Труды Карельского научного центра РАН
№ 4. 2014. С. 6–13

УДК 519.17,519.856,681.324

АЛГОРИТМЫ ГЛОБАЛЬНОЙ ОПТИМИЗАЦИИ ДЛЯ АНАЛИЗА ДАННЫХ

Б. З. Белашев[1,2], **А. В. Кабедев**[2]

[1] *Институт геологии Карельского научного центра РАН*
[2] *Петрозаводский государственный университет*

Алгоритмы глобальной оптимизации: генетические алгоритмы, поиск по шаблону совместно с методами максимума энтропии и наименьших квадратов применены для декомпозиции сложных распределений на компоненты. Цель применения – выявление вероятных структур размытых распределений и количественная оценка их параметров. Протестированные на моделях алгоритмы использовали для определения структур сложных ИК-спектров минералов, рентгенограмм аморфных материалов, селекции геофизических аномалий.

Ключевые слова: оптимизация, алгоритм, поиск, сходимость, энтропия, шаблон, распределение, спектр, геофизика.

B. Z. Belashev, A. V. Kabedev. GLOBAL OPTIMIZATION ALGORITHMS FOR DATA ANALYSIS

Global optimization algorithms genetic algorithms and patternsearch algorithms together with maximum entropy and least square methods were applied to decompose complex distributions into components. The goal of these algorithms is to reveal the probable structure of the blurred distributions and estimate their parameters. Algorithms were tested on the models and used to determine the structures of complex IR spectra of minerals X-ray diagrams of amorphous materials and a selection of geophysical anomalies.

Key words: optimization, algorithm, search, convergence, entropy, pattern, distribution, spectra, geophysics.

. . . в мире не происходит ничего, в чем бы не был виден смысл какого-нибудь максимума или минимума. . . *Л. Эйлер*

Введение

Предмет рассмотрения составляют алгоритмы и модели на основе вариационных методов. В отличие от дифференциальных или разностных уравнений, описывающих временной ход процессов, вариационные методы определяют установившиеся состояния, отвечающие экстремальным значениям функционалов. Анализ таких состояний помогает интерпретировать результаты моделирования сложных систем.

Экстремальные модели известны в физике, механике, термодинамике, экономике, теории управления [10]. В биологии их популярность растет с развитием эволюционных представлений [11]. Теория систем и синергетика в качестве базовых принципов используют методы максимальной энтропии, максимальной скорости изменения потока энергии через систему, наименьшей диссипации энергии, наискорейшего спуска [8].

В данной статье методы максимума энтропии и наименьших квадратов применены для декомпозиции эмпирического распределения на компоненты, имеющие понятный физический смысл. Алгоритмы глобальной оптимизации использованы для выявления числа таких компонент и определения их параметров. Обычно экстремальную задачу формулируют, вводя условный функционал, помимо целевой функции содержащий уравнения, описывающие исходное распределение, и неравенства, накладываемые на переменные. Ограничения присутствуют в условном функционале аддитивно и взвешены множителями Лагранжа. Приравнивание первых вариаций условного функционала по искомой функции к нулю дает уравнения, позволяющие выразить искомую функцию через множители Лагранжа. Множители Лагранжа находят, подставляя эту функцию в условие задачи и решая полученную систему уравнений [7]. Для работы алгоритма требуется дифференцируемость функций. Его сложность задана этапами вычислений.

Для выпуклых функционалов и данных, удовлетворяющих условиям [4], экстремальная задача имеет единственное решение. Оптимизация функционалов с несколькими экстремумами традиционным способом сопряжена с преждевременной сходимостью. В этом случае решение в виде локального экстремума может сильно отличаться от глобального экстремума. В такой ситуации предпочтительно иметь единственное решение задачи, близкое к глобальному экстремуму функции. Многие из алгоритмов, дающих такое решение, еще не нашли широкого применения в практике обработки данных.

Цель работы – демонстрация действия алгоритмов глобальной оптимизации на основе генетических алгоритмов и поиска по шаблону [14] для выявления скрытых структур распределений в области материаловедения и геофизики. На симулированных и реальных примерах с помощью этих алгоритмов реконструированы наиболее вероятные структуры сложных спектров. Полученная при реконструкциях информация полезна при разработке подходов к проблемам колебательной спектроскопии, аморфного состояния вещества, селекции аномалий потенциальных полей.

Алгоритмы глобальной оптимизации Дж. Холланда

Пример стохастической глобальной оптимизации дают генетические алгоритмы [9], основанные на идее естественного отбора Ч. Дарвина – совершенствования вида путем передачи потомкам лучших генов. Функцию приспособленности особей к среде определяют на множестве хромосом особей – последовательностей единиц и нулей, представляющих числа в коде Грея. Чем значение функции меньше, тем более особь считают приспособленной к среде.

Генетический алгоритм ищет глобальный минимум этой функции, начиная с произвольной совокупности особей, рассматриваемой в качестве популяции. На каждой итерации особи объединяют в пары и производят потомков путем кроссинговера – обмена хвостами хромосом и мутации – случайной инверсии значения в случайном разряде числа. По приспособленности потомков и родителей ведут отбор особей в новую популяцию. Алгоритм сходится, если новая популяция не отличается от предыдущей.

Преимущества генетических алгоритмов состоят в отсутствии требований непрерывности и дифференцируемости функций, нечувствительности к попаданию в локальные минимумы, в возможности многокритериальной оптимизации, многократного ускорения сходимости по сравнению со случайным поиском, простоте реализации. Их недостатками являются затрудняющая понимание биологическая терминология и невысокая точность определения глобального экстремума, которую повышают, выполняя алгоритм несколько раз и выбирая значение экстремума с наилучшей функцией приспособленности.

Меньшие затраты по сравнению с генетическими алгоритмами имеет алгоритм поиска по шаблону – множеству точек в виде вершин n-мерного куба, расширяющемуся или сжимающемуся в зависимости от того, имеет или нет точка шаблона меньшее значение, чем текущее значение функции. Минимальный размер шаблона является основанием для прекращения поиска.

В системе компьютерной математики «MATLAB» данные алгоритмы реализованы процедурами ga и patternsearch [15].

Реконструкция наиболее вероятной структуры размытого распределения

Данные алгоритмы применены при реконструкции структуры сигнала и шума методом максимума энтропии. По сигналу s_i и известной функции размытия h_{ij} требуется восстановить функцию распределения x_i и аддитивный шум x_{i+n}, связанные соотношением:

$$s_i = \sum_{j=1}^{n} h_{ij} x_j + x_{i+n}, \qquad (1)$$

где x_i, $x_{i+n} > 0$, $i, j = 1, 2, ..n$, а шум представлен последовательностью положительных случайных чисел.

Решение, отвечающее функционалу максимума энтропии и ограничениям (1), является максимально произвольным, наиболее вероятным [8]. Параметр ρ регулирует вклад энтропии шума в функционале (2).

$$-\sum_{i=1}^{n} x_i ln x_i - \rho \sum_{i}^{n} x_{i+n} ln x_{i+n} \rightarrow \ max, \qquad (2)$$

Симулированный сигнал s_i получали размытием двух δ-пиков амплитудами 0,5 и 0,7 функцией $h_{ij} = exp(-0,25(i-j)^2)$, с добавлением шума в диапазоне (0; 0,1). Условия задачи и неравенства, задающие диапазоны изменения переменных, вводили непосредственно в командную строку алгоритмов ga и patternsearch. Результаты реконструкции (рис. 1, а, б), полученной без использования множителей Лагранжа, демонстрируют близость исходных данных модели и полученных оценок.

В рассмотренных примерах для реконструкции структуры размытого распределения необходимо знать функцию размытия h_{ij}. Когда эта функция неизвестна, ее вид выбирают из дополнительных соображений. В задачах спектроскопии распространенной ее формой являются лоренциан или гауссиан.

Учитывая, что свертка лоренциана с лоренцианом дает лоренциан, а гауссиана с гауссианом – гауссиан, на первом этапе к экспериментальному распределению применяли процедуру с выбранной таким образом функцией h_{ij}, ширина которой равна меньшей ширине наиболее узкой компоненты распределения. В результате в оценке распределения проявлялись моды, указывающие на возможное число компонент распределения и диапазоны изменения их параметров [12]. Эту информацию использовали при поиске параметров компонент, который проводили, минимизируя невязку экспериментального и модельного распределений:

$$\sum_{i=1}^{n} (s_i - A_1 exp(-(\frac{i-t_1}{g_1})^2) - A_2 exp(-(\frac{i-t_2}{g_2})^2))^2 \qquad (3)$$

Результаты метода наименьших квадратов для двух компонент приведены на рис. 2.

Анализ значений функции приспособленности и параметров модели при разных прогонах алгоритма показал, что целесообразно выбирать результаты с наименьшим значением функции приспособленности. Оценки средних значений и дисперсий параметров при разных реализациях алгоритма дают ошибки параметров, связанные с процессом оптимизации. Ошибки параметров, зависящие от погрешностей измерений, определяли, разыгрывая исходное распределение в коридоре трех среднеквадратичных отклонений содержимого разрядов гистограммы и усредняя параметры модели, полученные в реализациях.

Примеры реконструкции структур размытых спектров и распределений

Задача декомпозиции сложных контуров распределений на компоненты возникает в разных областях исследований. В материаловедении ее актуальность определена существованием размытых спектров и рентгенограмм материалов, которые сложно поддаются анализу традиционными методами. На рис. 3 представлены результаты разложения ИК-спектра порошка кианита (а) и рентгенограммы дистиллированной воды (б) [1] на гауссовы компоненты, полученные алгоритмом patternsearch. Предварительно тем же методом были определены моды распределений и возможные диапазоны изменения их параметров. ИК-спектр порошка кианита был зарегистрирован на спектрофотометре «Спекорд М 80». Проведенная обработка подтвердила реальную сложность данных, выявила их структуру, продемонстрировала способность алгоритмов оптимизации оперировать большим числом компонент. С другой стороны, появилась проблема интерпретации, наполнения результатов моделирования физическим смыслом, отнесения полученных компонент к определенному типу колебаний или структурных характеристик. В этом плане показательна проблема воды. Несмотря на современные методы ее изучения [5], ясности в вопросе о структуре воды нет. Сложность проблемы частично может быть объяснена неадекватным методическим обеспечением исследований. Так, использование в анализе ближнего порядка функций радиального распределения электронной плотности, отражающих вклад разных координационных сфер [13], не учитывает характерную для воды направленность водородных связей.

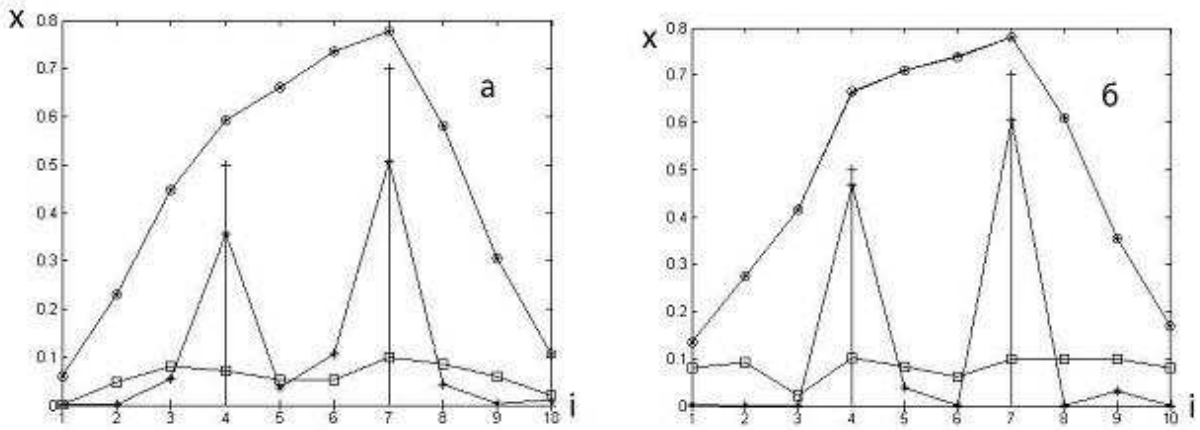

Рис. 1. Результаты реконструкции структуры сигнала и шума методом максимума энтропии с применением алгоритмов глобальной оптимизации ga (а) и patternsearch (б) системы компьютерной математики «MATLAB»: исходные δ-пики (+), данные (o), оценка данных (.), оценка функции (*) и шума (□). Параметр $\rho = 18$

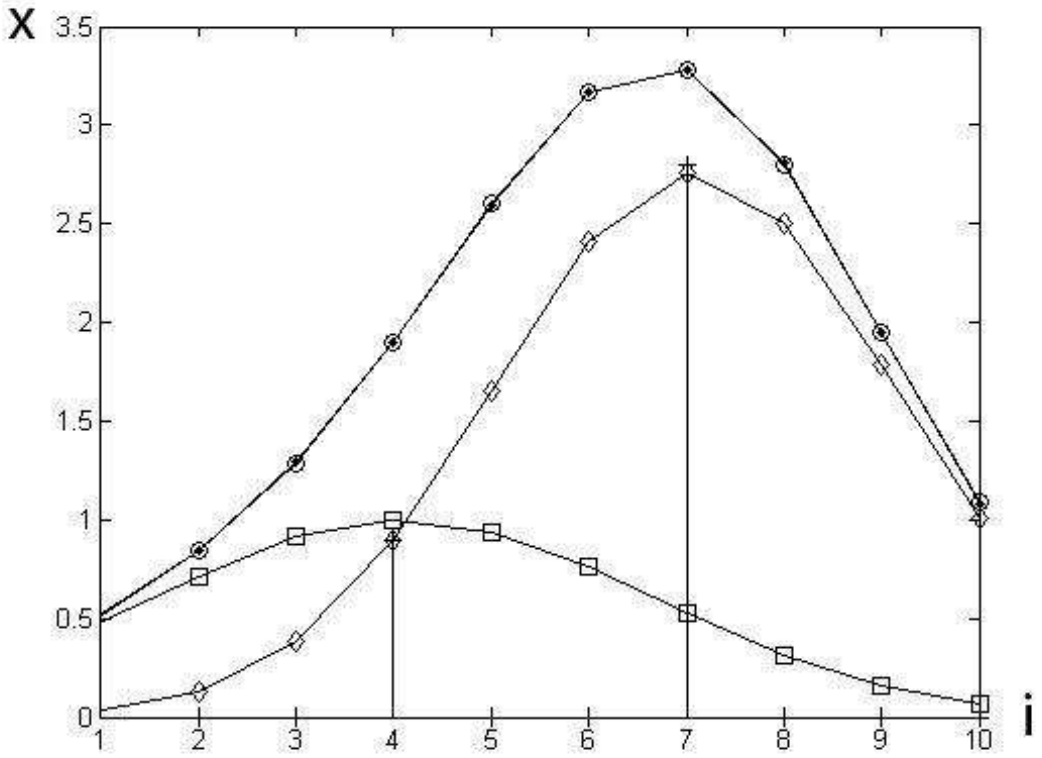

Рис. 2. Подгонка параметров модели по методу наименьших квадратов алгоритмом patternsearch: исходные δ-пики (+), данные (o), оценка данных (.), оценки компонент (□,◊)

.

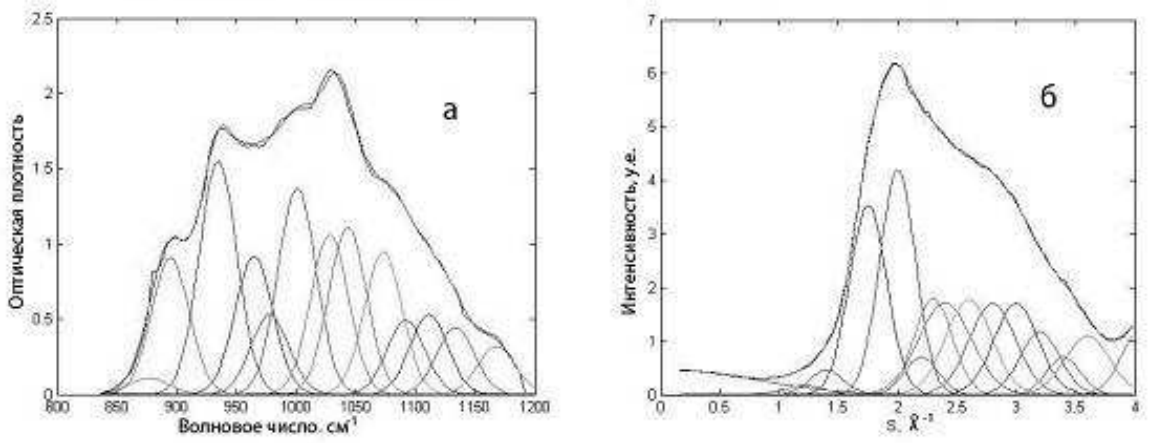

Рис. 3. Разложение на компоненты алгоритмом patternsearch ИК-спектра порошка кианита (а) и рентгенограммы дистиллированной воды (б)

Положения компонент рентгенограммы воды и аналоги межплоскостных расстояний

№	1	2	3	4	5	6	7	8	9	10	11	12	13
s,Å^{-1}	0,98	1,22	1,40	1,75	2,00	2,20	2,30	2,40	2,60	2,80	3,00	3,20	3,40
d,Å	6,43	5,10	**4,49**	3,59	3,14	2,86	**2,73**	2,62	2,42	2,24	2,09	1,96	1,85

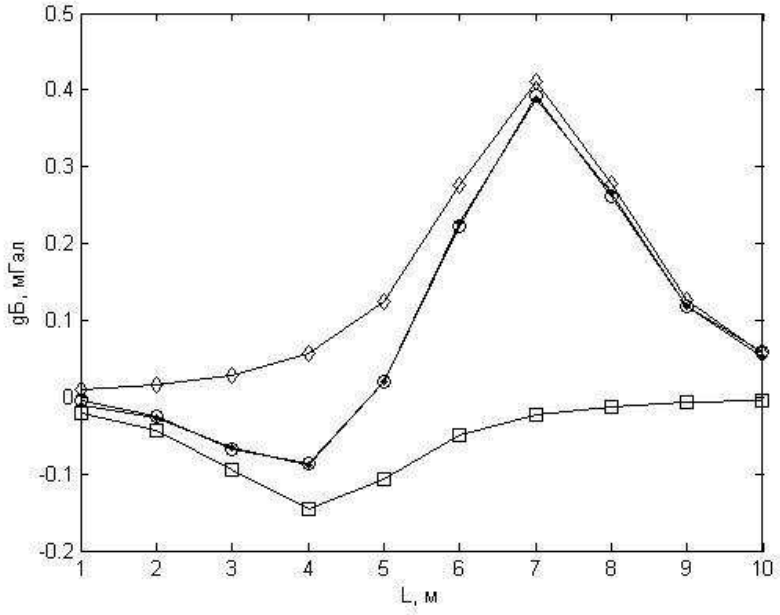

Рис. 4. Декомпозиция симулированного гравитационного профиля (о)на компоненты (□,◊) алгоритмом patternsearch 1: A1e = –1,0 (A1 = –1,2); g1e = 3,6 (g1 = 4,0); L1e = 4,0(L1 = 4,0); и компоненту 2: A2e = 2,2 (A2 = 2,1); g2e = 3,1 (g2 = 3,0); L2e = 7,0 (L2 = 7,0). Невязка между исходным контуром и контуром, построенным по реконструированным компонентам, составила (.) 0,00002

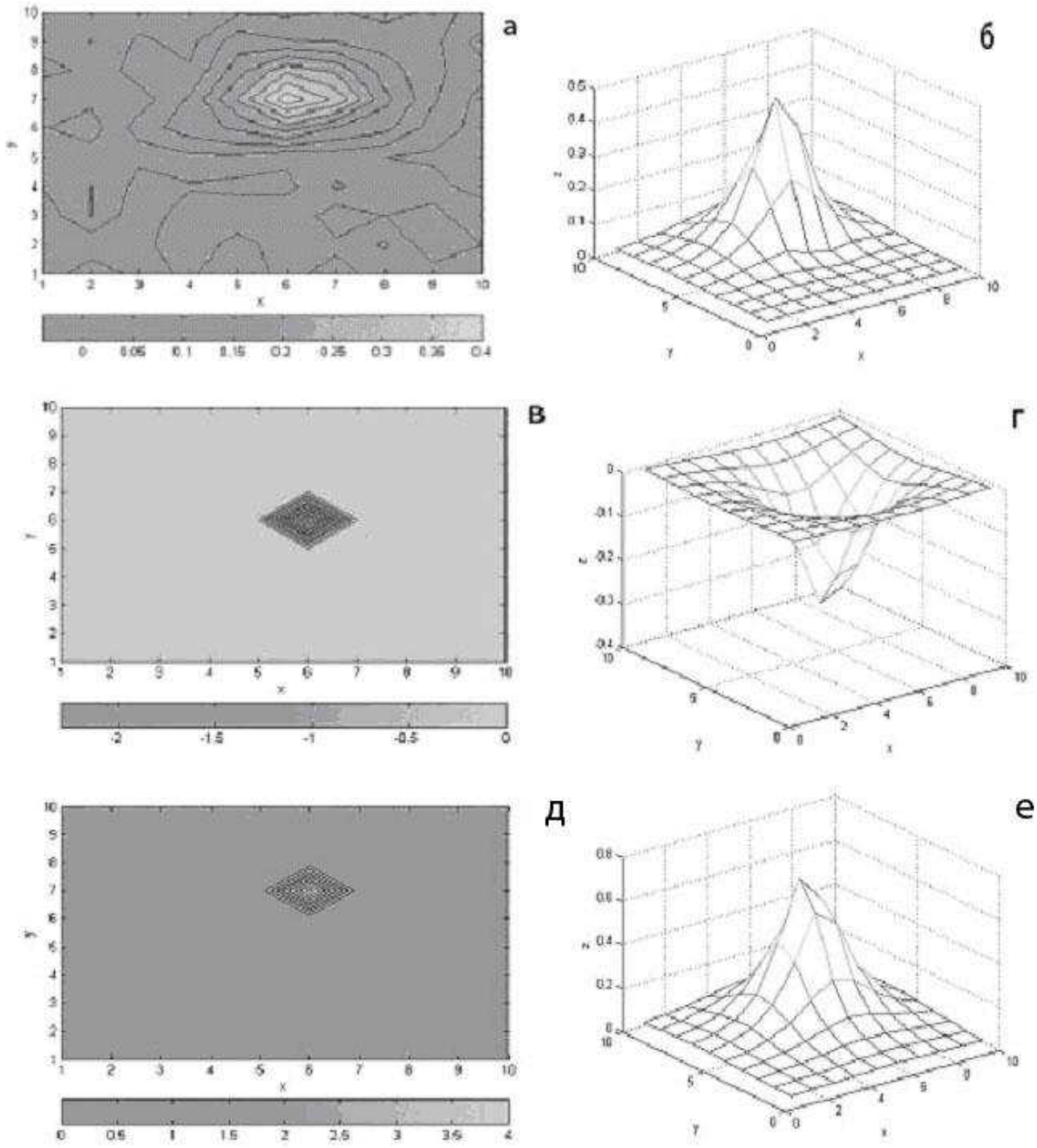

Рис. 5. Реконструкция гравиметрических данных алгоритмом patternsearch с выявлением отрицательной и положительной аномалий с источниками, лежащими вблизи одной вертикали: модельные данные (а) и площадные распределения первой (в) и второй (д) компонент и их 3D формы (б, г, е – соответственно)

Предлагаемый подход свободен от этого недостатка. Вычисленные по положениям компонент дифрактограммы s (см. рис. 3, б) по формуле $d = 2\pi/s$, аналоги межплоскостных расстояний d приведены в таблице. Аналоги, совпавшие с расстояниями между ближайшими водородосвязанными и неводородосвязанными молекулами воды в гексагональном льде I [6], выделены жирным шрифтом. Малый вклад этих компонент в рентгенограмму свидетельствует о том, что клатратные модели [5] не могут дать полного объяснения структуры воды.

Разобранные примеры содержали положительные компоненты распределений. Для расширения области применения алгоритмов глобальной оптимизации были опробованы распределения с компонентами разных знаков. Эта ситуация является типичной при селекции аномалий потенциальных полей в геофизике. Эффективность алгоритмов глобальной оптимизации в этом случае подтверждает разложение гравитационного профиля на компоненты (рис. 4).

При анализе гравитационного профиля параметр, задающий ширину компоненты, дает глубину залегания источника аномалии. Традиционными методами оказывается сложным разделить гравитационные аномалии в случаях, когда их источники находятся вблизи одной вертикали [3]. Алгоритмы глобальной оптимизации, использованные при селекции плоскостных модельных гравитационных распределений (рис. 5), справляются с этой ситуацией вполне приемлемо.

Благодаря единой методике обработки гравиметрических и магнитометрических данных [2] алгоритмы глобальной оптимизации распространены и на селекцию магнитных аномалий.

Заключение

С помощью генетических алгоритмов, алгоритма поиска по шаблону, адаптированных к методам максимума энтропии и наименьших квадратов, реконструированы скрытые структуры распределений случайных величин в разных предметных областях. Эвристические алгоритмы оптимизации имеют понятную основу, не детализированы, не подвержены преждевременной сходимости, дают единственное решение, близкое к глобальному экстремуму задачи, не требуют выполнения условий непрерывности и дифференцируемости функций, не используют множители Лагранжа, допускают многопараметрическую оптимизацию функций, позволяют получить устойчивые, наиболее вероятные оценки структур распределений, адекватны физической природе исследуемых объектов. Их применение способно повысить качество результатов и эффективность обработки данных.

Литература

1. *Алешина Л. А., Люханова И. В.* Рентгенографические исследования взаимодействия технических целлюлоз с водой // Ученые записки ПетрГУ. 2012. № 6 (111). С. 55–59.

2. *Блох Ю. И.* Обнаружение и разделение гравитационных и магнитных аномалий. М.: Изд-во МГГА, 1995, 80 с.

3. *Бычков С. Г.* Определение глубины аномалиеобразующих источников в системе «ВЕКТОР» // Материалы междунар. школы-семинара «Вопросы теории и практики геологической интерпретации гравитационных, магнитных и электрических полей». Апатиты: ОИФЗ РАН, 2002. С. 17–18.

4. *Грешилов А. А.* Прикладные задачи математического программирования: Учебное пособие. М.: Логос, 2006. С. 34–38.

5. *Захаров С. Д., Мосягина И. В.* Кластерная структура воды. Препринт ФИАН, М., 2011.

6. *Зацепина Г. Н.* Физические свойства и структура воды. М.: Изд-во МГУ, 1998. С. 70.

7. *Зейферт Г., Трельфалль В.* Вариационное исчисление в целом, 2-изд-е. М.: РХД, 2000.

8. *Копцик В. А.* Экстремальные принципы информационно-синергетической эволюции / Глобализация, синергетический подход // сайт С. П. Курдюмова «Синергетика». http://spkurdymov.ru/ (дата обращения: 12.09.2013).

9. *Панченко Т. В.* Генетические алгоритмы: Учебное пособие. Астрахань: Астраханский университет, 2007. 87 с.

10. *Полак Л. С.* Вариационные принципы механики их развитие и применения в физике. М.: ЛИБРОКОМ, 2010. 600 с.

11. *Фурсова П. В., Левич А. П., Алексеев В. Л.* Экстремальные принципы в математической биологии // Успехи современной биологии. 2003. Том 123. С. 115–137.

12. *Belashev B. Z.* Methods to reveal Hidden Structures of Signals and their Applications // Вестник РУДН. Серия Математика. Информатика. Физика. 2010. № 3(2). С. 132–135.

13. *Cochran T. W., Chiew Y. C.* Radial distribution function of freely jointed hard-sphere chains in the solid phase // Journal of Chemical Physics. 2006, Vol. 124, no 7 P.074901 http://dx.doi.org/10.1063/1.2167644. (дата обращения: 12.09.2013).

14. *Holland J.* Adaptation in natural and artificial systems. University of Michigan Press Ann Arbor, USA, 1975.

15. *http*://matlab.exponenta.ru/genalg/08.03.03.php. (дата обращения: 17.09.2013).

СВЕДЕНИЯ ОБ АВТОРАХ:

Белашев Борис Залманович
ведущий научный сотрудник, д. т. н.
Институт геологии Карельского научного центра РАН
ул. Пушкинская, 11, Петрозаводск,
Республика Карелия, Россия, 185910
профессор кафедры информационно-измерительных систем и физической электроники
Петрозаводский государственный университет
пр. Ленина, 33, Петрозаводск,
Республика Карелия, Россия, 185910
эл. почта: belashev@krc.karelia.ru
тел.: (8142) 782753

Belashev, Boris
Institute of Geology, Karelian Research Centre,
Russian Academy of Sciences
11 Pushkinskaya St., 185610 Petrozavodsk,
Karelia, Russia
Petrozavodsk State University
33 Lenina St., 185910 Petrozavodsk,
Karelia, Russia
e-mail: belashev@krc.karelia.ru
tel.: (8142) 719675

Кабедев Алексей Владимирович
аспирант кафедры информационно-измерительных систем и физической электроники
Петрозаводский государственный университет
пр. Ленина, 33, Петрозаводск,
Республика Карелия, Россия, 185910
эл. почта: akabedev@mail.ru
тел.: (8142) 719675

Kabedev, Alexey
Petrozavodsk State University
33 Lenina St., 185910 Petrozavodsk,
Karelia, Russia
e-mail: akabedev@mail.ru
tel.: (8142) 782753