УДК 004.056

Прогнозирование многомерных временны́х рядов

А.С. Лысяк, Б.Я. Рябко

В данной работе предложен новый подход к решению задачи прогнозирования временны́х рядов. Данный подход основан на принципе многомерного прогнозирования коррелирующих между собой временны́х рядов и позволяет учесть взаимное влияние и существующие взаимосвязи между двумя и более рядами. Показано теоретическое и практическое обоснование предложенного подхода, а также приведены экспериментальные результаты прогнозирования непрерывных экономических временны́х рядов, таких как цены на топливо США, внутренний валовый продукт США, индексы промышленных и потребительских цен, курсы Bitcoin. Также проведено исследование эффективности данных методов и способы выбора эффективных параметров работы для выбранных методов.

Ключевые слова: прогнозирование, многомерное прогнозирование, временны́е ряды, R-метод, экономические ряды.

1. Введение

Задача прогнозирования временны́х рядов представляет в наше время большой практический интерес и позволяет решать широкий класс проблем в науке, технике и экономике. В частности, одним из наиболее значимых приложений задачи прогнозирования является предсказание и анализ социальных, экономических, геофизических процессов, предсказание природных явлений и экономических событий. Кроме того, различные методы прогнозирования служат для исследования и выявления закономерностей и системных связей в процессе функционирования тех или иных процессов.

В последние два десятилетия появилось множество методов прогнозирования, показавших свою достаточно высокую эффективность. В частности, к таковым относятся модели машинного обучения [1], которые стали представлять собой серьёзную конкуренцию классическим статистическим моделям и методам прогнозирования [2, 3, 4].

В 1988 году была показана теоретическая взаимосвязь между сжатием данных и прогнозированием [5]. Для этого была применена теория универсальной меры и универсального кодирования. Данный метод позволял решать задачу прогнозирования для случая работы с дискретными источниками (порождающими элементы из конечного алфавита). Позже был разработан способ переноса данного метода на случай прогнозирования источников, порождающих элементы из непрерывного интервала, а также получено множество экспериментальных результатов, показавших высокую точность получаемых прогнозов [6].

В данной работе предложены новые подходы к прогнозированию временны́х рядов с использованием многомерного подхода, основанные на определённых моделях теории информации и сжатии данных. В качестве базовых методов решения задачи прогнозирования используются методы, построенные на основе теории сжатия информации. В частности, используются универсальные коды и универсальная мера. Для увеличения скорости работы предлагаемых методов в данной работе предложены способы оптимизации описываемых подходов с точки зрения их трудоёмкости.

2. Постановка задачи прогнозирования

В общем виде задача прогнозирования временны́х рядов может быть сформулирована следующим образом. Пусть имеется некоторый источник, порождающий последовательность элементов x_1x_2 ... из некоторого множества A, называемого алфавитом. Алфавит может быть как конечным, так и бесконечным (т.е. представлять собой некоторый ограниченный непрерывный интервал). Пусть при этом на момент времени t мы имеем конечную порождённую источником последовательность $x_1x_2 \dots, x_t$. Рассмотрим случайную величину x_{t+1}. Задача прогнозирования сводится к определению распределения вероятностей значений случайной величины x_{t+1}, т.е. к вычислению $p(x_{t+1} = a|x_1 \dots x_t)$ для любого допустимого a (т.е. из алфавита A). В случае если алфавит A является дискретным и конечным, задача сводится к перебору всех таких $a \in A$ и определению для них условных вероятностей.

В случае если алфавит A представляет собой непрерывный конечный интервал, поступим следующим образом. Разобьём заданный интервал на фиксированное количество непересекающихся подмножеств (в общем случае подмножества могут быть произвольного неравного размера), сопоставим им целочисленные номера в соответствии с их порядком в исходном интервале. Количество возможных номеров будет совпадать с числом интервалов. Множество всех номеров при этом будет представлять собой новый, уже конечный, дискретный алфавит A'. Далее преобразуем исходный временно́й ряд из терминов в алфавите A в ряд, записанный в терминах нашего нового алфавита A'. Таким образом получим некоторую конечную выборку (ряд) из уже конечного алфавита и будем работать с ним, как со случаем конечного дискретного алфавита. При этом после вычисления распределения вероятности для всех значений ряда естественным путём определяется плотность вероятности. В качестве прогнозного значения можно брать как весь непрерывный интервал, так и какую-либо точку, вычисленную на основе посчитанной плотности.

Количество букв алфавита обозначим через *N*. Предполагается, что процесс, или источник информации, является стационарным и эргодическим, т. е., говоря неформально, распределение вероятностей символов этого источника не изменяется со временем и не зависит от конкретной реализации процесса. Данное предположение связано с тем, что в работе [7] математически доказано, что метод на основе универсальной меры выявляет закономерности именно для таких видов рядов. Метод на основе решающих деревьев сходен по видам выявляемых закономерностей и некоторым принципам действия с методом на основе универсальной меры. Как ведут себя данные методы для других видов рядов, неясно. Одновременно многие реальные временны́е ряды могут и не являться стационарными и эргодическими, однако в задачи данной работы входит экспериментальное исследование применимости предложенных и описанных методов к реальным временны́м рядам.

Пусть источник порождает сообщение $x_1 \dots x_{t-1}x_t \mid x_i \in A,\ i = 1,2, \dots, t,$ и требуется спрогнозировать один следующий элемент. Тогда ошибкой прогноза E называется величина отклонения прогноза $x_{t+1}{}^*$ от истинного значения процесса x_{t+1} в рассматриваемый $(t+1)$-ый момент времени, т.е. следующая величина

$$E = |x_{t+1} - x_{t+1}{}^*|,$$

где $x_{t+1}{}^*$ – прогнозное значение, а x_{t+1} – истинное значение процесса. Под ошибкой прогноза при прогнозировании на *n* шагов вперёд будем понимать следующую величину:

$$E_n = |x_{t+n} - x_{t+n}{}^*|,$$

где *n* – число шагов, на которые мы осуществляем прогноз. Ошибка прогноза характеризует качество прогнозирования и является основным критерием определения эффективности работы выбранного метода прогнозирования.

Очевидно, что если распределение вероятностей исходов процесса известно заранее, то задача прогнозирования следующих элементов решается достаточно просто (в соответ-

ствии с известной закономерностью строится прогнозная функция либо просто выбираются значения исходя из условия удовлетворения известной плотности распределения вероятностей процесса). Однако в большинстве практических задач описанные априорные данные отсутствуют, да и не всегда заданное распределение явно существует. В этой работе мы будем рассматривать именно такой случай. В данной ситуации для решения задачи прогнозирования можно воспользоваться точными оценками указанных величин, полученными с помощью статистических методов, построенных на основе анализа взаимосвязи последовательных исходов процесса и выявления закономерностей.

В более общей постановке задачи прогнозирования элементы x_i могут быть не только конкретными числами (целыми или вещественными), а векторами размерности k, где первый элемент вектора – значение прогнозируемой характеристики ряда, а оставшиеся $(k-1)$ атрибутов – какие-либо характеристики процесса или величины, коррелирующие со значениями ряда и известные для всех элементов ряда. Приведём пример. Пусть имеется ряд значений ВВП какой-либо страны с интервалом между значениями в 1 месяц. Требуется спрогнозировать очередное значение данного ряда. Как известно, на ВВП влияют такие параметры, как уровень инфляции, индекс потребительских цен, объёмы промышленного производства, дефицит платёжного баланса и многие другие экономические показатели страны, ВВП которой мы рассматриваем. Допустим, нам известны значения всех таких характеристик на каждый месяц прогнозируемого ряда ВВП. Таким образом, мы можем естественным образом определить некоторый временно́й ряд векторов и прогнозировать уже не одно целое или вещественное значение ряда, а целый вектор. При этом интересовать нас будет лишь один – первый – элемент прогнозного вектора. Ряды такого типа и будем называть многомерными.

В итоге, задача прогнозирования может быть как классической – одномерной, так и многомерной.

3. Схема прогнозирования на базе универсальной меры

3.1. Схемы прогнозирования для источников из конечного и непрерывного алфавитов

Для решения задачи прогнозирования будем использовать метод на основе универсальных кодов и универсальной меры. Приведём определения данных понятий. По определению, мера μ называется универсальной, если для любого стационарного и эргодического источника P верны следующие равенства:

$$\lim_{t\to\infty}\frac{1}{t}(-\log_2 P(x_1 \dots x_t) - \log_2 \mu(x_1 \dots x_t)) = 0$$

с вероятностью 1 и

$$\lim_{t\to\infty}\frac{1}{t}\sum_{u\in A^t} P(u)\log_2 (P(u)/\mu(u)) = 0.$$

Данные равенства показывают, что, в некотором смысле, мера μ является непараметрической оценкой для (неизвестной) меры P. По этой причине универсальные меры могут быть использованы для оценки статистических характеристик процесса и прогнозирования.

Универсальная мера тесно связана с понятием универсального кода. Опишем взаимосвязь этих понятий. Код U называется универсальным, если для любого стационарного и эргодического источника P верны следующие равенства:

$$\lim_{t\to\infty} |U(x_1 \dots x_t)|/t = H(P),$$

с вероятностью 1 и

$$\lim_{t\to\infty} E_P(|U(x_1 \dots x_t)|)/t = H(P),$$

где $E_P(f)$ – среднее значение f по отношению к P, а $H(P)$ – энтропия P по Шеннону, т.е.

$$H(P) = \lim_{t\to\infty} -t^{-1} \sum_{u\in A^t} P(u)\log P(u).$$

Универсальные меры имеют глубокую взаимосвязь с универсальными кодами, и если есть универсальный код, то можно легко получить на его основе универсальную меру.

Следующая теорема, приведённая с доказательством в [8], говорит о том, что на базе любого универсального кода можно построить универсальную меру.

Теорема 1. Пусть U – универсальный код и

$$\mu_U(\omega) = 2^{-|U(\omega)|} \Big/ \sum_{u\in A^{|\omega|}} 2^{-|U(u)|},$$

тогда μ – это универсальная мера.

Теперь опишем универсальную меру R, которая была использована в качестве основы для метода прогнозирования в данной работе. Выбор именно этой меры связан с тем, что она построена на основе асимптотически оптимального универсального кода R, что доказано в [9].

В общем случае в качестве универсальной меры была взята мера Кричевского $K_m \geq 0$, которая является универсальной для множества марковских источников с памятью, или связностью, m, $m \geq 0$; если $m = 0$. то это источник независимых и одинаково распределённых символов. В некотором смысле эта мера является оптимальной для этого множества [7]. По определению,

$$K_m(x_1 \dots x_t) = \begin{cases} \dfrac{1}{|A|^t}, & t \leq m, \\ \dfrac{1}{|A|^m} \displaystyle\prod_{\vartheta\in A^m} \frac{\prod_{a\in A}(\Gamma(\nu_x(\vartheta a)+1/2)/\Gamma(1/2)))}{(\Gamma(\overline{\nu}_x(\vartheta)+|A|/2)/\Gamma(|A|/2))}, & t > m; \end{cases} \quad (1)$$

где $\nu_x(\vartheta)$ – число последовательностей ϑ, встречающихся в x, $\overline{\nu}_x(\vartheta) = \sum_{a\in A}\nu_x(\vartheta a)$,

$x = x_1 \dots x_t$, а $\Gamma()$ – это гамма-функция.

Определим также распределение вероятностей $\{\omega = \omega_1, \omega_2, \dots\}$ для целых {1,2,…} как

$$\omega_i = 1/\log(i+1) - 1/\log(i+2). \quad (2)$$

В дальнейшем будем использовать именно это распределение.

Мера R определяется как

$$R(x_1 \dots x_t) = \sum_{i=0}^{\infty} \omega_{i+1} K_i(x_1 \dots x_t). \quad (3)$$

Слагаемые ω_i играют в данном случае роль весовых коэффициентов, которые должны удовлетворять следующему условию:

$$\sum_{i=1}^{\infty} \omega_i = 1.$$

Одновременно с этим понятно, что слишком большие порядки в мере Кричевского должны иметь меньший вес и меньше влиять на прогноз. В результате, в качестве весовых коэффициентов было выбрано распределение (2). В общем случае весовые коэффициенты представляют собой варьируемый параметр метода и могут меняться в зависимости от ряда и метода.

Исходя из определения весовых коэффициентов (2) видно, что значение коэффициентов w_i стремится к нулю с ростом i (при $i > 5$ $w_{i+1} < 0.05$). Соответственно, вклад слагаемого $\omega_{i+1}K_i(x_1 \dots x_t)$ с увеличением i будет всё более незначительным. В целях уменьшения тру-

доёмкости вычислений меры R количество слагаемых в сумме (3) можно ограничить каким-либо параметром m, где $m = 1, \dots, t$. Назовём этот параметр глубиной вычислений метода.

Итак, значение меры R, вычисленное на основе формулы (3), может служить оценкой вероятности исхода процесса и использоваться для решения задачи прогнозирования.

Рассмотрим схему прогнозирования временны́х рядов, порождающих значения из конечного алфавита. В данном случае схема вычисления меры R достаточна проста. Пусть $x_1 \dots x_t$ – имеющаяся временна́я последовательность. Для каждого $a \in A$ построим последовательность $x_1 \dots x_t a$ и вычислим условную вероятность на основе меры R:

$$R(a|x_1 \dots x_t) = R(x_1 \dots x_t a)/R(x_1 \dots x_t).$$

Полученные таким образом для каждого $a \in A$ величины можно использовать в качестве оценок соответствующих неизвестных условных вероятностей $P(x_{t+1} = a|x_1 \dots x_t)$.

Рассмотрим теперь схему прогнозирования для источника из непрерывного интервала. Пусть имеется стохастический процесс, генерирующий последовательность X_t, каждый элемент которой принимает значения из стандартного борелевого пространства Ω, представляющего в нашем случае непрерывный интервал [A, B]. И пусть также $\{\Pi_n\}, n \geq 1$ – возрастающая последовательность конечных разбиений интервала [A, B] на n частей (назовём этот процесс квантизацией). В нашем случае разбиение интервалов производилось равномерно на равные подынтервалы, т.е. размер каждого подынтервала определяется как $h = (B - A)/n$. Обоснование выбора именно такого метода будет дано далее. Определим также $x^{[k]}$ как элемент Π_k, содержащий точку x.

Определим теперь оценку плотности вероятностей r следующим образом:

$$r(x_1 \dots x_t) = \sum_{s=1}^{\infty} \omega_s R\left(x_1{}^{[s]} \dots x_t{}^{[s]}\right). \quad (4)$$

Как показано в [8, 5], плотность $r(x_1 \dots x_t)$ является оценкой неизвестной плотности $p(x_1 \dots x_t)$, а соответствующая условная плотность

$$r(a|x_1 \dots x_t) = r(x_1 \dots x_t a)/r(x_1 \dots x_t) \quad (5)$$

является подходящей оценкой плотности вероятности $p(a|x_1 \dots x_t)$.

3.2. Схема вычислений прогнозных значений

Выше мы рассмотрели схему, на выходе которой мы получаем распределение вероятностей рассматриваемого процесса. Рассмотрим теперь проблему выбора прогнозного значения из найденного распределения. Пусть имеется ряд $x_1 x_2 \dots, x_t$ и какая-то оценка распределения вероятностей p для элемента x_{t+1}. Рассмотрим методы выбора численного прогнозного значения. В первом случае можем выбрать в качестве прогнозного значения середину интервала из разбиения, которому соответствует максимальную вероятность в полученном распределении p. Однако в данном случае не учитывается всё распределение p и соответственно, – бóльшая часть информации касательно поведения ряда. Пример. Если два соседних подынтервала имеют очень схожие вероятности, то стоило бы выбрать точку между ними, а не середину того из них, у кого вероятность больше. Для решения указанной проблемы предлагается следующий подход. Выберем в качестве прогнозного значения не середину подынтервала с максимальной вероятностью, а математическое среднее от середин всех подынтервалов. Таким образом, вычисление прогнозного значения будет сводиться к следующему соотношению:

$$x_{t+1} = \sum_{i=1}^{n} p_i \cdot k_i,$$

где p_i – вероятность i-го подынтервала, n – величина разбиения (число подынтервалов), k_i – середина i-го подынтервала. Как будет показано в экспериментальных результатах, данный подход повышает точность прогноза для случая небольших разбиений.

4. Подход на основе многомерного прогнозирования

Достаточно очевидным является факт взаимосвязи различных реальных процессов, происходящих в мире. К примеру, как ранее было показано, внутренний валовый продукт оказывает влияние на курс валюты рассматриваемой страны, а показатель уровня жизни – на индекс потребительских цен. Все эти показатели и процессы зачастую представляют собой отдельные временны́е ряды, которые нам также известны. Если бы можно было учесть корреляции хотя бы некоторого ограниченного набора временны́х рядов, то мы смогли бы существенно повысить точность и эффективность получаемых прогнозов. Пример наличия простой взаимосвязи между рядами: при увеличении значений одного временно́го ряда всегда происходит увеличение значений другого временно́го ряда. Конечно, такие влияния могут быть «запоздалыми» или наоборот «спешащими», но существующая корреляция между разными временны́ми рядами позволяет получить дополнительную информацию о том временно́м ряде, который мы хотим спрогнозировать.

Таким образом, качество прогнозирования временны́х рядов может быть существенно увеличено с использованием так называемого многомерного подхода, при котором в прогнозе учитываются другие временны́е ряды. Ранее методов, учитывающих сразу несколько различных коррелирующих временны́х рядов, не было.

В данной работе предлагается подход, который позволяет учесть при прогнозировании одного временно́го ряда другой временно́й ряд, который коррелирует с первым. Важно отметить, что данный подход не зависит от используемого метода (алгоритма) прогнозирования. В качестве основы мы можем использовать любой математический метод прогнозирования стационарных и эргодических источников.

Пусть имеется K временны́х рядов, коррелирующих каким-то образом между собой:

$$x_1^1, x_2^1, x_3^1, \dots, x_n^1;$$
$$x_1^2, x_2^2, x_3^2, \dots, x_n^2;$$
$$\dots$$
$$x_1^K, x_2^K, x_3^K, \dots, x_n^K.$$

При этом мы предполагаем, что все K временны́х рядов определены на одной и той же оси времени с едиными начальными и конечными точками. Также у них одинаковая квантизация (разбиение, т.е. одинаковый алфавит). Нам требуется спрогнозировать следующий элемент первого ряда, т.е. элемент x_{n+1}^1. Построим временно́й ряд $(K+1)$ на основе первых K по правилу:

$$x_i = x_i^1 + x_i^2 \cdot N + x_i^3 \cdot N^2 + \dots + x_i^K \cdot N^{K-1}, \quad (6)$$

где N – мощность алфавита, определённая в главе 1. Видно, что формула (6) является полиномиальным хешем от рассматриваемых K временны́х рядов. Это позволяет однозначно восстановить по полученному ряду все K временны́х рядов. Таким образом, мы получим уже один временно́й ряд, в котором содержится вся информация из всех K рядов. Далее применим какой-либо метод прогнозирования к последнему временно́му ряду и получим прогнозное значение x_{n+1}. Тогда элемент $x_{n+1}^1 = x_{n+1} \bmod N$.

В простейшем случае можно соединить всего два коррелирующих между собой временны́х ряда. Важно отметить, что существуют и другие способы соединения (слияния) временны́х рядов в один ряд. Например, можно соединять 2 временны́х ряда по принципу чередования значений одного и другого. Однако, как показали приведённые ниже экспериментальные

результаты, описанная выше методика соединения рядов является наиболее эффективной с точки зрения точности получаемых прогнозов.

Таким образом, для увеличения точности прогноза нам надо найти такие временны́е ряды, которые имеют ненулевую корреляцию между собой. Но этого мало. Известно, что некоторые процессы, не являющиеся абсолютно случайными, могут влиять на другие ряды с некоторым запозданием или, наоборот, опережением. Для нас важно найти такой временно́й ряд, который бы влиял на первый с опережением, т.е. такой, у которого факты увеличения / уменьшения значений, периоды, какие-либо ещё закономерности происходят раньше, чем у другого временно́го ряда. Только при соблюдении описанных условий мы сможем получить существенное увеличение эффективности работы выбранного метода прогнозирования.

Важно также отметить, что в данную схему можно ввести параметр сдвига временны́х рядов относительного первого. Данный параметр будет определяться в зависимости от предполагаемого среднего уровня опережения рассматриваемого временно́го ряда относительно первого временно́го ряда.

5. Экспериментальные результаты многомерного прогнозирования методом R

5.1. Методика экспериментальных исследований

Изложенный выше R-метод с модификацией группировки алфавита, описанной ранее в [6], был реализован на суперкомпьютере с учётом подхода на основе многомерного прогнозирования. Полученный метод был применён к некоторым базовым экономическим временны́м рядам США. В частности, было проведено прогнозирование уровня безработицы, ВВП, цен на топливо, курсов Bitcoin.

Все проведённые исследования проводилась в двух режимах. Первый режим – on-line – предполагает прогнозирование значений временно́го ряда на 1 шаг вперёд. Второй режим – на несколько шагов вперёд – прогнозирование значений ряда на более чем 1 шаг вперёд. Без ограничения общности, для второго режима было взято 10 шагов. При этом прогнозирование на 10 шагов считалось следующим образом. Прогнозируем очередной элемент ряда, после чего пополняем выборку прогнозным значением, далее ведём прогнозирование ещё на 1 шаг, но с использованием уже пополненного ряда и так далее продолжаем до 10-го прогнозного элемента, после чего считаем ошибку прогноза на последнем элементе. Все прогнозы выполнялись на 10 выборках одного ряда с различным сдвигом по временно́й оси, в итоговую таблице вносилась усреднённая ошибка. Выбор именно этого режима связан с целью получить оптимум между точностью прогнозирования на несколько шагов и трудоёмкостью метода. Описанный подход получен в ходе экспериментальных исследований, где он показал себя как оптимальный вариант между высокой точностью при разумной трудоёмкости. Существенное влияние в снижении трудоёмкости оказал и метод группировки алфавита, применённый здесь в качестве дополнительной модификации, и который позволяет в среднем на 1 порядок уменьшить сложность работы алгоритма R-метода при сохранении трудоёмкости, что было экспериментально показано в [6].

Важно заметить, что выполнялось прогнозирование не абсолютных значений ряда, а разницы между соседними элементами с последующей прибавкой спрогнозированной разницы к последнему элементу ряда (в результате получим значение следующего за последним элемента ряда): такой подход позволяет существенно снизить необходимый размер непрерывного интервала, в котором лежат прогнозные значения; а также позволяет выявлять линейные и квазилинейные тренды и периоды на них. Это было невозможно при прогнозировании абсолютных величин временно́го ряда. Определение границ интервала будем производить естественным образом: считаем величину максимального и минимального (с учётом знака) отклонения между соседними элементами ряда; полученные максимум и минимум берём в качестве левой и правой границы интервала, который далее и разбиваем

на количество частей, равное мощности алфавита (параметру разбиения n). При этом величину полученного интервала (максимальную разницу между соседними элементами) обозначим Δ. Данное значение будет определять максимально возможную теоретическую ошибку. Нетрудно понять, что при случайном выборе прогнозного значения и при прогнозировании на 1 шаг вперёд средняя ошибка будет стремиться к величине $\Delta/2$. Таким образом, зная Δ, мы можем оценить качество работы заданного метода прогнозирования.

5.2. Экспериментальные результаты

Рассмотрим прогнозирование индексов потребительских и промышленных цен США в период с 09.1983 по 03.2013. Их графики приведены на рис. 1 и 2, соответственно.

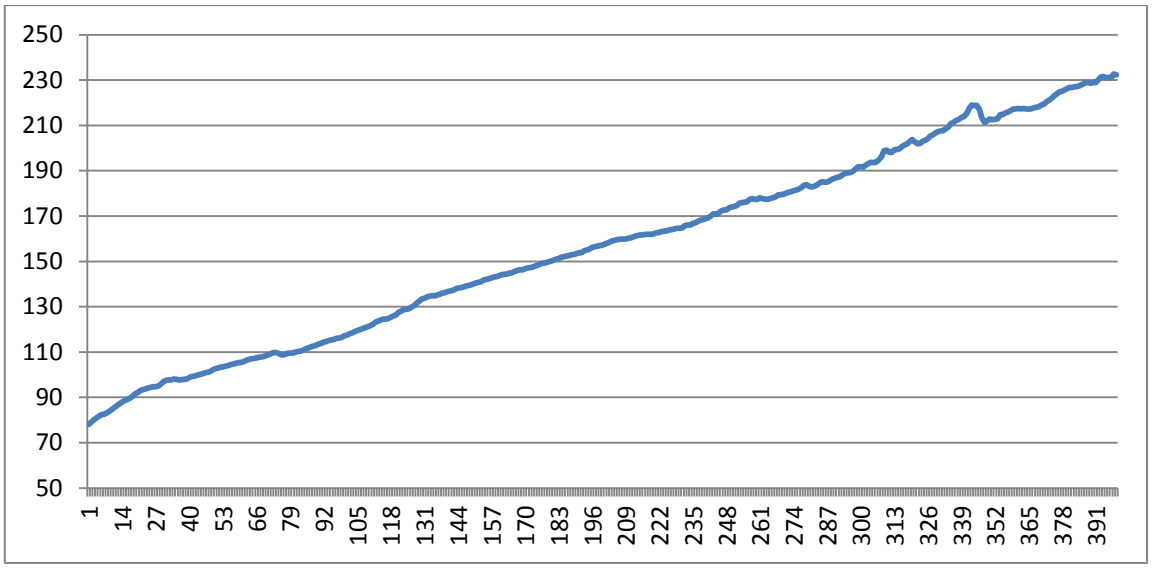

Рис. 1. Индекс потребительских цен США

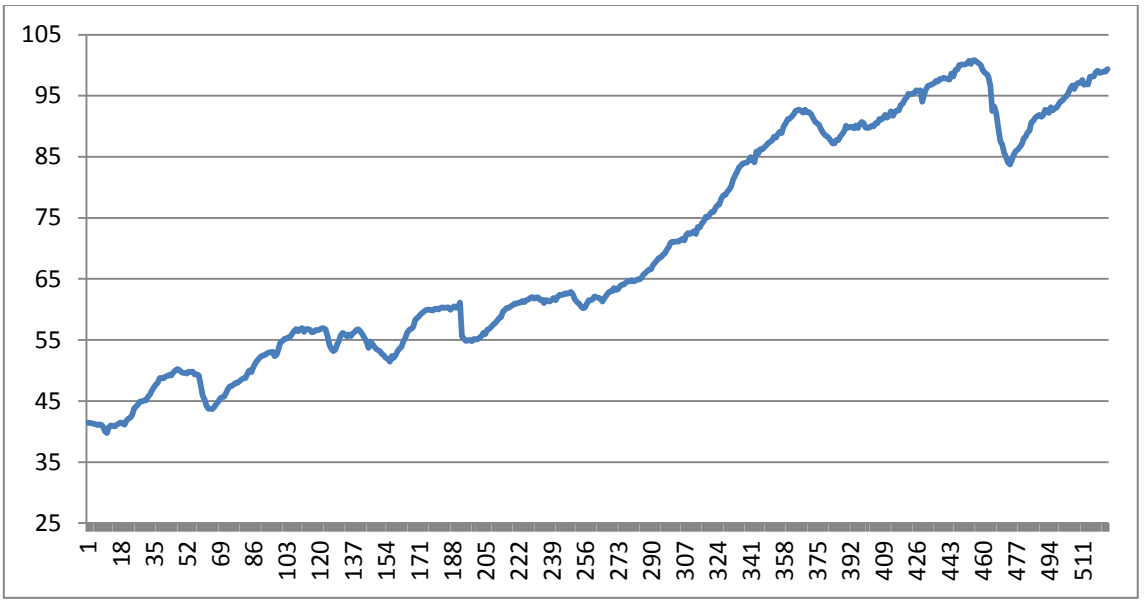

Рис. 2. Индекс промышленных цен США

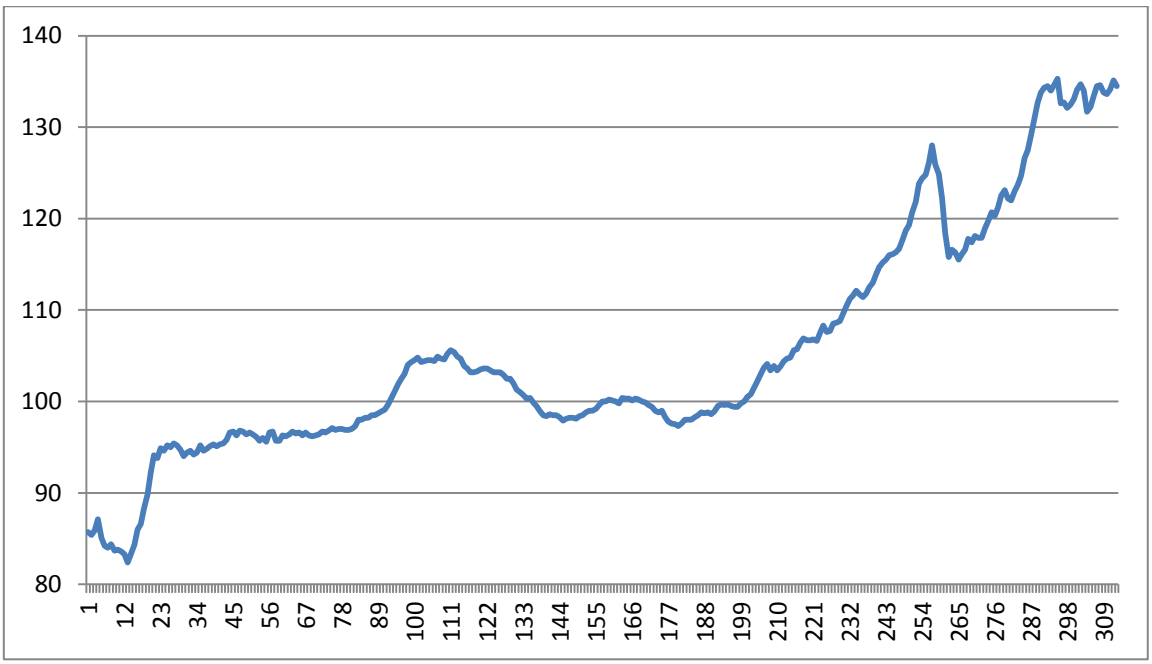

Рис. 3. Уровень экспорта США

Размеры обоих рядов составляют 360 элементов, период между измерениями равен 1 месяцу. Для осуществления многомерного прогнозирования к заданным двум рядам присоединялись другие экономические ряды с теми же временны́ми характеристиками (период; период между измерениями; длина). В частности, мы использовали дополнительно следующие ряды: уровень экспорта США (его график приведён на рис. 3), курсы валют американский доллар / британский фунт стерлингов и американский доллар / канадский доллар. Данные ряды были выбраны из соображений теоретической взаимосвязанности друг с другом и вполне могут иметь выраженные корреляции. Результаты прогнозирования индексов потребительских (CPI) и промышленных (PPI) цен приведены в табл. 1 и 2, соответственно. При этом в первой строке идут результаты одномерного прогнозирования временно́го ряда (без присоединения к нему других рядов), а далее идёт двумерное прогнозирование с обозначением вида A + B: значит, что прогнозируются значения ряда A с присоединением к нему ряда B. Глубина вычислений $m = 3$.

Таблица 1. Прогнозирование индекса потребительских цен США

№	Временно́й ряд	R-метод on-line	R-метод 10 шагов
1	CPI	0.3922	0.4670
2	CPI + PPI	0.4308	0.4537
3	CPI + USDGBP	0.4533	0.5703
4	CPI + USDCAD	0.4533	0.5703
5	CPI + Export	0.7468	1.4239

Таблица 2. Прогнозирование индекса промышленных цен США

№	Временно́й ряд	R-метод on-line	R-метод 10 шагов
1	PPI	1.3450	3.2575
2	PPI + CPI	1.0650	2.6825
3	PPI + USDGBP	1.2107	1.3571
4	PPI + USDCAD	1.2107	1.3571
5	PPI + Export	1.1393	2.9750

По графикам видно, что взятые дополнительные ряды не слишком сильно коррелируют друг с другом. В результате, в части прогнозирования индекса CPI получились результаты, в среднем хуже или сравнимые с одномерным случаем. В случае же прогнозирования индекса PPI результаты получились лучше, чем для одномерного ряда PPI, что говорит о нахождении предложенным подходом определённых корреляций.

Рассмотрим прогнозирование уровня безработицы в США (Unemployment). Его график представлен на рис. 4. Для этого в качестве дополнительных рядов возьмём явно коррелирующий с уровнем безработицы ряд количество обращений по безработице (claims of unemployment), график которого показан на рис. 5. Внешне хорошо видны корреляции и закономерности, однако ряды всё же отличаются и абсолютные значения и дисперсия (волатильность) у них совершенно различны. Для большей объективности возьмём также следующие ряды: уровень ВВП США (GDP; рис. 6), индекс промышленных цен (CPI; рис. 1) и индекс промышленного производства в США (IPI; рис. 7). Периоды данных рядов составляют с 01.1970 по 08.2012. Размер рядов: 512 элементов. Глубина вычислений $m = 3$. Величина Δ для ряда, представляющего уровень безработицы, равна 1.65.

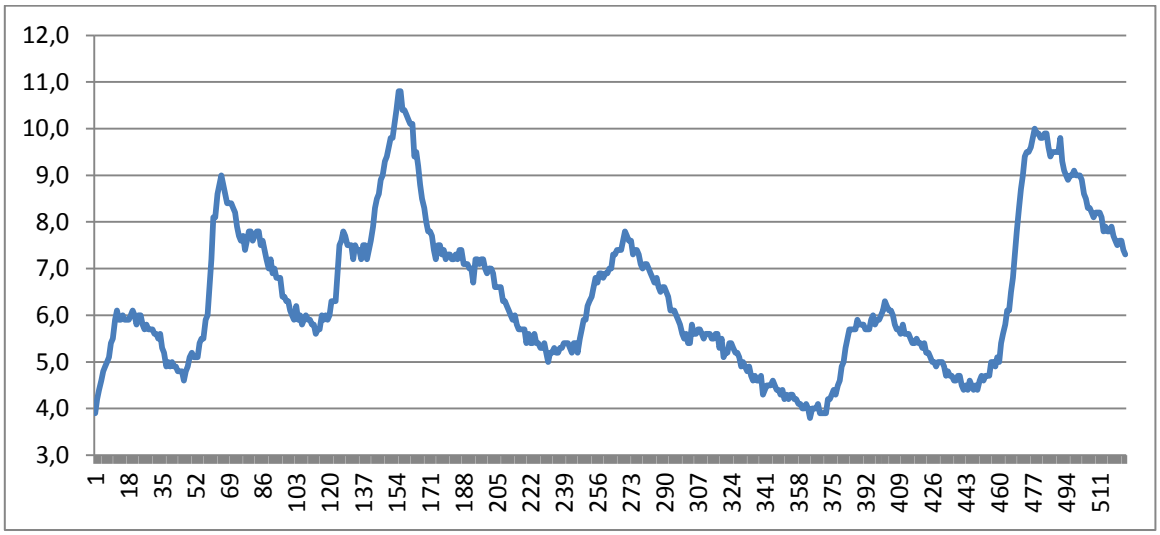

Рис. 4. Уровень безработицы в США

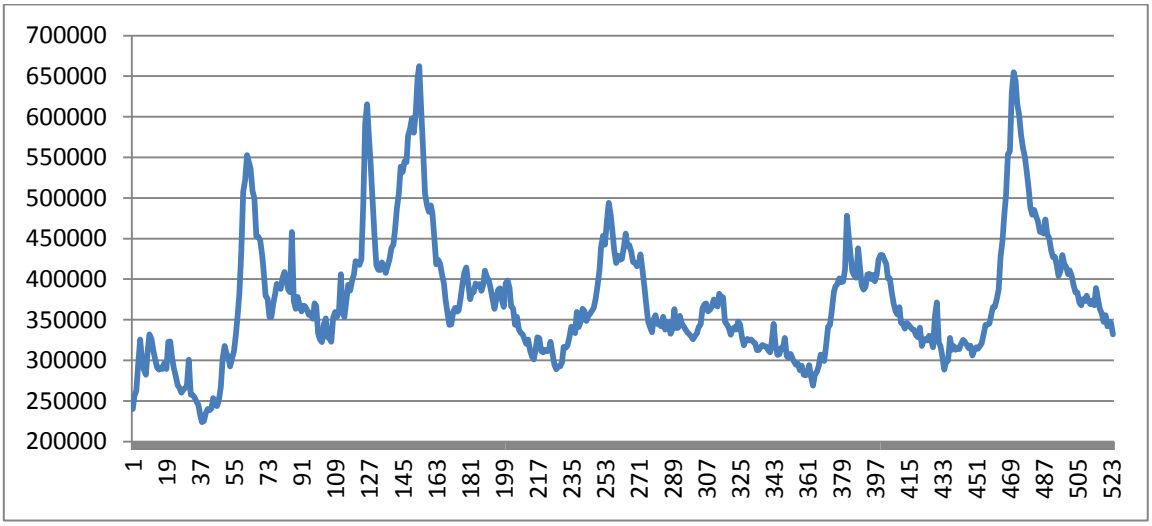

Рис. 5. Обращения по безработице в США

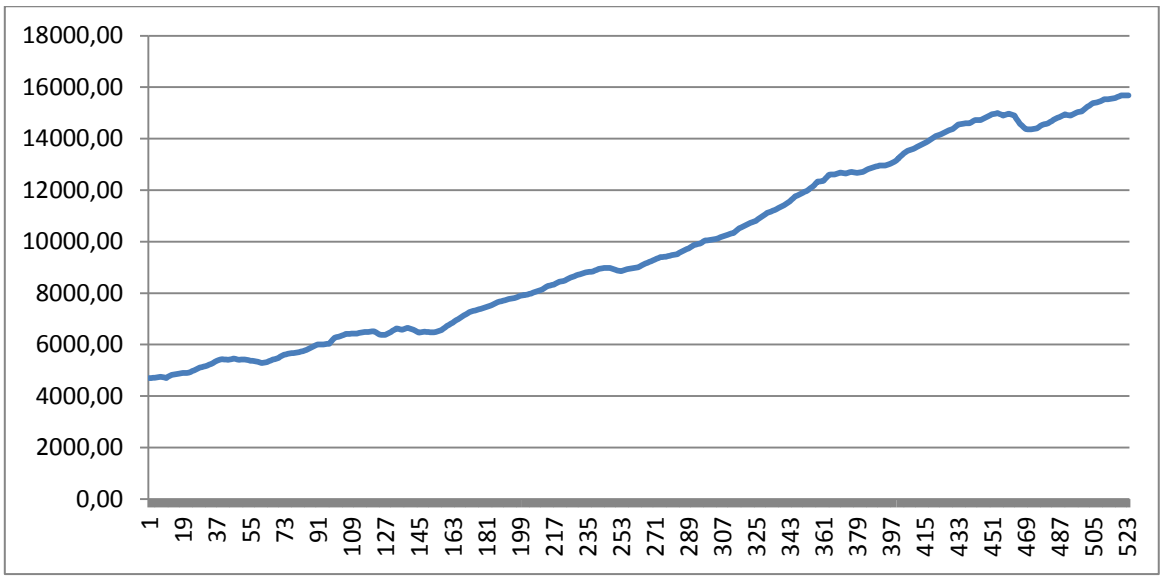

Рис. 6. Внутренний валовый продукт США

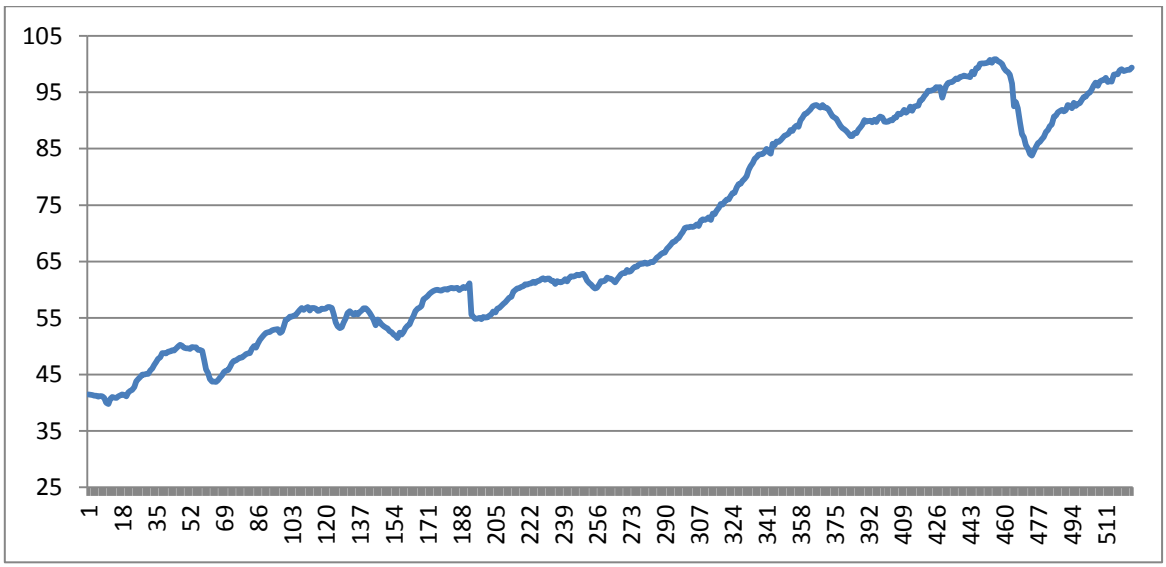

Рис. 7. Индекс промышленного производства США

Результаты прогнозирования ряда уровня безработицы в США приведены в табл. 3.

Таблица 3. Прогнозирование уровня безработицы в США

№	Временно́й ряд	R-метод on-line	R-метод 10 шагов
1	Unemployment	0.106	0.425
2	Unemployment + Claims of unemployment	0.098	0.136
3	Unemployment + IPI	0.106	0.460
4	Unemployment + GDP	0.116	0.370
5	Unemployment + CPI	0.134	0.7155

По представленным в табл. 3 данным хорошо видно, что добавление ряда обращений по безработице существенно увеличивает точность получаемых прогнозов, в особенности для случая прогнозирования на несколько шагов вперёд. Три других ряда прироста точности

не дают, оставляя ошибку прогноза на уровне одномерного подхода, что вполне закономерно ввиду отсутствия явных корреляций между рядами.

Рассмотрим прогнозирование уровня ВВП США с использованием в качестве дополнительных рядов уровень безработицы США, индекс промышленного производства США, а также индексы CPI и PPI. Их временны́е параметры аналогичны рассмотренным выше. Исходя из графиков данных временны́х рядов можно увидеть, что ВВП сильно коррелирует с индексом потребительских цен США, а также – немного с индексом промышленного производства США. Исходя из этого мы должны наблюдать прирост точности метода при соединении ряда GDP с двумя вышеназванными. Результаты прогнозирования приведены в табл. 4.

Таблица 4. Прогнозирование уровня ВВП США

№	Временно́й ряд	R-метод on-line	R- метод 10 шагов
1	GDP	9.363	79.132
2	GDP + Unemployment	16.988	50.475
3	GDP + IPI	9.209	78.133
4	GDP + Claims of unemployment	13.306	69.878
5	GDP + CPI	8.207	57.702
6	GDP + PPI	15.146	88.720

Как видно из вышеприведённых результатов, заметное увеличение точности по сравнению с одномерным подходом наблюдается в случае добавления рядов индексов промышленного производства и потребительских цен, что вполне закономерно.

6. Заключение

Предложенный метод многомерного прогнозирования, как показывают практические результаты, оказался достаточно эффективным и в случае подбора коррелирующих между собой рядов существенно уменьшает ошибку прогноза. Одновременно, в случае отсутствия каких-либо корреляций между соединяемыми рядами, данный метод показывает точность, сравнимую с классическим одномерным подходом.

Кроме того, описанный подход позволяет учитывать в прогнозировании не только 2 или более дополнительных ряда, но и какие-либо дополнительные атрибуты или свойства рассматриваемого процесса. Важно отметить, что рассмотренный метод может использоваться, как модификация по отношению к любым вероятностным методам прогнозирования (т.е. тем, которые дают на выходе распределение вероятностей).

Литература

1. *Bontempi G.* Local Learning Techniques for Modeling, Prediction and Control. Ph.d., IRIDIA-Universit de Libre de Bruxelles, BELGIUM, 1999.
2. *Ahmed N.* An empirical comparison of machine learning models for time series forecasting // Econometric Reviews. 2010,Vol. 29, Issue 5-6. P. 594-621.
3. *Palit A. K., Popovic D.* Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications (Advances in Industrial Control). Springer-Verlag New York: Secaucus, NJ, USA, 2005.
4. *Zhang G., Patuwo B. E., Michael Y. H.* Forecasting with articial neural networks: The state of the art // International Journal of Forecasting. 1998. Vol. 14, Issue 1. P. 35-62.
5. *Приставка П.А.* Экспериментальное исследование метода прогнозирования, основанного на универсальных кодах // Вестник СибГУТИ, 2010. №4, С. 26-35.

6. *А.С. Лысяк, Б.Я. Рябко.* Методы прогнозирования временны́х рядов с большим алфавитом на основе универсальной меры и деревьев принятия решений. Вычислительные технологии Т. 19, №2, 2014, с. 75-92.
7. *B. Ryabko*. Compression-Based Methods for Nonparametric Prediction and Estimation of Some Characteristics of Time Series. // IEEE Transactions on Information Theory, Vol. 55, № 9, 2009. P. 4309-4315.
8. *Рябко Б.Я.* Прогнозирование случайных последовательностей и универсальное кодирование. // Проблемы передачи информации. 1988. №24. С.3-14.
9. *Рябко Б. Я.* Дважды универсальное кодирование // Проблемы передачи информации. 1984. Т. 20, № 3. С. 24-28.
10. *Cheng H. et al.* Multistep-ahead time series prediction // Lecture Notes in Computer Science. 2006. V. 3918. P. 765-774.
11. *Рябко Б., Монарёв В.* Экспериментальное исследование методов прогнозирования, основанных на алгоритмах сжатия данных // Проблемы передачи информации. 2005. С. 65-69.
12. *Nevill-Manning C.G., Witten I.H., Paynter G.W.* Lexically-generated subject hierarchies for browsing large collections // International Journal of Digital Libraries. 1999. Vol. 2, Issue 3. P. 111-123.
13. *Nevill-Manning C.G., Witten I.H.* Identifying. Hierarchical Structure in Sequences: A linear-time algorithm // Journal of Artificial Intelligence Research. 1997. Vol. 7. P. 67-82.
14. *Poskitt D.S., Tremayne A.R.* The selection and use of linear and bilinear time series models // International Journal of Forecasting. 1986. Vol. 2, Issue 1. P. 101-114.
15. *Tong H.* Non-linear Time Series: A Dynamical System Approach. Oxford University Press, 1990.
16. *Tong H*. Threshold models in Nonlinear Time Series Analysis. Springer Verlag, Berlin, 1983.
17. *Tong H., Lim K. S*. Threshold autoregression, limit cycles and cyclical data //. Journal of the Royal Statistical. Series B (Methodological). 1980. Vol. 42, Issue 3. P. 245-292.
18. *Engle R.* Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom // Econometrica. 1982. Vol. 50. Issue 4. P. 987-1007.
19. *Clements M.P. et al*. Forecasting economic and financial time-series with non-linear models // International Journal of Forecasting. 2004. Vol. 20.Issue 2. P. 169-183.
20. *Krichevsky R.* Universal Compression and Retrival. Kluver Academic Publishers, 1993.

Статья поступила в редакцию 25.06.2014;
переработанный вариант — 09.09.2014

Лысяк Александр Сергеевич

Аспирант НГУ, ассистент кафедры компьютерных систем ФИТ НГУ (630090, Новосибирск, ул. Пирогова, 2), инженер-аналитик по информационной безопасности ЗАО «Центр финансовых технологий» (630090, Новосибирск, ул. Мусы-Джалиля, 11), тел. +79130110883; e-mail: accemt@gmail.com.

Рябко Борис Яковлевич

д.т.н., ректор СибГУТИ, профессор кафедры прикладной математики и кибернетики СибГУТИ (630102, Новосибирск, ул. Кирова, 86) тел. (383) 2-698-272, e-mail: boris@ryabko.net.

Prediction of multidimensional time series

A.S. Lysayak, B.Y. Ryabko

In this paper, a new approach to the problem solving of time series prediction is presented. This approach is based on multidimensional prediction principle of correlating among themselves time series and enables to take into account mutual influence and existing correlations between two or more series. Theoretical and practical reasons of the proposed approach are presented, as well as experimental results of prediction of continuous economic time series, such as fuel prices in the USA, gross domestic product in the USA, indexes of industrial and consumer prices, Bitcoin rates. Research of the efficiency of these methods and the ways of choosing the effective parameters of operation for the selected methods are also carried out.

Keywords: prediction, multidimensional prediction, time series, R-method, economic series.