발 간 등 록 번 호
11-1240245-000057-14

OPEN 공공누리 출처표시 상업용금지 변경금지
공공저작물 자유이용허락

2020년 연구보고서

통계지리정보서비스(SGIS) 제공 자료의 비밀보호 적용 방안

2021. 4.

http://kostat.go.kr/sri

9 772288 116605 08
ISSN 2288-1166(Print)
ISSN 2733-4120(Online)

통계청
통계개발원

연구보고서 2020-14

통계지리정보서비스(SGIS) 제공 자료의 비밀보호 적용 방안

홍영희 · 박민정

Statistics Korea
Statistics Research Institute

발간사

통계개발원(Statistics Research Institute, SRI)은 2006년에 개원한 이래 국내에 유일한 「국가통계 전문연구기관」으로서 국가통계・조사 방법론의 발전과 데이터기반 정책 연구・개발 분야에서 선도적인 역할을 하고 있습니다. 지난 한 해 동안 SRI는 "광출어혁(光出於革)"의 정신, 즉 혁신에(革) 기반한 실용적인 「팀연구」를 통해서 국가통계・데이터과학의 미래를 밝히고자(光) 노력하였습니다.

SRI의 「2020년 연구보고서」는 이러한 혁신연구의 결과물입니다. 특히 코로나19의 경제・사회 영향을 측정하는 지표 설계 연구를 비롯해서 「데이터기반 인구・사회・경제정책」을 뒷받침하는 심도 깊은 연구가 진행되었습니다. 예를 들어서 저혼인・저출산・초고령 사회에 대응하기 위한 데이터 심층 분석, 균형발전 정책수립을 지원하는 지역사회지표 체계구축, 그리고 SDG를 활용한 남북한 통합통계 방안 등입니다. 또한 2020년 2월에 출범한 「SDG 데이터연구센터」는 지속가능발전목표(SDG)의 이행현황 점검과 SDG 대국민 데이터 서비스를 준비하는 실용적인 연구를 진행하였습니다.

국가통계・데이터과학・조사방법론 분야에서는 「데이터 정보보호센터」를 중심으로 개인정보보호 방법의 통계지리정보서비스(SGIS) 적용 등 실용적인 연구가 진행되었습니다. 또한 「조사표연구센터」를 활용하여 빅데이터 활용 홈페이지 개선을 위한 사용성 평가를 비롯해서 「조사표 인지실험」을 적용한 국가통계의 품질개선이 이루어졌습니다. 특히 지난해에는 데이터과학(Data Science)을 국가통계에 활용하고자하는 기초 연구를 추진함으로써 「데이터 혁신 방법론」의 새로운 전기를 준비하였습니다.

금년에 개원 15주년을 맞는 SRI는 본 연구보고서가 증거기반정책 입안자의 데이터 활용에 실제적인 도움이 될 뿐만 아니라 국가통계 생산자의 혁신적인 조사방법론 개발에 유용한 자료로 활용될 수 있기를 기대합니다. SRI가 「국가통계 싱크탱크」로서 국내외적으로 선도적인 역할을 할 수 있도록 독자 여러분의 지속적인 격려를 부탁드립니다. 국가통계의 도약과 혁신적 개발을 위하여 제언이 있다면 언제든지 말씀하여 주십시오. 겸허히 귀 기울이겠습니다.

이번 「2020년 연구보고서」를 위하여 실용적이고 품질 높은 연구 결과를 도출하기 위해 광출어혁(光出於革)의 정신으로 최선을 다한 연구진과 대내외적으로 협력·공동연구에 참여한 민·관의 연구자들에게 따스한 감사를 전합니다. 끝으로 본 연구에서 제시된 내용 및 결과는 저자 개인의 의견이며, 통계청 또는 통계개발원의 공식적인 견해가 아님을 밝혀둡니다.

2021년 4월

통계청 통계개발원장

전 영 일

목 차

제1장 서론 ····· 1

제2장 SGIS 제공 자료와 노출 위험 ····· 2
제1절 SGIS 제공 자료 ····· 2
제2절 SGIS 자료의 노출 위험 ····· 9
제3절 빈도표에 대한 비밀보호 기법 ····· 14

제3장 SGIS 제공 빈도표의 비밀보호 방안 ····· 17
제1절 기저 빈도표에서의 비밀보호 방안 ····· 18
제2절 상위 빈도표에서의 비밀보호 방안 ····· 20
제3절 BSCA를 적용한 빈도표의 생성 ····· 37

제4장 비밀보호가 적용된 자료의 정합성 ····· 43
제1절 행정 구역 단위 자료의 정합성 ····· 43
제2절 격자 단위 자료의 정합성 ····· 56
제3절 현행 자료 제공 방식과의 비교 ····· 68

제5장 결론 및 기대효과 ····· 72

참고문헌 ····· 74

부 록 ····· 75
1. BSCA 알고리즘의 이해 ····· 75
2. 비밀보호가 적용된 SGIS 자료 안내서 ····· 101

Abstract ····· 122

요 약

본 연구에서는 통계지리정보서비스(SGIS)에서 제공하는 빈도표에 대하여 비밀보호 알고리즘이 적용되는 방법을 설명하였다. 해당 자료의 빈도표는 위계 구조를 가지고 있어 이를 기저 빈도표와 상위 빈도표로 구분하였다. 기저 빈도표에서는 랜덤 라운딩을 적용하고, 상위 빈도표에서는 중앙값 조정과 추가적인 조치를 통해 노출 위험을 제어할 수 있음을 예시를 활용하여 설명하였다. 또한 2018년 인구주택총조사 자료 중 대전 지역에 해당하는 자료에 비밀보호 알고리즘을 적용하여 빈도표를 생성하는데 소요되는 시간을 행정 구역과 격자 단위 자료로 구분하여 정리하였다.

다음으로 비밀보호 알고리즘 적용에 따른 빈도의 분포 변화를 행정 단위 및 격자 단위 빈도표로 나누어 검토하였다. 알고리즘 적용 후 기저 빈도표에서는 빈도가 1 또는 2인 셀의 비율이 없어지고, 대신 빈도가 0 또는 3인 셀의 비율이 증가하였다. 반면 상위 빈도표에서는 빈도가 3인 셀의 비율이 최대 11.2%p 증가하였고, 빈도가 4 이상인 셀의 비율은 5%p 내외로만 변화되어 알고리즘 적용 전과 상당히 유사한 분포를 보였다.

한편 알고리즘 적용으로 인한 정보손실은 필연적으로 발생하게 되는데, 그 범위는 3 이내임을 확인하였다. 특히 기저 빈도표에 가까운 형태일수록 정보손실이 발생하지 않을 확률이 높았다. 상위 빈도표에서는 정보손실이 1만큼 발생하는 경우가 가장 많았고, 최대 3만큼의 정보손실이 발생하는 셀 비율은 전체적으로 0.5%를 넘지 않는 것을 확인하였다.

따라서 이러한 정보손실을 빈도표에서의 노출 위험 제어를 위한 기회비용에 해당하는 것으로 이해할 수 있다면, 사용자에게 현행 SGIS의 자료 제공 방식에 비해 보다 많은 정보가 담겨있는 빈도표를 제공할 수 있을 것으로 기대된다.

주요용어 : 빈도표, 비밀보호 알고리즘, 노출 위험, 정보손실

제 1 장

서 론

2020년 온 지구는 코로나-19 바이러스로 인한 대혼란과 함께 변환의 시기를 맞이하고 있다고 해도 과언이 아닐 것이다. 많은 사람들이 구호처럼 외쳐왔던 제4차 산업혁명과 인공지능(AI)의 시대가 인간의 의지와는 다소 무관하게, 그러나 이전보다 훨씬 빠른 속도로 우리의 생활 속으로 찾아들었다. 학교 수업은 온라인 강의로 바뀌었고, 직장인의 재택근무가 보편화되는 등 반강제적인 디지털화가 진행되었고, 사회적 거리두기로 인한 비대면 인간관계가 사람의 생활과 습관을 바꾸어 가고 있다. 이렇듯 시대와 기술은 급변하는데 비해 우리의 적응 속도는 이를 따라가기에도 급급한 모습을 보이고 있다. 특히 방역 초기에 급증한 확진자의 동선을 공개하는 문제에 있어 개인의 프라이버시(privacy)보다는 감염병 예방이라는 공익적인 요인이 강조되었다. 이로 인해 확진자의 행적이 모두 공개될 뿐만 아니라, 해당하는 개인이 특정되는 노출(disclosure)이 발생할 수 있다는 취약점이 있었다. 이를 조금 더 확장시켜 생각해 보면 4차 산업혁명의 핵심이라고 할 수 있는 ICBM(IOT, Cloud, Bigdata, Mobile)에서도 강조되고 있는 (빅)데이터는 기하급수적으로 증가하는데 비해, 이를 적절하게 관리하고 활용하기 위한 기술과 제도 등에 있어 보완되어야 할 것이 있다고 하겠다. 데이터를 공개함에 있어 필요한 정보보호(data privacy)의 문제는 여러 측면에서 반드시 검토되어야 하며 이를 구현하기 위한 기술적 조치도 뒷받침되어야만 한다.

본 연구는 이러한 제반 사항과 맥락을 함께하고 있으며, 특히 통계지리정보서비스(SGIS)에서 빈도표(frequency table)의 형태로 제공되는 데이터에 대한 정보보호 문제에 대해 다루고자 한다. 이를 위해 제2장에서는 통계지리정보에서 제공하는 자료의 형태와 해당 자료에서 발생하는 노출 위험에 대해 설명하고, 선행 연구에 해당하는 빈도표에 대한 비밀보호 기법에 대해 간단히 리뷰하였다. 제3장에서는 빈도표에 대한 비밀보호 기법인 BSCA 알고리즘의 적용 절차에 대해 자세히 설명하였다. 제4장에서는 비밀보호 기법이 적용된 자료의 정합성을 검토하고, 현행 자료 제공 방식과 비교하였다. 마지막으로 BSCA 알고리즘의 원리와 코드에 대한 자세한 설명과 비밀보호 기법이 적용된 SGIS 자료 안내서를 부록에 제시함으로써 자료 제공자와 사용자 모두가 참조할 수 있도록 하여 편의를 돕고자 한다.

제 2 장

SGIS 제공 자료와 노출 위험

제1절 SGIS 제공 자료

1. SGIS 제공 자료의 형태

통계지리정보서비스(SGIS, Statistical Geographic Information Service)는 2006년 대전광역시를 대상으로 한 시범서비스를 시작으로, 2009년부터 대상을 전국으로 확대하여 서비스를 제공하고 있다. 현재 제공되고 있는 주요 서비스는 개인 생애주기와 관심분야에 초점을 맞춘 데이터 서비스인 My통계로(路), 인구·가구·주택·사업체 센서스 및 행정구역통계 등을 제공하는 대화형 통계지도, 생활·기술 업종별 통계를 제공하는 업종통계지도, 사회적 이슈 및 트렌드를 반영한 통계주제도 등으로 다양하다. 자료 사용자에게 필요한 통계 자료를 시각적 그래프와 함께 제공하여 편의성을 높이고자 하였다.

본 연구에서는 SGIS에서 제공하는 서비스 중 대화형 통계지도에 주목해 보려고 한다. 대화형 통계지도는 인구주택총조사, 농림어업총조사, 전국사업체조사 등의 전수 조사 자료에 대해 사용자가 원하는 항목을 선택하여 검색할 수 있는 맞춤형 소지역(small area) 통계를 서비스하고 있다. 대화형 통계지도에서는 조사연도에 따라 조금씩 다른 항목을 제공하고 있다. 각 항목들은 여러 개의 범주로 나뉘어져 있으며 사용자는 필요에 따라 항목을 조합하여 다양한 표를 만들어 볼 수 있다.

<표 2-1> SGIS 대화형 통계지도 인구주택총조사 제공 항목

조사년도	인구 조건	가구 조건	주택 조건
2000년, 2005년, 2010년	성별, 연령, 교육정도, 혼인상태	세대구성, 점유형태	주택유형, 건축연도, 연면적
2015년, 2016년, 2017년, 2018년	성별, 연령	세대구성	주택유형, 건축연도, 연면적

주: 2020.9월 현재 기준

본 연구에서는 2018년 인구주택총조사 자료를 중심으로 살펴보고자 하는데, 해당 조사에 대해서 성별, 연령, 세대구성, 주택면적, 주택유형, 건축연도 등 6개 항목에 대한 자료를 제공하고 있다.

<표 2-2> 2018년 인구주택총조사 자료의 제공 항목

제공 항목		선택지	범주 수
인구 조건	성별	남 여	2개 (다중 선택 가능)
	연령	0세 이상~ 5세 미만 5세 이상 ~ 10세 미만 … (추가) 95세 이상 ~ 100세 미만 100세 이상 ~	21개 (다중 선택 가능)
가구 조건	세대구성	1세대 가구 … (추가) 4세대 이상 가구 1인 가구 비혈연 가구	6개 (다중 선택 가능)
주택 조건	주택유형	단독주택 아파트 연립주택 대세대주택 비주거용 건물 내 주택	5개 (다중 선택 가능)
	건축연도	1979년 이전 1980년 ~ 1989년 1990년 ~ 1999년 2000년 ~ 2004년 2005년 ~ 2009년 2010년 … (추가) 2018년	14개
	주택면적	0㎡ 초과 ~ 20㎡ 이하 20㎡ 초과 ~ 40㎡ 이하 … (추가) 165㎡ 초과 ~ 230㎡ 이하 230㎡ 초과 ~	9개 (다중 선택 가능)

대화형 통계지도에서는 언급된 6개 항목에 대해 지역 수준별 자료를 조회해 볼 수 있다. 이때 지역 수준이라 함은 행정구역 수준과 격자 단위 수준을 의미하는 것으로 행정구역 단위란 집계구[1], 읍면동, 시군구, 시도 등으로 분류되고, 일종의 거리

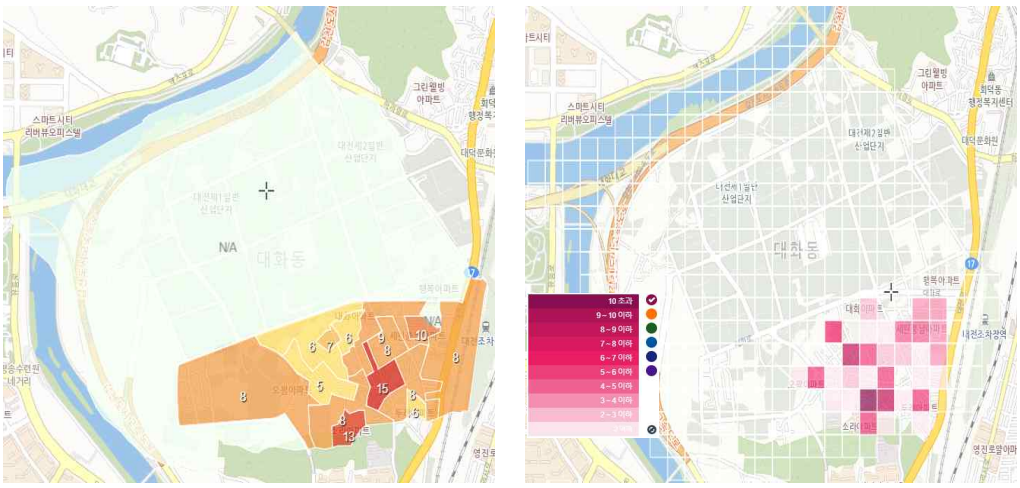
1) 기초단위구를 기반으로 인구규모(최적 500명), 사회경제적 동질성(주택유형, 지가), 집계구 형상을 고려하여 구축한 최소 통계 집계구역(출처 https://sgis.kostat.go.kr)

개념인 격자 단위에서는 100m, 1km, 10km, 100km 등으로 구분된다.

<표 2-3> SGIS 대화형 통계지도에서 제공하는 지역 수준(level)

level	OA	LA3	LA2	LA1
(행정구역 단위)	집계구	읍면동	시군구	시도
(격자 단위)	100m*100m	1km*1km	10km*10km	100km*100km

다음으로 SGIS에서 제공하는 자료의 형태를 살펴보도록 하겠다. 인구주택총조사 자료에서 대전광역시 대화동에 거주하는 10세 이상 ~ 15세 미만의 여자 인구 수를 조회해 보면, 해당 지역의 집계구 경계와 인구 수를 함께 볼 수 있다. 이때 화면 오른쪽에 표시되는 데이터 보기를 확대해 보면 집계구별 인구 수를 표 형태로 확인할 수 있다. 이러한 형태의 통계자료를 빈도표(frequency table)라 한다. 집계구별 빈도표에서는 5 미만의 빈도에 대해 “N/A”(not available)로 표시하여 자료를 제공하고 있다. 지도상에서 격자 단위의 자료 조회도 가능하며, 격자의 빈도는 구간 값(예, 11 초과 ~ 20 이하)으로 표시하고 있다. 현행 SGIS에서는 이러한 방법을 통해 자료에 대한 비밀보호 조치를 취하고 있다.



(1) 집계구 단위 자료 (2) 격자 단위 자료

<그림 2-1> SGIS 대화형 통계지도에서 제공하는 자료

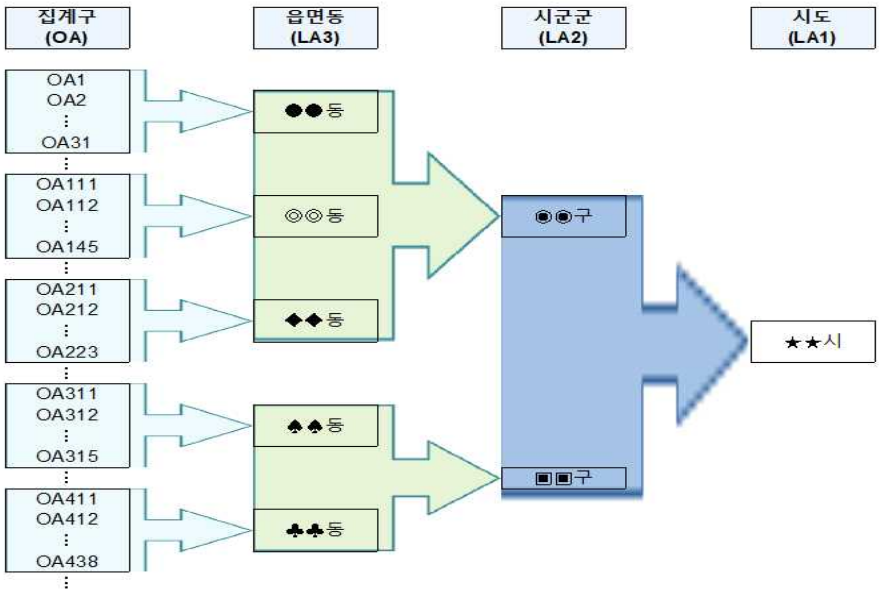
2. 빈도표의 위계(hierarchy) 구조

가. 빈도표에서 발생 가능한 위계 구조

SGIS에서 제공되는 자료를 이해하기 위해서는 빈도표의 위계(hierarchy) 구조를 파악할 필요가 있어 이에 관해 살펴보고자 한다.

SGIS 자료의 빈도표가 가지는 위계 구조는 2가지 측면에서 고려된다.

첫 번째는 지역 수준에 의한 위계 구조로 좁은 지역(small area) 수준에서 조사된 빈도(인구 수)를 집계하여 넓은 지역 수준의 빈도(인구 수)를 산출하는 경우에 발생한다. 예를 들어 대전광역시 서구 월평동의 경우에는 31개의 집계구가 모여 1개의 월평동이 되고, 11개의 동(洞)이 모여 대전광역시 서구를 형성하게 된다. 이와 유사하게 5개의 구(區)가 모여 대전광역시(市)를 이루는 위계 구조를 가진다.

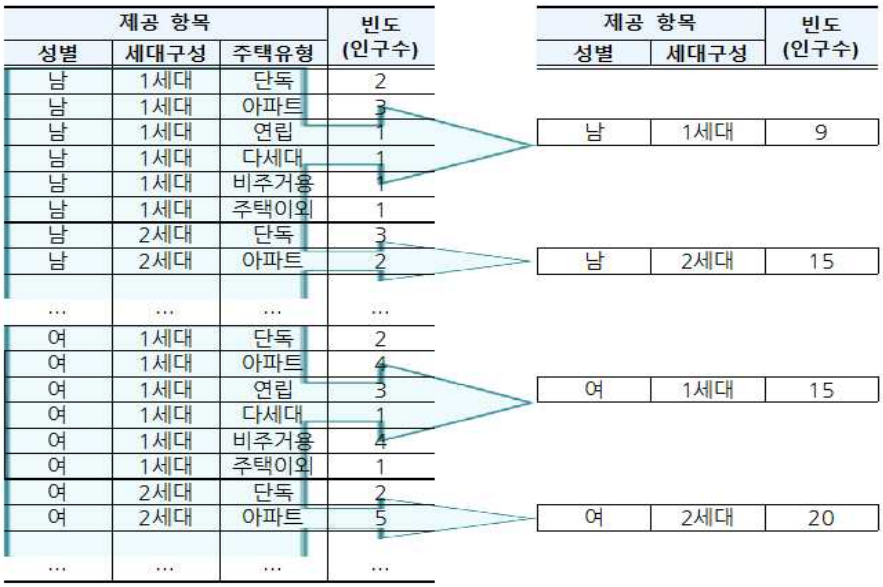


<그림 2-2> 지역 수준(행정 구역 단위)에 따른 위계 구조

위와 같은 행정 구역 단위에 의한 위계 구조는 격자 단위에서도 동일하게 적용된다. 100m 단위의 격자가 100개 모여서 1km 단위의 격자를 이루고, 1km 단위의 격자 100개가 모여 10km 단위의 격자를 이루는 등의 위계 구조를 보인다.

두 번째는 제공되는 항목 수에 의한 위계 구조로, 제공 가능한 모든 항목에서 조사된 빈도(인구 수)를 집계하여 항목의 수가 줄어든 경우의 빈도표를 생성함에 따라 발생한다.

예를 들어 제공하는 항목이 성별, 세대구성, 주택유형 등 3개 항목이라고 가정하자. 만약 자료 사용자가 성별과 세대구성에 따른 빈도(인구 수)를 요청한 경우, 자료 제공자는 주택유형별로 나누어진 빈도를 집계하여 제공하게 되므로 위계 구조가 생기게 된다.



제공 항목			빈도
성별	세대구성	주택유형	(인구수)
남	1세대	단독	2
남	1세대	아파트	3
남	1세대	연립	1
남	1세대	다세대	1
남	1세대	비주거용	1
남	1세대	주택이외	1
남	2세대	단독	3
남	2세대	아파트	2
…	…	…	…
여	1세대	단독	2
여	1세대	아파트	4
여	1세대	연립	3
여	1세대	다세대	1
여	1세대	비주거용	4
여	1세대	주택이외	1
여	2세대	단독	2
여	2세대	아파트	5
…	…	…	…

제공 항목		빈도
성별	세대구성	(인구수)
남	1세대	9
남	2세대	15
여	1세대	15
여	2세대	20

<그림 2-3> 제공 항목 수에 따른 위계 구조

나. 기저(base) 빈도표와 상위 빈도표

빈도표가 위계 구조를 가지고 있어 이를 기저(base) 빈도표와 상위 빈도표로 구분하게 된다. 기저 빈도표는 제공 가능한 모든 항목을 고려한 경우에 가장 좁은 지역(small area) 수준에서 만들어지는 빈도표에 해당한다. 상위 빈도표는 기저 빈도표를 집계하여 만드는데, 지역 수준이 넓어지거나 제공 항목의 수가 줄어드는 경우에 해당한다. 지역 수준이 넓어지고 동시에 항목의 수가 줄어드는 경우도 상위 빈도표에 해당한다.

예를 들어 2018년 인구주택총조사 자료에 해당하는 기저 빈도표는 성별, 연령, 세대구성, 주택유형, 주택면적, 건축연도 등 6개 항목을 모두 고려한 집계구 단위에서 만들어진 빈도표라고 이해할 수 있다. 이때 기저 빈도표는 행(row)의 수가 158,760개(=2×21×6×5×9×14)[2]개이고, 열(column)의 수는 집계구 수[3]만큼인 형태로 매우 많은 행과 열을 가지는 빈도표가 된다.

---

2) 항목별 선택지(범주)의 수는 성별(2개), 연령(21개), 세대구성(6개), 주택유형(5개), 주택면적(9개), 건축연도(14개) 이다.

3) 대한민국 전체 집계구 수는 총 106,347개임(2017년 인구주택총조사 기준)

<표 2-4> 인구주택총조사 자료의 기저 빈도표 예시

제공 항목						지역 수준				
성별	연령	세대 구성	주택 면적	주택 유형	건축 년도	1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	…
남	0~5세	1세대	~20㎡	다세대	2015년	3	0	0	0	…
여	0~5세	2세대	~40㎡	단독	2010년	10	4	6	14	…
남	5~10세	3세대	~60㎡	아파트	2018년	4	7	6	11	…
여	5~10세	4세대	~85㎡	연립	2018년	0	0	1	0	…
남	0~5세	1인	~100㎡	영업용	2010년	0	0	0	0	…
여	0~5세	비친족	~130㎡	비주거용	2019년	5	3	4	8	…
남	0~5세	1세대	~165㎡	다세대	2016년	5	5	5	11	…
…	…	…	…	…	…	…	…	…	…	…
남	100세+	1인	230㎡+	아파트	2016년	0	1	0	0	…
여	100세+	3세대	~100㎡	연립	2019년	2	0	0	0	…
남	100세+	4세대	~130㎡	연립	2016년	4	3	3	4	…
여	100세+	비친족	~165㎡	영업용	2016년	5	5	10	13	…
…	…	…	…	…	…	…	…	…	…	…
남	100세+	2세대	~60㎡	아파트	2016년	0	0	1	1	…
여	100세+	2세대	~100㎡	아파트	2019년	2	7	2	6	…
…	…	…	…	…	…	…	…	…	…	…

이러한 기저 빈도표에서 집계구별 자료를 집계하여 만든 다음의 표는 지역 수준이 읍면동으로 넓어진 상위 빈도표에 해당한다.

<표 2-5> 인구주택총조사 자료의 지역 수준이 넓어진 상위 빈도표 예시

제공 항목(6개)						지역 수준
성별	연령	세대구성	주택면적	주택유형	건축연도	ㄱㄴ동
남	0~5세	1세대	~20㎡	다세대	2015년	3
여	0~5세	2세대	~40㎡	단독	2010년	34
남	5~10세	3세대	~60㎡	아파트	2018년	28
여	5~10세	4세대	~85㎡	연립	2018년	0
남	0~5세	1인	~100㎡	영업용	2010년	0
여	0~5세	비친족	~130㎡	비주거용	2019년	20
…	…	…	…	…	…	…
남	100세+	1인	230㎡+	아파트	2016년	0
여	100세+	3세대	~100㎡	연립	2019년	0
남	100세+	4세대	~130㎡	연립	2016년	14
여	100세+	비친족	~165㎡	영업용	2016년	33
…	…	…	…	…	…	…
남	100세+	2세대	~60㎡	아파트	2016년	2
여	100세+	2세대	~100㎡	아파트	2019년	17
…	…	…	…	…	…	…

마찬가지로 기저 빈도표에서 성별 항목에 해당하는 남·여의 빈도를 집계하여 만들어진 다음의 표는 항목의 수가 6개에서 5개로 줄어든 상위 빈도표에 해당한다.

<표 2-6> 인구주택총조사 자료의 항목 수가 줄어든 상위 빈도표 예시

제공 항목(5개)					지역 수준				
연령	세대 구성	주택 면적	주택 유형	건축 년도	1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	…
0~5세	1세대	~20㎡	다세대	2015년	7	5	6	7	…
0~5세	2세대	~40㎡	단독	2010년	15	10	13	22	…
5~10세	3세대	~60㎡	아파트	2018년	10	14	14	20	…
5~10세	4세대	~85㎡	연립	2018년	7	8	10	10	…
0~5세	1인	~100㎡	영업용	2010년	8	9	10	11	…
0~5세	비친족	~130㎡	비주거용	2019년	14	13	15	20	…
0~5세	1세대	~165㎡	다세대	2016년	15	16	17	24	…
…	…	…	…	…	…	…	…	…	…
100세+	1인	230㎡+	아파트	2016년	13	15	15	16	…
100세+	3세대	~100㎡	연립	2019년	16	15	16	17	…
100세+	4세대	~130㎡	연립	2016년	19	19	20	22	…
100세+	비친족	~165㎡	영업용	2016년	21	22	28	32	…
…	…	…	…	…	…	…	…	…	…
100세+	2세대	~60㎡	아파트	2016년	18	19	21	22	…
100세+	2세대	~100㎡	아파트	2019년	21	27	23	28	…
…	…	…	…	…	…	…	…	…	…

기저 빈도표와 상위 빈도표의 구조를 간략하게 체계화하면 다음과 같은 형태로 정리할 수 있다. 지역 수준이 가장 넓은 범위인 LA1에서 제공 항목이 1개인 경우에 최상위 빈도표로 표시하였다.

<표 2-7> 기저 빈도표와 상위 빈도표의 위계 구조

지역 수준	OA	LA3	LA2	LA1
(행정 구역 단위)	집계구	읍면동	시군구	시도
(격자 단위)	100m*100m	1㎞*1㎞	10㎞*10㎞	100㎞*100㎞
제공 항목 6개	**기저 빈도**	→ 상위 빈도표	→ 상위 빈도표	→ 상위 빈도표
제공 항목 5개	↓상위 빈도표	↳ 상위 빈도표	↳ 상위 빈도표	↳ 상위 빈도표
제공 항목 4개	↓상위 빈도표	↳ 상위 빈도표	↳ 상위 빈도표	↳ 상위 빈도표
제공 항목 3개	↓상위 빈도표	↳ 상위 빈도표	↳ 상위 빈도표	↳ 상위 빈도표
제공 항목 2개	↓상위 빈도표	↳ 상위 빈도표	↳ 상위 빈도표	↳ 상위 빈도표
제공 항목 1개	↓상위 빈도표	↳ 상위 빈도표	↳ 상위 빈도표	↳ 최상위 빈도표

주: 화살표 방향은 집계 방향을 의미함.

제2절 SGIS 자료의 노출 위험

1. 빈도표에서의 노출 위험

가. 노출 위험의 개념

사용자의 요청에 의해 만들어진 빈도표에서는 노출의 위험이 존재한다. 이와 관련한 이론적 배경은 Hundepool 외(2010), 박민정과 성주현(2016), 박민정 외(2018)에서 자세히 다루고 있어 참고할 수 있다. 본 연구에서는 빈도표에서 발생할 수 있는 노출 위험에 대해 예시를 통해 살펴봄으로써 이해를 돕고자 한다.

선행 연구에서 밝혔듯이 통계표의 특정 셀(cell)에 속하는 개체가 하나 또는 두 개일 경우에 노출의 위험이 높다고 판단한다. 자료 제공자 입장에서는 사용자가 만든 빈도표에서 특정 빈도(인구 수)가 1명 또는 2명인 경우에 노출의 위험이 있다고 판단하게 된다.

예를 들어, 대전광역시 서구 둔산2동에 거주하는 사용자 A가 있다고 가정해 보자. 사용자 A는 자신과 비슷한 환경에 거주하는 인구 수가 얼마나 되는지를 궁금하게 여겨 SGIS의 인구주택총조사 자료를 통해 알아보기로 했다. 이때 사용자 A의 특성은 남자이면서 1인 가구이고 주택면적이 20 ㎡ 이하인 아파트에 거주하고 있다고 하자.

<표 2-8> 노출 위험이 있는 빈도표 예시1

성별	세대 구성	주택유형	주택면적	빈도(인구 수)
남자	1인 가구	아파트	20㎡ 이하	1

사용자 A가 만든 빈도표가 위와 같다면, 사용자 A는 대전시 서구 둔산2동에 거주하면서 해당 특성을 가진 사람은 본인밖에 없다는 사실을 알아차리게 됨과 동시에 노출 위험에 처하게 된다. 다시 말해 빈도표에서 조사된 빈도(인구 수)가 1명이었기 때문에 사용자 A는 노출 위험에 처하게 된다.

다음으로 동일한 항목에 대해 조사된 빈도(인구 수)가 2명이라고 가정해 보자. 이때 빈도(인구 수) 2명에 해당하는 사람이 사용자 C와 D라고 한다면, 각 개인은 빈도(인구 수) 2명 중 1명이 본인이라는 사실을 깨닫게 되고 나머지 1명은 상대방이 된다는 것을 알게 된다. 다시 말해 빈도표에서 조사된 빈도(인구 수)가 2명이었기 때문에 사용자 C와 D는 상호간에 식별 위험에 노출되게 된다.

<표 2-9> 노출 위험이 있는 빈도표 예시2

성별	세대 구성	주택유형	주택면적	빈도(인구 수)
남자	1인 가구	아파트	20㎡ 이하	2

이와 같이 빈도표에서 특정 셀에 속하는 빈도의 값이 1 또는 2인 경우 노출의 위험이 있다고 판단한다면, 빈도표를 제공해야 하는 SGIS 자료 제공자는 빈도가 얼마나 큰 값일 때 그 값을 제공할 수 있을지에 대해 고민하게 된다. 현재까지 논의된 바에 의하면 노출 위험이 없다고 판단할 수 있는 절대적인 기준이 있다기보다는, 자료의 특성에 따라 기준이 되는 값은 달라질 수 있다고 보는 시각이 우세한 편이다. 현재 SGIS에서는 5명 미만의 빈도(인구 수)는 노출 위험이 큰 것으로 판단하여, 해당 값에 대해 N/A 표시하여 제공하고 있다.

나. 노출 위험의 유형

앞서 다룬 빈도표에서의 노출 위험에 대해 Hundepool 등(2012)은 노출의 유형을 다음과 같이 정리한 바 있다.

<표 2-10> 빈도표에서 발생 가능한 노출의 유형

유형	설명
개별 속성 노출 (individual attributes disclosure)	빈도표의 부분 합(margin)이 1인 경우
그룹 속성 노출 (group attributes disclosure)	열이나 행의 빈도가 대부분 0이거나 작은 값인 경우
차이에 의한 노출 (disclosure by difference)	포함관계인 두 빈도표의 차이로 인해 특정 셀 값이 도출되는 경우
표 연계에 의한 노출 (disclosure by linking tables)	하나의 마이크로데이터에서 여러 개의 빈도표가 만들어지는 경우

주: 박민정과 성주현(2016)에서 재인용함

빈도표에서의 노출 유형에 대한 이해를 돕기 위해 예를 들어 보고자 한다.

만약 자료 사용자가 가나다광역시 ABC구 ㄱㄴ동에 거주하는 0세 이상 ~ 20세 미만의 남·여별 인구 수를 요청한 경우에 제공되는 빈도표가 다음과 같았다고 하자.

<표 2-11> 노출 위험이 있는 빈도표 예시1

성별	연령(5세 단위 제공)	인구 수	소계
남	0세 이상 ~ 5세 미만	1	55
	5세 이상 ~ 10세 미만	14	
	10세 이상 ~ 15세 미만	40	
	15세 이상 ~ 20세 미만	0	
여	0세 이상 ~ 5세 미만	0	2
	5세 이상 ~ 10세 미만	0	
	10세 이상 ~ 15세 미만	0	
	15세 이상 ~ 20세 미만	2	
소계		57	57

<표 2-11>에 제시된 빈도표에서는 2가지 유형의 노출 위험이 존재한다.

첫 번째는 0세 이상~5세 미만의 남자 빈도(인구 수)가 1명이기에 개별 속성 노출이 발생한다. 두 번째는 20세 미만의 여자는 연령이 모두 15세 이상~20세 미만에 해당한다는 그룹 속성 노출이 발생한다.

또, 사용자는 다양한 항목을 조합하여 상당히 많은 수의 빈도표를 만들 수 있다. 이처럼 특정 지역에 대해 동일한 항목의 조건을 달리하여 여러 개의 빈도표를 만들고 이를 조합해 봄으로써 표 연계에 의한 노출과 함께 차이에 의한 노출이 동시에 발생한다.

만약 자료 사용자가 가나다광역시 ABC구 ㄱㄴ동에 거주하는 남자 인구 수를 조사하고 있는데, 연령을 다르게 하여 2개의 빈도표를 만들었다고 하자. 이 경우 사용자는 2개의 빈도표를 조합해 봄으로써, 가나다광역시 ABC구 ㄱㄴ동에 거주하는 40세 이상 ~ 45세 미만의 남자 인구가 1명임을 알아차리게 된다.

<표 2-12> 노출 위험이 있는 빈도표 예시2

성별	연령(5세 단위 제공)	소계
남	40세 이상 ~ 60세 미만	55

성별	연령(5세 단위 제공)	소계
남	45세 이상 ~ 60세 미만	54

2. SGIS 제공 자료의 노출 위험

빈도표에서 발생할 수 있는 노출의 개념과 유형에 대해 살펴보았다. 그렇다면 빈도표에서 발생하는 노출 위험을 측정하는 방법은 무엇이 있을까? 이와 관련한 선행연구로 박민정과 성주현(2016)을 통해 기존에 제안된 노출 위험 척도와 빈도표의 노출 위험 척도가 갖추어야 할 성질에 대해 참고할 수 있다. 그러나 학계의 노력에도 불구하고 기존에 제안된 노출 위험 측도를 실무에 적용하기에는 해결해야 할 과제가 많이 있다(Shlomo 등, 2015).

이후 박민정 외(2018)에서는 노출 위험을 k-익명성(k-anonymity)으로 측정하였다. 빈도표에서의 k-익명성이란 특정 셀에 속하는 빈도 값으로 측정하게 되고, 그 빈도가 작은 경우에 노출 위험이 있다고 판단하게 된다. 앞서 노출 위험의 개념을 설명할 때, 빈도표의 특정 셀의 값이 1 또는 2인 경우 노출 위험이 있다고 판단한 것과 같은 맥락으로 이해할 수 있다. 여기서 작은 빈도를 갖는 셀에 해당하는 기준 값을 기호 B로 표시하고, 노출 위험은 $1/B$로 측정하였다. 예를 들어 $B=3$으로 정하게 되면 작은 빈도를 갖는 셀은 빈도가 0, 1, 2인 경우에 해당하며, 노출 위험은 $1/3$이 된다. 작은 빈도를 갖는 셀의 기준인 B의 값을 얼마로 하느냐에 대해 다양한 의견이 있을 수 있다. 통계작성기관에서 통상적으로 사용하는 기준으로 사업체 자료는 $B=3$, 인구 및 가구 자료는 $B=5$ 이다. 현재 SGIS에서는 인구 및 가구 자료에 대해서도 5 미만의 값을 N/A 표시하여 제공하고 있다.

SGIS 자료에서 발생하는 노출 위험에 대해 살펴보고자 한다. 대전광역시의 2018년 인구주택총조사 자료를 중심으로 기저 빈도표에서의 노출 현황을 분석해 보았다. 인구·가구·주택에서 제공하는 6개 항목을 모두 고려하고, 지역 수준은 집계구 단위로 한 기저 빈도표의 빈도 분포를 조사하였다.

2018년 당시 대전광역시의 집계구는 총 2,997개이며, 6개 항목을 모두 고려했을 때 생성되는 선택지 조합의 수는 총 158,760개(=2×21×6×5×9×14)[4]이다. 다시 말해 차원(dimension)이 158,760×2,997인 빈도표를 만들어 보면 노출 위험이 있는 셀의 규모를 파악할 수 있다. 해당 빈도표를 만들기 위해 생성해야 할 셀의 수는 475,803,720개(=158,760×2,997)에 달한다.

4) 항목별 선택지(범주)의 수는 성별(2개), 연령(21개), 세대구성(6개), 주택유형(5개), 주택면적(9개), 건축연도(14개) 이다.

<표 2-13> 2018년 인구주택총조사 자료의 기저 빈도 분포(대전시 기준, 집계구 단위)

	빈도						소계
	0	1	2	3	4	5 이상	
셀의 수	475,252,525	339,358	88,062	36,723	19,838	67,214	475,803,720
비율(%)	99.88	0.07	0.02	0.01	0.004	0.014	100

위 표에서 알 수 있는 바와 같이 사용자가 6개 항목을 모두 고려한 집계구 단위의 기저 빈도표를 고려했을 때, 빈도가 1 또는 2의 값을 가진 셀은 대전광역시 기준 대략 0.1% 수준이며 거의 대부분의 빈도는 0임을 확인할 수 있다.

이번에는 지역 수준을 100m 단위의 격자로 하여 인구·가구·주택에서 제공하는 6개 항목을 모두 고려했을 때의 빈도 분포를 살펴보았다. R에서 계산 가능한 범위로 한정하기 위해 대전광역시 지역에 해당하는 격자를 대상으로 하였으며, 이때 고려해야 할 100m 단위격자의 수는 10,060개이다. 따라서 해당 빈도표 생성을 위해 만들어야 하는 셀의 수는 1,597,125,600개(=158,760×10,060)이다.

<표 2-14> 2018년 인구주택총조사 자료의 기저 빈도 분포(대전시 기준, 100m 단위격자)

	빈도						소계
	0	1	2	3	4	5 이상	
셀의 수	1,596,514,330	408,845	87,335	34,524	18,938	61,628	1,597,125,600
비율(%)	99.962	0.026	0.005	0.002	0.001	0.004	100

사용자가 6개 항목을 모두 고려한 100m 단위격자의 빈도표를 요청한 경우, 빈도가 1 또는 2인 경우는 대전광역시 기준 약 0.03%임을 알 수 있다. 집계구 단위 빈도표와 비교해 보면 빈도가 0인 경우의 비율은 더 높은데 비해 빈도가 1 이상인 경우의 비율은 더 낮음을 확인할 수 있다.

정리하자면 기저 빈도표는 집계구 단위와 100m 단위격자에서 모두 매우 희박한(sparse) 형태를 띠고 있음을 알 수 있다. 앞서 다루었던 노출 위험의 유형 중에서 개별 속성 노출과 그룹 속성 노출 위험이 발생할 가능성을 확인할 수 있다. 만약 자료 제공자가 6개 이상의 항목을 제공하거나, 특정 항목의 선택지 수가 늘어나게 되면 빈도표는 더욱 희박한 형태를 보이게 될 것이고, 노출 위험의 증가는 필연적으로 동반된다.

제3절 빈도표에 대한 비밀보호 기법

1. 작은 셀 조정(SCA)

빈도표에서의 노출 위험을 방지하기 위한 방법들을 통계개발원에서는 지속적으로 연구해 왔다. 정동욱과 김경미(2008)에서는 집계표에서 민감한 셀을 찾는 방법과 전통적인 비밀보호 기법을 소개하였다. 반올림방법(rounding), 셀 감추기(cell suppression), CTA방법(controlled tabular adjustment) 등에 대해 자세히 소개하였고, 광업 및 제조업 통계조사의 집계표를 활용해 전통적인 비밀보호 기법들의 한계점을 밝혔다. 빈도표에서의 노출 제어와 관련한 해외 사례는 박민정과 성주현(2016)을 통해 확인할 수 있다. 특히 영국 통계청 ONS(Office for National Statistics)에서 사용하는 라운딩 기법을 활용한 SCA(small cell adjustments) 방법에 대한 자세한 설명을 참조할 수 있다.

본 연구에서 활용되는 빈도표에 대한 비밀보호 기법인 BSCA(Bounded small cell adjustments) 방법은 SCA 방법을 수정·보완한 형태로 이해할 수 있을 것이다. 따라서 이후 소개되는 BSCA 방법의 이해를 돕고자 SCA 방법에 대해 간략히 살펴보도록 하겠다.

SCA 방법은 가장 하위 수준의 빈도표에서 특정 셀의 빈도 값이 작은 경우 랜덤 라운딩(random rounding, 임의 반올림)을 적용하게 된다. 이때 가장 하위 수준의 빈도표라 함은 제공 가능한 모든 변수를 고려한 가장 좁은 지역(small area) 수준에서 만들어지는 기저 빈도표이다. 이 기저 빈도표에서 빈도 값이 1 또는 2인 경우에 다음과 같은 확률로 빈도 값을 조정하게 되는데, 이를 랜덤 라운딩이라고 한다. 즉, 작은 빈도를 갖는 셀에 대한 판단 기준을 $B = 3$으로 하였다. 기저 빈도에 대한 이러한 랜덤 라운딩 과정은 빈도의 총합에 대한 기댓값(expected value)을 그대로 보존할 수 있다는 장점이 있다.

<표 2-15> 기저 빈도(인구 수)의 랜덤 라운딩 과정

기저 빈도		반올림한 빈도	기저 빈도		반올림한 빈도
1	(2/3 확률) →	0	2	(1/3 확률) →	0
	(1/3 확률) →	3		(2/3 확률) →	3

상위 빈도표를 만들 때는 랜덤 라운딩이 적용된 값을 그대로 집계하여 계산하게 된다. 이렇게 반올림한 기저 빈도표를 집계한 상위 빈도표는 반올림하지 않은 기저 빈도표를 집계한 상위 빈도표의 참 값과 차이가 발생하게 되는데, 이를 정보손실

(information loss)이라고 한다. 정보손실은 최상위의 빈도표인 가장 넓은 지역 수준에 대한 빈도표 또는 항목의 수가 1개만 있는 빈도표에서 그 정도가 가장 커진다. 이러한 상황은 위계 구조가 있는 빈도표의 제공에 있어 큰 걸림돌이 될 수 있다.

SCA 방법의 또 한 가지 단점으로는 상위 빈도표를 집계하는 경로에 따라 서로 다른 값을 가지는 상위 빈도표를 생성하게 될 수도 있다는 것이다. 다시 말해 항목의 수를 먼저 줄인 상위 빈도표를 집계하여 보다 넓은 지역의 상위 빈도표를 생성한 결과와 먼저 지역 수준을 넓힌 상위 빈도표를 집계하여 항목의 수를 줄인 상위 빈도표를 생성한 결과가 다를 수 있다는 것이다.

이러한 SCA 방법의 단점을 보완하고자 BSCA(Bounded small cell adjustments) 알고리즘이 박민정 외(2018)에서 제안되었다.

2. 제한적 작은 셀 조정(BSCA)

박민정과 성주현(2016)에서는 SGIS 자료와 같이 위계 구조를 가지는 빈도표의 비밀보호를 위한 기법이 갖추어야 할 바람직한 성질에 대해 언급한 바 있다. 구체적으로는 가산성(additivity), 일치성(consistency), 정보손실(information loss)의 최소화, 노출위험(disclosure risk)의 제어 등이 필요하다고 하였다. 그러나 이러한 조건을 모두 만족시키면서 실무적 차원에서 활용 가능한 방법은 많지 않은 것으로 파악하였다.

박민정 외(2018)에서 제안된 BSCA(Bounded small cell adjustments) 알고리즘은 SCA 방법의 단점인 일치성 결여와 정보손실이 최소화되지 못하는 것을 보완하려 하였다. 기저 빈도표에 대한 랜덤 라운딩 방식은 그대로 유지하고, 상위 빈도표 생성 시에는 랜덤 라운딩한 기저 빈도표를 집계하여 만들되 정보손실과 노출 위험을 일정 수준 이하로 제어할 수 있도록 하였다.

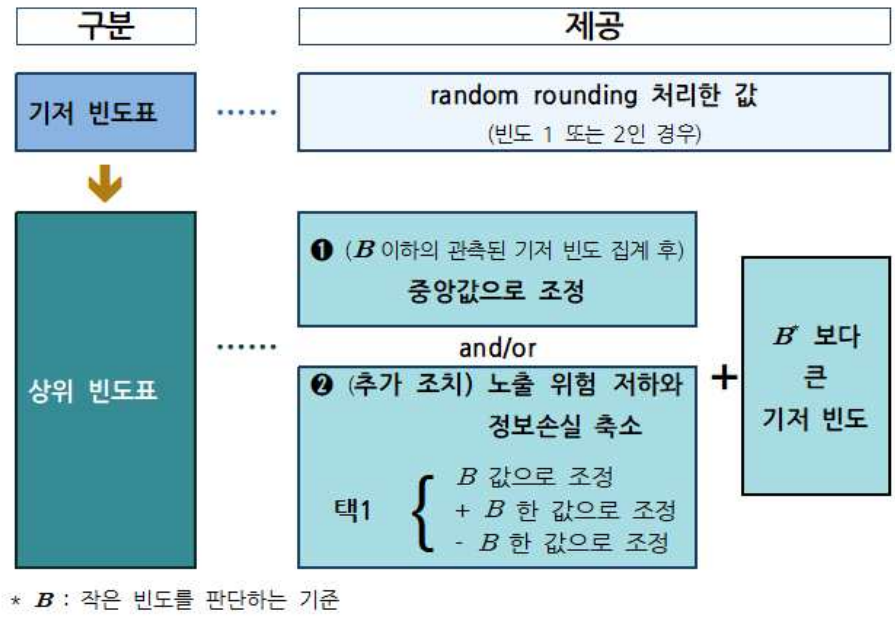
BSCA 알고리즘의 개략적인 흐름은 <그림 2-4>와 같다. 기저 빈도표와 상위 빈도표로 구분하여 각기 다른 방법을 적용하도록 하였다. 앞서 기저 빈도표는 제공 가능한 모든 항목을 고려한 가장 좁은 지역 수준에서의 빈도표로 매우 희박한(sparse) 형태임을 확인한 바 있다. 즉 조사된 대부분의 빈도가 0(zero) 값을 가지며, 빈도가 1 또는 2인 경우가 존재함으로 기저 빈도표 전체에서 노출의 위험은 거의 언제나 발생하게 마련이다.

만약 자료 제공자가 작은 빈도를 갖는 셀의 판단 기준을 $B = 3$으로 정하고, 기저 빈도표에서 빈도가 0인 경우에는 그 값을 그대로 제공한다고 가정해 보자. 그렇다면 기저 빈도표에서 비밀보호가 필요한 경우는 빈도가 1 또는 2인 경우로 해당 빈도에 대해 랜덤 라운딩(random rounding)을 적용하는 비교적 간단한 방법으로 노출 위험에

관한 문제를 해결할 수 있다.

이에 비해 상위 빈도표에서의 비밀보호 방법은 다소 복잡한 흐름을 가지고 있다. 먼저 기저 빈도표에서 작은 빈도를 집계하여 중앙값 조정이라는 단계를 기본적으로 거치게 되는데, 이를 통해 노출 위험을 제어할 수 있다. 그러나 모든 경우에 중앙값 조정으로 노출 위험이 제어되는 것은 아니어서, 노출 위험 정도를 추가적으로 검토하게 된다. 추가로 검토된 노출 위험에 따라 3가지의 처리 절차 중에서 1가지의 처리를 선택하여 취하게 되는데, 중앙값 조정을 거친 빈도 자체를 B 값으로 바꾸거나, 중앙값 조정을 거친 빈도에 B만큼 더하거나 빼는 절차 중 1가지를 취하게 되는 것이다.

이러한 절차를 거쳐 제공되는 상위 빈도표는 실제 조사된 빈도의 참 값을 정확하게는 유추할 수 없게 되고, 노출 위험은 $1/B$ 이하로 제어된다. 또 실제 조사된 빈도 참 값과의 차이인 정보손실을 3 이하로 제어할 수 있게 된다.



<그림 2-4> 빈도표에 대한 비밀보호 알고리즘 개요

제3장에서 BSCA 알고리즘의 세부적인 절차와 그에 따른 노출 위험의 제어, 정보손실 정도에 관해 살펴보고자 한다.

제 3 장

SGIS 제공 빈도표의 비밀보호 방안

SGIS에서 제공되는 빈도표 중에는 노출의 위험이 없어 비밀보호를 위한 특별한 조치가 필요하지 않는 경우도 존재한다. 예를 들어 인구밀도가 높은 도시지역의 인구 수를 집계구별로 제공한다거나, 사업체가 밀집해 있는 산업단지에서 사업체 수를 1km 단위의 격자로 제공하는 경우에는 빈도가 충분히 커서 노출 위험이 없다고 판단할 수 있다. 반면 빈도표 제공 시 노출의 위험이 있어 비밀보호를 위한 조치를 해야 하는 경우도 어렵지 않게 찾을 수 있다. 인구밀도가 낮은 농어촌지역의 인구 수를 100m 단위격자로 제공한다거나, 성별, 연령, 교육정도 등 다양한 항목 조건에 따른 집계구별 인구 자료를 제공하는 경우에는 노출의 위험이 발생하게 된다.

본 연구에서는 통계지리정보에서 제공하는 빈도표에서 비밀보호가 필요한 경우에 한정하여 보았다. 비밀보호가 필요한 경우를 행정 구역 단위와 격자 단위로 나누어 생각해 볼 수 있다. 먼저 행정 구역 단위에서 비밀보호가 필요한 경우로는 집계구, 읍면동, 시군구 등에 대한 빈도를 제공할 때, 제공하는 항목의 수가 많아질수록 노출의 위험이 증가한다고 볼 수 있다. 또 격자 단위에서 비밀보호가 필요한 경우로는 100m 단위 격자의 크기(100m×100m)가 작기 때문에, 제공하는 항목의 수가 많지 않아도 노출의 위험이 크다고 볼 수 있다. 따라서 SGIS에서 제공 가능한 모든 빈도표에 대한 노출 위험을 정확한 수치로 표현하는 것은 어렵다 할지라도, 많은 경우에 노출 위험 제어를 위한 조치가 필요하다고 할 수 있다.

물론 경우에 따라서는 노출 위험을 감수하더라도 매우 정확한 자료의 제공이 필요한 사안이 있을 수 있으나, 이러한 문제에 대한 것은 논외로 한다. 본 연구에서는 빈도표를 제공할 때 BSCA 알고리즘 적용을 통한 비밀보호가 어떠한 방식으로 적용되는가에 대해 자세히 다루고자 하였다.

위계 구조를 가지는 빈도표에 대한 비밀보호 방안으로 제안된 BSCA(Bounded Small Cell Adjustments) 알고리즘은 노출 위험을 $1/B$[5]로 제어하면서 실제 조사된 빈도의 참 값과의 차이인 정보손실은 3 이내가 되도록 하였다. 만약 $B=3$으로 정한다

5) B는 작은 빈도를 갖는 셀로 판단하는 기준 값

면, 빈도가 1 또는 2인 경우에 노출 위험이 있는 것으로 판단하여 노출 위험이 1/3 이하로 제어되도록 빈도를 조정하여 제공하는 방법이다.

제1절 기저 빈도표에서의 비밀보호 방안

앞서 여러 번 언급되었듯이, 기저 빈도표란 제공 가능한 모든 항목을 고려했을 때 가장 좁은 지역 수준에서 만든 빈도표이다. 여기서부터는 BSCA 알고리즘의 보다 상세한 이해를 돕고, 논의의 편의를 위해 기저 빈도표가 다음과 같다고 가정하겠다. 제공되는 항목은 성별과 주택유형 2가지이고, 집계구가 5개만 있다고 하자. 보다 현실성 있는 상황은 실제 SGIS에서 제공하는 6개의 항목과 전국의 모든 집계구를 반영한 기저 빈도표를 고려해야 하고, 집계구별 인구 수가 500명 내외가 되도록 해야 하지만 BSCA 알고리즘의 상세한 논의를 위해 축약하여 가정하였다.

<표 3-1> 조사된 기저 빈도표 예시

조사된 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	1	0	0	0	13
	아파트	18	3	2	0	36
	연립주택	4	4	1	7	22
	다세대	5	8	0	18	12
	비주거용	1	2	0	4	15
여	단독주택	4	5	8	3	1
	아파트	9	4	9	2	1
	연립주택	5	10	14	5	1
	다세대	10	6	7	8	1
	비주거용	4	5	11	11	1

위와 같이 조사된 기저 빈도표에서 빈도(인구 수)가 1명 또는 2명인 경우는 노출의 위험이 있기 때문에 조사된 빈도(인구 수)의 값을 그대로 제공하는 것은 문제가 될 수 있다. 이 경우 조사된 빈도(인구 수) 1 또는 2를 0 또는 3으로 바꾸어 제공해야만 노출의 위험이 제어되며, 그 가능성은 확률에 의해 결정된다. 이러한 과정을 랜덤 라운딩(random rounding, 임의 반올림)이라고 하였다.

<표 3-2> 조사된 빈도(인구 수)의 랜덤 라운딩 과정

조사된 빈도			반올림한 빈도	조사된 빈도			반올림한 빈도
1	{	(2/3 확률) →	0	2	{	(1/3 확률) →	0
		(1/3 확률) →	3			(2/3 확률) →	3

조사된 빈도(인구 수)는 랜덤 라운딩 과정을 거쳐 <표 3-3>과 같은 형태로 사용자에게 제공될 수 있다. 조사된 기저 빈도표에서 1명 또는 2명이었던 빈도가 모두 0명 또는 3명으로 바뀌어 노출 위험이 제어된 것으로 판단할 수 있다.

<표 3-3> 조사된 기저 빈도표와 반올림한 기저 빈도표 비교

조사된 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	1	0	0	0	13
	아파트	18	3	2	0	36
	연립주택	4	4	1	7	22
	다세대	5	8	0	18	12
	비주거용	1	2	0	4	15
여	단독주택	4	5	8	3	1
	아파트	9	4	9	2	1
	연립주택	5	10	14	5	1
	다세대	10	6	7	8	1
	비주거용	4	5	11	11	1

반올림한 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	0	0	0	0	13
	아파트	18	3	3	0	36
	연립주택	4	4	3	7	22
	다세대	5	8	0	18	12
	비주거용	0	3	0	4	15
여	단독주택	4	5	8	3	3
	아파트	9	4	9	3	3
	연립주택	5	10	14	5	3
	다세대	10	6	7	8	3
	비주거용	4	5	11	11	3

주: 조사된 기저 빈도표에서 조사된 빈도가 1 또는 2인 경우는 음영으로 표시됨.

위의 반올림한 기저 빈도표를 사용자에게 제공하는 경우에 정보손실을 고려해 보자. 정보손실은 조사된 기저 빈도와 반올림한 기저 빈도의 차이를 고려하는 것으로 조사된 빈도(인구 수)가 1 또는 2인 경우에만 최대 2만큼씩 발생한다. 또 조사된 빈도(인구 수)가 0명이거나 3명 이상인 경우에는 정보손실이 없으므로 사용자 입장에서 거부감이 크지 않을 수 있다.

따라서 자료 제공자는 기저 빈도표에 대한 사용자의 요청이 있는 경우 반올림한 기저 빈도표를 제공하도록 하며, 보다 정확한 정보의 전달을 위해 다음의 사항도 함께 공지하는 것을 고려해 볼 수 있다.

[공지] 공개된 자료를 통해 발생할 수 있는 개인정보보호를 위해 다음과 같이 조치하였습니다.

제공 받은 자료의 빈도(인구수)	실제 조사된 빈도(인구수)
3명	1, 2, 3명 中 1개 값
0명	0, 1, 2명 中 1개 값

<그림 3-1> SGIS 자료의 비밀보호를 위한 조치 사항 알림(예시)

다음으로 제2절에서는 상위 빈도표에서 비밀보호 알고리즘이 어떻게 적용되는지에 대해 자세히 다루고자 한다.

제2절 상위 빈도표에서의 비밀보호 방안

1. 제공 가능한 상위 빈도표에 대한 논의

상위 빈도표란 기저 빈도표에 비해 항목의 수가 줄어들거나, 지역 단위가 넓어지는 경우를 의미한다. 사용자가 상위 빈도표를 요청하는 경우에는 기저 빈도표를 항목별로 또는 지역 단위별로 집계하여 제공하게 된다. 상위 빈도표에서의 비밀보호 기법을 이해하기 위해 앞서 예로 제시된 기저 빈도표에서 항목의 수가 줄어드는 경우의 상위 빈도표를 자세히 다루고자 한다. 지역 단위가 넓어지는 경우의 상위 빈도표에서도 동일한 방식의 기법이 적용되는 것으로 이해할 수 있다.

앞서 제1절에서 사용자의 요청이 있는 경우 반올림한 기저 빈도표를 제공하기로 했으니, 상위 빈도표 제공 시에도 반올림한 기저 빈도표를 집계한 상위 빈도표를 그

대로 제공해도 무방한지에 대해 생각해 보려 한다.

예를 들어 사용자가 성별에 따른 빈도(인구 수)를 요청하였다고 가정하자. 이 경우에는 주택 유형별로 세분된 빈도(인구 수)를 남·여별로 집계하여 제공하면 된다. 구체적으로, 1번째 집계구에 거주하는 남·여의 빈도(인구 수)는 <표 3-4>와 같이 집계된다. 이러한 과정을 거쳐 구해진 성별에 따른 상위 빈도표는 <표 3-5>와 같다.

<표 3-4> 상위 빈도표 작성 예시

조사된 기저 빈도		1st 집계구
성별	주택 유형	
남	단독주택	1
	아파트	18
	연립주택	4
	다세대	5
	비주거용	1
여	단독주택	4
	아파트	9
	연립주택	5
	다세대	10
	비주거용	4

조사된 상위 빈도	1st 집계구
성별	
남	29
여	32

<표 3-5> 조사된 기저 빈도표를 집계한 상위 빈도표

조사된 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	1	0	0	0	13
	아파트	18	3	2	0	36
	연립주택	4	4	1	7	22
	다세대	5	8	0	18	12
	비주거용	1	2	0	4	15
여	단독주택	4	5	8	3	1
	아파트	9	4	9	2	1
	연립주택	5	10	14	5	1
	다세대	10	6	7	8	1
	비주거용	4	5	11	11	1

조사된 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	29	17	3	29	98
	여	32	30	49	29	5

조사된 기저 빈도표를 집계한 경우와 반올림한 기저 빈도표를 집계한 경우를 비교해 보자. 조사된 상위 빈도와 반올림한 상위 빈도는 얼마나 큰 차이를 보이는가? 1, 2, 3번 집계구의 여자 빈도(인구 수)처럼 동일한 경우도 있지만, 5번 집계구의 여자 빈도(인구 수)처럼 10만큼 차이가 날 수도 있다. 그 배경으로는 조사된 기저 빈도표에서 1 또는 2의 값을 가지는 셀의 수가 얼마나 되는가와 랜덤 라운딩된 값이 무엇인가에 따라 비교적 근사한 값을 보이거나, 아주 큰 차이를 보일 수도 있게 된다.

<표 3-6> 반올림한 기저 빈도표를 집계한 상위 빈도표

반올림한 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	0	0	0	0	13
	아파트	18	3	3	0	36
	연립주택	4	4	3	7	22
	다세대	5	8	0	18	12
	비주거용	0	3	0	4	15
여	단독주택	4	5	8	3	3
	아파트	9	4	9	3	3
	연립주택	5	10	14	5	3
	다세대	10	6	7	8	3
	비주거용	4	5	11	11	3

⇛

반올림한 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	27	18	6	29	98
	여	32	30	49	30	15

여기서 상기해야 할 사항은 노출 위험을 줄이기 위해 반올림한 기저 빈도표를 제공하기로 했다는 사실이다. 그렇다면 상위 빈도표로 어느 것을 제공해야 노출 위험이 제어되었다고 할 수 있을지 생각해 보자. 조사된 상위 빈도표와 반올림한 상위 빈도표 중에서 우선, 반올림한 상위 빈도표를 제공하기에는 실제 조사된 빈도의 참값인 조사된 상위 빈도와의 차이가 크기 때문에 사용자가 수용하기 힘든 측면이 있다고 판단된다.

그렇다면 반올림한 기저 빈도표와 조사된 상위 빈도표를 함께 제공한다고 가정하고, 이 경우 빈도표에서 발생할 수 있는 노출 위험에 대해 생각해 보려 한다.

<표 3-7> 반올림한 기저 빈도표와 조사된 상위 빈도표를 함께 제공할 경우

반올림한 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	0	0	0	0	13
	아파트	18	3	3	0	36
	연립주택	4	4	3	7	22
	다세대	5	8	0	18	12
	비주거용	0	3	0	4	15
여	단독주택	4	5	8	3	3
	아파트	9	4	9	3	3
	연립주택	5	10	14	5	3
	다세대	10	6	7	8	3
	비주거용	4	5	11	11	3

조사된 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	29	17	3	29	98
	여	32	30	49	29	5

<표 3-7>과 같은 자료를 제공받은 사용자는 일차적으로 자료의 집계 값이 일치하지 않다는 판단을 하게 된다. 5번 집계구의 여자 빈도(인구 수)는 반올림한 기저 빈도표에서는 15명인데 비해, 제공받은 상위 빈도표, 즉 조사된 상위 빈도는 5명으로 큰 차이를 보여 큰 혼란에 빠지게 될 지도 모른다. 더구나, 5번 집계구에서 주택 유형별로 거주하는 여자 빈도(인구 수)가 1명임이 노출된다. 왜냐하면 자료 제공자는 제공받은 자료의 빈도(인구 수)가 3인 경우 그 참 값은 1, 2, 3명 중 1개 값이라고 공지하였고, 제공받은 상위 빈도표에 집계된 빈도가 5이기 때문에 사용자는 비교적 쉽게 참값에 대한 유추가 가능하다.

따라서 자료 제공자는 반올림한 기저 빈도표와 조사된 상위 빈도표를 함께 제공할 경우 노출의 위험이 존재하게 된다는 것을 알 수 있으므로, 이러한 방식의 빈도표 제공은 불가하다는 판단에 이르게 된다. 또, 이 경우에는 정보손실이 크게 발생할 수도 있다는 것이 걸림돌이 된다. 정보손실이란 쉽게 말해 실제 조사된 값과 제공하는 값의 차이를 의미하는 것으로 자료 사용자의 수용 가능성도 함께 염두에 두어야 하겠다. 실제로 조사된 상위 빈도와 반올림한 상위 빈도의 차이를 구해보았다.

<표 3-8> 조사된 상위 빈도와 반올림한 상위 빈도의 차이

정보손실*		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	2	1	3	0	0
	여	0	0	0	1	10

* 정보손실 = | 조사된 기저 빈도를 집계한 상위 빈도 – 반올림한 기저 빈도를 집계한 상위 빈도 |

<표 3-8>에서 보듯이 5번 집계구의 여자 빈도(인구 수)에서 정보손실이 무려 10명만큼 발생하여, 사용자에게 유용한 정보를 제공한다고 할 수 없다. 자료 제공자는 정보손실을 보다 최소화하기 위해 노력할 필요가 있게 되었다.

따라서 자료 제공자는 상위 빈도표 제공 시 노출의 위험을 줄이면서 동시에 정보손실을 최소화하기 위한 조치를 취하려 한다.

2. 중앙값 조정을 통한 비밀보호

노출 위험이 제어된 상위 빈도표를 만들기 위한 첫 번째 단계는 조사된 기저 빈도표에서 빈도(인구 수)의 값이 **3**명 이하인 경우만 먼저 집계한다. 그리고 집계된 값을 중앙값으로 조정한 후에 집계하지 않았던, 3보다 큰 빈도(인구 수)를 더하여 제공하는 방식을 취하게 된다. 여기서 3명 이하의 빈도만 먼저 집계하는 것은 작은 빈도를 갖는 셀로 판단하는 기준을 $B=3$으로 한다는 가정을 두기 때문이다. 만약 작은 빈도를 갖는 셀로 판단하는 기준을 $B=5$로 한다면, 5명 이하의 빈도를 먼저 집계하여 다음에 제시된 순차적인 과정을 유사하게 따르면 된다.

작은 빈도를 집계한 값에 대한 중앙값 조정 과정을 자세히 설명해 보자. 만약 작은 빈도를 집계한 값이 1 ~ 3명에 해당한다면 집계된 값을 2명으로 바꾸고, 집계한 값이 4 ~ 6명에 해당하면 집계된 값을 5명으로 조정하면 된다. 다시 말해 집계된 값을 크기가 3인 구간(interval)으로 나누어 해당 구간의 가운데 있는 값, 즉 중앙값으로 조정하는 것이다. 만약 작은 빈도를 갖는 셀로 판단하는 기준을 $B=5$로 한다면, 크기가 5인 구간으로 나누어 해당 구간의 중앙값으로 조정하면 된다. 예를 들어 집계된 값이 1 ~ 5명에 해당하면 집계된 값을 3명으로 바꾸고, 집계된 값이 6 ~ 10명에 해당하면 집계된 값을 8명으로 간주하게 된다. 단, 0(zero)명은 그대로 0명으로 제공한다. 중앙값 조정에 대한 절차를 도식화하면 다음과 같이 나타낼 수 있다.

<표 3-9> 조사된 기저 빈도에서 작은 빈도 집계 후에 중앙값으로 조정하는 과정

① $B = 3$인 경우

조사된 작은 빈도 집계	0	1,2,3	4,5,6	7,8,9	10,11,12	13,14,15	…
중앙값 조정 빈도	0	2	5	8	11	14	…

② $B = 5$인 경우

조사된 작은 빈도 집계	0	1,2,3,4,5	6,7,8,9,10	11,12,13,14,15	…
중앙값 조정 빈도	0	3	8	13	…

이는 박민정 외(2018)에서 제안된 방식과 다소 차이가 있는 것으로 2018년에 제안된 기존의 방식은 다음과 같은 중앙값 조정 단계를 거쳤다. $B=3$으로 가정한 경우, 조사된 기저 빈도에서 3 이하의 빈도를 먼저 집계하여 조정하는 것은 동일하였다. 그러나 기존 방식에서는 작은 빈도의 집계값이 1 또는 2인 경우에는 0 또는 3으로 임의(random) 조정하였고, 작은 빈도의 집계값이 3인 경우에는 3으로 조정하였던 것이 가장 큰 차이점이라고 하겠다.

<표 3-9-1> 기존(2018년도) 방식의 중앙값 조정 과정($B = 3$인 경우)

조사된 작은 빈도 집계	0	1,2	3	4,5,6	7,8,9	10,11,12	…
중앙값 조정 빈도	0	0 or 3	3	5	8	11	…

2018년에 제안된 기존의 방식을 적용하였을 때 일부 상위 빈도가 1 또는 2로 제공되어 노출 위험이 제어되지 못하는 경우가 발생하여 수정되었다. 이상의 중앙값 조정 과정을 거침으로써 노출 위험이 제어되게 되는데, 관련 내용은 조금 뒤에서 자세히 설명하도록 하겠다.

앞서 예로 들었던 조사된 기저 빈도표의 중앙값 조정은 어떻게 이루어지는지 살펴보자. 설명한 바와 같이 조사된 기저 빈도표에서 빈도가 3명 이하인 경우만 먼저 집계한다. 조사된 작은 빈도를 집계한 값에서 5번 집계구의 남자와 1, 2, 3번 집계구의 여자는 빈도(인구 수)가 3명 이하인 경우가 없으므로 ‘ - ’로 표시하였다.

<표 3-10> 조사된 기저 빈도표에서 3 이하인 빈도(음영 표시)

조사된 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	1	0	0	0	13
	아파트	18	3	2	0	36
	연립주택	4	4	1	7	22
	다세대	5	8	0	18	12
	비주거용	1	2	0	4	15
여	단독주택	4	5	8	3	1
	아파트	9	4	9	2	1
	연립주택	5	10	14	5	1
	다세대	10	6	7	8	1
	비주거용	4	5	11	11	1

➠

조사된 작은 빈도 집계		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	2	5	3	0	-
	여	-	-	-	5	5

조사된 작은 빈도를 집계한 값을 중앙값으로 조정하게 되는데, 작은 빈도를 집계한 값이 1~3인 경우 2로, 4~6인 경우 5로 바꾸었다.

<표 3-11> 집계된 작은 빈도의 중앙값 조정

중앙값 조정 후 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	2	5	2	0	-
	여	-	-	-	5	5

다음 단계로 조사된 기저 빈도표에서 빈도가 3보다 큰 빈도를 중앙값 조정한 빈도에 더해주어 상위 빈도를 구해 보면 다음과 같다.

<표 3-12> 중앙값 조정 후 집계된 상위 빈도표

중앙값 조정 후 집계된 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	29[6]	17[7]	2	29[8]	98
	여	32	30	49	29[9]	5

6) 29 = 2+18+4+5

(비교)

조사된 빈도의 단순 집계		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	29	17	3	29	98
	여	32	30	49	29	5

중앙값 조정 후 집계된 상위 빈도와 조사된 기저 빈도를 단순 집계한 상위 빈도를 비교해 보자. 두 빈도표에서 거의 대부분의 빈도(인구 수)가 일치하며, 3번 집계구의 남자 빈도(인구 수)만 1만큼 차이 남을 알 수 있다. 따라서 중앙값 조정 후 집계된 상위 빈도를 제공한다면, 자료 사용자 입장에서는 정보의 손실이 작기 때문에 만족스러울 수 있다.

이러한 중앙값 조정 과정을 통해 일부 셀에 대한 노출의 위험은 제어된다고 할 수 있다. 구체적으로 2번 집계구의 남자 빈도(인구 수)를 살펴보자.

<표 3-13> 중앙값 조정으로 노출 위험이 제어된 경우

조사된 기저 빈도		2nd 집계구
성별	주택 유형	
남	단독주택	0
	아파트	3
	연립주택	4
	다세대	8
	비주거용	2

반올림한 기저 빈도		2nd 집계구
성별	주택 유형	
남	단독주택	0
	아파트	3
	연립주택	4
	다세대	8
	비주거용	3

중앙값 조정 후 집계된 상위 빈도		2nd 집계구
성별	남	17

사용자는 제공된 반올림한 기저 빈도표를 집계하면 남자의 빈도(인구 수)가 18명이 되므로, 제공되는 상위 빈도(인구 수) 17명과 차이가 난다는 것을 알 수 있다. 그러나 그 차이 값이 어디에서 발생했는지는 특정 지을 수 없기 때문에 노출의 위험이 제어되었다고 할 수 있다. 2번 집계구의 남자 빈도(인구 수)에 대해 사용자가 추측할 수 있는 정보는 다음과 같다.

7) 17 = 5+4+8

8) 29 = 0+7+18+4

9) 29 = 5+5+8+11

<표 3-14> 2번 집계구에 대한 사용자의 추측

반올림한 기저 빈도		2nd 집계구	제공 정보
성별	주택 유형		
남	단독주택	**0**	(0,1,2) 中 하나
	아파트	**3**	(1,2,3) 中 하나
	연립주택	4	4+8=12
	다세대	8	
	비주거용	**3**	(1,2,3) 中 하나

➠

사용자 추측
· 기저빈도 **0+3+3**=5 이다. · (0,3,3)의 참(True) 값 후보? (0,3,2), (0,2,3), (1,2,2), (2,1,2), (1,1,3), … → 후보가 다수 존재

결론적으로 이런 방식의 중앙값 조정을 통해 정보손실의 최소화와 노출 위험의 제어라는 2가지 목표를 일정 수준 달성한 것으로 판단할 수 있다. 따라서 자료 제공자는 중앙값 조정을 거친 후 상위 빈도표로 제공하는 것에 대해 고려할 수 있다. 그러나 중앙값 조정만으로 모든 경우에 노출 위험이 제어되는 것은 아니라는 것이 문제로 남아있다. 따라서 이러한 경우에는 노출 위험 축소를 위해 추가적인 조정 작업이 필요하게 된다.

3. 추가 조치를 통한 노출 위험 제어와 정보손실 축소

중앙값 조정만으로 언제나 노출 위험이 제어되는 것은 아니기에, 추가적인 노출 위험 조정을 위한 조치와 정보손실 축소를 위한 조치가 필요한 경우도 존재한다. 이 과정은 중앙값 조정 이후에도 노출 위험이 남아있거나, 정보손실이 큰 경우에만 적용된다. 반올림한 기저 빈도표에서 작은 빈도를 갖는 셀로 판단하는 기준 값인 B 이하의 작은 빈도만 집계하여 중앙값 조정을 취한 값에 대해 추가적으로 적용한다.

추가적인 노출 위험 조정 단계는 크게 3가지 조치로 유형화 할 수 있다.

가. 중앙값으로 조정한 빈도(인구 수)를 B로 바꾸는 경우

나. 중앙값으로 조정한 빈도(인구 수)에 B만큼 더하는 경우

다. 중앙값으로 조정한 빈도(인구 수)에 B만큼 빼는 경우

각 유형에 대한 판단은 다소 복잡한 기준에 의해 결정되기 때문에, 이것은 BSCA 알고리즘에 의해 컴퓨터가 판단하도록 하였다. 구체적인 판단 기준에 해당하는 알고리즘은 부록에 있는 BSCA 알고리즘의 이해를 참조하면 된다. 각 유형에서 취하는 조치들의 공통점이라고 한다면 숫자 3을 활용한다는 것이다. 앞서 소개된 제2장 제2절 빈도표의 노출 위험에서 언급하였듯이 빈도(인구 수)가 1명 또는 2명인 경우에는

노출의 위험이 있는 것으로 판단하였다. 이에 비해 빈도(인구 수)가 3명인 경우는 노출의 위험이 제어된 것으로 볼 수 있기에 숫자 3을 활용하였다. 자료 제공자의 판단에 따라 작은 빈도를 갖는 셀로 판단하는 기준을 $B=5$로 한다면, 이때는 숫자 5를 활용하여 동일한 알고리즘의 적용이 가능하다. BSCA 알고리즘에 대한 보다 상세한 이론적 증명은 Park(2018)을 참고할 수 있다.

각 유형별로 예시를 제시하여 어떤 노출의 위험이 있는지, 사용자가 추측할 수 있는 정보는 무엇인지, 각 경우마다 어떠한 조치를 취하게 되는지에 대해 설명하고자 한다. 이후 소개되는 내용의 이해를 위해 앞서 언급된 기저 빈도표를 다시 제시한다.

<표 3-15> 조사된 기저 빈도표와 반올림한 기저 빈도표

조사된 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	1	0	0	0	13
	아파트	18	3	2	0	36
	연립주택	4	4	1	7	22
	다세대	5	8	0	18	12
	비주거용	1	2	0	4	15
여	단독주택	4	5	8	3	1
	아파트	9	4	9	2	1
	연립주택	5	10	14	5	1
	다세대	10	6	7	8	1
	비주거용	4	5	11	11	1

반올림한 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	0	0	0	0	13
	아파트	18	3	3	0	36
	연립주택	4	4	3	7	22
	다세대	5	8	0	18	12
	비주거용	0	3	0	4	15
여	단독주택	4	5	8	3	3
	아파트	9	4	9	3	3
	연립주택	5	10	14	5	3
	다세대	10	6	7	8	3
	비주거용	4	5	11	11	3

가. (예시 1) 중앙값으로 조정한 빈도를 B로 바꾸는 경우

기저 빈도표에서 1번 집계구의 남자 빈도(인구 수)를 고려해 보자. 이 예시는 중앙값 조정 후에도 노출 위험이 남아 있는 경우에 해당하며, 중앙값으로 조정한 빈도(인구 수)를 $B = 3$명으로 바꾸는 경우에 해당한다.

그리고 여기서 다시 한번 상기해야 할 것은 자료 제공자는 반올림한 기저 빈도표를 사용자에게 제공하기로 했다는 점이다. 비교와 이해를 돕기 위해 조사된 기저 빈도표, 반올림한 기저 빈도표, 중앙값 조정 후 집계된 상위 빈도표를 함께 제시하였다.

<표 3-16> 중앙값 조정 후에도 노출 위험이 남아 있는 경우(예시1)

조사된 기저 빈도		1st 집계구
성별	주택 유형	
남	단독주택	1
	아파트	18
	연립주택	4
	다세대	5
	비주거용	1

반올림한 기저 빈도		1st 집계구
성별	주택 유형	
남	단독주택	0
	아파트	18
	연립주택	4
	다세대	5
	비주거용	0

중앙값 조정 후 집계된 상위 빈도		1st 집계구
성별	남	29

만약 상위 빈도를 29명으로 제공할 경우, 사용자는 반올림한 기저 빈도표에서 단독주택과 비주거용 건물에 거주하는 남자 빈도(인구 수)가 합산하여 2명이 된다는 사실을 알게 된다. 이때 사용자가 기저 빈도의 참 값에 대해 추측할 수 있는 경우의 수는 3가지이며, 그 중 1가지는 실제 조사된 기저 빈도의 참 값과 일치한다. 따라서, 중앙값 조정 후에도 노출 위험이 남아있다.

<표 3-17> (예시1)에 대한 사용자의 추측

반올림한 기저 빈도		1st 집계구	제공 정보
성별	주택 유형		
남	단독주택	0	(0,1,2) 中 하나
	아파트	18	18+4+5=27
	연립주택	4	
	다세대	5	
	비주거용	0	(0,1,2) 中 하나

➠

사용자 추측
· 기저빈도 0+0=2 이다. · (0,0)의 참(True) 값 후보? → (1, 1), (0, 2), (2, 0) → 참 값 노출 위험 1/3 존재

노출 위험이 제어된 상위 빈도를 집계하는 과정은 다음과 같다. 중앙값 조정 후 빈도(인구 수)인 2명을 노출 위험이 제어된 빈도인 **3명**으로 바꾸어 준다. 그리고 기저 빈도표에서 3명보다 큰 빈도에 해당하는 아파트 거주 인구 수 18명, 연립주택 거주 인구 수 4명, 다세대 거주 인구 수 5명을 합산하여 상위 빈도를 제공하였다.

<표 3-18> 노출 위험이 제어된 상위 빈도 집계 과정(예시1)

중앙값 조정 후 빈도		1st 집계구
성별	남	2

➡

노출 위험 제어된 빈도		1st 집계구
성별	남	3

➡

노출 위험 제어된 상위 빈도		1st 집계구
성별	남	3 0[10]

노출 위험이 제어된 상위 빈도를 제공할 경우, 사용자는 기저 빈도의 참 값을 정말 알아내지 못하는 것일까? 사용자 입장이 되어 추측해 보자.

1번 집계구의 남자 빈도(인구 수)를 30명으로 제공할 경우, 사용자가 추측할 수 있는 정보는 다음과 같다. 앞서 자료 제공자는 기저 빈도표 제공 시에 비밀보호를 위해 반올림한 기저 빈도표를 제공하기로 하였다. 또 보다 정확한 정보의 전달을 위해 다음의 사항을 함께 공지하기로 한 사실을 상기해 보자.

[공지] 공개된 자료를 통해 발생할 수 있는 개인정보보호를 위해 다음과 같이 조치하였습니다.

제공 받은 자료의 빈도(인구수)	실제 조사된 빈도(인구수)
3명	1, 2, 3명 中 1개 값
0명	0, 1, 2명 中 1개 값

제공받은 기저 빈도표에서 3보다 큰 빈도에 대해서는 참 값을 그대로 제공하기 때문에 의심할 여지는 없다. 그렇다면 단독주택에 거주하는 남자 빈도(인구 수)와 비주거용건물에 거주하는 남자 빈도(인구 수)를 합산한 실제 빈도가 3이 된다고 추측할 수 있다. 또 위와 같은 공지를 통해 두 개의 빈도 0을 합산하여 3이 되는 경우로 (1, 2)와 (2, 1)의 2가지 경우로 추측하게 된다. 그러나 사용자가 추측한 2가지 경우 모두 실제 조사된 기저 빈도표의 참 값은 아니기 때문에 노출 위험이 제어되었다고 판단할 수 있다.

10) 30=3+18+4+5

<표 3-19> 노출 위험이 제어된 상위 빈도에 대한 사용자의 추측

반올림한 기저 빈도		1st 집계구	제공 정보
성별	주택 유형		
남	단독주택	0	(0,1,2) 中 하나
	아파트	18	18+4+5=27
	연립주택	4	
	다세대	5	
	비주거용	0	(0,1,2) 中 하나

사용자 추측

· 기저빈도 0+0=3 이다.

· (0,0)에 대한 참(True) 값 후보?

(1, 2), (2, 1)

→ 참 값 노출 위험 제어

나. (예시 2) 중앙값 조정한 빈도(인구 수)에 $+B$ 하는 경우

이번에는 3번 집계구의 남자 빈도(인구 수)를 고려해 보자. 이 예시는 중앙값으로 조정한 빈도(인구 수)에 $B = 3$명을 더해주는 경우에 해당한다.

<표 3-20> 중앙값 조정 후에도 노출 위험이 남아 있는 경우(예시2)

조사된 기저 빈도		3rd 집계구
성별	주택 유형	
남	단독주택	0
	아파트	2
	연립주택	1
	다세대	0
	비주거용	0

반올림한 기저 빈도		3rd 집계구
성별	주택 유형	
남	단독주택	0
	아파트	3
	연립주택	3
	다세대	0
	비주거용	0

중앙값 조정 후 집계된 상위 빈도		3rd 집계구
성별	남	2

사용자는 제공받은 반올림한 기저 빈도를 집계하여 남자의 빈도(인구 수)가 6명이라고 판단하게 된다. 그런데 중앙값 조정 후 집계된 상위 빈도인 2명으로 제공하게 되면 정보의 손실이 크고, 사용자가 수용하기 어려울 수도 있다. 이 경우 사용자가 기저 빈도에 대해 추론할 수 있는 정보는 다음과 같다.

<표 3-21> (예시2)에 대한 사용자의 추측

반올림한 기저 빈도		3rd 집계구	제공 정보
성별	주택 유형		
남	단독주택	0	(0,1,2) 中 하나
	아파트	3	(1,2,3) 中 하나
	연립주택	3	(1,2,3) 中 하나
	다세대	0	(0,1,2) 中 하나
	비주거용	0	(0,1,2) 中 하나

사용자 추측

· 기저빈도 0+3+3+0+0=2 이다.

· 참(True) 값 후보?

(0, 1, 1, 0, 0)

→ 참 값 노출 위험 제어

but 정보손실 증가

상위 빈도를 2명으로 제공할 경우에 사용자는 조사된 기저 빈도의 참 값을 정확하게 추측할 수 없어, 노출 위험은 제어되었다고 할 수 있다. 그러나 기저 빈도와 상위 빈도의 차이가 4명이나 발생하게 된다. 따라서 자료 제공자는 중앙값 조정 후 빈도인 2명에 3명을 더해준 **5명**으로 제공하면, 노출 위험을 제어시키면서 제공받은 기저 빈도와 상위 빈도의 차이를 줄일 수 있다. 노출 위험이 제어된 상위 빈도를 집계하는 과정은 다음과 같다.

<표 3-22> 노출 위험이 제어된 상위 빈도 집계 과정(예시2)

중앙값 조정 후 빈도		3rd 집계구	노출 위험 제어된 빈도		3rd 집계구	노출 위험이 제어된 상위 빈도		3rd 집계구
성별	남	2	성별	남	5	성별	남	5

노출 위험이 제어된 상위 빈도를 제공하기 위해서는 노출 위험이 제어된 빈도 5명에 기저 빈도표에서 3명보다 큰 빈도를 합산하여 제공한다. 그러나 3번 집계구의 기저 빈도표에서는 3명보다 큰 빈도가 존재하지 않기 때문에, 노출 위험이 제어된 빈도 5명을 그대로 상위 빈도로 제공하게 된다.

추가적인 노출 위험 제어를 위해 조정된 상위 빈도에서 사용자가 추측할 수 있는 기저 빈도의 참 값은 어떤 것일까?

제공받은 기저 빈도를 합산하여 상위 빈도 5명이 되도록 하는 경우의 수는 많이 존재하기 때문에 참 값의 노출 위험은 제어되었다고 판단된다. 또 참 값과의 차이가 2만큼 존재하여, 중앙값 조정 후 빈도 제공 시보다 정보손실이 줄었다고 할 수 있다.

<표 3-23> 노출 위험이 제어된 상위 빈도에 대한 사용자의 추측(예시2)

반올림한 기저 빈도		3rd 집계구	제공 정보	사용자 추측
성별	주택 유형			
남	단독주택	0	(0,1,2) 中 하나	· 기저빈도 0+3+3+0+0=5 이다. · 참(True) 값 후보? (1,1,1,1,1), (0,2,2,0,1), … → 참 값 노출 위험 제어 and 정보손실 축소
	아파트	3	(1,2,3) 中 하나	
	연립주택	3	(1,2,3) 中 하나	
	다세대	0	(0,1,2) 中 하나	
	비주거용	0	(0,1,2) 中 하나	

이상과 동일한 작업은 5번 집계구의 여자 빈도(인구 수)에도 적용된다.

다. (예시 3) 중앙값 조정한 빈도(인구 수)에 $-B$ 하는 경우

마지막으로 4번 집계구의 여자 빈도(인구 수)를 살펴보자. 이 예시는 중앙값으로 조정한 빈도(인구 수)에 $B = 3$명을 빼는 경우에 해당한다.

<표 3-24> 중앙값 조정 후에도 노출 위험이 남아 있는 경우(예시3)

조사된 기저 빈도		4th 집계구
성별	주택 유형	
여	단독주택	3
	아파트	2
	연립주택	5
	다세대	8
	비주거용	11

반올림한 기저 빈도		4th 집계구
성별	주택 유형	
여	단독주택	3
	아파트	3
	연립주택	5
	다세대	8
	비주거용	11

중앙값 조정 후 집계된 상위 빈도		4th 집계구
성별	여	29

만약 상위 빈도를 29명으로 제공할 경우, 사용자는 단독주택과 아파트에 거주하는 여자 빈도(인구 수)가 합산하여 5명이 된다는 사실을 알게 된다. 이때 사용자가 기저 빈도의 참 값에 대해 추측할 수 있는 경우의 수는 2가지이며, 그 중 1가지는 실제 조사된 기저 빈도와 일치한다. 다시 말해 중앙값 조정 후 집계한 상위 빈도에 노출 위험이 여전히 남아있게 된다.

<표 3-25> (예시3)에 대한 사용자의 추측

반올림한 기저 빈도		4th 집계구	제공 정보
성별	주택 유형		
여	단독주택	**3**	(1,2,3) 中 하나
	아파트	**3**	(1,2,3) 中 하나
	연립주택	5	5+8+11=24
	다세대	8	
	비주거용	11	

➠

사용자 추측

· 기저빈도 **3**+**3**=5 이다.

· (**3**,**3**)에 대한 참(True) 값 후보?

(2, 3), (3, 2)

→ 참 값 노출 위험 1/2 존재

따라서 자료 제공자는 중앙값 조정 후 빈도 5명에서 3명을 빼준 **2명**으로 조정하는 작업을 한다. 이후 기저 빈도표에서 3명보다 큰 빈도에 해당하는 5명, 8명, 11명을 합산하여 상위 빈도를 제공하도록 조치하였다.

<표 3-26> 노출 위험이 제어된 상위 빈도 집계 과정(예시3)

중앙값 조정 후 빈도		4th 집계구	노출 위험 제어된 빈도		4th 집계구	노출 위험이 제어된 상위 빈도		4th 집계구
성별	여	5	성별	여	2	성별	여	26[11]

노출 위험이 제어된 상위 빈도표에서 사용자 입장에서 추측 가능한 기저 빈도표의 참 값은 어떻게 될까?

<표 3-27> 노출 위험이 제어된 상위 빈도에 대한 사용자의 추측(예시3)

반올림한 기저 빈도		4th 집계구	제공 정보	사용자 추측
성별	주택 유형			
여	단독주택	3	(1,2,3) 中 하나	· 기저빈도 3+3=2 이다. · (3,3)에 대한 참(True) 값 후보? (1, 1) → 참 값 노출 위험 제어
	아파트	3	(1,2,3) 中 하나	
	연립주택	5	5+8+11=24	
	다세대	8		
	비주거용	11		

제공받은 기저 빈도에서 3명, 3명인 값을 합산하여 상위 빈도 2명이 되도록 하는 경우의 수는 각각이 1명씩인 경우로 참 값의 노출 위험은 제어되었다고 판단된다. 조사된 기저 빈도표의 참 값과의 차이는 3명만큼 존재하게 되는데, 중앙값으로 조정 후 집계한 상위 빈도인 29명에 비해 정보손실은 다소 증가하였다. 그러나 이는 노출 위험 제어를 위해 불가피한 조치임을 상기할 필요가 있다.

라. 노출 위험이 제어된 상위 빈도표의 제공

상위 빈도표의 노출 위험 제어를 위해 다소 복잡한 조정 단계를 거쳤다. 첫 번째 단계로 기저 빈도표에서 3 이하인 빈도를 합산하여 해당 값이 속한 구간의 중앙값으로 조정하였다. 그리고 중앙값 조정된 빈도에 아직 합산하지 않은 빈도, 즉 기저 빈도표에서 3보다 큰 빈도를 합산하였다. 이 과정을 통해 노출 위험은 제어시키고, 정보손실도 줄일 수 있다. 그러나 모든 경우에 노출 위험이 제어되는 것은 아니어서, 경우에 따라서는 추가적인 노출 위험 조정이 필요하다. 추가적인 노출 위험 조정을 위한 두 번째 작업은 첫 번째 단계의 중앙값 조정된 값을 특정한 값으로 바꾸거나,

11) 26=2+5+8+11

특정한 값만큼 더하거나 빼는 것으로 하였다. 여기서 말하는 특정한 값이란, 자료 제공자가 판단하는 작은 빈도를 갖는 셀로 판단하는 기준 값 B를 의미한다. 다시 말해 중앙값 조정된 값을 B로 바꾸거나 중앙값 조정된 값에 $+B$ 또는 $-B$ 하여 상위 빈도표를 제공하는 것으로 하였다. 이러한 단계를 거쳐 최종적으로 제공되는 상위 빈도표는 다음과 같다.

<표 3-28> 노출 위험이 제어된 상위 빈도표

중앙값 조정 후 집계된 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	29	17	2	29	98
	여	32	30	49	29	5

노출 위험이 제어된 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	30	17	5	29	98
	여	32	30	49	26	8

그리고, 조사된 기저 빈도표를 단순 집계한 상위 빈도표와 노출 위험이 제어된 상위 빈도표의 정보손실은 다음과 같았다. 상위 빈도표에서 노출 위험 제어에 따른 정보손실은 3명 이내임을 확인할 수 있다.

<표 3-29> 노출 위험 제어에 따른 정보손실

정보손실*		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	1	0	2	0	0
	여	0	0	0	3	3

주: * 정보손실= | 참 값 빈도 − 제공 빈도 | = | 조사된 기저 빈도를 집계한 상위 빈도 − 노출 위험이 제어된 상위 빈도 |

제3절 BSCA를 적용한 빈도표 생성

제1절과 제2절에서 상세히 설명된 빈도표에 대한 비밀보호 방안인 BSCA 알고리즘을 SGIS 자료에 적용해 보았다. 2018년 인구주택총조사 자료 중 대전광역시에 해당하는 빈도표를 생성하는 시나리오를 작성하고, 빈도표 생성에 소요되는 시간을 측정하여 정리하였다. 20년 11월 현재, 구현된 BSCA 알고리즘은 통계패키지 R에서 구동되며 로컬 PC의 메모리 한계로 인해 통계청 R SERVER에 접근하여 작업하였다.

SGIS 자료의 빈도표는 기저 빈도표와 상위 빈도표의 위계 구조를 가지고 있다. 현재 구현된 BSCA 알고리즘에서 기저 빈도표를 만드는데 걸리는 시간은 행정 구역 단위에서는 2시간, 격자 단위에서는 7시간이 소요되었다. 반면 상위 빈도표의 생성은 빈도표로 만들 항목을 사용자가 직접 입력하면 이미 만들어져 있는 기저 빈도표를 바탕으로 상위 지역 수준인 OA, LA3, LA2, LA1에 해당하는 빈도표를 순차적으로 만드는 식으로 진행된다. 예를 들어 항목이 3개인 상위 빈도표를 요청하는 경우 항목이 6개인 기저 빈도표에서 항목의 수를 줄여 합산하는 방식으로 상위 빈도표를 만들되, OA, LA3, LA2, LA1 지역 수준에 해당하는 빈도표를 함께 작성하도록 설계되었다. 따라서 요청하는 항목의 수가 많을수록, 해당 항목의 선택지 수가 많을수록 계산해야 할 셀의 수가 많아진다. 생성된 각 셀에 대한 중앙값 조정과 추가적인 노출 위험 제어를 위한 과정을 거치는 등 상위 빈도표 생성을 위해 알고리즘상에서 고려해야 할 사항이 많아 시간을 필요로 한다.

기저 빈도표가 집계구 단위(행정 구역 단위)인 경우와 100m 단위 격자인 경우로 구분하여 정리하였다. 해당 내용은 추후 SGIS의 BSCA 알고리즘을 적용하기 위한 데이터 마트(data mart) 생성 작업 시에 참고 할 수 있다.

1. 행정 구역 단위에서의 빈도표

2018년 기준 대전광역시의 행정 구역 현황은 2,997개의 집계구, 79개의 동, 5개의 구로 구성되어 있다. 인구주택총조사 자료에 대한 제공 항목은 모두 6개이다. 6개 항목을 모두 고려한 집계구 단위의 기저 빈도표를 생성하기 위해 작성되어야 하는 셀의 수는 475,803,720개(=158,760×2,997)이다. 이때 제공 항목이 6개이고 지역 수준이 집계구인 기저 빈도표를 포함하여 지역 수준이 LA3, LA2, LA1에 해당하는 상위 빈도표의 생성까지 대략 2시간 정도 소요된다.

<표 3-30> 대전광역시 행정 구역 현황(2018년 기준)

행정 구역 단위		Number of level
(LA1)	시도	1
(LA2)	시군구	5
(LA3)	읍면동	79
(OA)	집계구	2,997
소계		3,083

<표 3-31> 2018년 인구주택총조사 자료의 제공 항목

제공 항목	선택지 수	제공 항목	선택지 수
성별	2	주택유형	5
연령	21	주택면적	9
세대구성	6	건축연도	14

항목의 수가 5개인 상위 빈도표는 6개의 제공 항목 중 5개를 선택하는 경우의 수에 해당하는 6가지가 존재하며 시나리오에 따라 최소 2시간에서 최대 약 7일 정도 소요되었다. 제공 항목 6가지 중 선택지의 수가 많은 항목은 연령(21가지), 건축연도(14가지)이며, 두 항목 모두 포함된 경우 상위 빈도표 생성에 많은 시간이 필요함을 알 수 있다. 제공 항목이 2개 또는 1개인 경우에는 대체로 10분 이내에서 상위 빈도표가 모두 만들어지는 것을 확인할 수 있다. 제공 항목의 수가 작은 경우는 항목의 수가 많은 경우보다 기저 빈도표로부터 집계된 빈도가 일정 수준 이상으로 커지기 때문에 알고리즘에서 고려해야 할 사항이 적어지기 때문이다.

<표 3-32> 행정 구역 단위별 빈도표 생성 소요 시간

구분	항목	소요 시간
빈도표* (항목 수 6개)	성별, 연령, 세대구성, 주택유형, 주택면적, 건축연도	2hr

주: * 기저 빈도표와 상위 빈도표를 포함한 경우임

구분	항목	소요 시간(분)	비고
상위 빈도표 (항목 수 5개)	연령, 세대구성, 주택유형, 주택면적, 건축연도	9,258	6.5일
	성별, 세대구성, 주택유형, 주택면적, 건축연도	123	2hr
	성별, 연령, 주택유형, 주택면적, 건축연도	781	13hr
	성별, 연령, 세대구성, 주택면적, 건축연도	1,031	17hr
	성별, 연령, 세대구성, 주택유형, 건축연도	392	6.5hr
	성별, 연령, 세대구성, 주택유형, 주택면적	204	4hr

구분	항목	소요 시간(분)	비고
상위 빈도표 (항목 수 4개)	세대구성, 주택유형, 주택면적, 건축연도	60	
	연령, 주택유형, 주택면적, 건축연도	282	5hr
	연령, 세대구성, 주택면적, 건축연도	337	6hr
	연령, 세대구성, 주택유형, 건축연도	171	3hr
	연령, 세대구성, 주택유형, 주택면적	84	
	성별, 주택유형, 주택면적, 건축연도	21	
	성별, 세대구성, 주택면적, 건축연도	25	
	성별, 세대구성, 주택유형, 건축연도	17	
	성별, 세대구성, 주택유형, 주택면적	13	
	성별, 연령, 주택면적, 건축연도	87	
	성별, 연령, 주택유형, 건축연도	45	
	성별, 연령, 주택유형, 주택면적	28	
	성별, 연령, 세대구성, 건축연도	49	
	성별, 연령, 세대구성, 주택면적	31	
	성별, 연령, 세대구성, 주택유형	21	

구분	항목	소요 시간(분)	비고
상위 빈도표 (항목 수 3개)	주택유형, 주택면적, 건축연도	16	
	세대구성, 주택면적, 건축연도	17	
	세대구성, 주택유형, 건축연도	12	
	세대구성, 주택유형, 주택면적	10	
	연령, 주택면적, 건축연도	38	
	연령, 주택유형, 건축연도	22	
	연령, 주택유형, 주택면적	16	
	연령, 세대구성, 건축연도	25	
	연령, 세대구성, 주택면적	18	
	연령, 세대구성, 주택유형	15	
	성별, 주택면적, 건축연도	10	
	성별, 주택유형, 건축연도	9	
	성별, 주택유형, 주택면적	9	
	성별, 연령, 건축연도	14	
	성별, 연령, 주택면적	12	
	성별, 연령, 세대구성	11	
	성별, 연령, 주택유형	10	
	성별, 세대구성, 주택유형	8	
	성별, 세대구성, 주택면적	9	
	성별, 세대구성, 건축연도	9	

구분	항목	소요 시간(분)	구분	항목	소요 시간(분)
상위 빈도표 (항목 수 2개)	성별, 연령	8	상위 빈도표 (항목 수 1개)	성별	8
	성별, 세대구성	8		연령	8
	성별, 주택유형	8		세대구성	8
	성별, 주택면적	8		주택유형	8
	성별, 건축연도	8		주택면적	8
	연령, 세대구성	9		건축연도	8
	연령, 주택유형	9			
	연령, 주택면적	10			
	연령, 건축연도	11			
	세대구성, 주택유형	8			
	세대구성, 주택면적	8			
	세대구성, 건축연도	9			
	주택유형, 주택면적	8			
	주택유형, 건축연도	8			
	주택면적, 건축연도	9			

2. 격자 단위에서의 빈도표

2018년 기준 대전광역시에 해당하는 격자 현황은 100m 단위 격자로는 10,060개가 해당하며, 1km 단위 격자로는 423개, 10km 단위 격자는 10개로 구성되어 있다. 격자 단위에서의 빈도표 생성 시나리오 해석 시, 대전광역시를 100m, 1km, 10km 단위로 구획한 것은 아니라는 점을 주의할 필요가 있다. 격자(그리드)는 국가지점번호 기준점에 따라 부여하는 기준이 별도로 존재하고 있다. 본 시나리오에서는 대전광역시에 해당하는 격자만 뽑아서 BSCA 알고리즘을 적용한 것으로, 경우에 따라 대전광역시의 일부만 포함된 격자도 있다. 마찬가지로 인구주택총조사 자료의 6개 항목을 모두 고려한 100m 단위 기저 빈도표를 생성하기 위해 작성되어야 하는 셀의 수는 1,597,125,600개(=158,760×10,060)로, 100m, 1km, 10km, 100km 격자 단위의 빈도표를 모두 생성하는데 대략 7시간 소요된다.

<표 3-33> 대전광역시 해당 격자 단위 현황

격자 단위	Number of level
(LA1) 100km	1
(LA2) 10km	10
(LA3) 1km	423
(OA) 100m	10,060
소계	10,494

항목의 수가 5개인 상위 빈도표는 마찬가지로 6가지가 있으며 최소 18시간에서 최대 약 23일이 소요되었다. 특히 연령, 세대구성, 주택유형, 주택면적, 건축연도를 포함한 상위 빈도표 생성에 약 23일이라는 굉장히 많은 시간이 소요된다. 행정 구역 단위의 경우와 마찬가지로 연령과 건축연도가 모두 포함된 경우의 상위 빈도표 생성에 많은 시간이 필요하고, 전반적으로 동일한 시나리오에서 행정 구역 단위의 상위 빈도표 생성 시간보다 격자 단위의 상위 빈도표 생성 시간이 약 2~3배 정도 더 많이 소요됨을 알 수 있다. 제공 항목이 2개 또는 1개인 경우에는 상위 빈도표 생성 시간이 비교적 안정적이며, 대체로 30분 정도 소요되었다.

<표 3-34> 격자 단위별 빈도표 생성 현황

구분	항목	소요 시간
빈도표* (항목 수 6개)	성별, 연령, 세대구성, 주택유형, 주택면적, 건축연도	7hr

주: * 기저 빈도표와 상위 빈도표를 포함한 경우임

구분	항목	소요 시간(분)	비고
상위 빈도표 (항목 수 5개)	연령, 세대구성, 주택유형, 주택면적, 건축연도	33,870	23days
	성별, 세대구성, 주택유형, 주택면적, 건축연도	5,211	4days
	성별, 연령, 주택유형, 주택면적, 건축연도	3,457	3days
	성별, 연령, 세대구성, 주택면적, 건축연도	14,024	10days
	성별, 연령, 세대구성, 주택유형, 건축연도	1,854	30hr
	성별, 연령, 세대구성, 주택유형, 주택면적	1,057	18hr

구분	항목	소요 시간(분)	비고
상위 빈도표 (항목 수 4개)	세대구성, 주택유형, 주택면적, 건축연도	241	4hr
	연령, 주택유형, 주택면적, 건축연도	1,062	17hr
	연령, 세대구성, 주택면적, 건축연도	2,595	43hr
	연령, 세대구성, 주택유형, 건축연도	702	12hr
	연령, 세대구성, 주택유형, 주택면적	350	6hr
	성별, 주택유형, 주택면적, 건축연도	85	
	성별, 세대구성, 주택면적, 건축연도	100	
	성별, 세대구성, 주택유형, 건축연도	68	
	성별, 세대구성, 주택유형, 주택면적	53	
	성별, 연령, 주택면적, 건축연도	374	6hr
	성별, 연령, 주택유형, 건축연도	171	3hr
	성별, 연령, 주택유형, 주택면적	121	
	성별, 연령, 세대구성, 건축연도	227	4hr
	성별, 연령, 세대구성, 주택면적	137	
	성별, 연령, 세대구성, 주택유형	88	

구분	항목	소요 시간(분)	비고
상위 빈도표 (항목 수 3개)	주택유형, 주택면적, 건축연도	58	
	세대구성, 주택면적, 건축연도	63	
	세대구성, 주택유형, 건축연도	47	
	세대구성, 주택유형, 주택면적	39	
	연령, 주택면적, 건축연도	161	2.6hr
	연령, 주택유형, 건축연도	88	
	연령, 주택유형, 주택면적	63	
	연령, 세대구성, 건축연도	101	
	연령, 세대구성, 주택면적	72	
	연령, 세대구성, 주택유형	50	
	성별, 주택면적, 건축연도	41	
	성별, 주택유형, 건축연도	34	
	성별, 주택유형, 주택면적	31	
	성별, 연령, 건축연도	53	
	성별, 연령, 주택면적	44	
	성별, 연령, 세대구성	34	
	성별, 연령, 주택유형	36	
	성별, 세대구성, 주택유형	26	
	성별, 세대구성, 주택면적	32	
	성별, 세대구성, 건축연도	31	

구분	항목	소요 시간(분)	구분	항목	소요 시간(분)
상위 빈도표 (항목 수 2개)	성별, 연령	31	상위 빈도표 (항목 수 1개)	성별	29
	성별, 세대구성	28		연령	28
	성별, 주택유형	28		세대구성	27
	성별, 주택면적	29		주택유형	26
	성별, 건축연도	29		주택면적	26
	연령, 세대구성	33		건축연도	25
	연령, 주택유형	32			
	연령, 주택면적	35			
	연령, 건축연도	39			
	세대구성, 주택유형	27			
	세대구성, 주택면적	28			
	세대구성, 건축연도	29			
	주택유형, 주택면적	27			
	주택유형, 건축연도	28			
	주택면적, 건축연도	30			

제 4 장

비밀보호가 적용된 자료의 정합성

지금까지 SGIS에서 제공하는 빈도표에 대한 비밀보호 방안의 적용 필요성과 함께 구체적인 적용 방법을 살펴보았다. 이러한 방법은 빈도표에서 빈도가 1 또는 2인 경우에는 필연적으로 그 값이 변경되고, 빈도가 3 이상인 경우에도 노출 위험이 있으면 그 값이 바뀌게 된다. 따라서 비밀보호 알고리즘 적용 전과 후에 빈도의 분포가 어떻게 달라지는지를 살펴볼 필요가 있어 행정 구역 단위의 자료와 격자 단위의 자료로 나누어 살펴보았다.

제1절 행정 구역 단위 자료의 정합성

1. 비밀보호 알고리즘 적용에 따른 빈도의 변화

가. 기저 빈도표의 변화

앞서 대전광역시 행정 구역 단위 자료의 기저 빈도표 제공 시 비밀보호가 필요한 경우는 약 0.1% 내외임을 확인한 바 있다. 만약 작은 빈도를 갖는 셀로 판단하는 기준을 $B = 3$으로 하고 조사된 빈도가 0(zero)인 경우는 그대로 제공한다고 가정하자. 그렇다면 기저 빈도표에서 비밀보호 조치가 필요한 경우는 조사된 빈도가 1 또는 2인 경우로 한정할 수 있다.

<표 4-1> 2018년 인구주택총조사 자료의 기저 빈도표 분포(대전시 기준)

제공 항목: 6개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
OA (집계구)	셀의 수	475,252,525	339,358	88,062	36,723	19,838	67,214	475,803,720
	비율(%)	99.88	0.07	0.02	0.01	0.004	0.014	100

제공 항목: 6개인 경우		빈도 (비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
OA (집계구)	셀의 수	475,508,406	-	-	208,262	19,838	67,214	475,803,720
	비율(%)	99.94	0	0	0.04	0.004	0.014	100

기저 빈도표에 대한 비밀보호 알고리즘 적용은 조사된 빈도가 1 또는 2인 경우, 랜덤 라운딩을 적용하여 0 또는 3으로 바꾸는 작업이 된다. 따라서 비밀보호 이후 제공되는 기저 빈도표는 빈도가 1 또는 2인 셀은 없어지고 빈도가 0인 셀의 수는 0.06%p, 빈도가 3인 셀의 수는 0.03%p 증가하는 형태를 보인다.

나. 상위 빈도표의 변화

다음으로 동일한 항목 수에 대해 지역 수준을 넓혀가는 방식으로 상위 빈도표를 만들어 빈도의 변화 양상을 관찰해 보았다. 비밀보호 알고리즘 적용 전·후 빈도표의 각 셀에서 나타나는 빈도를 0부터 5 이상인 경우로 나누어 셀의 수를 조사하였다.

알고리즘 적용 후에 빈도가 0인 셀이 증가하는 경우는 기저 빈도표에서만 발생하였고, 상위 빈도표에서는 발생하지 않았다. 알고리즘 적용 이전에는 제공 항목 수가 줄어들거나 지역 수준이 넓어지는 상위 빈도표에서 빈도가 0인 셀의 수는 감소하지만 빈도가 1 또는 2인 셀의 수는 점차 증가하는 경향을 보인다. 그러나 비밀보호 알고리즘 적용 후 상위 빈도표에서는 빈도가 1 또는 2인 셀은 나타나지 않고, 빈도가 3 이상인 셀의 비율이 증가함을 확인할 수 있다. 특히 제공 항목이 1개이고 지역 수준이 LA1(시도)인 최상위 빈도표에서는 모든 빈도가 5 이상의 값을 가짐을 확인할 수 있다. 행정 구역 단위 자료의 상위 빈도표에서 변화가 가장 큰 경우는 빈도가 3인 셀의 비율인 것으로 확인된다. 제공 항목의 수가 3~6 개인 경우에 지역 수준이 LA1(시도), LA2(시군구)인 경우 10%p 내외의 변화를 보이는 것으로 확인된다. 반면 제공 항목의 수가 2개인 경우에는 지역 수준이 LA3(읍면동), OA(집계구)에서 5%p 수준의 변화가 있음을 알 수 있다.

<표 4-2> 2018년 인구주택총조사 자료의 상위 빈도표 분포(대전시 기준)

제공 항목: 6개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	123,071	10,736	4,746	2,717	1,755	15,735	158,760
	비율(%)	77.5	6.8	3.0	1.7	1.1	9.9	100
LA2 (시군구)	셀의 수	707,136	31,379	12,366	6,809	4,826	31,284	793,800
	비율(%)	89.1	4.0	1.6	0.9	0.6	3.9	100
LA3 (읍면동)	셀의 수	12,298,495	118,338	40,453	20,631	12,608	51,515	12,542,040
	비율(%)	98.1	0.9	0.3	0.2	0.1	0.4	100

제공 항목: 6개인 경우		빈도(비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	123,071	-	-	17,553	66	18,070	158,760
	비율(%)	77.5	0	0	11.1	0.04	11.4	100
LA2 (시군구)	셀의 수	707,136	-	-	48,637	375	37,652	793,800
	비율(%)	89.1	0	0	6.1	0.05	4.7	100
LA3 (읍면동)	셀의 수	12,298,495	-	-	173,353	2,013	68,179	12,542,040
	비율(%)	98.1	0	0	1.38	0.02	0.5	100

제공 항목: 5개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	101,406	13,479	7,723	5,051	3,627	42,846	174,132
	비율(%)	58.2	7.7	4.4	2.9	2.1	24.6	100
LA2 (시군구)	셀의 수	657,501	54,797	26,907	16,061	11,534	103,860	870,660
	비율(%)	75.5	6.3	3.1	1.8	1.3	11.9	100
LA3 (읍면동)	셀의 수	12,948,737	305,300	131,971	73,061	47,670	249,689	13,756,428
	비율(%)	94.1	2.2	1.0	0.5	0.4	1.8	100
OA (집계구)	셀의 수	519,486,667	1,259,316	423,074	191,177	108,100	405,270	521,873,604
	비율(%)	99.54	0.24	0.08	0.04	0.02	0.08	100

제공 항목: 5개인 경우		빈도 (비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	101,406	-	-	24,531	45	48,150	174,132
	비율(%)	58.2	0	0	14.1	0.03	27.7	100
LA2 (시군구)	셀의 수	657,501	-	-	92,383	373	120,403	870,660
	비율(%)	75.5	0	0	10.6	0.04	13.8	100
LA3 (읍면동)	셀의 수	12,948,737	-	-	484,881	3,325	319,485	13,756,428
	비율(%)	94.1	0	0	3.5	0.02	2.3	100
OA (집계구)	셀의 수	519,486,667	-	-	1,849,696	56,932	480,309	521,873,604
	비율(%)	99.5	0	0	0.35	0.01	0.1	100

제공 항목: 4개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	22,049	4,368	2,924	2,121	17,000	35,544	68,706
	비율(%)	32.09	6.36	4.26	3.09	2.5	51.7	100
LA2 (시군구)	셀의 수	176,974	25,493	15,676	10,021	7,758	107,608	343,530
	비율(%)	51.52	7.42	4.56	2.92	2,3	31.3	100
LA3 (읍면동)	셀의 수	4,530,683	232,239	126,077	76,523	55,044	407,208	5,427,774
	비율(%)	83.47	4.28	2.32	1.41	1.0	7.5	100
OA (집계구)	셀의 수	202,054,366	1,604,448	704,777	358,520	223,477	966,294	205,911,882
	비율(%)	98.13	0.78	0.34	0.17	0.1	0.5	100

제공 항목: 4개인 경우		빈도 (비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	22,049	-	-	8,602	4	38,051	68,706
	비율(%)	32.1	0	0	12.5	0.01	55.4	100
LA2 (시군구)	셀의 수	176,974	-	-	47,500	81	118,975	343,530
	비율(%)	51.5	0	0	13.8	0.02	34.6	100
LA3 (읍면동)	셀의 수	4,530,683	-	-	405,993	1,339	489,759	5,427,774
	비율(%)	83.5	0	0	7.5	0.02	9.0	100
OA (집계구)	셀의 수	202,054,366	-	-	2,561,667	47,882	1,247,967	205,911,882
	비율(%)	98.1	0	0	1.2	0.02	0.6	100

제공 항목: 3개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	1,210	367	292	194	157	10,691	12,911
	비율(%)	9.4	2.8	2.3	1.5	1.2	82.8	100
LA2 (시군구)	셀의 수	14,162	3,313	2,495	1,697	1,535	41,353	64,555
	비율(%)	21.9	5.1	3.9	2.6	2.4	64.1	100
LA3 (읍면동)	셀의 수	618,043	58,973	39,259	25,951	21,241	256,502	1,019,969
	비율(%)	60.6	5.8	3.8	2.5	2.1	25.1	100
OA (집계구)	셀의 수	35,888,760	792,528	457,308	267,159	190,658	1,097,854	38,694,267
	비율(%)	92.8	2.0	1.2	0.7	0.5	2.8	100

제공 항목: 3개인 경우		빈도 (비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	1,210	-	-	764	-	10,937	12,911
	비율(%)	9.4	0	0	5.9	0	84.7	100
LA2 (시군구)	셀의 수	14,162	-	-	6,833	5	43,555	64,555
	비율(%)	21.9	0	0	10.6	0.01	67.5	100
LA3 (읍면동)	셀의 수	618,043	-	-	114,154	176	287,596	1,019,969
	비율(%)	60.6	0	0	11.2	0.02	28.2	100
OA (집계구)	셀의 수	35,888,760	-	-	1,424,893	15,565	1,365,049	38,694,267
	비율(%)	92.7	0	0	3.7	0.04	3.5	100

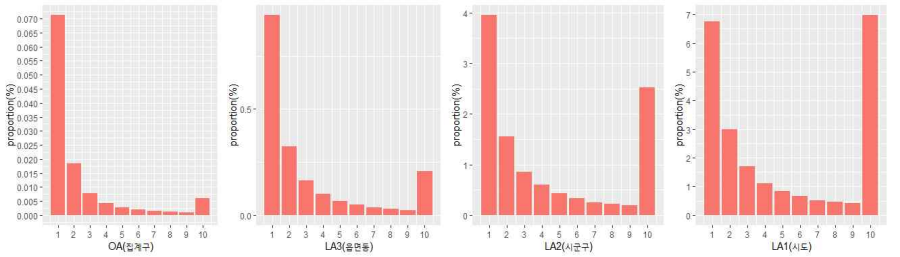
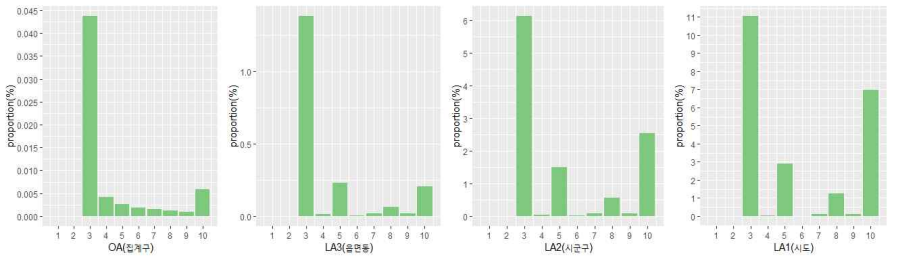
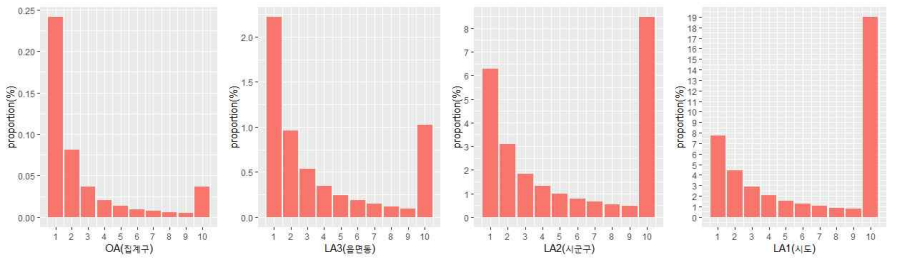
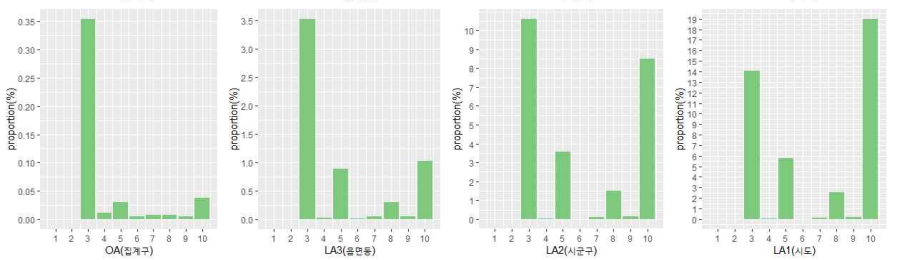
제공 항목: 2개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	11	5	4	5	4	1,204	1,233
	비율(%)	0.9	0.4	0.3	0.4	0.3	97.6	100
LA2 (시군구)	셀의 수	242	69	59	42	48	5,705	6,165
	비율(%)	3.9	1.1	1.0	0.7	0.8	92.5	100
LA3 (읍면동)	셀의 수	26,500	3,941	3,157	2,236	2,172	59,401	97,407
	비율(%)	27.2	4.0	3.2	2.3	2.2	61	100
OA (집계구)	셀의 수	2,804,711	127,115	95,544	64,367	54,680	548,884	3,695,301
	비율(%)	75.9	3.4	2.6	1.7	1.5	14.9	100

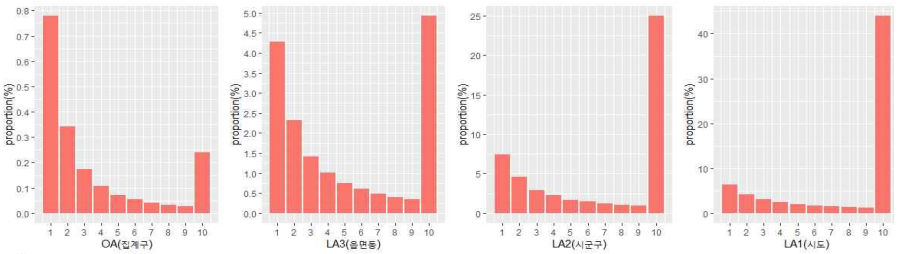
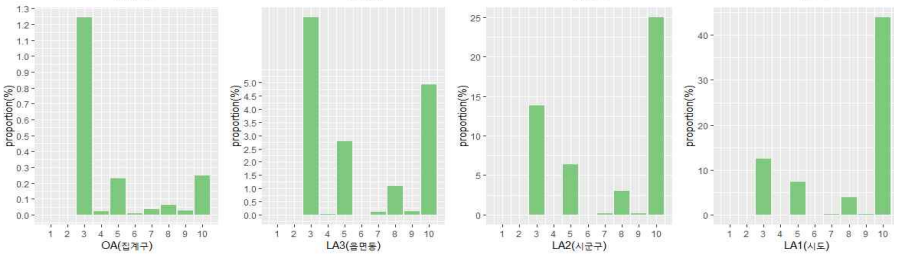
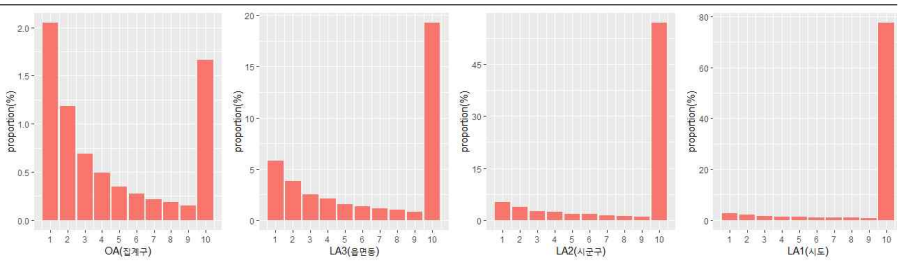
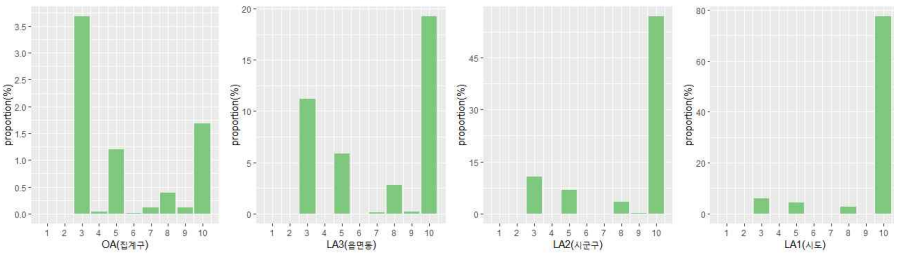
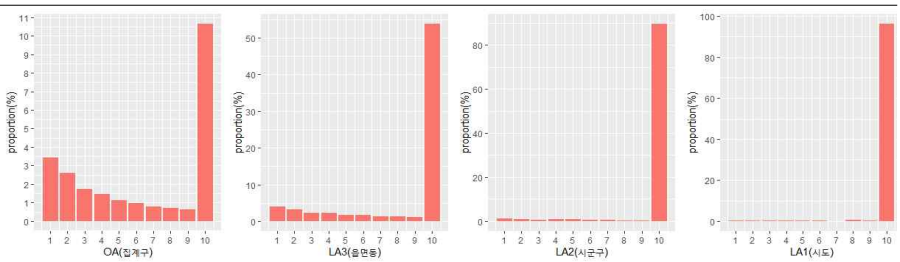
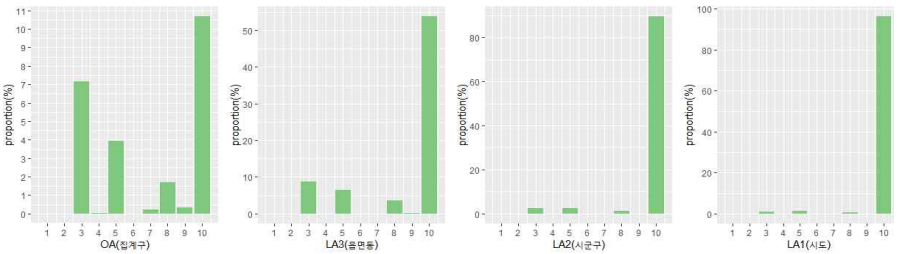
제공 항목: 2개인 경우		빈도 (비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	11	-	-	10	-	1,212	1,233
	비율(%)	0.9	0	0	0.8	0	98.3	100
LA2 (시군구)	셀의 수	242	-	-	151	-	5,772	6,165
	비율(%)	3.9	0	0	2.4	0	93.6	100
LA3 (읍면동)	셀의 수	26,500	-	-	8,446	14	62,447	97,407
	비율(%)	27.2	0	0	8.7	0.01	64.1	100
OA (집계구)	셀의 수	2,804,711	-	-	263,856	1,353	625,381	3,695,301
	비율(%)	75.9	0	0	7.1	0.04	16.9	100

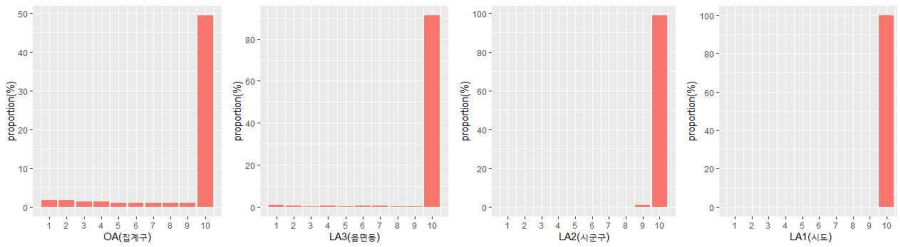
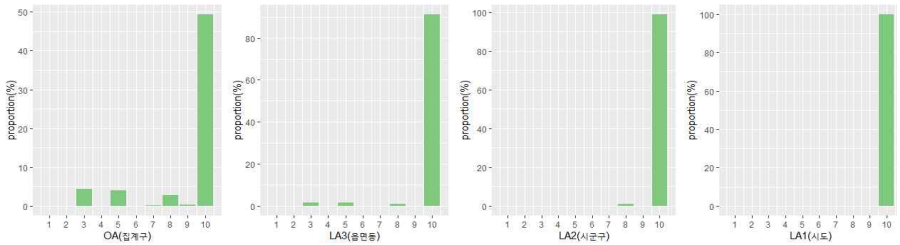
제공 항목: 1개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	-	-	-	-	-	57	57
	비율(%)	0	0	0	0	0	100	100
LA2 (시군구)	셀의 수	-	-	-	-	-	285	285
	비율(%)	0	0	0	0	0	100	100
LA3 (읍면동)	셀의 수	197	39	30	10	31	4,196	4,503
	비율(%)	4.4	0.9	0.7	0.2	0.7	93.2	100
OA (집계구)	셀의 수	67,039	3,001	2,900	2,208	2,299	93,382	170,829
	비율(%)	39.2	1.8	1.7	1.3	1.4	54.7	100

제공 항목: 1개인 경우		빈도 (비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	-	-	-	-	-	57	57
	비율(%)	0	0	0	0	0	100	100
LA2 (시군구)	셀의 수	-	-	-	-	-	285	285
	비율(%)	0	0	0	0	0	100	100
LA3 (읍면동)	셀의 수	197	-	-	72	-	4,234	4,503
	비율(%)	4.4	0	0	1.6	0	94	100
OA (집계구)	셀의 수	67,039	-	-	7,293	15	96,482	170,829
	비율(%)	39.2	0	0	4.3	0.01	56.5	100

최상위 빈도표에서는 비밀보호 알고리즘 적용에 다른 빈도의 변화가 가장 클 것으로 예상되는데, 위 제시된 표를 통해서는 그 변화를 파악하기가 쉽지 않았다. 최상위 빈도표에 가까울수록 각 셀에서 관측되는 빈도가 매우 큰 값일 가능성이 높기 때문이다. 따라서, 알고리즘 적용 전후에 상위 빈도표에서 변화가 없는 빈도가 0인 경우는 제외하고 빈도를 1부터 10 이상인 경우로 세분하여 각 빈도의 비율을 그래프로 나타내었다. 각 그래프에서 X축은 빈도를 나타내는 것으로 1부터 1씩 증가시켜 빈도가 10 이상인 경우로 표시하였고, Y축은 각 빈도표에서 해당 빈도를 가지는 셀의 수를 비율(%)로 표시하여 비교할 수 있도록 하였다.

level		OA	LA3	LA2	LA1
(지역 수준)		집계구	읍면동	시군구	시도
제공 항목 6개	비밀 보호 이전				
	비밀 보호 이후				
제공 항목 5개	비밀 보호 이전				
	비밀 보호 이후				

level		OA	LA3	LA2	LA1
(지역 수준)		집계구	읍면동	시군구	시도
제공 항목 4개	비밀 보호 이전				
	비밀 보호 이후				
제공 항목 3개	비밀 보호 이전				
	비밀 보호 이후				
제공 항목 2개	비밀 보호 이전				
	비밀 보호 이후				

level		OA	LA3	LA2	LA1
(지역 수준)		집계구	읍면동	시군구	시도
제공 항목 1개	비밀 보호 이전				
	비밀 보호 이후				

<그림 4-1> 행정 구역 단위별 조사된 상위 빈도의 분포(대전시)

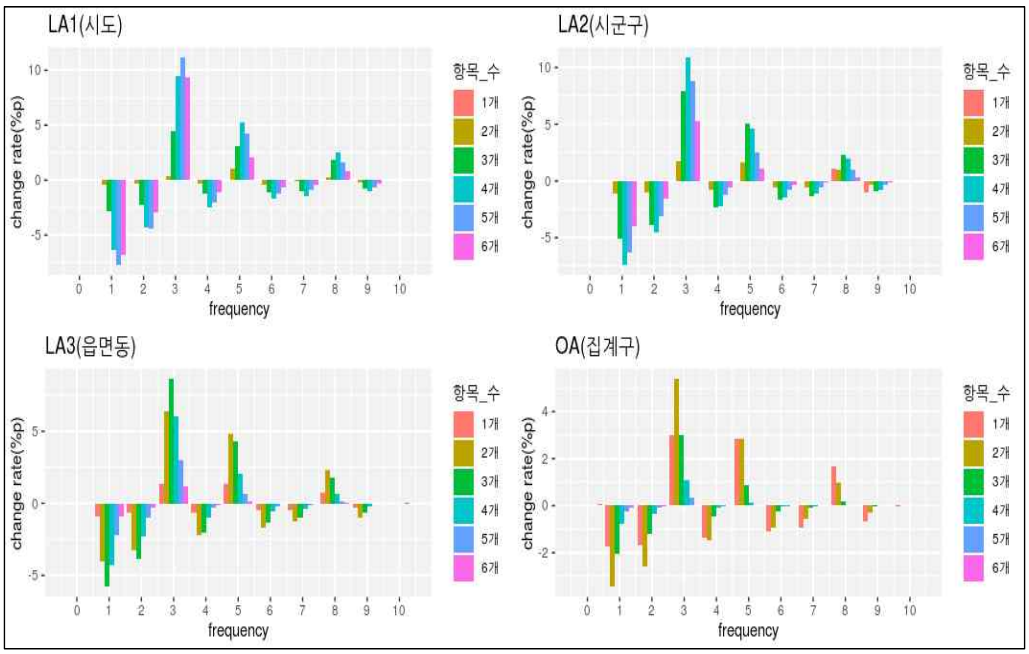
위 그래프를 통해 비교적 작은 값에 해당하는 빈도, 즉 빈도가 1부터 9인 셀의 변화를 한눈에 파악해 보고자 하였다. 알고리즘 적용 이전의 자료에서는 빈도가 1인 셀의 비율이 상대적으로 높으면서 빈도가 9인 경우까지 점진적으로 감소하는 형태를 보인다면, 알고리즘 적용 이후에는 특정 빈도를 가지는 셀의 비율 편차가 심한 형태로 나타남을 알 수 있다.

다음으로 알고리즘 적용에 따른 빈도의 비율 변화를 고려해 보았다. 빈도의 비율 변화는 특정 빈도를 가지는 셀의 비율이 알고리즘 적용 이후에 얼마나 증감하였는가를 계산한 값으로 이해할 수 있다. 비율 변화가 음수인 경우는 해당 빈도를 가지는 셀의 비율이 알고리즘 적용 이후 감소하였음을 의미하는 것이고, 양수인 경우는 알고리즘 적용 이후 해당 빈도를 가지는 셀이 비율이 증가하였음을 의미한다. 빈도의 범위는 0부터 10 이상인 경우로 하였다.

그래프상의 공통점으로 빈도가 3인 셀이 가장 큰 폭으로 증가하였고 다음으로 빈도가 5인 셀이 증가함을 확인할 수 있다. 반면 빈도가 4, 6, 7인 셀의 비율은 소폭 감소하였다는 것을 알 수 있다. 또 빈도가 10 이상인 경우는 변화가 거의 감지되지 않음도 확인할 수 있다.

보다 구체적으로 빈도가 3인 셀의 비율 변화가 10%p 이상인 경우는 제공 항목이 5개인 LA1(시도)에서 11.2%p로 가장 큰 폭의 변화를 보였고, 제공 항목이 4개인 LA2(시군구)에서 10.9%p 증가를 보였다. 다음으로 변화가 큰 경우는 제공 항목이 6개인 경우 LA1(시도)에서 9.3%p, 제공 항목이 5개인 경우 LA2(시군구)에서 8.8%p, 제공 항목이 4개인 경우 LA1(시도)에서 9.43%p, LA3(읍면동)에서 6.1%p, 제공 항목이 3개인 경

우 LA3(읍면동)에서 8.7%p, LA2(시군구)에서 8.0%p, LA1(시도)에서 4.4%p 증가하였다.



<그림 4-2> 알고리즘 적용에 따른 빈도 비율 변화(%p, 행정 구역 단위)

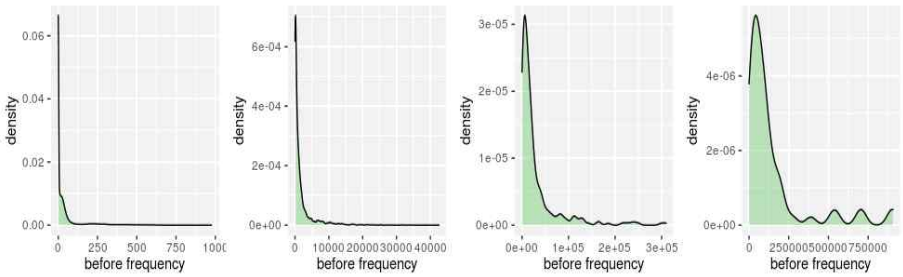
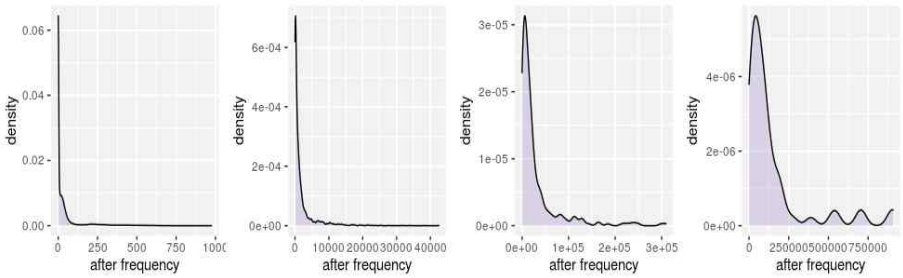
빈도가 4인 셀은 큰 폭의 비율 변화는 없으며 제공 항목이 4개인 경우 LA1(시도)에서 최대 2.5%p 감소한 것으로 파악된다.

빈도가 5인 셀의 비율 변화는 제공 항목이 4개인 경우 LA1(시도)에서 최대 5.3%p, 제공 항목이 3개인 경우 LA2(시군구)에서 5%p 증가를 보인 것이 가장 큰 폭의 변화로 감지된다.

빈도가 6 이상인 셀의 비율 변화는 대체로 ± 3%p 이내이며, 특히 빈도가 10 이상인 셀의 비율은 ± 0.1%p 이내로 크지 않음을 확인할 수 있다.

다. 최상위 빈도표의 밀도함수

빈도의 분포 변화를 전체적으로 조망하기 위해 최상위 빈도표에서 등장하는 빈도에 대한 확률밀도함수(probability density function)를 확인해 보았다. 여기서 최상위 빈도표는 제공 항목의 수가 1개인 경우에 LA1(시도)인 경우로 이해할 수 있고, 비교를 위해 LA2(시군구), LA3(읍면동), OA(집계구) 수준의 그래프를 함께 제시하였다. 그래프에서 X축은 빈도를 나타내고, Y축은 해당 빈도의 출현 확률로 이해할 수 있다. 따라서 density plot을 통해 확인되는 것처럼 알고리즘 적용에 따른 빈도의 전체적인 분포는 크게 다르지 않다고 할 수 있겠다.

level	OA	LA3	LA2	LA1
행정 구역	집계구	읍면동	시군구	시도
비밀 보호 이전				
비밀 보호 이후				

<그림 4-3> 제공 항목의 수가 1개인 경우 빈도의 분포(행정 구역 단위)

라. 비밀보호 알고리즘 적용이 필요한 최소 비율

마지막으로 <표 4-2>에서 제시된 행정 구역 단위의 조사된 빈도표를 바탕으로 비밀보호가 필요한 셀의 수를 조사해 보면 다음과 같다. 이때 비밀보호가 필요한 셀의 수는 조사된 빈도가 1 또는 2인 경우로 한정하였다. 상위 빈도표에서의 비밀보호 알고리즘은 다소 복잡한 판단 기준을 가지고 있어 모든 경우의 수를 파악하기에는 무리가 있으므로 경우를 한정하였다. 따라서 다음에 제시된 비율은 빈도표에서 비밀보호 알고리즘 적용이 필요한 최소한의 비율로 이해할 수 있다.

<표 4-3> 행정 구역 단위별 빈도표에서 비밀보호가 필요한 셀의 비율(대전시 기준)

구분	제공 항목 수					
	6개	5개	4개	3개	2개	1개
LA1(시도)	9.8%	12.2%	10.6%	5.1%	0.7%	0%
LA2(시군구)	5.5%	9.4%	12.0%	9.0%	2.1%	0%
LA3(읍면동)	1.3%	3.2%	6.6%	9.6%	7.3%	1.5%
OA(집계구)	0.1%	0.3%	1.1%	3.2%	6.0%	3.5%

기저 빈도표에 해당하는 OA(집계구) 단위에서 제공 항목 수가 6개인 경우에 비밀보호가 필요한 경우가 0.1%로 매우 낮은 이유는 빈도가 0인 셀이 전체의 99.88%를

차기하기 때문으로 파악된다. 마찬가지 이유로 OA(집계구) 단위에서 제공 항목 수가 3~5개인 경우에도 빈도가 1 또는 2인 비율이 낮음을 알 수 있다.

2. 비밀보호 알고리즘 적용에 따른 정보손실

다음으로 비밀보호 알고리즘 적용에 따른 정보손실 정도를 살펴보았다. 정보손실이란 실제 조사된 자료의 빈도와 비밀보호 알고리즘 적용 이후의 빈도의 차이로 이해할 수 있다. 기저 빈도표에서는 빈도가 1 또는 2인 경우, 0 또는 3으로 랜덤 라운딩을 적용하기 때문에 필연적으로 정보손실이 발생한다. 상위 빈도표 작성 시에는 작은 셀을 합산하여 해당 구간의 중앙값으로 조정하고, 노출 위험을 따져 추가적인 조치를 취한다는 사실을 앞서 밝힌 바 있다. 따라서 상위 빈도표에서도 정보손실은 발생할 수밖에 없는 구조를 가지고 있다. 이에 제공 항목수별로 지역 수준에 따른 정보손실을 조사해 보았다. 아래의 표를 통해 비밀보호 알고리즘 적용으로 인한 정보손실은 모두 3 이내임을 확인할 수 있다.

<표 4-4> 행정 구역 단위별 빈도표의 정보손실(대전시 기준)

제공 항목: 6개인 경우		정보손실				소계
		0	1	2	3	
LA1 (시도)	셀 개수	131,544	15,614	11,120	482	158,760
	비율(%)	82.86	9.83	7.0	0.3	100
LA2 (시군구)	셀 개수	725,583	34,090	33,487	640	793,800
	비율(%)	91.41	4.29	4.22	0.08	100
LA3 (읍면동)	셀 개수	12,343,195	71,793	124,625	2,427	12,542,040
	비율(%)	98.41	0.57	0.99	0.02	100
OA (집계구)	셀 개수	475,376,300	285,479	141,941	-	475,803,720
	비율(%)	99.91	0.06	0.03	0	100

제공 항목: 5개인 경우		정보손실				소계
		0	1	2	3	
LA1 (시도)	셀 개수	120,894	37,766	14,721	751	174,132
	비율(%)	69.43	21.69	8.45	0.43	100
LA2 (시군구)	셀 개수	709,200	99,577	59,818	2,065	870,660
	비율(%)	81.46	11.44	6.87	0.24	100
LA3 (읍면동)	셀 개수	13,114,274	302,051	329,640	10,463	13,756,428
	비율(%)	95.33	2.20	2.40	0.08	100
OA (집계구)	셀 개수	520,000,932	540,280	1,315,356	17,036	521,873,604
	비율(%)	99.64	0.104	0.25	0.003	100

제공 항목: 4개인 경우		정보손실				소계
		0	1	2	3	
LA1 (시도)	셀 개수	36,094	27,379	4,925	308	68,706
	비율(%)	52.53	39.85	7.17	0.45	100
LA2 (시군구)	셀 개수	222,734	90,973	28,393	1,430	343,530
	비율(%)	64.84	26.48	8.27	0.42	100
LA3 (읍면동)	셀 개수	4,745,036	414,461	256,701	11,576	5,427,774
	비율(%)	87.42	7.64	4.73	0.21	100
OA (집계구)	셀 개수	202,935,718	1,171,924	1,756,329	47,911	205,911,882
	비율(%)	98.55	0.57	0.85	0.02	100

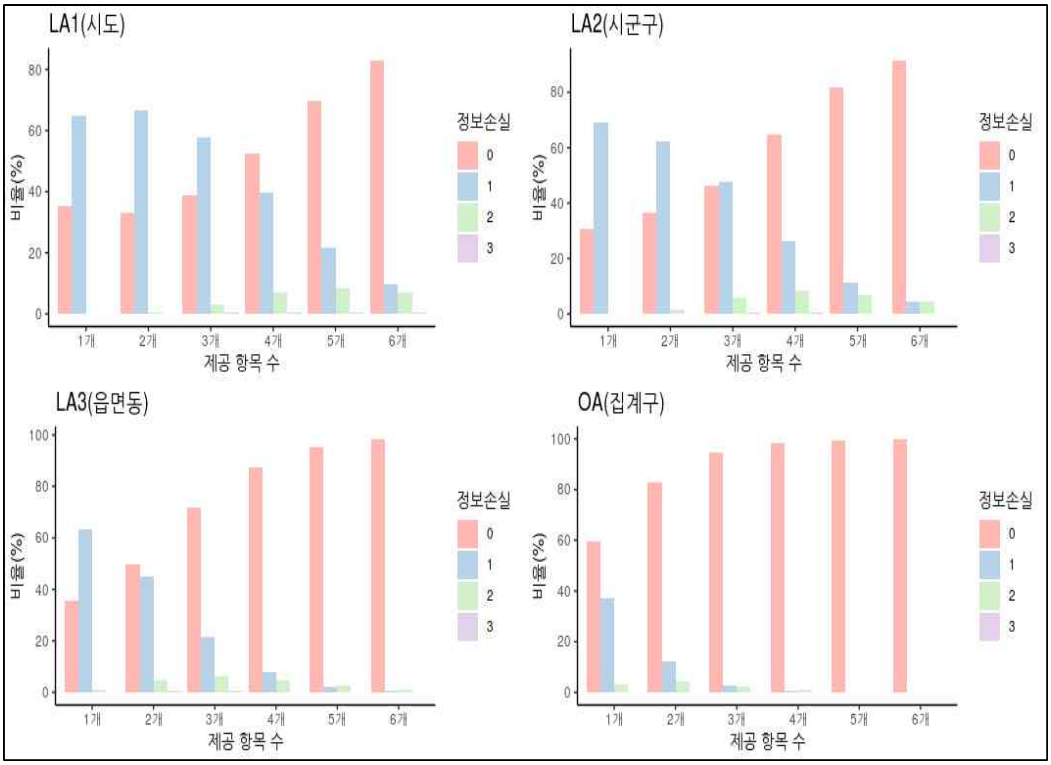
제공 항목: 3개인 경우		정보손실				소계
		0	1	2	3	
LA1 (시도)	셀 개수	4,998	7,456	421	36	12,911
	비율(%)	38.71	57.75	3.26	0.28	100
LA2 (시군구)	셀 개수	29,758	30,748	3,791	258	64,555
	비율(%)	46.10	47.63	5.87	0.40	100
LA3 (읍면동)	셀 개수	729,554	219,915	66,617	3,883	1,019,969
	비율(%)	71.53	21.56	6.53	0.38	100
OA (집계구)	셀 개수	36,617,917	1,112,089	924,417	39,844	38,694,267
	비율(%)	96.63	2.87	2.39	0.10	100

제공 항목: 2개인 경우		정보손실				소계
		0	1	2	3	
LA1 (시도)	셀 개수	406	818	8	1	1,233
	비율(%)	32.93	66.34	0.65	0.08	100
LA2 (시군구)	셀 개수	2,236	3,841	81	7	6,165
	비율(%)	36.27	62.30	1.31	0.11	100
LA3 (읍면동)	셀 개수	48,569	43,946	4,533	359	97,407
	비율(%)	49.86	45.12	4.65	0.37	100
OA (집계구)	셀 개수	3,072,155	447,566	165,845	9,735	3,695,301
	비율(%)	83.14	12.11	4.49	0.26	100

제공 항목: 1개인 경우		정보손실				소계
		0	1	2	3	
LA1 (시도)	셀 개수	20	37	-	-	57
	비율(%)	35.09	64.91	0	0	100
LA2 (시군구)	셀 개수	88	197	-	-	285
	비율(%)	30.88	69.12	0	0	100
LA3 (읍면동)	셀 개수	1,601	2,855	42	5	4,503
	비율(%)	35.55	63.40	0.93	0.11	100
OA (집계구)	셀 개수	101,287	63,369	5,758	415	170,829
	비율(%)	59.29	37.09	3.37	0.24	100

기저 빈도표인 제공 항목이 6개이고 OA(집계구)에서는 정보손실이 0인 경우가 99.9%를 차지하였는데 그 이유는 관측된 기저 빈도표에서 대부분의 빈도가 0이었고, 빈도가 1 또는 2인 셀의 비율이 0.1%에 불과하였기 때문이다.

최상위 빈도표인 제공 항목이 1개인 경우 LA1(시도), LA2(시군구)에서 정보손실이 2 이상인 경우는 발생하지 않았고, 거의 대부분인 65~70%의 셀에서 정보손실이 1만큼만 발생하였다. 또 모든 경우의 빈도표에 있어 정보손실이 3만큼 발생하는 셀의 비율은 0.5%를 넘지 않는 수준에서 제어되고 있음을 확인할 수 있다.



<그림 4-4> 행정 구역 단위별 정보손실(대전시)

OA(집계구) 수준에서는 제공 항목의 수가 3~6개인 경우 정보손실이 없는 경우가 90% 이상을 차지하고, 제공 항목의 수가 2개인 경우 83% 수준, 제공 항목의 수가 1개인 경우 59% 수준으로 낮아짐을 확인할 수 있다. 전체적으로 정보손실 정도가 2 또는 3인 경우는 10% 이내임이 확인되고,, 정보손실 정도가 1인 경우는 제공 항목의 수가 줄어들수록 점차 증가하는 경향을 보인다. 특히, 제공 항목의 수가 1개인 경우 LA2(시군구) 수준에서 약 69.12%에서 정보손실이 1만큼 발생하여 최대치를 보였다.

제2절 격자 단위 자료의 정합성

1. 비밀보호 알고리즘 적용에 따른 빈도의 변화

가. 기저 빈도표의 변화

격자 단위 자료의 빈도표에 대해서도 비밀보호 적용에 따른 빈도의 분포를 검토하였다. 마찬가지로 기저 빈도표에서는 빈도가 0인 셀의 수가 99.96%로 거의 대부분을 차지함을 알 수 있고, 빈도가 1 또는 2인 경우에 비밀보호가 필요하다고 한다면 그 비율은 0.031%로 행정 구역 단위 자료에 비해 낮음을 알 수 있다. 비밀보호 이후에는 빈도가 3인 셀의 수가 5배 이상 증가하였고, 빈도가 4 이상인 셀의 수는 변동이 없음을 알 수 있다.

<표 4-5> 2018년 인구주택총조사 자료의 기저 빈도표 분포(대전시 해당)

제공 항목: 6개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
OA (100m)	셀의 수	1,596,514,330	408,845	87,335	34,524	18,938	61,628	1,597,125,600
	비율(%)	99.96	0.026	0.005	0.002	0.001	0.004	100

제공 항목: 6개인 경우		빈도 (비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
OA (100m)	셀의 수	1,596,816,226	-	-	228,808	18,938	61,628	1,597,125,600
	비율(%)	99.98	0	0	0.01	0.001	0.004	100

나. 상위 빈도표의 변화

동일한 항목 수에 대해 지역 수준을 넓혀가는 방식으로 상위 빈도표를 만들어 빈도의 변화 양상을 관찰하기 위해 빈도가 0부터 5 이상인 경우로 나누어 조사하였다.

격자 단위에서의 빈도표에서도 비밀보호 알고리즘 적용 후에 빈도가 0인 셀이 증가하는 경우는 기저 빈도표에서만 발생함을 알 수 있다. 또 비밀보호 알고리즘 적용 전·후 나타나는 빈도의 변화 역시 매우 흡사함을 확인할 수 있다. 행정 구역 단위 자료와 차이점을 찾는다면 제공 항목이 1개인 경우 LA2 수준에서 관측된 빈도가 0 ~ 3인 셀이 존재한다는 점이다.

<표 4-6> 2018년 인구주택총조사 자료의 상위 빈도표 분포(대전시 해당 격자)

제공 항목: 6개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	123,071	10,736	4,746	2,717	1,755	15,735	158,760
	비율(%)	77.5	6.8	3.0	1.7	1.1	9.9	100
LA2 (시군구)	셀의 수	1,513,207	28,341	10,507	5,675	3,842	26,028	1,587,600
	비율(%)	95.3	1.8	0.7	0.4	0.2	1.6	100
LA3 (읍면동)	셀의 수	66,867,290	150,343	46,887	22,805	13,984	54,171	67,155,480
	비율(%)	99.6	0.2	0.1	0.03	0.02	0.1	100

제공 항목: 6개인 경우		빈도(비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	123,071	-	-	17,609	56	18,024	158,760
	비율(%)	77.5	0	0	11.1	0.04	11.4	100
LA2 (시군구)	셀의 수	1,513,207	-	-	42,989	255	31,149	1,587,600
	비율(%)	95.3	0	0	2.7	0.02	2	100.02
LA3 (읍면동)	셀의 수	66,867,459	-	-	212,907	2,354	72,760	67,155,480
	비율(%)	99.6	0	0	0.3	0.003	0.1	100

제공 항목: 5개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	101,406	13,479	7,723	5,051	3,627	42,846	174,132
	비율(%)	58.2	7.7	4.4	2.9	2.1	24.6	100
LA2 (시군구)	셀의 수	1,552,388	54,441	25,114	14,294	9,946	85,137	1,741,320
	비율(%)	89.2	3.1	1.4	0.8	0.6	4.9	100
LA3 (읍면동)	셀의 수	72,626,732	444,668	169,316	87,195	55,922	274,003	73,657,836
	비율(%)	98.6	0.6	0.2	0.1	0.1	0.4	100
OA (집계구)	셀의 수	1,748,953,561	1,690,737	460,500	188,672	104,785	369,665	1,751,767,920
	비율(%)	99.8	0.1	0.03	0.01	0.01	0.02	100

제공 항목: 5개인 경우		빈도 (비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	101,406	-	-	24,510	37	48,179	174,132
	비율(%)	58.2	0	0	14.1	0.02	27.7	100
LA2 (시군구)	셀의 수	1,552,388	-	-	88,764	268	99,900	1,741,320
	비율(%)	89.2	0	0	5.1	0.02	5.7	100
LA3 (읍면동)	셀의 수	72,626,732	-	-	669,638	4,367	357,099	73,657,836
	비율(%)	98.6	0	0	0.9	0.01	0.5	100
OA (집계구)	셀의 수	1,748,953,561	-	-	2,314,905	58,693	440,761	1,751,767,920
	비율(%)	99.8	0	0	0.1	0.003	0.03	100

제공 항목: 4개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	22,049	4,368	2,924	2,121	1,700	35,544	68,706
	비율(%)	32.1	6.4	4.3	3.1	2.5	51.7	100
LA2 (시군구)	셀의 수	527,194	32,428	17,558	10,632	8,037	91,211	687,060
	비율(%)	76.7	4.7	2.6	1.6	1.2	13.3	100
LA3 (읍면동)	셀의 수	27,765,038	432,226	199,720	108,801	74,975	481,878	29,062,638
	비율(%)	95.5	1.5	0.7	0.4	0.3	1.7	100
OA (집계구)	셀의 수	686,165,274	2,575,609	910,398	402,574	236,718	891,787	691,182,360
	비율(%)	99.3	0.4	0.1	0.1	0.03	0.1	100

제공 항목: 4개인 경우		빈도 (비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	22,049	-	-	8,648	3	38,006	68,706
	비율(%)	32.1	0	0	12.6	0.04	55.3	100
LA2 (시군구)	셀의 수	527,194	-	-	56,404	74	103,388	687,060
	비율(%)	76.73	0	0	8.21	0.01	15	100
LA3 (읍면동)	셀의 수	27,765,038	-	-	697,754	2,287	597,559	29,062,638
	비율(%)	95.5	0	0	2.4	0.01	2.1	100
OA (집계구)	셀의 수	686,165,274	-	-	3,761,630	56,651	1,198,805	691,182,360
	비율(%)	99.3	0	0	0.5	0.01	0.2	100

제공 항목: 3개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	1,210	367	292	194	157	10,691	12,911
	비율(%)	9.4	2.8	2.3	1.5	1.2	82.8	100
LA2 (시군구)	셀의 수	72,466	7,438	4,643	2,965	2,367	39,231	129,110
	비율(%)	56.1	5.8	3.6	2.3	1.8	30.4	100
LA3 (읍면동)	셀의 수	4,754,184	171,103	93,209	53,092	39,738	350,027	5,461,353
	비율(%)	87.1	3.1	1.7	1.0	0.7	6.4	100
OA (집계구)	셀의 수	125,653,604	1,694,432	788,529	392,605	257,915	1,097,575	129,884,660
	비율(%)	96.7	1.3	0.6	0.3	0.2	0.8	100

제공 항목: 3개인 경우		빈도 (비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	1,210	-	-	786	-	10,915	12,911
	비율(%)	9.4	0	0	6.1	0	84.5	100
LA2 (시군구)	셀의 수	72,466	-	-	13,803	6	42,835	129,110
	비율(%)	56.1	0	0	10.7	0.005	33.2	100
LA3 (읍면동)	셀의 수	4,754,184	-	-	295,464	402	411,303	5,461,353
	비율(%)	87.1	0	0	5.4	0.01	7.5	100
OA (집계구)	셀의 수	125,653,604	-	-	2,727,353	22,393	1,481,310	129,884,660
	비율(%)	96.7	0	0	2.1	0.02	1.1	100

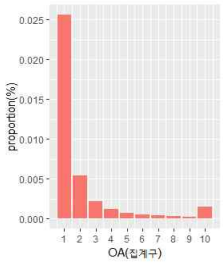
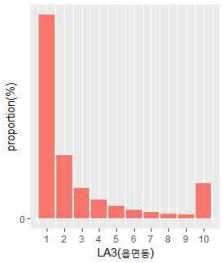
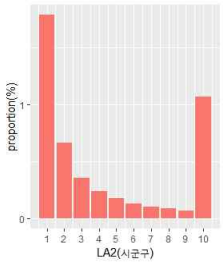
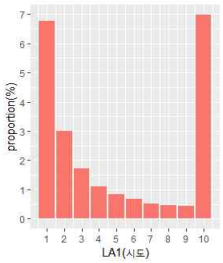
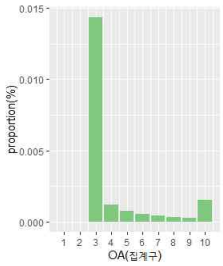
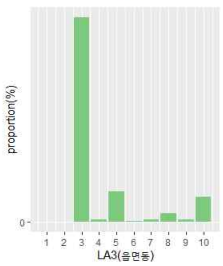
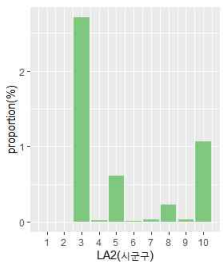
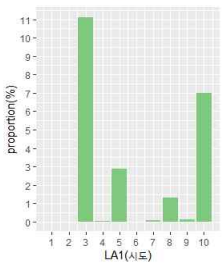
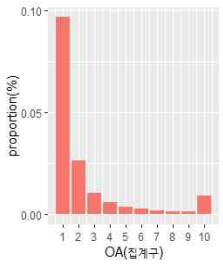
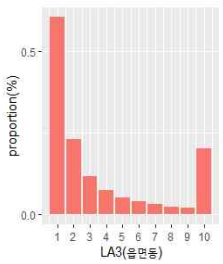
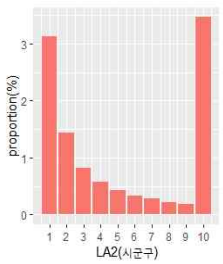
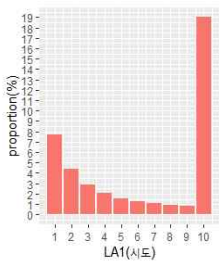
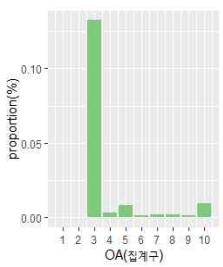
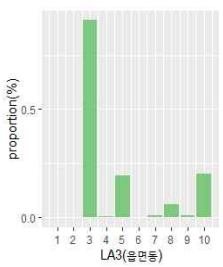
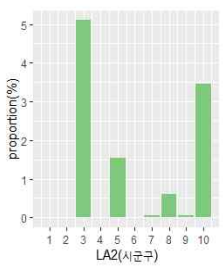
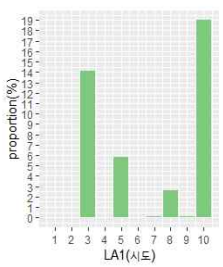
제공 항목: 2개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	11	5	4	5	4	1,204	1,233
	비율(%)	0.89	0.41	0.32	0.41	0.3	97.6	100
LA2 (시군구)	셀의 수	3,688	609	470	296	259	7,008	12,330
	비율(%)	29.91	4.94	3.81	2.40	2.1	56.8	100
LA3 (읍면동)	셀의 수	351,847	27,344	18,024	10,733	8,585	105,026	521,559
	비율(%)	67.46	5.24	3.46	2.06	1.7	20.1	100
OA (집계구)	셀의 수	10,750,927	436,561	269,930	155,095	117,985	673,482	12,403,980
	비율(%)	86.67	3.52	2.18	1.25	1.0	5.4	100

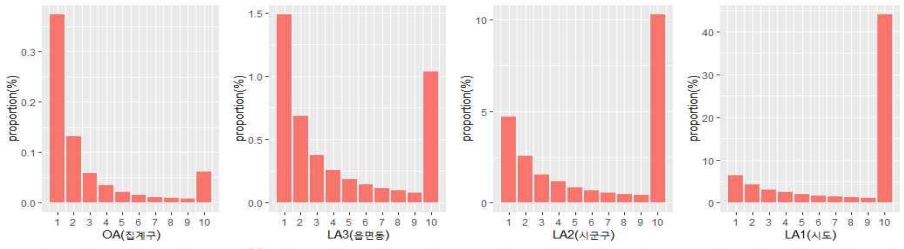
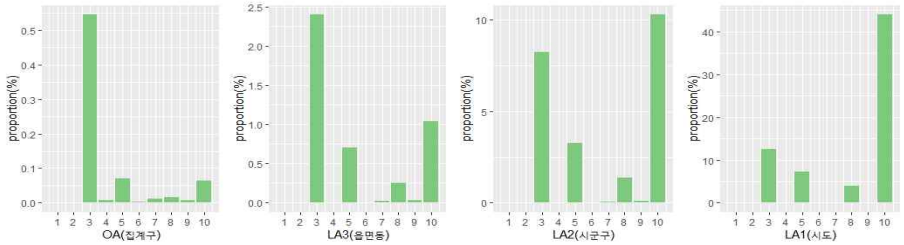
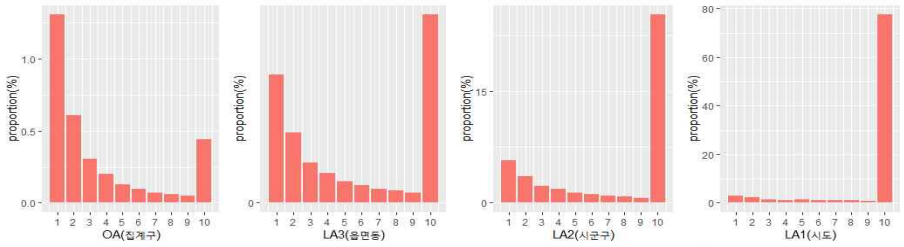
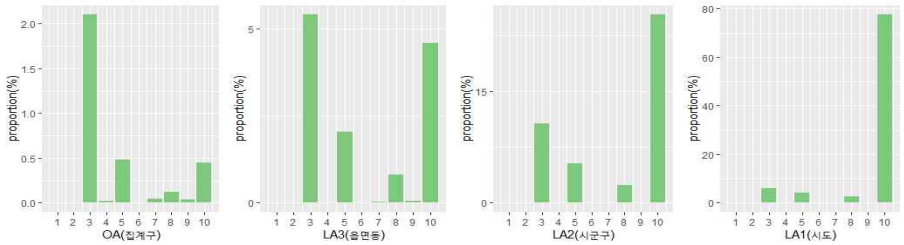
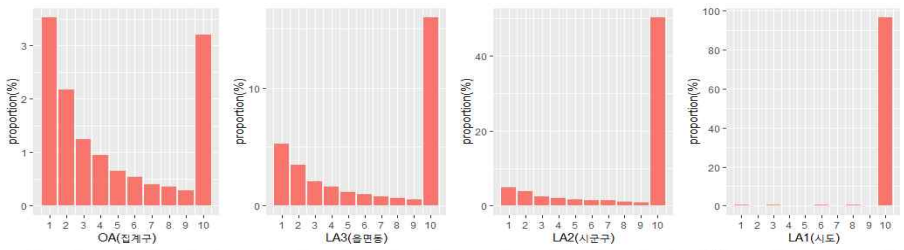
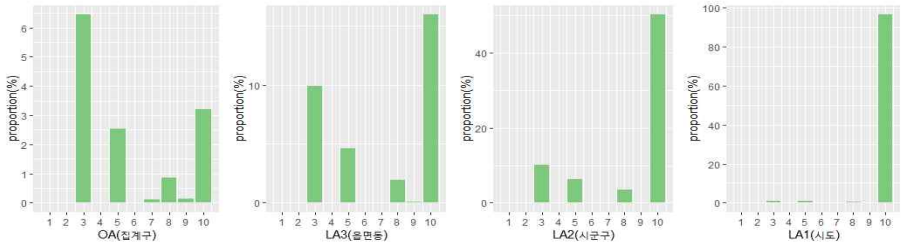
제공 항목: 2개인 경우		빈도 (비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	11	-	-	12	-	1,210	1,233
	비율(%)	0.89	0	0	0.97	0	98.1	100
LA2 (시군구)	셀의 수	3,688	-	-	1,250	-	7,392	12,330
	비율(%)	29.91	0	0	10.14	0	60	100
LA3 (읍면동)	셀의 수	351,847	-	-	51,597	30	118,085	521,559
	비율(%)	67.46	0	0	9.89	0.01	22.6	100
OA (집계구)	셀의 수	10,750,927	-	-	799,258	2,797	850,998	12,403,980
	비율(%)	86.67	0	0	6.44	0.02	6.9	100

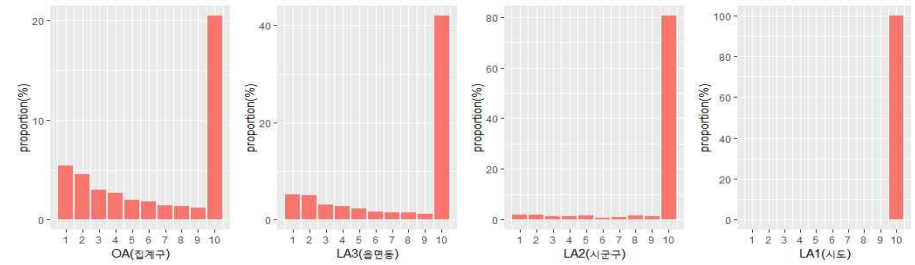
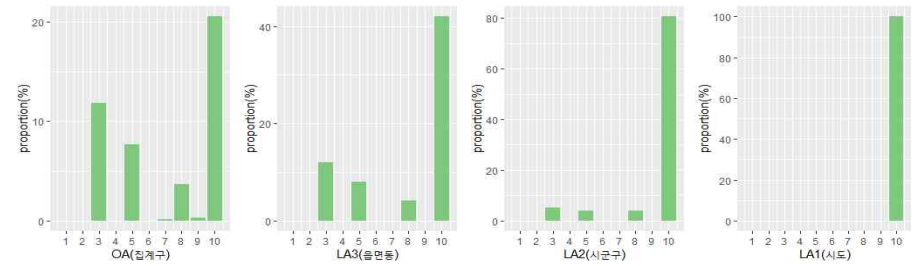
제공 항목: 1개인 경우		빈도 (비밀보호 이전)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	-	-	-	-	-	57	57
	비율(%)	0	0	0	0	0	100	100
LA2 (시군구)	셀의 수	35	12	12	7	8	496	570
	비율(%)	6.14	2.11	2.11	1.23	1.4	87	100
LA3 (읍면동)	셀의 수	8,152	1,241	1,205	763	675	12,075	24,111
	비율(%)	33.81	5.15	5.00	3.16	2.8	50.1	100
OA (집계구)	셀의 수	320,258	31,112	26,252	17,427	15,613	162,758	573,420
	비율(%)	55.85	5.43	4.58	3.04	2.7	28.4	100

제공 항목: 1개인 경우		빈도 (비밀보호 이후)						소계
		0	1	2	3	4	5 이상	
LA1 (시도)	셀의 수	-	-	-	-	-	57	57
	비율(%)	-	0	0	-	0	100	100
LA2 (시군구)	셀의 수	35	-	-	30	-	505	570
	비율(%)	6.14	0	0	5.26	0	88.6	100
LA3 (읍면동)	셀의 수	8,152	-	-	2,893	1	13,065	24,111
	비율(%)	33.81	0	0	12.00	0.004	54.2	100
OA (집계구)	셀의 수	320,258	-	-	67,777	71	185,314	573,420
	비율(%)	55.85	0	0	11.82	0.01	32.3	100

마찬가지로 비밀보호 알고리즘 적용에 따른 빈도의 변화를 상세히 관찰하기 위해 빈도가 0인 경우는 제외하고 빈도를 1부터 10 이상인 경우로 세분하여 해당하는 셀의 수를 조사한 그래프로 나타내었다. 각 그래프에서 X축은 1부터 10 이상에 해당하는 빈도를 나타내었고, Y축은 해당 빈도를 가지는 셀의 수를 비율(%)로 표시하였다.

level		OA	LA3	LA2	LA1
(지역 수준)		100m	1km	10km	100km
제공 항목 6개	비밀 보호 이전				
	비밀 보호 이후				
제공 항목 5개	비밀 보호 이전				
	비밀 보호 이후				

level		OA	LA3	LA2	LA1
(지역 수준)		100m	1km	10km	100km
제공 항목 4개	비밀 보호 이전				
	비밀 보호 이후				
제공 항목 3개	비밀 보호 이전				
	비밀 보호 이후				
제공 항목 2개	비밀 보호 이전				
	비밀 보호 이후				

level		OA	LA3	LA2	LA1
(지역 수준)		100m	1km	10km	100km
제공 항목 1개	비밀 보호 이전				
	비밀 보호 이후				

<그림 4-5> 격자 단위별 BSCA 적용 전·후 상위 빈도의 분포(대전시 해당)

격자 단위 자료에서도 빈도가 1~9 정도로 작은 값인 경우에는 관측된 빈도가 점진적으로 감소하는 형태를 보이지만, 알고리즘 적용 이후에는 특정 빈도로 변환되는 양상이 동일하게 관측되었다.

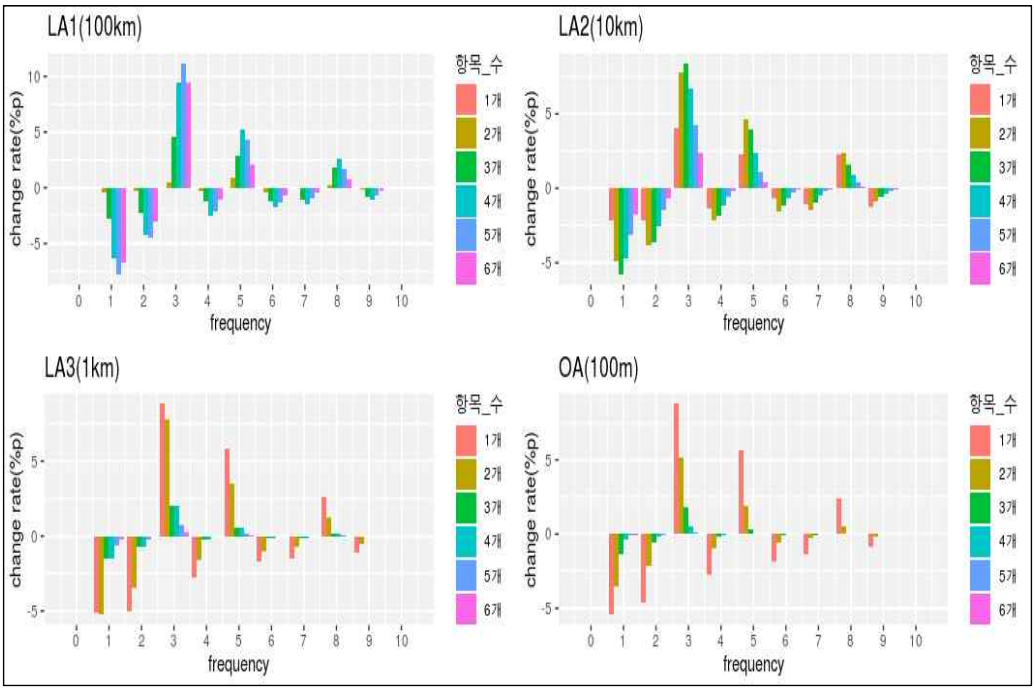
알고리즘 적용에 따른 빈도의 비율 변화를 살펴봄으로써 보다 자세히 관찰해 보았다. 행정 구역 단위의 빈도표와 마찬가지로 빈도가 3인 셀이 가장 큰 폭으로 증가하였고, 빈도가 5 또는 8인 셀의 비율이 증가하였다. 반면 빈도가 4, 6, 7, 9인 셀은 감소하였으며, 빈도가 10 이상인 경우에는 변화가 거의 감지되지 않았다.

구체적으로는 비밀보호 알고리즘 적용 이후 빈도가 3인 셀의 비율이 가장 큰 폭으로 증가한 경우는 제공 항목이 5개인 LA1(100km)에서 11.2%p, 제공 항목이 4개인 LA1(100km)에서 9.5%p, 제공 항목이 6개인 LA1(100km)에서 9.4%p 이다. 제공 항목이 3개인 LA2(10km)에서 8.4%p, LA1(100km)에서 4.6%p, LA3(1km)에서 4.4%p 증가하였다.

빈도가 4인 셀의 비율 변화는 크지 않으며 제공 항목이 1개인 경우 LA3(1km)에서 최대 2.8%p 감소한 것으로 파악된다.

빈도가 5인 셀의 비율 변화는 제공 항목이 4개인 경우 LA1(100km)에서 최대 5.23%p, 제공 항목이 5개인 경우 LA1(100km)에서 4.3%p 증가를 보인 것이 비교적 큰 폭의 변화로 감지된다.

빈도가 6 이상인 셀의 비율 변화는 대체로 ± 3%p 이내이며, 특히 빈도가 10 이상인 셀의 비율은 ± 0.02%p 이내로 크지 않음을 확인할 수 있다.



<그림 4-6> 알고리즘 적용에 따른 빈도 비율 변화(%p, 격자 단위)

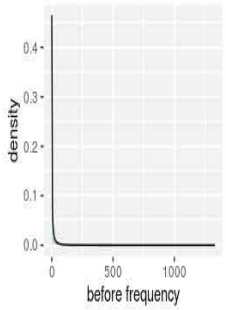
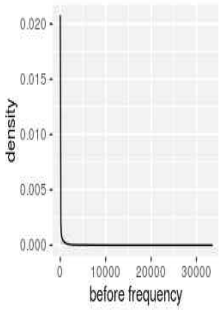
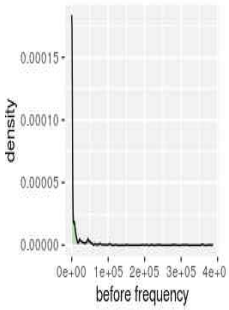
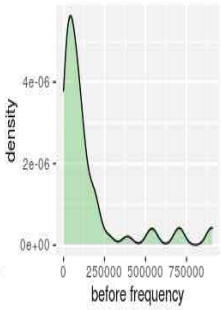
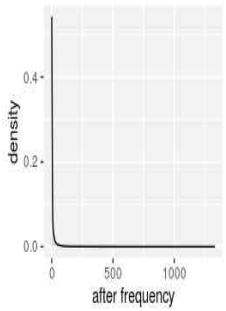
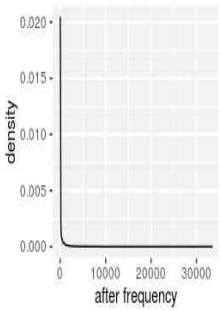
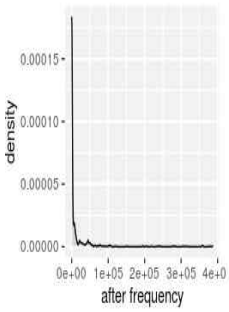
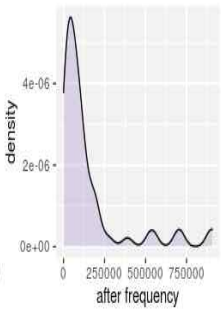
다. 행정 구역 단위 빈도표와 격자 단위 빈도표의 변화 비교

행정 구역 단위 자료와 격자 단위 자료에서 공통적으로 변화된 것은 빈도가 3, 5인 셀의 비율은 증가하고, 빈도가 4, 6, 7인 셀의 비율은 감소하였다. 특히 빈도가 3인 셀의 비율이 큰 폭으로 증가하게 되는데 최대 11.2%p까지 증가함을 확인하였고, 빈도가 4 이상인 경우에는 대체로 5%p 수준에서만 변화되었다. 전체적으로 두 빈도표의 변화 양상은 상당히 유사하다고 볼 수 있다.

두 빈도표의 차이점으로는 제공 항목이 1개인 경우 OA 수준에서 빈도가 3인 셀의 수는 비밀보호 알고리즘 적용 이후 격자 단위 빈도표에서는 약 8%p 이상 증가하여, 행정 구역 단위 빈도표에서 약 3%p 증가한데 비해 그 변화가 훨씬 더 크다는 것을 확인할 수 있다. 다시 말해 최상위 빈도표에 가까운 제공 항목이 1개인 경우 행정 구역 단위 빈도표에서는 특정 빈도를 가지는 셀의 비율에 거의 변화가 없는데 비해, 상대적으로 격자 단위 빈도표에서는 비교적 큰 변화를 보였다.

라. 최상위 빈도표의 밀도함수

격자 단위 자료에 대해서도 제공 항목의 수가 1개인 경우에 대한 확률밀도함수를 확인해 보았다. 앞서와 동일하게 X축은 빈도를 나타내고, Y축은 해당 빈도의 출현 확률로 이해하면 된다. 위 그래프에서 확인되는 바와 같이 알고리즘 적용에 따른 빈도의 전체적인 분포는 크게 다르지 않다고 할 수 있다. 행정 구역 단위 자료와 비교하여 최상위 빈도표인 LA1 수준에서의 밀도함수는 매우 흡사한 형태를 보이며, LA2 수준에서는 변화를 감지하기가 더 어려운 측면이 있다.

level	OA	LA3	LA2	LA1
격자 단위	100m	1Km	10Km	100Km
비밀 보호 이전				
비밀 보호 이후				

<그림 4-7> 제공 항목의 수가 1개인 경우 빈도의 분포(격자 단위)

2. 비밀보호 알고리즘 적용에 따른 정보손실

격자 단위 자료에서 실제 조사된 빈도와 비밀보호 알고리즘 적용 이후의 빈도의 차이를 의미하는 정보손실 정도를 살펴보았다. 제공 항목 수별로 지역 수준에 따른 정보손실을 조사해 보았다. 제공 항목 수가 동일한 경우 지역 수준이 넓어지면서 정보손실은 커지게 된다. 구체적으로 살펴보면 제공 항목이 6개인 경우에는 지역 수준이 넓어지더라도 80% 이상에서 정보손실이 발생하지 않는다. 제공 항목이 5개인 경우에 지역 수준이 LA1으로 넓어지면 정보손실이 1만큼 발생하는 셀이 20%를 초과하게 되고, 제공 항목이 4개인 경우에는 그 비율이 약 40%에 달하게 된다. 이후 제공 항목이 3개인 경우에는 그 비율이 약 60%에 달함과 동시에, 지역 수준이 LA2인 경우에도 그 비율이 20%를 초과하게 된다. 이후 제공 항목이 2개, 1개로 줄어들수록 해당 비율은 급격히 증가함을 알 수 있다.

<표 4-7> 격자 단위별 빈도표의 정보손실(대전시 해당 격자 기준)

제공 항목: 6개인 경우		정보손실				소계
		0	1	2	3	
LA1 (100km)	셀 개수	131,501	15,784	11,017	458	158,760
	비율(%)	82.83	9.94	6.94	0.29	100
LA2 (10km)	셀 개수	1,528,499	28,671	29,857	573	1,587,600
	비율(%)	96.28	1.81	1.88	0.04	100
LA3 (1km)	셀 개수	66,916,096	79,348	157,102	2,934	67,155,480
	비율(%)	99.64	0.12	0.23	0.004	100
OA (100m)	셀 개수	1,596,629,420	331,299	164,881	-	1,597,125,600
	비율(%)	99.97	0.02	0.01	0	100

제공 항목: 5개인 경우		정보손실				소계
		0	1	2	3	
LA1 (100km)	셀 개수	120,933	37,777	14,627	795	174,132
	비율(%)	69.45	21.69	8.40	0.46	100
LA2 (10km)	셀 개수	1,595,528	85,082	58,682	2,028	1,741,320
	비율(%)	91.63	4.89	3.37	0.12	100
LA3 (1km)	셀 개수	72,817,568	353,141	473,473	13,654	73,657,836
	비율(%)	98.86	0.48	0.64	0.02	100
OA (100m)	셀 개수	1,749,447,615	560,202	1,741,627	18,476	1,751,767,920
	비율(%)	99.87	0.032	0.10	0.001	100

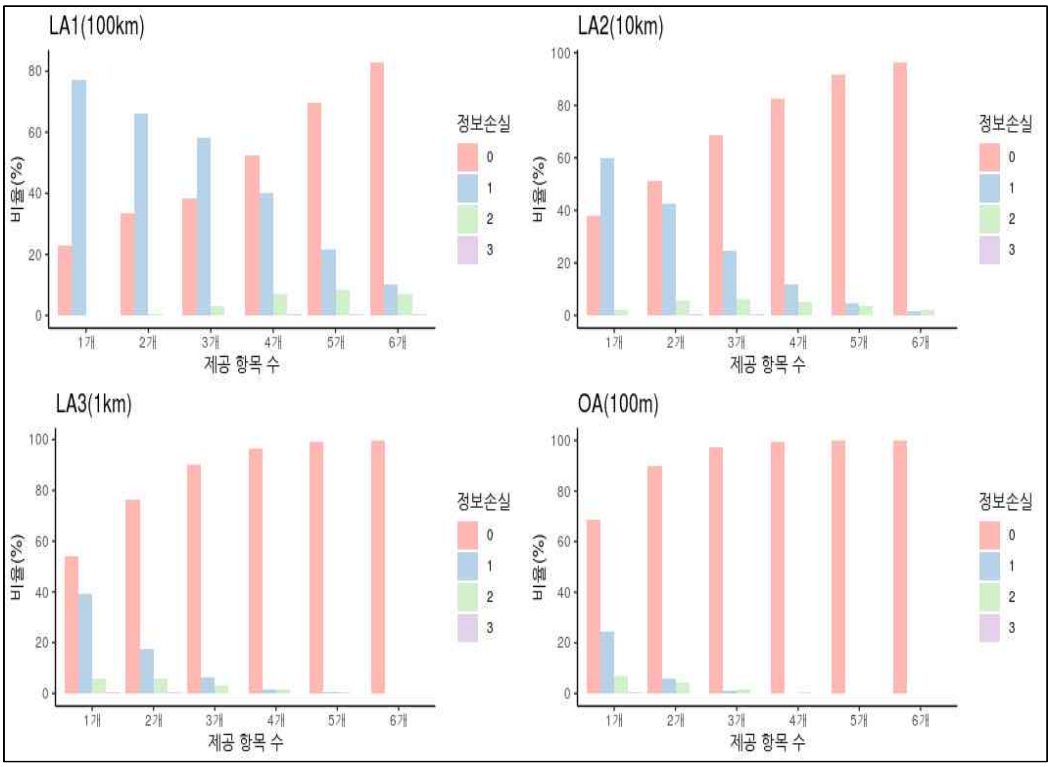
제공 항목: 4개인 경우		정보손실				소계
		0	1	2	3	
LA1 (100km)	셀 개수	36,113	27,408	4,892	293	68,706
	비율(%)	52.56	39.89	7.12	0.43	100
LA2 (10km)	셀 개수	568,091	81,787	35,469	1,713	687,060
	비율(%)	82.68	11.90	5.16	0.25	100
LA3 (1km)	셀 개수	28,039,404	537,538	467,076	18,620	29,062,638
	비율(%)	96.48	1.85	1.61	0.06	100
OA (100m)	셀 개수	687,063,130	1,326,503	2,730,605	62,122	691,182,360
	비율(%)	99.40	0.19	0.40	0.01	100

제공 항목: 3개인 경우		정보손실				소계
		0	1	2	3	
LA1 (100km)	셀 개수	4,973	7,503	405	30	12,911
	비율(%)	38.52	58.11	3.14	0.23	100
LA2 (10km)	셀 개수	88,422	31,912	8,256	520	129,110
	비율(%)	68.49	24.72	6.39	0.40	100
LA3 (1km)	셀 개수	4,924,078	340,851	187,201	9,223	5,461,353
	비율(%)	90.16	6.24	3.43	0.17	100
OA (100m)	셀 개수	126,498,048	1,456,258	1,861,064	69,290	129,884,660
	비율(%)	97.39	1.12	1.43	0.05	100

제공 항목: 2개인 경우		정보손실				소계
		0	1	2	3	
LA1 (100km)	셀 개수	412	814	7	-	1,233
	비율(%)	33.41	66.02	0.57	0	100
LA2 (10km)	셀 개수	6,334	5,256	687	53	12,330
	비율(%)	51.37	42.63	5.57	0.43	100
LA3 (1km)	셀 개수	397,505	91,871	30,333	1,850	521,559
	비율(%)	76.21	17.61	5.82	0.35	100
OA (100m)	셀 개수	11,145,222	728,755	502,736	27,267	12,403,980
	비율(%)	89.85	5.88	4.05	0.22	100

제공 항목: 1개인 경우		정보손실				소계
		0	1	2	3	
LA1 (100km)	셀 개수	13	44	-	-	57
	비율(%)	22.81	77.19	0	0	100
LA2 (10km)	셀 개수	215	342	13	-	570
	비율(%)	37.72	60.00	2.28	0	100
LA3 (1km)	셀 개수	13,063	9,480	1,442	126	24,111
	비율(%)	54.18	39.32	5.98	0.52	100
OA (100m)	셀 개수	392,721	139,398	38,450	2,851	573,420
	비율(%)	68.49	24.31	6.71	0.50	100

다음으로 동일한 지역 수준 내에서 제공 항목의 수가 줄어드는 방식의 상위 빈도표에서 정보손실이 얼마나 증가하는지를 그래프로 나타내었다. 최상위 빈도표인 제공 항목이 1개이고 격자 단위가 LA1(100km)인 빈도표에서 정보손실이 2 이상인 경우는 발생하지 않았고, 거의 대부분의 경우(77.19%)에서 정보손실이 1만큼만 발생하였다. 또 모든 경우의 빈도표에 있어 정보손실이 3만큼 발생하는 셀의 비율은 0.5%를 넘지 않는 수준에서 제어되고 있음을 확인할 수 있다. 이러한 현상은 행정 구역 단위 빈도표의 정보손실에서도 동일한 패턴임을 확인한 바 있다.



<그림 4-8> 격자 단위별 제공 항목 수에 따른 정보손실

제3절 현행 자료 제공 방식과의 비교

현행 자료 제공 방식과 비밀보호 알고리즘 적용 후 제공되는 자료의 차이를 확인해 보자. SGIS 대화형 통계 지도의 현행 자료 제공 방식은 사용자 요청에 의한 제공 항목별 조합과 지역 수준에 따라 기저 빈도표와 상위 빈도표 모두에서 5 미만의 값에 대해 N/A 표시하여 제공하고 있다. 반면 비밀보호 알고리즘을 적용하면 기저 빈도표는 랜덤 라운딩하여 제공하고, 상위 빈도표는 중앙값 조정 및 추가적인 노출 위험 제어를 위한 조치를 취하여 제공하기로 하였다.

이와 관련하여 보다 상세한 이해를 돕고자 제3장에서 소개된 예시 자료를 인용하여 설명하려고 한다. 원 자료(raw data)에 해당하는 조사된 빈도표와 현행 제공 방식의 빈도표를 비교해 보자.

<표 4-8> 조사된 빈도표와 현행 제공 방식의 빈도표 비교

조사된 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	1	0	0	0	13
	아파트	18	3	2	0	36
	연립주택	4	4	1	7	22
	다세대	5	8	0	18	12
	비주거용	1	2	0	4	15
여	단독주택	4	5	8	3	1
	아파트	9	4	9	2	1
	연립주택	5	10	14	5	1
	다세대	10	6	7	8	1
	비주거용	4	5	11	11	1

* 조사된 빈도가 1 또는 2인 경우는 음영으로 표시됨.

조사된 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	29	17	3	29	98
	여	32	30	49	29	5

현행 제공 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	N/A	N/A	N/A	N/A	13
	아파트	18	N/A	N/A	N/A	36
	연립주택	N/A	N/A	N/A	7	22
	다세대	5	8	N/A	18	12
	비주거용	N/A	N/A	N/A	N/A	15
여	단독주택	N/A	5	8	N/A	N/A
	아파트	9	N/A	9	N/A	N/A
	연립주택	5	10	14	5	N/A
	다세대	10	6	7	8	N/A
	비주거용	N/A	5	11	11	N/A

* 제공 빈도가 5 미만인 경우는 음영으로 표시됨.

현행 제공 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	29	17	N/A	29	98
	여	32	30	49	29	5

예시에서 보듯이 현행 제공 방식의 기저 빈도표에서는 전체의 50%에 해당하는 25개 셀에 대해 N/A로 제공 되고, 상위 빈도표에서는 10%에 해당하는 1개 셀에 N/A로 제공된다. 기저 빈도표는 상위 빈도표에 비해 훨씬 희박한(sparse) 형태이므로 N/A로 제공되는 비율이 높을 수밖에 없다.

조사된 빈도표를 그대로 제공한다면 정보손실이 없는 정확한 자료를 제공받을 수 있다는 측면에서 사용자 입장에서 가장 환영할 일이지만, 비밀보호를 위해 조사된 빈도를 그대로 제공할 수 없는 이유에 대해서는 앞서 많은 논의가 이루어졌다고 생각된다. 따라서 자료 제공자는 빈도표를 제공함에 있어 비밀보호를 위해 몇 가지 조치를 취한 빈도표를 제공하기로 하였다. 현행 제공 방식의 빈도표에서는 비밀보호를 위해 N/A 표시하여 제공하고 있기 때문에 노출 위험이 제어되었다고 할 수 있다. 그러나 5 미만의 셀이 많을수록 현행 제공 방식의 빈도표에서 사용자가 얻을 수 있는 정보는 매우 빈약해질 수밖에 없는 상황이 된다. 다음의 <표 4-9>는 <표 4-1>과 <표 4-2>, <표 4-5>와 <표 4-6>에서 비밀보호 이전에 빈도가 5 미만인 셀의 비율을 조사한 것으로 현행 자료 제공 방식에서 N/A 표시되는 셀의 비율로 이해할 수 있다.

<표 4-9> 현행 자료 제공 방식에서 N/A 표시 되는 셀의 비율*(%)

행정 구역 단위	OA	LA3	LA2	LA1
(지역 수준)	집계구	읍면동	시군구	시도
제공 항목 6개	99.9	99.6	96.1	90.1
제공 항목 5개	99.9	98.2	88.1	75.4
제공 항목 4개	99.5	92.5	68.7	48.3
제공 항목 3개	97.2	74.9	35.9	17.2
제공 항목 2개	85.2	39	7.5	2.4
제공 항목 1개	45.3	6.8	-	-

격자 단위	OA	LA3	LA2	LA1
(지역 수준)	100m	1km	10km	100km
제공 항목 6개	99.9	99.9	98.4	90.1
제공 항목 5개	99.9	99.6	95.1	75.4
제공 항목 4개	99.9	98.3	86.7	48.3
제공 항목 3개	99.2	93.6	69.6	17.2
제공 항목 2개	94.6	79.9	43.2	2.4
제공 항목 1개	71.6	49.9	13	-

* 빈도가 5 미만인 경우 N/A 표시되어 제공되는 셀의 비율로, 특정 조건에 대한 복수 선택은 허용하지 않은 경우에 계산된 수치임

위 표에서 보듯이 OA 수준에서는 제공 항목의 수가 6~3개인 경우 99% 이상의 셀이 N/A 표시되어 제공되고, 심지어 LA1 수준에서조차 제공 항목의 수가 6개인 경우에는 90% 이상의 셀이 N/A 표시됨을 알 수 있다. 이는 사용자가 요청한 특정 조건을 만족하는 빈도를 그대로 제공하는 경우보다 제공하지 못하는 경우가 더 많을 수 있음을 의미한다. 이러한 현상이 역전되기 위해서는 지역 수준이 LA1 또는 LA2로 넓어지고 제공 항목의 수도 3개 이하로 작아져야만 한다는 것을 알 수 있다. 다만, 위 표에서 제시된 수치는 사용자가 특정 조건에 대해 복수 선택을 하지 않았다는 가정을 하는 것으로, 현재 SGIS에서 제공되는 N/A 비율의 체감도와는 차이가 있을 수 있다. 여기서 특정 조건에 대한 복수 선택의 허용이란, 예를 들자면 가구 조건(세대구성)에 대해 1세대 가구와 2세대 가구를 동시에 선택할 수 있음을 의미한다. 따라서 복수 선택을 허용하지 않았다는 것은 특정 조건에 대해 1가지 선택지만 채택함을 가정하였다는 의미가 된다.

이상으로 제1, 2절을 통해 빈도표에 대한 비밀보호 알고리즘 적용으로 인해 작은 빈도를 가지는 셀의 분포에 차이가 존재함을 확인하였다. 빈도의 분포가 달라지는 경우는 빈도가 1 또는 2인 경우에 노출 위험이 존재한다고 판단되어 비밀보호 조치가 취해졌기 때문이다. 실제 관측된 빈도와 알고리즘 적용 후 빈도의 차이에 해당하는 정보손실은 (±) 1 만큼인 경우가 대부분이며 최대 (±) 3 이내로만 발생함을 확인하였다. 제3절에서는 5 미만의 빈도에 대해 N/A 표시하는 현행 지료 제공 방식과 비교하여 보다 많은 정보를 사용자에게 제공할 수 있음을 확인하였다.

이에 비밀보호 알고리즘을 적용한 SGIS 자료에 대한 안내서를 부록에 제시하여 사용자의 이해를 돕고자 한다. 제시된 안내서는 자료 제공자의 필요에 의해 가감될 수 있으며, 이를 위한 기초 자료로 활용되기를 기대한다.

부록에 제시된 안내서의 주요 내용으로는 SGIS에서 제공하는 빈도표의 형태와 빈도표에서 발생하는 노출 위험에 대해 설명하였다. 또 빈도표의 위계 구조와 비밀보호 알고리즘에 대해 설명하고, 예시 자료를 통해 기저 빈도표와 상위 빈도표에서 비밀보호 알고리즘이 어떻게 적용 되는지를 설명하였다. 마지막으로 현행 자료 제공 방식과 비교하여 비밀보호 알고리즘을 적용함으로써 보다 많은 정보를 사용자에게 제공할 수 있다고 하였다.

제 5 장

결론 및 기대효과

본 연구에서는 통계지리정보서비스(SGIS)에서 제공하는 빈도표의 비밀보호 방안에 대해 다루었다. 2018년 인구주택총조사 자료로 생성된 빈도표를 검토하고, 비밀보호 기법인 BSCA 알고리즘이 적용되는 방식과 알고리즘 적용에 따른 빈도의 분포 변화를 상세히 살피고자 하였다.

SGIS에서 제공하는 빈도표는 위계 구조를 가지고 있어 이를 기저 빈도표와 상위 빈도표로 구분하였다. 기저 빈도표는 제공 가능한 모든 항목을 고려했을 때 가장 좁은 지역 수준에서 만들어지는 빈도표이고, 상위 빈도표는 기저 빈도표를 집계하여 지역 수준을 넓혀가거나 제공 항목의 수를 줄여가는 방식으로 만들어진 빈도표에 해당한다. 본 연구에서 활용된 BSCA 알고리즘은 기저 빈도표와 상위 빈도표에 대해 각각 다른 방식으로 비밀보호를 적용하게 되는데, 예시를 활용해 이를 자세히 설명하였다. 이때 작은 빈도를 갖는 셀로 판단하기 위한 기준 값을 B로 표시하였고, 여기서는 $B = 3$으로 가정하되 빈도가 0인 경우에는 그대로 제공하는 것으로 하였다. 따라서 기저 빈도표에서는 B 미만의 빈도, 즉 빈도가 1 또는 2인 경우에 대해 노출의 위험이 있는 것으로 판단하여 랜덤 라운딩 하여 제공하면 셀 값에 대한 노출의 위험을 $1/B$로 제어한다고 보았다. 상위 빈도표에서는 작은 빈도를 합산하여 중앙값 조정이라는 과정을 거침으로써 정보손실과 노출 위험을 제어하였고, 조치 후에도 제어되지 않은 셀에 대하여는 추가적인 조치를 취하였다. 상위 빈도표에서의 중앙값 조정 과정은 박민정 외(2018)에서 제안된 방식을 소폭 수정하여 적용하였음을 밝힌 바 있다.

2018년 인구주택총조사 자료 중 대전 지역에 해당하는 자료를 제공받아 비밀보호 알고리즘을 적용한 빈도표를 생성하는데 소용되는 시간을 행정 구역 단위 자료와 격자 단위 자료를 구분하여 시나리오별로 정리해 두었다. 2018년 인구주택총조사 자료의 제공 항목 중 선택지의 수가 많은 연령과 건축연도를 포함하는 빈도표 생성에 상대적으로 많은 시간이 필요함을 알 수 있다.

다음으로 비밀보호 알고리즘이 적용된 빈도표의 정합성을 행정 구역 단위 자료와 격자 단위 자료로 나누어 검토하였다. 행정 구역 단위 자료와 격자 단위 자료에서

공통적으로 변화된 것은 빈도가 3, 5인 셀의 비율은 증가하고, 빈도가 4, 6, 7인 셀의 비율은 감소하여 두 자료의 변화 양상은 상당히 유사하다고 볼 수 있다. 특히 빈도가 3인 셀의 비율이 큰 폭으로 증가하게 되는데 최대 11.2%p까지 증가하였고, 빈도가 4 이상인 셀의 비율은 대체로 5%p 수준에서만 변화되었다.

두 빈도표의 차이점으로는 제공 항목이 1개인 경우 OA 수준에서 빈도가 3인 셀의 수는 비밀보호 알고리즘 적용 이후 격자 단위 빈도표에서는 약 8%p 이상 증가하여, 행정 구역 단위 빈도표에서 약 3%p 증가한데 비해 그 변화가 훨씬 더 크다는 것을 확인할 수 있다. 다시 말해 최상위 빈도표에 가까운 제공 항목이 1개인 경우 행정 구역 단위 빈도표에서는 특정 빈도를 가지는 셀의 비율에 거의 변화가 없는데 비해, 상대적으로 격자 단위 빈도표에서는 비교적 큰 변화를 보였다.

비밀보호 알고리즘 적용에 의해 정보손실이 필연적으로 발생하게 되는데, 그 범위는 3 이내임을 확인하였다. 기저 빈도표에 가까운 형태일수록 정보손실이 없는 경우가 많고, 최상위 빈도표에 가까운 형태일수록 정보손실이 1만큼인 셀의 비율이 증가하였다. 정보손실이 0이라는 것은 실제 조사된 빈도와 알고리즘 적용 이후의 빈도가 일치한다는 것이고, 정보손실이 1이라는 것은 알고리즘 적용 전·후의 빈도 차이가 ± 1만큼 발생한다는 것이다. 다시 말해 조사된 빈도표와 큰 차이가 없는 빈도표를 제공할 수 있다는 것을 의미하므로, 노출 위험을 제어하면서 사용자에게 거부감 없는 자료를 제공할 수 있다는 의의를 가질 수 있다.

비밀보호 알고리즘에 의한 정보손실의 최대값은 3이며 그 비율은 전체적으로 0.5%를 넘지 않는 것으로 파악된다. 이는 빈도표에서의 노출 위험 제어를 위한 기회비용에 해당하는 것으로, 사용자는 조사된 빈도표와는 다소 차이가 있는 빈도표를 제공받게 될 것이며 그 차이가 최대 3 이내라는 것을 의미한다.

현행 SGIS에서는 빈도가 5 미만인 경우 N/A 표시하여 제공하는 방식을 취하고 있는데, 기저 빈도표와 같은 매우 희박한 형태의 빈도표에서는 빈도가 5 미만인 셀이 많기 때문에 N/A 표시 비율이 높음을 알 수 있다. 이는 사용자가 요청한 특정 조건에서 빈도를 제대로 제공하지 못하고, N/A 표시하여 제공해야 함을 의미하는 것이 된다. 그러나 N/A 처리만으로는 노출 위험을 제어할 수 없으며(Hoodepool et al., 2012), 노출 위험을 주어진 수준 이하로 유지하기 위해서는 정보손실을 추가로 발생시켜야 한다. 이는 현재 SGIS 격자 서비스에서 구간 범위의 빈도를 제공하고 있는 이유이기도 하다. 따라서 작은 빈도에 대해서도 정확한 값을 필요로 하는 사용자에 대한 제공을 제외한다면, 자료 제공자는 빈도표에 대한 비밀보호 알고리즘을 적용함으로써 노출 위험과 정보손실을 제어하면서 보다 많은 정보를 사용자에게 제공할 수 있을 것으로 기대할 수 있다.

참고문헌

박민정, 성주현(2016). 통계지리정보(인총) 집계구 단위 빈도표의 통계적 노출제어., **2016년 연구보고서**, 통계개발원.

박민정, 신우람, 박천영, 서영오(2018). 그리드별 통계서비스의 비밀보호 방안, **2017년 연구보고서**, 통계개발원.

정동욱, 김경미(2008). 매크로데이터의 비밀보호방법에 대한 고찰, **2007년 연구보고서**, 통계개발원.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Nordhort, E. S., Seri, G., & de Wolf P. P. (2010). ESSNet Handbook on Statistical Disclosure Control, ver 1.2.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Nordhort, E. S., Spicer, K., & de Wolf P. P. (2012). *Statistical Disclosure Control,* John Wiley $ Sons Ltd.

Park. M-J. (2018). Bounded small cell adjustments for flexible frequency table generators, *Privacy in Statistical Database 2018*.

Shlomo, N., Antal, L., & Elliot, M. (2015). Measuring disclosure risk and data utility for flexible table generators, *Journal of Official Statistics, 31(2),* 305-324.

부 록 1

대량의 빈도표를 일반인이 이용하기 편리하도록 제공할 때 비밀보호는 중요한 문제이다. 빈도표가 제공하는 정보로부터 이용자가 개인을 식별하거나 통계기관이 의도하지 않은 사항을 특정할 수도 있기 때문이다. 몇 개의 빈도표를 제공할 때는 정보가 함축되어 있으므로 큰 문제가 아니나, 다양한 조건에 대하여 대량의 빈도표를 만들면 마이크로데이터를 복원할 수 있을 정도로 많은 정보를 제공하게 될 수도 있다. 이로 인하여 통계기관이 의도하는 수준에서 노출위험 제어가 이루어지지 않을 수 있다.

한편 몇 개의 빈도표는 만들기 쉬운 통계이지만 다양한 조건에 대한 대량의 빈도표를 제공하는 플랫폼을 구축하는 데는 많은 계산 비용이 발생한다. 여기에 비밀보호 처리를 추가하는 것은 상당히 복잡한 문제이며, 해외 통계기관에서도 꾸준히 연구되어 온 이슈이기도 하다. 본 보고서의 본문은 일반적인 이해를 위한 알고리즘 이해와 활용방안 안내를 목적으로 하지만, 부록에서는 빈도표 비밀보호 알고리즘인 BSCA를 전문적으로 이해할 수 있도록 이론적인 설명을 수록한다.

1. BSCA 알고리즘의 이해

BSCA(Bounded Small Cell Adjustments)는 기존의 작은 셀 조정(SCA) 알고리즘이 의도하는 수준으로 노출 위험을 제어하면서 정보손실 또한 일정 수준 이하로 제어하도록 고안된 알고리즘이다. 이는 SCA 처리된 빈도표를 집계하여 상위 수준 빈도표를 작성할 때 사용하기 위하여 제안되었다.

먼저 작은 셀 조정(SCA)이란 빈도표에서 큰 값들은 그대로 두고, 값이 작은 셀만 랜덤 라운딩하는 방법을 말한다. 참값을 제공하는 큰 셀의 기준을 $B = 3$이라 하면, 랜덤 라운딩은 다음과 같은 방식으로 작은 셀의 값을 조정한다. 이러한 방식은 주어진 빈도표 전체 합의 기댓값을 보존하는 장점이 있다.

원래 빈도	조정 결과	
1	2/3 확률로	0
	1/3 확률로	3
2	1/3 확률로	0
	2/3 확률로	3

작은 셀 조정이 필요한 이유는 다음과 같다. 빈도표에서 주어진 셀의 값이 1이라는 것은 셀을 규정하는 변수의 조합을 가지는 개체가 자료에서 하나 존재한다는 의미를 가진다. 따라서 그 셀의 노출확률을 1 또는 100%라고 말할 수 있다. 셀의 값이 2이면 노출 확률은 0.5 혹은 50%가 된다. 다만 해당 셀에 속하는 한 개체가 다른 개체를 식별할 위험은 100%라고 할 수 있다. 따라서 셀의 값이 1 또는 2라면 이를 공표하는 것은 바람직하지 않으며, 해외 통계기관의 현황을 살펴보아도 그러한 사례를 찾기는 어렵다. 현실적으로 사용 가능한 가장 낮은 공표 기준값은 3이라고 할 수 있으며, 이 기준을 높일수록 노출위험이 낮아지고 정보손실이 커진다. SCA 이외에도 대부분의 빈도표를 위한 매스킹 기법들은 이러한 기준값을 사용하고 있다. 참고로 통계청은 인구·가구 자료에 대해서는 빈도표 제공 시 매스킹 기준값으로 5를, 사업체 자료에 대해서는 3을 주로 사용해왔다.

일반적으로 통계기관은 어떤 개별정보가 노출될 확률을 $1/B$ 이하로 관리하기 위해 공표하는 셀의 기준을 B라고 정한다. 그러나 큰 셀과 작은 셀에서 노출확률의 개념이 서로 다름을 주의할 필요가 있다. 큰 셀의 경우에는 통계기관의 의도대로 '개체의 노출확률'을 $1/B$ 이하로 제어한다. 반면, 작은 셀들에서는 개체의 노출확률이 아닌 '셀 값이 노출될 확률'을 $1/B$ 이하가 되도록 제어한다. 즉 작은 셀의 경우 셀에 속하는 개체가 노출될 확률이 아니라, 셀 값 자체의 노출확률을 일컫는 것임을 주의할 필요가 있다.

예를 들어 기준값을 $B=3$이라 하자. 빈도표에서 값이 1 또는 2인 셀이 랜덤 라운딩을 통해 0으로 변환되었다면, 셀의 참값은 0, 1, 2 중 하나이므로 정확한 값이 노출될 확률은 1/3이다. 동일한 셀이 랜덤 라운딩을 통해 3으로 변환되었다면 셀의 참값은 1, 2, 3 중 하나이며 정확한 값이 노출될 확률이 1/3이다. 따라서 랜덤 라운딩을 통해 정확한 셀 값이 노출될 확률을 $1/B(=1/3)$ 이하로 제어할 수 있다. 반면 값이 4인 셀은 셀 값은 정확하게 알려지지만, 셀에 속한 개체가 노출될 확률이 1/4이므로 의도한 노출위험인 $1/B(=1/3)$보다 낮게 제어된다.

한편 노출위험 기준 중 하나인 k-익명성으로 기준값 B의 의미를 설명할 수도 있다. 빈도표에서 B 미만인 셀의 값을 불확실하게 제공하여 익명성 B를 만족시킬 수 있기 때문이다. 다만 작은 셀에서는 셀 값 자체가 익명성 B를 만족하는 것이며, 큰 셀에서는 셀에 속한 개체가 익명성 B를 만족하는 것임을 유의해야 한다. 다만 크기가 B인 셀은 양쪽 모두에 속한다. 부록에서는 혼란을 방지하기 위하여 작은 셀은 셀의 크기가 B 이하인 셀들을, 큰 셀은 셀의 크기가 B를 초과하는 셀들을 지칭하도록 한다.

하나의 마이크로데이터로부터 위계구조를 가지는 여러 빈도표를 생성하는 문제, 노출위험과 정보손실을 제어하는 비밀보호 알고리즘인 BSCA의 원리, 알고리즘 및

코드, JAVA 시스템에서 구현하는 방법을 차례대로 살펴보겠다.

1.1 빈도표 구조와 SCA 적용

마이크로데이터에 포함된 여러 변수들 중에서 빈도표를 생성할 때 사용되는 변수들을 키(key)변수 라고 하자. 개인이나 가구 단위로 구성된 DB의 경우 지역, 성별, 연령대 등의 변수를, 사업체 단위로 구성된 DB에서는 업종, 범주화된 기업규모 등의 변수를 이용해 빈도표를 생성하는 경우가 많다. 한편 키변수 중에서 위계구조(hierarchical structure)가 존재하는 경우가 있다. 예를 들어 지역의 경우, 시도, 시군구, 읍면동, 집계구라는 변수들을 고려할 수 있으며, 이 변수들 사이에서 포함관계, 위계구조가 성립한다. 읍면동마다 빈도표를 생성하여 합하면 해당 시군구에 대한 빈도표를 만들 수 있기 때문이다. 마찬가지로 업종 대분류, 중분류, 소분류 각 변수의 경우에도 위계구조가 성립한다. 이제 빈도표를 생성할 때 사용하는 변수들 중에서 위계구조를 가지는 변수를 H키변수(hierarchical key)라고 부르고, 위계구조가 없는 변수들은 키변수라 하자. 본 연구에서 고려한 H키변수로는 행정구역 변수, 그리드 지역변수, 업종분류 등이 있다.

다음과 같은 인구주택총조사 마이크로데이터에서 빈도표를 생성한다고 가정하자. 이 마이크로데이터는 총 4개의 H키변수와 4개의 키변수로 구성되어 있다. 지역수준을 나타내는 변수들의 값은 각 지역 코드로 구성되어 있으며, 4개 H키변수를 모두 붙이면 각 지역을 중복없이 나타낼 수 있다. 이 예에서 H키변수는 시도(LA1), 시군구(LA2), 읍면동(LA3) 및 집계구(OA)가 있으며, 키변수는 성별(G), 연령(A), 혼인상태(M), 교육수준(E)이 있다.

<표 1-1> 마이크로데이터 예제

id	H키변수				키변수			
	LA1 시도	LA2 시군구	LA3 읍면동	OA 집계구	G 성별	A 연령	M 혼인	E 교육
1	25	10	68	10001	1	1	1	1
…	25	10	12	40003	2	17	5	1
…	25	20	11	20402	1	15	1	3
…	25	50	01	110001			…	
N	25	30	09	10001	2	18	5	8

이 마이크로데이터로부터 각 지역별로 키변수 조합에 대한 빈도표를 생성할 수 있다. 예를 들어 G, A, M, E 모든 조합에 대한 빈도표를 집계구마다 생성할 수 있으며, 시도에 대한 빈도표를 생성할 수도 있다. 물론 각 집계구에 대한 빈도표를 다 더하면 시도에 대한 빈도표와 일치하게 된다. 한편 각 지역에 대하여 키변수가 3개, 2개, 1개인 빈도표를 만들 수도 있다. 아래 <표 1-2>는 이 마이크로데이터로부터 생성할 수 있는 빈도표의 예를 보여준다. 2010년 인구주택총조사 대전지역의 경우 각 1개 시도, 5개 시군구, 78개 읍면동, 2506개 집계구에 대하여 4개 변수 1종, 3개 변수 4종, 2개 변수 6종, 1개 변수 1종의 빈도표를 생성할 수 있다. 빈도표의 총 개수는 38,850개이며, 빈도(빈도표에서 각 셀의 값)의 개수는 총 7,969,430으로 약 800만 개가 된다. 만약 변수 범주별로 중복 선택을 허용하면 각 지역별로 다루어야 할 빈도의 개수의 총합은 $2^{(2+18+5+8)}-1$로 85억 개가 넘는다.

<표 1-2> 빈도표 예제 (키변수가 각각 4개, 3개, 2개, 1개인 경우)

	G	A	M	E	빈도
1	1	1	1	1	$f_{1,1,1}=f_1$
...		...			...
K (1440)	2	18	5	8	$f_{2,18,5}=f_K$

	G	A	E	빈도
1	1	1	1	$f_{1,1,1}=f_1$
...		...		...
K (288)	2	18	8	$f_{2,18,8}=f_K$

	G	M	빈도
1	1	1	$f_{1,1,1}=f_1$
...	...		...
K (10)	2	5	$f_{2,5,8}=f_K$

	A	빈도
1	1	$f_{1,1,1}=f_1$
...	...	...
K (18)	18	$f_{18,5,8}=f_K$

하나의 마이크로데이터에서 생성되는 빈도표를 유형화하면 다음과 같다. 다양한 빈도표들을 H키변수 개수 및 키변수 개수에 따라 유형화할 수 있다. 먼저 가장 상세한 수준의 H키변수에 대하여 모든 키변수로 구성된 빈도표를 기저 빈도표(full table)라 부르겠다. 다음 <표 1-3>에서 기저 빈도표는 T_{H4}^{v4}이다. T_{H4}^{v4}를 제외한 다른 빈도표는 T_{H4}^{v4}를 재집계하여 산출할 수 있으며, 상위 빈도표라고 부르겠다. 변수의 개수가 적어지면서 상위 빈도표를 얻을 수도 있고, 상위 수준 H키변수에 대한 빈도표가 상위 빈도표가 될 수도 있다.

<표 1-3> 빈도표 유형

T	H4	H3	H2	H1
키변수 4개	T_{H4}^{v4}	T_{H3}^{v4}	T_{H2}^{v4}	T_{H1}^{v4}
키변수 3개	T_{H4}^{v3}	T_{H3}^{v3}	T_{H2}^{v3}	T_{H1}^{v3}
키변수 2개	T_{H4}^{v2}	T_{H3}^{v2}	T_{H2}^{v2}	T_{H1}^{v2}
키변수 1개	T_{H4}^{v1}	T_{H3}^{v1}	T_{H2}^{v1}	T_{H1}^{v1}

알고리즘을 설명하기 위해 첨자를 다음과 같이 정한다. 상위 빈도표의 i번째 셀의 값을 f_i라고 하고, 이 셀을 구성하는 하위 빈도표 셀의 집합을 $\{f_{ij}, j = 1, ..., n_i\}$라고 하자. 이때 n_i는 상위 빈도표의 i번째 셀을 구성하는 기저 빈도표 셀의 개수가 된다. <표 1-1> 마이크로데이터 예제의 경우, T_{H2}^{v4}는 시군구(H2)에 대하여 4개 변수로 구성된 빈도표를 나타낸다. 어느 시군구의 빈도표에서 i번째 셀의 값은 그 시군구에 속하는 n_i개 집계구 빈도표의 동일한 셀을 모두 합하여 얻어진다.

빈도표 구조 예제

기저 빈도표 (i번째 시군구에 속한 집계구)

집계구 1

G	A	M	E	빈도
1	1	1	1	2
1	1	1	2	0
1	1	1	3	1
1	1	1	4	3
	...			...
1	1	2	1	0
	...			...
2	18	5	8	3

...

집계구 n_i

G	A	M	E	빈도
1	1	1	1	0
1	1	1	2	0
1	1	1	3	3
1	1	1	4	0
	...			...
1	1	2	1	1
	...			...
2	18		8	2

상위 빈도표 (i번째 시군구의 빈도표들)

시군구 i

G	A	M	E	빈도
1	1	1	1	9
1	1	1	2	0
1	1	1	3	7
1	1	1	4	52
	...			...
1	1	2	1	2
	...			...
2	18		5	11

...

시군구 i

G	A	M	빈도
1	1	1	230
1	1	2	0
1	1	3	55
1	1	4	78
	...		...
1	2	1	2
	...		...
2	18	5	25

...

이제 각 상위 빈도표 i번째 셀에 대한 비밀보호 과정을 살펴보자. 우선 기저 빈도표의 비밀보호를 위해서는 작은 셀에 랜덤 라운딩을 적용하고, 큰 셀은 그대로 제공한다. n_i개의 셀 값 중에서 작은($f_{ij} \le B$) 셀의 개수를 K라고 하자(K_i라고 하는 것이 정확하나, 편의상 첨자를 생략하였다. 이후 사용되는 기호 k, v, w, r, w, a, b 모두 편의를 위하여 첨자 i를 생략하도록 한다). 한편 설명을 단순하게 만들기 위하여 큰 셀들이 없다고 가정하자($K = K_i = n_i$). 실제로 f_i를 구할 때, 큰 셀들은 단순 집계하여 더하면 되므로 이러한 가정은 무리가 없다. 이제 f_i를 구성하는 기저 빈도표 셀들의 집합은

$$\left\{f_{ij} \ : j = 1, 2, \ ..., K\right\}, \quad f_{ij} \in \{0, 1, 2, \ ..., B\}$$

이며, $f_i = \sum_j f_{ij}$가 성립한다. 한편 기저 빈도표 셀들에 SCA를 적용한 결과를 다음과 같이 표기하자.

$$\left\{\tilde{f}_{ij} \ : j = 1, 2, \ ..., K\right\}, \quad \tilde{f}_{ij} \in \{0, B\}$$

기저 빈도표의 비밀보호는 SCA를 적용한다면 기저 빈도표를 집계하여 작성하는 상위 빈도표의 비밀보호는 BSCA를 적용한다. 즉, BSCA는 $\{f_{ij}\}$, $\left\{\tilde{f}_{ij}\right\}$ 및 f_i를 알고 있는 상태에서 $\tilde{f}_i$의 값을 결정하는 알고리즘이며, 모든 빈도에 대하여 노출위험을 $1/B$ 이하로 유지하고 정보손실을 최소화하는 목적을 가진다. 이를 도식화하면 다음과 같다.

기저 빈도표	$f_{ij} \in \{0, 1, \ ..., B\} \quad \xrightarrow{RR} \quad \tilde{f}_{ij} = 0 \text{ or } B$
상위 빈도표	$f_i = \sum_j^K f_{ij} \quad \xrightarrow{BSCA} \quad \tilde{f}_i = ?$

다만 BSCA는 $K > 1$인 경우에 적용되며, $K = 1$인 경우에는 SCA 결과를 그대로 사용한다. 이후 BSCA 알고리즘 적용에 관한 모든 논의는 $K > 1$을 가정한다.

시스템이 상위 빈도표를 제공할 때 노출위험을 증가시키지 않도록 알고리즘을 설계하기 위하여 이용자가 공표되는 셀 값들로부터 얻게 되는 정보를 반영할 필요가 있다. 이용자가 얻는 정보를 정리하면 다음과 같다. 기저 빈도표에서 공표되는 셀의 값, $\left\{\tilde{f}_{ij}\right\}$는 0 또는 B로 구성되어 있다. 공표되는 기저 빈도표에서 값이 B인 셀의 개수를 k개라고 하면, 이 k개 셀들은 실제로는 1에서 B 사이의 값을 가지므로 k개 셀들의 합은 k에서 kB 사이의 값을 가진다. 마찬가지로 기저 빈도표에서 값이 0인 셀의 개수는 $K-k$개이며, 이 $K-k$개 셀들은 실제로는 0에서 $(B-1)$ 사이

의 값을 가지므로 $K-k$개 셀들의 합은 0에서 $(K-k)(B-1)$ 사이의 값을 가진다. 따라서 기저 빈도표에서 공표되는 셀 값들으로부터 상위 빈도표 셀의 참값 f_i가 가질 수 있는 최소값 및 최대값은 다음과 같이 정리된다. 최대값 및 최소값으로 정의되는 집합 $D=[v,w]$는 이용자가 시스템을 통해 습득하는 정보가 된다.

$$f_i \in D = [0 \cdot (K-k) + 1 \cdot k, \quad (B-1) \cdot (K-k) + B \cdot k]$$

$$= [k, \qquad\qquad k + K(B-1)] = [v, w]$$

이상의 내용을 정리하면 아래와 같다.

요약

기저 빈도표

참값

셀 번호	빈도
1	f_{i1}
2	f_{i2}
...	...
j	f_{ij}
...	...
K_i	f_{iK_i}

$f_{ij} \in \{0, 1, ..., B\}$

공표값

셀 번호	빈도
1	$\tilde{f}_{i1}$
2	$\tilde{f}_{i2}$
...	...
j	$\tilde{f}_{ij}$
...	...
K_i	$\tilde{f}_{iK_i}$

$\tilde{f}_{ij} = 0 \text{ or } B$

상위 빈도표

참값

셀 번호	빈도
i	f_i

$f_i = \sum_{j=1}^{K} f_{ij}$

공표값

셀 번호	빈도
i	$\tilde{f}_i$

$\tilde{f}_i = ?$

이용자의 관찰

$$\left\{\tilde{f}_{ij} \ : j = 1, 2, ..., K\right\} = \underbrace{0, 0, ..., 0}_{(K-k)}, \underbrace{B, ..., B}_{k}$$

: $(K-k)$개의 0 및 k개의 B

이용자의 추론

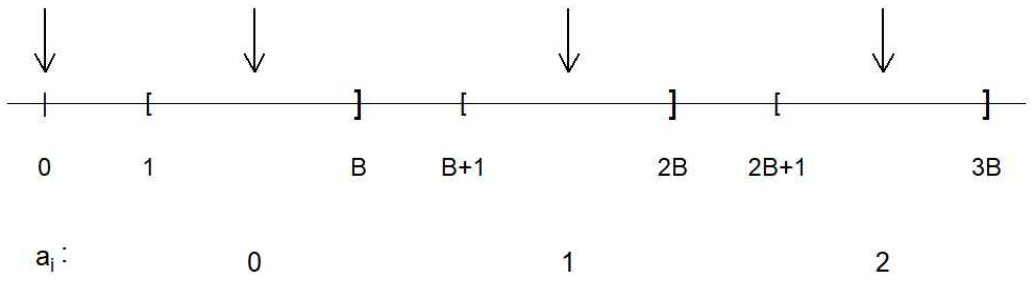
$$f_i \in D = [0 \cdot (K-k) + 1 \cdot k, \quad (B-1) \cdot (K-k) + B \cdot k]$$

$$= [k, \qquad\qquad k + K(B-1)] = [v, w]$$

1.2 BSCA 알고리즘의 원리

박민정(2018) 및 Park(2018)이 제안하였듯이, BSCA의 핵심 아이디어는 노출 위험이 아니라 정보손실을 먼저 줄이자는 것이다. 대다수의 비밀보호 방법론이 노출 위험을 우선 제어한 이후에 정보손실을 줄이기 위한 단계를 제시하는 것과 반대로, BSCA에서는 정보손실을 우선 축소시킨 이후에 노출 위험을 제어하는 단계를 구상하였다. 다만 노출 위험을 제어하면서 추가적으로 약간의 정보손실이 발생하는 것을 허용하였다. 참고로 본 연구에서는 기존 BSCA의 주요 아이디어를 그대로 사용하지만, 약간의 오류를 바로 잡고 실제 시스템에 적용하기 용이하도록 기존 알고리즘의 일부 세부 사항을 수정하였다. 기존의 R패키지는 본 연구의 수정된 알고리즘을 반영하여 향후 공표하도록 한다.

먼저 상위 빈도표 셀의 값 $\tilde{f}_i$을 결정할 때, 먼저 적은 정보손실을 가지도록 다음과 같은 단계를 도입한다. 앞의 설명과 마찬가지로 모든 셀은 작은 셀($\leq B$)이라고 가정한다. 이제 주어진 작은 셀 기준 B를 이용하여, 길이 B로 이루어진 구간들을 구성한다. 그리고 상위 빈도표 셀의 참값 f_i가 속한 구간의 중앙값을 공표 셀 값인 $\tilde{f}_i$에 할당한다. <그림 1-1>은 각 구간에서 중앙값을 할당하는 것을 보여준다. 다만 값이 0인 경우에는 값을 그대로 보존한다. 이처럼 BSCA는 고정된 구간(Fixed Segmentation)의 중앙값을 사용하는 간단한 아이디어로 정보손실을 축소한다. 이를 BSCA의 1단계라고 하자.



<그림 1-1> 고정된 구간 설정 및 중앙값 할당

이를 수식으로 표현하기 위해서는 모드(mod) 함수를 사용할 수 있다. 참값과 공표값은 다음과 같이 표현된다. 한편 <표 1-4>는 처음 몇 구간에 대하여 BSCA 1단계에서 상위 빈도표의 셀 값이 어떻게 할당되는지를 구체적으로 보여준다.

$$f_i = aB + b + 1, \qquad \tilde{f}_i = aB + \left[\frac{B}{2}\right] + 1$$

단, $a \in Z,\ b \in \{0, 1, 2, \ldots, (B-1)\}$.

<표 1-4> 1단계 구간 중앙값 제공을 통한 정보손실 제어

f_i	a	b	$\tilde{f}_i$	$\tilde{f}_i$ $(B=3)$
0	-1	$B-1$	$-B + [B/2] + 1$	0
1		0		
⋮	0	⋮	$[B/2] + 1$	2
B		$B-1$		
$B+1$		0		
⋮	1	⋮	$B + [B/2] + 1$	5
$2B$		$B-1$		
$2B+1$		0		
⋮	2	⋮	$2B + [B/2] + 1$	8
$3B$		$B-1$		
⋮	⋮	⋮	⋮	⋮

이처럼 모드 함수를 사용하여 빈도를 표현하면 다음과 같은 두 성질을 얻게 된다. 작은 셀의 전체 개수 K 및 B로 변환된 작은 셀의 개수 k에 관하여 다음 성질이 각각 성립한다.

작은 셀의 개수 K는 주어진 참값 f_i에 대하여 생성되는 구간의 개수보다 많으므로

$$\text{(P1)}\ \ a + 1 \le K$$

이며, SCA에 의하여 B로 변환된 작은 셀의 개수 k는 참값 f_i가 속한 구간의 최대값보다 작으므로 다음이 성립한다.

$$\text{(P2)}\ \ 0 \le k \le (a+1)B$$

이 두 성질은 추후 BSCA와 관련한 수학적 증명에 필요하여 여기에 기술한다.

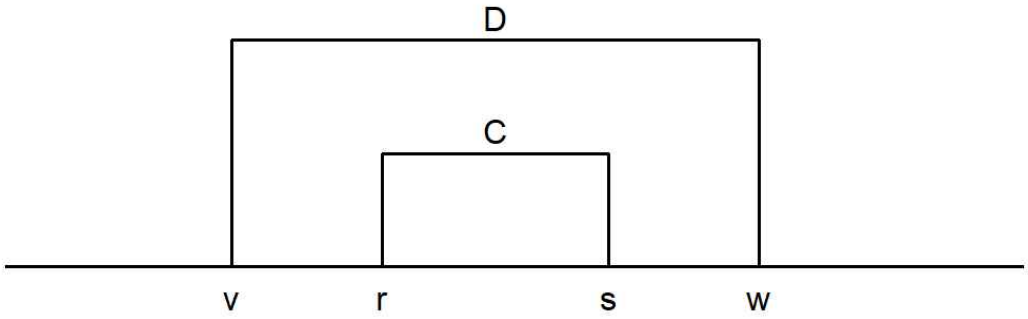
이제 노출위험을 축소하는 단계를 만들기 위하여 이용자가 제공받는 정보를 검토하자. 앞 소절에서 설명하였듯이 이용자는 기저 빈도표의 정보를 통해 $f_i \in D = [v, w]$를 알 수 있다. 또한 정보손실을 축소하는 1단계에서 할당된 $\tilde{f}_i$를 공표값으로 사용하면 다음과 같이 참값이 주어진 구간을 나타내는 집합 $C = [r, s]$

에 속하는 것 역시도 알게 된다. 이때 정보 C는 시스템이 직접적으로 제공하는 정보이다.

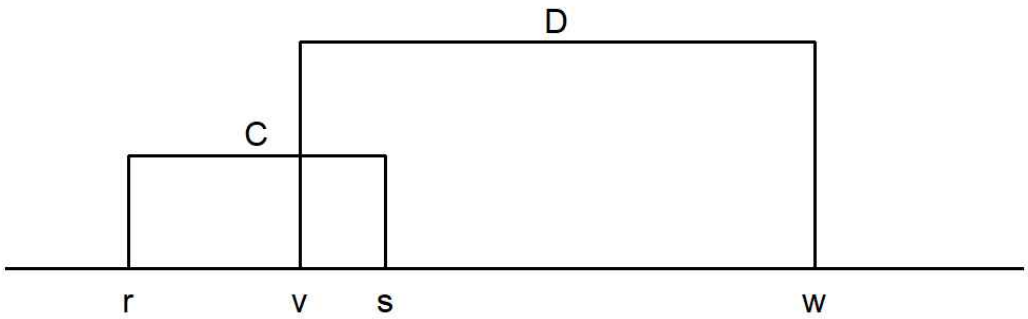
$$f_i \in C = [aB+1, (a+1)B] = [r, s]$$

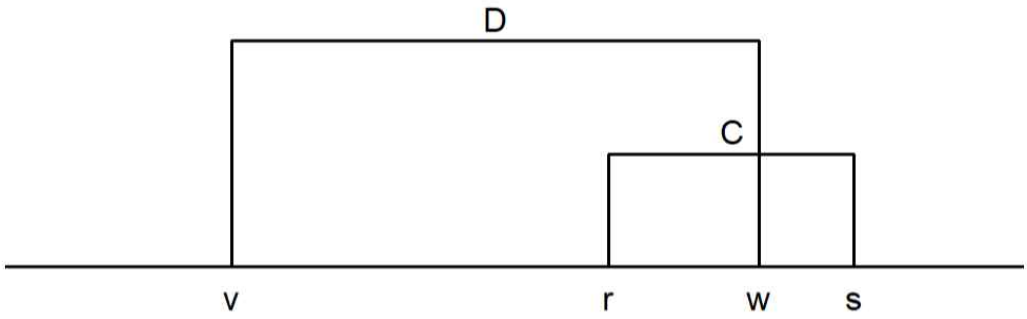
만약 $C \subset D$가 성립하면, 이용자는 시스템이 제공하는 정보만을 얻게 되므로 참값의 후보는 B개 존재하며 노출위험은 $1/B$로 유지된다. 그러나 $C \not\subset D$인 경우, 참값의 후보는 $f_i \in C \cap D = [v, s]$ or $[r, w]$로 B개 보다 적어져 노출위험이 높아진다. <그림 1-2>는 노출위험에 변화가 없는(노출이 일어나지 않는) 경우를, <그림 1-3> 및 <그림 1-4>는 $C \not\subset D$가 되어 노출위험이 증가하는(노출이 일어나는) 경우를 모두 나타낸다.

시스템이 참값 f_i 대신 $\tilde{f}_i$를 결정하는 핵심 아이디어는 이용자가 추가적으로 습득하는 정보를 주지 않아 노출위험은 제어하면서, 최대한 정확한 값을 제공하는 것이다.

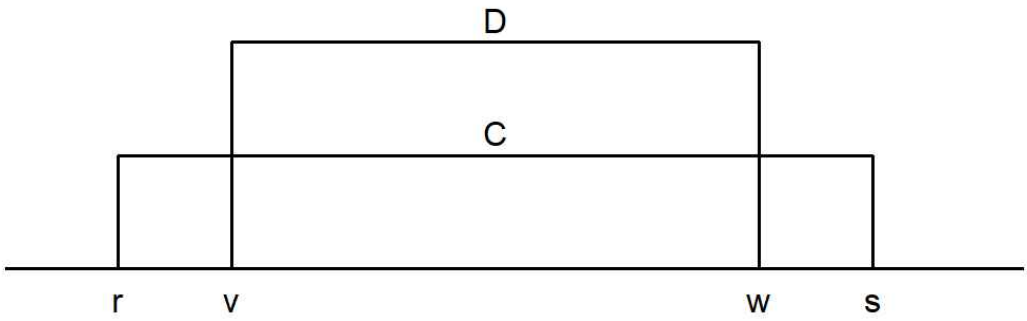


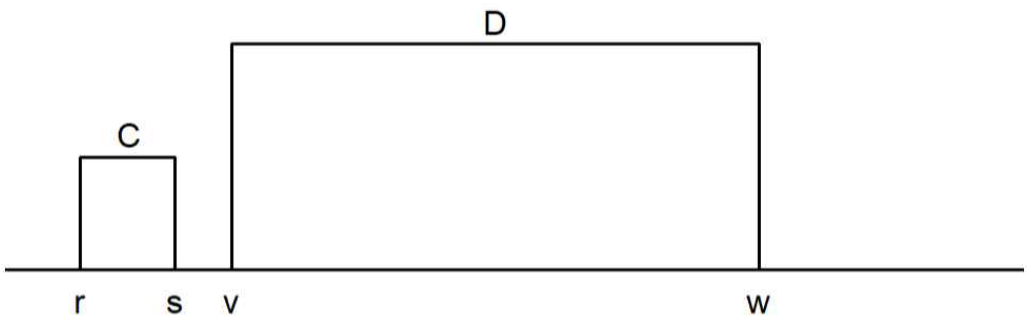
<그림 1-2> 노출위험에 변화가 없는(노출이 발생하지 않는) 경우

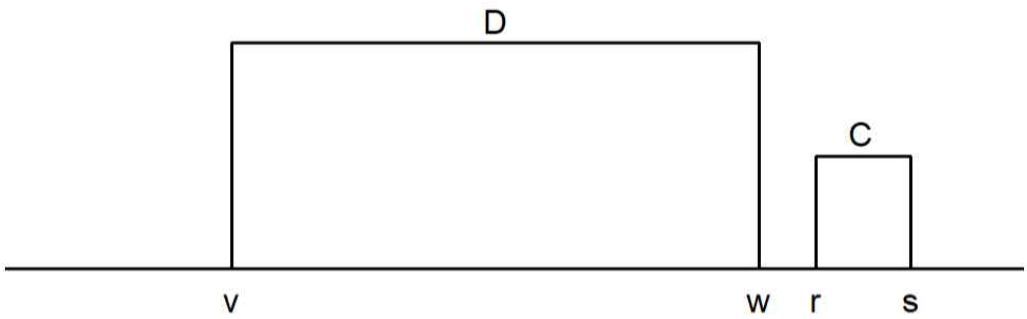




<그림 1-3> 노출위험이 증가하는(노출이 발생하는) 경우







<그림 1-4> 노출위험이 증가하는(노출이 발생하는) 경우 (일어나지 않음)

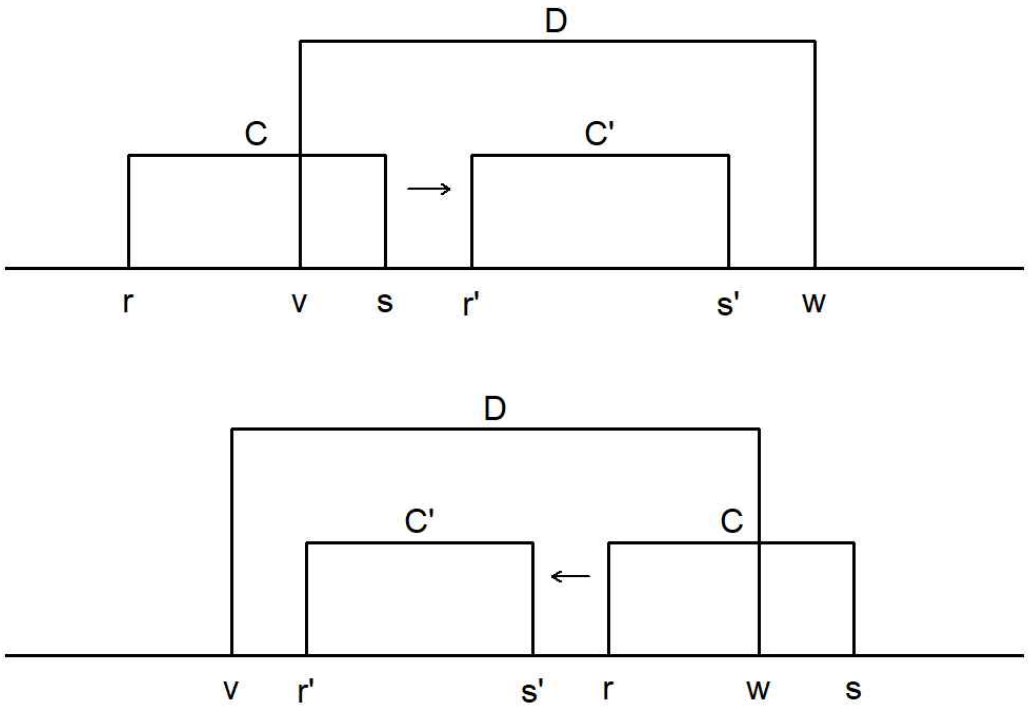
한편 <그림 1-4>는 다음의 노출위험이 증가하는(노출이 일어나는) 경우를 나타낸다.

$$\text{i) } r < v \text{ 이고 } w < s,\ \text{ii) } s < v,\ \text{iii) } w < r$$

다만 이는 실제로는 일어나지 않으며 그 이유는 다음과 같다. 첫 번째 경우는 D의 길이가 B보다 크고 C의 길이는 B이므로 발생할 수 없으며, 두 번째 경우는 (P2)에 의하여 $v \le s$ 이므로 일어날 수 없다. 마지막 경우 역시 일어날 수 없으며 그 이유는 다음과 같다.

$$\begin{aligned} w = k + K(B-1) &\ge k + (a+1)(B-1) \quad \because K \ge a+1 \text{ (P1)} \\ &= aB + 1 + B - 2 + k \\ &> r \qquad\qquad\qquad\qquad \because B \ge 3,\ k \ge 0 \end{aligned}$$

이제 BSCA 2단계에서는 <그림 1-3>과 같이 $C \not\subset D$인 경우의 노출위험을 제어하는 단계를 도입한다. 이 단계의 주요 아이디어는 단순하다. $C \not\subset D$인 경우 이웃 구간의 중앙값으로 $\tilde{f}_i$의 값을 조정하는 것이다. 다만 조정하는 값에 대해서는 정보손실이 B만큼까지 추가로 증가할 수 있다. <그림 1-5>는 노출위험 증가를 방지하기 위하여 공표값을 조정하는 것을 나타내고 있다. C'가 여전히 D에 속하는 것은 이 소절의 끝부분에 증명되어 있다.



<그림 1-5> 노출위험이 증가하는(노출이 발생하는) 공표값의 조정

이상을 수식으로 나타내면 다음과 같다. $C \subset D$로 노출이 발생하지 않는 경우, 다음과 같이 셀의 값을 유지한다. 정보손실은 최대 $[B/2]$로 유지된다.

$$\tilde{f}_i = aB + \left[\frac{B}{2}\right] + 1$$

$C \not\subset D$로 노출이 발생하는 경우는 <그림 1-3>처럼 $r < v \leq s$ 및 $r \leq w < s$인 경우이다. 먼저 $r < v \leq s$인 경우에는 B를 더하여 이웃 구간으로 이동시킨다.

$$\tilde{f}_i = aB + \left[\frac{B}{2}\right] + 1 + B$$

한편 $r \leq w < s$인 경우에는 값을 B만큼 감소시켜 이웃 구간으로 이동한다.

$$\tilde{f}_i = aB + \left[\frac{B}{2}\right] + 1 - B$$

BSCA의 마지막 세 번째 단계는 첫 번째 구간의 중앙값으로 결정이 된 공표값을 조정하는 것이다. 이용자는 기저 빈도표에서 0 또는 B의 값을 제공받으므로, 상위 빈도표에서 공표값이 첫 번째 구간 $[1, B]$의 중앙값으로 결정되는 대신에 기저 빈도표와 동일한 값을 가지는 것이 바람직하기 때문이다. 즉 모든 빈도표가 제공하는 값들이 통일성을 가져서 이용자에게 혼란을 주지 않도록 제공값이 작은 셀 구간의 중앙값으로 결정되는 ($\tilde{f}_i = [B/2] + 1$) 경우 $\tilde{f}_i = B$가 되도록 $\tilde{f}_i$의 값을 조정한다.

이제 BSCA 알고리즘 세 단계의 주요 목표를 다음과 같이 정리할 수 있다.

단계	내용
1	정보손실 축소 (Fixed Segmentation)
2	노출 위험 축소 (최대 [B/2]만큼 정보손실 범위 증가)
3	이용자 친화성 증가

마지막으로 $C \not\subset D$일 때 구간을 이동하면 $C \subset D$가 성립함을 나타내기 위한 간단한 증명을 제시한다. 증명해야 할 사항은 다음과 같다.

$r < v \leq s$인 경우	A1) $r' = r + B \geq v$ A2) $s' = s + B \leq w$
$r \leq w < s$인 경우	B1) $s' = s - B \leq w$ B2) $r' = r - B \geq v$

증명을 위해 필요한 사항은 다음 구간의 정의 및 앞에 설명한 두 가지 성질이다.

$$C = [r,\, s] = [aB+1,\, (a+1)B], \quad D = [v,\, w] = [k,\, k+K(B-1)]$$

$$\text{(P1)}\ a+1 \le K, \quad \text{(P2)}\ 0 \le k \le (a+1)B$$

증명은 다음과 같다. $K(\ge 2)$ 및 $B(\ge 3)$ 임을 주의하라.

A1) 의 증명

$r + B = aB + 1 + B = (a+1)B + 1 > k = v$ $\because$ (P2)

A2) 의 증명

i) $a_i = 0$ 일 때,

$$\begin{aligned} w - s &= k + K(B-1) - B \\ &\ge k + B - 2 && \because K \ge 2 \\ &> B - 1 && \because r < v \Leftrightarrow 1 < k \end{aligned}$$

$\therefore\ w - s \ge B$

ii) $a_i \ge 1$ 일 때,

$$\begin{aligned} w - s &= k + K(B-1) - (a+1)B \\ &\ge k + (a+1)(B-1) - (a+1)B && \because K \ge a+1\ (\text{P1}) \\ &= k - (a+1) \\ &> aB + 1 - (a+1) = a(B-1) && \because r < v \Leftrightarrow aB+1 < k \\ &\ge B - 1 && \because a \ge 1 \end{aligned}$$

$\therefore\ w - s \ge B$

B1)의 증명

$r \le w < s$ 이며 $s - r = B - 1$ 이므로 성립

B2)의 증명

$$\begin{aligned} w < s &\Leftrightarrow k + K(B-1) < (a+1)B \\ &\Leftrightarrow aB + 1 - k > 1 - B + K(B-1) = (K-1)(B-1) \end{aligned}$$

$$\begin{aligned} r - v = aB + 1 - k &> (K-1)(B-1) && \because w < s \\ &\ge (B-1) && \because K \ge 2 \end{aligned}$$

$\therefore\ r - v \ge B$

1.3 BSCA 알고리즘 및 코드 설명

앞 소절에서 BSCA의 원리를 설명하였다. 이를 바탕으로 BSCA 알고리즘의 순서도를 살펴보자. 앞에서와 마찬가지로 설명을 단순하게 만들기 위하여 기저 빈도표에는 큰 셀들이 없다고 가정하며, 큰 셀들이 있다면 단순 집계하여 마지막 단계에서 더하면 된다. 또한 단순한 표현을 위하여 수식의 k, K, r, s, v, w 등에서 첨자 i를 생략하였다.

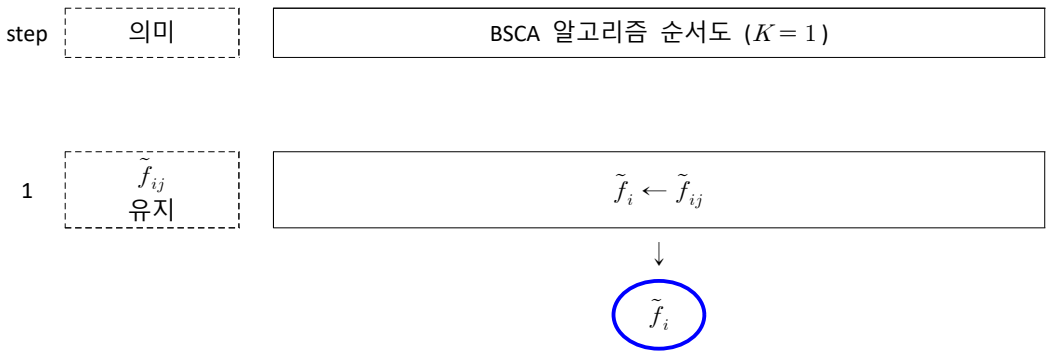
먼저 기저 빈도표에서 SCA를 적용하여 공표할 셀 값은 다음과 같다.

$$\left\{\tilde{f}_{ij} \ : j = 1, 2, \ ..., K\right\}, \quad \tilde{f}_{ij} \in \{0, B\}$$

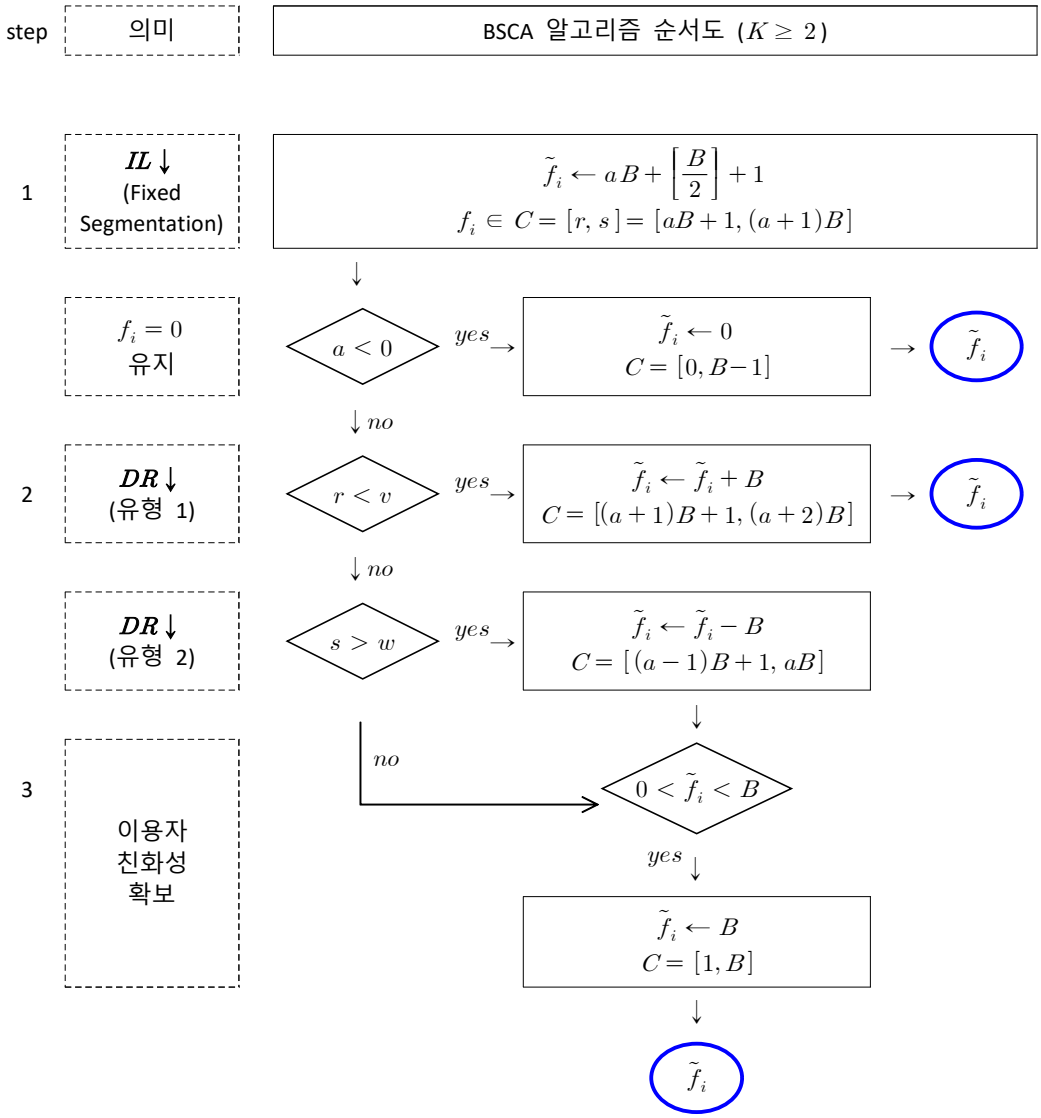
이러한 기저 빈도표의 공표 셀 값을 기준으로 다음과 같은 조건이 발생한다.

$$f_i \in D = [v, \ w] = [k, \ k + K(B-1)]$$

이제 상위 빈도표의 공표 셀 값 $\tilde{f}_i$은 다음 <그림 1-7>과 같은 순서도를 통해 산출된다. 이 순서도는 참값으로 이루어진 기저 빈도표와 랜덤 라운딩 처리된 기저 빈도표를 입력받아 상위 빈도표의 셀값을 계산하는 과정을 보여주는 것이며, IL은 정보손실(information loss)을, DR은 노출 위험(disclosure risk)을 나타낸다. 만약 작은 셀이 하나만 있다면($K = 1$) <그림 1-7>의 과정을 거치지 않으며, 대신에 <그림 1-6>과 같이 SCA를 적용한 기저 빈도표 셀의 값을 그대로 가져온다.



<그림 1-6> BSCA 알고리즘 순서도 ($K = 1$)



<그림 1-7> BSCA 알고리즘 순서도 ($K \geq 2$)

한편 BSCA를 실무적으로 활용할 수 있도록 R패키지를 개발하였다. 현재 R패키지를 이용하여 빈도표를 생산하는 과정은 다음과 같은 간단한 함수들로 구현되어 있다.

```
# 기저 빈도표 저장
saveFullTables(hkey, key, wght, B)

# 주어진 키변수에 대한 H키변수 수준별 빈도표 저장
saveAllHLevels(key.nm)

# 주어진 키변수 및 H키변수에 대한 빈도표 저장
saveSelected(key.nm, lv = 2, hkeySelected.nm = "서구")
```

BSCA 알고리즘 및 R패키지의 이해를 돕기 위하여 위 세 개 함수와 관련된 코드를 설명하면 다음과 같다. 프로그램을 활용하기 위하여 먼저 입력하는 데이터의 형식을 살펴보자. 보통의 마이크로데이터는 아래와 같은 형식으로 구성되어 있다.

	시도	시군구	읍면동	집계구	성별	연령	혼인상태	교육수준
1	25	10	53	10001	1	13	2	4
2	25	10	53	10001	2	13	2	2
3	25	10	53	10001	2	13	3	3
4	25	10	53	10001	2	8	1	6
...								

이제 마이크로데이터로부터 키변수 및 H키변수를 분리하여 저장한다. 예제에서는 hkey 및 key 라는 변수명을 사용하였으며, 마이크로데이터는 aa라는 변수명을 사용하였다.

hkey

	시도	시군구	읍면동	집계구
1	25	2510	251053	25105310001
2	25	2510	251053	25105310001
3	25	2510	251053	25105310001
4	25	2510	251053	25105310001
...				

key

	성별	연령	혼인상태	교육수준
1	1	13	2	4
2	2	13	2	2
3	2	13	3	3
4	2	8	1	6
...				

이 예제에서는 H키변수가 총 4개의 수준을 가지면 키변수는 4개로 구성되어 있다. 마이크로데이터에 빈도표를 계산하기 위한 가중값이 존재하면 wght라는 변수에 그 가중값을 할당하고, 가중값이 없다면 1의 값으로 구성된 가중값 벡터를 다음과 같이 할당한다.

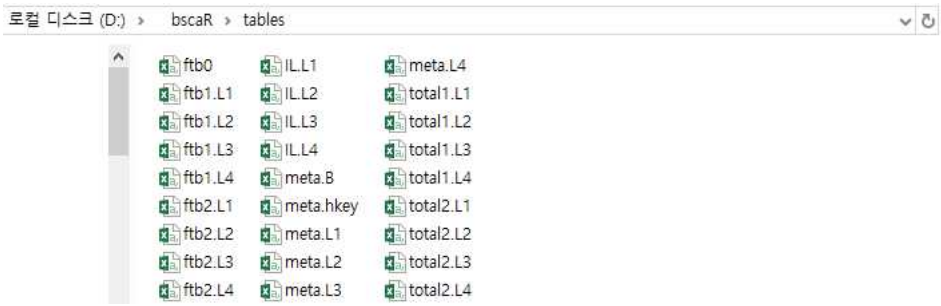
```
wght = rep(1, length(key[[1]]))
```

이제 다음과 같이 saveFullTables 함수를 실행하면 함수는 특정 폴더(예. bscaR/tables)

에 H키변수 수준별로 모든 키변수로 구성된 기저 빈도표를 저장한다.

```
# 기저 빈도표 저장
saveFullTables(hkey, key, wght, B)
```

기저 빈도표의 이름은 ftb.L로 구성되어 있다. 참값 빈도표는 ftb1 이라는 이름의 파일들로, BSCA가 적용되어 비밀보호 처리된 빈도표는 ftb2 라는 이름의 파일로 이루어져 있다. 키변수 조합의 정보는 ftb0 라는 이름의 파일로 저장된다. 이용자는 두 파일을 결합하여 원하는 빈도표를 만들 수 있다. 한편 정보손실은 IL로 시작하는 파일에, 총계는 total로 시작하는 이름의 파일에 저장되어 있다. 이용자는 ftb2의 결과를 수령한다.



한편 IL.L 파일에는 H키변수 수준별로 정보손실이 어느 정도 발생하였는지 정리되어 있다. $B=3$을 기준으로 SCA의 경우 정보손실이 1 또는 2만큼 발생하지만, BSCA를 적용하면 1만큼 추가로 발생할 수 있다. 일반적으로 정보손실은 SCA의 경우 1에서 $(B-1)$까지, BSCA의 경우 $[B/2]$만큼 정보손실이 추가로 발생할 수 있다. 아래 그림은 IL파일 예제로 정보손실의 크기에 따른 셀의 개수와 비율을 보여준다.

IL	-3	-2	-1	0	1	2	3
Freq	462	9193	12346	83428	6706	151	34
Percent	0.41	8.18	10.99	74.28	5.97	0.13	0.03

빈도표 파일의 예제는 다음과 같다.

키변수 조합 (ftb0.csv)

	A	B	C	D	E
1	성별	연령	혼인상태	교육수준	
113	1	3	4	8	
114	1	3	9	1	
115	1	3	9	2	
116	1	3	9	3	
117	1	3	9	4	
118	1	3	9	5	
119	1	3	9	6	
120	1	3	9	7	
121	1	3	9	8	
122	1	4	1	1	
123	1	4	1	2	
124	1	4	1	3	
125	1	4	1	4	
126	1	4	1	5	
127	1	4	1	6	
128	1	4	1	7	
129	1	4	1	8	
130	1	4	2	1	
131	1	4	2	2	
132	1	4	2	3	
133	1	4	2	4	

지역별 빈도 (ftb2.L2.csv)

	A	B	C	D	E
1	2510	2520	2530	2540	2550
113	0	0	0	0	0
114	3	0	3	3	8
115	3825	4548	10255	6137	4239
116	3300	4067	9308	5095	3779
117	17	29	38	23	11
118	0	0	0	0	0
119	0	0	0	0	0
120	0	0	0	0	0
121	0	0	0	3	0
122	5	14	8	8	5
123	14	11	14	14	23
124	1247	1466	3410	1710	1323
125	5100	6092	13659	6835	5209
126	997	538	1195	996	499
127	2022	812	2839	3488	1160
128	0	0	0	0	0
129	0	0	0	0	0
130	0	0	0	0	0
131	3	0	0	0	0
132	5	5	14	5	3
133	17	20	26	8	17

saveFullTables 함수는 내부적으로 makeFullTableRcsv, sca, bsca라는 세 개의 함수를 사용한다. makeFullTableRcsv 함수[1]는 재귀적 방식으로 모든 H키변수 수준에 대하여 빈도표 및 자료처리를 위한 메타 정보를 생성한다. sca 및 bsca는 앞에서 논의한 BSCA알고리즘을 구현한 함수이다. makeFullTableRcsv 함수는 리스트 형식의 결과물을 반환하며 이 리스트는 두 개의 성분으로 구성되어 있다. 첫 번째 성분에는 빈도 값들이 저장되어 있고, 두 번째 성분에는 메타 정보가 수록되어 있다. 메타 정보는 각 H키변수 수준별로 최소 단위(예, 집계구)의 개수이다. BSCA 프로그램은 메타 정보에 따라 빈도 행렬을 끊어 재집계하는 원리로 빈도표들을 재귀적으로 생성하도록 구현되어 있다.

BSCA 알고리즘 적용을 위하여 먼저 makeFullTableRcsv 함수가 반환한 빈도 행렬에 대하여 SCA를 적용하여 기저 빈도표를 생성한다. 다음으로 메타 정보를 따라 행렬을 나누고, 한 부분씩 이동하면서 BSCA를 적용하여 새로운 빈도를 생성하고 반환한다. 더불어 참 빈도표와 BSCA를 적용한 빈도표 사이의 정보손실을 파일로 저장한다. 코드는 다음과 같다.[2]

1) makeFullTableRcsv 함수의 이용은 통계청 SGIS 시스템에서 데이터 마트(data mart)를 생성하는 것과 동일하다. 다만 재귀함수를 사용하여 속도를 상당히 개선하였음을 참고할 만하다.

2) 시스템 운영을 위하여 참고로 활용하도록 메타 정보도 함께 저장한다.

```
saveFullTables = function(hkey, key, wght, B) {

  pp0 = dim(key)[2] # 키변수 개수

  # construct the key frame : data.frame consisting of all combinations in the key(microdata)
  ctgr = getCategoryTb(key)

  # when hkey does not exist
  if (is.null(hkey)) {hkey = data.frame(rep(1, length(key[[1]]))); names(hkey) = "name"}

  # construct full tables using makeFullTableRcsv function
  full = makeFullTableRcsv(hkey, key, ctgr, wght)

  freq1 = freq2 = full$freq
  meta1 = full$meta
  total1 = total2 = list(NULL) # total sum at each hierarchical level

  # construct masked full tables
  lv0 = length(freq1)

  freq2[[lv0]] = t(sapply(as.data.frame(t(freq1[[lv0]])), sca, B))
  if (dim(freq2[[lv0]])[1] != dim(ctgr)[1]) freq2[[lv0]] = t(freq2[[lv0]])
  rownames(freq2[[lv0]]) = NULL
  if (lv0 > 1) for (ii in 1:(lv0-1)) freq2[[ii]] = bsca(freq1[[lv0]], freq2[[lv0]], meta1[[ii]], B)
  for (ii in 1:lv0) colnames(freq2[[ii]]) = colnames(freq1[[ii]])

  # calculate total sums
  for (ii in 1:lv0) total1[[ii]] = colSums(freq1[[ii]])
  for (ii in 1:lv0) total2[[ii]] = bscaTotal(freq1[[lv0]], freq2[[lv0]], meta1[[ii]], B)

  # save full tables
  write.csv(ctgr, file = "tables/ftb0.csv", row.names = F)
  for (ii in 1:lv0) write.csv(freq1[[ii]], file = paste("tables/ftb1.", "L", ii, ".csv", sep = ""), row.names = F)
  for (ii in 1:lv0) write.csv(freq2[[ii]], file = paste("tables/ftb2.", "L", ii, ".csv", sep = ""), row.names = F)

  # save meta information
  for (ii in 1:lv0) write.csv(meta1[[ii]], file = paste("tables/meta.", "L", ii, ".csv", sep = ""))
  write.csv(aggregate(wght, by = as.list(hkey), FUN="sum"), file = "tables/meta.hkey.csv", row.names = F)
  write.csv(B, file = "tables/meta.B.csv", row.names = F)

  # save total sums
  for (ii in 1:lv0) write.csv(total1[[ii]], file = paste("tables/total1.", "L", ii, ".csv", sep = ""), row.names = F)
  for (ii in 1:lv0) write.csv(total2[[ii]], file = paste("tables/total2.", "L", ii, ".csv", sep = ""), row.names = F)

  # save information loss
  for (ii in 1:lv0) {
    rst = as.data.frame(table(freq1[[ii]]-freq2[[ii]])); names(rst)[1] = "IL"
    Percent = round(prop.table(rst$Freq)*100, 2)
    write.csv(cbind(rst, Percent), file = paste("tables/IL.L", ii, ".csv", sep = ""), row.names = F)
  }
}
```

BSCA 알고리즘을 구현하는 주요 함수인 sca 및 bsca를 설명하면 다음과 같다. 함수 sca는 기저 빈도표에서 H키변수 중 최소 단위(예. 집계구)마다 SCA, 작은 셀에 대한 랜덤 라운딩을 적용하고 그 결과를 반환한다. 즉, 집합 $\{f_{ij}\}$를 freq이라는 벡터에 저장했다면 주어진 작은 셀 기준값 B에 대하여 다음과 같은 함수를 사용하면 SCA를 적용한 결과인 $\{\tilde{f}_{ij}\}$를 반환한다.

```
sca = function(freq, B) {

  #' @param freq Numeric vector : a frequency vector
  #' @param B Number : the value defining small cells
  #'
  #' @return Numeric vector ; a transformed frequency vector in whose small cells are adjusted
                                by random rounding

  if(!is.vector(freq)) freq = as.vector(freq)
  mm = length(freq)

  for (ii in 1:(B-1)) {
    indx = c(1:mm)[freq == ii]
    lth = length(indx)

    if (lth < B) {
      for (jj in 1:lth) freq[indx[jj]] = ifelse(runif(1) < ii/B, B, 0)
    } else {
      indxB = indx[order(runif(lth))][1:round(lth*ii/B)]
      freq[indxB] = B
      freq[setdiff(indx, indxB)] = 0
    }
  }

  return(freq)
}
```

함수 bsca는 주어진 메타 정보를 이용하여 상위 빈도표 범위에 맞추어 기저 빈도표에서 최소 단위의 범위(예. 읍면동 1에 속한 집계구들의 모임)를 나눈다. 주어진 범위에 대하여 다음과 같이 bscaVec 함수를 실행하여 상위 빈도표의 셀값을 계산해 나간다. bscaVec 함수는 $\{f_{ij}\}$ 및 $\{\tilde{f}_{ij}\}$를 입력값으로 하여 BSCA를 적용하고 $\tilde{f}_i$를 반환한다.

한편 saveAllHLevels 함수는 주어진 키변수에 대한 모든 H키변수 수준에서의 빈도표를 특정 폴더(예. bscaR/done)에 저장한다. saveSelected 함수는 주어진 키변수 및 H키변수에 대한 빈도표를 특정 폴더에 저장한다. 이 두 함수는 saveFullTables 함수가 생성한 tables의 결과를 사용한다. 특히 saveSelected 함수의 경우 정보손실 파일로

tables의 결과를 그대로 사용한다. 한편 bscaVec 함수의 코드는 다음과 같다.

```
bscaVec = function(dd1, dd2, B) {

  #' @param dd1 Numeric vector : a true frequency vector
  #' @param dd2 Numeric vector : a masked frequency vector by SCA
  #' @param B Number : the value defining small cells
  #'
  #' @return Number : a transformed/masked frequency value by BSCA

  if(!is.vector(dd1)) dd1 = as.vector(dd1)
  if(!is.vector(dd2)) dd2 = as.vector(dd2)

  indx_dd1_zsb = (dd1 <= B)
  indx_dd2_b    = (dd2 == B)

  K = sum(indx_dd1_zsb)
  k = sum(indx_dd2_b)

  d1 = sum(dd1[indx_dd1_zsb])
  d2 = sum(dd2[indx_dd1_zsb])

  if(K > 1) {
    v = k
    w = k + K*(B-1)

    a = (d1-1)%/%B
    d2 = a*B + trunc(B/2) + 1

    r = a*B + 1
    s = (a + 1)*B

    if(a < 0) {
      d2 = 0
    }else if(r < v){
      d2 = d2 + B
    }else if(s > w){
      d2 = d2 - B
    }

    if(d2 > 0 & d2 < B) d2 = B
  }

  # adding large cells
  d2 = d2 + sum(dd1[!indx_dd1_zsb])

  return(d2)
}
```

한편 BSCA는 사전에 정의된 구조의 빈도표를 범주의 중복을 허락하지 않고 제공하는 것을 가정하여 고안되었다(박민정 외, 2018). 그러나 빈도표를 제공하는 SGIS 시스템은 이용자의 요구를 반영하여 다양한 선택을 허용하는 방향으로 진화하고 있다. 실시간 선택을 허용할 때 노출 위험을 가장 높이는 경우는 기저 빈도표를 집계할 때 셀이 하나만 추가되는 경우이다. 대다수의 경우 BSCA는 SCA가 의도하는 셀 값 혹은 개체에 대한 익명성(B)을 보장한다. 그러나 다음과 같은 사례에서는 BSCA가 보장하는 셀 값에 대한 익명성이 하나 감소할 수 있다. 아래 예에서는 이용자가 기저 빈도표의 셀을 임의로 선택할 수 있고, 작은 셀 기준 $B = 3$ 이라고 가정한다.

이 경우에 파란색으로 0이라고 표시된 셀 값은 원래의 의도로는 $\{0, 1, 2\}$ 중 하나라는 정보를 제공하는 것이 목적이었으나, 두 요청에 대하여 시스템이 제공한 값의 차이로 인하여 $\{1, 2\}$ 중 하나라는 것을 이용자가 알 수 있다. 즉, 셀 값에 대한 익명성을 $B = 3$ 으로 보장하지 못하고 익명성 2를 확보하는 데 그치게 된다. 한편 이용자가 특정 위치의 셀에 대하여 다음과 같이 요청할 때도 동일한 현상이 발생할 수 있다.

요청 1

참값 기저 빈도표

연령	0~4세	5~9세	10~14세	15~19세	20~24세	25~29세	30~34세	-
빈도	1	1	1	1	1	1	1	-

SCA 처리된 기저 빈도표

연령	0~4세	5~9세	10~14세	15~19세	20~24세	25~29세	30~34세	-
빈도	0	0	0	0	0	0	0	-

BSCA 처리된 빈도표

연령	0~4세	5~9세	10~14세	15~19세	20~24세	25~29세	30~34세	-
빈도	5							-

요청 2

참값 기저 빈도표

연령	0~4세	5~9세	10~14세	15~19세	20~24세	25~29세	30~34세	35~39세
빈도	1	1	1	1	1	1	1	1

SCA 처리된 기저 빈도표

연령	0~4세	5~9세	10~14세	15~19세	20~24세	25~29세	30~34세	35~39세
빈도	0	0	0	0	0	0	0	***0***

BSCA 처리된 빈도표

연령	0~4세	5~9세	10~14세	15~19세	20~24세	25~29세	30~34세	35~39세
빈도	8							

유계한 정보손실을 범위를 유지하면서 이러한 현상까지 모두 방어하여 셀 값의 익명성을 유지하는 방법을 찾는 것은 대단히 어려운 일이라고 판단된다. 사전에 정의된 구조에 한정하여 자료 분석 결과를 제공하는 것을 넘어서서, 이용자의 모든 요청에 응하면서도 사전에 결정한 노출위험 수준을 유지하는 방법은 현재로서는 차등정보보호(differential privacy)의 적용만이 있을 것이라고 생각된다. 그러나 차등정보보호를 적용하면 정보손실이 크게 발생한다는 것은 잘 알려진 사실이다.

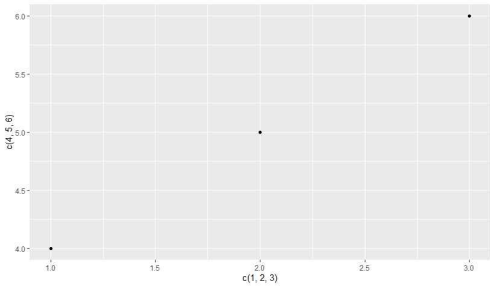
BSCA는 작은 셀을 정의하는 값 B를 기준으로 셀 값 혹은 셀에 속한 개체의 익명성을 보장하도록 설계되어 있다. 다만 이는 사전에 규정된 지역 범위에서 변수 범주 선택의 중복을 허락하지 않음을 가정한다. 이용자의 다양한 선택을 허용하면 BSCA가 보장하는 익명성은 $B-1$로 감소한다. 따라서 충분한 크기의 B값을 선택할 필요가 있다. 참고로 통계청이 전통적으로 사용해 온 사업체 3, 개인 및 가구 5라는 익명성 기준을 사용하면서 이용자의 다양한 선택을 허용한다면, BSCA가 적용된 셀 값 일부에 대하여 익명성이 각각 2 및 4로 감소할 수 있다.

1.4 JAVA 시스템에서 BSCA 알고리즘의 구현 방식

BSCA는 R을 이용하여 구현되어 있다. 반면 대개의 시스템은 JAVA로 구현되어 있다. 이에 통계청 데이터센터 또는 SGIS와 같은 자료제공 플랫폼에서 빈도표를 제공할 때 BSCA알고리즘을 바로 사용하기 위해서는 JAVA와 R을 연동할 필요가 있다. R과 JAVA를 연동하는 것은 Rserve가 제공하는 REngine.jar 혹은 RserveEngine.jar와 같은 JAVA Library를 활용하면 된다. 주요 참조 사이트는 https://www.rforge.net/Rserve 이다.

현재 통계청 데이터센터에서 시범적으로 구현하고 있는 방법은 R과 JAVA를 연동한 이후에 JAVA에서 R로 특정 명령문을 직접 전달하는 것이다. 명령문 전달을 위하여 JAVA에서 c.eval 함수를 이용할 수 있다. 이 함수의 간단한 사용 예는 다음과 같다.

함수 (JAVA 시스템)	결과 (특정 위치, 예. 'done/ex.jpg')

c.eval("library(ggplot2)");
 	
c.eval("ggplot(data.frame(a=c(1,2,3), b=c(4,5,6)),
aes(x=c(1,2,3),y=c(4,5,6)))+geom_point()");
 	
c.eval("ggsave('done/ex.jpg')");
	

BSCA를 구현하기 위하여 위의 방법을 활용할 수 있다. 다만 효율성을 위하여 기존의 JAVA 시스템에서 최소한의 옵션을 R로 전달하고, R은 특정 위치에 사전에 규정해 둔 대로 빈도표를 저장한다. 이제 JAVA 시스템은 특정 위치에 저장된 빈도표를 불러오고, 기존 시스템을 활용하여 이용자에게 결과를 서비스한다. 다만 이렇게 저장장치를 이용하여 두 프로그램을 연계하는 방식은 DB를 구축하는 데는 문제가 없으나, 이용자의 선택에 따른 빈도표를 실시간으로 서비스할 때는 속도가 느리다는 문제가 발생한다.

실시간 서비스를 위해서는 JAVA 시스템에서 참값으로 구성된 기저 빈도표와 SCA(작은 셀에 랜덤 라운딩) 처리된 기저 빈도표를 DB로 운영하면서, 상위 빈도표를 제공할 때 BSCA 알고리즘을 실시간으로 적용하는 것이 대안이 될 수 있다. 이를 위해서는 BSCA를 JAVA 시스템에 직접 구현할 필요가 있어, 이를 지원하기 위하여 부록에서 알고리즘과 코드를 상세히 설명하였다. 현재 SGIS 시스템은 기저 빈도표를 데이터마트(datamart)로 이용하여 지도상에 이용자의 요청에 따른 빈도를 제공하고 있

다. 따라서 시스템이 기저 빈도표의 참값을 단순 집계하는 대신에, bscaVec 함수를 이용하여 집계하도록 수정하여 BSCA를 구현할 수 있다.

부 록 2

비밀보호가 적용된 통계지리정보서비스(SGIS) 자료 안내서

통계개발원 통계방법연구실

차 례

1. SGIS의 빈도표 제공 1

2. 빈도표의 노출 위험 2

2.1 빈도표에서 발생하는 노출 위험의 개념 2

2.2 빈도표에서 발생하는 노출 위험의 유형 3

3. 빈도표의 비밀보호 4

3.1 빈도표의 위계 구조 4

3.2 빈도표의 비밀보호(BSCA) 알고리즘 6

3.3 기저 빈도표에서의 비밀보호 8

3.4 상위 빈도표에서의 비밀보호 9

3.4.1 제공 가능한 상위 빈도표에 대한 논의 10

3.4.2 중앙값 조정을 통한 비밀보호 12

3.4.3 추가 조치를 통한 노출 위험 제어와 정보 손실 축소 14

3.5 현행 자료 제공 방식과의 비교 18

1. SGIS의 빈도표 제공

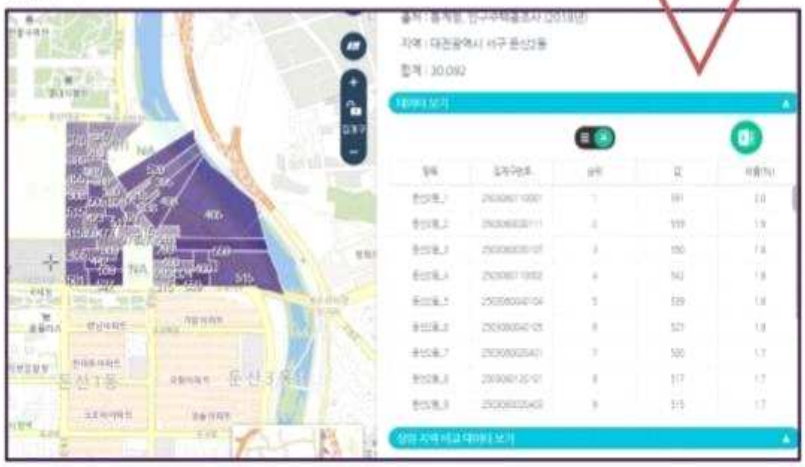
▶ 통계지리정보서비스에서 제공하는 빈도표

현재 SGIS에서 제공하고 있는 다양한 서비스 중 '대화형 통계지도'에 주목해 보려고 합니다. 대화형 통계지도는 인구주택총조사, 농림어업총조사, 전국사업체조사 등의 조사 자료를 사용자가 원하는 조건에 맞게 검색할 수 있는 맞춤형 소지역(small area) 통계 서비스입니다.

예를 들어, 인구주택총조사 자료에서 대전광역시 서구 둔산2동의 10세 이상~64세 미만인 인구 수를 요청하면 다음과 같은 결과를 얻을 수 있습니다. 자료를 요청한 지역의 집계구 경계와 인구 수를 함께 볼 수 있습니다.

[그림 1] SGIS 대화형 통계지도 예시

이 때 화면 오른쪽에 표시되는 데이터 보기를 확대해 보면 다음과 같은 표를 확인할 수 있습니다.



[그림 2] SGIS 대화형 통계지도에서 제공하는 빈도표

사용자에게 제공되는 이러한 형태의 통계자료를 빈도표(frequency table)라 하고, SGIS의 대화형 통계지도에서 제공하는 자료는 빈도표인 경우가 많이 있습니다.

2. 빈도표의 노출 위험

2.1 빈도표에서 발생하는 노출 위험의 개념

SGIS의 인구주택총조사 자료에서 제공하는 항목은 2018년 자료를 기준으로 성별, 연령, 세대구성, 주택면적, 주택유형, 건축연도 총 6개입니다. 각 항목들은 여러 개의 범주로 나뉘어져 있으며, 사용자는 필요에 따라 항목을 조합하여 다양한 빈도표를 만들어 볼 수 있도록 서비스하고 있습니다.

이렇게 사용자의 요청에 의해 만들어진 빈도표에서는 특정 개인이 식별될 위험이 존재합니다. 특히 자료 제공자 입장에서는 만들어진 빈도표에서 특정 빈도(인구 수)가 1명 또는 2명인 경우에 노출 위험이 있다고 판단하게 됩니다.[1]

[1] Hundepool 외(2010)

예를 들어, 대전시 서구 둔산2동에 거주하는 사용자 A가 있다고 가정해 보겠습니다.

사용자 A는 자신과 비슷한 환경에 거주하는 인구는 얼마나 되는지를 궁금하게 생각하여 SGIS에서 조사해 보기로 하였습니다. 사용자 A의 특성은 남자이면서 1인 가구이고 주택면적이 20 ㎡ 이하인 아파트에 거주하고 있습니다.

사용자 A가 만든 빈도표가 <표 1>과 같다고 하겠습니다.

<표 1> 노출 위험이 있는 빈도표 예시1

성별	세대 구성	주택유형	주택면적	빈도(인구 수)
남자	1인 가구	아파트	20㎡ 이하	1

사용자 A는 대전시 서구 둔산2동에 거주하면서 위와 같은 특성을 가진 사람은 본인밖에 없다는 사실을 깨닫게 됩니다. 이때 만약 사용자 A를 알고 있는 다른 사용자 B가 동일한 항목으로 빈도표를 만든다고 가정해 보겠습니다. 이 경우 사용자 B 또한 <표 1>을 얻게 될 것이고, 빈도(인구 수) 1명에 해당하는 사람이 사용자 A임을 알아차리게 됩니다. 다시 말해, <표 1>에서 빈도(인구 수)가 1명이었기 때문에 사용자 A는 개인 식별 위험에 노출되게 됩니다.

이번에는 동일한 항목에 대해 조사된 빈도(인구 수)가 2명이라고 가정해 보겠습니다.

이때 빈도(인구 수) 2명에 해당하는 사람이 사용자 C와 D라고 한다면, 각 개인은 빈도(인구 수) 2명 중 1명이 본인이라는 사실을 깨닫게 되고 나머지 1명은 상대방이 된다는 것을 알아차리게 됩니다. 즉, <표 2>에서 빈도(인구 수)가 2명이었기 때문에 사용자 C와 D는 상호간에 식별 위험에 노출되게 됩니다.

<표 2> 노출 위험이 있는 빈도표 예시2

성별	세대 구성	주택유형	주택면적	빈도(인구 수)
남자	1인 가구	아파트	20㎡ 이하	2

따라서 자료 제공자는 특정 빈도(인구 수)가 얼마나 큰 값일 때 조사된 자료를 그대로 제공할 수 있을지에 대해 고민하게 됩니다. 여기서 노출 위험이 없다고 단정할 수 있는 절대적인 기준이 있다기보다는, 자료의 특성에 따라 기준이 되는 값은 달라질 수 있다고 보는 시각이 우세합니다. 현재 SGIS에서는 5명 미만의 빈도(인구 수)는 노출 위험이 큰 것으로 판단하여, 해당 값에 대해 N/A 표시하여 제공하고 있습니다.

2.2 빈도표에서 발생하는 노출 위험의 유형

앞서 빈도표에서 노출의 위험이 발생하는 간단한 예를 살펴보았습니다. 빈도표에서 노출의 위험이 발생하는 몇 가지 경우를 더 고려해 보도록 하겠습니다.

예를 들어 살펴보겠습니다.

자료 사용자가 가나다광역시 ABC 구 ㄱㄴ동에 거주하는 0 세 이상 ~ 20 세 미만의 남·여별 인구 수를 요청한 경우, 제공되는 빈도표가 다음과 같다고 가정해 보겠습니다.

〈표 3〉 노출 위험이 있는 빈도표 예시1

성별	연령(5세 단위 제공)	인구 수	소계
남	0세 이상 ~ 5세 미만	1	55
	5세 이상 ~ 10세 미만	14	
	10세 이상 ~ 15세 미만	40	
	15세 이상 ~ 20세 미만	0	
여	0세 이상 ~ 5세 미만	0	2
	5세 이상 ~ 10세 미만	0	
	10세 이상 ~ 15세 미만	0	
	15세 이상 ~ 20세 미만	2	
소계		57	57

[2] 개별 속성 노출

[3] 그룹 속성 노출

〈표 3〉에 제시된 빈도표에서는 2 가지 유형의 노출 위험이 존재합니다.

첫 번째는 0 세 이상~5 세 미만의 남자 빈도(인구 수)가 1 명이기에 노출 위험이 있습니다.[2]

두 번째는 20 세 미만의 여자는 연령이 모두 15 세 이상~20 세 미만에 해당한다는 사실이 노출됩니다.[3]

[4] 차이에 의한 노출

한편, SGIS 자료 사용자는 다양한 항목을 조합하여 상당히 많은 수의 빈도표를 만들 수 있습니다. 이 경우 특정 지역에 대해 동일한 항목의 조건을 달리하여 여러 개의 빈도표를 만들고 이를 조합해 봄으로써 노출이 발생할 수 있게 됩니다.[4]

만약 자료 사용자가 가나다광역시 ABC 구 ㄱㄴ동에 거주하는 남자 인구 수를 조사하고 있는데, 연령을 다르게 하여 2 개의 빈도표를 만들었다고 가정하겠습니다. 이 경우 사용자는 〈표 4〉를 조합해 봄으로써, 가나다광역시 ABC 구 ㄱㄴ동에 거주하는 40 세 이상~45 세 미만의 남자 인구는 1 명임을 알아차리게 됩니다.

〈표 4〉 노출 위험이 있는 빈도표 예시2

성별	연령(5세 단위 제공)	소계
남	40세 이상 ~ 60세 미만	55

성별	연령(5세 단위 제공)	소계
남	45세 이상 ~ 60세 미만	54

[5] 박민정 외(2018)

이상에서 다루어진 빈도표에서의 노출 위험을 최소화하면서 자료를 제공할 수 있는 방법으로 제안된 **BSCA(Bounded Small Cell Adjustments)**[5] 알고리즘에 대해 살펴보겠습니다.

3. 빈도표의 비밀보호

3.1 빈도표의 위계 구조

빈도표에서의 노출 위험을 줄인 자료를 제공하기 위한 BSCA 알고리즘을 이해하기 위해서는 먼저 빈도표의 위계(hierarchy) 구조를 이해할 필요가 있습니다. 다시 말해 SGIS 에서 제공되는 빈도표는 위계 구조를 가지고 있기 때문에 이에 관해 살펴보겠습니다.

SGIS 자료의 빈도표가 가지는 위계 구조는 2 가지 측면에서 고려해야 합니다.

첫 번째는 지역 수준에 의한 위계 구조로, 좁은 지역(small area) 수준에서 조사된 빈도(인구 수)를 집계하여 보다 넓은 지역 수준의 빈도(인구 수)를 구하게 되는 것입니다.

예를 들어 대전광역시 서구 월평동의 경우에는 31 개의 집계구[6]가 모여 1 개의 월평동이 되고, 11 개의 동이 모여 대전시 서구를 형성하게 됩니다. 또 5 개의 구가 모여 대전광역시를 이루는 위계 구조를 보입니다.

두 번째는 제공되는 항목 수에 의한 위계구조로, 제공 가능한 모든 항목에서 조사된 빈도(인구 수)를 집계하여 항목의 수가 줄어든 경우의 빈도(인구 수)를 구하게 됩니다.

20 년 9 월 현재 SGIS 대화형 통계지도의 인구주택총조사 자료에서 제공하는 항목은 성별, 연령, 세대구성, 주택면적, 주택유형, 건축연도 등 총 6 개 항목입니다. 만약 사용자가 6 개 항목 중 5 개 항목에 대한 빈도표를 요구한 경우에는 선택되지 않은 1 개 항목의 빈도를 집계하여 제공하게 되므로 위계 구조가 생기게 됩니다.

이러한 빈도표의 위계구조에서 제공 가능한 모든 항목을 고려한 경우에 가장 좁은 지역(small area) 수준에서 만들어지는 빈도표가 **기저(base) 빈도표**가 됩니다. 따라서 인구주택총조사 자료에 해당하는 기저 빈도표는 성별, 연령, …, 건축연도 등 6 개 항목을 고려한 집계구 단위에서 만들어진 빈도(인구 수)를 조사한 표라고 이해할 수 있습니다. 이때 인구주택총조사 자료의 기저 빈도표는 행(row)의 수가 158,760 개[7] 이고, 열(column)의 수는 집계구 수[8] 만큼인 형태로 매우 많은 행과 열을 가지는 빈도표가 됩니다.

〈표 5〉 기저(base) 빈도표 예시

제공 항목						지역 수준				
성별	연령	세대 구성	주택 면적	주택 유형	건축 년도	1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	…
남	0~5세	1세대	~20㎡	다세대	2015년	3	0	0	0	…
여	0~5세	2세대	~40㎡	단독	2010년	10	4	6	14	…
남	5~10세	3세대	~60㎡	아파트	2018년	4	7	6	11	…
여	5~10세	4세대	~85㎡	연립	2018년	0	0	1	0	…
남	0~5세	1인	~100㎡	영업용	2010년	0	0	0	0	…
여	0~5세	비친족	~130㎡	비주거용	2019년	5	3	4	8	…
남	0~5세	1세대	~165㎡	다세대	2016년	5	5	5	11	…
…	…	…	…	…	…	…	…	…	…	…
남	100세+	1인	230㎡+	아파트	2016년	0	1	0	0	…
여	100세+	3세대	~100㎡	연립	2019년	2	0	0	0	…
남	100세+	4세대	~130㎡	연립	2016년	4	3	3	4	…
여	100세+	비친족	~165㎡	영업용	2016년	5	5	10	13	…
…	…	…	…	…	…	…	…	…	…	…
남	100세+	2세대	~60㎡	아파트	2016년	0	0	1	1	…
여	100세+	2세대	~100㎡	아파트	2019년	2	7	2	6	…
…	…	…	…	…	…	…	…	…	…	…
남	100세+	3세대	~130㎡	아파트	2016년	0	0	0	1	…
여	100세+	1인	~165㎡	단독	2016년	0	2	0	0	…

6 기초단위구를 기반으로 인구규모(최적 500명), 사회경제적 동질성(주택유형, 지가), 집계구 형상을 고려하여 구축한 최소 통계 집계구역 (https://sgis.kostat.go.kr)

7 성별(남·여, 2개 범주), 연령(0세~100세+, 5세 단위 21개 범주), 세대구성(6개 범주), 주택유형(5개 범주), 주택면적(9개 범주), 건축연도(14개 범주), (158,760 = 2×21×6×5×9×14)

8 2017년 인구주택총조사 기준, 총 106,347개

〈표 5〉의 기저 빈도표에서 집계구별로 자료를 집계하여 만든 〈표 6〉은 지역 수준이 집계구에서 읍면동으로 넓어진 상위 빈도표에 해당합니다.

〈표 6〉 상위 빈도표 예시1(지역의 범위가 넓어진 빈도표)

제공 항목						지역 수준
성별	연령	세대 구성	주택 면적	주택 유형	건축 년도	ㄱ.ㄴ동
남	0~5세	1세대	~20㎡	다세대	2015년	3
여	0~5세	2세대	~40㎡	단독	2010년	34
남	5~10세	3세대	~60㎡	아파트	2018년	28
여	5~10세	4세대	~85㎡	연립	2018년	0
남	0~5세	1인	~100㎡	영업용	2010년	0
여	0~5세	비친족	~130㎡	비주거용	2019년	20
…	…	…	…	…	…	…
남	100세+	1인	230㎡+	아파트	2016년	0
여	100세+	3세대	~100㎡	연립	2019년	0
남	100세+	4세대	~130㎡	연립	2016년	14
여	100세+	비친족	~165㎡	영업용	2016년	33
…	…	…	…	…	…	…
남	100세+	2세대	~60㎡	아파트	2016년	2
여	100세+	2세대	~100㎡	아파트	2019년	17
…	…	…	…	…	…	…
남	100세+	3세대	~130㎡	아파트	2016년	1
여	100세+	1인	~165㎡	단독	2016년	2

마찬가지로 〈표 5〉에서 성별 항목에 해당하는 남·여의 빈도(인구 수)를 집계하여 만들어진 〈표 7〉은 항목의 수가 6 개에서 5 개로 줄어든 상위 빈도표에 해당합니다.

〈표 7〉 상위 빈도표 예시2(항목의 수가 줄어든 빈도표)

제공 항목					지역 수준				
연령	세대 구성	주택 면적	주택 유형	건축 년도	1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	…
0~5세	1세대	~20㎡	다세대	2015년	7	5	6	7	…
0~5세	2세대	~40㎡	단독	2010년	15	10	13	22	…
5~10세	3세대	~60㎡	아파트	2018년	10	14	14	20	…
5~10세	4세대	~85㎡	연립	2018년	7	8	10	10	…
0~5세	1인	~100㎡	영업용	2010년	8	9	10	11	…
0~5세	비친족	~130㎡	비주거용	2019년	14	13	15	20	…
0~5세	1세대	~165㎡	다세대	2016년	15	16	17	24	…
…	…	…	…	…	…	…	…	…	…
100세+	1인	230㎡+	아파트	2016년	13	15	15	16	…
100세+	3세대	~100㎡	연립	2019년	16	15	16	17	…
100세+	4세대	~130㎡	연립	2016년	19	19	20	22	…
100세+	비친족	~165㎡	영업용	2016년	21	22	28	32	…
…	…	…	…	…	…	…	…	…	…
100세+	2세대	~60㎡	아파트	2016년	18	19	21	22	…
100세+	2세대	~100㎡	아파트	2019년	21	27	23	28	…
…	…	…	…	…	…	…	…	…	…
100세+	3세대	~130㎡	아파트	2016년	21	22	23	25	…
100세+	1인	~165㎡	단독	2016년	22	25	24	25	…

3.2 빈도표의 비밀보호(BSCA) 알고리즘

BSCA(Bounded Small Cell Adjustmetns) 알고리즘은 위계 구조를 가지는 빈도표에서 노출 위험을 최소화하면서 동시에 실제 조사된 값과의 차이는 3 이내가 되도록 마련되었습니다. 만약 자료 제공자가 조사된 빈도(인구 수)가 **3명**[9] 미만인 경우에 노출 위험이 있는 것으로 판단한다면, 이 경우 노출 위험이 제어되도록 빈도(인구 수)를 조정하되 실제 조사된 값과의 차이를 3 이내로 최소화하여 제공하는 방법으로 이해할 수 있습니다.

9 작은 빈도를 갖는 셀로 판단하는 기준

이제부터 BSCA 알고리즘의 보다 상세한 이해를 돕고, 논의의 편의를 위해서 아래와 같이 가정해 보겠습니다. 자료에서 제공되는 항목은 성별과 주택유형 2 가지만 있고, 집계구가 5 개만 존재하여 기저 빈도표가 다음과 같이 조사되었다고 하겠습니다.[10]

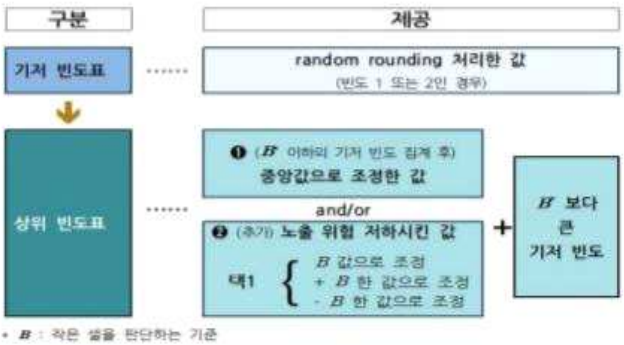
10 실제로는 SGIS에서 제공하는 6개의 항목과 모든 집계구를 반영한 기저 빈도표가 생성됨. 또한, 집계구별 인구수는 500명 내외가 되지만, 이 또한 논의를 위해 조정함.

<표 8> 조사된 기저(base) 빈도표

조사된 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	1	0	0	0	13
	아파트	18	3	2	0	36
	연립주택	4	4	1	7	22
	다세대	5	8	0	18	12
	비주거용	1	2	0	4	15
여	단독주택	4	5	8	3	1
	아파트	9	4	9	2	1
	연립주택	5	10	14	5	1
	다세대	10	6	7	8	1
	비주거용	4	5	11	11	1

이렇게 조사된 기저 빈도표와 이를 집계한 상위 빈도표에서 노출의 위험은 거의 언제나 존재합니다. 위 표에서 보듯이 빈도가 1 또는 2 인 셀에서 노출의 위험이 있다고 이해하면 됩니다. 또 제공하는 항목의 수가 더 많아질수록 노출의 위험이 증가한다는 것도 예상할 수 있습니다.

이러한 문제를 해결하기 위한 방법으로 BSCA 알고리즘이 제안되었고, 해당 알고리즘은 기저 빈도표와 상위 빈도표에 따라 그 방법이 다르게 적용되므로 이를 구분하여 설명하려 합니다. 빈도표에 대한 비밀보호 알고리즘의 개요를 나타내면 다음과 같습니다.



[그림 3] 빈도표에 대한 비밀보호 알고리즘 개요

기저 빈도표는 제공 가능한 모든 항목을 고려한 가장 좁은 지역(small area) 수준에서 작성된 빈도표입니다. 기저 빈도표에서는 관측된 빈도가 0(zero)인 경우가 많고, 빈도가 1 또는 2 인 경우도 존재하기 때문에 노출의 위험은 거의 언제나 발생합니다. 만약 자료 제공자가 작은 빈도를 갖는 셀로 판단하는 기준을 3 으로 정하고, 빈도가 0 인 경우에는 그 값을 그대로 제공한다고 가정해 보겠습니다. 그렇다면 기저 빈도표에서 비밀보호가 필요한 경우는 빈도가 1 또는 2 인 경우로 해당 빈도에 대한 랜덤 라운딩(random rounding)이라는 비교적 간단한 방법으로 노출 위험을 제어시킬 수 있습니다.

이에 비해 상위 빈도표에서의 비밀보호 방법은 다소 복잡한 흐름을 가지고 있습니다. 기저 빈도표에서 작은 빈도의 값을 집계하여 중앙값 조정이라는 단계를 기본적으로 거치게 됩니다. 이 과정을 통해 노출 위험을 제어할 수 있게 되는데, 일부 경우에는 노출 위험이 여전히 남아 있기도 합니다. 그래서 중앙값 조정된 값의 노출 위험 정도를 추가적인 검토하게 됩니다. 추가로 검토된 노출 위험에 따라 3 가지 처리 절차, 중앙값 조정을 거친 빈도 자체를 B값으로 바꾸거나, 중앙값 조정을 거친 빈도에 B만큼 더하거나 빼는 절차, 중에서 1 가지를 취하게 됩니다.

이러한 절차를 거쳐 제공되는 상이 빈도표는 실제 조사된 빈도의 참 값을 정확하게는 유추할 수 없게 되고, 노출 위험이 $1/B$ 이하로 제어됩니다. 또한 실제 조사된 빈도의 참 값과 차이가 3 이하로만 발생하게끔 설계되었습니다.

이러한 빈도표에 대한 비밀보호 알고리즘의 상세한 적용 절차는 다음 페이지에서 예시를 통해 확인할 수 있습니다.

그리고 BSCA 알고리즘 적용 후에는 사용자에게 적절한 정보를 제공한다는 측면에서 다음의 사항을 공지해야 할 것입니다.

[공지] 공개된 자료를 통해 발생할 수 있는 개인정보 노출을 방지하기 위해 다음과 같이 조치하였습니다.

제공 받은 자료의 빈도(인구수)	실제 조사된 빈도(인구수)
3명	1, 2, 3명 中 1개 값
0명	0, 1, 2명 中 1개 값

3.3 기저 빈도표에서의 비밀보호

조사된 기저 빈도표에서 빈도(인구 수)가 1명 또는 2명인 경우 노출의 위험이 있기 때문에 조사된 빈도(인구 수)의 값을 그대로 제공하는 것은 문제가 될 수 있습니다. 이 경우 조사된 빈도(인구 수) 1명 또는 2명을 0명 또는 3명으로 바꾸어 제공해야만 노출의 위험이 제어되며, 그 가능성은 다음과 같은 확률에 의해 결정됩니다.[11]

11 이러한 과정을 랜덤 라운딩(random rounding, 임의 반올림)이라고 함

〈표 9〉 조사된 빈도(인구 수)의 반올림 과정

조사된 빈도		반올림한 빈도	조사된 빈도		반올림한 빈도
1	*(2/3 확률)* →	0	2	*(1/3 확률)* →	0
	(1/3 확률) →	3		*(2/3 확률)* →	3

조사된 빈도(인구 수)는 반올림 과정을 거쳐 〈표 10〉과 같은 형태로 사용자에게 제공될 수 있습니다. 조사된 기저 빈도표에서 1명 또는 2명이었던 값이 모두 0명 또는 3명으로 바뀌어 노출 위험이 제어된 것으로 판단할 수 있습니다. 또, 빈도(인구 수)가 0명이거나 3명 이상인 경우에는 조사된 빈도(인구 수)를 그대로 제공하기 때문에 자료 사용자 입장에서 거부감이 크지 않을 수 있습니다.

〈표 10〉 조사된 기저 빈도표와 반올림한 기저 빈도표 비교

조사된 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	1	0	0	0	13
	아파트	18	3	2	0	36
	연립주택	4	4	1	7	22
	다세대	5	8	0	18	12
	비주거용	1	2	0	4	15
여	단독주택	4	5	8	3	1
	아파트	9	4	9	2	1
	연립주택	5	10	14	5	1
	다세대	10	6	7	8	1
	비주거용	4	5	11	11	1

반올림한 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	0	0	0	0	13
	아파트	18	3	3	0	36
	연립주택	4	4	3	7	22
	다세대	5	8	0	18	12
	비주거용	0	3	0	4	15
여	단독주택	4	5	8	3	3
	아파트	9	4	9	3	3
	연립주택	5	10	14	5	3
	다세대	10	6	7	8	3
	비주거용	4	5	11	11	3

지금부터 자료 제공자는 사용자의 요청이 있는 경우 **반올림한 기저 빈도표**를 **제공**하는 것으로 하겠습니다.

3.4 상위 빈도표에서의 비밀보호

상위 빈도표란 기저 빈도표에 비해 항목의 수가 줄어들거나, 지역 수준이 넓어지는 경우를 의미합니다. 사용자가 상위 빈도표를 요청하는 경우에는 <표 10>의 기저 빈도표에서 항목별로 또는 지역 단위별로 집계하여 제공하게 됩니다.

예를 들어 사용자가 성별에 따른 빈도(인구 수)를 요청하였다고 가정하겠습니다.[12]

이 경우에는 주택 유형 별로 세분된 빈도(인구 수)를 남·여별로 집계하여 제공하면 되는데, 구체적으로 보면 1 번째 집계구에 거주하는 성별 빈도(인구 수)는 다음과 같습니다.

12 항목의 수가 줄어드는 상위 빈도표에 해당

<표 11> 상위 빈도표 작성 예시

조사된 기저 빈도		1st 집계구
성별	주택 유형	
남	단독주택	1
	아파트	18
	연립주택	4
	다세대	5
	비주거용	1
여	단독주택	4
	아파트	9
	연립주택	5
	다세대	10
	비주거용	4

조사된 상위 빈도	1st 집계구
성별	
남	29
여	32

위와 동일한 과정을 거쳐 구해진 <표 10>에 해당하는 상위 빈도표는 <표 12>, <표 13>과 같습니다.

<표 12> 조사된 기저 빈도표를 집계한 상위 빈도표

조사된 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	1	0	0	0	13
	아파트	18	3	2	0	36
	연립주택	4	4	1	7	22
	다세대	5	8	0	18	12
	비주거용	1	2	0	4	15
여	단독주택	4	5	8	3	1
	아파트	9	4	9	2	1
	연립주택	5	10	14	5	1
	다세대	10	6	7	8	1
	비주거용	4	5	11	11	1

조사된 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
➡ 성별	남	29	17	3	29	98
	여	32	30	49	29	5

〈표 13〉 반올림한 기저 빈도표를 집계한 상위 빈도표

반올림한 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	0	0	0	0	13
	아파트	18	3	3	0	36
	연립주택	4	4	3	7	22
	다세대	5	8	0	18	12
	비주거용	0	3	0	4	15
여	단독주택	4	5	8	3	3
	아파트	9	4	9	3	3
	연립주택	5	10	14	5	3
	다세대	10	6	7	8	3
	비주거용	4	5	11	11	3

➠

반올림한 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	27	18	6	29	98
	여	32	30	49	30	15

3.4.1 제공 가능한 상위 빈도표에 대한 논의

자료 제공자는 〈표 12〉와 〈표 13〉 중에서 어떤 것을 상위 빈도표로 제공해야 할지에 대해 고민하게 됩니다. 이 때 상기해야 할 사항은 노출 위험을 줄이기 위해 반올림한 기저 빈도표를 제공하기로 했다는 사실입니다. 만약 반올림한 기저 빈도표와 조사된 상위 빈도표를 함께 제공한다면 〈표 14〉와 같이 됩니다.

〈표 14〉 반올림한 기저 빈도표와 조사된 상위 빈도표를 함께 제공할 경우

반올림한 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	0	0	0	0	13
	아파트	18	3	3	0	36
	연립주택	4	4	3	7	22
	다세대	5	8	0	18	12
	비주거용	0	3	0	4	15
여	단독주택	4	5	8	3	3
	아파트	9	4	9	3	3
	연립주택	5	10	14	5	3
	다세대	10	6	7	8	3
	비주거용	4	5	11	11	3

조사된 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	29	17	3	29	98
	여	32	30	49	29	5

〈표 15〉 반올림한 기저 빈도표와 조사된 상위 빈도표를 함께 제공할 때의 노출 위험

반올림한 기저 빈도		5th 집계구
성별	주택 유형	
여	단독주택	3
	아파트	3
	연립주택	3
	다세대	3
	비주거용	3

조사된 상위 빈도		5th 집계구
성별	여	5

〈표 15〉를 제공받은 사용자는 일차적으로 자료의 집계 값이 일치하지 않다는 판단을 하게 됩니다. 또, 5번 집계구에서 특정 주택 유형에 거주하는 여자 빈도(인구 수)가 1명 또는 2명임이 노출되게 됩니다. 왜냐하면 자료 제공자가 다음과 같이 공지하였기에 자료 사용자는 쉽게 유추가 가능해 집니다.

[공지] 공개된 자료를 통해 발생할 수 있는 개인정보 노출을 방지하기 위해 다음과 같이 조치하였습니다.

제공 받은 자료의 빈도(인구수)	실제 조사된 빈도(인구수)
3명	1, 2, 3명 中 1개 값
0명	0, 1, 2명 中 1개 값

따라서 자료 제공자는 노출의 위험이 존재하는 조사된 상위 빈도표는 제공 불가라는 판단에 이르게 되고, 반올림한 기저 빈도를 집계한 상위 빈도표의 제공에 대해 고민합니다.

여기서 자료 제공자가 고려해야 할 또 다른 문제로 정보손실(information loss)이 있습니다. 정보손실이란 쉽게 말해 실제 조사된 값과 제공하는 값의 차이를 의미하는 것으로, 자료 제공자는 정보손실을 최소화하기 위한 노력이 필요합니다.

실제로 〈표 12〉의 조사된 상위 빈도와 〈표 13〉의 반올림한 상위 빈도는 차이가 있을까요?

〈표 16〉 조사된 상위 빈도와 반올림한 상위 빈도의 차이

정보손실*		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	2	1	3	0	0
	여	0	0	0	1	10

* 정보손실 = | 조사된 기저 빈도를 집계한 상위 빈도 - 반올림한 기저 빈도를 집계한 상위 빈도 |

〈표 16〉에서 보듯이 5번 집계구의 여자 빈도(인구 수)에서는 정보손실이 무려 10명만큼 발생하여, 사용자에게 유용한 정보를 제공한다고 할 수 없습니다. 자료 제공자는 상위 빈도표를 제공하는 경우, 정보손실을 보다 최소화하기 위해 노력할 필요가 있게 되었습니다.

자료 제공자는 상위 빈도표 제공 시 노출의 위험을 줄이면서 동시에 정보손실을 최소화하기 위해 어떤 특별한 조치를 취하려 합니다. 지금부터 상위 빈도표에 대한 어떤 특별한 조치란 무엇인지 자세히 알아보겠습니다.

3.4.2 중앙값(median) 조정을 통한 비밀보호

[13] 작은 빈도를 갖는 셀로 판단하는 기준

노출 위험이 제어된 상위 빈도표를 만들기 위한 첫 번째 단계는 조사된 기저 빈도표에서 빈도(인구 수)의 값이 3^{13} 명 이하인 경우만 먼저 집계합니다. 그리고 집계된 값을 중앙값으로 조정한 이후, 3 명보다 큰 빈도(인구 수)를 더하여 제공하는 방식입니다.

이때 작은 빈도를 집계한 값에 대한 중앙값 조정 과정은 이렇습니다. 만약 집계된 값이 1, 2, 3 명 중에 하나인 경우에는 집계된 값을 2 명으로 바꾸고, 집계된 값이 4, 5, 6 명인 경우에는 5 명으로 바꾸는 것입니다. 다시 말해 집계된 값을 크기가 3^{13} 인 구간(interval)으로 나누어, 해당 구간의 가운데 있는 값(중앙값)으로 조정을 하게 됩니다. 단, 0(zero)명은 그대로 0 명으로 제공합니다. 이 과정을 거침으로써 노출 위험이 제어 됩니다.

〈표 17〉 조사된 기저 빈도표에서 3 이하인 빈도(인구 수) 집계 후 중앙값으로 조정

조사된 작은 빈도 집계	0	1,2,3	4,5,6	7,8,9	10,11,12	13,14,15	…
↓	↓	↓	↓	↓	↓	↓	
중앙값 조정 빈도	**0**	**2**	**5**	**8**	**11**	**14**	**…**

앞서 예로 들었던 조사된 기저 빈도표에서 중앙값 조정은 어떻게 이루어지는지 살펴보겠습니다. 조사된 기저 빈도표에서 빈도(인구 수)가 3 명 이하인 경우만 먼저 집계합니다.

〈표 18〉 조사된 기저 빈도표에서 3 이하인 빈도(음영으로 표시됨)

조사된 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	1	0	0	0	13
	아파트	18	3	2	0	36
	연립주택	4	4	1	7	22
	다세대	5	8	0	18	12
	비주거용	1	2	0	4	15
여	단독주택	4	5	8	3	1
	아파트	9	4	9	2	1
	연립주택	5	10	14	5	1
	다세대	10	6	7	8	1
	비주거용	4	5	11	11	1

➡

조사된 작은 빈도 집계		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	2	5	3	0	-
	여	-	-	-	5	5

조사된 작은 빈도 집계값에서 5 번 집계구의 남자와 1, 2, 3 번 집계구의 여자는 빈도(인구 수)가 3 명 이하인 경우가 없으므로 ' - '로 표시하였습니다. 〈표 18〉의 집계된 빈도를 중앙값으로 조정합니다. 그리고 3 명보다 큰 빈도(인구 수)를 합산하여 상위 빈도표를 구해보면 다음과 같습니다.

〈표 19〉 집계된 작은 빈도의 중앙값 조정

중앙값 조정 후 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	2	5	2	0	-
	여	-	-	-	5	5

〈표 20〉 중앙값 조정 후 집계된 상위 빈도표

중앙값 조정 후 집계된 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	29[14]	17[15]	2	29[16]	98
	여	32	30	49	29[17]	5

(비교)

조사된 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	29	17	3	29	98
	여	32	30	49	29	5

14 29 = 2+18+4+5

15 17 = 5+4+8

16 29 = 0+7+18+4

17 29 = 5+5+8+11

〈표 20〉의 중앙값 조정된 상위 빈도표를 사용자에게 제공한다고 가정해 보면, 조사된 상위 빈도와 거의 동일한 값을 제공하는 것이므로 자료 사용자 입장에서는 만족스러울 수 있습니다. 그러나 중앙값 조정에도 불구하고 노출의 위험이 여전히 남아 있는 경우도 있습니다.

먼저 중앙값 조정만으로 노출의 위험이 제어된 경우로, 2번 집계구의 남자 빈도(인구 수)를 살펴보겠습니다.

〈표 21〉 중앙값 조정으로 노출 위험이 제어된 경우

조사된 기저 빈도		2nd 집계구
성별	주택 유형	
남	단독주택	0
	아파트	3
	연립주택	4
	다세대	8
	비주거용	2

반올림한 기저 빈도		2nd 집계구
성별	주택 유형	
남	단독주택	0
	아파트	3
	연립주택	4
	다세대	8
	비주거용	3

중앙값 조정 후 집계된 상위 빈도		2nd 집계구
성별	남	17

18 반올림한 기저 빈도표에서 3이하인 경우가 여러 개 있음

사용자는 반올림한 기저 빈도표를 집계하면 남자의 빈도(인구 수)가 18명이 되고, 제공되는 상위 빈도(인구 수)는 17명이 되어 1명 차이가 난다는 것을 알 수 있습니다. 그러나 그 차이 값이 어디에서 발생했는지는 특정 지을 수 없기 때문에[18] 노출의 위험이 줄었다고 할 수 있습니다.

3.4.3 추가 조치를 통한 노출 위험 제어와 정보 손실 축소

중앙값 조정만으로 언제나 노출 위험이 제어되는 것은 아니기에, 자료 제공자는 추가적인 조치를 하게 됩니다. 이 단계는 중앙값 조정 이후에도 노출 위험이 남아있거나, 정보손실이 크다고 판단되는 경우에만 적용되며, 중앙값 조정을 취한 값에 대하여 적용합니다.

이 단계는 다음과 같이 3 가지 유형으로 나뉘게 됩니다.

① 중앙값으로 조정한 빈도(인구 수)를 3^{19} 명으로 바꾸는 경우

② 중앙값으로 조정한 빈도(인구 수)에 3^{19} 명을 더하는 경우

③ 중앙값으로 조정한 빈도(인구 수)에 3^{19} 명을 빼는 경우

19 작은 빈도를 갖는 셀로 판단하는 기준

각 유형에 대한 판단은 다소 복잡한 기준에 의해 결정되기 때문에, 이것은 BSCA 알고리즘에 의해 컴퓨터가 판단하도록 하였습니다. 각 유형에서 취하는 조치들의 공통점이라고 한다면 숫자 ' 3 '을 활용하였다는 것입니다. 앞서 소개된 「2. 빈도표의 노출 위험」에서 언급되었듯이 빈도(인구 수)가 1 명 또는 2 명인 경우에는 노출의 위험이 있는 것으로 판단하였습니다. 이에 비해 빈도(인구 수)가 3 명인 경우는 노출의 위험이 비교적 낮은 것으로 판단할 수 있기에 숫자 '3'을 활용하였습니다[20]. 이러한 내용과 관련된 BSCA 알고리즘에 대한 보다 상세한 이론적 증명은 통계개발원 보고서를 참고하실 수 있습니다.

20 자료 제공자의 판단에 따라 조정 가능 (예, $B = 5$)

본 자료에서는 각 유형별 예시를 통해 사용자가 추측할 수 있는 정보는 무엇이고, 해당 경우마다 어떠한 조치를 취하게 되는지에 대해 설명하는 것에 주안점을 두었습니다. 이후 소개되는 내용의 이해를 위해 앞서 언급된 기저 빈도표를 제시하면 다음과 같습니다.

〈표 22〉 조사된 기저 빈도표와 반올림한 기저 빈도표

조사된 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	1	0	0	0	13
	아파트	18	3	2	0	36
	연립주택	4	4	1	7	22
	다세대	5	8	0	18	12
	비주거용	1	2	0	4	15
여	단독주택	4	5	8	3	1
	아파트	9	4	9	2	1
	연립주택	5	10	14	5	1
	다세대	10	6	7	8	1
	비주거용	4	5	11	11	1

반올림한 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	0	0	0	0	13
	아파트	18	3	3	0	36
	연립주택	4	4	3	7	22
	다세대	5	8	0	18	12
	비주거용	0	3	0	4	15
여	단독주택	4	5	8	3	3
	아파트	9	4	9	3	3
	연립주택	5	10	14	5	3
	다세대	10	6	7	8	3
	비주거용	4	5	11	11	3

① (예시 1)중앙값으로 조정한 빈도(인구 수)를 3 명으로 바꾸는 경우

1 번 집계구의 남자 빈도(인구 수)를 고려해 보겠습니다. 여기서 상기해야 할 것은 자료 제공자는 반올림한 기저 빈도표를 사용자에게 제공하기로 했다는 점입니다.

〈표 23〉 중앙값 조정 후에도 노출 위험이 남아 있는 경우(예시1)

조사된 기저 빈도		1st 집계구
성별	주택 유형	
남	단독주택	1
	아파트	18
	연립주택	4
	다세대	5
	비주거용	1

반올림한 기저 빈도		1st 집계구
성별	주택 유형	
남	단독주택	0
	아파트	18
	연립주택	4
	다세대	5
	비주거용	0

중앙값 조정 후 집계된 상위 빈도		1st 집계구
성별	남	29

만약 상위 빈도를 29 명으로 제공할 경우, 사용자는 단독주택과 비주거용 건물에 거주하는 남자 빈도(인구 수)가 합산하여 2 명이 된다는 사실을 알게 되어, 노출의 위험이 남아있게 됩니다. 따라서 이 경우에는 중앙값 조정 후 빈도(인구 수)인 2 명을 **3 명**으로 바꾼 후 상위 빈도를 집계하여 제공하면 노출의 위험을 제어 시킬 수 있습니다.

〈표 24〉 (예시1)에 대한 사용자의 추측

반올림한 기저 빈도		1st 집계구	제공 정보	사용자 추측
성별	주택 유형			
남	단독주택	**0**	(0,1,2) 中 하나	· 기저빈도 **0**+**0**=2 이다. · (**0**,**0**)의 참(True) 값 후보? (1, 1), (0, 2), (2, 0) → 참 값 노출 위험 1/3 존재
	아파트	18	18+4+5=27 ➡	
	연립주택	4		
	다세대	5		
	비주거용	**0**	(0,1,2) 中 하나	

〈표 25〉 노출 위험이 제어된 상위 빈도 집계 과정(예시1)

중앙값 조정 후 빈도		1st 집계구		노출 위험 제어된 빈도		1st 집계구		노출 위험 제어된 상위 빈도		1st 집계구
성별	남	2	➡	성별	남	3	➡	성별	남	30[21]

[21] 30 = 3+18+4+5

〈표 26〉 노출 위험이 제어된 상위 빈도에 대한 사용자의 추측

반올림한 기저 빈도		1st 집계구	제공 정보	사용자 추측
성별	주택 유형			
남	단독주택	**0**	(0,1,2) 中 하나	· 기저빈도 **0**+**0**=3 이다. · (**0**,**0**)에 대한 참(True) 값 후보? (1, 2), (2, 1) → 참 값 노출 위험 제어
	아파트	18	18+4+5=27 ➡	
	연립주택	4		
	다세대	5		
	비주거용	**0**	(0,1,2) 中 하나	

② (예시 2) 중앙값 조정한 빈도(인구 수)에 3 명을 더하는 경우

이번에는 3 번 집계구의 남자 빈도(인구 수)를 고려해 보겠습니다.

〈표 27〉 중앙값 조정 후 정보손실이 큰 경우(예시2)

조사된 기저 빈도		3rd 집계구
성별	주택 유형	
남	단독주택	0
	아파트	2
	연립주택	1
	다세대	0
	비주거용	0

반올림한 기저 빈도		3rd 집계구
성별	주택 유형	
남	단독주택	0
	아파트	3
	연립주택	3
	다세대	0
	비주거용	0

중앙값 조정 후 집계된 상위 빈도		3rd 집계구
성별	남	2

사용자는 제공받은 반올림한 기저 빈도를 집계하여 남자의 빈도(인구 수)가 6 명임을 알 수 있게 되고, 이 경우 사용자가 기저 빈도에 대해 추론할 수 있는 정보는 다음과 같습니다.

〈표 28〉 (예시2)에 대한 사용자의 추측

반올림한 기저 빈도		3rd 집계구	제공 정보
성별	주택 유형		
남	단독주택	**0**	(0,1,2) 中 하나
	아파트	**3**	(1,2,3) 中 하나
	연립주택	**3**	(1,2,3) 中 하나
	다세대	**0**	(0,1,2) 中 하나
	비주거용	**0**	(0,1,2) 中 하나

사용자 추측

- 기저빈도 **0+3+3+0+0**=2 이다.
- 참(True) 값 후보?
  (0, 1, 1, 0, 0)
  → 참 값 노출 위험 제어
  but 정보손실 증가

상위 빈도를 2 명으로 제공할 경우 노출 위험은 제어되었으나, 사용자 입장에서는 기저 빈도와 상위 빈도의 차이가 4 명 발생하게 됩니다. 그래서 이러한 경우에 자료 제공자는 중앙값 조정 후 빈도인 2 명에 3 명을 더해준 **5 명**으로 제공하게 되면 노출 위험을 제어시키면서 기저 빈도와 상위 빈도의 차이를 줄일 수 있게 됩니다. 이와 동일한 작업은 5 번 집계구의 여자 빈도(인구 수)에도 적용 됩니다.

〈표 29〉 노출 위험이 제어된 상위 빈도 집계 과정(예시2)

중앙값 조정 후 빈도		3rd 집계구
성별	남	2

➡

노출 위험 제어된 빈도		3rd 집계구
성별	남	5

➡

노출 위험이 제어된 상위 빈도		3rd 집계구
성별	남	5^{22}

$\overset{22}{5} = 0+5+0+0$

〈표 30〉 노출 위험이 제어된 상위 빈도에 대한 사용자의 추측(예시2)

반올림한 기저 빈도		3rd 집계구	제공 정보
성별	주택 유형		
남	단독주택	**0**	(0,1,2) 中 하나
	아파트	**3**	(1,2,3) 中 하나
	연립주택	**3**	(1,2,3) 中 하나
	다세대	**0**	(0,1,2) 中 하나
	비주거용	**0**	(0,1,2) 中 하나

사용자 추측

- 기저빈도 **0+3+3+0+0**=5 이다.
- 참(True) 값 후보?
  (1,1,1,1,1), (0,2,2,0,1), …
  → 참 값 노출 위험 제어
  and 빈도 차이 축소

③ (예시 3) 중앙값 조정한 빈도(인구 수)에 3 명을 빼는 경우

마지막으로 4 번 집계구의 여자 빈도(인구 수)를 살펴보겠습니다.

〈표 31〉 중앙값 조정 후에도 노출 위험이 남아 있는 경우(예시3)

조사된 기저 빈도		4th 집계구
성별	주택 유형	
여	단독주택	3
	아파트	2
	연립주택	5
	다세대	8
	비주거용	11

반올림한 기저 빈도		4th 집계구
성별	주택 유형	
여	단독주택	3
	아파트	3
	연립주택	5
	다세대	8
	비주거용	11

중앙값 조정 후 집계된 상위 빈도		4th 집계구
성별	여	29

만약 상위 빈도를 29 명으로 제공할 경우, 사용자는 단독주택과 아파트에 거주하는 여자 빈도(인구 수)가 합산하여 5 명이 된다는 사실을 알게 됩니다. 이때 사용자가 기저 빈도의 참 값에 대해 추측할 수 있는 경우의 수는 2 가지이며, 그 중 1 가지는 실제 조사된 기저 빈도와 일치하게 됩니다. 따라서 이 경우에도 중앙값 조정 후 노출 위험이 남아있게 됩니다.

〈표 32〉 (예시3)에 대한 사용자의 추측

반올림한 기저 빈도		4th 집계구	제공 정보
성별	주택 유형		
여	단독주택	**3**	(1,2,3) 中 하나
	아파트	**3**	(1,2,3) 中 하나
	연립주택	5	5+8+11=24
	다세대	8	
	비주거용	11	

➡

사용자 추측

· 기저빈도 **3+3**=5 이다.
· (**3**,**3**)에 대한 참(True) 값 후보?
(2, 3), (3, 2)
→ 참 값 노출 위험 1/2 존재

따라서 자료 제공자는 중앙값 조정 후 빈도 5 명에서 3 명을 뺀 2 **명**으로 바꾼 후 상위 빈도를 집계하여 제공하도록 하였습니다.

〈표 33〉 노출 위험이 제어된 상위 빈도 집계 과정(예시3)

중앙값 조정 후 빈도		4th 집계구
성별	여	5

➡

노출 위험 제어된 빈도		4th 집계구
성별	여	2

➡

노출 위험이 제어된 상위 빈도		4th 집계구
성별	여	26[23]

23
26 = 2+5+8+11

〈표 34〉 노출 위험이 제어된 상위 빈도에 대한 사용자의 추측(예시3)

반올림한 기저 빈도		4th 집계구	제공 정보
성별	주택 유형		
여	단독주택	**3**	(1,2,3) 中 하나
	아파트	**3**	(1,2,3) 中 하나
	연립주택	5	5+8+11=24
	다세대	8	
	비주거용	11	

➡

사용자 추측

· 기저빈도 **3+3**=2 이다.
· (**3**,**3**)에 대한 참(True) 값 후보?
(1, 1)
→ 참 값 노출 위험 제어

이상의 언급된 단계를 거쳐 최종적으로 제공되는 상위 빈도표는 다음과 같습니다.

〈표 35〉 노출 위험이 제어된 상위 빈도표

노출 위험이 제어된 상위 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	30	17	5	29	98
	여	32	30	49	26	8

또 조사된 기저 빈도표를 단순 집계한 상위 빈도표와 노출 위험이 제어된 상위 빈도표에서의 정보손실은 다음과 같으며, 노출 위험 제어에 따른 정보손실은 3명 이내임을 확인할 수 있습니다.

〈표 36〉 노출 위험 제어에 따른 정보손실

정보손실*		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	남	1	0	2	0	0
	여	0	0	0	3	3

* 정보손실= | 조사된 기저 빈도를 집계한 상위 빈도 - 노출 위험이 제어된 상위 빈도 |= | 참 값 빈도 - 제공 빈도 |

3.5 현행 자료 제공 방식과의 비교

현행 자료 제공 방식과 비밀보호 알고리즘 적용 후 제공되는 자료의 차이를 확인해 보겠습니다. 현행 SGIS 대화형 통계 지도에서는 기저 빈도표와 상위 빈도표에서 5 미만의 값에 대해 N/A 표시하여 제공하고 있습니다. 앞서 예시로 제시된 자료에서 현행 제공 방식의 빈도표를 구해 보면 다음과 같습니다.

〈표 37〉 조사된 빈도표와 현행 제공 방식의 빈도표 비교

조사된 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	1	0	0	0	13
	아파트	18	3	2	0	36
	연립주택	4	4	1	7	22
	다세대	5	8	0	18	12
	비주거용	1	2	0	4	15
여	단독주택	4	5	8	3	1
	아파트	9	4	9	2	1
	연립주택	5	10	14	5	1
	다세대	10	6	7	8	1
	비주거용	4	5	11	11	1

* 조사된 빈도가 1 또는 2인 경우는 음영으로 표시됨.

현행 제공 기저 빈도		1st 집계구	2nd 집계구	3rd 집계구	4th 집계구	5th 집계구
성별	주택 유형					
남	단독주택	N/A	N/A	N/A	N/A	13
	아파트	18	N/A	N/A	N/A	36
	연립주택	N/A	N/A	N/A	7	22
	다세대	5	8	N/A	18	12
	비주거용	N/A	N/A	N/A	N/A	15
여	단독주택	N/A	5	8	N/A	N/A
	아파트	9	N/A	9	N/A	N/A
	연립주택	5	10	14	5	N/A
	다세대	10	6	7	8	N/A
	비주거용	N/A	5	11	11	N/A

* 제공 빈도가 5 미만인 경우는 음영으로 표시됨.

위 빈도표에서 보듯이 현행 제공 방식의 기저 빈도표에서는 전체의 50%에 해당하는 25개 셀에 대해 N/A 로 제공되고 있습니다. 현행 제공 방식 역시 비밀보호를 위한 불가피한 조치였음 감안한다 하더라도, 사용자가 원하는 정보를 충분히 제공한다고 보기에는 다소 무리가 있는 것을 알 수 있습니다.

따라서 지금까지 다루어진 빈도표에 대한 비밀보호 알고리즘 적용을 통해 정보손실은 3 이하로 제어하면서 노출 위험이 제어된 빈도표를 제공한다면, 사용자에게 보다 많은 정보를 제공하는 하나의 방법이 될 것으로 기대합니다.

Abstract

A study on the Application of Confidentiality to frequency tables provided by SGIS

Younghee Hong, Min-Jeong Park

This study describes a method under which a confidentiality algorithm is applied to the frequency tables provided by the Statistical Geographic Information Service(SGIS). The frequency tables have a hierarchy which is divided into the base frequency table and the upper frequency table. The study explains with examples that the base frequency table uses random rounding whereas the upper frequency table can lead to a decline in disclosure risks through the adjustment of median values and additional measures. In addition, the study utilized Daejeon city applicable data derived from the 2018 Population and Housing Census and calculated the time required to create a frequency table that applies the confidentiality algorithm in units of administrative districts and grids.

Next, it reviewed how the distribution of frequencies had been changed before and after the application of the confidentiality algorithm. After applying the algorithm, the ratio of cells with a frequency of 1 or 2 disappeared from the base frequency table, and the ratio of cells with a frequency of 0 or 3 increased instead. On the other hand, in the upper frequency table, the ratio of cells with a frequency of 3 increased by up to 11.2%p, and the ratio of cells with frequency of 4 or more changed only to around 5%p, showing a fairly similar distribution before and after algorithm application.

Such an application of the algorithm could lead to information loss and it has been confirmed that such a loss was within the range of (±)3. In particular, the closer to the base frequency table, the higher possibility of no information loss. In the upper frequency table, cells with an information loss of 1 topped the list whereas the ratio of cells with an information loss of a maximum of 3 did not excess 0.5% of the total.

If this is understood as the opportunity cost for a reduction in disclosure risks from the frequency table, it is expected to facilitate the provision of a frequency table with more information than the those through which the SGIS currently provides data to users.

Key word : frequency table, confidentiality algorithm, disclosure risk, information loss

집필진

- 홍영희 (통계청 통계개발원 통계방법연구실 주무관)
- 박민정 (통계청 사무관)

연구보고서 2020-14

통계지리정보서비스(SGIS) 제공 자료의 비밀보호 적용 방안

인 쇄	2021년 4월 12일
발 행	2021년 4월 13일
발 행 인	통계개발원장 전영일
발 행 처	통계청 통계개발원 35220 대전광역시 서구 한밭대로 713 TEL.(042)366-7100 FAX.(042)366-7123
홈페이지	http://sri.kostat.go.kr
ISSN(Online)	2733-4120

OPEN 출처표시 상업용금지 변경금지
공공누리 공공저작물 자유이용허락

통계청
통계개발원