Мониторинг метрик и детекция аномалий

Курс «Продуктовые метрики»

Зачем нужен мониторинг?

Оперативное выявление аномалий в продуктовых и бизнес метриках помогает:

- предотвратить финансовые потери;
- защитить пользовательский опыт;
- учитывать внешние факторы (праздники, парсинг).

Задачи:

- мониторинг активности пользователей;
- поиск ошибок после релизов;
- фильтрация искусственных аномалий.

Временные ряды: основные понятия

Временной ряд — это последовательность значений изменяющегося во времени признака, собранных и упорядоченных через равные промежутки времени.

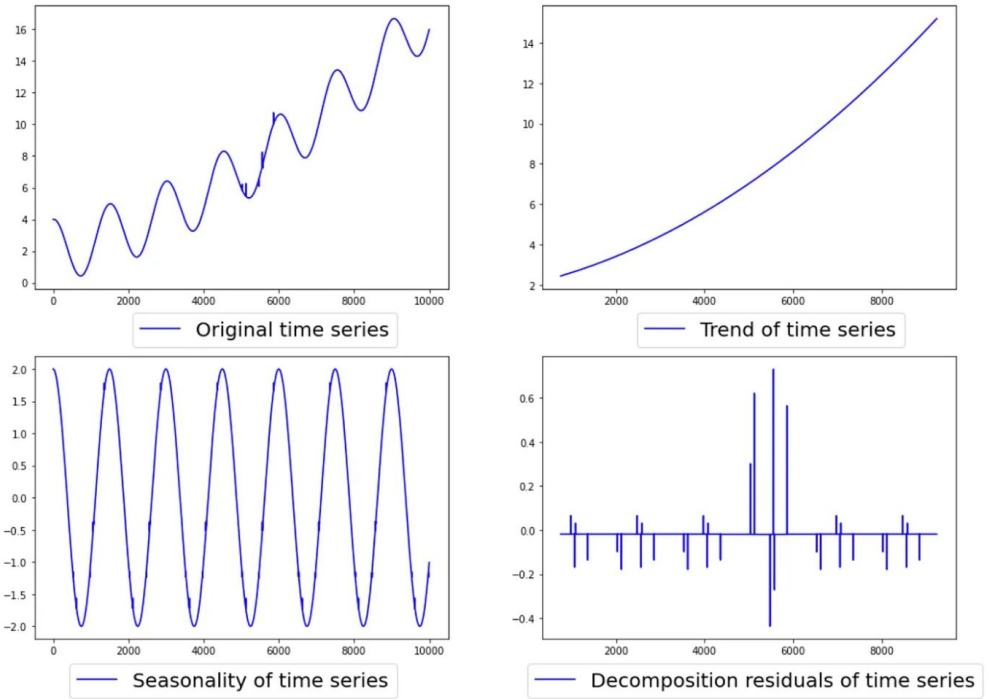
Временные компоненты:

- Тренд
- Сезонность
- Ошибки (при разложении — остатки)

Cleveland, Cleveland, McRae & Terpenning — STL: A Seasonal-Trend Decomposition Procedure Based on Loess (1990)
Box & Jenkins — Time Series Analysis: Forecasting and Control (1970)

Временные ряды: основные понятия

Тренд — долгосрочная тенденция.
Сезонность — циклические изменения с постоянным периодом.
Ошибки — непрогнозируемые изменения без конкретного паттерна.



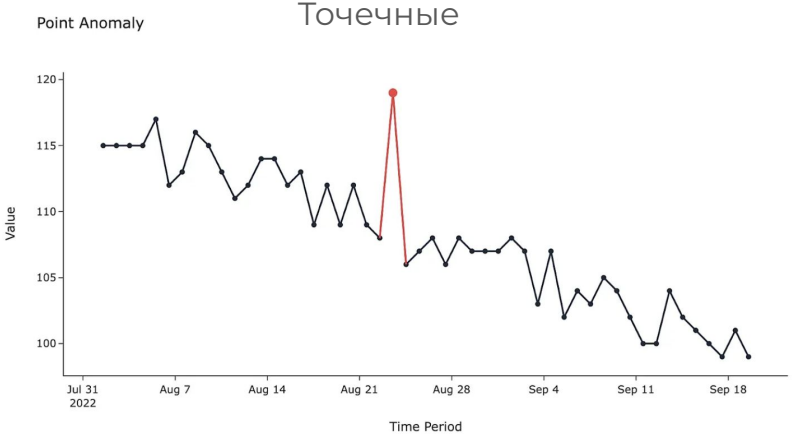
Cleveland, Cleveland, McRae & Terpenning — STL: A Seasonal-Trend Decomposition Procedure Based on Loess (1990)
Box & Jenkins — Time Series Analysis: Forecasting and Control (1970)
источник изображения

Временные ряды: основные понятия

Аномалия — необычное, несвойственное поведение временного ряда.

Бывают:

Точечные

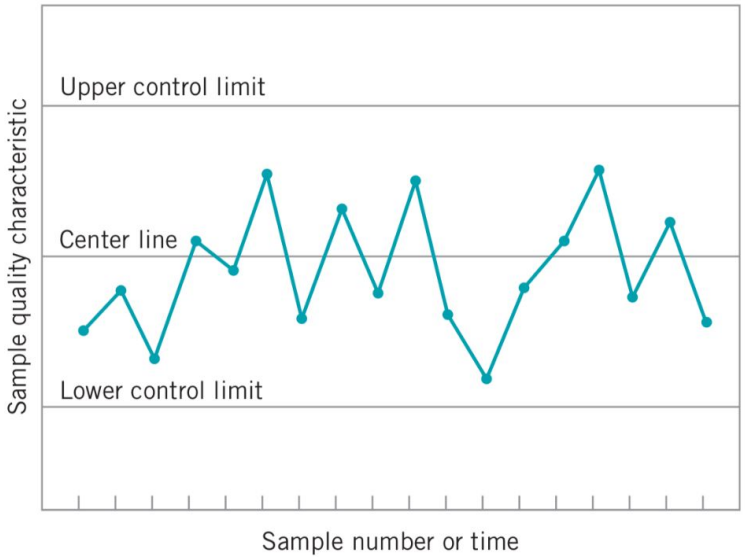


Коллективные

Chandola, Banerjee & Kumar — Anomaly Detection: A Survey (2009)
источник изображений

Онлайн методы: с чего все началось?

Контрольные карты — первый формализованный инструмент онлайн-мониторинга процессов (часть Statistical Process Control).



- Center line — среднее значение наблюдаемой характеристики
- Upper control limit и Lower control limit — допустимая вариация значения вокруг среднего, в рамках которой значение не считается аномальным

Montgomery — introduction to statistical quality control (2020)

Онлайн методы: с чего все началось?

В основе контрольных карт лежит математическая статистика.

$$UCL = \mu_w + L \cdot \sigma_w$$

$$Center\ Line = \mu_w$$

$$LCL = \mu_w - L \cdot \sigma_w$$

- L=3 (3σ) — лимиты действия
- L=2 (2σ) — лимиты предупреждения

Для среднего

$$UCL = \bar{X} + 3 \cdot \frac{\sigma}{\sqrt{n}}$$

$$Center\ Line = \bar{X}$$

$$LCL = \bar{X} - 3 \cdot \frac{\sigma}{\sqrt{n}}$$

Для доли

$$UCL = p + 3 \cdot \sqrt{\frac{p(1-p)}{n}}$$

$$Center\ Line = p$$

$$LCL = p - 3 \cdot \sqrt{\frac{p(1-p)}{n}}$$

Montgomery — introduction to statistical quality control (2020)

Онлайн методы: с чего все началось?

Контрольные карты для детекции накопленных сдвигов.

CUMSUM

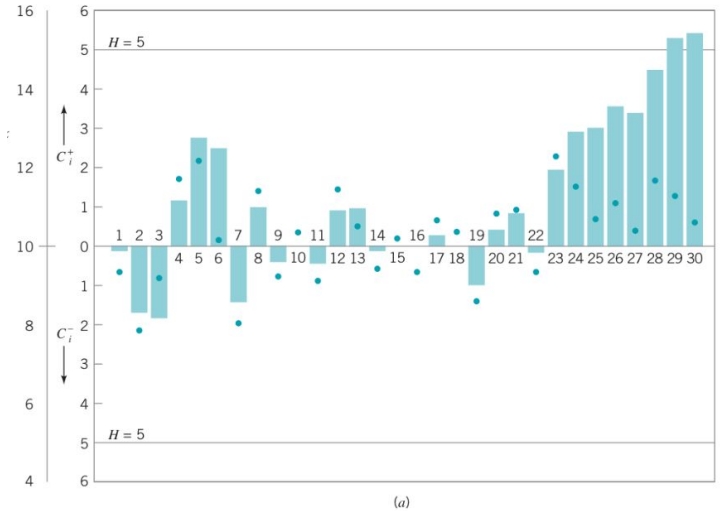
Для положительных и отрицательных сдвигов рассчитываются две статистики:

$$C_t^+ = \max\left\{0,\ C_{t-1}^+ + (x_t - \mu - k)\right\}$$

$$C_t^- = \max\left\{0,\ C_{t-1}^- + (\mu - x_t - k)\right\}$$

где:

- x_t — наблюдаемое значение в момент времени t
- μ — ожидаемое среднее
- k — референсное значение минимального отклонения, который хотим обнаружить



Сигнал тревоги $C_t^+ \geq h$ или $C_t^- \geq h$

NIST — CUMSUM Control Charts. Engineering Statistics Handbook
Montgomery — introduction to statistical quality control (2020)

Онлайн методы: с чего все началось?

Контрольные карты для детекции накопленных сдвигов.

EWMA (Exponentially Weighted Moving Average)

Сглаженное значение:

$$z_t = \lambda x_t + (1-\lambda) z_{t-1}, \quad 0 < \lambda \leq 1$$

где:

- x_t — наблюдаемое значение в момент времени t
- z_t — сглаженное значение
- λ — референсное значение минимального отклонения, который хотим обнаружить

Контрольные границы:

$$UCL_t = \mu + L \cdot \sigma \sqrt{\frac{\lambda}{2-\lambda}\left(1-(1-\lambda)^{2t}\right)}$$

$$LCL_t = \mu - L \cdot \sigma \sqrt{\frac{\lambda}{2-\lambda}\left(1-(1-\lambda)^{2t}\right)}$$

При больших t — асимптотическое приближение:

$$\sigma_z = \sigma \sqrt{\frac{\lambda}{2-\lambda}}$$

NIST — EWMA Control Charts. Engineering Statistics Handbook
Montgomery — introduction to statistical quality control (2020)

Онлайн методы: сегодня

Детекция аномалий на основе 3σ подхода не теряет актуальности:

- интерпретируемая;
- легко реализовать;
- не сложно поддерживать.

Для средних

$$\mu_t = \frac{1}{n}\sum_{i=t-n}^{t-1} x_i, \quad \sigma_t = \sqrt{\frac{1}{n-1}\sum_{i=t-n}^{t-1}(x_i - \mu_t)^2}$$

$$UCL_t = \mu_t + 3\sigma_t, \quad LCL_t = \mu_t - 3\sigma_t$$

По дельтам

$$\Delta_t = x_t - x_{t-1}$$

$$\mu_{\Delta,t},\ \sigma_{\Delta,t}$$

$$UCL_t = \mu_{\Delta,t} + 3\sigma_{\Delta,t}, \quad LCL_t = \mu_{\Delta,t} - 3\sigma_{\Delta,t}$$

Среднее и дисперсия рассчитываются скользящим окном за N последних дней до анализируемой даты

Методы ретроспективного анализа

Эти тесты применяются к историческим данным, когда нужно проверить, является ли отдельное значение (или несколько) статистически аномальными.

Grubbs's test

Проверяет гипотезу о том, что самая экстремальная точка в выборке является выбросом.

$$G = \max_i \frac{|x_i - \bar{x}|}{s}$$

где:

- x_i — значение в выборке
- $\bar{x}$ — среднее по выборке
- s — стандартное отклонение

Критерий проверки:

$$G_{crit} = \frac{(n-1)}{\sqrt{n}} \cdot \sqrt{\frac{t^2_{\alpha/(2n),\, n-2}}{n-2+t^2_{\alpha/(2n),\, n-2}}}$$

где:

- n — число наблюдений
- $t_{\alpha/(2n),\, n-2}$ — квантиль t-распределения
- α — уровень значимости

NIST — Grubbs' Test for Outliers. Engineering Statistics Handbook

Методы ретроспективного анализа

Что если в выборке несколько аномалий?

Generalized ESD (Extreme Studentized Deviate)
Метод для поиска до k выбросов в данных, когда заранее неизвестно их количество. На каждом шаге удаляется самая экстремальная точка и пересчитываются параметры.

$$R_j = \max_i \frac{|x_i - \bar{x}_j|}{s_j}$$

где:

- x_i — значение в выборке
- $\bar{x}_j$ — среднее по выборке без ранее удаленных значений
- s_j — стандартное отклонение по выборке без ранее удаленных значений

Критерий проверки:

$$\lambda_j = \frac{(n-j) \cdot t_{p,\, n-j-1}}{\sqrt{(n-j-1+t^2_{p,\, n-j-1})(n-j+1)}}$$

где:

- n — число наблюдений
- $t_{p,\, n-j-1}$ — квантиль t-распределения
- $p = 1 - \alpha/(2(n-j+1))$
- α — уровень значимости

NIST — Generalized ESD Test for Outliers. Engineering Statistics Handbook

Временные ряды: STL-разложение

Seasonal-Trend decomposition using Loess — представляет временной ряд как сумму трех компонентов:

$$x_t = T_t + S_t + R_t$$

где:

- T_t — тренд
- S_t — сезонность
- R_t — остатки, шум и выбросы

Часто используют, чтобы убрать тренд и сезонность и запустить тест на выбросы уже на остатках. Так работает метод S-H-ESD (Seasonal Hybrid ESD), применяемый в Twitter для детекции аномалий в метриках.

Cleveland, Cleveland, McRae & Terpenning — STL: A Seasonal-Trend Decomposition Procedure Based on Loess (1990)

Инфраструктура мониторинга и детекции

1. **Слой данных** — источник, может быть Event-трекером и DWH.
2. **Слой метрик** — единый слой вычисления продуктовых и бизнес-метрик:
   - Строгое описание формул, срезов, источников
   - Все расчеты централизованы
   - На этом же уровне рассчитываются агрегации
3. **Сервис детекции аномалий** — работает риал-тайм или батчами, применяет методы детекции, возвращает рассчитанные результаты и флаги аномалий.
4. **Хранилище** — история расчетов детекций.
5. **Алертинг** — триггерная система, рассылающая уведомления о найденных аномалиях в интерфейс для пользователя.