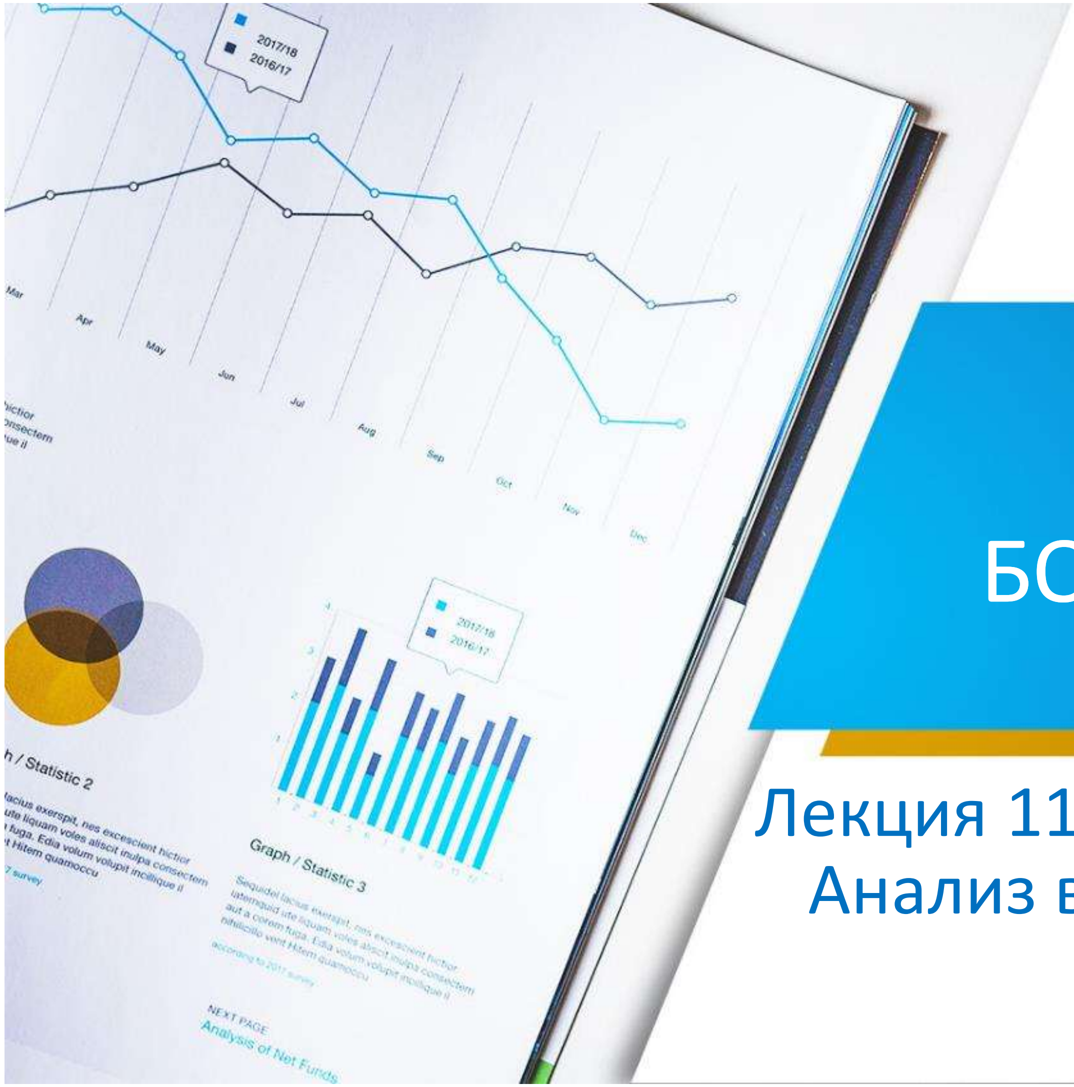


СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 11. Предобработка данных. Анализ временных рядов. Часть 3

Метод Прони

Барон Гаспар де Прони (1755—1839)

Метод Прони позволяет аппроксимировать последовательность комплексных данных $x_1,\ldots,x_n$ моделью, состоящей из m затухающих комплексных экспонент:

$$x_k = \sum_{i=0}^{m} A_i \exp\big(-\lambda_i \Delta t k + j(\omega_i \Delta t k + \varphi_i)\big), \qquad (1)$$

$$k = 1,2,\ldots n.$$

Здесь m — модальная глубина модели, $m \le n/2$, A_i — амплитуда; ω_i — частота; φ_i — начальная фаза; λ_i — коэффициент затухания; Δt — период дискретизации сигнала; k — номер отсчета; n — число отсчетов сигнала, j — мнимая единица.

Формула Эйлера $\exp(\lambda t + j\varphi t) = \exp(\lambda t)\,(\cos\varphi t + j\sin\varphi t)$

Метод Прони

Аппроксимация вида (1) можно представить в виде следующего полинома:

$$(2) \quad x_k = \sum_{i=1}^{m} h_i z_i^{k-1}, \quad k = 1,2,\dots,n,$$

где $h_i = A_i \exp(j\varphi_i)$, $z_i = \exp[(\lambda_i + 2\pi j\omega_i)\Delta t]$.

Метод Прони основан на том, что коэффициенты a_k характеристического полинома

$$P(z) = \prod_{i=1}^{m} (z - z_i) = z^m + a_1 z^{m-1} + \dots + a_{m-1} z + a_m$$

корни которого равны z_i, удовлетворяют системе линейных уравнений

$$\sum_{k=1}^{m} a_k x_{i-k} = -x_i, \qquad i = m+1, m+2, \dots, n. \qquad (3)$$

Метод Прони

Три этапа процедуры Прони:

1. Решается система (3) линейных уравнений: $\sum_{k=1}^{m} a_k x_{i-k} = -x_i$ относительно коэффициентов $a_1, a_2, \dots, a_m$ характеристического полинома $P(z)$.

2. Вычисляются корни $z_1, z_2, \dots, z_m$ полинома $P(z)$, по которым находятся коэффициенты затухания λ_k и частоты ω_k:

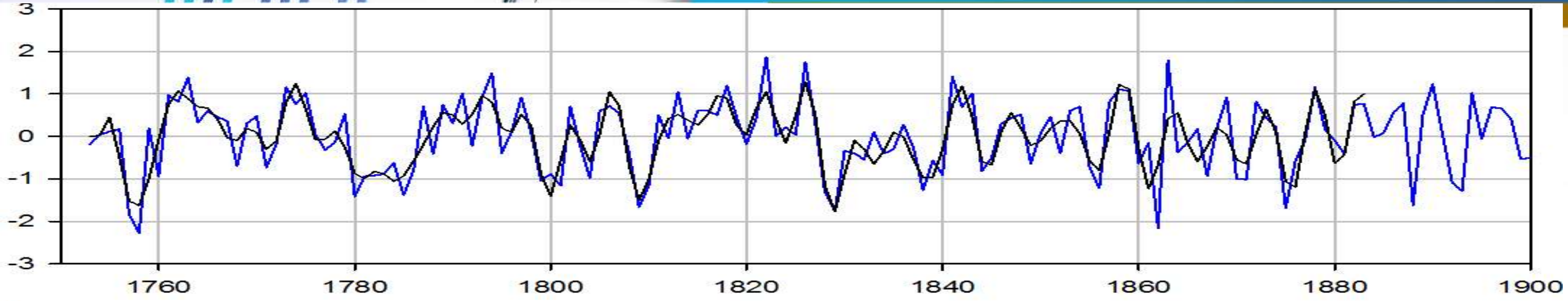
$$\lambda_k = \frac{\ln|z_k|}{\Delta t}, \quad \omega_k = \frac{1}{2\pi\Delta t}\tan^{-1}\frac{\mathrm{Im}(z_k)}{\mathrm{Re}(z_k)}.$$

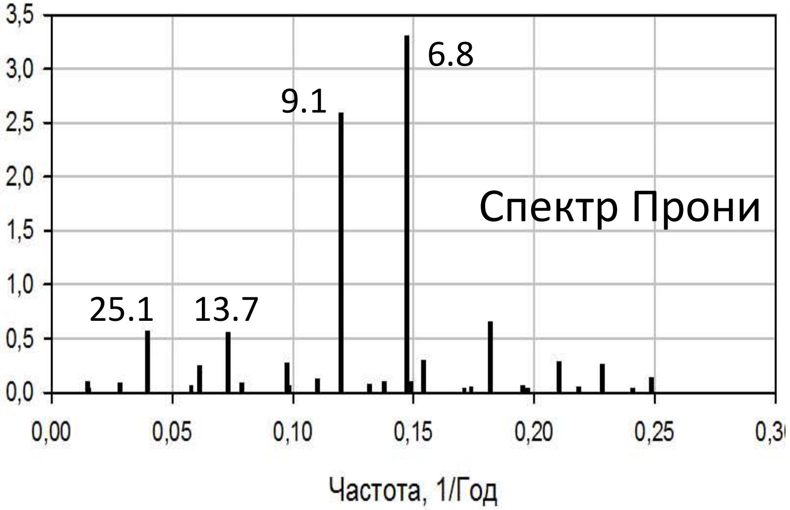
3. Решается система (2) линейных уравнений $x_k = \sum_{i=1}^{m} h_i z_i^{k-1}$ относительно комплексных амплитуд h_k. Вещественные амплитуды A_k и фазы φ_k равны

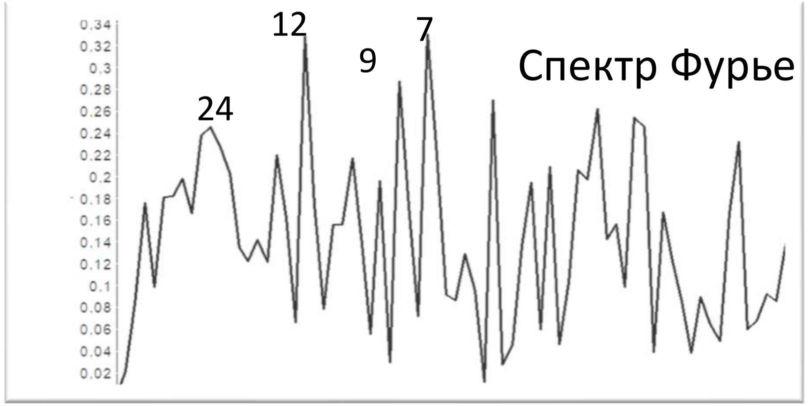
$$A_k = |h_k|, \qquad \varphi_k = \tan^{-1}\frac{\mathrm{Im}(h_k)}{\mathrm{Re}(h_k)}.$$

При $n > 2m$ системы (2) и (3) будут переопределенными. Для их решения применяют метод наименьших квадратов.

Метод Прони.Пример







Показатель Херста

Многие изучаемые системы (солнечные пятна, среднегодовых значений выпадения осадков, финансовые рынки…) не являются нормально-распределенными или близкими к ней.

Для анализа таких систем Херст предложил новую статистику – показатель Херста (H). Данный метод позволяет различить случайный и фрактальный временные ряды, а также делать выводы о наличии непериодических циклов, долговременной памяти и т.д.

Гарольд Эдвин Хёрст
(1880—1978)

Показатель Херста

Броуновское движение - хаотической движение микрочастиц, взвешенных в жидкости или газе (Броун, 1827).

В теории случайных процессов используется **Винеровский процесс** W_t как **математическая модель броуновского движения** (случайного блуждания с непрерывным временем).

Среднеквадратичное смещение частицы пропорционально корню квадратному из времени наблюдения (Формула Эйнштейна-Смолуховского):

$$\bar{X} = Ct^{0.5}.$$

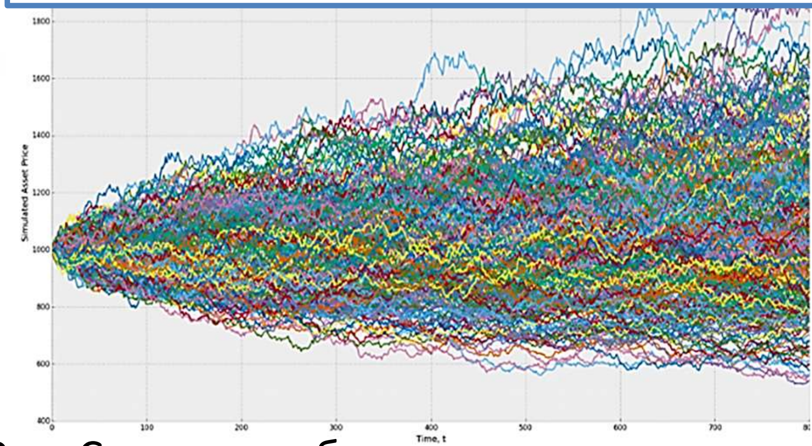


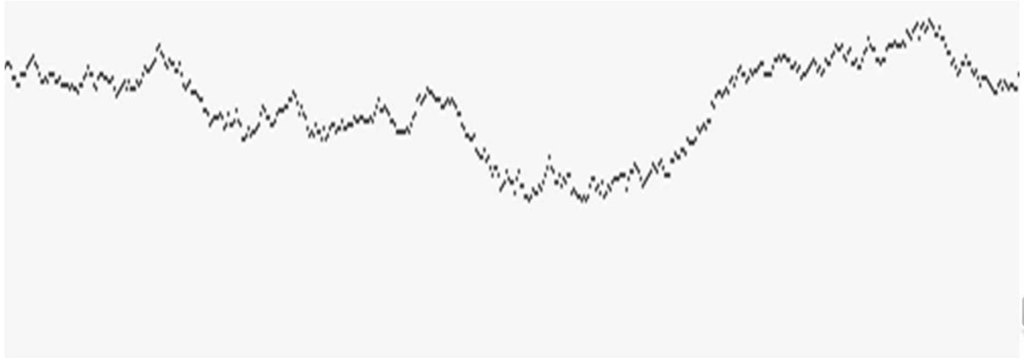
Рис. Симуляция броуновского движения.

Определение винеровского процесса.

1. $W_0 = 0$
2. W_t — процесс с независимыми приращениями.
3. Разность W_t-W_s распределена по нормальному закону $N(0, \sigma^2(t-s))$.

Винеровский процесс масштабно инвариантен или самоподобен. Если W_t - винеровский процесс, и c>0, то $V_t = \frac{1}{\sqrt{c}} W_{ct}$ также является винеровским процессом.

Демонстрация масштабной инвариантности винеровского процесса:



Показатель Херста Н

Показатель Херста *Н* определяется в терминах асимптотического поведения временного ряда следующим образом:

$$E\left(\frac{R(n)}{s(n)}\right) = Cn^H,$$

где $R(n)$ — размах накопленных отклонений первых n значений от среднего значения ряда, $s(n)$— стандартное отклонение; n — величина промежутка времени (количество точек в отрезке временного ряда), C — константа.

Размах $R(n)$: $$R(n) = \max_{m=1,n} X(m,n) - \min_{m=1,n} X(m,n)\,, \quad 1 \le m \le n.$$

$x(t)$ случайная величина, $x_i = x(t_i), n$ - длина выборки, $\bar{x}$— ее среднее,

$s(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ - стандартное отклонение, $X(m,n) = \sum_{i=1}^{m}(x_i - \bar{x})$ – накопившееся отклонение значений x_i от ее среднего значения за время m.

Свойства показателя Херста

Для **случайного процесса с независимыми приращениями и конечной дисперсией** было строго доказано (Feller W. The asymptotic distribution of the range of sums of independent variables // Ann. Math. Statist. 1951. V. 22)., что показатель **H равен 0,5.**

Имеются три различных классификации для показателя Херста:

1) **H ≈ 0.5**

Указывает на случайный ряд. События случайны и некоррелированы. Настоящее не влияет на будущее. Функция плотности вероятности может быть нормальной кривой, однако это не обязательное условие. R/S-анализ может классифицировать произвольный ряд, безотносительно к тому, какой вид распределения ему соответствует.

2) $0.5 < H < 1.0$

Имеем **персистентные,** или сохраняющие тренд ряды. Если ряд возрастает (убывает) в предыдущий период, то вероятно, что он будет сохранять эту тенденцию какое-то время в будущем. Тренды очевидны. Трендоустойчивость поведения или сила персистентности, увеличивается при приближении H к 1. Чем ближе H к 0.5, тем более зашумлен ряд и тем менее выражен его тренд. Персистентный ряд – это обобщенное броуновское движение, или смещенные случайные блуждания. Сила этого смещения зависит от того, насколько H больше 0.5. Персистентные временные ряды являют собой более интересный класс, так как оказалось, что они не только в изобилии обнаруживаются в природе, – это открытие принадлежит Херсту, – но и свойственны рынкам капитала.

Свойства показателя Херста

3**) 0 ≤ H< 0.5**

Данный диапазон соответствует **антиперсистентным**. Такой тип системы часто называют – «возврат к среднему». Если система демонстрирует рост в предыдущий период, то скорее всего, в следующем периоде начнется спад. И наоборот, если шло снижение, то вероятен близкий подъем. Устойчивость такого антиперсистентного поведения зависит от того, насколько H близко к нулю. Такой ряд более изменчив, или волатилен, чем ряд случайный, так как состоит из частых реверсов спад-подъем. Несмотря на широкое распространение концепции возврата к среднему в экономической и финансовой литературе, до сих пор было найдено мало антиперсистентных рядов.

Вычисление показателя Херста

1. Для первые *n* членов исходного ряда рассчитаем среднее значение и стандартное отклонение:

$$E(n) = \frac{1}{n}\sum_{k=1}^{n} x_k,\; s(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - E(n))^2}\,.$$

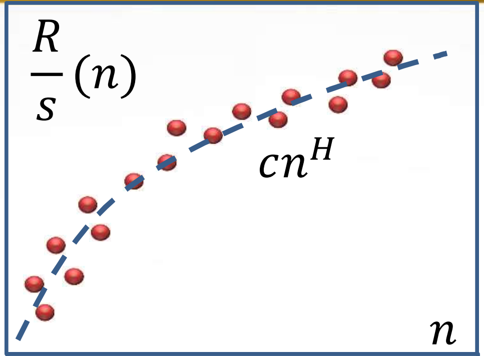
2. Отклонения от среднего значения: $X(k,n) = \sum_{i=1}^{k}(x_i - E(n))$.

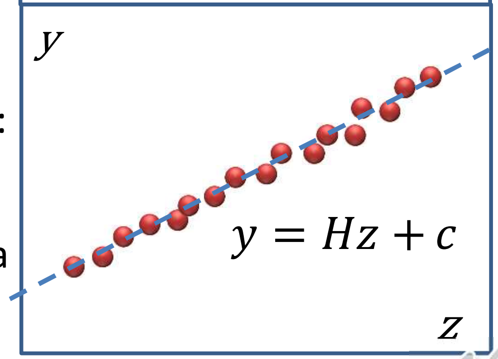
3. Размах: $R(n) = \max\limits_{k=1,n} X(k,n) - \min\limits_{k=1,n} X(k,n)$. Делим $R(n)$ на $s(n)$

4. Увеличиваем *n* повторяем шаги 1-3 и строим две выборки значений:

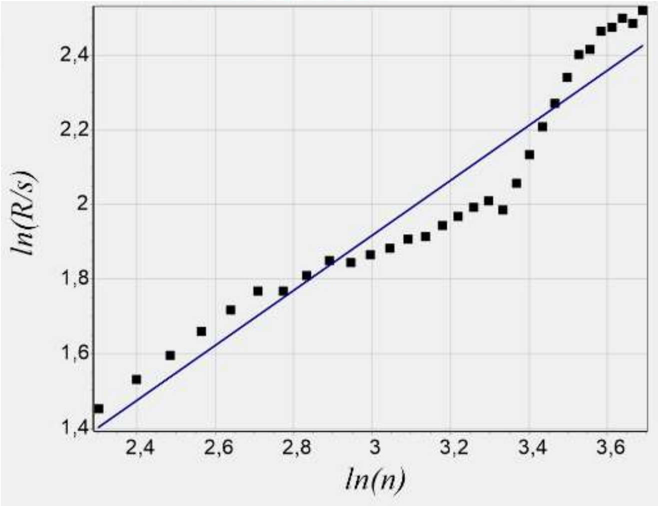
$$y(n) = \ln(R(n)/s(n)) \text{ и } z(n) = \ln(n)$$

5. Находим регрессию вида: $y = Hz + c$. Коэффициент H – оценка показателя Херста.

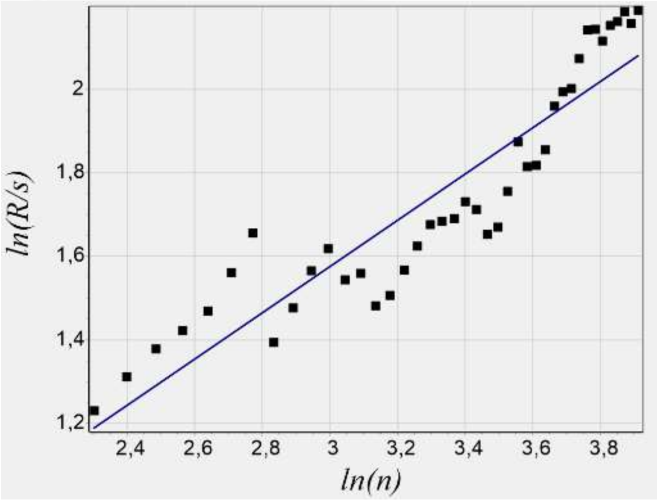




Пример ряда годовых температур СПб

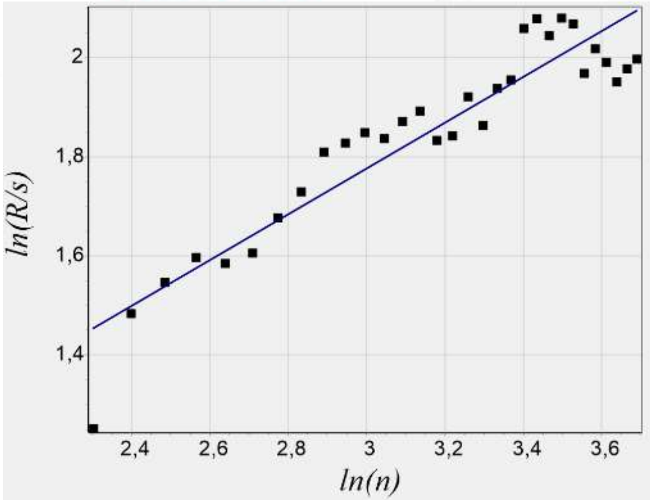


Вычисление показателя Херста для начала ряда среднегодовых температур Спб (первые 40 точек, 1753-1793г).
H=0.737

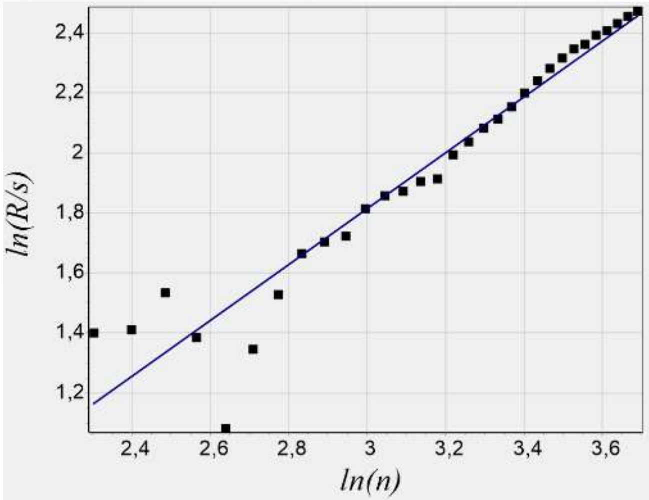


Вычисление показателя Херста для середины ряда среднегодовых температур Спб (точки с 120 по 170, 1873-1923г).
H=0.554

Пример ряда годовых температур СПб

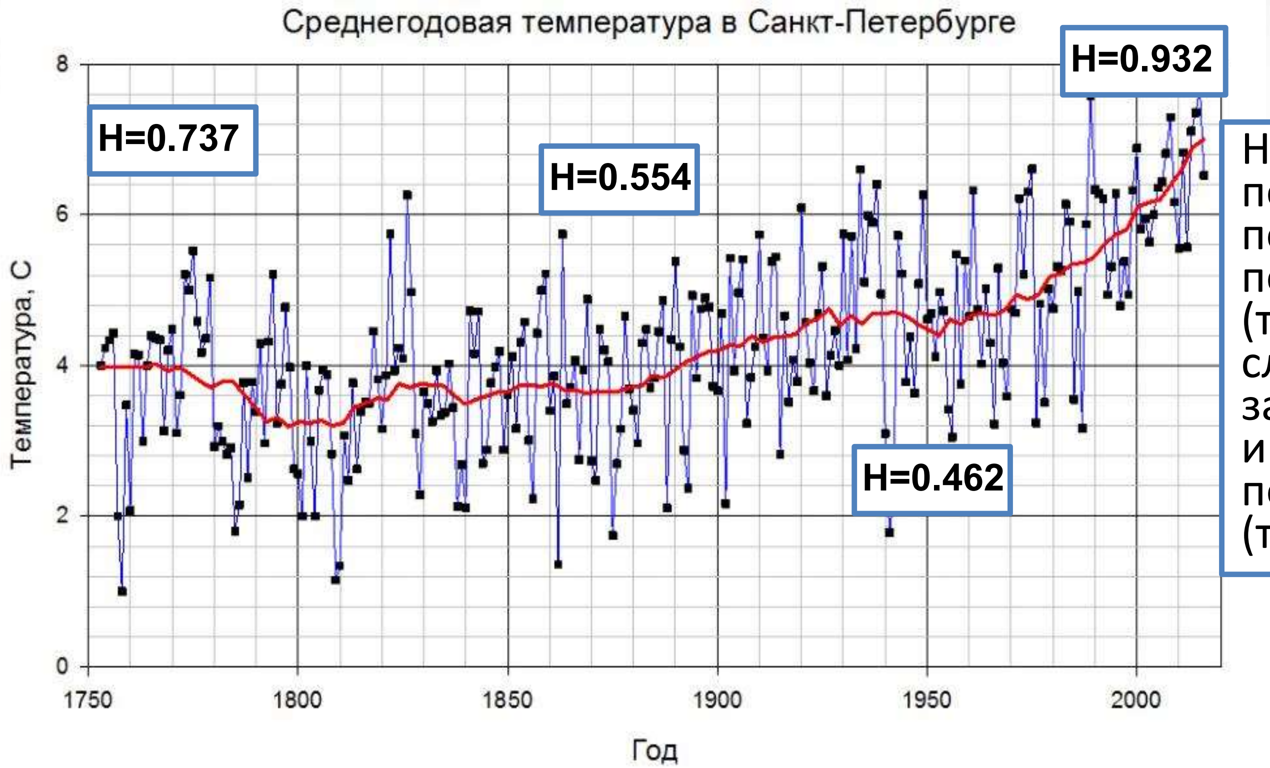


Вычисление показателя Херста для ряда среднегодовых температур Спб точки с 180 по 220, 1933-1973г). **H=0.462**



Вычисление показателя Херста для конца ряда среднегодовых температур Спб (последние 40 точек, 1976-2016г). **H=0.932**

Пример ряда годовых температур СПб



Наблюдается качественный переход в характере поведения температуры: от персистентности (тенденция к понижению) к случайному поведению, затем антиперсистентность и переход к существенной персистентности (тенденция к повышению)

Контрольные вопросы и задания

1. Составить на языке R программу разложения ряда по методу Прони. Для ее проверки сгенерировать модельный ряд $x_i = \sum_{k=1}^{3} k \exp\left(-\frac{hi}{k}\right) \cos\left(4\pi khi + \frac{\pi}{k}\right), i = 1, \ldots, 200, h = 0.02$ и применить к нему метод Прони.
2. Найти в интернете данные по среднесуточным температурам за два года вашего населенного пункта и провести их анализ на наличие тренда и сезонных колебаний.