EESTI NSV TEADUSTE AKADEEMIA TOIMETISED. 27. KÖIDE
BIOLOOGIA. 1978, NR. 4

ИЗВЕСТИЯ АКАДЕМИИ НАУК ЭСТОНСКОЙ ССР. ТОМ 27
БИОЛОГИЯ. 1978, № 4

https://doi.org/10.3176/biol.1978.4.04

Велло КАСК

УДК 631:466

ОБ ИСПОЛЬЗОВАНИИ МЕТОДА КОРРЕЛЯЦИИ ПРИ АНАЛИЗЕ РЯДОВ ДИНАМИК

Метод корреляционного анализа впервые был применен К. Пирсоном (Pearson, 1903; Pearson, Lee, 1903) при исследовании проблем наследственности. В дальнейшем этот метод нашел применение в экономике, сельском хозяйстве, естественных и общественных науках, а в настоящее время почти во всех областях науки и техники. Вначале для определения тесноты связи использовался в основном простой метод корреляции.

Во многих случаях, когда пары данных накапливаются во временной последовательности (например, в биологии, медицине, экономике, астрономии и т. д.), исследователь имеет дело с рядами динамик. В начале и в середине 20-х годов ученые не имели представления о всех проблемах, связанных с выборочным значением рядов динамик, и после выхода работы Дж. Э. Юла (Yule, 1926) о «бессмысленной корреляции» пришли к выводу, что корреляционный метод как таковой неприменим к рядам динамик. Сам Дж. Э. Юл совершенно отрицательной позиции в этом отношении не занимал, а говорил лишь об определенных условиях, при которых корреляция между рядами динамик может ввести в заблуждение. Это не исключало существования других условий, при которых корреляция между рядами динамик могла иметь то обычное значение, какое она имела в выборке (Езекиэл, Фокс, 1966).

Корреляционная модель предполагает случайное распределение данных, т. е. данные со значением ниже и выше среднего могут следовать друг за другом. В рядах динамик такое условие обычно не соблюдается, если последовательные моменты времени (часы, месяцы, годы) рассматриваются как последовательные наблюдения. Кроме того, во многих рядах динамик обнаруживается автокорреляция, т. е. имеет место корреляция между рядом стоящими членами, которая существенно отличается от нуля. Это означает, что основные условия простого отбора данных не соблюдались, и возникает сомнение в реальности значений коэффициентов корреляции между двумя такими рядами динамик.

Учитывается ли, однако, в настоящее время такое своеобразие корреляционного анализа? К сожалению, надо признаться, что очень часто в трудах по сельскому хозяйству, биологии, медицине, а также астрономии возможность автокорреляции не учитывается. Такого рода ошибки могли быть простительными в 20—30-х гг., а сейчас они, безусловно, недопустимы.

В настоящее время использование вычислительных центров при переработке экспериментальных данных в условиях, когда отсутствует контакт между программистом и исследователем, привело к тому, что исследователь, не зная сущности статистических методов, полученный из вычислительного центра результат (коэффициент корреляции) принимает за неопровержимую истину, поскольку достоверность связи доказана на ЭВМ. На самом деле, в вычислительном центре для анализа материала по некомпетентности экспериментатора могла быть использована неверная модель, приводящая к ложным выводам.

Особенно часто ложные выводы делаются при исследовании связей между двумя периодическими процессами и тогда, когда один переменный процесс увеличивается или уменьшается во времени постоянно, а другой имеет периодический характер, и если при вычислении не элиминируется фактор времени.

Таким образом, чтобы продемонстрировать, какая связь может существовать между двумя случайно взятыми процессами, не имеющими ничего общего, кроме парных последовательных наблюдений, приведем несколько примеров.

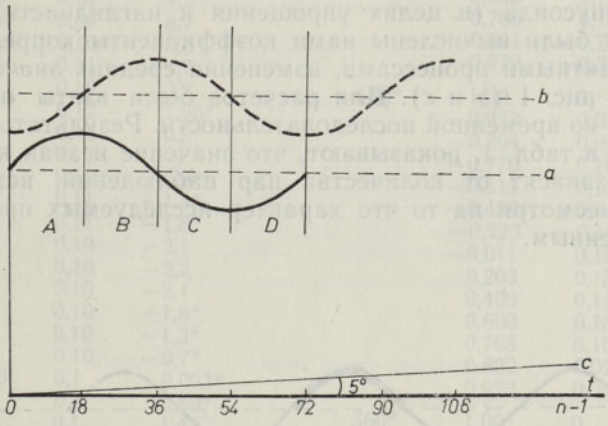

Рис. 1.

С помощью ЭВМ нами изучалось изменение коэффициента корреляции в условиях, когда оба рассматриваемых процесса имели периодический характер (рис. 1*a, b*), а средние величины этих процессов в течение длительного времени оставались постоянными. В принципе точки этих двух исследуемых совокупностей могут быть точками двух параллельно идущих синусоид. Если для корреляционного анализа пары точек наблюдений взять во временной последовательности, то значение и знак коэффициента корреляции будут зависеть от наблюдаемого промежутка времени. В действительности многие явления в определенном промежутке времени наблюдения имеют криволинейный характер, частным случаем которого может быть синусоида. При выяснении связей между такими явлениями, как, например, годовая температура воздуха, почвы и воды, годовой световой режим, режим влажности в почве, уровень воды в реках и озерах, количество вредителей, заболеваемость инфекционными болезнями во время эпидемий и другие, исследователь, используя обычный метод корреляционного анализа, может прийти к ложным выводам, поскольку средние значения этих процессов в рассматриваемом промежутке времени не являются постоянными (увеличиваются или уменьшаются). Сказанное нами хорошо иллюстрируется коррелятивной связью, определенной между процессами *a* и *b* (рис. 1). На рис. 1 приведены 3 изменяющихся во времени процесса:

a и b — периодические процессы с одинаковой частотой, но сдвинутые в фазах (две синусоиды параллельные оси абсцисс), c — процесс, среднее которого постоянно увеличивается.

Анализ рис. 1 показывает, что значение и знак коэффициента корреляции между процессами a и b зависят от изучаемого промежутка времени. Если пары данных взяты в промежутке времени A, то корреляционная связь будет противоположной связи между парами данных, взятыми в промежутке времени B. Такое же противоречие имеет место и в том случае, когда из изучаемых процессов один периодичен, а другой — постоянно увеличивается или уменьшается. Для примера проследим изменение корреляционной связи между процессами b и c (рис. 1). В данном случае среднее значение процесса b имеет характер синусоиды, идущей параллельно оси абсцисс $\left[b=10+\sin\left(i\frac{\pi}{36}-\frac{\pi}{2}\right)\right]$, а среднее значение процесса c постоянно увеличивается ($c=0{,}0875\cdot i$). Так как абсолютно постоянными во времени считаются только константы, то любой процесс, или явление, (грубо говоря) в определенный отрезок времени можно охарактеризовать как прямо- или криволинейный (конечно, с вариацией), частными случаями которых могут быть прямая и синусоида (в целях упрощения и наглядности). На основе сказанного и были вычислены нами коэффициенты корреляции между двумя абстрактными процессами, изменения средних значений которых показаны на рис. 1 (b и c). Для расчетов были взяты пары данных, накопленные во временной последовательности. Результаты вычислений, приведенные в табл. 1, показывают, что значение и знак коэффициента корреляции зависят от количества пар наблюдений, использованных в расчетах, несмотря на то что характер исследуемых процессов оставался неизменным.

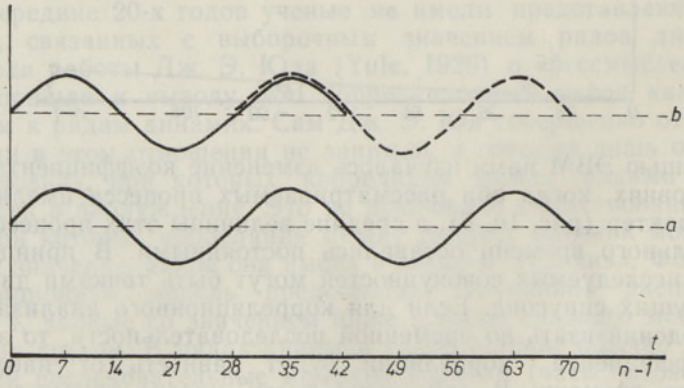

Рис. 2.

Приведем еще один пример. На рис. 2 приведены две находящиеся в одинаковой фазе синусоиды, характеризующие средние значения процессов a и b. На основе 42 пар последовательных наблюдений нами были вычислены коэффициенты корреляции между ними и прослежены изменения последних в случае, когда для каждого последующего вычисления одна синусоида (b) сдвинута в отношении другой на шаг $\pi/14$ радиана (табл. 2).

Как видно из табл. 2, расчеты проведены при сдвиге одной синусоиды до 360° (конечное положение синусоиды изображено пунктиром на рис. 2, b). Значение коэффициента корреляции в данном случае зависит от фазового сдвига между синусоидами и колеблется от -1 до $+1$.

Таблица 1

Коэффициенты корреляции между процессами *b* и *c*

Кол-во пар данных	r	s_r	t
19	0,978	0,05	19,4
22	0,981	0,04	22,6
25	0,985	0,04	27,4
28	0,989	0,03	33,4
31	0,991	0,02	40,9
34	0,993	0,02	47,6
37	0,992	0,02	46,2
40	0,986	0,03	36,9
43	0,974	0,04	27,4
46	0,951	0,05	20,3
49	0,912	0,06	15,2
52	0,852	0,07	11,5
55	0,766	0,09	8,7
58	0,654	0,10	6,5
61	0,521	0,11	4,7
64	0,377	0,12	3,2
67	0,237	0,12	1,96*
70	0,108	0,12	0,9*
75	0	0,12	—
76	−0,088	0,12	−0,8*
79	−0,155	0,11	−1,4*
82	−0,201	0,11	−1,8*
85	−0,227	0,10	−2,1
88	−0,233	0,10	−2,2
91	−0,218	0,10	−2,1
94	−0,183	0,10	−1,8*
97	−0,131	0,10	−1,3*
100	−0,068	0,10	−0,7*
103	0,0003	0,1	0,003*
106	0,068	0,1	0,69*
109	0,130	0,1	1,4*

* — коэффициент корреляции не достоверен.

Таблица 2

Коэффициенты корреляции между двумя синусоидами, параллельными оси абсцисс, в зависимости от фазового сдвига

Фазовый сдвиг	r	s_r	t
0°	1,00	0	∞
	0,973	0,04	26,1
	0,893	0,71	12,6
	0,770	0,10	7,6
	0,611	0,12	4,9
	0,426	0,14	3,0
	0,222	0,15	1,4*
	0,015	0,16	0,07*
	−0,203	0,15	−1,3*
	−0,409	0,14	−2,8
	−0,600	0,12	−4,7
	−0,765	0,10	−7,5
	−0,891	0,72	−12,4
	−0,972	0,37	−26,4
180°	−1,00	0	−∞
	−0,973	0,04	−26,5
	−0,894	0,07	−12,6
	−0,770	0,10	−7,64
	−0,611	0,13	−4,9
	−0,426	0,14	−3,0
	−0,223	0,15	−1,4*
	−0,011	0,15	−0,07*
	0,203	0,15	1,3*
	0,409	0,14	2,8
	0,600	0,13	4,7
	0,765	0,10	7,5
	0,892	0,07	12,4
	0,972	0,04	26,4
360°	1,00	0	∞

* — коэффициент корреляции не достоверен.

Приведенные выше примеры очень хорошо, хотя и несколько грубо, позволяют объяснить целый ряд «достоверных» корреляций между двумя явлениями. Противоречивые выводы получаются тогда, когда хотя бы один из процессов цикличен, и исследователи используют в одних случаях промежуток времени с увеличивающимся, а в других с уменьшающимся средним признака. Типичным примером такого рода связей могут служить связи, найденные в гелиобиологии, так как солнечная активность явно цикличный процесс, имеющий циклы разной длины, в том числе 11-летний и вековой. «Глайсбергу удалось проследить эти циклы (по полярным сияниям) до III века нашей эры [17]. Средняя продолжительность векового цикла оказалась по данным Глайсберга равной 79 годам, но эта величина имеет очень большой разброс: число 11-летних циклов, образующих вековой цикл, колеблется от 5 до 11» (Оль, 1967; с. 75).

Большинство исследователей в связи с недостатком данных на несколько столетий так или иначе используют в своих работах часть векового цикла, и часто из-за неправильной трактовки корреляционного анализа получают важные, но противоречивые результаты. Таким путем

выявляются «прямые связи», которых скорее всего вообще не существует.

Приведем несколько примеров. На основе данных основателя гелиобиологии А. Л. Чижевского (1973; табл. 12, 13) нами были вычислены коэффициенты корреляции между заболеваемостью гриппом по кварталам с 1923 по 1929 гг. и средними числами Вольфа (солнечная активность) также по кварталам за этот же период. Вычисленный коэффициент корреляции оказался достоверным ($0{,}42 \leqslant r \leqslant 0{,}76$) при любых сочетаниях числа Вольфа с заболеваемостью в пределах трех лет (сдвиг налево и направо от соответствующего квартала на 1,5 года). Далее, для элиминирования временного фактора был определен коэффициент частной корреляции; он оказался незначимым ($r = -0{,}07$). Таким образом, на основе данных, приведенных в табл. 3 и 4 (Чижевский, 1973; табл. 12 и 13), связь между солнечной активностью и заболеваемостью гриппом обнаружить не удалось.

Таблица 3

Средняя месячная заболеваемость гриппом

Кварталы	Годы						
	1923	1924	1925	1926	1927	1928	1929
I	87058	158593	345106	596824	589306	778098	1160472
II	56727	121427	189115	423837	431959	487386	369093
III	57836	125556	168663	212911	246243	302494	—
IV	167982	212553	240659	305471	356069	476235	—

Таблица 4

Средние числа Вольфа — Вольфера

Кварталы	Годы						
	1923	1924	1925	1926	1927	1928	1929
I	3,1	2,5	15,6	68,0	80,2	80,8	61
II	6,1	18,7	30,6	58,7	77,6	83,0	—
III	5,7	24,2	35,5	58,5	58,5	90,5	—
IV	8,1	21,5	75,4	70,5	54,6	53,6	—

Примерами противоречивых результатов могут служить данные о зависимости урожая от солнечной активности, а также выводы разных авторов о корреляции непосредственно между заболеваниями сердечно-сосудистой системы и солнечной активностью (Цимахович, 1973). Таких случаев неправильной трактовки коэффициента корреляции можно привести много не только из гелиобиологии, но и из других отраслей науки.

Надеемся, что настоящая статья в какой-то мере облегчит понимание появления возможных ошибок при использовании и интерпретации корреляционного анализа, а тем самым сэкономит в исследовательской работе время и средства.

ЛИТЕРАТУРА

Езекиэл М., Фокс К. Методы анализа корреляций и регрессий. М., 1966.
Оль А. И. О фазах векового цикла солнечной активности. — Солн. данные, 1967, № 9.
Чижевский А. Л. Земное эхо солнечных бурь. М., 1973.
Цимахович Н. П. Исследования по проблеме «солнце-биосфера» в СССР (краткий обзор). — Бюл. Радиоастрофизической обсерват., АН ЛатвССР, Рига, 1973, I.
Pearson, K. The law of ancestral heredity. — Biometrica, 1903, v. 11, p. 211—236.
Pearson, K., Lee, A. On the laws of inheritance in man. I. Inheritance of physical characters. — Biometrica, 1903, v. 11, p. 357—468.
Yule, G. U. Why do we sometimes get nonsense correlations between time-series? — J. Roy. Statistic. Soc., 1926, v. 89, N 1, p. 1—64.

Институт экспериментальной биологии
Академии наук Эстонской ССР

Поступила в редакцию
6/III 1978

Vello KASK

KORRELATSIOONMEETODI KASUTAMISEST AEGRIDADE ANALÜÜSI PUHUL

Resümee

Kasutades korrelatsioonikoefitsienti abstraktsete protsesside vahelise seose tugevuse kriteeriumina, on uuritud selle seose tekke võimalikke põhjusi ja selliste koefitsientide alusel tehtud järelduste tõesust. On näidatud, et kui kahest protsessist vähemalt ühel on perioodiline iseloom ja andmed, mille alusel hiljem tehakse korrelatsioonanalüüs, kogutakse ajalises järjestuses paaridena, siis sõltub saadud korrelatsioonikoefitsiendi väärtus analüüsitavast ajavahemikust. Korrelatsioonikoefitsiendi märk sõltub sellest, millises suunas antud vahemikus toimub perioodilise protsessi muutus.

Eesti NSV Teaduste Akadeemia
Eksperimentaalbioloogia Instituut

Toimetusse saabunud
6. III 1978

Vello KASK

ON THE USE OF THE CORRELATION METHOD IN THE TIME-SERIES ANALYSIS

Summary

Using the correlation coefficient as a criterion for determining the strength of the connection between abstract processes, probable reasons of formation of such connection were studied, as well as the statistical significance of the conclusions made on the basis of these correlation coefficients.

Academy of Sciences of the Estonian SSR,
Institute of Experimental Biology

Received
March 6, 1978