ỨNG DỤNG HÀM SPLINE ĐỂ XỬ LÝ CÁC QUAN HỆ TƯƠNG QUAN TRONG KHÍ TƯỢNG THỦY VĂN

PTS. Nguyễn Hữu Khải
KS. Lê Xuân Cầu
Trường cán bộ KTTV Hà Nội

Các qúa trình khí tượng thủy văn có cấu trúc không gian phức tạp, khi nghiên cứu nó thường phải dùng phương pháp thống kê xét mối tương quan giữa chúng. Trong số đó phương pháp tương quan tuyến tính nhiều chiều thường được dùng nhất. Tuy nhiên, khi mối quan hệ là phi tuyến thì phương pháp trên kém hiệu qủa, có thể khắc phục điều này bằng cách sau:

1. Biến đổi các biến số sao cho liên hệ giữa chúng là tuyến tính, như Alếchxâyev G. A. đã thực hiện [1]. Phương pháp này chỉ dùng khi liên hệ giữa các đại lượng đơn điệu.

2. Hồi quy tuyến tính từng bước: các nhân tố được phân chia thành số lượng tối ưu các nhóm, trên mỗi nhóm xây dựng hàm hồi quy tuyến tính [3].

Trong những năm gần đây, khi phân tích các quan hệ nhiều chiều thường dùng hàm spline và đối với các bài toán KTTV thường dùng hàm spline bậc 3 (spline3).

1. Khái niệm về spline

Lý thuyết về spline được nhiều các tác giả nghiên cứu [1, 2, 3, 4,]. Spline xuất hiện tự nhiên trong các bài toán cơ học. Chẳng hạn dầm đàn hồi với tải trọng điểm có dạng spline.

Các quan hệ trong KTTV thường là trơn, do vậy thích hợp nhất là áp dụng spline đa thức. Spline loại này được Schoenberg I.J. [4] nghiên cứu và phát triển. Có thể hiểu spline một cách tóm tắt như sau:

Chia đoạn [a, b] ra N phần bởi các điểm chia:

$$\omega_N = \{a = x_1 < x_2 < \ldots < x_N = b\} \qquad (1)$$

Hàm $\varphi(x)$ mà trên mỗi đoạn $\{X_j, X_{j+1}\}$; $j = 1,2, \ldots, N$ được biểu thị bằng đa thức p(x) bậc m gọi là spline đa thức bậc m với N điểm liên kết tại các nút.

Các hệ số của đa thức p(x) phù hợp và tương ứng với nhau sao cho đảm bảo điều kiện liên tục của hàm và đạo hàm bậc (m-1) của chúng tại các nút liên kết, nghĩa là:

$$\varphi^{(k)}(x_1 - 0) = p_{j-1}^{(k)}(x_j) = p_j^{(k)}(x_j) = \varphi^{(k)}(x_{j+1}) \qquad (1^*)$$

$$j = 1, 2, \ldots, N; \quad k = 0,1, \ldots, N$$

ω_n được gọi là lưới và các điểm chia x_j được gọi là các nút của spline.

Như vậy, hàm spline φ (x) có (m - 1) lần vi phân liên tục trên toàn đoạn [a, b] và trên mỗi đoạn $[x_i, x_j]$ là một đa thức bậc m. Nói riêng hàm spline bậc 3 là hàm hai lần vi phân liên tục hình thành từ các mẩu riêng biệt của parabol bậc 3. Spline bậc 3 thường được biểu thị dưới dạng:

$$\varphi(x) = S_3(x) = a_j + b_j (x-x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3 \qquad (2)$$

$$x_j < x < x_{j+1} ; j = 1, 2, ..., N$$

Như vậy, hàm $S_3(x)$ được cho bởi 4(N - 1) hệ số. Các hệ số này được xác định từ điều kiện liên kết (1*) cho các parabol (2) trên lưới (1).

Các hệ số của spline có ý nghĩa như sau:

a_j - giá trị $S_3(x)$ tại các nút lưới (1)

b_j - giá trị các đạo hàm bậc 1

c_j - giá trị các đạo hàm bậc 2

d_j - giá trị các đạo hàm bậc 3

2. Ứng dụng của spline

Spline và đặc biệt spline bậc 3 được áp dụng trong nhiều ngành khoa học và lần đầu tiên được Vager B. G. [2] áp dụng cho khí tượng thủy văn.

a) Nội suy bằng spline bậc 3

Trước hết spline được ứng dụng cho bài toán nội suy. Nếu trên lưới (1) ta cho các giá trị y_i được xác định bởi một hàm f(x) nào đó, nghĩa là $y_i = f(x_i)$. Một hàm $\varphi(x)$ sao cho $\varphi(x_i) = y_i$ gọi là nội suy của f(x), tức là một gần đúng của f(x), nói một cách khác φ nội suy các giá trị $\{x_i, y_i\}$. Nếu trên lưới (1) cho các giá trị gần đúng y_i của hàm f, nghĩa là $y_i \neq f(x_i)$ thì bài toán xác định hàm φ gọi là xấp xỉ gần đến f theo số liệu $\{x_i, x_j\}$. Khi đó cần lưu ý đến tính tiệm cận của φ đến f. Nói riêng khi sai số chứa trong y_i là đủ nhỏ, có thể chấp nhận gần đúng $\varphi(x_i) = y_i$ và khi đó bài toán đưa đến bài toán nội suy và $\varphi(x)$ là dạng đặc trưng mong đợi của f(x). Xấp xỉ bằng phương pháp bình phương bé nhất xác định được loại hàm $\varphi(x)$. Tuy nhiên, khó tìm được loại hàm $\varphi(x)$ gần f(x). Spline bậc 3 là một công cụ có hiệu qủa để giải quyết vấn đề này. Cho rằng trên các nút của lưới (1) xác định được các số thực $\{y_i\}$. Từ tất cả các hàm g(x) trên [a, b] có đạo hàm bậc 2 liên tục và $g(x_i) = y_i$ thì spline bậc 3 mà nó có $S_3''(a) = S_3''(b) = 0$ làm cực tiểu tích phân sau:

$$I(g) = \int_a^b |g''(x)|^2 dx$$

$$I(\varphi) = \min \qquad (3)$$

Tích phân (3) là độ đo mức trơn của hàm g(x). Giá trị tích phân càng nhỏ thì độ trơn của hàm càng lớn. Từ tất cả các hàm tiến tới f(x) thì spline bậc 3 là hàm trơn nhất.

b) Làm trơn bằng spline bậc 3

Vấn đề làm trơn bằng spline bậc 3 được quan tâm nhiều hơn vì trong thực tế các quan hệ hồi quy cần được thiết lập.

Chúng ta xuất phát từ trường hợp đơn giản nhất. Giả thiết có một dãy quan trắc đồng thời hai biến x_i và y_i, giữa chúng có mối quan hệ $y = f(x)$, nhưng dạng cụ thể của nó là chưa biết, chỉ biết các giá trị cụ thể của nó kèm theo sai số, nghĩa là có dãy $\{x_i, y_i\}$, $i = 1,2, ..., n$ và $\{x_i\} \in [a, b]$.

Nếu xem xét dãy $\{x_i\}$ và $\{y_i\}$ như là khai triển của đại lượng ngẫu nhiên X và Y. Gọi $\theta(x)$ là kỳ vọng toán học có điều kiện của đại lượng ngẫu nhiên y khi X nhận giá trị x_i

$$y = \theta(x) = M(Y/X = x) \tag{4}$$

Phương trình (4) là phương trình hồi quy của Y theo X

Giả sử rằng trị số quan trắc y_i được tạo thành từ tổng của thành phần tất định và sai số ∂_i của nó:

$$y_i = \varphi(x_i) + \partial_i \tag{5}$$

Giả thiết thêm rằng hàm $\varphi(x)$ trên [a, b] là đủ trơn, còn sai số ∂_i không tương quan và có trung bình bằng không. Thành phần này không quan trắc được, nó là một đại lượng lý thuyết.

Có thể lặp lại các quan trắc một số tùy ý lần và nhận được một tập quan trắc mới. Khi đó hàm $\theta(x)$ được gọi là trend (khuynh hướng), và rõ ràng trend $\theta(x)$ và $\varphi(x)$ từ (4) là như nhau. Nhiệm vụ của chúng ta là tìm hàm f(x), gần đúng tốt nhất của $\theta(x)$ trên đoạn [a, b] theo tài liệu quan trắc đã có.

Trên lưới (1) xác định dãy y_i, nghĩa là tạo thành dãy các cặp số hạng $\{y_i, x_i\}$, $i = 1, ..., n$. Và $x_i \in [a, b]$. Gọi W_2^m là tập hợp các hàm có đạo hàm bậc $(m - 1)$ liên tục trên [a, b]. Trong tất cả các hàm $g(x) \in W_2^m[a, b]$ yêu cầu tìm được hàm f(x) trên đó làm cực tiểu hàm sau:

$$I(g) = \sum_{i=1}^{n} \left(g(x) - y_i\right)^2 + p\int_a^b \left(g^m(x)\right)^2 dx$$

$$I(f) = \min \tag{6}$$

Schoenberg chỉ ra rằng nghiệm bài toán này tồn tại và chính là spline bậc $(2m - 1)$ trên lưới (1), nghĩa là $f(x) = S_{2m-1}(x, \omega_n)$. Khi p nhỏ thì số hạng đầu là cơ bản và giá trị hàm f(x) khi đó rất gần y_i.

Tìm nghiệm từ (6) khi m lớn là khó khăn. Vì vậy, thường giới hạn xây dựng spline làm trơn với số bậc không lớn lắm (m=2). Spline bậc 3 đảm bảo độ chính xác chấp nhận được và thỏa mãn điều kiện (6).

c) Xây dựng hồi quy nhiều chiều bằng spline bậc 3 trung bình

Các mô hình hồi quy được trình bày trong rất nhiều tài liệu. Có nhiều phương pháp xây dựng các quan hệ hồi quy nhiều chiều. Tuy nhiên, như trên đã phân tích, hồi quy nhiều chiều bằng spline bậc 3 có ưu thế hơn. Đây là một bài toán riêng khá phức tạp. Ở đây có thể tóm tắt mấy nét đại cương như sau:

Cho rằng đại lượng nghiên cứu được đặc trưng bằng các đại lượng x^1, x^2, ..., y. Chúng ta cần tìm liên hệ giữa biến y với các biến x^j, j = 1,2, ..., k. Nếu liên hệ tồn tại ta có:

$$y = \Phi(x^1, x^2, ..., x^k) = M(Y/x = x_i, x = x_2, ..., x = x_k) \quad (7)$$

nghĩa là các giá trị y_i của đại lượng ngẫu nhiên Y được xác định bởi các giá trị x_i^j của các đại lượng ngẫu nhiên X. Việc xấp xỉ hàm Φ được thực hiện qua các giá trị quan trắc đồng bộ x^j

$$\{y_i, x_i^1, ..., x_i^k\} \quad i = 1,2,..., N \quad (8)$$

Ta thay quan hệ (7) bằng mô hình dạng:

$$y = F_i(x^1) + F_2(x^2) + ...+ F_k(x^k) \quad (9)$$

Xác định các hàm số $F_j(x^j)$ là nội dung bài toán hồi quy. Hàm $F_j(x^j)$ cần được chọn sao cho thỏa mãn một số yêu cầu thể hiện bản chất vật lý của hiện tượng nghiên cứu. Thường thì quan hệ (9) là phi tuyến và do đó chọn các hàm $F_j(x^j)$ là những spline cho hiệu qủa tốt hơn.

3. Xác định hàm hồi quy một chiều bằng spline bậc 3 trung bình

Trong phần này chúng tôi chỉ giới hạn trong việc xác định hàm hồi quy một chiều bằng spline bậc 3 trung bình.

Hàm này có thể biểu thị dưới dạng:

$$\varphi(x) = \sum_{j=1}^{N+4} \alpha_j \varphi_j(x) \quad (10)$$

Trong đó $\varphi_j(x)$ là các spline bậc 3 cơ sở.

$$\varphi_j(x) = M_{ij}\left(\frac{(x_{i+1}-x)^3}{6h_{i+1}}\right) + M_{ij}\frac{(x-x_i)^3}{6h_{i+1}} +$$

$$\left(\varphi_j(x_i) - \left(\frac{M_{ij}h_{i+1}^2}{6}\right)\right)\frac{x_{i+1}-x}{h_{i+1}} + \left(\varphi_j(x_{i+1}) - \left(\frac{M_{ij}h_{i+1}^2}{6}\right)\right)\frac{x-x_i}{h_{i+1}} \quad (11)$$

Ma trận M_{ij} là ma trận đạo hàm bậc 2 của spline bậc 3 cơ sở tại các nút liên kết và hai đầu mút của [a, b].

$$h_{i+1} = x_{i+1} - x_i$$

Để xác định (11) trên đoạn [a, b] cần xác định N + 2 giá trị chưa biết M_{ij}, j = 0,1, ...,N+1. Đối với các đoạn phân chia đều nhau, hệ thống phương trình xác định M_{ij} có dạng:

$$\begin{vmatrix} 2 & 1 & & & \\ \frac{1}{2} & 2 & \frac{1}{2} & & \\ \dots & \dots & \dots & \dots & \\ \dots & \dots & \dots & \dots & \frac{1}{2} \end{vmatrix} \quad M = \left(\frac{6}{H}\right) \quad \begin{vmatrix} -h & -1 & -1 & & & \\ & \frac{1}{2} & -1 & \frac{1}{2} & & \\ & \dots & \dots & \dots & \dots & \\ & & & & 1 & -1 \end{vmatrix} \qquad (12)$$

Ma trận M có kích thước (N + 2). (N + 2)

$$H = h_{i+1} = x_{i+1} - x_i$$

Giải hệ phương trình này ta được:

$$\{M_{ij}\} = \left(\frac{6}{H^2}\right) C \begin{vmatrix} -H & -1 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \dots & & & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -1 & H \end{vmatrix} \qquad (13)$$

với

$$c_{ij} = (B_j b_1, \dots B_j b_j, \alpha_j b_{N+2-j}, \dots \alpha_j b_i) \qquad (14)$$

Ở đây:

$$B_j = \frac{b_{N+3-j}}{b_{j-1} b_{N+3-j} - b_j b_{N+4-j}}$$

$$\alpha_j = \frac{b_j}{b_{j-1} b_{N+3-j} - b_j b_{N+4-j}} \qquad (15)$$

$$j = 1,2,\dots N+2$$

Các giá trị b_j tính bằng công thức trúy hồi:

$$b_{j+i} = -4\, b_j - b_{j-i}$$

$$b_o = -2;\ b_i = 1 \qquad (16)$$

Vì vậy, để tính các giá trị của spline bậc 3 cơ sở với N liên kết trên lưới phân chia đều nhau cần tính ma trận C_{ij} kích thước (N+2)(N+2) theo (14) và ma trận M_{ij} theo (13). Cuối cùng theo (11) xác định được $\varphi_j(x)$.

Tuy nhiên, khi số điểm N lớn thì việc tính toán không thuận lợi. Trong thực tế thường tìm spline bậc 3 trên một lưới hẹp hơn, nhưng vẫn bảo đảm độ chính xác theo yêu cầu:

$$\omega_q = \{a = x^*_1 < x^*_2 < \dots < x^*_q = b\} \text{ với } q \le N \qquad (17)$$

Lưới (17) cần giữ cho các điểm chia trùng với một trong các điểm của lưới (1).

Xây dựng trên lưới (17) spline bậc 3 với q liên kết $\varphi^q(x)$. Lưới (17) được chọn khi bảo đảm quan hệ:

$$|\varphi(x) - \varphi^q(x)| < \varepsilon = \varepsilon_o \tag{18}$$

Trong đó: $$\varepsilon_o = \frac{\sigma_\Delta}{\sqrt{N-1}} \tag{19}$$

với σ_Δ là độ lệch quân phương của hồi quy với N liên kết. Để hình thành lưới (17) ta làm như sau:

- Trước hết cho q = 2 nghĩa là khi ấy có:

$$\omega_2 = \{a = x^*_1 < x^*_2 = b\} \tag{20}$$

Xác định spline bậc 3 $\varphi^2(x)$ trên lưới (20) kiểm tra điều kiện (18) nếu thỏa mãn được thì có spline bậc 3 và lưới cần tìm.

- Nếu không thực hiện được (18) ta lại chia đôi đoạn [a, b] (q = 3) ta có lưới:

$$\omega_3 = \{a = x^*_1 < x^*_2 < x^*_3 = b\} \tag{21}$$

Trong đó $x^*_2 = \frac{a+b}{2}$

Xác định hàm $\varphi^3(x)$ và lại kiểm tra (18). Nếu thỏa mãn thì $\varphi^3(x)$ là hàm cần tìm. Nếu không thì lại chia đôi đoạn của lưới (21) và lại làm như trên. - Qúa trình được lặp lại đến khi nào điều kiện (18) được thỏa mãn.

4. Phân tích quan hệ lưu lượng - mực nước Q = f(H) bằng spline bậc 3 trung bình

Quan hệ Q = f(H) nói chung là phi tuyến. Trong chỉnh biên, việc xác định quan hệ này được thực hiện bằng tay trên cơ sở phân tích sai số quân phương σ. Nhưng cách làm này mang nặng tính chủ quan, khó xác định hồi quy tốt nhất. Mặt khác, rất khó triển khai trên máy tính. Một số tác giả đề nghị khái quát quan hệ này theo dạng [5, 6]:

$$Q = a.H^2 + b.H + c \text{ hay } Q = a_1 . (H - H_0)^m \tag{22}$$

Trong đó a, b, c, a_1, m là các tham số.

Tuy nhiên, các dạng này không bảo đảm sự phù hợp của đường hồi quy trên toàn bộ khoảng dao động của Q = f(H).

Để khắc phục nhược điểm này và để dễ dàng dùng máy tính nên dùng spline bậc 3 trung bình vì khi dùng hàm này ta dễ dàng đạt được điều sau:

Trên mỗi khoảng dao động của H, đường Q = f(H) được gần đúng bằng spline bậc 3. Thêm vào đó, tại các điểm liên kết các mẩu spline liên tục, ta có sai số quân phương σ nhỏ nhất và đây là đường cong trơn nhất, vì nó làm cực tiểu (6).

Kết qủa tính thử cho một số quan hệ Q=f(H) của một số trạm được đưa ra trong bảng 1.

Bảng 1

Đặc trưng	Trạm Bản Giốc (S. Quy Sơn), 1974		Trạm Thanh Sơn (S. Bửa), 1990	
	Theo chỉnh biên	Theo hàm spline	Theo chỉnh biên	Theo hàm spline
Sai số quân phương (%)	3,16	4,11	3,96	3,07
Sai số lớn nhất (%)	12,4	10,75	9,87	7,11
Số điểm đo	13	13	40	42
Số điểm lệch âm	25	26	15	15
Số điểm lệch dương	24	23	16	16
Tổng sai số	+3,51	+0,016	+0,35	-0,0008

Như vậy, có thể thấy rằng spline bậc 3 có ưu thế rõ rệt trong việc phân tích các quan hệ tương quan và nói riêng quan hệ phi tuyến một chiều. Đây là một công cụ có hiệu qủa để chỉnh biên, chỉnh lý số liệu KTTV. Nó cũng có thể mở rộng cho dự báo, tính toán các yếu tố KTTV.

TÀI LIỆU THAM KHẢO

1.Alêcxâyev. G. A. Phương pháp khách quan làm trơn và chuẩn hóa các quan hệ tương quan. NXB KTTV, Leningrat, 1971 (Tiếng Nga).

2. Constanchinov A.P. Ứng dụng spline và phương pháp độ lệch dư trong KTTV. NXB KTTV, Leningrat, (Tiếng Nga).

3. Thuật toán và các chương trình thiết lập các quan hệ. (Dưới sự lãnh đạo Vapnhic B. N), NXB khoa học, Matxcơva, 1984. (Tiếng Nga).

4. Schoenberg I.J. Spline function and the problem of graduation. Pro. Nat. USA. 1964.

5. Karaxev I. F., Sumkov I. G. Đo đạc thủy văn NXB KTTV, Leningrat, 1985. (Tiếng Nga).

6. Mekong Secretariat. Hymos. Delft hydraulics, 1989.

7. Chỉnh biên tài liệu thủy văn trạm Thanh Sơn 1990.

8. Chỉnh biên tài liệu thủy văn trạm Bản Giốc 1974.