

**О ПРИМЕНЕНИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА
ВРЕМЕННЫХ РЯДОВ В ЗАДАЧАХ СОКРАЩЕНИЯ
РАЗМЕРНОСТИ ПРОСТРАНСТВА ПРИЗНАКОВ НА ОСНОВЕ
ВЫЯВЛЕНИЯ ЗАВИСИМОСТЕЙ¹**

Гришина Е.Н., Рыжова М.Н.
Кафедра информационных технологий

Поступила в редакцию 14.08.2014, после переработки 08.09.2014.

В статье развивается подход к выявлению зависимостей между исходными параметрами исследуемой системы управления, основанный на расчете меры ассоциации. Проведены расчеты меры ассоциации и коэффициента корреляции Пирсона. Полученные результаты применены к решению задачи сокращения размерности пространства признаков с целью дальнейшего применения к решению проблемы прогнозирования неисправностей системы управления электрооборудованием вагона. Представлен сравнительный анализ полученных результатов.

Ключевые слова: интеллектуальный анализ данных, мера сходства, преобразование скользящих аппроксимаций, временной ряд, железнодорожный транспорт.

Вестник ТвГУ. Серия: Прикладная математика. 2014. № 3. С. 95–104.

Введение

Анализ технического состояния подвижного состава железнодорожного транспорта играет ключевую роль в совершенствовании транспортной инфраструктуры [8].

Современный железнодорожный транспорт в настоящее время комплектуется электрооборудованием, позволяющим выполнять мониторинг текущего состояния соответствующего оборудования. Однако, так как количество установленных на вагоне датчиков достаточно велико, то возникает проблема сокращения размерности получаемых исходных данных. С целью выявления зависимостей в работе датчиков устройств вагонного оборудования, а также решения задачи сокращения размерности данных, представляется целесообразным применение методов интеллектуального анализа данных.

Методы интеллектуального анализа данных, в частности методы интеллектуального анализа временных рядов, находят свое применение в анализе экономических и финансовых баз данных, метеорологических и геофизических баз данных, медицинских баз данных, а также при анализе нефтяных месторождений [4].

¹Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-13160-офи_м_РЖД.

Несмотря на то, что анализ состояния транспортного оборудования является малоисследованной областью, в настоящее время активно изучается возможность применения методов интеллектуального анализа данных в сфере железнодорожного транспорта.

1. Описание математической модели данных

Будем рассматривать структуру данных системы контроллера управления электрооборудованием вагона (КУЭВ), приведенную в статье [6].

Система электрооборудования, установленного на вагоне, имеет сложную структуру, в связи с чем необходим индивидуальный подход для подготовки и анализа данных с датчиков различных устройств.

КУЭВ принимает от датчиков контроллеров вагонного оборудования дискретные и аналого-цифровые сигналы обратной связи и представляет их в виде значений параметров, каждый из которых имеет в системе свое имя и принимает значения из заданного диапазона.

Значения каждого датчика, поступающие от КУЭВ с заданной периодичностью, определяют собой временной ряд. Обозначим i -ый временной ряд как

$$X^i = (x_j^i, t_j^i), \quad i = \overline{1, m}, \quad j = \overline{1, n}, \quad (1)$$

где x_j^i – значение i -го датчика в j -ый временной отсчет, t_j^i – значение j -го временного отсчета для i -го датчика, m – количество датчиков в системе, n – количество временных отсчетов в анализируемом интервале времени. В предположении, что ряды являются равноотстоящими и отсчеты времени для всех рядов являются одинаковыми ($t_j^{i_k} = t_j^{i_{k+1}}, \forall k, i, j$), можно опустить вторую составляющую ряда и оперировать только его значениями x_j^i .

В зависимости от области принимаемых значений ряды в модели данных бывают двух типов:

- дискретные временные ряды $x_j^i \in T^i = \{T_1^i, T_2^i, \dots, T_Q^i\}$, где T^i – дискретное множество значений данного ряда, $i = \overline{1, Q}$, $j = \overline{1, n}$,
- непрерывные временные ряды $x_j^i \in R$, где R – вещественная ось, $i = \overline{1, m}$, $j = \overline{1, n}$.

Количество параметров системы достаточно велико, что приводит к проблеме оперирования данными большой размерности. Для выявления связей между рассматриваемыми временными рядами целесообразно проанализировать всю исходную совокупность имеющихся временных рядов. Применение описанного в статье подхода направлено на выявление существующих зависимостей между исходными временными рядами и сокращение размерности пространства признаков.

2. Сокращение размерности на основе расчета меры ассоциации между временными рядами

В работе [1] определяются базовые понятия, а также вводится понятие ассоциативной функции, являющейся гибким инструментом, позволяющим выявлять ассоциации между временными рядами, имеющими сложную структуру.

Рассмотрим аппарат, предложенный в работе [1], и применим его к задаче анализа данных, поступающих с датчиков вагонного оборудования.

Введем необходимые понятия и определения для ряда вида (1).

Окно M_i длины $w > 1$ – это последовательность индексов $M_i = (i, i+1, \dots, i+w-1)$, $i \in \{1, \dots, n-w+1\}$, n – количество временных отсчетов в анализируемом интервале времени. Тогда $x_{M_i} = (x_i, x_{i+1}, \dots, x_{i+w-1})$ – значения ряда X в окне M_i .

Определение 1. Последовательность $J = (M_1, M_2, \dots, M_{n-w+1})$ всех окон фиксированной длины w ($1 < w \leq n$) называется скользящим окном.

Определение 2. Линейная функция $f_i(t) = a_i t + b_i$ с параметрами a_i, b_i , минимизирующая критерий

$$Q(f_i, x_{M_i}) = \sum_{j=i}^{i+w-1} (f_i(t_j) - x_j)^2 = \sum_{j=1}^{i+w-1} (a_i t_j + b_i - x_i)^2, \quad (2)$$

называется линейной регрессией или линейной аппроксимацией значения x_{M_i} временного ряда X в окне M_i .

Значение a_i может быть вычислено по следующей формуле:

$$a_i = \frac{\sum_{j=i}^{i+w-1} t_j x_j - \left(\sum_{j=i}^{i+w-1} t_j \right) \sum_{j=i}^{i+w-1} x_j}{\sum_{j=i}^{i+w-1} t_j^2 - \left(\sum_{j=i}^{i+w-1} t_j \right)^2}. \quad (3)$$

Определение 3. Преобразованием скользящих аппроксимаций (САП преобразованием) называется преобразование

$$MAP_w(x, t) = a, \quad (4)$$

где $a = (a_1, \dots, a_{n-w+1})$ – последовательность угловых коэффициентов a_i линейных аппроксимаций временного ряда в скользящем окне длины w . Значения $a = (a_1, \dots, a_{n-w+1})$ называются локальными трендами.

Определение 4. Мерой ассоциации локальных трендов временных рядов y и x называется функция

$$\begin{aligned} \text{coss}_w(x, y) &= \cos(\angle MAP_w(y, t), \angle MAP_w(x, t)) = \\ &= \frac{MAP_w(y, t) \cdot MAP_w(x, t)}{|MAP_w(y, t)| \cdot |MAP_w(x, t)|} = \frac{\sum_{i=1}^m a_{yi} \cdot a_{xi}}{\sqrt{\sum_{i=1}^m a_{yi}^2 \cdot \sum_{j=1}^m a_{xi}^2}}. \end{aligned} \quad (5)$$

Наиболее полную информацию об ассоциациях между временными рядами дает последовательность ассоциации локальных трендов, подсчитанных в общем случае для некоторого подмножества значений скользящих окон $K \subseteq \{1, \dots, n\}$.

Определение 5. Ассоциативной функцией временных рядов y и x называется функция

$$AF(y, x) = (\text{coss}_1(y, x), \dots, \text{coss}_n(y, x)). \quad (6)$$

Максимальное или среднее значение ассоциативной функции может быть использовано как мера ассоциации между временными рядами:

$$AM(y, x) = \max(AF_w(y, x)) = (\text{coss}_w(y, x)), \quad (7)$$

$$AM(y, x) = \frac{\sum_{w \in K} AF_w(y, x)}{|K|} = \frac{\sum_{w \in K} \text{coss}_w(y, x)}{|K|}. \quad (8)$$

Данная мера имеет преимущества по сравнению с известными мерами ассоциации и сходства между временными рядами благодаря ее инвариантности и способности оценивать ассоциации между временными рядами для разных временных интервалов [1].

3. Численный эксперимент

Описанный выше подход применим к расчету ассоциации между временными рядами для анализа взаимодействия датчиков устройств вагонного оборудования. Рассмотрим 700 измерений, полученных с 12 датчиков устройства типа КОНТ_КУЕВ. Вычисление парных ассоциации между временными рядами осуществим на основе меры (7).

Для выявления силы связи между временными рядами представляется целесообразным расчет меры ассоциации для групп окон различной длины. Также представляется интересным сравнение полученных результатов с результатами корреляционного анализа применительно к тем же данным.

3.1 Результаты расчета коэффициента корреляции Пирсона

Корреляционный анализ является одним из способов выявления связей между параметрами исходной выборки [7]. Корреляционная зависимость представляет собой статистическую взаимосвязь двух или нескольких случайных величин. При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин [2].

Степень взаимозависимости случайных величин определяется коэффициентом корреляции, наиболее известным из которых является коэффициент Пирсона [5].

В Таблице 1 приведены результаты расчета значений парной корреляции временных рядов датчиков устройства КОНТ-КУЕВ.

По результатам расчетов, приведенным в Таблице 1, можно судить о наличии связи между рядами V1 и V5, V1 и V9, V2 и V10, V2 и V11, V3 и V5, V3 и V9, V4 и V9, V5 и V9, V7 и V8, V8 и V10, V8 и V11 (на уровне значения коэффициента корреляции больше 0,5). Особо необходимо отметить выявленную сильную связь между временными рядами V10 и V11.

Отметим, что наличие корреляции позволяет судить о существовании некоторой статистической связи в анализируемой выборке, при этом необходимо учитывать, что такая связь не всегда присуща другой выборке и имеет причинно-следственный характер [3].

Таблица 1: Матрица значений парной корреляции временных рядов датчиков устройства КОИТ-KUEV

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
V1	1	0,07	0,45	-0,3	0,64	0,15	-0,04	0,04	0,64	0,05	0,05
V2	0,07	1	-0,2	0,42	-0,2	0,11	-0,3	0,26	-0,4	-0,5	-0,5
V3	0,45	-0,2	1	0,02	0,55	0,06	0,27	-0,1	0,63	0,21	0,21
V4	-0,3	0,42	0,02	1	-0,2	0	0	-0,1	-0,5	-0,1	-0,1
V5	0,64	-0,2	0,55	-0,2	1	0,05	0,21	-0,1	0,59	0,23	0,23
V6	0,15	0,11	0,06	0	0,05	1	-0,3	0,46	0,14	-0,2	-0,2
V7	-0,04	-0,3	0,27	0	0,21	-0,3	1	-0,5	0,07	0,44	0,44
V8	0,04	0,26	-0,1	-0,1	-0,1	0,46	-0,5	1	0,13	-0,5	-0,5
V9	0,64	-0,4	0,63	-0,5	0,59	0,14	0,07	0,13	1	0,12	0,11
V10	0,05	-0,5	0,21	-0,1	0,23	-0,2	0,44	-0,5	0,12	1	1
V11	0,05	-0,5	0,21	-0,1	0,23	-0,2	0,44	-0,5	0,11	1	1

3.2 Результаты расчета меры ассоциации между временными рядами

В Таблицах 2, 3, 4 приведены значения меры ассоциации временных рядов для групп окон различной длины. Ассоциативная мера рассчитывается по формуле (7).

Таблица 2: Матрица значений меры ассоциации временных рядов для окон длины $2 \div 3$

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
V1	1	0,03	0,05	0,08	-0,04	0,02	0,08	0,07	0,41	-0,01	-0,02
V2	0,03	1	0,02	-0,01	0,03	-0,06	0,04	-0,02	0,04	-0,02	-0,04
V3	0,05	0,02	1	0,29	0,06	0,05	0,01	0,01	0,04	-0,55	-0,55
V4	0,08	-0,01	0,29	1	0,03	0,06	0	0	0,1	-0,19	-0,19
V5	-0,04	0,03	0,06	0,03	1	0,07	-0,06	0,02	0,06	-0,01	-0,02
V6	0,02	-0,06	0,05	0,06	0,07	1	0,06	0,27	0	-0,02	-0,02
V7	0,08	0,04	0,01	0	-0,06	0,06	1	0,05	0,03	-0,07	-0,07
V8	0,07	-0,02	0,01	0	0,02	0,27	0,05	1	0,05	-0,03	-0,02
V9	0,41	0,04	0,04	0,1	0,06	0	0,03	0,05	1	-0,03	-0,03
V10	-0,01	-0,02	-0,55	-0,19	-0,01	-0,02	-0,07	-0,03	-0,03	1	0,98
V11	-0,02	-0,04	-0,55	-0,19	-0,02	-0,02	-0,07	-0,02	-0,03	0,98	1

Для более наглядного представления сведем воедино результаты расчетов для пар рядов с выявленными связями (на уровне значения коэффициента корреляции/меры ассоциации выше 0,5). Сводные результаты представлены в Таблице 5.

Таким образом, на основе анализа приведенных выше таблиц можно говорить о том, что по результатам расчета коэффициента корреляции выявлены зависимости между датчиками устройства КОИТ-KUEV вагона, представленными временными рядами V1 и V5, V1 и V9, V2 и V10, V2 и V11, V3 и V5, V3 и V9, V4 и V9, V5 и V9, V7 и V8, V8 и V10, V8 и V11, V10 и V11. При этом при расчете меры ассоциации для окон длины $2 \div 3$ также было выявлено наличие связи между временными рядами V3 и V10, V3 и V11. Однако данная связь пропадает при

Таблица 3: Матрица значений меры ассоциации временных рядов для окон длины $2 \div 10$

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
V1	1	0,04	0,13	0,11	-0,03	0,03	0,08	0,07	0,41	0,01	0
V2	0,04	1	0,04	0	0,03	-0,05	0,13	0,14	0,05	-0,02	-0,04
V3	0,13	0,04	1	0,36	0,06	0,06	0,03	0,01	0,05	-0,08	-0,08
V4	0,11	0	0,36	1	0,03	0,06	0	0,04	0,1	-0,01	-0,01
V5	-0,03	0,03	0,06	0,03	1	0,09	0,05	0,03	0,06	0,01	0
V6	0,03	-0,05	0,06	0,06	0,09	1	0,06	0,28	0,06	0,01	0,02
V7	0,08	0,13	0,03	0	0,05	0,06	1	0,06	0,03	-0,07	-0,07
V8	0,07	0,14	0,01	0,04	0,03	0,28	0,06	1	0,05	-0,03	-0,02
V9	0,41	0,05	0,05	0,1	0,06	0,06	0,03	0,05	1	-0,03	-0,03
V10	0,01	-0,02	-0,08	-0,01	0,01	0,01	-0,07	-0,03	-0,03	1	1
V11	0	-0,04	-0,08	-0,01	0	0,02	-0,07	-0,02	-0,03	1	1

Таблица 4: Матрица значений меры ассоциации временных рядов для окон длины $2 \div 20$

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11
V1	1	0,07	0,14	0,11	0,1	0,21	0,08	0,07	0,41	0,01	0
V2	0,07	1	0,04	0,03	0,03	0,06	0,13	0,41	0,18	-0,02	-0,04
V3	0,14	0,04	1	0,36	0,06	0,09	0,04	0,01	0,05	-0,04	-0,04
V4	0,11	0,03	0,36	1	0,03	0,25	0	0,1	0,1	0,03	0,03
V5	0,1	0,03	0,06	0,03	1	0,12	0,16	0,03	0,12	0,02	0,02
V6	0,21	0,06	0,09	0,25	0,12	1	0,06	0,28	0,06	0,06	0,06
V7	0,08	0,13	0,04	0	0,16	0,06	1	0,06	0,03	-0,07	-0,07
V8	0,07	0,41	0,01	0,1	0,03	0,28	0,06	1	0,1	-0,03	-0,02
V9	0,41	0,18	0,05	0,1	0,12	0,06	0,03	0,1	1	-0,03	-0,03
V10	0,01	-0,02	-0,04	0,03	0,02	0,06	-0,07	-0,03	-0,03	1	1
V11	0	-0,04	-0,04	0,03	0,02	0,06	-0,07	-0,02	-0,03	1	1

Таблица 5: Сводная таблица результатов

Пары рядов	Коэффициент корреляции	Мера ассоциации при окнах длины 2 ÷ 3	Мера ассоциации при окнах длины 2 ÷ 10	Мера ассоциации при окнах длины 2 ÷ 20
V1, V5	0,64	-0,04	-0,03	0,1
V1, V9	0,64	0,41	0,41	0,41
V2, V10	-0,5	-0,02	-0,02	-0,02
V2, V11	-0,5	-0,04	-0,04	-0,04
V3, V5	0,55	0,06	0,06	0,06
V3, V9	0,63	0,04	0,05	0,05
V4, V9	-0,5	0,1	0,1	0,1
V5, V9	0,59	0,06	0,06	0,12
V7, V8	-0,5	0,05	0,06	0,06
V8, V10	-0,5	-0,03	-0,03	-0,03
V8, V11	-0,5	-0,02	-0,02	-0,02
V10, V11	1	0,98	1	1
V3, V10	0,21	-0,55	-0,08	-0,04
V3, V11	0,21	-0,55	-0,08	-0,04

расчете меры ассоциации для окон большей длины.

Отметим, что оценка степени связи между временными рядами с помощью коэффициента корреляции Пирсона в некоторых случаях значительно отличается от значений меры ассоциации. Также значение меры ассоциации на окнах малой длины может уточняться путем включения в расчеты окон большей длины. Однако, начиная с определенной совокупности длин окон, значение меры ассоциации для временных рядов стабилизируется или изменяется незначительно.

По результатам численного эксперимента можно говорить о том, что стабильной сильной связью является только связь между датчиками устройства КОНТ-КУЕВ вагона, представленными временными рядами V10 и V11 (Рис. 1).

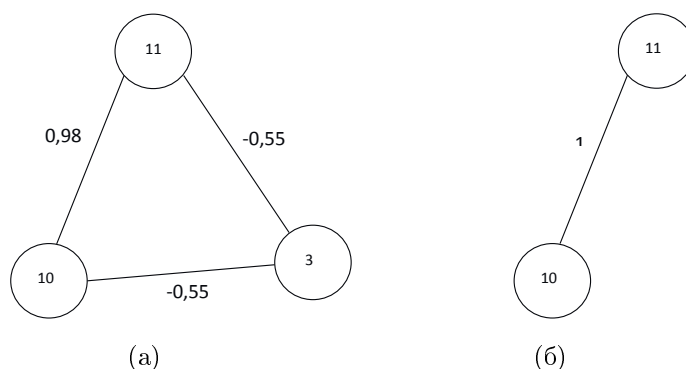


Рис. 1: Результаты расчетов меры ассоциации на уровне выше 0,5 для (а) окон длины 2 ÷ 3, (б) окон длины 2 ÷ 10

Рассмотренный подход к выявлению зависимостей между датчиками устройств, представленных временными рядами, основанный на расчете меры ассоциации, дает возможность для решения задачи сокращения размерности исходных данных, поступающих с электрооборудования вагона.

Так, делая вывод о наличии сильной связи между датчиками устройства КОИТ-КУЕВ вагона, представленными временными рядами V10 и V11, мы можем судить об их схожем поведении. Таким образом, в дальнейшем при решении задачи прогнозирования поведения электрооборудования вагона представляется целесообразным рассмотреть только одного из указанных датчиков.

Заключение

Таким образом, в статье рассмотрен подход, основанный на расчете меры ассоциации между временными рядами, направленный на выявление связи между датчиками вагонного оборудования.

Сравнительный анализ результатов, полученных в ходе численного эксперимента с использованием реальных данных, позволяет сделать вывод о наличии связи между датчиками вагонного оборудования, а также решить задачу сокращения размерности анализируемых данных.

В дальнейшем на основе использования результатов данной работы, а также результатов, представленных в работе [9], планируется проведение исследований, направленных на решение задачи прогнозирования возникновения неисправностей вагонного оборудования.

Список литературы

- [1] Batyrshin I., Herrera-Avelar R., Sheremetov L., Panova A. Moving approximation transform and local trend associations in time series data bases // In: Studies in Computational Intelligence. Vol. 36. Perception-Based Data Mining and Decision Making in Economics and Finance. Springer, 2007. Pp. 55–83.
- [2] Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989. 607 с.
- [3] Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. М.: Мир, 1982. 488 с.
- [4] Батыршин И.З., Кошульски А., Шереметов Л.Б., Климова А.С., Панова А.М. Анализ взаимодействия нефтяных скважин на основе гибридной кластеризации временных рядов продуктивности скважин // Нечеткие системы и мягкие вычисления. 2007. Т. 2, № 4. С. 63–73.
- [5] Гмурман В.Е. Теория вероятностей и математическая статистика: Учебное пособие для вузов. 10-е издание, стереотипное. М.: Высшая школа, 2004. 479 с.
- [6] Гришина Е.Н., Солдатенко И.С. О сокращении размерности пространства признаков в задачах прогнозирования неисправностей вагонного оборудования // Нечеткие системы и мягкие вычисления. 2012. Т. 7, № 2. С. 73–80.

- [7] Елисеева И.И., Юзбашев М.М. Общая теория статистики: Учебник / Под ред. чл.-корр. РАН И.И. Елисеевой. 4-е издание, переработанное и дополненное. М.: Финансы и статистика, 2001. 480 с.
- [8] Иванова Е.И., Гордеев Р.Н., Михайлов В.В., Северов А.В., Язенин А.В. Модель централизованной интеллектуальной информационной системы для решения задач диагностики и прогнозирования неисправностей вагонного оборудования и управления им на железнодорожном транспорте // Нечеткие системы и мягкие вычисления. 2012. Т. 7, № 2. С. 51–72.
- [9] Солдатенко И.С., Гришина Е.Н. О некоторых подходах к построению интеллектуальных моделей прогноза возникновения неисправностей вагонного оборудования // Нечеткие системы и мягкие вычисления. 2012. Т. 7, № 2. С. 81–88.

Библиографическая ссылка

Гришина Е.Н., Рыжова М.Н. О применении интеллектуального анализа временных рядов в задачах сокращения размерности пространства признаков на основе выявления зависимостей // Вестник ТвГУ. Серия: Прикладная математика. 2014. № 3. С. 95–104.

Сведения об авторах

1. **Гришина Елена Николаевна**

начальник отдела аспирантуры и докторантуры, доцент кафедры информационных технологий Тверского государственного университета.

Россия, 170100, г. Тверь, ул. Желябова, д. 33, ТвГУ.

E-mail: mail.grishina@gmail.com.

2. **Рыжова Маргарита Николаевна**

магистрант кафедры информационных технологий Тверского государственного университета.

Россия, 170100, г. Тверь, ул. Желябова, д. 33, ТвГУ.

E-mail: mryzhova@tversu.ru.

**ON APPLICATION OF TIME SERIES DATA MINING
IN A PROBLEM OF DIMENSION REDUCTION
OF CHARACTERISTICS BASED ON DEPENDENCIES REVELATION**

Grishina Elena Nikolaevna

Head of department of Postgraduate and Doctoral Studies,
associate professor of Information Technology department, Tver State University
Russia, 170100, Tver, 33 Zhelyabova str., TSU. E-mail: mail.grishina@gmail.com

Ryzhova Margarita Nikolaevna

Master student of Information Technology department, Tver State University
Russia, 170100, Tver, 33 Zhelyabova str., TSU. E-mail: mnryzhova@tversu.ru

Received 14.08.2014, revised 08.09.2014.

This paper develops an approach to interaction identification of initial system parameters based on association measure calculation. Calculation of association measure and Pearson correlation coefficient are conducted. Obtained results are applied to the problem of dimension reduction. Comparative analysis of obtained results is presented.

Keywords: data mining, similarity measure, moving approximation transform, time series, rail transport.

Bibliographic citation

Grishina E.N., Ryzhova M.N. On application of time series data mining in a problem of dimension reduction of characteristics based on dependencies revelation. *Vestnik TvGU. Seriya: Prikladnaya matematika* [Herald of Tver State University. Series: Applied Mathematics], 2014, no. 3, pp. 95–104. (in Russian)