Tạp chí Tin học và Điều khiển học, T.22, S.1 (2006), 84–98

NGHIÊN CỨU THỐNG KÊ ĐỘNG THÁI DÂN SỐ VIỆT NAM THỜI KỲ 1878 - 2003 QUA 150 MÔ HÌNH XU THẾ

VŨ HOÀI CHƯƠNG

Viện Công nghệ Thông tin, Viện KH&CN Việt Nam

Abstract. This article presents the computation and sorting results of 150 trend-function using Vietnam population's data from 1878 up to 2003. The suprised fact that the well-known functions in demography such as exponent function, logistic function and Gompertz function do not approximated the Vietnam population's changes appropriately. A question raised: can a trend-functions be found, which is suitable for description of several countries, or regions population? The results of correlation analysis at the end of the article give the hope of a positive answer.

Tóm tắt. Bài báo trình bày kết quả tính toán và xếp hạng 150 hàm xu thế theo dữ liệu dân số Việt Nam từ năm 1878 đến năm 2003. Điều bất ngờ là các hàm quen thuộc trong nhân khẩu học như hàm mũ, hàm logistic, hàm Gompertz lại không xấp xỉ tốt biến động dân số nước ta, mà cần đến các dạng hàm khác. Câu hỏi được đặt ra là: có thể tìm được các hàm xu thế chung, thích hợp với dân số nhiều nước, nhiều khu vực trên thế giới hay không? Kết quả phân tích tương quan ở cuối bài cho phép hy vọng rằng câu trả lời là có.

1. MỞ ĐẦU

Chúng tôi nghiên cứu các mô hình xu thế trên số liệu dân số Việt Nam thời kỳ 1878-2003 nhằm tìm ra những hàm xu thế thích hợp nhất và trên cơ sở đó để:

- Hiệu chỉnh các số liệu đã có nhưng chưa đáng tin cậy.
- Ước lượng các số liệu chưa biết.
- Dự báo trong tương lai.

Tận dụng các ưu thế của máy tính điện tử, của phương pháp hồi quy theo trọng số và phương pháp hiệu chỉnh bằng thăm dò ngẫu nhiên, chúng tôi đã tính toán theo 150 hàm xu thế khác nhau và xếp hạng theo thứ tự giảm dần *độ sai tiêu chuẩn của ước lượng* (SEE = Standard Error of Estimate), cũng tức là thứ tự của *sai số bình phương trung bình* (MSE = Mean Square Error), vì MSE = $(\text{SEE})^2$. Muốn tính các đại lượng này, ta lấy *tổng bình phương sai số* (SSE = Sum of Squares Error) chia cho bậc tự do ([14]):

$$\text{MSE} = \text{SSE}/(n-k) = \sum_{i=1}^{n}(Y_i - Z_i)^2/(n-k)$$

rồi tính $\text{SEE} = \sqrt{\text{MSE}}$, trong đó Y_i là giá trị quan sát, Z_i là giá trị ước lượng, n là số quan sát và k là số tham số cần ước lượng.

Phần tiếp theo của bài báo trình bày các kết quả hiệu chỉnh theo 3 hàm xu thế thích hợp nhất và ước lượng dân số nước ta từ năm 1880 đến năm 2010 theo trung bình có trọng số

của 12 hàm xu thế đầu tiên.

Chúng tôi cũng đã tính 150 hàm xu thế trên dữ liệu 5 nước ở các châu lục khác nhau, trong những thời kỳ khác nhau. Các kết quả tính toán có tương quan khá chặt chẽ với nhau và với dân số Việt Nam. Điều đó chứng tỏ rằng động thái dân số của các nước có những điểm tương đồng và do vậy có thể tìm được các hàm xu thế chung cho các quá trình nhân khẩu học.

2. NGUỒN SỐ LIỆU

Các số liệu dân số Việt Nam sau năm 1975 được lấy từ tài liệu chính thức của Tổng cục Thống kê ([22]).

Số liệu dân số Việt Nam trong những năm thuộc Pháp và những năm chiến tranh có nhiều chỗ không thống nhất giữa các tài liệu của Việt Nam ([21]), của Pháp ([4, 5, 7]), của Liên Xô cũ ([25]), của Liên Hiệp Quốc ([23]) và của Hoa Kỳ ([24]). Khi có khác biệt thì chúng tôi lấy giá trị trung bình mỗi năm.

Theo thông lệ, dân số được tính vào giữa năm, tức là ngày 1 tháng 7. Vì thế các thời điểm điều tra dân số của ta được chuyển đổi như sau: ngày 1-10-1979 thành 1979,25; ngày 1-4-1989 thành 1988,75; ngày 1-4-1999 thành 1998,75.

Các số liệu về dân số Việt Nam được ghi trong các cột đầu của Bảng 5, trong đó ba thời điểm điều tra dân số nói trên được in đậm.

3. CÁC DẠNG HÀM XU THẾ ĐÃ CHỌN

Chúng tôi chọn các hàm xu thế dạng

$$G(j, Y) = F(k, T); j = 1, ..., 6, k = 1, ..., 25,$$

trong đó Y ký hiệu dân số (nghìn người), X ký hiệu năm, T là khoảng thời gian tính từ một năm gốc X_g nào đó: $T = X - X_g$ (T phải dương và khác 1 để $\ln T, \ln\ln T$ có nghĩa). Trong các tính toán dưới đây $X_g = 1825$.

Như vậy tổng cộng có $6.25 = 150$ hàm xu thế. Đồ thị của một số hàm cơ bản có trong [6].

Sáu hàm $G(j, Y)$ là:

$$G(1, Y) = Y; G(2, Y) = 1/Y; G(3, Y) = Y^2; G(4, Y) = \sqrt{Y}; G(5, Y) = 1/\sqrt{Y}; G(6, Y) = \ln Y.$$

Hai mươi lăm hàm $F(k, T)$ được chia làm bốn nhóm:

a) Sáu hàm có hai tham số

$F(1, T) = a.T + c; F(2, T) = a/T + c; F(3, T) = a.T^2 + c; F(4, T) = a.\sqrt{T} + c;$
$F(5, T) = a.\ln T + c; F(6, T) = a.\ln\ln T + c.$

b) Bảy hàm có ba tham số

$F(7, T) = a + b.T + c/T; F(8, T) = a + b.T + cT^2; F(9, T) = a + b.T + c\sqrt{T};$
$F(10, T) = a + b.T + c.\ln T; F(11, T) = a + b.T + c.\ln\ln T;$
$F(12, T) = a + b.\ln T + c.(\ln T)^2; F(13, T) = a + b.\ln\ln T + c.\ln T.$

c) Sáu hàm dạng mũ

$F(14,T) = a.\exp(b.T) + c$; $F(15,T) = a.\exp(b/T) + c$;
$F(16,T) = a.\exp(b.T^2) + c$; $F(17,T) = a.\exp(b.\sqrt{T}) + c$;
$F(18,T) = a.\exp(b.\ln T) + c$; $F(19,T) = a.\exp(b.\ln\ln T) + c$.

d) Sáu hàm đa thức bậc hai bổ sung (có bốn tham số)

$F(20,T) = a + b.T + c.T^2 + d.T^3$; $F(21,T) = a + b.T + c.T^2 + d/T$;
$F(22,T) = a + b.T + c.T^2 + d.\sqrt{T}$; $F(23,T) = a + b.T + c.T^2 + d.T^{3/2}$;
$F(24,T) = a + b.T + c.T^2 + d.\ln T$; $F(25,T) = a + b.T + c.T^2 + d.\ln\ln T$.

4. MÔ TẢ SƠ LƯỢC VỀ THUẬT TOÁN TỔNG QUÁT

Cách tính toán mỗi hàm xu thế tuỳ thuộc vào dạng của $G(j,Y)$ và $F(k,T)$, theo các kỹ thuật của thống kê ([11,14]), hồi quy phi tuyến ([16]), các khái niệm về đường xu thế ([8,9]), công thức thực nghiệm ([10,13]) và ý tưởng hiệu chỉnh tham số ([19]). Bài này không trình bày được tất cả các thuật toán, nhưng có thể mô tả các bước sau đây:

a) Các tham số của mỗi hàm xu thế $G(j,Y) = F(k,T)$, viết tắt là hàm (j,k), được ước lượng sơ bộ bằng phương pháp bình phương tối thiểu. Trong trường hợp có các dạng hàm mũ $F(k,T), (k = 14, ..., 19)$ thì cần phải nội suy ([12]) để có được các giá trị trung gian.

b) Khi tính toán các phương trình hồi quy với $G(j,Y) = F(k,T), j > 1$, chúng tôi đưa vào các trọng số thích hợp, tỷ lệ nghịch với bình phương giá trị của đạo hàm (xem [1,17]):

$$w = \left[\frac{1}{G'(j,Y)}\right]^2.$$

Chẳng hạn với hàm $G(6,Y) = \ln Y$ thì

$$w_k = \left[\frac{1}{(\ln Y)'_k}\right]^2 = \left[\frac{1}{(1/Y)_k}\right]^2 = (y_k)^2.$$

Các kết quả tính toán cho thấy chỉ nhờ dùng trọng số như trên (chưa hiệu chỉnh ngẫu nhiên) mà sai số bình phương trung bình (MSE) giảm từ 1450,1581 xuống còn 1341,5292 (xem cột A và cột B Bảng 1), tức là khoảng 1,08 lần. Tuy vậy khi $j = 3$ thì trọng số đem lại kết quả tồi hơn.

Các cột 2-3 và 4-5 Bảng 2 cho thấy 5 hàm (j,k) có MSE giảm nhiều nhất (từ 6 đến 21 lần) do trọng số.

c) Các phương trình hồi quy được đưa về dạng hiện $Y = G^{-1}(F(k,T))$. Ví dụ: hàm $G(5,Y) = 1/\sqrt{Y} = F(12,T) = a + b.\ln T + c.(\ln T)^2$ được đưa về dạng $Y = 1/(a + b.\ln T + c.(\ln T)^2)^2$.

Cho các tham số a, b, c dao động ngẫu nhiên quanh giá trị tốt nhất hiện thời a_m, b_m, c_m của chúng và tính sai số bình phương trung bình MSE tương ứng. Nếu MSE giảm thì a, b, c nhận các giá trị mới. Quá trình hiệu chỉnh cho mỗi hàm gồm nhiều vòng lặp với các biên độ ngẫu nhiên. Nếu sau khoảng 2000 vòng lặp mà không thấy MSE giảm thì chuyển sang hàm tiếp theo.

Hiệu quả của phương pháp này rất lớn. Sau hiệu chỉnh ngẫu nhiên, tổng MSE giảm từ 1450,1581 (A) xuống 817,8312 (C) hoặc từ 1341,5292 (B) xuống 811,8931 (Z), tức là trung

bình A/C = 1,77 lần, B/Z = 1,65 lần (xem Bảng 1). Đặc biệt có những trường hợp MSE giảm đi hàng chục lần, và nhờ đó mà ta có được thứ tự chính xác của các hàm xu thế. Năm hàm (j,k) có MSE giảm nhiều nhất (từ 15 đến 59 lần) sau các bước dùng trọng số và hiệu chỉnh ngẫu nhiên được ghi trong các cột 4-9 ở Bảng 2.

Bảng 1. Các giá trị TổngMSE/$\text{MSE}_{(1,8)}$ theo 4 cách tính khác nhau

Cách tính	A	B	C	Z
Trọng số	không	có	không	có
Hiệu chỉnh	không	không	có	có
j = 1	257,4847	246,7114	245,0796	245,0771
j = 2	298,7724	200,5027	63,7395	60,0521
j = 3	544,2399	650,5277	360,3813	359,1255
j = 4	94,1337	90,4907	69,1625	69,0483
j = 5	137,7767	92,1060	42,0591	41,3468
j = 6	117,4508	61,1907	37,4092	37,2434
Tổng	1450,1581	1341,5292	817,8312	811,8931
Mức độ giảm MSE	A/B = 1,081 A/C = 1,773 A/Z = 1,786	B/Z =1,652	C/Z = 1,007	

Bảng 2. Các hàm có MSE giảm nhiều nhất

Thứ tự	Hàm (j,k)	Tỷ lệ A/B	Hàm (j,k)	Tỷ lệ A/C	Hàm (j,k)	Tỷ lệ B/Z	Hàm (j,k)	Tỷ lệ A/Z
1	(2,7)	21,46	(3,15)	58,72	(3,19)	20,34	(3,15)	58,90
2	(1,15)	6,93	(2,7)	28,73	(3,15)	17,12	(2,7)	28,73
3	(5,8)	6,87	(2,18)	16,24	(2,25)	16,39	(2,25)	16,39
4	(6,15)	6,17	(2,19)	16,12	(2,18)	16,19	(2,18)	16,24
5	(3,18)	5,69	(1,15)	14,99	(2,19)	16,05	(2,19)	16,19
Chú thích	do trọng số		do hiệu chỉnh từ A		do hiệu chỉnh từ B		do trọng số và do hiệu chỉnh	

d) Để đo được hiệu quả của các bước nói trên, ta tính hệ số biến thiên % giữa các cách tính A, B, C, Z bằng công thức $100.s/P$, trong đó P là trị số trung bình của 4 cách tính: $P = (Z_A + Z_B + Z_C + Z_Z)/4$;

s là độ lệch chuẩn tương ứng, tức là căn của phương sai mẫu:

$$s^2 = (Z_A^2 + Z_B^2 + Z_C^2 + Z_Z^2)/4 - P^2.$$

Hệ số biến thiên lớn chứng tỏ các cách tính A, B, C, Z có kết quả rất khác biệt nhau, cũng có nghĩa là hiệu chỉnh rất hữu hiệu. Khi cả 4 cách tính đều cho kết cục hoàn toàn như nhau thì hệ số biến thiên bằng 0.

Trong Bảng 3, hàm (3, 19) là một trường hợp đặc biệt. Khi dùng trọng số, MSE không những không giảm mà còn tăng lên gần 13 lần ($B/A = 12,7$ hay là $A/B = 0,08$). Nhưng quá trình hiệu chỉnh ngẫu nhiên sau đó đã giảm MSE được $B/Z = 20,34$ lần, cho nên kết cục MSE vẫn giảm được $A/Z = 1,61$ lần so với lúc khởi đầu và do đó xếp thứ 80/150.

Bảng 3. Các hàm có độ biến thiên lớn nhất

Thứ tự	Hàm (j,k)	Hệ số biến thiên	Tỷ lệ và thứ tự A/B	Tỷ lệ và thứ tự A/C	Tỷ lệ và thứ tự B/Z	Tỷ lệ và thứ tự A/Z
1	(2,7)	87,56	21,46 (1)	28,73 (2)	1,34 (61)	28,73 (2)
2	(3,15)	79,72	3,44 (13)	58,72 (1)	17,12 (2)	58,90 (1)
3	(3,19)	76,31	0,08 (150)	1,61 (76)	20,34 (1)	1,61 (80)
4	(1,15)	64,89	6,93 (2)	14,99 (5)	2,16 (40)	14,99 (7)
5	(3,18)	62,53	5,69 (5)	13,38 (7)	2,80 (30)	15,91 (6)

5. KẾT QUẢ TÍNH TOÁN VÀ THỨ TỰ CÁC HÀM

Bảng 4 cho thấy thứ tự sắp xếp theo Độ sai tiêu chuẩn của ước lượng (SEE) tăng dần (theo cách tính chính xác nhất là Z).

Đa thức bậc hai $Y = a + bT + cT^2$ $(j = 1, k = 8)$ được chọn làm chuẩn vì tính toán đơn giản và có $\mathrm{MSE}_{(1,8)}, \mathrm{SEE}_{(1,8)}$ bất biến đối với năm gốc X_g. Hàm này xếp hạng thứ 47 và được in đậm. Cột 2 của bảng ghi $\mathrm{SEE}_{(j,k)}$, tức là căn bậc 2 của $\mathrm{MSE}_{(j,k)}$. Cột 3 ghi các giá trị của hệ số xác định R^2. Cột 4 ghi tỷ lệ giữa $\mathrm{MSE}_{(j,k)}$ và $\mathrm{MSE}_{(1,8)}$ của đa thức bậc 2. Đây là một số đo tương đối, có thể dùng để so sánh các bộ dữ liệu khác nhau. Cột 5 ghi hệ số biến thiên % giữa các cách tính A, B, C, Z. Một số hàm dạng $(1, k)$ tức là $G(1, Y) = Y = F(k, T)$ có hệ số biến thiên bằng 0 bởi vì có trọng số $w = 1$ và không thể hiệu chỉnh được. Hai cột cuối cùng của Bảng 4 là các giá trị j và k tương ứng.

Bảng 4. Kết quả tính toán và xếp hạng các hàm xu thế theo SEE
(Độ sai tiêu chuẩn của ước lượng)

Thứ tự	$\mathrm{SEE}_{(j,k)} = \sqrt{\mathrm{MSE}_{j,k}}$	Hệ số xác định R^2	Tỷ lệ $\mathrm{MSE}_{(j,k)}/\mathrm{MSE}_{(1,8)}$	Hệ số biến thiên	j	k
1	757,225	0,99883	0,1990	35,08	5	20
2	791,484	0,99872	0,2185	56,21	2	25
3	821,808	0,99862	0,2344	29,03	2	20
4	828,435	0,99860	0,2382	39,52	2	23
5	833,404	0,99858	0,2411	50,45	2	22
6	861,225	0,99849	0,2575	55,24	2	24
7	921,250	0,99827	0,2946	39,06	5	23
8	951,138	0,99816	0,3140	56,41	2	21
9	1062,481	0,99770	0,3919	79,72	3	15
10	1090,543	0,99758	0,4128	20,86	6	20
11	1098,922	0,99754	0,4192	38,81	5	22
12	1132,562	0,99739	0,4453	17,32	2	8
13	1201,567	0,99706	0,5012	44,27	5	21
14	1236,376	0,99688	0,5306	64,89	1	15
15	1258,253	0,99677	0,5496	0,00	1	21
16	1280,348	0,99666	0,5691	10,52	1	19
17	1281,200	0,99665	0,5698	22,43	6	23

Thứ tự	$SEE_{(j,k)} = \sqrt{MSE_{j,k}}$	Hệ số xác định R^2	Tỷ lệ $MSE_{(j,k)}/MSE_{(1,8)}$	Hệ số biến thiên	j	k
18	1289,725	0,99661	0,5774	0,00	1	25
19	1301,957	0,99654	0,5884	1,38	3	20
20	1303,154	0,99654	0,5895	0,00	1	24
21	1303,827	0,99653	0,5901	34,12	5	24
22	1328,685	0,99640	0,6128	0,00	1	22
23	1383,314	0,99610	0,6643	0,00	1	23
24	1407,907	0,99596	0,6881	62,53	3	18
25	1440,807	0,99577	0,7206	20,27	6	22
26	1447,544	0,99573	0,7274	5,84	4	20
27	1455,110	0,99568	0,7350	19,53	4	15
28	1465,743	0,99562	0,7458	0,00	1	20
29	1465,867	0,99562	0,7459	32,54	2	16
30	1469,552	0,99560	0,7497	31,42	5	25
31	1476,566	0,99556	0,7568	23,04	6	25
32	1478,392	0,99554	0,7587	13,11	2	9
33	1484,580	0,99551	0,7651	21,20	6	24
34	1551,194	0,99509	0,8353	4,64	4	23
35	1558,162	0,99505	0,8428	22,13	6	21
36	1565,124	0,99501	0,8503	0,16	4	9
37	1607,104	0,99473	0,8966	5,22	4	24
38	1610,784	0,99471	0,9007	0,56	4	10
39	1622,387	0,99463	0,9137	27,03	5	16
40	1622,709	0,99463	0,9141	11,44	2	10
41	1626,758	0,99461	0,9186	5,26	4	25
42	1639,802	0,99452	0,9334	5,29	4	21
43	1650,475	0,99445	0,9456	3,36	4	22
44	1655,760	0,99441	0,9517	0,93	4	11
45	1656,810	0,99440	0,9529	37,40	1	18
46	1659,474	0,99439	0,9560	26,29	3	17
47	**1697,270**	**0,99413**	**1,0000**	**0,00**	**1**	**8**
48	1706,907	0,99406	1,0114	9,46	2	11
49	1710,952	0,99403	1,0162	1,76	4	8
50	1717,434	0,99399	1,0239	51,60	5	8
51	1768,380	0,99362	1,0855	52,93	6	15
52	1773,282	0,99359	1,0916	17,25	3	22
53	1776,540	0,99357	1,0956	15,71	4	19
54	1788,583	0,99348	1,1105	0,60	6	13
55	1793,149	0,99345	1,1162	18,22	6	16
56	1810,949	0,99331	1,1384	8,72	3	23
57	1842,192	0,99308	1,1781	14,52	4	18
58	1844,101	0,99307	1,1805	2,22	4	7
59	1849,789	0,99302	1,1878	2,03	6	12
60	1858,173	0,99296	1,1986	11,27	1	17

Thứ tự	$\text{SEE}_{(j,k)} = \sqrt{\text{MSE}_{j,k}}$	Hệ số xác định R^2	Tỷ lệ $\text{MSE}_{(j,k)}/\text{MSE}_{(1,8)}$	Hệ số biến thiên	j	k
61	1913,307	0,99254	1,2708	9,64	2	12
62	1927,301	0,99243	1,2894	19,08	3	14
63	1929,128	0,99241	1,2919	87,56	2	7
64	1962,738	0,99215	1,3373	11,86	2	13
65	1968,430	0,99210	1,3450	12,11	4	16
66	1975,439	0,99204	1,3546	21,46	3	21
67	1975,905	0,99204	1,3553	16,83	4	17
68	1978,346	0,99202	1,3586	2,79	4	12
69	1988,847	0,99194	1,3731	3,83	6	7
70	1993,932	0,99190	1,3801	53,04	2	14
71	2003,537	0,99182	1,3935	44,38	5	9
72	2015,087	0,99172	1,4096	24,51	6	19
73	2016,393	0,99171	1,4114	3,13	1	14
74	2019,551	0,99169	1,4158	7,13	6	11
75	2027,225	0,99162	1,4266	8,38	6	10
76	2027,319	0,99162	1,4267	40,19	5	14
77	2031,612	0,99159	1,4328	8,53	6	1
78	2038,880	0,99153	1,4431	10,92	6	9
79	2042,549	0,99149	1,4482	17,70	4	14
80	2045,172	0,99147	1,4520	18,00	6	8
81	2045,914	0,99147	1,4530	26,07	6	18
82	2048,842	0,99144	1,4572	28,03	6	14
83	2082,456	0,99116	1,5054	28,57	6	17
84	2097,742	0,99103	1,5276	41,68	5	10
85	2126,781	0,99078	1,5702	0,39	4	3
86	2134,712	0,99071	1,5819	7,72	1	16
87	2143,954	0,99063	1,5956	13,19	3	24
88	2147,056	0,99060	1,6002	39,88	5	11
89	2150,165	0,99058	1,6049	14,61	3	25
90	2152,647	0,99055	1,6086	33,11	5	15
91	2160,160	0,99049	1,6198	42,90	5	17
92	2198,750	0,99014	1,6782	31,41	5	12
93	2218,963	0,98996	1,7092	3,71	4	13
94	2221,419	0,98994	1,7130	58,12	2	17
95	2222,327	0,98993	1,7144	27,23	5	13
96	2226,023	0,98990	1,7201	23,64	6	4
97	2243,216	0,98974	1,7468	41,88	5	18
98	2248,361	0,98969	1,7548	41,96	5	19
99	2275,383	0,98945	1,7972	0,02	5	6
100	2306,955	0,98915	1,8475	35,60	5	7
101	2313,004	0,98909	1,8572	9,42	5	2
102	2356,857	0,98868	1,9283	0,48	5	5
103	2375,677	0,98849	1,9592	60,20	2	18
104	2394,370	0,98831	1,9901	17,54	3	16
105	2435,369	0,98791	2,0589	60,08	2	19

Thứ tự	$SEE_{(j,k)} = \sqrt{MSE_{j,k}}$	Hệ số xác định R^2	Tỷ lệ $MSE_{(j,k)}/MSE_{(1,8)}$	Hệ số biến thiên	j	k
106	2527,880	0,98697	2,2183	53,92	2	15
107	2596,629	0,98625	2,3406	3,88	6	3
108	2648,492	0,98570	2,4350	0,00	1	9
109	2657,404	0,98560	2,4514	33,65	6	5
110	2718,942	0,98493	2,5662	0,22	5	4
111	2832,122	0,98365	2,7843	5,36	2	2
112	2874,620	0,98315	2,8685	36,55	6	6
113	3042,785	0,98113	3,2140	0,00	1	10
114	3233,737	0,97868	3,6300	4,67	5	1
115	3249,137	0,97848	3,6647	0,00	1	11
116	3450,620	0,97573	4,1333	2,87	4	1
117	3585,842	0,97379	4,4636	4,98	2	6
118	3767,154	0,97107	4,9263	42,68	6	2
119	3808,233	0,97044	5,0344	14,99	2	5
120	3841,643	0,96991	5,1231	0,00	1	7
121	4103,070	0,96568	5,8441	0,00	1	12
122	4103,775	0,96567	5,8461	6,06	4	4
123	4398,122	0,96056	6,7148	21,45	2	4
124	4477,166	0,95914	6,9583	5,51	5	3
125	4569,987	0,95742	7,2498	0,00	1	13
126	4685,830	0,95524	7,6220	10,09	4	5
127	4857,050	0,95191	8,1892	0,00	1	3
128	4985,464	0,94933	8,6280	76,31	3	19
129	5019,305	0,94864	8,7455	24,15	2	1
130	5045,713	0,94810	8,8378	10,68	4	6
131	5200,809	0,94486	9,3894	18,14	3	8
132	6282,661	0,91953	13,7020	26,76	2	3
133	6895,024	0,90308	16,5033	10,94	3	9
134	7447,383	0,88693	19,2533	8,27	3	10
135	7501,157	0,88529	19,5324	0,00	1	1
136	7714,183	0,87869	20,6575	5,46	4	2
137	7730,324	0,87818	20,7441	7,01	3	11
138	8243,034	0,86148	23,5870	0,14	3	3
139	8265,987	0,86071	23,7185	5,18	3	7
140	8525,551	0,85183	23,2315	4,75	3	12
141	8750,064	0,84392	26,5779	4,31	3	13
142	7999,076	0,83491	28,1121	0,00	1	4
143	9165,314	0,82875	29,1604	3,04	3	1
144	9601,048	0,81208	31,9989	5,51	3	4
145	10035,942	0,79467	34,9635	8,25	3	5
146	10211,728	0,78742	36,1990	9,50	3	6
147	10596,318	0,77110	38,9770	0,00	1	5
148	10798,917	0,76227	40,4817	14,48	3	2
149	11305,347	0,73945	44,3686	0,00	1	6
150	13932,541	0,60428	67,6843	0,00	1	2

12 hàm xu thế thích hợp nhất đối với dân số nước ta trong 125 năm qua, dưới dạng hiện là:

1) Hàm (5,20): $Y = 1/(a + b.T + c.T^2 + d.T^3)^2$
2) Hàm (2,25): $Y = 1/(a + b.T + c.T^2 + d.\ln\ln T)$
3) Hàm (2,20): $Y = 1/(a + b.T + c.T^2 + d.T^3)$
4) Hàm (2,23): $Y = 1/(a + b.T + c.T^2 + d.T^{3/2})$
5) Hàm (2,22): $Y = 1/(a + b.T + c.T^2 + d.\sqrt{T})$
6) Hàm (2,24): $Y = 1/(a + b.T + c.T^2 + d.\ln T)$
7) Hàm (5,23): $Y = 1/(a + b.T + c.T^2 + d.T^{3/2})^2$
8) Hàm (2,21): $Y = 1/(a + b.T + c.T^2 + d/T)$
9) Hàm (3,15): $Y = \sqrt{a.\exp(b/T) + c}$
10) Hàm (6,20): $Y = \exp(a + b.T + c.T^2 + d.T^3)$
11) Hàm (5,22): $Y = 1/(a + b.T + c.T^2 + d.\sqrt{T})^2$
12) Hàm (2,8): $Y = 1/(a + b.T + c.T^2)$

Những hàm xu thế thích hợp nhất phần lớn là những dạng hàm mới lạ, chưa từng được nhắc đến trong các tài liệu về nhân khẩu học ([2, 3, 15, 18, 20, 25–28]). Các hàm quen thuộc đối với các nhà dân số học lại đứng ở vị trí khá thấp:

• Hàm đa thức bậc 2 ($j = 1, k = 8$) xếp thứ 47; hàm đa thức bậc 3 ($j = 1, k = 20$) xếp thứ 28.

• Hàm mũ 2 tham số ($j = 6, k = 1$) : $\ln Y = aT + c$, tức là $Y = (e^c).\exp(a.T)$ xếp thứ 77; hàm mũ 3 tham số ($j = 1, k = 14$) : $Y = a.\exp(b.T) + c$ xếp thứ 73.

• Hàm logistic ($j = 2, k = 14$) : $1/Y = a.\exp(bT) + c$ xếp thứ 70.

• Hàm Gompertz ($j = 6, k = 14$) : $\ln Y = a.\exp(bT) + c$ xếp thứ 82.

Kết quả tính toán đưa đến nhận xét sơ bộ sau đây: đối với dân số Việt Nam thời kỳ 1878 - 2003 không nên dùng các hàm xu thế quen thuộc mà cần thử nghiệm các hàm xu thế dạng mới.

6. KẾT QUẢ HIỆU CHỈNH

Bảng 5 trình bày kết quả hiệu chỉnh theo 3 hàm xu thế đầu tiên cùng với sai số % tương đối của chúng tại các giá trị $n = 10, 20, 30,$ Các sai số nhỏ nhất trong mỗi dòng được in đậm. Qua đó ta thấy không có hàm xu thế nào tốt nhất tại tất cả các thời điểm. Vì thế khi nội suy hoặc ngoại suy nên dùng đồng thời nhiều hàm xu thế.

7. ƯỚC LƯỢNG VÀ DỰ BÁO

Dân số Việt Nam từ 1880 đến 2010 được ước lượng theo 12 hàm xu thế đầu tiên với những trọng số khác nhau. Trọng số này tỷ lệ nghịch với MSE của mỗi hàm (xem cột 4 Bảng 4). Từ nhiều giá trị ước lượng có thể tính được độ phân tán của chúng qua độ lệch quân phương (s) và hệ số biến sai (s/P). Mặt khác việc tính trung bình các giá trị ước lượng có tác dụng làm trơn một phần và giảm bớt những nhiễu loạn có thể.

Bảng 6 cho kết quả tính toán từng thập kỷ. Cột 3 ghi trung bình có trọng số (P) của 12 ước lượng. Cột 4 ghi độ lệch quân phương giữa các ước lượng đó và cột 5 ghi hệ số biến sai. Ba cột cuối cùng ghi dân số thực tế (nếu có) và các sai số tương ứng. Phần lớn các sai

số này đều có giá trị tuyệt đối nhỏ hơn độ lệch quân phương, chỉ trừ hai trường hợp ngoại lệ được in đậm, tương ứng với các năm 1960 và 1980. Có lẽ trong những năm 1960 nước ta còn bị chia cắt cho nên tổng dân số 2 miền Bắc-Nam không phải là số liệu đáng tin cậy. Và những biến động xã hội trước sau năm 1980 cũng có thể gây nên những nhiễu loạn bất thường về dân số.

Bảng 5. Hiệu chỉnh theo 3 hàm xu thế thích hợp nhất

Thứ tự	Năm	Dân số thực tế (Y)	Hàm (5,20) Z_1	Hàm (2,25) Z_2	Hàm (2,20) Z_3
1	1878	10500,0	12278,5	12871,6	10556,0
2	1901	13196,2	12195,4	12598,3	12456,1
3	1906	14097,9	12598,7	13076,6	13087,4
4	1911	14705,8	13152,9	13705,5	13819,9
5	1913	15582,5	13419,3	14000,1	14144,5
6	1914	14349,5	13562,5	14156,9	14314,2
7	1921	15644,5	14770,3	15442,1	15653,8
8	1926	17067,5	15872,0	16580,6	16797,9
9	1930	17491,5	16918,9	17642,8	17847,2
10	1931	17698,4	17206,0	17931,3	18130,4
			(- 2,78%)	(**1,32**%)	(2,44%)
11	1934	17501,0	18132,7	18857,0	19034,8
12	1935	18337,0	18464,6	19186,7	19355,7
13	1936	18975,6	18808,9	19527,4	19687,0
14	1938	19500,0	19535,0	20243,6	20381,9
15	1939	21066,7	19917,9	20619,8	20746,4
16	1940	20868,7	20314,5	21008,7	21122,8
17	1943	22290,4	21591,4	22255,3	22328,2
18	1944	22571,5	22048,7	22699,0	22757,0
19	1945	22644,0	22522,2	23157,7	23200,2
20	1950	25884,0	25150,3	25692,8	25650,7
			(- 2,83%)	(-**0,74**%)	(- 0,90%)
21	1951	25279,5	25731,8	26252,1	26191,9
22	1954	26704,4	27597,8	28045,2	27929,7
23	1955	27282,1	28262,2	28683,3	28549,3
24	1956	28652,2	28948,7	29342,8	29190,3
25	1957	29572,4	29657,9	30024,3	29853,4
26	1958	30033,2	30390,2	30728,3	30539,4
27	1959	31098,2	31146,4	31455,6	31249,0
28	1960	31130,6	31926,7	32206,9	31983,0
29	1961	32714,6	32731,8	32982,6	32743,2
30	1962	32907,7	33562,1	33783,6	33527,4
			(1,99%)	(2,66%)	(**1,88** %)
31	1965	35421,7	36208,2	36343,8	36046,3
32	1967	37768,7	38104,8	38186,9	37868,8

Thứ tự	Năm	Dân số thực tế (Y)	Hàm (5,20) Z_1	Hàm(2,25) Z_2	Hàm(2,20) Z_3
33	1968	36594,0	39093,6	39150,8	38825,0
34	1970	41275,6	41152,4	41164,8	40829,9
35	1971	41913,0	42222,3	42215,3	41879,5
36	1974	45715,7	45591,2	45540,6	45218,7
37	1975	47638,0	46765,5	46706,1	46395,0
38	1976	49160,0	47964,4	47899,2	47602,5
39	1977	50413,0	49186,9	49119,1	48840,2
40	1978	51421,0	50431,7	50364,6	50107,1
			(-**1,92**%)	(- 2,05%)	(- 2,56%)
41	1979	52462,0	51697,6	51634,4	51402,1
42	**1979,25**	**52741,766**	52017,2	51955,4	51730,0
43	1980	53722,0	52983,0	52926,7	52723,4
44	1981	54927,0	54286,0	54239,7	54068,8
45	1982	56170,0	55604,6	55570,8	55435,9
46	1983	57373,0	56936,4	56917,4	56821,7
47	1984	58653,0	58279,0	58276,4	58222,6
48	1985	59872,0	59629,3	59644,1	59634,6
49	1986	61109,0	60984,4	61016,6	61053,1
50	1987	62452,0	62340,7	62389,3	62472,7
			(- 0,18%)	(- 0,10%)	(**0,03**%)
51	1988	63727,0	63694,4	63757,4	63887,6
52	**1988,75**	**64375,762**	64705,7	64777,1	64941,7
53	1989	64774,0	65041,6	65115,4	65291,2
54	1990	66016,7	66377,9	66457,5	66676,3
55	1991	67242,4	67698,6	67777,4	68035,2
56	1992	68450,1	68998,9	69068,4	69359,6
57	1993	69644,5	70273,6	70323,5	70640,4
58	1994	70824,5	71517,3	71535,3	71868,3
59	1995	71995,5	72724,4	72696,2	73033,7
60	1996	73156,7	73889,1	73798,5	74126,7
			(1,00%)	(**0,88**%)	(1,33%)
61	1997	74306,9	75005,5	74834,2	75137,2
62	1998	75456,3	76067,6	75795,7	76055,5
63	**1998,75**	**76323,173**	76824,8	76463,6	76677,7
64	1999	76596,7	77069,2	76675,4	76871,7
65	2000	77635,4	78004,4	77466,0	77576,8
66	2001	78685,8	78867,0	78160,5	78162,4
67	2002	79727,4	79651,3	78752,6	78620,7
68	2003	80902,4	80351,5	79236,7	78945,3
			(-**0,68**%)	(-2,06%)	(-2,42%)

Bảng 6. Ước lượng và dự báo theo 12 hàm xu thế đầu tiên

Thứ tự	Năm	Dân số ước lượng (P)	Độ lệch quân phương (s)	Hệ số biến sai (100.s/P) %	Dân số thực tế (Y)	Sai số tuyệt đối (Y-P)	Sai số % tương đối
1	1880	10691,6	1900,1	17,77	-	-	-
2	1890	10969,5	1316,7	12,00	-	-	-
3	1900	11722,2	926,3	7,90	-	-	-
4	1910	12946,9	638,2	4,93	-	-	-
5	1920	14737,8	507,1	3,44	-	-	-
6	1930	17278,7	529,3	3,06	17491,5	212,8	1,22
7	1940	20837,4	520,5	2,50	20868,7	31,3	0,15
8	1950	25760,6	444,7	1,73	25884,0	117,4	0,45
9	1960	32476,8	427,0	1,31	31130,6	**-1346,2**	-4,32
10	1970	41447,4	401,0	0,97	41275,6	-171,8	-0,42
11	1980	52898,0	238,3	0,45	53722,0	**824,0**	1,53
12	1990	66050,0	472,7	0,72	66016,7	-33,3	-0,05
13	2000	77880,5	367,2	0,47	77635,4	-245,1	-0,32
14	2010	83369,4	4034,8	4,84	-	-	-

8. KHẢ NĂNG ỨNG DỤNG CÁC HÀM XU THẾ MỚI

Câu hỏi được đặt ra là các hàm xu thế xấp xỉ tốt nhất biến động dân số Việt Nam có thích hợp với dân số các nước khác, các khu vực khác trên thế giới hay không?

Để có thể kết luận về mức độ phổ dụng của các dạng hàm xu thế mới chúng tôi đã tính toán thêm 2 lần trên dân số Việt Nam (với các năm gốc X_g khác) và dữ liệu dân số của 5 nước thuộc 5 châu lục khác nhau: Australia (châu Đại Dương), Ba Lan (châu Âu), Brazil (châu Mỹ), Jordan (châu Á) và Senegal (châu Phi).

Bảng 7. Thông tin về các bộ dữ liệu dùng để so sánh

Ký hiệu	Tên nước	Thời kỳ	Số dữ liệu N	Năm gốc X_g	Dân số năm 2003 (triệu người)	Mật độ dân số (người/ km^2)	Tỷ lệ % tăng dân số
1	Việt Nam	1878-2003	68	1800			
2	**Việt Nam**	**1878-2003**	**68**	**1825**	80,9	243,37	1,3
3	Việt Nam	1878-2003	68	1851			
4	Australia	1851-2003	69	1825	19,9	2,56	1,0
5	Ba Lan	1850-2003	65	1825	38,6	119,3	-0,1
6	Brazil	1850-2003	69	1825	176,5	20,62	1,2
7	Jordan	1950-2003	58	1915	5,5	61,79	2,8
8	Senegal	1950-2003	58	1922	11,0	55,83	2,5

Dân cư các nước này nói các thứ tiếng khác nhau (Anh, Ba Lan, Bồ Đào Nha, A Rập, Pháp), theo các tôn giáo khác nhau (Anh giáo, Thiên chúa giáo, Hồi giáo), có quy mô dân

số, mật độ dân số và tỷ lệ tăng dân số rất khác biệt ([5]). Sự đa dạng này bảo đảm cho tính tổng quát và tính khách quan của việc khảo sát các hàm xu thế (Bảng 7). Số liệu dân số được lấy từ các tài liệu ([5, 23, 24, 25]).

Bảng 8 và Bảng 9 ghi các hệ số tương quan Pearson và hệ số tương quan hạng Spearman giữa các giá trị $\mathrm{SSE}_{(j,k)}$ tương ứng của 8 bộ dữ liệu nói trên ($n = 150$). Các giá trị lớn nhất và nhỏ nhất trong mỗi bảng được in đậm.

Bảng 8. Hệ số tương quan Pearson giữa các giá trị $\mathrm{SEE}_{(j,k)}$

R	1	2	3	4	5	6	7
2	0,9691						
3	**0,9429**	0,9849					
4	0,9288	0,9591	0,9518				
5	0,6644	0,7028	0,7199	0,7725			
6	0,8298	0,8756	0,9143	0,8599	0,7621		
7	0,7852	0,8021	0,7787	0,8297	0,7023	**0,6457**	
8	0,8590	0,8887	0,8842	0,8722	0,6901	0,7542	0,9419

Bảng 9. Hệ số tương quan hạng Spearman giữa các giá trị $\mathrm{SEE}(j, k)$

R_h	1	2	3	4	5	6	7
2	0,9464						
3	0,9182	**0,9624**					
4	0,7446	0,7631	0,6940				
5	0,5984	0,6414	0,5892	0,6150			
6	0,8252	0,8396	0,8209	0,7798	0,4863		
7	0,7417	0,7372	0,7050	0,7936	0,6990	0,7831	
8	0,6099	0,6097	0,5873	0,5297	**0,3151**	0,7083	0,5569

Để kiểm định ý nghĩa của các hệ số tương quan, trong cả hai trường hợp người ta đều dùng phép thử

$$t = r.\sqrt{\frac{n-2}{1-r^2}} \quad \text{với } (n-2) \text{ bậc tự do.}$$

Có sự khác biệt ở chỗ: phép thử là hai phía đối với R và một phía đối với R_h ([1, 14]).

Ta chỉ cần kiểm định với hai giá trị nhỏ nhất: $r = 0{,}6457$ trong Bảng 8 và $r_h = 0{,}3151$ trong Bảng 9. Đầu tiên chúng tôi đặt mức ý nghĩa 0,05 với bậc tự do $(n-2) = 148$. Giá trị tới hạn tương ứng hai phía là 1,9763 và một phía là 1,6552.

Với $r = 0{,}6457$ thì $t_1 = 10{,}2862 > 1{,}9763$; với $r_h = 0{,}3151$ thì $t_2 = 4{,}0393 > 1{,}6552$. Như vậy giả thiết "hệ số tương quan = 0" bị bác bỏ.

Ngay cả giá trị tới hạn lớn nhất của t trong bảng số (với bậc tự do 148) là 3,3581, tương ứng với mức ý nghĩa 0,001 (hai phía) và 0,0005 (một phía) cũng nhỏ hơn hai trị số tính toán t_1 và t_2 nói trên. Có thể thấy rằng sự tương quan giữa 8 bộ dữ liệu đều có ý nghĩa, và do đó có hy vọng tìm được các hàm xu thế chung cho nhiều nước, nhiều khu vực khác nhau. Đó cũng chính là hướng nghiên cứu sẽ được tiếp tục sau này.

TÀI LIỆU THAM KHẢO

[1] S. Aivazian, *Étude Statistique des Dépendances*, Éditions Mir, Moscou, 1978.

[2] E. M. Andreev, A. G. Volkov (chủ biên), *Các mô hình nhân khẩu học*, NXB Statistika, Moskva, 1977 (tiếng Nga).

[3] J. Bourgeois-Pichat, *La Dynamique des Populations*, Institut National d'Études Démographiques, Presses Universitaires de France, Paris, 1994.

[4] P. Brocheux, D. Hemery, *Indochine le Colonisation Ambigue (1858-1954)*, Éditions La Découverte, Paris, 1995.

[5] D. Frémy, M. Frémy, *Quid 2004*, Éditions Robert Laffont, Paris, 2003.

[6] Frey Tamás, *Egyváltozós Elemi Fuggvények*, Tankonyvkiadó, Budapest, 1952.

[7] F. Gendreau, V. Fauveau, Đặng Thu, *Démographie de la Péninsule Indochinoise*, AUPELF-UREF, ESTEM, Paris, 1997. (Bản dịch: *Dân số Bán đảo Đông Dương*, NXB Thế giới, Hà Nội, 1997).

[8] J. V. Gregg, C. H. Hossell, J. T. Richardson, *Mathematical Trend Curves: An Aid to Forecasting*, Oliver & Boyd, Edinburgh, 1964.

[9] P. G. Guest, *Numerical Methods of Curve Fitting*, Cambridge University Press, 1961.

[10] Hunováci Hugó, *Empirikus formulák felállítása a kísérleti megfigyelés eredményei alapján.* Szakmérnoki diplomaterv. BME Villamosmérnoki Kar, Matematikai Tanszék, Budapest 1973.

[11] R. I. Jennrich, *An Introduction to Computational Statistics, Regression Analysis*, Prentice-Hall International, Inc., Englewood Cliffs, New Jersey, 1995.

[12] Kis Ottó, Kovács Margit, *Numerikus módszerek*, Muszaki Konyvkiadó, Budapest, 1973.

[13] E. N. Lvovskii, *Các phương pháp thống kê để xây dựng các công thức thực nghiệm*, NXB Vyshaya Shkola, Moskva, 1988 (tiếng Nga).

[14] R. D. Mason, D. A. Lind, *Statistical Techniques in Business and Economics*, Irwin/McGraw-Hill, 1999.

[15] C. Newell, *Methods and Models in Demography*, Belhaven Press, London, 1988.

[16] D. A. Ratkowski, *Nonlinear Regression Modeling: A Unified Practical Approach*, Marcel Dekker, Inc. , New York and Basel, 1983.

[17] L. Z. Rumshiskii, *Xử lý toán học các kết quả thực nghiệm*, NXB Nauka, Moskva, 1971 (tiếng Nga).

[18] D. Smith, N. Keyfitz, *Mathematical Demography, Selected Papers*, Spinger-Verlag, Berlin, 1977.

[19] I. M. Sobol, R. B. Statnikov, *Những giải pháp tốt nhất - tìm chúng ở đâu?* NXB Znanie, Moskva, 1982 (tiếng Nga).

[20] O. V. Staroverov, *Những nguyên tắc cơ bản của nhân khẩu học toán học*, NXB Nauka, Moskva , 1997 (tiếng Nga).

[21] Tổng cục Thống kê, *Số liệu thống kê 1930-1984*, NXB Thống kê, Hà Nội, 1985.

[22] Tổng cục Thống kê, *Số liệu thống kê dân số và kinh tế - xã hội Việt Nam 1975-2001*, NXB Thống kê, Hà Nội, 2002.

[23] United Nations (Statistical Division), *Demographic Yearbook 1997, Historical Supplement (1948-1997)*, UN Publications, New York - Geneve, 1999.

[24] U. S. Census Bureau, *International Data Base(IDB), Summary Demographic Data.* (http://www.census.gov/ipc/www.idbsum.html)

[25] D. I. Valentei (chủ biên), *Từ điển bách khoa nhân khẩu học*, NXB Bách khoa thư Xô viết, Moskva, 1985 (tiếng Nga).

[26] I. G. Venetskii, *Các phương pháp toán học trong nhân khẩu học*, NXB Statistika, Moskva, 1971 (tiếng Nga).

[27] I. G. Venetskii, *Các phương pháp thống kê trong nhân khẩu học*, NXB Statistika, Moskva, 1977 (tiếng Nga).

[28] I. G. Venetskii, *Các phương pháp xác suất trong nhân khẩu học*, NXB Finansy i Statistika, Moskva, 1981 (tiếng Nga).

Nhận bài ngày 7 - 7 - 2005