연구논문

인구통계학적 분석을 이용한 우리나라의 센서스 및 동태자료에 대한 질적 평가*

Quality Evaluation for Census and Vital Statistics of Korea Using Demographic Analysis

전새봄** · 김성용*** · 박유성****
Saebom Jeon · Seongyong Kim · Yousung Park

인구통계학적 분석(Demographic Analysis: DA)은 센서스 및 동태자료의 포함오류에 대한 질적 평가를 위해 센서스 후 조사(Post Enumeration Survey: PES)와 함께 널리 이용되는 방법이다. 그러나 우리나라에서는 인구통계학적 분석이 시행된 적이 없으며, 해외 인구통계학 논문에서는 인구통계학적 분석을 통해 우리나라 사망자 등록의 후진성을 지적하고 있다.

본 논문에서는 다양한 인구통계학적 분석방법을 소개하고, 한국의 현실에 맞도록 변형시킨 방법을 제시한다. 또한 이를 이용하여 1985년부터 2010년까지의 센서스 및 사망자 자료를 분석한 결과, 우리나라의 사망자 자료는 해외 논문의 결과와 달리 완성도(completeness)가 높은 것으로 나타났다. 센서스의 연도별 연령별 누락률을 추정하여, 센서스의 비정상적인 코호트의 증가 현상, 성비, 그리고 연령비의 이상 현상에 대해 분석하였다.

주제어 : 인구통계학적 분석, 완성도, 센서스, 동태자료, 질적 평가

Demographic Analysis(DA) as well as Post Enumeration Survey(PES) are typical methods for evaluating completeness of census and vital statistics. In spite of its popularity, DA has never been attempted in Korea, while other international journals of demography have pointed out the backwardness of death registration in Korea in

* 이 논문은 2010년 정부재원(교육과학기술부 인문사회연구역량강화사업비)으로 한국연구재단의 지원을 받아 연구되었음(NRF−2010−411−B00028).

** 고려대학교 통계연구소 연구교수

*** 고려대학교 세종캠퍼스 경제통계연구소 연구교수

**** 교신저자(corresponding author): 고려대학교 통계학과 교수 박유성.
E−mail: yspark@korea.ac.kr

terms of DA approach. This paper introduces various DA methods and modifies them to be adequate for Korea census and vital statistics. Our method are also applied to reconstruct year−age−sex specific population and estimate their omission rates for year and age. Empirical analysis of census and vital statistics of Korea from 1985 to 2010 demonstrates high completeness of death registration in Korea, contrary to existing literatures. We also investigates abnormal patterns in census by comparing with reconstruction data in view of cohort, sex ratio and age ratio.

Key words : DA, completeness, census, vital statistics, quality evaluation

Ⅰ. 서 론

센서스(census)의 결과는 수많은 오류(error)를 포함하고 있다. 이러한 오류는 크게 포함오류(coverage error)와 내용오류(content error)로 구분될 수 있다. 포함오류는 가구나 개인의 센서스에서 누락(omission)되거나 중복(duplication)때문에 발생하는 오류를 말하며 내용오류는 가구나 개인의 특성에 대한 부정확한 보고나 기록(incorrect reporting or recording)때문에 발생하는 오류를 말한다. 센서스에 대한 전반적인 질적 평가는 이러한 포함오류와 내용오류의 추정을 통해 이루어지며 오류 추정치들은 좀 더 향상된 센서스 결과를 위해 필연적으로 필요하다.

오류의 추정방법은 크게 인구통계학적 분석(Demographic Analysis: DA)과 센서스 후 조사(Post Enumeration Survey: PES)로 나눌 수 있다. DA를 거시적인 관점에서의 오류추정방법이라고 하면 PES는 미시적 관점에서의 오류추정방법이라고 할 수 있다(Robinson et al. 1993). DA는 주로 출생, 사망, 이민 등의 동태자료(vital statistics)를 이용하여 소위 '인구 방정식'을 이용한 분석방법이며, PES는 표본의 자료를 센서스 자료와 case−by−case로 비교(matching)하여 오류를 추정하는 방법이기 때문이다. 이 두 가지의 방법은 추정방법뿐만 아니라 사용하는 자료의 출처(source)가 완전히 다르기 때문에, 이 두 추정치의 불일치는 센서스의 질적 신뢰도에 문제가 있다는 것을 의미하며 그렇지 않으면 DA나 PES가 방법론적인 문제가 있음을 시사하게 된다(U.S. Census Bureau 2004).

특히 DA는 센서스의 포함오류 추정, 센서스간 오류의 추이(trend), 성별, 연령간 오류의 차이를 측정하는 매우 유용한 도구이다. 미국의 경우, 1960년 이래로 10년마다 시행되는 센서스의 질적 평가를 위해 PES와 함께 DA 추정치가 사용되고 있으며(U.S. Census Bureau 2004), WHO(World Health Organization)는 약 100여개 국가의 성인사망률(adult mortality)를 점검하기 위해 DA 방법을 사용하고 있다(WHO 2010). 그러나 한국에서는 센서스의 질적 평가를 위해 DA 방법이 체계적으로 활용되지 못하고 있는 실정이다.

본 연구에서는 5년마다 시행된 한국 센서스 자료(1985~2010년)를 이용하여, 센서스의 포함오류와 동태자료의 질적 평가를 위한 인구통계학적 분석을 하고자 한다. 한국 자료에 대한 인구통계학적 분석을 실시하는 또 하나의 이유는 국제적인 인구통계학 논문에서 한국에 대한 DA 추정치의 왜곡 또는 오류를 바로잡기 위함이다. Bennett & Horiuchi (1981)과 Hill (1987)은 한국의 1970~1975년 센서스 자료를 이용하여 한국의 사망등록률이 62%~65%라고 추정하여 한국의 사망등록 시스템의 후진성을 간접적으로 나타내고 있다. 한편, Murry et al.(2010)은 이러한 오류를 어느 정도 수정했으나 1970~2005년에 걸친 0~4세에 대한 사망등록률을 50% 내외로 추정하는 심각한 오류를 범하고 있다.

인구통계학적 분석은 출생, 사망, 국제이동 등의 행정등록자료(administrative register data)로부터 구한 연령별 인구추정치를 센서스 결과와 비교함으로써, 주로 센서스와 사망등록률의 완성도(completeness)를 측정한다. 센서스의 완성도가 1보다 작다는 것은 센서스가 참값보다 과소집계(undercount)되었다는 것을 의미하며 1보다 크면 센서스가 중복(duplicate)에 의해 과다집계(overcount)되었다고 말한다. 사망 자료의 경우도 유사하게 해석할 수 있다.

DA에 필수적인 자료는 성별·연령별 센서스 인구자료, 사망자수, 그리고 국제인구이동 자료이다. 한국의 경우, 1983년부터 성별연령별 사망자수가 존재하므로 1985부터 2010년까지 시행된 센서스 자료, 즉 6개의 센서스 자료를 이용하여 DA를 하고자 한다. 이 6개의 센서스 자료는 성별로 0세부터 84세까지 연령별 인구를 제공하고 있으며 85세 이상은 85+로 열린 연령구간(open-ended interval)으로 정리되어 있다. DA의 출발점은 다음의 인구방정식이다.

$$N_c(t+x) = N_c(t) - D_c + I_c - O_c \quad (1)$$

여기에서 $N_c(t)$는 시점 t에서 코호트 c의 인구이며, D_c는 시점 $t \sim t+x$사이에서 발생한 코호트 c의 사망자수, I_c와 O_c는 각각 시점 $t \sim t+x$사이에 발생한 코호트 c의 입국자수(immigration)과 출국자수(outmigration)이며 $I_c - O_c$를 순이민자수(net immigration)라고 정의한다. 이민자수에 대한 자료는 2000년부터 존재하며 <표 1>은 순이민자수를 연령별로 정리한 표이다.

<표 1>에서 음의 값은 출국자가 입국자보다 많다는 것을 의미한다. 그러므로 순이민자수에 대한 가장 큰 특징은 남자의 경우 39세까지, 여자의 경우 44세까지 출국자수가 입국자수보다 많으며 특히 24세까지는 이 차이가 매우 크게 나타난다는 것이다. 이러한 현상은 인구의 국제이동이 없다는 폐쇄인구(closed population) 가정하에서 이론을 전개하는 DA 방법론에서 유의할 사항이다.

<표 1> 성별 연령별 순이민자수

(단위: 명)

연 령	남 자			여 자		
	2000	2005	2010	2000	2005	2010
0~4	-4,771	-17,721	-17,581	-3,920	-15,242	-15,461
5~9	-3,698	-17,116	-13,053	-2,967	-15,361	-11,998
10~14	-4,938	-27,247	-16,428	-4,188	-22,648	-14,170
15~19	-4,206	-27,591	-18,081	-5,467	-26,202	-19,437
20~24	-8,159	-29,917	-19,384	-9,154	-48,651	-39,853
25~29	-3,433	-26,287	-15,225	-6,240	-21,636	-1,039
30~34	-1,468	-8,198	-871	-3,637	-17,983	-6,701
35~39	-446	-4,885	-1,514	-3,757	-16,895	-8,215
40~44	-868	-4,843	181	-2,991	-15,142	-4,731
45~49	-296	-2,503	4,661	-893	-6,455	2,322
50~54	204	1,623	6,651	-314	-301	5,022
55~59	336	2,794	4,902	-230	0	3,119
60~64	96	1,823	3,957	-255	-232	2,317
65~69	-41	684	2,020	-254	-228	1,681
70~74	-34	-9	1,110	-35	108	1,359
75~79	-50	-71	434	-58	-87	713
80~84	-6	-101	138	-35	-69	505
Total	-31,797	-159,547	-78,020	-44,393	-206,971	-104,301

〈표 2〉 센서스 여성인구

(단위: 천명)

연 도	0~4	5~9	10~14	15~19	20~24	25~29	30~34	35~39	40~44
2000	1,489	1,613	1,449	1,778	1,820	2,040	2,025	2,069	1,966
2005	1,145	1,515	1,619	1,474	1,746	1,814	2,036	2,047	2,040
2010	1,077	1,151	1,518	1,612	1,430	1,736	1,829	2,039	2,060

〈표 2〉는 2000~2010년 44세 이하 센서스 여성인구이다. 이 여성인구의 어떤 코호트(cohort)도 〈표 1〉의 음의 순이민자수 때문에 시간이 지남에 따라 인구가 단조감소(monotone decreasing)해야 하지만 이러한 규칙을 따르는 코호트는 2000년 15~19세와 35~39세 두 개에 불과한 것을 쉽게 알 수 있다. 이는 센서스에서 포함오류가 발생하였거나 동태자료에서 문제가 있음을 시사하고 있다. 특히 각 센서스의 0~4세 코호트의 크기가 5년 후(즉, 5~9세)에 모두 증가한다는 것은 센서스의 포함오류가 아니면 출생자 등록 및 영아사망자 등록시스템의 완성도(completeness)가 낮을 가능성이 높다고 할 수 있다. 이러한 0~4세에서 5~9세의 인구증가 현상은 〈표 2〉의 2000년~2010년 센서스뿐만 아니라 1985년과 1995년 센서스(1990년만 제외하고)에도 모두 발생하는 현상이다.

센서스 및 동태자료의 완성도 문제는 남성 대비 여성의 성비에서도 발견될 수 있다. 일반적으로 출생성비는 여성 100에 남성 104~106으로부터 출발하여 20세 후반~30세 전반에 비율이 100:100으로 맞추어진 후 이후 점차적으로 여성 100에 남성 50~60으로 감소하게 된다(Poston 2006). 〈표 3〉의 한국의 성비에서도 이러한 일반적인 규칙이 나타나고 있다. 그러나 1995년 0~19세까지 4개 코호트의 성비가 증가하고 있으며, 특히 모든 센서스 연도(1985년과 1990년 포함)의 15~19세 코호트의 성비가 20~24세에 증가하였다는 것은 20~24세 센서스 인구의 포함오류가 성별로 큰 차이를 보일 가능성이 높다는 것을 의미한다.

DA관점에서 본 연구의 주 관심사를 정리하면 다음과 같이 7가지 의문점으로 요약할 수 있다. 한국 센서스의 포함오류는 어느 정도인가? 이 포함오류의 추이는 어떠한가? 한국의 사망자 등록시스템을 통한 사망자 등록률은 우수한가? 사망자 등록률의 성별·연령별 패턴과 추이는 어떠한가? 센서스별·성별·연령별 포함오류는 어느 정도인가? 연령선호(age preference) 또는 연령집적(年齡集積; age heaping) 등에 의한 내용오류는 있는가? 마지막으로 이러한 의문점들은 앞에서 열거한 비정상적인 코호트별 인구증가 현상과 성비에서 지적된 문제를 해결할 수 있는가? 이다.

〈표 3〉 년도별 연령별 성비

연 령	연 도					
	1985	1990	1995	2000	2005	2010
0~4	1.0803	1.1120	1.1340	1.1021	1.0607	1.1071
5~9	1.0711	1.0727	1.1074	1.1357	1.0921	1.0958
10~14	1.0670	1.0604	1.0643	1.1142	1.1222	1.0900
15~19	1.0662	1.0391	1.0589	1.0766	1.1033	1.1327
20~24	1.0614	1.0915	1.0830	1.1144	1.0972	1.1366
25~29	0.9922	0.9946	1.0092	1.0087	1.0247	1.0384
30~34	1.0419	1.0377	1.0300	1.0213	1.0116	1.0205
35~39	1.0538	1.0613	1.0355	1.0232	1.0091	1.0105
40~44	1.0277	1.0744	1.0594	1.0318	1.0205	1.0056
45~49	0.9969	1.0233	1.0488	1.0276	1.0118	1.0078
50~54	0.9142	0.9793	0.9942	1.0174	0.9985	0.9884
55~59	0.7927	0.8830	0.9331	0.9513	0.9788	0.9679
60~64	0.7774	0.7473	0.8202	0.8783	0.9051	0.9394
65~69	0.7371	0.7163	0.6754	0.7594	0.8180	0.8512
70~74	0.6133	0.6448	0.6264	0.6110	0.6963	0.7534
75~79	0.4963	0.5131	0.5437	0.5430	0.5454	0.6097
80~84	0.3563	0.3906	0.4074	0.4491	0.4625	0.4542

이 논문은 총 4개의 절로 구성되어 있으며 제 2절에서는 인구통계학적 분석방법에 대한 기존의 연구방법을 살펴보고, 제3절에서는 이러한 DA 방법에서 발생하는 연령구간의 선택문제와 모수추정 및 계산문제를 해결하기 위한 수정된 DA 방법을 제안한다. 제 4절에서는 1985~2010년 동안 5년마다 실시된 센서스 자료를 이용하여 한국 자료에 가장 잘 적합되는 DA의 연령구간을 제시하고, 기존의 4가지 DA 방법과 새로운 2가지 DA 방법을 비교한 후, 위의 7가지 의문점에 대한 DA 결과를 제시하고 논의할 것이다.

Ⅱ. 인구통계학적 분석방법(Demographic Analysis)

센서스 자료의 질적 평가나 성인사망률(adult mortality)를 추정하기 위한 DA 방법은 크게 센서스 자료 대비 사망 자료의 완성도를 측정하는 사망분포방법(death distribution method), 센서스간(intercensal) 인구를 비교하는 방법, 표준생명표함수(standard life

table function) 등 세 가지로 분류할 수 있다(Hill et al. 2009). 이 중 가장 널리 사용되는 방법은 사망분포방법이며, 이 방법은 다시 General Growth Balance(GGB)와 Synthetic Extinct Generations(SEG)로 구분할 수 있는데 GGB는 센서스와 사망자 자료의 완성도를 측정하기 위한 방법이며, SEG는 사망자 자료의 완성도를 측정하기 위한 방법이다.

GGB는 인구의 국제이동이 없다는 폐쇄인구(closed population)가정과 인구증가율이 연령에 의존하지 않는다는 가정에서 인구증가율 = 출생률 − 사망률의 관계로부터 도출된

$$r = \frac{N(a)}{N(a+)} - \frac{D(a+)}{N(a+)} \quad \text{또는} \quad \frac{N(a)}{N(a+)} = r + \frac{D(a+)}{N(a+)} \tag{2}$$

의 관계식을 이용한다(Brass 1975). 여기에서 $N(a)$는 연령 a에서의 인구, $N(a+)$는 연령 a 이상 인구, 그리고 $D(a+)$는 연령 a 이상의 총 사망자수를 말한다. Martin(1980)은 인구증가율이 연령에 무관하다는 가정을 완화한

$$\frac{N(a)}{N(a+)} = r(a+) + \frac{D(a+)}{N(a+)} \tag{3}$$

을 제안하였고, Hill(1987)은 인접한 두 센서스 자료에 대하여 N_1°와 N_2°를 각각 첫 번째와 두 번째의 관찰된 센서스 인구라고 정의하고 D°를 두 센서스 사이에서 발생한 관찰된 사망자수라고 정의한 후,

$$N_1 = N_1^\circ / K_1, \; N_2 = N_2^\circ / K_2, \; D = D^\circ / K_3 \tag{4}$$

으로 참값(N_1, N_2, D)과 관찰값의 관계를 설정하였다. 그러므로 K_1은 첫 번째 센서스의 완성도, K_2는 두 번째 센서스의 완성도, 그리고 K_3를 사망자등록 완성도로 정의하였다. 두 센서스의 시점 차이를 t라고 할 때, 연령 a 이상의 증가율을

$$r(a+) = \frac{1}{t} log \frac{N_2(a+)}{N_1(a+)}$$

으로 정의하고, 관찰된 증가율인 $r^\circ(a+)$ 및 식(4)를 식(3)에 대입하면

$$\frac{N^\circ(a)}{N^\circ(a+)} - r^\circ(a+) = \frac{1}{t} log \frac{K_1}{K_2} + \frac{(K_1 K_2)^{1/2}}{K_3} \frac{D^\circ(a+)}{N^\circ(a+)} \tag{5}$$

가 됨을 보일 수 있다(Hill 1987). 여기서 $N^{\circ}(a+)$ 및 $N^{\circ}(a)$는 인구의 연령분포가 정상성이라는 가정하에 기하평균인

$$N^{\circ}(a+) = t[N_1^{\circ}(a+)N_2^{\circ}(a+)]^{1/2},$$
$$N^{\circ}(a) = \frac{t}{5}[N_1^{\circ}(a-5,a)N_2^{\circ}(a-5,a)N_1^{\circ}(a,a+5)N_2^{\circ}(a,a+5)]^{1/4}$$

이다. Hill(1987)은 식(5)를 회귀분석을 통해 추정된 절편과 기울기를 이용하여 K_1 대비 K_2와 K_3를 구하는 방법을 제안하고 있다. 그러므로 GGB방법을 이용하기 위해서는 폐쇄인구이며 연령분포가 정상성이고 센서스와 사망자등록의 완성도가 연령에 무관하게 동일하다는 가정이 필요하다.

폐쇄인구 가정하에서 전개되는 SEG방법은 매우 간단한 항등식으로부터 출발한다. $N_t(a)$와 $D_t(a)$를 각각 시점 t에서의 연령 a인구와 사망수라고 할 때

$$N_t(a) = \int_a^{\infty} D_{t+x-a}(x)\,dx \tag{6}$$

의 관계를 가지고 있다(Vincent 1951). 그러나 이 관계식을 이용하기 위해서는 코호트 $N_t(a)$가 완전히 죽어서 없어질 때까지 기다려야 하기 때문에 현실적으로 응용할 수 없다. 이를 극복하기 위해 Bennett and Horiuchi (1981, 1984)는 인구의 연령분포가 정상성이라는 가정에서 관계식(6)으로부터

$$N_t(a) = \int_a^{\infty} D_t(x) e^{\int_a^x r(u)du} dx \tag{7}$$

를 도출하였다. $N_t^{\circ}(a)$와 $D_t^{\circ}(x)$를 각각 시점 t에서의 센서스 및 사망자수 관찰치라고 정의할 때 (7)의 추정치를 $\widehat{N}_t(a) = \int_a^{\infty} D_t^{\circ}(x)\exp\left[\int_a^x r(u)du\right]dx$ 로 놓으면

$$\frac{\widehat{N}_t(a)}{N_t^{\circ}(a)} = \frac{\int_a^{\infty} D_t^{\circ}(x)\, e^{\int_a^x r(u)du} dx}{N_t^{\circ}(a)} = \frac{K_3 \int_a^{\infty} D_t(x)\, e^{\int_a^x r(u)du} dx}{N_t^{\circ}(a)} \tag{8}$$
$$= \frac{N_t^{\circ}(a)K_3/K_1}{N_t^{\circ}(a)} = \frac{K_3}{K_1}$$

이 된다. 식(8)의 두 번째, 세 번째 항등식은 식(4)에 의해 성립한다. 연령별 인구증가율 $r(a)$는 두 개의 인접한 센서스 자료로부터 $K_1 = K_2$라는 가정하에서 구하게 된다. 그러므로 SEG는 GGB의 가정에 인접한 두 개 센서스의 완성도(즉, $K_1 = K_2$)가 같다는 가정이 추가적으로 필요하다. 그러나 $K_1 \neq K_2$이면 $r(a)$의 추정치가 편향(bias)되어 사망등록의 완성도 K_3가 편향되게 추정되게 된다(Bennett & Horiuch 1981; Hill et al. 2009). 이를 극복하기 위해 Bennett & Horiuch(1981)은 $\widehat{N}_t(a)/N_t^{\circ}(a)$가 연령 a에 의존하지 않도록 연령별 인구증가율 $r(a)$에 적절한 상수를 더해주는 확장된 SEG(extended SEG: ESEG)를 제안하였다.

한편, Hill et al.(2009)는 GGB를 이용하여 K_1대비 K_2를 구한 후, SEG를 이용하여 K_3를 구하는 GGB−SEG 방법을 제안하였다. 첫 번째 센서스 N_1°에 K_2/K_1을 곱하면 $N_1^{\circ} = K_1 N_1$이므로 $N_1^{\circ} \times K_2/K_1 = K_2 N_1$이 되어 $N_2^{\circ} = K_2 \cdot N_2$와 동일한 완성도 K_2가 된다.그러므로 GGB−SEG는 첫 번째 센서스와 두 번째 센서스의 완성도가 같아야 한다는 SEG의 조건을 만족하게 된다.

지금까지 논의한 GGB, SEG 그리고 GGB−SEG는 모두 연령선호 또는 연령집적이 없다는 가정과 폐쇄인구 가정으로 인해 인구의 국제이동이 없다는 가정을 하고 있다. 이러한 가정을 근사적으로나마 충족시키기 위해서는 필연적으로 모형(5) 또는 (8)에 적합할 연령구간(즉, 연령 a)의 선택이 매우 중요하다.

Hill et al.(2009)는 GGB나 SEG의 가정을 위배한 유형에 따라 총 98개의 시나리오 시뮬레이션을 실시한 결과, 5~65세보다는 15~55세가 GGB와 SEG의 적합구간으로 좀 더 좋은 결과를 보였다. 국제이동이 없을 경우, GGB−SEG가 가장 좋은 결과를 보여주었으며 국제이동이 있는 경우에는 GGB와 SEG가 매우 불안정하여 30~65의 연령구간을 제안하고 있다. Murry et. al (2010)은 78개의 연령구간을 세 개의 서로 다른 시뮬레이션 환경에 적용하여 GGB는 40~70세, SEG는 55~80세, 그리고 GGB−SEG는 50~70세 연령구간에서 가장 좋은 결과를 보인다고 보고하고 있다. 본 연구에서는 Hill et al.(2009)가 제안한 3개의 연령구간과 Murry et al.(2010)이 제안한 3개의 연령구간을 한국자료에 적용하여 이들 연령구간을 비교하고 그 특성을 살펴보고자 한다.

센서스의 포함오류나 사망자등록률의 완성도를 측정하기 위한 GGB, SEG 또는 GGB−SEG는 인구의 연령분포가 정상성이고 폐쇄인구이며 센서스의 완성도가 연령에 의존하지 않는다는 가정을 하고 있다. Preston et al.(1998)은 연령분포의 정상성이나 폐

쇄인구의 가정이 없이 연령별 센서스 완성도를 측정하고 코호트별 인구를 재구축(reconstruction)하기 위한 Age - Period - Cohort(APC)모형을 고려하였다. 이 APC 모형의 적용은 코호트 자료의 생성을 위해 다수의 센서스 자료가 존재해야만 하고 사망자와 국제인구이동 자료의 신뢰도가 높다는 전제가 있어야 한다. X_{it}를 시점 t에서 i번째 코호트의 인구라고 정의하면

$$X_{it} = \gamma_i - D_{it} \tag{9}$$

가 된다. 여기에서 γ_i는 i번째 코호트가 자료구간에서 처음으로 나타났을 때의 인구이며 D_{it}는 γ_i가 측정된 시점부터 t 시점까지 사망한 코호트 i의 누적사망자수와 누적순이민자수의 합이다. Preston et. al (1998)은 X_{it}°를 센서스 시점 t에서 관찰된 코호트 i의 인구수라고 할 때 연령에 의존하는 센서스의 완성도를 측정하기 위해 아래와 같은 승법모형(multiplicative model)을 제안하였다.

$$E(X_{it}^{\circ}|\alpha_j,\tau_t,\gamma_i) = \alpha_j\tau_t X_{it} = \alpha_j\tau_t(\gamma_i - D_{it}) \tag{10}$$

여기에서 α_j는 j번째 연령그룹의 효과를 나타내며 τ_t는 시점 t에서의 센서스 완성도를 나타낸다. 그러므로 $\alpha_j\tau_t$는 시점 t에서 j번째 연령그룹의 완성도를 나타내게 된다. 식 (10)은

$$X_{it}^{\circ} = \alpha_j\tau_t(\gamma_i - D_{it}) + \epsilon_{it}$$

로 재표현할 수 있으며(단, ϵ_{it}는 평균이 0인 오차항이다)

$$SSE = \sum_{i,t}(X^{o} - \alpha_j\tau_t(\gamma_i - D_{it}))^2 \tag{11}$$

를 최소화하는 α_j, τ_t 그리고 γ_i를 추정한다. SSE를 최소화하는 APC의 최대 난점은 자료에 비해 추정해야 할 모수가 너무 많다는 것이다. 이를 해소하기 위해 Preston et al.(1998)은 미국의 경우 Medicare 자료 등의 보조 자료를 이용하였으나, 우리나라의 경우 이와 같은 자료가 존재하지 않아 APC 모형을 직접 사용하기에는 문제가 있다고 할 수 있다.

Ⅲ. 새로운 인구통계학적 분석방법

Bennett & Horiuchi(1981, 1984)가 제안한 SEG에 의한 사망자 등록률의 완성도는 식(7)에서 주어진

$$N(a)=\int_a^{\infty} D_t(x)\exp[\int_a^x r(u)\,du]dx$$

를 얼마나 정확하게 구할 수 있느냐에 달려있다. $k \geq 1$에 대해

$$\begin{aligned} N(a-k) &= \int_{a-k}^{\infty} D(x)\exp[\int_{a-k}^{x} r(u)du]dx \\ &= \int_a^{\infty} D(x)\exp[\int_a^x r(u)du+\int_{a-k}^{a} r(u)du]dx+\int_{a-k}^{a} D(x)\exp[\int_{a-k}^{x} r(u)du]dx \\ &= N(a)\exp[\int_{a-k}^{a} r(u)du]+\int_{a-k}^{a} D(x)\exp[\int_{a-k}^{x} r(u)du]dx \end{aligned} \tag{12}$$

이 된다. Bennett & Horiuchi(1981)은 $k=5$(즉, 5세 단위 인구)를 취하고

$$\exp\left[\int_{a-5}^{a} r(u)du\right] \approx \exp[5\times {}_5r_{a-5}]$$

$$\int_{a-5}^{a} D(x)\exp\left[\int_{a-5}^{x} r(u)du\right]dx \approx {}_5D_{a-5}\exp[2.5\times {}_5r_{a-5}]$$

로 근사하여

$$N(a-5)=N(a)\exp[5\times {}_5r_{a-5}]+{}_5D_{a-5}\exp[2.5\times {}_5r_{a-5}] \tag{13}$$

을 이용하였다. 그러나 우리나라와 같이 사망률이 급격하게 감소할 경우, 근사식(13)은 정확도가 상당히 떨어지는 것으로 지적되고 있다(Bennett &Horiuchi 1984). 식(12)에서 $k=1$로 놓으면

$$N(a-1)=N(a)\exp[r(a-1)]+D(a-1)\exp[r(a-1)] \quad (14)$$

이 되어 식(13) 과 같은 근사식이 필요없게 된다. 즉, 1세 단위로 인구를 계산하게 되면 근사식이 필요없는 항등식(14)를 얻게 된다. Bemmett & Horiuchi(1981)은 5세 단위의 인구를 $2.5[N(a)+N(a+5)]$로 추정하여 K_3를 구하였으나, 식(14)를 이용하면 5세 단위 인구를 $\sum_{k=0}^{4}N(a+k)$로 직접 계산하여 추정오차를 없앨 수 있는 장점이 있다.

Bennet & Horiuchi(1981)가 $k=5$를 선택하여 5세 단위 인구를 사용한 근본적인 이유는 연령선호, 연령집적, 또는 연령과장(年齡誇張; age overstatement) 등으로 인한 연령보고오류(age misreporting)을 최소화하기 위해서이다(United Nations 1983; Hill et al. 2009). 그러므로 항등식(14)를 사용하기 위해서는 연령보고오류가 거의 없다는 전제조건이 필요하다. 식(14)를 사용한 SEG를 수정된 SEG(Modified SEG: MSEG)라고 정의하여 SEG나 ESEG와 구별하고자 한다.

Preston et al.(1998) 이 제안한 APC(Age - Period - Cohort) 모형을 적용하기 위해서는 자유도 문제 때문에 미국의 medicare 자료 등과 같은 보조정보가 충분히 있거나 추정해야 할 모수를 줄이는 방법을 고려해야 한다.

〈표 4〉는 APC 방법을 적용할 때 추정해야 할 모수를 보여주고 있다. 우리나라의 경우, 1985년 자료가 분석에 사용되는 첫 연도이기 때문에 $\gamma_1 \sim \gamma_{17}$과 5세 단위 전 연령(즉, $\gamma_{18} \sim \gamma_{22}$) 역시 추정해야 할 코호트들이다. $\tau_1 \sim \tau_6$은 6개 센서스의 완성도를 나타내며, $\alpha_1 \sim \alpha_{17}$은 연령그룹의 효과를 나타내고 있다.

〈표 4〉 APC에서 추정해야 할 모수

연 령	1985(τ_1)	1990(τ_2)	1995(τ_3)	2000(τ_4)	2005(τ_5)	2010(τ_6)
0-4(α_1)	γ_1	γ_{18}	γ_{19}	γ_{20}	γ_{21}	γ_{22}
5-9(α_2)	γ_2					
⋮	⋮					
80-84(α_{17})	γ_{17}					

Preston et al.(1998)이 제안한 목적함수인 $\sum_{i,t}\left(X_{it}^{\circ}-\alpha_j\tau_t(\gamma_j-D_{it})\right)^2$을 최소화하는 모수추정방법은 두 가지 문제점을 가지고 있다. 첫 번째는 X_{it}°의 값이 작은 연령대(즉, 60대 이후 연령대)의 경우 이 목적함수는 오차가 크게 나타날 가능성이 크다. 두 번째는 45개의 모수를 추정하기에는 자유도가 $102-45=57$개로 너무 작아 모수추정치의 신뢰성을 확보하기 어렵다. 이를 극복하기 위해 목적함수를

$$\sum_{i,t}\frac{\left(X_{it}^{\circ}-\alpha_j\tau_t(\gamma_j-D_{it})\right)^2}{X_{it}^{\circ}} \tag{15}$$

로 하고, 이 목적함수를 최소화하는 모수를 추정하여 인구가 많거나 작은 연령에 동일한 비중(weight)를 두고자 한다. 두 번째 문제인 자유도를 늘이기 위해 GGB에 추정된 τ_t를 목적함수 (15)에 대입하고 코호트의 길이가 2개 이하인 코호트(즉, γ_{16}, γ_{17}, γ_{21}, γ_{22})는 모수에서 제외한다. 이 모수들은 Bennett & Horiuchi(1981)의 가정에 따라

$$E(X_{it}^{\circ})=\alpha_j\tau_t(\gamma_i-D_{it})$$

이므로

$$\hat{\gamma_i}=\frac{X_{it}^{\circ}}{\alpha_j\tau_t}+D_{it}$$

으로 추정한다.

이와 같이 GGB와 결합된 APC 방법을 GGB-APC 라고 명명하고, 이를 통해 연령별 센서스 완성도를 측정하고 추정된 γ_i와 α_j를 기반으로 인구방정식을 통해 연령별 인구를 재구축하고자 한다. 재구축된 인구는 센서스 인구와의 비교를 통해 연도별 연령별 누락률(omission rate)를 제시하고 제 2절에서 지적한 <표 2>의 여성인구에서의 이상현상과 <표 3>의 남녀성비에서 발생하는 문제 등이 얼마나 해결되는지를 살펴보고자 한다.

Ⅳ. 자료 분석

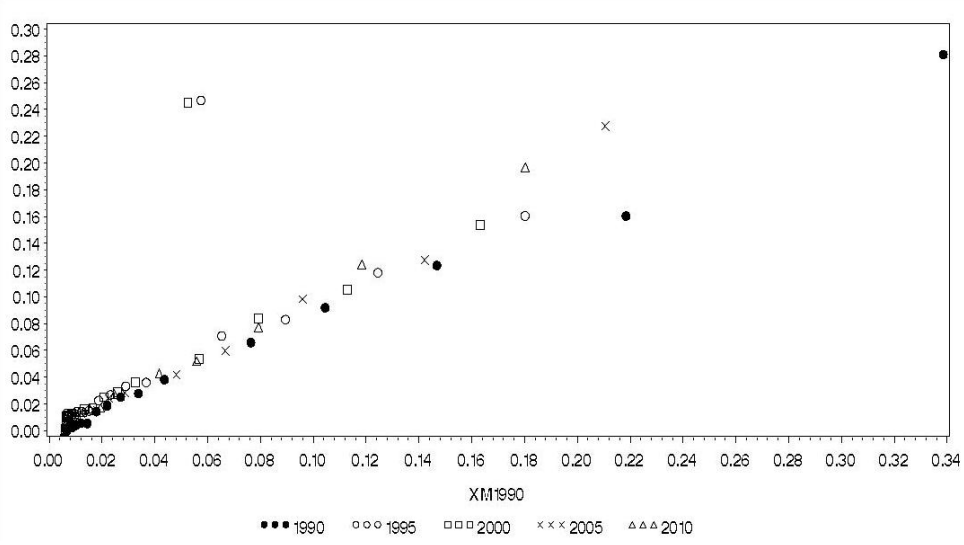
1. GGB 와 SEG를 이용한 완성도

1985년부터 2010년까지 5년마다 실시된 센서스 자료와 통계청이 발표한 월별 사망자

자료, 그리고 2000년부터 제공되는 국제인구이동 자료를 이용하여 인구통계학적 분석을 하고자 한다.

먼저 식(5)에 제시된 Hill (1987)의 GGB를 적용하기 위하여 $(D^{\circ}(a+)/N^{\circ}(a+),\ N^{\circ}(a)/N^{\circ}(a+)-r^{\circ}(a+))$의 플롯을 1985~2010년에 걸쳐 $a=5\sim a=75$세까지 그린 것이 <그림 1>이다.

<그림 1>에서 볼 수 있듯이 어린 연령에서는 직선에서 많이 벗어나 있고, 연도별로 플롯의 값들의 변동이 심한 것으로 보인다. Hill et al.(2009)의 5~65세, 15~55세, 그리고 30~65세 연령구간과 Murray et al.(2010)의 40~70세, 55~80세, 그리고 50~70세 연령구간 등 총 6개의 연령구간을 고려한 모든 경우 중에서 GGB 모형의 결정계수가 가장 큰, 즉 설명력이 가장 우수한 것으로 나타난 연령구간은 30~65세와 40~70세 연령구간으로 나타났다. <표 5> 는 30~65세와 40~70세의 연령구간에 대한 단순회귀의 결정계수를 정리한 표이다. 또한 제1절의 <표 1>의 순이민자수를 살펴보면 남자의 경우 30세 이상, 여자의 경우는 최소한 40세 이상일 때 순이민자수가 급격히 줄어들게 됨을 알 수 있다. 이는 폐쇄인구를 가정하고 있는 GGB의 적용을 위해서 남자의 경우 최소한 30세 이상, 여자의 경우는 40세 이상을 고려하는 것이 타당함을 의미한다.



<그림 1> 1985부터 2010년의 GGB플롯

〈표 5〉 30~65세와 40~70세 연령구간에 대한 적합도

		1985~1990	1990~1995	1995~2000	2000~2005	2005~2010
남	30~65	0.975	0.984	0.979	0.971	0.986
	40~70	0.990	0.981	0.977	0.983	0.982
여	30~65	0.980	0.982	0.979	0.972	0.976
	40~70	0.995	0.993	0.986	0.975	0.986

〈표 6〉 첫 번째 센서스 대비 두 번째 센서스 및 사망자 등록률 완성도

성 별	연령구간	완성도	1985 ~1990	1990 ~1995	1995 ~2000	2000 ~2005	2005 ~2010
남자	30~65	K_2	1.018	0.993	0.974	0.989	1.000
		K_3	1.048	0.971	1.150	1.130	0.969
	40~70	K_2	1.025	0.982	0.979	0.987	0.997
		K_3	1.027	1.016	1.122	1.139	0.988
여자	30~65	K_2	0.997	0.988	0.984	0.999	1.005
		K_3	1.080	1.129	1.146	1.103	1.035
	40~70	K_2	0.992	0.981	0.993	0.999	0.999
		K_3	1.088	1.183	1.051	1.106	1.133

〈표 5〉에 따르면 남성의 경우 30~65세 구간과 40~70세 구간의 결정계수 가운데 어느 것이 좋다고 말할 수 없으나, 여성의 경우 40~70세 구간의 결정계수가 30~65세 구간보다 높게 나와 GGB의 기본가정을 좀더 잘 충족하는 연령구간으로 판단된다.

〈표 6〉은 첫 번째 센서스 대비 두 번째 센서스 및 사망자 등록률의 완성도를 나타낸다. 예를 들어 1995~2000년인 경우 K_2는 1995년 대비 2000년 센서스의 완성도를 나타내며, K_3는 1995~2000년 사이에 발생한 사망자 등록률의 완성도를 나타낸다. K_2가 1보다 작다는 것은 두 번째 센서스가 과소집계(undercount)되었다는 것을 의미하며, 1보다 크다는 것은 과다집계(overcount)되었다는 뜻이다. K_3도 동일한 해석을 할 수 있다.

완성도 K_2나 K_3가 30~65세 연령구간과 40~70세 연령구간에서 매우 비슷한 값을 가지고 있으며 특히 K_2는 연령구간과 성별에 관계없이 $1 \pm 3\%$ 내의 값을 가지고 있어 한국 센서스의 완성도는 매우 우수한 것으로 나타났다. 그러나 사망자 등록률의 완성도는 남자의 경우(30~65세 기준), 1995~2000년과 2000~2005년의 사망자 등록률이 과다집계된 것으로 보이며, 여자의 경우(40~70세 기준) 1990~1995년, 2000~2005년, 그리고 2005~2010년에서 다소 과다집계된 것으로 판단할 수 있다.

사망자 등록률의 완성도를 측정하는 또 다른 방법은 SEG이다. Bennett & Horiuchi (1981)의 제안에 따라 5세 단위 인구증가율을 이용하여 센서스 인구집계의 끝 연령인 85세에 대해

$$N(85) = D(85+)\left[\exp\{r(85+)e(85)\} - \{r(85+)e(85)\}^2/6\right] \tag{16}$$

를 구한다. 여기서 $e(85)$는 인구 증가율을 구하기 위해 사용된 첫 번째 센서스에서의 85세의 기대여명(life expectancy)이다. 각 센서스 연도별 85세의 기대여명은 1985년은 남녀별로 2.64년, 4.44년이며, 1990년은 3.22년, 4.57년, 1995년은 4.27년, 5.32년, 2000년은 4.50년, 5.60년, 2005년은 5.17년, 6.42년이며 2010년의 경우 6.66년 및 8.73년으로 계산되었다. 식(13)과 식(16)을 이용하여 5세 단위로 $N(a)$를 구하고 5세 단위 인구를

$$\hat{}_5N_a = 2.5\left[N(a) + N(a+5)\right]$$

로 추정한 후, $\hat{}_{10}N_a = \hat{}_5N_a + \hat{}_5N_{a+5}$ 로 10세 단위 인구 추정치를 구한다. 이 값을 첫 번째 센서스의 10세 단위인구 $_{10}N_a$로 나누어, 이 비율들의 중위수로 사망자 등록률의 완성도(즉, K_3)를 구하게 된다.

<표 7>은 이러한 계산과정을 통해 구한 SEG와 ESEG의 K_3이다. 1985~1990년 남자의 SEG 추정치를 제외하고 모두 K_3가 1보다 현격하게 낮은 값을 보이고 있다. 이러한 결과는 사망자 등록을 해야만 매장이 가능한 우리나라 사망자 등록시스템을 고려할 때 거의 불가능한 사망자 등록률의 완성도일 뿐만 아니라 <표 6>에서 제시된 GGB의 K_3와 비교하더라도 문제점이 많은 것으로 보인다. 제 3절에서 지적하였듯이 근사식(13)의 정확성 때문으로 판단되는 근거를 다음 절에서 보여주게 될 것이다.

〈표 7〉 Bennett & Horiuchi의 SEG 와 ESEG 에 의한 K_3

		1985~1990	1990~1995	1995~2000	2000~2005	2005~2010
남	SEG	1.053	0.745	0.689	0.702	0.646
	ESEG	0.867	0.745	0.689	0.702	0.544
여	SEG	0.718	0.716	0.705	0.759	0.850
	ESEG	0.615	0.716	0.705	0.759	0.687

2. 연령 보고오류(age misreporting)와 수정된 SEG(MSEG)

인구통계학적 방법은 연령선호, 연령집적, 그리고 연령과장 등의 연령 보고오류가 없다는 가정을 하고 있다. 연령 보고오류 측정의 가장 일반적인 방법은 연령비(age ratio)이다. 연령비는 인접한 연령의 인구는 선형적으로 증가 또는 감소한다는 이론(Shrock & Siegel 1976)에 따라 5세별 인구를 이용하여 특정 연령인구의 2배로 이 연령의 선행 연령인구와 후행 연령인구의 합을 나눈 값이다. 이 값이 1에 가까우면 연령 보고오류가 없다는 것을 의미한다.

〈표 8〉의 남녀별 연령비를 살펴보면 남녀 모두 1995년 5~9세의 코호트, 1985년 25~29세 코호트, 그리고 1985년 남성 55~59세 코호트를 제외하고는 100%를 크게 벗어나지 않는 것으로 보인다. 1985년 남성 55~59세 코호트는 1950년 20~24 세로 6.25 전쟁에 의한 남성인구의 감소가 주 요인인 것으로 보이며(Poston 2006), 1985년 25~29세 코호트는 1956~1960년생으로 소위 베이비 부머(baby boomer)의 핵심연령이고, 1995년 5~9세 코호트는 1986~1990년생으로 우리나라의 출산률이 급격하게 줄어든 연령이다. 이러한 관점에서 볼 때, 연령비가 100%에서 크게 벗어나는 경우는 코호트적인 특성으로 해석할 수 있으므로 우리나라는 연령 보고오류가 없거나 심각하지 않은 것으로 판단된다. 다만, 남성의 경우 1985년 5~9세 코호트가 2000년에 대폭 증가하는 현상, 1985년 15~19세 코호트가 1995년에 대폭 감소하는 현상, 그리고 1985년 여성 20~24세 코호트가 1990년에 대폭 증가하는 현상은 좀더 면밀한 검토가 필요하다. 이러한 현상은 뒤에서 논의될 GGB-APC 방법의 적용에서 재논의될 것이다.

〈표 8〉 연령비

성 별	연령비	1985	1990	1995	2000	2005	2010
남성	5~9	95.7	105.7	87.1	112.5	108.3	88.9
	10~14	108.7	96.3	105.9	86.2	110.7	107.8
	15~19	99.1	104.3	95.7	105.1	87.2	111.3
	20~24	102.7	103.6	110.1	102.1	110.0	89.6
	25~29	107.4	97.4	94.8	100.4	93.5	103.3
	30~34	94.9	112.5	102.7	99.1	105.0	96.6
	35~39	98.2	95.3	112.9	103.4	99.7	104.6
	40~44	93.7	95.7	93.9	112.3	103.4	100.9
	45~49	108.7	95.3	96.7	93.1	111.8	103.3
	50~54	101.0	106.8	94.2	96.5	92.4	110.9
	55~59	89.7	102.2	108.5	94.9	97.0	92.4
	60~64	101.6	87.1	100.2	107.7	95.3	96.4
	65~69	97.2	103.2	87.0	100.3	107.1	96.3
	70~74	92.9	92.6	101.0	86.5	100.2	108.2
	75~79	91.3	88.8	88.0	95.6	83.2	95.6
여성	5~9	95.9	106.8	86.3	109.8	109.6	88.7
	10~14	108.8	95.8	107.5	85.5	108.3	109.9
	15~19	98.9	108.0	97.1	108.7	87.6	109.4
	20~24	99.7	96.5	105.0	95.4	106.2	85.4
	25~29	114.0	104.3	99.2	106.1	95.9	106.5
	30~34	92.5	110.8	101.9	98.6	105.5	96.9
	35~39	96.5	94.4	113.6	103.7	100.4	104.9
	40~44	93.7	93.1	92.2	111.6	102.4	101.3
	45~49	106.5	96.1	95.2	93.0	111.8	102.2
	50~54	101.0	104.8	94.4	94.5	92.5	111.2
	55~59	97.4	102.7	106.6	95.3	95.2	92.6
	60~64	100.9	95.5	101.8	106.4	95.5	94.4
	65~69	94.9	102.5	96.6	102.8	106.8	97.0
	70~74	99.5	93.5	102.1	97.3	104.0	108.1
	75~79	101.2	99.3	91.7	99.9	96.0	103.4

〈표 9〉 연령비 점수(ARS)와 연령정밀지수(AAI)

		1985	1990	1995	2000	2005	2010
ARS	남	5.17	5.98	6.93	6.13	7.22	6.30
	녀	4.24	5.05	5.72	5.97	6.16	6.78
AAI		4.70	5.52	6.32	6.05	6.69	6.54

〈표 9〉는 Shryock & Siegel(1976)가 제안한 연령비 점수(Age Ratio Score; ARS)와 Smith(1992)가 제안한 연령정밀지수(Age Accuracy Index; AAI)를 정리한 표이다.

ARS는 |연령비−100|의 평균으로 기준점인 100으로부터 평균적으로 얼마나 벗어나는지를 측정하는 측도이고, AAI는 남자와 여자의 ARS의 단순평균으로 성별에 관계없이 연령인구의 전반적인 정확도에 대한 측도이다. 여성보다는 남성의 ARS가 약간 크게 나타나고 있으며 AAI는 최대 6.69%으로 연령 보고오류(age misreporting)에 대한 우려는 하지 않아도 될 것으로 보인다(Poston 2006).

〈표 10〉 MSEG, GGB−SEG에 의한 사망자 등록률의 완성도

성 별	방 법	연 령	1985 ~1990	1990 ~1995	1995 ~2000	2000 ~2005	2005 ~2010
남자	MSEG		1.300	1.016	0.967	0.967	0.916
	GGB-MSEG	30~65	1.079	0.934	1.093	1.091	0.913
	GGB-MSEG	40~70	1.141	1.018	1.049	1.067	0.937
	GGB	30~65	1.048	0.971	1.150	1.130	0.969
	GGB	49~70	1.027	1.016	1.120	1.139	0.988
	Mean		1.119	0.991	1.076	1.079	0.945
여자	MSEG		1.110	1.039	1.018	1.056	1.039
	GGB-MSEG	30~65	1.025	1.03	1.12	1.117	0.989
	GGB-MSEG	40~70	1.003	1.135	1.06	1.061	1.047
	GGB	30~65	1.080	1.129	1.146	1.103	1.035
	GGB	49~70	1.088	1.183	1.051	1.106	1.133
	Mean		1.061	1.103	1.079	1.088	1.049

이러한 결과는 식(14)를 이용할 수 있는 이론적 근거를 제시하여 새롭게 제시한 MSEG를 적용할 수 있다. MSEG를 이용한 K_3값은 <표 10>의 남성과 여성의 첫 번째 행에 각각 제시되어 있다. 1985~1990년의 K_3를 제외하고는(특히, 남자의 경우) 사망자 등록률의 완성도에 큰 문제가 없는 것으로 보인다. 이는 <표 7>에 제시된 SEG의 K_3값과 비교할 때 큰 차이를 보이고 있어, Bennett & Horiuchi(1981)의 SEG를 적용하기에는 우리나라의 경우 문제가 있는 것으로 보인다.

Hill et al.(2009)가 제안한 GGB-SEG의 방법에 따라, GGB에 의해 구한 K_2/K_1 을 첫 번째 센서스에 곱해서 MSEG를 적용한 GGB-MSEG의 K_3가 <표 10>에 제공되어 있다. GGB를 30~65세에 적용했을 때와 40~70세에 적용했을 때의 K_2/K_1값의 차가 아주 작은데도 불구하고(지면 절약을 위해 생략, 요청시 제공 가능) K_3의 값이 민감하게 변하는 것을 볼 수 있다.

그러므로 Hill et al.(2009)가 K_3의 추정값으로 GGB와 SEG의 평균값으로 하자는 제안에 따라 <표 5>의 GGB 방법으로 구한 K_3값(<표 10>의 제4행과 5행에 제공)을 추가로 하여 MSEG, GGB-MSEG, 그리고 GGB의 5개 평균치를 K_3의 추정치로 하는 것이 합리적으로 보인다. 이 평균의 값은 <표 10>의 마지막 행에 제공되어 있으며, K_3의 값이 이상치인 1에 모두 근접해 있어 우리나라 사망자 등록률의 완성도가 비교적 우수한 것으로 나타난다.

3. GGB-APC에 의한 연도별 연령별 누락률의 추정

본 논문에서는 GGB-APC 모형에 사용할 연도효과(period effect)인 τ_t로 1절에서 제시한 GGB의 K_2를 이용하고자 한다. 그러나 GGB의 K_2는 인접한 두 센서스를 기준으로 구한 값이므로 기준년도가 모두 다르다. 따라서 하나의 연도를 기준으로 K_2를 재산출하여 이를 τ_t의 값으로 고정하고자 한다. 2010년이 가장 최근의 센서스 시행연도이나, 인터넷 센서스를 병행했으므로 2005년의 센서스를 기준으로 1절의 <표 6>의 값을 재산출하면 다음과 같다.

〈표 11〉 GGB−APC 모형에서의 연도효과(period effect) τ_t

성 별	연령구간	1985	1990	1995	2000	2005	2010
남자	30~65	1.027	1.045	1.038	1.012	1	1.000
	40~70	1.028	1.054	1.035	1.013	1	0.997
여자	30~65	1.033	1.029	1.017	1.000	1	1.005
	40~70	1.035	1.027	1.008	1.001	1	0.999

〈표 11〉은 2005년 센서스 대비 각 해당년도의 센서스의 완성도이다. 그러므로 2005년 센서스의 누락률(omission rate)가 0%이라고 가정한다면, 예를 들어 남성의 경우 30~65세 기준으로 1985년은 2.7%의 중복률을 보이고 있으며, 40~70세 기준으로 2.8%의 중복률을 보이고 있다고 해석할 수 있다. 또한 GGB 적합에 사용한 두 개의 연령구간 30~65세와 40~70세에서 구한 각각의 τ_t가 매우 유사하여 40~70세 연령구간으로만 논의를 진행하고자 한다. 목적함수인

$$\sum_{i,t} \frac{\left(X_{it}^{\circ} - \alpha_j \tau_t (\gamma_j - D_{it})\right)^2}{X_{it}^{\circ}}$$

를 최소화하는 α_j의 추정치를 정리하면 〈표 12〉와 같다.

〈표 12〉 연령효과(α_j)의 성별 추정치

연 령	남	녀	연 령	남	녀
0~4	1.017	0.982	45~49	0.989	0.995
5~9	1	1	50~54	0.984	0.981
10~14	1.022	1.025	55~59	0.980	0.982
15~19	1.040	1.028	60~64	0.941	0.946
20~24	1.060	1.010	65~69	0.921	0.963
25~29	0.997	1.017	70~74	0.929	0.988
30~34	0.980	0.978	75~79	0.920	0.998
35~39	0.966	0.992	80~84	0.886	1.004
40~44	0.979	0.999			

연령효과는 5~9세를 1로 했을 때의 상대적인 연령효과이다. Preston et al.(1998)이 5~9세를 기준으로 하였듯이, 한국의 경우도 5~9세가 학령인구로 진입하는 연령으로서 비교적 정확한 인구통계로 인식되고 있다. 남자의 경우 0~24세까지는 과다집계 또는 중복이 있는 것으로 나타났으며, 25세 이후는 과소집계 또는 누락이 있는 것으로 나타났다. 특히 60세 이후에 누락률이 매우 높게 나타난 것은 향후 센서스와 센서스 후 조사의 설계 및 정확도 검증 시 유의해야 할 부분이다. 여성의 경우 0~4세는 누락이, 5~29세까지는 중복이, 그리고 30세 이후는 누락이 있는 것으로 나타나고 있다. 그러나 여성의 경우는 남성과 달리 60세 이후에 누락률이 높아지는 경향은 나타나지 않는 것으로 보인다.

〈표 13-1〉 인구의 재구성(남성)

(단위: 명)

연 령	1985	1990	1995	2000	2005	2010
0~4	1,861,488	1,584,389	1,757,958	1,587,997	1,215,600	1,125,146
5~9	1,929,789	1,850,806	1,578,300	1,750,496	1,567,859	1,246,572
10~14	2,132,872	1,923,054	1,846,246	1,571,829	1,721,896	1,549,641
15~19	2,127,885	2,122,490	1,915,331	1,837,426	1,541,533	1,700,146
20~24	2,148,266	2,110,039	2,107,729	1,898,323	1,803,142	1,518,729
25~29	2,096,658	2,127,137	2,092,889	2,091,324	1,863,405	1,778,200
30~34	1,597,711	2,071,602	2,105,636	2,075,494	2,072,261	1,854,245
35~39	1,314,126	1,571,485	2,043,056	2,082,036	2,054,397	2,057,829
40~44	1,137,369	1,282,808	1,539,685	2,008,002	2,049,929	2,033,891
45~49	1,064,057	1,091,228	1,245,446	1,501,310	1,964,880	2,019,966
50~54	819,818	1,004,586	1,039,291	1,201,946	1,459,557	1,923,524
55~59	569,216	760,129	938,939	980,945	1,156,051	1,416,198
60~64	457,992	507,551	694,368	864,969	917,018	1,105,602
65~69	316,543	384,363	440,042	620,421	782,563	844,949
70~74	202,267	241,731	306,673	366,207	537,115	688,763
75~79	109,421	134,889	170,924	227,302	287,759	446,604
80~84	39,708	58,786	80,135	107,499	153,680	209,789

〈표 13-2〉 **인구의 재구성(여자)**

(단위: 명)

연 령	1985	1990	1995	2000	2005	2010
0~4	1,775,318	1,466,717	1,591,994	1,487,320	1,166,239	1,097,769
5~9	1,845,251	1,766,984	1,462,328	1,586,468	1,469,850	1,152,395
10~14	2,070,911	1,840,333	1,764,079	1,457,378	1,563,058	1,454,466
15~19	2,058,452	2,065,064	1,836,408	1,756,791	1,430,094	1,542,040
20~24	2,084,485	2,050,206	2,059,231	1,824,477	1,706,076	1,387,508
25~29	2,012,759	2,074,957	2,043,170	2,048,530	1,798,551	1,698,937
30~34	1,505,803	2,002,417	2,066,604	2,033,429	2,025,018	1,786,068
35~39	1,228,582	1,495,671	1,992,198	2,054,396	2,009,348	2,009,423
40~44	1,090,939	1,216,941	1,485,338	1,978,146	2,028,918	1,994,959
45~49	1,055,479	1,073,894	1,204,216	1,472,520	1,957,659	2,018,040
50~54	900,223	1,033,062	1,054,995	1,189,400	1,457,857	1,946,180
55~59	710,192	873,117	1,007,409	1,033,222	1,172,415	1,444,357
60~64	570,360	677,136	840,711	976,880	1,007,005	1,154,358
65~69	417,604	527,790	634,463	798,872	937,587	976,448
70~74	315,725	364,385	469,812	576,654	742,832	887,149
75~79	201,995	249,947	294,743	392,773	498,828	669,943
80~84	97,619	136,130	171,838	210,042	297,793	405,029

APC의 가장 큰 특징은 연령효과(α_j), 시작 코호트의 값(γ_i), 그리고 연도효과(τ_t)를 이용하여 인구를 재구성할 수 있다는 것이다. 2000년 이전에는 이민자에 대한 자료가 없어, 이민자가 없는 닫힌 인구로 가정하여 재구성한 인구는 〈표 13-1〉 및 〈표 13-2〉와 같다. 〈표 13-1〉 및 〈표 13-2〉를 기초로 하여 연령별 센서스 자료의 연도별 연령별 누락률(즉, (센서스-재구성인구)/센서스)를 구한 결과는 〈표 14-1〉과 〈표 14-2〉와 같다.

남성의 누락률을 살펴보면 50세를 기준으로 50세보다 젊은 경우에는 대부분 중복(과다집계)이 발생하고 있고, 50세 이상의 나이에서는 누락(과소집계)이 발생하고 있다. 시간이 지나면서 10~19세의 중복률은 낮아지다가 다시 증가하는 추세를 보이고 있으며,

80~84세의 누락률은 심각할 정도로 크게 나타나고 있다. 2010년 남성의 경우 55~64세, 75~84세의 누락률이 다른 해보다 높게 나타나고 있다. 한편, 여성의 누락률을 살펴보면 45세를 기준으로 45세 이하는 중복이, 그 이상은 누락이 많은 것으로 나타났다. 특히 2010년 센서스의 경우 65세 이상이 모두 중복이 일어나고 있으며, 이는 다른 센서스에서는 나타나지 않는 현상이다. 이와 같은 2010년 누락률의 이상 현상이 2010년에 처음으로 실시된 인터넷 센서스의 영향때문이 아닌지 면밀한 검토가 필요하다.

〈표 14-1〉 연도별 연령별 센서스 누락률(남성)

(단위: %)

연 령	1985	1990	1995	2000	2005	2010
0~4	-3.187	-8.250	-3.480	-3.240	-1.754	-1.495
5~9	-4.718	-7.413	-2.989	-4.420	-5.221	0.264
10~14	-7.691	-6.398	-3.530	-2.674	-5.199	-6.364
15~19	-4.464	-6.380	-3.609	-3.995	-5.217	-6.901
20~24	-1.714	-8.031	-5.818	-6.404	-5.885	-6.561
25~29	3.427	-1.563	0.696	1.653	0.273	-1.365
30~34	0.510	-3.324	-1.897	0.353	0.599	-0.651
35~39	-0.773	-4.655	-2.851	-1.674	-0.546	-0.117
40~44	2.587	-2.462	-2.542	-1.055	-1.561	-1.812
45~49	2.020	-0.885	-1.273	0.348	0.154	-1.207
50~54	1.260	1.013	1.011	1.410	2.310	1.883
55~59	1.540	-0.114	1.658	2.216	2.578	4.075
60~64	3.998	2.568	3.065	3.408	2.188	4.595
65~69	3.206	2.292	4.555	4.453	3.521	1.405
70~74	6.148	3.610	4.418	5.164	4.448	2.358
75~79	5.707	5.461	6.496	7.549	6.328	8.735
80~84	9.802	7.155	12.444	14.196	12.417	12.785

〈표 14-2〉 연도별 연령별 센서스 누락률(여성)

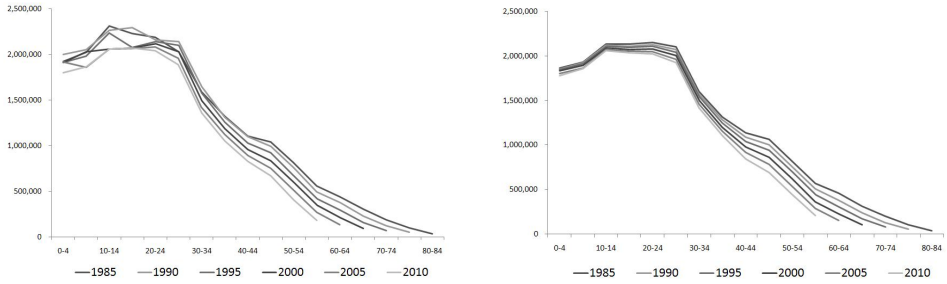
(단위: %)

연 령	1985	1990	1995	2000	2005	2010
0~4	-0.252	-5.551	-0.876	-0.119	1.851	1.941
5~9	-2.419	-5.180	-0.467	-1.621	-2.958	0.089
10~14	-4.364	-5.011	-1.896	0.548	-3.430	-4.202
15~19	-1.460	-5.353	-2.134	-1.176	-2.988	-4.354
20~24	1.220	-2.465	-0.349	0.247	-2.299	-2.975
25~29	-1.491	-4.494	-0.793	0.435	-0.825	-2.143
30~34	-1.299	-3.025	-0.829	0.415	-0.557	-2.345
35~39	-2.246	-3.692	-1.903	-0.728	-1.845	-1.446
40~44	1.123	-0.584	-0.397	0.571	-0.573	-3.157
45~49	0.885	-0.189	0.119	1.140	0.960	-0.526
50~54	1.647	1.729	1.944	2.093	2.041	1.886
55~59	0.426	1.306	1.775	2.422	1.822	2.732
60~64	0.683	2.253	2.356	2.572	1.567	2.591
65~69	0.360	0.615	1.823	2.138	1.457	-0.253
70~74	1.617	0.712	0.206	1.186	0.588	-0.669
75~79	-3.155	0.273	-0.146	0.905	0.522	-0.549
80~84	-3.821	-3.076	-1.764	0.200	0.758	-1.092

4. 인구 특성의 이상 현상은 해결되었는가?

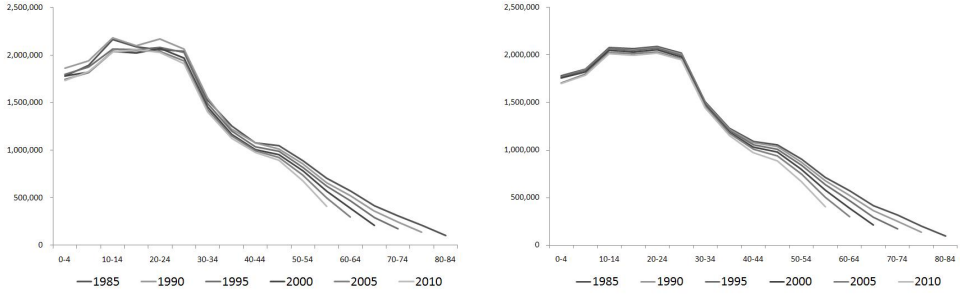
여성의 경우 44세까지 음의 순이민자수 때문에 코호트별 인구가 단조 감소해야 함에도 그렇지 않은 현상이 발생하는 것을 서론에서 지적하였다. 이러한 현상은 재구축된 여성인구인 〈표 13-2〉에서는 사라졌다.

(1) 센서스 (2) 재구성



〈그림 2-1〉 연도별 연령별 인구(남성)

(1) 센서스 (2) 재구성



〈그림 2-2〉 연도별 연령별 인구(여성)

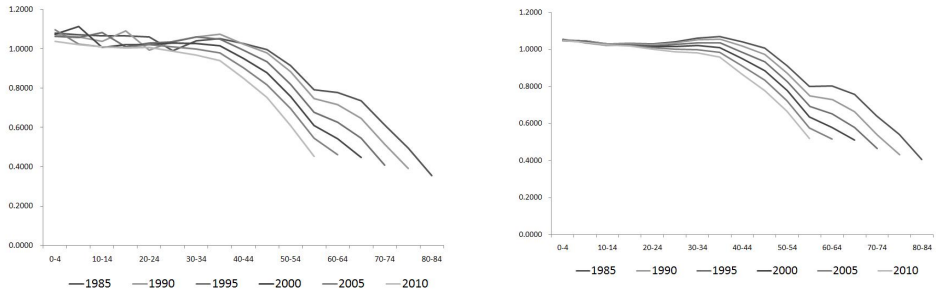
〈그림 2-1〉 및 〈그림 2-2〉는 연도별로 성별 연령별 인구를 그린 것으로 종축은 1985년에서의 각 연령을 나타내며, 횡축은 1985년에서의 코호트들의 인구가 시간이 지남에 따라 어떻게 변하는지를 나타낸다. 이에 따르면 센서스 자료의 경우 남녀 모두 45세 미만에서 역전현상(corss-over)가 발생하지만 재구성한 인구의 경우 이와 같은 역전현상이 발생하지 않음을 알 수 있다.

〈표 15〉 **재구성한 인구로부터의 성비**

연 령	재구성한 인구 성비					
	1985	1990	1995	2000	2005	2010
0~4	1.049	1.080	1.104	1.068	1.042	1.025
5~9	1.046	1.047	1.079	1.103	1.067	1.082
10~14	1.030	1.045	1.047	1.079	1.102	1.065
15~19	1.034	1.028	1.043	1.046	1.078	1.103
20~24	1.031	1.029	1.024	1.040	1.057	1.095
25~29	1.042	1.025	1.024	1.021	1.036	1.047
30~34	1.061	1.035	1.019	1.021	1.023	1.038
35~39	1.070	1.051	1.026	1.013	1.022	1.024

두 번째 의문사항은 비정상적인 성비의 증가현상이다. 〈표 15〉는 〈표 13-1〉 및 〈표 13-2〉의 재구성된 인구를 기초로 다시 계산한 성비이다(지면상 39세까지의 성비만 제공). 서론에서 지적한 1995년 0~19세까지의 4개 코호트의 성비 증가현상이 〈표 15〉에서는 사라졌으며, 2005년 및 2010년을 제외한 모든 연도에서 15~19세의 코호트 성비가 20~24세에서 증가하는 현상 역시 없어진 것으로 나타났다. 〈그림 3〉은 이를 그림으로 그린 것이다.

(1) 센서스 (2) 재구성 인구



〈그림 3〉 **연도별 연령별 성비**

센서스 자료에서 이와 같은 성비의 이상 현상이 일어난 이유는 <표 14>에서 알 수 있듯이, 2000년 0~19세의 남성의 중복률이 2000년 0~19세의 여성의 중복률보다 다른 연도에 비해 상대적으로 매우 높으며, 동일한 현상이 20~24세에서 역시 발생하기 때문으로 판단된다. 그러나 2005년 및 2010년의 경우 재구성된 인구를 기초로 계산한 성비 역시 15~19세의 코호트 성비가 20~24세에서 증가하였는데, 이는 <표 1>에서 알 수 있듯이 여자의 순이민자수가 남자의 순이민자수보다 훨씬 작기 때문인 것으로 보인다.

2절에서 논의한 연령비에 대한 이상 현상의 원인도 <표 14>의 누락률에 의해 설명이 가능하다. 즉, 1985년 남성 5~9세의 코호트 연령비가 2000년에 대폭 증가하는 현상은 2000년 20~24세의 중복률이 가장 높게 나타났기 때문인 것으로 보인다. 1985년 남성 15~19세 코호트의 연령비가 1995년에 대폭 감소하는 현상은 1995년 25~29세에서만 누락이 발생하고, 그 이외의 0~49세 연령대는 모두 중복이 발생하는 것으로 설명이 가능하다. 1985년 여성 20~24세 코호트가 1990년에 대폭 증가하는 현상은 25~30세의 누락률이 다른 연도에 비해 1990년에 매우 높게 나타났기 때문으로 해석할 수 있다. 이 외에 75~79세 남성의 연령비가 코호트별로 갑자기 줄어드는 현상 역시 75~79세 누락률의 대폭적인 증가로 인한 것으로 해석 가능하며, 2005년 여성 60~64세와 65~69세 코호트의 연령비가 2010년에 증가하는 현상은 다른 연도와 달리 음의 누락률(즉, 중복)이 발생했기 때문으로 해석할 수 있다.

끝으로 0~4세의 사망자 등록률의 완성도를 MSEG와 GGB-MSEG로 추정된 K_3를 정리하면 <표 16>과 같다.

<표 16> 0~4세의 사망자 등록률의 완성도

성별	방법	연령	1985 ~1990	1990 ~1995	1995 ~ 2000	2000 ~2005	2005 ~2010
남자	MSEG		1.297	0.773	0.802	0.791	0.838
	GGB-MSEG	30~65	0.912	0.606	1.004	0.949	0.832
	GGB-MSEG	40~70	1.015	0.753	0.924	0.925	0.873
여자	MSEG		1.081	0.772	0.821	0.917	0.971
	GGB-MSEG	30~65	0.958	0.741	1.042	1.009	0.890
	GGB-MSEG	40~70	0.928	0.940	0.889	0.922	0.986

Murray et al.(2010)이 추정한 우리나라 사망자 등록률의 완성도인 K_3가 0.5 내외라는 연구결과는 <표 16>을 보면 타당한 결과라고 보기에는 어려워 보인다. 그러나 1990~1995년의 K_3의 값이 특히 작게 나오는 현상은 문제가 있는 것으로 보이며, 이로 인해 <표 14>에서 1990년 남녀 모두 0~4세의 중복률이 높게 나오는 것으로 판단된다.

Ⅴ. 결론

본 논문에서는 현재의 인구는 과거의 인구에서 그동안의 사망자수와 순이민자수를 차감한 값이라는 단순한 인구방정식을 기반으로 하는 인구동태학적 분석방법을 소개하고, 한국의 현실에 맞도록 변형시킨 DA 방법을 제안하여 한국의 인구자료를 분석하였다. GGB 방법에 적합한 연령구간은 30~65세 또는 40~70세 연령구간으로 나타났고, Bennett & Horiuchi(1981, 1984)가 제안한 SEG 방법은 근사식의 문제점이 있는 것으로 나타났으며, MSEG가 좀 타당한 것으로 나타났다. GGB에 의한 센서스 완성도는 비교적 만족할 만한 수준으로 나타났으며, 사망자 등록률의 완성도 K_3는 MSEG와 GGB-MSEG를 이용한 결과 이 역시 우수한 완성도를 보였다. 그러나 이들 방법론은 정상성 조건에 의해 인구증가율이 시간에 의존하지 않는 폐쇄인구를 가정하였으며, 완성도가 연령에 의존하지 않는다는 다소간 비현실적인 가정을 했다는 한계가 있다. Preston et al.(1998)은 APC 모형을 응용하여 제시한 GGB-APC 모형을 통해 구한 연령별 완성도와 이를 통해 구한 연도별 연령별 누락률은 비정상적인 코호트별 인구증가 현상과 성비, 그리고 연령비의 이상 현상을 설명하는 데 매우 유용한 것으로 나타났다. GGB-APC 모형은 동태자료가 정확해야 한다는 전제조건을 충족해야 하므로, GGB-APC 모형에 의해 추정된 누락률은 센서스의 완성도에 의해 발생할 수 있으나 한편으로는 사망자 등록률의 완성도 또는 인구이동의 정확성 때문에 발생할 가능성도 있다.

이러한 관점에서 중복률이 높게 나타나는 0~9세, 누락률(특히, 남성)이 매우 높은 75세 이상에 대해 센서스의 누락률 조사를 위한 센서스 후 조사(PES)시 면밀한 조사설계가 진행되어야 하며, 동시에 해당 연령대의 동태자료의 정확성을 재조사할 필요가 있다. 이를 토대로 과거의 센서스 자료의 수정 및 재구축을 해야 할 것이다. 2010년의 센서스 자료는 GGB-APC 모형과 연령비에 의해 이상 현상이 드러나고 있다. 이러한 이상 현

상이 2010년 처음으로 실시된 인터넷 센서스 때문에 발생한 것이 아닌지 면밀한 점검과 검증이 필요하다 하겠다. 끝으로, 현행의 센서스가 중단되고 등록 센서스가 실시될 경우 등록 센서스의 완성도와 질적인 우수성을 평가할 수 있는 유일한 방법은 인구통계학적 방법이며 이에 대한 활용이 적극적으로 시행되어야 할 것이다.

참고문헌

Brass, W. 1975. *Methods for Estimating Fertility and Mortality from Limited and Defective Data*. Chapel Hill: International Program of Laboratories for Population Statistics.

Bennett, G. and S. Horiuchi. 1981. "Estimating the Completeness of Death Registration in a Closed Population." *Population Index* 47(2): 207-221.

Bennett, G. and S. Horiuchi. 1984. "Mortality Estimation from Registered Deaths in Less Developed Countries." *Demography* 21(2): 217-233.

Hill, K. 1987. "Estimating Census and Death Registration Completeness." *Asian and Pacific Population Forum* 1(3): 8-24.

Hill, K., D. You, and Y. Choi. 2009. "Death Distribution Methods for Estimating Adult Mortality: Sensitivity Analysis with Simulated Data Error." *Demography Research* 21(9): 235-254.

Martin, L. 1980. "A Modification for Use in Destabilized Population of Brass's Technique for Estimating Completeness of Death Registration." *Population Studies* 34(2): 381-396.

Murray, C.J.L., J.K. Rajaratnam, J. Marcus, T. Laakso, and A.D. Lopez. 2010. "What Can We Conclude from Death Registration? Improved Methods for Evaluating Completeness." *PLoS Medicine* 7(4): 1-17.

Poston, D.L. 2006. "Age and Sex." In D.L Poston and M. Micklin (eds.), *Handbook of Population*. New York: Springer Science+Business Media.

Preston, S.H., I.T. Elo, A. Foster, and H. Fu. 1998. "Reconstructing the size of the African American Population by Age and Sex, 1930-1990." *Demography* 35(1): 1-21.

Robinson, J.G., B. Ahmed, P.D. Gupta, and K.A. Woodrow. 1993. "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis." *Journal of the American Statistical Association* 88(423): 1061-1071.

Shryock, H.S. and J.S. Siegel. 1976. *The Methods and Materials of Demography*. New York: Academic Press.

Smith, D.P. 1992. *Formal Demography*. New York: Plenum Press.

United Nations. 1983. *Mannual X: Indirect Techniques for Demographic Estimation. ST/ESA/SER.A/81. Population Studies 81. Department of International Economic and Social Affairs*. New York: United Nations.

U.S. Census Bureau. 2004. *Accuracy and Coverage Evaluation of Census 2000: Design and Methodogy*. Washington: U.S. Census Bureau.

Vincent, P. 1951. "La Mortalitė des Vieillards." *Population* 6: 182-204.

WHO. 2010. *Post Enumeration Surveys -Operational Guideline-*. New York: United Nations.

<접수 2011/11/05, 수정 2012/2/8, 게재확정 2012/2/10>