ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 26 за 2020 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Кислицын А.А.

Программный комплекс для анализа статистики согласованного уровня стационарности временных рядов

Рекомендуемая форма библиографической ссылки: Кислицын А.А. Программный комплекс для анализа статистики согласованного уровня стационарности временных рядов // Препринты ИПМ им. М.В.Келдыша. 2020. № 26. 22 с. http://doi.org/10.20948/prepr-2020-26
URL: http://library.keldysh.ru/preprint.asp?id=2020-26

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В.Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

А.А. Кислицын

Программный комплекс для анализа статистики согласованного уровня стационарности временных рядов

Москва — 2020

Кислицын А.А.

Программный комплекс для анализа статистики согласованного уровня стационарности временных рядов

В работе построен алгоритм вычисления многомерного индикатора разладки в нестационарном временном ряде, доказаны теоремы о свойствах его аппроксимации, позволяющие значительно сократить время счета, и построен программный комплекс с интерфейсом, реализующий предложенную методику. Метод построения эквивалентной по Чернову полугруппы был применен для обоснования алгоритма приближенного вычисления индикатора разладки для предсказания приступа эпилепсии по данным электроэнцефалограмм.

Ключевые слова: нестационарный временной ряд, индикатор разладки, встык-выборка, согласованный уровень стационарности

Kislitsyn A.A.

Software complex for statistical analysis of consistent stationary level of time-series

In this paper the numerical method of calculation of multidimensional disorder indicator for non-stationary time-series is constructed. The theorem of approximation for this indicator is proved. The program software for statistical analysis of consistent stationary level is constructed. The partial example of disorder as an epilepsy predictor of electroencephalogram data is presented.

Key words: non-stationary time-series, disorder indicator, joining sample, consistent stationary level

Работа выполнена при поддержке гранта РНФ № 19-71-30004.

Содержание

1. Введение

Задача разработки высокопроизводительных численных алгоритмов и программных комплексов для обработки больших массивов нестационарных данных, возникающих во многих областях практической деятельности, в настоящее время приобрела большое практическое значение в связи с развившимися возможностями вычислительной техники и значительно увеличившейся статистической детализацией описания самих процессов. Одной из актуальных задач является проблема статистического распознавания разладки в нестационарных рядах. Для стационарных случайных процессов эта задача имеет решение в виде классических критериев – либо параметрических, когда оцениваются доверительные интервалы принадлежности параметра, характеризующего класс распределений наблюдаемой величины, либо непараметрических, основанных на близости выборочных функций распределения в различных нормах.

Существует много примеров временных рядов, в которых одним из инструментов анализа является модель разладки с учетом нестационарных свойств ряда. Это биржевые ряды цен сделок на финансовые инструменты, а также ряды данных о доходности соответствующих торговых операций, совершенных в рамках определенных стратегий (алгоритмов), где потоки событий составляют до нескольких десятков в секунду по отдельному инструменту. Другим примером являются потоки событий, связанные с анализом трафика беспроводной связи, достигающие десятков тысяч событий в секунду. Менее мощные, но тоже достаточно большие потоки данных требуется обрабатывать при анализе различной биометрической информации о состоянии живых организмов, а также телеметрической информации о состоянии технических систем. Анализ данных электроэнцефалограмм в медицине, статистика показаний сейсмограмм и различных счетчиков радиоактивности, последовательности символов в текстах различной природы представляют типичный набор задач, в которых требуется отнести результат наблюдения к определенным классам «норма» или «авария». Практически важной задачей является определение промежуточного состояния, которое, собственно, и представляет собой разладку «нормы», чтобы успеть принять меры по недопущению состояния «авария». Модель разладки должна при этом давать вероятностную картину перехода из одного состояния в другое, причем ошибка второго рода (пропуск цели) должна быть минимальна при заданном уровне ошибки первого рода (ложная тревога).

Для стационарных рядов, функции распределения которых известны, вероятности ошибок при принятии или отклонении статистических гипотез могут быть вычислены теоретически. Однако если ряд нестационарный и его выборочные распределения не принадлежат определенному классу, оценить ошибку затруднительно. Тем не менее, решения в таких случаях принимаются на основе классических критериев, только уровень значимости постфактум

оказывается хуже, чем утверждается критерием. Следовательно, необходимо скорректировать уровень значимости с учетом нестационарности изучаемой системы.

В настоящей работе описывается программный комплекс для расчета непараметрического индикатора разладки, называемого согласованным уровнем стационарности, и для анализа статистики этого индикатора, которая может служить обоснованием предлагаемого критерия разладки в нестационарном временном ряде.

Основу теоретического анализа составляют методы исследования асимптотических свойств стационарной точки уровня значимости, на котором принимается гипотеза об однородности двух выборочных распределений. Исходной статистикой для этого анализа является выборочная функция распределения расстояний между выборочными функциями распределения так называемых встык-выборок изучаемого временного ряда.

В статистическом смысле классической разладкой называется изменение одного стационарного распределения случайной величины на другое (тоже стационарное), произошедшее в случайный момент времени. В этом случае задачей математической статистики является обнаружение соответствующего момента с наименьшей погрешностью на основе анализа наблюдаемых данных.

Одним из основных методов обнаружения разладки является метод кумулятивных сумм (см. работы [1-4]), который основывается на методе максимального правдоподобия.

Среди других стандартных подходов к анализу разладки в стационарных процессах отметим также байесовские подходы [5, 6] и процедуры на основе методов разложения многомерных данных, такие как анализ главных компонент [7-10] и анализ сингулярного спектра [11-14].

Однако, как было показано в [7, 15, 16], использование традиционных методов и процедур для обнаружения разладок в системах контроля с интенсивным программным обеспечением часто неэффективно по причине нестационарности потоков данных. Необходимость включения в стандартные пакеты новых алгоритмов, позволяющих повысить точность статистических оценок при проверке вероятностных гипотез, отмечалась в работах [17-19].

Таким образом, существует проблема создания индикаторов, пригодных именно для нестационарных временных рядов.

В настоящей работе описывается вычислительный алгоритм для построения предиктора разладки достаточно общего вида, т.е. он может быть применен к процессам различной физической или экономической природы. Нетривиальным фактом является обнаруженная эмпирически стационарность распределения значений этого идикатора, в частности, для рядов электроэнцефалограмм, которые представляют практический пример применения разработанного метода.

2. Согласованный уровень стационарности

В стационарном случае задача о принадлежности двух выборок длины N одной генеральной совокупности решается непараметрической статистикой Колмогорова-Смирнова:

$$S_N = \sup_x \left| F_{1,N}(x) - F_{2,N}(x) \right|. \quad (1)$$

Для статистики (1) имеет место асимптотика

$$\lim_{N\to\infty} P\left\{ 0 < \sqrt{\frac{N}{2}}\, S_N < z \right\} = K(z), \quad (2)$$

где $K(z)$ есть табулированная функция Колмогорова [20, 21]:

$$K(z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}, \quad z > 0. \quad (3)$$

На практике формула (2) применяется следующим образом. Задавая уровень значимости α для функции распределения статистики (1), требуется вычислить соответствующий α-квантиль, зависящий от длины выборки N. Обозначим его $\varepsilon_N(\alpha)$. Эта величина имеет следующее асимптотическое представление [22]:

$$\varepsilon_N(\alpha) = \sqrt{\frac{z}{2N}} - \frac{1}{6N} + o(1/N), \quad (4)$$

где z есть корень уравнения

$$1 - K\left(\sqrt{\frac{z}{2}} \right) = \alpha. \quad (5)$$

Однако имеется некоторая неопределенность такого статистического вывода. Она состоит в том, что надо априори задать желаемый уровень малости критерия $1 - K(z)$. Но для этого надо знать, как часто две независимых выборки длины N отличаются в смысле (1) больше, чем на некоторое число ε, такое, что $1 - K\left(\sqrt{\frac{N}{2}}\varepsilon \right) < \alpha$.

Для вывода согласованного с экспериментом уровня значимости в (5) заметим следующее. В экспериментах по сравнению выборок случайной величиной является расстояние между парой выборок. В частности, такой величиной является квантиль функции распределения этих расстояний, который следует выбрать в качестве нужного уровня значимости. Как известно (см., напр., [23]), если случайная величина ξ (здесь это расстояние между выборками) имеет функцию распределения $F(x)$ (для расстояний между выборками асимптотически это есть функция Колмогорова), то случайная величина $\eta = F(\xi)$ (здесь это уровень значимости) имеет равномерное распределение на $[0;1]$. Следовательно, согласованный с экспериментом уровень значимости α есть функция, линейно зависящая от расстояния ε

между выборками. Поскольку же в норме C это расстояние меняется от нуля до единицы, то следует положить $\alpha = \varepsilon$. В результате получаем, что критическое расстояние $\varepsilon^*(N)$ разделения выборок на уровне значимости, согласованном в вышеописанном смысле с экспериментом, определяется из уравнения

$$1 - K\left(\sqrt{\frac{N}{2}}\varepsilon\right) = \varepsilon. \tag{6}$$

Решение уравнения (6) единственно, поскольку правая часть как функция ε монотонно возрастает от нуля до единицы, а левая монотонно убывает от единицы до нуля в соответствии со свойствами $K(z)$ как функции распределения. Найденное решение было введено в [24] и названо согласованным отклонением для стационарных выборочных распределений. Оно представляет собой стационарную точку уровня значимости критерия Колмогорова-Смирнова.

Применим данный результат к нестационарным распределениям. Построим статистику расстояний между так называемыми встык-выборками, т.е. между выборочными функциями распределения (далее ВФР) $F_N(x,t)$ и $F_N(x,t+N)$, сдвинутыми одна относительно другой на величину N окна выборки (время *t* измеряется порядковым номером элемента ряда):

$$\rho(N;t) = \left\| F_N(x,t) - F_N(x,t+N) \right\|. \tag{7}$$

Далее строится функция распределения (тоже выборочная) $G_N(\rho)$ расстояний (7), которая представляет эмпирическую вероятность того, что расстояние между распределениями не больше ρ. Определим теперь согласованный уровень стационарности (далее СУС) $\rho^*(N)$ так, что соответствующее расстояние равно значимости критерия, т.е. является решением уравнения

$$G_N(\rho) = 1 - \rho. \tag{8}$$

В стационарном случае уравнение (8) переходит в уравнение (6), поскольку тогда функция распределения $G_N(\rho)$ переходит в функцию из критерия Колмогорова-Смирнова.

Индекс нестационарности временного ряда определяется как отношение решения $\rho^*(N)$ уравнения (8) к стационарной точке $\varepsilon^*(N)$ уровня значимости стационарного критерия (6):

$$J(N) = \frac{\rho^*(N)}{\varepsilon^*(N)}. \tag{9}$$

Этот индекс показывает, во сколько раз доля расстояний, больших СУС, превосходит аналогичный показатель для стационарных рядов. Если $J(N) \le 1$, ряд считается стационарным, а если $J(N) > 1$, то ряд нестационарный.

3. Формулировка вычислительной задачи

Для анализа нестационарных временных рядов имеется программный комплекс [25], в котором реализован расчет СУС в зависимости от длины выборки для анализа пошаговой эволюции выборок внутри заданного одномерного фрагмента данных. Это означает, что для заданной длины встык-выборки программа выдает ровно одно число. После вариации длины выборки результатом работы программы [25] является график СУС в виде $\rho^*(N)$, где N меняется в заданных пределах в соответствии с длиной L_{tot} полного набора анализируемых данных.

Для задачи распознавания разладки с указанием вероятности ошибки первого рода (ложной тревоги) необходимо дополнить имеющийся численный алгоритм расчета СУС возможностью проводить анализ по различным фрагментам исходного ряда. Это связано с практическим применением СУС как индикатора разладки. Если ряд нестационарный, то разладка будет состоять в изменении его уровня нестационарности, каковым и является СУС.

На практике задача усложняется тем, что необходимо сравнивать СУС для встык-выборок одной и той же длины, но для разных значений L_{tot}. Дело в том, что настройка уровня $\rho^*(N)$ определяется по эталонному ряду данных достаточно большой длины L_{tot}, в котором этой разладки нет. Такой ряд экспертно отбирается на стадии обучения алгоритма. Разладка же должна выявляться оперативно, то есть окно наблюдения должно иметь длину $L << L_{tot}$, но быть тем не менее достаточно большим для того, чтобы в нем можно было выделить встык-выборки длины N в количестве, достаточном для построения распределения расстояний между ними. В этом сравнительно малом окне длины L разладка фиксируется следующим образом. Если окно длины L содержит в себе n выборок длины N, то в нем будет вычислено $n-1$ значений расстояний между встык-выборками. Каждое конкретное расстояние может быть как меньше эталонного значения $\rho^*(N)$, так и больше него. Отмечаются последовательные значения расстояний между встык-выборками, которые больше $\rho^*(N)$. Как только их количество k превзойдет число

$$K_{cr} = (n-1)\cdot\rho^*(N), \qquad (10)$$

фиксируется разладка. Формула (10) следует из определения СУС: вероятность превышения критического расстояния, т.е. доля превышающих его событий, равна самому критическому расстоянию.

Однако естественно, что если разбить исходный обучающий ряд длины L_{tot} на фрагменты длины L (пусть таких фрагментов M) и в каждом из них вычислить величину $\rho_m^*(N)$, $m=1,2,...,M$, то полученные значения $\rho_m^*(N)$ будут, вообще говоря, отличаться от величины $\rho^*(N)$ как в меньшую, так и в большую стороны, что приведет к возникновению ложной разладки – ведь в обучающем ряде ее нет. Возникают вопросы: как зависит от длины L различие

между СУС полной совокупности и СУС части совокупности? Иными словами, насколько устойчиво распределение СУС $F_L\left(\rho^*(N)\right)$, построенное по достаточно большому набору фрагментов длины L, каждый из которых играет локально роль некоторого полного набора данных L_{tot}? Как устойчивость зависит от длины L? При каких соотношениях между длинами L и N регистрация разладки имеет наименьшую ошибку? Если распределение $F_L\left(\rho^*\right)$ стационарно, то ответы на эти вопросы могут быть получены аналитически.

Следовательно, необходимо построить двухпараметрический алгоритм определения СУС, который в качестве выходных данных выдает функцию распределения случайной величины СУС на основе статистики значений последовательности СУС $\rho_m^*(N,L)$, построенной для встык-выборок длины N в окнах длины L. Подчеркнем, что параметры N и L произвольны. Таким образом, Задача 1 при разработке программного комплекса формулируется как увеличение размерности блока статистического анализа СУС по сравнению с относительно простым одномерным алгоритмом [25].

Кроме того, надо предусмотреть возможность многоканального анализа, для чего требуется провести распараллеливание алгоритма. Это важно в задачах анализа высокочастотных биометрических данных, которые записываются одновременно по нескольким отведениям (вплоть до нескольких тысяч различных рядов). Следовательно, возникает Задача 2 об использовании высокопроизводительных вычислений и проведении оценки их эффективности на основе вычислительных экспериментов для конкретных прикладных задач. Эта задача естественным образом сопряжена с развитием так называемых систем с интенсивным программным обеспечением. Согласно стандарту ISO/IEC/IEEE 42010:2011(E) это системы, функциональность которых определяется в основном их программными средствами [26]. Согласно Стратегии развития отрасли информационных технологий в Российской Федерации на 2014–2020 годы и на перспективу до 2025 года такими системами являются аппаратно-программные комплексы с большим удельным весом программной части [27].

Задача 3 состоит в разработке метода сокращенных вычислений для построения статистики СУС, поскольку в условиях больших данных алгоритмы, основанные на полном переборе, неэффективны, а часто и нереализуемы за разумное время в силу естественно ограниченных возможностей вычислительной техники. В следующем разделе описывается метод вычисления двухпараметрической статистики $\rho_m^*(N,L)$ на основе вычисления $\rho_m^*(N,L_{\min})$ только для одной минимальной длины $L_{\min}$ окна наблюдения и последующего пересчета в окно произвольной длины L.

4. Эквивалентный по Чернову согласованный уровень значимости

В этом разделе метод так называемых конечнократных аппроксимаций для построения решающих операторов задачи Коши для дифференциальных уравнений в частных производных применяется к анализу уровня нестационарности выборочных функций распределения. Этот метод устанавливает связь между статистикой и динамикой, что позволяет построить математически корректную процедуру аппроксимации случайных процессов с помощью динамических систем.

Ключевым утверждением является теорема Чернова [28], устанавливающая сходимость определенного итерационного процесса к полугруппе, которая представляет решение задачи Коши. Эта теорема состоит в следующем.

Пусть X – банахово пространство, $B(X)$ – банахово пространство линейных ограниченных операторов в X и пусть функция $\mathbf{F}$: $[0;+\infty) \to B(X)$ удовлетворяет условиям $\mathbf{F}(0)=\mathbf{I}$, $\mathbf{F}(t)$ непрерывна в сильной операторной топологии и имеет место оценка $\|\mathbf{F}(t)\|_{B(X)} \le \exp(\alpha t), t \ge 0$ при некотором $\alpha \ge 0$. Тогда, если оператор $\mathbf{F}(t)$ замыкаем и его замыкание является генератором сильно непрерывной полугруппы операторов G_t, $t>0$, то для любого $u \in X$ и любого $T>0$ существует предел

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \left\| G_t u - \left(\mathbf{F}\left(\frac{t}{n}\right) \right)^n u \right\|_X = 0. \qquad (11)$$

Утверждение (11) можно распространить на сумму операторов, что позволяет трактовать полугруппу G_t как результат статистического усреднения независимых случайных полугрупп. Указанная трактовка теоремы Чернова была рассмотрена в [29], что позволило придать математический смысл усреднению полугрупп. В общем случае линейная комбинация полугрупп не обязана быть полугруппой. Но если в результате итераций, аналогичных описанным выше, можно получить сходимость к полугруппе, то такая полугруппа объявляется эквивалентной по Чернову данной линейной комбинации полугрупп. Согласно [29] операторнозначная функция $\hat{f}(t)$, действующая из некоторой правой полуокрестности нуля на числовой оси в банахово пространство $B(X)$ линейных ограниченных операторов, действующих в банаховом пространстве X, называется эквивалентной по Чернову полугруппе $\hat{g}(t)$, если $\forall T>0$ и $\forall u \in X$ выполняется условие:

$$\lim_{n\to\infty} \sup_{t\in[0;T]} \left\| \left(\left(\hat{f}(t/n)\right)^n - \hat{g}(t) \right) u \right\|_X = 0. \qquad (12)$$

Сильно непрерывная однопараметрическая полугруппа U ограниченных линейных преобразований банахова пространства X называется обобщенным средним значением случайной полугруппы ξ, если полугруппа U эквивалентна по Чернову математическому ожиданию $M\xi$.

Также в работе [29] была доказана теорема об эквивалентности по Чернову для средних значений случайных полугрупп, в которых сформулированы достаточные условия того, чтобы эти средние значения, не будучи сами полугруппами, порождали полугруппу, эквивалентную им по Чернову. Теорема об эквивалентности состоит в следующем.

Пусть $\{\hat{H}_n\}$ – последовательность генераторов сильно непрерывных полугрупп в банаховом пространстве X. Пусть также $\{p_n\}$ – весовая последовательность неотрицательных чисел, сумма ряда из которых равна единице: $\sum_{n=1}^{\infty} p_n = 1$. Пусть также существует линейное подпространство $D \subset X$, являющееся существенной областью определения каждого из генераторов $\hat{H}_n$ и такое, что для любого $g \in D$ ряд $\sum_{n=1}^{\infty} p_n \left\| \hat{H}_n g \right\|_X$ сходится. Тогда, если оператор $\hat{H}$ определен на D формулой $\hat{H}g = \sum_{n=1}^{\infty} p_n \hat{H}_n g$ и его замыкание является генератором сильно непрерывной полугруппы $\hat{U}(t) = e^{t\hat{H}}$, $t \geq 0$, то среднее значение $\mathrm{M}\hat{U}$ случайной полугруппы $\hat{U}_n(t) = e^{t\hat{H}_n}$, определяемое формулой $\mathrm{M}\hat{U} = \sum_{n=1}^{\infty} p_n \hat{U}_n$, эквивалентно по Чернову полугруппе $\hat{U}(t) = e^{t\hat{H}}$.

Описанные конструкции позволяют сформулировать единый подход к решению многих эволюционных задач. Он состоит в том, что вместо поиска точного решения ищется подходящая операторнозначная функция, не обладающая, вообще говоря, полугрупповым свойством, после чего строится эквивалентная ей по Чернову полугруппа, которая и является решением соответствующей задачи. Рассмотрим аналог этого метода применительно к задаче математической статистики по определению СУС.

Эвристические соображения по использованию эквивалентных по Чернову функций для вычисления СУС основаны на близости определений случайных величин и динамических систем. Фактически СУС построен по совокупности расстояний между эволюционирующими выборками, плотности распределения которых изменяются вследствие некоторой неизвестной нам динамики. Если бы эта динамика была известной, можно было бы построить для этих функций уравнение Лиувилля, т.е. фактически определить полугруппу. Как сумма полугрупп не является полугруппой, так и сумма величин СУС не является их средним значением по полному объему данных L_{tot}. Однако ВФР расстояний между выборками есть по определению среднее значение ВФР, построенных по подвыборкам. Тогда возможно, что эквивалентный по Чернову средний согласованный уровень значимости по

данным из окон длины L будет близок к тому СУС, который получается из анализа выборок в окне L_{tot}. Докажем существование эквивалентного СУС.

Рассмотрим функцию

$$\Psi_N(\rho) = 1 - G_N(\rho). \tag{13}$$

Предположим, что ВФР $G_N(\rho)$ аппроксимирует дифференцируемую функцию распределения $G(\rho)$ соответствующей генеральной совокупности. Поскольку $G_N(\rho)$ монотонно возрастает от 0 до 1, то функция $\Psi_N(\rho)$ обладает свойствами:

$$\Psi_N(0) = 1,\ \Psi'_N(0) = -a_N \le 0.$$

В таком случае существует предел

$$\lim_{n\to\infty}\left(\Psi_N\left(\frac{\rho}{n}\right)\right)^n = e^{-\rho a_N} \equiv \Phi_N(\rho). \tag{14}$$

Следуя определениям, данным выше для операторных функций, будем называть предельную функцию $\Phi_N(\rho)$ эквивалентной по Чернову функции $\Psi_N(\rho)$ (то есть уровню значимости распределения $G_N(\rho)$). Будем обозначать эту эквивалентность как $\Phi_N(\rho) \overset{Ch}{\propto} \Psi_N(\rho)$. Очевидно, $\left|\Phi_N(\rho) - \Psi_N(\rho)\right| = o(\rho)$.

Согласно теореме об эквивалентности по Чернову средней полугруппы, если имеется (конечный или бесконечный) набор функций $\Psi_N^k(\rho)$ вида (13), эквивалентных в смысле (14) соответствующим функциям $\Phi_N^k(\rho)$ с коэффициентами $-a_N^k$ в показателе экспоненты, и если задан набор неотрицательных коэффициентов p_k, таких что $\sum_k p_k = 1$, то средняя функция $\overline{\Psi}_N(\rho) = \sum_k p_k \Psi_N^k(\rho)$ эквивалентна в смысле (14) функции $\overline{\Phi}_N(\rho) = e^{-\rho\overline{a}_N}$, где средний показатель экспоненты, играющий в данном случае роль генератора средней полугруппы случайных сдвигов в пространстве ρ, равен

$$\overline{a}_N = \sum_k p_k a_N^k. \tag{15}$$

Рассмотрим далее последовательность непересекающихся промежутков длины L. Для каждого k-го промежутка построим эмпирическое распределение $G_N^k(\rho, L)$ расстояний между встык-выборками длины N. Пусть таких промежутков M, так что $ML = L_{tot}$ есть полная длина изучаемого фрагмента ряда. Тогда распределение расстояний, построенное по всему объему данных, есть среднее распределение, полученное усреднением по отдельным выборкам:

$$G_N(\rho, L_{tot}) = \frac{1}{M}\sum_{k=1}^{M} G_N^k(\rho, L),$$
$$\overline{\Psi}_N(\rho) = \sum_k p_k \Psi_N^k(\rho) = 1 - G_N(\rho, L_{tot}),\ p_k = \frac{1}{M}. \quad (16)$$

Пусть $\rho_k^*(N)$ есть стационарная точка для функции $\Psi_N^k(\rho)$, а $\tilde{\rho}_k(N)$ отвечает эквивалентной ей функции $\Phi_N^k(\rho)$. В силу того, что $a_N^k \geq 0$, эта стационарная точка единственная. Обозначим через $\tilde{\overline{\rho}}(N)$ стационарную точку эквивалентной средней функции $\overline{\Phi}_N(\rho) = e^{-\rho \overline{a}_N}$. Тогда с точностью $o(\rho)$ из (14), (15) следует:

$$\tilde{\rho}_k(N) = \frac{1}{1 + a_N^k},\ \tilde{\overline{\rho}}(N) = \frac{1}{1 + \overline{a}_N} = \frac{1}{\frac{1}{M}\sum_{k=1}^{M} \frac{1}{\tilde{\rho}_k(N)}}. \quad (17)$$

Поскольку эквивалентные функции отличаются от оригиналов на величину $o(\rho)$, то на такую же величину отличаются и соответствующие стационарные точки.

Таким образом, доказана следующая теорема о стационарных точках распределений с непрерывными плотностями.

Теорема. *Пусть распределения случайных величин имеют непрерывные плотности и пусть на множестве этих случайных величин задана некоторая неотрицательная мера. Тогда стационарная точка функции, эквивалентной по Чернову среднему уровню значимости данных распределений, с точностью до бесконечно малой второго порядка совпадает с обратной величиной к среднему значению обратных величин стационарных точек функций, эквивалентных по Чернову уровням значимости данных распределений:*

$$\frac{1}{\tilde{\overline{\rho}}(N)} = \sum_{k=1}^{M} \frac{p_k}{\tilde{\rho}_k(N)}. \quad (18)$$

В силу сказанного ранее соотношение (18) с точностью $o(\rho)$ может быть распространено и на стационарные точки распределений $\Psi_N^k(\rho)$. Формула (18) принципиальна, поскольку она позволяет значительно уменьшить количество вычислительных процедур, если требуется изучить поведение СУС на множествах, представляющих собой объединение некоторых других множеств, для которых, собственно, и считается СУС. То есть для каждой длины N встык-выборок СУС вычисляется всего один раз в наименьших доступных для этой цели непересекающихся окнах длины L, покрывающих все исследуемое множество данных L_{tot}.

Такой подход существенно сокращает время расчета в задачах анализа Больших Данных. Он позволяет оценивать важную статистическую

характеристику нестационарного временного ряда по аппроксимационной процедуре, сходящейся к эталонному значению с гарантированной точностью.

5. Алгоритм вычисления СУС и программный комплекс

Поскольку данные для анализа временных рядов могут быть получены в различных форматах и с использованием различного количества электродов, то для унификации работы с такими данными используется определенный формат. (При обработке данных, поступающих в формате edf, использовался модуль библиотеки EDFbrowser, который служит для конвертации исходных файлов в кодировку ASCII (github.com/Teuniz/EDFbrowser).)

В начале работы алгоритма либо задаются, либо считываются из исходного файла параметры, характеризующие входные данные, и дополнительные настройки для алгоритма. К таким параметрам, в частности, относятся: общее число используемых в расчетах выборок; количество данных по каждой компоненте; шаг для встык-выборок внутри заданного расчетного окна; длина одной встык-выборки; размер интервала, в котором считается одно значение СУС. На Рис. 1 представлена блок-схема комплекса по вычислению СУС применительно к анализу электроэнцефалограмм (далее ЭЭГ).

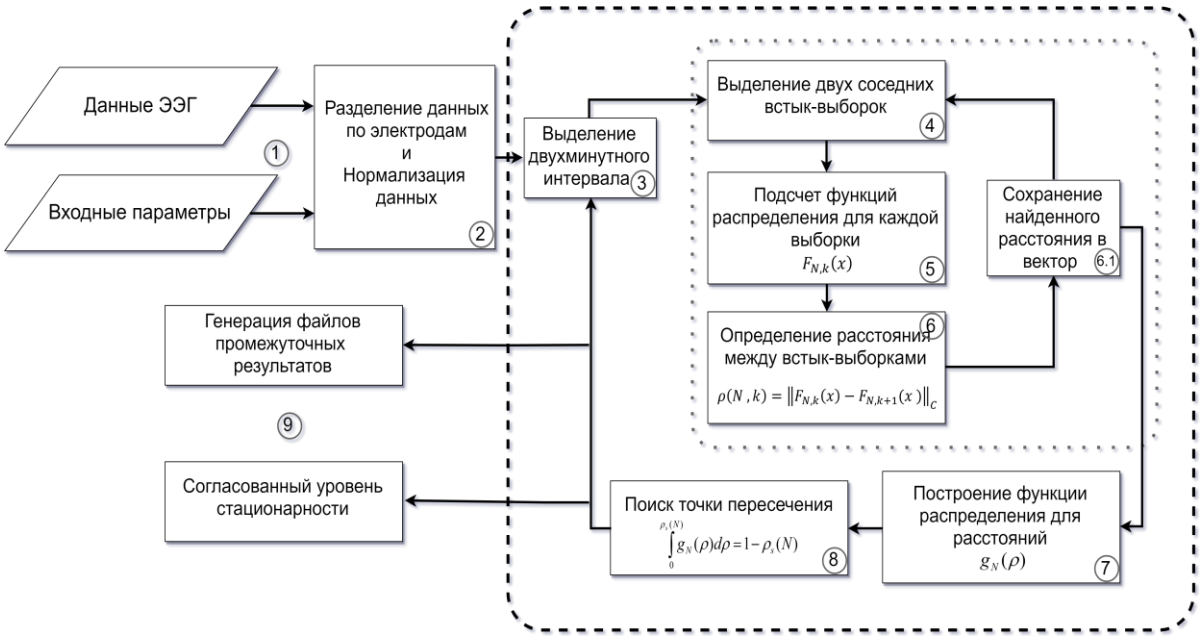

Рис. 1 – Блок-схема вычислительного алгоритма
для согласованного уровня стационарности

Программный комплекс NSSAT [30] имеет возможность работы в двух режимах: во-первых, подбор оптимальной ширины окна для дальнейшего анализа нестационарного случайного процесса; во-вторых, реализация расчета СУС по экспериментальным данным, с задаваемыми параметрами разбиения выборок данных. Схемы работы каждого из перечисленных выше режимов приведены на Рис. 2 и Рис. 3.

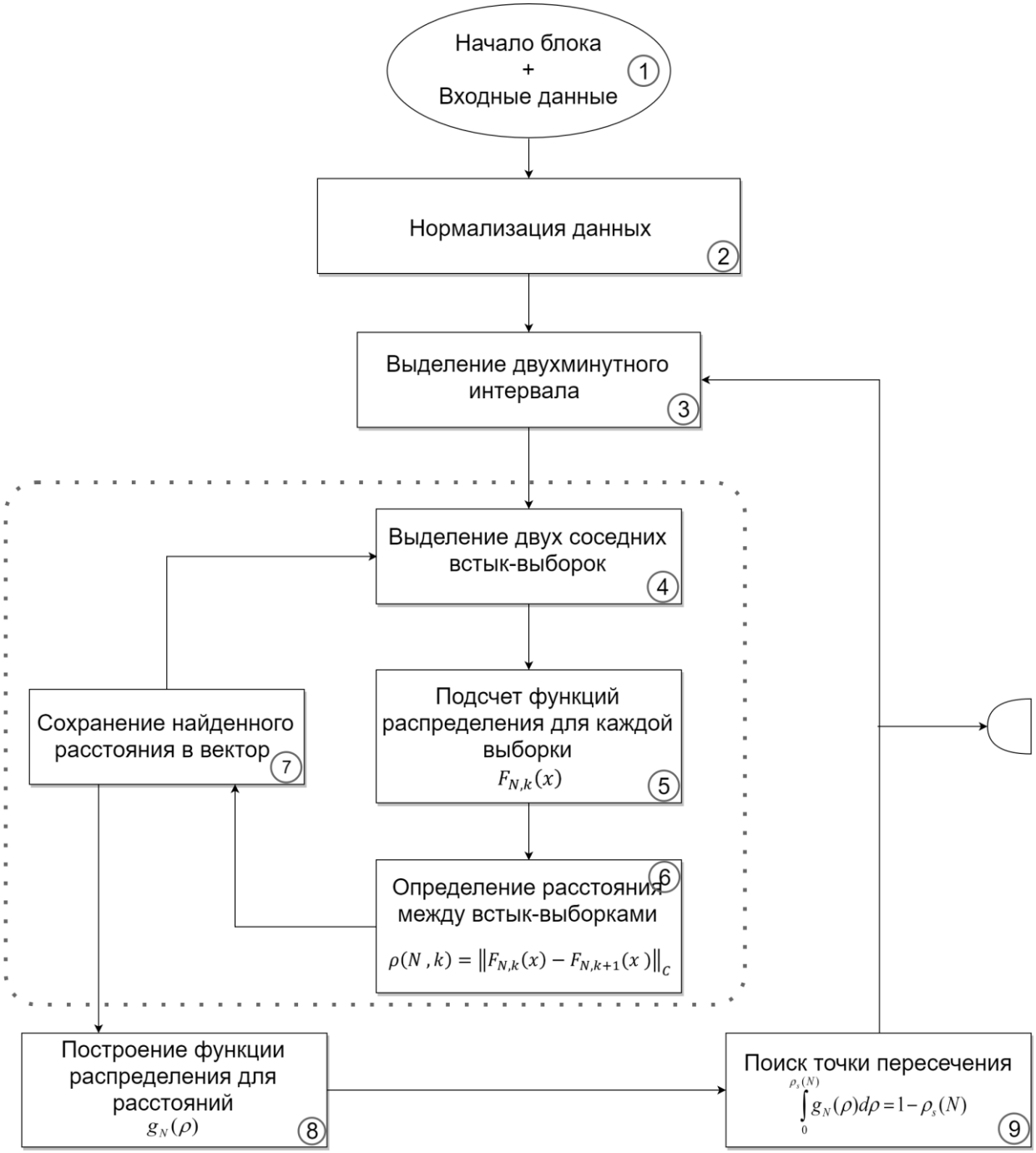

Рис 2 – Схема работы расчетного модуля комплекса NSSAT
в режиме расчета СУС с задаваемыми параметрами
разбиения выборок данных

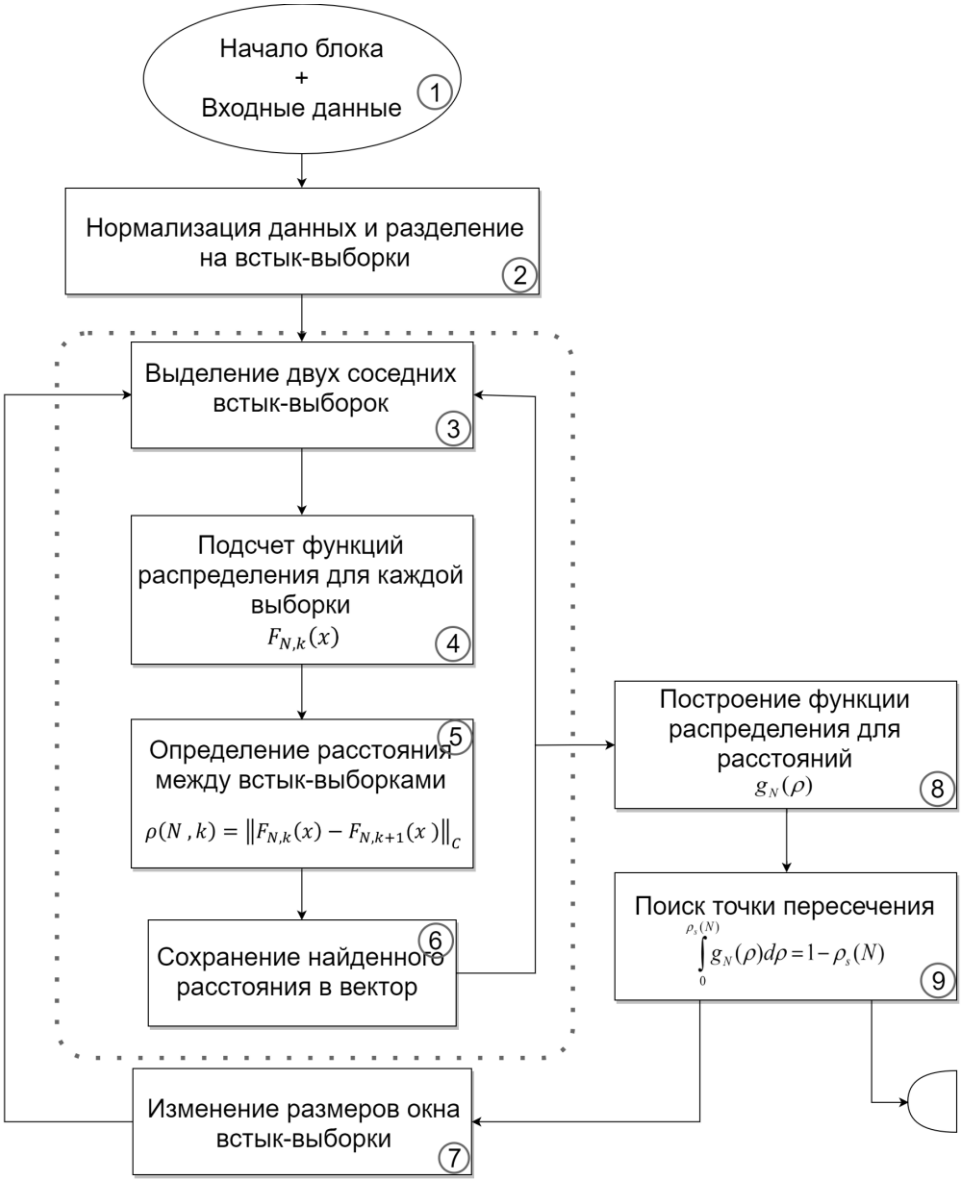

Рис. 3 – Схема работы расчетного модуля комплекса NSSAT в режиме подбора оптимальной ширины окна

Вычислительный алгоритм и пользовательский интерфейс написаны на языке C++/CLI в среде разработки MS Visual Studio 2017. Для работы программного комплекса необходима операционная система Windows 7 или более новые версии с поддержкой ASP.NET Core Runtime 2.1.509; 10 Mb свободного места на диске.

Скриншот рабочего окна программы представлен на Рис. 4. В верхней части окна программы располагается стандартный элемент интерфейса программ, предназначенных для работы в операционных системах семейства Windows, строка меню.

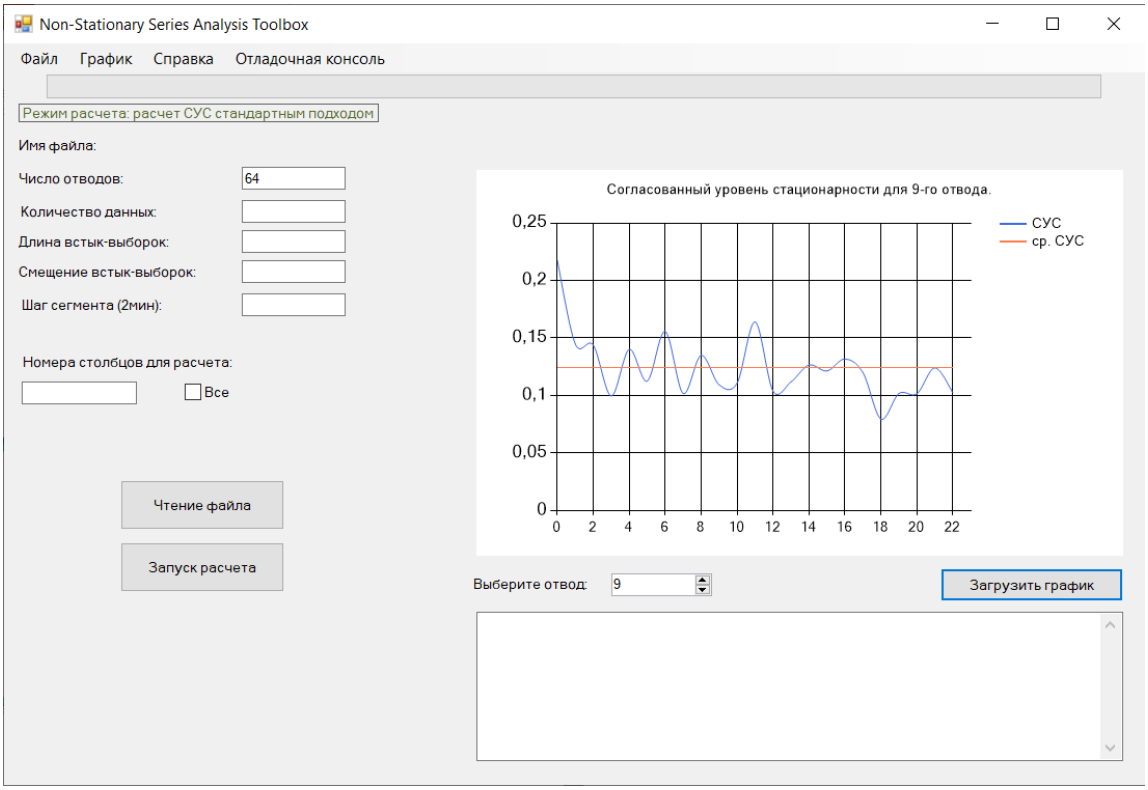

Рис. 4 – Скриншот рабочего окна программы NSSAT

В пункте меню «Файл» находятся элементы управления, относящиеся к работе с текущей сессией расчетов.

В пункте меню «График» находятся элементы управления, относящиеся к работе с графической областью, где отображаются изображения, построенные по полученным результирующим данным.

Пункт меню «Справка» позволяет получить детальную информацию о каждом элементе управления программного комплекса. Также в пункте меню «Справка» можно найти информацию о том, в каких форматах должны быть входные данные.

Пункт меню «Отладочная консоль» позволяет вызвать командную консоль программного комплекса для детального анализа отладочной информации по происходящим расчетам в программном комплексе.

6. Пример расчета разладки для ряда ЭЭГ

Построение методики предсказания приступа эпилепсии является одной из актуальных задач практической нейрохирургии. Эта задача в настоящее время не имеет ясного решения в виде установления причинно-следственных связей, поэтому применяются различные статистические методы, основанные на анализе данных, получаемых с отведений (электродов) при снятии ЭЭГ.

Биоэлектрическая активность фиксируется в достаточно широкой частотной полосе. В то же время низкие и высокие частоты подвергаются весьма сильным воздействиям из-за изменения проводимости кожного покрова и при непроизвольной мышечной активности. Для снятия клинической ЭЭГ

обычно используется диапазон частот от 0,5 до 70 Гц. Компьютерные энцефалографы работают на высокой частоте в 256 Гц, 512 Гц или 1024 Гц.

Предположим, что нестационарность свойственна самой природе наблюдаемой системы. Возможно, что она вызвана случайным вложением нескольких стационарных случайных процессов, отвечающих за определенные скрытые состояния этой системы. Хотя эти состояния на относительно коротких промежутках наблюдения не могут быть выявлены статистически, гипотеза об их соответствии скрытым параметрам может быть проверена. Если смена состояния системы соответствует изменению состава случайных процессов, то, предположительно, уровень нестационарности изменится. Индикацией разладки в таком случае может служить СУС, вычисляемый в скользящем окне определенной длины.

Если допустить, что состояние пациента перед приступом эпилепсии меняется не мгновенно, то индикатор изменения распределения СУС мог бы тогда использоваться как предиктор.

Тестирование программы проводилось на системе со следующими характеристиками: Процессор Intel® Core™ i7-2670QM 6 МБ кэш-памяти, 2,20 ГГц, ОЗУ 16 ГБ.

На рис. 5 приведена зависимость индекса нестационарности (9) рядов ЭЭГ от длины выборки. То, что этот индекс существенно больше единицы, означает высокую нестационарность изучаемого процесса.

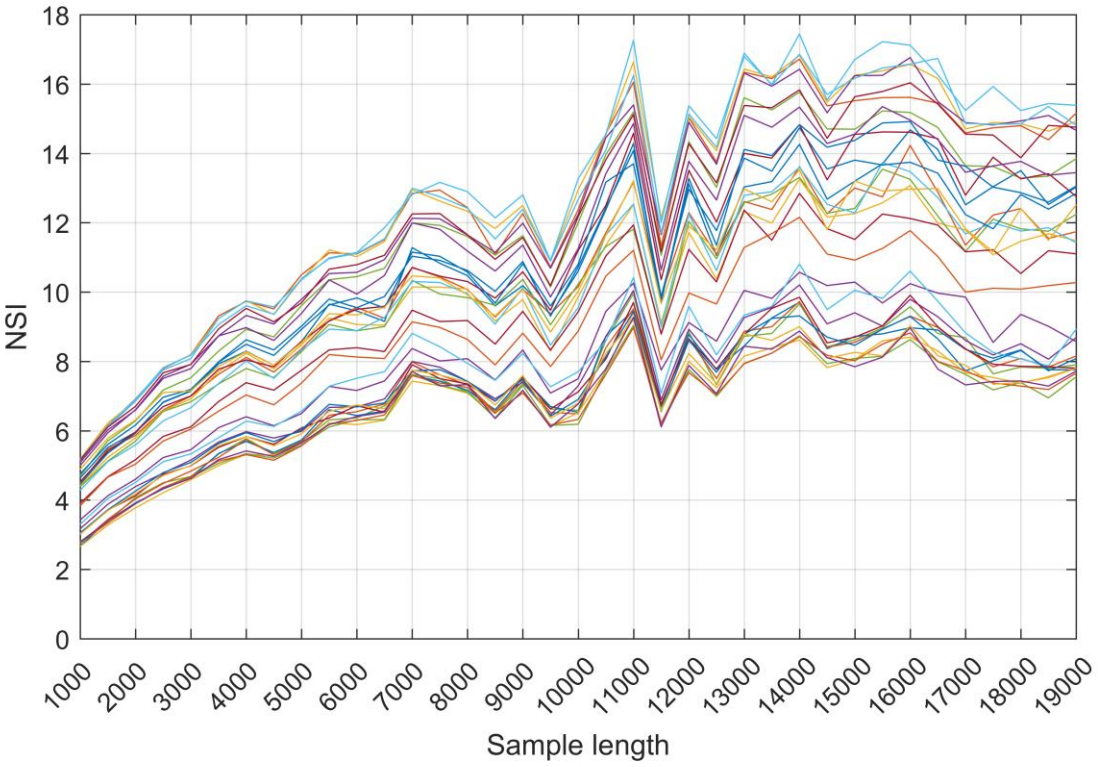

Рис. 5 – Зависимость индекса нестационарности от длины выборки для различных отведений

Из анализа, проведенного в [31, 32], следует, что оптимальная длина встык-выборок для построения СУС равна $N = 5000$, чему отвечает временной

промежуток примерно 20 сек., а длина окна наблюдения выбирается такой, чтобы относительная статистическая погрешность оценки СУС не превосходила самого СУС, для чего длина окна выбирается равной $L = 30$ тыс. (примерно 2 минуты наблюдений), а выборки сдвигаются с шагом 500.

Исследования были проведены для 10 разных пациентов, и результаты качественно совпали. Приведем для примера результаты обработки данных для одного из пациентов. Общий объем данных, отвечающих визуально не меняющемуся состоянию бодрствования пациента, составил 2,5 часа, т.е. примерно 2,3 млн данных или M = 78 независимых двухминутных отрезков. Для каждого из этих двухминутных промежутков и для каждого отведения вычислялся СУС $\rho_k^*(5000, 30000)$, $k = 1,...M = 78$. Типичный пример ряда указанных величин СУС для одного из электродов приведен ниже на Рис. 6.

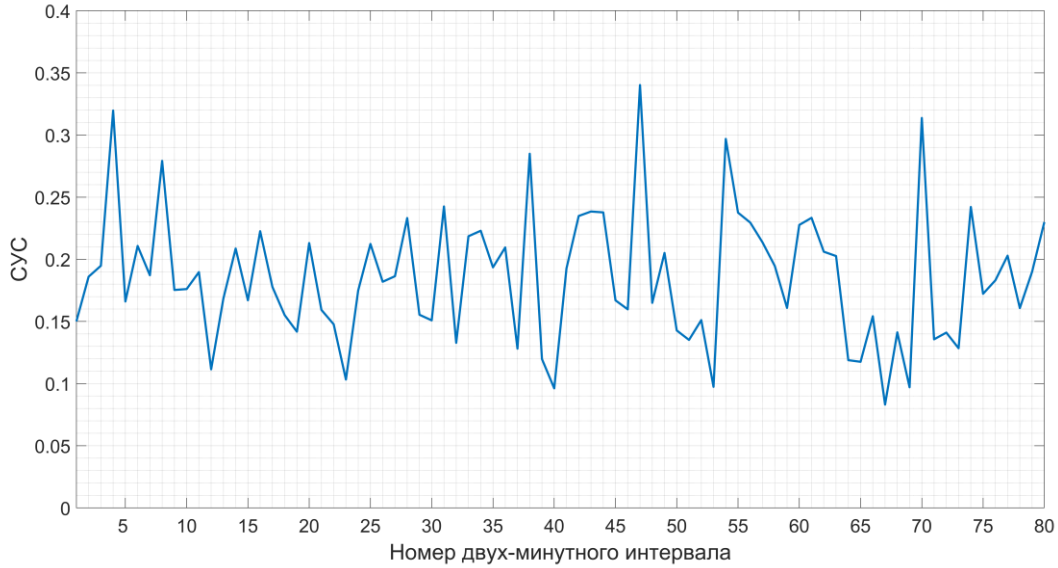

Рис. 6 – Пример ряда СУС $\rho_k^*(5000, 30000)$

Чтобы выяснить, действительно ли СУС является индикатором состояния пациента или же последовательность $\rho_k^*(N)$ просто представляет собой реализацию некоторого случайного процесса, не связанного с этим состоянием напрямую, а вызванного случайным изменением окружающей среды, необходимо проверить на стационарность выборку $\rho_k^*(N)$ на промежутке, где состояние пациента стабильно. Расчет по разработанной программе NSSAT показал, что ряд СУС стационарный, хотя сам исходный ряд показаний ЭЭГ, для которых этот СУС посчитан, является сильно нестационарным. Это означает, что построенный индикатор представляет собой статистический закон сохранения, идентифицирующий состояние человека.

Плотность функции распределения СУС представлена на Рис. 7. Она имеет приближенно симметричный треугольный вид.

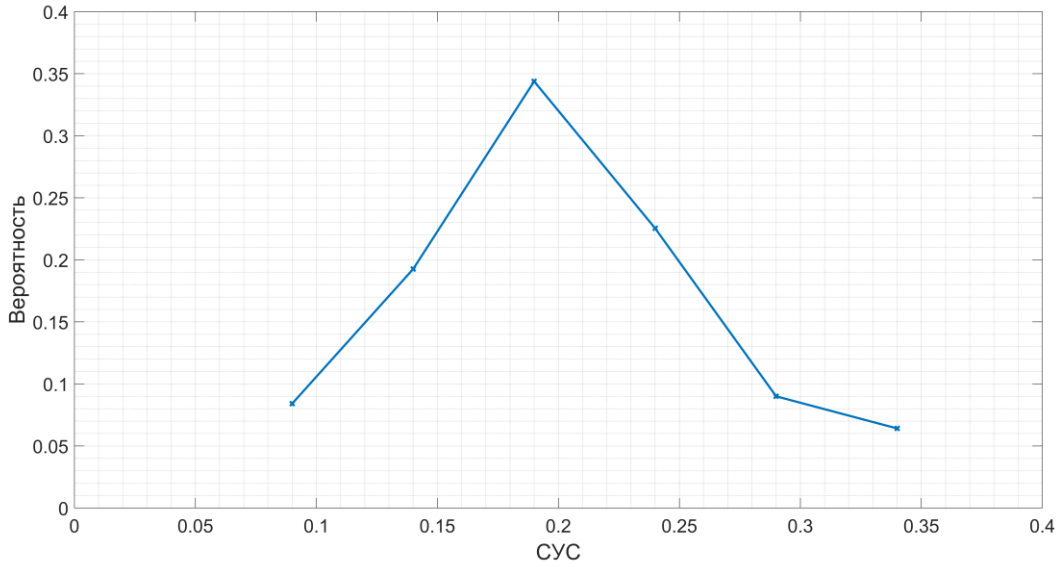

Рис. 7 – Стационарная плотность распределения СУС в окне 30 тыс. данных

Выяснилось, что вероятность превышения уровня $\rho^*_{tot}(5000) \approx 0{,}22$ в окне длины L на величину, равную $\left(\rho^*_{tot}(5000)\right)^2 \approx 0{,}05$, есть $1 - df(0{,}27) \approx 0{,}23$. Эта величина равна вероятности ошибки ложной тревоги.

Полезно сравнить уровень СУС на всем промежутке (СУС-tot в легенде Рис. 8) с эквивалентной по Чернову стационарной точкой функции значимости.

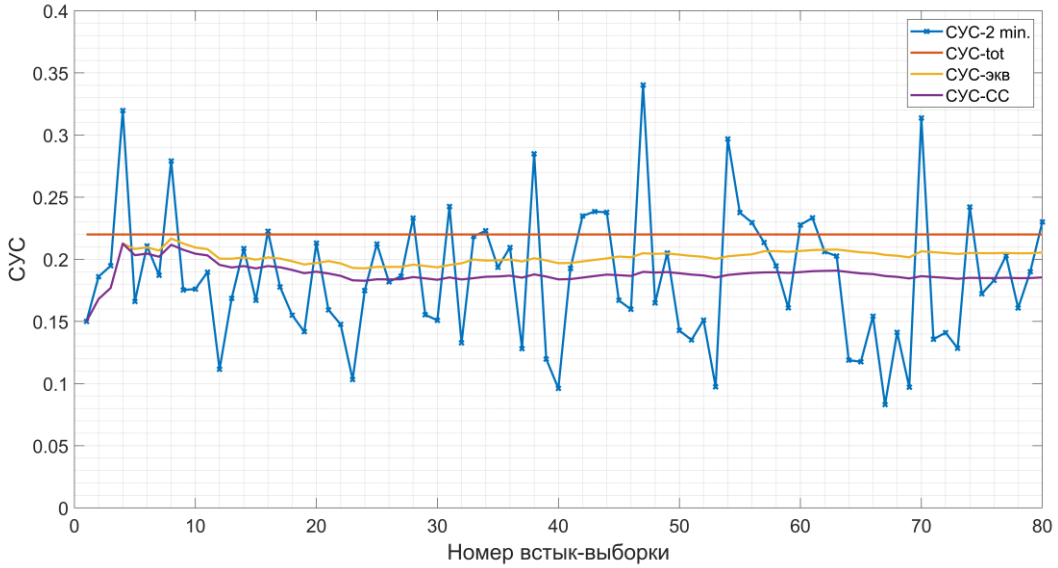

Рис. 8 – Сравнение стационарных точек уровня значимости для данных, представленных на рис. 6

Для всего множества данных стационарная точка уровня значимости $\Psi_{5000}(\rho)$ равна $\rho^*_{tot}(5000) \approx 0{,}22$. Значение же стационарной точки эквивалентной ей по Чернову функции $\Phi_{5000}(\rho)$ составило величину

$\tilde{\rho}_{tot}(5000) \approx 0{,}20$. Для сравнения на Рис. 8 показана также кривая скользящего среднего уровня СУС (обозначение СУС-СС). Видно, что она заметно уступает в точности модели, использующей аппроксимацию по Чернову. Следовательно, предложенное в п.4 приближение более точное, чем кажущееся естественным выборочное среднее.

7. Заключение

В работе описан численный код, генерирующий по фрагменту траектории нестационарного временного ряда величину предиктора разладки и определяющий одновременно оптимальную длину сканирования ряда.

Новизна работы заключается в том, что впервые теоретический метод построения эквивалентной по Чернову полугруппы был применен в математической статистике для обоснования алгоритма приближенного вычисления индикатора разладки. Также впервые был проведен статистический анализ выборочных функций распределения этого индикатора, что являлось в вычислительном плане серьезным техническим затруднением, связанным с увеличением размерности задачи. Практический результат, полученный при тестировании предложенного метода на примерах электроэнцефалограмм, состоит в улучшении распознавания приближения приступа эпилепсии. Этот результат получен с использованием нового критерия разладки, в качестве которого предложено использовать статистику, называемую согласованным уровнем стационарности ряда.

Теоретическая и практическая ценность работы состоит в том, что в ней построен алгоритм вычисления многомерного индикатора разладки в нестационарном временном ряде, доказаны теоремы о свойствах его аппроксимации, позволяющие значительно сократить время счета, и построен программный комплекс с интерфейсом, реализующий предложенную методику.

Дальнейшей задачей в области построения и анализа индикаторов разладки в теоретической области является исследование возможности уменьшения порога ошибки первого рода, который в данной работе равен СУС, за счет анализа самого ряда СУС в окне индикации и построения модели соответствующего временного ряда.

В практической области следует рассмотреть возможность увеличения размерности вычислительной задачи за счет анализа расстояний между выборками, сдвинутыми на расстояния, превосходящие окно выборки. Этот аспект также может снизить частоту ложных срабатываний индикатора разладки, не увеличивая при этом вероятность ошибки второго рода.

Литература

1. Бассвиль М. и др. Обнаружение изменения свойств сигналов и динамических систем (пер. с англ.) – М.: Мир, 1989. – 280 с.

2. Жиглявский А.А., Красковский А.Е. Обнаружение разладки случайных процессов в задачах радиотехники. – Л.: Изд-во ЛГУ, 1988. – 224 с.
3. Ширяев А.Н. Минимаксная оптимальность метода кумулятивных сумм в случае непрерывного времени // Успехи математических наук, 1996. Т. 51, № 4. С. 173-174.
4. Lorden G. Procedures for Reacting to a Change in Distribution. 1971.
5. Ширяев А. Задача скорейшего обнаружения нарушения стационарного режима // ДАН СССР, 1961. Т. 138. С. 1039-1042.
6. Girshick M.A., Rubin H. A Bayes approach to a quality control model // The Annals of Mathematical Statistics. – 1952. P. 114 – 125.
7. Casas P., Vaton S., Fillatre L., Nikiforov I. Optimal volume anomaly detection and isolation in large-scale IP networks using coarse-grained measurements // Computer Networks, 2010. Vol. 54, no. 11. P. 1750 – 1766.
8. Lakhina A., Crovella M., Diot C. Characterization of network-wide anomalies in traffic flows // Proceedings of the 4th ACM SIGCOMM conference on Internet measurement - IMC '04. – 2004. Vol. 6. P. 201.
9. Lakhina A., Crovella M., Diot C. Detecting distributed attacks using networkwide flow traffic // Proceedings of FloCon 2005 Analysis Workshop, 2005.
10. Lakhina A., Crovella M., Diot C. Diagnosing network-wide traffic anomalies // ACM SIGCOMM Computer Communication Review, 2004. Vol. 34, no. 4. P. 219.
11. Hassani H. Singular spectrum analysis: methodology and comparison // Journal of Data Science, 2007. Vol. 5, no. 2. P. 239 – 257.
12. Vautard R., Ghil M. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series // Physica D: Nonlinear Phenomena, 1989. Vol. 35, no. 3. P. 395–424.
13. Vautard R., Yiou P., Ghil M. Singular-spectrum analysis: A toolkit for short, noisy chaotic signals // Physica D: Nonlinear Phenomena, 1992. Vol. 58, no. 1. P. 95–126.
14. Yigitbasi N., Gallet M., Kondo D., Iosup A., Epema D. Analysis and modeling of time-correlated failures in large-scale distributed systems // Proceedings IEEE/ACM International Workshop on Grid Computing. 2010. P. 65–72.
15. Cook M.J., O'Brien T.J., Berkovic S.F., et al. Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study // Lancet Neurol., 2013. Vol. 12. P. 563 – 571.
16. Erramilli A., Narayan O., Willinger W. Experimental queueing analysis with long-range dependent packet traffic // IEEE/ACM Transactions on Networking (TON), 1996. Vol. 4, no. 2. P. 209 – 223.
17. Айвазян С.А. Методы эконометрики: учеб. – М.: Магистр. ИНФРА-М, 2014.
18. Казанцева Н.Н. Статистический контроль и статистические методы управления качеством. – Томск: Изд-во ТПУ, 2004. – 116 с.
19. Pham D.-S., Venkatesh S., Lazarescu M., Budhaditya S. Anomaly detection in large-scale data stream networks // Data Mining and Knowledge Discovery. 2014. Vol. 28. P. 145–189.
20. Гнеденко Б.В. Курс теории вероятностей. – М.: Физматлит, 1961. – 406 с.

21. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985. – 640 с.
22. Большев Л.Н. Асимптотически пирсоновские преобразования // Теория вероятностей и ее применения, 1963. Т.8. С. 129-155.
23. Уилкс С. Математическая статистика. – М.: Наука, 1967. – 632 с.
24. Орлов Ю.Н. Кинетические методы исследования нестационарных временных рядов. – М.: МФТИ, 2014. – 276 с.
25. Орлов Ю.Н., Федоров С.Л. Свидетельство № 2017619117 от 15.08.2017 о государственной регистрации программы для ЭВМ «Программный комплекс NonStatBox для статистического анализа и моделирования нестационарных временных рядов».
26. ISO/IEC/IEEE Systems and software engineering – Architecture description // ISO/IEC/IEEE 42010:2011(E) (Revision of ISO/IEC 42010:2007 and IEEE Std 1471-2000). – 2011. – Jan. – P. 1–46.
27. Об утверждении Стратегии развития отрасли информационных технологий в Российской Федерации на 2014–2020 годы и на перспективу до 2025 года. – 2013. URL: http://government.ru/docs/8024/.
28. Chernoff P. Note on product formulas for operator semigroups // J. Funct. Anal., 84, 1968. P. 238-242.
29. Орлов Ю.Н., Сакбаев В.Ж., Смолянов О.Г. Формулы Фейнмана как метод усреднения случайных гамильтонианов // Труды МИРАН, 2014. Т. 285. С. 232-243.
30. Кислицын А.А., Орлов Ю.Н. Свидетельство о государственной регистрации № 2019660374 от 05.08.2019 программы для ЭВМ «Программный комплекс NSSAT (Non-Stationary Series Analysis Toolbox) для определения и визуализации характеристик нестационарности временных рядов».
31. Кислицын А.А., Козлова А.Б., Корсакова М.Б., Машеров Е.Л., Орлов Ю.Н. Стационарная точка уровня значимости для нестационарных функций распределения // Препринты ИПМ им. М.В. Келдыша. 2018. № 113. 20 с.
32. Кислицын А.А., Козлова А.Б., Корсакова М.Б., Орлов Ю.Н. Индикатор разладки для нестационарных случайных процессов // Доклады РАН, сер. математическая, 2019. Т. 484. № 4. С. 393-396.