УДК 519.246.8

Н.В. Комарова

Днепропетровский национальный университет им. Олеся Гончара

МЕТОД БЛИЖАЙШИХ СОСЕДЕЙ С НОРМАЛИЗОВАННЫМ ЕВКЛИДОВЫМ РАССТОЯНИЕМ

Пропонується використання нормалізованої евклідової відстані у методі найближчих сусідів для прогнозування часових рядів, приводиться порівняль-ний аналіз застосування методу до прогнозу цін на метали та металопродукцію.

Предлагается использование нормализованного евклидового расстояния в методе ближайших соседей для прогнозирования временных рядов, приводится сравнительный анализ результатов применения метода для прогноза цен на металлы и металлопродукцию.

In the article the using of normalized Euclidian distance is considered in the nearest neighbor method for time series forecasting; the results of the method application to metal and steel product prices forecasting are analyzed in comparison with the results of other methods.

Ключевые слова: нормализованное евклидово расстояние, метод ближайших соседей, прогнозирование, временные ряды.

Вступление. Анализ и прогнозирование временных рядов – важная задача во многих областях исследований. Основными инструментами ее решения являются статистические методы, среди которых наиболее широкое распространение получили методы авторегрессии и скользящего среднего. Однако их применение для прогноза многих временных рядов, описывающих сложные системы, не дает удовлетворительных результатов.

Новый подход предложила нелинейная динамика, изучающая структуру и свойства эволюционных процессов в нелинейных динамических системах.

Постановка задачи. Нестационарные временные ряды встречаются довольно часто. Известно, что классические методы авторегрессии, скользящего среднего и индексов сезонности часто не являются адекватным инструментом их прогнозирования. Поэтому возникает задача поиска метода прогнозирования, который можно было бы использовать для широкого класса нестационарных временных рядов. Одним из методов, используемых для прогнозирования нестационарных временных рядов, является метод ближайших соседей, предложенный нелинейной динамикой. Зачас-тую в методе используется евклидово расстояние, однако результаты прогноза при этом не всегда удовлетворительны, в частности, когда ряд имеет ярко выраженный тренд. Решению этой задачи посвящена настоящая работа.

Подход к решению

Метод ближайших соседей прогнозирования временных рядов. В последнее время при решении задачи прогнозирования временных рядов широко используется метод ближайших соседей, предложенный нелинейной динамикой (нелинейная динамика изучает структуру и свойства эволюционных процессов в нелинейных динамических системах).

Пусть $\{x_i,\ i=1,2,...,N\}$ – временной ряд. По значениям x_i, $i=1,2,...,M-1,$ временного ряда необходимо построить прогноз $\tilde{x}_M$ его значения x_M в момент времени M. Для прогнозирования воспользуемся классическим методом ближайших соседей в его простейшем варианте одного ближайшего соседа нулевого порядка [1]. Строим фазовую траекторию временного ряда $\{x_i,\ i=1,2,...,N\}$ – последовательность точек с координатами $\mathbf{x}_i=(x_i, x_{i+\tau},..., x_{i+(m-1)\tau}),\ \ i=1,2,...,N-(m-1)\tau;\ (m,\ \tau<<N)$ в пространстве $\mathbb{R}^m$. Для точки $\mathbf{x}_{M-1-(m-1)\tau}=(x_{M-1-(m-1)\tau}, x_{M-1-(m-2)\tau},..., x_{M-1})$ находим точку $\mathbf{x}_{k-1-(m-1)\tau}=(x_{k-1-(m-1)\tau}, x_{k-1-(m-2)\tau},..., x_{k-1})$ фазовой траектории, ближайшую к точке $\mathbf{x}_{M-1-(m-1)\tau}$. Следующей за точкой $\mathbf{x}_{k-1-(m-1)\tau}$ фазовой траектории является точка $\mathbf{x}_{k-(m-1)\tau}=(x_{k-(m-1)\tau}, x_{k-(m-1)\tau+1},..., x_k)$. Согласно методу ближайших соседей в качестве прогноза $\tilde{x}_M$ значения x_M временного ряда в момент M принимается

значение x_k. Если необходимо построить прогноз на несколько точек $x_M, x_{M+1}, .., x_{M+L}$, то, как правило, в качестве прогнозных значений $\tilde{x}_M, \tilde{x}_{M+1}, ..., \tilde{x}_{M+L}$ рассматривают значения $x_k, x_{k+1}, .., x_{k+L}$ временного ряда.

Прогноз можно строить по двум и большему числу ближайших соседей, и в этом случае в качестве прогнозных значений рассматривается среднее или средневзвешенное значение прогнозов по каждому из ближайших соседей.

Выбор параметров *m* фазового пространства $\mathbb{R}^m$ и параметра и τ является отдельной задачей. На практике не худшим вариантом оказывается подбор этих параметров.

Нормализованное евклидово расстояние. В методе ближайших соседей в качестве меры близости между точками фазового пространства, как правило, используют евклидово расстояние. В настоящей работе в качестве меры близости предложено использовать так называемое нормализованное евклидово расстояние.

Введем обозначения:

$$\mathbf{x} = (x_1, x_2, ..., x_n), \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{\mathbf{x}} = (\bar{x}, \bar{x}, ..., \bar{x}), \quad \|x\|^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad d_E \ - \text{евклидова метрика в } \mathbb{R}^n.$$

Определим нормализованное евклидово расстояние d_{NE} между точками $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ равенством

$$d_{NE}(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{n} \left(\frac{\mathrm{x_i} - \bar{\mathrm{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|} - \frac{\mathrm{y_i} - \bar{\mathrm{y}}}{\|\mathbf{y}-\bar{\mathbf{y}}\|} \right)^2 \right)^{1/2} = d_E\left(\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|}, \frac{\mathbf{y}-\bar{\mathbf{y}}}{\|\mathbf{y}-\bar{\mathbf{y}}\|} \right). \quad (1)$$

Нормализованное евклидово расстояние, хотя и определяется через евклидову метрику, но само на $\mathbb{R}^n$ метрикой не является, уже хотя бы потому, что из $d_{NE}(\mathbf{x},\mathbf{y}) = 0$ не следует равенство $\mathbf{x} = \mathbf{y}$. Например, если $\mathbf{y} = a\mathbf{x} + \mathbf{b}$, то $d_{NE}(\mathbf{x},\mathbf{y}) = 0$. Заметим также, что d_{NE} не определено для пары точек, среди которых хотя бы одна имеет вид $\mathbf{c} = (c, c, ..., c)$. Поэтому далее, когда мы будем говорить о нормализованном евклидовом расстоянии между точками $\mathbf{x}$ и $\mathbf{y}$, всегда будем предполагать, что они не равны константам, т. е. $\mathbf{x}$ и $\mathbf{y}$ отличны от $\mathbf{c} = (c, c, ..., c)$.

Нормализованное евклидово расстояние d_{NE} обладает следующими свойствами:

1. $d_{NE}(\mathbf{x},\mathbf{y}) \geq 0$.
2. $d_{NE}(\mathbf{x},\mathbf{y}) = 0$ тогда и только тогда, когда $\mathbf{y} = a\mathbf{x} + b$.
3. $d_{NE}(\mathbf{x},\mathbf{y}) = d_{NE}(\mathbf{y},\mathbf{x})$.
4. $d_{NE}(\mathbf{x},\mathbf{z}) \leq d_{NE}(\mathbf{x},\mathbf{y}) + d_{NE}(\mathbf{y},\mathbf{z})$.

Свойство 1 очевидно.

Свойство 2. Из $\mathbf{x} = \mathbf{y}$ очевидно следует $d_{NE}(\mathbf{x},\mathbf{y}) = 0$.

Пусть $\mathbf{x}, \mathbf{y}$ таковы, что

$$d_{NE}(\mathbf{x},\mathbf{y}) = d_E\left(\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|}, \frac{\mathbf{y}-\bar{\mathbf{y}}}{\|\mathbf{y}-\bar{\mathbf{y}}\|} \right) = 0,$$

тогда $\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|} = \frac{\mathbf{y}-\bar{\mathbf{y}}}{\|\mathbf{y}-\bar{\mathbf{y}}\|}$. Из последнего равенства для $\mathbf{y}$ получаем представление

$$\mathbf{y} = \frac{\|\mathbf{y}-\bar{\mathbf{y}}\|}{\|\mathbf{x}-\bar{\mathbf{x}}\|}\cdot\mathbf{x} + \left(\bar{\mathbf{y}} - \frac{\|\mathbf{y}-\bar{\mathbf{y}}\|}{\|\mathbf{x}-\bar{\mathbf{x}}\|}\cdot\bar{\mathbf{x}}\right) = a\mathbf{x}+\mathbf{b},$$

где $a = \dfrac{\|\mathbf{y}-\bar{\mathbf{y}}\|}{\|\mathbf{x}-\bar{\mathbf{x}}\|}$, $\mathbf{b} = \bar{\mathbf{y}} - \dfrac{\|\mathbf{y}-\bar{\mathbf{y}}\|}{\|\mathbf{x}-\bar{\mathbf{x}}\|}\cdot\bar{\mathbf{x}}$.

Свойство 3 следует из определения (1) d_{NE}.

Свойство 4 следует из соотношений

$$d_{NE}(\mathbf{x},\mathbf{y}) = d_E\left(\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|}, \frac{\mathbf{y}-\bar{\mathbf{y}}}{\|\mathbf{y}-\bar{\mathbf{y}}\|}\right) \le \le d_E\left(\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|}, \frac{\mathbf{z}-\bar{\mathbf{z}}}{\|\mathbf{z}-\bar{\mathbf{z}}\|}\right) + d_E\left(\frac{\mathbf{z}-\bar{\mathbf{z}}}{\|\mathbf{z}-\bar{\mathbf{z}}\|}, \frac{\mathbf{y}-\bar{\mathbf{y}}}{\|\mathbf{y}-\bar{\mathbf{y}}\|}\right) = d_{NE}(\mathbf{x},\mathbf{z}) + d_{NE}(\mathbf{z},\mathbf{y}).$$

Метод ближайших соседей с нормализованным евклидовым рассто-янием. Для построения прогноза $\tilde{x}_M$ значения x_M временного ряда $\{x_i,\ i = 1,2,...,N\}$ сначала для точки $\mathbf{x}_{M-1-(m-1)\tau} = (x_{M-1-(m-1)\tau}, x_{M-1-(m-2)\tau},..., x_{M-1})$ находим ближайшего соседа $\mathbf{x}_{k-1-(m-1)\tau}$, принадлежащего фазовой траектории в $\mathbb{R}^m$ (ближайшего в смысле нормализованного евклидового расстояния). Ближайший сосед в смысле нормализованного евклидового расстояния существует (выбор производится из конечного множества), но, вообще говоря, не единственный.

Точки вида $\mathbf{y}_{k-1-(m-1)\tau} = a\mathbf{x}_{k-1-(m-1)\tau} + \mathbf{b}$ из пространства $\mathbb{R}^m$ имеют такое же расстояние (в смысле нормализованного евклидового расстояния) до точки $\mathbf{x}_{M-1-(m-1)\tau}$, как и точка $\mathbf{x}_{k-1-(m-1)\tau}$. Действительно,

$$d_{NE}(a\mathbf{x}_{k-1-(m-1)\tau} + \mathbf{b}, \mathbf{x}_{M-1-(m-1)\tau}) = = d_E\left(\frac{a\mathbf{x}_{k-1-(m-1)\tau} + \mathbf{b} - a\bar{\mathbf{x}}_{k-1-(m-1)\tau} - \mathbf{b}}{\left\|a\mathbf{x}_{k-1-(m-1)\tau} + \mathbf{b} - a\bar{\mathbf{x}}_{k-1-(m-1)\tau} - \mathbf{b}\right\|}, \frac{\mathbf{x}_{M-1-(m-1)\tau} - \bar{\mathbf{x}}_{M-1-(m-1)\tau}}{\left\|\mathbf{x}_{M-1-(m-1)\tau} - \bar{\mathbf{x}}_{M-1-(m-1)\tau}\right\|}\right) =$$

$$= d_E\left(\frac{a(\mathbf{x}_{k-1-(m-1)\tau} - \bar{\mathbf{x}}_{k-1-(m-1)\tau})}{a\left\|\mathbf{x}_{k-1-(m-1)\tau} - \bar{\mathbf{x}}_{k-1-(m-1)\tau}\right\|}, \frac{\mathbf{x}_{M-1-(m-1)\tau} - \bar{\mathbf{x}}_{M-1-(m-1)\tau}}{\left\|\mathbf{x}_{M-1-(m-1)\tau} - \bar{\mathbf{x}}_{M-1-(m-1)\tau}\right\|}\right) =$$

$$= d_E\left(\frac{\mathbf{x}_{k-1-(m-1)\tau} - \bar{\mathbf{x}}_{k-1-(m-1)\tau}}{\left\|\mathbf{x}_{k-1-(m-1)\tau} - \bar{\mathbf{x}}_{k-1-(m-1)\tau}\right\|}, \frac{\mathbf{x}_{M-1-(m-1)\tau} - \bar{\mathbf{x}}_{M-1-(m-1)\tau}}{\left\|\mathbf{x}_{M-1-(m-1)\tau} - \bar{\mathbf{x}}_{M-1-(m-1)\tau}\right\|}\right) = = d_{NE}(\mathbf{x}_{k-1-(m-1)\tau}, \mathbf{x}_{M-1-(m-1)\tau}).$$

Обозначим $L(\mathbf{x}_{M-1-(m-1)\tau}) = \{\mathbf{y}_{k-1-(m-1)\tau} = a\mathbf{x}_{k-1-(m-1)\tau} + \mathbf{b} : a \in \mathbb{R}^1\setminus\{0\},\ \mathbf{b} = (b,b,...,b) \in \mathbb{R}^m\}$.

Из множества $L(\mathbf{x}_{M-1-(m-1)\tau})$ выбираем точку $\tilde{\mathbf{x}}_M$, ближайшую точку к $\mathbf{x}_{M-1-(m-1)\tau}$ (в смысле евклидового расстояния):

$$\tilde{\mathbf{x}}_M = \tilde{a}\mathbf{x}_{k-1-(m-1)\tau} + \tilde{\mathbf{b}}.$$

$$\tilde{a} = \frac{n\left(\sum_{i=1}^{L} x_{M-1-(m-i)\tau}\cdot x_{k-1-(m-i)\tau} - \sum_{i=1}^{L} x_{M-1-(m-i)\tau}\sum_{i=1}^{L} x_{k-1-(m-i)\tau}\right)}{n\sum_{i=1}^{L} x^2_{k-1-(m-i)\tau} - \left(\sum_{i=1}^{L} x_{k-1-(m-i)\tau}\right)^2};$$

$$\tilde{b} = \frac{\left(\sum_{i=1}^{L} x_{M-1-(m-i)\tau} \cdot \sum_{i=1}^{L} x^2_{k-1-(m-i)\tau} - \sum_{i=1}^{L} x_{M-1-(m-i)\tau} \cdot x_{k-1-(m-i)\tau} \sum_{i=1}^{L} x_{k-1-(m-i)\tau}\right)}{n\sum_{i=1}^{L} x^2_{k-1-(m-i)\tau} - \left(\sum_{i=1}^{L} x_{k-1-(m-i)\tau}\right)^2}.$$

Значение $\tilde{x}_M = \tilde{a}x_k + \tilde{b}$ будем рассматривать в качестве прогноза значения x_M временного ряда $\{x_i,\ i = 1,2,...,N\}$.

Сравнение различных методов прогнозирования цен. Метод ближайших соседей с нормализованным евклидовым расстоянием был применен для прогноза цен на металлы. Прогноз строился на 6 месяцев – июль-декабрь 2010 г. В качестве прогноза значений $x_M, x_{M+1},..., x_{M+5}$ брались значения $\tilde{x}_M, \tilde{x}_{M+1},..., \tilde{x}_{M+5}$.

Интерес представляет сравнение прогнозов, построенных методом ближайших соседей с нормализованным евклидовым расстоянием и другими методами (в частности, методом ближайших соседей с евклидовым расстоянием).

На рис. 1 приведены исторические (март 2001 г – июнь 2010 г) значения очищенных от инфляции цен (в $/тонну) на стальной г/к рулон, алюминий и цинк. Для этих показателей далее разными методами строились прогнозы на 6 месяцев (июль – декабрь 2010 г).

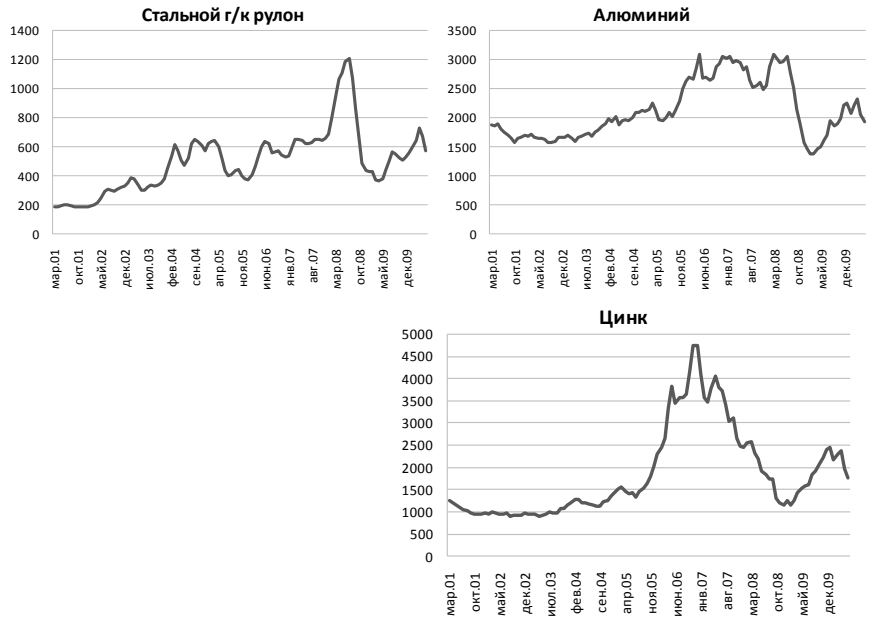

Рис. 1. Динамика цен на различные виды металлов (очищенных от инфляции), март 2001 – июнь 2010, $/т

Были применены следующие методы прогноза: метод ближайших соседей с использованием евклидового расстояния (БСЕ), и метод ближайших соседей с использованием нормализованного евклидового расстояния (БСНЕ); метод авторегрессии и скользящего среднего (АРСС); метод индексов сезонности (ИС). В случае метода ближайших соседей были построены прогнозы при m=9 и τ=3.

Для каждого показателя с помощью перечисленных методов прогнозирования построены по два прогноза: первый, когда значения цен в 2008 г, соответствующие периоду кризиса, не использовались для построения прогноза (с исключением кризиса), и второй, когда использовались (без исключения кризиса). Для метода АРСС в первом случае использовалась модель с интервенциями.

Результаты прогнозирования показаны на графиках (см. рис. 2 – 4).

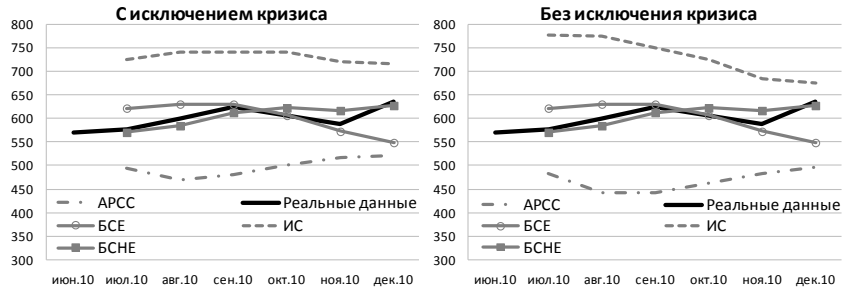

Рис. 2. Реальные значения цены на г/к рулон и прогноз на июль–декабрь 2010 г., полученный методами АРСС, БСЕ, БСНЕ, ИС

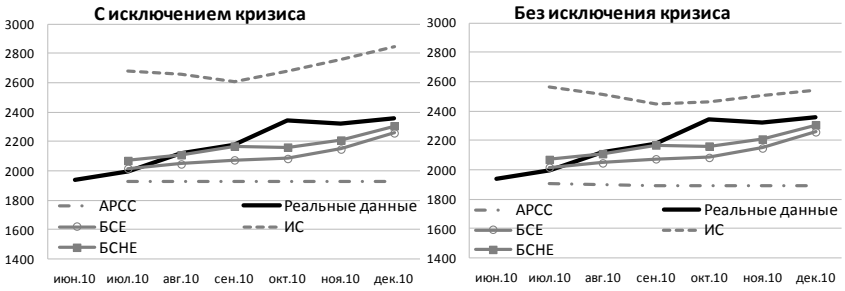

Рис. 3. Реальные значения цены на алюминий и прогноз на июль–декабрь 2010 г., полученный методами АРСС, БСЕ, БСНЕ, ИС

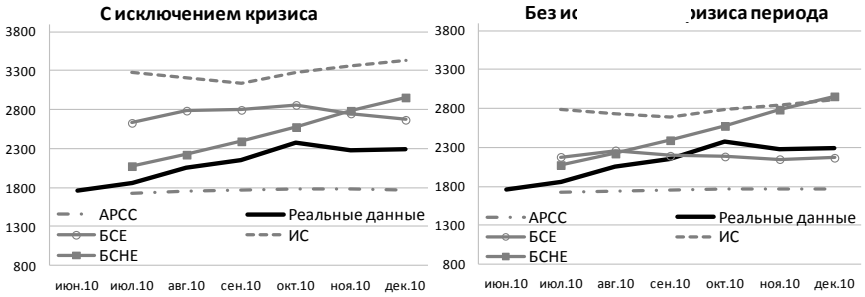

Рис. 4. Реальные значения цены на цинк и прогноз на июль–декабрь 2010 г., полученный методами АРСС, БСЕ, БСНЕ, ИС

В табл. 1 приведены средние погрешности прогнозирования цен:

$$\Delta_1 = 100\,\% \cdot \frac{1}{6}\sum_{i=0}^{5}(\tilde{x}_{M+i} - x_{M+i}) / x_{M+i};$$

и средние абсолютные погрешности:

$$\Delta_2 = 100\,\% \cdot \frac{1}{6}\sum_{i=0}^{5}|\,\tilde{x}_{M+i} - x_{M+i}\,| / x_{M+i}.$$

Таблица 1

Погрешности прогнозов, полученных различными методами

Показатель	Метод	Погрешность			
		С исключением кризиса		Без исключения кризиса	
		Δ_1	Δ_2	Δ_1	Δ_2
Цена на г/к рулон	БСЕ	– 0,4 %	5 %	– 0,4 %	5 %
	БСНЕ	0,1 %	2 %	0,1 %	2 %
	ИС	21 %	21 %	21 %	21 %
	АРСС	– 18 %	18 %	– 23 %	23 %
Цена на алюминий	БСЕ	– 5 %	5 %	– 5 %	5 %
	БСНЕ	– 2 %	3 %	– 2 %	3 %
	ИС	22 %	22 %	13 %	13 %
	АРСС	– 13 %	13 %	– 14 %	14 %
Цена на цинк	БСЕ	28 %	28 %	2 %	8 %

	БСНЕ	15 %	15 %	15 %	15 %
	ИС	52 %	52 %	30 %	30 %
	АРСС	– 18 %	18 %	– 19 %	19 %

Во всех рассмотренных случаях лучшие прогнозы давал метод ближайших соседей с использованием нормализованного евклидова расстояния.

Замечание. Предложенный метод также апробировался для прогноза цен на г/к рулон на периоды продолжительностью в 6 мес. начиная с периода [июнь 09, июль 09,…, дек. 09] и заканчивая периодом [июнь 10, июль 10,…, дек. 10]. При этом погрешности имеют такой же порядок, как и приведенные в таблице.

Выводы. В статье показаны преимущества метода ближайших соседей с нормализованным евклидовым расстоянием при прогнозировании нестационарных временных рядов на примерах нестационарных рядов цен на металлы и металлопродукцию.

Библиографические ссылки

1. **Зульпукаров М.-Г.М.** Метод русел и джокеров на примере исследования системы Розенцвейга–Макартура. / М.-Г.М. Зульпукаров, Г.Г. Малинецкий, А.В. Подлазов // Институт прикладной математики им. М.В. Келдыша РАН (препринт) – 2006. – № 21.
2. **Лоскутов А.Ю.** Временные ряды: анализ и прогноз. / А.Ю. Лоскутов, О.Л. Котляров, Д.И. Журавлев // Математика. Компьютер. Образование. Сборник трудов XI международной конференции. – Ижевск, 2004. – С. 9-46.
3. **Малинецкий Г.Г.** Современные проблемы нелинейной динамики. / Г.Г. Малинецкий, А.Б. Потапов. – М., 2000.
4. **Малинецкий Г.Г.** Русла и джокеры: о новых методах прогноза поведения сложных систем. / Г.Г. Малинецкий, А.Б. Потапов. // ИПМ им. М.В.Келдыша РАН (препринт). – 1998. – № 32.
5. **Комарова Н.В.** Некоторые инструменты нелинейной динамики в анализе и прогнозировании временных рядов. / Н.В. Комарова // Питання прикладної математики і математичного моделювання. Зб. наукових праць. – Д., 2011. – С. 191-199.

Надійшла до редколегії 08.06.2012