MỘT SỐ LƯU Ý KHI SỬ DỤNG TƯƠNG QUAN PEARSON

TS. **Nguyễn Đăng Quang**
Trung tâm Dự báo khí tượng thủy văn Trung ương

Bài báo này nêu ra một số lưu ý khi sử dụng hệ số tương quan Pearson. Dựa trên một số chuỗi số liệu giả lập, chúng tôi đã chỉ ra ảnh hưởng của các giá trị quan trắc bất thường tới chất lượng của hệ số Pearson. Chúng tôi cho rằng hiển thị số liệu theo dạng đồ thị là một trong những cách thức đơn giản và hiệu quả để khảo sát chuỗi số liệu trước khi tính toán hệ số Pearson.

1. Đặt vấn đề

Hệ số tương quan Pearson là một trong những trị số được sử dụng phổ biến nhất trong khí tượng thủy văn để xác định mối liên hệ tuyến tính giữa hai chuỗi số liệu. Ví dụ, ta muốn xác định mối liên hệ giữa chuỗi 40 năm (1965-2014) số liệu nhiệt độ bề mặt đất tại trạm Phù Liễn với chuỗi số liệu tương ứng tại Hà Nội, công thức tính hê số tương quan Pearson giữa Phù Liễn và Hà Nội được biểu diễn như sau:

$$r_{xy} = \frac{COV(x,y)}{s_x s_y} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

trong đó x, y tương ứng với chuỗi nhiệt độ tại Phù Liễn và Hà Nội.

Hệ số tương quan giữa x và y, kí hiệu là r_{xy}, cho ta biết mức độ phụ thuộc tuyến tính giữa x và y. Hệ số này được tính bằng tỉ số giữa hiệp phương sai của hai biến x, y và tích của độ lệch chuẩn sx sy của chúng; n là độ dài của chuối số liệu, trong ví dụ nêu trên thì n = 40.

Hệ số tương quan r_{xy} dao động trong khoảng [-1, 1]. Trị số tuyệt đối của hệ số càng lớn thì tương quan giữa hai chuỗi số liệu càng lớn và ngược lại. Nếu hai biến là độc lập thống kê thì hệ số tương quan r_{xy} bằng 0 [1].

Trong phạm vi bài viết này, chúng tôi đưa ra một số lưu ý khi thao tác, tính toán hệ số tương quan Pearson trên các chuỗi số liệu. Những lưu ý này được nhà toán học người Anh Francis John Frank Anscombe mô tả [2]. Nay chúng tôi giả lập một số chuỗi số liệu khí tượng để minh họa một cách tường minh hơn.

2. Tiến hành thử nghiệm

Chuỗi số liệu giả định chuẩn sai nhiệt độ tháng 1 (^{0}C) tại 6 trạm được dẫn ra trong bảng 1.

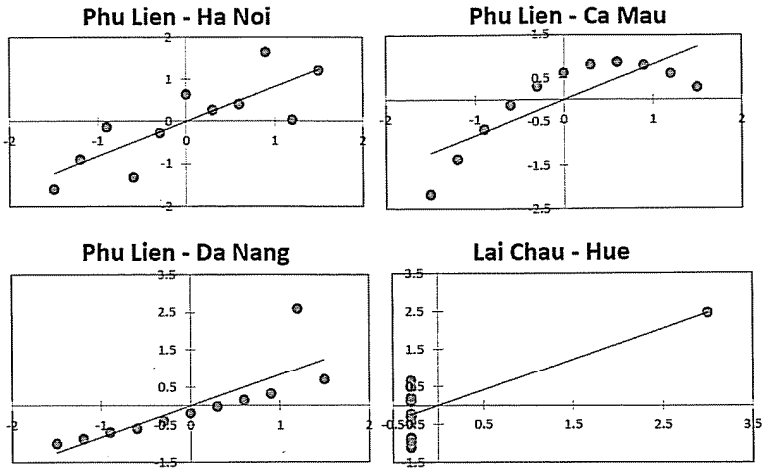
Sáu chuỗi số liệu giả định này tuy có những đặc trưng thống kê rất giống nhau, nhưng chúng lại trở nên rất khác nhau khi hiển thị trên đồ thị. Điều này thể hiện tầm quan trọng của việc kiểm tra số liệu trước khi tìm kiếm mối tương quan tuyến tính giữa các tập số liệu và ảnh hưởng của các số liệu quan trắc bất thường (hàm ý là sai số) tới kết quả tính toán.

Đặc trưng thống kê của số liệu giả định

Theo bảng 1, tất cả các chuỗi số liệu giả định này (Phù Liễn, Hà Nội, Cà Mau, Đà Nẵng, Lai Châu và Huế) đều có giá trị trung bình là 0 và độ lệch chuẩn là 1. Hệ số tương quan giữa Phù Liễn - Hà Nội, Phù Liễn - Cà Mau, Phù Liễn - Đà Nẵng, Lai Châu – Huế đều bằng 0,82 (làm tròn đến 0,01). Đường hồi quy tuyến tính của từng cặp quan hệ nêu trên có chung dạng: y = 0,82 * x – 0,0009.

Bảng 1. Chuẩn sai nhiệt độ tháng 1 tại 6 trạm quan trắc trong thời kì 2004-2014

Tháng 1	Phù Liễn	Hà Nội	Cà Mau	Đà Nẵng	Lai Châu	Huế
2004	0,3	0,27	0,81	-0,02	-0,3	-0,45
2005	-0,3	-0,27	0,31	-0,4	-0,3	-0,86
2006	1,2	0,04	0,61	2,6	-0,3	0,1
2007	0	0,64	0,62	-0,19	-0,3	0,66
2008	0,6	0,41	0,87	0,15	-0,3	0,48
2009	1,5	1,21	0,29	0,7	-0,3	-0,23
2010	-0,9	-0,13	-0,67	-0,7	-0,3	-1,11
2011	-1,5	-1,6	-2,17	-1	3	2,46
2012	0,9	1,64	0,8	0,32	-0,3	-0,96
2013	-0,6	-1,32	-0,12	-0,6	-0,3	0,2
2014	-1,2	-0,9	-1,36	-0,87	-0,3	-0,3



Hình 1. Đồ thị biểu diễn quan hệ giữa các chuỗi số liệu

3. Nhận xét

Đồ thị minh họa mối quan hệ nhiệt độ Phù Liễn – Hà Nội thể hiện mối quan hệ tuyến tính đơn giản, ứng với hai chuỗi số liệu tuân theo phân bố chuẩn. Đồ thị Phù Liễn – Cà Mau thể hiện mối quan hệ đa thức bậc 2 giữa hai chuỗi số liệu, chúng không tuyến tính, và do đó không thể sử dụng cách tính hệ số tương quan Pearson cho cặp số liệu này. Ở đồ thị thứ 3, đồ thị Phù Liễn – Đà Nẵng, ta thấy số liệu năm 2006 của trạm Đà Nẵng là rất bất thường, nhiều khả năng ẩn chứa sai số. Sự xuất hiện của số liệu Đà Nẵng năm 2006 đã làm giảm hệ số tương quan của cả chuỗi. Nếu loại bỏ số liệu này, hệ số tương quan Pearson thu được là 0,998 (so sánh với giá trị 0,82 khi chưa loại bỏ số liệu). Minh họa này cho ta thấy ảnh hưởng rất lớn của trị số quan trắc bất thường tới tương quan Pearson.

Đồ thị cuối cùng minh họa quan hệ giữa hai trạm Lai Châu và Huế cũng thể hiện tác động của trị số quan trắc bất thường. Mặc dù hai chuỗi số liệu Lai Châu và Huế là độc lập, nhưng chỉ với sự xuất hiện của một trị số bất thường (trạm Lai Châu năm 2011) trong chuỗi số liệu thì cũng thiết lập được một hệ số tương quan khống là 0,82.

4. Kết luận

Để xác định mối tương quan tuyến tính giữa hai tập số liệu khí tượng thủy văn một cách chính xác, một trong những thao tác đầu tiên mà chúng ta nên thực hiện đó là hiển thị các chuỗi số liệu theo dạng đồ thị, nhằm tìm ra các giá trị quan trắc bất thường, hoặc các phân bố phi tuyến. Các giá trị bất thường sẽ được loại bỏ nếu xác định được đó là trị số sai. Cũng như vậy, nếu chuỗi số liệu không tuân theo luật tuyến tính thì hệ số tương quan Pearson không được sử dụng trong các trường hợp này.

Tài liệu tham khảo

1. Wilks, D. S. (2011), Statistical Methods in the Atmospheric Sciences, 3rd edn, Academic Press, pp704.

2. Anscombe, F. J. (1973), Graphs in Statistical Analysis, American Statistician, 27 (1): 17–21. JSTOR 2682899.