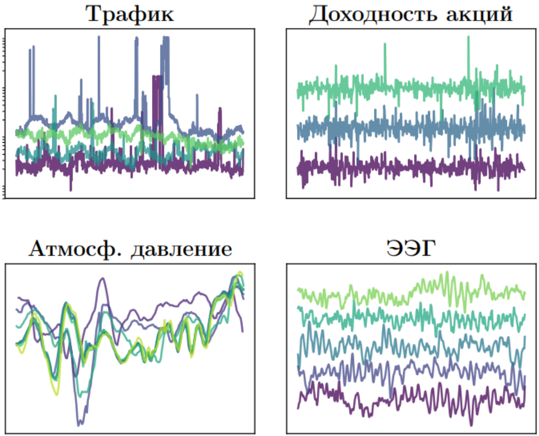
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Лекция 4. Разладки

Анализ данных с временной структурой

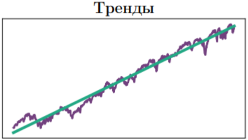
Что такое разладки?

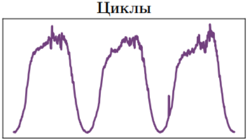
Многие реальные процессы описываются (многомерными) временными рядами

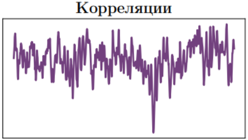


Что такое разладки?

В этих временных рядах выделяют компоненты (статистические характеристики)





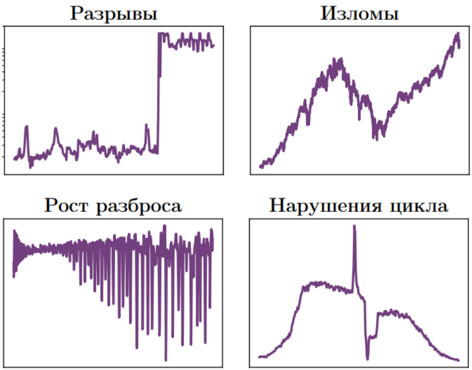


Описание, **оценивание**, **выводы** о наблюдаемых временных рядах: теория и статистика случайных процессов

- фильтрация, сегментация, шумоподавление, анализ трендов, корреляционный, дисперсионный, регрессионный, морфологический анализ, ...

Что такое разладки?

Разладка: изменение статистических свойств ряда



Задача «о разладке»: выявить возникающее изменение

Насколько это важно? Примеры приложений I

- Обнаружение внедрений в компьютерные сети (атак, ведущих к изменению объема передаваемого трафика) [Kim и др. 2004; Alexander G Tartakovsky 2003; Alexander G. Tartakovsky и др. 2006]
- Обнаружение аномалий в сетях передачи данных (видеопотоки в системах видеонаблюдения, сетевой трафик и др.) [Casas и др. 2010; Lakhina, Crovella и Christiphe Diot 2004; Lakhina, Crovella и Christophe Diot 2004; Pham и др. 2014; A. Tartakovsky 2013]
- Обнаружение и изоляция отказов узлов систем управления транспортными средствами [Malladi и др. 1999; Willsky 1976]
- Мониторинг целостности системы геопозиционирования [Basseville и др. 2002] I

Насколько это важно? Примеры приложений II

- Обнаружение изменений структуры породы при бурении скважин [Adams и др. 2007]
- Обнаружение начала рецессии или экономического роста [Andersson и др. 2002]
- Обнаружение изменений волатильности индекса Dow Jones [Adams и др. 2007]
- Обнаружение сигнала при наблюдении подводных целей [Streit и др. 1999]
- Автоматическое обнаружение аномального человеческого поведения при видеонаблюдении [Pham и др. 2014]
- Автоматический контроль качества выпускаемой продукции [Ben-Gal и др. 2003; Girshick, Meyer A and Rubin 1952; Shewhart 1931]

Насколько это важно? Примеры приложений III

- Мониторинг и анализ смертности и заболеваемости раком легких [Dass 2009; Dass и др. 2011; Taweab и др. 2015] I Обнаружение возникновения эпидемий [MacNeill и др. 1995]
- Обнаружение аритмии (внезапных изменений ритма биения) сердца [Willsky 1976]
- Предсказание транзиторных ишемических атак (преходящих нарушений мозгового кровообращения) [Cerutti S. и др. 1993]
- Диагностика задержки внутриутробного роста [Petzold и др. 2004]
- Анализ несчастных случаев на угольных шахтах [Adams и др. 2007]
- Мониторинг уровня хлора в питьевой воде [Guépié и др. 2012]

Насколько это важно? Объем публикуемой литературы

Поиск в системе индексации Google Scholar выдает, начиная с 2000 года:

- change point detection — 10 200 статей
- anomaly detection — 53 900 статей
- break detection — 3 980 статей
- обнаружение разладок, обнаружение аномалий, обнаружение изменений — 765 статей

Первые работы по разладкам: 1931 год, W. A. Shewhart (цель — контроль качества выпускаемой продукции).

Математический формализм в задачах о разладке

Структура наблюдаемых данных

Наблюдаемый случайный процесс $\xi = (\xi_1)_1 \geq 0$ задан на пространстве (Ω, F, P) и имеет структуру

$\xi_1,\ \xi_2, \ldots,\ \xi_{\theta-1},\ \xi_\theta,\ \xi_{\theta+1}, \ldots$ $\underbrace{\xi_1, \xi_2, \ldots, \xi_{\theta-1},}_{\text{до разладки } P=P_\infty}$ $\underbrace{\xi_\theta,\ \xi_{\theta+1}}_{\text{до разладки } P=P_0}$

и описывается данными $X = (X_1)_1 \geq 0$

$X_1,\ X_2, \ldots,\ X_{\theta-1},\ X_\theta,\ X_{\theta+1}, \ldots$ $X_1,\ X_2, \ldots,\ X_{\theta-1},$ $X_\theta, X_{\theta+1}$

θ — момент появления **разладки**, который требуется оценить по данным

Классическая модель в непрерывном времени:

$$\xi_t = \mu \mathbb{I}_{\{t \geq \theta\}}(t) + W_t, \quad W = (W_t)_{t \geq 0} - \text{БД}$$

Роль распределений процесса P_∞, P_0 и P_θ

Наблюдаемый случайный процесс $\xi = (\xi_1)_{1 \geq 0}$ имеет структуру

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1} \quad \xi_\theta, \xi_{\theta+1}, \ldots$$

P_∞: распределение ξ в предположении, что разладка не появляется никогда ($\theta = \infty$)

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \xi_\theta, \xi_{\theta+1}, \ldots \quad \sim \quad P_\infty$$

P_0: распределение ξ в предположении, что разладка произошла в момент старта наблюдений ($\theta = 0$)

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1}, \xi_\theta, \xi_{\theta+1}, \ldots \quad \sim \quad P_\infty$$

У мер P_∞ и P_0 есть плотности $f_\infty(\cdot)$ и $f_0(\cdot)$ и соответствующие матожидания E_∞ и E_0

Роль распределений процесса P_∞, P_0 и P_θ

Наблюдаемый случайный процесс $\xi = (\xi_t)_t \geq 0$ имеет структуру

$$\xi_1, \xi_2, \ldots, \xi_{\theta-1} \quad \xi_\theta, \xi_{\theta+1}, \ldots$$

P_θ: распределение ξ в предположении, что разладка произошла в момент θ

Плотность $f_\theta^n(x) \in \mathbb{R}^n$ меры P_∞ имеет специальный вид

$$f_\theta^n(X_1, \ldots, X_n) = f_\infty(X_1, \ldots, X_{\theta-1}) \cdot f_0(X_\theta, \ldots, X_n)$$

Пример: бракованные изделия

- $\xi_1, \xi_2, \ldots$: длина выпускаемых изделий, $\xi_i \in \mathbb{R}_+$
- Нормальный ход индустриального процесса:
  $\xi_1, \xi_2, \ldots : -\text{i.i.d.}, \xi_i \sim \mathcal{N}(\mu_\infty, \sigma^2), (\theta = \infty)$
- Изначально производятся бракованные изделия:
  $\xi_1, \xi_2, \ldots : -\text{i.i.d.}, \xi_i \sim \mathcal{N}(\mu_\infty, \sigma^2), (\theta = 0)$
- Типичный случай: сначала имеет место нормальный ход, но в момент θ наступает «сбой» («разладка»):

$$\underbrace{\xi_1, \xi_2, \ldots, \xi_{\theta-1}}_{N(\mu_\infty, \sigma^2),} \quad \underbrace{\xi_\theta,\ \xi_{\theta+1}, , \ldots}_{N(\mu_\infty, \sigma^2)}$$

Обнаружение разладки: момент подачи сигнала тревоги

Пусть до момента времени n доступны наблюдения

$$\mathbf{X}_n = (X_1, \ldots, X_n)$$

Требуется подать *сигнал тревоги* в момент $\tau = n$, если есть доказательства появления разладки

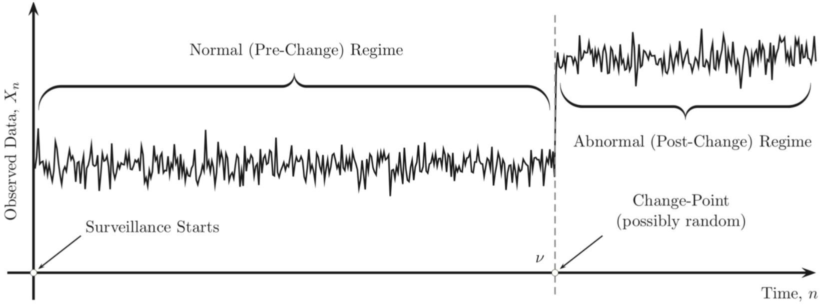
Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$

$$\tau \in \{0, 1, \ldots, \infty\}, \qquad \{\mathbf{X}_n : \tau(\mathbf{X}_n) = n\} \in \sigma(\mathbf{X}_n)$$

Используется лишь накопленная к *настоящему времени* информация (не используется будущее)

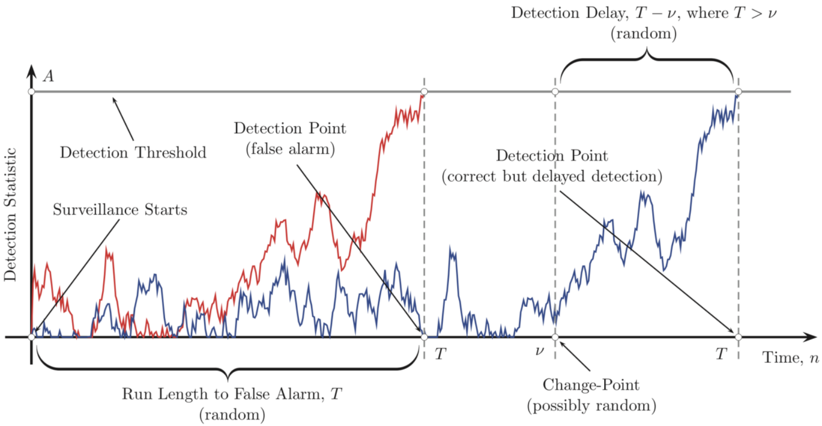
Момент остановки τ строится на основе некоторой статистики наблюдений $\gamma = (\gamma_n)_{t \geq 0}, \gamma_n = \gamma_n(\mathbf{X}_n)$

Типичный сценарий обнаружения разладки



Изображение: Polunchenko, Alexey S., and Alexander G. Tartakovsky. "State-of-the-art in sequential change-point detection."Methodology and computing in applied probability 14.3 (2012): 649-684.

Типичный сценарий обнаружения разладки, $\upsilon = \theta, T = \tau$



Изображение: Polunchenko, Alexey S., and Alexander G. Tartakovsky. "State-of-the-art in sequential change-point detection."Methodology and computing in applied probability 14.3 (2012): 649-684.

Высококачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$
- $E_\infty \tau$: среднее время до ложной тревоги (false detection delay, FDD(τ))
- **Хорошо** : $E_\infty \tau \to \infty$ (редкие ложные тревоги)
- $E_0 \tau$ или $E_\infty[\tau - \theta | \tau > \theta]$: средняя задержка в обнаружении (average detection delay, ADD(τ))
- **Хорошо** : $E_0 \tau \to 0$ (быстрое обнаружение)

Некачественный момент $\tau(\mathbf{X}_n)$

- Момент остановки: статистика $\tau = \tau(\mathbf{X}_n)$
- $P_\theta(\tau < \theta)$: вероятность ложной тревоги (probability of false alarm, $\mathrm{PFA}(\tau)$)
- **Плохо** : $\mathrm{P}_\theta(\tau < \theta) \to 1$ (частые ложные тревоги)
- $E_0\tau$ или $E_\theta[\tau - \theta | \tau > \theta]$: средняя задержка в обнаружении (average detection delay, $\mathrm{ADD}(\tau)$)
- **Плохо** : $\mathrm{E}_\theta(\tau < \theta) \to \infty$ (медленное обнаружение)
- **В практике**: задержка срабатывания $\mathrm{ADD}(\tau)$ и время без ложных тревог $\mathrm{FDD}(\tau)$ — **конфликтующие критерии** (строится зависимость одного от другого)

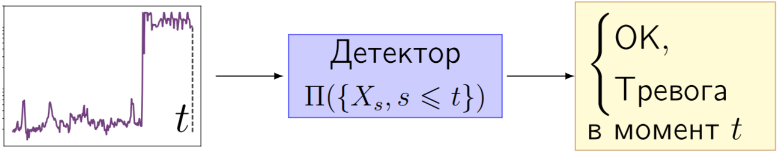
«Наивные» методы обнаружения разладки

Подходы к обнаружению разладки временного ряда

Некоторые процедуры обнаружения разладки, использующие частные особенности данных:

- Контрольные карты [Shewhart 1931]
- Алгоритм кумулятивных сумм (CUSUM) [Page 1954]
- Экспоненциально взвешенное скользящее среднее [Roberts 1959]
- Фильтр Калмана [Kalman 1960]
- Байесовские методы [Girshick, Meyer A and Rubin 1952; А. Ширяев 1961]
- Процедура Ширяева-Робертса [Roberts 1966; Альберт Николаевич Ширяев 1961]
- Метод обобщенного отношения правдоподобия [Willsky 1976]
- Методы на основе контекстных деревьев [Ben-Gal и др. 2003]
- Ядерные методы [Desobry и др. 2005]
- Энтропийный подход [Дарховский 2013]

Построение процедур обнаружения разладки



Π: детектор, процедура обнаружения (должен учитывать некоторые предположения о модели сигнала и разладки)

Π: момент тревоги $\tau = \inf\{t \geq 0 : \gamma_t \geq h\}$

Общие обозначения

Пусть $X_1, \dots, X_n$ – наблюдения, доступные до момента времени n

Основная статистика – отношение правдоподобия

$$L_n = \frac{f_0(X_1, \dots, X_n)}{f_\infty(X_1, \dots, X_n)}$$

Удобно также использовать статистику $Z_n = \log L_n$

Если наблюдения $X_1, \dots, X_n$ независимы:

$$L_n = \prod_{k=1}^{n} \frac{f_0(X_k)}{f_\infty(X_k)} = \prod_{k=1}^{n} l_k, \quad Z_n = \sum_{k=1}^{n} \log \frac{f_0(X_k)}{f_\infty(X_k)} = \sum_{k=1}^{n} \zeta_k$$

Простой пример

Пусть $\xi_1, \ldots, \xi_n$ – нормальные i.i.d.r.v., причем
$f_0(x) = \frac{1}{\sqrt{2}\pi\sigma} e^{-\frac{(x-r_0)^2}{2\sigma^2}}$ $\qquad f_\infty(x) = \frac{1}{\sqrt{2}\pi\sigma} e^{-\frac{(x-r_\infty)^2}{2\sigma^2}}$
Тогда правдоподобие выборки $X_1, \ldots, X_n$

$$L_n = \prod_{k=1}^{n} exp\left\{ \frac{r_\infty - r_0}{\sigma^2} \left[X_k - \frac{r_0 + r_\infty}{2} \right] \right\},$$

а его логарифм —

$$Z_n = \frac{r_\infty - r_0}{\sigma^2} \left[\overline{X_n} - \frac{r_0 + r_\infty}{2} n \right]$$

Контрольные карты Шухарта

Наблюдения $X_1, X_1, \ldots$ разбиваются на группы (батчи) размера N (N — параметр алгоритма)

Для каждой группы $\mathbf{X}_K = (X_{N\cdot(K-1)+1}, \ldots, X_{N\cdot K}),\ K = 1, 2 \ldots$ подсчитывается логарифм правдоподобия:

$$S_i^k = \sum_{j=i}^{k} \zeta_j$$

Момент остановки — первый момент выхода статистики

$S_{N\cdot(K-1)+1}^{N\cdot K}$ на заданный уровень h:

$$\tau_{SH} = N \cdot min\{K : S_{N\cdot(K-1)+1}^{N\cdot K} \geq h\}$$

Экспоненциально взвешенное скользящее среднее

Задается рекурсивная оценка среднего значения

$$\hat{m}_k = (1-\lambda)\hat{m}_{k-1} + \lambda X_k, \quad k = 1, 2 \ldots,$$

где λ («вес» новых данных) — параметр алгоритма.

Момент остановки — момент первого выхода статистики $\hat{m}_k$ на заданный уровень h:

$$\tau_{EWMA} = min\{k \geq 1 : \hat{m}_k \geq h\}$$

«Оптимальные» методы обнаружения разладки

Кумулятивные суммы

- О параметре θ не делается никаких предположений
- Фиксируется $T > 0$ и задается класс

$$\mathcal{M}_T = \{\tau : E_{\infty\tau} \geq T\}$$

  тех моментов остановки, для которых среднее время до ложной тревоги не меньше T
- Качество алгоритма задается величиной

$$\mathbf{D}(T) = \sup_{\theta \geq 0} \operatorname{ess\,sup}_{\omega} \left((\tau - \theta)^+ \mid \mathcal{F}_\theta \right) (\omega) \mathbf{D}(T)$$

Кумулятивные суммы

- Вводятся статистики

$$\gamma_n = \sup_{\theta \geq 0} \frac{f_0(X_1, \ldots, X_n)}{f_\infty(X_1, \ldots, X_n)} \qquad \text{и} \qquad T_n = \log \gamma_n$$

- Если случайные величины $\xi_1, \ldots, \xi_n$ независимы, то

$$\gamma_n = \max\left\{1, \max_{1 \leq \theta \leq n} \prod_{k=\theta}^{n} \frac{f_0(X_k)}{f_\infty(X_k)}\right\},$$

$$T_n = \max\left\{0, \max_{1 \leq \theta \leq n} \sum_{k=\theta}^{n} \log \frac{f_0(X_k)}{f_\infty(X_k)}\right\} =$$

$$= \max\left\{0, \max_{1 \leq \theta \leq n} \sum_{k=\theta}^{n} \zeta_k\right\}$$

Кумулятивные суммы

- Статистика T_n обладает свойством $T_n = \max\left(0, T_{n-1} + \zeta_k\right)$ и называется статистикой кумулятивных сумм (CUmulative SUMs, CUSUM).
- Остановка в момент τ_{CUSUM} минимизирует величину $\mathbf{D}(T)$

$$\tau_{CUSUM} = \inf\{n \geq 0 : T_n \geq h\}$$

Статистика Ширяева-Робертса

Разработана независимо А.Н.Ширяевым (Альберт Николаевич Ширяев 1961, 1963) и S.W.Roberts (Roberts 1966)

Условия для моментов остановки и параметра θ в точности совпадают с условиями для процедуры кумулятивных сумм

Качество алгоритма задается величиной

$$\mathbf{C}(T) = \sup_{\theta \geq 0} \mathbb{E}_\theta \left((\tau - \theta)^+ \mid \tau \geq \theta \right) \mathbf{C}(T)$$

Статистика Ширяева-Робертса

Вводится статистика $R_n = \sum\limits_{\theta=1}^{n} \frac{f_0(X_1,\ldots,X_n)}{f_\infty(X_1,\ldots,X_n)}$

Если случайные величины $\xi_1, \ldots, \xi_n$ независимы, то

$$R_n = \sum_{\theta=1}^{n} \prod_{k=1}^{n} \frac{f_0(X_k)}{f_\infty(X_k)} = \sum_{\theta=1}^{n} \prod_{k=1}^{n} l_k.$$

Статистика R_n обладает свойством $R_n = (1 + R_{n-1}) l_k$ и называется статистикой Ширяева-Робертса (Shiryaev-Roberts, SR).

Остановка в момент τ_{SR} минимизирует $\mathbf{C}(T)$

$$\tau_{SR} = \inf\{n \geq 0 : R_n \geq h\}$$

Статистика апостериорной вероятности π_t I

Пусть $\theta = \theta(\omega)$ — случайная величина, $\theta \perp Z_t$, имеющая (геометрическое) распределение

$$P(\theta = 0) = \pi, \quad P(\theta = n \mid \theta > 0) = pq^{n-1}$$

причем $\pi \in [0, 1)$ и $p \in (0, 1)$ известны, $q = 1 - p$.

Фиксируется некоторое $\alpha \in (0, 1]$ и задается класс

$$\mathcal{M}_\alpha = \{\tau : P(\tau < \theta) \le \alpha\}$$

моментов остановки, для которых вероятность ложной тревоги не выше α.

Статистика апостериорной вероятности π_t II

Качество алгоритма задается величиной

$$E(\tau - \theta \mid \tau > \theta) \quad \sim \quad \inf_{\tau \in \mathcal{M}_\alpha}$$

Этот критерий эквивалентен безусловному

$$\mathbf{A}(c) = \underbrace{P(\tau < \theta)}_{\text{вероятность ложной тревоги}} +$$

$$+ \underbrace{cE(\tau - \theta \mid \tau > \theta)}_{\text{(условное) среднее время обнаружения разладки}} \quad \sim \quad \inf_{\tau}$$

Статистика апостериорной вероятности

Вводится статистика

$$\pi_n = P(\theta \leq n | X_1, \dots, X_n)$$

представимая в виде

$$\pi_n = \frac{\varphi_n}{1 - \varphi_n}, \quad \varphi_{n+1} = \frac{p + \varphi_n}{q} l_k$$

Остановка в момент τ_π минимизирует $\mathbf{A}(c)$:

$$\tau_\pi = \inf\{n \geq 0 : \tau_\pi \geq h\}$$

π_n — апостериорная вероятность появления разладки до момента времени n в предположении, что получены наблюдения $X_1, \dots, X_n$.

Использованные источники

1. Дарховский Б.С. Обнаружение разладки случайной последовательности при минимальной априорной информации.: Теория вероятностей и ее применения 58.3. – 2013. – с. 585—590.

2. Ширяев А.Н. Задача скорейшего обнаружения нарушения стационарного режима.: Докл. АН СССР. Т. 138. 5. – 1961. – с. 1039—1042.

3. Ширяев А.Н. Обнаружение спонтанно возникающих эффектов». В: Докл. АН СССР. Т. 138. 4. – 1961. – с. 799—801.

4. Basseville M., Nikiforov I.V. Detection of abrupt changes: theory and application. – 1993.

5. Kalman R.E. A new approach to linear filtering and prediction problems.: Journal of Fluids Engineering 82.1. – 1960. – pp. 35-45.

6. Desobry F., Davy M., Doncarli C. An online kernel change detection algorithm.: IEEE Transactions on Signal Processing 53.8. – 2005. – с. 2961—2974.

Использованные источники

7. Dass, Sarat C, Chae Young Lim, Tapabrata Maiti. Change Point Analysis of Cancer Mortality Rates for US States using Functional Dirichlet Processes. – 2011.

8. Dass, Sarat C. Hierarchical Spatial Regression Models for Change Point Analysis». – 2009. – с. 136-144.

9. Cerutti S. и др. Time variant power spectrum analysis for the detection of transient episodes in HRV signal.: IEEE Transactions. – 1993. – с. 136-144. on biomedical engineering 40.2, с. 136—144.

10. Irad Ben-Gal, Morag G., Shmilovici A. Context-Based Statistical Process Control.: Technometrics 45.4. – 2003. – с. 293—311.

11. Hongjoong K., Rozovskii B.L., Tartakovsky A.G. A Nonparametric Multichart CUSUM Test for Rapid Detection of DOS Attacks in Computer Networks. – 2004. – с. 149—158.

Использованные источники

12. Casas P. и др. Optimal volume anomaly detection and isolation in large-scale IP networks using coarse-grained measurements.: Computer Networks 54.11 – 2010. – с. 1750-1766.

13. Roberts S.W. Control chart tests based on geometric moving averages.: Technometrics 1.3. – 1959. – с. 239-250.

14. Pham, Duc-Son и др. Anomaly detection in large-scale data stream networks.: Data Mining and Knowledge Discovery 28.1. – 2014. – с. 145—189.

15. Malladi D.P., Speyer J.L. A generalized Shiryayev sequential probability ratio test for change detection and isolation.: Automatic Control, IEEE Transactions. – 1999. – с. 1522—1534.

16. MacNeill I.B., Mao Y. Change-point analysis for mortality and morbidity rate.: Applied Change Point Problems in Statistics. – 1995. – с. 37—55

Использованные источники

17. Petzold, Max и др. Surveillance in longitudinal models: Detection of intrauterine growth restriction.: Biometrics 60.4. – 2004. – c. 1025—1033.

18. Lakhina A., Crovella M., Diot C. Characterization of network-wide anomalies in traffic flows.: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement. – 2004. – c. 201. URL: http://portal.acm.org/citation.cfm?doid=1028788.1028813.

19. Lakhina A., Crovella M., Diot C. Diagnosing network-wide traffic anomalies.: ACM SIGCOMM Computer Communication Review 34.4. – 2004. – c. 219.

20. Basseville M., Nikiforov I. Fault isolation for diagnosis: nuisanece rejection and multiple hypothesis testing.: Annual Reviews in Control 26. – 2002. – c. 189-202.

21. Roberts S.W. A comparison of some control chart procedures».: Technometrics 8.3. – 1966. – c. 411—430.

Использованные источники

22. Adams, Ryan Prescott, David J. C. MacKay. Bayesian Online Changepoint Detection. – 2007. – с. 7. URL: http://arxiv.org/abs/0710.3742.

23. Shewhart W.A. Economic control of quality of manufactured product. – 1931.

24. Andersson, Eva, David Bock, Marianne Frisen. Department of Statistics Goteborg University Sweden with application to turns in business cycles. – 2002.

25. Tartakovsky A.G. Quickest Change Detection in Distributed Sensor Systems.: Proceedings of the 6th International Conference on Information Fusion. – 2003. – с. 756—763.

26. Tartakovsky A.G. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods.: IEEE Transactions on Signal Processing 54.9. – 2006. – с. 3372—3381.

Использованные источники

27. Streit, Roy L., Peter K. Willett. Detection of random transient signals via hyperparameter estimation.: IEEE Transactions on Signal Processing 47.7. – 1999. – с. 1823—1834.

28. Tartakovsky A.G. Efficient computer network anomaly detection by changepoint detection methods.: IEEE Journal of Selected Topics in Signal Processing 7.1. – 2013. – с. 7—11.

29. Taweab, Fauzia, Noor Akma Ibrahim, Jayanthi Arasan. A Bounded Cumulative Hazard Model with A change-Point According to a Threshold in a covariate for Right-Censored Data». В: 74.1. – 2015. – с. 69—74.

30. Willsky A.S. A survey of design methods for failure detection in dynamic systems.: Automatica 12. – 1976. – с. 601—611. URL: `http://www.sciencedirect.com/science/article/pii/0005109876900418`.