УДК 51.7:532.546

# Методика определения коэффициента корреляции для нестационарных временных рядов

© Д. С. Кириллов[1], Л. В. Клочкова[2], Ю. Н. Орлов[3], В. Ф. Тишкин[4]

**Аннотация.** В работе строится метод определения выборочного коэффициента корреляции для двух нестационарных временных рядов. Отличие от стационарного случая состоит в том, что одновременно с величиной коэффициента корреляции строятся эмпирические статистики оптимальной длины выборки и доверительного интервала, содержащего коэффициент корреляции в наибольшем числе случаев.

**Ключевые слова:** нестационарный коэффициент корреляции, оптимальная длина выборки, оптимальный уровень достоверности

## 1. Введение

Во многих задачах долгосрочного прогнозирования величин, модель изменения которых во времени предполагает наличие случайных процессов, требуется выделить главные векторы корреляционной матрицы, чтобы определить группы величин, находящиеся между собой в статистической зависимости. Таковы модели и прогнозы макроэкономических показателей, модели факторов, влияющих на экологическое состояние, статистические модели в социологии и др. В частности, в работе [1] была поставлена задача о методах оценки корреляции эпидемического состояния мегаполиса в зависимости от уровня загрязненности воздуха. Настоящая работа продолжает начатое исследование.

Традиционно практические исследования с использованием статистических методов в этих областях опираются на модель реакции (линейной или нелинейной) сложной системы при изменении параметров, представляемых как внешние или управляющие воздействия. При этом цель анализа данных за определенный исторический период состоит в определении коэффициентов в такой модели, предполагая, что они образуют стационарный временной ряд. Чаще всего анализируются линейные функции реакции, когда предполагается, что два наблюдаемых нестационарных ряда данных находятся между собой в стационарной статистической взаимосвязи. В реальности стационарность связи выполняется лишь приближенно, поскольку в сложных системах коэффициенты в параметрических моделях являются неизвестными функционалами от анализируемых рядов данных. Поэтому все модели коинтеграции нестационарных рядов содержат коррелированные остатки регрессий. В результате оказывается, что очевидный качественный результат о том, что некоторые два ряда находятся в корреляционной зависимости - например, уровень загрязненности атмосферы определенным типом вещества и заболеваемость населения тем или иным видом заболевания, не имеет четкого численного подтверждения. Именно, если зафиксировать лаг корреляции и момент времени наблюдения, и начать затем варьировать длину выборки, то часто обнаруживается, что выборочный коэффициент корреляции

[1] Аспирант Института прикладной математики им. М.В.Келдыша РАН, г. Москва; ov3159fd@yandex.ru.

[2] Старший научный сотрудник Института прикладной математики им. М.В.Келдыша РАН, г. Москва; klud@imamod.ru.

[3] Ведущий научный сотрудник Института прикладной математики им. М.В.Келдыша РАН, г. Москва; ov3159fd@yandex.ru.

[4] Заместитель директора Института прикладной математики им. М.В.Келдыша РАН, г. Москва; v.f.tishkin@mail.ru.

изменяется при этом в весьма широких пределах и даже может менять знак. И обратно, если зафиксировать длину выборки, то корреляция в разные моменты времени может быть сильно нестационарной. Следовательно, актуальной является задача корректного определения интересующей исследователя корреляционной связи.

В настоящей работе излагается методика определения величины корреляции между двумя нестационарными временными рядами. Математически проблема состоит в том, что выборочный коэффициент корреляции между двумя рядами данных стабилизируется к своему генеральному значению с увеличением объема выборки только для стационарных в широком смысле процессов, среднее и дисперсия которых не зависят от времени. Для нестационарных процессов наблюдается непостоянство корреляции как функции времени для выборки определенного объема, а также и непостоянство ее в один и тот же момент времени, но для выборок разных объемов. Как определить величину корреляции и ее достоверность для нестационарных рядов? Необходимым условием для корректного определения нестационарной корреляции является нахождение оптимальной длины выборки, на которой следует вычислять выборочные моменты распределений. Именно, требуется найти длину выборки, на которой определенное значение корреляции между двумя рядами проявляется с наибольшей достоверностью.

Формально нестационарный выборочный коэффициент корреляции представляет собой обычный выборочный коэффициент корреляции, зависящий от длины выборки и текущего момента времени, от которого назад в прошлое отсчитывается эта выборка.

Для каждого интервала, ширина которого определяет точность определения коэффициента корреляции, можно найти длину выборки, на которой доля корреляций, попадающих в этот промежуток, наибольшая. Эта доля и представляет собой уровень достоверности корреляционной связи. Тот интервал, для которого эта доля абсолютно наибольшая, принимается за промежуток, содержащий коэффициент корреляции. Этот методический подход и описывается в настоящей работе.

## 2. Определение статистической функциональной связи.

Опишем сначала общий метод отыскания корреляционной статистической зависимости между случайными величинами. Идея метода была предложена в [2] для анализа нелинейных хаотических динамических систем. Для краткости мы ограничимся далее рассмотрением двух случайных величин, равномерно ограниченных по времени на отрезке $[0;1]$. Для каждой из них можно построить распределение вероятностей попадания значений в определенные классовые интервалы, на которые на практике разбивается промежуток $[0;1]$. Будем для простоты считать разбиение на промежутков равномерным. В этом случае вероятность и плотность вероятности аппроксимируются одной и той же гистограммой.

Предположим сначала, что ряды $\{x_i\}$ и $\{y_i\}$ стационарны, и между ними есть прямая функциональная связь, т.е. $y = \varphi(x)$. Тогда, построив совместное распределение вероятностей $F_N(x, y, t)$ по выборке длины $N$ в момент времени $t$, мы обнаружим, что с точностью $1/\sqrt{N}$ в норме суммируемых функций оно не меняется, и его носитель находится в квадратах $y_i = \varphi(x_j)$ , в соответствии с разбиением гистограммы. Точность, с которой мы можем говорить о такой функциональной связи, равна точности позиционирования случайных величин в классовых интервалах, т.е. $1/n$ . Действительно, точность оценки функциональной связи -— это мера носителя совместного распределения в единичном квадрате. При фиксированной мелкости разбиения эта доля, т.е. точность оценки, не может быть меньше, чем $1/n$ ($n$ квадратов с площадью $1/n^2$). Уровень достоверно-

сти полученной оценки равен интегралу от плотности распределения по выбранной доле носителя. Поскольку в данном случае вне отмеченных квадратов нет точек носителя совместного распределения, уровень достоверности равен единице.

Если функциональной связи нет, то при фиксированном номере $j$ интервала по первому аргументу мы обнаружим отличные от нуля значения функции $F_N(j,i,t)$ для нескольких номеров $i$ интервалов по второму аргументу. При этом с увеличением длины выборки носитель совместного распределения занимает все большую долю области разбиения гистограммы. Это означает, что путем потери точности можно получить достоверную оценку функциональной связи даже в отсутствие таковой, но будет ли это удовлетворять исследователя? Насколько точно нужно позиционировать искомое значение, чтобы и вероятность его принадлежности определенному интервалу была не исчезающе малой, и сам интервал существенно отличался бы от всего множества значений случайной величины? Представляется, что вместо априори задаваемого уровня значимости следует ввести согласованный критерий совместной оценки точности и уровня значимости.

Пусть $\delta = \int\limits_{\Omega} dxdy$ есть мера множества $\Omega(x,y)$, принадлежащего носителю совместного распределения, на котором можно однозначно говорить о связи между $x$ и $y$. Величина $\delta$ будет точностью, с которой установлена эта связь, а величина $\alpha = \int\limits_{\Omega} F(x,y)dxdy$ будет давать уровень достоверности найденной связи. И $\delta$, и $\alpha$ зависят от множества $\Omega(x,y)$. Тогда введем функционал, минимизирующий совокупную ошибку оценки корреляционной связи (не обязательно линейной), которая находится на том множестве $\Omega(x,y)$, где

$$U^2 = (1-\alpha(\Omega))^2 + \delta(\Omega)^2 \to \min. \tag{2.1}$$

Заметим, что условие (2.1) не позволяет в общем случае однозначно определить множество $\Omega(x,y)$. Для унимодальных распределений $F(x,y,t)$ это множество при фиксированном $x$ содержит $y = \arg\max F(x,y)$. Будем поэтому для определенности считать, что $\Omega(x,y)$ содержит полосу разбиения гистограммы, содержащую локальные (при фиксированных номерах -ячеек) максимумы распределения $F_N(j,i,t)$. Собственно значением $i(j)$ будет называться номер ячейки, содержащей условное среднее значение номера $i$ по множеству ячеек с $j$-ым номером вертикальной полосы, которое с точностью $\delta$ охватывает $\arg\max F(j,i)$.

Функционал $U$ зависит от длины выборки $N$. В стационарном случае с увеличением происходит лишь уточнение множества $\Omega(x,y)$ и, соответственно, уменьшение значения $U$. В нестационарном случае увеличение длины выборки сверх оптимального значения может привести к увеличению функционала $U$. Пусть $U(N,t)$ есть результат оптимизации (2.1) в данный момент времени $t$ по выборкам произвольных объемов, при котором находятся локально-оптимальные значения $\delta(N,t)$ и $\alpha(N,t)$. Тогда

$$N_{opt}(t) = \arg\min U(N,t). \tag{2.2}$$

Подчеркнем, что функциональная связь, определяемая множеством $\Omega(x,y;t)$, полученном в результате оптимизации (2.1), в разные моменты времени может быть совершенно различной. Также и оптимальные длины $N_{opt}(t)$ образуют в совокупности некоторое распределение с плотностью $\nu(N)$. В результате глобально оптимальным за рассматриваемый промежуток времени будет некоторое среднее из оптимальных длин и некоторая средняя функциональная связь. Точность этой связи будет определяться ее условной дисперсией (при условии, что имеется заданная неточность в определении локально-оптимальной длины выборки) и чувствительностью функциональной связи к длине выборки. Пример этого общего подхода будет далее рассмотрен подробно для конкретного вида искомой

функциональной связи – а именно, линейной. Меняющимся параметром в этом случае (т.е. множеством $\Omega(x, y; t)$ ) будет выборочный коэффициент корреляции.

## 3. Определение оптимальной корреляции между двумя временными рядами

Пусть $x(t)$ и $y(t)$ — два исследуемых временных ряда на промежутке времени $[1; T]$, и $R(k, t)$ есть их выборочный коэффициент корреляции по выборке объема $k$ в момент времени $t$, т.е. коэффициент корреляции, вычисленный по данным, взятым в моменты времени $t - k + 1$, $t - k + 2, \ldots, t$.

Корреляция $R(k, t)$ выборки $\{x_1, \ldots, x_k\}$ объема $k$ в момент времени $t$ на выборку $\{y_1, \ldots, y_k\}$ определяется $r(t)$ по формуле (3.3)

$$R(k,t) = \frac{k \sum\limits_{i=t-k+1}^{t} x_i y_i - \left(\sum\limits_{i=t-k+1}^{t} x_i\right)\left(\sum\limits_{i=t-k+1}^{t} y_i\right)}{\sqrt{k \sum\limits_{i=t-k+1}^{t} x_i^2 - \left(\sum\limits_{i=t-k+1}^{t} x_i\right)^2}\sqrt{k \sum\limits_{i=t-k+1}^{t} y_i^2 - \left(\sum\limits_{i=t-k+1}^{t} y_i\right)^2}}. \tag{3.3}$$

Определим два взаимодополняющих временных ряда: ряд $r(t)$ максимальных алгебраических значений корреляции $R(k, t)$, и ряд $n(t)$ соответствующих объемов выборки, на которых корреляция в данный момент времени $t$ максимальна:

$$r(t) = \max_k R(k,t) \quad n(t) = \arg\max_k R(k,t). \tag{3.4}$$

При вычислении $r(t)$ объем выборки должен превосходить некоторый минимальный объем, на котором вообще разумно вычислять корреляцию, поэтому будем считать, что $k \geq 3$.

Определим также среднее значение $\bar{r}$ и дисперсию $\sigma_r^2$ максимальных значений корреляции по всем моментам времени, и аналогично $\bar{n}$ и $\sigma_n^2$. Пусть также $\gamma$ есть коэффициент корреляции рядов $r(t)$ и $n(t)$ на промежутке времени $[1; T]$. Этот коэффициент подсчитывается по количеству данных $T - N + 1$, где $N \leq T$ есть максимальный объем среди всех $n(t) : N = \max\limits_t n(t)$.

Величина $\sigma_r/\bar{r}$ характеризует относительную ширину разброса средних по времени значений максимумов корреляций. Это условие для стационарных рядов является основным показателем точности оценки $R(n, t) = \bar{r}$. Если же объем выборки $n(t)$ также меняется со временем, то фактическая неточность в оценке корреляции $r(t)$ должна определяться с учетом вариации $\sigma_n/\bar{n}$. Однако при вариации объема выборки может оказаться, что корреляция на отрезке $\Delta_\sigma(n(t)) = [n(t) - \sigma_n; n(t) + \sigma_n]$ меняется незначительно. Поэтому, чтобы учесть вариацию объема выборки, необходимо оценить чувствительность корреляции по объему выборки.

Введем величину

$$\lambda(t) = \frac{n(t)}{2\sigma_n r(t)}\left(\max_{k \in \Delta_\sigma(n(t))} R(k,t) - \min_{k \in \Delta_\sigma(n(t))}\right), \tag{3.5}$$

которая представляет собой разностный аналог логарифмического размаха корреляции по логарифму объема выборки в окрестности точки максимума, который вычисляется на отрезке $\Delta_\sigma(n(t)) = [n(t) - \sigma_n; n(t) + \sigma_n]$ ширины $2\sigma_n$. Среднее по времени значение

$\bar{\lambda}$ характеризует среднюю крутизну графика корреляции как функции объема выборки в окрестности своего максимального значения. Если $\bar{\lambda}$ близко к нулю, т.е. график $R(n(t),t)$ представляет слабо меняющуюся функцию объема выборки, приблизительно параллельную оси абсцисс, то знание именно того объема $n(t)$, на котором достигается максимум корреляции, не обязательно, поскольку близкие к нему значения корреляции достигаются и на других объемах. Тем самым $\bar{\lambda}$ можно трактовать как чувствительность максимума корреляции к объему выборки. Тогда в качестве ориентировочной оценки относительной неточности $O^2$ в установлении корреляционной связи на уровне $\bar{r}$ между двумя рядами можно использовать показатель

$$O^2 = \left(\frac{\sigma_r}{\bar{r}}\right)^2 + (1-\gamma^2)(\bar{\lambda})^2 \left(\frac{\sigma_n}{\bar{n}}\right)^2 \tag{3.6}$$

Величина $\gamma^2$ определяет ту долю дисперсии максимальной корреляции, которая может быть «объяснена» дисперсией объема выборки в регрессионном приближении. Остаток представляет собственный вклад вариации объема выборки в неточность оценки корреляции. Тогда ширина доверительного интервала, содержащего оценку максимальной корреляции на уровне значимости $O$, равна $2O\bar{r}$.

Критерий $\ll O^2 \gg$ (3.6) дает уровень значимости оценки корреляционной связи, характеризуемой числом $\bar{r}$. Но объем выборки, на котором следует вычислять эту корреляцию, не фиксируется критерием (3.6). Этот объем может быть любым в границах $[\bar{n}-\sigma_n;\bar{n}-\sigma_n]$, что учитывается введением коэффициента чувствительности $\bar{\lambda}$. Поскольку же не обязательно наибольшая корреляция с наибольшей же достоверностью достигается на среднем объеме $\bar{n}$, дающем максимумы корреляций, то следует найти соответствующий оптимальный объем $n_{opt}$.

Введем достоверность $\alpha$ в определении объема выборки, при котором корреляция между рядами не ниже определенного уровня. Определим множество $M_n(t)$ тех объемов выборки $n$, корреляция по которым в момент времени $t$ лежит в определенной области $\Delta$ значений, т.е.

$$M_n(t) : \bigcup_n \{n : R(n,t) \in \Delta\}. \tag{3.7}$$

Корреляционная связь между двумя рядами на промежутке времени $[1;T]$ называется достоверной по объему выборки на уровне $\alpha$ с абсолютной ошибкой $\delta$, если пересечение $\bigcap\limits_{t=1}^{T} M_n^{\alpha}(t)$ не пусто. При этом должно быть не менее, чем $[\alpha(T-n(t)+1)]$ носителей $n(t)$ значений корреляции этих рядов как функции объема выборки, содержащихся в полосе минимальной ширины $\delta$ ($[1;1-\delta]$ для максимальной корреляции, $[-1;-1+\delta]$ для минимальной, $[-\delta/2;\delta/2]$ для нулевой корреляции).

Интерес представляют три случая: достоверная корреляция, достоверная антикорреляция, а также достоверное отсутствие корреляции (нулевая корреляция). Разберем подробно случай достоверной корреляции, т.е. устойчивого по времени достаточно высокого положительного коэффициента выборочной корреляции, вычисленного по одинаковому объему данных. Остальные случаи рассматриваются аналогично.

В каждый момент времени $t \in [1;T]$ строится выборочный коэффициент корреляции $R(n,t)$ как функция объема выборки $n$, $n \le t$, задается произвольное значение $0<\delta<1$ и рассматриваются те и только те значения $R(n,t)$, которые попали в полосу $[1;1-\delta]$. Пусть $\{n_i(t)\}$ — соответствующие значения объемов выборок. Множество $M_n(t)$ определяется как $M_n(t) = \bigcup\limits_i n_i(t)$. Значение $n_i(t) = m$ как элемент множества $M_n$ может

появиться в различные моменты времени не более чем $T-m+1$ раз. Если оно появилось именно столько раз, то во всех случаях (с достоверностью $\alpha=1$) объем выборки удовлетворяет условию принадлежности корреляции множеству $[1;1-\delta]$.

Зададим некоторое $0 \le \alpha \le 1$. Пусть значение $n_i(t)=m$ появилось $k$ раз. Если $k/(T-m+1) \ge \alpha$, то пересечение соответствующей части $\bigcap\limits_{t=1}^{T} M_n^{\alpha}(t)$ будем считать непустым с достоверностью $\alpha$. Взяв затем объединение всех таких непустых пересечений, получим множество (не обязательно связное) объемов выборки, на каждом из которых удовлетворяется условие наличия корреляционной связи на выбранном уровне значимости. Далее для определенности берется тот объем выборки $n_{opt}$, для которого достоверность, т.е. отношение $k/(T-m+1)$ выше, а при равных значениях таких отношений — наибольший из объемов.

Минимальное значение $\delta(\alpha)$, при котором вышеописанное объединение пересечений не пусто, и будет представлять абсолютную ошибку в определении корреляции, а величина $1-\delta(\alpha)$ — наименьшую из возможных оценок максимума корреляции. Относительная ошибка в определении максимума корреляции будет тогда не больше, чем $\delta/(1-\delta)$.

Устойчивость корреляционной связи по отношению к объему выборки характеризуется достоверностью $\alpha$, которую желательно сделать как можно больше. Однако с увеличением $\alpha$ начинает возрастать и минимальное $\delta(\alpha)$, так что оптимальным следует считать такое значение $\alpha$, при котором

$$U^2=\delta^2+(1-\alpha)^2 \to \min. \tag{3.8}$$

Подчеркнем, что критерий (3.8) не обобщает критерий (3.6), а уточняет объем, по которому следует вычислять выборочную корреляцию. Величина критерия $U$ не является уровнем значимости корреляционной связи, т.е. его числовое значение не имеет самостоятельного смысла, а важность имеют лишь аргументы, при которых этот критерий минимален. Доверительная вероятность того, что корреляция при выбранном объеме $n_{opt}$ принадлежит промежутку $[1;1-\delta(\alpha)]$, определяется выборочным распределением $F(R,n_{opt})$, и по построению равна $\alpha$.

Описанный подход может быть обобщен для определения наиболее вероятной корреляции между двумя рядами. Зададим некоторое произвольное число $-1<r<1$ и рассмотрим отрезок, его содержащий, шириной $\delta$:

$$\Delta_\delta(r)=[a;b] \subset [-1;1], \quad b-\alpha=\delta. \tag{3.9}$$

Для каждого $r$ рассматриваются те, и только те значения $R(n,t)$, которые попали в отрезок $\Delta_\delta(r)$. Пусть $\{n_i(t)\}$ — соответствующие значения объемов выборок. Корреляцию $R$ между двумя рядами на промежутке времени $[1;T]$ называем достоверной по объему выборки на уровне $\beta$ в промежутке $\Delta_\delta(r)$, если пересечение $\bigcap\limits_{t=1}^{T} M_n^{\beta}(t)$ не менее чем $[\beta(T-n(t)+1)]$ носителей $n(t)$ значений корреляции этих рядов как функции объема выборки, и оно не пусто. Из всех возможных пар $\alpha,\beta$, для которых это пересечение не пусто, выбираются те, которые определяют отрезок (3.9) наименьшей длины. Варьируя затем величину $\beta$, находим минимум критерия (3.8): $U^2(r)=\delta^2+(1-\beta)^2 \to \min$. В результате находим $\beta(r)$, $\delta(\beta)$ и $n_{opt}(r)$. Наиболее достоверной будем считать ту корреляцию, для которой уровень доверия наибольший:

$$r_{opt}=\arg\max \beta(r). \tag{3.10}$$

## 4. Заключение

В работе построена методика определения нестационарной корреляционной связи между двумя рядами, которая формально без изменений обобщается на случай большего числа случайных величин. Введены два критерия, совместная оптимизация которых позволяет найти наилучший объем выборки и доверительный интервал, с наибольшей вероятностью содержащий коэффициент корреляции.

Для краткости изложения мы провели оптимизацию для фиксированного лага между рядами, поскольку методика не зависит от этого параметра. Если требуется определить также и наиболее достоверный лаг, то, параметризуя результат (3.10), находится максимальное значение из наибольших уровней доверия, которому и отвечает этот лаг.

Примеры эффективного применения методики анализа нестационарных корреляций с переменным лагом к задачам макроэкономического анализа в нефтегазовой сфере содержится в [4]–[5]. В последующих работах описанная методика будет применена для определения корреляционной матрицы для многофакторных задач экологического и эпидемиологического мониторинга и прогнозирования.

Работа выполнена при поддержке гранта РФФИ №11–01–00444–а

## Список литературы

1. Клочкова Л. В., Орлов Ю. Н., Тишкин В. Ф., “Математическое моделирование корреляции эпидемической обстановки в мегаполисах от состояния воздуха”, *Журнал Средневолжского математического общества*, 2012, № 3, 34–43.

2. Орлов Ю. Н., Осминин К. П., *Нестационарные временные ряды: методы прогнозирования с примерами анализа финансовых и сырьевых рынков*, Эдиториал УРСС/Книжный дом "ЛИБРОКОМ", М., 2011, 384 с.

3. Королюк В. С., Портенко Н. И., Скороход А. В., Турбин А. Ф., *Справочник по теории вероятностей и математической статистике*, Наука, М., 1985, 640 с.

4. Вовк В. С., Новиков А. И., Глаголев А. И., Орлов Ю. Н., Бычков В. К., Удалов В. А., *Мировая индустрия и рынки сжиженного природного газа: прогнозное моделирование*, ООО "Газпром экспо", М., 2009, 312 с.

5. Абрамов С. Э., Босов Д. Б., Орлов Ю. Н., Першуков В. А., “Некоррелированность цен на рынках СПГ Европы, США и Юго-Восточной Азии”, *Газовая промышленность*, 2010, № 5, 10–14.

# Methodology of correlation coefficient calculation for non-stationary time series

© D.S. Kirillov[5], L.V Klochkova[6], J.H. Orlov[7], V.F. Tishkin[8]

**Abstract.** In this paper we construct a method of determination of correlation coefficient for two non-stationary time series. In comparison with stationary case we need to determine the so-called optimal set length and confidence interval as empirical statistics.

**Key Words:** non-stationary correlation coefficient, optimal set length, optimal confidence level.

[5] Postgraduate student of the Institute of applied mathematics by name M.V.Keldysh of RAS, Moscow; ov3159fd@yandex.ru.

[6] Senior Research Fellow of Keldysh Institute of Applied Mathematics of RAS, Moscow; klud@imamod.ru.

[7] Senior Researcher Officer of the Institute of applied mathematics by name M.V.Keldysh of RAS, Moscow; ov3159fd@yandex.ru.

[8] Deputy Director of Keldysh Institute of Applied Mathematics of RAS, Moscow; v.f.tishkin@mail.ru.