. Отсюда, в частности, следует полезный факт, вытекающий из независимости возникновения событий и центральной предельной теоремы [1], что плотность распределения случайной величины:

$$\eta(t)=\frac{N(t)-M\{N(t)\}}{\sqrt{D\{N(t)\}}}=\frac{N(t)-\mu t}{\sqrt{\mu t}} \tag{17}$$

при $t\to\infty$ стремится к стандартному нормальному распределению:

$$\Pr\{a \le \eta(t) < b\} \underset{t\to\infty}{\to} \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2} du \tag{18}$$

Обозначим через τ_n непрерывную случайную величину, равную времени ожидания n-го события для пуассоновского процесса и пусть $\varphi_n(t)$ – плотность вероятности ее распределения. Имеем:

$$\Pr\{\tau_n < t\} = 1 - \Pr\{\tau_n \ge t\} = 1 - \sum_{k=0}^{n-1} P_k(t) = 1 - \sum_{k=0}^{n-1} \frac{e^{-\mu t}(\mu t)^k}{k!} \tag{19}$$

Отсюда:

$$\varphi_n(t) = \frac{d \Pr\{\tau_n < t\}}{dt} = \mu e^{-\mu t} \sum_{k=0}^{n-1} \frac{(\mu t)^k}{k!} - \mu e^{-\mu t} \sum_{k=0}^{n-1} k \frac{(\mu t)^{k-1}}{k!} = \mu e^{-\mu t} \frac{(\mu t)^{n-1}}{(n-1)!} \tag{20}$$

Заметим, что $\varphi_n(t)$ есть ничто иное, как плотность вероятности для величины $\chi^2_{2n}/(2\mu)$. Напомним, что χ^2_m есть распределение суммы квадратов независимых случайных величин $u_i, i = 1,...,m$, одинаково распределенных согласно стандартному нормальному закону: $u_i \sim \mathbb{N}(0,1)$, $\sum_{i=1}^{m} u_i^2 \sim \chi^2_m$. Таким образом $2\mu\tau_n \sim \chi^2_{2n}$. Поскольку $M\{\chi^2_m\} = m, \quad D\{\chi^2_m\} = 2m$, то отсюда следует, что

$$M\{\tau_n\} = \frac{n}{\mu}, \quad D\{\tau_n\} = \frac{n}{\mu^2} \tag{21}$$

Формулы (21) могут быть выведены также непосредственно из определения математического ожидания и дисперсии с использованием формулы интегрирования по частям:

$$M\{\tau_n\} = \int_0^\infty t\varphi_n(t)dt, \quad D\{\tau_n\} = \int_0^\infty t^2\varphi_n(t)dt - (M\{\tau_n\})^2 \tag{22}$$

Согласно формуле (20) время ожидание 1-го события распределено с плотностью $\mu e^{-\mu t}$. По своему смыслу это время есть длина интервалов между событиями, поскольку каждое событие не зависит от предыдущих. Поэтому функция $\mu e^{-\mu t}$ является плотностью

распределения длин интервалов Δt между событиями, то интеграл от нее, $1-e^{-\mu t}$ равен функции распределения этих величин:

$$\Pr\{\Delta t < t\} = 1 - e^{-\mu t} \tag{23}$$

3. Рассмотрим случай, когда интенсивность $\lambda(t)$ в формулах (2) не является постоянной и может меняться со временем. Обозначим через $\Psi(t,s)$ вероятность того, что на интервале $(t,s]$ не произойдет ни одного события. Тогда вероятность того, что на интервале $(t,s+\varepsilon]$ по прежнему не будет ни одного события, в силу независимости моментов времени событий, равна произведению вероятности $\Psi(t,s)$ на вероятность, что на $(s,s+\varepsilon]$ не будет ни одного события. Используя условие (2), получаем:

$$\Psi(t,s+\varepsilon) = \Psi(t,s)\cdot(1-\lambda(s)\cdot\varepsilon + o(\varepsilon)) \tag{24}$$

или:

$$\frac{\partial\Psi(t,s)}{\partial s} = -\lambda(s)\cdot\Psi(t,s) \tag{25}$$

Поскольку $\Psi(t,t)=1$, то из (25) получаем:

$$\Psi(t,s) = \exp\left(-\int_t^s \lambda(u)du\right) \tag{26}$$

Отсюда следует, что функция распределения длин интервалов между событиями для нестационарного пуассоновского процесса равна $1-\Psi(t,s)$, а плотность вероятности распределения длин интервалов

$$\varphi(t,s) = \frac{\partial}{\partial s}\big(1-\Psi(t,s)\big) = \lambda(s)\cdot\exp\left(-\int_t^s \lambda(u)du\right) \tag{27}$$

Поэтому плотность вероятности того, что события произойдут в моменты времени $t_j > 0,\ j=1,...,N$ равна, в силу независимости распределения длин интервалов между событиями, произведению:

$$p(t_1,...,t_N)=\prod_{j=1}^{N}\lambda(t_j)\cdot\exp\left(-\int_{t_{j-1}}^{t_j}\lambda(u)du\right),\quad t_0=0 \tag{28}$$

Эта формула является, по сути, функцией правдоподобия для оценки параметров модели интенсивности точечных процессов. Пусть имеется некоторая функция для интенсивности, известная с точностью до вектора параметров θ: $\lambda=\lambda(t\,|\,\theta)$ и последовательность моментов времени $(t_1,...,t_N)$, наблюденных на интервале $(0,T]$. Тогда, согласно методу максимума правдоподобия, вектор θ может быть оценен путем максимизации выражения (28), рассматриваемого как функции от θ. Удобнее рассматривать его логарифм (логарифмическая функция правдоподобия):

$$\ln(L(\theta\,|\,t_1,...,t_N))=\sum_{j=1}^{N}\ln(\lambda(t_j\,|\,\theta))-\int_0^T\lambda(u\,|\,\theta)du\to\max_{\theta} \tag{29}$$

Например, для пуассоновского процесса с постоянной интенсивностью $\lambda(t\,|\,\theta)=\mu$ и в качестве вектора параметров θ следует рассматривать само значение интенсивности. Тогда (29) имеет вид:

$$\ln L=N\ln(\mu)-\mu\cdot T\to\max_{\mu},\Rightarrow\mu=\hat{\mu}_0=N/T \tag{30}$$

4. *Выделение периодических компонент интенсивности точечных процессов с помощью оценки приращения логарифмической функции правдоподобия*.

Метод был предложен в работе [2]. Пусть $t_i, i=1,...,N$ – времена последовательности событий, наблюдаемых на интервале $(0,T]$. Рассмотрим следующую модель интенсивности, содержащую периодическую компоненту:

$$\lambda(t)=\mu\cdot(1+a\cos(\omega t+\varphi)) \tag{31}$$

где частота ω, амплитуда $a, 0\le a\le 1$, фазовый угол φ, $\varphi\in[0,2\pi]$ и множитель $\mu>0$ (описывающий пуассоновскую часть интенсивности) являются параметрами модели. Таким образом, пуассоновская часть интенсивности модулируется гармоническим колебанием.

Зафиксируем какое-то значение частоты ω. Логарифмическая функция правдоподобия (29) в этом случае для серии наблюденных событий равна:

$$\begin{aligned}\ln L(\mu, a, \varphi \mid \omega) &= \sum_{t_i} \ln(\lambda(t_i)) - \int_0^T \lambda(u)du = \\ &= N\ln(\mu) + \sum_{t_i} \ln(1 + a\cos(\omega t_i + \varphi)) - \mu T - \frac{\mu a}{\omega}[\sin(\omega T + \varphi) - \sin(\varphi)]\end{aligned} \tag{32}$$

Взяв максимум выражения (32) по отношению к параметру μ нетрудно найти что:

$$\mu = \hat{\mu}(a, \varphi \mid \omega) = \frac{N}{T + a(\sin(\omega T + \varphi) - \sin(\varphi))/\omega} \tag{33}$$

Подставляя (33) в формулу (32) получаем

$$\ln(L(\hat{\mu}, a, \varphi \mid \omega)) = \sum_{t_i} \ln(1 + a\cos(\omega t_i + \varphi)) + N \cdot \ln(\hat{\mu}(a, \varphi \mid \omega)) - N \tag{34}$$

Следует заметить, что выражение $\hat{\mu}(a = 0, \varphi \mid \omega) \equiv \hat{\mu}_0 = N/T$ является оценкой (30) интенсивности процесса при условии, что он является однородным пуассоновским (чисто случайным).

Таким образом, приращение логарифмической функции правдоподобия вследствие рассмотрения более богатой, чем для чисто случайного потока событий, модели интенсивности с гармонической компонентой с заданной частотой ω равно:

$$\Delta \ln L(a, \varphi \mid \omega) = \sum_{t_i} \ln(1 + a\cos(\omega t_i + \varphi)) + N \cdot \ln(\hat{\mu}(a, \varphi \mid \omega)/\hat{\mu}_0) \tag{35}$$

Пусть

$$R(\omega) = \max_{a,\varphi} \Delta \ln L(a, \varphi \mid \omega), \;\; 0 \le a \le 1, \varphi \in [0, 2\pi] \tag{36}$$

Функция (36) может рассматриваться как обобщение спектра для последовательности событий. График этой функции показывает насколько «более выгодна» периодическая

модель интенсивности по сравнению с чисто случайной моделью. Максимальные значения функции (36) выделяют частоты, присутствующие в потоке событий.

Следующим очевидным обобщением метода является вычисление функции (36), используя наблюденные моменты времени не на всем интервале $(0,T]$, но внутри скользящего временного окна заданной длины L. Пусть τ – время правого конца скользящего временного окна. Тогда выражение (36) становится функцией от 2-х аргументов: $R(\omega,\tau \mid L)$, которая может быть визуализирована в виде 2-мерных карт или 3-мерных рельефов на плоскости аргументов (ω,τ). Эта частотно-временная диаграмма позволяет исследовать динамику возникновения и развития периодических компонент внутри исследуемого потока событий.

Важным вопросом применения этого метода к реальным данным является выяснение статистической значимости полученных пиковых значений статистик $R(\omega)$ или $R(\omega,\tau \mid L)$. Для его решения можно применить два подхода. Первый состоит в применении классической асимптотической теории Уилкса [3]. Пусть для одного и того же набора данных $X^{(N)}$, состоящего из N независимых наблюдений, рассматриваются 2 гипотезы:

1) гипотеза H_0: $X^{(N)}$ распределена в соответствии с плотностью $p_0(X^{(N)} \mid \theta_0)$;

2) гипотеза H_1: $X^{(N)}$ распределена в соответствии с плотностью $p_1(X^{(N)} \mid \theta_1)$.

где θ_0 и θ_1 - векторы неизвестных параметров, имеющих размерности m_0 и m_1, причем гипотеза H_1 является более «богатой»: $m_1 > m_0$, а вектор параметров θ_1 полностью включает в себя компоненты вектора θ_0. Рассмотрим разницу между логарифмами правдоподобий для этих двух гипотез, при условии, что для векторов параметров взяты их оценки метода максимального правдоподобия:

$$\Delta \ln L(X^{(N)}) = \ln\left(\max_{\theta_1} p_1(X^{(N)} \mid \theta_1)\right) - \ln\left(\max_{\theta_0} p_0(X^{(N)} \mid \theta_0)\right) \qquad (37)$$

Очевидно, что $\Delta \ln L(X^{(N)}) \geq 0$. Согласно теореме Уилкса, при условии справедливости гипотезы H_0 величина (37) имеет асимптотическое распределение:

$$\Delta \ln L(X^{(N)}) \sim \frac{\chi^2_m}{2}, \quad m = m_1 - m_0, \quad N \to \infty \qquad (38)$$

В нашем случае $m=2$ и, следовательно, удвоенная величина (36) имеет асимптотическую плотность распределения χ_2^2, равную $e^{-x/2}/2$, а сама величина (36) распределена асимптотически с экспоненциальной плотностью e^{-x} или

$$\Pr\{R(\omega)<x\}=1-e^{-x}, \quad N\to\infty \tag{39}$$

при условии, что анализируемая последовательность моментов времени распределена согласно пуассоновскому закону с постоянной интенсивностью.

Недостатком этого подхода является его асимптотический характер. Следовательно, область его применимости ограничивается в основном «статическими» оценками, когда периодические компоненты ищутся с использованием информации от всей выборки и эта выборка содержит «достаточно большое» число событий. Если же вычисляется «динамическая» оценка в скользящем временном окне, то число событий варьируется от окна к окну и может достигать лишь нескольких десятков, что ставит вопрос о применимости формулы (39). В этом случае более надежным является подход, основанный на статистическом моделировании. Именно, вычисляется оценка средней интенсивности точечного процесса по формуле $\hat{\mu}_0=N/T$ и генерируется длинная выборка моментов времени, интервалы между которыми распределены по закону Пуассона с интенсивностью $\hat{\mu}_0$. Согласно формуле (23) для этого последовательно вызывается датчик псевдослучайных чисел ξ, равномерно распределенных на $[0,1]$ и для каждого значения ξ_j находится длина интервала Δt_j до следующего события по формуле:

$$\Delta t_j=\ln(1-\xi_j)/\hat{\mu}_0, \quad t_j=t_{j-1}+\Delta t_j, \quad t_0=0 \tag{40}$$

Формула (40) применяется многократно для независимых реализаций псевдослучайных чисел ξ_j и генерируются моменты времени t_j. Общее число таких искусственных моментов времени определяется из условия, чтобы их было «много больше» реального числа наблюдений, например, $10^2N \div 10^3N$. Далее к этой выборке применяется метод вычисления $R(\omega,\tau\,|\,L)$ с той же длиной окна L, что и для реальных данных, но при смещении окон, равных длине окна (чтобы получить независимые значения (36) для не перекрывающихся окон), и рассматриваются полученные результаты. Пиковые значения $R(\omega,\tau\,|\,L)$ на искусственных данных дают пороги статистической значимости максимумов статистики (36). Для формализации этого подхода можно построить эмпирическую функцию

распределения значений $R(\omega,\tau\,|\,L)$ для заданной частоты (или периода) на искусственных моментах времени и задать уровень значимости как квантиль этого распределения с соответствующей доверительной вероятностью.

5. *Программная реализация метода*.

Программа PPPeriod (**P**oint **P**rocesses **Period**icity) написана на языке Compaq Visual Fortran и представляет собой консольное приложение. Диалог с пользователем организован по принципу последовательности «вопрос-ответ» на английском языке. Целью программы является вычисление приращений максимума логарифмической функции правдоподобия для модели интенсивности точечного процесса с заданным периодом по отношению к максимуму логарифмической функции правдоподобия для модели чисто пуассоновского процесса с постоянной интенсивностью (для нулевой гипотезы). Значения периода сканируются в заданных пределах и те периоды, для которых эта разность имеет пик, существенно возвышающийся над уровнем фоновых статистических флуктуаций оценки, выделяют периодические компоненты потока событий.

Входным файлом программе может служить любая символьная таблица чисел, в первой колонке которой стоят монотонно неубывающие числа, интерпретируемые как значения моментов времени. Прочие колонки входного файла, если таковые имеются, игнорируются.

По выбору пользователя программа может работать в 2-х режимах:

1) оценка по всей имеющейся выборке;
2) оценка в скользящем временном окне заданной длины и с заданным взаимным смещением (в размерном времени).

После запуска программы необходимо ответить на следующие вопросы:

1. Имя входного файла. Файл должен находится в той же директории, что и загрузочный модуль программы. Если файла нет или он имеет неправильную структуру (например, в первой колонке входной таблицы значения убывают для каких-то строк), то программа завершает работу с соответствующим сообщением.
2. Ввести значения $T_{\min}$ и $T_{\max}$ - минимальных и максимальных значений периодов сканируемого диапазона периодов в размерном времени, то есть в тех же единицах времени, которые присутствуют во входном файле.
3. Ввести значение N_p - числа пробных значений периодов, покрывающих диапазон $[T_{\min}, T_{\max}]$ с шагом, равномерным в логарифмическом масштабе, то есть рассматриваются периоды:

$$T_j = 10^{\varphi_j},\ \varphi_j = \lg(T_{\min}) + (j-1)\cdot\frac{\lg(T_{\max}) - \lg(T_{\min})}{(N_p - 1)},\quad j = 1,\ldots,N_p \qquad (41)$$

4. Задать режим вычислений – по всей выборке или в скользящем временном окне.
5. Если задан режим оценки в скользящем окне, то необходимо ответить на дополнительные вопросы:

   5.1. Задать длину T_L скользящего временного окна в размерном времени.

   5.2. Задать смещение ΔT_L скользящих временных окон в размерном времени.

   5.3. Задать смещение T_{shift} временных меток в выводном grd-файле.

Если задан режим оценки по всей выборке, то в текущей директории создается выводной файл с именем "**PPP_out.dat**", который представляет собой таблицу из 3-х колонок длиной N_p строк. В первой колонке последовательно идут тестированные значения периодов T_j, во второй – разность $\Delta\ln(L)$ между максимумами логарифмических функций правдоподобия, а в 3-й – значение безразмерного параметра $a, 0 < a \le 1$ амплитуды гармонического колебания интенсивности процесса в модели (31).

Если задан режим оценки в скользящем временном окне то, в текущей директории создается выводной файл с именем "**PPP_out.grd**" в символьном grd-формате, готовый для построения 2-мерных или 3-мерных рельефов частотно-временных диаграмм для статистики $\Delta\ln(L)$ в пакете Surfer. При этом временные метки соответствует значениям

$$T_{shift} + T_L + (k-1)\cdot\Delta T_L,\ k = 1,\ldots,N_W,$$

где индекс k нумерует временные окна, N_W - общее число временных окон. Таким образом, временные метки соответствуют правым концам скользящих временных окон, взятых со смещением T_{shift}. Параметр T_{shift} может быть полезен, например, в ситуации, когда времена событий указываются в годах от начала 1900-го года, а на диаграмме полезно иметь метки в общепринятых годах – тогда задается T_{shift} =1900.

6. *Примеры применения.*

Во всех рассматриваемых ниже примерах, если они касаются анализа последовательности землетрясений, данные взяты из источника: http://neic.usgs.gov/neis/epic/ .

6.1. *Периодические компоненты интенсивности сильнейших землетрясений для всего мира.* На рис.1. представлена последовательность сильнейших сейсмических событий для глобального сейсмического процесса с начала эпохи инструментальной сейсмологии по

настоящее время. Серые и черные вертикальные линии (1369 событий) дают последовательность событий с магнитудами $M \geq 7.0$ и глубинами эпицентров не более 100 км, 1901-2005 гг., и для магнитуд $M \geq 8.0$ – черные линии (126 событий). Энергия землетрясения $E = 10^{Class}$ джоулей, $Class = 4.8 + 1.5 \cdot M$ - формула Гуттенберга-Рихтера.

На рис.2 представлен график оценки статистики (36), причем рядом с каждым пиком поставлены значения периодов в годах. Для оценки значимости пиков можно применить асимптотическую формулу (39), считая, что число 1369 «достаточно велико». Если назначить доверительную вероятность 90%, то доверительным порогом будет значение $-\ln(0.1) \approx 2.3$. Видно, что все помеченные пики являются значимыми с этой вероятность.

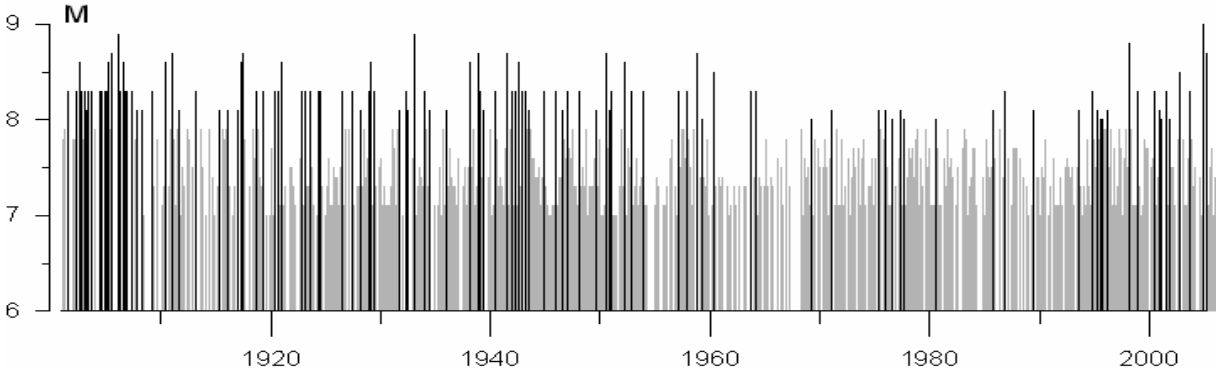

Рис.1. Последовательность моментов времени сильнейших землетрясений глобальной сейсмичности.

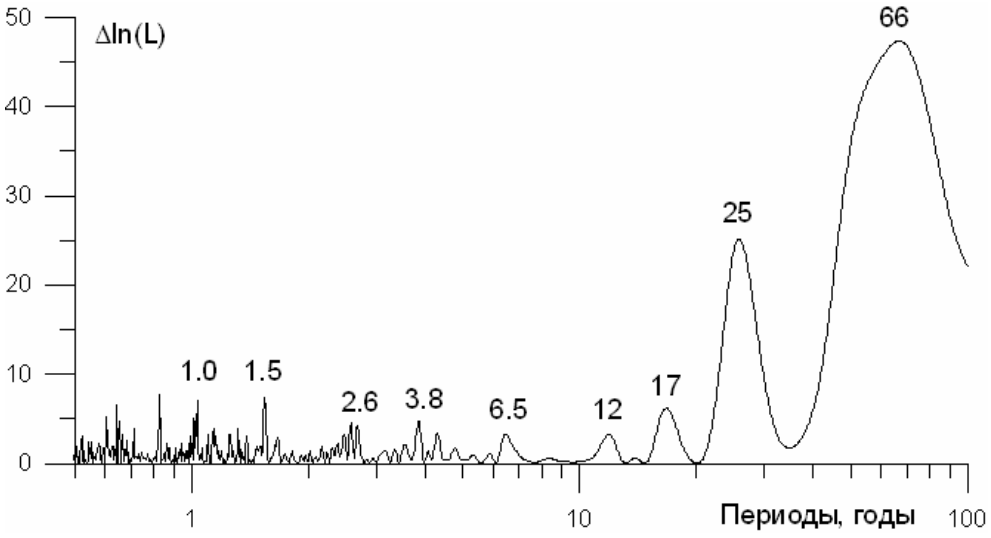

Рис.2. Оценка приращения логарифмической функции правдоподобия для всей последовательности моментов землетрясений $M \geq 7.0$ на рис.1, $T_{\min} = 182$ сут., $T_{\max} = 36500$ сут., $N_p = 1000$.

6.2. *Сейсмический режим в окрестности эпицентра землетрясения на Суматре 26.12.2004.*

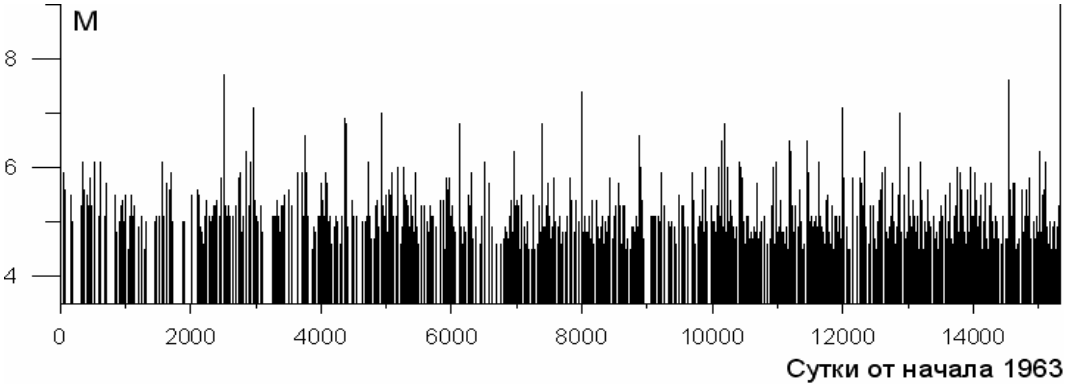

Рис.3. Последовательность сейсмических событий $M \geq 4.5$ за период 1963-2004 гг. внутри прямоугольной рамки на рис.4. Общее число событий строго до события 26.12.2004 равно 1387.

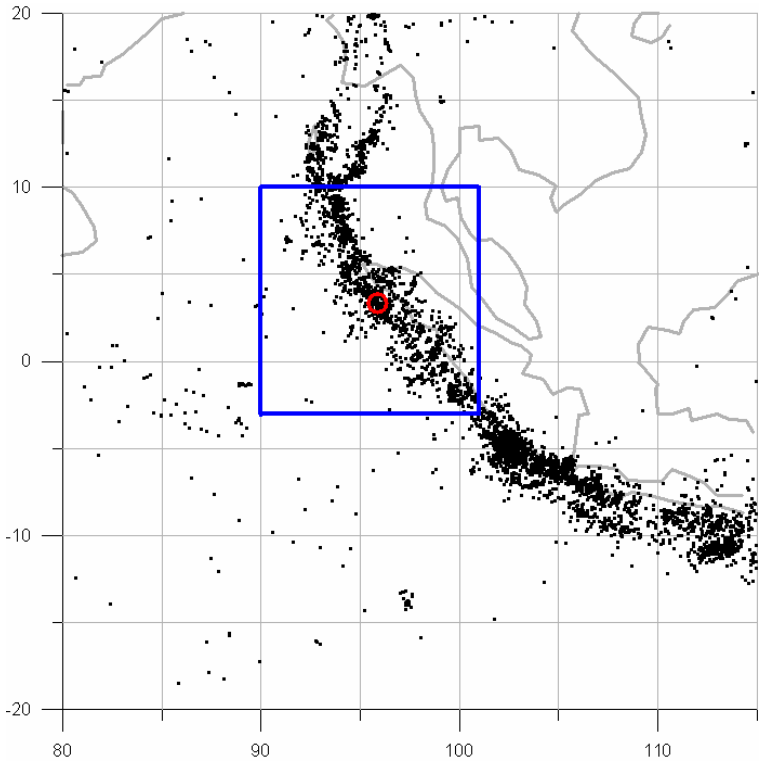

Рис.4. Распределение эпицентров землетрясений с магнитудами не менее 4.5 и глубинами эпицентров не более 100 км в Юго-Восточной Азии за период 1963-2004 гг. (строго до 26.12.2004), кружок – эпицентр землетрясения на Суматре 26.12.2004, M=9; прямоугольником выделена область рассмотрения сейсмического режима до момента толчка.

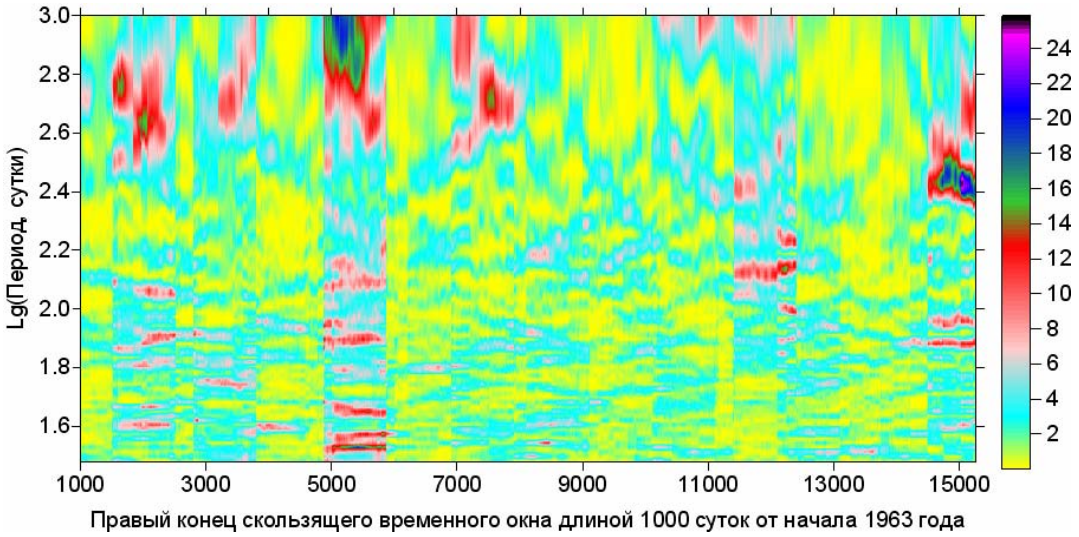

Рис.5. Эволюция приращения логарифмической функции правдоподобия $R(\omega,\tau\,|\,L)$ при оценке в скользящем окне длиной 1000 дней для периодов от 30 до 1000 суток.

На рис.5 изображена частотно-временная диаграмма статистики $R(\omega,\tau\,|\,L)$, которая оценивалась в скользящем временном окне длиной 1000 суток со смещением 10 суток для пробных значений периодов (соответствующих частоте ω в формуле (31)) от 30 до 1000 суток. Эти периоды образовывали равномерную логарифмическую шкалу. Поскольку из временной последовательности не устранялись афтершоки (серии событий после сильных землетрясений), то частотно-временная диаграмма имеет характерную линейчатую структуру с умеренными всплесками статистики $R(\omega,\tau\,|\,L)$ в афтершоковых сериях сильных землетрясений. Диаграмма на рис.5 имеет два существенных всплеска. Один – на максимальных периодах 800-1000 суток для временных меток правого конца окна от 5000 до 6000 дней от начала 1963 г. Величина этого всплеска приращения логарифмической функции правдоподобия равна 20. Второй и наиболее значительный (амплитуда всплеска = 25.8) возник внутри афтершоковой серии предыдущего сильного землетрясения (02.11.2002, M=7.6, Lat=2.82, Lon=96.08), происшедшего за 2 года до катастрофы. Характерный период этого второго всплеска – примерно 250-320 суток. На этих периодах за все время рассмотрения не было ни одного столь же значительного увеличения, несмотря на то, что землетрясения такой силы были в реализации, представленной на рис.4. Таким образом, эту частотно-временную аномалию можно рассматривать как своего рода предвестник землетрясения на Суматре, детектированного за 2 года до события, после того как правая часть скользящего окна достаточно глубоко вошла в серию афтершоков предыдущего толчка. Для проверки статистической значимости этого предвестника была сгенерирована серия из 10^5 событий, моменты времени которых распределены случайно, согласно распределению Пуассона с интенсивностью, равной средней интенсивности событий и для

этой серии оценена эволюция приращения логарифмической функции правдоподобия с теми же параметрами, что и для диаграммы на рис.5. В результате максимальный всплеск статистики $R(\omega,\tau\,|\,L)$ оказался равным 10.8 – это свидетельствует о статистической значимости всплеска перед землетрясением на диаграмме рис.5.

6.3. *Временной ряд ежедневных расходов воды в Рейне, измеренных в Кельне за период с 01.11.1816 по 31.05.1997.* Данные принадлежат Центру по глобальным данным расхода рек в г. Кобленце, Германия и были любезно предоставлены автору доктором Питером Ван Гельдером из Технического университета г. Дельфты, Нидерланды. На рис.6 представлен график временного ряда, который содержит 65956 отсчетов. Рассмотрим моменты времени максимальных расходов воды, превосходящих уровень, равный медиане ряда плюс четыре медианных отклонения – этот уровень обозначен на рис.6 толстой горизонтальной линией. Моменты времени этих пиковых значений расхода воды формируют некоторую последовательность событий, к изучению свойств которой может быть применен вышеизложенный метод.

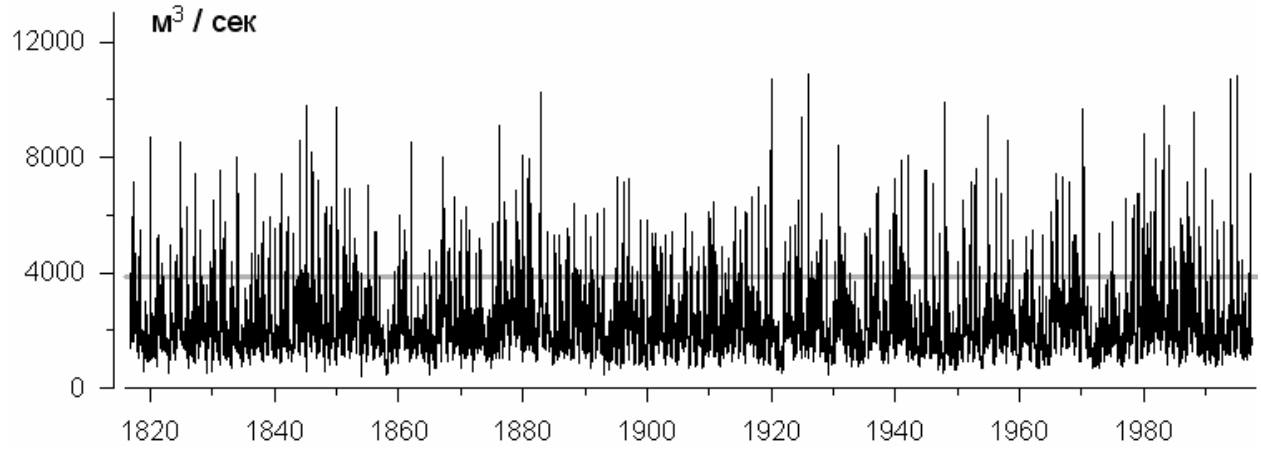

Рис.6. График ежесуточных расходов воды в Рейне (Кельн) за период 1816-1997 гг., толстой горизонтальной линией отмечен уровень, моменты времени пиковых выбросов за который рассматриваются как точечный процесс – всего 652 момента времени.

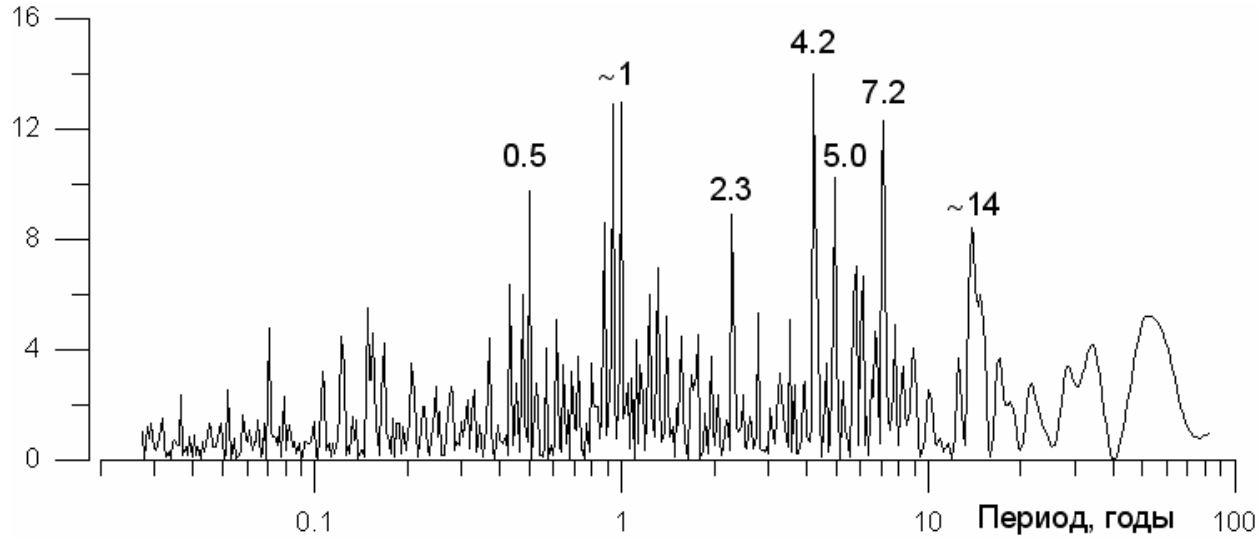

Рис.7. Оценка приращения логарифмической функции правдоподобия для всей последовательности моментов времени локальных максимумов, превышающих уровень на рис.6, $T_{\min} = 100$ сут., $T_{\max} = 30000$ сут., $N_p = 500$.

На рис.7 изображен график оценки приращения логарифмической функции правдоподобия для всех 652 моментов времени пиковых превышений порога. Считая 652 достаточно большим числом, аналогично оценке на рис.2, используя асимптотическое распределение (39), можно утверждать, что порог 2.3 является доверительным с вероятностью 0.9.

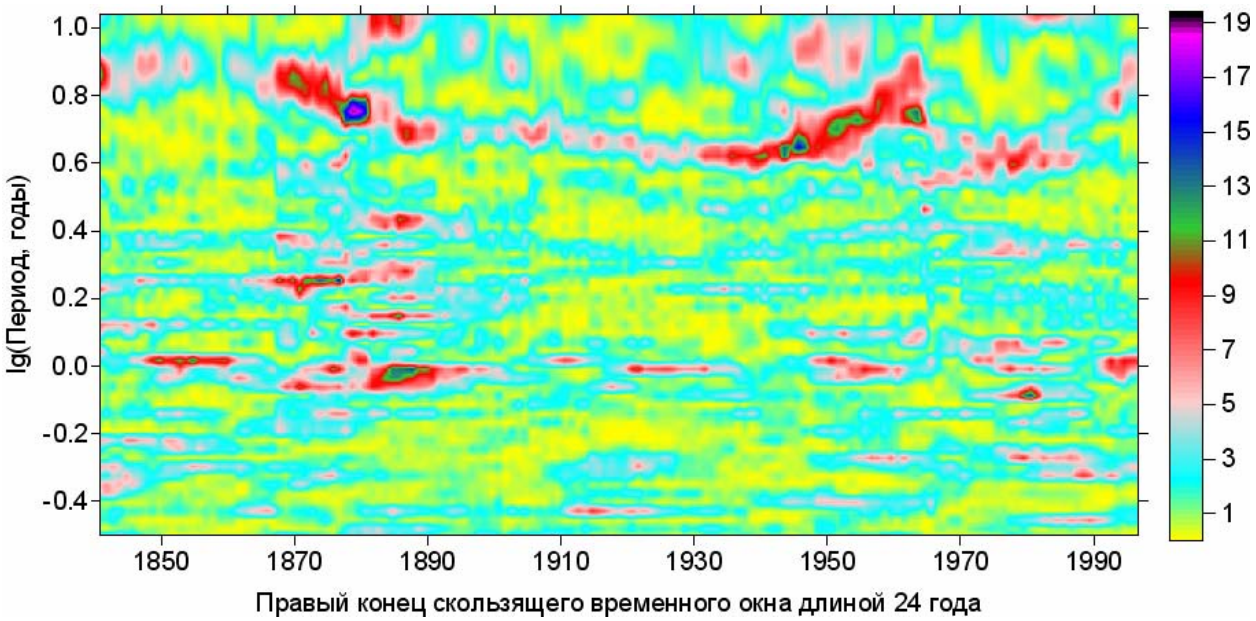

Рис.8. Эволюция приращения логарифмической функции правдоподобия $R(\omega, \tau \mid L)$ при оценке в скользящем окне длиной 24 года для периодов от 10 до 40000 суток.

На рис.8 изображена частотно-временная диаграмма приращения логарифмической функции правдоподобия при оценке в скользящем окне длиной 24 года (8766 дней) со

смещением 365 дней (1 год). Можно выделить основную особенность гидрологического режима: наличие периодических компонент в максимальных расходах с периодами от 4 до 8 лет (к этому интервалу принадлежит «библейский гидрологический цикл» Нила, равный 7 годам), причем максимальными эти периодические компоненты были в интервалы времени 1825-1880 и 1905-1970 гг. (с учетом 24-летней длины окна). Определение начала такого цикла может быть прогностическим признаком для усиления опасности наводнений.

ЗАДАНИЯ.

1. Сгенерировать искусственную пуассоновскую последовательность моментов времени t_k, заданного объема N для процесса с постоянной интенсивностью $\mu > 0$ согласно формуле (40). Оценить для нее, используя программу PPPeriod, приращение логарифмической функции правдоподобия для всей выборки и в скользящем временном окне для различных объемов N и длин временных окон. Учесть, что временная длина выборки $T = \sum_{k=0}^{N} \Delta t_k$, $\Delta t_k = t_k - t_{k-1}$, $t_0 = 0$, является случайной величиной, зависящей от случайных реализаций величин Δt_k длин интервалов между событиями.
2. Сгенерировать искусственную пуассоновскую последовательность моментов времени t_k, заданного объема N для процесса с периодической интенсивностью $\lambda(t) = \mu \cdot (1 + a\cos(\omega t + \varphi))$. Модель интенсивности содержит 4 параметра: интенсивность фона $\mu > 0$, амплитуду гармонических вариаций интенсивности $0 < a < 1$, период τ, через который надо выразить частоту: $\omega = 2\pi / \tau$, и фазу φ. Последовательность моментов времени t_k находится через значения длин интервалов между событиями: $t_k = t_{k-1} + \Delta t_k$, $t_0 = 0$. Значения Δt_k являются случайными величинами, генерируемыми аналогично предыдущему заданию, но с использованием функции распределения:

$$\Pr\{\Delta t_k < s\} = 1 - \Psi(t_{k-1}, t_{k-1} + s) = 1 - \exp\left(- \int_{t_{k-1}}^{t_{k-1}+s} \lambda(u)du \right), \quad t_k = t_{k-1} + \Delta t_k, \quad t_0 = 0 \qquad (42)$$

Оценить для последовательности t_k, используя программу PPPeriod, приращение логарифмической функции правдоподобия для всей выборки и в скользящем временном окне для различных значений параметров a, τ, объемов N и длин временных окон.
3. Сгенерировать искусственную последовательность моментов времени, состоящую из объединения нескольких (двух и более) нестационарных пуассоновских последовательностей моментов времени $\{ t_k^{(\alpha)}, k = 1, ..., N_\alpha; \alpha = 1, ..., m; m \geq 2 \}$. Здесь α -

индекс, нумерующий последовательности, $m \geq 2$ - общее число таких последовательностей, N_α - число событий в последовательности с номером α. Параметр $\mu > 0$ для всех последовательностей положить одинаковым, но значения периодов τ выбрать разными. При создании объединенной последовательности событий можно использовать тот же метод, что и в п.2, но перед объединением значения моментов времени в каждой последовательности, кроме первой, увеличить на величину момента времени последнего события в предыдущей серии. Оценить приращения логарифмической функции правдоподобия в скользящем временном окне различной длины. Убедиться, что при пересечении временным окном границ между сериями событий пик максимума функции (36) переключается с одного периода на другой.

4. Это задание может быть основой для дипломной работы. Зайти в интернете на сайт глобального сейсмического каталога по адресу http://neic.usgs.gov/neis/epic/ . Выбрать сейсмоактивный регион в виде прямоугольной или круговой области и скачать каталог сейсмических событий, начиная с 1963-го года (начало работы глобальной сети сейсмических наблюдений). Самостоятельно преобразовать даты событий в значения дней, прошедших от начала 1963-го года для магнитуд землетрясений, начиная с различных минимальных значений. Исследовать динамику периодических компонент сейсмического режима в выбранной области, используя программу PPPeriod для различных длин временных окон.

Литература.

1. Кокс Д., П.Льюис (1969) Статистический анализ последовательностей событий. М., Мир. 312с.
2. Любушин А.А., Писаренко В.Ф., Ружич В.В., Буддо В.Ю. (1998) Выделение периодичностей в сейсмическом режиме. Вулканология и сейсмология. 1998, № 1, С. 62-76.
3. Рао С.Р. (1968) Линейные статистические методы и их применение. М., Наука. 548 с.