NHIỆT LIỆT CHÀO MỪNG NGÀY BÁO CHÍ VIỆT NAM 21 — VI

TỐI ƯU HÓA VIỆC LỰA CHỌN CÁC NHÂN TỐ CHO MÔ HÌNH SỬ DỤNG PHƯƠNG PHÁP PHÂN TÍCH PHÂN BIỆT TUYẾN TÍNH ĐỂ DỰ BÁO SỰ PHÁT SINH CỦA XOÁY THUẬN NHIỆT ĐỚI

IUTORTRUC IU. V.,
NGUYỄN SĨ THANH
Phòng Nghiên cứu liên hợp Việt-Xô về khí tượng nhiệt đới

MỞ ĐẦU

Các quá trình nảy sinh và phát triển của xoáy thuận nhiệt đới (XTNĐ) chịu ảnh hưởng của rất nhiều loại quá trình diễn ra trong khí quyển cũng như trên đại dương. Bởi vậy, về nguyên tắc, để nhận được các kết luận dự báo với độ tin cậy cao hơn, nên đưa được vào mô hình thống kê XTNĐ càng nhiều nhân tố dùng làm dự báo càng tốt. Nhưng, khi tăng chiều của véc tơ các nhân tố dùng làm dự báo, sai số trong việc tính các hệ số mẫu của hàm phân biệt cũng tăng lên đáng kể, và điều đó lại dẫn đến việc hạ thấp độ tin cậy của dự báo. Vì vậy, trong việc lựa chọn các nhân tố dùng làm dự báo, chúng ta buộc phải tìm một giải pháp dung hòa giữa hai yếu tố đối nghịch nhau này.

Việc lựa chọn các nhân tố dùng làm dự báo có thể tiến hành hoặc trên mẫu phụ thuộc, nghĩa là mẫu mà theo các số liệu của nó, chúng ta tiến hành tính các hệ số của hàm phân biệt, hoặc tiến hành trên mẫu độc lập trên cơ sở các số liệu khí tượng chưa được sử dụng trong quá trình tính toán các thông số của mô hình thống kê.

Tuy nhiên, cần lưu ý rằng, việc lựa chọn các nhân tố dựa trên mẫu phụ thuộc gắn liền với những khó khăn nhất định. Vấn đề là ở chỗ khi tăng chiều của véc tơ các nhân tố dùng làm dự báo, độ tin cậy của các kết quả dự báo tính trên mẫu phụ thuộc sẽ tăng đơn điệu và khi vượt quá một số lượng nào đó các nhân tố, nó sẽ làm sai lệch đáng kể độ tin cậy thực sự của sơ đồ dự báo. Ví dụ, nếu số lượng nhân tố dùng làm dự báo bằng dung lượng mẫu bớt đi một đơn vị thì không phụ thuộc vào việc các nhân tố có mang một lượng thông tin thực tế nào đó về quá trình đang được dự báo hay không, độ tin cậy của sơ đồ dự báo trên mẫu phụ thuộc vẫn là 100%, trong khi đó, nếu đưa vào áp dụng trong công tác nghiệp vụ, độ tin cậy của sơ đồ có lẽ gần với mức tin cậy của một dự báo ngẫu nhiên. Từ lý thuyết thống kê toán học, đã biết rằng độ tin cậy trung bình của các dự báo phụ thuộc là quá cao, nghĩa là có độ chệch dương.

Độ tin cậy của các kết quả dự báo tính trên mẫu độc lập là một ước lượng không chệch, vì vậy, có thể sử dụng để xác định thành phần tối ưu các nhân tố dùng làm dự báo. Mặc dù vậy, do khối lượng thông tin khí tượng hiện có bị hạn chế, các nhà nghiên cứu nói chung, không thể dành cho mẫu độc lập một số lượng cần thiết các véc tơ trạng thái. Bởi vì, điều đó sẽ dẫn tới việc giảm dung lượng mẫu phụ thuộc, như vậy, sẽ kéo theo việc giảm độ tin cậy của các ước lượng cho hàm phân biệt, và tiếp sau là của các kết quả dự báo. Còn nếu sử dụng mẫu độc lập với dung lượng bé để tiến hành tối ưu hóa thành phần các nhân tố dùng làm dự báo thì sẽ không thể nhận được các kết quả thỏa đáng, vì trong trường hợp này, sự ước lượng độ tin cậy tuy là một ước lượng không chệch nhưng dù sao vẫn có độ sai số ngẫu nhiên đáng kể.

Chúng ta hãy điểm qua ngắn gọn một phương pháp tối ưu hóa nữa gọi là phương pháp «dao xếp», mà thời gian gần đây được sử dụng khá rộng rãi, mặc dù đã được đề xướng từ những năm 50. Nội dung cơ bản của phương pháp như sau: Từ một mẫu có N phần tử, lần lượt lấy ra mỗi lần một véc tơ trạng thái. Sau đó, trên phần còn lại của mẫu có dung lượng là (N – 1), người ta tính các thông số của mô hình thống kê, còn trạng thái đã lấy ra được xem như mẫu độc lập để tính sự ước lượng chất lượng dự báo. Kết quả là trên cơ sở một mẫu có chứa N phần tử, chúng ta nhận được N dự báo độc lập, mỗi một trong số đó được tính theo mô hình sử dụng (N – 1) phần tử. Mỗi dự báo nhận được bằng cách đó, nếu xét riêng rẽ, đúng là không phụ thuộc, nhưng tổ hợp các dự báo đó, nói chung, không tương đương với tổ hợp một số lượng như thế các dự báo nhận được trên mẫu độc lập. Vấn đề là ở chỗ việc bỏ đi một phần tử của mẫu không thể có ảnh hưởng đáng kể đến giá trị của các ước lượng các thông số của mô hình thống kê. Cho nên, trong trường hợp này, tất cả các dự báo được tính trên cơ sở các bộ thông số hầu như không đổi của mô hình thống kê, và các thông số này chẳng khác gì lắm so với các thông số tương ứng nhận được trên mẫu phụ thuộc. Như vậy, các dự báo nhận được bằng phương pháp «dao xếp» hầu như không khác gì lắm so với các dự báo phụ thuộc, bởi thế, không thể sử dụng các dự báo đó để tối ưu hóa thành phần các nhân tố dùng làm dự báo.

1. ĐẶT VẤN ĐỀ

Mục đích của bài này là đưa ra một cách tiếp cận mới trong việc giải quyết bài toán tối ưu hóa thành phần và độ dài của véc tơ các nhân tố dùng làm dự báo, dựa trên phương pháp thử nghiệm thống kê, hay còn gọi là phương pháp Mônglơ-Karlô.

Phương pháp thử nghiệm thống kê cho phép tạo ra các mẫu các số ngẫu nhiên có các tính chất thống kê đã cho trước, vì vậy, trong cách tiếp cận này, thực chất không nảy sinh vấn đề gì liên quan tới tính hạn chế về dung lượng của mẫu độc lập. Đồng thời, khó khăn chủ yếu của cách tiếp cận này gắn liền với tính chất phức tạp của việc thiết lập sự tương quan giữa các kết quả thử nghiệm thống kê với các tính toán bằng các nhân tố thực, tiến hành trên mẫu phụ thuộc. Để giải quyết vấn đề này, chúng ta sử dụng đại lượng chỉ số thông tin η đã được đưa ra trong công trình của một trong các tác giả [1], đặc trưng cho tỉ lệ giữa tín hiệu và nhiễu của một tổ hợp các nhân tố.

Giả sử chúng ta có m giá trị các ước lượng mẫu $\tau_1, \tau_2 \ldots \tau_m$ tương ứng với các tập hợp tổng quát của các đại lượng thống kê với các giá trị thực là $\Theta_1, \Theta_2 \ldots \Theta_m$, và được đặc trưng bởi các sai số mẫu trung bình bình phương là $\delta[\xi_1], \delta[\xi_2], \ldots, \delta[\xi_m]$. Khi đó;

$$\eta = \frac{\delta^2[\Theta]}{\delta^2[\xi]} = \frac{\delta^2[\tau]}{\delta^2[\xi]} \qquad (1)$$

Ở đây $\delta^2[\tau]$ và $\delta^2[\Theta]$ tương ứng là phương sai của các ước lượng mẫu và các giá trị thực của chúng.

Vì các giá trị $\delta[\xi]$ của các sai số mẫu trung bình bình phương đối với phần lớn các ước lượng mẫu đã biết trước từ lý thuyết thống kê toán học, cho nên, từ (1), tỉ số giữa tín hiệu và nhiễu có thể được ước lượng theo các số liệu thực nghiệm.

Như vậy, bài toán tối ưu hóa có thể được đưa về việc xây dựng trên cơ sở phương pháp Môngte-Karlô các nhân tố dùng làm dự báo với độ thông tin của tổ hợp các nhân tố và với dung lượng của mẫu phụ thuộc.

2. SƠ ĐỒ DỰ BÁO SỰ NẢY SINH BÃO

Việc dự báo sự hình thành bão từ một áp thấp nhiệt đới với thời hạn dự báo là τ ngày có thể biểu diễn dưới dạng một bài toán loại trừ có lời giải là có hoặc không. Cũng có thể giải quyết bài toán dự báo sự tiến triển của XTNĐ — sau một khoảng τ, XTNĐ sẽ đầy lên hay sẽ phát triển? — bằng cách đặt vấn đề mang tính chất định tính như vậy.

Để giải quyết các bài toán loại trừ, trong thống kê toán học rất phát triển công cụ phân tích phân biệt nhiều chiều, cho phép xây dựng trong không gian m chiều một mặt siêu phẳng để tách một cách tối ưu nhất các điểm thuộc về hai lớp khác nhau.

Với một dung lượng không lớn lắm của mẫu, dùng phương pháp phân tích phân biệt tuyến tính là có hiệu quả nhất. Trong trường hợp này, mặt siêu phẳng để tách các điểm có dạng sau:

$$Y_{t+\tau} = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_m X_{m,t} \qquad (2)$$

Ở đây, $X_{i,t}$; $i = \overline{1,m}$ là các giá trị của véc tơ các nhân tố dùng làm dự báo ở thời điểm t:

$Y_{t+\tau}$ — giá trị của hàm phân biệt, nếu $Y \geqslant 0$ thì dự báo ở thời điểm $t+\tau$ sẽ xảy ra sự kiện thuộc lớp 1, còn nếu $Y < 0$ thì dự báo sự kiện thuộc lớp 2;

β_i — hệ số của hàm phân biệt, được chọn ra theo điều kiện thoả mãn việc tách tuyến tính tốt nhất các điểm thuộc lớp 1 và lớp 2.

Giả sử chúng ta có 2 mẫu con các nhân tố dùng làm dự báo có dung lượng tương ứng là n_1 và n_2, ứng với các trạng thái của nhân tố được dự báo thuộc lớp 1 và lớp 2:

$$X^{(1)}_{i\,t}: i = \overline{1,m}\ ;\ t = \overline{1,\ n_1}\ ;$$

$$X^{(2)}_{i\,t}: i = \overline{1,m}\ ;\ t = \overline{1,\ n_2}\ ;$$

Khi đó, hệ số của hàm phân biệt sẽ được tính theo công thức:

$$\beta = S_*^{-1}\left(\bar{X}^{(1)} - \bar{X}^{(2)}\right) \qquad (3)$$

Ở đây, S_* — ma trận hiệp biến liên hợp.

$$S_* = \frac{1}{n_1 + n_2 - 2}\left(X^{(1)T} X^{(1)} + X^{(2)T} X^{(2)}\right) \qquad (4)$$

Trong (3) và (4) sử dụng các ký hiệu sau:

X^T — dấu chuyển vị;

$\bar{X}^{(1)}$, $\bar{X}^{(2)}$ — Các vác tơ các trung bình toán học của các nhân tố dùng làm dự báo ứng với các mẫu con 1 và 2;

$x_{it} = X_{it} - \bar{X}_i$ — dị thường của các nhân tố dùng làm dự báo. Phần tử tự do β_o của toán tử tuyến tính được xác định như sau:

$$\beta_o = -\frac{1}{2}\beta^T\left(\bar{X}^{(1)} + \bar{X}^{(2)}\right) \qquad (5)$$

Có thể chứng minh (xem [2]) rằng các hệ số β được tính theo công thức (3), sẽ cực đại hóa khoảng cách Makhalônôbis:

$$l = \left(\bar{X}^{(1)} - \bar{X}^{(2)}\right)^T S_*^{-1}\left(\bar{X}^{(1)} - \bar{X}^{(2)}\right) \qquad (6)$$

Là một trong các chỉ số để tách hai nhóm trong không gian các nhân tố dùng làm dự báo.

Bây giờ chúng ta xét sơ đồ sử dụng thủ tục sàng lọc để chọn các nhân tố dùng làm dự báo.

Ở giai đoạn thứ nhất, tiến hành việc xem xét từng nhân tố một. Đối với mỗi nhân tố đều xây dựng hàm phân biệt một chiều và tính khoảng cách Makhalônôbis. Nhân tố ứng với khoảng cách lớn nhất sẽ được chọn là nhân tố thứ nhất

Ở giai đoạn hai, nhân tố được chọn ra kết hợp với từng nhân tố còn lại và đối với mỗi cặp như vậy xây dựng hàm phân biệt hai chiều. Nhân tố nào trong cặp với nhân tố thứ nhất cho ta giá trị lớn nhất của khoảng cách Makhalônôbis sẽ được chọn làm nhân tố thứ hai.

Tương tự như vậy, tiến hành việc chọn các nhân tố thứ 3, thứ 4 và tiếp sau nữa. Như đã nói ở trên, khi ta bổ sung thêm nhân tố tiếp sau, khoảng cách Makhalônôbis tính trên mẫu phụ thuộc sẽ không thể giảm, vì vậy, khó có thể sử dụng nó để xác định số lượng tối ưu các nhân tố dùng làm dự báo.

3. THỬ NGHIỆM TÍNH HIỆU QUẢ CỦA SƠ ĐỒ DỰ BÁO BẰNG PHƯƠNG PHÁP MÔNGTE-KARLÔ.

Cơ sở của phương pháp Môngte-Karlô là bộ tạo số ngẫu nhiên cho phép thu được dãy các số ngẫu nhiên không tương quan ξ_i có hàm phân bố cho trước. Ở đây, chúng ta sẽ không xét các vấn đề liên quan đến các phương pháp tạo số ngẫu nhiên mà chỉ lưu ý rằng các vấn đề này đã được nghiên cứu rất đầy đủ, [3].

Sử đụng bộ tạo số ngẫu nhiên chuẩn. hiện có trong bảo đảm toán học của máy tính điện tử ES – 1022, các tác giả đã tạo 2 mẫu con mô phỏng các tính chất thống kê của các nhân tố dùng làm dự báo ứng với 2 lớp của quá trình được dự báo:

$$X_{i,t}^{(1)} = \xi_j \; ; \; i = \overline{1,m} \; ; \; t = \overline{1,N} \; ; \; j = (t-1).m + i$$

$$X_{i,t}^{(2)} = \xi_j + \alpha_i \; ; i = \overline{1,m}; \; t = \overline{1,N}; \; j = N.m + (t-1).m + i \qquad (7)$$

Ở đây chỉ số thứ tự của nhân tố dùng làm dự báo, còn chỉ số t biểu thị số thứ tự của véc tơ trạng thái trong mỗi mẫu con.

Từ (7), thấy rõ rằng giá trị trung bình của các nhân tố ở mẫu con thứ nhất khác với giá trị trung bình tương ứng ở mẫu con thứ hai bởi một đại lượng là (α.i). Bằng cách thay đổi hệ số α, chúng ta có thể thu được các giá trị chỉ số thông tin khác nhau của tổ hợp các nhân tố. Dễ thấy rằng, trong mô hình (7), lượng thông tin của các nhân tố dùng làm dự báo tỉ lệ với số thứ tự của chúng.

Mỗi mẫu con lại được chia làm 2 phần: n phần tử đầu tiên dành cho mẫu phụ thuộc, còn (N–n) phần tử còn lại dành cho các thử nghiệm độc lập. Như vậy, mẫu con phụ thuộc chứa 2n phần tử, còn mẫu con độc lập chứa 2(N–n).

Theo phương pháp đã mô tả ở phần 2, trên cơ sở mẫu con phụ thuộc tiến hành tính các hệ số của hàm phân biệt theo công thức (3) và (5) và sau đó đưa ra các kết quả dự báo đối với các mẫu con độc lập và phụ thuộc theo (2). Chất lượng của các dự báo được đánh giá cho từng mẫu con theo công thức:

$$P = \left(\frac{n_+}{n_+ + n_-} .100 \right)\% \qquad (8)$$

Ở đây: n_+ – số các dự báo mà số thứ tự của lớp dự báo và lớp thực là trùng nhau;

n_- – số các dự báo sai.

Ngoài ra, trên cơ sở mẫu phụ thuộc, cũng đánh giá tỉ lệ giữa tín hiệu với nhiễu theo công thức:

$$\eta = \frac{1}{m} \sum_{i=1}^{m} \frac{\left(\overline{X}_i^{(1)} - \overline{X}_i^{(2)} \right)^2}{\frac{1}{n_1} \delta^2_{X_i^{(1)}} + \frac{1}{n_2} \delta^2_{X_i^{(2)}}} - 1 \qquad (9)$$

Sau đó, quá trình này, bắt đầu từ việc tạo các mẫu con mới (7), xây dựng các hàm phân biệt, đánh giá lượng thông tin của các nhân tố dùng làm dự báo đến việc tính độ tin cậy của các dự báo, được lặp lại k lần. Kết quả cuối cùng là các ước lượng thu được bằng cách tính trung bình theo k. Một phần các kết quả đó được biểu diễn trên các hình 1 – 3.

Trong khi tiến hành các tính toán, các hằng số trong các công thức (1) – (9) đã được gán các giá trị sau:

Dung lượng mẫu được tạo: $2N = 400$;

Dung lượng mẫu phụ thuộc: $n_1 + n_2 = 2n = 30$;

Dung lượng mẫu độc lập: 370;

Số các nhân tố được tạo ra: $m_{max} = 20$;

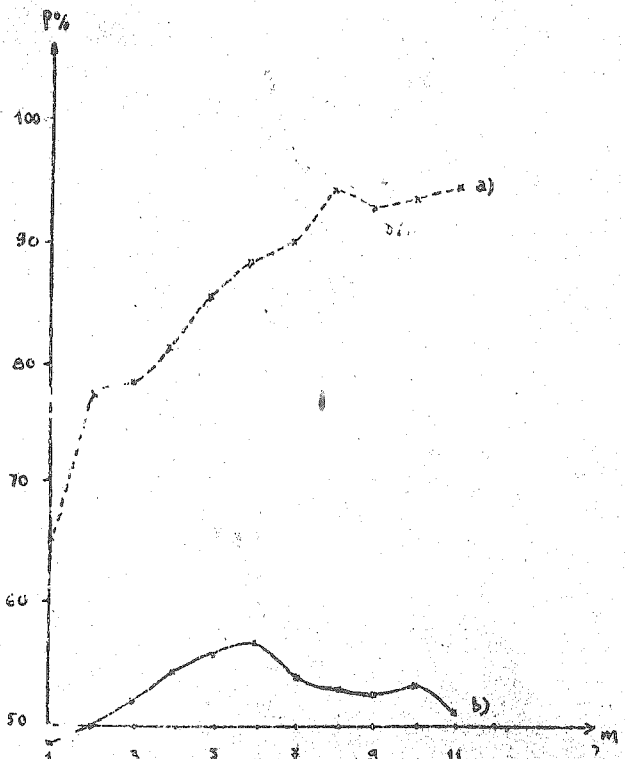
Số lần lặp lại quá trình thử nghiệm Mônte-Karlô: $k = 4$;

Hệ số α trong biểu thức (7): $\alpha = 0{,}0025 \div 0{,}002$.

Để minh họa, trên hình 1 đã dẫn ra sự phụ thuộc giữa độ tin cậy của sơ đồ dự báo như một hàm số đối với số lượng các nhân tố dùng làm dự báo đã được lựa chọn bằng thủ tục sàng lọc, ứng với giá trị $\eta = 0{,}15$. Trước hết, tập trung sự chú ý là sự khác nhau đáng kể giữa độ tin cậy của các kết quả dự báo trên mẫu phụ thuộc và mẫu độc lập, tăng từ 15% đối với một nhân tố được chọn đầu tiên đến 40% đối với hàm phân biệt chứa 10 nhân tố đã sàng lọc.

Khi các giá trị của tỉ số giữa tín hiệu và nhiễu bé, độ tin cậy cao, thậm chí gần bằng 100%, của các dự báo trên mẫu phụ thuộc thực chất không nói lên điều gì, bởi vì khi chuyển sang các thử nghiệm độc lập, độ chính xác hạ xuống đến mức dự báo ngẫu nhiên (hình 1,b). Cho nên, trong trường hợp này, đặc biệt quan trọng là chọn $m_{tối\ ưu}$ — số lượng tối ưu các nhân tố dùng làm dự báo.

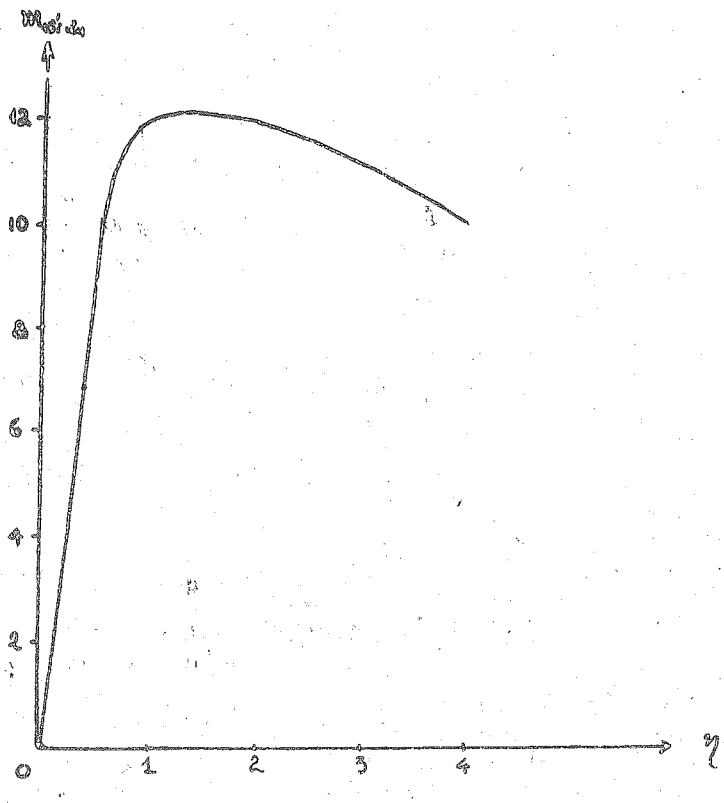
Với các giá trị đã chọn để tiến hành thử nghiệm thì $m_{tối\ ưu}$ là 6. Việc tồn tại các điểm cực đại trên đường cong b được giải thích bởi tác động của hai yếu tố trái ngược nhau quyết định độ tin cậy của các dự báo độc lập. Một mặt thì khi tăng số lượng các nhân tố dùng làm dự báo, lượng thông tin về nhân tố dự báo sẽ tăng lên theo, nhưng mặt khác, lượng nhiễu cũng sẽ tăng lên. Bởi vậy, khi bổ sung thêm nhân tố dùng làm dự báo tiếp sau, nếu yếu tố thứ hai chiếm ưu thế thì độ tin cậy sẽ giảm xuống.



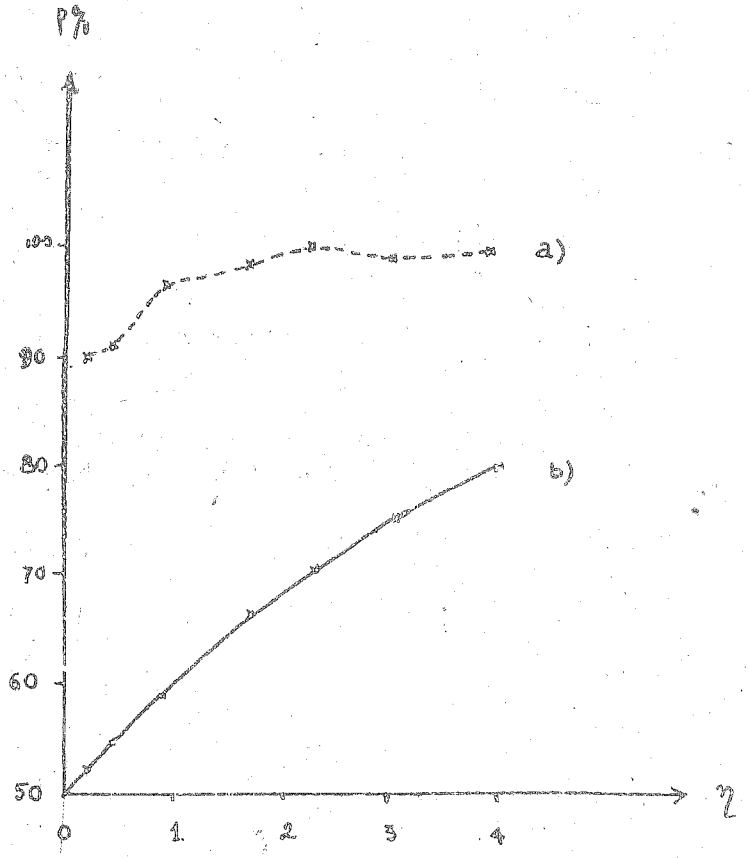
Hình 1 — Sự phụ thuộc giữa độ tin cậy của dự báo và số lượng nhân tố dùng làm dự báo ứng với giá trị $\eta = 0{,}15$.
a) Mẫu phụ thuộc;

Trên hình 2 biểu diễn sự phụ thuộc giữa số lượng tối ưu các nhân tố dùng làm dự báo và tỉ số giữa tín hiệu và nhiễu của tổ hợp các nhân tố. Với những giá trị η bé: $\eta = 0 - 0{,}5$, số lượng tối ưu các nhân tố dùng làm dự báo tăng vọt lên từ 0 đến 12 cùng với việc tăng chỉ số thông tin, nhưng bắt đầu từ giá trị $\eta = 2$, số lượng đó hơi giảm xuống. Điều này có lẽ liên quan đến yếu tố là khi lượng thông tin của tổ hợp các nhân tố dùng làm dự báo tăng lên, trong những điều kiện như nhau, tương quan giữa các nhân tố cũng tăng lên thành thử với giá trị m như nhau, số các nhân tố tương đương không phụ thuộc sẽ giảm xuống. Như vậy, việc bổ sung các nhân tố tiếp sau sẽ đem thêm vào một lượng thông tin ít hơn về quá trình được dự báo, trong khi đó thì lượng nhiễu được đem vào vẫn ở mức như cũ.

Để có thể nhận định về độ tin cậy của các dự báo đối với mẫu phụ thuộc và mẫu độc lập, các tác giả đã xây dựng đồ thị $P = f(\eta)$ (hình 3). Tập trung sự chú ý là độ chệch của các ước lượng độ tin cậy của các dự báo trên mẫu phụ thuộc giảm xuống khi lượng thông tin dùng làm dự báo tăng lên. Sử dụng đồ thị dẫn ra ở hình 3 có thể đưa ra các kết luận về mức độ tin cậy của các dự báo nghiệp vụ trên cơ sở số liệu của mẫu phụ thuộc. Chẳng hạn, để đảm bảo độ tin cậy của các dự báo ở mức 70%, cần có một tổ hợp các nhân tố dùng làm dự báo được đặc trưng bởi tỉ số giữa tín hiệu và nhiễu là $\eta \geqslant 2.$



Hình 2 — Sự phụ thuộc giữa số lượng tối ưu các nhân tố dùng làm dự báo và lượng thông tin của tổ hợp các nhân tố.



Hình 3 — Sự phụ thuộc giữa độ tin cậy của dự báo về lượng thông tin của các nhân tố được làm dự báo.

a) Mẫu phụ thuộc ;

b) Mẫu độc lập

TÀI LIỆU THAM KHẢO

1. Jutortruk A.T, Jutortruk Iu.V. Utrot apriorhoi informaxi v regresionnre modelax pragnoza pagodur – Trudы GGO, 1983, vup.466. (tiếng Nga)

2. Kramer G. Matematitreski metodu statictriki. M., «Mir», 1975. (tiếng Nga)

3. Xabol I.M.Trislennuri metodu Monte – Karlo,–M., «Nauka», 1983 (tiếng Nga)