



제1장

2010 인구주택총조사 표본조사 결과 가중치 적용 방안 I

- 가구원 가중치를 대상으로 -

권순필·원정연

제1절 서론

1. 연구 목적

2010년은 5년 만에 돌아오는 인구주택총조사(이하 ‘인총’)의 해이다. 인총은 우리나라의 인구 규모, 분포 및 구조와 주택에 관한 특성을 파악하여 각종 정책 입안의 기초 자료를 제공하고, 경제활동인구조사·가계동향조사 등 가구 관련 경상조사를 위한 표본틀(Sampling Frame)의 기초 자료로 활용하는 것을 목적으로 한다.

인총은 조사기준 시점에 대한민국에 거주하는 모든 내국인과 외국인을 상주지에서 조사한다. 일반적이고 응답 부담이 적은 항목들은 조사인구 모두를, 상대적으로 심층적인 추가 항목들은 약 10%의 표본 가구 거주 인구를 대상으로 조사를 실시한다. 따라서, 이러한 심층 항목에 대한 결과를 공표할 때에는 모집단을 잘 대표할 수 있는 추정 결과를 제공해야 한다. 이를 위해 가장 중요한 일은 표본의 가중치를 산정하는 일이다. 가중치 산정은 표본 추출을 통해서 얻어진 표본에 적절한 가중치를 부여해 추정량의 효율성과 신뢰성을 높이려는 것으로 표본조사에서 필수적인 요소이다.

1925년 근대적 의미의 인구총조사가 처음으로 실시되고, 1960년에 주택에 관한 조사가 함께 시행되는 등 인총은 5년마다 16차례 실시되어 왔다. 그 동안, 조사 항목은 끊임없이 변화해 왔으며(통계청, 2007), 가중치 산정 방식도 조금씩 바뀌어 왔다.

이 연구는 이러한 변화에 맞춰 2010년 인총 표본에 적절한 가중치를 부여하는 것을 목적으로 한다. 이상적 가중치는 모집단의 모든 특성을 가장 잘 대표할 수 있어야 하겠

지만, 이러한 가중치를 작성하는 것은 현실적으로 불가능하다. 따라서 본 연구는 2005년 인총 가중치 산정 결과를 토대로 기존 가중치 작성에 이용했던 것보다 많은 특성을 대표할 수 있는 가중치를 산정하는 것에 초점을 맞춘다.

가중치 산정 못지 않게 고려되어야 할 것은 가중치 산정 과정에 대한 상세 기록의 유지와 자동화다. 오랜 기간의 인총 표본조사에도 불구하고, 가중치 산정 방법에 대한 상세 매뉴얼이 미흡하여, 업무 담당자가 바뀔 경우 이전과 동일한 가중치를 산출해 내는데 어려움이 있어 왔다. 또한, 가중치 산정의 비자동화는 업무의 효율성을 떨어뜨리고 가중치 산정 과정에 업무 담당자의 주관성을 강하게 반영할 수 있다. 따라서, 가중치 산정에 대한 기록을 유지하고 최대한의 자동화를 실현한다면 업무의 효율성과 객관성을 향상시키고, 향후 가중치 작성 방안 연구의 토대로 활용할 수 있을 것이다.

2. 연구 내용과 범위

2010년 인총의 표본조사 가중치 적용 방안을 검토하기 위해 2005년 인총 표본조사 결과와 캐나다 센서스의 표본조사 가중치 작성 사례를 검토한다.

2005년 인총 표본조사의 가구원 가중치 작성 과정을 살펴본다. 작성된 가중치의 분포를 분석하고, 전수와 표본 추정 결과가 불일치하는 주요 항목에 대하여 불일치를 보완할 방법을 강구한다. 2005년 인총 결과와의 시계열 유지를 위해 2005년 가중치 적용에 사용된 주요 변수들을 최대한 활용하고, 과대 가중치에 대해서는 그 범위를 조정한다.

가중치 작성은 레이킹비(Raking Ratio) 혹은 반복비례적합(Iterative Proportional Fitting; IPF)이라고 하는 반복비례가중법(이하 “레이킹”)과 사후층화(Post-stratification, 혹은 Cell Weighting; 칸 가중법)를 통해 검토한다. 추정 과정에 이용할 주요 변수는 각 변수가 전수의 성별, 연령그룹별 인구 분포와 표본의 성별, 연령그룹별, 통근·통학, 경제활동상태에 따른 인구 분포에 영향을 미치는지 여부를 검토하여 선정한다.

또한, 비표집 특성 항목과 과소 표본에 따른 과대 가중치 처리를 검토한다. 레이킹과 사후층화 가중법의 가장 큰 문제는 표본이 없는 빈 셀이 발생할 경우 가중치를 계산할 수 없다는 것이다. 또한 표본 설계 당시 고려되지 않은 특성 항목은 너무 적은 표본 때문에 가중치가 크게 부여될 수 있다. 이러한 문제는 셀 병합을 통해 해결하는 것이 가장 일반적이다. 따라서, 적은 노력을 들여서 합리적으로 셀을 병합하는 방안을 찾아본다.

본 연구의 분석을 위해 사용한 자료는 2005년 인총 전수조사와 표본조사 자료 중 조사구특성이 아파트조사구(A), 보통조사구(1), 섬조사구(2)인 자료¹⁾이다. 그 외의 조사구(기숙시설조사구, 특수사회시설조사구, 관광호텔 및 외국인조사구)는 전수조사이므로 가

1) 아파트·보통·섬조사구에서 “외국인조사표”로 조사된 외국인도 제외



중치 산정이 필요하지 않아 분석에서 제외한다.

새로운 가중치 평가는 주요 특성들에 대한 추정인구와 모집단인구 사이의 불일치 (discrepancy) 정도를 이용한다. 불일치 정도가 가장 작은, 즉, 모집단 분포와 가장 유사한 추정 결과를 제공하는 가중치 산정 방법을 2010년 인총에서 이용될 가중치 적용 방안으로 한다.

3. 연구 기대 효과와 제한

전수조사 결과와 표본조사 추정값을 모든 특성에 대해 정확하게 일치시킬 수는 없다. 따라서, 본 연구는 최대한 많은 특성을 대표할 수 있는 가중치를 산정하고, 누가 가중치를 계산해도 같은 값이 나오는 재현성을 확보하는 데 중점을 둔다. 이를 위해 최대한 산정 과정을 자동화 가능한 로직으로 구현한다. 가중치 산정의 자동화를 통해 업무를 효율적으로 처리하고 객관적인 가중치를 생성할 것으로 기대된다.

본 연구 결과를 이용함에 있어 유의할 점은 연구를 위해 사용된 자료가 조사구특성 중 아파트조사구, 보통조사구, 섬조사구 자료이기 때문에, KOSIS 등에 공표된 결과와 일치하지 않는다는 것이다.

제2절 인총 표본조사 가중치 사례

1. 2005년 인총

가. 표본설계

2005년 인총은 전수조사표(short form)와 표본조사표(long form)를 사용하여 조사하였다. 표본조사표는 보다 심층적인 자료를 얻기 위해 전수조사 항목(21개)에 통근·통학, 인구이동, 경제활동, 임차료 등 23개 항목을 추가하였다. 인총은 원칙적으로 전국의 모든 가구를 대상으로 인구, 가구 및 주택에 관련한 내용을 조사하며 전체 가구의 10%를 표본으로 선정하여 표본조사표로 조사한다.

2005년 인총은 평균 60가구로 구성된 조사구(enumeration district)를 조사단위로 하였다. 조사구는 조사의 중복과 누락을 방지하기 위해 모든 지역을 대상으로 행정 읍·면·동 경계 내에서 도로, 산, 강, 하천 등의 변화가 적은 지형지물을 이용하여 일정한 가구수가 포함되도록 분할하였다. 조사구는 출입의 통제여부에 따라 보통조사구와 특별조사구로 나누었다. 보통조사구는 아파트조사구(A), 보통조사구(1), 섬조사구(2), 기숙시설조사구(3), 특수사회시설조사구(4), 관광호텔 및 외국인조사구(5)로, 특별조사구는 군대, 교도

등으로 이루어졌다.

10% 표본을 위한 조사모집단 자료는 관광호텔 및 외국인조사구를 제외한 보통조사구만을 대상으로 하였다. 보통조사구 중 관광호텔 및 외국인조사구와 특별조사구는 그 특성상 표본조사로 자료를 수집하고자 하는 특성과 관계가 없으므로 제외하였다. 표본조사구 추출을 위해 우선, 모집단의 조사구자료를 행정구역에 따라 읍·면·동별로 1차 층화한 후 각 읍·면·동 내에서 조사구의 특성에 따라 아파트조사구(제1층), 보통조사구·섬조사구(제2층), 기숙시설조사구·특수사회시설조사구(제3층)로 2차 층화하였다. 2차 층화 후 층별로 조사구 번호순으로 자료를 나열한 다음 제1층과 제2층에서 조사구를 계통추출법에 따라 추출하였다. 이때, 인구 및 가구수가 적은 일부 지역에 대하여 조사구 규모에 따라 추출 비율을 차등 적용하여 대표성을 높이도록 하였다(<표 1-1>).

마지막으로, 제3층인 기숙시설조사구와 특수사회시설 조사구는 원칙적으로 특정 목적을 위한 집단조사구이며 모집단조사구의 약 2.7%에 불과하여 제3층에서 표본조사구를 추출하는 것은 조사 및 자료처리 등 모든 면에서 오히려 비효율적이므로 제3층의 조사구는 모두 표본조사구로 하였다.

따라서 전국 265,317개의 아파트조사구, 보통조사구 및 섬조사구 중에서 26,505개의 표본조사구가 추출되었다.

<표 1-1> 읍면동별 표본추출비율

전체 조사구수	읍면동수	추출 조사구수
	1개	
2~8개	42개	2개
9~49개	1,560개	4개
50~199개	1,837개	10% 내외
200개 이상	126개	8% 내외, 최소 18개 이상

나. 가중치 작성 및 적용 방법

모수(parameter)는 표본조사를 통해 얻어진 자료를 통해 추정하게 된다. 2005년 인총에서는 각 행정구역별(시군구의 동부, 읍부, 면부 포함)로 성별, 연령별 인구수, 거처의 종류(단독주택, 아파트, 연립주택, 다세대주택, 비거주용건물내주택, 주택이외의거처)별 가구수가 전수 집계 결과와 일치하도록 가중치(승수)를 작성하여 추정치를 구하였다.

즉, 아파트조사구, 보통조사구 및 섬조사구의 전수자료와 표본자료를 전국 442개 행정구역별(시군구의 동부, 읍부, 면부를 각 1개로 산정), 성별에 따른 5세 계급으로 구분된 21개의 연령항목별 가구원수로 집계하고, 전국 442개 행정구역별, 가구원수(1인, 2인,

3인, 4인, 5인, 6인 이상)별에 따른 거처종류 항목별 가구수를 집계하여 추정치가 전수 집계 결과와 일치하도록 가중치를 산출하였다(<표 1-2>, <표 1-3>, <부표 1> ~ <부표 3>). 전국 442개 행정구역별로 인구부문은 각 40개, 가구부문은 각 36개의 가중치가 만들어졌다.

<표 1-2> 인구부문 자료 집계표

인구부문			0-4세	5세	6-9세	10-11세	12-14세	15-19세	...	80-84세	85세 이상	미상
서울	종로구	남										
		여										
	중구	남										
		여										
∫												
제주	남제주군	읍부	남									
			여									
		면부	남									
			여									

<표 1-3> 가구부문 자료 집계표

가구부문			단독	아파트	연립	다세대	영업용	주택이외
서울	종로구	1인 가구						
		2인						
		3인						
		4인						
		5인						
		6인 이상						
	중구							
∫								
제주	남제주군	읍부						
		면부	1인 가구					
			2인					
			3인					
			4인					
			5인					
6인 이상								

이때 가중치는 정수로 산출하며, 아파트조사구, 보통조사구 및 섬조사구의 각 레코드(인구, 가구)별로 적용하였다.

여기에 전체 표본조사된 기숙사시설조사구, 사회특수시설조사구 및 외국인·관광호텔조사구를 합산하여 최종 추정치를 작성하였다. 추정치는 시군구 행정구역단위로 산정하여 시도별 및 전국으로 합산하였다(통계청, 2007).

다. 가중치 분포

2005년에 산정된 가중치는 정수로 제공되었으며 각 시도별 가중치 분포를 보면 <표 1-4>와 같다. 평균적으로 10에 가까운 가중치가 부여되었으며, 가장 작은 가중치는 2, 가장 큰 가중치는 40의 값을 갖는다.

<표 1-4> 2005년 인총 시도별 가구원 가중치 분포

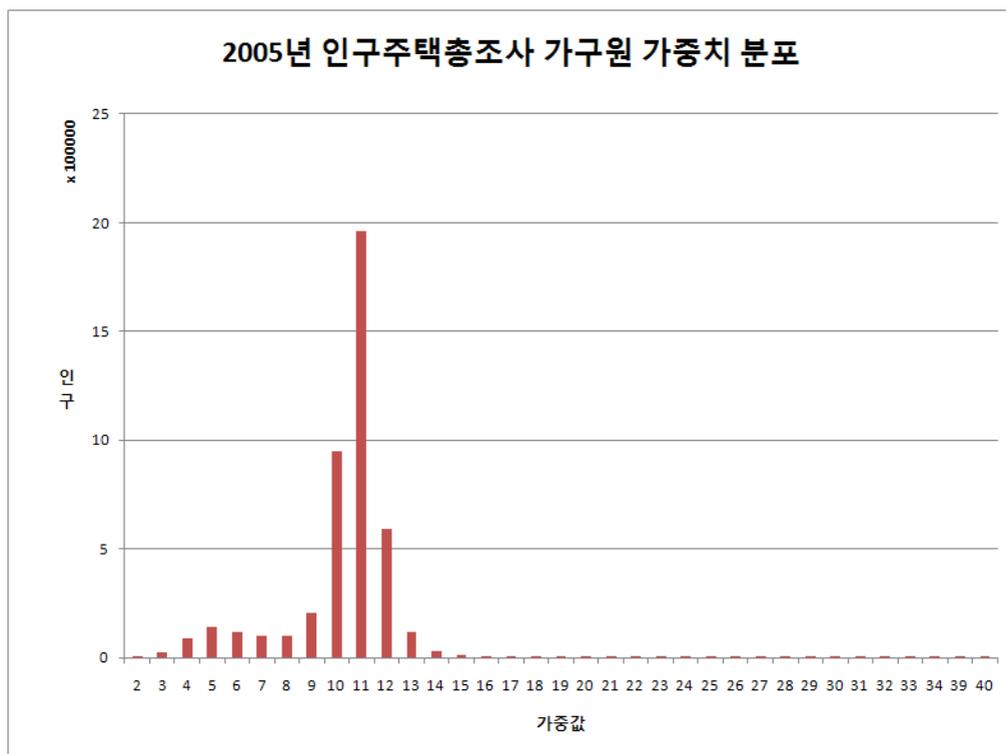
시도	전수(명)	표본(명)	가중치			
			평균값	표준편차	최소값	최대값
서울	9,555,567	880,181	10.86	0.66	8	19
부산	3,430,829	320,285	10.71	0.86	5	19
대구	2,403,106	221,920	10.83	0.81	5	17
인천	2,472,773	235,783	10.49	1.39	2	19
광주	1,372,490	129,923	10.56	0.90	7	15
대전	1,399,806	128,990	10.85	0.71	8	15
울산	1,018,275	91,866	11.08	0.97	6	20
경기	10,109,197	933,104	10.83	1.16	3	24
강원	1,407,144	152,837	9.21	2.75	2	31
충북	1,403,479	148,481	9.45	2.90	2	26
충남	1,798,397	189,505	9.49	2.77	2	28
전북	1,720,504	197,360	8.72	3.03	2	28
전남	1,759,483	209,214	8.41	2.99	2	34
경북	2,486,076	274,067	9.07	2.91	2	31
경남	2,956,823	298,848	9.89	2.95	2	40
제주	518,528	50,014	10.37	1.47	5	22
전국	45,812,477	4,462,378	10.27	2.02	2	40

경기와 제주를 제외한 7개 도(道)의 가중치는 평균적으로는 9이지만, 표준편차가 2 이상이다. 가중치가 평균적으로 10보다 작다는 것은 인구와 가구수가 적은 일부 지역에 대하여 과소 표집을 방지하기 위하여 추출률을 높여준 때문이다. 그럼에도 불구하고 표준편차가 2 이상, 즉 상대적으로 평균에 비해 큰 가중치의 값이 나타나는 이유는 가중치 작성에 사용된 특성 항목들(예를 들어, 연령그룹이 2(5세)거나 21(85세 이상)인 경우 등)



은 여전히 일부 과소 표집된다는 것을 보여준다.

2부터 40까지 주어진 정수 가중치의 히스토그램을 살펴보면 가중치의 분포를 보다 확연히 알 수 있다([그림 1-1]). 전체 가중치들 중에서 약 83%의 가구원이 9, 10, 11, 12의 가중치를 부여받는다. 30을 초과하는 가중치를 부여받은 인구는 아주 작은 수이다.



[그림 1-1] 2005년 인총 가구원 가중치 분포

2. 캐나다 센서스의 표본조사 가중치 작성

캐나다 센서스에서 표본조사 가중치는 1986년에 레이킹이라고 불리는 반복비례가중법(Iterative Proportional Fitting), 1991년과 1996년에는 이단계일반화회귀(two-step Generalized Regression), 2001년과 2006년에는 유사최적회귀(Pseudo-optimal Regression) 방법을 통해 작성되었다.

이하에서는 2006년 캐나다 센서스의 표본조사 연혁과 절차 그리고 표본조사 가중치 작성 절차에 대해 기술한다. 보다 자세한 내용은 캐나다 통계청(2009)의 보고서를 참고하기 바란다.

가. 표본조사

1) 표본조사 연혁

캐나다 센서스에서 표본조사는 1941년에 처음 시행되었다. 주택조사는 각 센서스 부지구(census subdistrict) 내의 10번째 거처(dwelling)마다 시행되었다. 주택조사의 27개 질문에 대한 정보는 같은 거처의 인구조사 가구원·가구부문 자료와 병합되어 표본 특성과 기초 특성에 관한 교차표로 만들어졌다. 1951년에 다시 주택 센서스가 표본조사로 시행되었는데, 24개의 질문을 포함하는 주택조사를 위해 매 다섯 번째 거처(고유번호가 2나 7로 끝나는 것)를 표본으로 선택하였다. 1961년 센서스에서는 개인 가구의 20%에 대해 15세 이상 인구의 국내인구이동, 출산력과 수입을 포함하는 표본조사를 하였다. 그러나 1956년과 1966년에는 센서스 규모가 작아 표본조사를 시행하지 않았다.

1971년 센서스는 조사와 관련하여 몇 개의 주요한 변화가 있었다. 우선, 근본적인 변화는 전통적인 면접원 조사방식을 자기기입식으로 변경했다는 것이다. 방문 조사에서 조사원 효과(effect of the enumerator)가 센서스 값의 분산에 크게 영향을 미친다는 연구 결과가 나왔기 때문이다(Fellegi, 1964; Hansen et al., 1959). 자기기입식의 사용은 이런 조사원 효과를 감소시켜 센서스 분산값도 감소시킬 것으로 기대되었다. 더불어 질문에 대답하기 위한 보다 많은 시간과 사생활 보호를 응답자에게 제공하였다. 두 번째는 항목의 변화이다. 포함하고 있는 주제와 요구된 질문의 수가 이전 센서스보다 많았다. 왜냐하면, 자기기입과 표본조사로 기대되는 자료의 질적 수준에 비해, 비용과 응답자 부담이 줄어들 것이고, 시의성도 고려하였기 때문이다. 표본은 전수의 3분의 1을 표집하였고, 이를 통해 가능한 모든 것을 수집하도록 했다. 표본 가구는 센서스 질문 모두를 포함하는 표본조사표를 받았고 나머지 가구는 성명, 가구주와의 관계, 성, 생년월일, 혼인상태, 모국어, 거주형태, 소유여부, 방의 수, 수도공급, 화장실 등 기본적인 질문을 포함하는 전수조사표를 받았다.

1976년 센서스 항목은 1971년에 비해서 줄었다. 1976년 조사표는 자료를 수집하는데 곤란(예, 수입)하거나 코드화하기에 비용이 많이 드는 질문(예, 직업, 산업, 일하는 장소)을 포함하지 않았다. 1971년과 1976년 센서스에서 사용된 대부분의 방법은 표본이 1/3에서 1/5로 줄었다는 것 외에 1981년 센서스에서 그대로 유지됐다. 그 당시 수행된 연구는 자료의 질(분산)의 저하는 용인 가능한 수준이지만, 비용과 응답 부담 경감, 시의성의 향상은 기대만큼 충분하지는 않다는 것을 보여주었다. 5분의 1 표본조사는 1981년 이후 매 센서스에서 유지되었다.

2) 2006년 센서스에서 사용된 표본조사 계획

센서스 정보의 대부분은 표본조사에서 얻어졌다. 모든 자기기입 조사 지역²⁾에서 개



인 거처의 5분의 1을 표본으로 선택했고, 나머지는 전수조사하였다. 거처의 종류, 나이, 성, 혼인상태, 모국어, 가구주와의 관계 등에 대한 기본적인 질문이 전수조사되었고, 거처³⁾에 대한 부가정보와 사회경제적 질문이 표본조사에서 추가로 조사됐다.

센서스 표본은 모든 층(Collection Unit; CU)에서 항상 5분의 1의 비율로 거처를 표집하는데 표집 방법으로는 층화계통추출을 사용한다. 반면에, 거처 내의 가구원에 대한 설계는 거처를 집락으로 한 층화계통집락추출이다.

나. 표본조사 추정

1) 추정절차 개발

가중치는 표본의 각 단위(가구원, 가족 또는 가구)에 할당된다. 가장 간단한 가중치 부여 방법은 표본이 1/5의 확률로 추출되어졌기 때문에 각 표본 단위에 5의 가중치를 주는 것이다. 이 가중치를 이용한 추정량은 비편향이다. 그리고, 만약 표본자료 외에 다른 정보가 알려져 있지 않다면 최적의 절차이기도 하다. 그러나 표본이 모든 거처의 1/5을 거의 정확하게 포함할 것이라는 것을 알지만, 그것이 모든 가구원의 1/5 혹은 가구 각 형태의 1/5 또는 25에서 34세 사이 여성의 1/5을 정확하게 포함하는지 확신할 수는 없다. 따라서, 각 단위에 5의 가중치를 부여하는 것은 인구의 가장 중요한 하위 그룹(subgroup)에서 일치성을 확보하는 데 어려움이 있다. 즉, 큰 그룹에 대해서는 이들 확률이 1/5에 매우 근접할 것이지만, 보다 작은 하위 그룹에서는 1/5과 크게 다를 수 있다.

다음으로 가장 간단한 절차는 중요한 하위 그룹(예, 지역 내 연령-성 그룹)을 정의하고 각 하위 그룹의 전수조사 단위수(N)와 표본조사수(n)를 센다. 그리고 하위 그룹의 각 표본 단위에 N/n 의 가중치를 할당하는 것이다. 이들 하위 그룹을 사후층(post-strata)이라 한다. 예를 들어 만약 Prince Edward Island에 20-24세 남자 전수 인구가 5,000명이 있고, 이들 중 1,020명이 표본이라면, 그때 (사후)가중치 $4.9(5,000/1,020)$ 가 Prince Edward Island의 20-24세 남성 표본에 할당되어질 것이다. 이런 형태의 추정을 비추정(ratio estimation)이라고 한다. 만약 앞에서 언급한 초기 가중치 5를 그냥 사용한다면, 표본 추정은 $5,100(1,020 \times 5)$ 일 것이다.

인구수와 추정 결과 사이의 완벽한 일치를 위해 5의 간단한 가중치를 조금씩 조정하는 것을 계수조정(calibration)이라 한다. 1991년 이전에는 계수조정을 위해 레이킹 비추정을 사용했다. 가구 수준 추정은 가구 수준 가중치에 의해 이루어지며, 가구원 수준 추정은 가구원 수준 가중치에 의해 이루어진다.

1991년에는 이단계일반화회귀추정량(two-step Generalized REGression estimator; GREG)이

2) 인디안 거주 지역 등 일부 지역은 여전히 면접조사 방식으로 센서스가 수행되었다.
3) 이후로는 가구와 거처를 혼용하여 사용

도입되었다. 그것은 기존의 레이킹 추정보다 인구수와 추정 결과 사이에서 높은 수준의 일치성을 달성했다. 추가적으로, 1인 가구의 가중치는 가구와 가구원 모두의 추정치를 생산하는 데 사용되었다. 이것은 1991년 이전의 일부 추정량에서 보였던 불일치성을 제거했다. GREG 추정량은 1996년에도 사용되었다.

2001년과 2006년에는 유사최적회귀추정량(pseudo-optimal regression estimator)이 사용되었다. 이 추정량은 GREG보다 인구수와 추정 결과 사이에 더 좋은 일치성을 달성했다. 유사최적회귀추정량은 대략 5의 초기 가중치로부터 가능한 적게, 그러나 가능한 많은 특성에 대해 추정값과 전수조사 인구가 완벽하게 일치할 수 있도록 조정되었다. 완벽한 일치하는 가중치 지역(Weighting Area; WA) 수준에서 이루어진다.

2006년 캐나다는 센서스를 위해 전국을 대략 50,000개의 CU로 나누었다. 거처 1/5표본은 센서스 가중치 처리에 이용될 대부분의 CU로부터 선별되었다. DA(Dissemination Areas)는 CU와 크기에 있어서 유사한 또 다른 지리학적 수준으로, 모든 DA는 WA를 만들기 위해 결합된다. 평균적으로 1개의 WA는 8개의 DA와 7개의 표본 CU를 포함한다.

2) 이단계일반화회귀추정량(GREG)

회귀 절차에서는 34개의 보조변수를 사용한다. 보조변수는 5세 범위의 연령, 혼인상태, 성, 가구크기, 거처형태 등을 포함한다. 2006년 센서스 가중치 산정 목표는 다음과 같은 4가지 조건을 만족시키는 것이다.

- (a) 가능한 34개의 보조변수를 활용해 WA수준에서 정확한 인구/추정값 일치성을 갖는 것
- (b) 34개의 보조변수를 활용해 보다 큰 DA에서의 대략적인 인구/추정값 일치성을 갖는 것
- (c) 가능한 많은 DA에서 인구의 총수와 가구의 총수에 대한 정확한 인구/추정값 일치성이 있을 것
- (d) 최종 센서스 가중치는 1에서 25 범위 내에 있을 것. 각 표본은 적어도 자기 자신을 대표해야 하기 때문에 가장 작은 값은 1이어야 함.

가중치는 자동화된 시스템을 통해 WA별로 계산된다. 초기 가중치는 WA에서 각 표본 개별 가구에 할당된다. 그리고 이들 가중치는 그들에게 적용될 2개 또는 3개의 가중치 조정항을 갖는다. 우선, 가구는 가구 크기에 기초되어 WA수준에서 사후총화된다. 왜냐하면, 작거나 큰 가구는 표본에서 과소 포함되기 때문이다. 두 번째 조정은 (b)에서 기술된 것처럼 DA수준에서 개략적인 인구/추정값 일치성을 이루기 위해 가중치에 적용된다. 마지막으로 세 번째 조정이 (a)와 (c)에서 묘사된 것처럼 WA와 DA수준에서 정확한 인구/추정값 일치성을 이루기 위해 적용된다. 초기 CU수준의 가중치가 WA에서 각 개인 가구에 할당된다. 가중치는



CU수준에서의 개인 가구수를 표본 가구수로 나눈 값이다. 대략, 다섯 가구 중 하나가 표본 추출되면 초기 가중치는 5 근처인 경향이 있다. 가중치 조정 단계를 살펴보면 다음과 같다.

- 단계 1) 가구는 WA수준에서 가구 크기에 기초해 사후층화된다. 초기 가중치는 사후층화가중치를 만들기 위해 조정항에 의해 배가된다. 예를 들어 사후층화가중치에 기초한 WA에서의 1인가구의 추정값은 WA인구의 1인가구수와 일치할 것이다. 사후층화가중치는 1에서 20의 범위 내에 있도록 만들어진다. 25보다 20의 상한이 추가조정을 위한 약간의 여지를 주기 위해 사용된다.
- 단계 2) 단계 1)의 회귀가중치 조정항이 DA수준에서 계산된다. 두 번째 조정단계에서 WA수준에 적용될 34개의 보조변수(나이, 성, 혼인상태, 가구크기, 거처형태)를 이용한 가중치 조정항이 계산된다. 이 과정에서 DA수준에서의 인구/추정값 차이는 단계 1)의 가중치를 사용함으로 감소된다.
- 단계 3) 단계 2)의 회귀 가중치 조정항이 WA수준에서 계산된다. 34개의 보조변수는 WA내의 각 DA에 대해 두개의 보조변수(가구와 가구원수)와 함께 WA수준에서 적용된다. 이들은 최종 가중치를 만들기 위해 단계 1) 가중치에 적용된다. 34개 보조변수에 대한 WA수준에서의 인구/추정값 차이는 최종 가중치를 사용해 크게 감소되거나 제거된다.

추정을 위한 몇 가지 제약은 단계1)과 2)에서 해소되어야 한다. 왜냐하면, 제약에 걸리는 단위들은 그 비중이 상당히 작은 소수(SMALL)이고, 회귀추정 과정에서 거의 불필요(linearly dependent; LD 혹은 nearly linearly dependent; NLD)하기 때문이다. 또한 이 단위들은 1에서 25의 범위를 벗어나는 가중치의 이상치를 야기한다.

센서스 가중치는 각 WA에서 독립적으로 계산된다. 이것은 각 WA에서 가중치 를 위한 자료가 서로 다른 형태로 사용되는 것이 가능하게 했다. (예를 들어, 사후층화할지 아닐지, 거처 형태의 제약을 사용할지 아닐지 등). 1996년 이전에는 모든 WA에서 동일한 자료 형태를 사용했지만, 2001년 어떤 지역에서 발견된 큰 차이 때문에 2006년에 도입된 방법이다.

[그림 1-2]는 2006년 센서스에서 초기 가중치와 사후층화 가중치, 사후층화 가중치와 단계 1) 가중치, 단계 1) 가중치와 최종 가중치를 비교한 그림이다. 초기 가중치는 전반적으로 5 근처에 강하게 모여 있지만, 가중치 보정 단계가 지나면서 점진적으로 가중치가 퍼지는 경향을 알 수 있다.

3) 표본 추정값과 전수조사 인구와의 차이

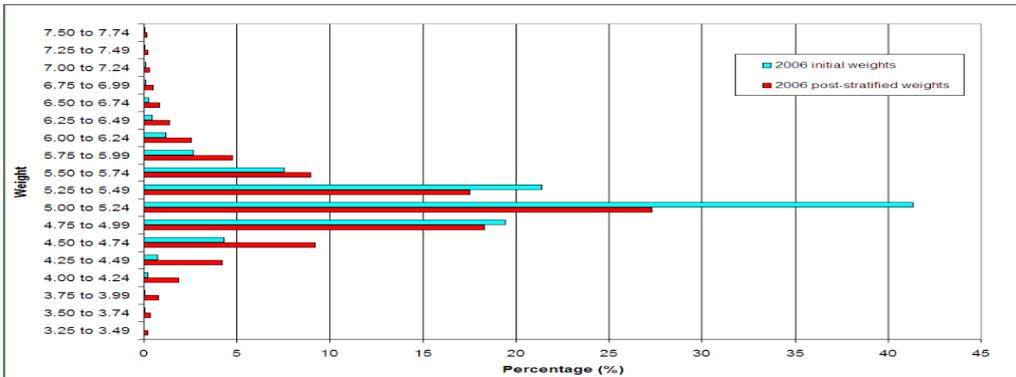
최종 가구 가중치는 모집단과 가능한 많은 특성에 대해 가중치 추정값이 일치하도록

만들어졌다. 전수조사와 표본조사에서 이용할 수 있는 특성은 5세 연령그룹, 성, 혼인상태, 가구크기 및 거처의 종류가 포함된다. 가중치 절차는 1,000에서 3,000개의 단위 거처를 포함하는 WA에서 인구/추정값 차이를 제어하려는 것이다.

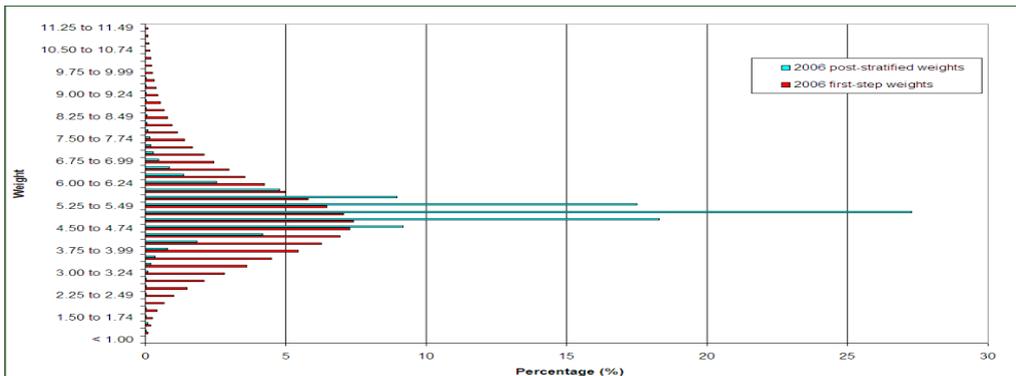
그러나 이런 최대한의 노력에도 불구하고, 특히 작은 지역에서 표본 추정값과 전수 인구가 다를 수 있는 몇 가지 이유가 있다.

- (a) 회귀과정에서 제거된 제약들 : SMALL, LD나 NLD, 혹은 이상치를 야기시킬 수 있는 단위들에 대한 제약은 일반적으로 추정값과 인구수 사이에 약간의 차이를 발생시킨다.
- (b) WA의 하위 지역 : WA는 가중치 시스템이 추정값과 인구수를 일치시키려는 가장 작은 지리적 지역이다. 그래서 WA 내에 포함된 작은 지역(DA 혹은 매우 작은 지방자치단체와 같은)은 인구수와 추정값 사이에 일반적으로 차이를 보인다.

(a) 2006년 센서스에서 초기 가중치와 사후층화 가중치 비교

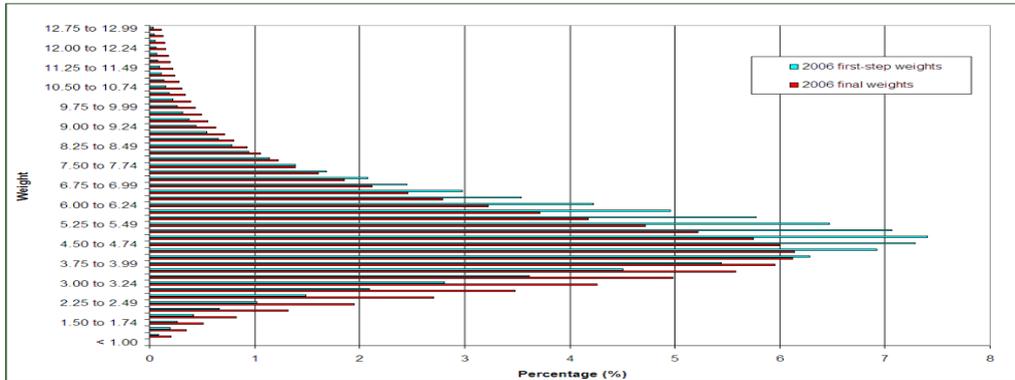


(b) 2006년 센서스에서 사후층화 가중치와 단계 1) 가중치 비교





(c) 2006년 센서스에서 단계 1) 가중치와 최종 가중치 비교



※ 자료 : 캐나다 통계청 2006 센서스

[그림 1-2] 2006년 캐나다 센서스에서 각 단계별 가중치와 최종 가중치 비교

제3절 가중치 적용 방안 검토

인총 표본이 적절한 설계 절차에 따라 뽑혔다면, 가중치는 원칙적으로 각 개체에 10(10% 표본)을 주면 된다. 그러나 표본이 거의 정확하게 1/10의 확률로 뽑혔다고 해도 모든 가구원이나 가구, 특정 연령대의 여성들이 1/10의 확률로 뽑힌 것은 아니다. 단순히 10의 가중치를 부여하는 방법으로는 가장 중요한 모집단의 하위 그룹들에 대해 전수 인구조와 추정 결과 사이의 일관성을 유지할 수 없다. 따라서 최대한 중요 하위 그룹에 대한 일치성을 달성할 수 있는 적절한 가중치 적용 방안이 필요하다.

1. 추정 방법

가. 사후층화

사후층화(post-stratification)는 자료가 설계 단계에서 이용 가능하지 않은 경우, 표본을 선택한 후 신뢰성 있는 최신의 모집단에 대한 층화 정보가 이용 가능해진 경우, 또는 설계 단계에서는 사용되지 않았지만 층화에 적합한 변수로 조사의 가중치를 수정하기 위하여 사용된다. 사후층화는 보조 자료가 총수의 형식(예를 들어, 모집단의 남자와 여자의 수)으로 되어 있을 때 사용된다. 이는 관심 변수의 모집단 평균이 사후 층에 따라 다를 경우에 표본추출 분산을 줄이는 데 매우 효율적이다(캐나다 통계청, 2003). 또한 이 방법은 모집단을 다차원 분할표로 분류했을 때 각 칸의 모집단수를 안다면 모집단수 대

표본수의 비율을 가중치로 부여함으로써, 각 칸에서 가중치의 합을 모집단 수에 일치시켜 표본의 내적 일치성(internal consistency)을 만족시킬 수 있다(허명희 등, 2005). 그러나 이 방법은 다차원 칸에서 모집단수를 모르는 경우 사용이 어렵다. 모집단의 다차원 결합 분포가 알려진 경우에도 표본의 다차원 결합분포의 칸이 비어있는 경우 이용이 어렵다는 단점이 있다(임경은, 2010).

인총은 모집단을 조사구 특성에 따라 층화하여 조사구를 계통추출하고, 표본으로 선정된 조사구 내의 모든 가구와 가구원을 대상으로 표본조사를 시행한다. 즉, 설계 단계에서 가구와 가구원의 정보가 보조 정보로 사용되지 않는다. 그러므로 가구와 가구원에 대한 설계 가중치 계산이 용이하지 않다. 게다가 표본 추출 당시와 실제 조사 과정 사이에 표본 구성에 많은 변화가 발생하기 때문에 인총 추정 시 설계가중치를 이용하는 것이 오히려 오차를 크게 할 수도 있다. 또한 인총과 같이 관심 항목의 수가 많은 경우에는 사후층화처럼 조사 이후의 정보를 이용하는 추정이 보다 편리하고 유리하다. 우리나라뿐만 아니라 많은 국가들이 인총 가중치 계산시 우선적으로 사후층화를 고려하는 이유이기도 하다.

그러나 앞서 언급한 바와 같이 사후층화 가중치는 다차원 결합분포의 칸이 비어있는 경우 계산할 수가 없다. 그러므로 가능한 많은 변수를 추정에 이용하는 데 어려움이 있고(선택 가능한 변수의 수가 제한적), 빈 셀(혹은 칸)을 가중치 계산 전에 미리 처리해주어야 하는 문제가 남는다. 또한 사후층화 가중치를 이용하게 되면, 각각의 사후층 내에서 특성값의 비율은 변하지 않지만, 한 개 이상의 사후층을 포함하고 있는 전체 모집단에서의 특성값의 비율은 변화한다는 점을 유의해야 한다(캐나다 통계청, 2003).

사후층화를 위한 좋은 인구학적 특성을 찾는 것은 때로는 어려운 일이다. 가장 대표적으로 사용되는 인구학적 특성은 성, 연령그룹, 혼인상태, 교육정도 등이며 외국의 경우에는 추가적으로 인종도 주요 특성으로 활용하기도 한다.

나. 레이킹(반복비례가중법)

레이킹이란, 모집단의 인구사회학적 특성의 주변 분포를 알고 있는 경우 적용할 수 있는 가중치 보정 방법으로 각 특성 변수의 주변 분포를 이용하여 반복적으로 셀 내의 모집단수를 구하는 방법이다. 이 방법은 모집단의 주변 분포만 알면 적용이 가능하므로 고려할 특성 변수가 많은 경우에도 활용이 가능하고, 사후가중에서 발생하는 빈 셀 문제에도 보다 적극적으로 대응할 수 있다.

레이킹은 Deming과 Stephan(1940)이 처음 제안한 이래 현실적인 유용성 때문에 실제 문제에 활용되기 시작하였다. 이 방법은 가중값을 반복적으로 수정해 가므로 수정된 가중치의 수렴 여부가 중요한데, 모든 칸의 표본수가 0이 아니라면 수렴한다는 결과가 있다



(Ireland와 Kullback, 1968). 그러나 만일 일부 칸에서 표본수가 관측되지 않으면 수렴하지 않을 수도 있다(Thompson, 1981). 레이킹이 수렴하는 경우 이 방법으로 구한 최종 가중치는 설계 가중치를 보조정보로 이용하여 수정을 한 형태이기 때문에 설계 비편향은 아니고 일치 추정량도 아니다. 그러나 레이킹은 주요 보조변수에 관하여 표본의 추정 결과를 모집단 비율과 일치시킴으로 추정의 관점에서 신뢰성을 확보하는 장점이 있다. 또한 단순임의추출에 의한 표본뿐 아니라 일반적인 추출법에 의한 표본에도 적용이 가능하고, 모집단의 1차원 주변 분포만 알려져 있어도 적용이 가능하다는 장점이 있다(허명회 등, 2005).

추정 방법은 모집단이 2개 이상의 특성 변수에 의해 분할표 형태로 나누어지고, 모집단의 특성 변수에 대한 1차원 주변 분포가 알려져 있는 경우에 각 칸의 표본수와 모집단의 주변 분포를 이용하여 1번에 1변수씩 표본의 비율을 모집단의 비율과 일치시키는 과정을 반복적인 알고리즘을 이용하여 구하는 방법이다.

레이킹 추정 과정에 대한 이해를 돕기 위해, 일반적인 표집방법에 대한 2차원 분할표를 이용한 레이킹 추정 과정을 간단히 살펴본다. 추정 방법은 허명회 등(2005)과 임경은(2010)을 기반으로 기술한다.

모집단 U 는 $I \times J$ 개의 셀로 분할되고, ij 번째 셀의 모집단을 $U_{ij}(i=1, \dots, I, j=1, \dots, J)$ 라고 하자. 이때 U_{ij} 의 크기는 N_{ij} 이다. 표본 s 는 임의의 표집방법에 의해 선정되며 $s_{ij} = s \cap U_{ij}$ 는 ij 번째 셀에 속하는 표본이고 크기는 n_{ij} 이다. 그리고 w_k 는 k 번째 조사 단위에 부여되는 가중치로서 표집 방법에 의해 결정되며, 단순임의표집의 경우는 $w_k = N/n$ 이다. 여기서 N 은 모집단 크기, n 은 표본 크기이다. 이 때 우리의 목적은 표본수 n_{ij} 로부터 미지인 각 칸의 모집단수 N_{ij} 를 추정하는 것이다. 이 때 모집단의 1차원 주변분포가 알려져 있으므로 다음의 제약식을 이용할 수 있다.

$$N_{i+} = \sum_{j=1}^J \hat{N}_{ij}, \quad N_{+j} = \sum_{i=1}^I \hat{N}_{ij}$$

이들 제약식은 각 칸의 보정된 모집단수의 주변합을 실제 모집단수와 일치되도록 한다.

알고리즘의 t 번 반복이 경과한 시점에서 모집단수에 대한 ij 번째 셀의 레이킹 추정량은 $\hat{N}_{ij}^{(t)}$ 는 다음과 같이 표현된다.

$$\hat{N}_{ij}^{(0)} = \sum_{k \in s_{ij}} w_k,$$

$$\hat{N}_{ij}^{(t)} = \hat{N}_{ij}^{(t-1)} \times \frac{N_{i+}}{\hat{N}_{i+}^{(t-1)}}, \quad t = 2m - 1, \quad m = 1, 2, 3, \dots$$

$$\hat{N}_{ij}^{(t)} = \hat{N}_{ij}^{(t-1)} \times \frac{N_{i+}}{\hat{N}_{+j}^{(t-1)}}, \quad t = 2m, \quad m = 1, 2, 3, \dots$$

여기서 $\hat{N}_{i+}^{(t-1)} = \sum_j \hat{N}_{ij}^{(t-1)}$, $\hat{N}_{+j}^{(t-1)} = \sum_i \hat{N}_{ij}^{(t-1)}$ 이다. 따라서 레이킹 추정치는 추정량 $\hat{N}_{ij}^{(t)}$ 이 수렴할 때까지 두 개의 변수에 번갈아 가면서 반복된다.

사후층화가중치와 레이킹 가중치의 차이를 보다 분명하게 알기 위해 다음과 같이 성과 연령의 2개 변수를 가진 간단한 자료를 이용하여 실제 추정을 해본다. 표본조사 결과와 각 특성 변수에 대한 모집단의 주변분포는 <표 1-5>, <표 1-6>과 같다. 각 층에서 10% 확률로 표본을 추출했지만, 남성 10대는 표본이 추출되지 않았다고 가정한다.

<표 1-5> 표본조사 결과

i \ j	10대(1)	20대(2)	합계
남성(1)	$n_{11} = 0$	$n_{12} = 30$	$n_{1+} = 30$
여성(2)	$n_{21} = 50$	$n_{22} = 70$	$n_{2+} = 120$
합계	$n_{+1} = 50$	$n_{+2} = 100$	$n = 150$

<표 1-6> 모집단의 주변분포

i \ j	10대(1)	20대(2)	합계
남성(1)	$N_{11} = 100$	$N_{21} = 300$	$N_{1+} = 400$
여성(2)	$N_{21} = 500$	$N_{22} = 700$	$N_{2+} = 1,200$
합계	$N_{+1} = 600$	$N_{+2} = 1,000$	$N = 1,600$

□ 사후층화가중치

Step0) 사후층을 선정 : 성, 연령그룹

Step1) 사후층화가중치 계산 : $w_{ij}^{pst} = N_{ij}/n_{ij}$

□ 레이킹 가중치

Step0) 레이킹 변수 선정 : 성, 연령그룹

Step1) 초기 가중치 $w_0 = N/n = 1600/150 = 10.7$ 계산

Step2) 표본의 각 칸에 초기(혹은 수정) 가중치를 곱한 후, 연령그룹에 대한 주변분포



계산

Step3) Step2에서의 연령그룹의 주변분포와 모집단의 연령그룹에 대한 주변분포의 비로 각 칸을 재조정 한 후, 성별에 대한 주변분포 계산

Step4) Step2와 Step3이 다음의 수렴조건을 만족할 때까지 반복

$$\left| \frac{\hat{N}_{ij}^{(t)} - \hat{N}_{ij}^{(t-1)}}{\hat{N}_{ij}^{(t-1)}} \right| \times 100 < 0.01(\%)$$

위 예제는 7번 반복하면 레이킹 가중치 w_{ij}^{rk} 가 계산됨

위 예의 사후층화가중치와 레이킹 가중치 결과는 <표 1-7>과 같다. 표본이 없는 10대 남성에 대해서는 두 추정법 모두 가중치를 계산할 수 없다. 결과를 비교해보면, 모집단 1,600명에 대해 사후층화는 1,500명, 레이킹 추정은 1,600명으로 모집단 인구를 추정하였다. 사후층화는 표본이 있는 셀에 대해서는 모집단을 정확하게 추정하지만, 표본이 없는 셀이 대표하는 모집단 100명은 추정하지 못하기 때문이다. 반면 레이킹은 표본이 있는 셀에 사후가중보다 큰 가중치를 부여하여 빈 셀에서 대표하지 못한 모집단을 나머지 표본 셀들이 대표할 수 있도록 하였다. 또한 10대, 20대, 그리고 남성, 여성처럼 각각의 주변 분포에 대한 추정 결과도 사후층화가중치보다 정확히 일치하였다. 이처럼 레이킹은 각 변수들의 주변 분포 범주에만 빈 셀이 없으면 추정이 잘 수행된다. 따라서 사후층화에 비해 빈 셀 문제에 보다 유연한 대응이 가능하다.

<표 1-7> 사후층화와 레이킹을 이용한 추정값

i \ j	10대(1)		20대(2)		합계	
	사후층화	레이킹	사후층화	레이킹	사후층화	레이킹
남성(1)	.	.	300	400	300	400
여성(2)	500	600	700	600	1,200	1,200
합계	500	600	1,000	1,000	1,500	1,600

그러나 레이킹이 사후가중에 비해 빈 셀 발생이 덜 문제가 되는 것 뿐이지 전혀 문제가 되지 않는 것은 아니다. 예를 들어, 2×2 이원 분할표에서 0값이 한 개의 셀에서 존재하거나 서로 대각선에 존재(2개)하면 레이킹 가중치는 수렴하지만, 0값이 세 개 이상의 셀에 존재하면 수렴하지 않는다(조태휘, 2005). 이런 경우에는 표본수가 0이 되는 일부 변수의 셀을 인근 셀과 병합하여 표본수가 0인 셀 자체를 제거한 후 레이킹 추정을 시도할 수 있다.

2. 변수 선택

가중치 산정에 활용될 주요 변수는 전수와 10% 표본조사 자료를 이용하여 선택한다. 이를 위해 2005년 인총 항목과 결과 항목을 비교한다(<표 1-8>).

<표 1-8> 2005년 인총 조사 항목과 결과 항목

	전수조사	표본조사
조사 항 목	▶ 가구원(인구)에 관한 사항	
	(1) 성명 (2) 성별 (3) 나이 (4) 가구주와의 관계 (5) 남북이산가족 (6) 교육정도 (7) 혼인상태 (8) 종교	전수조사 항목((1) ~ (8))에 추가 ① 아동보육 ⑨ 종사상의 지위 ② 5년 전 거주지 ⑩ 산업 ③ 활동계약 ⑪ 직업 ④ 통근·통학 여부 ⑫ 근로장소 ⑤ 통근·통학 장소 ⑬ 혼인년월 ⑥ 이용교통수단 ⑭ 총 출생아 수 ⑦ 통근·통학 소요시간 ⑮ 추가 계획 자녀 수 ⑧ 경제활동상태 ⑯ 고령자 생활비 원천
	▶ 가구에 관한 사항	
	(1) 가구 구분 (2) 사용 방 수 (3) 주거 시설 형태 (4) 거주층 (5) 난방 시설 (6) 점유 형태 (7) 주인가구 및 주택 소유 여부	전수조사 항목((1) ~ (7))에 추가 ① 거주 기간 ② 자동차 보유 대수 ③ 주차시설 ④ 임차료
	▶ 주택에 관한 사항	
	(1) 거처의 종류 및 건물층수 (2) 연건평 (3) 대지 면적 (4) 총 방 수 (5) 건축년도 (6) 편익 시설 수	좌동
결 과 항 목	▶ 가구원(인구)에 관한 사항	
	인구, 가구 및 주택(읍면동) 등 16건	① 경제활동(16건) ② 고령자(27건) ③ 국내인구이동통계(9건) ④ 산업·직업(9건) ⑤ 여성·아동(17건) ⑥ 통근·통학(20건) ⑦ 활동계약(17건)
	▶ 가구에 관한 사항	
	가구형태별 가구 및 가구원(읍면동) 등 21건	① 1인가구(14건) ② 경제활동(4건) ③ 고령자가구(6건) ④ 임차료(15건) ⑤ 주거실태(17건) ⑥ 활동계약(4건)
▶ 주택에 관한 사항 : 전수조사 결과로 발표		



표본조사는 전수조사 21개 항목에 23개의 항목을 추가하여 모두 44개 항목을 조사한다. 표본 항목들 중 KOSIS에서 발표되는 가구원에 대한 주요 표본조사 결과는 통근·통학 관련 교통량 조사, 국내인구이동통계, 경제활동, 산업, 직업, 활동제약, 여성, 아동, 고령자 등에 대한 조사 결과이다. 현실적으로 이 모든 특성들을 고려한 가구원의 단일 가중치를 구하는 것은 불가능하지만, 최대한 많은 특성에 대해서 전수와 추정 결과를 일치시키는 노력은 가능하다. 이를 위해, 표본조사 결과 특성에 가장 크게 영향을 미칠 수 있는 주요 변수를 선택해야 한다. 고려된 변수는 성, 연령, 조사구특성, 교육정도, 혼인상태다.

성과 연령은 주요 인구학적 특성 변수로 표본자료의 다른 항목들에 영향을 주고 분산을 최소화 시킬 것이라는 가정으로 사용한다. 그리고 이들 변수는 가중치 보정의 마지막 단계에서 다시 사용될 수 있다. 연령은 모든 연령을 고려할 수가 없기 때문에 2005년과 마찬가지로 5세 계급으로 구분된 21개 연령그룹(<부표 2>)을 고려한다. 마찬가지로 혼인상태와 교육정도는 표본조사의 통근·통학, 경제활동 및 여성 관련 항목에 큰 영향을 미치고, 그 외의 항목에도 유의미한 영향을 미칠 것으로 가정된다. 조사구특성은 공표 항목은 아니지만, 표본조사구 추출 시 고려된 사전 층으로 아파트, 일반, 섬 같은 주거 환경이 표본조사 항목에 유의미한 영향을 미칠 것으로 보인다. 따라서 새로운 가중치 적용을 위해 2005년 인총 가중치 산정 시 고려된 성, 연령그룹 외에 조사구특성, 혼인상태와 교육정도 변수가 가중치 산정에 이용될 수 있을 지 여부를 검토한다.

검토 방법은 조사구특성, 혼인상태와 교육정도에 따라 전수의 성, 연령그룹의 인구특성비에 차이가 있거나, 표본의 성, 연령그룹의 인구특성비, 통근·통학 비중, 경제활동상태 비중에 차이가 있는지를 전수와 표본자료⁴⁾를 대상으로 확인한다. 이를 위해 범주형 변수들 간 상호 관련성 여부에 대한 독립성 검정⁵⁾, 즉, 카이스퀘어(χ^2) 검정⁶⁾을 실시하였다.

가. 조사구특성

2005년 인총은 조사구특성에 따라 아파트조사구를 제1층, 보통조사구와 섬조사구를 제2층으로 사전 층화한 후 표본을 계통 추출하였다. 그리고 통근·통학, 경제활동 및 여성 관련 항목 등 주요 표본 조사 결과는 주거 환경에 따라 차이가 있을 것이다. 따라서,

-
- 4) 표본자료는 전수는 아니지만, 표본수 자체가 상당히 크기 때문에 전체적으로는 모집단과 비슷한 분포를 가질 것으로 기대됨
 - 5) 2개의 범주형 변수가 통계적으로 서로 관련이 있는지 검정. 이에 대한 귀무가설은 “두 변수는 서로 독립이다”, 대립가설은 “두 변수는 서로 독립이 아니다”임
 - 6) 유의수준(0.01, 0.05 등)에 따라 다르지만, χ^2 통계량은 클수록 귀무가설을 기각하게 됨. 즉 χ^2 통계량이 클수록 두 변수는 서로 독립이 아님

조사구특성이 가중치 산정에 이용되는 것이 유용할 것인지 확인해야 한다.

전수와 표본자료에서 조사구특성에 따른 시도별 인구 비중을 비교해보면 <표 1-9>와 같다. 섬조사구처럼 일부 규모가 작은 읍면동 지역의 경우, 과소 표집에 따른 문제를 사전에 방지하기 위해 원래 비중보다 다소 많은 수의 표본을 추출했지만(<표 1-1> 참고), 특정 조사구특성이 불균등하게 뽑히지는 않은 것으로 보인다. 조사구가 기본적인 표집단이지만, 가구원도 대략적으로는 모집단에서 약 1/10의 확률로 뽑힌 것을 확인할 수 있다.

<표 1-9> 조사구특성에 따른 조사인구와 비중

시도	전수(천 명, %)						표본(백 명, %)					
	A		1		2		A		1		2	
	인구	비중	인구	비중	인구	비중	인구	비중	인구	비중	인구	비중
서울	3,815	8.33	5741	12.53	-	-	3,624	8.12	5,178	11.60	-	-
부산	1,566	3.42	1863	4.07	2	0.00	1,479	3.32	1,719	3.85	5	0.01
대구	1,132	2.47	1271	2.77	-	-	1,070	2.40	1,149	2.58	-	-
인천	1,243	2.71	1213	2.65	18	0.04	1,173	2.63	1,138	2.55	47	0.10
광주	850	1.86	522	1.14	-	-	815	1.83	484	1.09	-	-
대전	752	1.64	648	1.41	-	-	708	1.59	582	1.30	-	-
울산	537	1.17	481	1.05	-	-	493	1.10	426	0.95	-	-
경기	5,438	11.87	4671	10.20	0	0.00	5,045	11.30	4,285	9.60	1	0.00
강원	582	1.27	825	1.80	-	-	550	1.23	978	2.19	-	-
충북	622	1.36	781	1.70	-	-	559	1.25	926	2.07	-	-
충남	688	1.50	1107	2.42	3	0.01	644	1.44	1,248	2.80	3	0.01
전북	794	1.73	923	2.02	3	0.01	761	1.71	1,203	2.70	9	0.02
전남	614	1.34	1044	2.28	102	0.22	577	1.29	1,348	3.02	167	0.37
경북	944	2.06	1542	3.37	-	-	884	1.98	1,856	4.16	-	-
경남	1,331	2.91	1616	3.53	10	0.02	1,184	2.65	1,789	4.01	15	0.03
제주	99	0.22	416	0.91	4	0.01	86	0.19	404	0.91	10	0.02
	21,007	45.9	24,665	53.8	144	0.3	19,652	44.0	24,714	55.4	259	0.6

※ 1. 비중 = (해당 시도 내의 해당 조사구특성 조사인구 / 전국 조사인구) × 100

2. "0", "0.00" : 조사인구는 있지만, 적은 수, "-" : 조사인구가 없음.

조사구특성이 인구구성비, 통근·통학 비중 및 취업·실업 등 경제활동상태 비중에 미치는 영향을 확인하기 위해 카이스퀘어(χ^2) 검정을 실시하였다. 검정 결과는 <표 1-10>과 같이 나타났다. 표의 첫 번째 행은 아파트조사구, 보통조사구, 섬조사구를 각각 하나의 범주로 놓고 검정한 결과이고, 두 번째 행은 보통조사구와 섬조사구만을 대상으로 독립성 검정을 시행한 결과이다.

조사구특성과 성별 간 독립성 검정을 위한 χ^2 통계량 157.425의 유의확률이 0.01보다



작아, 유의수준 0.01에서 조사구특성(A, 1, 2)과 성별이 서로 관련이 있다. 즉, 조사구특성에 따라 성별의 분포가 다르다고 말할 수 있다. 마찬가지로 조사구특성은 연령그룹, 통근·통학, 경제활동상태에서 모두 유의한 영향을 미치는 것으로 나타났다.

두 번째로는 보통조사구와 섬조사구 사이에 인구 특성 분포에 대한 차이가 있는지를 확인해 보았다. 검정 결과 표본에서의 성별 분포를 제외하고는 보통조사구와 섬조사구는 인구 특성 분포에서 차이를 보이는 것으로 나타났다. 그러나, 섬조사구는 부산, 인천, 경기 등 특정 광역시도에만 존재하고, 전체에서 차지하는 비중이 너무 작기 때문에 섬조사구를 하나의 조사구특성으로 분류하여 가중치를 산정할 경우 항목에 대한 빈 셀이 상당수 발생하고 특정 하위 그룹에서는 상당히 불안정한, 즉, 표본이 너무 작기 때문에 과대한 가중치가 계산될 수 있다. 따라서 보통조사구와 섬조사구의 주요 특성이 크게 차이가 나지 않는다면, 하나의 조사구특성으로 처리하는 것이 효율적이다. 결과적으로 인구 가중치 계산을 위한 조사구특성은 아파트조사구(A)와 일반(1)·섬(2)조사구 두 개로 나누어 사용한다.

〈표 1-10〉 조사구특성이 인구구성, 통근통학 여부, 경제활동상태에 미치는 영향(χ^2 검정 통계량)

전수			표본			
조사구 특성 ×	성별	연령그룹	성별	연령그룹	통근·통학	경활상태
A,1,2	157,425**	1,838,652**	107,782**	235,887**	19,453,757**	11,151,996**
1,2	25,345**	59,527,214**	2,261**	8,323,851**	1,701,819**	1,390,756**

※ 1. **: 유의수준 $p < 0.01$

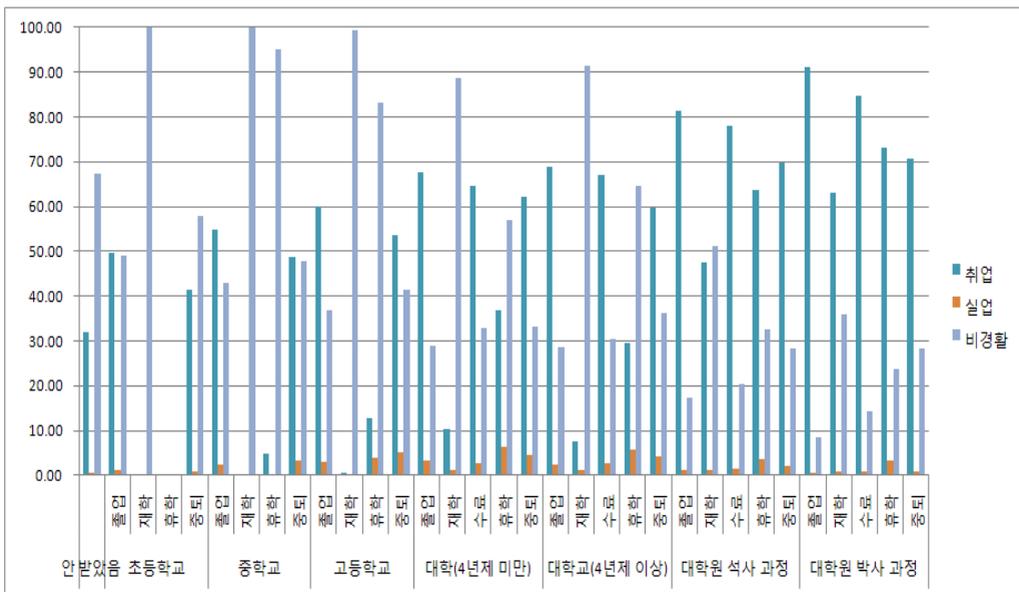
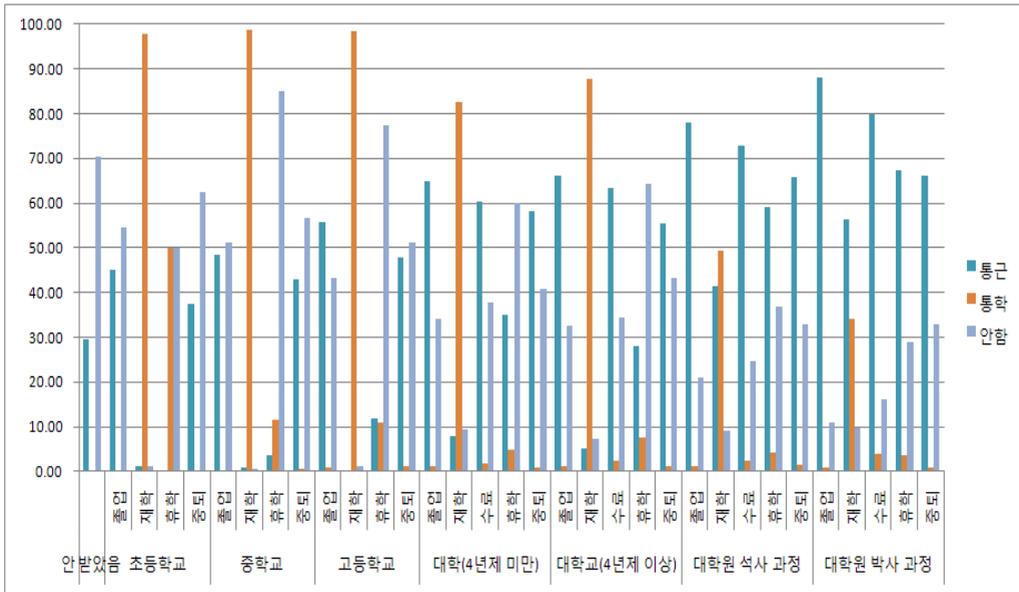
2. A : 아파트조사구, 1 : 보통조사구, 2 : 섬조사구

나. 교육정도

교육정도는 정규교육을 안 받았음(미취학 포함)부터 대학원 박사과정 수료까지 모두 33개의 세부 항목으로 조사된다. 교육정도에 따른 통근·통학 여부와 경제활동상태에 대한 비중을 살펴보면 [그림 1-3]과 같다. 초등학교, 중학교, 고등학교의 재학생들은 통학 비율이 월등히 높으면서 비경활 비율이 거의 100%에 달하는 등 비슷한 특성을 가지고 있으므로, 하나의 교육정도 항목으로 묶을 수 있다. 마찬가지로, 초등학교, 중학교, 고등학교를 졸업했거나 중퇴한 사람들은 통근·통학 비율과 취업·실업 등 경제활동 특성이 서로 비슷한 경향이 있다. 이런 경향들을 정리해서 원래 33개의 교육정도 항목을 다음과 같이 10개의 항목으로 재분류하였다.

- | | |
|------------------|--------------|
| ① 안 받았음(미취학 포함), | ② 초중고 재학/휴학, |
| ③ 초등학교 졸업/중퇴, | ④ 중학교 졸업/중퇴, |

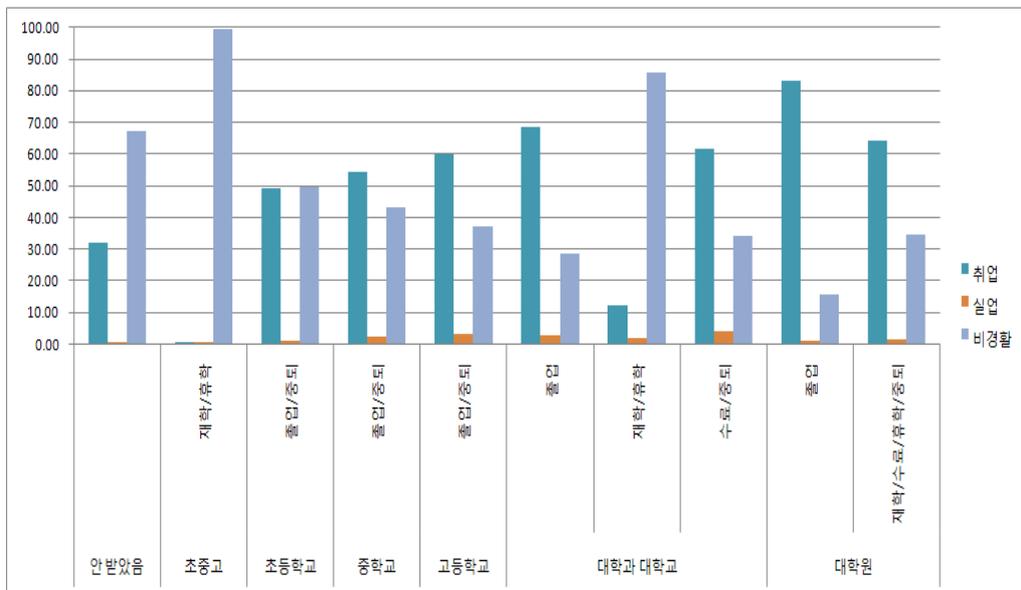
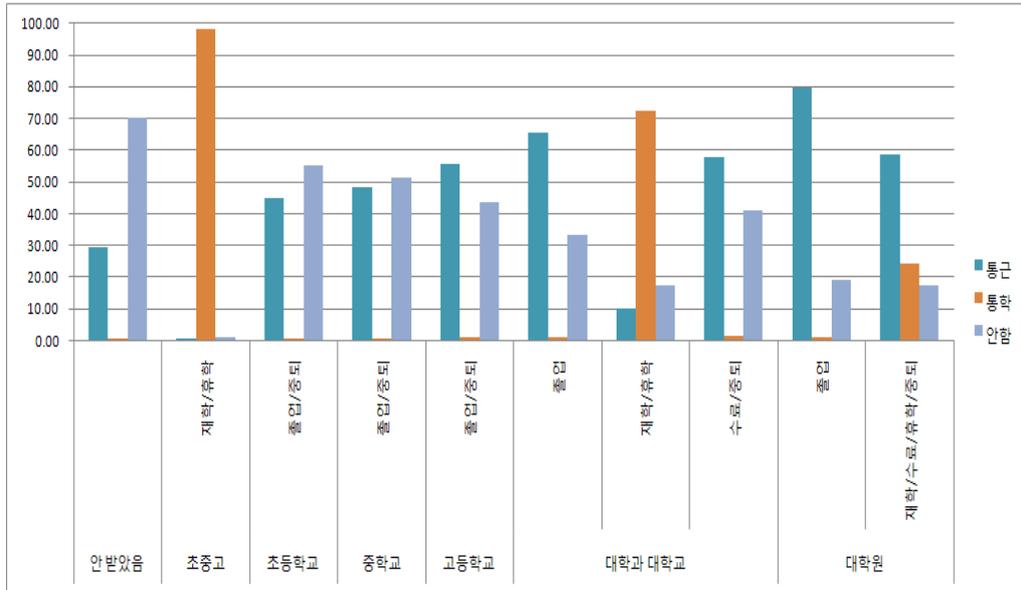
- ⑤ 고등학교 졸업/중퇴,
- ⑥ 대학과 대학교 졸업,
- ⑦ 대학과 대학교 재학/휴학,
- ⑧ 대학과 대학교 수료/중퇴,
- ⑨ 대학원 졸업,
- ⑩ 대학원 재학/수료/휴학/중퇴



[그림 1-3] 교육정도에 따른 통근·통학과 취업·실업 비중



재분류된 교육정도에 따른 통근·통학 비율과 경제활동 특성은 [그림 1-4]와 같다.



[그림 1-4] 새로 분류된 교육정도에 따른 통근·통학과 취업·실업 비중

새로 분류된 교육정도가 인구구성, 통근·통학 여부, 경제활동상태에 미치는 영향을 확인해보기 위한 독립성 검정 결과는 <표 1-11>과 같다. 새로 분류된 교육정도와 성별 간 독립성 검정을 위한 χ^2 검정 통계량 870,296의 유의확률이 0.01보다 작아, 유의수준 0.01에서 교육정도에 따라 성비에 차이가 있다고 말할 수 있다. 마찬가지로 교육정도는 연령그룹, 통근·통학, 경제활동상태에 모두 유의한 영향을 미치는 것으로 나타났다.

<표 1-11> 교육정도가 인구구성, 통근통학 여부, 경제활동상태에 미치는 영향(χ^2 검정 통계량)

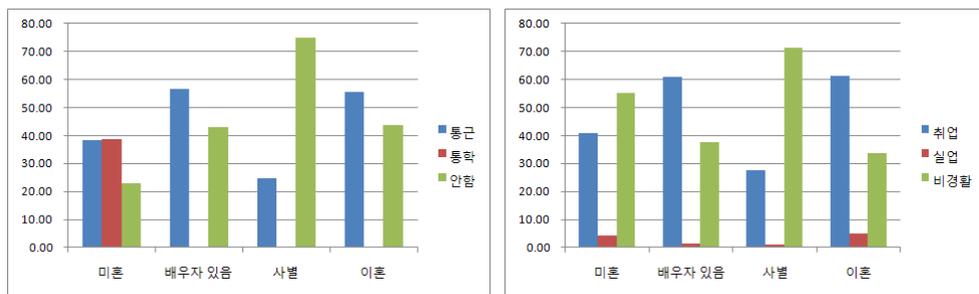
	전수		표본			
	성별	연령그룹	성별	연령그룹	통근·통학	경활상태
교육정도 ×	870,296**	100,362,166**	90,271,072**	9,663,094**	3,367,680**	589,917**

※ ** : 유의수준 p<0.01

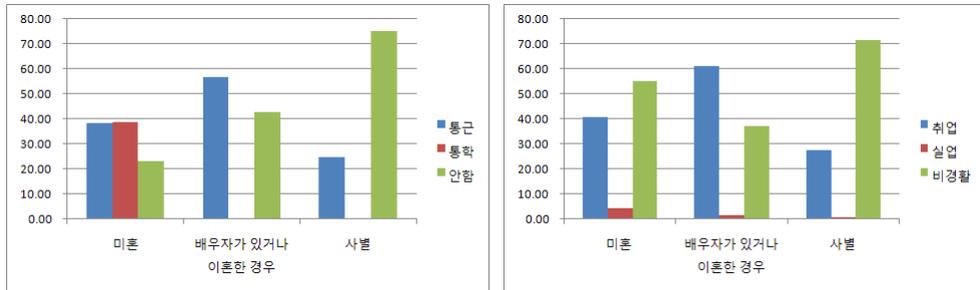
다. 혼인상태

혼인상태는 미혼, 배우자 있음, 사별, 이혼의 4가지 범주로 조사된다. 혼인상태에 따른 통근·통학 여부와 경제활동상태에 대한 비중을 보면 [그림 1-5]와 같다. 각 혼인상태에 따른 특성을 확인해보면, 미혼은 다른 범주와 확연히 다른 통근·통학 경향을 보인다. 이는 미혼 범주에는 연령대가 어린 학생 인구가 많기 때문이다. 학생 인구가 많기 때문에 실업과 비경활 인구의 비중도 여타의 경우에 비해 크다. 그리고 배우자가 있거나 이혼한 경우에는 통근·통학과 경제활동상태가 비슷한 경향을 보인다. 또한 이혼은 전체 중에서 차지하는 비중(3.04%)이 작다. 따라서, 혼인상태에 대해서는 미혼, 배우자 있음과 이혼, 사별의 3개 범주로 나누어서 분석에 사용하기로 한다.

재분류된 혼인상태에 따른 통근·통학과 경제활동상태에 대한 비중은 [그림 1-6]과 같다.



[그림 1-5] 혼인상태에 따른 통근·통학과 취업·실업 비중



[그림 1-6] 새로 분류된 혼인상태에 따른 통근·통학과 취업·실업 비중

3개의 범주로 재분류된 혼인상태가 인구구성, 통근·통학 여부, 경제활동상태에 미치는 영향을 확인해보기 위한 독립성 검정 결과는 <표 1-12>와 같다. 새로 분류된 혼인상태와 성별 간 독립성 검정을 위한 χ^2 검정 통계량은 1,578,154(유의수준 $p < 0.01$)로, 유의수준 0.01에서 혼인상태에 따라 성비에 차이가 있다고 말할 수 있다. 마찬가지로 혼인상태는 연령그룹, 통근·통학, 경제활동상태에 모두 유의한 영향을 미치는 것으로 나타났다.

<표 1-12> 혼인상태가 인구구성, 통근·통학 여부, 경제활동상태에 미치는 영향(χ^2 검정 통계량)

	전수		표본			
	성별	연령그룹	성별	연령그룹	통근·통학	경활상태
혼인상태 ×	1,578,154**	33,743,227**	183,445**	8,546,309**	2,061,859**	216,860**

※ **: 유의수준 $p < 0.01$

이상에서 살펴본 바와 마찬가지로, 전수에서 조사되는 항목은 기본적으로 주요 인구특성에 많은 영향을 끼친다. 그만큼 중요도가 높고 일반적인 특성을 반영하는 조사 항목이기 때문이다. 그러나 이런 변수와 변수들에 관한 응답 항목을 모두 가중치 산정에 반영할 수 없기 때문에 주요 변수로 조사구특성, 교육정도, 혼인상태 변수를 추가로 분석해보았다. 이들 변수는 다양한 가중치 산정 방안 고려 시 사후 층이나 레이킹 변수로 활용된다.

3. 셀 병합

사후층화나 레이킹을 추정 방법으로 이용하기 위해서는 각 방법에 따라 조금 차이는 있지만, 기본적으로 빈 셀이 발생하지 않아야 한다. 그러나 주요 변수의 항목 간에는 빈 셀이 발생하는 경우가 종종 있다. 예를 들면, 모집단에는 A읍에 85세 이상의 남성이 있

지만, 표본에는 이 범주에 속하는 조사인구가 없을 수 있다. 이런 경우, 허명희 등(2005)은 표본수가 0인 셀은 모집단수도 매우 작은 경우가 대부분이므로 해당하는 셀의 모집단수 손실을 감내하는 것도 현실적으로는 괜찮은 방법이라고 제안하였다. 다른 대안은 이런 빈 셀을 인근 셀과 병합하여 표본수가 0인 셀을 제거한 후 제안된 방법을 사용하는 것이다. 인총 가중치 산정의 주요 원칙은 최대한 모집단 인가와 추정 인구 사이에 일치(consistency)를 이루는 것이기 때문에, 본 연구에서는 표본수가 0인 셀을 인근 셀과 병합하여 가중치를 산정하는 방법을 선택한다.

셀 병합을 위해서는 우선 변수들 간 교차표나 주변 분포에서 발생하는 빈 셀 현황을 파악해야 한다. 2005년 인총의 가중치 산정을 위해 사용된 사후층은 442개 행정구역(시도·시군구·동읍면부), 2개 성, 21개 연령그룹에 대한 18,564개 층(셀)이었다. 이 중 빈 셀은 18개로 전체의 0.1%를 차지한다.

빈 셀의 특성을 보면, 충북을 제외한 나머지 도(道) 지역에서 모두 발생하였고, 특히 읍단위 지역에서 발생 빈도가 높다. 표본이 뽑히지 않은 빈 셀의 전수인구는 40명 이하이고, 여성층보다 남성층에서 빈 셀이 더 많이 발생하였다. 연령그룹은 2, 4, 5, 21일 때 발생하였다. 연령그룹의 내용은 <부표 2>를 참고할 수 있다.

<표 1-13> 2005년에 적용된 사후층 중 빈 층(셀) 현황

시도	동읍면부		전수인구		성		연령그룹	
	빈셀수	빈셀수	빈셀수	빈셀수	빈셀수	빈셀수	빈셀수	
서울		동 2	~10 3	남 14	2 (5세)	4		
부산		읍 15	~20 6	여 4	4 (10~11세)	1		
대구		면 1	~30 5		5 (12세)	3		
인천			~40 4			21 (85세 이상)	10	
광주								
대전								
울산								
경기	1							
강원	1							
충북								
충남	2							
전북	4							
전남	4							
경북	4							
경남	2							
제주								
계	18							



본 연구에서는 빈 셀을 인근 셀과 병합하기 위해 행정구역, 성별보다는 연령그룹에서 빈 셀 병합을 수행한다. 즉, 빈 셀 연령 그룹이 2(5세)인 경우에는 3(6~9세), 4(10~11세)인 경우에는 5(12세), 5(12세)일 경우에는 4(10~11세), 21(85세 이상)일 경우에는 20(80~84세)으로 병합하였다. 이는 빈 셀이 인근 셀로 병합되어도 연령그룹이 최대한 5세별 분류에서 벗어나지 않는 것을 기준으로 하였다. 연령그룹이 5(12세)일 경우에 4(10~11세)로 병합한 이유는 연령그룹 5는 초등학교 재학생이 다수인데 반해 연령그룹 6(13~14세)은 중학교 재학생이므로 6보다는 같은 초등학교 연령그룹인 4와 그 특성이 더 비슷할 것이기 때문이다.

조사구특성을 사후층으로 이용하면, 사후층은 442개 행정구역(시도·시군구·동읍면부), 2개 조사구특성(아파트조사구, 보통·섬조사구), 2개 성, 21개 연령그룹에 대한 35,314개가 된다. 이 중 빈 셀은 2,485개로 전체의 7.0%를 차지한다. 이는 읍면단위에서 표집되지 않은 아파트조사구의 특성들이 빈 셀로 나타났기 때문이다. 동읍면부 내의 비표집 아파트조사구는 45개인데, 그 중 읍 지역이 16개, 면 지역이 29개로 주로 면 지역의 아파트조사구가 표집되지 않았다. 아파트조사구가 표집되지 않은 45개 읍면 지역에서 전수자료를 이용하여 아파트조사구와 나머지 조사구간 성별, 연령그룹별 구성비를 살펴보면(<표 1-14>) 아파트조사구에는 저연령대 인구가 많이 분포하고, 다른 조사구에는 고연령대 인구가 많이 분포했다. 그러나 해당 지역에서 비표집된 아파트조사구 인구가 차지하는 비중은 상당히 작다. 따라서 다른 셀과 병합하기보다는 해당 동읍면 내의 보통·섬조사구와 아파트조사구를 구분하지 않고 가중치를 부여하는 것이 보다 합리적이고 간편한 방법으로 보인다.

<표 1-14> 비표집 아파트조사구와 표집 조사구 간 성별, 연령그룹별 구성비

(명, %)

조사구 특성	성별		연령그룹								
	남	여	1	2	3	4	5	6	7	8	9
A	10,721	10,558	4,567	2,863	2,452	5,095	3,459	1,506	731	450	156
	1.15	1.13	0.49	0.31	0.26	0.54	0.37	0.16	0.08	0.05	0.02
1,2	425,551	489,248	65,399	81,623	64,408	82,636	119,895	137,494	175,761	136,856	50,727
	45.46	52.27	6.99	8.72	6.88	8.83	12.81	14.69	18.78	14.62	5.42

* 연령그룹 : 10세 단위

4. 가중치 평가

제안된 가중치에 대한 평가는 시도, 시군구, 동읍면부 내의 주요 항목(성별, 연령그

률, 교육정도, 혼인상태, 조사구특성, 가구주와의 관계)에 대해 추정된 인구수와 모집단의 인구 사이의 불일치(discrepancy) 정도 D 를 이용한다. 불일치 정도 D 는 다음과 같은 식으로 계산될 수 있다.

$$D = \frac{\text{표본 추정인구} - \text{모집단 인구}}{\text{모집단 인구}} \times 100$$

분자는 표본 추정인구와 모집단 인구 사이의 차이를 나타내고, 이 차이가 모집단의 크기에 비례하여 어느 정도의 비중으로 나타나는지 확인할 수 있다(캐나다 통계청, 2009).

불일치 정도를 전체적으로 살펴보기 위해 결과 비교는 다음과 같은 불일치 정도의 절대값의 평균 \bar{D} 로 하기로 한다.

$$\bar{D} = \sum_{i=1}^m |D_i| / m$$

여기서 m 은 항목의 수를 나타낸다. 예를 들어 시군구 내의 동읍면부별 성별 불일치 정도를 나타낼 때 m 은 동읍면부 수 $\times 2$ 이다. 불일치 정도가 작을수록 좋은 추정 결과를 나타낸다.

5. 가중치 지역

가중치 계산을 위해서는 가중치 지역(Weighting Areas)을 설정해야 한다. 가중치 지역이란 가중치 절차가 적용되고, 주요 특성에 대해 통계값들이 일치성을 갖기를 원하는 제일 작은 지리적 지역을 말하는데 일반적으로 시도, 시군구와 같은 행정 구역을 말한다.

미국 센서스국은 1980년 센서스 가중치 계산을 위해 다음과 같은 3가지 조건을 만족시키는 가중치 지역을 설정했다(Woltman 등, 1981).

1. 가중치 지역은 최소 400명의 표본 인구를 가져야 한다.
2. 최소 표본이 400명이 되지 않더라도 가중치 지역은 카운티의 경계를 넘을 수 없다
3. 1,2의 조건 아래서 가중치 지역은 카운티 경계 내의 모든 지역을 포함해야 한다.

캐나다 통계청도 2006년 센서스 가중치 계산을 위한 가중치 지역을 다음과 같이 설정했다(캐나다 통계청, 2009).

1. 가중치 지역은 센서스 구역(census divisions)의 경계를 넘지 않는다.



2. 가중치 지역은 전형적으로 1,000~3,000 단위 거처(dwelling)를 모집단에 포함한다.
3. 가중치 지역은 가능하면, 센서스 하위구역(census subdivision)과 센서스 트랙(census track) 경계를 넘지 않는다.
4. 가중치 지역은 가능하면 인접하는 DAs(Dissemination Areas)로 구성된다.

즉, 가중치 지역을 선정함에 있어 가장 중요한 요소는 가중치 지역이 서로 겹치지 않으면서 전국을 모두 포함하는 것이다. 그리고 추정에 필요한 적절한 인구수를 가지고 있어야 한다. 이런 조건을 2005년 인총 자료에 적용해보기 위해 <표 1-15>와 같이 시도, 시군구, 동읍면부의 조사구특성까지 고려하는 가중치 예상 지역 수를 검토해보았다. 전국적으로 시도, 시군구, 동읍면부, 조사구특성까지 고려한 지역은 886개로 그 중 121개 지역은 전수 인구 규모가 2,000명 미만이다. 또한, 일부 읍면 지역의 아파트조사구는 아예 표집되지 않는 경우도 있다. 조사구특성까지 가중치 지역으로 선정함으로써, 추정 과정에서 추정값이 계산되지 않거나, 가중치의 이상치가 발생할 가능성이 높다. 따라서, 본 연구는 우리나라의 가중치 지역으로 시도, 시군구, 동읍면까지 고려하고, 사후가중치 계산 시 조사구특성에 따른 영향을 고려하여 계산하도록 한다.

<표 1-15> 시도, 시군구, 동읍면, 조사구특성을 고려한 가중치 예상 지역 수

시도	시군구 수	가중치 지역 수	그 중 전수인구가 2,000명 미만인 지역 수(비중 %)
전국	250	886	121 (13.7)
서울	25	50	
부산	16	34	
대구	8	18	
인천	10	23	1 (4.3)
광주	5	10	
대전	5	10	
울산	5	12	
경기	44	142	4 (2.8)
강원	18	74	12 (16.2)
충북	13	48	5 (10.4)
충남	16	83	16 (19.3)
전북	15	62	15 (24.2)
전남	22	96	27 (28.1)
경북	24	110	19 (17.3)
경남	20	101	18 (17.8)
제주	4	13	4 (30.8)

6. 가중치 부여 원칙

가중치를 가구원이나 가구에 부여하기 위해서는 몇 가지 원칙이 필요하다. 미국 센서스국에서는 1980년 센서스 가중치 부여를 위해 다음과 같은 3가지 원칙을 제시했다 (Woltman, H. 등, 1981).

1. 하나의 가구원, 하나의 가구에는 각각 오직 한 개의 가중치만 부여한다. 모든 표본 추정값을 생산하기 위해서 다양한 가중치를 부여할 수는 있다. 그러나 다양한 가중치가 부여된다면, 이를 저장하고 제어하고 활용하는 것이 불가능하다.
2. 부여된 가중치는 정수여야 한다. 이는 자료 이용자의 편의를 증진시키기 위함인데, 만약 소수점을 가지는 가중치를 제공한다면 소수점의 어느 위치까지를 사용할 것인가의 결정에 따라 결과가 달라지기 때문이다.
3. 특정 특성에 대한 표본 추정값은 해당 모집단의 총계와 맞아야 한다. 특히 총 인구와 가구수에서 일치해야 하고, 그 외에도 나이, 인종, 성별과 태생(origin) 등에서도 일치할 수 있다. 이것 역시 자료 이용자의 편의를 증진시킨다.

캐나다 통계청은 가중치를 부여함에 있어 최종 센서스 가중치를 1에서 25 범위 내로 제한함으로써 특정 레코드의 과대 가중치가 표본조사 항목에 큰 영향을 미치는 것을 사전에 제한하였다.

미국과 캐나다의 이와 같은 가중치 부여 원칙은 그 동안 우리의 인종 가중치 산정 시에도 지속적으로 적용되어 왔었다. 즉, 각 레코드(가구원, 가구)별로 하나의 정수 가중치를 부여해왔다. 그러나 본 연구에서는 가중치의 정수화는 고려하지 않도록 한다. 가중치 정수화가 주는 이용상의 많은 장점이 인종처럼 규모가 큰 조사 자료에서 가중치를 획일적으로 정수화시킴으로 추정 결과가 왜곡될 수 있는 단점을 모두 극복할 수 있는지 확신할 수 없기 때문이다.

따라서, 본 연구에서는 가중치 부여 원칙을 재정리하고, 향후 가중치 산정에 활용하고자 다음과 같은 4가지 가중치 부여 원칙을 제시한다.

1. 하나의 가구원, 하나의 가구에는 각각 오직 한 개의 가중치만 부여한다.
2. 부여된 가중치는 소수점을 포함⁷⁾할 수 있다.
3. 특정 특성에 대한 표본 추정값은 해당 모집단의 총계와 맞아야 한다. 특히 총인구수와 가구수는 일치해야 한다.
4. 최소 가중치는 1, 최대 가중치는 가급적 40으로 제한한다.

7) 허용 가능 소수점 자릿수 지정 필요



제4절 가중치 방안 적용 및 결과 분석

1. 가중치 방안

가중치 산정을 위한 주요 변수로 시도, 시군구, 동읍면부, 조사구특성, 성, 연령그룹, 혼인상태를 고려해서 5가지 가중치 산정 방안을 적용해 보았다. 변수 선택을 위해 고려한 교육정도 항목의 사용은 최종 후보 방안들에 포함되지 않았다. 다양한 측면으로 활용 방안을 검토했으나 교육정도의 범주가 많아 빈 셀이 많이 발생하는 등 가중치 개선 정도에 비해 산정 과정의 효율성이 떨어지는 단점이 있었다.

5가지 가중치의 구체적인 산정 과정은 다음과 같다.

<방안 1> 성, 연령그룹 분류에 따른 사후가중치(mth1)

성, 연령그룹별로 사후 총화하여 다음과 같이 사후가중치를 계산한다.

$$mth1 = \frac{\text{성.연령그룹 전수인구}}{\text{성.연령그룹 표본인구}}$$

<방안 2> 성, 연령그룹, 조사구특성 분류에 따른 사후가중치(mth2)

성, 연령그룹별로 사후 총화하여 사후가중치를 계산한 후 조사구특성 인구에 비례하여 다음과 같이 가중치를 조정한다.

$$mth2 = \frac{\text{성.연령그룹 전수인구}}{\text{성.연령그룹 표본인구}} \times \text{조사구특성 조정계수}$$

여기서, 조사구특성 조정계수란, 전수자료에서 계산된 동읍면부 내의 아파트조사구와, 보통·섬조사구의 인구 비중이다.

<방안 3> 성, 연령그룹, 조사구특성을 이용한 레이킹 적용(mth3)

성, 연령그룹, 조사구특성의 1차원 주변 분포를 이용한 레이킹 추정을 한다. 이 때 35를 넘는 가중치가 계산될 경우 35로 고정하고 다시 성, 연령그룹만 고려한 1차원 주변 분포로 레이킹 추정하여, 초과된 가중치 부분 만큼의 인구를 다른 그룹으로 분배해서 보정해준다.

<방안 4> 성, 연령그룹, 조사구특성, 혼인상태를 이용한 레이킹 적용(mth4)

성, 연령그룹, 조사구특성, 혼인상태 등 4개 변수의 1차원 주변 분포를 이용한 레이킹 추정을 한다. 이 때 35를 넘는 가중치가 계산될 경우 35로 고정하고 다시 성, 연령그

룹만 고려한 1차원 주변 분포로 레이킹 추정하여, 초과된 가중치 부분 만큼의 인구를 다른 그룹으로 분배해서 보정해준다.

<방안 5> 성, 연령그룹의 사후가중치와 조사구특성, 혼인상태를 이용한 레이킹 적용(mth5)

<방안 1>에서 계산된 사후가중치(mth1)를 초기 가중치로 하여 레이킹 추정을 하는데, 초기 가중치가 35를 초과할 경우에는 35로 고정하여 사용한다. 이때 레이킹 변수로는 조사구특성과 혼인상태를 고려한다. 사후가중 시 이미 성, 연령그룹에 대해서 고려했기 때문에 레이킹에서는 이 변수를 추가적으로 이용하지 않는다. 레이킹 수행 후 최대 가중치가 35를 초과할 경우, 35로 고정하고 다시 성, 연령그룹만 고려한 1차원 주변 분포로 레이킹 추정하여, 초과된 가중치 부분 만큼의 인구를 다른 그룹으로 분배해서 보정해준다.

2. 가중치 방안 적용 결과

각 방안별 추정 결과를 비교해본다. 우선 산출된 가중치들의 전국 분포는 <표 1-16>과 같다.

<표 1-16> 방안별 가구원 가중치 분포

방안	평균값	최소값	최대값	합계
mth2005	10.27	2.0	40.00	45,812,477
mth1	10.27	2.0	139.00	45,812,477
mth2	10.27	0.6	139.00	45,812,477
mth3	10.27	0.9	59.76	45,812,477
mth4	10.27	0.9	73.02	45,812,477
mth5	10.27	0.7	62.56	45,812,477

※ mth2005는 2005년에 적용된 가중치, mth1~5는 각각 방안 1~5에 의해 산출된 가중치임

가중치 평균은 비교된 방안 모두 10.27이고 전체 인구수는 모집단과 추정 결과가 일치하였다. 2005년에 산정된 가중치는 최소값 2, 최대값은 40이고 사후가중치를 이용한 가중치 mth1과 mth2는 각각 최소값이 2, 0.6이고, 최대값은 139이다. 레이킹 추정을 하면서 초기 가중치를 이용하지 않은 mth3~4는 사후가중치에 비해서는 조금 작은 최대값 59.76과 73.02를 가지며 레이킹 추정 수행 시 사후가중치를 초기 가중치로 이용한 mth5는 최대 가중치가 62.56으로 나타났다. mth2~mth5는 최소 가중치가 1보다 작은 경우가 발생했는데, 이는 조사구특성과 혼인상태가 산정 과정에 추가되어 나타난 경향이다.

주요 특성에 대하여 전수 인구와 추정된 인구수가 어느 정도 일치되었는지 확인한다.



일치성 확인을 위해서 성, 연령그룹, 성×연령그룹, 교육정도, 혼인상태, 조사구특성, 가구주와의 관계와 같은 특성을 이용한다. 비교를 위한 최소 단위는 가중치 지역, 즉, 시도·시군구·동읍면부이다.

가. 성

성에 대한 불일치 정도를 계산한 결과는 <표 1-17>과 같다. mth2005, mth1과 mth3은 모두 시도·시군구·동읍면부 수준에서 전수 인구나 표본 추정 인구가 일치하였다. mth4는 다른 지역은 모두 일치하였지만, 서울 노원구에서 전수 인구나 표본 추정 인구가 남·여 각 ±1명씩 차이가 발생하였다. 이는 추정 알고리즘상 원칙적으로는 전수 인구나 표본 추정 인구가 성별에 대해 정확히 일치해야 하지만 현재 SAS 시스템에서 제공하는 소수점 가능 자리수(9자리) 때문에 발생한 차이로 생각된다.⁸⁾ mth2와 mth5는 초기 가중치 산정 시에는 성별 인구를 고려하였지만, 결과적으로 추정 과정에서 마지막에 사용된 특성, 즉, mth2는 조사구특성, mth5는 조사구특성과 혼인상태에 맞추어 가중치가 보정되었기 때문에 성별에 대해서는 전수 인구나 추정 인구가 불일치하는 결과를 많이 보여준다.

<표 1-17> 시도·시군구·동읍면부성 평균 불일치 정도

시도	mth2005	mth1	mth2	mth3	mth4	mth5
서울	0.00	0.00	0.02	0.00	0.00	0.10
부산	0.00	0.00	0.02	0.00	0.00	0.23
대구	0.00	0.00	0.02	0.00	0.00	0.13
인천	0.00	0.00	0.05	0.00	0.00	0.16
광주	0.00	0.00	0.05	0.00	0.00	0.17
대전	0.00	0.00	0.01	0.00	0.00	0.08
울산	0.00	0.00	0.01	0.00	0.00	0.08
경기	0.00	0.00	0.07	0.00	0.00	0.21
강원	0.00	0.00	0.19	0.00	0.00	0.45
충북	0.00	0.00	0.13	0.00	0.00	0.36
충남	0.00	0.00	0.20	0.00	0.00	0.48
전북	0.00	0.00	0.17	0.00	0.00	0.37
전남	0.00	0.00	0.12	0.00	0.00	0.53
경북	0.00	0.00	0.12	0.00	0.00	0.46
경남	0.00	0.00	0.19	0.00	0.00	0.51
제주	0.00	0.00	0.19	0.00	0.00	0.51
전국	0.00	0.00	0.12	0.00	0.00	0.36

※ **0.00** : 불일치 정도가 실제로 0은 아니고 소수점 때문에 0.00으로 표현

8) 연령그룹, 혼인상태, 조사구특성에서도 비슷한 사례 발생. 가중치 정수화와 마찬가지로 소수점의 위치를 어느 정도까지 허용할 것인지도 주요 검토 사항임

나. 연령그룹

mth2005와 mth1은 광역시 지역에서는 시군구·동읍면부 수준에서 전수 인구나 표본 추정 인구가 일치하였고, 도(道) 지역에서는 다소 불일치하는 경향이 나타났다. mth1은 불일치 정도가 0.07로 mth2005에 비해서는 많이 작았다. mth3과 mth4는 불일치 정도가 전국적으로 모두 0.01로 6개의 방안들 중 가장 작다. 이 두 방법 역시 추정 알고리즘으로는 전수 인구나 표본 추정 인구가 연령그룹에 대해 정확히 일치해야 하지만 그렇지 않은 이유는 성별에서 일치하지 않은 이유와 마찬가지로 소숫점 문제로 보인다. mth3과 mth4에서 전남과 경북지역에서 연령그룹별 인구가 일치하지 않은 경우는 일부 지역에서 연령그룹이 빈 셀인 경우에 인근 연령그룹으로 병합하였기 때문이다. mth2와 mth5는 초기 가중치 산정 시에는 연령그룹별 인구를 고려하였지만, 성별에서와 마찬가지로 마지막에 사용된 특성에 맞추어 가중치가 보정되었기 때문에 연령그룹에 대해 전수 인구나 표본 추정 인구 불일치 정도가 크다.

〈표 1-18〉 시도·시군구·동읍면부·연령그룹 평균 불일치 정도

시도	mth2005	mth1	mth2	mth3	mth4	mth5
서울	0.00	0.00	0.30	0.00	<u>0.00</u>	0.62
부산	0.00	0.00	0.57	0.00	<u>0.00</u>	1.10
대구	0.00	0.00	0.41	0.00	<u>0.00</u>	0.72
인천	0.00	0.00	0.67	0.00	<u>0.00</u>	0.58
광주	0.00	0.00	1.03	0.00	<u>0.00</u>	1.14
대전	0.00	0.00	0.59	0.00	<u>0.00</u>	0.82
울산	0.00	0.00	0.75	0.00	<u>0.00</u>	0.70
경기	0.05	0.03	1.08	0.00	<u>0.00</u>	1.35
강원	0.24	0.05	2.77	0.00	<u>0.00</u>	2.86
충북	0.00	0.00	2.00	0.00	<u>0.00</u>	2.11
충남	0.22	0.10	3.08	0.00	<u>0.00</u>	2.92
전북	0.84	0.35	2.55	0.00	<u>0.00</u>	2.97
전남	0.60	0.10	1.47	0.04	0.04	2.24
경북	0.79	0.08	1.73	0.02	0.02	2.43
경남	0.41	0.08	2.49	0.00	<u>0.00</u>	2.37
제주	0.00	0.00	0.53	0.00	<u>0.00</u>	1.29
전국	0.32	0.07	1.73	0.01	0.01	2.03

※ 0.00 : 불일치 정도가 실제로 0은 아니고 소수점 때문에 0.00으로 표현



다. 성×연령그룹

meth2005는 광역시 지역에서는 시군구·동읍면부 수준에서 전수 인구나 표본 추정 인구가 일치하였고, 도(道) 지역에서는 다소 불일치하는 경우가 나타났다. meth1은 불일치 정도가 0.04로 meth2005에 비해서는 많이 작았다. 반면, meth3과 meth4는 다른 추정량에 비해 불일치 정도가 각각 6.78, 6.72로 높았다. 이는 meth2005, meth1과 meth5는 성×연령그룹을 사후층으로 한 사후가중치를 사용하였는데, 이런 경우 사후가중치는 사후층(성×연령그룹)의 모집단 인구에 대해서는 정확한 추정이 가능하기 때문이다. 반면, meth3과 meth4는 성, 연령그룹 각각에 대한 1차원 주변분포만 추정에 이용하였기 때문에, 성×연령그룹의 교차표에서는 일치성이 떨어질 수밖에 없었다. meth3과 meth4에서 성×연령그룹에 대해 일치성을 높이기 위해서는 레이킹 추정에 성×연령그룹의 2차원 주변 분포를 이용하는 것을 고려할 수 있다.

〈표 1-19〉 시도·시군구·동읍면부성×연령그룹 평균 불일치 정도

시도	meth2005	meth1	meth2	meth3	meth4	meth5
서울	0.00	0.00	0.32	2.67	2.66	0.63
부산	0.00	0.00	0.59	3.99	4.08	1.07
대구	0.00	0.00	0.42	4.03	4.03	0.72
인천	0.00	0.00	0.70	4.24	4.22	0.67
광주	0.00	0.00	1.08	2.82	2.80	1.19
대전	0.00	0.00	0.61	2.54	2.39	0.84
울산	0.00	0.00	0.77	4.30	4.25	0.79
경기	0.05	0.02	1.12	4.83	4.82	1.45
강원	0.30	0.02	2.96	8.42	8.39	3.15
충북	0.00	0.00	2.12	7.22	7.15	2.31
충남	0.25	0.07	3.20	7.38	7.21	3.25
전북	0.63	0.16	2.58	8.55	8.52	3.09
전남	0.65	0.08	1.52	8.53	8.40	2.38
경북	0.85	0.04	1.82	8.70	8.62	2.68
경남	0.47	0.02	2.61	7.71	7.69	2.69
제주	0.00	0.00	0.55	6.73	6.79	1.27
전국	0.33	0.04	1.81	6.78	6.72	2.20

라. 혼인상태

mth2005, mth1, mth2, mth3은 혼인상태에 대해서 전국적으로 모두 불일치 정도가 5를 넘겼는데, mth4와 mth5는 불일치 정도가 4보다 작게 나타났다. 왜냐하면 mth2005, mth1, mth2, mth3은 추정에 혼인상태 변수를 사용하지 않았지만, mth4와 mth5는 추정 과정에서 혼인상태를 주요 변수로 사용했기 때문이다. 또한, 광역시에서의 추정 결과가 도 단위 지역들에서 보다는 혼인상태에 대해 좋은 일치성을 보여주고 있다. 결과적으로 mth4와 mth5는 혼인상태에 따른 전수와 추정 인구 사이의 불일치 정도를 크게 개선시켰다.

〈표 1-20〉 시도·시군구·동읍면부·혼인상태 평균 불일치 정도

시도	mth2005	mth1	mth2	mth3	mth4	mth5
서울	2.29	2.30	2.63	2.64	1.75	1.77
부산	2.67	2.67	2.61	2.65	1.30	1.27
대구	2.81	2.82	2.93	2.90	1.93	1.92
인천	3.04	2.94	3.10	2.91	2.27	2.26
광주	3.68	3.66	4.04	3.98	2.55	2.50
대전	2.87	2.90	3.25	3.57	2.49	2.42
울산	1.72	1.74	2.19	2.04	1.18	1.19
경기	4.25	4.27	4.47	4.54	2.77	2.74
강원	6.69	6.74	7.14	6.74	4.00	3.90
충북	6.02	6.00	6.17	6.17	4.18	4.12
충남	6.43	6.42	6.48	6.26	3.93	3.80
전북	7.50	7.44	8.02	7.92	5.84	5.62
전남	6.76	6.71	6.73	6.82	4.11	4.10
경북	7.16	7.10	7.38	7.62	4.95	4.80
경남	5.75	5.75	6.38	5.88	3.79	3.84
제주	4.21	4.21	4.09	4.08	2.36	2.21
전국	5.52	5.50	5.77	5.71	3.65	3.59



마. 교육정도

교육정도에서는 mth2005, mth1~mth5 모두 시도·시군구·동읍면부 수준에서 전수 인가와 추정 인구가 일치하지 않았다. 전체적인 불일치 정도도 42 이상으로 서로 우열을 가리기 힘든 수준이다. 이는 혼인상태와는 달리 교육정도 변수가 각 방안들의 추정 과정에 포함되지 않았기 때문이다. 그럼에도 전국적으로는 mth3, mth4와 mth5가 mth2005보다는 다소 개선된 결과를 보여준다.

〈표 1-21〉 시도·시군구·동읍면부·교육정도 평균 불일치 정도

시도	mth2005	mth1	mth2	mth3	mth4	mth5
서울	19.01	19.04	19.21	19.09	19.12	19.14
부산	34.89	34.94	35.06	35.40	35.18	34.72
대구	32.70	32.80	32.72	32.58	32.57	32.78
인천	31.86	31.72	31.59	31.41	31.11	31.10
광주	28.48	28.54	29.04	29.05	29.18	29.01
대전	32.46	31.99	32.36	32.24	32.40	32.58
울산	31.53	31.74	31.35	31.49	31.82	31.51
경기	32.79	32.80	32.82	32.65	32.64	32.71
강원	47.31	47.33	46.63	46.50	46.47	45.83
충북	51.75	51.82	50.94	49.99	50.14	51.14
충남	52.08	51.66	51.62	50.96	51.46	51.65
전북	46.28	46.27	45.97	45.77	45.97	45.79
전남	53.36	53.44	54.66	54.86	54.58	54.42
경북	50.73	51.10	50.77	50.73	51.13	50.81
경남	46.99	47.26	47.59	46.46	46.40	47.51
제주	50.03	50.72	50.76	49.01	49.61	50.22
전국	42.49	42.55	42.55	42.26	42.34	42.41

바. 조사구특성

조사구특성별 평균 불일치 정도는 mth2~mth5가 3.63으로 mth2005의 13.96과 mth1의 13.94에 비해 크게 개선되었다. 이는 혼인상태와 마찬가지로 추정 과정에 조사구특성을 활용했는지 여부에 따른 것이다. 일부 읍면부 지역에서 표집되지 않은 아파트조사구를 보통·섬조사구와 병합하였기 때문에 mth2~mth5도 완전히 일치되는 가중치를 산정할 수는 없었다. 조사구특성은 공표 항목은 아니지만, 아파트, 보통, 섬과 같은 주거 환경이 표본조사 항목에 유의미한 영향을 미치기 때문에 이를 가중치 산정에 반영한다면 표본조사 결과의 설명력이 보다 높아질 것으로 기대된다.

〈표 1-22〉 시도·시군구·동읍면부·조사구특성 평균 불일치 정도

시도	mth2005	mth1	mth2	mth3	mth4	mth5
서울	3.47	3.47	0.00	0.00	0.00	0.00
부산	7.40	7.40	2.80	2.80	2.78	2.76
대구	2.71	2.70	0.00	0.00	0.00	0.00
인천	16.90	16.86	11.94	12.19	12.20	12.10
광주	5.49	5.49	0.00	0.00	0.00	0.00
대전	3.00	2.99	0.00	0.00	0.00	0.00
울산	2.72	2.71	0.00	0.00	0.00	0.00
경기	12.95	12.95	7.35	7.34	7.31	7.39
강원	12.87	12.89	0.36	0.36	0.36	0.36
충북	8.22	8.21	0.02	0.02	0.02	0.02
충남	21.49	21.48	0.47	0.47	0.46	0.47
전북	26.47	26.42	2.95	2.95	2.96	2.96
전남	21.56	21.55	13.10	13.02	12.99	13.02
경북	12.63	12.54	0.37	0.37	0.37	0.37
경남	11.97	11.98	1.65	1.68	1.67	1.67
제주	23.85	23.82	12.36	12.35	12.35	12.36
전국	13.96	13.94	3.63	3.63	3.62	3.63



사. 가구주와의 관계

가구주와의 관계 항목에서는 mth2005, mth1~mth5 모두 시도·시군구·동읍면부 수준에서 전수 인가와 추정 인구가 일치하지 않았다. 전체적인 불일치 정도도 27 전후로 교육정도와 마찬가지로 우열을 가리기 힘든 수준이다. 마찬가지로 가구주와의 관계 항목도 추정 과정에 포함되지 않았기 때문이다. 그럼에도 전국적으로는 mth3과 mth4가 각각 26.85, 26.79로 mth2005에 비해 다소 개선된 결과를 보여준다.

〈표 1-23〉 시도·시군구·동읍면부·가구주와의 관계 평균 불일치 정도

시도	mth2005	mth1	mth2	mth3	mth4	mth5
서울	13.22	13.26	13.15	13.11	13.17	13.30
부산	16.46	16.55	16.34	16.29	16.59	16.67
대구	15.39	15.34	15.26	15.03	15.42	15.76
인천	15.50	15.49	15.36	15.77	15.71	15.42
광주	13.23	13.10	13.04	12.76	12.53	12.45
대전	14.60	14.65	14.69	14.49	14.62	14.72
울산	13.69	13.51	13.33	13.34	13.36	13.16
경기	19.85	19.92	19.85	19.68	19.80	19.93
강원	35.29	35.15	34.31	33.57	33.11	34.17
충북	32.38	32.33	31.86	31.83	31.76	32.08
충남	31.20	31.28	30.58	30.44	30.37	30.11
전북	33.75	33.50	34.35	33.50	33.31	34.05
전남	38.04	38.58	38.97	38.79	38.75	39.41
경북	30.34	30.70	31.60	31.35	31.33	31.76
경남	28.13	27.92	27.82	28.28	27.96	27.20
제주	23.03	23.15	22.84	23.09	23.18	23.26
전국	26.93	27.00	27.01	26.85	26.79	27.00

그러나 mth4와 mth5는 혼인상태 변수를 가중치 산정에 이용했기 때문에 특별히 가구주와의 관계 중에서 가구주와 가구주의 배우자 관계에 대해서는 보다 정확한 추정이 가

능하였다. <표 1-24>를 보면 mth2005의 불일치 정도가 2.03인데 비해 mth4와 mth5는 1.75와 1.72로 가구주와의 관계를 전체적으로 보았을 때보다 일치도의 개선 정도가 큰 것으로 나타났다.

<표 1-24> 시도·시군구·동읍면부·가구주와의 관계 중 가구주와 가구주의 배우자 평균 불일치 정도

시도	mth2005	mth1	mth2	mth3	mth4	mth5
서울	0.85	0.86	0.95	0.95	0.87	0.89
부산	1.03	1.04	1.06	0.99	1.12	1.10
대구	1.03	1.04	1.14	1.06	0.87	0.93
인천	1.13	1.12	1.15	1.33	1.29	1.22
광주	1.62	1.62	1.51	1.41	1.12	1.29
대전	1.03	1.03	1.31	0.94	0.86	1.09
울산	0.86	0.87	0.95	0.91	0.84	0.78
경기	1.59	1.60	1.69	1.73	1.43	1.40
강원	3.07	3.06	3.24	3.16	2.45	2.62
충북	2.25	2.23	2.34	2.45	2.06	2.05
충남	1.96	1.95	1.84	1.98	1.79	1.73
전북	2.35	2.35	2.60	2.45	2.22	2.03
전남	2.52	2.51	2.62	2.72	1.97	1.94
경북	2.48	2.48	2.55	2.66	2.04	1.94
경남	2.20	2.21	2.43	2.11	1.71	1.75
제주	1.86	1.90	1.96	2.09	2.39	2.00
전국	2.03	2.03	2.13	2.12	1.75	1.72

3. 가중치 방안 적용 결과 분석

제시된 각 방안들의 추정 결과를 전국에 대하여 간단히 요약해보면 <표 1-25>와 같다. 표에서 음영이 들어간 칸은 기존의 mth2005에 비해 전수 인구나 추정 인구의 일치 정도가 개선된 것으로 mth4가 전반적으로 가장 향상된 일치성을 보이는 것을 확인할 수 있다. 특히 연령그룹, 혼인상태, 조사구특성, 가구주와의 관계 항목에서 그 경향이 뚜렷하다. mth4는 성×연령그룹 항목에서는 불일치 정도가 크게 나타났는데, 이는 사후가중



방법이 성별의 하부 연령그룹, 즉, 남성의 10대와 같은 계층 구조에 대해서는 보다 모집단 인구수를 정확하게 추정하기 때문이다. 그러나 이와 같은 사후가중은 추정에 활용 가능한 변수 선택이 수적인 부분에서 극히 제한적이기 때문에 인총과 같은 자료에 적용하기에는 부족함이 있다.

<표 1-25> 특성별 시도·시군구·동읍면부 평균 불일치 정도

특성 항목	mth2005	mth1	mth2	mth3	mth4	mth5
성	0.00	0.00	0.12	0.00	0.00	0.36
연령그룹	0.32	0.07	1.73	0.01	0.01	2.03
성×연령그룹	0.33	0.04	1.81	6.78	6.72	2.20
혼인상태	5.52	5.50	5.77	5.71	3.65	3.59
교육정도	42.49	42.55	42.55	42.26	42.34	42.41
조사구특성	13.96	13.94	3.63	3.63	3.62	3.63
가구주와의 관계	26.93	27.00	27.01	26.85	26.79	27.00
가구주와 가구주의 배우자	2.03	2.03	2.13	2.12	1.75	1.72

※ 1. **0.00** : 불일치 정도가 실제로 0은 아니고 소수점 때문에 0.00으로 표현
 2. 음영이 들어간 칸 : mth2005에 비해 일치 정도가 개선된 추정 방안

제시된 각 방안들의 특징을 요약하여 살펴보면 <표 1-26>과 같다. 사후가중법을 이용한 mth1과 mth2는 추정법이 가장 단순하지만 빈 셀의 발생 빈도가 높고 레이킹에 비해 가중치 최대값이 크다. 레이킹을 이용한 mth3, mth4와 mth5는 추정법이 단순, 사후가중에 비해서는 빈 셀에 덜 민감하고 마지막으로 고려 가능한 변수가 많다는 장점이 있지만, 계산 결과 1보다 작은 가중치가 발생할 수 있다. 사후가중법을 약간 변형한 mth2도 같은 현상이 나타나는데, 이는 조사구특성을 추정 과정에 이용하였기 때문이다. 결과적으로 레이킹은 고려 가능한 특성 항목이 사후가중에 비해 많기 때문에 보다 많은 항목에서 전수 인구수와 표본 추정 인구 사이의 일치도를 높일 수 있다는 것이 최대 장점이다.

그러나 5가지 방법 모두 최대 가중치를 자동적으로 제어하는 데는 한계가 있었으며 심지어 사후적인 가중치 보정 과정에서 더 큰 가중치가 발생하는 경우도 있었다. 최대 가중치를 제어하기 위해서는 추정 결과에 따라 최대 가중치가 발생하는 일부 항목(셀)에 대해서 셀 병합을 하는 등의 사후 제어를 할 수 있을 것이다.

〈표 1-26〉 추정 방안별 특징

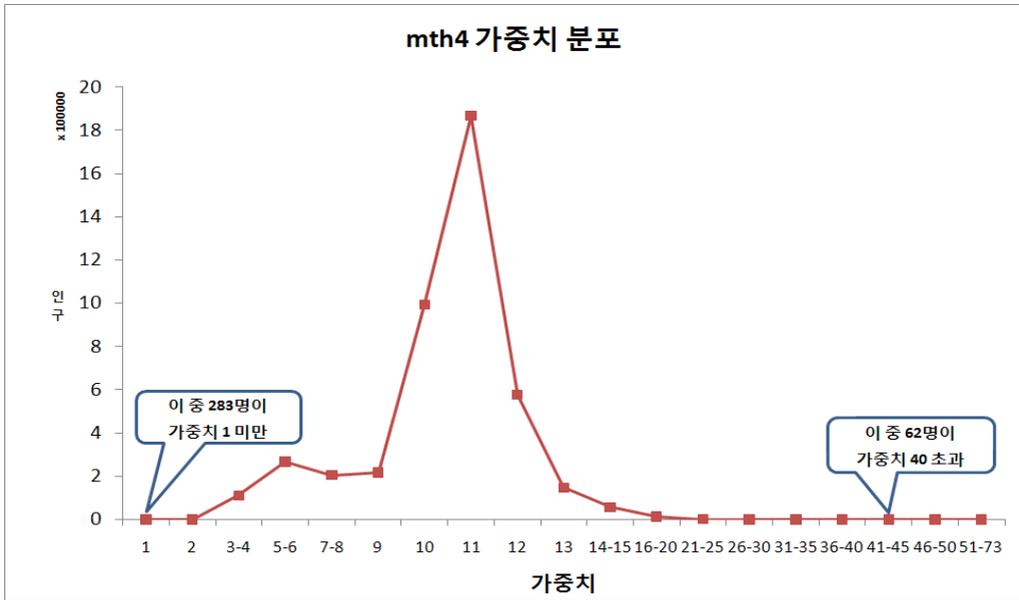
방안	장점	단점	평균 반복 횟수
mth1	○ 추정법이 가장 단순 ○ 성, 연령그룹 고려 ○ 성·연령그룹 모집단 인구 정확히 추정	○ 이용 가능한 변수가 제한 ○ 빈 셀 발생 빈도 높음 ○ 레이킹에 비해 가중치 최대값이 큼	
mth2	○ 추정법이 가장 단순 ○ 성·연령그룹·조사구특성 고려	○ 빈 셀 발생 빈도 높음 ○ 레이킹에 비해 가중치 최대값이 큼 ○ 1보다 작은 가중치 발생 ○ 성·연령그룹의 기본 항목에서 불일치 정도가 높음	
mth3	○ 추정법이 단순 ○ 성·연령그룹·조사구특성 고려 ○ 사후가중보다 빈 셀에 덜 민감	○ 사후가중에 비해 가중치 최대값이 작음 ○ 1보다 작은 가중치 발생	5
mth4	○ 추정법이 단순 ○ 성·연령그룹·조사구특성·혼인상태 고려 ○ 사후가중보다 빈 셀에 덜 민감	"	15
mth5	○ 추정법이 단순 ○ 성·연령그룹·조사구특성·혼인상태 고려	○ 1보다 작은 가중치 발생 ○ 성·연령그룹의 기본 항목에서도 불일치 정도가 높음	5

4. 가중치가 1 미만이거나 40 초과인 경우 가중치 재산정

제시된 방안들 중 가장 좋은 추정 결과를 보인 mth4의 가중치 결과를 상세히 분석하고, 1 미만의 가중치와 40을 초과하는 가중치의 처리 방안에 대하여 검토한다.

mth4의 가중치 분포를 보면 [그림 1-7]과 같다. 총 4,462,378명 중 345명에 대한 가중치가 가중치 부여 원칙에 어긋나는 값인 1 미만(283명)⁹⁾이거나 40을 초과(62명)한다. 이런 가중치를 부여받은 가구원이 전체 인구에서 차지하는 비중은 상당히 작지만, 소지역이나 소영역에서의 추정에는 큰 영향을 미친다. 그러므로 이들 가중치를 가지는 가구원에 대한 가중치를 재산정해서 그 영향을 줄일 필요가 있다.

9) 가중치를 정수화하면 이 값들은 모두 1이 됨.



[그림 1-7] mth4 가중치 분포

345명에 대한 특성을 분석한 결과, <표 1-27>처럼 크게 11개 사례로 나타난다. 가중치가 1 미만인 경우(①~⑤)는 동읍면부에서 해당 조사구특성의 모집단이 차지하는 비중 에 비해 표본이 많이 뽑힌 경우이다. 40을 초과한 경우는 해당 조사구특성의 모집단이 동읍면부에서 차지하는 비중 에 비해 표본이 적게 뽑히거나(⑦, ⑨, ⑪), 특정 특성 항목 (연령그룹이나 혼인상태 등)이 너무 적게 뽑힌 경우(⑥, ⑧, ⑩)이다.

이들 지역에 대해서는 각각 추정에 조사구특성을 활용하지 않거나 연령그룹을 인근 연령그룹과 병합¹⁰⁾하여 가중치 산정에 mth4 방법을 다시 적용하였다. 적용 결과, 1 미만의 가중치는 1보다 커졌고, 40을 초과하던 가중치도 대체로 40이하로 잘 제어되었다.

<표 1-27> 가중치가 1 미만이거나 40을 초과한 경우

가중치	사례 번호	시도	시군구	동읍면부	조사구특성	인구	처리
1 미만	①	충남	계룡시	동부	보통·섬	18	추정에 조사구특성 활용 안함
	②	전북	남원시	면부	아파트	83	
	③	"	무주군	"	"	123	
	④	"	임실군	읍부	"	4	
	⑤	전남	구례군	면부	"	55	

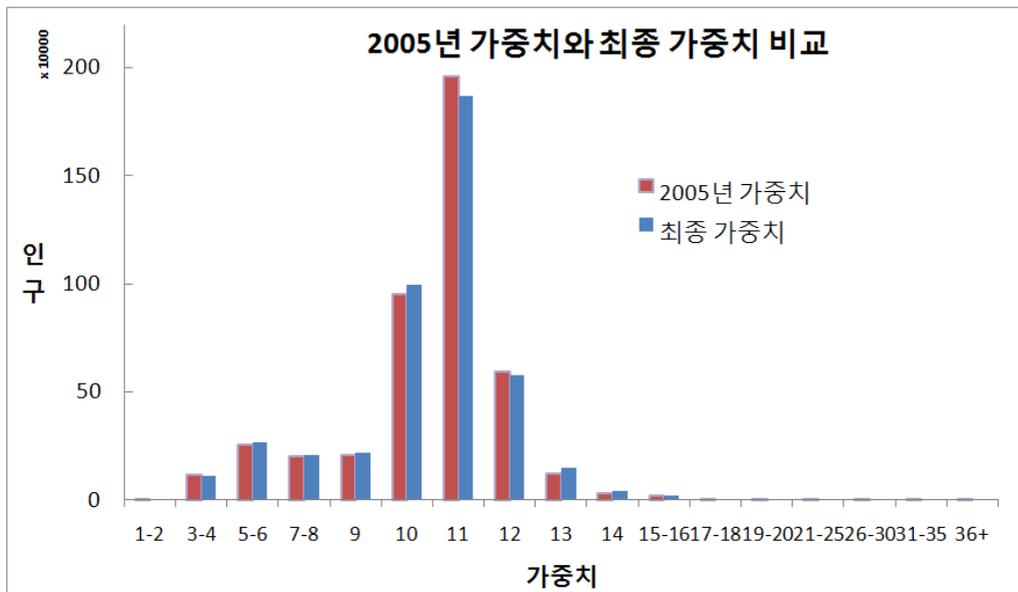
10) 우선 조사구특성을 활용하지 않는 방법을 적용하고 가중치 보정 효과가 크지 않은 경우, 다음으로 연령 그룹 병합 적용

가중치	사례 번호	시도	시군구	동읍면부	조사구특성	인구	처리
40 초과	⑥	강원	원주시	읍부	보통·섬	1	연령그룹 21→20으로 병합
	⑦	"	고성군	"	아파트	47	조사구특성 활용 안함
	⑧	충남	공주시	"	보통·섬	1	연령그룹 9→10으로 병합
	⑨	"	아산시	"	아파트	2	조사구특성 활용 안함
	⑩	전남	진도군	"	보통·섬	3	연령그룹 2→3으로 병합
	⑪	경북	상주시	"	"	8	조사구특성 활용 안함

최종적으로 산정된 가중치 결과와 2005년 가중치 결과의 분포를 살펴보면 <표 1-28> 과 [그림 1-8]과 같다. 최종 가중치는 최소값이 1.4, 최대값이 40.65이고 표준편차는 2.04로 2005년에 비하여 가중치가 가지는 값의 범위가 조금 늘어났으며, 최소값과 최대값 사이에서 보다 고르게 분포한다.

<표 1-28> 2005년 가중치와 최종 가중치 분포

방안	평균값	최소값	최대값	표준편차
2005년 가중치	10.27	2.0	40.00	2.02
최종 가중치	10.27	1.4	40.65	2.04



[그림 1-8] 2005년 가중치와 최종 가중치 비교



주요 특성에 대하여 2005년 가중치와 최종 가중치를 이용한 표본 인구가 전수 인구와 어느 정도 일치되었는지 <표 1-29>와 그림으로 확인해 보기로 한다. 수치상의 결과는 <표 1-17> ~ <표 1-24>와 크게 다르지 않지만, 조사구특성과 연령그룹에서 일부 병합이 있었기 때문에 해당 특성에서 일치성이 다소 떨어질 수 있다. 주요 특성은 성, 연령그룹, 성×연령그룹, 교육정도, 혼인상태, 조사구특성, 가구주와의 관계를 이용한다. 비교를 위한 최소 단위는 가중치 지역(시도·시군구·동읍면부)이다.

<표 1-29> 2005년 가중치, mth4와 최종 가중치의 특성별 시도·시군구·동읍면부 평균 불일치 정도

특성 항목	2005년 가중치	mth4	최종 가중치
성	0.00	<u>0.00</u>	<u>0.00</u>
연령그룹	0.32	0.01	0.03
성×연령그룹	0.33	6.72	6.69
혼인상태	5.52	3.65	3.63
교육정도	42.49	42.34	42.22
조사구특성	13.96	3.62	6.25
가구주와의 관계	26.93	26.79	26.84
가구주와 가구주의 배우자	2.03	1.75	1.74

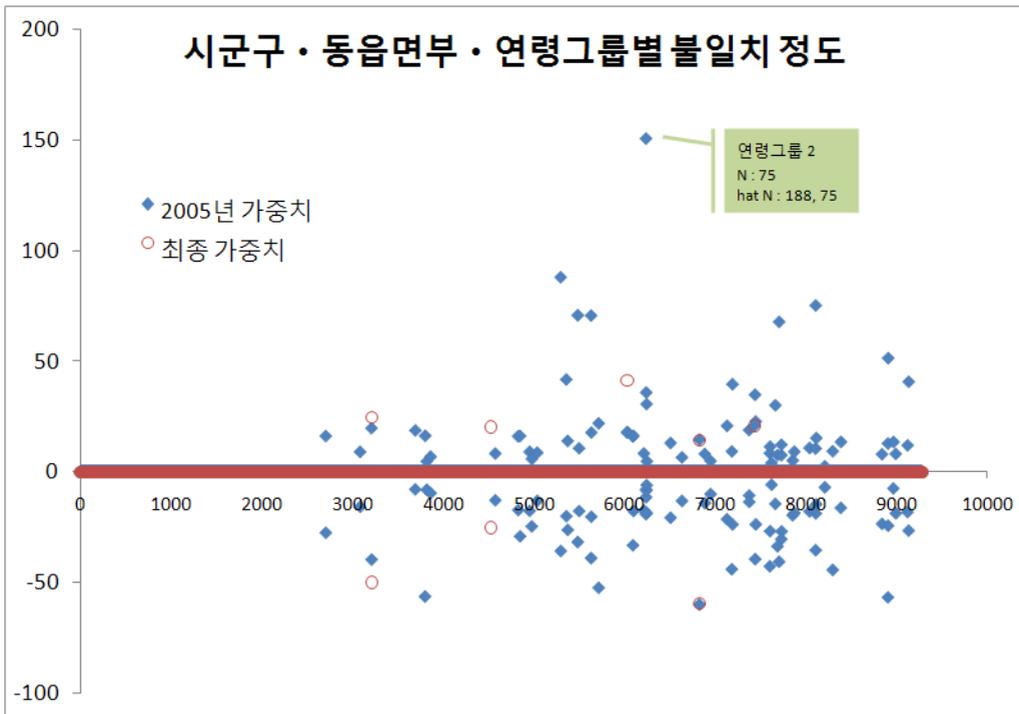
- ※ 1. 0.00 : 불일치 정도가 실제로 0은 아니고 소수점 때문에 0.00으로 표현
 2. 음영이 들어간 칸 : mth2005에 비해 일치 정도가 개선된 추정 방안

가중치가 1 미만이거나 40을 초과하는 가중치를 다시 산정한 결과 연령그룹과 조사구특성에서는 불일치 정도가 다소 상승하였지만, 성×연령그룹, 혼인상태, 교육정도, 가구주와의 관계 등에서는 불일치 정도가 오히려 감소하였다. 이는 소지역에서 과대 가중치가 가지는 영향력이 제어됐기 때문이다.

최종적으로 산정된 최종 가중치가 주요 특성에 대하여 전수 인구와 추정 인구를 어느 정도 일치시켰는지 시각적으로 확인해본다. 우선, 성에 대한 불일치 정도는 <표 1-17>에 나타난 바와 같이 서울의 노원구에서 발생한 1명의 차이를 빼고는 모두 0이다 (따라서, 성별 불일치 정도에 대한 그림은 생략).

가. 연령그룹

math4에 의해 산정된 1 미만과 40 초과 가중치를 재산정한 최종 가중치와 2005년 가중치의 시군구·동읍면부·연령그룹별 불일치 정도를 살펴보면 [그림 1-9]와 같다. 시도·시군구·동읍면부에서 연령그룹에 대해 최종 가중치는 불일치 정도가 거의 0에 가깝지만 2005년 가중치는 그보다는 조금 더 많이 0 이외의 값을 갖는 것으로 나타났다. 즉, 평균 불일치 정도가 2005년에는 0.32였는데, 최종 가중치는 0.03으로 연령그룹에 대한 불일치 정도를 크게 개선시켰다. 불일치 정도의 경향을 알아보기 위해서 실제 추정 결과를 하나 살펴보면, 어떤 지역의 연령그룹 2에 대해 모집단 인구는 75명인데 2005년 가중치는 188명(불일치 정도는 150.7%), 최종 가중치는 75명(불일치 정도는 0.0%)으로 추정하였다.

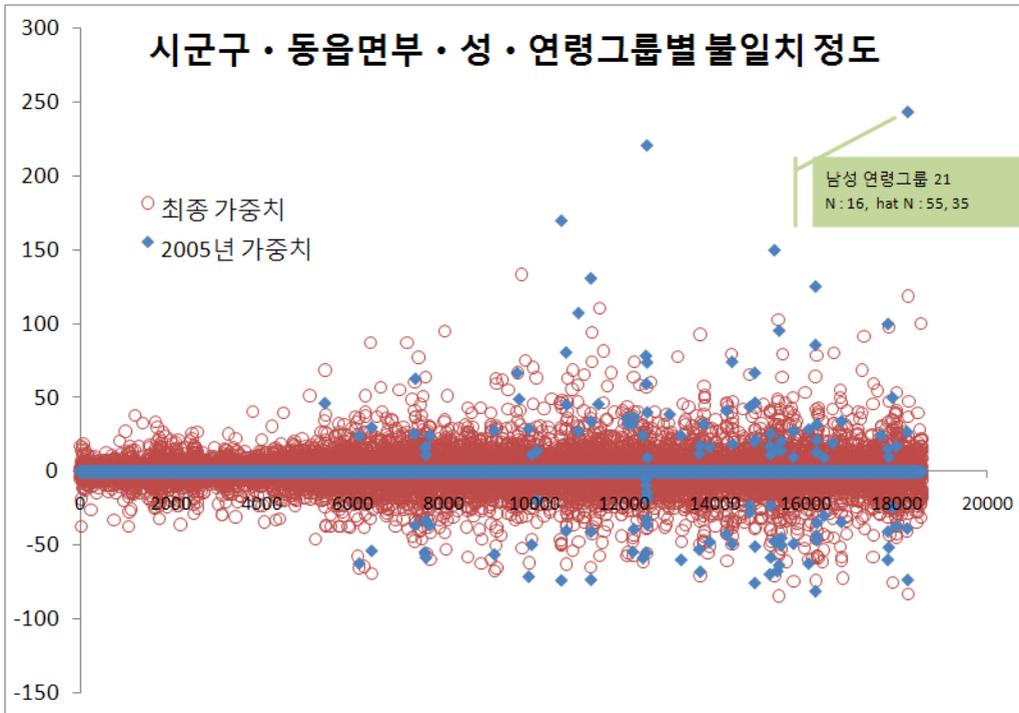


[그림 1-9] 2005년 가중치와 최종 가중치의 연령그룹별 불일치 정도 산점도



나. 성×연령그룹

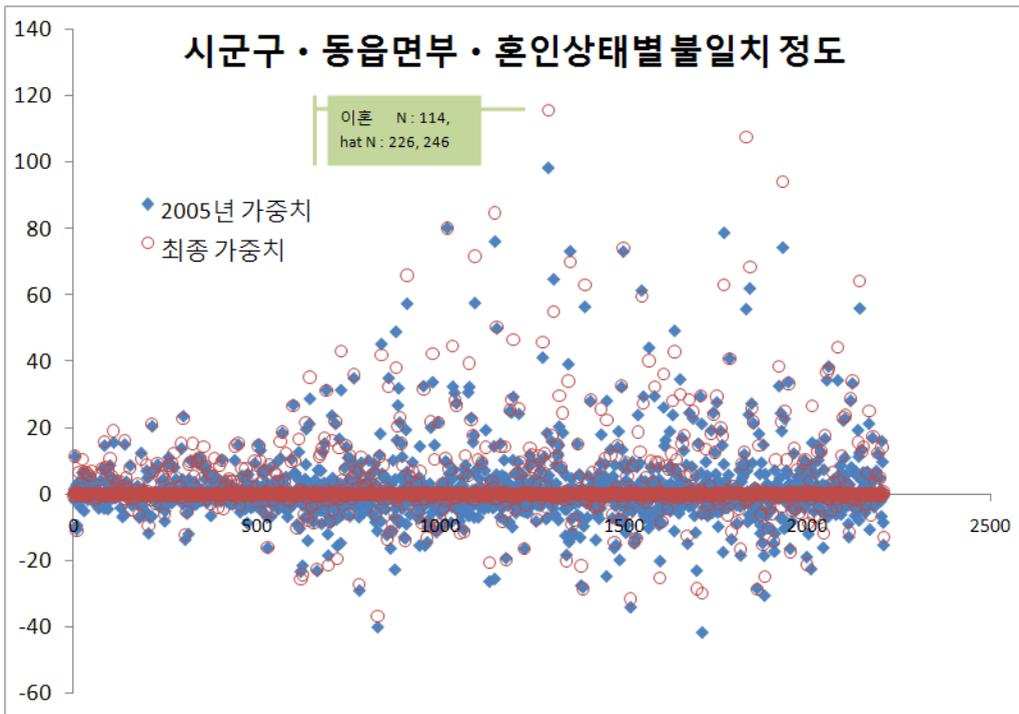
최종 가중치와 2005년 가중치의 시군구·동읍면부·성×연령그룹별 불일치 정도를 살펴보면 [그림 1-10]과 같다. 시도·시군구·동읍면부에서 성×연령그룹에 대해 2005년 가중치는 불일치 정도가 거의 0에 가깝지만 최종 가중치는 더 많이 0 이외의 값을 갖는 것으로 나타났다. 즉 평균 불일치 정도가 2005년에는 0.33이었는데, 최종 가중치는 6.69로 성×연령그룹에 대한 불일치 정도를 개선시키지 못했다. 왜냐하면, 앞에서 언급한 대로 성×연령그룹에 대한 사후가중을 이용한 2005년 가중치는 성×연령그룹에 대해 모집단을 거의 정확하게 추정하지만, 최종 가중치는 성과 연령그룹의 1차원 주변분포를 이용하였기 때문에 성×연령그룹의 층 모집단 인구를 정확하게 추정할 수 없었다. 성×연령그룹에 대해서 불일치 정도가 가장 컸던 경우는 어떤 지역의 남성의 연령그룹 21에 대해 모집단 인구는 16명인데 2005년 가중치는 55명(불일치 정도는 243.8%), 최종 가중치는 35명(불일치 정도는 118.8%)으로 추정한 경우다.



[그림 1-10] 2005년 가중치와 최종 가중치의 성×연령그룹별 불일치 정도 산점도

다. 혼인상태

최종 가중치와 2005년 가중치의 시군구·동읍면부·혼인상태별 불일치 정도를 살펴 보면 [그림 1-11]과 같다. 시도·시군구·동읍면부에서 혼인상태에 대해 최종 가중치는 불일치 정도가 거의 0에 가깝지만 2005년 가중치는 더 많이 0 이외의 값을 갖는 것으로 나타났다. 즉, 평균 불일치 정도가 2005년에는 5.52였는데, 최종 가중치는 3.63으로 혼인 상태에 대한 불일치 정도를 크게 개선시켰다. 최종 가중치는 2005년 가중치와 달리 혼인 상태를 추정 과정에서 전수와 추정 인구가 최대한 일치되도록 레이킹 변수로 사용했기 때문이다. 혼인상태에 대해서 불일치 정도가 가장 컸던 경우는 어떤 지역의 이혼한 경우에 대해 모집단 인구는 114명인데 2005년 가중치는 226명(불일치 정도는 98.2%), 최종 가중치는 246명(불일치 정도는 115.8%)으로 추정한 경우이다.

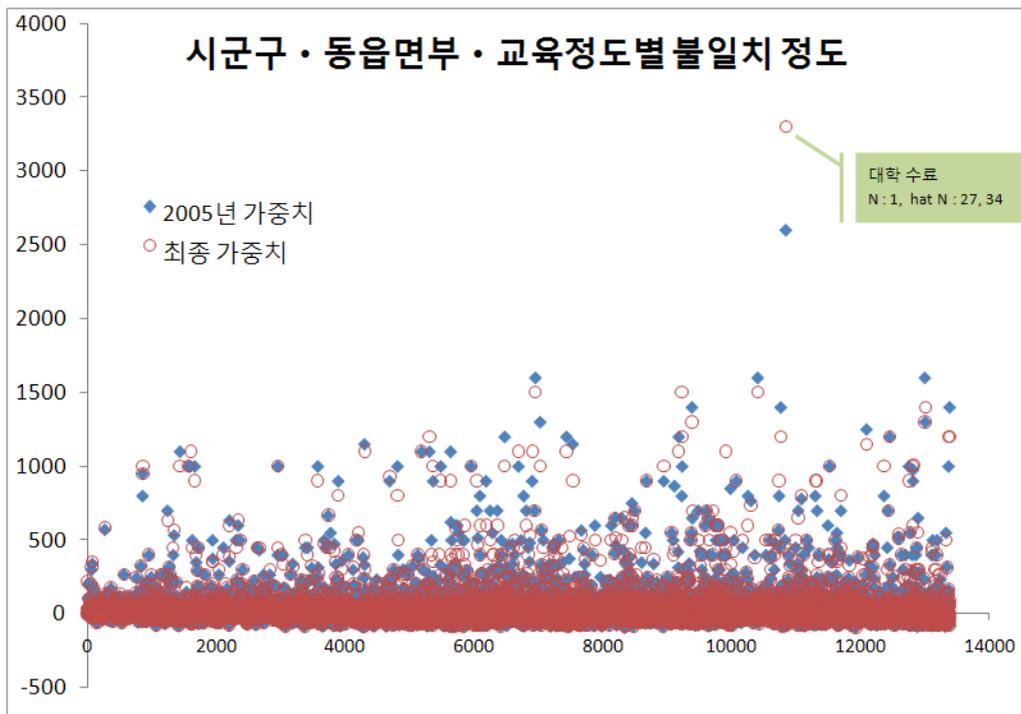


[그림 1-11] 2005년 가중치와 최종 가중치의 혼인상태별 불일치 정도 산점도



라. 교육정도

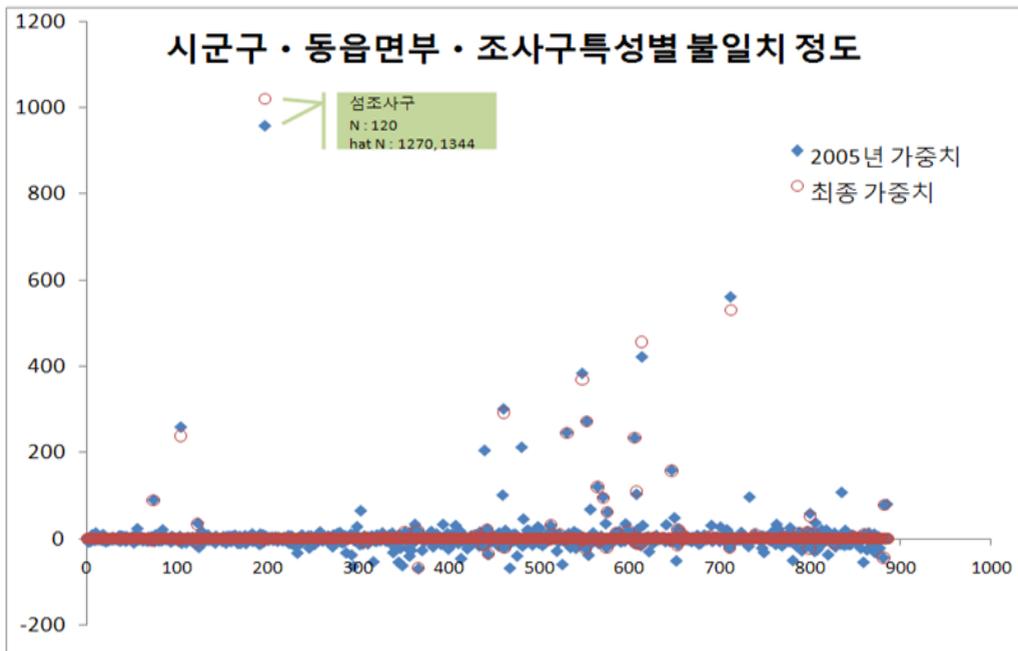
최종 가중치와 2005년 가중치의 시군구·동읍면부·교육정도별 불일치 정도를 살펴 보면 [그림 1-12]와 같다. 시도·시군구·동읍면부에서 교육정도에 대해서는 최종 가중치와 2005년 가중치가 0 이외의 값을 많이 갖는 것으로 나타났다. 즉, 평균 불일치 정도가 2005년에는 42.49였는데, 최종 가중치는 42.22로 교육정도에 대한 불일치 정도를 크게 개선시키지는 못했다. 왜냐하면, 최종 가중치와 2005년 가중치 모두 교육정도를 추정 에 이용하지 않았기 때문이다. 교육정도에 대해서 불일치 정도가 가장 컸던 경우는 어떤 지역의 대학을 수료한 1명의 모집단 인구를 2005년 가중치는 27명(불일치 정도는 2,600%), 최종 가중치는 34명(불일치 정도는 3,300%)으로 추정한 경우이다.



[그림 1-12] 2005년 가중치와 최종 가중치의 교육정도별 불일치 정도 산점도

마. 조사구특성

최종 가중치와 2005년 가중치의 시군구·동읍면부·조사구특성별 불일치 정도를 살펴보면 [그림 1-13]과 같다. 시도·시군구·동읍면부에서 조사구특성에 대해 최종 가중치는 불일치 정도가 거의 0에 가깝지만 2005년 가중치는 더 많이 0 이외의 값을 갖는 것으로 나타났다. 즉, 평균 불일치 정도가 2005년에는 13.96이었는데, 최종 가중치는 6.25로 조사구특성에 대한 불일치 정도를 크게 개선시켰다. 최종 가중치는 2005년 가중치와 달리 조사구특성을 추정 과정에서 전수와 추정 인구가 최대한 일치되도록 레이킹 변수로 사용했기 때문이다. 조사구특성에 대해서 불일치 정도가 가장 컸던 경우는 어떤 지역의 섬조사구의 모집단 인구 120명을 2005년 가중치는 1,270명(불일치 정도는 958.3%), 최종 가중치는 1,344명(불일치 정도는 1,020.0%)으로 추정한 경우이다.

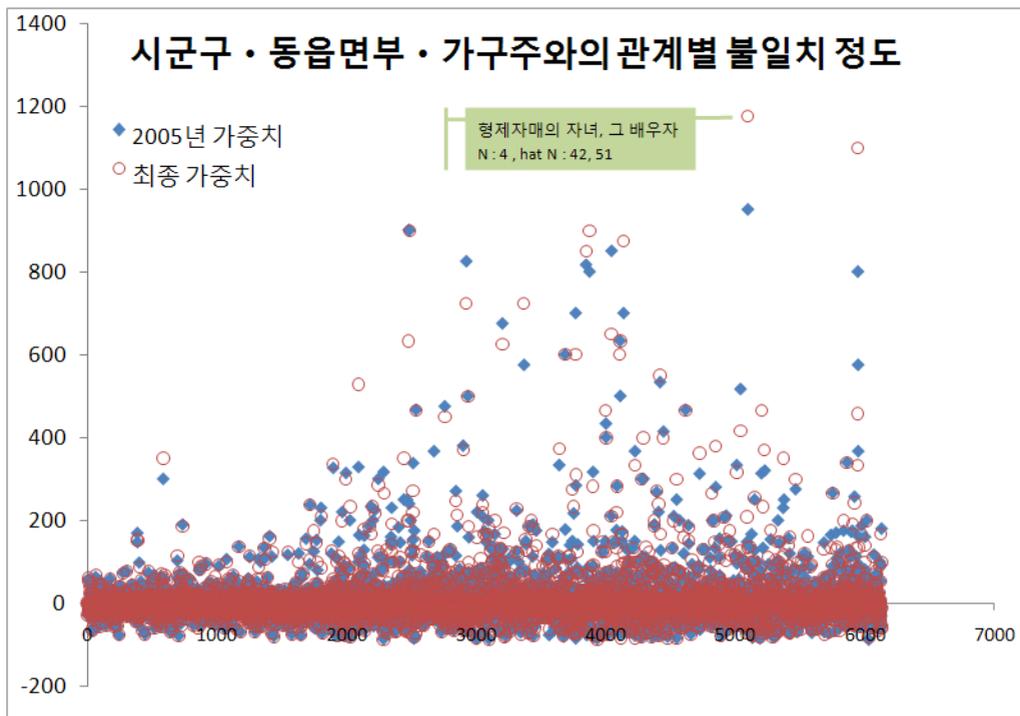


[그림 1-13] 2005년 가중치와 최종 가중치의 조사구특성별 불일치 정도 산점도



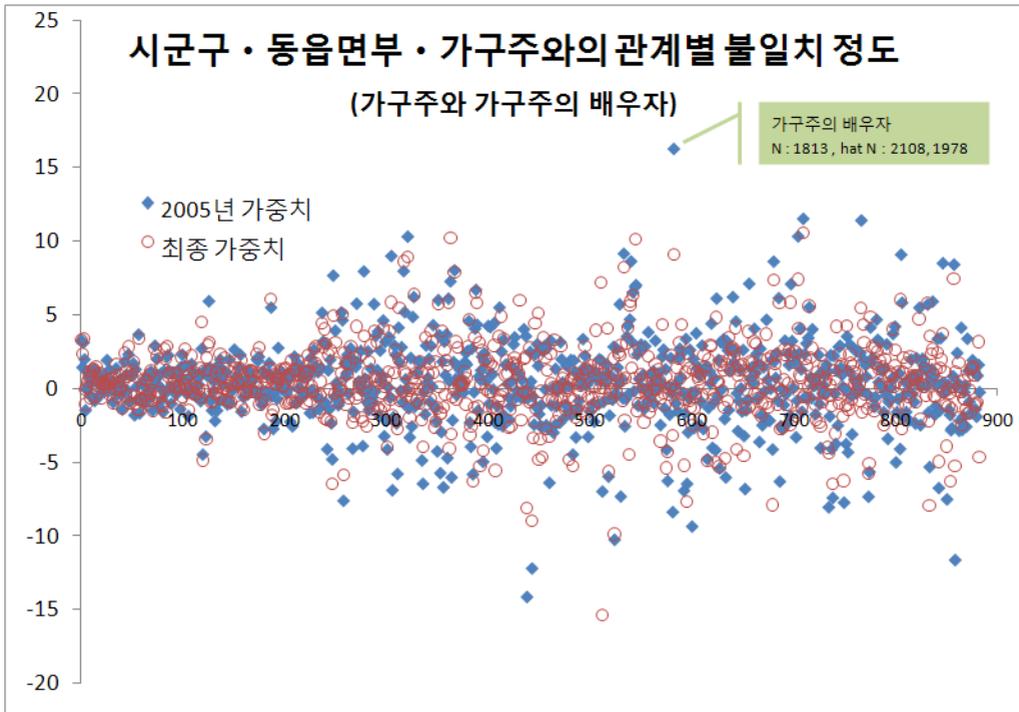
바. 가구주와의 관계

최종 가중치와 2005년 가중치의 시군구·동읍면부·가구주와의 관계별 불일치 정도를 살펴보면 [그림 1-14]와 같다. 시도·시군구·동읍면부에서 가구주와의 관계에 대해 최종 가중치와 2005년 가중치의 불일치 정도가 대체로 0과 100 사이에 위치한다. 즉, 평균 불일치 정도가 2005년에는 26.93이었는데, 최종 가중치는 26.84로 가구주와의 관계에 대한 불일치 정도를 크게 개선시키지는 못했다. 왜냐하면, 최종 가중치와 2005년 가중치 모두 가구주와의 관계를 추정에 이용하지 않았기 때문이다. 가구주와의 관계에 대해서 불일치 정도가 가장 컸던 경우를 살펴보면, 형제자매의 자녀, 그 배우자 항목에 대해서 모집단 인구 4명을 2005년 가중치는 42명(불일치 정도는 950%), 최종 가중치는 51명(불일치 정도는 1,175%)으로 추정하였다.



[그림 1-14] 2005년 가중치와 최종 가중치의 가구주와의 관계별 불일치 정도 산점도

그러나 최종 가중치는 혼인상태 변수를 가중치 산정에 이용했기 때문에 특별히 가구주와의 관계 중에서 가구주와 가구주의 배우자에 대해서는 이전에 비해 정확한 추정이 가능하였다. <표 1-29>를 보면 2005년 가중치의 불일치도가 2.03인데 비해 최종 가중치는 1.74로 일치도가 크게 개선된 것으로 나타났다.



[그림 1-15] 2005년 가중치와 최종 가중치의 가구주와의 관계 중 가구주와 가구주의 배우자 불일치 정도 산점도

제5절 결론

지금까지 2005년 인총 자료를 이용하여 2010년 가구원 가중치 산정 방안을 검토해 보았다. 실험 결과 모두 5가지의 산정 방안 중 방안 4(mth4)가 가장 좋은 결과를 보여주었다. 방안 4는 성, 연령그룹, 조사구특성, 혼인상태 등 4개 변수의 1차원 주변 분포를 이용하여 레이킹 추정한 방법이다. 즉, 레이킹 추정을 이용함으로써 특성 항목의 주변 분포에 표본 추정 결과가 가장 잘 맞도록 가중치를 산정했다. 특히 기존에 고려하지 않았던 조사구특성과 혼인상태 변수가 추가되어 그 정도가 향상되었다. 더불어 가구주와



가구주의 배우자 항목에 대한 추정 성능도 함께 향상되었다. 반면에 성×연령그룹의 교차 항목에 대한 일치성은 다소 떨어졌다. 이는 레이킹 추정 시 성과 연령그룹의 1차원 주변 분포를 각각 이용하기 보다는 성×연령그룹의 2차원 주변 분포를 이용함으로써 개선시킬 수 있을 것이다.

세부적으로는 소수점을 유지하면서 가중치를 산정했기 때문에 1보다 작은 가중치가 발생하는 경우가 있었고 최대 가중치를 40 근처로 완벽히 제어하지는 못했다. 그래서 이들 가중치를 처리해주기 위해 본 연구에서는 셀 병합을 다시 이용하였다. 최대한 공표 항목이 아닌 조사구특성을 병합해주고, 조사구특성 병합이 효과가 없는 경우 연령그룹을 병합하는 순서로 가중치를 재산정하였다. 그리고 소수점이 있는 가중치를 사용할 것인지 정수로 처리된 가중치를 사용할 것인지 여부는 이후로도 지속적인 논의의 대상이 될 것으로 보인다. 이와 관련해서는 가중치 산정 전문가, 실무 담당자들 사이에서 끊임 없는 의견 교환이 필요하다.

가중치 지역은 시도·시군구·동읍면부로 설정하였는데, 분석 결과 광역시도와 같이 인구 밀집도가 높은 경우에는 보다 작은 단위, 즉 각 동읍면에 대해서도 추정이 가능할 것이라 판단된다. 따라서 가중치 지역의 설정을 일률적으로 하지 않고 광역시도의 특성에 따라서 조정한다면 보다 일치성 높은 가중치를 산정할 수 있을 것으로 기대된다.

본 연구에서는 초기에 목표했던 가중치 산정 과정의 상세 기록 유지, 가중치 산정 과정의 자동화에 충실하려고 노력하였으나 완전할 수는 없었다. 그러나 최대한 가중치 산정 과정에서 고려해야 하는 다양한 상황, 즉, 추정 방법, 변수의 선택, 셀 병합 등에 대해서 실제 자료를 활용하여 이해하기 쉽게 기술하였다. 이를 통해, 이후 연구의 기초 자료나 인총 가중치 산정의 매뉴얼로써 본 연구 결과가 활용될 것으로 기대한다.

더불어 2010년 인총은 2005년 인총 결과와는 상당히 다를 수 있다. 빠르게 변화하는 사회 환경속에서 5년은 짧지 않은 기간이기 때문이다. 따라서, 2010년 인총 표본조사를 위한 가중치 작업 산정을 위해서 본 연구 결과를 직접 적용하기보다는 충분한 사전 검토 후 적용하는 것이 합리적인 것이다.

참고문헌

- 임경은 (2010), “두 통계자료를 이용한 병합 추정 방법 연구”, 통계개발원.
- 조태휘 (2005), “표본수가 0인 칸을 포함하는 분할표에서 래킹비 조정법의 적용”, 서울시립대학교 학위논문(석사).
- 통계청 (2007), 2005 인총보고서.
- 허명희, 윤영아, 김규성 (2005). 2차원 기준 반복비례가중법 연구, 통계연구, 10(1), 1-22.
- Deming, W.E. and Stephan, F.F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *Annals of Mathematical Statistics*, 11, 427-444.
- Fellegi, I.P. (1964). Response Variance and its Estimation. *Journal of the American Statistical Association*. 59, December: 1016-1041.
- Hansen, M.H., W.N. Hurwitz and M.A. Bershada. (1959). Measurement Errors in Censuses and Surveys. *Bulletin of the International Statistical Institute*. 38, 359 - 374.
- Ireland, C.T. and Kullback, S. (1968). *Contingency tables with given marginals*. *Biometrika*, 55, 179-188.
- Thompson, J.H. (1981). Convergence properties of the iterative 1980 census estimator. *American Statistical Association Proceedings, Section on Survey Research Methods*, 182-185.
- Woltman, H., Miskura, S., Thompson and J., Bounpane, P. (1981). 1980 Census Weighting and Variance Estimation Studies, Design and Methodology, U.S. Bureau of the Census.
- 2006 Census Technical Report: Sampling and Weighting, (2009), Statistics Canada.
- Survey Methods and Practice (2003), Statistics Canada.



<부 록>

<부표 1> 행정구역별 동읍면부 구분

일련번호	시도	시군구	동읍면부	일련번호	시도	시군구	동읍면부
1	서울특별시	종로구	동부	41	부산광역시	기장군	읍부
2	서울특별시	중구	동부	42	부산광역시	기장군	면부
3	서울특별시	용산구	동부	43	대구광역시	중구	동부
4	서울특별시	성동구	동부	44	대구광역시	동구	동부
5	서울특별시	광진구	동부	45	대구광역시	서구	동부
6	서울특별시	동대문구	동부	46	대구광역시	남구	동부
7	서울특별시	중랑구	동부	47	대구광역시	북구	동부
8	서울특별시	성북구	동부	48	대구광역시	수성구	동부
9	서울특별시	강북구	동부	49	대구광역시	달서구	동부
10	서울특별시	도봉구	동부	50	대구광역시	달성군	읍부
11	서울특별시	노원구	동부	51	대구광역시	달성군	면부
12	서울특별시	은평구	동부	52	인천광역시	중구	동부
13	서울특별시	서대문구	동부	53	인천광역시	동구	동부
14	서울특별시	마포구	동부	54	인천광역시	남구	동부
15	서울특별시	양천구	동부	55	인천광역시	연수구	동부
16	서울특별시	강서구	동부	56	인천광역시	남동구	동부
17	서울특별시	구로구	동부	57	인천광역시	부평구	동부
18	서울특별시	금천구	동부	58	인천광역시	계양구	동부
19	서울특별시	영등포구	동부	59	인천광역시	서구	동부
20	서울특별시	동작구	동부	60	인천광역시	강화군	읍부
21	서울특별시	관악구	동부	61	인천광역시	강화군	면부
22	서울특별시	서초구	동부	62	인천광역시	옹진군	면부
23	서울특별시	강남구	동부	63	광주광역시	동구	동부
24	서울특별시	송파구	동부	64	광주광역시	서구	동부
25	서울특별시	강동구	동부	65	광주광역시	남구	동부
26	부산광역시	중구	동부	66	광주광역시	북구	동부
27	부산광역시	서구	동부	67	광주광역시	광산구	동부
28	부산광역시	동구	동부	68	대전광역시	동구	동부
29	부산광역시	영도구	동부	69	대전광역시	중구	동부
30	부산광역시	부산진구	동부	70	대전광역시	서구	동부
31	부산광역시	동래구	동부	71	대전광역시	유성구	동부
32	부산광역시	남구	동부	72	대전광역시	대덕구	동부
33	부산광역시	북구	동부	73	울산광역시	중구	동부
34	부산광역시	해운대구	동부	74	울산광역시	남구	동부
35	부산광역시	사하구	동부	75	울산광역시	동구	동부
36	부산광역시	금정구	동부	76	울산광역시	북구	동부
37	부산광역시	강서구	동부	77	울산광역시	울주군	읍부
38	부산광역시	연제구	동부	78	울산광역시	울주군	면부
39	부산광역시	수영구	동부	79	경기도	수원시 장안구	동부
40	부산광역시	사상구	동부	80	경기도	수원시 권선구	동부

〈부표 1〉 행정구역별 동읍면부 구분(계속)

일련번호	시도	시군구	동읍면부	일련번호	시도	시군구	동읍면부
81	경기도	수원시 팔달구	동부	121	경기도	이천시	면부
82	경기도	수원시 영통구	동부	122	경기도	이천시	동부
83	경기도	성남시 수정구	동부	123	경기도	안성시	동부
84	경기도	성남시 중원구	동부	124	경기도	안성시	면부
85	경기도	성남시 분당구	동부	125	경기도	안성시	동부
86	경기도	의정부시	동부	126	경기도	김포시	읍부
87	경기도	안양시 만안구	동부	127	경기도	김포시	면부
88	경기도	안양시 동안구	동부	128	경기도	김포시	동부
89	경기도	부천시 원미구	동부	129	경기도	화성시	읍부
90	경기도	부천시 소사구	동부	130	경기도	화성시	면부
91	경기도	부천시 오정구	동부	131	경기도	화성시	동부
92	경기도	광명시	동부	132	경기도	광주시	읍부
93	경기도	평택시	읍부	133	경기도	광주시	면부
94	경기도	평택시	면부	134	경기도	광주시	동부
95	경기도	평택시	동부	135	경기도	양주시	읍부
96	경기도	동두천시	동부	136	경기도	양주시	면부
97	경기도	안산시 상록구	동부	137	경기도	양주시	동부
98	경기도	안산시 단원구	동부	138	경기도	포천시	읍부
99	경기도	고양시 덕양구	동부	139	경기도	포천시	면부
100	경기도	고양시 일산동구	동부	140	경기도	포천시	동부
101	경기도	고양시 일산서구	동부	141	경기도	여주군	읍부
102	경기도	과천시	동부	142	경기도	여주군	면부
103	경기도	구리시	동부	143	경기도	연천군	읍부
104	경기도	남양주시	읍부	144	경기도	연천군	면부
105	경기도	남양주시	면부	145	경기도	가평군	읍부
106	경기도	남양주시	동부	146	경기도	가평군	면부
107	경기도	오산시	동부	147	경기도	양평군	읍부
108	경기도	시흥시	동부	148	경기도	양평군	면부
109	경기도	군포시	동부	149	강원도	춘천시	읍부
110	경기도	의왕시	동부	150	강원도	춘천시	면부
111	경기도	하남시	동부	151	강원도	춘천시	동부
112	경기도	용인시 처인구	읍부	152	강원도	원주시	읍부
113	경기도	용인시 처인구	면부	153	강원도	원주시	면부
114	경기도	용인시 처인구	동부	154	강원도	원주시	동부
115	경기도	용인시 기흥구	동부	155	강원도	강릉시	읍부
116	경기도	용인시 수지구	동부	156	강원도	강릉시	면부
117	경기도	파주시	읍부	157	강원도	강릉시	동부
118	경기도	파주시	면부	158	강원도	동해시	동부
119	경기도	파주시	동부	159	강원도	태백시	동부
120	경기도	이천시	읍부	160	강원도	속초시	동부



〈부표 1〉 행정구역별 동읍면부 구분(계속)

일련번호	시도	시군구	동읍면부	일련번호	시도	시군구	동읍면부
161	강원도	삼척시	읍부	201	충청북도	영동군	면부
162	강원도	삼척시	면부	202	충청북도	진천군	읍부
163	강원도	삼척시	동부	203	충청북도	진천군	면부
164	강원도	홍천군	읍부	204	충청북도	괴산군	읍부
165	강원도	홍천군	면부	205	충청북도	괴산군	면부
166	강원도	횡성군	읍부	206	충청북도	음성군	읍부
167	강원도	횡성군	면부	207	충청북도	음성군	면부
168	강원도	영월군	읍부	208	충청북도	단양군	읍부
169	강원도	영월군	면부	209	충청북도	단양군	면부
170	강원도	평창군	읍부	210	충청북도	증평군	읍부
171	강원도	평창군	면부	211	충청북도	증평군	면부
172	강원도	정선군	읍부	212	충청남도	천안시	읍부
173	강원도	정선군	면부	213	충청남도	천안시	면부
174	강원도	철원군	읍부	214	충청남도	천안시	동부
175	강원도	철원군	면부	215	충청남도	공주시	읍부
176	강원도	화천군	읍부	216	충청남도	공주시	면부
177	강원도	화천군	면부	217	충청남도	공주시	동부
178	강원도	양구군	읍부	218	충청남도	보령시	읍부
179	강원도	양구군	면부	219	충청남도	보령시	면부
180	강원도	인제군	읍부	220	충청남도	보령시	동부
181	강원도	인제군	면부	221	충청남도	아산시	읍부
182	강원도	고성군	읍부	222	충청남도	아산시	면부
183	강원도	고성군	면부	223	충청남도	아산시	동부
184	강원도	양양군	읍부	224	충청남도	서산시	읍부
185	강원도	양양군	면부	225	충청남도	서산시	면부
186	충청북도	청주시 상당구	동부	226	충청남도	서산시	동부
187	충청북도	청주시 흥덕구	동부	227	충청남도	논산시	읍부
188	충청북도	청주시	읍부	228	충청남도	논산시	면부
189	충청북도	청주시	면부	229	충청남도	논산시	동부
190	충청북도	청주시	동부	230	충청남도	계룡시	면부
191	충청북도	제천시	읍부	231	충청남도	계룡시	동부
192	충청북도	제천시	면부	232	충청남도	금산군	읍부
193	충청북도	제천시	동부	233	충청남도	금산군	면부
194	충청북도	청원군	읍부	234	충청남도	연기군	읍부
195	충청북도	청원군	면부	235	충청남도	연기군	면부
196	충청북도	보은군	읍부	236	충청남도	부여군	읍부
197	충청북도	보은군	면부	237	충청남도	부여군	면부
198	충청북도	옥천군	읍부	238	충청남도	서천군	읍부
199	충청북도	옥천군	면부	239	충청남도	서천군	면부
200	충청북도	영동군	읍부	240	충청남도	청양군	읍부

〈부표 1〉 행정구역별 동읍면부 구분(계속)

일련번호	시도	시군구	동읍면부	일련번호	시도	시군구	동읍면부
241	충청남도	청양군	면부	281	전라북도	부안군	읍부
242	충청남도	홍성군	읍부	282	전라북도	부안군	면부
243	충청남도	홍성군	면부	283	전라남도	목포시	동부
244	충청남도	예산군	읍부	284	전라남도	여수시	읍부
245	충청남도	예산군	면부	285	전라남도	여수시	면부
246	충청남도	태안군	읍부	286	전라남도	여수시	동부
247	충청남도	태안군	면부	287	전라남도	순천시	읍부
248	충청남도	당진군	읍부	288	전라남도	순천시	면부
249	충청남도	당진군	면부	289	전라남도	순천시	동부
250	전라북도	전주시 완산구	동부	290	전라남도	니주시	읍부
251	전라북도	전주시 덕진구	동부	291	전라남도	니주시	면부
252	전라북도	군산시	읍부	292	전라남도	니주시	동부
253	전라북도	군산시	면부	293	전라남도	광양시	읍부
254	전라북도	군산시	동부	294	전라남도	광양시	면부
255	전라북도	익산시	읍부	295	전라남도	광양시	동부
256	전라북도	익산시	면부	296	전라남도	담양군	읍부
257	전라북도	익산시	동부	297	전라남도	담양군	면부
258	전라북도	정읍시	읍부	298	전라남도	곡성군	읍부
259	전라북도	정읍시	면부	299	전라남도	곡성군	면부
260	전라북도	정읍시	동부	300	전라남도	구례군	읍부
261	전라북도	남원시	읍부	301	전라남도	구례군	면부
262	전라북도	남원시	면부	302	전라남도	고흥군	읍부
263	전라북도	남원시	동부	303	전라남도	고흥군	면부
264	전라북도	김제시	읍부	304	전라남도	보성군	읍부
265	전라북도	김제시	면부	305	전라남도	보성군	면부
266	전라북도	김제시	동부	306	전라남도	회순군	읍부
267	전라북도	완주군	읍부	307	전라남도	회순군	면부
268	전라북도	완주군	면부	308	전라남도	장흥군	읍부
269	전라북도	진안군	읍부	309	전라남도	장흥군	면부
270	전라북도	진안군	면부	310	전라남도	강진군	읍부
271	전라북도	무주군	읍부	311	전라남도	강진군	면부
272	전라북도	무주군	면부	312	전라남도	해남군	읍부
273	전라북도	장수군	읍부	313	전라남도	해남군	면부
274	전라북도	장수군	면부	314	전라남도	영암군	읍부
275	전라북도	임실군	읍부	315	전라남도	영암군	면부
276	전라북도	임실군	면부	316	전라남도	무안군	읍부
277	전라북도	순창군	읍부	317	전라남도	무안군	면부
278	전라북도	순창군	면부	318	전라남도	함평군	읍부
279	전라북도	고창군	읍부	319	전라남도	함평군	면부
280	전라북도	고창군	면부	320	전라남도	영광군	읍부



〈부표 1〉 행정구역별 동읍면부 구분(계속)

일련번호	시도	시군구	동읍면부	일련번호	시도	시군구	동읍면부
321	전라남도	영광군	면부	361	경상북도	경산시	면부
322	전라남도	장성군	읍부	362	경상북도	경산시	동부
323	전라남도	장성군	면부	363	경상북도	군위군	읍부
324	전라남도	완도군	읍부	364	경상북도	군위군	면부
325	전라남도	완도군	면부	365	경상북도	의성군	읍부
326	전라남도	진도군	읍부	366	경상북도	의성군	면부
327	전라남도	진도군	면부	367	경상북도	청송군	읍부
328	전라남도	신안군	읍부	368	경상북도	청송군	면부
329	전라남도	신안군	면부	369	경상북도	영양군	읍부
330	경상북도	포항시 남구	읍부	370	경상북도	영양군	면부
331	경상북도	포항시 남구	면부	371	경상북도	영덕군	읍부
332	경상북도	포항시 남구	동부	372	경상북도	영덕군	면부
333	경상북도	포항시 북구	읍부	373	경상북도	청도군	읍부
334	경상북도	포항시 북구	면부	374	경상북도	청도군	면부
335	경상북도	포항시 북구	동부	375	경상북도	고령군	읍부
336	경상북도	경주시	읍부	376	경상북도	고령군	면부
337	경상북도	경주시	면부	377	경상북도	성주군	읍부
338	경상북도	경주시	동부	378	경상북도	성주군	면부
339	경상북도	김천시	읍부	379	경상북도	칠곡군	읍부
340	경상북도	김천시	면부	380	경상북도	칠곡군	면부
341	경상북도	김천시	동부	381	경상북도	예천군	읍부
342	경상북도	안동시	읍부	382	경상북도	예천군	면부
343	경상북도	안동시	면부	383	경상북도	봉화군	읍부
344	경상북도	안동시	동부	384	경상북도	봉화군	면부
345	경상북도	구미시	읍부	385	경상북도	울진군	읍부
346	경상북도	구미시	면부	386	경상북도	울진군	면부
347	경상북도	구미시	동부	387	경상북도	울릉군	읍부
348	경상북도	영주시	읍부	388	경상북도	울릉군	면부
349	경상북도	영주시	면부	389	경상남도	창원시	읍부
350	경상북도	영주시	동부	390	경상남도	창원시	면부
351	경상북도	영천시	읍부	391	경상남도	창원시	동부
352	경상북도	영천시	면부	392	경상남도	마산시	읍부
353	경상북도	영천시	동부	393	경상남도	마산시	면부
354	경상북도	상주시	읍부	394	경상남도	마산시	동부
355	경상북도	상주시	면부	395	경상남도	진주시	읍부
356	경상북도	상주시	동부	396	경상남도	진주시	면부
357	경상북도	문경시	읍부	397	경상남도	진주시	동부
358	경상북도	문경시	면부	398	경상남도	진해시	동부
359	경상북도	문경시	동부	399	경상남도	통영시	읍부
360	경상북도	경산시	읍부	400	경상남도	통영시	면부

〈부표 1〉 행정구역별 동읍면부 구분(계속)

일련번호	시도	시군구	동읍면부	일련번호	시도	시군구	동읍면부
401	경상남도	통영시	동부	431	경상남도	함양군	읍부
402	경상남도	사천시	읍부	432	경상남도	함양군	면부
403	경상남도	사천시	면부	433	경상남도	거창군	읍부
404	경상남도	사천시	동부	434	경상남도	거창군	면부
405	경상남도	김해시	읍부	435	경상남도	합천군	읍부
406	경상남도	김해시	면부	436	경상남도	합천군	면부
407	경상남도	김해시	동부	437	제주도	제주시	동부
408	경상남도	밀양시	읍부	438	제주도	서귀포시	동부
409	경상남도	밀양시	면부	439	제주도	북제주군	읍부
410	경상남도	밀양시	동부	440	제주도	북제주군	면부
411	경상남도	거제시	읍부	441	제주도	남제주군	읍부
412	경상남도	거제시	면부	442	제주도	남제주군	면부
413	경상남도	거제시	동부				
414	경상남도	양산시	읍부				
415	경상남도	양산시	면부				
416	경상남도	양산시	동부				
417	경상남도	의령군	읍부				
418	경상남도	의령군	면부				
419	경상남도	함안군	읍부				
420	경상남도	함안군	면부				
421	경상남도	창녕군	읍부				
422	경상남도	창녕군	면부				
423	경상남도	고성군	읍부				
424	경상남도	고성군	면부				
425	경상남도	남해군	읍부				
426	경상남도	남해군	면부				
427	경상남도	하동군	읍부				
428	경상남도	하동군	면부				
429	경상남도	산청군	읍부				
430	경상남도	산청군	면부				



〈부표 2〉 성별, 연령그룹, 혼인상태와 교육정도 구분

연번	성별	연번	연령그룹	고려 사항	연번	혼인상태	연번	교육정도
1	남자	1	0-4세		1	미혼	1	안 받았음 (미취학 포함)
2	여자	2	5세	보육, 거주지	2	배우자가 있거나 이혼	2	초중고 재학/휴학
		3	6-9세	교육정도	3	사별	3	초등학교 졸업/중퇴
		4	10-11세	아동보육상태			4	중학교 졸업/중퇴
		5	12세	아동보육상태, 통근학 관련			5	고등학교 졸업/중퇴
		6	13-14세	통근학 관련			6	대학과 대학교 졸업
		7	15-19세				7	대학과 대학교 재학/휴학
		8	20-24세				8	대학과 대학교 수료/중퇴
		9	25-29세				9	대학원 졸업
		10	30-34세				10	대학원 재학/수료/휴학/중퇴
		11	35-39세					
		12	40-44세					
		13	45-49세					
		14	50-54세					
		15	55-59세					
		16	60-64세					
		17	65-69세					
		18	70-74세					
		19	75-79세					
		20	80-84세					
		21	85세이상					

※ 1. 연령그룹 : 기본적으로 5세 단위로 연령을 묶지만, 표본조사 항목과 관련하여 보다 세분화했음.
 2. 연령그룹 0세의 경우 1년전 거주지 관련 항목 있을 경우 따로 추정하기도 함

〈부표 3〉 조사구특성, 가구규모와 거처종류별 구분

연번	조사구특성	연번	구분	연번	구분
1	보통	1	1인가구	1	단독주택
2	섬	2	2인	2	아파트
A	아파트	3	3인	3	연립주택
		4	4인	4	다세대주택
		5	5인	5	영업용건물내주택
		6	6인 이상	6	주택이외의거처