УДК 336.76

СЕГМЕНТАЦИЯ И ХЭШИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ В ЗАДАЧАХ ПРОГНОЗИРОВАНИЯ НА ФОНДОВОМ РЫНКЕ[1]

А.Г. Спиро, М.Д. Гольдовская, Н.Е. Киселева, И.В. Покровская

Предложено каждому анализируемому временному ряду цены торгуемого на бирже актива (ВР-Ц) поставить в соответствие временной ряд хэш-кодов (ВР-ХК), которые для каждого элемента ВР-Ц будут показывать рост или падение цены. Отмечено, что в данном случае хэш-коды представляют собой целые числа, и их последовательность позволяет выделить в динамике изменения цены биржевого актива одинаковые (типовые) группы членов ряда ВР-Ц. Описаны процедуры преобразования исходного временного ряда и определения соответствующих хэш-кодов. Сформулированы основные свойства последовательности хэш-кодов. Предложена методика анализа траектории и прогнозирования цены биржевого актива по данным сегментации и хэширования.

Ключевые слова: фондовый рынок, процедура скользящего окна, хэш-коды, сегментация временных рядов, типовые сегменты, прогнозирование роста/падения котировок акций.

ВВЕДЕНИЕ

Профессиональные участники рынка используют в техническом анализе ряд различных индикаторов для прогнозирования цен [1, 2]. Поскольку принципы анализа достаточно подробно изложены в литературе, то обратим внимание лишь на тот факт, который сформулирован в одной из «аксиом» технического анализа, а именно, что в цене актива в данный момент времени отражены все возможные ситуации, которые известны участникам фондового рынка относительно данного актива. Это в свою очередь предполагает, что при одних и тех же внешних условиях (ситуациях) участники рынка будут вести себя одинаково. Заметим, что участники при работе на фондовом рынке располагают данными в виде ценовых графиков или временных рядов (ВР-Ц), отражающих цену торгуемого на бирже актива. При работе на фондовом рынке часто требуется анализировать не только входные данные, но и делать прогноз цены исследуемого актива (в статье рассматривается прогнозирование на ближайшее будущее). В настоящей работе для решения задач анализа и прогнозирования временных рядов на фондовом рынке предлагается метод, использующий процедуры хэширования [3, 4] и сегментации исследуемых временных рядов. Несколько слов о процедуре хэширования. *Хэширование* (от англ. *hashing*) — преобразование по заданному алгоритму входной последовательности чисел в выходную бинарную последовательность. Элементы бинарной последовательности (0,1) либо элементы последовательности, полученной из бинарной с помощью дополнительного преобразования, называют *хэш-кодом* или *хэшем* (подробнее см. далее § 1).

Основная идея этого метода состоит в следующем. Предлагается анализ исходной траектории заменить анализом последовательности (траектории) хэш-кодов, полученной из исходной траектории. Это сильно упрощает задачу анализа и прогнозирования, но благодаря тому, что теперь по последовательности хэш-кодов уже нельзя прогнозировать значение характеристики, а только ее знак, т. е. «увеличение» или «уменьшение» (а точнее, уменьшение или неизменность) значения этой характеристики, а именно такой прогноз и интересует большинство «игроков» фондового рынка. Поскольку хэш-коды в данном случае представля-

[1] Работа выполнена при частичной финансовой поддержке РНФ (грант 14-29-00309) и РФФИ (проекты 14-07-00463, 15-07-06713, 16-07-00895, 16-07-00896).

ют собой целые числа, то их последовательность позволяет выделить в динамике изменения цены биржевого актива одинаковые (типовые) группы членов (сегменты) ряда ВР-Ц. Анализируемые сегменты ряда ВР-Ц получаются с помощью скользящего окна заданного размера. Типовая группа — это группа сегментов, для которой рост или падение котировочной цены в каждом сегменте происходит в одной и той же последовательности по времени. Другими словами, хэш-коды позволяют разделить исходный ряд ВР-Ц на сегменты одинаковой длины (равной размеру скользящего окна) и разделить все множество сегментов на типовые группы (в каждую типовую группу входят сегменты, имеющие одно и то же значение хэш-кода, которое и служит индексом типовой группы).

На этапе прогнозирования определяется типовая группа для последнего по времени сегмента временного ряда. Для каждого сегмента из этой типовой группы (всего таких сегментов V_0) определяется хэш-код для следующей по времени точки временного ряда относительно последней точки этого сегмента. Подсчитывается число сегментов V_2, для которых такой хэш-код соответствует падению котировок. Тогда величина $P_{02} = V_2/V_0$ является оценкой значения вероятности падения котировок в следующий момент времени. По величинам P_{02} и V_0 можно рассчитать доверительный интервал значения такой вероятности, а значит, принять или отвергнуть гипотезу «в следующий момент времени будет падение котировок». Размер оценки значения вероятности роста котировок P_{01} определяется соотношением $P_{01} = V_1/V_0$, где V_1 — число сегментов, для которых соответствующий хэш-код определяет рост котировок. На практике для принятия решения используется правило Байеса — по максимальному значению P_{0i}, $i = 1, 2$.

1. ПОСТАНОВКА ЗАДАЧИ

Необходимо осуществить сегментацию исходного ряда ВР-Ц для поиска типовых групп. Для этого используем прием, рассмотренный в работе [5].

1. На основе исходного временного ряда ВР-Ц строится двоичный кортеж (последовательность), состоящий из единиц и нулей. Значение «1» соответствует повышению или неизменности котировок в данный момент времени, а «0» — понижению.

2. Для этой последовательности применяется пошаговая процедура скользящего окна шириной m с шагом одна позиция. Каждому окну, полученному в рамках этой процедуры, соответствует m-разрядный двоичный код.

3. Этот код преобразуется в десятичное целое число, которое и будет представлять собой хэш-код для соответствующего сегмента ряда ВР-Ц.

4. Полученная последовательность хэш-кодов и является искомым временным рядом хэш-кодов (ВР-ХК).

5. Используя ВР-ХК, выделяются и индексируются типовые группы сегментов ряда ВР-Ц.

2. ПРЕОБРАЗОВАНИЕ ИСХОДНОГО ВРЕМЕННОГО РЯДА И ОПРЕДЕЛЕНИЕ ХЭШ-КОДОВ

Котировки биржевого актива являются временным рядом цены актива, элементы которого представляют собой динамическую переменную $C(t_j)$, которая наблюдается (измеряется) с некоторым постоянным шагом τ по времени, при этом выполняется условие: $t_j = t_0 + (j-1)\tau$, $C_j = C(t_j)$, $j = \overline{1, N}$, где N — объем выборки.

Определим ВР-ХК в результате преобразования ряда, который получен из исходного ВР-Ц следующим образом:

$$X(t_j) = \begin{cases} 1, \text{ если } \Delta C(t_j) > 0 \\ 0, \text{ если } \Delta C(t_j) \le 0 \end{cases},$$

$$\Delta C(t_j) = C(t_j) - C(t_{j+1}). \qquad (1)$$

Ряд (1) фактически представляет собой кортеж, компонентами которого служат 1 и 0. Для анализа кортежа (1) используется скользящее окно шириной $m \ll N$, где $(N-1)$ — длина (размерность) кортежа (1). Такое окно позволяет ввести в рассмотрение двоичный позиционный код, образованный из единиц и нулей ряда (1). Старший разряд этого кода в момент времени t_j совпадает с двоичной переменной кортежа (1) $X(t_j)$, младший разряд совпадает, соответственно, с двоичной переменной $X(t_j - (m-1))$. Таким образом, в окне шириной m будет находиться двоичный позиционный код, имеющий взаимно однозначное соответствие с целым числом $d^m(T)$, для момента времени $t_i = T$

$$d^m(T) = \sum_{i=0}^{m-1} X(T-i)2^{m-i-1}, \qquad (2)$$

где m — ширина скользящего окна; $i = (0, 1, ..., m)$. Из формулы (2) непосредственно следует общий вид j-го члена ряда (1): $d_j^m(T-j) = \sum_{i=0}^{m-1} X(T - -(j+i))2^{m-i-1}$, где $j = 0, 1, ..., N$. Применяя процедуру «скользящего окна» для кортежа (1) и смещая окно на один шаг (одну позицию) от $t = 0$ к

Таблица 1

Сегменты и их хэш-коды для ВР-Ц акций Сбербанка (за период с 20.02 по 05.03.2008 г.)

Дата	Сегмент				ВР-Ц	Разность	Знак разности	Значение хэш-кода сегмента			
	1	2	3	4				1	2	3	4
20.02					83,09	—	—	—	—	—	—
21.02	×				83,75	0,66	1	5	—	—	—
22.02	×	×			82,99	−0,76	0		2	—	—
26.02	×	×	×		85,10	2,11	1			1	—
27.02	×	×	×	×	83,78	−1,32	0				32
28.02	×	×	×	×	82,95	−0,83	0				
29.02	×	×	×	×	80,20	−2,75	0				
03.03		×	×	×	78,70	−1,50	0	—			
04.03			×	×	76,50	−2,20	0	—	—		
05.03				×	78,32	1,82	1	—	—	—	

$t = T$, получим последовательность целых чисел, которые и являются хэш-кодами окон, соответствующих сегментам анализируемого ВР-Ц. В качестве примера в табл. 1 приведены значения ВР-Ц для ПАО «Сбербанк», разность цен $\Delta C(t_j)$, вычисленная для смежных дней на основе ВР-Ц, знак этой разности, кортеж из единиц и нулей, характеризующий знак разности, и хэш-коды, которые рассчитаны для четырех шестидневных сегментов. Будем считать, что местоположение старшего разряда двоичного позиционного кода, который определяет хэш-код, всегда соответствует последней дате шестидневного сегмента. Заметим, что для скользящего окна шириной $m = 6$ будет 64 хэш-кода, т. е. 64 различных типовых сегмента.

3. ИСПОЛЬЗОВАНИЕ ХЭШ-КОДОВ ДЛЯ АНАЛИЗА И ПРОГНОЗИРОВАНИЯ

Исходные ВР-Ц и ВР-ХК позволяют совместно провести анализ траектории движения цены биржевого актива. Предлагается методика такого анализа, суть которой состоит в следующем.

1. В распоряжении инвестора всегда имеются данные о котировочной цене биржевого актива за интересующий интервал времени вплоть до текущего момента T.

2. Динамику движения цены отражает существующий ВР-Ц на всех интервалах $(T - j)$, где $j = 0, 1, ..., N - 1$, N — объем выборки.

3. На всех интервалах, кроме начального, вместе с ценой известно значение хэш-кода (элемент ВР-ХК). Другими словами, между сегментами ВР-Ц и ВР-ХК существует соответствие — каждому сегменту ВР-Ц соответствует элемент ВР-ХК.

4. В текущий момент времени T всегда известен хэш-код $d_0^m(T)$, который определим как *текущий хэш-код*, а сегмент ВР-Ц, соответствующий текущему хэш-коду, определим как *текущий сегмент* ВР-Ц. Текущий хэш-код является последним по времени хэш-кодом в выборке.

Методика состоит из двух этапов — набора статистики и прогнозирования.

На первом этапе набираются биржевые данные, которые используются для получения статистики частоты вхождения в выборку типовых сегментов и хэш-кодов. С этой целью для каждого типа хэш-кода определяются два параметра. Первый — число вхождений в выборку самого хэш-кода и второй — это число вхождений в выборку пары, состоящей из хэш-кода данного типа и хэш-кода, непосредственно следующего за ним, например, при падении котировок.

В начале второго этапа (прогнозирование) последний в выборке хэш-код рассматриваемого типа принимается искусственно за текущий и относительно него осуществляется прогноз. Правильность прогноза, т. е. значение прогнозируемого хэш-кода, проверяется на имеющейся выборке. Таким образом, вначале все хэш-коды разделяются на два класса: первый, у которого результаты прогноза совпали с историей котировок, и второй, — там, где этого не произошло. По численности первого и второго класса можно сделать вывод о целесообразности применения этого метода прогнозирования.

Затем начинает работать процедура прогнозирования для текущего хэш-кода. Далее приведено более детальное ее описание.

1. Прогнозирование движения цены (рост или падение) сводится к выделению в последовательности ВР-ХК всех хэш-кодов, равных текущему $d_0^m(T)$.

2. После этого необходимо выделить пары хэш-кодов $d_0^m(T)$, $d_2^m(T+1)$, следующие за $d_0^m(T)$ (индекс 2 указывает на падение или неизменность котировок, а индекс 1 — соответствует их росту).

3. Определить наиболее вероятный хэш-код из этой пары можно с помощью следующего алгоритма.

3.1. Определим число вхождений V_0 хэш-кода $d_0^m(T)$ в ряд ВР-ХК.

3.2. Аналогично п. 3.1 определим число вхождений V_2 для пары хэш-кодов, непосредственно следующих один за другим $d_0^m(T)$, $d_2^m(T+1)$.

3. Оценим вероятность P_{02} наступления события $d_2^m(T+1)$:

$$P_{02} = V_2/V_0. \quad (3)$$

4. Оценим вероятность наступления события $d_1^m(T+1)$:

$$P_{01} = 1 - P_{02}. \quad (4)$$

Соотношения (3) и (4) позволяют на основе имеющихся статистических данных сделать вывод о возможном направлении движения цены актива: если $P_{01} > P_{02}$, то выбираем хэш-код «рост цены»; если $P_{01} \le P_{02}$, то выбираем хэш-код «падение цены».

4. ПРОВЕРКА МЕТОДИКИ НА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Разработанная методика проверялась на данных как для срочного рынка FORTS биржи РТС, так и для рынка акций ММВБ.

В табл. 2 представлены фрагменты ВР-Ц (индекс РТС) и ВР-ХК, полученного для скользящего окна шириной $m = 6$.

В табл. 3 представлен результат обработки массива данных индекса РТС с шагом по времени (таймфреймом), равном одному дню (период с 29.11.2011 по 29.01.2013 г.) для хэш-кода 25. Как было сказано ранее, каждому хэш-коду можно поставить в соответствие типовую группу сегментов, полученных с помощью процедуры скользящего окна. Тип этого сегмента определяется хэш-кодом

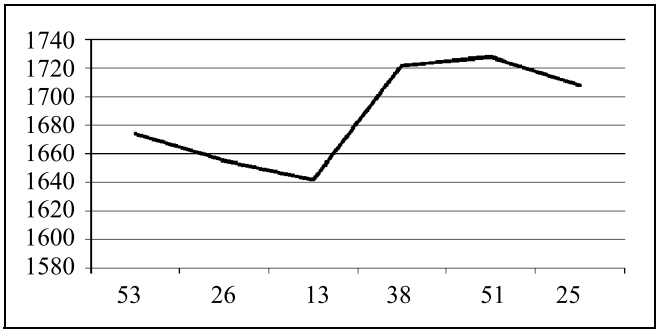

Рис. 1. Типовой сегмент 1, 28.02.2012 г., хэш-код 25 индекса РТС

последнего по времени элемента ВР-Ц, входящего в окно. На рис. 1 показан типовой сегмент хэш-кода 25 для окна шириной $m = 6$ (в табл. 3 это третья строка).

По оси ординат указано значение индекса РТС, по оси абсцисс — последовательность хэш-кодов. Последним по времени является хэш-код 25. Этот хэш-код и определяет типовую группу. Пример другого сегмента из этой же группы показан на рис. 2 (в табл. 3 — это вторая строка).

Визуально эти сегменты отличаются друг от друга, однако общим признаком для этой типовой группы (как и для всех остальных типовых групп) служит рост или падение котировочной цены при

Таблица 2

Фрагменты ВР-Ц и ВР-ХК

Дата	Индекс РТС	Хэш-код
17.10.2012	1513,96	50
18.10.2012	1512,01	25
19.10.2012	1494,44	12
22.10.2012	1497,63	38
23.10.2012	1456,73	19
24.10.2012	1462,43	41
25.10.2012	1456,91	20
26.10.2012	1441,38	10
29.10.2012	1435,05	05

Таблица 3

Выборка с использованием хэш-кода 25

Дата	Индекс РТС	Хэш-код
21.01.2013	1600,13	25
18.10.2012	1512,01	25
28.02.2012	1708,16	25
29.11.2011	1466,36	25

переходе от одной временной точки в пределах окна (сегмента) к другой. Так, например, для окна ширины 6 всего будет 64 различные типовые группы, которые можно пронумеровать от 0 до 63, и в каждой такой группе одноименные позиции будут «вести» себя одинаково (цены будут расти или падать), но при этом сами размеры роста или падения будут различаться. Максимальное число сегментов $R_{\max}$ при ширине окна m и выборке N равно $N - m - 1$.

Данная методика сегментации и прогнозирования была проверена на акциях Газпрома. Анализируемый период составлял 1416 дней с 10.12.2007

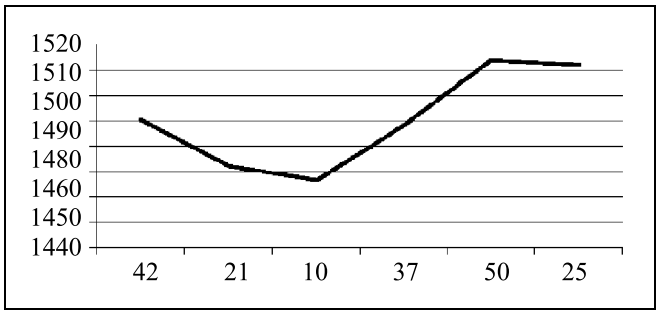

Рис. 2. Типовой сегмент 2, 18.10.2012 г., хэш-код 25 индекса РТС

Таблица 4

Результаты прогнозирования движения котировок акций Газпрома за период с 10.12.2007 по 16.08. 2013 г.

1	2	3	4	5	6	7	1	2	3	4	5	6	7
0	17	6	0,35	**0,65**	1		33	21	8	0,38	0,62	1	
1	22	10	0,45	0,55	1		34	25	16	**0,64**	0,36	1	
2	19	9	0,47	0,53	1		35	29	18	**0,62**	0,38	1	
3	23	12	0,52	0,48	1		36	23	13	0,57	0,43	1	
4	26	11	0,42	0,58		1	37	23	12	0,52	0,48		1
5	18	7	0,39	**0,61**	1		38	27	14	0,52	0,48	1	
6	30	15	0,50	0,50			39	17	6	0,35	**0,65**	1	
7	22	7	0,32	**0,68**		1	40	24	6	0,25	**0,75**	1	
8	26	16	**0,62**	0,38	1		41	24	11	0,46	0,54	1	
9	23	9	0,39	**0,61**		1	42	14	5	0,36	**0,64**	1	
10	24	8	0,33	**0,67**	1		43	18	14	**0,78**	0,22	1	
11	18	10	0,56	0,44	1		44	20	7	0,35	**0,65**		1
12	32	15	0,47	0,53	1		45	28	14	0,50	0,50		
13	26	15	0,58	0,42	1		46	16	10	**0,63**	0,38	1	
14	16	10	**0,63**	0,38	1		47	11	2	0,18	**0,82**		1
15	24	11	0,46	0,54		1	48	27	14	0,52	0,48		1
16	15	7	0,47	0,53		1	49	19	8	0,42	0,58	1	
17	35	18	0,51	0,49		1	50	19	10	0,53	0,47		1
18	26	12	0,46	0,54		1	51	23	16	**0,70**	0,30	1	
19	21	10	0,48	0,52	1		52	30	19	**0,63**	0,37	1	
20	18	10	0,56	0,44	1		53	12	6	0,50	0,50		
21	20	13	**0,65**	0,35	1		54	27	15	0,56	0,44	1	
22	22	9	0,41	0,59		1	55	14	11	**0,79**	0,21		1
23	13	7	0,54	0,46		1	56	23	11	0,48	0,52	1	
24	23	14	**0,61**	0,39		1	57	15	7	0,47	0,53	1	
25	28	16	0,57	0,43		1	58	16	6	0,38	**0,63**	1	
26	26	13	0,50	0,50			59	13	6	0,46	0,54	1	
27	28	12	0,43	0,57		1	60	19	7	0,37	**0,63**	1	
28	19	9	0,47	0,53	1		61	18	7	0,39	**0,61**	1	
29	13	6	0,46	0,54		1	62	20	12	**0,60**	0,40	1	
30	15	8	0,53	0,47		1	63	11	7	**0,64**	0,36	1	
31	20	15	0,75	0,25	1						**Итого**	**40**	**20**
32	22	7	0,32	**0,68**	1								

по 16.08.2013 г. В качестве основного таймфрейма был принят день. Объем выборки составлял 1416 значений. В табл. 4 приведены результаты прогнозирования.

Результаты анализа и прогнозирования роста и падения котировок акций с использованием хэш-кодов, представленных в табл. 4, можно интерпретировать следующим образом.

- Предложенный алгоритм прогнозирования носит вероятностный характер, и на примере акций Газпрома за указанный период вероятность совпадения реального роста и падения котировок с прогнозом составляет 0,66, что по оценкам экспертов-брокеров является достаточно высоким результатом, поскольку при прогнозировании не использовалось никакой другой информации, кроме временного ряда цен исследуемого актива (инсайдерской информации, цен на нефть, золото или другие базовые активы, биржевые индексы и др.).
- Вопреки устоявшемуся мнению, что значения вероятностей роста и падения должны быть близки к 0,5, из табл. 4 следует, что более 42 % значений равны или превышают 0,6 (в табл. 4 они помечены полужирным шрифтом). Результат 0,5 (значения вероятностей роста и падения равны) получены только для хэш-кодов 6, 26, 45 и 53.
- По данным, для которых значения соответствующих вероятностей равны или превышают 0,6, совпадение реального движения цены с прогнозом составляет около 80 %.

По столбцам в табл. 4 указано: 1 — все хэш-коды, полученные в последовательности BP-XK; 2 — число вхождений V_0 каждого хэш-кода в BP-XK за указанный период; 3 — число вхождений V_1 пары из двух хэш-кодов: хэш-код, указанный в первом столбце и хэш-код смежный с ним при падении цены; 4 — значение оценки вероятности при падении цены (см. формулу (3)); 5 — значение оценки вероятности при росте цены (см. формулу (4)); 6 — результат прогноза и реальное движение цены за указанный период совпали; 7 — результаты не совпали.

ЗАКЛЮЧЕНИЕ

Предложен новый метод прогнозирования на фондовом рынке, базирующийся на процедурах сегментации и хэширования соответствующих временных рядов. Благодаря использованию хэш-кодов удается существенно упростить задачу прогноза — прогнозируется не значение характеристики, а только ее знак, т. е. «увеличение» или «уменьшение» значения этой характеристики, а именно такой прогноз и интересует большинство «игроков» фондового рынка. Поскольку сами хэш-коды представляют собой целые числа, то их последовательность позволяет выделить в динамике изменения цены биржевого актива одинаковые (типовые) группы членов временного ряда (сегменты). Анализируемые сегменты временного ряда цены торгуемого на бирже актива получаются с помощью процедуры скользящего окна заданного размера. Типовая группа — это группа сегментов, для которой рост или падение котировочной цены в каждом сегменте происходит в одной и той же последовательности во времени. Другими словами, хэш-коды позволяют разделить исходный временной ряд цены актива на сегменты одинаковой длины (равной размеру скользящего окна) и разделить все множество сегментов на типовые группы.

Прогнозирование осуществляется на базе информации о хэш-кодах сегментов типовой группы для последнего по времени сегмента временного ряда.

Разработанная методика проверялась на данных как для срочного рынка FORTS биржи РТС (индекс РТС), так и для рынка акций ММВБ (акции Газпрома). Полученные результаты подтвердили эффективность предложенного метода.

ЛИТЕРАТУРА

1. *О'Нил У.Дж.* Как делать деньги на фондовом рынке. Стратегия торговли на росте и падении. — М.: Изд. дом «Альпина», 2003. — 328 с.
2. *Толли Т.* Игра на понижение или техника «коротких» продаж. Правила игры финансовых топ-менеджеров Уолл-стрит на фондовом рынке. — М.: Торговый дом «Грант», 2004. — 366 с.
3. *Ахо А., Хопкрофт Д.* Структуры данных и алгоритмы. — М.: Изд. дом «Вильямс», 2000. — С. 116—127.
4. *Кормен Т.Х., Лейзерсон Ч.И., Ривест Р.Л., Штайн К.* Алгоритмы: Построение и анализ: 2-е изд. — М.: Изд. дом «Вильямс», 2005. — С. 282—315.
5. *Спиро А.Г., Дорофеюк Ю.А.* Структурно-графовой подход к анализу фондового рынка // Проблемы управления. — 2011. — № 6. — С. 61—65.

Статья представлена к публикации членом редколлегии А.С. Манделем.

Спиро Арнольд Григорьевич — канд. техн. наук, ст. науч. сотрудник, ✉ arn.spi@mail.ru,

Гольдовская Марина Дмитриевна — науч. сотрудник, ✉ mdgold54@mail.ru,

Киселева Нелли Евсеевна — науч. сотрудник, ✉ lab55@ipu.ru,

Покровская Ирина Вячеславовна — науч. сотрудник, ✉ ivp750@mail.ru,

Институт проблем управления им. В.А. Трапезникова РАН, г. Москва.