Proceedings of Joint Conference of Korean Data And Information Science Society
and The Korean Data Analysis Society,
April 28 - 29, 2006, pp. 77-85

간이생명표 확장 기법을 통한 사망확률 계산

김기환[1] · 이동희[2] · 정승환[3]

요 약

본 논문에서는 간이생명표 확장기법인 HP8(Heligman and Pollard 8-parametric) 모형과 spline 내삽법을 이용한 사망확률 계산 결과를 비교하고 HP8 모형을 우리나라 간이생명표 자료에 적용하여 각 연령별, 연도별로 사망확률을 계산하였다. 그리고 HP8 모형의 8개 모수와 사망확률을 계산하는데 있어 SAS/OR의 NLP procedure를 이용한 결과와 UN(United Nation)에서 인구통계분석을 목적으로 만든 소프트웨어인 MORTPAK을 이용한 결과를 비교하였다. 분석에 사용한 자료는 통계청에서 제공되는 1971년부터 2003년까지 우리나라 간이생명표 자료이다.

주요용어 : 간이생명표, 사망확률, 확장 기법, HP8.

1. 서론

생명표(life table)는 현재의 사망수준이 그대로 지속된다는 가정 하에, 어떤 출생 집단이 연령이 많아짐에 따라 소멸되어 가는 과정을 정리한 표로(통계청, 2005), 인구통계학이나 보험통계학 등 여러 분야에 걸쳐서 매우 유용한 자료이다. 생명표의 종류는 크게 두 가지로 나누어지는데, 첫 번째는 인구집단의 사망확률, 기대여명 등을 각각의 연령별로 기록한 완전생명표(complete life table)이고, 두 번째는 연령구간을 묶어서 기록한 간이생명표(abridged life table)이다. 생명표를 작성하는데 있어서 가장 기초가 되는 것이 사망률이기 때문에 사망률을 추정하고 분석하는 것이 가장 중요하다고 할 수 있다. 완전생명표를 사용해서 분석하는 것이 가장 좋은 방법이지만 우리나라 완전생명표는 1997년부터 2003년까지 5개만 작성되어 있어서 분석하는데 어려움이 있다. 외국의 경우, age heaping 현상 때문에 간이생명표를 완전생명표로 확장시키는 기법을 통해 사망률을 분석하는 경우가 많다. age heaping 현상이 생기게 되는 주된 이유는 자료를 등록할 때 연령을 잘못 기입하는데 있고, 또 다른 이유는 충분하지 않은 적은 표본수로 만든 인구동태 통계량들이 신뢰할만하지 못하다는 것에 있다. 외국에 비해 우리나라는 사망자수가 비교적 정확하게 보고된다고 알려져 있어 외국의 경우처럼 age heaping 문제는 거의 발생하지 않지만 위에서 언급했듯이 사망률 분석을

[1]339-700 충청남도 연기군 조치원읍 서창동 208, 고려대학교 정보통계학과 교수. E-mail: korpen@korea.ac.kr
[2]136-701 서울시 성북구 안암동 5가 1, 고려대학교 통계연구소 박사후 과정. E-mail : ld0351@korea.ac.kr
[3]339-700 충청남도 연기군 조치원읍 서창동 208, 고려대학교 정보통계학과 대학원 석사과정.
E-mail : verylong@korea.ac.kr

위한 과거의 자료가 부족하기 때문에 간이생명표 확장기법을 이용하여 사망률을 분석하는 방법을 사용하였다.

2. 간이생명표 확장기법

2.1. 우리나라 간이생명표 자료설명

본 논문에서 사용한 자료는 통계청 통계정보시스템(KOSIS)에서 제공하고 있는 1971년부터 2003년까지의 우리나라 남녀별 간이생명표 자료이다.

표 1 : 우리나라 간이생명표의 연령구간별 사망확률 자료제공 형태(남자 사망확률)

	1971	1993	1999	2001
0	0.04075	0.01034	0.00613	0.00591
1	0.01338	0.0034	0.00204	0.00192
5	0.01187	0.00278	0.00156	0.00139
10	0.00852	0.00211	0.0013	0.0011
15	0.01378	0.00576	0.0035	0.00266
20	0.01863	0.0063	0.00474	0.00379
25	0.01634	0.008	0.00552	0.00457
30	0.01898	0.01051	0.00727	0.00614
35	0.02192	0.01567	0.01153	0.01001
40	0.03887	0.02434	0.01873	0.01657
45	0.06006	0.03574	0.02937	0.02619
50	0.09082	0.05285	0.04151	0.0378
55	0.13858	0.07793	0.06282	0.0565
60	0.19336	0.11475	0.09455	0.08632
65	0.27328	0.17174	0.1425	0.13109
70	0.34395	0.2522	0.22005	0.20584
75	0.49978	0.36414	0.33073	0.31498
80	1	0.51054	0.47451	0.45267
85+		1		
85			0.6268	0.60535
90			0.76055	0.74533
95+			1	
95				0.856
100+				1

표 1은 우리나라 간이생명표에서 연령별로 사망확률을 제공하고 있는 형태를 요약한 결과이다. 1971년부터 1992년까지는 0세부터 80세 이상, 18개의 연령구간으로 제공하고 있고, 1993년부터 1998년까지는 0세부터 85세 이상, 19개의 연령구간, 1999년과 2000년은 0세부터 95세 이상, 21개의 연령구간으로 제공하고 있으며, 2001년부터 2003년까지는 0세부터 100세 이상, 22개의 연령구간으로 제공하고 있다.

표 1에서 볼 수 있듯이 연도별로 연령구간이 일정하지 않다. 사망률에 관한 일관된 분석을 하기

위해서는 모두 동일한 연령구간으로 조정할 필요가 있으며, 특히 Coale, Demeny, Vaughan (1983)에 의하여 100세 이상까지의 생명표 작성에 대한 연구가 있었고, 우리나라의 경우 급속한 고령화가 진행 중이므로 사망률자료를 100세 이상까지 조정하는 것이 필요하다고 하겠다.

2.2. 간이생명표 확장기법

간이생명표 확장을 위한 내삽법(內揷法, interpolation formula)에 대한 최근 연구는 Kostaki and Panousis(2001)에 의해 수행되었으며, 이 논문은 여러 가지 간이생명표 확장기법들을 비교하고 그 결과를 제시하였는데, HP8(Heligman and Pollard 8-Parametric) 모형과 spline 내삽법이 좋은 결과를 산출하는 것으로 나타났다. HP8 모형의 기본적인 형식은 다음과 같다.

$$q_x/p_x = A^{(x+B)^C} + Dexp(-E(\ln(x/F))^2) + GH^x \quad (1)$$

식 1의 q_x는 특정연령 x세에 있는 사람들의 사망확률을 나타내는 것이고 $p_x = 1 - q_x$이다. 그리고 A, B, … ,H는 추정된 8개의 모수들을 나타내는 것이다. 위의 식을 살펴보면 크게 세 부분으로 구성되어있는 것을 볼 수 있는데 첫 번째 부분($A^{(x+B)^C}$)은 급격한 지수적 감소를 나타내는 부분으로써 유아연령층에서 나타나는 사망률의 감소 현상에 영향을 주는 부분이라고 할 수 있다. 두 번째 부분($Dexp(-E(\ln(x/F))^2)$)은 사고로 인한 사망률에 영향을 주는 부분으로, 다른 두 개의 부분으로 표현된 일반적인 사망률 곡선 형태에 추가적으로 첨가되는 사망률을 나타낸다고 할 수 있다. 마지막으로 세 번째 부분(GH^x)은 성인 연령대 사망률의 기하급수적인 증가를 표현해주는 것으로써 노화 또는 신체의 악화 등으로 사망률이 증가하는 것을 표현하는 것이다.

Spline method는 $a = x_1 < x_2 < \cdots < x_n = b$와 같이 (n-1)개의 부분 구간으로 이루어진 구간 [a,b]가 주어지고 그 구간 위에서 정의된 2차 미분가능 함수를 $g(u)$라고 했을 때, 다음의 식 2를 최소화하는 $g(x_i)$와 λ를 추정하는 것이다. 식 2는 3차(cubic) spline의 경우이고 통계 소프트웨어 중 하나인 R을 이용하여 계산하였다.

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \left[\lambda \int_u (g''(u))^2 du\right], \lambda < \infty \quad (2)$$

표 2는 HP8 모형과 spline 내삽법으로 추정한 사망확률 값과 실제 생명표상의 사망확률과의 적합성을 비교한 결과이다. 비교기준은 각각의 기법으로 추정한 연령별 추정사망확률과 실제 완전생명표 연령별 사망확률의 오차제곱합, 즉 $\sum(q_x - \hat{q}_x)^2$(SSE, Sum of Squared Error)로 하였다. 표 2에서 HP8 모형을 이용한 확장기법이 spline 내삽법을 이용한 확장기법보다 SSE가 작은 것을 확인했으므로 본 논문에 사용될 간이생명표 확장기법으로 HP8 모형을 선택하였다.

표 2 : HP8 모형과 Spline method의 SSE값 비교

	2001		2002		2003	
	HP8	Spline	HP8	Spline	HP8	Spline
SSE	0.279347	0.475665	0.302190	0.614541	0.312237	0.613740

2.3. 간이생명표 확장에서 HP8 formula의 적용

서론에서 언급했듯이, 외국의 경우 완전생명표 사망확률의 age heaping 현상 때문에 1세 단위의 사망확률을 일정한 단위(보통 5세 단위)로 묶은 사망확률을 구성하고, 그것을 간이생명표 확장기법을 통해 다시 1세 단위로 확장시킨 사망확률 값을 사용한다. 우리나라의 경우 age heaping 현상은 잘 나타나지 않지만 완전생명표가 1997년부터 현재까지 5개만 작성되어있어 사망률을 분석하기 어렵다. 그러나 간이생명표는 1971년부터 작성되어있어서 간이생명표 확장기법을 이용한 분석이 가능하다. 간이생명표 확장기법에 HP8 모형이 적용되는 과정은 다음과 같다.

HP8 모형의 기본형식인 식 1에서 $C=(A,B,\cdots,H)$라 하고 우변을 $F(x;C)$라고 표현한다면, $q_x/p_x=F(x;C)$로 쓸 수 있고, 특정한 연령 x세에서 1년 동안의 사망확률은 식 3을 이용하여 구할 수 있다.

$$q_x=\frac{F(x;C)}{1+f(x;C)}=G(x;C) \tag{3}$$

식 4는 1세 단위 사망확률을 5세 단위의 사망확률로 변환하는 식이다. 완전생명표의 1세 단위 사망확률을 식 4에 대입하여 1세 단위의 사망확률을 5세 단위의 간이 생명표 사망확률로 변환한다.

$$\begin{aligned} {}_nq_x &= 1-\prod_{i=0}^{n-1}(1-q_{x+i}) \\ &= 1-\prod_{i=0}^{n-1}(1-G(x+i;C) \\ &= {}_nG(x;C) \end{aligned} \tag{4}$$

그리고 식 4를 통해 생성된 5세 단위의 사망확률과 기존 간이생명표 사망확률과의 차이를 최소화하는, 즉 식 5를 최소화하는 C를 추정한다. 식 5를 최소화하는 8개의 모수들을 추정하는 방법으로 비선형최소제곱(nonlinear least square) 알고리즘을 사용하는데 이 부분은 SAS/OR의 NLP procedure를 사용하였다.

$$\sum_x\left(\frac{{}_nG(x;C)}{{}_nq_x}-1\right)^2 \tag{5}$$

이렇게 추정된 C 를 식 3에 대입하여 5세 단위의 사망확률을 1세 단위로 확장시키고, 1세 단위 사망확률을 추정한다. 표 3은 1971년부터 2001년까지 HP8 모형을 사용한 8개 모수 추정 결과를 남

녀별 10년 단위로 요약한 것이다.

표 3 : 8개 모수 추정 결과(1971~2001)

		1971	1981	1991	2001
남자	A	0.013870	0.009122	0.003581	0.001818
	B	0.004406	0.036249	0.050335	0.094675
	C	0.055916	0.085825	0.083678	0.088223
	D	0.008526	0.004896	0.002957	0.001602
	E	13.932229	11.656197	12.024484	10.305131
	F	19.097868	18.530304	18.046985	18.923548
	G	0.000630	0.000812	0.000735	0.000305
	H	1.103059	1.097132	1.092067	1.102239
여자	A	0.013437	0.009216	0.003373	0.001333
	B	0.006912	0.032809	0.021184	0.000656
	C	0.061119	0.081474	0.059261	0.033114
	D	0.008284	0.002999	0.001355	0.001081
	E	8.268247	6.830042	6.950357	5.168759
	F	22.474919	21.342713	21.774014	24.099223
	G	0.000503	0.000337	0.000143	0.000037
	H	1.091682	1.096198	1.105897	1.123473

2.3절의 과정을 통하여 추정된 1세 단위 사망확률은 5세 단위 추정사망확률을 단순히 확장한 값이기 때문에 실제 완전생명표 사망확률과 차이가 발생하므로 이를 보정하는 과정이 필요하다. 이러한 보정에 관한 내용은 Kostaki(1991)에서 언급되었다. 추정된 사망확률 값에 대한 보정은 상수 K를 계산함으로써 가능한데, 그 방법은 식 6과 같다.

$$K=\frac{\ln(1-{}_nq_x)}{\sum_{i=0}^{n-1}\ln(1-\hat{q}_{x+i})} \tag{6}$$

그리고, 식 6에서 계산한 K를 식 7에 삽입하여 보정된 사망확률을 추정한다.

$$\hat{q}'_{x+i}=1-(1-\hat{q}_x)^K \tag{7}$$

식 7에서 최종적으로 추정된 사망확률은 식 8을 만족하게 된다.

$$1-\prod_{i=0}^{n-1}(1-\hat{q}'_{x+i})={}_nq_x \tag{8}$$

2.4. MORTPAK과의 결과 비교

MORTPAK은 UN(United Nation)에서 인구통계분석을 목적으로 만든 소프트웨어로 추계인구작성, 생명표작성, HP8등과 관련된 총 17개의 분석 모듈을 갖고 있다. MORTPAK에서 제공하는 HP8 방

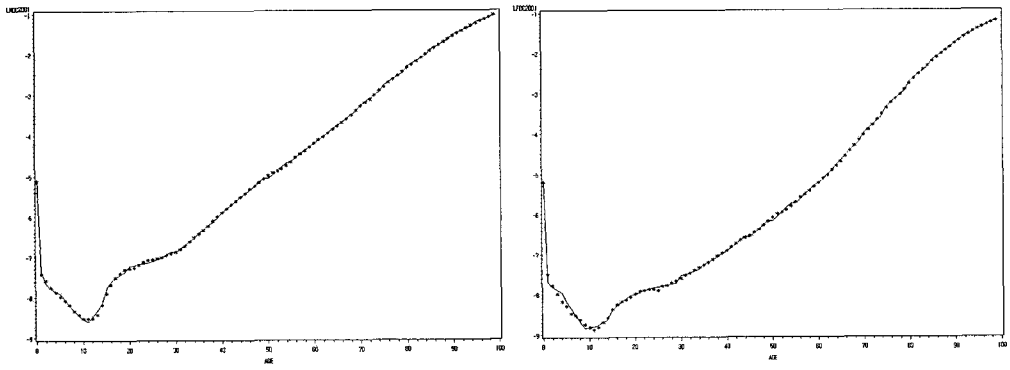
법은 매우 간단하게 5세 단위의 사망확률을 1세 단위의 사망확률로 바꾸어주지만, Kostaki(1991)의 연구에 의하면, 정확도는 상대적으로 떨어지는 것으로 지적하고 있으며, 실제 우리나라 자료에 적용해본 결과 Kostaki의 지적과 같은 결과를 얻을 수 있었다. MORTPAK과의 비교는 MORTPAK의 계산결과와 MORTPAK에서 제공되는 HP8 방법의 A-H까지 8개의 모수 추정치를 초기값으로 하여 SAS/OR의 NLP procedure에 의한 최적화 결과로 재추정한 결과를 비교하여 어느 정도나 개선되었는지를 평가하였다. 그리고 초기 연령대에서 약간 smooth하지 못한 부분이 생기는 문제는 약간의 보정과정을 거쳐 해결하였다. 아래의 표 4는 MORTPAK과 SAS/OR의 NLP procedure를 사용하여 추정한 사망확률 값과 1997년부터 2003년까지의 완전생명표 값의 차이를 살펴본 결과이다. 비교 결과를 보면 NLP를 이용한 결과가 더 좋은 결과임을 알 수 있다.

표 4 : MORTPAK과 NLP procedure의 SSE값 비교(1997~2003년, 남자)

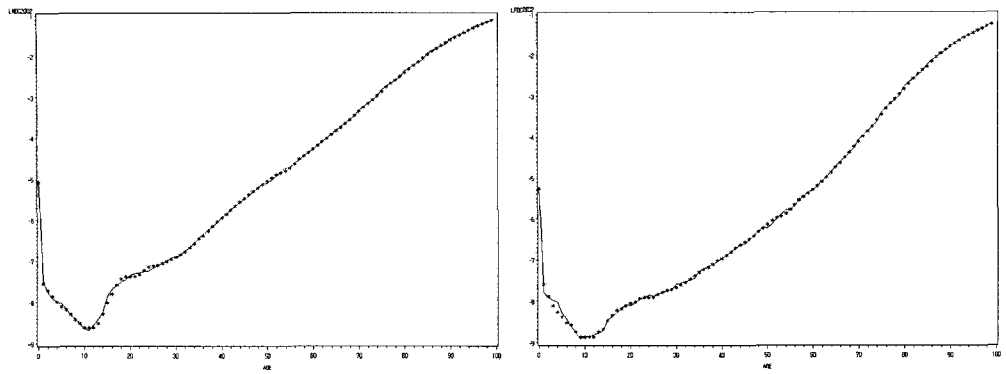
	1997년	1999년	2001년	2002년	2003년
NLP	0.190930	0.098833	0.118585	0.151049	0.159325
MORTPAK	0.735148	0.936156	0.837235	1.001255	1.025971

위의 과정을 통해서 추정된 사망확률이 분석하기에 적합한 추정치인지 확인해보기 위해서는 실제 완전생명표와 비교해보는 절차가 필요하다. 그림 1은 추정된 사망확률과 실제 완전생명표 상의 사망확률을 비교한 결과이고 사망확률의 패턴을 명확하게 파악하기 위해 사망확률에 로그를 취한 로그사망확률을 사용하였다. 그림 1에서 '*'로 표시된 LMDC2001~LMDC2003과 LFDC2001~LFDC2003은 완전생명표 상의 남녀별 사망확률 값을 나타낸 것이고 직선으로 표시된 것은 HP8 모형을 이용하여 추정된 사망확률 값이다. 그림 1을 통해 남녀 모두 추정된 사망확률이 실제 완전생명표의 사망확률 패턴을 잘 따라가고 있음을 알 수 있다.

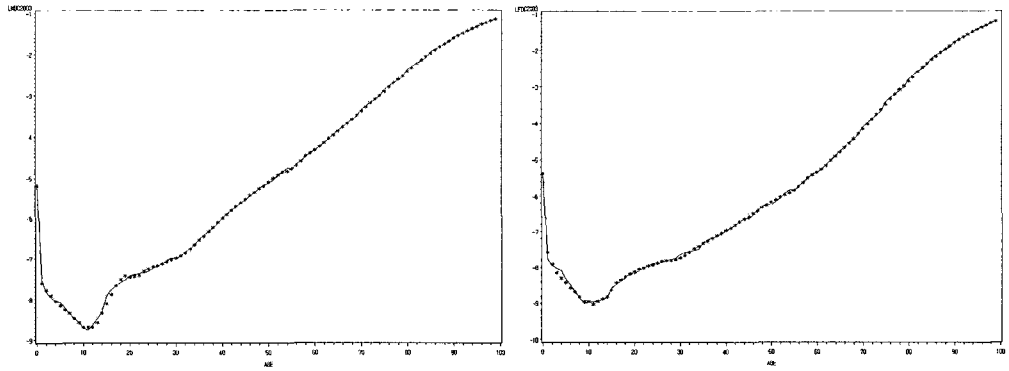
그림 2는 HP8 모형을 이용하여 구한 1971년부터 2003년까지의 추정사망확률을 그래프로 나타낸 것이다. 그림 2를 통해 사망확률은 과거로부터 현재까지 미래로 갈수록 감소하는 추세라는 것을 알 수 있다.



< 2001년 >



< 2002년 >



< 2003년 >

그림 1 : 남녀별 추정사망확률과 실제 생명표(2001~2003년) 사망확률 비교

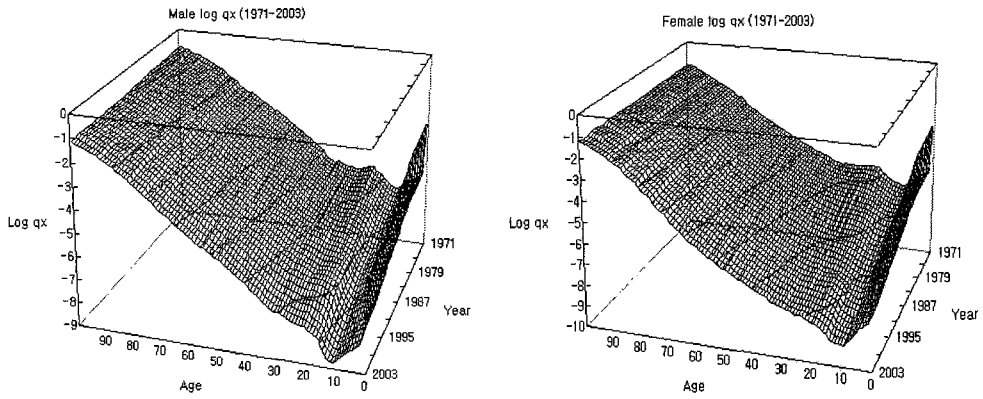


그림 2 : HP8 모형을 이용한 남녀별 추정사망확률(1971~2003년)

3. 결론

본 논문에서는 통계청에서 제공하는 우리나라 간이생명표 자료를 사용하여 간이생명표 확장기법인 HP8 모형을 통해서 1971년부터 2003년까지의 남녀별 사망확률을 계산하였다. 각 나라의 생명표는 그 나라 국민들의 종합적인 사망수준을 보여줌으로써 국가의 보건 · 의료정책 수립, 보험료율, 인명피해 보상비 산정, 장래인구추계 작성, 국가 간 경제 · 사회 · 보건 수준 비교 등 기타 여러 분야의 정책들을 수립하는데 중요한 지표 역할을 하는 자료이다. 그러한 생명표의 기초가 되는 사망률을 분석하여 앞으로 일어날 일에 대해 미리 대비하는 것은 매우 중요한 일이 아닐 수 없다. 간이생명표 확장기법을 이용해서 각 연령별 사망확률을 계산하는 방법은 그 동안 여러 가지가 제안되었다. 본 논문에서는 간이생명표 확장기법들에 대한 판단 기준을 세우고 대표적으로 HP8 모형을 사용하였지만 그 외의 방법들과 계산 과정에서의 정교성에 대해서도 좀 더 깊이 있는 연구를 할 필요가 있다고 판단된다. 이러한 것들은 추후 연구과제로 남겨놓기로 한다. 간이생명표 확장기법을 이용한 사망확률 계산에 관한 연구로 인해 향후 사망률에 관한 연구, 특히 사망률 예측에 관한 연구에 많은 도움을 줄 수 있으리라 기대하는 바이다.

참고문헌

[1] 구자흥 (2002). *인구통계학의 이론과 실제*, 교우사, 서울.

[2] 박유성, 김기환 (2004). *SAS/ETS를 이용한 시계열자료분석 I*, 자유아카데미, 서울.

[3] 통계청 (2005). *2003년 생명표 작성결과*.

[4] Carriere J. F. (1992). Parametric models for life tables, *Transactions of Society of Actuaries*, Vol. 44, pp. 77-99.

[5] Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*, Chapman&Hall ed., New York.

[6] Coale, A. J., Demeny, P. and Vaughan, B. (1983). *Regional Model Life Tables and Stable Populations* (Second Edition), Academic Press, New York/London.

[7] Elandt-Johnson, R. and Johnson, N. (1980). *Survival Models and Data Analysis*, John Wiley, New York.

[8] Heligman, L. and Pollard, J. H. (1980). The Age pattern of mortality, *Journal of the Institute of Actuaries*, 107, pp. 49-77.

[9] Kostaki, A. (1991). The Heligman-Pollard Formula as a Tool for Expanding an Abridged Life Table, *Journal of Official Statistics*, Vol. 7, No. 3, pp. 311-323.

[10] Kostaki, A. and Panousis, V. (2001). Expanding an abridged life table, *Demographic Research*, Vol. 5. Article 1.

[11] John Maindonald and John Braun (2003). *Data Analysis and Graphics Using R*, Cambridge University Press, New York.

[12] McNeil, D. R., Trussel, T. J. and Turner, J. C. (1977). Spline Interpolation of Demographic Data, *Demography*, Vol. 14, No. 2, pp. 245-252.

[13] SAS Institute Inc. (2004). *SAS/IML 9.1 User's Guide*, Cary, North Carolina: SAS Institute Inc.

[14] SAS Institute Inc. (2004). *SAS/OR 9.1.2 User's Guide : Mathematical Programming*, Cary, North Carolina: SAS Institute Inc.

[15] Wegman, J. E. and Wright, W. I. (1983). Splines in Statistics, *Journal of the American Statistical Association*, Vol. 78, No. 382, pp. 351-365.