Математические вопросы кибернетики и вычислительной техники 19, 1998, 32–39.

Программная реализация алгоритма обнаружения моментов изменения состояния временных рядов

П. А. Петросян

Институт проблем информатики и автоматизации НАН РА и ЕрГУ

Аннотация

Предлагается к рассмотрению компьютерная программа "Геостат" реализующая алгоритм непараметрического статистического обнаружения моментов изменения свойств случайных последовательностей [1, 2, 3]. Программа позволяет удобно манипулировать данными в ASCII-формате (считывать, сохранять, вводить, редактировать, осуществлять поиск и замену), производить вычисление статистик Вилкоксона, Муда, а также полиномиальной статистики, просматривать данные в виде графиков. Программа применена для идентификации предвестников землетрясений на основе результатов обработки гидрогеохимических временных рядов. Предварительные результаты успешны [1, 3].

1 Введение

В результате появления мощных пакетов программ, предназначеных для анализа данных на персональных компьютерах, резко расширился круг потребителей методов анализа данных. Многие новые, весьма актуальные с точки зрения практического применения методы анализа данных ждут своей реализации в современных статистических программных пакетах [4]. Описываемая ниже компьютерная программа "Геостат" представляет собой программную реализацию одного из таких актуальных методов-разработанного в Армении алгоритма непараметрического статистического обнаружения моментов изменения свойств случайных последовательностей.

2 Математическая формулировка алгоритма

Для последовательности результатов наблюдений $x_1, ...x_N$ вычисляется вектор рнгов $R_{x_1}, ...R_{x_N}$ по формуле

$$R_{x_i} = \sum_{j=1}^{N} z_{ij} \qquad \text{где } z_{ij} = \left\{ \begin{array}{l} 1, \quad x_j \le x_i \\ 0, \quad x_j > x_i \end{array} \right\}, \qquad i = 1, ..., N$$

Затем вычисляется последовательность статистик Чернова-Севиджа с заданными коэффициентами.

$$W_N(n) = \frac{N}{N-n} \left(\frac{1}{n} \sum_{i=1}^{n} \left(J\left(\frac{R_{x_i}}{N+1} \right) - A(J) \right) \right) \quad n = 1, ..., N-1, \quad \Delta = 0.05,$$

где $J(u) = \gamma_4 u^4 + \gamma_3 u^3 + \gamma_2 u^2 + \gamma_1 u$, $A(J) = \int_0^1 J(u)\, du$, $A(J) = \frac{\gamma_4}{5} + \frac{\gamma_3}{4} + \frac{\gamma_2}{3} + \frac{\gamma_1}{2}$.

Для статистики Вилкоксона $\gamma_1 = 1$, $\gamma_2, \gamma_3, \gamma_4 = 0$, для статистики Муда $\gamma_1 = -1$, $\gamma_2 = 1$, $\gamma_3, \gamma_4 = 0$, для полиномиальной статистики целочисленные коэффициенты задаются пользователем.

Далее вычисляется нормированное значение статистик, по формуле

$$\left(\sqrt{\frac{(N-n)\,n}{N\,C(J)}} W_n(n) \right) \mid H_0 \sim N(0,1)$$

и сравнивается с равным ± 1.96 критическим значением стандартного нормального распределения, соответствующим уровню значимости 0.95. Максимальное значение статистики, превышающее это значение указывает момент изменения. В случае, если все значения лежат в границах ± 1.96, принимается гипотеза о том, что изменения нет.

3 Особенности программы

Программа "Геостат" написана на языке Pascal с применением Turbo Pascal Professional, откомпилирована при помощи компилятора Borland Pascal 7.0, длина исходного текста программы - 1909 строк, размер исполняемого модуля - 101104 байт. Программа может работать на IBM PC - совместимых компьютерах под управлением MS-DOS.

С целью достижения совместимости программы Геостат с другими статистическими программными пакетами, в качестве входного и выходного формата данных был принят самый распространенный и универсальный ASCII-формат.

Программа интегрирована в оригинальный текстовый редактор, позволяющий при помощи системы ниспадающих и всплывающих меню, в сочетании с действием некоторых переопределенных функциональных клавиш удобно манипулировать данными, запускать процессы вычисления статистик и графической визуализации данных.

Благодаря применению модуля TPVARRAY из Turbo Pascal Professional для создания виртуальной памяти на жестком диске, стало возможным использовать строки фиксированной длины, при довольно большом (1,83 Мбайт) максимальном размере редактируемого файла с данными. Это упростило, а значит и повысило надежность некоторых, часто употребляемых в программе функций, связанных с операциями над массивами данных.

Для минимизации потерь в быстродействии, в результате использования более медленной по сравнению с оперативной, виртуальной памяти, в программе применена индексация строк данных, позволяющая производить операции удаления, добавления и перестановки строк данных на логическом уровне, с физической фиксацией изменений лишь во время сохранения файла на диске.

Аналогичный метод индексации применен в дополнение к методу "поплавка" во время вычисления рангов статистик, что также благотворно сказалось на быстродействии, особенно во время работы программы на компьютерах без математического сопроцессора (кстати распознавание и инициализация сопроцессора производится программой автоматически, при запуске) [6].

Во время операций вычисления статистик, данные преобразуются в тип EXTENDED, что позволяет работать с числами, имеющими 19-20 значащих разрядов, при

максимальной длине выборки в 1000 элементов. Последнее ограничение обусловлено особенностями алгоритма.

Программа позволяет визуализировать до 5 массивов данных по 7500 элементов в виде графиков. При отображении данных в виде графиков производится их автоматическое масштабирование, в зависимости от количества отображаемых массивов данных, и с учетом разброса между наибольшим и наименьшим значенями элементов каждого из отображаемых массивов в отдельности. В случае визуализации части массива, при масштабировании учитывается разброс значений лишь из отображаемого диапазона.

Для ускорения отображения информации на экране монитора, повсеместно, как в текстовом, так и в графическом режимах, применяется прямой ввод/вывод в память видеоадаптера. Допускается работа с видеоадаптерами EGA, VGA, SVGA [5].

4 Описание работы пакета

Работу программы удобно начинать передавая имя файла с данными в качестве параметра, при запуске программы. При этом имена программы и предполагаемого файла с данными должны быть разделены в командной строке символом пробела (например GEOSTAT.EXE ARRAYS.DAT). После этого на экране появляется изображенное на (рис.1) окно программы.

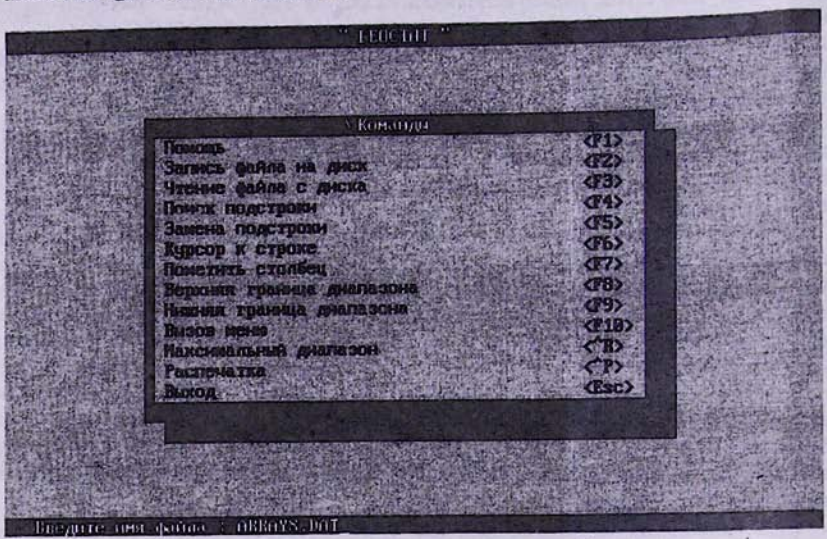

Рис.1 Окно программы с всплывающим окном помощи и срокой состояния.

В случае запуска без параметров, программа переходит в режим запроса имени файла с данными. Имя файла необходимо указать в строке состояния-последней строке окна программы (Рис.2). Следует отметить, что в случае указания имени не существующего файла, будет выдан запрос на создание нового файла с указанным именем (Рис.3).

Рис.2 Строка состояния в режиме запроса имени файла.

Рис.3 Строка состояния в режиме создания нового файла.

В результате перечисленных действий программа переходит в состояние редактирования данных. В этом состоянии нажатие функциональных клавиш F1 - F10, Ctrl R и Ctrl P вызывает выполнение следующих операций.

- F1-вызывает всплывание в центре экрана окна помощи, с указанием "горячих" клавиш (Рис. 1).
- F2-осуществляет запись файла на диск без запроса на изменение имени файла.
- F3-осуществляет считывание файла с диска. Имя файла необходимо указать в строке состояния. Если фаил не найден, то будет выдан запрос на создание нового файла с указанным именем (Рис.3). При наличии изменений в замещаемом файле, предварительно будет выдан запрос на сохранение последнего (Рис.4).

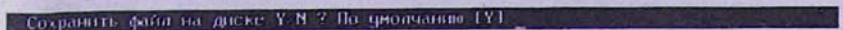

Рис.4 Строка состояния в режиме подтверждения сохранения файла.

- F4-вызывает процедуру поиска подстроки. Искомая последовательность ASCII-символов (шаблон) указывается в строке состояния (Рис.5). При обнаружении в редактируемом файле подстроки соответствующей шаблону, изображение ее на экране инвертируется, а в строке состояния появляется сообщение о возможности продолжить поиск, отменить его, либо указать в строке состояния новый шаблон (Рис.6).

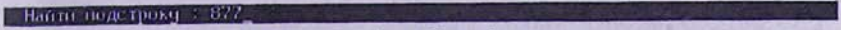

Рис.5 Строка состояния в режиме ввода шаблона для поиска.

Рис.6 Фрагмент окна программы после успешного поиска подстроки. Найденное значение выделено цветом. В строке состояния указаны горячие клавиши для повтора, отмены поиска, либо замены шаблона.

- F5-вызывает процедуру замены подстроки. Вводятся две последовательности ASCII-символов, первая из которых служит шаблоном для поиска (Рис.7), а вторая - для замены (Рис.8). Далее предоставляется возможность отменить режим подтверждения замены (Рис.9). В случае отказа, т.е. во время работы в режиме с подтверждением, процедуру замены можно прервать во время каждого очередного запроса, нажатиеем функциональной клавиши ESC (Рис.10).

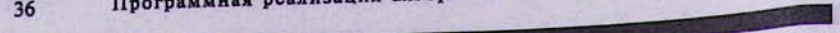

Рис.7 Строка состояния в момент запроса шаблона для поиска (режиме замены подстроки).

Рис.8 Строка состояния в момент запроса шаблона для замены (режиме замены подстроки).

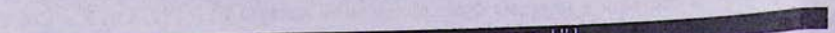

Рис.9 Строка состояния в момент подтверждения замены без уведомления (режиме замены подстроки).

Рис.10 Строка состояния в режиме замены подстроки с уведомлением.

- F6-певод курсора в строку, номер которой хеобходимо указать в строке состояния (Рис.11).

Рис.11 Строка состояния в режиме перехода к строке по номеру.

- F7-пометка столбца с данными для осуществления в дальнейшем операций удаления данных, ычисления статистик, а также отабражения данных в виде графиков. Ширину столбца определяет количество расположенных подряд в обе стороны от курсора символов, отличных от пробела или табуляции (которые считаются разделителями). Левой границей столбца считается позиция, следующая после первого, расположенного слева от курсора разделителя (либо при отсутствии такового - первая позиция в строке). Верхняя и нижняя границы столбца переустанавливаются клавишами F8, F9 и Ctrl R (см. пункты 8, 9).
- F8-установка верхней границы помечаемой области данных.
- F9-установки нижней границы помечаемой области данных.
- Ctrl R - восстановление исходных значения нижней и верхней границы помечаемой области данных.
- Ctrl P - распечатка данных из вводимого в строку состояния диапазона строк (Рис.12, 13).

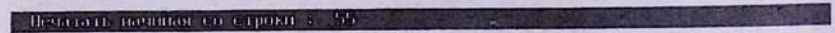

Рис.12 Строка состояния в момент запроса верхней границы распечатываемого диапазона строк.

Рис.13 Строка состояния в момент запроса нижней границы распечатываемого диапазона строк.

- ESC-отказ от продолжения операций, выход из режима графического представления данных, и наконец выход из программы. В последнем случае, при наличии изменений в редактируемом файле предварительно выдается запрос на сохранение последнего (Рис. 4).

- F10-отображение в верхней строке окна программы ниспадающего меню, с возможностью активизации как дополнительных, так и некоторых уже перечисленных операций. (Рис.14-16).

Рис.14 Фрагмент окна программы с ниспадающим меню "Файл".

Меню "Файл" содержит 4 пункта (Рис14).

1. "Открыть файл"- аналогично F3.
2. "Записать файл"- аналогично F2.
3. "Изменить имя"- позволяет записать файл под новым именем.
4. "Конец работы"- завершает работу программы.

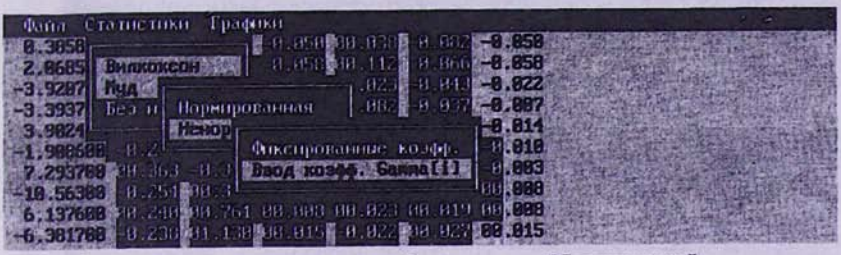

Рис.15 Фрагмент окна программы с ниспадающим меню "Статистики".

Меню "Статистики" содержит 3 пункта (Рис. 15).

1. "Вилкоксон"- вычисляет статистику Вилкоксона.
2. "Муд"- вычисляет статистику Муда.
3. "Без названия"- вычисляет полиномиальную статистику.

Каждый из пунктов меню "Статистики" содержит по два подпункта "Нормированная" и "Ненормированная", вычисляющие соответсвующие, описанные ранее статистики. При этом выбор "Нормированная" для статистики Вилкоксона и Муда, сопровождается

отображением в строке состояния справочной информации (Рис.16) о фиксированных коэффициентах, а в случае полиномиальной статистики - возможностью ввода целочисленных нормирующих коэффициентов.

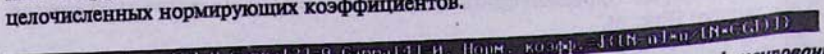

Рис.16 Строка состояния отображающая справочную информацию о фиксированных коэффициентах для статистик Вилкоксона и Муда.

Меню "Графики" содержит два пункта - "Обычный график" и "Сопоставление с 1.96". Каждый из пунктов содержит два подпункта - "Точечный" и "Линейный", которые позволяют производить просмотр предварительно отмеченных данных на экране дисплея в виде точечных или линейных графиков (Рис.17).

Рис.17 Фрагмент окна программы с ниспадающим меню "Графики".

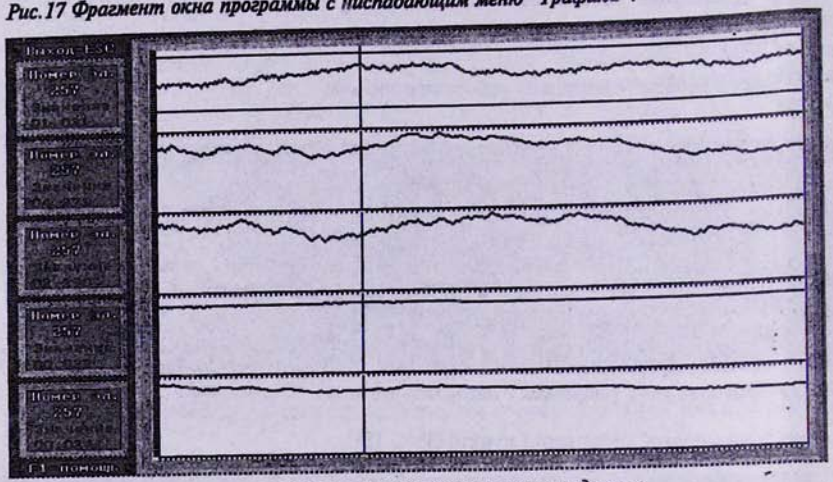

Рис.18 Окно программы в режиме графической визуализации данных.

Число отображаемых на экране графиков колеблется от 1 до 5, (но не более) в зависимости от числа помеченных столбцов данных. Масштабирование производится с таким рассчетом, чтобы наибольшее и наименьшее значения выборки могли быть отображены в окне графического представления данных одновременно. При выборе режима сопоставления, на графике выделяется также и соответствующий доверительный интервал (±1.96). Кроме того окно графического представления

данных (Рис.18), снабжено специальной подвижной риской, перемещение которой относительно графика по горизонтали, осуществляется соответствующими клавишами управления курсором (шаг перемещения регулируется нажатием клавиши "Shift"). При этом слева от графика отображается информация о точке, на которую указывает подвижная риска.

В таком виде программа была применена для идетификации предвестников землетрясений на основе результатов обработки гидрогеохимических временных рядов. Предварительные результаты экспериментов оказались успешными [1, 2]. В настоящее время ведется работа по наращиванию и совершенствованию программы, а также по переводу ее на 32-разрядную платформу операционных систем WINDOWS-95 / WINDOWS-NT.

Литература

1. Е. А. Арутюнян, И. А. Сафарян, П. А. Петросян, А. В. Нерсесян. "Статистический анализ гидрогеохимических данных для обнаружения надежных предвестников землетрясений". Тезисы доклада, V-я Школа-семинар стран СНГ, "Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа", Цахкадзор, 1995 г.

2. E. A. Haroutunian, I. A. Safarian, P. A. Petrossian, H. V. Nersessian. "Earthquake Percursors Identification on the Base of Statistical Analysis of Hydrogeochemical Time Series", Yerevan, "Mathematical problems of computer science", №-18, 1997, p. 33–39.

3. Е. А. Арутюнян, И. А. Сафарян. "Непараметрическое состоятельное оценивание момента изменения свойств случайных последовательностей" Ереван, "Математические вопросы кибернетики и вычислительной техники", №-17, 1997 г., стр. 76–85.

4. Ю. Н. Тюрин, А. А. Макаров. "Анализ данных на компьютере" Москва, изд. "Финансы и статистика" 1995 г.

5. Р. Джордан. "Справочник программиста на персональном компьютере фирмы IBM", Москва, 1992 г.

6. В. В. Фаронов. "Программирование на персональных ЭВМ в среде турбо-паскаль", Москва, изд. МГТУ-1991 г.