한국시뮬레이션학회 논문지
Vol. 25, No. 4, pp. 13-20 (2016. 12)
http://doi.org/10.9709/JKSS.2016.25.4.013
ISSN 1225-5904

Iterative Proportional Updating 방법을 이용한 한국 가상 인구 데이터 생성

손우식† · 권오규 · 이상희

Generating Korean synthetic populations by using the iterative proportional updating method

Woo-Sik Son† · Okyu Kwon · Sang-Hee Lee

ABSTRACT

Microsimulation model has aimed to simulate the impact of policy at the level of individual and household. Recently, microsimulation model has been widely accepted in OECD countries for evaluating their economic and social policies. For improving the availability of microsimulation model, the population data which shows good accordance with the official statistics should be required. In this paper, we generate Korean synthetic populations by using the iterative proportional updating method. For the validation of Korean synthetic populations, we compute the difference between the generated synthetic populations and the summary table of Korean census. Then, we confirm that it shows good accordance with the summary table.

Keywords : Microsimulation model, Synthetic populations, Iterative proportional updating

요 약

마이크로시뮬레이션 모델은 거시적 수준의 인구, 사회, 경제 변화를 각 개인과 가구 단위의 미시적 수준의 사건들로부터 기술하고자 하는 것을 목적으로 하며, 최근 OECD 국가들을 중심으로 정책 시뮬레이션 도구로서 많은 관심을 받고 있다. 마이크로시뮬레이션 모델의 활용도를 높이기 위해서는 해당 국가의 인구 구조를 잘 반영하는 인구 데이터가 필요한데, 우리는 반복비례갱신 (iterative proportional updating) 방법을 이용하여 한국 가상 인구를 생성하였다. 생성된 가상 인구 데이터의 검증을 위하여 인구센서스 집계 결과와의 오차를 계산하였으며, 가구와 인구 모두에 대해서 실제 집계 결과와 작은 오차를 보이는 것을 확인하였다.

주요어: 마이크로시뮬레이션 모델, 가상 인구, 반복비례갱신

1. 서론

이 연구는 한국과학기술정보연구원 (초고성능 컴퓨팅 기반 건강한 고령사회 대응 빅데이터 기술개발 과제)의 지원으로 수행되었습니다.

Received: 24 June 2016, **Revised**: 4 October 2016, **Accepted**: 4 October 2016

† **Corresponding Author**: Woo-Sik Son
E-mail: wsson@nims.re.kr
Division of Integrated Mathematics, National Institute for Mathematical Sciences, Daejeon, Korea

마이크로시뮬레이션 모델 (Microsimulation model, MSM)은 거시적 수준의 인구, 사회, 경제 변화를 각 개인과 가구 단위의 미시적 수준의 사건들로부터 기술하는데 폭넓게 활용되고 있다 (Van Imhoff and Post, 1998). 최근 OECD 국가들은 정책 시뮬레이션 도구로서 자국의 경제, 사회 상황에 맞는 MSM을 운용하고 있으며, MSM의 예로는 세금 관련 정책 시뮬레이션, 연금 정책 시뮬레이션, 그리고 토지 이용 정책 시뮬레이션 등이 존재한다

(Sutherland and Figari, 2013; Hancock et al, 1992; Waddell, 2002).

MSM을 통한 사회, 경제 정책 시뮬레이션을 위해서는 해당 국가의 인구 구조 (MSM에서 고려되는 속성들: 지역, 성별, 연령, 혼인 상태, 교육 정도 등)를 잘 반영하는 인구 데이터가 필요하다. 현재 한국을 포함한 많은 국가들은 국가 공식 통계기관을 통하여 인구센서스 샘플을 공개하고 있다. 인구센서스 샘플은 인구 및 주택에 대한 다양한 사회, 경제 현상을 분석하고자 하는 연구자 또는 기업의 마케팅 전략에 이용될 수 있다. 최근 인구센서스 샘플을 활용하여 MSM 연구에 이용될 수 있는 가상 인구 데이터를 생성하는 연구가 많은 관심을 받고 있다. Beckman 등은 미국 Census Bureau가 제공하는 요약된 인구센서스 집계 (summary) 결과와 인구센서스 샘플 (public use microdata sample)을 이용하여 가상 인구 데이터를 생성하였다 (Beckman et al, 1996). Beckman 등은 인구센서스 샘플의 각 가구에 가중치를 부여하고 각 가구가 가중치만큼 포함되도록 가상 인구 데이터를 생성하였는데, 가중치가 인구센서스 집계 결과의 제약 조건 (constraint)을 만족하도록 반복비례적합 (Iterative proportional fitting, IPF) 방법을 이용하여 가중치를 적합(fitting)하였다. Beckman 등의 연구 결과는 가상 인구 데이터 생성 분야를 개척한 것으로 평가받고 있지만, 생성된 가상 인구 데이터가 가구원과 관련한 제약 조건들을 만족하지 않는다는 단점이 존재한다. 이 단점을 극복하기 위한 최근의 연구들이 존재한다 (Barthelemy and Toint, 2012; Muller, 2011). 특히, Ye 등은 Beckman 등의 방법을 변형하여 가구 그리고 가구원에 대한 제약 조건들을 모두 만족시킬 수 있는 반복비례갱신 (Iterative proportional updating, IPU) 방법을 발표하였다 (Ye et al, 2009).

한국의 사회, 경제 정책 효과를 MSM으로 시뮬레이션하기 위해서는 한국 가상 인구 데이터가 필요하다. 본 연구에서, 우리는 IPU 방법을 활용하여 한국 가상 인구 데이터를 생성하였다. IPU를 이용하여 인구센서스 집계 결과의 제약 조건이 만족되도록 인구센서스 샘플의 각 가구에 가중치를 부여하고, 각 가구가 가중치만큼 포함되도록 가상 인구 데이터를 생성하였다. 이 논문의 남은 부분은 다음과 같이 구성되었다. 2장은 가상 인구 생성 방법을 설명하고, 3장은 생성된 한국 가상 인구 데이터에 대한 결과를 분석하며, 4장은 토의 및 결론을 담고 있다.

2. 한국 가상 인구 생성 방법

한국 가상 인구 생성을 위해서는 인구센서스 샘플 데이터와 집계 결과가 필요하다. 한국은 5년 주기로 인구센서스 조사를 시행하고 있으며, 시행년도 다음해에 인구센서스 전체 자료에 대한 집계 결과와 전체 자료에서 추출된 샘플을 공개하고 있다. 우리는 현재 공개된 가장 최근 그리고 가장 큰 규모의 인구센서스 샘플인 2010년 인구센서스 2% 샘플 (357,828 가구; 933,950 가구원) 그리고 2010년 인구센서스 집계 결과를 이용하였다. 인구센서스 샘플은 마이크로데이터 통합서비스에서 제공하며 인구센서스 집계 결과는 국가통계포털이 공개하고 있다 (Microdata Integrated Service, 2015; Korean Statistical Information Service, 2015). 우리가 생성하고자 하는 한국 가상 인구 데이터는 2010년 인구센서스 집계의 내국인 (17,339,422 가구; 47,990,761 가구원)과 동일한 크기를 갖고 있다. 이는, 인구센서스 집계 결과에서 가구와 가구원에 대한 속성들 (예; 지역, 성별, 연령, 혼인상태, 교육정도 등)이 내국인 기준으로 발표되기 때문이다.

우리가 이용한 인구센서스 집계 결과는 가구에 대한 제약 조건 16개 그리고 가구원(인구)에 대한 제약 조건 13,824개 [13,824 = 시도별 인구(16) × 성별 인구(2) × 연령 그룹별 인구(18) × 혼인상태별 인구(4) × 교육정도별 인구(6)]이다. 가구에 대한 제약 조건은 16개 시도별 (서울, 부산, 대구, 인천, 광주, 대전, 울산, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주) 가구 수이다. 가구원에 대한 제약 조건은 2010년 인구센서스에서 조사된 모든 가구원을 시도(16), 성별(2), 연령 그룹(18), 혼인상태(4), 교육 정도(6)에 따라서 13,824개 집단으로 분류한 뒤, 각 집단별 가구원의 합계이다. 가구원에 대한 구체적 속성은 다음과 같다. (가) 성별: [1] 남자, [2] 여자, (나) 연령 그룹: [1] 0~4세, [2] 5~9세, …, [17] 80~84세, [18] 85세 이상, (다) 혼인상태: [1] 미혼, [2] 배우자 있음, [3] 사별, [4] 이혼, (라) 교육정도: [1] 안 받았음(미취학 포함), 초등학교(재학 + 중퇴), [2] 초등학교(졸업), 중학교(재학 + 중퇴), [3] 중학교(졸업), 고등학교(재학 + 중퇴), [4] 고등학교(졸업), 대학교(4년제 미만 + 4년제 이상) (재학 + 수료 + 휴학 + 중퇴), [5] 대학교(4년제 미만 + 4년제 이상) (졸업), 대학원(석사과정 + 박사과정) (재학 + 수료 + 중퇴), [6] 대학원(석사과정 + 박사과정) (졸업)

Table 1과 2는 각각 인구센서스 집계 결과와 인구센서

Table. 1. Example of summary table of 2010 population and housing census

지역	성별	연령	혼인상태	교육정도	총합
서울	남자	30~34세	미혼	대학원 박사 (졸업)	931
서울	남자	35~39세	미혼	대학원 박사 (졸업)	995
서울	여자	30~34세	미혼	대학원 박사 (졸업)	964
서울	여자	35~39세	미혼	대학원 박사 (졸업)	1,153

Table. 2. Example of microdata sample of 2010 population and housing census

가구 번호	지역	가구 원번호	가구 주와의 관계	성별	연령	혼인 상태	…	경제 활동 상태
10	11	1	1	1	43	1	…	1
10	11	2	2	2	39	1	…	2
10	11	3	3	1	7	4	…	2
11	21	1	1	1	36	1	…	1
12	23	1	1	1	34	1	…	1
12	23	2	2	2	33	1	…	1

Table. 3. Code for public use microdata sample of 2010 population and housing census

항목명	코드	코드명
지역	11	서울특별시
	21	부산광역시
	22	대구광역시
	23	인천광역시
	24	광주광역시
	25	대전광역시
	26	울산광역시
	31	경기도
	32	강원도
	33	충청북도
	34	충청남도
	35	전라북도
	37	경상북도
	38	경상남도
	39	제주도
성별	1	남자
	2	여자
연령	0~84	0세~84세 (연령 그룹이 아닌 각 세 단위로 제공)
	85	85세 이상
혼인상태	1	미혼
	2	배우자 있음
	3	사별
	4	이혼
경제활동상태	1	취업 (일하였음 + 임시휴직)
	2	미취업
가구주와의 관계	1	가구주
	2	가구주의 배우자
	3	자녀
	4	자녀의 배우자
	5	가구주의 부모, 배우자의 부모, 조부모
	6	손자녀 및 그 배우자, 증손자녀 및 그 배우자
	7	형제자매, 그 배우자
	8	기타 친인척
	9	기타 동거인

Table. 4. Example of summary table of 2010 population and housing census & weight

가구 번호	지역	가구 원번호	가구 주와의 관계	성별	연령	혼인 상태	…	경제 활동 상태	가중치
10	11	1	1	1	43	1	…	1	48
10	11	2	2	2	39	1	…	2	48

10	11	3	3	1	7	4	⋯	2	48
11	21	1	1	1	36	1	⋯	1	35
12	23	1	1	1	34	1	⋯	1	52
12	23	2	2	2	33	1	⋯	1	52

스 샘플의 예를 보여준다. 인구센서스 샘플의 각 행은 한 명의 가구원에 해당하며, 각 열은 속성을 나타내는데 가구번호가 같을 경우 동일가구에 속한다. 인구센서스 샘플의 각 속성별 코드는 마이크로데이터 통합서비스에서 제공하는 2010 인구주택총조사 2% 마이크로데이터이용설명서의 코드표와 동일한 방식으로 기록되었는데 지역, 가구주와의 관계, 성별, 연령, 혼인상태, 경제활동상태에 대한 코드를 살펴보면 Table 3과 같다 (Microdata Integrated Service, 2015). 인구센서스 샘플의 각 가구에 인구센서스 집계 결과의 제약 조건이 만족되도록 가중치를 부여하는데 동일 가구는 동일한 가중치를 갖는다. Table 4는 인구센서스 샘플과 IPU로 생성된 가중치의 예를 보여준다. Table 4의 예처럼, 10번 가구의 가중치가 48이면, 가상 인구 데이터에는 10번 가구가 48번 포함된다.

IPU의 과정을 설명하면 다음과 같다. IPU의 첫 번째 단계는 인구센서스 집계 결과의 제약 조건들로부터 constraint vector C $(1 \times m$ dim)를 생성한다. m은 제약 조건의 수이며, C의 원소 c_j는 j번째 제약 조건을 나타낸다. 우리는 국가통계포털이 제공하는 제약 조건을 이용했지만, Ye 등의 연구 결과와 같이 경우에 따라서는 각 속성별 제약 조건으로부터 IPF를 이용해서 최종 제약 조건들을 생성할 수도 있다 (Ye et al, 2009). 예를 들어, 각 속성에 대한 114개 제약 조건 [114 = 시도별 인구(16) + 성별 인구(2) + 각 연령별 인구(86) + 혼인상태별 인구(4) + 교육 정도별 인구(6)]으로부터 IPF를 이용하여 66,048개의 제약 조건 [66,048 = 시도별 인구(16) × 성별 인구(2) × 각 연령별 인구(86) × 혼인상태별 인구(4) × 교육정도별 인구(6)]을 생성하여 이용할 수 있다. 이 경우, IPF의 초기 추정치(initial guess)는 인구센서스 샘플로부터 구하게 되므로 우리가 이용한 2% 샘플 보다 큰 규모의 인구센서스 샘플이 필요하다.

IPU의 두 번째 단계는 인구센서스 샘플로부터 위 제약 조건에 해당하는 frequency matrix D $(N \times m$ dim) 를 생성한다. N은 인구센서스 샘플의 가구 수이며, D의 원소 $d_{i,j}$는 인구센서스 샘플 i번째 가구의 j번째 제약 조건에 대한 기여도를 나타낸다. 구체적으로 기여도는 다음과 같이 계산된다. 가구에 대한 제약 조건일 경우, i번째 가구가 해당 제약 조건에 속하면 $d_{i,j} = 1$ 아니면 0이다. 가구원에 대한 제약 조건일 경우, $d_{i,j}$는 i번째 가구의 가구원 중 해당 제약 조건에 속하는 사람의 수를 나타낸다.

IPU의 세 번째 단계는 위 두 과정을 통해서 생성된 constraint vector와 frequency matrix에 zero cell 문제가 있는지 검사하고 zero cell 문제가 있다면 이를 해결하는 과정이다. Zero cell 문제는 constraint vector의 원소는 0이 아닌데 frequency matrix의 해당 열은 모두 0인 경우를 나타낸다. 즉, $c_j \neq 0$, $d_{i,j} = 0 \ \ \forall \ i$. Zero cell 문제는 constraint의 개수는 크고 인구센서스 샘플의 크기는 작을 때 발생한다. 가구원에 대한 제약 조건 중 굉장히 작은 값을 갖는 제약 조건의 경우, 예를 들어 제주에 사는 85세 이상 미혼, 박사 여성에 대한 제약 조건이 1명이라고 가정하면 인구센서스 샘플에는 이 가구원을 포함한 가구가 포함되지 않을 가능성이 높기 때문이다. Zero cell 문제가 있을 경우, IPU는 수렴하지 않기 때문에 다음의 방법들을 통하여 zero cell 문제를 해결해야 한다. 첫 번째 방법으로 constraint의 개수를 줄여서 zero cell 문제를 사라지게 할 수 있다. 두 번째 방법은 (제주가 아닌) 다른 시도에 대한 인구센서스 샘플 중, 해당 제약 조건에 속하는 가구원을 포함하고 있는 가구가 있다면 이 가구를 (제주의) 인구센서스 샘플에 포함시키는 것이다. 이 과정에서 constraint vector와 frequency matrix는 새로 생성되어야 한다.

우리 연구 결과에서 zero cell 문제는 다음과 같이 해결되었다. 우리가 이용하고자 했던 가구원에 대한 제약 조건은 13,824개가 아닌 25,344개 [시도(16) × 성별(2) × 연령 그룹(18) × 혼인상태(4) × 교육정도(11)] 제약 조건이었다. 앞에서 언급된 제약 조건들과 교육 정도 속성만 다른데 구체적으로는 다음과 같다; [1] 안 받았음(미취학 포함), [2] 초등학교 재학 + 수료 + 휴학 + 중퇴, [3] 초등학교 졸업, [4] 중학교 재학 + 수료 + 휴학 + 중퇴, [5] 중학교 졸업, [6] 고등학교 재학 + 수료 + 휴학 + 중퇴, [7] 고등학교 졸업, [8] 대학교(4년제 미만 + 4년제 이상) 재학 + 수료 + 휴학 + 중퇴, [9] 대학교(4년제 미만 + 4년제 이상) 졸업, [10] 대학원(석사과정 + 박사과정) 재학 + 수료 + 휴학 + 중퇴, [11] 대학원(석사과정 + 박사과정) (졸업)

우리는 가구원에 대하여 25,344개 제약 조건을 이용할

경우 5,131개 속성에서 zero cell 문제가 발생하는 것을 확인하였다. 교육 정도의 속성을 6개로 축소시킨 13,824개 제약 조건을 이용할 경우, zero cell 문제가 발생하는 속성의 개수는 2,937개로 감소된다. 2,937개의 zero cell 문제에 대해서 zero cell 문제의 두 번째 해법, 즉 다른 시도 중 해당 제약 조건에 속하는 가구원을 포함하는 가구를 인구센서스 샘플에 포함시키는 과정을 거치면 zero cell 문제가 발생하는 속성의 개수는 692개로 감소된다. 아직 남은 692개의 zero cell 문제를 완벽하게 정리하기 위해서는 6개 교육 정도 속성을 한 개의 속성으로 축소시켜야 한다. 즉, 가상 인구 데이터에 교육 정도 속성은 포함되지 않게 되는데 이 경우, 가상 인구 데이터의 활용도는 떨어지게 된다. zero cell 문제가 아직 남아 있는 692개 제약 조건의 총가구원 수는 3,670명으로 총인구에 비하여 작기 때문에 우리는 zero cell 문제가 아직 남아있는 692개 제약 조건들을 무시하고 IPU 작업을 진행하였다. 즉, 우리는 13,824개 제약 조건에서 692개를 뺀 13,132개의 제약 조건을 이용하였다.

Initial guess for $\omega_i = 1$

While ($\Delta > \epsilon$)

  For $j = 1$ to m

    $\omega s_j = \sum_{i=1}^{N} \omega_i \cdot d_{i,j}$

    For $i = 1$ to N

      If $d_{i,j} \neq 0$ Then $\omega_i = \frac{c_j}{\omega s_j} \cdot \omega_{i,\,prev}$

    EndFor

  EndFor

  $\delta = \frac{1}{m} \sum_{j=1}^{m} \frac{|\omega s_j - c_j|}{c_j}$

  $\Delta = |\delta - \delta_{prev}|$

EndWhile

Fig. 1. Pseudo code for iterative proportional updating

IPU의 네 번째 단계에서는 constraint vector C와 frequency matrix D로부터 i번째 가구의 가중치 w_i를 계산한다. Fig. 1의 의사 코드 (pseudo code)는 가중치의 구체적인 계산 과정을 나타낸다. ωs_j는 j번째 제약 조건에 대한 weighted sum 그리고 δ는 모든 제약 조건에 대한 적합도 (goodness of fit)의 평균을 나타낸다. 우리는 두 연속된 적합 과정에서 δ의 이득, 즉 $\Delta = |\delta - \delta_{prev}|$가 역치 (threshold) ϵ보다 작게 될 때까지 IPU를 진행하였다. 역치는 1.0×10^{-7}로 설정하였다. 위 과정을 통해 생성된 가중치는 실수이므로 반올림 (rounding)을 통해 가중치를 정수화하고, 가중치만큼 해당 가구를 가상 인구 데이터에 포함하는 것으로 IPU의 모든 과정은 마무리된다.

3. 결과 분석

우리는 IPU를 통해서 생성한 한국 가상 인구 데이터의 검증을 위하여 인구센서스 집계 결과와의 오차를 계산하였다. 이 과정에서 비교를 위하여 IPU가 아닌 다른 방법들, 즉 모든 가구에 동일한 가중치를 부여한 데이터 세트 1과 마이크로데이터 통합서비스에서 제공하는 2010년 가구센서스 2% 샘플의 가구 가중치를 적용한 데이터 세트 2와 IPU를 통해서 계산된 가중치를 적용한 데이터 세트 3에 대해서 각각 인구센서스 집계 결과와의 오차를 계산하였다. 데이터 세트 1은 인구센서스 샘플의 모든 가구에 가중치 48을 부여하였는데, 이 값은 2010년 일반 가구 수 / 인구센서스 2% 샘플의 가구 수 $\simeq$ 48.457691를 정수화한 값이다.

Table. 5. Summary for Korean synthetic populations

데이터 세트	(1) 전체 가구 수	(2) 총인구	(3) 전체 가구 상대오차	(4) 총인구 상대오차	(5) 전체 가구 절대오차	(6) 총인구 절대오차
1	17,175,648	44,829,312	-0.009440	-0.065876	-163,774	-3,161,449
2	17,339,422	46,641,143	0	-0.028122	0	-1,349,618
3	17,339,194	47,986,646	-0.000013	-0.000085	-228	-4,115

Table 5는 각 데이터 세트에 대해서 (1) 전체 가구 수, (2) 총인구, (3) 전체 가구 상대오차, (4) 총인구 상대오차, (5) 전체 가구 절대오차, (6) 총인구 절대오차를 보여준다. 전체 가구 오차와 총인구 오차는 2010년 인구센서스 집계의 내국인(17,339,422 가구, 47,990,761 가구원)을 기준으로 계산되었다. Table 5의 결과를 보면, 데이터 세트 1은 가구와 인구 모두에 대해서 제약 조건과 큰 오

Table. 6. Error about the number of households for 16 administrative divisions

지역	가구 상대오차 (데이터 세트 1)	가구 절대오차 (데이터 세트 1)	가구 상대오차 (데이터 세트 2)	가구 절대오차 (데이터 세트 2)	가구 상대오차 (데이터 세트 3)	가구 절대오차 (데이터 세트 3)
서울	-0.120293	-421,545	0	0	0.000007	26
부산	-0.045927	-57,128	0	0	0.000159	198
대구	-0.133948	-116,311	0	0	0.000251	218
인천	-0.044440	-40,834	0	0	0.000273	251
광주	-0.095838	-49,439	0	0	0.000181	93
대전	-0.062065	-33,059	0	0	-0.000257	-137
울산	-0.106888	-39,937	0	0	-0.000524	-196
경기	-0.154009	-590,030	0	0	-0.000082	-317
강원	0.303119	169,065	0	0	0.000089	50
충북	0.188240	105,188	0	0	0.000282	158
충남	0.140347	105,125	0	0	-0.000351	-263
전북	0.316180	208,662	0	0	-0.000127	-84
전남	0.419859	286,105	0	0	0.000278	190
경북	0.225938	227,147	0	0	0.000541	543
경남	0.065140	74,988	0	0	-0.000398	-459
제주	0.043929	8,229	0	0	-0.002663	-499

Table. 7. Quantiles on the absolute error for 13,824 constraints

데이터 세트	최솟값	1분위값	중앙값	3분위값	최댓값
1	-136,933	-45	0	23	20,750
2	-105,446	-41	0	20	9,142
3	-676	-1	0	0	280

차를 보이는 것을 확인할 수 있다. 데이터 세트 2는 가구 수에 대한 오차는 없는 반면, 인구에 대해서는 큰 오차를 보인다.

Table 6은 전체 가구 수에 대한 오차를 16개 시도별로 보여준다. Table 6의 데이터 세트 1에 대한 결과를 보면 지역별 편차가 매우 큰 것을 확인할 수 있는데 이는 2010년 인구센서스 2% 샘플에 지역별 bias가 존재하기 때문이다. 총인구에 대한 절대 오차를 13,824개 제약 조건 [시도(16) × 성별(2) × 연령 그룹(18) × 혼인상태(4) × 교육정도(6)]으로 분류하고 4분위로 표현하면 Table 7과 같다. 13,824개 제약 조건 모두에 대해서 오차를 표현하기에는 조건 수가 너무 많기 때문에, 우리는 Table 7부터 Table 9까지 4분위수를 이용하여 오차의 최솟값과 최댓값 그리고 1분위, 중앙값, 3분위 값을 표현하였다. Table 7의 결과에서 최솟값은 20~24세의 연령 그룹에 해당한다. 20~24세 연령 그룹에서 큰 오차를 보이는 이유는 인구센서스 샘플에 포함되지 않는 국군, 전투경찰대 등의 특별조사구 거주인구와 기숙사 등의 집단가구 집단시설가구 거주인구가 이 연령 그룹에 많이 존재하기 때문이다. Table 8은 13,824개 제약 조건별 상대 인구 오차를 4분위로 보여준다. Table 8에서 -1의 값을 갖는 상대 인구 오차는 해당 제약 조건에 포함되는 가구원이 가상 인구 데이터에 전혀 존재하지 않는 경우이므로 이에 대한 자세한 언급이 필요하다. -1의 값을 갖는 상대 인구 오차가 발생하는 첫 번째 이유는 2장의 한국 가상 인구 생성 방법에서 언급한 것처럼 아직 남아 있는 692개 zero

cell 때문이며, 두 번째 이유는 실수 (real number)로 계산된 가중치가 반올림되면서 0이 되는 가중치가 발생하기 때문이다. Table 9는 각 데이터 세트에 대해서 13,824개 제약 조건별 상대 인구 오차 중 (1) -1의 값을 갖는 제약 조건 수, (2) 이에 해당하는 총 가구원 수, 그리고 각 제약 조건에 속하는 가구원 수의 (3) 최솟값, (4) 1분위값, (5) 중앙값, (6) 3분위값, (7) 최댓값을 보여준다. 우리는 Table 5~8의 결과들을 통해서 IPU를 이용하여 생성된 데이터 세트 3이 가구와 인구 모두에 대해서 제약 조건과 작은 오차를 보이는 것을 확인하였다. 또한 -1의 값을 갖는 상대 인구 오차에 대한 Table 9의 결과에서도 데이터 세트 3은 다른 데이터 세트에 비하여 상대적으로 양호한 결과를 보여준다.

Table. 8. Quantiles on the relative error for 13,824 constraints

데이터 세트	최솟값	1분위값	중앙값	3분위값	최댓값
1	-1	-0.445086	0	0.078651	47
2	-1	-0.470085	0	0.034661	56
3	-1	-0.000918	0	0	4.1

Table. 9. Analysis on the relative error about population, whose value is minus one

데이터 세트	(1) 제약 조건 수	(2) 총 가구원 수	(3) 최솟값	(4) 1분위 값	(5) 중앙값	(6) 3분위 값	(7) 최댓값
1	2,937	59,650	1	3	8	24	407
2	2,937	59,650	1	3	8	24	407
3	782	3,871	1	1	2	5	162

4. 논의 및 결론

우리는 MSM에 이용될 수 있는 한국 가상 인구 데이터 생성에 대하여 연구하였다. IPU를 이용하여 인구센서스 집계 결과의 제약 조건이 만족되도록 2010년 인구센서스 2% 샘플의 각 가구에 가중치를 부여하고, 각 가구가 정수화된 가중치만큼 포함되도록 한국 가상 인구 데이터를 생성하였다. 생성된 가상 인구 데이터의 검증을 위하여 인구센서스 집계 결과와의 오차를 계산하였으며, 가구와 인구 모두에 대해서 집계 결과와 작은 오차를 보이는 것을 확인하였다.

MSM을 이용한 사회, 경제 정책 시뮬레이션의 신뢰도와 활용도를 높이기 위해서는 해당 국가의 인구구조를 잘 반영하는 인구 데이터가 반드시 필요하다. 최근 한국 사회의 주요 이슈 중 하나인 청년 실업 문제에 대한 정부 정책 시뮬레이션에 데이터 세트 1 또는 2를 이용한 MSM을 적용한다고 가정해보자. 우리는 Table 7의 결과를 통하여 데이터 세트 1과 2는 20~24세 연령 그룹에서 실제 인구 데이터와 큰 오차를 보이는 것을 확인하였다. 데이터 세트 1 또는 2를 이용한 MSM의 경우, 초기 입력 데이터부터 심각한 오류를 포함하게 되므로 청년 실업 문제에 대한 정부 정책 효과 시뮬레이션의 출력 결과 또한 신뢰하기 어렵다. 이와 다르게, 본 연구를 통하여 생성된 가상 인구 데이터는 초기 입력 오류를 방지할 수 있으며 사회, 경제 정책 시뮬레이션의 신뢰도와 활용도를 높일 수 있을 것으로 기대된다.

우리는 가구와 가구원에 대한 13,840개 제약 조건을 이용하여 가상 인구 데이터를 생성하였다. 만약 더 많은 수의 제약 조건을 이용하여 더 상세한 속성들을 포함한 가상 인구 데이터를 생성할 수 있다면 MSM을 통하여 보다 정교한 정책 시뮬레이션을 할 수 있을 것이다. 이를 위해서는 공공 기관으로부터 더 상세한 집계 결과의 공개가 요구된다. 그리고 우리는 인구센서스 2% 샘플을 이용하였는데 만약 이보다 큰 규모인 10% 내외의 인구센서스 샘플이 공개된다면, 2장에서 언급되었던 것처럼 각 속성별 제약 조건으로부터 IPF를 통하여 보다 상세한 제약 조건을 만들 수 있을 것으로 기대된다.

References

Barthelemy J. and Toint P.L. (2012) “Synthetic Population Generation Without a Sample”, *Transportation Science*, 47(2), 266-279.

Beckman R.J., Baggerly K.A., and McKay M.D. (1996) “Creating Synthetic Baseline Populations”, *Transportation Research Part A: Policy and Practice*, 30(6), 415-429.

Hancock R., Mallender J., and Pudney S. (1992) “Constructing a Computer Model for Simulating the Future Distribution of Pensioners Incomes for Great Britain”, In: Hancock, R. and Sutherland H. (eds.) *Microsimulation Models for Public Policy Analysis*: New Frontiers. London: London School

of Economics.
Muller K. (2011) “Hierarchical IPF: Generating a synthetic population for Switzerland” *Eidgenossische Technische Hochschule Zurich*, IVT.
Sutherland H. and Figari F. (2013) “EUROMOD: the European Union tax-benefit microsimulation model”, *International Journal of Microsimulation,* 6(1), 4-26.
Van Imhoff, E. and Post W. (1998) “Microsimulation methods for population projection”, *Population: An English Selection*, 10(1), 97-138.
Waddell, P. (2002) “UrbanSim: Modeling urban development for land use, transportation, and environmental planning”, J*ournal of the American Planning Association*, 68(3), 297-314.
Ye X., Konduri K., Pendyala R.M., Sana B., and Waddell P. (2009) “A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic
Populations”, *88th Annual Meeting of the Transportation Research Board*.
Microdata Integrated Service (2015) available at: mdis.kostat.go.kr.
(마이크로데이터 통합서비스 (2015)).
Korean Statistical Information Service (2015) available at kosis.kr.
(국가통계포털 (2015)).

손 우 식 (wsson@nims.re.kr)

2000 서강대학교 물리학과 학사
2002 서강대학교 물리학과 석사
2007 서강대학교 물리학과 박사
2007~ 2011 서강대학교 물리학과 박사후 연구원
2012~ 현재 국가수리과학연구소 연구원

관심분야 : 데이터 과학, 복잡계, 비선형 동역학

권 오 규 (okw@nims.re.kr)

1999 KAIST 물리학과 학사
2001 KAIST 물리학과 석사
2006 KAIST 물리학과 박사
2009 고려대 세포동력학연구센터 박사후 연구원
2010 Northwestern Institute on Complex Systems 박사후 연구원
2011 APCTP 박사후 연구원
2011~ 현재 국가수리과학연구소 선임연구원

관심분야 : 복잡계, 데이터 과학, 네트워크

이 상 희 (sunchaos@nims.re.kr)

2005 부산대학교 물리학과 박사
2009~ 2013 한국수리생물학회 운영위원
2014~ 현재 한국수리생물학회 부회장
2008~ 현재 국가수리과학연구소 선임연구원

관심분야 : 생물행동 모델링, 생태계 모델링, 사회성곤충 전략모델링