KHẢ NĂNG DỰ BÁO HOẠT ĐỘNG MÙA BÃO BIỂN ĐÔNG VIỆT NAM: XÂY DỰNG MÔ HÌNH DỰ BÁO VÀ KIỂM NGHIỆM

PGS. TS. **Nguyễn Văn Tuyên**
Trung tâm KHCN KTTV&MT

Tiếp theo phần trước, phần này giới thiệu kết quả sử dụng phương pháp hồi quy từng bước tối ưu cùng với thuật toán jackknife để thiết lập các sơ đồ dự báo hoạt động mùa bão trên cơ sở những nhân tố sơ tuyển ở phần trước. Qua các kết quả tính toán cho thấy vai trò của các chỉ số khí hậu, đặc biệt là các chỉ số thuộc nhóm ENSO, gió mùa, dao động và nhóm các chỉ số quan hệ từ xa đều có mặt trong các sơ đồ dự báo. Kết quả thiết lập được 32 sơ đồ dự báo, trong đó 26 sơ đồ dự báo giải thích được 70÷83% phương sai, 2 sơ đồ xấp xỉ 70%, chỉ có 4 sơ đồ đạt 56÷63% phương sai. Các phương trình dự báo được thẩm định, kiểm nghiệm và đánh giá độ tin cậy qua hindcast trên số liệu phụ thuộc, kiểm tra chéo trên số liệu độc lập jackknife và dự báo thử nghiệm. Cuối cùng tác giả đã chỉ ra khả năng tổ hợp dự báo và những sơ đồ dự báo có khả năng áp dụng vào thực tế dự báo nghiệp trong tương lai cũng như những nghiên cứu cần được tiếp tục.

1. Những ưu việt và hạn chế của phép đổi biến HS1

Như ta đã thấy trong bài báo trước [11], phép đổi biến HS1 đã mang lại khả năng rất lớn trong việc tìm kiếm các nhân tố dự báo có thể cho mô hình dự báo thống kê. Sau đây ta sẽ chỉ ra tính ưu việt đó không phải là một ngẫu nhiên, mà thuộc về bản chất. Bên cạnh ưu điểm của phép đổi biến HS1 cũng kèm theo những hạn chế nhất định như kéo theo những khó khăn trong việc tính toán và giảm thiểu độ chính xác khi phải xấp xỉ qua các lần chuyển đổi biến.

a. Bản chất của phép biến đổi HS1

Để dễ theo rõi và quản lý các biến cũng như các file số liệu, từ đây chúng ta sẽ thêm ký tự "F" (hoặc "f") vào cuối mỗi tên biến, tên file hoặc tên sơ đồ để chỉ biến, file hoặc sơ đồ đó là đã được biến đổi theo phép đổi biến HS1, ký tự "H" (hoặc "h") để chỉ dự báo hindcast, còn ký tự "J" (hoặc "j") để chỉ dự báo jackknife; đồng thời cũng xin nói thêm ở đây là do trong tính toán chúng tôi sử dụng nhiều phần mềm khác nhau nên ký hiệu chữ hoa và chữ thường trong tên biến, file hoặc tên sơ đồ được hiểu như nhau (xem Phụ lục).

Bản chất phép đổi biến HS1 là tạo ra biến mới có phân bố chuẩn hoặc xấp xỉ chuẩn so với biến nguyên gốc, nhờ đó mà nó đã biến một chuỗi không có tính dừng thành một chuỗi có tính dừng hoặc xấp xỉ như thế. Ta có thể chỉ ra điều này như sau:

Giả thiết ta có biến ban đầu là X={x1, x2,..., xn}, sau khi đổi biến ta được biến mới, ký hiệu là XF={XF1,XF2,...,XFn-1, trong đó phép đổi biến được thực hiện như sau:

$$\left.\begin{aligned} x_2 - x_1 &= XF_1 \\ x_3 - x_2 &= XF_2 \\ &\cdots\cdots\cdots\cdots \\ x_n - x_{n-1} &= XF_{n-1} \end{aligned}\right\} \qquad (1)$$

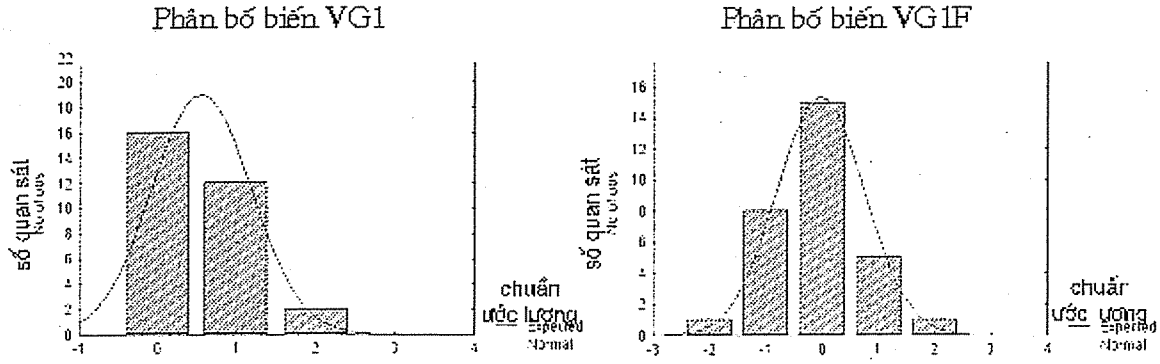
Nếu ta cộng 2 vế của các đẳng thức trên từ trên xuống dưới, rồi tách vế trái ra thành 2 thành phần ta sẽ có:

$$(x_n + x_2 + x_3 + \ldots + x_{n-1}) - (x_1 + x_2 + x_3 + \ldots + x_{n-1}) = XF_1 + XF_2 + XF_3 + \ldots + XF_{n-1} \qquad (3)$$

Người phản biện: GS.TSKH. **Nguyễn Đức Ngữ**

Nếu ta lấy trung bình của biểu thức (2) theo $(n-_1)$ thành phần ta sẽ thấy vế trái xấp xỉ bằng khi0 (≈ 0) mà x_1 và x_n khác nhau nhưng chúng có cùng bậc đại lượng ($\frac{x_1 - x_n}{n-1} \neq 0$, hoặc đúng bằng 0 h ($\equiv 0$) khi mà x_1 và x_n. Từ đó suy ra vế phải của biểu thức (2) cũng xấp xỉ bằng 0 hoặc đúng bằng 0. Chính nhờ đó mà biến mới XF có phân bố gần chuẩn /chuẩn, và do đó nó đồng thời cũng có tính dừng hoặc gần như dừng. Để chứng tỏ điều đó trên số liệu thực tế, ta có thể lấy trung bình bất kỳ một biến mới nào trong số các biến đã đổi biến HS1 cho ở phần II ta cũng thấy chúng có phân bố gần chuẩn.

Ta lấy ví dụ từ số liệu bão vào vùng 1, VG1, hình 1 dưới đây chứng tỏ biến mới VG1F (hình bên phải) có phân bố chuẩn hơn biến VG1 nguyên gốc (hình bên trái).



Hình 1. Phân bố biến VG1 (trái) và phân bố biến VG1F (phải)

b. Phép đổi biến HS1 đã làm tăng số nhân tố dự tuyển

- Nhờ phép đổi biến HS1 ta đã thu được một số lượng lớn các nhân tố dự báo dự tuyển, mà nếu không có nó chúng ta không thể thiết lập được tất cả các sơ đồ dự báo khả dĩ thỏa mãn yêu cầu về chất lượng sơ đồ dự báo. Nhìn vào cột 4 bảng 1 ta thấy một số yếu tố dự báo chỉ có dưới 10 nhân tố dự tuyển, thậm chí chỉ có vài ba nhân tố dự tuyển. Nhưng sau khi đổi biến thì ở cột 6 ta thấy đa phần đã có trên 10 nhân tố dự tuyển, thậm chí có yếu tố dự báo có tới trên 100 nhân tố dự tuyển. Đó là thực tế làm tăng số nhân tố dự báo có thể cho việc thiết lập các sơ đồ dự báo.

Tuy vậy, nhìn vào cột 6 bảng 1 ta thấy vẫn còn tới 7 sơ đồ chỉ có 2 đến 6 nhân tố dự tuyển. Với những sơ đồ này ta có thể thấy trước là khó mà thiết lập được sơ đồ dự báo với chất lượng và độ tin cậy có thể sử dụng được. Để cho số nhân tố dự tuyển của mỗi sơ đồ dự báo không quá ít, buộc chúng ta phải nới rộng cửa cho những nhân tố dự tuyển có hệ số tương quan $< |0,4|$ lọt vào đội ngũ những nhân tố dự tuyển như sẽ nói đến ở phần dưới.

Bảng 1. Số nhân tố sơ tuyển cho từng sơ đồ dự báo có thể

STT	Yếu tố Dự Báo	Sơ đồ Dự Báo	số nhân tố dự báo có thể (sơ tuyển với $r \geq \|0,4\|$)			Ghi chú
			X	XF	tg số	
1	2	3	4	5	6	7
1	VG1	VG1	2	0	2	Do hạn chế số ký tự nên tên file và tên biến có khi phải bỏ bớt dấu gạch nối, song chúng được hiểu như nhau.
		VG1F	4	0	4	
2	VG1-11	vg111	6	0	6	
		vg111f	3	0	3	

STT	Yếu tố Dự Báo	Sơ đồ Dự Báo	số nhân tố dự báo có thể (sơ tuyển với $r \geq \|0,4\|$)			Ghi chú
			X	XF	tg số	
3	VG2	vg2	8	13	21	
		vg2f	13	39	52	
4	VG2-10	VG210	30	16	46	
		VG210F	32	37	69	
5	VG3	VG3	6	0	6	
		VG3F	2	9	11	
6	VG3-9	VG39	2	20	22	
		VG39F	4	12	16	
7	VN	VN	6	0	6	
		VNF	9	15	24	
8	VNC3	VNC3	2	1	3	
		VNC3F	3	10	13	
9	VN-8T	VN8T	9	0	9	
		VN8TF	17	14	31	
10	VN-10	VN10	6	8	14	
		VN10F	34	12	46	
11	vg4	vg4	29	9	38	
		vg4f	6	40	46	
12	VG4-7	VG47	7	8	15	
		VG47F	1	9	10	
13	VG6	VG6	24	37	61	
		VG6F	47	87	134	
14	VG6C3	VG6C3	24	3	27	
		VG6C3F	62	76	138	
15	VG6-8T	VG68T	11	3	14	
		VG68TF	28	67	95	
16	VG6-8	VG68	14	6	20	
		VG68F	8	19	27	
Tổng số		32	459	570	1029	

c. Phép đổi biến HS1 làm tăng gấp đôi số sơ đồ dự báo cho cùng một yếu tố dự báo

- Nhờ phép đổi biến mà mỗi yếu tố dự báo có thể có tới 2 sơ đồ dự báo riêng rẽ (cột 3 bảng 1), làm tăng gấp đôi cơ hội lựa chọn sơ đồ dự báo có chất lượng tốt hơn, cụ thể là khi cho yếu tố dự báo ban đầu Y, ta có thể tạo ra yếu tố dự báo đã đổi biến YF, khi ấy ta sẽ có 2 sơ đồ dự báo tương ứng là:

$$\left.\begin{aligned} &1)\ Y = f(x_1, x_2,..., x_n; x_2-x_1, x_3-x_2,..., x_n-x_{n-1}),\\ &2)\ YF = f(x_1, x_2,..., x_n; x_2-x_1, x_3-x_2,..., x_n-x_{n-1}) \end{aligned}\right\} \quad (3)$$

Các biểu thức của (3) chứng tỏ đối với môi yêu tố dự báo Y và YF đều có thể chứa 2 loại nhân tố dự báo X và XF.

Khi biến (hay yếu tố) dự báo là nguyên gốc thì ta nhận được ngay yếu tố dự báo Y như thông thường. Còn khi biên dự báo là YF, là biến đã biến đổi HS1, thì ta dùng phép tính ngược lại với phép đổi biến HS1 ta sẽ có được biến dự báo nguyên gốc. Vì $YF_n = y_n - y_{n-1}$, do đó $y_n = YF_n + y_{n-1}$. Thí dụ: vào năm thứ (n-1) ta dự báo "số cơn bão sẽ xảy ra vào

năm thứ n", ta dự tính được Yn cho năm thứ n nào đó, chẳng hạn YFn=5, có nghĩa là $y_n - y_{n-1}=5$, trong đó y_{n-1} là số liệu đã biết, giả dụ là 2, ta có $y_n = 5 + y_{n-1}=5 + 2=7$, là trị số của yếu tố dự báo năm thứ n, ta sẽ phát báo: "Năm thứ n số cơn bão dự báo sẽ là 7 cơn".

Nếu không biến đổi ngược trở lại để được biến dự báo nguyên gốc ta vẫn có thể dùng chúng làm dự báo với ý nghĩa so sánh với năm trước đó. Thí dụ: vào năm thứ (n-1) ta đi dự báo hoạt động bão cho năm thứ n, được $YF_n=5$, ta có thể phát báo: "Năm thứ n số cơn bão sẽ nhiều hơn năm trước là 5 cơn" (mà không cần làm phép biến đổi về nguyên gốc như trên). Trong thực tế có những trường hợp người dùng thông tin dự báo không phân biệt được nhiều bão hay ít bão nếu ta không so sánh với năm trước đó, chứ không phải là so sánh với trung bình nhiều năm (vì người ta nhớ rõ số cơn bão vừa xảy ra năm trước hơn là số cơn bão trung bình nhiều năm).

- Nhờ phép đổi biến HS1 ta có 2 sơ đồ dự báo cho cùng một yếu tố dự báo nên ta có thêm cơ hội làm dự báo tổ hợp: đối với mỗi yếu tố dự báo ta có thể có 2 sơ đồ dự báo riêng rẽ/độc lập, như chỉ ra trên cột 3 bảng1, nên ta có cơ hội làm dự báo tổ hợp. Do tính riêng rẽ của các sơ đồ dự báo nên ta có thể coi chúng độc lập với nhau, nhờ đó ta có khả năng làm dự báo tổ hợp từ 2 dự báo thành phần độc lập, mà theo lý thuyết đã được chứng minh ở công trình [9, 10] của chúng tôi thì dự báo tổ hợp từ 2 dự báo thành phần độc lập sẽ có khả năng gia tăng chất lượng dự báo của mô hình. Tuy nhiên, để có thể dùng làm dự báo thành phần cho dự báo tổ hợp, chất lượng của các dự báo thành phần tốt nhất là phải tương đương. Trong trường hợp chất lượng các sơ đồ chênh lệch nhau nhiều, ta sẽ dễ dàng lựa chọn sơ đồ tốt nhất trong 2 sơ đồ đó.

Theo cách này thì mỗi yếu tố dự báo sẽ có 2 sơ đồ dự báo tương ứng, nên ứng với 16 yếu tố dự báo ta sẽ có tất cả 32 sơ đồ dự báo.

Ở đây tính từ "khả năng" hoặc "có thể" được dùng để chỉ nghĩa tương đối hay chưa chắc chắn của sự kiện. Cụ thể ở đây không phải cứ có nhiều nhân tố dự báo dự tuyển (có thể) là chắc chắn thu được sơ đồ dự báo có ý nghĩa/chất lượng cao, hoặc ít nhân tố dự tuyển sẽ không thu được sơ đồ dự báo có ý nghĩa. Tất cả chỉ là có thể ! Thuật ngữ này được dùng cả với trường hợp tổ hợp dự báo theo từng cặp sơ đồ, không phải cứ có các dự báo thành phần là tổ hợp dự báo cho ngay chất lượng khả quan hơn, mà có thể bị kéo chất lượng xuống mức trung bình cộng.

d. Những hạn chế kèm theo của phép đổi biến

Nhờ phép đổi biến HS1 mà số nhân tố dự báo có thể được tăng lên đáng kể, như liệt kê ở bảng 1, trong đó nhiều trường hợp số nhân tố dự báo có thể lên đến hàng chục, thậm chí hàng trăm như VG6F có tới 134 nhân tố dự báo có thể, hay VG6C3F có đến 138 nhân tố dự báo có thể. Số biến quá lớn sẽ làm cho việc tính toán trở lên rất phức tạp. Thực tế ngày nay người ta xây dựng những phần mềm thống kê cho phép xử lý tới hàng nghìn biến, thậm chí lên đến hàng vạn biến, nhưng số biến quá lớn thường phát sinh các vấn đề sau:

- Thứ nhất là vấn đề quá tải của phần mềm thống kê hoặc bộ nhớ máy tính, chẳng hạn phần mềm STATISTICA phiên bản thấp không chạy được khi số biến 100.

- Thứ hai là số biến quá lớn so với số trường hợp, dẫn đến ma trận hiệp phương sai suy biến và bài toán không giải được bằng cách thông thường.

- Thứ ba là vấn đề đa cộng tuyến (multicollinearity) hoặc cộng tuyến (collinearity), cũng làm cho ma trận hiệp phương sai bị suy biến, không giải được bằng cách thông thường, buộc phải xử lý riêng mất rất nhiều công sức.

2. Quan hệ giữa các nhân tố dự tuyển trong các nhóm chỉ số khí hậu và vấn đề đa cộng tuyến

a. Quan hệ bên trong giữa những chỉ số khí hậu

Những chỉ số khí hậu được sử dụng trong phân tích nhân tố dự báo có thể từ những tác giả khác nhau, nhưng chúng có nhiều điểm tương đồng, được tạo ra từ cùng nguồn yếu tố cơ bản là nhiệt, áp và gió, nên cũng có hệ số tương quan rất cao, thậm chí đạt đến mức gần như cộng tuyến hoàn hảo (perfect collinearity), làm cho độ dư thừa thông tin

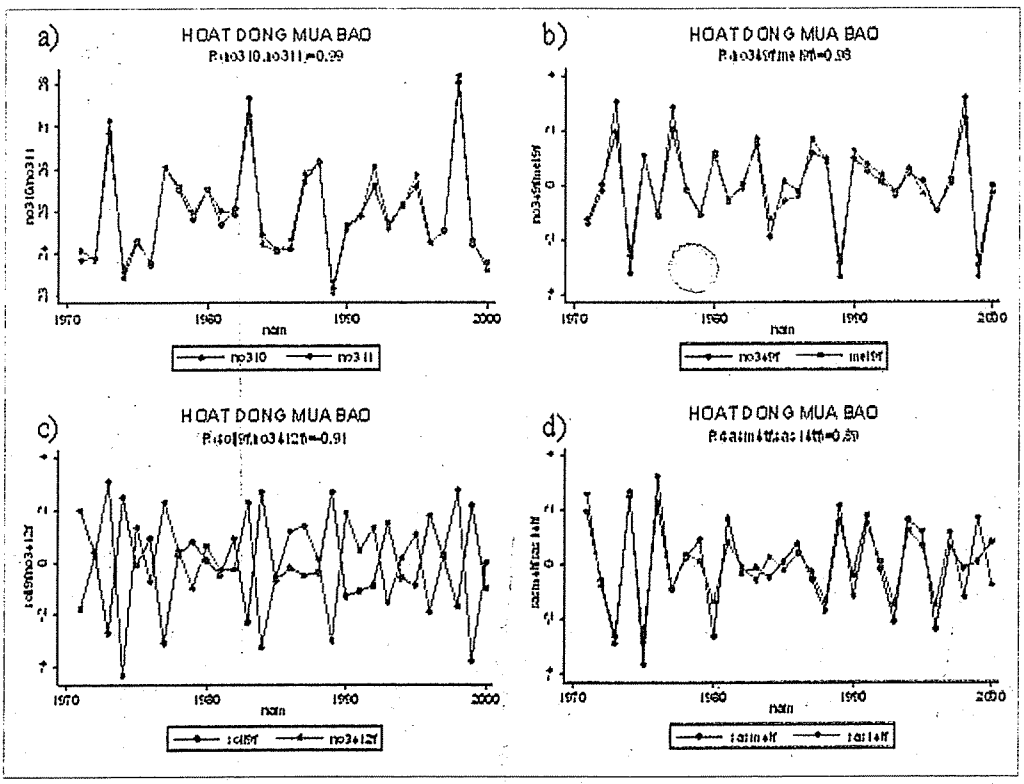
lớn, cũng dẫn đến hiện tượng ma trận suy biến, mà quan trọng nhất là làm cho các hệ số của phương trình dự báo không ổn định (instable).

Quan hệ chặt chẽ nhất là giữa các chỉ số thuộc các nhóm NINO (NINO12, NINO3, NINO4 và NINO34), các chỉ số MEI, SOI, ONI, TNI (mọi ký hiệu được cho ở Phụ lục).

Thí dụ điển hình như các cặp biến được rút ra từ VG6 dưới đây:

- Giữa những chỉ số thuộc cùng nhóm, cùng chỉ số nhưng khác tháng nhau: hệ số tương quan r(tni7, tni8)=0.99, r(no310, no311)=0.99 (được biểu diễn trên hình 2a),...

- Giữa 2 chỉ số khác nhau nhưng cùng nhóm chỉ số: hệ số tương quan r(mei9f,oni8f)=0.98, r(no349f, mei9f)=0.98 (được biểu diễn trên hình 2b), r(mei10f, no349f)=0.98, r(no310, no1211)=0.91, r(soij9f, no3412f)=-0.91 (được biểu diễn trên hình 2c), r(soij9f, no311f)=-0.91, ...



Hình 2. Tương quan giữa các chỉ số khí hậu

- Đối với các chỉ số thuộc nhóm gió mùa thì giữa các tháng có quan hệ chặt, nhưng kém nhóm NINO, ngay giữa chỉ số SAS1 và SASM có quan hệ khá chặt cũng chỉ ở mức <0.9, như r (sasm4tf, sas14tf)=0.89 (được biểu diễn trên hình 2d).

Đặc biệt là nếu ta dùng hồi quy một nhân tố để dự báo trước vài ba tháng một chỉ số ENSO cũng có thể đạt độ chính xác cao§, thí dụ như dùng chỉ số SOI tháng 9 để dự báo cho NINO3.4 tháng 12 với phương sai khá cao, là 0.83, theo phương trình sau:

$$no3412f = -0.03712 - 1.11648 * soij9f$$

Trong đó no3412f là NINO3.4 tháng 12 đã đổi biến theo HS1t, soij9f là chỉ số SOIJ tháng 9 đã đổi biến theo HS1.

b. Xử lý đa cộng tuyến

- Việc dùng phép đổi biến HS1 tạo ra nhiều biến dự tuyển có khả năng tương quan với nhau sẽ gây ra đa cộng tuyến, đồng thời những chỉ số khí hậu nói ở tiểu mục trên cũng chứng tỏ ở những mô hình thống kê sử dụng các chỉ số khí hậu làm nhân tố dự báo, khả năng xảy ra đa cộng tuyến là vấn đề hệ trọng cần được đề cập và xử lý. Ngoài ra, tập dữ liệu kích thước nhỏ cũng dễ dẫn đến đa cộng tuyến.

- Lợi điểm của đa cộng tuyến là làm tăng phương sai của ước lượng hồi quy ổn định như thí dụ của Robert M. Lynch và Brian Kim [8] đưa ra, nhưng là cực kỳ hiếm gặp, mà chủ yếu là tác hại. Tác hại của đa cộng tuyến trước hết là làm cho ma trận hiệp phương sai trong phép hồi quy bị suy biến hay kỳ dị, gọi là "tính cộng tuyến kỳ dị" (singular collinearity), không giải được hoặc phải giải bằng phượng pháp đặc thù, gây mất nhiều công sức. Sau nữa nó có thể làm tăng phương sai của ước lượng hồi quy một cách giả tạo và làm cho các hệ số của phương trình hồi quy không ổn định, thậm chí đổi dấu trong những trường hợp cận kỳ dị và không kỳ dị (near-singular và non -singular collinearity).

- Để kiểm tra đa cộng tuyến ở các phần mềm thống kê khác nhau có thể cho kết quả khác nhau nên chúng ta buộc phải dùng phép thử và xem xét trực tiếp để giữ lại đúng biến tốt nhất và bỏ đi biến tồi nhất. Cách xử lý trong các phần mềm khác nhau có thể cho chất lượng giảm đi so với thông thường, thí dụ trong phần mềm STATISTICA tính cộng tuyến được xử lý bằng "hồi quy xống cao" (ridge regression), mà hồi quy này cho ta phương sai thấp hơn thông thường; còn phần mềm STATA lại lọc bỏ theo thứ tự biến, có thể cũng dẫn đến thiếu chuẩn xác vì có khi bỏ đi biến tốt nhất. Trong thực tế, sau khi hồi quy đều phải kiểm tra bằng nhân tố tăng phương sai VIF (Variance Inflation Factor)

$$VIF = \frac{1}{1-r_i^2} \qquad (4)$$

Trong đó là hệ số tương quan giữa biến độc lập i với ước lượng của nó trên các biến độc lập còn lại của phương trình hồi quy. Lưu ý là r_i ở đây không phải là hệ số tương quan riêng phần của cặp biến độc lập trong phương trình hồi quy.

Thường các phần mềm thống kê đều lấy ngưỡng tolerance r_i^2 =0.1 làm ngưỡng đa cộng tuyến, vì khi ấy ma trận hiệp biến trở thành kỳ dị. Ngưỡng đó tương ứng với VIF =10, nghĩa là nếu VIF $\geq$10 hay tolerance $\leq$ 0.1 thì ma trận hiệp phương sai kỳ dị, không giải được. Đây là những trường hợp đa cộng tuyến kỳ dị (hay đa cộng tuyến hoàn hảo) và máy thường thông báo ma trận hiệp phương sai xấu (ill-conditinned).

Những trường hợp đa cộng tuyến gần kỳ dị và không kỳ dị các phần mềm máy tính thường bỏ qua, vì vậy buộc ta phải kiểm tra và xử lý theo kinh nghiệm thực tế.? đây kinh nghiệm của chúng tôi là phải kiểm tra cả VIF và sự đổi dấu của hệ số phương trình hồi quy. Vì nếu chỉ riêng VIF ở mức không kỳ dị mà không gây ra tác động gì thì chẳng cần xử lý; còn chỉ riêng hiện tượng đổi dấu của phương trình hồi quy thì có thể không do nguyên nhân đa cộng tuyến. Tóm lại là phải kiểm tra đồng thời 2 điều kiện sau:

+ VIF $>$3;

+ Hệ số b của phương trình hồi quy đổi dấu.

Còn với VIF$\leq$ 3, thường là không có đa cộng tuyến ảnh hưởng đến phương trình hồi quy đã thu được. Những bất ổn định như đổi dấu hệ số phương trình hồi quy có thể do những nguyên nhân khác như do biến đó không có ý nghĩa thống kê khi đưa vào phương trình hồi quy.

Sau khi xác định được biến đa cộng tuyến chúng ta xử lý bằng cách bỏ bớt các biến độc lập đa cộng tuyến, đó là cách gạt bỏ được mọi ngờ vực về tính bất ổn định của sơ đồ dự báo do đa cộng tuyến gây ra. Trong công trình này mọi cách xử lý khả thi khác (như thay dữ liệu, thay đổi kích cỡ tập mẫu,...) đều không khả thi hoặc thiếu tin tưởng và không được áp dụng.

Đó là lý thuyết và kinh nghiệm của tác giả bài này muốn chia sẻ cùng bạn đọc, còn thực tế như nhiều người đã biết, xử lý nó "tốn không ít mồ hôi" và phải luôn luôn nhớ rằng: đừng tin cả vào máy tính!

3. Thiết lập các sơ đồ dự báo

Trong khi thiết lập các sơ đồ dự báo bằng phương trình hồi quy tuyến tính ta sẽ sử dụng phương pháp hồi quy lọc từng bước tối ưu. Để kiểm soát được quá trình lọc nhân tố ta phải đưa ra những chỉ tiêu như số nhân tố tối ưu hay tối đa, ngưỡng phương sai cần đạt, trong đó ngưỡng phương sai chính là chất lượng dự báo hindcast của sơ đồ hay mô hình dự báo.

a. Về chất lượng hay khả năng dự báo và số nhân tố dự báo

Khi thiết lập phương trình cho từng sơ đồ dự báo, các nhân tố được tuyển chọn bằng phương pháp từng bước tối ưu (stepwise regression) dựa trên hệ số tương quan giữa thực tế và ước lượng. Vì vậy người ta lấy R2, là phương sai, đặc trưng cho khả năng dự báo của sơ đồ dự báo. Để quá trình tuyển chọn nhân tố dự báo đảm bảo cho sơ đồ dự báo đạt được một mức độ chất lượng nhất định nào đó, chúng ta lấy R2 làm chỉ tiêu tuyển chọn nhân tố. Như trên đã nói, trong mô hình của ta mỗi yếu tố dự báo ta có 2 sơ đồ dự báo tương ứng, nên chúng ta đặt chỉ tiêu là đối với mỗi yếu tố dự báo, ta cố gắng chọn được ít nhất một sơ đồ có phương sai giải thích được khoảng 70% độ biến động của yếu tố dự báo (tức R2 $\geq$ 70%), với số nhân tố dự báo đưa vào $\leq$ 7 (tối đa là 7). Chỉ tiêu này không hề thấp mà là khá cao, vì để đạt được $R^2 \geq 70\%$, hệ số tương quan giữa thực tế và dự báo hindcast phải là ≥ 0.84.

Ở đây chúng ta không thể xác định số nhân tố tối đa cần chọn theo phương pháp biểu đồ diễn tả chất lượng tối đa bị chặn trên khi số nhân tố dự báo tăng dần lên, thường áp dụng cho các chuỗi lớn với số nhân tố tối đa có thể tới 12 hoặc hơn nữa .

Về mặt lý luận, việc chọn số nhân tố dự báo tối đa ở đây bằng 7 là dựa trên phương pháp mà Davis [4, 5] đưa ra. Như ta thấy, sơ đồ dự báo hindcast được thiết lập trên số liệu phụ thuộc, khi áp dụng vào dự báo trên số liệu độc lập trong tương lai, chất lượng hay kỹ năng dự báo thường thấp hơn, vì tương quan trong số liệu độc lập khác trong số liệu phụ thuộc, người ta gọi phần chênh lệch này là phần kỹ năng giả tạo SA , được biểu diễn dưới dạng hàm số:

$$S_A = \frac{m}{n}(1 - S_H)(1 - \frac{m}{n})^{-1} \qquad (5)$$

Trong đó m là số nhân tố dự báo, n là tổng số trường hợp trong tập hindcast, SH là kỹ năng dự báo hindcast, SH =R2 (variance). Kỹ năng dự báo thực tế S được xác định bằng hiệu số:

$$S = S_H - S_A \qquad (6)$$

Ở đây số trường hợp trong tập hindcast là 30, với số nhân tố tối đa m=7, ta có SA=8%, với số nhân tố dự báo bằng 5, ta có SA=5%, nghĩa là sai khác

a chất lượng dự báo hindcast và dự báo thực tế n số liệu độc lập có thể nằm trong khoảng 5 8% nếu số nhân tố nằm trong khoảng 5 7.

Thực tế ta còn thấy những người có kinh nghiệm trong dự báo hoạt động mùa bão như Johnny Chan còn lấy tới 8 nhân tố cho tập 30 trường hợp [3]. Kinh nghiệm cũng cho thấy những sơ đồ với vài ba nhân tố dự báo không thể cho chất lượng dự báo hindcast có R2 70%, mặc dù kỹ năng giả tạo SA giảm xuống.

b. Kiểm tra chéo chất lượng sơ đồ dự báo

Phương pháp kiểm tra chéo (cross-validation) thường dùng để ước lượng sai số tổng quát nhằm chọn mô hình tốt nhất- có sai số nhỏ nhất trong số các mô hình được thiết lập. Kiểm tra chéo cũng nhằm xác nhận độ tin cậy hay độ ổn định của mô hình, vì vậy ở đây phương pháp "bỏ ra một" (leave-one-out) được áp dụng để kiểm tra chéo chất lượng phương trình dự báo được thiết lập từ tập dữ liệu có dung lượng hạn chế, được hiểu như đánh giá trên số liệu độc lập. Vì số liệu dùng để thiết lập sơ đồ dự báo và số liệu dùng đánh giá giao nhau nên gọi là chéo hay giao nhau. Tuy vậy, khi áp dụng phương pháp này, mỗi tác giả thực hiện mỗi khác; ở đây chúng tôi sử dụng thuật toán jackknife kết hợp với xử lý ma trận để thục hiện kiểm tra chéo như sau:

1) Giả sử ta có một tập số liệu với số trường hợp là n (n quan sát), trước tiên ta xây dựng phương trình hồi quy với m nhân tố nào đó, ta quy định gọi là phtr(n), nhóm ký tự (n) dùng để chỉ số quan sát được sử dụng là n quan sát; hệ số tương quan của dự tính và thực tế là R; tập số liệu với m nhân tố đó tạo thành ma trận a. Sau đó ta có thể thực hiện lặp lại n lần thiết lập phương trình hồi quy cũng với m nhân tố đó, nhưng lần lượt mỗi lần bỏ ra 1 trường hợp, nghĩa là khi tính phương trình hồi quy dự báo ta chỉ sử dụng tập luyện gồm (n-1) quan sát, còn 1 quan sát bỏ ra làm số liệu kiểm tra, được xem như số liệu độc lập. Như vậy tuần tự sau n lần thiết lập được n phương trình hồi quy, ta ký hiệu phtr(n-1)1, phtr(n-1)2, ..., phtr(n-1)n , ta thiết lập được ma trận b theo m nhân tố với n phương trình hồi quy; đồng thời ta cũng bỏ ra được n trường hợp số liệu độc

lập tương ứng với chúng, tạo thành chuỗi số liệu ta gọi là chuỗi jackknife. Về thực chất chuỗi này chính là chuỗi quan trắc n quan sát ban đầu. Mỗi quan sát chỉ độc lập với một phương trình trong số n phương trình phtr(n-1)1, phtr(n-1)2, ..., phtr(n-1)n, và không độc lập với phương trình phtr(n). Mọi tính toán sau đó được xử lý trên ma trận a và b.

2) Dùng từng trường hợp số liệu độc lập tương ứng với từng phương trình dự báo đã được thiết lập để tính dự báo và sai số dự báo cho yếu tố dự báo. Như vậy ta sẽ có n trị số dự báo, tạo thành một chuỗi ta qui ước gọi là chuỗi dự báo jackknife. Tính hệ số tương quan giữa chuỗi jackknife với chuỗi dự báo jackknife, ký hiệu là Rjac. Tất nhiên ở đây ta có thể tính những sai số tổng quát như MSE (Mean squared error), RMSE (Root mean squared error), phương sai (estimated variance), SD (standard deviation).

3) So sánh hệ số tương quan R và Rjac, nếu chất lượng không giảm đáng kể thì sơ đồ được chấp nhận là có độ tin cậy cao, nếu giảm đáng kể thì xem thêm sai số dự báo SE (Standard Error) đối với từng nhân tố dự báo giữa hồi quy jackknife (SEjac) và hồi quy không jackknife (SE). Nếu ứng với nhân tố nào đó có SEjac quá khác biệt so với SE thì cần loại bỏ nhân tố đó và chọn nhân tố khác để thiết lập lại phương trình dự báo, sau đó lại lặp lại quá trình kiểm tra chéo nói trên; nếu cuối cùng không có sơ đồ nào thỏa mãn điều kiện kiểm tra chéo thì sơ đồ dự báo có thể bị loại bỏ hoặc bị ghi nhận rằng đó là sơ đồ có độ tin cậy hạn chế.

Thực nghiệm cho thấy nếu quá trình thiết lập phtr(n) bằng phương pháp từng bước tối ưu có độ tin cậy thống kê cao thì các hệ số trung bình của n phương trình phtr(n-1)1, phtr(n-1)2, ..., phtr(n-1)n gần đúng bằng các hệ số của phương trình phtr(n), và ý nghĩa kiểm tra chéo bằng phương pháp jackknife là ý nghĩa đầy đủ.

Để quá trình kiểm tra chất lượng phương trình hồi quy của sơ đồ dự báo không quá phức tạp, ta có thể chỉ sử dụng hệ số tương quan R và Rjac thay vì sử dụng phương sai, sao cho giảm được quan hệ phi tuyến trong chỉ tiêu tuyển chọn. Từ hệ số tương quan R và Rjac ta đặt hệ số suy giảm tương quan kr như sau:

$$kr = \frac{R - R_{jac}}{R} \qquad (7)$$

Trong đó R là hệ số tương quan giữa thực tế quan trắc và dự báo hindcast, Rjac là hệ số tương quan giữa thực tế quan trắc và dự báo trên tập jackknife, kr chính là độ suy giảm hệ số tương quan khi chuyển từ số liệu phụ thuộc sang số liệu độc lập kiểm tra chéo "bỏ ra một". Hệ số kr càng lớn chứng tỏ sự sai khác càng lớn giữa dự báo trên số liệu phụ thuộc và số liệu kiểm tra chéo, nghĩa là chất lượng dự báo của phương trình hồi quy càng kém ổn định.

Vì R và Rjac là hệ số tương quan giữa thực tế với dự tính hindcast và với dự tính kiểm tra chéo "bỏ ra một", nên thực tế chúng luôn luôn dương, đồng thời R ≥ Rjac, vì số quan sát của R luôn luôn lớn hơn của Rjac một quan sát. Nếu R>>Rjac thì $\frac{R_{Jac}}{R} \to 0$ và kr=1, khi ấy chứng tỏ mô hình hoàn toàn không ổn định. Còn khi $R_{jac} \to R$(hay Rjac=R) thì $kr \to 0$ (hay kr=0), khi ấy độ ổn định của mô hình đạt mức hoàn hảo.

c. Thiết lập các sơ đồ dự báo

1) Đối với những yếu tố dự báo nguyên gốc

Trên bảng 1 ta thấy có một số sơ đồ có số nhân tố sơ tuyển quá ít, chỉ từ 2 đến 4, hoặc 6 biến, dưới cả mức 7 biến mà ta định tuyển chọn đã được nói ở mục trên, đó là các sơ đồ VG1, VG1F, VG111, VG111F, VG3, VN và VNC3. Kinh nghiệm cho thấy từ những nhân tố sơ tuyển ít ỏi đó khó có thể thiết lập được sơ đồ dự báo thỏa mãn chỉ tiêu nêu ra ở mục 2, mà cần có sự bổ sung nhân tố sơ tuyển. Bằng phép thử chúng tôi đã thấy phán đoán đó là đúng và chúng tôi đã bổ sung những nhân tố sơ tuyển có hệ số tương quan với yếu tố dự báo $|0.4| > r \geq |0.35|$. Bằng cách đó số nhân tố dự tuyển của chúng tăng lên thành $15 \div 22$.

Sau khi đã loại các biến cộng tính, chạy hồi quy từng bước tối ưu, phương trình hồi quy thu được dựa trên việc thử hàng loạt tiêu chí về quan hệ tuyến tính giữa biến phụ thuộc và độc lập, tính dự báo hindcast cho yếu tố dự báo và dư sai/sai số dự báo hindcast (predict e, residual) trên biến phụ thuộc,

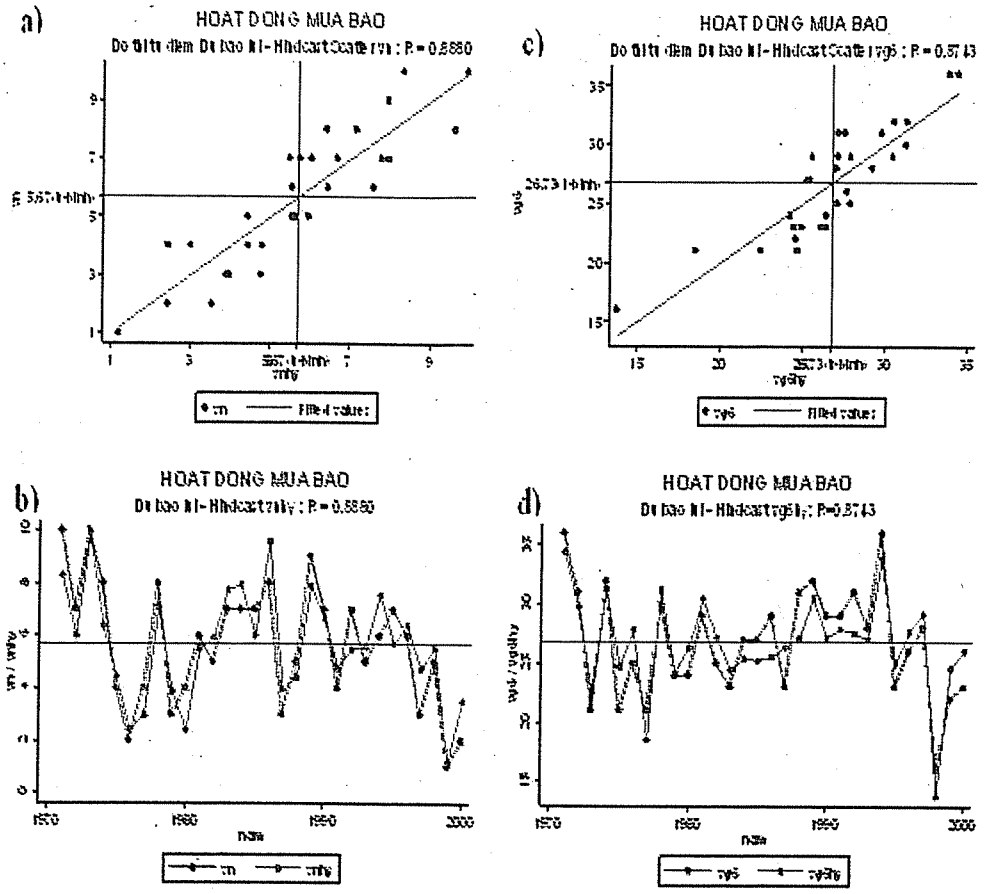
phân tích phương sai, phân tích dư sai, vẽ các biểu đồ phân bố, biểu đồ quan hệ giữa thực tế và dự đoán, ... Những tính toán và phân tích tương tự như thế được lặp lại trên 2 phần mềm khác nhau, nhằm phân tích và điều chỉnh những tính toán cần thiết khác nhau để ấn định việc chọn các biến độc lập cho mỗi sơ đồ dự báo, sau đó mới chuyển sang hồi quy với thuật toán jackknife nhằm thẩm định/kiểm tra chéo để xác nhận độ tin cậy của từng sơ sồ/phương trình dự báo thu được.

Bảng 2. Các sơ đồ dự báo cho các yếu tố dự báo nguyên gốc

Số TT	Sơ đồ dự báo	n	Các chỉ số khí hậu được chọn làm nhân tố dự báo				R	R^2	R^2 Adj	RMSE
			ENSO	Gió mùa	Dao động	Chỉ số Khác				
1	2	3	4	5	6	7	8	9	10	11
1	vg1	7		nasm3t swmo12	ao10f	pna10 wp3f tni3 tsa4	0.89	0.79	0.72	.33
2	vg111	7		nasm3t	pdo1 ao7f ao5 ao10f	tni4 tsa4	0.90	0.81	0.75	.21
3	vg2	7	soij4 no124f	wang6	epo4f	amm2f pna5f wh12f	0.91	0.83	0.77	.58
4	vg210	7	soij7		noi3 ao10f	pna5f gmss9f wp10 tsa1f	0.91	0.81	0.75	.42
5	vg3	7	nowe11f no124f	nasm7		amm10 wp9 wp11 tni4	0.88	0.77	0.70	1.07
6	vg39	7	no35f	swmo2f	nao4f	tsa10f pna9f tsa3f pacw6	0.86	0.73	0.64	.44
7	vn	7	soij6	nasm8f nasm7	epo4f	wp9 pna11f wp11	0.89	0.79	0.72	1.24
8	vnc3	7		wasm7f sasm6f	nao5 nao6	wp11 wp7 pna4f	0.86	0.73	0.65	.79
9	vn8t	7	soil7		epo3	wp11 wp9 pna11f tna9f amm9	0.86	0.75	0.67	1.36
10	vn10	7	soil7	wasm8f	epo10f	amm1 pna5f tsa1 gmss9f	0.88	0.77	0.70	.68
11	vg4	7	soij2f	sas26f	noi12 noi2	wh10 wp12 amm10f	0.88	0.78	0.71	.97
12	vg47	7	nowe5f	sasm7 easm6	noi5f	gmss9f wp2 sspo7f	0.89	0.77	0.70	.31
13	vg6	7		swmo5f	noi9 nao3f	wp10 tni6 gmss12f pna5f	0.87	0.76	0.69	2.60
14	vg6c3	7		sas19f swmo5f easm7		tni6 tni12 wh4 pna5f	0.89	0.8	0.73	1.98
15	vg68t	7	soij7	easm7 sas19		wp10 wp4 wp7 tni7	0.86	0.74	0.66	2.45
16	vg68	7	mei5 soij7		nao4 nao3f	tsa7f wp7 tni7f	0.88	0.78	0.71	.88

Sau khi tính toán, phân tích và chọn các nhân tố, kết quả xác lập các sơ đồ dự báo cho các yếu tố nguyên gốc được cho trên bảng 2, trong đó toàn bộ các sơ đồ dự báo đều thỏa mãn kỳ vọng ban đầu là phương sai giải thích được khoảng 70% độ biến động của yếu tố dự báo, thể hiện ở cột 9 bảng 2.

Để quan sát trực giác hơn chất lượng các sơ đồ dự báo trên số liệu phụ thuộc, ta có thể dẫn ra đây 2 sơ đồ điển hình, một cho quy mô địa phương là dự báo VN (Việt Nam) và một cho quy mô lớn khu vực là dự báo VG6 (TB TBD), đồng thời biểu diễn chúng dưới 2 dạng biểu đồ như trên hình 3.



Hình 3. Quan hệ giữa thực tế và dự báo cho VN (a, b), VG6 (c, d) (Đường song song với trục hoành là trung bình theo tập luyện, áp dụng cho tất cả các hình)

Trên hình 3a cho vùng VN ta thấy, hàng năm bão vào /ảnh hưởng trực tiếp đến khu vực ven biển và đất liền Việt Nam trung bình chỉ có 5, 67 cơn mỗi năm, năm ít nhất là 1 cơn, năm nhiều nhất tới 10 cơn. Kết quả ước lượng khá tốt, với hệ số tương quan 0,888, đường thẳng ước lượng gần trùng với đường chéo hình chữ nhật, về trung bình thực tế là 5,67, ước lượng cũng là 5, 67. Sang hình 3b ta thấy nhịp điệu biến thiên của thực tế và dự báo trên số liệu phụ thuộc khá phù hợp nhịp nhàng, ước lượng chính xác 1 (năm 1973) trong 2 năm có bão cực đại (năm 1971 và 1973) và có 1 năm bão cực tiểu (1999) cũng ước lượng đúng.

Trên hình 3c cho khu vực TB TBD ta thấy tụ điểm khá chụm, tương quan giữa thực tế và dự báo trên số liệu phụ thuộc cũng khá cao, tới 0,8743, ước lượng trung bình là 26, 73 cơn thì trung bình thực tế cũng là 26, 73 cơn. Vì thế mà nhịp điệu biến thiên giữa dự báo và thực tế trên hình 3d cũng rất phù hợp, đặc biệt trong đó 2 năm có số cơn bão biến động lớn gồm 1 cực đại vào năm 1971 và 1 cực tiểu vào năm 1998 đều dự báo được. Riêng cực đại năm 1994 dự báo hơi thấp, nhưng vẫn cùng xu thế trên trung bình.

2) Đối với những yếu tố dự báo đã đổi biến

Đối với những yếu tố dự báo là biến đã đổi biến, ta ký hiệu là YF, thì vấn đề phức tạp hơn nhiều, vì ngoài tiến trình thiết lập sơ đồ dự báo cho YF giống như với biến Y, ta còn phải chuyển đổi biến YF đã dự báo được trở về yếu tố dự báo nguyên gốc Y, ký hiệu là YFY, và tính các hệ số kèm theo một lần nữa.

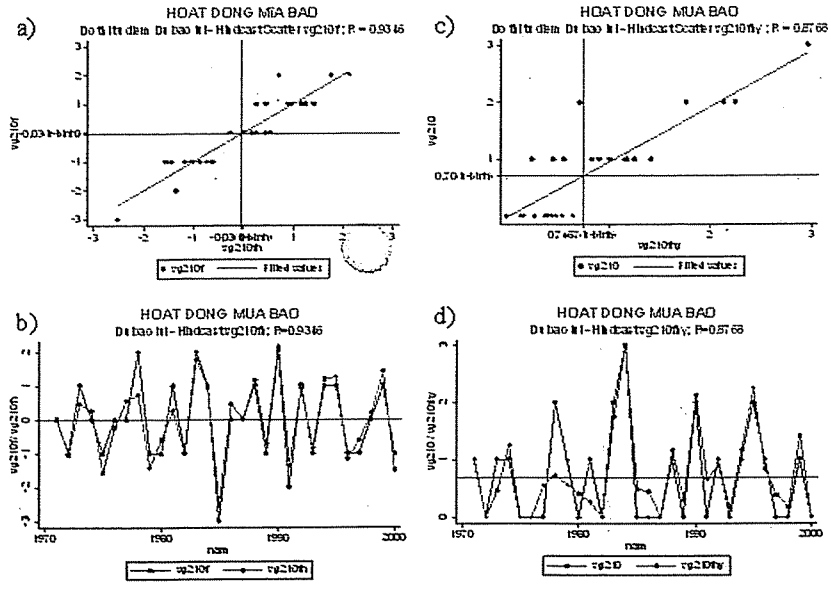
Bảng 3. Các sơ đồ dự báo cho các yếu tố dự báo đổi biến

Số TT	Sơ đồ dự báo	n	Các chỉ số khí hậu được chọn				Biến dự báo				
			ENSO	Gió mùa	Dao động	Chỉ số khác	YF			YFY	
							R	R^2	RMSE	Ry	Ry^2
1	2	3	4	5	6	7	8	9	10	11	12
1b	vg1f	7		nasm7f swmo12	nao8 nao3f	wp10 tni2f tsa3f	0.88	0.78	.46	0.86	0.73
2b	vg111f	7	soij12		nao1	wp9 wp7f tna9f pna7f sspo3f	0.89	0.79	.32	0.83	0.7
3b	vg2f	7			noi6 nao1 noi12 epo4f	tna1f pna3 tsa4f	0.86	0.74	1.03	0.75	0.56
4b	vg210f	6		scsm9f	noi6 epo2	wp10f gmss9f amm1f	0.94	0.87	0.49	0.88	0.76
5b	vg3f	7	nowe11f	sasm7f	noi3f noi7	wp7f wp9f wp7	0.87	0.76	1.41	0.79	0.63
6b	vg39f	7	no124f	swmo8f swmo8	nao7 nao4f	wp6f wp11	0.91	0.83	.49	0.84	0.71
7b	vnf	6	soij7	wasm8f sas17f	noi8 nao2f	wp7f	0.91	0.83	1.24	0.89	0.79
8b	vnc3f	6		sasm6f wasm7f swmo2f	noi8	wp9f pna1f	0.87	0.76	0.97	0.79	0.62
9b	vn8tf	7	soij7	wasm8f	noi8 nao2f noi12f noi3f	wp7f	0.93	0.87	1.16	0.91	0.82
10b	vn10f	7	soij5	wasm8 fscsm7	qbo7	gmss6f pacw12f amm1f	0.92	0.85	.78	0.85	0.73
11b	vg4f	7	soij2f	wasm6f	pdo1f	tna10f tna4f wh4f wp12f	0.91	0.83	1.38	0.75	0.56
12b	vg47f	7	nowe5 fnowe10f	sas26f		tsa9f wp2f pacw11f gmss9f	0.88	0.77	.41	0.81	0.65
13b	vg6f	7			nao5f noi9f	wp2f pna2 gmss12f	0.91	0.82	3.15	0.82	0.68

Như vậy buộc ta phải hai lần tính các hệ số tương quan và phương sai và các tham số khác. Trên cơ sở các quá trình phân tích và xử lý như đối với biến Y, thêm phần chuyển đổi YF về YFY, ta thiết lập được các sơ đồ dự báo cho các biến phụ thuộc đã đổi biến, được cho trên bảng 3. Để dễ phân biệt kết quả tính cho biến YFY, ta thêm ký tự "y" vào các hệ số, tức R thành Ry, R2 thành Ry2, nghĩa là đối với biến YF ta có các hệ số R, R2 (cột 8 và 9 bảng 3) khi sơ dồ dự báo YF, và Ry, Ry2 (cột 11 và 12 bảng 3) khi YF được chuyển đổi về YFY.

Chất lượng các sơ đồ đạt được ở bảng 3 khi chưa chuyển đổi về biến nguyên gốc được thể hiện ở cột 8-10, sau khi chuyển về biến nguyên gốc được thể hiện ở cột 11, 12. Số nhân tố được chọn ≤ 7, (cột 3), gồm các chỉ số thuộc 4 nhóm (cột 4-7). Ta thấy khi yếu tố dự báo là YF thì kết quả đều tốt và thỏa mãn chỉ tiêu phương sai giải thích được≥70% độ biến động của yếu tố dự báo (cột 9). Song khi chuyển đổi về biến nguyên gốc, thể hiện trên các cột 11-12 thì có tới 6 sơ đồ không đạt được phương sai ≥ 70%, đó là các sơ đồ VG2F, VG3F, VNC3F, VG4F, VG47F, vg6f.

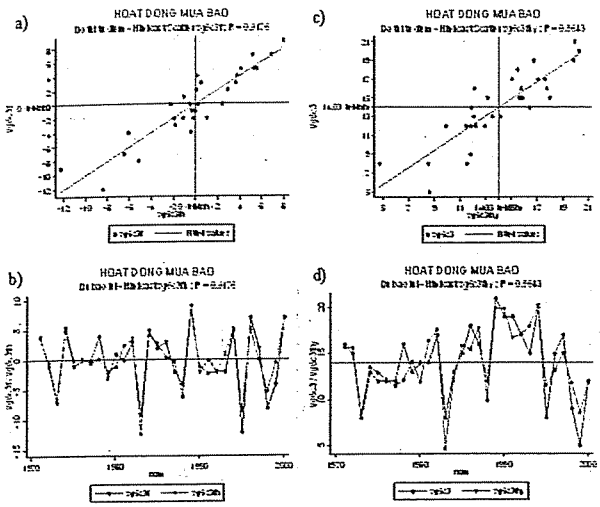
Để trực giác thấy rõ hơn chất lượng dự báo trên số liệu phụ thuộc các biến đã đổi biến, ta mô tả VG210F (số cơn bão tháng 10 vùng 2 đã đổi biến) trên biểu đồ ở hình 4. Hình 4a là biểu đồ tụ điểm quan hệ giữa VG210F thực tế và VG210FH dự báo hindcast, còn hình 4b biểu diễn quan hệ giữa VG210F thực tế và VG210FH –dự báo hindcast theo thời gian (1971-2000). Sau khi đổi biến về nguyên gốc ta có 2 hình tương ứng là 4c và 4d, trong đó hình 4c là biểu đồ tụ điểm quan hệ giữa VG210 thực tế và VG210FHY dự báo hindcast, hình 4d biểu diễn quan hệ giữa VG210 thực tế và VG210FHY hindcast theo thời gian (1971-2000).



Hình 4. Biểu đồ thực tế và dự báo cho VG210F (hình a, b), VG210FHY (hình c, d)

Số cơn bão vào khu vực miền Trung tháng 10 trung bình hàng năm chỉ có 0, 7 cơn, năm nhiều bão có tới 3 cơn, năm ít bão thì không có cơn nào. Trên hình 4a ta thấy các biểu đồ tụ điểm chụm đẹp, đường thẳng ước lượng gần như trùng với đường chéo hình chữ nhật, về trung bình ước lượng và thực tế trùng khớp nhau, đều xấp xỉ bằng 0 (=-0.03). Tương ứng với nó là biểu đồ 4b biểu thị quan hệ thực tế và dự báo trên biến đổi biến VG210F cũng dao động nhịp nhàng theo cùng một xu thế, với phương sai là 87%, trong đó dự báo đúng 2 trong 3 cực đại dương và 1 cực tiểu âm của VG210F.

Sau khi quy về biến nguyên gốc ta có đồ thị 4c và 4d, trong đó sự phù hợp giữa dự báo và thực tế có phần nào kém hơn, với phương sai chỉ còn 76%, tuy vậy cũng dự báo đúng 3 trong 4 năm nhiều bão (2-3 cơn). Nếu coi trị số dự báo dưới trung bình là không có bão thì trong 15 năm không có bão tháng 10 vùng 2 cũng dự báo đúng 13 năm.



Hình 5. Biểu đồ thực tế và dự báo cho VG6C3F (hình a, b), VG6C3FHY (hình c, d)

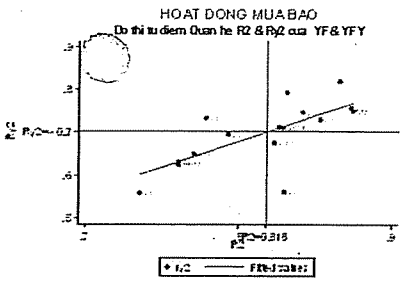
Ta dẫn ra thêm sơ đồ dự báo VG6C3F trên hình 5, trong đó hình 5a là biểu đồ tụ điểm biểu diễn quan hệ giữa VG6C3F thực tế và VG6C3FH hindcast; hình 5b biểu diễn quan hệ giữa VG6C3F thực tế và VG6C3FH hindcast theo thời gian (1971-2000). Sau khi đổi biến về nguyên gốc ta có 2 hình tương ứng là 5c và 5d, trong đó hình 5c là biểu đồ tụ điểm quan hệ giữa VG6C3 thực tế và VG6C3FHY dự báo hindcast, còn hình 5d biểu diễn quan hệ giữa VG6C3 thực tế và VG6C3FHY hindcast theo thời gian (1971-2000). Tương tự như ở hình 4, ở đây ta cũng quan sát thấy sự suy giảm phương sai từ 84% xuống còn 75%.

Nhìn vào bảng 3 và qua 2 ví dụ trên ta thấy rằng phương sai của yếu tố dự báo bị suy giảm đáng kể khi ta chuyển đổi yếu tố dự báo YF về nguyên gốc. Sự suy giảm chất lượng khi chuyển từ YF sang YFY do những nguyên nhân chủ yếu sau đây:

- Khi tính toán thiết lập sơ đồ dự báo ta chỉ sử dụng biến YF, còn khi chuyển YF về YFY ta phải sử dụng 2 biến là YF với số thành phần từ k đến n, ta ký hiệu , và biến YF với số thành phần từ (k-1) đến (n-1), ta ký hiệu tức là phụ thuộc vào 2 biến, làm cho độ chính xác bị suy giảm.

- Các biến YF là những số nguyên dương hoặc âm, còn biến YFY phải là biến dương (nếu lấy số cơn bão đến phần thập phân) hay nguyên dương (nếu lấy số cơn bão là số nguyên), buộc ta phải xấp xỉ khi loại bỏ số âm, làm cho độ chính xác bị suy giảm một lần nữa.

Để có thể hình dung ra sự suy giảm này, ta biểu diễn quan hệ giữa Ry2 với R2 và xấp xỉ bằng đường thẳng trên hình 6.



Hình 6. Sự suy giảm phương sai khi chuyển YF về YFY

Từ hình 6 ta nhận thấy về trung bình phương sai suy giảm 12%. Vậy khi chuyển đổi từ YF về YFY, về trung bình, muốn có phương sai ≥70% đối với YFY thì sơ đồ dự báo YF phải đạt được phương sai ≥ 81,8%. Đây là nhược điểm nữa của sơ đồ với yếu tố dự báo là biến YF, nên nói chung các sơ đồ với yếu tố dự báo YF khó đạt chất lượng cao so với các sơ đồ yếu tố dự báo nguyên gốc.

Cách dùng các sơ đồ dự báo đối với yếu tố dự báo YF trên đây có thể có 2 khả năng:

- Nếu ta phát dự báo theo cách so sánh số cơn bão nhiều hơn (YF > 0) hay ít hơn (YF < 0) giữa năm nay với năm trước thì ta không phải chuyển đổi YF về YFY, nghĩa là chất lượng sơ đồ dự báo tương ứng với cột 8 và 9 bảng 3.

- Nếu ta không phát báo theo cách trên, mà phát báo theo số cơn bão cụ thể hoặc so với trung bình nhiều năm thì buộc ta phải chuyển YF về YFY, và chất lượng sơ đồ dự báo sẽ thấp hơn, tương ứng với cột 11-12 bảng 3. *(còn nữa).*