SỬ DỤNG PHƯƠNG PHÁP ĐỘ LỆCH DƯ VÀ KHÔNG GIAN SPLINE BẬC 3 CƠ SỞ ĐỂ THIẾT LẬP HỒI QUY PHI TUYẾN NHIỀU CHIỀU

KS. Lê Xuân Cầu
Trường cán bộ KTTV Hà Nội

I- Đặt vấn đề

Các mô hình hồi quy được mô tả nhiều trong các nghiên cứu về toán thống kê [1] còn các ứng dụng của nó trong khí tượng thủy văn nói riêng được trình bầy trong [5, 9, 10, 11]. Phương pháp độ lệch dư được A.P. Constanchinov sử dụng rộng rãi trong các nghiên cứu về khí tượng thủy văn, như nghiên cứu về bay hơi trong thiên nhiên [11], về năng suất lúa mì trong thủy văn nông nghiệp [2, 5] và nhiều vấn đề khác. A.P.Constanchinov dùng phương pháp độ lệch dư để thiết lập mô hình hồi quy chủ yếu bằng vẽ đồ thị, nó đòi hỏi người nghiên cứu phải tốn nhiều công sức và kết quả cuối cùng phụ thuộc vào kinh nghiệm của người vẽ. Gần đây trong [5] ông đã dùng phương pháp độ lệch dư và nội suy spline bậc 3 nhằm đưa mô hình vào máy tính, đó là bước đi đầu tiên nhằm nâng mô hình hồi qui lên một nấc mới. Nó có nhược điểm là nội suy phải lấy các điểm thực nghiệm làm nút lưới nội suy, trong khi đó các điểm thực nghiệm có một số sai số nào đó và phương pháp chọn các spline chưa thật khách quan. Tuy nhiên, ông đã đóng góp nhiều công sức để xây dựng và hoàn thiện mô hình này. Dựa vào ý đồ trên, tác giả đã thiết lập được mô hình hồi quy phi tuyến nhiều chiều với nội dung hoàn toàn mới và hoàn thiện hơn để nghiên cứu các quá trình khí tượng thủy văn. Khi thiết lập hồi quy phi tuyến nhiều chiều tác giả đã dùng nhiều công cụ toán sau đây:

1. Hồi quy phi tuyến một chiều bằng không gian spline bậc 3 cơ sở.
2. Phương pháp độ lệch dư.
3. Phương pháp cực tiểu hóa mạo hiểm trung bình thực nghiệm theo cấu trúc.

Mô hình hồi quy này có tính hiệu quả và tính phổ thông. Nó không những cho kết quả số trị với độ chính xác cao mà còn cho phép phân tích tác động của các nhân tố lên nhân tố ta quan tâm và giải một lớp rộng các bài toán. Tác giả đã lập chương trình tính và sẽ chỉ ra tính hiệu quả bằng ví dụ minh họa ở cuối bài này.

Một hiện tượng nghiên cứu được đặc trưng bởi chuỗi các đại lượng vật lý: Y,X1,X2.....Xn. Ta phải lập một mối quan hệ tương quan giữa Y

và X1,X2....., Xn. Giả sử Y quan hệ với các biến Xj (j= 1,2,...,n) bằng hàm sau:

$$Y = \varphi (X1,X2, ..., Xn)$$

Quan hệ này thường được biểu diễn dưới dạng một tổng các hàm:

$$y = \sum_{j=1}^{N} F_j (X_j) \qquad (1)$$

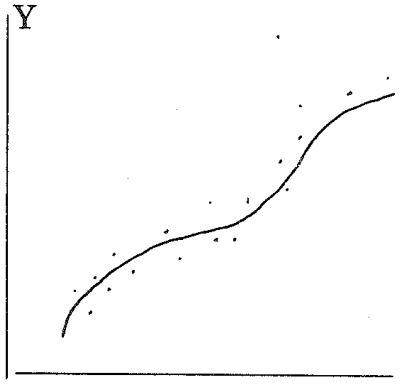
Nếu F_j là hàm tuyến tính bậc nhất theo X_j : $F_j (X_j) = a._j X._j + b_j$ thì khi thạy thế các biểu thức này vào (1) ta có phương trình hồi quy tuyến tính nhiều chiều. Nếu cho rằng F_j là các đa thức hoặc các hàm khác thì ta có phương trình hồi quy phi tuyến nhiều chiều.

Các mối quan hệ tương quan trong khí tượng thủy văn thường phi tuyến và nhiều chiều. Hồi quy phi tuyến sau đây sẽ là một công cụ khá mạnh để nghiên cứu các quá trình khí tượng thủy văn.

II- Xây dựng hồi quy phi tuyến nhiều chiều

1. Hồi quy phi tuyến một chiều bằng spline bậc 3 trung bình

Một quan hệ tương quan giữa biến Y và X như hình 1 là phi tuyến và diễn tả nó bằng hàm spline bậc 3 cơ sở là tốt hơn cả. Điều đó có được vì nó biểu diễn đường cong đi qua trung tâm các điểm thực nghiệm và đạt độ cong trơn lớn nhất trong tập tất cả các đường cong có đạo hàm bậc hai liên tục đi qua trung tâm các điểm thực nghiệm, nghĩa là nó cho sai số quân phương sicma nhỏ và độ cong trơn lớn [12].



Hình 1

Muốn biết mức độ tương quan giữa Y và X ta tính hệ số tương quan giữa hai biến này. Để biết quan hệ giữa chúng có phi tuyến hay không ta tính hệ số tương quan của Y khi nó được sắp xếp theo X tăng hoặc dùng phân tích phương sai để kiểm tra.

Tính cong trơn của đường hồi quy có một ý nghĩa rất quan trọng trong các mô hình hồi quy. Dùng spline ta có thể thiết lập được đường hồi quy có độ cong trơn tùy ý. Trong thực tế, người nghiên cứu từ chối dùng đa thức bậc cao để lập hồi quy giữa các biến vì chúng có các cực trị địa phương và những điểm uốn không phải do bản chất của hiện tượng mà do sự hạn chế độ dài của chuỗi số liệu. Tính chất cong trơn giới hạn khả năng dao động của đường hồi quy trong trường các điểm thực nghiệm, nó điều chỉnh theo hướng chung sự biến đổi của biến Y theo X. Một so sánh quan trọng nữa là khi độ dài chuỗi thực nghiệm tăng vô hạn thì hàm spline bậc 3 trung bình

tiến dần đều tới hàm phải tìm f (x), trong khi đó các hàm đa thức thì không [6]. Nói cụ thể gọi s3 (X,α1) là spline bậc ba trung bình và P (x, α2) là đa thức với α1, α2 là các vectơ tham số, khi đó có:

$$\sup_{\alpha 1 \in \{a,b\}} |f(x) - s3(x,\alpha 1)| \xrightarrow{p} 0 \quad (l \to \infty)$$

$$\sup_{\alpha 2 \in \{a,b\}} |f(x) - p(x,\alpha 2)| \xrightarrow{p} 0 \quad (l \to \infty)$$

trong đó: l: độ dài chuỗi, p: xác suất.

Do đó, hồi quy bằng spline bậc 3 trung bình tốt hơn hồi quy bằng đa thức.

2. Phương pháp độ lệch dư

Xác định hàm Fj (j = 1, ...,n) trong (1) là nhiệm vụ của bài toán hồi quy. Trong nhiều trường hợp, quan hệ Fj (j = 1, ..., n) là phi tuyến thì hồi quy tuyến tính là không phù hợp.

Có những cách mô tả hàm Fj khác nhau. Nhưng ở đây chỉ trình bày phương pháp độ lệch dư. Phương pháp độ lệch dư như sau:

Giả sử các số liệu quan trắc được là (y, x1, x2, ..., xn). Từ đó ta dựng đường hồi quy y1 = f1 (x1) từ các cặp giá trị (x1, y), thế thì độ lệch dư Δy = y - y1 sẽ do các yếu tố còn lại (x2, x3,...., xn) gây nên. Trong nhiều trường hợp , một số yếu tố còn chưa biết hoặc không có số liệu thì bước đầu tiên ta cũng phải làm như vậy. Ta tiếp tục dựng đường hồi quy Δy1 = f2 (x2) từ các cặp (x2, Δy), từ đó ta được độ lệch dư Δy2 = Δy1-Δy sẽ do các biến x3, x4, ..., xn gây nên. Tiếp tục như vậy ta sẽ có :

$$y = f1(x1) + f2(x2) + \ldots\ldots + fn(xn) \qquad (2)$$

Trước đây [5] các hàm fj (j = 1, ..., n) được xác định bằng đồ thị hoặc bằng nội suy spline bậc 3, do đó nó chưa mang tính khách quan và kết quả chưa cao. Nay ta thay nội suy spline bậc 3 bằng hồi quy spline bậc 3 để xây dựng các hàm fj (j = 1, ..., n). Từ lý thuyết toán cũng như từ thực nghiệm thấy rằng hồi quy spline bậc 3 cho sai số sicma nhỏ và độ cong trơn lớn.

Việc chọn thứ tự các biến và loại bỏ biến nào trong mô hình hồi quy phi tuyến nhiều chiều này như sau:

Tính ma trận hệ số tương quan biến x để loại bỏ biến có hệ số tương quan cao hoặc dùng chúng để tạo biến mới. Tính các hệ số tương quan giữa biến y và các biến x1, x2, ..., xn ta được Ry - x1, Ry - x2, ..., Ry - xn và các hệ số tương quan của y khi y được sắp xếp theo mỗi biến Xj (j = 1, ..., n) tăng dần Rs1, Rs2, ...Rsn. Chọn biến Xj nào đạt giá trị Max (Ry -Xk, Rsk) (k = 1, ...,n). Cách làm này sẽ cho phép đánh giá các hàm fj (xj) là phi tuyến hay không. Quá trình lựa chọn biến cứ tiếp tục tương tự đối với các độ lệch dư còn lại [5].

Quá trình tính sẽ dừng khi kiểm nghiệm giả thuyết thống kê thấy rằng tương quan giữa độ lệch dư và các biến còn lại không còn đáng tin cậy.

3. Phương pháp cực tiểu hóa mạo hiểm trung bình thực nghiệm theo cấu trúc

Cực tiểu hóa mạo hiểm trung bình thực nghiệm chỉ có thể trong điều kiện tồn tại các thông tin gần đúng về xác suất. Lý thuyết rất phức tạp, ở đây chỉ trình bầy một cách làm cực tiểu mạo hiểm trung bình khi sử dụng thông tin ít hơn, nhưng lại cho phép chọn một lớp đủ nhỏ các hàm để lập mối quan hệ tương quan hồi quy. Nói đơn giản như sau:

Giả sử cần lập phương trình hồi quy:

$$y = \varphi(x,\alpha) \qquad \text{với } \alpha \text{ là véc tơ các tham số.} \qquad (3)$$

Cần phải cực tiểu hóa mạo hiểm trung bình

$$I(\alpha) = \int (y - \varphi(x,\alpha)^2 \, p(x,y)\, dx\, dy \qquad (4)$$

Trong đó p - (x,y) mật độ phân bố thực nghiệm.

Từ các cặp giá trị thực nghiệm (x_1, y_1), (x_2, y_2), ..., (x_n,y_n), ta cực tiểu hóa hàm thực nghiệm.

$$Itn(\alpha) = (1/\ell) \sum_{i=\ell}^{\ell} (y_i - \varphi(x1,\alpha))^2 \qquad (5)$$

Điểm cực tiểu $\varphi(x, \alpha tn)$ của hàm này là điểm cực tiểu $\varphi(x,\alpha o)$ hàm (4).

Nếu như ta lập một cấu trúc trên một tập các đường hồi quy gần đúng với các điểm thực nghiệm thì sẽ có thêm các khả năng cực tiểu hóa mạo hiểm thực nghiệm. Sử dụng phương pháp cực tiểu hóa mạo hiểm thực nghiệm theo cấu trúc cho phép với một số hữu hạn các thông tin ta tìm được số các biến cố tối ưu trong phương trình hồi quy và tìm độ dài chuỗi tối ưu nhất.

Giả sử ta tìm phương trình hồi quy $y = \varphi(x1, ..., xn, \alpha)$ với α là véctơ các tham số từ L điểm thực nghiệm. Từ lý thuyết cực tiểu mạo hiểm trung bình thực nghiệm theo cấu trúc sẽ có hàm giải tích sau:

$$j(\alpha) = \Psi(Itn(\alpha), N, L, \mu) \qquad (6)$$

Trong đó

$j(\alpha)$ là cận trên của hàm thực nghiệm (5),

$1-\mu$ là xác suất đánh giá này là đúng,

L là số điểm thực nghiệm.

Rõ ràng là nếu chỉ tìm α sao cho sai số quân phương sicma nhỏ thì chưa đủ. Vậy tốt hơn cả là tìm α sao cho $J(\alpha)$ nhỏ nhất. Trong từng bài toán cụ thể nó sẽ có dạng riêng [6]. Ta phải tìm các giá trị α,N,L sao cho

J (α) = min là được. Nhờ đánh giá (6) mà ta có thể chọn được số biến tối ưu Nopt và có thể chọn lọc được số các điểm thực nghiệm, tức là tăng hay giảm độ dài chuỗi L để tìm Lopt sao cho J (α) = min) để mô hình hồi quy cho kết quả tốt. Đây là một phương pháp mới tỏ ra rất có hiệu quả để tìm một phương trình hồi quy khi độ dài chuỗi số liệu ngắn. Phân tích (6) thấy rằng đây là phương pháp khách quan cho phép chọn lọc số biến và chọn lọc số điểm thực nghiệm sao cho hồi quy là tốt nhất. Thật vậy, nhiều khi trong phương trình hồi quy thêm biến vào thấy sai số quân phương nhỏ đi nhưng điều đó chưa chắc đã tốt hơn nếu ta lấy số biến ít hơn với sai số quân phương lớn hơn. Tương tự trong thực tế thì vấn đề có bỏ đi một số điểm thực nghiệm hay thêm vào để lập đường hồi quy tốt hơn sẽ được giải quyết nhờ vào đánh giá (6).

Do đó, nhờ có lý thuyết cấu trúc mà ta chọn được hàm hồi quy tốt nhất khi thêm bớt tham số và lọc những điểm thực nghiệm có sai số hệ thống lớn.

Vậy ta đã xây dựng xong một mô hình hồi quy phi tuyến nhiều chiều. Sau đây là ví dụ dùng mô hình trên để tính Qmax p% và phân tích sơ bộ về sự hình thành Qmax p%.

III. Một vài kết quả ban đầu khi thử nghiệm mô hình

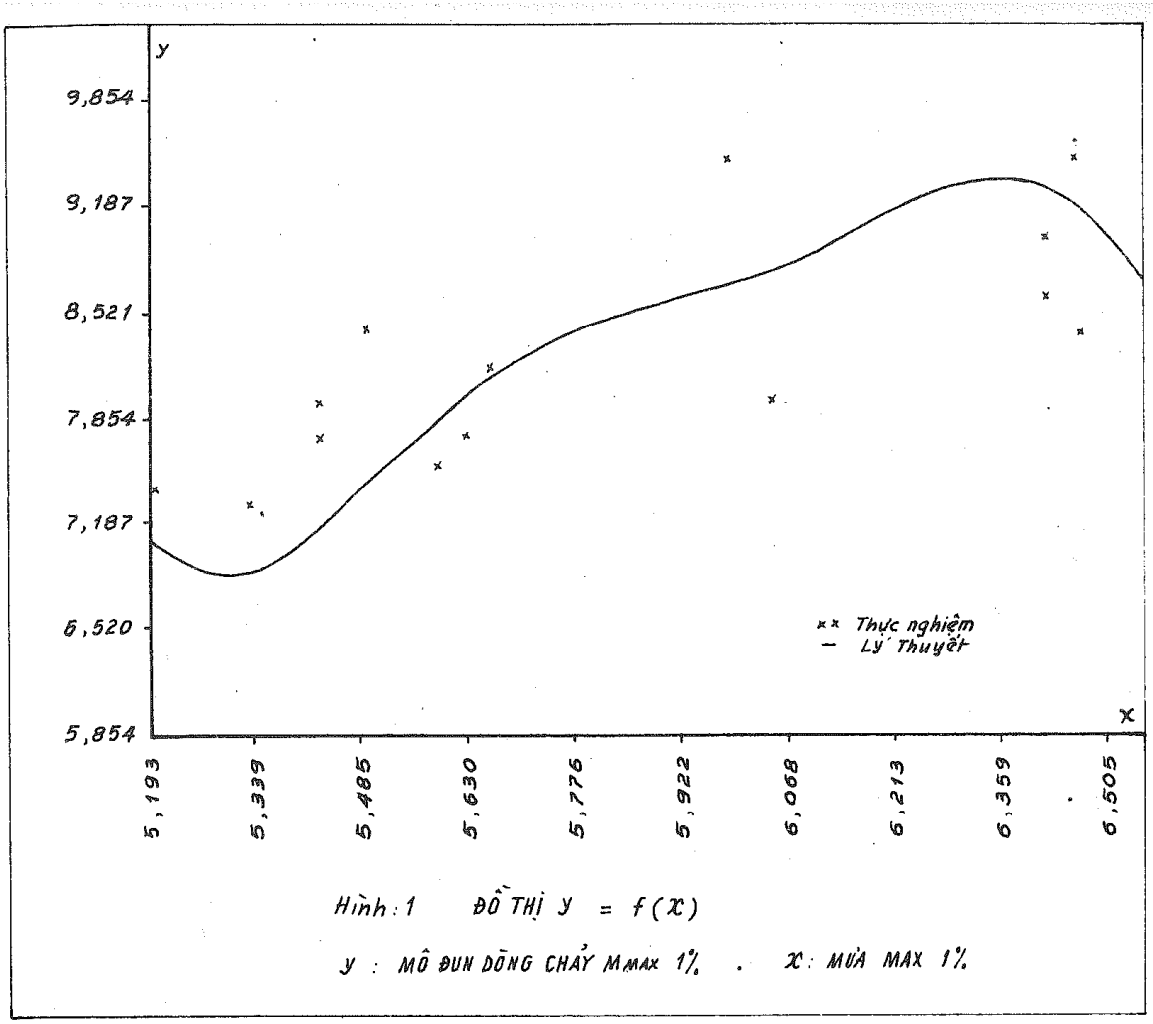
1. Số liệu cho mô hình

Ta có số liệu trên 19 lưu vực sông: modun dòng chảy Mmax1%, lượng mưa trên lưu vực Xmax 1%, diện tích lưu vực F (km^2), độ dốc bình quân lưu vực j (%), độ dốc lòng sông I (‰), độ dài lưu vực L (km), độ cao trung bình lưu vực H (m), độ rộng lưu vực B (km), mật độ lưới sông D (km/ km^2), hệ số đối xứng Kd. Vì mối quan hệ giữa dòng chảy Qmax với các biến khác thường được biểu diễn dưới dạng một tích cho nên các biến được biểu diễn dưới dạng lôgarit.

2.Thử nghiệm mô hình

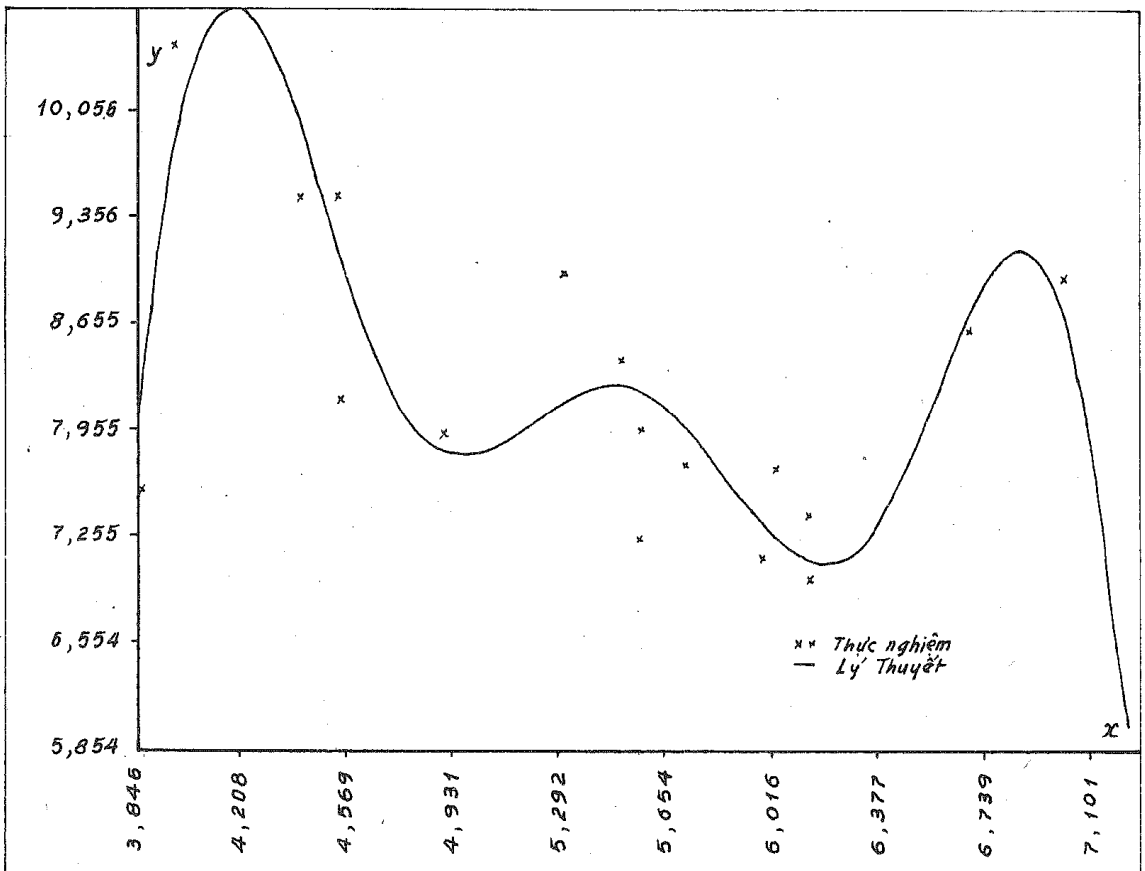
Đầu tiên, tính ma trận hệ số tương quan biến X, các hệ số tương quan giữa Mmax1% với các biến Xj. Vì hệ số tương quan giữa biến F với L lớn nên loại bỏ biến L trong danh sách các biến. Tiếp theo là lập các quan hệ riêng Mmax1% = f (Xmax1%), Mmax1% = f (F), Mmax1% = f(H), Mmax1% = f(B). Trong số các hệ số tương quan giữa biến Mmax1% với các biến khác thì hệ số tương quan giữa Mmax với Xmax là lớn nhất.

Hình 1 cho thấy rằng Mmax1% tăng theo Xmax1% và đạt giá trị lớn tại lân cận Xmax1% = 580 mm. Mmax1% đạt giá trị lớn tại lân cận F = 66,6km^2, còn cực trị thứ hai đạt tại F = 879km^2 (Hình 2). Mmax1% đạt giá trị lớn tại lưu vực nằm ở độ cao H = 403m và độ rộng của nó B = 5km. Kết quả tính cho thấy rằng Mmax1% đạt giá trị lớn tại các lưu vực có độ dốc



Hình : 1 ĐỒ THỊ y = f(x)

y : MÔ ĐUN DÒNG CHẢY MMAX 1% . x : MƯA MAX 1%



Hình : 2 ĐỒ THỊ Y = f(x)

y MÔ ĐUN DÒNG CHẢYMMAX 1% . x : DIỆN TÍCH LƯU VỰC SÔNG

J > 22,2% và mật độ sông khoảng D= 1,04 km/km^2 hoặc D > 1,59 km/km^2. Tại các lưu vực có hệ số đối xứng rất nhỏ
Kd = 0,002 ta cũng thấy như vậy.

Quá trình chọn lọc các biến như sau:

a. Với một biến mưa Xmax1% tại các trạm Tài Chi, Bằng Cả, Nghĩa Đô, Bản Diệp, Pa Há có độ lệch dư dương lớn. Lưu vực trạm Tài Chi có diện tích nhỏ hơn 100 km^2, còn tại Bản Diệp, Sa Pả có diện tích F từ 200 đến 400 km^2 nhưng chúng nằm ở nơi có độ cao H > 1200m, hệ số tương quan giữa Mmax1% và Xmax1%: $R_y - y_t = 0,79$, sai số quân phương sicma = 8,1%. Tại các lưu vực mà ở đó độ lệch dư dương lớn chỉ có thể giải thích là do ảnh hưởng của các yếu tố khác lên sự hình thành Mmax1% làm cho chúng lệch đáng kể so với đường trung bình Mmax1% = f (Xmax1%).

b. Khi thêm biến diện tích F vào thì hệ số Ry - yt = 0,88 [hệ số tương quan giữa biến y quan trắc và biến y tính tức là giữa Mmax1% cho trước và Mmax1% tính], sai số sicma = 6,4%, lúc này độ lệch dư dương lớn tại Tài Chi, Nghĩa Đô, Bản Diệp, Pa Há. Tại trạm Bằng Cả độ lệch dư không còn nhiều nữa, vậy khi thêm biến F thì hệ số tương quan tăng lên rõ rệt và sai số quân phương giảm đi đáng kể.

c. Quá trình tính và phân tích tương tự tiếp tục như vậy cho các biến khác. Khi thêm các biến B, i, J, Kd, D ta có sicma = 3,3% và Ry - yt = 0,966

3. Nhận xét

a. Mô hình cho kết quả tính đạt yêu cầu:Ry - yt = 0,966 và sai số sicma = 3,3%. Nó cho phép phân tích các kết quả thu được. Mô hình phi tuyến nhiều biến trên cho phép xác định một quan hệ rất phức tạp giữa Mmax1% với các biến khác. Kết quả tính toán cuối cùng được thể hiện bằng đồ thị giữa Y cho trước (Mmax1%) và Yt tính (Mmax1% tính). Quan hệ này gần là một đường thẳng.

b. Giá trị Mmax1% tăng theo Xmax1% và nó có cực trị tại Xmax1% = 588mm. Trong hình thành Mmax1% thì cùng một lượng mưa Xmax1%, modun dòng chảy Mmax1% lớn ở các lưu vực có diện tích nhỏ khoảng F = 66km^2 hoặc với các lưu vực nằm ở độ cao khoảng H ≈ 1200m và độ rộng lưu vực hẹp B = 5km hoặc B = 10km.

c. Kết quả tăng rõ rệt nếu trong công thức tính Mmax1% có hai biến Xmax1% và F.

d. Tại các lưu vực có độ lệch dư dương lớn ngay cả khi ta thêm khá nhiều biến để tính, ở đó có thể xảy ra lũ đặc biệt.

Nếu có chuỗi số liệu dài hơn ta sẽ có những nhận xét chính xác hơn về hình thành lũ và khi có đủ số liệu thì có thể dùng mô hình này để tính toán và phân tích kết quả theo cách tương tự như trên. Nó có thể dùng để đánh giá năng suất cây trồng trong thủy văn nông nghiệp, tính toán cân bằng nước v.v.

IV. Kết luận

Mô hình hồi quy phi tuyến nhiều chiều mà tác giả đã lập như trên tỏ ra hữu hiệu khi tính toán số trị, phân tích, đánh giá sự tác động của các yếu tố khác nhau. Nó chặt chẽ hơn các mô hình tương tự đã có về lý thuyết và đã thể hiện khả năng của mô hình qua kết quả số trị và cách phân tích tương quan như trên. Khi dùng hồi quy tuyến tính nhiều chiều thì rất khó khăn phân tích kết quả. Vậy mô hình này là một công cụ mới hữu hiệu để nghiên cứu các quá trình khí tượng thủy văn nhiều chiều.

Tài liệu tham khảo

1. Seber D.G. Giải tích hồi quy tuyến tính, Matxcơva, 1980, (Tiếng Nga).

2. Constanchinov A.P. Thời tiết, thổ nhưỡng và năng suất lúa mì. Lêningrat, 1978, (Tiếng Nga)

3. Constanchinov A.P. Khimin H.N. Sử dụng phương pháp độ lệch dư của phân tích thống kê để nghiên cứu các quá trình thủy văn, Tạp chí khí tượng thủy văn 1980, No.2 (Tiếng Nga).

4. Constanchinov A.P. Khimin H.N. Làm trơn các tương quan khí tượng thủy văn bằng spline bậc 3, Tạp chí khí tượng thủy văn, 1980, No.7.

5. Constanchinov A.P. Khimin H.N. Ứng dụng spline và phương pháp độ lệch dư trong KTTV. Lêningrat, 1983, (Tiếng Nga).

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

7. Schoenberg I.J Spline function and problem of graduation Pro. Nat. USA, 1964.

8. Alêcxâyev.G.A. Phương pháp khách quản làm trơn và chuẩn hóa các quan hệ tương quan, Lêningrat, 1971, (Tiếng Nga).

9. Panovski G.A, Braier G.B. Các phương pháp thống kê trong khí tượng. Lêningrat, 1971, (Tiếng Nga).

10. Rodestvenski A.V, Trebotarev A.P. Các phương pháp thống kê trong thủy văn. Lêningrat, 1974, (Tiếng Nga).

11. Constanchinov A.P. Bay hơi trong tự nhiên. Lêningrat, 1968, (Tiếng Nga).

12. Nguyễn Hữu Khải, Lê Xuân Cầu. Ứng dụng hàm spline bậc 3 để xử lý các quan hệ tương quan trong khí tượng thủy văn. Tập san khí tượng thủy văn, 1994, No406.