УДК 550.3 550.343 004.6 004.9 51.7

WINABD – ПАКЕТ ПРОГРАММ ДЛЯ СОПРОВОЖДЕНИЯ И АНАЛИЗА ДАННЫХ ГЕОФИЗИЧЕСКОГО МОНИТОРИНГА

А.В. Дещеревский[1], В.И. Журавлев[1], А.Н. Никольский[2]

[1] *Институт физики земли РАН,*
[2] *ООО "КМК Консалтинг"*

Введение

Современная геофизика изучает большое число явлений, развивающихся во времени. Экспериментальные ряды данных используются для построения статистической, а затем и физической модели процесса, для контроля происходящих процессов и прогнозирования возможных явлений. При обработке данных мониторинга исследователь решает такие задачи, как:

1) Прием данных, их размещение в базе данных;

2) Формальный и визуальный контроль данных, выбраковка (корректировка) сомнительных наблюдений и различных дефектов данных – скачков, выбросов, узкополосных или высокочастотных помех и т.д.;

3) Выявление и компенсация различных мешающих эффектов, связанных с известными внешними факторами, включая построения моделей их влияния;

4) Анализ данных с использованием различных алгоритмов и методов. Наиболее часто используются:

– методы разведочного анализа, позволяющие построить "базовую" статистическую модель сигнала и оценить ее параметры. На этом этапе необходимы инструменты для выделения и фильтрации трендов различного вида, периодических компонент, оценки функций распределения, автокорреляционных и структурных функций, спектров, фрактальных статистик и т.д.

– методы поиска зависимостей и оценки их параметров: диаграммы корреляционного поля, регрессионные инструменты, взаимные фильтры и др.;

– различные узкоспециализированные методы обработки, применяющиеся для отдельных видов наблюдений;

– инструменты для верификации построенных моделей и оценки значимости эффектов, дополнительные по отношению к обычным критериям значимости – как связанные с численным моделированием, так и иные.

Для решения этих задач часто используются универсальные пакеты статистического анализа временных рядов. Однако в таких пакетах обычно отсутствуют многие важные функции – такие, как сопровождение базы данных, обработка пропусков или календарная синхронизация различных рядов.

При использовании полуоткрытой среды, – такой, как матлаб, – исследователь может использовать различные готовые функции (такие, как расчет спектра или рисование графика), а также самостоятельно разрабатывать те алгоритмы, которые необходимы для решения нестандартных задач. Однако это требует как хороших навыков программирования, так и глубокого освоения среды, и все равно не избавляет от различных трудностей и проблем. Так, в среде матлаба нет встроенных инструментов для организации базы данных, возникают серьезные трудности при анализе рядов с пропусками, а эффективные инструменты для календарной синхронизации рядов данных необходимо программировать самому, так как имеющиеся в пакете средства крайне ресурсоемки.

Многие научные коллективы в России предпочитают использовать для анализа получаемых данных программы собственной разработки. Но в этом случае сложно реализовать большой спектр методов усилиями компактного коллектива. Такие программы часто недостаточно универсальны, плохо документированы и требуют постоянного авторского сопровождения. Их сложно приспособить к решению новых задач, а ведь при исследовательской работе с рядами экспериментальных данных такая потребность возникает достаточно регулярно.

Анализ показывает, что, несмотря на все богатство выбора, ни один из существующих инструментов для работы с рядами не обеспечивает достаточно полную реализацию всего необходимого функционала, во всяком случае в среде Windows. Поэтому нами был создан программный пакет WinABD, который решает весь комплекс упомянутых выше задач.

Структура пакета

Хранение первичных данных.

Все данные в WinABD хранятся в собственной базе данных. Каждый ряд представляет собой бесконечную последовательность ячеек, каждая из которых имеет точную календарную привязку. Частота опроса данных может составлять от 0.000001с (1 мегагерц) до 1 наблюдения в 999лет. При импорте данных необходимо указать точный хронометраж для каждого измерения. Например, внешние данные могут быть записаны в файле в формате «дата-время-значение».

Первичное оформление данных в базу требует определенных усилий, однако позволяет впоследствии работать не с именами файлов (столбцов), а с хорошо паспортизированными выборками. При совместной обработке рядов с несовпадающими интервалами и/или периодичностью наблюдений WinABD автоматически пересчитывает данные в единую шкалу времени. Разумеется, при всех операциях с данными, настройке параметров обработки и визуализации результатов анализа WinABD использует реальную календарную шкалу времени, а не условные «номера точек», что существенно повышает удобство работы.

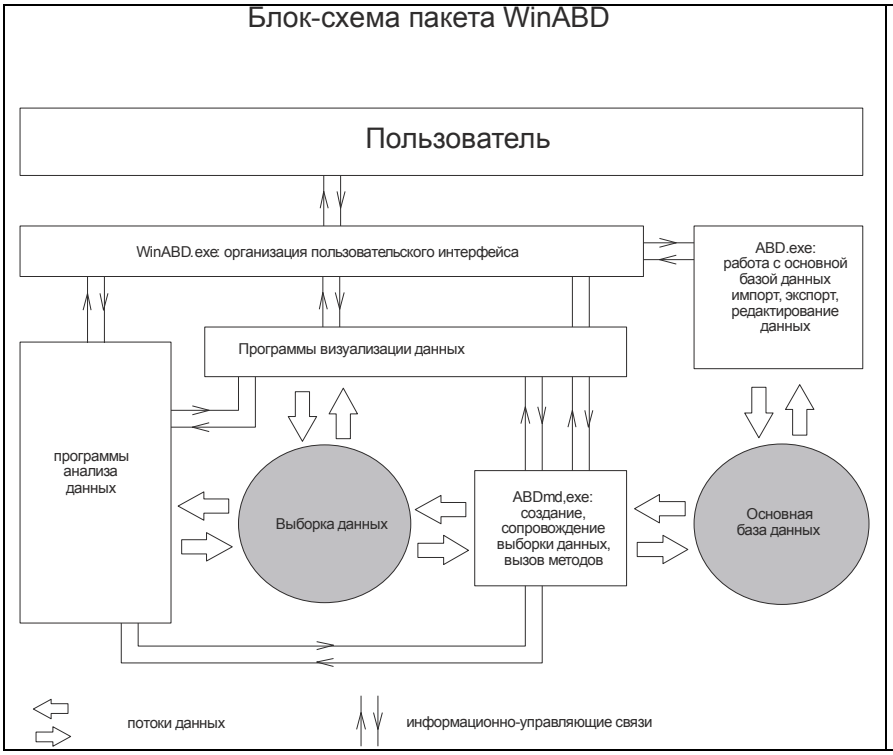

Рис.1 Блок-схема пакета WinABD

Двухуровневая система доступа к данным

При выполнении различных операций над данными происходит их изменение. Создаются новые ряды, уничтожаются старые. Чтобы обезопасить первичные данные от случайного изменения, в WinABD реализована двухуровневая система доступа к данным. Первичная информация – те ряды, которые были получены при наблюдениях и затем импортированы в среду пакета – хранятся в основной базе данных. Однако доступ к ней имеют только методы, связанные с сопровождением данных (импорт, экспорт, редактирование и т.д.). Все вычислительные процедуры оперируют не с первичными данными, а со специально создаваемыми копиями этих рядов (рис.1). Обратное копирование обработанного (видоизмененного) ряда из рабочего пространства в основную базу данных происходит только по особой команде, при этом новый ряд добавляется к основной базе данных.

В отличие от большинства других пакетов статистической обработки, рабочее пространство WinABD хранится не в оперативной памяти, а на диске. Это позволяет загружать в выборку любые объемы данных (сколько хватит места на диске). Созданная выборка не теряется даже при аварийном окончании сеанса работы, а вновь автоматически открывается при следующем запуске программы.

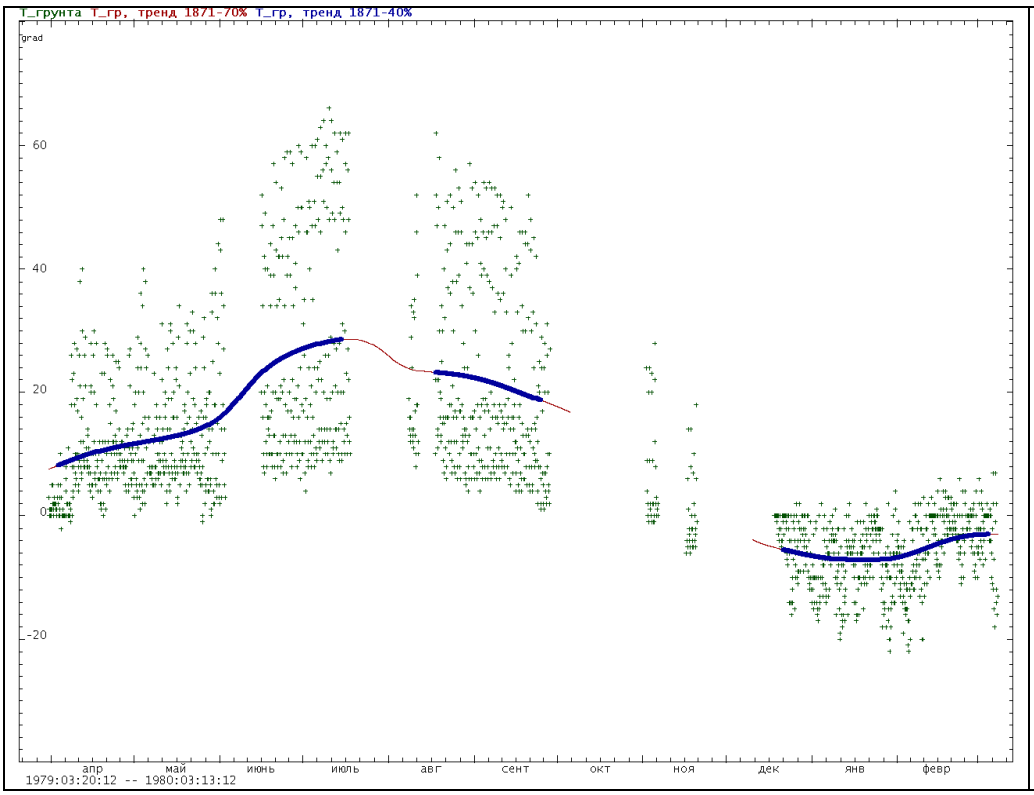

Рис.2. Данные температуры грунта на ст.Лайрон (точки) с интервалами пропусков данных и тренд, выделенный методом ядерного скользящего сглаживания в окне шириной 1871ч с гауссовой весовой функцией окна при разрешенной доле пропусков 70% (тонкая линия) и 40% (жирная линия).

Концепция пропусков данных

Экспериментальные ряды почти неизбежно содержат пропуски данных. Обычно пропущенные наблюдения заполняются каким-либо способом на этапе предварительной подготовки данных. Однако явное заполнение пропусков данных далеко не всегда является оптимальным решением.

Модель данных WinABD допускает наличие пропусков и перерывов в наблюдениях на любом этапе анализа. Для заполнения пропусков в WinABD имеется большое количество инструментов, однако их применение не является обязательным ни в каких ситуациях. Например, можно оценить и удалить тренд по ряду с пропусками. Особая настройка задает максимально допустимый процент пропусков в пределах окна обработки – если заданный уровень превышен, то тренд в данной точке не вычисляется, а в ряд результата записывается пропуск (рис.2). Аналогичные алгоритмы применяются и во всех остальных процедурах WinABD. В предельном случае – при отсутствии пропусков данных – расчеты выполняются по стандартным формулам, предполагающим равномерный по времени шаг между измерениями. При наличии пропусков данных соответствующие моменты времени исключаются из расчетов, а для вычислений используются специальные формулы, оптимизированные с учетом специфики конкретного метода.

Благодаря описанной технологии проверка и отбраковка сомнительных наблюдений может быть выполнена не только на стадии предварительной подготовки ряда, но и на любом этапе анализа. Только в простых случаях брак визуально выглядит как резкий выброс. Часто встречаются ситуации, когда некачественные данные легче всего обнаруживаются при проверке уже отфильтрованного сигнала, прошедшего через целую серию преобразований. Использование «конвейера» обработки данных позволяет разрабатывать итеративные или настраиваемые процедуры выбраковки дефектных наблюдений, применимые на всех стадиях процессинга данных.

Предусмотрена возможность работы с масками пропусков, есть команды «растекания» и «высушивания» пропусков, что позволяет строить достаточно гибкие алгоритмы заполнения пропусков, комбинируя для этого различные методы.

Методы визуализации данных. Оценка статистик

В WinABD средства визуализации данных – это концептуальный, а не вспомогательный инструмент. Предполагается, что любая творческая операция начинается с анализа графика исходного ряда и заканчивается просмотром графика обработанного сигнала. Любые процедуры анализа данных могут применяться как к ряду в целом, так и к его фрагменту, выделенному на

графике. Ряды длиной миллионы точек отображаются практически без задержки, можно «на лету» развернуть любой участок сигнала со сколь угодно подробной детальностью. Ось времени размечается в реальной календарной шкале, оптимизированной в соответствии с интервалом развертки данных и параметрами сигнала. Все вспомогательные надписи и поля минимизированы, чтобы наиболее полно использовать площадь экрана для отображения графиков (см.рис.2).

Методы визуализации данных получают на вход один ряд или группу рядов, а на выходе выдают соответствующую диаграмму. В экране функций распределения можно не только проверить распределение на нормальность, но и удалить выбросы, выделив их прямо на графике. Можно строить корреляционные и структурные функции, периодограммы и спектры [Дещеревский, Журавлев, 1996; Дещеревский, Сидорин, 2011а,б], оценивать фрактальные характеристики ряда (R/S-анализ, метод фрактальных длин [Дещеревский, 1997], изучать траектории Шустера [Сидорин, 2009] и т.д.

Один из наиболее часто используемых методов визуализации – это метод «диаграммы рассеяния» или «корреляционного поля». WinABD позволяет не только визуализировать зависимость одной переменной от другой и оценивать параметры регрессии, но и показывать траекторию процесса в фазовом пространстве в динамике, выполнять другие необходимые операции. Подчеркнем, что при построении взаимной диаграммы, как и при всех других операциях с данными, WinABD использует правило синхронизации наблюдений. Это означает, что всегда сопоставляются именно те значения данных, которые измерены в один и тот же момент времени. При этом анализируемые ряды не обязаны иметь совпадающие начало и конец наблюдений, не обязательна и одинаковая частота наблюдений. Требуется только пересечение рядов (совпадение моментов измерений) на каком-то интервале времени.

Обработка в скользящем окне

Одна из основных задач геофизического мониторинга – это отслеживание изменений, происходящих в контролируемой системе. Например, может изменяться фрактальная размерность сигнала или коэффициент отклика на вариации известного внешнего фактора – такого, как атмосферное давление или прилив. Для обнаружения таких изменений надо оценивать свойства сигнала не по всему ряду, а в пределах некоторого временного окна. Затем окно сдвигается вправо и все вычисления повторяются. Подобная технология применяется и при адаптивной фильтрации различных помех, позволяя "подгонять" параметры фильтра к текущим свойствам ряда. Степень адаптивности можно менять, варьируя ширину окна.

В WinABD имеется большое число методов скользящего окна. Например, можно оценивать в скользящем окне такие статистики, как дисперсию, медиану или заданный квантиль, и использовать их в алгоритмах выделения и подавления выбросов. Можно оценивать динамику показателя Херста, отношения R/S и фрактальной размерности временного ряда [Дещеревский, 1997], отслеживать изменения амплитуды, фазы и зашумленности несинусоидального квазипериодического сигнала, использовать различные регрессионные алгоритмы и др.

В WinABD принято, что скользящее окно, как правило, сдвигается на одну точку, поэтому временной шаг дискретизации обработанного сигнала равен шагу дискретизации исходного ряда (а не размеру окна). Это позволяет наиболее точно отслеживать момент изменения контролируемых параметров.

Известный недостаток методов скользящего окна связан с тем, что такие методы трудно комбинировать друг с другом. Строго говоря, при вычислениях в скользящем окне требуется, чтобы окно всегда целиком помещалось «внутрь» ряда. Из-за этого полный «пробег» окна всегда меньше, чем длина исходного ряда. Если ряд имеет ограниченную длину, а окно достаточно широкое, то уже после применения нескольких методов от сигнала «ничего не останется».

В WinABD проблема уменьшения длины отфильтрованного сигнала решается с помощью описанного выше механизма пропусков данных. При обработке ряда любым методом он автоматически дополняется пропусками справа и слева так, чтобы полный пробег окна равнялся бы длине исходного ряда. Варьируя разрешенное количество пропусков, можно регулировать предельно допустимую величину «выезда» границы окна за пределы ряда для каждого метода. Аналогичные правила действуют и при обработке интервалов пропусков внутри ряда (рис.2).

Заключение

В отличие от многих других пакетов статистического анализа, WinABD обеспечивает полный цикл операций, необходимых при работе с экспериментальными временными рядами. В состав пакета входит система управления базой данных временных рядов, мощный исследовательский комплекс и интерактивная среда визуализации данных.

Пакет позволяет анализировать структуру рядов, выявлять зависимости и взаимосвязи между сигналами. Имеется большое количество нестандартных инструментов и методов, необходимых в повседневной работе с неидеальными данными. Широко используется технология скользящего временного окна, что позволяет изучать развитие всех процессов во времени и выявлять изменения, связанные с какими-либо событиями. Специальная технология "схлопывания" окна на границах ряда позволяет выполнять такую обработку без уменьшения длины ряда, что дает возможность произвольного комбинирования применяемых методов. Все процедуры допускают наличие пропусков в наблюдениях.

При всех операциях с данными при их отображении используется шкала календарного времени, что существенно повышает удобство работы. Обеспечивается корректная совместная обработка рядов, имеющих неодинаковые даты начала и несовпадающую периодичность наблюдений.

Список литературы

Дещеревский А.В. Фрактальная размерность, показатель Херста и угол наклона спектра временного ряда. М.: ОИФЗ РАН, 1997. 36 с.

Дещеревский А.В., Журавлев В.И. Тестирование методики оценки параметров фликкер-шума. М.: ОИФЗ РАН, 1996. 12 с.

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, N3, с.44-70.

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