УДК 004.42:519.25

ПРИМЕНЕНИЕ ВЫБОРОЧНОГО КОЭФФИЦИЕНТА ДЕТЕРМИНАЦИИ ДЛЯ ПОСТРОЕНИЯ И АНАЛИЗА КРОСС-КОРРЕЛЯЦИОННЫХ ФУНКЦИЙ

В.Е. Бахрушин, В.Е. Павленко, С.В. Петрова

Классический приватный университет, ул. Жуковского, 70-б, Запорожье, Украина, 69002
Vladimir.Bakhrushin@zhu.edu.ua

Предложена методика построения и анализа кросс-корреляционных функций многомерных временных рядов, основанная на использовании выборочного коэффициента детерминации в качестве показателя наличия связи. Показано преимущество используемого подхода по сравнению с традиционными при наличии нелинейных связей в исследуемых системах.

Введение

При исследовании динамических систем часто возникает ситуация, когда одни процессы, протекающие в них, оказывают влияние на другие с некоторым отставанием (временным лагом). Одним из методов выявления таких связей является кросс-корреляционный анализ, который широко используют при исследовании динамических систем разной природы [1 – 3].

Кросс-корреляционную функцию для двух стационарных временных рядов (компонент многомерного временного ряда) x_t, y_t определяют как зависимость значения коэффициента парной корреляции Пирсона между рядами x_t и y_{t+k} от величины лага k [4]:

$$r_k = r(k) = \frac{\sum_{t=1}^{n-k} X_t Y_{t+k} - \sum_{t=1}^{n-k} Y_t \sum_{t=1}^{n-k} X_{t+k} / (n-k)}{\sqrt{\left[\sum_{t=k}^{n-k} Y_t^2 - \sum_{t=k}^{n-k} Y_t^2 / (n-k)\right]\left[\sum_{t=k+1}^{n} X_t^2 - \sum_{t=k+1}^{n} X_t^2 / (n-k)\right]}}, \qquad (1)$$

где n – количество элементов в полных рядах x_t, y_t.

При отсутствии связи между исследуемыми рядами все значения кросс-корреляционной функции близки к нулю. Если при некоторых значениях k на ней наблюдаются максимумы, это свидетельствует о наличии связи с величиной запаздывания, равной соответствующему временному лагу.

Коэффициент корреляции Пирсона и, соответственно, кросс-корреляционная функция являются мерами линейной связи [5]. Их значения могут варьироваться от −1 до +1. Значения, близкие по абсолютной величине к единице, свидетельствуют о наличии прямой или обратной линейной связи. Значения, близкие к нулю, могут наблюдаться, как при отсутствии какой-либо связи, так и при наличии нелинейных связей. Поэтому существует необходимость в разработке средств диагностики многомерных временных рядов, которые были бы чувствительны к наличию нелинейных связей.

Универсальной характеристикой тесноты связи между количественными признаками является выборочный коэффициент детерминации [5]. Для его вычисления данные предварительно упорядочивают по возрастанию значений одной из компонент, например ряда x. Затем весь диапазон от x_{min} до x_{max} делят на интервалы равной ширины так, чтобы количество точек в каждом интервале было не менее 5 -10, а связь между x и y внутри интервала

была близка к линейной. После этого коэффициент детерминации рассчитывают по формуле:

$$K_d(y;\mathbf{X}) = 1 - \frac{\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{\nu_j}\left(y_{ij} - \bar{y}_{j*}\right)^2}{\sum_{i=1}^{n}\frac{\left(y_i - \bar{y}\right)^2}{n}}, \qquad (2)$$

где m – количество интервалов, ν_j − количество данных, попавших в j-й інтервал; y_{ji} − значение i-го наблюдения компоненты y, попавшего в j-й интервал; $\bar{y}_{j*} = \frac{\sum_{i=1}^{\nu_j} y_{ji}}{\nu_j}$ − среднее значение компоненты y по результатам, попавшим в j-й интервал.

Величина коэффициента детерминации находится в пределах от нуля до единицы и отображает долю общей дисперсии компоненты y, которая обусловлена влиянием компоненты x. Нулевое значение соответствует отсутствию какой-либо связи, а единичное – наличию строго функциональной связи между исследуемыми компонентами. Описанный выше способ расчета коэффициента детерминации соответствует приближению неизвестной связи ломаной линией. Недостатком этого показателя является влияние способа группирования данных на получаемый результат. Однако при наличии сильной связи это влияние будет не очень существенным.

Для реализации основных методов анализа данных широко используют статистические пакеты SPSS, Statistica и другие [6, 7]. Авто- и кросс-корреляционный анализ временных рядов в этих пакетах можно проводить только с использованием коэффициента корреляции Пирсона.

Целью данной работы являлась разработка алгоритма и программного продукта, предназначенных для проведения кросс-корреляционного анализа многомерных временных рядов на основе использования коэффициента детерминации в качестве показателя наличия связи.

Алгоритм и программа кросс-корреляционного анализа

Для хранения данных выделяем в памяти место для двух исходных массивов с заданным количеством элементов. Кроме того формируем два дополнительных массива для хранения сдвинутых временных рядов на каждом этапе анализа. Выделяем память и формируем переменные для хранения результатов анализа и промежуточных данных. При импорте данных каждое значение вносим в соответствующий массив.

Формируем цикл по величине сдвига. Максимальное значение лага задаем в пределах n/4 – n/3 в зависимости от целей исследования и длины анализируемого ряда. На каждом шаге для расчета коэффициента детерминации находим общее среднее ряда значений зависимого признака. Затем формируем интервалы и для каждого вычисляем количество точек и среднее арифметическое значений зависимого признака, попавших в соответствующий интервал. Далее по формуле (2) вычисляем коэффициент детерминации.

По результатам расчетов строим диаграмму, где по оси абсцисс откладываем величину лага, а по оси ординат – соответствующее значение коэффициента детерминации.

Программа написана на языке C++ в компиляторе Borland C++ Builder 6. Сначала был создан новый проект, форма которого приведена на рис. 1.

При написании программы были созданы такие функции:

float max(float a[], int msize) – функция нахождения максимального элемента из msize

первых элементов массива a[]. При вызове функции ей передаются такие данные: массив действительных чисел в в котором будет выполняться поиск наибольшего элемента и длина этого массива. В теле функции порождаются переменные im целочисленного типа и max действительного типа.

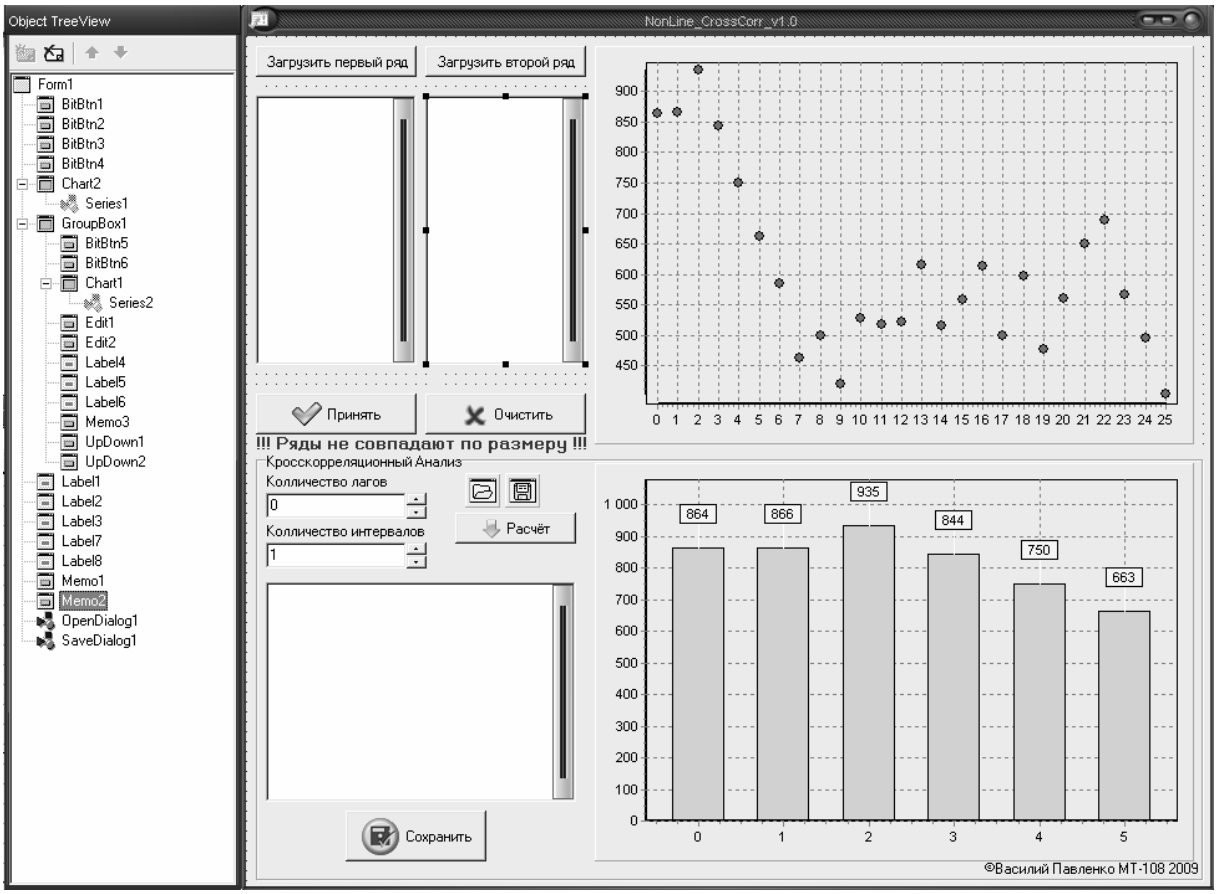

Рис. 1. Внешний вид формы с компонентами

Переменной max присваивается значение первого элемента массива. Переменная im является счетчиком при выполнении цикла и изменяется от единицы до msize. На каждом шаге цикла будем сравнивать элемент массива a[im] с переменной max. В случае, если значение элемента будет больше, переменная принимает это значение, в противном случае она остается неизменной. По окончании цикла в переменной max будет храниться значение наибольшего элемента массива, возвращаемое в программу, из которой вызывается функция.

– float min(float a[], int msize) – функция нахождения минимального элемента из msize первых элементов массива a[]. На первом шаге порождается переменная min, которой присваивается значение первого элемента массива. Входные данные и дальнейшая работа функции такие же, как и в предыдущем случае.

– float SRED(float a[], int msize) – функция нахождения среднего значения msize первых элементов массива a[]. На вход функции подаются массив и количество его элементов. Сначала порождается и аннулируется переменная SUM. Далее в цикле по номерам элементов на каждом шаге к ней прибавляется значение соответствующего элемента массива. На выходе функция возвращает среднее арифметическое элементов массива.

– float KoefDet(float a[], float b[], int size, int k) – функция для вычисления коэффициента детерминации, где a[], b[] – массивы значений исходных временных рядов, size – количество значений каждого ряда, k – количество интервалов, формируемых при расчетах. Ко-

эффициент детерминации рассчитываем по формуле (2). Для удобства в программе разработаны отдельные функции для вычисления числителя и знаменателя (2). В переменную Ro после выполнения функции записывается значение коэффициента детерминации. Значение переменной Bsr вычисляется через вызов функции нахождения среднего элемента массива и передается как входной параметр в функции расчета числителя и знаменателя (2). Результаты вызова соответствующих функций записываются в переменные ZNAM и CHIS.

– float Znamenat(float a[], int msize, float sr) – функция для вычисления знаменателя формулы (2), где sr – среднее значение элементов ряда. В функции порождается переменная SUM, значение которой на первом шаге равно нулю, а далее в цикле по номерам элементов входного массива на каждом шаге к нему прибавляется квадрат разности соответствующего элемента и среднего значения массива. Результат делится на количество элементов и подается на выход функции.

– int KolVInt(float Ryad[], int size, float left, float right) – функция для подсчета количества элементов, попадающих в интервал (left; right) ряда Ryad[] размера size. При порождении целочисленной переменной KOL функции вначале присваивается значение 0. Затем в цикле по номерам элементов ряда проверяется выполнение условия о том, что значение текущего элемента больше левой и меньше правой границы, которые подаются на вход функции. При выполнении условия переменная KOL увеличивается на единицу. По завершении цикла в переменной KOL будет храниться число элементов массива, попадающих в заданный интервал.

– float SrInt(float aRyad[], float bRyad[], int size, float left, float right) – функция, вычисляющая среднее по интервалу значение ординат bRyad[] двумерных выборок размера size, на интервале (left; right). В цикле по номерам элементов проверяем выполнение условия попадания текущего элемента в заданный интервал. При его выполнении к переменной SUM прибавляется значение соответствующего элемента, вследствие чего по завершении цикла в ней будет храниться значение суммы элементов, попавших в заданный интервал. Разделив его на количество элементов, вычисляемое функцией KolVInt, получаем среднее арифметическое для рассматриваемого интервала.

– float Chislit(float a[], float b[], int msize, float sr, int kol) – функция вычисления числителя формулы (2). Переменным Xmin и Xmax присваиваются результаты вычисления функций нахождения минимального и максимального элементов массива. Ширина интервала рассчитывается, как отношение их разности к числу интервалов. Переменной L, в которую записывается значение левой границы каждого интервала, присваивается значение минимального элемента ряда, после чего выполняется цикл по номерам интервалов. На каждом шаге вначале вычисляется правая граница прибавлением к левой границе ширины интервала. Затем проверяется условие, что интервал не является крайним правым. Если оно выполняется, то к правой границе прибавляется небольшое число для того, чтобы при дальнейших расчетах не было потеряно значение самой правой точки. Переменной sredint присваивается среднее по интервалу значение, получаемое вызовом соответствующей функции. К переменной аккумулятору SUM на каждом шаге цикла прибавляем число, равное количеству элементов, попавших в соответствующий интервал, умноженному на квадрат разности среднего по интервалу и общего среднего. На последнем шаге цикла переменной, в которой хранится значение левой границы отрезка, присваивается текущее значение правой границы, после чего цикл повторяется. После прохождения цикла kol раз, значение переменной SUM делится на количество элементов массива и подается на выход функции.

Примеры анализа кросс-корреляции модельных временных рядов

Для тестирования программы были сформованы и сохранены в файлы такие модельные временные ряды:

1. Два ряда с параметрами: количество элементов – 200, закон распределения – нор-

мальный, среднее значение – 100, стандартное отклонение – 50.

2. Два ряда по 200 элементов каждый, полученные по формулам $y = 1,535 \cdot x + b$ и $z = 0,1 \cdot x^2 - 20 \cdot x + b \cdot 20$, где x – первый временной ряд, b – временной ряд случайных чисел с нормальным законом распределения, нулевым средним и стандартным отклонением равным пяти;

3. Два ряда по 190 элементов каждый, полученные удалением из ряда x первых десяти элементов, а из ряда – последних десяти.

В случае 1 корреляционная связь отсутствует (рис. 1). Значения коэффициента детерминации в разработанном программном продукте не превышает 0,1, а коэффициента корреляции Пирсона, рассчитанного в пакетах SPSS и Statistica, – 0,2.

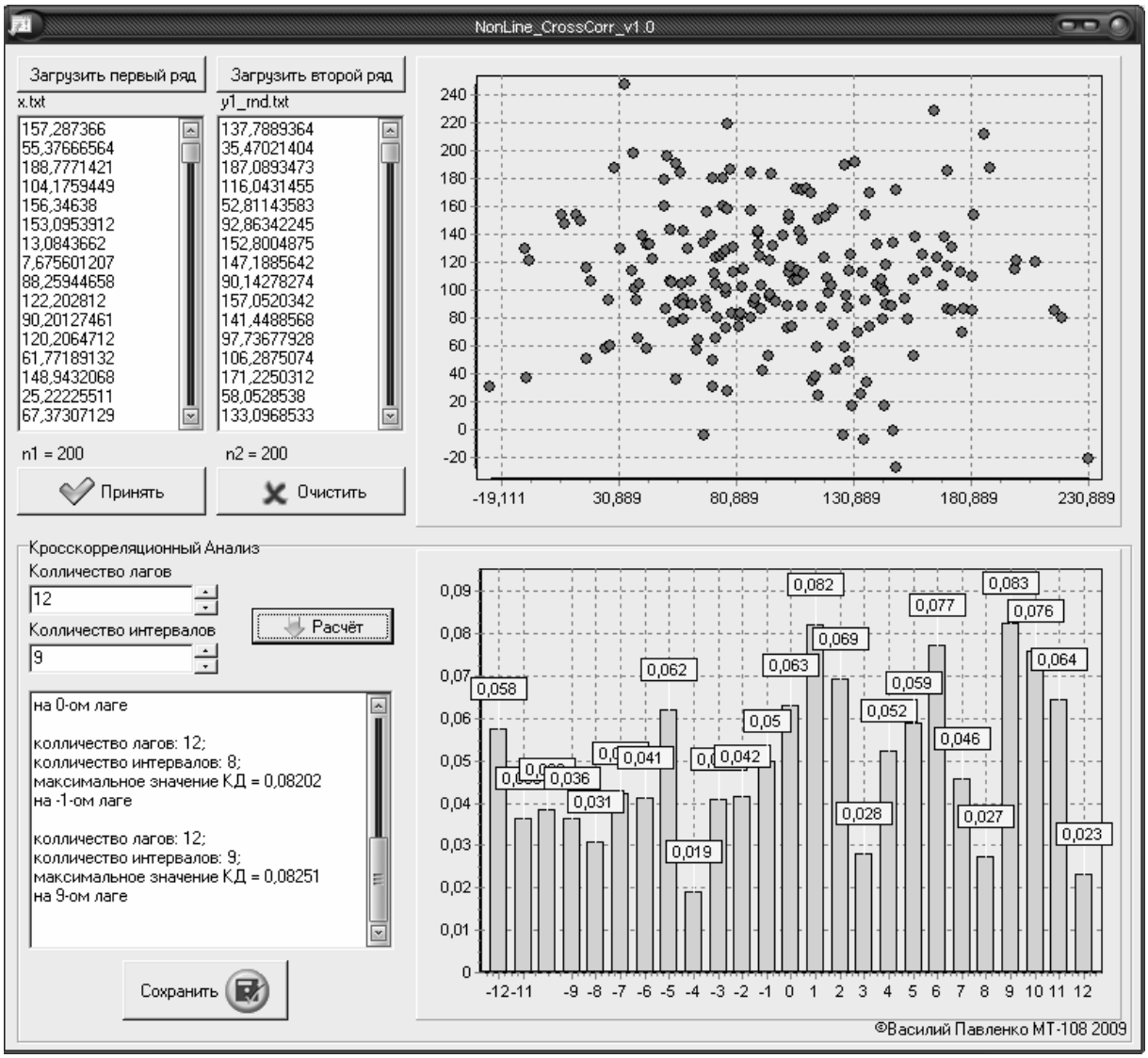

Рис. 2. Результаты крос-кореляционного анализа для случая 1, выполненного с применением разработанной программы

Между рядами x, y наблюдается сильная корреляционная связь. Коэффициент детерминации (для разработанного продукта) и коэффициент корреляции Пирсона (для пакетов SPSS, Statistica) близки к единице, что соответствует методике построения этих рядов.

Коэффициент детерминации между рядами x, z равен 0,9 на нулевом лаге (рис. 3) при расчете с помощью разработанной программы. Анализ в пакетах SPSS и Statistica не выявляет наличия кросс-корреляции. Значение коэффициента корреляции Пирсона не превышает

0,2. Это подтверждает то, что указанные пакеты не могут использоваться для выявления нелинейных связей в многомерных временных рядах.

В случае 3 коэффициент детерминации между исследуемыми рядами равен 0,89 на 10 лаге (рис. 4), что соответствует способу построения указанных рядов. Пакеты SPSS и Statistica, как и в предыдущем случае, не выявляют наличия кросс-корреляции (рис. 4).

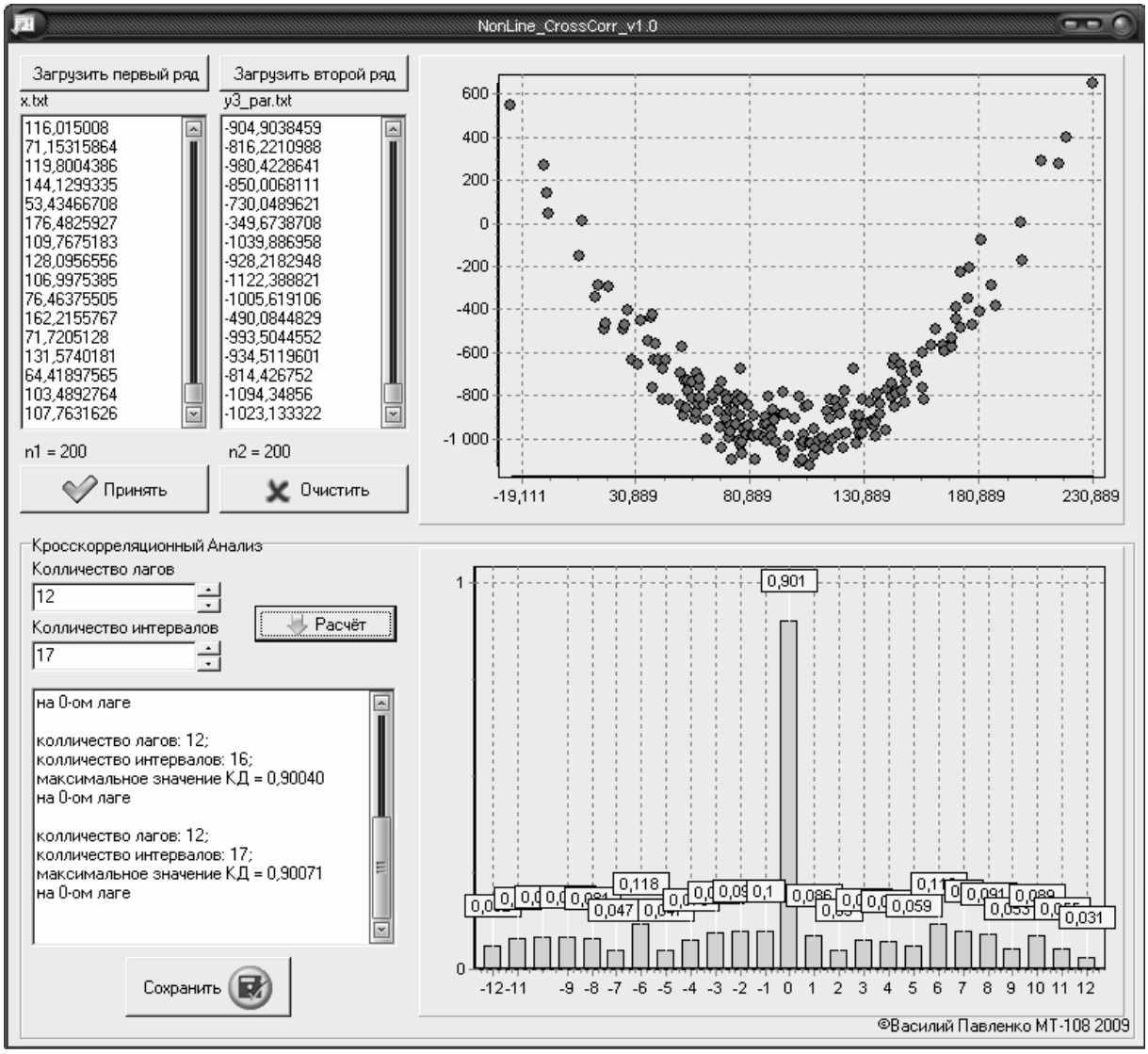

Рис. 3. Результаты крос-кореляционного анализа для рядов x, z, выполненного с применением разработанной программы

Выводы

Имеющиеся универсальные пакеты статистического анализа (SPSS, Statistica и др.) нецелесообразно использовать для анализа кросс-корреляции в многомерных временных рядах, если есть основания предполагать существование нелинейных связей между компонентами. Это обусловлено тем, что в основу анализа положено вычисление коэффициента корреляции Пирсона, чувствительного только к наличию линейных связей.

Нами разработаны алгоритм и программный продукт, дающие возможность проведения кросс-корреляционного анализа при наличии нелинейных связей. В основу алгоритма положено использование выборочного коэффициента детерминации, как универсального показателя нелинейной связи.

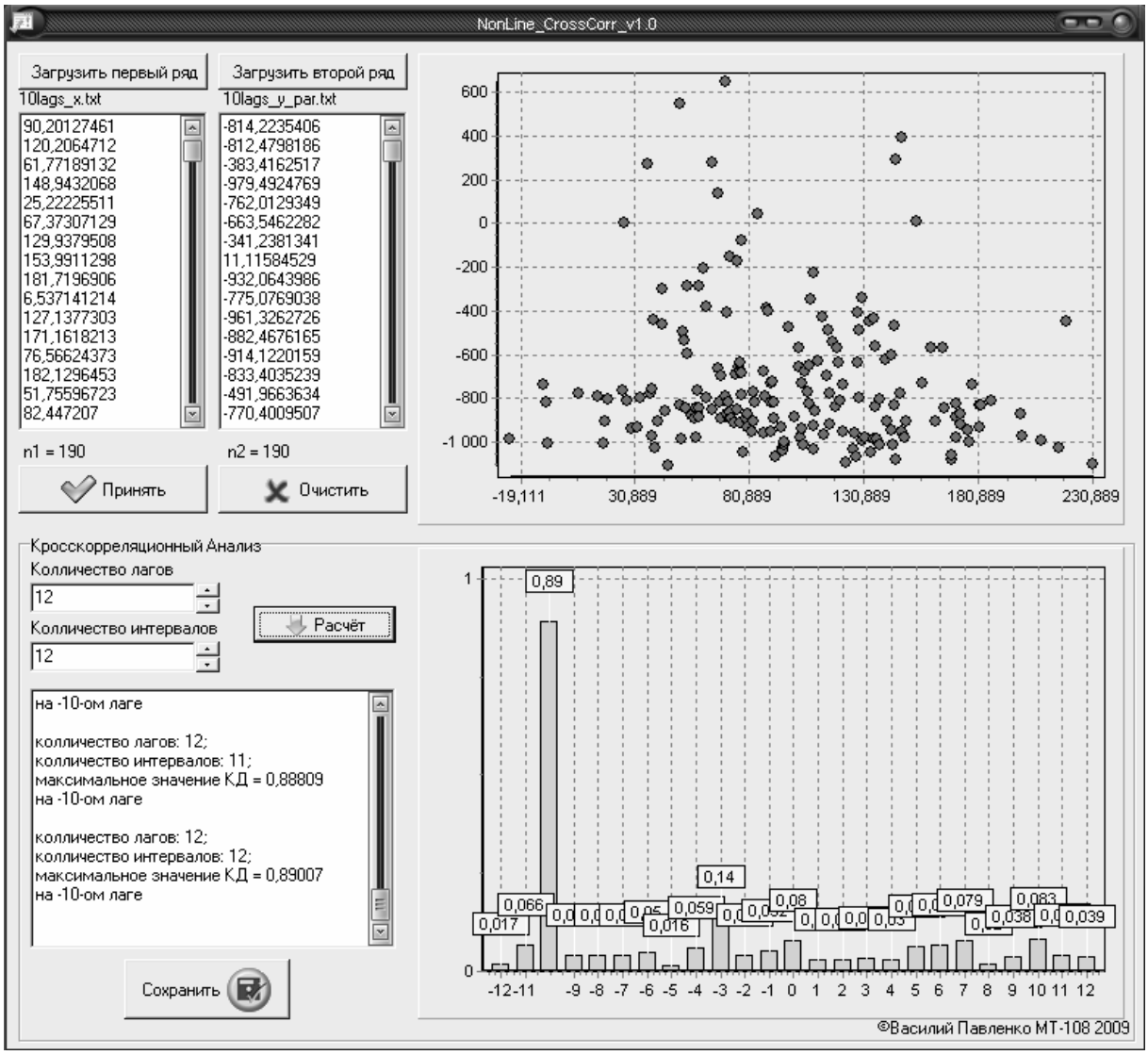

Рис. 4. Результаты крос-кореляционного анализа случая 3, выполненного с применением разработанной программы

Література

1. Бесконтактный спекл-интерферометрический измеритель малых смещений / Е.А. Аксенов, А.А. Шматко, В.И. Зворский, А.С. Кравчук // Радіоелектронні і комп'ютерні системи. – 2008. – № 1(28). – С. 15 – 19.

2. Кросс-корреляции в живых системах: Анализ нейромагнитных сигналов коры головного мозга человека / С.А. Демин, Р.Р. Зарипов, О.Ю. Панищев, Р.М. Юльметьев // Научно-технический вестник СП-бГУ информационных технологий, механики и оптики. – 2007. – № 37. – С. 202 – 212.

3. Фортус М.И. Анализ корреляционных связей между временными рядами с помощью фазового спектра // Известия РАН. Физика атмосферы и океана. – 2007. – Т. 43, № 5. – С. 602 – 616.

4. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей: Справ. изд. – М.: Финансы и статистика, 1985. – 487 с.

5. Бахрушин В.Є. Аналіз даних. – Запоріжжя: ГУ „ЗІДМУ”, 2006. – 128 с.

6. Бююль А., Цёфель П. SPSS: Искусство обработки информации. – СПб.: ООО ДиаСофтЮП, 2005. – 608 с.

7. Электронный учебник по статистике. [Электронный ресурс]: Режим доступа: http://www.statsoft.ru/home/textbook/default.htm.