УДК [004.67+004.9]:[550.3+57.045]

ПРОБЛЕМЫ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ С ПРОПУСКАМИ И МЕТОДЫ ИХ РЕШЕНИЯ В ПРОГРАММЕ WINABD

© 2016 г. А.В. Дещеревский[1], В.И. Журавлев[1], А.Н. Никольский[2], А.Я. Сидорин[1]

[1] Институт физики Земли им. О.Ю. Шмидта РАН, г. Москва, Россия
[2] ООО «КМК Консалтинг», г. Москва, Россия

Рассмотрены технологии, применяемые при анализе временных рядов с пропусками. Обсуждаются некоторые алгоритмы выделения (очистки) сигнала и оценки его характеристик, в частности ритмических составляющих, для рядов с пропущенными наблюдениями. Приведены примеры анализа данных, полученных при долговременных наблюдениях на Гармском геофизическом полигоне и в других регионах. Рассмотрены используемые в программе WinABD технические решения, позволяющие наиболее эффективно организовать работу обсуждаемых алгоритмов при наличии дефектов наблюдений.

Ключевые слова: данные геофизического мониторинга, анализ временных рядов с пропусками, алгоритм, программа, обработка в скользящем окне, ритмы.

Введение

Проблема обработки временных рядов данных с дефектами является одной из самых обсуждаемых среди специалистов в этой области. Это не удивительно: идеальных систем наблюдений не бывает. На практике сплошь и рядом встречаются ситуации, при которых некоторые измерения по каким-то причинам пропущены. Это может случиться из-за различных организационных проблем или отказа техники, а также из-за помех и других аномальных внешних воздействий, искажающих регистрируемый сигнал без возможности его восстановления. Например, при удаленном наблюдении за состоянием здоровья пациентов, не находящихся в медицинском учреждении, связь на какое-то время может быть потеряна, сами датчики могут быть отключены или потерять контакт, либо пациент может подвергаться лечебным или иным воздействиям, влияние которых существенно искажает отслеживаемый ритм. В первом случае наблюдения не выполняются, во втором в данных появляется выброс, или дефект, который необходимо как-то откорректировать. Даже при наличии единственного выброса значительной амплитуды значения многих рассчитываемых статистик могут существенно исказиться

[*Хампель и др.,* 1989]. Пропуск данных сильно влияет, например, на точность оценок параметров регрессии [*Baghi et al.,* 2015].

Сигналы с пропусками встречаются в самых разных областях знаний: астрофизике [*Scargle et al.,* 2013], геофизике [*Love,* 2009], биологии [*Богдасарян,* 1980], медицине [*Reffinetti et al.,* 2007], социологии и экономике [*Giles,* 1998], экологии [*Губанов,* 2004], при работе с GIS и решении других научных задач [*Макс,* 1983]. Такие сигналы приходится анализировать при оптимизации технологических процессов и автоматизации производств [*Скрипник и др.,* 1988; *Злоба, Яцкив,* 2002; *Карлов, Проворова,* 2012], распознавании речи и идентификации говорящего [*Разумихин,* 2000], а также в других технических приложениях.

Большинство публикаций, посвященных анализу сигналов с пропусками, нацелены на решение отдельных задач, относящихся к конкретной предметной области. Так, значительное число работ посвящено модификации известных методов анализа временных рядов типа АРИМА [*Бокс, Дженкинс,* 1974] с учетом дефектов наблюдений [*Klingenberg,* 2007]. Медики часто пытаются выделить в наблюдениях признаки развивающейся болезни и ее источник [*Katinas,* 2012]. Похожие подходы к выделению полезного сигнала используются в геофизике и климатологии [*Dergachev et al.,* 2001]. Эти исследования часто содержат весьма развитую, но, как правило, узкоспециализированную математическую основу, т.е. рассматриваются лишь отдельные алгоритмы или даже отдельные аспекты применения конкретного алгоритма при строго определенных характеристиках обрабатываемого сигнала и четких целевых критериях обработки [*Torres-Reyna,* 2014]. Аналогичные вопросы поднимаются и в многочисленных интернет-обсуждениях этой тематики (например, [Time…, 2015]).

Анализ литературы показывает, что при обработке сигналов с пропусками стандартом де-факто является подход, основанный на каком-либо способе заполнения (интерполяции) пропущенных наблюдений. Это позволяет построить ряд данных, имеющий равномерный шаг во времени между опросами. Для таких временных рядов существует огромное количество методов статистического анализа, хорошо обоснованных математически. Они включаются во все распространенные пакеты статистического анализа и легко доступны для применения исследователями, специализирующимися на решении самых разных задач.

Различные способы заполнения пропущенных наблюдений обсуждаются не только в публикациях [*Грачев,* 2004; *Снитюк,* 2006], но и в рамках многочисленных интернет-форумов (см., например, известный форум Research Gate [*Seelam,* 2015]). В этих материалах можно найти сотни ссылок на публикации и советы, как заполнить пропуски данных в том или ином сигнале. В России при обработке временных рядов с дефектами также преобладает подход, основанный на заполнении пропущенных наблюдений «правдоподобными» значениями данных. В частности, именно этот метод наиболее популярен среди пользователей широко известного пакета MATLAB. Дело в том, что лишь небольшая часть функций этого популярного пакета позволяет корректно обрабатывать массивы данных с пропусками, причем эти функции реализуют далеко не все нужные исследователю инструменты и разбросаны по разным надстройкам, каждая из которых устанавливается отдельно. Если же сначала заполнить пропуски данных с помощью тех или иных алгоритмов, можно без каких-либо ограничений использовать весь инструментарий, имеющийся как в пакете MATLAB, так и в других программных продуктах, предназначенных для статистического анализа временных рядов.

Существует множество методов, предназначенных для заполнения пропущенных наблюдений. Основная идея таких алгоритмов довольно проста и состоит в построении

некоторого функционала, проходящего через все известные точки данных или наиболее близко к ним. Далее предполагается, что в промежутках между отсчетами значение построенной функции будет близко к истинному значению сигнала. Одно из первых систематических описаний таких алгоритмов дано в известной книге Р.В. Хемминга [1972]. Различные способы заполнения пропусков активно изучаются и сейчас [*Литтл, Рубин,* 1990; *Россиев,* 2000; *Gorban et al.,* 2002*; Абраменкова, Круглов,* 2005; *Рыженкова,* 2011; *Концевая,* 2012; *Маркин, Щербаков,* 2013; *Pashova et al.,* 2013]. В программном обеспечении, имеющемся на рынке и предлагающем исследователям готовые решения для различных задач, реализованы самые разные алгоритмы для заполнения пропущенных наблюдений.

Однако любой способ заполнения пропусков всегда использует те или иные допущения о свойствах временного ряда (определенной морфологии, информационной избыточности, гладкости и т.д.). Фактически необходимо построить полную статистическую модель сигнала [*Моисеев,* 1998]. Это приводит к тому, что для каждого типа наблюдений необходимо разрабатывать свои методы восстановления потерянных данных. Именно по этой причине многие исследователи вынуждены отказаться от использования готовых алгоритмов, имеющихся в матобеспечении, и разрабатывать собственные методики, наиболее подходящие, по их мнению, для рассматриваемых ими временных рядов. Однако такие методики чаще всего не реализуются в виде программных продуктов, доступных другим исследователям.

Если исследуемый процесс не вполне соответствует той модели, на которой основаны алгоритмы заполнения пропусков, или его статистические характеристики попросту неизвестны (что довольно часто встречается при долговременном мониторинге геофизических и иных показателей, когда каждый получаемый ряд уникален и невоспроизводим), то качественное заполнение пропусков чаще всего остается недостижимой мечтой. Хуже того, при неаккуратном заполнении пропусков в данные могут привноситься различные паразитные эффекты и артефакты, что существенно осложняет интерпретацию результатов статистической обработки [*Рыженкова,* 2011].

Сложности и искажения данных, возникающие при попытках искусственного восстановления потерянной информации (что, строго говоря, невозможно), привели к появлению альтернативного подхода к обработке рядов с дефектами, не предусматривающего заполнение пропусков наблюдений. Идея состоит в том, чтобы анализировать только реальные наблюдения, не внося в них никаких дополнительных предположений [*Литтл, Рубин,* 1991]. Так, несмотря на весьма ограниченные возможности пакета MATLAB по работе с такими сигналами, для него разработаны специальные рекомендации по работе с пропусками, которые обозначаются стандартным значением NaN [*Smith,* 2014]. Ограниченное использование пустых значений возможно и в Excel, например предусмотрены способы их игнорирования при построении графиков. Более широкие возможности в обработке подобных данных имеются в пакете SAS (см. сайт в Интернете: http://support.sas.com/kb/22/921.html). Однако его применение относится больше к экономике, чем другим областям знаний.

Актуальность проблемы привела к тому, что многие исследователи разрабатывают собственные алгоритмы для обработки сигналов без заполнения пропусков или используют те возможности, которые реализованы в немногих специализированных программах, предназначенных для обработки подобных рядов. Так, для ответа на вопрос, есть ли в анализируемых данных какой-то полезный сигнал и какие периоды могут быть уверенно выделены, часто используется косинор-преобразование [*Маркелов,* 2007; *Refinetti et al.,* 2007; *Cornelissen,* 2014]. Этот метод анализа, предложенный Ф. Халбер-

гом [*Halberg,* 1969, 1980], подробно описан в работах [*Емельянов,* 1976; *Карп, Катинас,* 1989; *Оранский, Царфис,* 1989]. Несмотря на довольно громоздкие формулы, метод основан на простом предположении о возможности аппроксимация экспериментальных сигналов гармониками с заданным периодом. После определения амплитуд и фаз таких гармоник идет уточнение вклада других гармоник и остатка.

Большая популярность данного преобразования среди медиков и хронобиологов связана с тем, что многие медицинские приборы (например, кардиографы) часто комплектуются компьютерным рабочим местом, на котором установлено готовое матобеспечение для расчетов косинор-преобразования. Так, оно реализовано в имеющейся на рынке программе Cosinor Ellipse 2006 [*Корягина, Нопин,* 2006; Cosinor…, 2015], применение которой не требует высокой квалификации пользователя. В цитированной выше работе Р.А. Багдасаряна она используется для диагностики заболеваний. Большой международный проект БИОКОС под эгидой ученых из США [*Halberg et al.,* 1997] несколько десятилетий приводит данные измерений под косинор-модель. Преимуществом такого подхода является возможность сравнить результаты разных исследователей в одной шкале.

Широкое использование косинор-алгоритма, несмотря на известные недостатки и ограничения такого подхода, свидетельствует о существующей острой потребности в алгоритмах и программных продуктах для работы с данными, содержащими пропуски. Реальные временные ряды часто устроены намного сложнее, чем упомянутая косинор-модель. Ясно, что для полноценной работы с такими рядами необходим полнофункциональный продукт, позволяющий не только выполнять косинор-анализ, но и применять без каких-либо ограничений другие методы обработки сигналов, включая средства визуализации данных, инструменты для ведения базы данных и т.д. Очень важно, чтобы все применяемое матобеспечение изначально допускало наличие пропусков наблюдений, так как без этого невозможно реализовать многие важные функции.

Упомянутые выше программы, предназначенные для косинор-анализа данных с пропусками, этому требованию не удовлетворяют. Подготовка данных, а также визуализация полученных результатов выполняются с помощью продуктов третьих фирм, т.е. используемая среда не является интерактивной, что сильно снижает эффективность работы. При хранении исходных данных в обычно используемых СУБД возникает проблема корректной обработки пропусков, поскольку стандартные запросы часто не обеспечивают необходимой функциональности. В результате исследователь вынужден разрабатывать свои методы для работы с такими данными. Фактически появился отдельный раздел знаний – работа с базами данных, содержащими пропущенные значения [*Schlüter,* 2012].

Анализ показывает, что готовых алгоритмов, позволяющих обрабатывать временные ряды с пропусками данных, не так уж много. На рынке практически не представлены программные средства, в которых этот подход был бы реализован с достаточной полнотой. Это связано как со сложностью строгого математического обоснования таких алгоритмов (без чего их включение в промышленные пакеты проблематично), так и с существенным возрастанием различных погрешностей и искажений при оценивании статистик по данным с пропусками [*Love,* 2009; *Baghi et al.,* 2015; *Sandip et al.,* 2015]. Однако не следует забывать, что и процедура заполнения пропусков может серьезно влиять на результаты расчетов. Если свойства процесса изучены недостаточно, то оценка погрешностей и искажений, вносимых в сигнал при восстановлении потерянных наблюдений, становится нетривиальной задачей. Попытки игнорировать этот факт и сводить обработку сигналов с заполненными пропусками к формальному применению типовых алгоритмов крайне опасны, поскольку при этом может создаваться иллю-

зия точности и математической строгости результатов, которая на самом деле не имеет под собой никаких оснований [*Урбах,* 1963].

Авторы настоящей работы в течение многих лет изучали самые разные временные ряды, проводили режимные наблюдения на геофизических полигонах. В силу разных причин получаемые данные неизбежно содержали довольно многочисленные дефекты. Для анализа этих наблюдений мы использовали как опубликованные алгоритмы, так и различные модификации стандартных методов, специально сконструированные для работы с такими рядами и не требующие предварительного заполнения пропусков данных. Эти алгоритмы позволяют оценивать характеристики сигналов, выполнять адаптивную и иную фильтрацию, изучать взаимосвязи между процессами и т.д. Все они могут применяться не только в геофизических приложениях, но и в медицине, биологии, экономике и т.д.

Цель настоящей работы – представить некоторые способы анализа временных рядов с пропусками и продемонстрировать особенности их работы на примере различных данных. Некоторые из представленных ниже алгоритмов позволяют работать одновременно с обычными, условно непрерывными, рядами и рядами, состоящими из отдельных событий. Для геофизики ярким примером ряда, описывающего дискретный процесс, являются каталоги землетрясений. Весь такой ряд состоит только из времен, координат и магнитуд событий, между которыми информации нет. Подобные «штучные» события существуют и в других областях знаний, имеющих дело с измерением параметров во времени. В медицине это, например, моменты выхода человека из сна, резкое начало заболевания, выбросы эндорфинов, скачки давления и изменения сердечного ритма, в экономике скачки цен на бирже, промышленные катастрофы, выход новых производств на рынок, в климатологии различные катастрофы. Поскольку такие события могут влиять на изучаемые процессы, необходимо иметь возможность одновременно работать с обычными рядами и рядами таких событий, отображать их на одном графике, искать возможные предвестники и последствия штучных событий в условно непрерывных рядах.

Все обсуждаемые в работе алгоритмы реализованы авторами в программе WinABD. Этот программный комплекс включает специализированную базу данных для хранения исходных сигналов, интерактивную среду визуализации и широкий набор инструментов обработки, т.е. представляет собой полноценную среду для работы с экспериментальными временными рядами [*Дещеревский и др.,* 2016а, б].

Технологии работы с временными рядами данных, содержащими пропуски

В наиболее общем случае временной ряд, или ряд наблюдений, – это последовательность измерений некоторой величины, выполненных в произвольные моменты времени. Однако всегда удобнее иметь дело с наблюдениями, которые выполняются через равные промежутки времени (если конечно, природа наблюдаемой величины допускает такую возможность) [*Теребиж,* 1992]. Поэтому при режимных наблюдениях различных параметров интервал между измерениями, как правило, постоянен. Но из-за пропуска некоторых наблюдений шаг отсчета оказывается лишь почти равномерным.

Напрашивающееся решение в такой ситуации состоит в заполнении пробелов некоторыми вымышленными значениями и последующем анализе рядов с равномерным опросом во времени. Однако заполнение пропусков не является панацеей, так как оно всегда приводит к искажению информации. Эти искажения могут непредсказуемо возрастать, если ряд имеет сложные свойства и/или его характеристики плохо изучены.

Другое возражение против заполнения пропусков состоит в том, что при выполнении многих расчетов явное заполнение утерянных данных не является необходимым. Например, оценить среднее значение или автоковариацию вполне можно и без предварительного заполнения пропусков.

Чтобы работать с сигналом «как есть», не вводя в данные искусственных наблюдений, необходимо решить два вопроса. Технический аспект требует, чтобы используемая в программном обеспечении модель данных обеспечивала сохранение информации о моментах пропусков данных и позволяла любой процедуре однозначно идентифицировать забракованные наблюдения. Обычно такие значения кодируются значением NaN или NULL [*Guiles,* 2007; Filling…, 2013; Working…, 2015], что обеспечивает более компактное использование памяти, чем при создании отдельного вектора с маской дефектов.

Алгоритмический аспект подразумевает, что вместо стандартных алгоритмов, предполагающих непрерывность сигнала, во всех случаях должны использоваться альтернативные алгоритмы, допускающие пробелы в данных. Для многих задач такие алгоритмы известны либо могут быть получены путем тривиальной модификации стандартных вычислительных процедур. Например, при оценке среднего значения пропуски просто игнорируются, и оценка выполняется по имеющимся наблюдениям [*Yuan et al.,* 2004]. Аналогичная техника может быть применена при оценке ковариаций или коэффициентов регрессии. В других случаях необходимы более сложные процедуры, связанные с особой обработкой отсутствующих наблюдений.

Ниже рассмотрены некоторые алгоритмы анализа временных рядов с пропусками, реализованные нами в программе WinABD. Общий принцип их построения состоит в следующем. В отсутствие пропусков данных расчеты выполняются по стандартным формулам, предназначенным для обработки непрерывных рядов. При наличии пропусков отсутствующие значения просто исключаются из вычислений. В формулы вводятся необходимые поправки: пересчитываются весовые и нормировочные коэффициенты, учитываются другие эффекты. По возможности, формулы корректируются так, чтобы обеспечить получение несмещенных оценок.

Заметим, что далеко не для всех модифицированных таким образом алгоритмов имеется достаточно строгое математическое обоснование. Однако следует подчеркнуть, что проблема математической строгости актуальна для любого подхода, если данные не вполне соответствуют той модели, в рамках которой строится используемый алгоритм. Практика показывает, что экспериментальные ряды, получаемые в результате длительных непрерывных наблюдений в геофизике, метеорологии, медицине и в других приложениях, почти всегда не стационарны, засорены меняющимися во времени помехами, для них характерны сильные внутренние корреляции, а функции распределения далеки от гауссовой. При этом исследователь часто располагает только единичными реализациями изучаемого процесса, а информация о моментах пропусков становится доступной лишь после проведения наблюдений. В такой ситуации невозможно построить строгую статистическую модель изучаемого процесса и провести полноценное теоретическое исследование работоспособности любого нетривиального алгоритма: исследовать его свойства, рассчитать погрешности различных оценок и т.д. Все сказанное в полной мере относится и к алгоритмам заполнения пропусков.

Исследователи-практики часто закрывают глаза на неполное соответствие обрабатываемых данных условиям применимости алгоритма, считая, что использование методов и критериев с доказанной оптимальностью автоматически гарантирует качество результатов [*Урбах,* 1963]. Но такой подход часто приводит к серьезным проблемам. Обзор, приведенный в [*Дещеревский и др.,* 2015а], показывает, что ошибки в применении

статистических критериев и моделей массово допускаются не только в биомедицинских, но и технических публикациях. При обработке неидеальных данных крайне рискованно полагаться на любые критерии значимости, основанные на внутренней сходимости алгоритма, поскольку такие критерии почти всегда излишне оптимистичны.

Выход из этого положения мы видим в использовании дополнительных инструментов, позволяющих контролировать достоверность получаемых результатов и оценивать значимость обнаруживаемых эффектов с помощью внешних, т.е. независимых от применяемого алгоритма верифицирующих процедур. Несмотря на то, что такие процедуры далеко не всегда могут быть строго формализованы, а успешность их применения сильно зависит от квалификации исследователи и от того внимания, которое он уделяет этим вопросам, такое решение представляется наиболее приемлемым с методологической точки зрения.

К числу таких верифицирующих процедур можно отнести самые разные методы и приемы: численное моделирование, модификацию данных, сравнение результатов обработки одного массива наблюдений разными алгоритмами и т.д. Дополнительные возможности верификации появляются, если используемая среда в полной мере поддерживает работу с данными, содержащими пропуски, и не требует их обязательного заполнения уже на начальном этапе анализа. Во-первых, исследователь получает возможность комбинировать различные алгоритмы поиска и выбраковки дефектных данных с другими инструментами статистической обработки. Это существенно расширяет возможности контроля качества данных и оценки степени влияния сомнительных наблюдений на результаты анализа. Во-вторых, открывается возможность прямого учета эффектов, обусловленных пропусками, непосредственно в ходе вычислений, путем встраивания соответствующих процедур в алгоритмы обработки сигнала.

В WinABD реализованы обе эти возможности. Так, при использовании любого алгоритма исследователь может задать максимально допустимую долю пропусков η в любом обрабатываемом подмножестве данных. Это условие проверяется на каждом шаге вычислений. Например, при оценке фрактальных статистик проверка количества пропусков выполняется отдельно для каждого масштаба. Если в какой-то момент доля пропусков превышает критическую величину, то вычисления прекращаются, и результат объявляется пропуском. Аналогично, при вычислении скользящего среднего тренд рассчитывается только для тех моментов времени, когда внутри окна сглаживания есть достаточно данных. Это никак не мешает последующей обработке сглаженного сигнала другими методами, поскольку модель данных WinABD допускает наличие пропусков и перерывов в исследуемых рядах на любом этапе анализа.

Меняя разрешенный процент пропусков η и отслеживая происходящие при этом изменения, можно оценить, насколько сильно влияет на результаты неполнота ряда. Подчеркнем, что при определении критерия η исследователь может руководствоваться исключительно соображениями научной целесообразности, а вовсе не техническими ограничениями алгоритмов. Это позволяет детально контролировать работу каждого метода и точно оценивать величину эффектов, вносимых пропусками наблюдений.

Ниже даются примеры использования предложенного подхода при обработке данных, полученных на Гармском полигоне в ходе длительного эксперимента по автоматизированному мониторингу различных параметров. Этот эксперимент выполнялся с целью поиска предвестников землетрясений [Гармский…, 1990; Автоматизированная…, 1991]. Представленные алгоритмы могут быть использованы для изучения временной структуры любых подобных процессов в медицине, метеорологии, астрономии и других приложениях.

Ядерное скользящее сглаживание

Функция ядра и ее нормировка

Один из наиболее распространенных непараметрических алгоритмов, используемых для выделения трендов, – это алгоритм скользящего среднего. Для оценки тренда $\hat{S}$ в момент времени t рассчитывается среднее значение ряда S в скользящем окне некоторой длительности (ширины) T:

$$\hat{S}(t) = (1/T) \cdot \sum S(i), \quad (1)$$

где индекс суммирования i меняется от $t - T + 1$ до t, т.е. пробегает все значения в пределах окна. В других обозначениях это можно записать как

$$\hat{S}(t) = S(t - t1) * F(t1), \quad (2)$$

где символ «*» обозначает свертку, а $F(t1)$ – это нормированная функция окна. Для случая прямоугольного окна, когда все значения в окне учитываются с одинаковым весом, условие нормировки можно записать в виде

$$F(t) = (1/T) \cdot f(t), \quad (3)$$

где

$$f(t) = (\Theta(t) - \Theta(t - T)), \quad (4)$$

а $\Theta(t)$ – это функция Хевисайда. Фактически выражение (2) представляет собой обычную нормированную свертку во времени.

Функцию $F(t)$, которая задает значения коэффициентов фильтра (т.е. определяет форму окна), принято называть ядром сглаживающего фильтра [*Хардле,* 1989; *Любушин,* 2007; *Лагутин,* 2009]. При расчете обычного скользящего среднего все весовые коэффициенты сглаживающего фильтра одинаковы, т.е. используется окно прямоугольной формы. Однако во многих случаях вместо прямоугольного окна целесообразно использовать более сложные окна [*Валеев,* 2001; *Концевая,* 2011, 2013]. В частности, именно такие ядра используются при частотной фильтрации. Путем специального подбора коэффициентов окна можно добиться того, чтобы фильтр пропускал только колебания с определенными периодами, варьировать крутизну среза АЧХ и т.д. [*Kanasevich,* 1981; *Хемминг,* 1987]. Например, такое ядро может иметь треугольную форму:

$$f(t) = (t \cdot (\Theta(t) - \Theta(t - T)), \quad (5)$$

Согласно (5), коэффициенты фильтра линейно спадают в пределах окна, которое расположено ранее по времени, чем рассчитываемая точка. Для нормировки весовой функции $f(t)$ ее значения делятся на общую площадь:

$$F(t) = f(t) / \sum f(i). \quad (6)$$

Если весовая функция окна отличается от прямоугольной, такое сглаживание принято называть ядерным [*Хардле,* 1989; *Любушин,* 2007; *Лагутин,* 2009], чтобы подчеркнуть отличие функции ядра от прямоугольника.

Симметричное и причинное окно

В выражениях (1)–(6) значение тренда $\hat{S}$ оценивается на правом краю окна, т.е. при его вычислении используются только данные за предшествующие моменты времени. Однако при ретроспективной обработке часто бывает удобнее использовать не «причинное», а симметричное окно. В этом случае оцененное значение $\hat{S}$ сопоставляется не с правым краем окна, а с его серединой, т.е. с моментом времени $t - T/2$. При этом ис-

пользуется симметричная функция ядра. Ниже при обработке мы будем применять гауссово ядро:

$$f(t) = (1/\sqrt{2\pi}) \cdot \exp(-(t - t0)^2 / (2\sigma^2)), \quad (7)$$

где σ – это стандартное отклонение анализируемой величины $t0$ – время, соответствующее середине окна в случае симметричного ядра или его правой части в случае причинного окна $f(t)$ – ненормированное ядро сглаживания. Ширина гауссиана обычно подбирается так, чтобы края окна соответствовали трем «сигма»: $\sigma = T/6$ (симметричное окно).

Оценка скользящего среднего при наличии пропусков

При наличии пропусков данных амплитудно-частотная характеристика сглаживающего фильтра меняется, причем эти изменения зависят от того, в какие именно моменты времени пропущены наблюдения [*Kanasevich,* 1981; *Хемминг,* 1987]. Строго говоря, пересчет характеристик фильтра в этом случае необходимо проводить заново для каждого t, поскольку для каждого нового положения окна набор пропусков будет другим. В WinABD при сглаживании данных с пропусками используется упрощенный эмпирический алгоритм, а именно: для расчетов используется то же самое выражение (2) с той разницей, что в свертке участвуют только реальные значения данных. Нормировка весовой функции в этом случае выполняется согласно выражению (6), при этом из суммы $\sum f(i)$ исключаются те моменты времени i, для которых наблюдения отсутствуют.

Заметим, что при подсчете общего количества пропусков η в пределах окна в этом случае учитывается не количество отсутствующих значений данных, а суммарный вес коэффициентов фильтра f(i), исключенных из свертки из-за того, что соответствующие им значения данных являются пропусками:

$$\eta = \sum f(k) / \sum f(i), \quad (8)$$

где индекс i пробегает все моменты времени в пределах окна, а индекс k – только те моменты времени t_k, для которых $S(t_k)$ не является пропуском.

Обычно функция ядра более или менее быстро спадает от середины окна к его краю. Поэтому возможность расчета скользящего среднего при заданном η зависит не только от количества пропусков в пределах окна, но и от того, как именно эти пропуски расположены в пределах окна, а также от того, какое сглаживающее ядро выбрано для расчетов. Применение этого правила на практике почти всегда приводит к хорошим результатам, поскольку описанный алгоритм соответствует здравому смыслу. Значения, расположенные по соседству с вычисляемым скользящим средним, оказывают большее влияние на результат, чем удаленные от него. Это относится и к критерию «доля пропусков». Правомерность такого подхода для большинства сигналов подтверждается анализом автокорреляционной функции, чьи значения на малых лагах времени обычно заметно выше, чем на удаленных.

Вычисления в скользящем окне без уменьшения длины ряда

Известный недостаток методов скользящего окна связан с тем, что такие алгоритмы трудно комбинировать друг с другом. Начальное положение окна обычно выбирается так, чтобы его левая граница совпадала с началом ряда, а конечное – так, чтобы с окончанием ряда совпадала правая граница окна. Понятно, что в этом случае полный пробег окна будет меньше, чем длина исходного ряда. Соответственно длина отфильтрованно-

го ряда (или ряда, показывающего динамику какого-то расчетного показателя) будет при каждой фильтрации уменьшаться на размер окна. Если ряд имеет ограниченную длину, а окно достаточно широкое, то уже после применения нескольких методов от сигнала «ничего не останется». А ведь с сигналом, как правило, необходимо выполнить целую серию операций. Достаточно типичным выглядит алгоритм, при котором из ряда адаптивно удаляются выбросы, отфильтровывается высокочастотная помеха, вычитается низкочастотный тренд. Затем выделяется полезный сигнал (например, суточный ритм), оценивается его амплитуда и, наконец, рассчитывается динамика изменений коэффициента регрессии амплитуды на некоторый внешний фактор.

В WinABD проблема уменьшения длины ряда решается путем схлопывания окна на границах ряда. Суть этой процедуры состоит в том, что при обработке ряда любым методом он автоматически дополняется пропусками справа и слева так, чтобы полный пробег окна равнялся бы длине ряда. Затем используются те же самые алгоритмы, которые применяются при наличии пропусков данных внутри сигнала. Рассмотрим, как это работает, для окна шириной пять точек. Пусть, для определенности, наблюдения начаты в момент времени $t = 1$. В этом случае для расчета $\hat{S}(1)$ в симметричном окне используются значения $S(-1)$, $S(0)$, $S(1)$, $S(2)$ и $S(3)$. Значения $S(-1)$ и $S(0)$, формально находящиеся вне ряда, автоматически заполняются пропусками. После этого массив данных $S(-1{:}3)$, состоящий из двух «искусственных» пропусков и трех реальных значений, передается в функцию свертки (2). Данная функция проверяет общее количество пропусков согласно формуле (8). Если это соотношение удовлетворяет заданному в настройках алгоритма критическому уровню η, то $\hat{S}(1)$ вычисляется по формуле (2); в противном случае значение $\hat{S}(1)$ объявляется пропуском. Совершенно аналогично вычисляются функции $\hat{S}$ в конце ряда, а также в начале и конце длительных интервалов пропущенных наблюдений.

Легко видеть, что если значения $S(1)$, $S(2)$ и $S(3)$ не являются пропусками, а разрешенная доля пропусков η в пределах окна размером $T = 5$ точек составляет 40 % или выше, то расчет функции $\hat{S}$ для момента времени $t = 1$ будет выполнен. Фактически это означает, что длина ряда $\hat{S}$ будет такой же, как и у ряда S. То есть вычисления в скользящем окне выполняются без уменьшения длины отфильтрованного сигнала.

Аналогичная технология используется в WinABD и при выполнении любых других вычислений в скользящем окне.

Выбор оптимального значения параметра η *при обработке сигналов с пропусками*

Пример оценки скользящего среднего с треугольной функцией ядра приведен на рис. 1. Видно, что интервал времени, для которого удается рассчитать функцию $\hat{S}(t)$, зависит от значения η. При $\eta = 80$ % длина отфильтрованного сигнала $\hat{S}$ (линия *2*) практически соответствует длине исходного ряда S. Заметим, что в этом случае описанный алгоритм автоматически заполняет пропуски внутри ряда, рассчитывая скользящее среднее даже для тех моментов времени, когда наблюдения не выполнялись.

Иная картина наблюдается при $\eta = 40$ %. В этом случае скользящее среднее оценивается только для тех положений окна, когда обеспечивается достаточная представительность данных (линия *1*). Анализ показывает, что линия 80 %-го тренда содержит заметно больше случайных особенностей (флуктуаций), чем линия 40 %-го тренда, и, по-видимому, не все эти флуктуации вполне отвечают интуитивным представлениям об оптимальном сглаживании исходной кривой (см., например, особенности линии тренда в начале августа или в конце января).

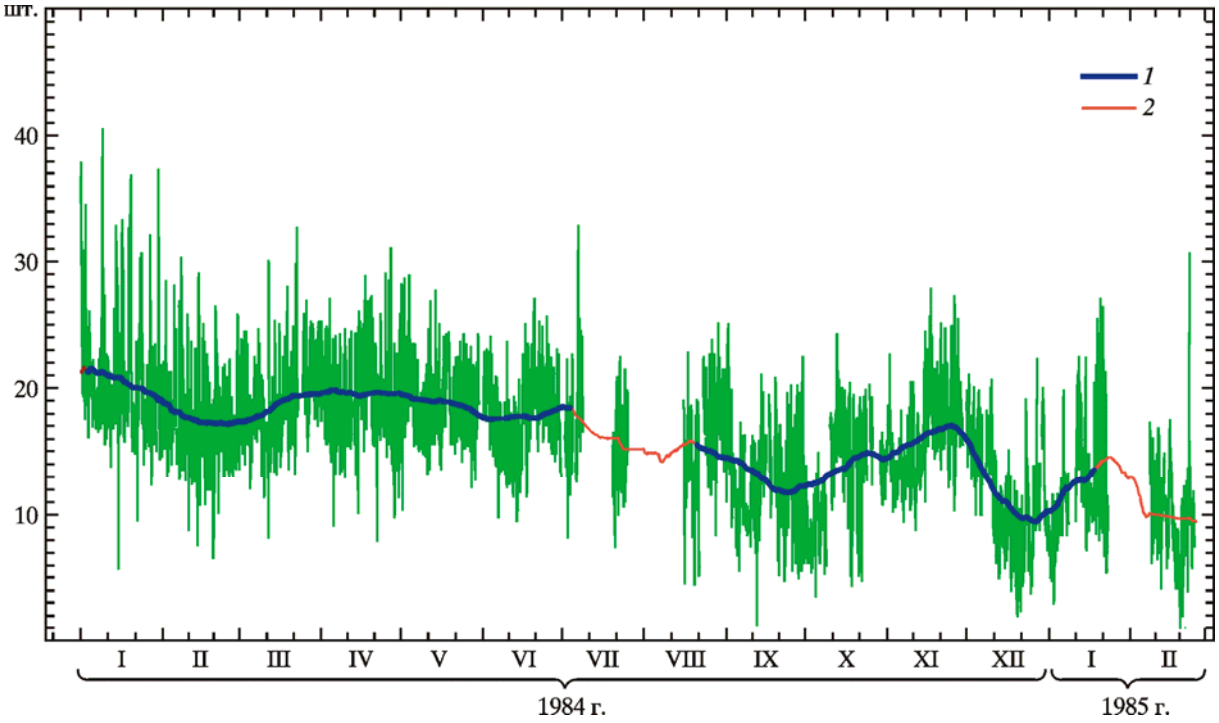

Рис. 1. Экспоненциальное сглаживание ряда SLON1B в скользящем окне шириной 731 ч.
1 – тренд, выделенный при разрешенной доле пропусков η = 40 %; *2* – η = 80 %

Частотная фильтрация

Стандартный алгоритм частотной фильтрации

Частотная фильтрация – это один из наиболее часто применяемых инструментов при обработке экспериментальных сигналов. Он используется при фильтрации помех, выделении ритмов и в других случаях. Стандартный алгоритм частотной фильтрации [*Kanasevich,* 1981; *Хемминг,* 1987] основан на расчете коэффициентов фильтра в соответствии с заданными параметрами амплитудно-частотной характеристики фильтра (АЧХ). Обычно задаются границы полосы подавления/пропускания и крутизна среза фильтра. Затем выполняется свертка исходного ряда с фильтрующей последовательностью $F(t)$, реализующей оптимальный частотный фильтр, согласно формуле (2).

В общем случае последовательность $F(t)$ содержит как положительные, так и отрицательные коэффициенты, что позволяет улучшить АЧХ фильтра [*Kanasevich,* 1981; *Хемминг,* 1987]. Однако при наличии резких особенностей у функции ядра $F(t)$ рассмотренная выше процедура исключения пропусков неприменима, так как на выходе фильтра будут наблюдаться знакопеременные колебания, повторяющие форму функции $F(t)$. Поэтому обычно перед частотной фильтрацией выполняется заполнение пропусков данных. Более корректная процедура состоит в пересчете коэффициентов фильтра, составляющих функцию $F(t)$, на каждом шаге с учетом фактического распределения пропусков в пределах окна.

Чтобы продемонстрировать работу стандартного алгоритма частотной фильтрации при наличии пропусков, мы выбрали ряды биологической природы. Ряды представляют собой число электрических импульсов слабоэлектрической рыбы нильский слоник *Gnathonemus leopoldianus*, генерируемых для электрической локации [*Дещеревский,*

Сидорин, 2002]. Чем выше двигательная активность рыбки, тем выше поток импульсов. Возможный спектральный диапазон электрических импульсов такого объекта может быть довольно высок, однако в силу специфики измерений рассматривались только относительно медленные вариации. Анализировалась активность двух экземпляров рыбы, обозначаемых ниже как SLON1B и SLON2M.

Результаты выделения полосы периодов 168–336 ч для ряда двигательной активности SLON1B, заполненного линейной интерполяцией, показаны на рис. 2, *в*. Длина фильтрующей последовательности была выбрана равной 336 точек, что соответствует длине фильтрующей последовательности рассмотренного ниже сглаживающего фильтра. Видно, что на участках пропусков данных наблюдаются вариации отфильтрованного сигнала, которые не имеют аналогов на исходном ряде. Кроме того, из отфильтрованного сигнала исключается его начальный участок, что хорошо видно на рис. 2, *в*.

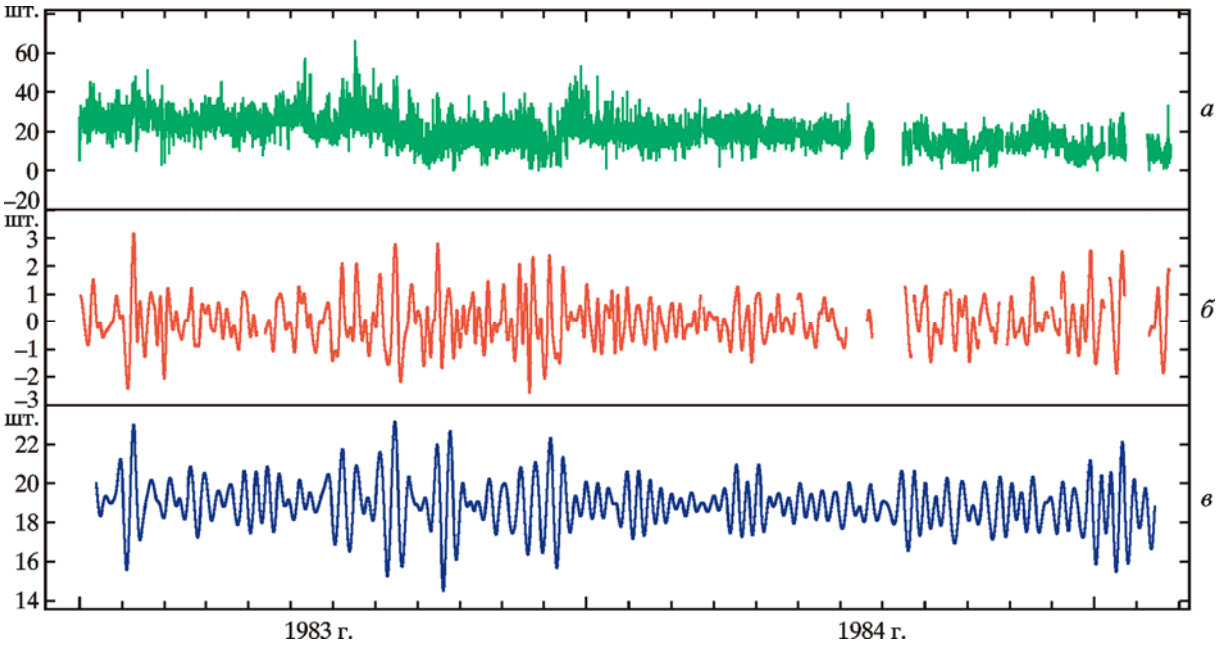

Рис. 2. Ряд SLON1B (*а*), а также вариации с периодами от 168 до 336 ч, отфильтрованные методом ядерного сглаживания с гауссовым ядром (7) (*б*) и методом частотной фильтрации (*в*)

Использование ядерного сглаживания в качестве частотного фильтра

Альтернативный метод выделения (вырезания) полосы частот основан на последовательной обработке ряда сглаживающим фильтром для удаления сначала низких, а затем высоких частот. Этот алгоритм очень редко применяется при обработке данных геофизического мониторинга, поскольку при обработке сигналов с заполненными пропусками он не имеет особых преимуществ по сравнению со стандартной частотной фильтрацией, а крутизна среза АЧХ у сглаживающего фильтра значительно меньше.

В WinABD обработка сигнала в скользящем окне не приводит к уменьшению длины ряда, а любой массив данных может содержать пропуски. Это создает условия для более широкого применения ядерного скользящего сглаживания [*Хардле,* 1989; *Лагутин,* 2009] для частотной фильтрации. Пример полосовой частотной фильтрации ряда двигательной активности биоиндикатора SLON1B в диапазоне периодов 168–336 ч приведен на рис. 2, *б*. Спектр отфильтрованного ряда показан на рис. 3 (линия *2*).

Как видно из данных, приведенных на рис. 2, отфильтрованный сигнал не содержит ни высокочастотных вариаций, ни трендов. При этом на границах интервалов пропусков данных не возникает никаких паразитных эффектов. Это связано с тем, что ядро (7) имеет только положительные коэффициенты, которые монотонно спадают от середины окна к его краю. АЧХ такого фильтра имеет довольно пологий срез, что хорошо видно по спектру отфильтрованного сигнала (см. рис. 3, линия *2*). Амплитуда низкочастотных вариаций подавлена лишь на порядок, а высокочастотных – примерно на два порядка по сравнению с исходным сигналом. Достоинством фильтра с гауссовым ядром является отсутствие побочных максимумов (дополнительных лепестков) АЧХ. Это уменьшает возможность просачивания в отфильтрованный сигнал монохроматических периодичностей высокой амплитуды (например, суточной, сезонной или приливной), которые часто присутствуют в экспериментальных сигналах.

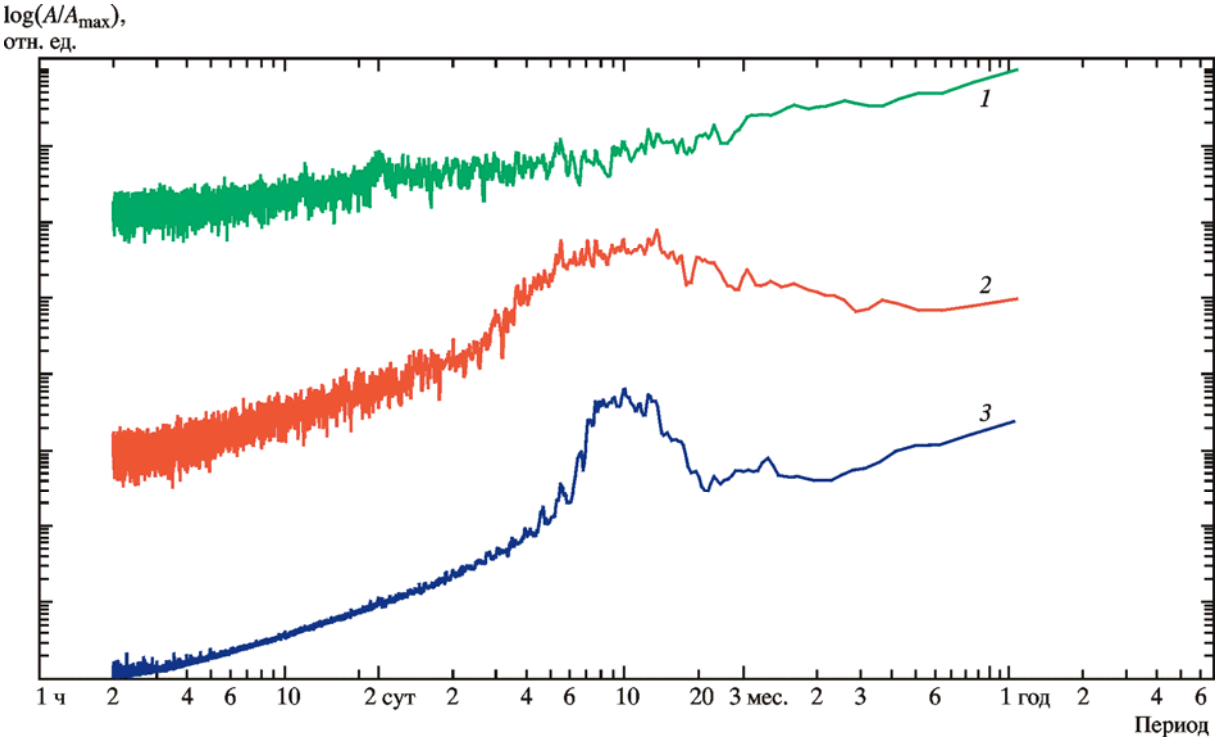

Рис. 3. Спектр ряда SLON1B до (*1*) и после (*2*, *3*) частотной фильтрации: *2* – фильтрация методом ядерного сглаживания с гауссовым ядром (7); *3* – фильтрация с помощью полосового частотного фильтра

Сравнение двух алгоритмов частотной фильтрации

Анализ отфильтрованных сигналов (см. рис. 2, *б*, *в*) и их спектров (ср. линии *2* и *3* на рис. 3) показывает, что стандартный алгоритм частотной фильтрации обеспечивает более сильное подавление высоких и низких частот. Крутизна среза АЧХ у частотного фильтра оказалась равной 22 дБ, что заметно выше, чем у сглаживающего фильтра (длина фильтрующей последовательности $F(t)$ в обоих случаях была одинаковой). С другой стороны, сглаживающий фильтр сохранил длину ряда без изменений. При большом количестве пропусков это может быть существенным преимуществом.

Подчеркнем, что выбор оптимального алгоритма частотной фильтрации должен осуществляться, исходя из соображений научной целесообразности, а не технических

ограничений пакета. Это возможно только в рамках программной среды, допускающей работу с реальными исходными данными, без обязательного заполнения пропусков.

Выделение ритма

При решении многих задач важное место уделяется изучению различных ритмов – квазипериодических изменений каких-либо показателей. Ритмом принято называть периодически повторяющиеся изменения характера и интенсивности различных процессов и явлений, причем форма вариации, период и амплитуда ритма могут меняться в известных пределах с течением времени. В биологии подобные квазипериодические процессы принято называть биоритмами [*Кучеров и др.,* 1970]. Сигналы с аналогичными свойствами достаточно широко распространены и в других предметных областях: геофизике, метеорологии, социологии, медицине, экономике и т.п. Ритмические вариации различных параметров могут происходить как под воздействием внешних причин – например, изменения активности животных вследствие суточного хода температуры, освещенности или других параметров, так и в силу эндогенных эффектов

Оценка среднего ритма в скользящем окне

Рассмотрим в качестве примера метод, основанный на оценке среднего ритма в скользящем окне. Этот алгоритм представляет собой развитие метода наложения эпох [*Дещеревский, Сидорин,* 1999]. Функция среднего ритма $\mathbf{R}(t)$ с периодом P рассчитывается, как

$$\mathbf{R}(t) = (1/N) \cdot \sum(\mathrm{S}(t + iP)), \qquad (9)$$

где суммирование ведется по всем целым i, при которых момент времени $t + iP$ оказывается в пределах симметричного скользящего окна шириной T_1, а N – число суммируемых значений (не-пропусков). Таким образом, ряд в пределах окна делится по длине периода на некоторое число частей, или эпох, длительностью P, а затем все эти эпохи когерентно накладываются друг на друга и усредняются.

При отсутствии пропусков данных значение N примерно равно числу периодов в пределах окна (точнее, оно может быть равно $N = \mathrm{int}(T_1/P)$ или $N = \mathrm{int}(T_1/P) + 1$ в зависимости от фазы t). При наличии пропусков значение N будет соответственно меньше, причем оно может сильно отличаться для разных моментов времени t, если пропуски концентрируются преимущественно на определенных фазах ритма. Например, при мониторинге суточного ритма наблюдения могут систематически пропускаться в ночные часы. Чтобы быть уверенным, что все значения функции среднего ритма $\mathbf{R}(t)$ оценены по достаточному числу наблюдений, следует указать в настройках алгоритма более строгое ограничение предельной доли пропусков η. В этом случае повышается вероятность того, что некоторые значения среднего ритма $\mathbf{R}(t)$ окажутся пропусками. Обычно характеристики ритма на таких фазах в дальнейшем оцениваются особо. Так, если известно, что $\mathbf{R}(t)$ представляет собой плавную функцию, то не вычисленные значения $\mathbf{R}(t)$ можно впоследствии интерполировать каким-либо способом. Интерес может представлять и сравнение функций $\mathbf{R}(t)$, рассчитанных при разных ограничениях η.

Подгонка амплитуды среднего ритма

На втором шаге построенный таким образом «средний ритм» используется в качестве образца, или репера. В центрированном окне меньшей ширины $T_2 << T_1$ оценивается регрессия фактического сигнала **S** на средний ритм **R** и рассчитывается аппроксимирующая функция $\hat{\mathbf{S}}$:

$$\mathbf{S} = a \cdot \mathbf{R} + b + \mathbf{Z}, \quad (10а)$$

$$\hat{\mathbf{S}} = a \cdot \mathbf{R} + b, \quad (10б)$$

где a и b – коэффициенты регрессионной модели, а **Z** – не скомпенсированный остаток, или разница между моделью и реальным сигналом. (Жирным шрифтом в этой формуле выделены векторные величины, обозначающие фрагменты соответствующих временных рядов в пределах окна). Коэффициенты a и b оцениваются методом наименьших квадратов, так чтобы минимизировать **Z** (т.е. минимизируется сумма квадратов $\sum Z_i^2$, где индекс i пробегает все значения в пределах окна). Отметим, что регрессионная модель (10) строится по фактически выполненным наблюдениям. Предварительное заполнение пропусков в данном случае только ухудшит оценки параметров ритма.

Оценив значения коэффициентов a и b и зная реперную функцию **R**, можно рассчитать значение аппроксимирующей функции $\hat{\mathbf{S}} = a^*\mathbf{R} + b$ для любого t. Однако на практике целесообразно вычислять значение $\hat{\mathbf{S}}$ не для всего окна в целом, а только для какого-то одного момента времени t. Если смещать окно сразу на всю его ширину, то значения коэффициентов a и b могут резко меняться при переходе от одного окна к следующему. В результате аппроксимирующая функция $\hat{\mathbf{S}}$ потеряет гладкость и непрерывность, так как она будет построена из «кусков», величина которых равна размеру окна, но которые плохо сопрягаются друг с другом на границах окон. По этой причине скользящее окно WinABD при подобных расчетах всегда смещается только на одну точку, а не на всю ширину окна. Для каждого положения скользящего окна параметры регрессионной модели оцениваются заново. Для уменьшения вычислительных затрат при этом используется специальная библиотека функций, обеспечивающая возможность поточечного добавления и удаления данных к массивам, по которым рассчитывается регрессия. Это позволяет при смещении окна не пересчитывать полностью все регрессионные суммы, а просто корректировать их. Таким образом, на каждом шаге работы алгоритма вычисляется значение аппроксимирующей функции $\hat{\mathbf{S}}$ в одной-единственной точке. Если эта точка располагается на правой границе окна или даже еще правее (в последнем случае реперная функция **R** экстраполируется вперед по времени за пределы того окна, в котором выполняется оценка коэффициентов a и b регрессионной модели), алгоритм позволяет прогнозировать будущие значения ряда **S**.

При ретроспективном анализе данных, когда приоритетом является качество оценки параметров ритма, расчет аппроксимирующей функции $\hat{\mathbf{S}}$ обычно выполняется в точке, располагающейся в середине окна. Это позволяет не только выполнять обработку без уменьшения длины ряда, но и гарантирует отсутствие систематических искажений фазы у отфильтрованного сигнала.

Выделение суточного ритма двигательной активности биоиндикаторов

Для иллюстрации работы описанного выше алгоритма мы отфильтровали суточный ритм активности биоиндикатора SLON2M. По исходному сигналу (рис. 4, линия *1*) оценивался средний суточный ритм **R** в скользящем окне шириной 168 ч (модель 9). Затем этот ритм был сглажен скользящим средним с гауссовым ядром. Согласно [*Дещеревский, Сидорин,* 1999], ширина окна сглаживания была выбрана равной 9 ч. Полученный

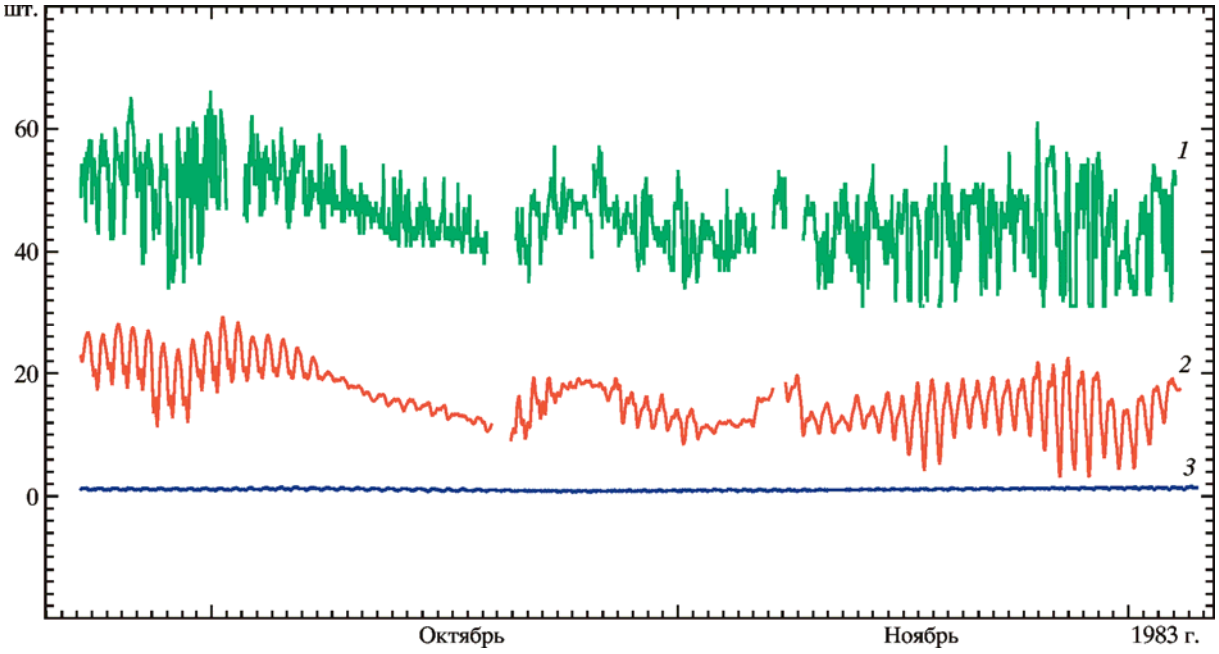

Рис. 4. Выделение суточного ритма активности биоиндикатора SLON2M
1 – исходный сигнал; *2* – суточный ритм, выделенный методом наложения эпох в скользящем окне шириной 731 ч (модель (9)) с подгонкой амплитуды в скользящем окне шириной 73 ч (модель (10)); *3* – погрешность оценки среднего ритма. Для наглядности кривая *1* смещена вверх

сглаженный средний ритм использовался для аппроксимации наблюдаемых вариаций активности согласно модели (10). Коэффициенты модели (10) оценивались в скользящем окне шириной 72 ч. Параметр η при всех вычислениях был задан равным $\eta = 60$ %.

Результат выделения суточного ритма $\hat{\mathbf{S}}$ показан на рис. 4 (линия *2*). Хорошо видно, что форма ритма непостоянна, а его амплитуда испытывает значимые изменения. При этом погрешность оценки суточного ритма, оцениваемая как стандартное отклонение каждого значения $\hat{\mathbf{S}}$ (линия *3*), на удивление невелика. Большую часть времени она составляет около 10% от амплитуды ритма. Хотя вполне очевидно, что для данного сигнала вполне правомерным выглядело бы и использование других алгоритмов выделения суточного ритма (например, основанных на частотной или адаптивной фильтрации), и что такие алгоритмы привели бы к построению иных моделей суточной ритмики, существенно отличающихся от представленной.

При наличии пропусков данных различия между моделями могут существенно возрастать. В этой связи интересно рассмотреть интервал 6–9 ноября. Как видно из данных, приведенных на рис. 4, средний уровень отфильтрованного сигнала в этот период выше, чем в другие дни, так как использованная модель ритма включает трендовую составляющую в ритм. Очевидно, повышение уровня связано с тем, что ширина окна при построении модели (10) была выбрана равной 72 ч, что очень близко к длительности интервала пропусков. По этой причине уровень среднего ритма (кривая *2*) на данном интервале определяется уровнем исходного сигнала в небольшом промежутке времени между пропусками (линия *1*). Если увеличить ширину окна оценивания модели (10) или уменьшить значение параметра η хотя бы до 40 %, эффект пропадает, что говорит о его недостаточной устойчивости. Однако погрешность оценки ритма, рассчитанная согласно модели (10) (см. рис. 4, линия *3*), в этот период лишь незначительно превышает

обычный уровень. Этот пример наглядно показывает, что критерии значимости, основанные на внутренней сходимости алгоритма, не всегда достаточно эффективны.

Анализ фазы суточного ритма

Одна из важнейших характеристик ритма – его фаза. На рис. 5 показаны акрофазы максимальной активности биоиндикатора SLON2M (для наглядности выбран небольшой фрагмент ряда). При расчетах использован параметр $\eta = 40$ %, т.е. расчет акрофазы выполнялся даже для тех суток, когда пропуски составляли 2/5 от всех значений. В случае неустойчивого зашумленного ритма такая высокая терпимость к пропускам обычно приводит к сильному дребезгу графика акрофазы, так как ее значения постоянно перескакивают с одного часа на другой. Этот эффект хорошо виден по акрофазе исходного (неотфильтрованного) сигнала (рис. 5, *а*). После выделения ритма по описанному выше алгоритму его акрофаза оценивается гораздо стабильнее (рис. 5, *б*).

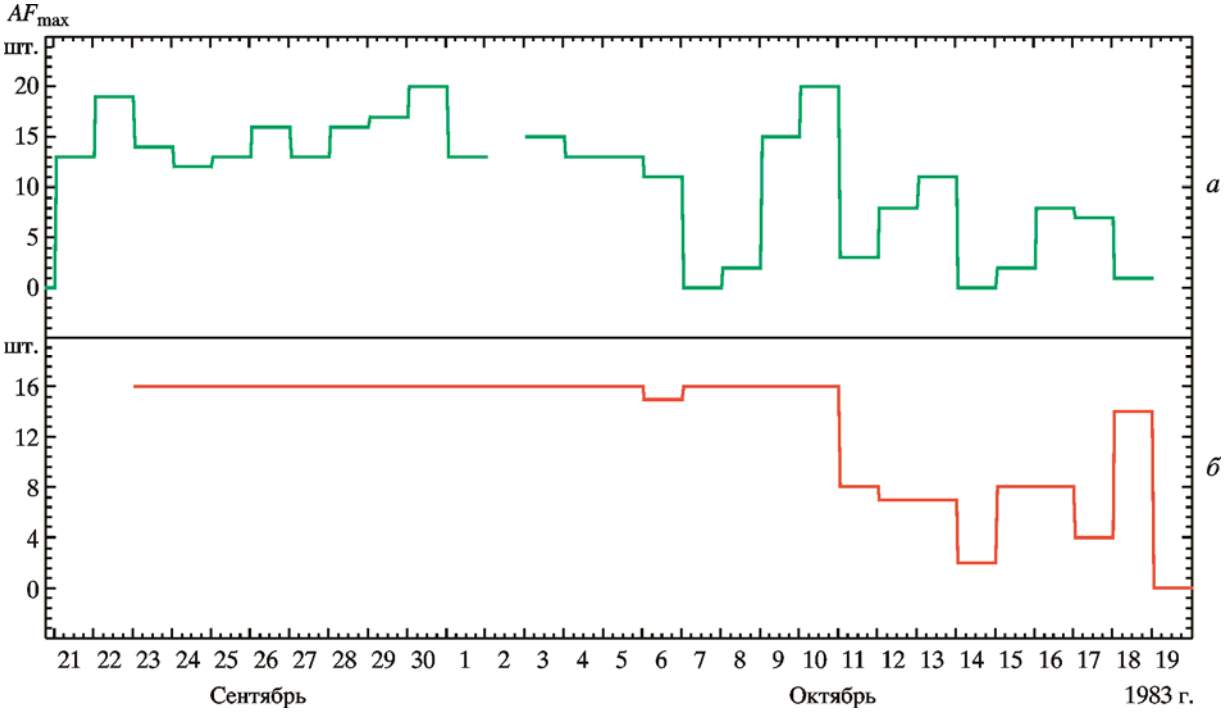

Рис. 5. Акрофаза суточной активности для исходного SLON2M (*1*) и отфильтрованного (*2*) суточного ритма

Однако графики акрофаз иллюстрируют лишь одну из фазовых характеристик сигнала. Наряду с акрофазой интерес могут представлять батифаза, смещение времени проявления других особенностей ритма, изменение формы вариации и т.д. Кроме того, анализ графиков акрофаз не всегда удобен из-за того, что периодическая функция отображается в линейной шкале. Изменение акрофазы на величину, близкую к 2π, отображается на таком графике сильным скачком, хотя истинная фаза при этом почти не меняется.

Более информативный и удобный для визуального анализа способ отображения фазовых характеристик ритма дают так называемые годографы Рэлея–Шустера [*Сидорин,* 2009]. Суть этого метода состоит в том, что фаза каждого колебания откладывается на единичной плоскости в виде вектора, направление которого показывает значение фазы.

Затем от конца этого вектора строится следующий вектор, показывающий фазу ритма в следующий момент времени, и т.д. В результате вычислений строится годограф Рэлея–Шустера, показывающий изменения фазы ритма на всем протяжении ряда.

Мы применили для построения годографов усовершенствованный алгоритм Рэлея–Шустера, предусматривающий условную нормировку суточных векторов [*Дещеревский, Сидорин,* 2015а]. Построенный годограф отражает исключительно вариации фазы ритма, полностью исключая влияние амплитуды сигнала. В то же время использование указанной нормировки обеспечивает робастность, т.е. устойчивость работы алгоритма не нарушается даже в тех случаях, когда амплитуда суточного вектора близка к нулю и возникает опасность мультипликации случайных шумов.

Рассчитанные годографы приведены на рис. 6. На графиках помечены даты изломов годографа. Каждый такой излом соответствует определенному изменению фазовых характеристик ритма. Видно, что годограф исходного сигнала (рис. 6, *а*) более зашумлен, а после фильтрации ритма годограф становится более плавным (рис. 6, *б*). Имеются и другие отличия. В настоящей работе мы не имеем возможности обсуждать, насколько значимы показанные на рис. 6 изменения фазы ритма, поскольку для этого требуется более глубокий анализ [*Дещеревский, Сидорин,* 2015а]. Но можно с уверенностью утверждать, что изменения годографа, произошедшие после фильтрации ритма, обусловлены не только подавлением шума, но и спецификой использованной модели ритма.

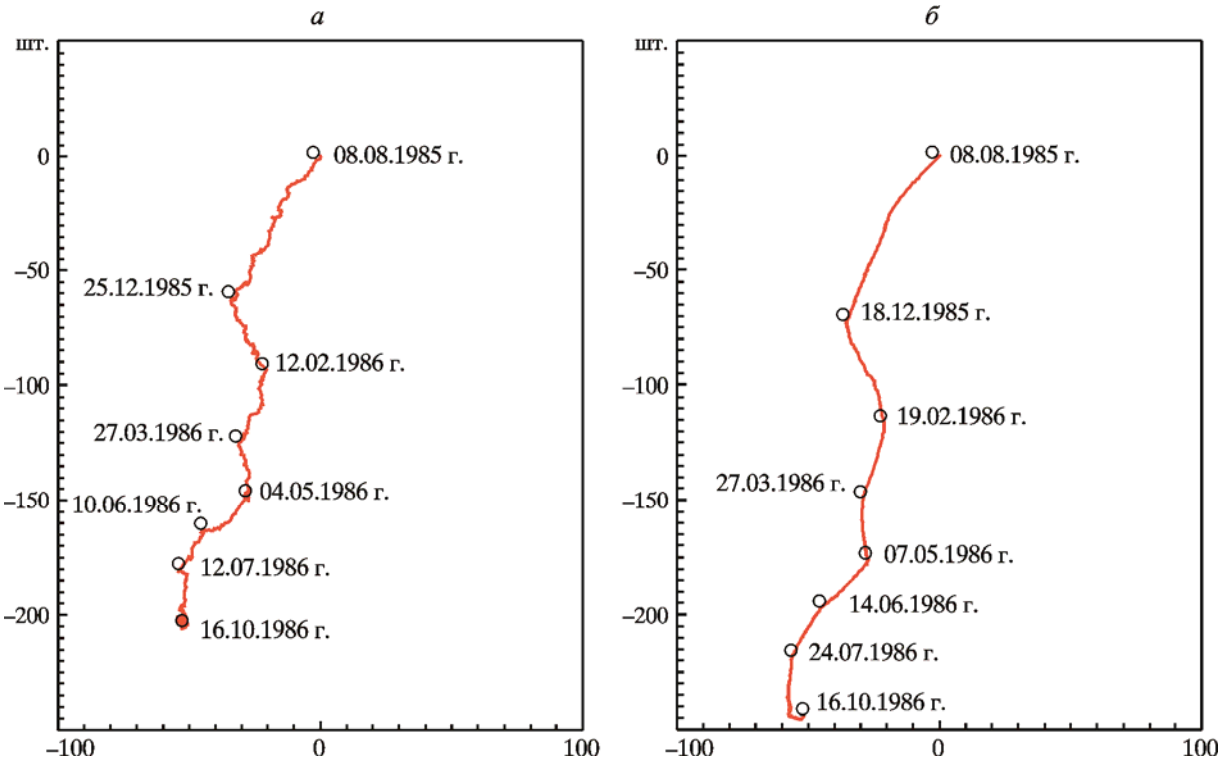

Рис. 6. Годограф Рэлея–Шустера для суточной активности биоиндикатора SLON2M
а – исходный ряд; *б* – отфильтрованный суточный ритм. Помечены даты изломов годографа

Выше уже отмечалось, что существует большое количество различных методов выделения ритмов, причем каждый такой алгоритм реализует собственную модель ритма. Поэтому оценка любой характеристики ритма, выполняемая в рамках некоторой модели, является условной. Чтобы дать максимально полное и объективное описание ритма, необходимо опробовать несколько разных моделей.

Работа с каталогами землетрясений

Каталог землетрясений – это набор записей о временах, координатах и силе событий. Временной интервал между событиями достаточно произволен. Формально такие ряды можно превратить в непрерывные сглаживанием, после чего для их обработки пригодны все методы, применяемые для обычных временных рядов. Но это противоречит структуре такого ряда. Так, в мировом каталоге сильных землетрясений среднее число событий в сутки колеблется от одного до нескольких штук. Поэтому осреднение внутри текущих суток или вычитание среднесуточного хода тут не имеет строго смысла, а при более сильном окне сглаживания невозможно оценить параметры суточной вариации потока землетрясений. Конечно, можно оценить средний суточный ход по всему каталогу методом наложения эпох. В этом случае на каждый час придется достаточное число наблюдений, т.е. можно будет сформировать весьма плотный каталог, для которого будет иметь смысл операция усреднения даже при небольшой ширине окна. Этот подход использовался, например, при расчете БПФ-спектров для оценки сезонных эффектов и фрактальности [*Щербина,* 2004], а также моделей типа ARIMA [*Журавлев,* 1983; *Любушин,* 2007]. При такой работе, однако, полностью выпадает исходно дискретная сущность сейсмических событий.

Модель данных, допускающая наличие пропусков, позволяет преодолеть конфликт между исходно дискретной структурой сигнала и желанием обрабатывать подобные наблюдения как обычный временной ряд. Современные персональные компьютеры обладают достаточной мощностью, чтобы без каких-либо проблем анализировать многолетние временные ряды с периодичностью измерений один раз в минуту и даже чаще. При решении подавляющего большинства практически важных задач такая точность временной привязки дискретных событий более чем достаточна. Представление каталога в форме временного ряда позволяет применять для анализа сейсмических данных огромное количество неспецифических алгоритмов, разработанных для обычных временных рядов. Одновременно возникают широчайшие возможности для совместного анализа сейсмических данных и обычных сигналов геофизического мониторинга.

Придерживаясь принципа, что программа должна обеспечивать все основные потребности исследователя-геофизика в области временных рядов, мы включили в WinABD оба типа инструментов для работы с данными типа каталогов землетрясений. Каталог можно импортировать во встроенную базу данных и хранить в формате WinABD. Для обработки каталогов в программе имеются специальные методы – такие, как выделение групповых землетрясений, оценка графика повторяемости, построение различных карт и разрезов. Весьма полезна и возможность наложения землетрясений на график обычного ряда.

Кроме того, сейсмический каталог можно представить в виде временного ряда с равномерным опросом во времени с любой периодичностью наблюдений. Для сохранения полной информации о землетрясениях в этом случае генерируется пакет временных рядов, каждый из которых содержит информацию об одной из характеристик сейсмичности: количестве событий, произошедших в течение интервала дискретизации, энергии и координатах события и т.д. Для их анализа можно использовать любые методы, предназначенные для работы с обычными временными рядами. И хотя при обработке подобных сигналов не все стандартные алгоритмы достаточно эффективны, такой подход вполне оправдывает себя при решении многих важных задач [*Дещеревский, Сидорин,* 2011а, б]. В частности, он позволяет совместно анализировать сейсмичность и временные ряды, вести обработку в скользящем окне, использовать специальные алгоритмы для поиска периодичностей, строить модели авторегрессии – скользящего среднего,

вычислять корреляции и т.д. Для примера на рис. 7, *в* приведен график коэффициента корреляции в скользящем окне шириной 1 год для сейсмичности Гармского полигона (рис. 7, *а*) и солнечной активности (рис. 7, *б*). Видно, что в некоторые годы (1955, 1966, 1969, 1987) коэффициент корреляции значительно выше обычного. В другие годы (1971, 1985) корреляция, наоборот, отрицательна. Вопрос о значимости этих эффектов мы в данном случае оставляем за рамками рассмотрения; цель примера состоит в демонстрации совместной обработки дискретных (сейсмичность) и непрерывных (солнечная активность) данных с помощью обычных методов анализа временных рядов.

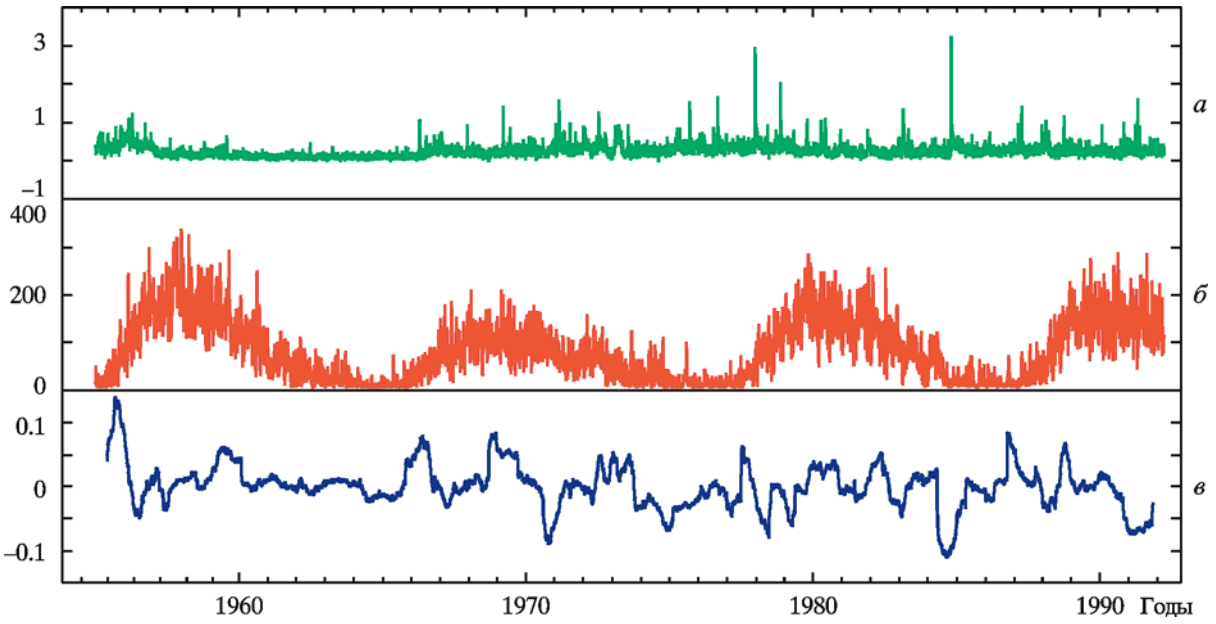

Рис. 7. Почасовое количество землетрясений в каталоге Гармского полигона (*а*), солнечная активность (*б*) и коэффициент корреляции между ними в скользящем окне (*в*)

Остановимся теперь на некоторых особенностях расчетов, выполняемых непосредственно по сейсмическим данным. При оценке спектра каждое событие каталога представляется дельта-функцией:

$$\sum_{j}^{N} A_j e^{i\omega_k t_j}, \tag{11}$$

где t_j – время сейсмического события j, общее число которых N; ω_k – частоты, рассчитываемые через выбираемые периоды T_k как $\omega_k = 2\pi/T_k$. Величина A_j для сейсмических событий считается равной единице. Затем выполняется обычное фурье-преобразование полученного сигнала.

При расчете периодограмм Ломба [*Lomb,* 1976] используется преобразование

$$P_x(\omega) = \frac{1}{2}\left(\frac{\left[\Sigma_j X_j \cos\omega\left(t_j - \tau\right)\right]^2}{\Sigma_j \cos^2 \omega\left(t_j - \tau\right)} + \frac{\left[\Sigma_j X_j \sin\omega\left(t_j - \tau\right)\right]^2}{\Sigma_j \sin^2 \omega\left(t_j - \tau\right)} \right), \tag{12}$$

где τ вычисляется как

$$\tan 2\omega\tau = \frac{\Sigma_j \sin 2\omega t_j}{\Sigma_j \cos 2\omega t_j}, \tag{13}$$

а X принимается равным единице. Понятно, что оба эти подхода не вполне соответствуют классическим представлениям о спектре процессов. Однако практика показывает, что получаемые при этом результаты достаточно информативны и допускают содержательную физическую интерпретацию [*Журавлев, Сидорин,* 2005; *Сидорин,* 2008, 2013, 2015; *Журавлев, Лукк,* 2011]. Примеры использования такого подхода в других областях обсуждаются, например, в работах [*Press, Rybicki,* 1989; *Moody,* 1993; *Press et al.,* 2002].

Для примера на рис. 8 приведены спектры сейсмичности, построенные непосредственно по каталогу землетрясений Греции за 2011 г. (линии *1* и *2*) и по этим же данным, но преобразованным к формату временных рядов с равномерным шагом по времени (линии *3*–*5*). Как видно на рисунке, обычные фурье-спектры, рассчитанные непосредственно по ряду дискретных событий (линия *2*) и по ряду часовых количеств землетрясений (линия *3*), практически идентичны. В этом нет ничего удивительного, ведь построенный временной ряд фактически и представляет собой серию дельта-функций, описывающих сейсмические события, только с чуть более грубой дискретизацией времени.

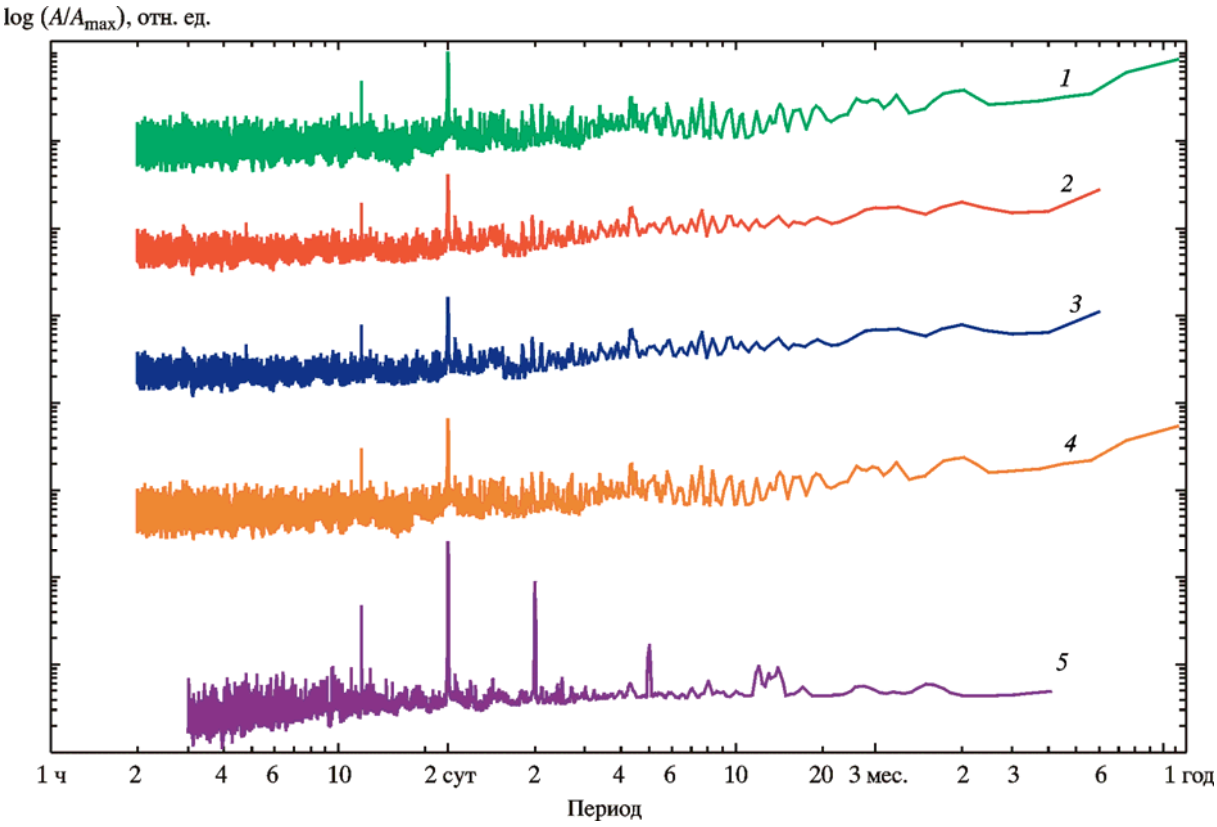

Рис. 8. Спектры Ломба (*1*) и Фурье (*2*), рассчитанные по сейсмическому каталогу дискретных событий, спектр Фурье (*3*), БПФ-спектр (*4*) и периодограмма Аббе (*5*), рассчитанные по временному ряду почасового количества землетрясений

Гораздо более сильные различия в спектрах могут появляться при использовании разных модификаций спектрального метода. Так, спектр, оцененный методом БПФ (линия *4* на рис. 8), определенно не идентичен классическому фурье-спектру (линия *3* на рис. 8), несмотря на то, что оба они рассчитаны по одному и тому же временному ряду. В частности, на БПФ-спектре отсутствует особенность на периоде 4.8 ч (кратная гармоника 1/5 суток). Еще сильнее отличается спектр, оцененный методом Ломба

[*Lomb,* 1976] (см. рис. 8, линия *1*). Он не только не содержит особенностей на периоде 4.8 ч, но и более зашумлен, чем обычный фурье-спектр (см. рис. 8, линия *2*), что хорошо видно в области высоких частот. Впрочем, это частично компенсируется более высокой амплитудой суточного и полусуточного пика на спектре Ломба. Поэтому отношение сигнал/шум для этих пиков на периодограмме Ломба по сравнению со спектром Фурье ухудшается незначительно: для спектра Ломба (см. рис. 8, линия *1*) оно равно 1.4 и 1.7, а для обычного фурье-спектра (см. рис. 8, линия *2*) – 1.8 и 2.6. Очевидно, это связано с тем, что спектральная модель, на которой основан алгоритм Ломба, выполняет неявную интерполяцию ряда между событиями, а для реальных каталогов это не всегда приводит к улучшению отношения сигнал/шум. Вместе с тем основные параметры спектров Фурье и периодограмм Ломба, в частности суточная и полусуточная гармоники, общий вид спектра, практически идентичны, что свидетельствует об объективности результатов.

Наиболее сильно отличается от остальных спектров периодограмма Аббе (см. рис. 8, линия *5*). Это, впрочем, неудивительно, поскольку алгоритм Аббе основан на принципиально иной модели периодичности [*Дещеревский, Сидорин,* 2011а, б]. Стоит заметить, что метод периодограмм, так же как и упоминаемый выше косинор-метод, позволяет выполнять альтернативное фурье-разложению разложение сигнала на периодические компоненты [*Дещеревский, Лукк,* 2002]. Однако метод периодограмм является более общим, поскольку базис разложения не обязательно формируется из гармонических функций, а строится итеративно непосредственно по анализируемому сигналу.

Завершая этот раздел, отметим, что реальные каталоги, как и данные других видов мониторинга, очень часто содержат многочисленные дефекты [*Дещеревский, Сидорин,* 2014в]. Это надо учитывать при выборе алгоритмов обработки таких наблюдений.

Обсуждение

Как правило, при организации геофизического и иного долговременного мониторинга выдвигается требование, чтобы измерения выполнялись с заданной регулярностью, через равные промежутки времени. Однако на практике приблизиться к этому идеалу почти невозможно. Периодически приходится останавливать наблюдения, чтобы выполнить калибровку, наладку и ремонт аппаратуры, замену отдельных узлов и т.д. Даже если измерения ведутся по утвержденному графику, некоторые значения (выбросы и т.п.) приходится выбраковывать из-за резких искажений сигнала, возникающих под влиянием различных экстремальных помех. В результате экспериментальный сигнал выглядит как ряд с равномерным шагом опроса данных, но с пропуском некоторой части измерений, иногда довольно значительной. Опыт показывает, что доля пропущенных измерений при геофизическом или биологическом мониторинге обычно варьирует от нескольких единиц до десятков процентов.

Большинство алгоритмов, используемых при обработке экспериментальных сигналов, предполагает, что временной интервал между всеми наблюдениями одинаков. Такие вычислительные схемы не могут применяться для рядов, содержащих пропуски данных. По этой причине обработка данных мониторинга обычно начинается с заполнения пропусков. Затем применяются типовые алгоритмы статистического анализа, реализованные в стандартных программных средствах.

Такой подход является фактически общепринятым, однако он имеет очевидные недостатки. Требование «полноты» обрабатываемого сигнала означает, что выбраковка сомнительных и аномальных значений и их заполнение «правдоподобными» данными должны выполняться уже на начальном этапе анализа наблюдений. Но опыт показыва-

ет, что ошибочные значения гораздо точнее и надежнее выявляются по отфильтрованному и преобразованному сигналу, т.е. на поздних стадиях обработки. Если заполнение пропусков предшествует любым вычислениям, очень сложно определить, какие из используемых процедур наиболее чувствительны к дефектам данных и насколько сильно результаты выполнения тех или иных операций зависят от алгоритма интерполяции пропусков. Эти проблемы актуальны даже в том случае, если пропуски заполнены идеально. Однако свойства многих экспериментальных процессов довольно сложны и/или плохо изучены, что приводит к серьезным ошибкам и искажениям сигнала при интерполяции пропущенных наблюдений [*Рыженкова,* 2011].

На самом деле широкое распространение подхода, основанного на заполнении пропусков данных, обусловлено не столько какими-то содержательными преимуществами, сколько простотой его технической реализации [*Моисеев,* 1998]. При всем разнообразии возможных подходов к заполнению пропусков в коммерческом программном обеспечении реализована лишь небольшая часть этих методов, а на практике чаще всего используются лишь самые простые алгоритмы, не обеспечивающие нужного качества интерполяции [*Злоба, Яцкив,* 2002]. Мы считаем, что при обработке сигналов с пропусками меньшим из зол является альтернативный подход, предусматривающий не заполнение пробелов в данных, а исключение таких точек из вычислений [*Скрипник и др.,* 1988; *Литтл, Рубин,* 1991]. Для достижения этой цели пропущенные наблюдения кодируются особым числом, а каждый используемый алгоритм модифицируется таким образом, чтобы обеспечить оптимальную обработку сигнала с пропусками в рамках данной вычислительной процедуры.

Разумеется, этот подход также не свободен от недостатков. Основной состоит в том, что теоретически исследовать свойства подобных модифицированных алгоритмов в общем случае практически невозможно. Более того, при отсутствии ограничений на количество и структуру пропусков почти для всех типовых алгоритмов можно сконструировать контрпримеры, доказывающие отсутствие тех или иных оптимальных свойств у модифицированного алгоритма.

Впрочем, аналогичные контрпримеры часто можно сконструировать и для алгоритмов, обрабатывающих сигналы с заполненными пропусками наблюдений. При исследовании уникальных экспериментальных рядов, получаемых при долговременном мониторинге, сплошь и рядом оказывается, что статистические характеристики процесса доподлинно неизвестны, а в данных имеются явные признаки нестационарности [*Дещеревский и др.,* 1987; *Дещеревский, Сидорин,* 2003а, 2004] и другие нетривиальные особенности. В такой ситуации любая процедура искусственного заполнения пропусков может вносить в обрабатываемые сигналы непредсказуемые искажения и эффекты [*Моисеев,* 1998]. В результате даже классические алгоритмы с доказанными оптимальными свойствами часто не могут обеспечить получение достоверного результата. Понятно, что это связано не с недостатками таких алгоритмов, а с тем, что обрабатываемые данные даже после заполнения пропусков не удовлетворяют тем требованиям, которые обусловливают указанные оптимальные свойства [*Хампель и др.,* 1989].

На практике анализ подобных рядов невозможно свести к формальному применению строгих алгоритмов и методов. Исследователь в любом случае вынужден принимать априорные допущения и верифицировать их, опробовать различные модели и алгоритмы, творчески интерпретировать результаты. Подход, предусматривающий исключение дефектных значений из вычислений (вместо их заполнения), дает исследователю большое количество дополнительных инструментов анализа данных, которые очень сложно реализовать в рамках стандартных вычислительных схем, ориентированных исключи-

тельно на работу с сигналами без дефектов. В частности, обеспечивается возможность гораздо более гибкой обработки пропусков данных, включая индивидуальную подстройку параметров каждого алгоритма в зависимости от характеристик сигнала. Дополнительный «бонус» описанной технологии – это возможность выполнения различных вычислений в скользящем окне (например, частотной фильтрации) без уменьшения длины ряда.

Описанный подход реализован нами в программе WinABD, предназначенной для сопровождения и анализа данных режимных наблюдений. При умеренном количестве пропусков качество работы таких оптимизированных (адаптированных к наличию пропусков) алгоритмов обычно не вызывает сомнений. По мере роста количества пропусков дисперсия оценок растет, а иногда становится заметным и их смещение. Это неудивительно – при обработке неидеальных данных невозможно получать идеальные результаты. В конечном счете, никакая программа не может сама решить все вопросы, возникающие при анализе таких наблюдений. Подход, предусматривающий исключение дефектных значений из вычислений (вместо их заполнения), требует от исследователя повышенного внимания к вопросам оценки значимости различных эффектов, но одновременно открывает дополнительные возможности верификации получаемых результатов.

Заключение

Большая часть имеющегося на рынке программного обеспечения может работать только с бездефектными временными рядами. Прежде чем приступить к анализу данных, необходимо заполнить пропуски наблюдений вычисляемыми значениями. В статье рассмотрен альтернативный подход, предусматривающий исключение дефектных значений из вычислений (вместо заполнения пропусков), и продемонстрированы особенности применения некоторых таких алгоритмов на реальных рядах. Этот подход может быть реализован только при наличии специальных программных средств, к числу которых относится разработанная нами программа WinABD. В каждом методе задача обработки пропущенных наблюдений решается по-своему, в зависимости от применяемого алгоритма. Это приводит к дополнительным затратам ресурсов при вычислениях и требует известного усложнения алгоритмов. Взамен исследователь получает такие инструменты анализа данных, которые очень сложно реализовать при традиционных подходах.

Для полнофункциональной работы с сигналами, содержащими пропуски наблюдений, необходимо обеспечить особое кодирование пропусков и их корректную обработку при любых операциях, начиная от хранения данных и кончая выполнением вычислений и визуализацией результатов. Если каждая группа задач решается с помощью специализированных программ, возникает проблема организации их взаимодействия.

В WinABD весь перечисленный функционал реализован в рамках одной программы. Система управления базой данных обеспечивает работу с временными рядами, имеющими периодичность наблюдений от лет до микросекунд. При всех операциях с данными используется календарное время, а не условные «точки». Программа позволяет анализировать структуру рядов, выявлять зависимости и взаимосвязи между процессами [*Дещеревский, Сидорин,* 1996, 2003а, 2004, 2012, 2013а, б, 2014б; *Дещеревский и др.,* 1996, 2005, 2009]. Имеется большое количество нестандартных инструментов и методов, необходимых в повседневной работе с неидеальными данными [*Сидорин,* 2009, 2013; *Дещеревский, Сидорин,* 2011а, б, 2015а, б]. Особое развитие получила технология

скользящего временного окна, что позволяет изучать развитие всех процессов во времени и выявлять изменения, связанные с какими-либо событиями [*Дещеревский, Сидорин,* 1996, 2003б; *Дещеревский и др.,* 1996]. Данное программное обеспечение уникально еще и тем, что позволяет одновременно анализировать обычные временные ряды и ряды, состоящие из одиночных событий, такие как каталоги землетрясений [*Дещеревская, Сидорин,* 2004, 2005, 2015; *Журавлев, Сидорин,* 2005; *Дещеревский, Сидорин,* 2008, 2014а, 2015в; *Журавлев, Лукк,* 2011; *Дещеревский и др.,* 2016в].

Опробование подхода, опирающегося на работу только с реальными измерениями, может представлять интерес при решении широкого круга научных задач. Однако наиболее перспективно его применение при анализе экспериментальных временных рядов, получаемых при мониторинге различных процессов [*Дещеревский и др.,* 2016а].

Литература

Абраменкова И.В., Круглов В.В. Методы восстановления пропусков в массивах данных // Программные продукты и системы. 2005. № 2. С. 18–22.

Автоматизированная обработка данных на Гармском полигоне / Отв. ред. А.Я. Сидорин. М.: ИФЗ АН СССР, 1991. 216 с.

Богдасарян Р.А. Частотно-индивидуальный косайнор анализ. Частота биоритмов – критерий раннего выявления патологии: Методические рекомендации. Ереван, 1980. 55 с.

Бокс Дж., Дженкинс Г. Анализ временных рядов. В 2-х т. М.: Мир, 1974. 405 с. 197 с.

Валеев Р.Т. Метод взвешенного скользящего среднего и математическая модель «японских свечек» в условиях фондового рынка и их применение для его анализа: Дис. … канд. техн. наук. Томск: Том. гос. ун-т, 2001. 113 с.

Гармский геофизический полигон / Отв. ред. А.Я. Сидорин. М.: ИФЗ АН СССР, 1990. 240 с.

Грачев А.В. К восстановлению пропусков в экспериментальных данных // Вестн. ННГУ им. Н.И. Лобачевского. Сер. Радиофизика. 2004. Вып. 2. С. 15–23.

Губанов В.А. Анализ воздействия выбросов на результат сезонной корректировки временных рядов // Научные труды: Институт народнохозяйственного прогнозирования РАН. 2004. № 2.

Дещеревская Е.В., Сидорин А.Я. Некоторые результаты изучения суточной периодичности землетрясений Гармского полигона // Сейсмические приборы. 2004. Вып. 40. С. 57–70.

Дещеревская Е.В., Сидорин А.Я. Внутрисезонные колебания сейсмичности Гармского полигона и их связь с атмосферными процессами // Докл. РАН. 2005. Т. 401, № 1. С.80–83.

Дещеревская Е.В., Сидорин А.Я. Пространственно-временные особенности суточной периодичности слабых землетрясений Гармского полигона // Вопросы инженерной сейсмологии. 2015. Т. 42, № 4. С. 77–84.

Дещеревский А.В., Лукк А.А. Выделение регулярных составляющих во временных вариациях геофизических параметров методом разложения на негармонические компоненты // Вулканология и сейсмология. 2002. № 5. С. 65–78.

Дещеревский А.В., Сидорин А.Я. Алгоритмы и результаты формализованного поиска предвестников землетрясений во временных рядах геоэлектрического мониторинга // Федеральная система сейсмологических наблюдений: Инф.-аналит. бюл. 1996. Т. 3, № 3. С. 11–27.

Дещеревский А.В., Сидорин А.Я. Некоторые вопросы методики оценки среднесезонных функций для геофизических данных. М.: ОИФЗ РАН, 1999. 40 с.

Дещеревский А.В., Сидорин А.Я. База данных биологического мониторинга на Гармском полигоне // Геофизические процессы и биосфера. 2002. Т. 1, № 2. С. 3–15.

Дещеревский А.В., Сидорин А.Я. Проблема фликкер-шума при изучении причинно-следственных связей между природными процессами // Докл. РАН. 2003а. Т. 392, № 3. С. 392–396.

Дещеревский А.В., Сидорин А.Я. Параметризация временных рядов активности животных для геофизических исследований // Моделирование геофизических процессов. М.: ОИФЗ РАН, 2003б. С. 137–155.

Дещеревский А.В., Сидорин А.Я. Исследование значимости корреляции электрической активности рыб и электротеллурического поля // Биофизика. 2004. Т. 49, вып. 4. С. 715–722.

Дещеревский А.В., Сидорин А.Я. Поиск корреляции сейсмичности Гармского полигона с атмосферным давлением и скоростью ветра // Геофизические исследования. 2008. Т. 9, № 1. С. 3–15.

Дещеревский А.В., Сидорин А.Я. Периодограммы наложенных эпох при поиске скрытых ритмов в экспериментальных рядах // Сейсмические приборы. 2011а. Т. 47, № 2. С. 21–43.

Дещеревский А.В., Сидорин А.Я. Сравнение периодограмм наложенных эпох и спектров Фурье экспериментальных рядов // Сейсмические приборы. 2011б. Т. 47, № 3. С. 44–70.

Дещеревский А.В., Сидорин А.Я. Поиск влияния гравитационных приливов на региональную сейсмичность Греции разными методами: 1. Спектральный и периодограммный анализ // Сейсмические приборы. 2012. Т. 48, № 4. С.5–26.

Дещеревский А.В., Сидорин А.Я. Поиск влияния гравитационных приливов на региональную сейсмичность Греции разными методами: 2. Корреляционный анализ // Сейсмические приборы. 2013а. Т. 49, № 1. С. 35–59.

Дещеревский А.В., Сидорин А.Я. Поиск влияния гравитационных приливов на региональную сейсмичность Греции разными методами: 3. Корреляция с солнечным и лунным компонентами прилива // Сейсмические приборы. 2013б. Т. 49, № 3. С. 41–53.

Дещеревский А.В., Сидорин А.Я. Суточная периодичность землетрясений Южной Калифорнии // Сейсмические приборы. 2014а. Т. 50, № 1. С. 27–50.

Дещеревский А.В., Сидорин А.Я. Оценка корреляции потока землетрясений Аляски с лунно-солнечным приливом // Наука и технологические разработки. 2014б. Т. 93, № 1. С. 29–44.

Дещеревский А.В., Сидорин А.Я. Технические проблемы и ошибки при работе с каталогами землетрясений // Наука и технологические разработки. 2014в. Т. 93, № 4. С. 32–41.

Дещеревский А.В., Сидорин А.Я. Повышение робастности и устойчивости оценок параметров годографов Рэлея–Шустера с помощью различных способов нормировки векторов // Сейсмические приборы. 2015а. Т. 51, № 2. С. 56–80.

Дещеревский А.В., Сидорин А.Я. Тестирование метода годографов Рэлея–Шустера на модельных временных рядах и потоках землетрясений // Сейсмические приборы. 2015б. Т. 51, № 3. С. 59–84.

Дещеревский А.В., Сидорин А.Я. Анализ суточной периодичности в Душанбино-Вахшском каталоге землетрясений методом годографов Рэлея–Шустера // Вопросы инженерной сейсмологии. 2015в. Т. 42, № 3. С. 71–92.

Дещеревский А.В., Журавлев В.И., Сидорин А.Я. Некоторые алгоритмы фильтрации для геофизических временных рядов // Физика Земли. 1996. № 2. С. 56–67.

Дещеревский А.В., Лукк А.А., Сидорин А.Я. Признаки фликкер-шумовой структуры во временных реализациях геофизических полей // Физика Земли. 1997. № 7. С. 3–19.

Дещеревский А.В., Сидорин А.Я., Харин Е.П. Исследование влияния гелиогеофизических факторов на активность животных в лабораторных условиях // Докл. РАН. 2005. Т. 401, № 6. С. 837–841.

Дещеревский А.В., Сидорин А.Я., Харин Е.П. Геомагнитные возмущения и активность животных в лабораторных условиях // Биофизика. 2009. Т. 54, вып. 3. С. 554–562.

Дещеревский А.В., Журавлев В.И., Никольский А.Н., Сидорин А.Я. Технологии анализа геофизических временных рядов. Ч. 1. Требования к программе обработки // Сейсмические приборы. 2016а. Т. 52, № 1. С. 61–82

Дещеревский А.В., Журавлев В.И., Никольский А.Н., Сидорин А.Я. Технологии анализа геофизических временных рядов. Ч. 2. WinABD – пакет программ для сопровождения и анализа данных геофизического мониторинга // Сейсмические приборы. 2016б. Т. 52, № 3. С.51–80.

Дещеревский А.В., Мирзоев К.М., Лукк А.А. Критерии группирования землетрясений с учетом пространственной неоднородности сейсмичности // Физика Земли. 2016в. № 1. С. 79–97.

Емельянов И.П. Формы колебаний в биоритмологии. Новосибирск: Наука, 1976. 128 с.

Журавлев В.И. Моделирование сейсмического режима уравнением авторегрессии // Экспериментальная сейсмология. М.: Наука, 1983. С. 99–108.

Журавлев В.И., Лукк А.А. Полуденная активизация сейсмичности в Турции и ряде других регионов мира // Геофизические исследования. 2011. Т. 12, № 4. С. 31–57.

Журавлев В.И., Сидорин А.Я. Спектральные исследования суточной периодичности землетрясений Гармского полигона // Геофизические исследования. 2005. Вып. 1. С. 48–57.

Злоба Е., Яцкив И. Статистические методы восстановления пропущенных данных // Computer Modelling & New Technologies. 2002. V. 6, N 1. С. 51–61.

Инструкция к программе «Cosinor Ellipse 2006». Омск: ООО Научно-методический центр «Аналитик», 2015. 13 с.

Карлов И.А., Проворова О.Г. Новый подход к исследованию устойчивости алюминиевого электролизера // Вестн. Краснояр. гос. ун-та. Физико-математические науки. 2002. № 1. С. 116–120.

Карп В.П., Катинас Г.С. Математические методы исследования биоритмов // Хронобиология и хрономедицина / Под ред. Ф.И. Комарова. М.: Медицина, 1989. С. 29–45.

Концевая Н.В. Оптимизация процедур сглаживания показателей финансовых рынков // Аудит и финансовый анализ. 2011. № 1. С. 122–127.

Концевая Н.В. Анализ методов заполнения пропусков во временных рядах показателей финансовых рынков // Вестн. Воронеж. гос. техн. ун-та. 2012. Т. 8, № 8. С. 18–20.

Концевая Н.В. Скользящий β-коэффициент как инструмент оптимизации торговых стратегий на примере валютного рынка // Вестн. Финансового ун-та. 2013. № 1. С. 73–81.

Корягина Ю.В., Нопин С.В. Cosinor Ellipse 2006. № 2006611345 // Программы для ЭВМ… (офиц. бюл.). 2006. № 3 (56). С. 42.

Кучеров И.С., Ткачук В.Г., Волков А.В. Длительные биологические ритмы в динамике мышечной работоспособности человека // Кибернетика и вычислительная техника. М.: Наука, 1970. Вып. 7. С. 71–77.

Лагутин М.Б. Наглядная математическая статистика. 2-е изд., испр. М.: БИНОМ, 2009. 472 с.

Литтл Р.Дж.А., Рубин Д.Б. Статистический анализ данных с пропусками. М.: Финансы и статистика, 1990. 336 с.

Любушин А.А. Анализ данных систем геофизического и экологического мониторинга. М.: Наука, 2007. 228 с.

Макс Ж. Методы и техника обработки сигналов при физических измерениях. В 2-х т. М.: Мир, 1983. Т. 1. 312 с.

Маркелов О.А. Система информационной поддержки принятия решений врача при лечении вегетативных расстройств: Дис. … канд. техн. наук. СПб. гос. электротехн. ун-т «ЛЭТИ» им. В.И. Ульянова (Ленина), 2007.

Маркин А.В., Щербаков М.В. Метод автоматического восстановления значений в потоках данных на основе взвешенной модели // Прикаспийский журнал: управление и высокие технологии. 2013. № 3. С. 49–54.

Моисеев С.Н. Заполнение пропусков в случайно-цензурированных временных рядах // Автометрия. 1998. № 1. С. 61–66.

Оранский И.Е., Царфис П.Г. Биоритмология и хронотерапия. М.: Высш. шк., 1989. 159 с.

Разумихин Д.В. Использование нейронных сетей на уровне семантики в системе распознавания речи // IV Всерос. конф. «Нейрокомпьютеры и их применение», г. Москва, 6–18 февраля 2000 г. М.: ИПУ РАН, 2000. С. 208–210.

Россиев А.А. Итерационное моделирование неполных данных с помощью многообразий малой размерности. Красноярск: КГТУ, 2000. 84 с.

Рыженкова К.В. Методы восстановления пропуска данных при проведении статистических исследований // Интеллект. Инновации. Инвестиции. 2011. № 3. С. 127–133.

Сидорин А.Я. Суточная периодичность сильных землетрясений Гармского полигона // Сейсмические приборы. 2008. Т. 44, № 3. С. 70–76.

Сидорин А.Я. О применении метода Рэлея–Шустера в исследованиях периодичности землетрясений // Сейсмические приборы. 2009. Т. 45, № 3. С. 29–40.

Сидорин А.Я. Различия внутрисуточных фазовых диаграмм потоков землетрясений разной энергии // Сейсмические приборы. 2013. Т. 49, № 2. С. 71–84.

Сидорин А.Я. Техногенная суточная периодичность сейсмических событий в районе Нурекского водохранилища // Наука и технологические разработки. 2015. Т. 94, № 2. С. 28–44.

Скрипник В.М., Назин А.Е., Приходько Ю.Г., Благовещенский Ю.Н. Анализ надежности технических систем по цензурированным выборкам. М.: Радио и связь, 1988. 183 с

Снитюк В.Е. Эволюционный метод восстановления пропусков в данных // Тр. VI Междунар. конф. «Интеллектуальный анализ информации». Киев, 2006. С. 262–271.

Теребиж В.Ю. Анализ временных рядов в астрофизике. М.: Наука, 1992. 392 с.

Урбах В.Ю. Математическая статистика для биологов и медиков. М.: Изд-во АН СССР, 1963. 322 с.

Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. Робастность в статистике: Подход на основе функций влияния. М.: Мир, 1989. 512 с.

Хардле В. Прикладная непараметрическая регрессия. М.: Мир, 1993. 349 с.

Хемминг Р.В. Цифровые фильтры. 2-е изд. М.: Недра, 1987. 221 с.

Хемминг Р.В. Численные методы для научных работников и инженеров. М.: Наука, 1972. 400 с.

Щербина С.В. Экспериментальное исследование динамического хаоса в сейсмогенной среде // Геофиз. журн. 2004. Т. 26, № 3. С. 125–131.

Baghi Q., Gilles M., Berge J., Christophe B., Touboul P., Rodrigues M. Regression analysis with missing data and unknown colored noise: Application to the MICROSCOPE space mission // Physical Rev. D. 2015. V. 91, N 6.

Cornelissen G. Cosinor-based rhythmometry // Theoretical Biology and Medical Modelling. 2014. V. 11:16.

Dergachev V.A., Makarenko N.G., Karimova L.N., Danilkina E.B. Nonlinear methods of analysis of data with gaps // Geochronometria. 2001. V. 20. P. 45–50.

Filling data gaps in a periodic timeseries in MATLAB. 2014. URL: http://www.mathworks.com/matlabcentral/newsreader/view_thread/337892

Filling gaps in time series with Nan. 2013. URL: http://www.mathworks.com/matlabcentral/answers/76164-filling-gaps-in-time-series-with-nan

Giles D.E.A. The underground economy: Minimizing the size of government // How to spend the fiscal dividend: The optimal size of government / Ed. by H. Grubel. Vancouver: Fraser Institute, 1998. P. 93–110.

Gorban A., Rossiev A., Makarenko N., Kuandykov Y., Dergachev V. Recovering data gaps through neural network methods // Intern. J. of Geomagnetism and Aeronomy. 2002. V. 3, N 2. P. 191–197.

Guiles M.D. Effect of diurnal data gaps on regression and FFT analysis. SOEST, University of Hawaii at Manoa, 2007. 9 p.

Halberg F., Cornélissen G., Tarquini B., Grafe A., Syutkina E.V., Otsuka K., Watanabe Y., Siegelova J., Sanchez de la Pena S., Carandente F., Schwartzkopff O. Pineals, cancers and feedsidewards in the biosphere and the cosmos (BIOCOS) // Cancer Biotherapy and Radiopharmaceuticals. 1997. V. 12. P. 421–422.

Halberg F. Chronobiology // Ann. Rev. Physiol. 1969. V. 31. P. 675–725.

Halberg F. Chronobiology: Methodological problems // Acta Med. Rom. 1980. V. 18. P. 399–440.

Kanasewich E.R. Time sequence analysis in geophysics. Edmonton [Alta.]: Univ. of Alberta Press, 1981. 480 p.

Katinas G.S. Logistic informative complex of time series analysis // Здоровье и образование в XXI. Сер. Медицина. 2012. Т. 14, № 2. С. 128–133.

Klingenberg B. Regression models for binary time series with gaps // Computational Statistics and Data Analysis. 2008. V. 52. P. 4076–4090.

Lomb N.R. Least-squares frequency analysis of unequally spaced data // Astrophys. and Space Sci. 1976. V. 39. P. 447–462.

Love J.J. Missing data and the accuracy of magnetic-observatory hour means // Ann. Geophys. 2009. V. 27. P. 3601–3610.

Moody G.B. Spectral analysis of heart rate without resampling // Proc. of «Computers in cardiology» conf., London, 5–8 Sept. 1993. IEEE Computer Society, 1993. P. 71–718.

Pashova L., Koprinkova-Hristova P., Popova S. Gap filling of daily sea levels by artificial neural networks // TransNav. 2013. V. 7, N 2. P. 225–232.

Press W.H., Rybicki G.B. Fast algorithm for spectral analysis of unevenly sampled data // Astrophysical J. 1989. V. 338. P. 277–280.

Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P. Numerical recipes in C: The art of scientific computing. 2-nd ed. Cambridge–New York–Port Chester–Melbourne–Sydney: Cambridge Univ. Press, 2002. 925 p.

Refinetti R., Cornelissen G., Halberg F. Procedures for numerical analysis of circadian rhythms // Biological Rhythm Res. 2007. V. 38, N 4. P. 275–325.

Ryan K.F., Giles D.E.A. Testing for unit roots in economic time-series with missing observations // Advances in Econometrics / Ed. by T.B. Fomby & R. Carter Hill. Emerald Group Publ. Ltd., 1998. V. 13 P. 203–242.

Sandip V. George, Ambika G. Effect of data gaps on correlation dimension computed from light curves of variable stars // Astrophys. Space Sci. 2015. V. 360:5.

SAS. Knowledge base. Usage Note 22921. URL: http://support.sas.com/kb/22/921.html

Scargle J.D., Norris J.P., Jackson B., Chiang J. Studies in astronomical time series analysis. VI. Bayesian block representations // The Astrophys. J. V. 764, N 2:167.

Schluter T. Knowledge discovery from time series: Inaugural-dissertation zur Erlangung des doktorgrades der mathematisch-naturwissenschaftlichen fakultat der Heinrich-Heine-Universität Dusseldorf. 2012. 169 p.

Seelam M.R. How do I fill gap in time series data? 2015. URL: https://www.researchgate.net/post/How_do_I_fill_gap_in_time_series_data

Smith J. How to fill data gaps in a time series with NaN. 2014. URL: http://www.mathworks.com/matlabcentral/answers/160053-how-to-fill-data-gaps-in-a-time-series-with-nan

Time series analysis and data gaps. URL: http://epchan.blogspot.ru/2015/07/time-series-analysis-and-data-gaps.html

Torres-Reyna O. Data analysis notes: Links and general guidelines: Online Stata Tutorial. 2014. URL: https://www.princeton.edu/~otorres/Stata/statnotes

Working with missing data. 2015. URL: http://pandas.pydata.org/pandas-docs/stable/missing_ data.html.

Yuan G.-C., Lozier M.S., Pratt L.J., Jones C.K.R.T., Helfrich K.R. Estimating the predictability of an oceanic time series using linear and nonlinear methods // J. Geophys. Res. 2004. V. 109, C08002. doi: 10.1029/2003JC002148.

Сведения об авторах

ДЕЩЕРЕВСКИЙ Алексей Владимирович – кандидат физико-математических наук, ведущий научный сотрудник, Институт физики Земли им. О.Ю. Шмидта РАН. 123242, г. Москва, ул. Большая Грузинская, д. 10, стр. 1. Тел.: +7 (499) 254-90-35. E-mail: adeshere@ifz.ru

ЖУРАВЛЕВ Владимир Ильич – кандидат физико-математических наук, ведущий научный сотрудник, Институт физики Земли им. О.Ю. Шмидта РАН. 123242, г. Москва, ул. Большая Грузинская, д. 10, стр. 1. Тел.: +7 (499) 254-90-35. E-mail: vzhtvertsa@gmail.ru

НИКОЛЬСКИЙ Александр Николаевич – инженер, ООО «КМК Консалтинг». 117420, г. Москва, ул. Наметкина, д. 14. E-mail: anickol@yahoo.com

СИДОРИН Александр Яковлевич – кандидат физико-математических наук, заведующий лабораторией, Институт физики Земли им. О.Ю. Шмидта РАН. 123242, г. Москва, ул. Большая Грузинская, д. 10, стр. 1. E-mail: al_sidorin@hotmail.com

PROBLEMS IN ANALYSIS OF TIME SERIES WITH GAPS AND THEIR SOLUTIONS IN WINABD SOFTWARE PACKAGE

A.V. Desherevskii[1], V.I. Zhuravlev[1], A.N. Nikolsky[2], A.Ya. Sidorin[1]

[1] Schmidt Institute of Physics of the Earth, Russian Academy of Sciences, Moscow, Russia
[2] OOO «CMC Consulting», Moscow, Russia

Abstract. The technologies used for the analysis of time series with gaps are considered. Algorithms for a signal discrimination and estimation of its parameters, in particular discriminated periods, are discussed for time series with missing observation data. Analysis examples of the data obtained during long-term monitoring at the Garm research area and in other regions are given. The solutions used in WinABD software package are considered. They allow a researcher to optimize studies of time series with defects using the considered algorithms.

Keywords: geophysical monitoring data, analysis of time series with gaps, algorithm, software, processing in sliding window, rhythms