РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

АВТОМЕТРИЯ

2009, том 45, № 1

УДК 621.372 : 519.72

АВТОМАТИЧЕСКАЯ ПЕРИОДИЗАЦИЯ СЛУЧАЙНЫХ ВРЕМЕННЫХ РЯДОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДА ОБЕЛЯЮЩЕГО ФИЛЬТРА

В. В. Савченко, Д. А. Пономарев

Нижегородский государственный лингвистический университет,
603155, г. Нижний Новгород, ул. Минина, 31а
E-mail: svv@lunn.ru

В соответствии с общей формулировкой задачи о разладке ставится и решается задача автоматической периодизации случайного временного ряда на однородные отрезки данных длиной в один кластер. На основе авторегрессионной модели и критерия минимума информационного рассогласования разработан новый алгоритм с нормировкой кластеров по дисперсии порождающего шума. Показано, что его главное преимущество по сравнению с известными аналогами состоит в высоких динамических свойствах. Приведены результаты экспериментальных исследований алгоритма в задаче анализа динамики фондовых рынков США и России. Получены оценки для допустимого (порогового) уровня разладки временного ряда в пределах одного кластера в информационной метрике Кульбака – Лейблера.

Ключевые слова: случайный временной ряд, модель авторегрессии, оценка прогнозирования, метод обеляющего фильтра, критерий минимума информационного рассогласования.

Введение. К числу центральных задач статистического анализа, особенно в экономике, относится задача прогнозирования случайных временных рядов (СВР) по конечным выборкам наблюдений. Современные методы прогнозирования динамики временных рядов [1, 2], объединенные общей идеей теоретико-информационного подхода и авторегрессионной модели (АР-модели) данных, обладают высокой скоростью сходимости и хорошо зарекомендовали себя на практике в условиях однородных выборок [3]. К сожалению, в большинстве задач экономического анализа наблюдателям неизвестны заранее точные границы однородных выборок наблюдений и они (наблюдатели) вынужденно (страхуясь от ошибок) сужают их до минимума в своих дальнейших оценках прогнозирования. В результате достигаемая точность прогнозирования часто не реализует всех возможностей современных методов, и этим сильно тормозится их практическое распространение. Поэтому дальнейшим и естественным направлением развития теории практики применения современных методов является повышение точности прогнозиро-

вания за счет максимального расширения используемой выборки в режиме on-line, т. е. по текущему состоянию временного ряда. При этом ключевыми для большинства задач прогнозирования становятся проблема проверки однородности данных и задача автоматической периодизации СВР [4].

Одним из наиболее перспективных инструментов для решения задачи периодизации СВР на однородные отрезки данных, или кластеры в терминологии работы [5], является метод обеляющего фильтра (МОФ) [6]. Его центральное звено – критерий минимального информационного рассогласования (МИР) в метрике Кульбака – Лейблера [7]. Задача в данном случае сводится к проверке сложных гипотез о разладке некоторого (гипотетического) случайного процесса по конечным (малым) выборкам наблюдений [8, 9]. Наиболее актуальной в такой задаче является проблема выбора и обоснования допустимого уровня (порога) разладки временного ряда в пределах каждого кластера. Ее исследованию на основе теоретико-информационного подхода и посвящена предлагаемая работа.

Задача о разладке. Следуя общей теории статистического критерия МИР в задачах о разладке [8], воспользуемся гауссовой (нормальной) аппроксимацией временного ряда $X \subset \mathbf{N}(\mathbf{K})$, где $\mathbf{K}$ – его автокорреляционная матрица (ряд центрирован). Известно [6], что в этом случае критерий МИР строго эквивалентен общесистемному байесовскому критерию максимума правдоподобия. Возьмем некоторую выборку $X_0 = (X_1, X_2)$ из такого ряда и мысленно разобьем ее на две последовательные части $X_1 = \mathrm{col}(\mathbf{x}_{1,1}, \mathbf{x}_{1,2}, \ldots$ $\ldots, \mathbf{x}_{1,M_1})$ и $X_2 = \mathrm{col}(\mathbf{x}_{2,1}, \mathbf{x}_{2,2}, \ldots, \mathbf{x}_{2,M_2})$ объемом M_1 и M_2 соответственно. Здесь $\mathbf{x}_{i,j} = \mathrm{col}(x_{1,j}(1), x_{i,j}(2), \ldots, x_{i,j}(n))$ – n-вектор (столбец) отсчетов сигнала в j-м цикле наблюдений в пределах выборки X_i, $i = 1,2$, со свойством $\mathbf{M}_x(\mathbf{x}_{i,j}\mathbf{x}^T_{i,j}) = \mathbf{K}_i$ ($\mathbf{M}_x$ – символ математического ожидания, T – символ транспонирования векторов). Причем в общем случае количество циклов наблюдений в выборках M_1, M_2 может быть различным. Далее будем полагать, что выполняется соотношение $M_1 \gg M_2$ как очевидное условие для оперативного принятия решений о разладке в выборке X_1 по наблюдениям X_2. Задача формулируется в терминах проверки двух статистических гипотез в отношении автокорреляционных матриц (АКМ) $\mathbf{K}_1$ и $\mathbf{K}_2$: проверяетя сложная гипотеза об их равенстве

$$W_0 \colon \mathbf{K}_1 = \mathbf{K}_2 \stackrel{\Delta}{=} \mathbf{K}_0$$

против сложной альтернативы об их неравенстве

$$W_1 \colon \mathbf{K}_1 \neq \mathbf{K}_2.$$

В указанной формулировке задача была впервые поставлена и решена в работе [9]. Ключевым звеном оптимальной обработки сигнала в ней установлена стандартная процедура корреляционного выборочного оценивания двух АКМ по классифицированным выборкам наблюдений:

$$S_k = \frac{1}{M_k}\sum_{i=1}^{M_k} \mathbf{x}_{k,i}\mathbf{x}^T_{k,i}, \qquad k = 1,2.$$

А оптимальное правило принятия решений об обнаружении разладки в объединенной выборке X_0 имеет вид

$$W_1\colon \lambda(X_0) = M_1\gamma_{1,0} + M_2\gamma_{2,0} > \lambda_0. \tag{1}$$

Здесь

$$\gamma_{k,0} = \frac{1}{2}\left[\operatorname{tr}(S_k S_0^{-1}) - \ln\left|S_k S_0^{-1}\right| - n\right] \tag{2}$$

– величина информационного рассогласования (по Кульбаку – Лейблеру) гипотетического гауссова процесса с автокорреляционной матрицей S_k, $k = 1,2$, по отношению к гауссову же процессу, но с матрицей S_0;

$$S_0 = \frac{M_1}{M_0} S_1 + \frac{M_2}{M_0} S_2$$

– выборочная оценка максимального правдоподобия для АКМ $\mathbf{K}_0$ объединенной выборки $X_0 = (X_1, X_2)$ ($|\cdot|$ обозначает определитель $(n \times n)$-матрицы, $\operatorname{tr}(\cdot)$ – ее след). При этом пороговый уровень $\lambda_0 = \text{const}$ устанавливается в зависимости от требований к уровню значимости принимаемого решения:

$$P\{\lambda(X_0) > \lambda_0 / W_0\} \le \alpha_0 = \text{const}. \tag{3}$$

Решение (1) отвечает критерию минимума взвешенной суммарной величины информационного рассогласования между гипотетическими гауссовыми распределениями с автокорреляционными матрицами S_1 и S_2 относительно закона $\mathbf{N}(S_0)$. Чем ближе в теоретико-информационном смысле выборки X_1 и X_2 расположены друг к другу, тем меньше информационное рассогласование их распределений.

Отметим, что с учетом существенных различий в объемах выборок X_1 и X_2 при $M_1 \gg M_2$ выражение для оптимальной решающей статистики (1) существенно упрощается:

$$W_1\colon \lambda(X_0) = \gamma_{2,1} = \frac{1}{2}\left[\operatorname{tr}(S_2 S_1^{-1}) - \ln\left|S_2 S_1^{-1}\right| - n\right] > \tilde{\lambda}_0. \tag{4}$$

Решение здесь принимается по принципу минимума информационного рассогласования между двумя рассматриваемыми выборками – это стандартная формулировка критерия МИР [7].

Метод обеляющего фильтра. В асимптотическом случае $n \to \infty$, когда в качестве объектов статистического анализа рассматриваются два стационарных гауссовых процесса $X_1(t)$ и $X_2(t)$, и синтезированный алгоритм (4) может быть переписан в частотной области в эквивалентном виде [6]:

$$W_1\colon \lambda(X_0) = \frac{1}{F}\sum_{f=1}^{F}\left(\frac{G_x(f)}{G_1(f)} + \ln\frac{G_1(f)}{G_x(f)}\right) - 1 > \tilde{\lambda}_0, \tag{5}$$

где $G_x(f)$ – выборочная оценка спектральной плотности мощности (СПМ) анализируемого сигнала $X_2(t)$ в функции дискретной частоты f; $G_1(f)$ – аналогичная оценка СПМ «опорного» сигнала $X_1(t)$; F – верхняя граница частотного диапазона сигнала или используемого для его передачи канала связи.

При дополнительном и актуальном для задач прогнозирования условии нормировки АР-модели сигналов типа кластеров по дисперсии их порождающего шума второе слагаемое в правой части (5) оказывается тождественно равным нулю и выражение для решающей статистики еще более упрощается:

$$W_1\colon \lambda(X_0) = \frac{1}{F}\sum_{f=1}^{F} \frac{\left|1 + \sum\limits_{m=1}^{p} a_1(m)\exp(-j\pi mf/F)\right|^2}{\left|1 + \sum\limits_{m=1}^{p} a_x(m)\exp(-j\pi mf/F)\right|^2} - 1 > \tilde{\lambda}_0. \qquad (6)$$

Это стандартная формулировка метода обеляющего фильтра [10]. Решение в нем в пользу разладки (1) принимается по признаку превышения некоторого порогового уровня $\tilde{\lambda}_0$ средней мощностью или дисперсией отклика на сигнал $X_2(t)$ адаптивного (настроен на сигнал $X_1(t)$) обеляющего фильтра. При этом выражение для решающей статистики (6) описывает выборочную оценку величины информационного рассогласования (ВИР) между сигналом X на входе и опорным сигналом в частотной области [6]. Здесь $\{a_1(m)\}$, $\{a_x(m)\}$ – k-векторы коэффициентов линейной среднеквадратической авторегрессии сигналов $X_1(t)$ и $X_2(t)$ соответственно. Таким образом, при сделанных выше допущениях МОФ – это одновременно и эффективный, и экономный способ реализации критерия МИР в рассматриваемой задаче о разладке в СВР. Одновременно это первый шаг к решению задачи автоматической периодизации экономических временных рядов. Применяя решающее правило (6) последовательно к очередным коротким (1–3 месяца) сегментам данных $X_3(t)$, $X_4(t),\ldots$, в каждом случае будем иметь два варианта решения: сигнал сохраняет/не сохраняет свой первоначальный закон распределения. В первом варианте фиксируем продолжение первоначального кластера, во втором – начало нового кластера во временном ряду.

Нетрудно понять, что длительность кластеров, а значит, и значение ВИР между ними в каждом отдельном случае прямо связаны с величиной порогового уровня $\tilde{\lambda}_0$ из решающего правила (5). Поэтому проблема выбора порога разладки СВР на выходе обеляющего фильтра является ключевой для решаемой задачи. Будем оптимизировать величину порогового уровня по принципу стабилизации количества кластеров, полученных в результате работы алгоритма периодизации.

Программа и результаты эксперимента. Для исследования были выбраны три крупнейшие компании: «General Motors», «General Electrics» (США) и РАО «ЕЭС России», акции которых относятся к наиболее ликвидным финансовым инструментам в мире. В качестве исследуемых временных рядов использовались официальные данные Нью-Йоркской фондовой биржи и Московской межбанковской валютной биржи по динамике приращения цены дневного закрытия для акций указанных компаний за период с 2005 по 2007 гг.

Каждый из указанных рядов предварительно разбивался на короткие сегменты данных длиной N отсчетов, к которым применялось решающее правило (6). При этом значения порога разладки и длины (объема) каждого сегмента варьировались в широких пределах: $\widetilde{\lambda}_0 = 0{,}05$–$1{,}00$; $N = 20$–100. А для расчета коэффициентов авторегрессии в (6) применялась рекуррентная процедура Берга – Левинсона с высокой скоростью сходимости [11]:

$$
\begin{gathered}
a_m(i) = a_{m-1}(i) + c_m a_{m-1}(m-i), \quad i = \overline{1, m}, \\
c_m = S_{m-1}^{-2} \sum_{t=m}^{N-1} \eta_{m-1}(t)\, \mathrm{v}_{m-1}(t-1), \\
S_{m-1}^{-2} = 0{,}5(n-m)^{-1} \sum_{t=m}^{N-1} [\eta_{m-1}^2(t) + \mathrm{v}_{m-1}^2(t-1)], \\
\eta_m(t) = \eta_{m-1}(t) - c_m \mathrm{v}_{m-1}(t-1), \\
\mathrm{v}_m(t) = \mathrm{v}_{m-1}(t-1) - c_m \eta_{m-1}(t), \quad t = 0, 1, \ldots, N-1; \\
\sigma_m^2 = (1 - c_m^2)\sigma_{m-1}^2, \quad \sigma_0^2 = S_0^2, \quad m = \overline{1, p},
\end{gathered}
\tag{7}
$$

при ее инициализации системой равенств $\mathrm{v}_0(n) = \eta_0(n-1) = x(n)$. Порядок АР-модели $p = 5$–30 устанавливался в зависимости от объема сегмента данных N по известной методике [2]. Полученные результаты отображаются в виде ряда графиков на следующих рисунках.

Зависимость количества выявленных кластеров в динамике цен на акции американской компании «General Motors» от двух основных параметров обработки (объема сегмента данных N и порога по ВИР $\widetilde{\lambda}_0$) показана на рис. 1. Видно, что при увеличении значений обоих параметров сначала происходит

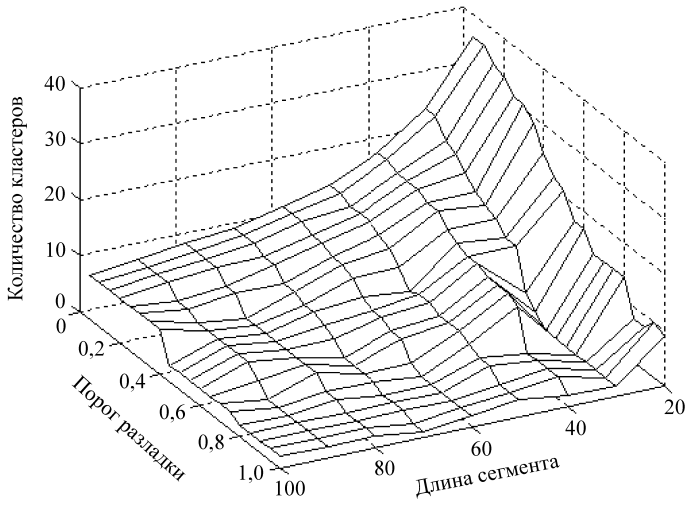

Рис. 1

резкое уменьшение количества кластеров. Проблема выбора оптимальных значений параметров ($N^* =$ 50–70 рабочих дней, или примерно 2–3 месяца, и $\widetilde{\lambda}_0^* = 0{,}5$–$0{,}7$) решается здесь очевидным путем – по принципу относительной стабилизации кластерного состава рассматриваемого временного ряда. С одной стороны, при малых значениях параметров мы получаем чрезмерно (по логике) большое количество кластеров с пренебрежимо малыми различиями между собой в теоретико-информационном смысле. С другой стороны, при слишком больших значениях параметров принципиально различающиеся отрезки данных будут объединяться в один кластер, что противоречит здравому смыслу. Соответственно значения параметров $\widetilde{\lambda}_0$ и N следует выбирать по тем точкам на графике, где количество выделенных кластеров (до 10 в нашем примере) одновременно достаточно представительно и устойчиво.

Временная диаграмма приращения цены закрытия $\Delta C(t)$ для акций компании «General Motors» представлена на рис. 2. Здесь вертикальными линиями отмечены границы выявленных кластеров. При этом порог разладки был установлен $\widetilde{\lambda}_0 = 0{,}55$, длина сегмента $N = 50$. Видно, что за период с 2005 по 2007 гг. выделяются четыре продолжительных кластера длиной от 4 до 6,5 месяца и два коротких кластера между ними длиной примерно по 50 рабочих дней каждый, или ровно два календарных месяца. Последние представляют собой, по-видимому, некие переходные процессы в экономике США за соответствующие периоды времени. Это типичный пример практического применения предложенного алгоритма периодизации СВР в задачах экономического анализа.

На следующем этапе исследований была рассмотрена устойчивость полученных оценок основных параметров алгоритма по отношению к ценам на

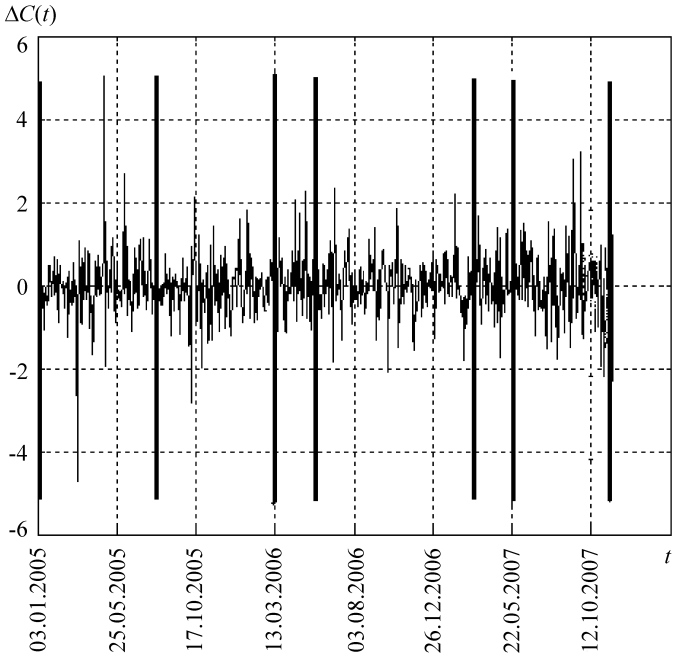

Рис. 2

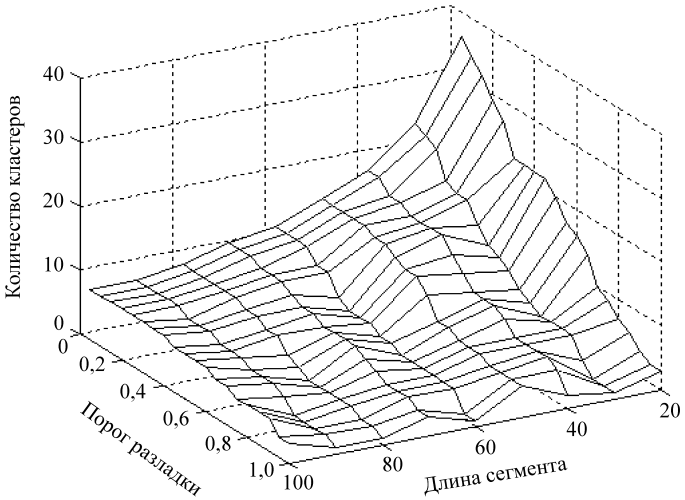

Рис. 3

акции разных компаний на разных торговых площадках. При этом все вычисления проводились по той же схеме (7), что и с акциями «General Motors». Было показано, что временные границы кластеров в динамике цен на акции разных компаний на рынке США в общем случае могут не совпадать, но их количество и длительность являются весьма устойчивыми величинами.

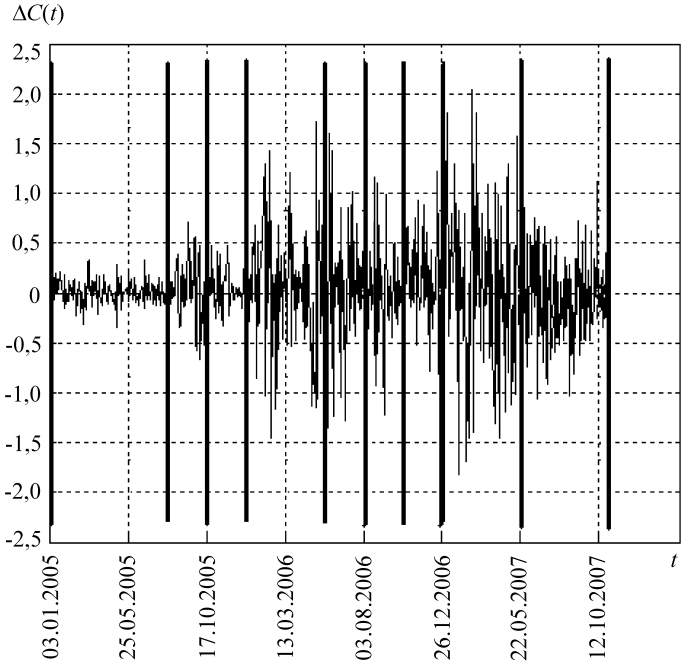

Рис. 4

Данный вывод распространяется и на российский фондовый рынок. Для его иллюстрации на рис. 3 показана зависимость количества выявленных кластеров за тот же период (3 года) в динамике цен на акции РАО «ЕЭС России» от величины порога разладки и длины сегмента.

Хорошо видно, что характер экспериментальных зависимостей (ср. рис. 1 и 3) во всех случаях остается неизменным. Следует особо отметить, что и оптимальный порог разладки $\widetilde{\lambda}_0^* = 0{,}5$–$0{,}7$ и длина сегмента $N^* = 50$–70 одновременно являются в большой степени устойчивыми по отношению к разным видам ценных бумаг на разных фондовых рынках мира. Значение сделанного вывода как для теории, так и для практики экономического анализа представляется очевидным.

Временная диаграмма приращения цены закрытия Московской межбанковской валютной биржи по акциям компании РАО «ЕЭС России» при пороге разладки $\widetilde{\lambda}_0 = 0{,}55$ и длине сегмента $N = 50$ дней приведена на рис. 4. Сопоставляя ее с диаграммой на рис. 2, можно сделать вывод об относительной нестабильности развивающегося российского рынка ценных бумаг, его сильной зависимости от американского рынка, а также отметить, что за 2007 г. российский рынок в значительной степени стабилизировался.

Заключение. Таким образом, благодаря проведенному в данной работе исследованию предложен новый эффективный алгоритм автоматической периодизации случайных временных рядов. Его математическую основу составляет адаптивный метод обеляющего фильтра с высокими динамическими свойствами. Указанная особенность обеспечивает одновременно две важные цели: во-первых, этим достигается высокая точность автоматической настройки алгоритма в результате компактности и эффективности используемых при адаптивном спектральном анализе вычислительных процедур; во-вторых, значительно расширяется область практического применения линейных АР-оценок прогнозирования в расчете на конечную инерционность анализируемых процессов в экономике.

К числу наиболее перспективных задач в данном направлении наряду с собственно прогнозированием СВР следует отнести многофакторный анализ или в более общей формулировке задачу экономической диагностики. Рассмотренные выше примеры относятся именно к этой области исследований.

СПИСОК ЛИТЕРАТУРЫ

1. **Савченко В. В.** Прогнозирование социально-экономических процессов на основе адаптивных методов спектрального оценивания // Автометрия. 1999. № 3. С. 99–108.
2. **Савченко В. В.** Теоретико-информационное обоснование линейных оценок прогнозирования // Автометрия. 2001. № 5. С. 68–77.
3. **Савченко В. В.** Использование линейной авторегрессионной модели для прогнозирования динамики биржевых котировок // Автометрия. 2004. **40**, № 4. С. 117–128.
4. **Савченко В. В.** Проверка однородности выборочных данных в задачах спектрального оценивания // Радиотехника и электроника. 1999. **44**, № 1. С. 65–69.
5. **Костерин А. Г.** Практика сегментирования рынка. С.-Пб.: Питер, 2002.
6. **Савченко В. В.** Различение случайных сигналов в частотной области // Радиотехника и электроника. 1997. **42**, № 4. С. 426–431.

7. **Кульбак С.** Теория информации и статистика. М.: Наука, 1967.

8. **Савченко В. В.** Обнаружение и прогнозирование разладки случайного процесса на основе спектрального оценивания // Автометрия. 1996. № 2. С. 77–84.

9. **Акатьев Д. Ю., Савченко В. В.** Обнаружение разладки случайного процесса на основе принципа минимума информационного рассогласования // Автометрия. 2005. **41**, № 2. С. 68–74.

10. **Савченко В. В.** Автоматическая обработка речи по критерию минимума информационного рассогласования на основе метода обеляющего фильтра // Радиотехника и электроника. 2005. **50**, № 3. С. 309–314.

11. **Марпл С. Л.** Цифровой спектральный анализ и его приложения. М.: Мир, 1990.

Поступила в редакцию 15 января 2008 г.