발 간 등 록 번 호
11-1240245-000057-14

OPEN 공공누리 출처표시 상업용금지 변경금지 공공저작물 자유이용허락

2021년 연구보고서

통계주제별 통계지리정보 격자 서비스의 공표기준 산정 연구

2022. 4.

http://kostat.go.kr/sri

ISSN 2288-1166(Print)
ISSN 2733-4120(Online)

통계청
통계개발원

연구보고서 2021-02

통계주제별 통계지리정보 격자 서비스의 공표기준 산정 연구

홍영희 · 오영주 · 정재준

Statistics Korea
Statistics Research Institute

발간사

21세기 데이터 혁명시대에 빅데이터와 공공데이터의 홍수 속에서 통계는 더 이상 전문가들만의 전유물이 아닌 국민의 삶에 밀접한 영향을 미치는 중요한 요소로 보편화되고 있습니다. 특히 국가통계는 개인과 기업의 의사결정뿐만 아니라 정부의 정책결정에 필수적인 도구로서 그 중요성이 증대되고 있습니다. 이러한 변화에 따라 통계청은 빅데이터의 활용, 조사자료와 행정자료 간의 연계 등과 같은 통계생산방식의 혁신을 통해서 응답자 부담은 최소화하면서 동시에 보다 정확하고 사용자 친화적인 통계를 만들고자 끊임없이 노력하고 있습니다.

이러한 통계청의 싱크탱크로서 전략적인 연구를 수행하고 있는 통계개발원은 국내의 유일한「국가통계 전문연구기관」입니다. 통계개발원은 2006년에 창립한 이래 기존의 조사통계를 보다 효율적으로 작성하기 위한 각종 기법과 관련된 통계방법론적 연구는 물론 데이터에 기반한 국가정책이 수립될 수 있도록 경제·사회현상에 대한 심층 분석 연구를 강화하고 있습니다. 또한 저출산·고령사회 현상 등으로 인해 대내외적으로 관심이 높아지고 있는 인구집단 및 인구동향에 관한 분석연구 및 인구동태 관련 방법론 연구를 밀도 깊게 수행하고 있습니다. 이러한 연구의 구체적인 결과를 중심으로 통계개발원은「2021년 연구보고서」를 발간하게 되었습니다.

이번「2021년 연구보고서」는 데이터과학 활용통계 승인 개선방안 연구, 통계주제별 통계지리정보 격자 서비스의 공표기준 산정 연구 등「국가통계 방법론」연구, 감염병 위기 대응 경제·사회 측정지표 구축: 코로나19를 중심으로 등「정책통계」관

련 연구, 데이터경제 측정기반 구축을 위한 프레임워크 기초연구, 통계 및 행정자료를 이용한 고용효과 분석방안 등「경제통계」관련 연구, SDGs 이주지표 모니터링을 위한 국내 데이터 가용성 연구 등「SDGs 지표」관련 연구 등을 제시하고 있습니다.

통계개발원은 본 연구보고서가 데이터 이용자의 통계활용에 실제적인 도움이 되고 국가통계 생산자의 통계개발 및 개선에 유용한 자료로 활용될 수 있기를 기대합니다. 앞으로도 통계개발원이 "국가통계 전문연구기관"으로서 대내외적으로 선도적인 역할을 할 수 있도록 독자 여러분의 지속적인 관심을 부탁드립니다. 아울러 실용적이고 품질 높은 연구 결과를 도출하기 위해 최선을 다한 연구진에게 따스한 감사를 전합니다. 끝으로 본 연구에서 제시된 내용 및 결과는 저자 개인의 의견이며, 통계청 또는 통계개발원의 공식적인 견해가 아님을 밝혀둡니다.

2022년 4월

통계청 통계개발원장 직무대리

김 진

목 차

제1장 서론 1

제2장 격자통계 제공 사례 3

1. 고정형 격자(fixed size grid) 체계 4
2. 변동형 격자(varying size grid) 체계 39
3. 격자통계 제공의 시사점 42

제3장 격자 체계별 노출제어와 정보보호 51

1. 지리 데이터(geographic data)의 정보보호 51
2. 고정형 격자의 노출제어와 정보보호 53
3. 변동형 격자의 노출제어와 정보보호 75
4. 격자 체계에 따른 비교 100
5. 격자통계 공표기준에 대한 제언 105

제4장 결론 및 시사점 107

참고문헌 113

부 록 115

Abstract 124

요 약

본 연구에서는 통계지리정보서비스(SGIS)의 격자통계 다양화를 위해 국내·외 사례를 조사하여 벤치마킹하고 소단위 격자통계 제공의 가능성을 검토하였다.

격자통계 제공 사례는 고정형 격자 체계와 변동형 격자 체계로 구분하였다. 대부분의 국가 통계 기관에서 사용하는 고정형 격자는 크기가 규격화되어 있고, 최소 격자의 크기를 100m, 250m로 설정한 경우가 많다. 변동형 격자는 다양한 크기의 격자를 허용하는 방식으로 일부 국가에서 연구 차원으로 진행되었고, 최소 격자의 크기는 125m, 250m, 500m 등으로 다양하다. 데이터 정보보호의 방법으로 고정형 격자에서는 익명성 기준 이하에 해당하면 N/A로 표시하고, 변동형 격자에서는 보다 큰 격자로 집계된다. 이때 익명성 기준 값으로 2 또는 3을 사용하는 경우가 다수 있어, 통계청의 현행 기준을 완화할 필요가 있는 것으로 판단된다.

실제 '다바' 격자의 2019년 인구주택총조사 자료와 전국사업체조사 자료를 입수하여 k-익명성에 기반한 고정형 격자와 변동형 격자 체계를 적용·분석하였다. 두 격자 체계 모두에서 노출위험을 줄이기 위해 제공 항목별로 전반적 재코딩이 필요하고, 정보손실을 고려하여 공표기준을 산정하였다. 그 결과 고정형 격자에서는 N/A 표시 비율 30% 수준을 기준으로 하고, 변동형 격자에서는 기저 격자 비율 70% 수준을 기준으로 결정하였다. 익명성 기준 2 또는 3에서 제공 가능한 소 격자는 항목에 따라 다소 차이가 있고, 동일한 조건에서 검토하면 고정형 체계보다 변동형 체계에서 더 작은 격자로 제공이 가능하다.

마지막으로 본 연구는 향후 다양한 격자통계 제공을 위한 기초자료로 활용될 수 있을 것으로 기대한다.

주요 용어 : 고정형 격자(fixed size grid), 변동형 격자(varying size grid), 전반적 재코딩(global recoding), 노출위험(disclosure)

제 1 장

서 론

□ 통계지리정보서비스(SGIS)에 대한 이용자들의 수요가 증가함에 따라 제공 자료의 다양성 및 활용성 제고에 대한 요구가 증가하고 있다.

○ 통계지리정보서비스를 통한 자료 제공 건수는 2015년 약 250만 건에서 2018년 약 500만 건으로 2배 이상 증가하였고, 2020년에는 약 630만 건 이상을 기록하였다.

○ 통계지리정보서비스 내 격자통계 서비스[1]는 경계의 변동이 없어 표준화된 크기로 다(多)출처 정보를 공간상에서 상호 연계할 수 있고, 시계열 분석에도 용이하기 때문에 제공 자료에 대한 다양성 및 활용성 제고 요구를 충족시킬 수 있다.

○ 그러나 100m 등의 소(small) 단위 격자에서는 통계 자료(예: 인구 수, 사업체 수 등)가 매우 작은 값을 가지는 경우가 많기 때문에, 정보의 노출위험을 최소화시키기 위해 작은 값을 가지는 통계 값은 N/A(not available)로 표시하여 서비스를 제공한다.

□ 본 연구에서는 향후 SGIS 격자 서비스의 확대를 위해 소 격자 단위의 제공 가능성을 검토하고자 한다.

○ 현재 통계청에서 격자통계로 제공하는 자료는 인구주택총조사, 전국사업체조사 자료가 있으며, 격자의 크기는 100m, 1㎞, 10㎞, 100㎞ 단위이다.

○ 반면 유럽의 일부 국가에서는 인구 또는 사업체 자료를 1㎞ 이하의 격자인 125m, 250m 등의 단위에 해당하는 통계로 제공하는 것으로 알려져 있다.

○ 이에 1km 이하의 소 격자에서 발생하는 노출위험과 정보손실을 검토하여 제공 가능한 격자 단위를 알아보고자 한다.

○ 또 현행 익명성 기준 값으로 인구 자료의 경우 5, 사업체 자료의 경우 3을 적용하고 있는데, 이 기준을 완화하여 적용이 가능한지에 대해 검토하고자 한다.

---

1) 동일 간격으로 구획된 바둑판형의 격자(grid) 영역에 통계 자료를 집계하여 제공하는 서비스

□ 이러한 연구 목적을 달성하기 위해 격자통계 제공에 관한 국내·외 사례를 조사하고, 통계주제별로 격자통계를 산출하여 노출위험을 분석해 본다.

○ 우선 국내에서 격자통계를 제공하고 있는 국토지리정보원의 사례를 조사하여 벤치마킹하고자 한다.

○ 국외 사례 연구로 미국, 일본, 유럽 등의 격자통계 제공 현황을 조사하여 서비스 격자의 크기와 정보보호 방안을 검토하고자 한다.

○ 또 현행 서비스 DB 자료를 입수하여, 소단위 격자통계를 산출하고 노출위험을 분석하여 공표기준을 산정하고자 한다.

○ 마지막 결론에서 향후에 격자통계 생산이 가능할 것으로 예상되는 행정자료에 대해 간략히 언급하고자 한다.

○ 또 제공 항목의 다양화 측면에서 향후 격자 서비스로 제공이 가능한 자료에 대해 간략히 검토하고자 한다.

□ 본 연구는 향후 소단위 격자통계 제공을 위한 기초자료로 활용될 수 있을 것으로 기대한다.

○ (실무적 차원) 다양한 국내·외 사례를 수집하고 검토함으로써 소단위 격자통계 제공에 필요한 백업(back up) 자료로 활용 가능하다.

○ (연구적 차원) 빈도표(frequency table) 형태로 제공하는 자료의 노출 제어에 관련하여, 2020년 연구과제(통계지리정보서비스 제공 자료의 비밀보호 적용 방안)와 함께 2 Track 검토가 완성된다.

제 2 장

격자통계 제공 사례

□ 격자통계 제공 사례를 고정형 격자(fixed size grid) 체계와 변동형 격자(varying size grid) 체계라는 두 가지 유형으로 분류하였다.

○ 고정형 격자 체계는 격자의 크기를 100km, 10km 등 사전에 규격화함으로써 해당 격자별 통계를 제공하는 체계를 의미하는 것으로 정하였다. 대부분의 국가통계 기관들은 고정형 격자 체계를 이용하여 통계를 제공한다.

○ 변동형 격자 체계는 사전에 격자 크기의 기저(base)만 결정하고, 자료 제공을 위한 격자의 크기를 다양하게 바꾸는 것을 허용하는 것으로 정하였다. 해외에서는 일부 국가의 연구에서 시도된 사례를 찾아볼 수 있었다.

□ 고정형 격자 체계를 국가적 차원에서 사용하고 있는 경우는 한국, 유럽 대륙, 미국 등을 찾아볼 수 있다.

○ 대표적인 사례로 한국의 국토지리정보원과 통계청, 유럽연합통계청, 핀란드, 오스트리아, 네덜란드, 노르웨이, 독일연방, 에스토니아, 일본, 영국, 미국 등을 조사하였다.

○ 국토 면적이 넓은 러시아, 캐나다, 중국, 호주 등에서는 연구 차원에서 격자통계를 제작하는 수준으로 이를 국가 공식 통계로 보기에는 무리가 있었다.

○ 해외 국가별 격자통계 서비스를 제공하는 자세한 웹사이트 주소(URL)는 부록을 참조하면 된다.

□ 변동형 격자 체계를 연구한 사례는 스페인 북동부 지방인 카탈로니아 통계청(Idescat)으로 2014년 인구등록 데이터에 적용하여 온라인 맵에서 통계 데이터를 제공하기 위해 시도하였다.

□ 격자통계는 인구밀도와 토지 사용의 집약도가 높은 나라에서 특히 많이 사용된다.

○ 밀도가 높은 영역은 격자의 크기가 작아도 집계할 수 있는 데이터가 충분하여

보다 유용한 정보를 제공할 수 있기 때문으로 판단된다.

1. 고정형 격자(fixed size grid) 체계

1.1 한국 국토지리정보원 국토통계지도

□ 한국의 국토지리정보원은 2014년 이후 피라미드 구조의 격자 체계를 도입하여 격자 기반 통계지도인 '국토통계지도'를 제작하고 있다.

○ 국토지리정보원은 국토정보플랫폼(map.ngii.go.kr)을 통해 국토통계지도를 온라인 상에서 서비스하며, 국토통계지도의 격자 크기는 현재 6종(100km, 10km, 1km, 500m, 250m, 100m)으로 구분·제작된다.

○ 국토조사와 관련한 근거 법령, 격자의 기준이 되는 국가지점번호 부여 방법, 국토통계지도의 다운로드 서비스 단위 및 파일 형태에 관한 자세한 내용은 부록을 참조할 수 있다.

가. 국토통계지도의 격자 구성

□ 국가지점번호는 국토 및 이와 인접한 해양을 격자형으로 구획한 지점(10×10m)마다 부여한 위치 표시 번호로, 한글 2자와 숫자 8자로 표기된다. 국토통계지도에서 사용되는 격자는 국가지점번호를 변형하여 고안·구성된다.

○ 국토통계지도에서 사용되는 격자는 국가지점번호 내 숫자의 자릿수*를 통해 다양한 크기로 생성될 수 있다.

* 0-자릿수: 100km, 2-자릿수: 10km, 4-자릿수: 1km, 6-자릿수: 100m, 8-자릿수: 10m

○ 그러나 토지가 집약적으로 사용되는 대도시 권역의 소지역에서는 기존 격자만 가지고 정보를 제공할 경우 여러 가지 한계점이 발생한다.

○ 따라서 500m 또는 250m 단위의 격자를 별도 생성하여 정보를 추가 제공하며, 이를 위해 1km 격자에서 동서·남북방향으로 각각 2등분한 후 a, b 영문자를 통해 표시한다.

○ 추가적으로, 민간 유동인구 데이터와의 상호 운용성을 확보하는 차원에서 100m 격자를 동서·남북방향으로 각각 2등분한 후 50m 격자를 추가 생성하였으나 현

재 인터넷 서비스는 제공되지 않는다.

<표 2-1> 국토지리정보원의 격자 구성

격자 크기	서비스	문자		숫자		폴리곤 id	비고
		동서	남북	동서	남북		
100km	○	마	사			마사	
10km	○	마	사	0	3	마사03	
1km	○	마	사	02	32	마사0232	
500m	○	마	사	02a	32a	마사02a32a	신규격자에 해당 (대도시 권역의 소지역)
250m	○	마	사	02ab	32ab	마사02ab32ab	
100m	○	마	사	024	324	마사024324	
50m		마	사	024b	324b	마사024b324b	신규격자에 해당 (서비스되지는 않음)
10m		마	사	0245	3248	마사02453248	국가지점번호 원형

□ 국토통계지도의 격자 단위 통계지도는 유용성이 뛰어난 것으로 판단된다.

○ 장점으로는 ① 행정구역 단위 통계지도에서 벗어나 격자 단위로 보다 세부적인 데이터의 제공이 가능하다는 점, ② 행정 경계의 변화와 무관하게 시계열적인 분석을 가능하게 한다는 점, ③ 통일된 공간 단위의 데이터 구축으로 데이터의 융합 및 비교 분석을 가능하게 한다는 점이다.

나. 국토통계지도의 격자형 제공 자료

□ 국토통계지도는 4개 분야(인구, 건물, 토지, 국토지표)로 구분하여 자료를 제공하고 있으며, 국토통계지도에서 제공하는 대부분의 자료는 격자통계지도에서도 제공된다.

○ 단 인구, 건물, 토지의 경우 1km(또는 500m) 이하의 격자통계는 시도와 시군구를 선택하여야만 조회가 가능하고, 국토지표의 경우에는 500m 격자로만 정보가 제공된다.

○ 인구정보는 행전안전부 주민정보 자료를, 건물정보는 국토교통부 건축물 대장 자료를 각 월 1회, 토지정보는 국토교통부 개별공시지가와 개별주택가격 정보를

6개월 단위로 수집하여 정보를 갱신한다.

○ 국토지표의 경우 인구와 사회, 토지와 주택, 생활과 복지, 국토 인프라, 환경과 안전 5개 분야에 대해 여러 지표를 설정하고 이를 통계적으로 계산하여 그 결과를 제공한다.

<표 2-2> 국토통계지도 제공 자료의 자료 출처

분야	대 분류	하위분류	자료 출처
인구	총인구	전체, 남자, 여자	행정안전부 주민정보
	유소년	전체, 남자, 여자	
	생산가능인구	전체, 남자, 여자	
	고령	전체, 남자, 여자	
	유아	전체, 남자, 여자	
	학생	전체, 남자, 여자	
	연령별	전체, 남자, 여자	
건물	일반	면적, 건폐율, 용적률 등	국토교통부 건축물 대장
	건축물 수	시기별 건축물 수	
	주용도	공동, 단독, 구용도	
	구조	목 구조, 철근콘크리트 구조 등	
	개별주택	개별주택 가격, 주거용도 면적	
토지	공시지가	공시지가	국토교통부 개별공시지가
국토 지표	인구와 사회	인구과소지역 비율	(격자 서비스 하지 않음)
	토지와 주택	노후건물, 토지이용압축도 등	국토정보플랫폼
	생활과 복지	공원, 공공시설 등에 대한 접근도	LH, 지자체, 교육부, 보건복지부, 심평원 등
	국토 인프라	고속국도, 철도, 주차장 접근성	국가교통DB, 지자체 등
	환경과 안전	경찰서, 소방서, 지진대피소 접근성	공공데이터 포털 등

□ 인구 분야 제공 자료는 인구 DB의 도로명 주소를 기준 DB(건물)의 도로명 주소와 매칭함으로써 자료가 생산된다.

○ 인구 분야 제공 자료는 7개 대 분류에 대해 성별로 구분하여 자료를 제공하고 있다. 학생은 초등학생, 중학생, 고등학생으로 구분되며, 연령별 인구는 20대부터 100세 이상 인구까지 총 9개의 연령대로 구분하여 제공한다.

○ 6개의 격자 단위가 모두 제공되며, 격자별 인구가 5 이하인 경우에는 정보보호를 위해 N/A로 처리하여 제공한다.

<표 2-3> 격자통계지도 중 인구 분야 제공 자료

대분류	소분류(세부내용)	격자 크기	비고
총인구	성별(전체, 남자, 여자)	100km, 10km, 1km, 500m, 250m, 100m (6단계)*	5이하 N/A 처리
유소년인구	성별(전체, 남자, 여자)		
생산가능인구	성별(전체, 남자, 여자)		
고령인구	성별(전체, 남자, 여자)		
유아인구	성별((전체, 남자, 여자)		
학생인구	초등학생_성별(전체, 남자, 여자) 중학생_성별(전체, 남자, 여자) 고등학생_성별(전체, 남자, 여자)		
연령별인구	20대_성별(전체, 남자, 여자) 30대_성별(전체, 남자, 여자) 40대_성별(전체, 남자, 여자) 50대_성별(전체, 남자, 여자) 60대_성별(전체, 남자, 여자) 70대_성별(전체, 남자, 여자) 80대_성별(전체, 남자, 여자) 90대_성별(전체, 남자, 여자) 100세 이상_성별(전체, 남자, 여자)		

* 1km(또는 500m) 이하의 격자통계지도는 시도와 시군구를 선택하여야 조회 가능

□ 건물 분야 제공 자료는 국토교통부의 건축물대장 내 건물 DB의 도로명 주소를 기준 DB(건물)의 도로명주소와 매칭함으로써 자료를 생산한다.

<표 2-4> 격자통계지도 중 건물 분야 제공 자료

대분류	소분류(세부내용)	격자 크기
일반	건축면적, 건폐율, 높이, 대지면적, 연 면적, 용적률, 지상 층수, 지하 층수	100km, 10km, 1km, 500m, 250m, 100m (6단계)*
건축물 수	건축물 수 합계 통계, 사용 승인일, 시기별(9년 이하), 시기별(15년 이상 19년 이하), 시기별(20년 이상 24년 이하), 시기별(25년 이상 29년 이하), 시기별(30년 이상 34년 이하), 시기별(35년 이상)	
주 용도	건축물 수(공동), 건축물 수(단독), 주 용도 코드	
구조	건축물 수(일반목 구조, 철근콘크리트 구조), 주 용도 코드별 건축물 수	
개별주택	개별주택 가격, 주거용도 면적	

* 1km(또는 500m) 이하의 격자통계지도는 시도와 시군구를 선택하여야 조회 가능

□ 토지 분야 자료인 공시지가는 국토교통부의 공시지가를 활용하며, 토지 DB의 지번 정보(PNU)와 기준 DB(지적)의 지번 정보를 매칭하여 자료를 생산한다.

<표 2-5> 격자통계지도 중 토지 분야 제공 자료

대분류	소분류	격자 크기
공시지가	-	100km,⋯,100m (6단계)*

* 1km(또는 500m) 이하의 격자통계지도는 시도와 시군구를 선택하여야 조회 가능

□ 국토지표 분야는 지자체와 공공기관의 데이터를 활용하여 5개 분야에 대해 활용 가능성이 있는 지표를 선정한 후 500m 격자통계를 제공한다.

○ 인구, 건물, 토지 자료가 원시 데이터(raw data)의 성격인데 반해, 국토지표 분야 자료는 가공 데이터로 활용도는 상황에 따라 다를 것으로 판단된다.

<표 2-6> 격자통계지도 중 국토지표 분야 제공 자료

대분류	소분류(세부내용)	격자 크기	비고
인구와 사회	인구과소지역 비율	500m 격자만 제공	· 격자 서비스 없음
토지와 주택	노후건물 (노후건물 비율 20년/30년, 노후건물 수 20년/30년)		· 비율= 건물 수 / 격자 내 건물 수
	노후주택 (노후주택 비율 20년/30년, 노후주택 수 20년/30년)		
	토지이용 압축도		· 압축도= 연면적 / 격자면적
	토지이용 복합도		· 복합도= 격자 내 용도 수
생활과 복지	생활권공원, 주제공원, 공연문화시설, 도서관, 공공 체육시설, 어린이집, 유치원, 초등학교, 노인여가복지시설, 종합병원, 응급의료시설, 병원, 의원, 보건기관, 경로당, 노인교실, 노인복지관, 생활체육시설, 약국, 종합사회복지관		· 시설까지의 도로 이동 거리 (접근성)
국토 인프라	고속도로IC, 고속화철도, 주차장		· 시설까지의 도로 이동 거리 (접근성)
환경과 안전	경찰서, 소방서, 지진옥외대피소		· 시설까지의 도로 이동 거리 (접근성)

다. 국토통계지도의 격자통계 검색 사례

□ 국토지리정보원 국토정보플랫폼(map.ngii.go.kr)의 국토통계지도에서 인구, 건물, 토지, 국토지표 등에 대한 격자통계 검색 사례를 살펴보면 다음과 같다.

○ 첫 번째로, 인구통계 검색 사례로 대전광역시의 총 인구 수에 대한 1㎞ 격자통계지도를 조회하면 다음과 같은 결과를 볼 수 있다.

- 항목 : 총 인구 수 (대전광역시)
- 통계구역기준 : 1㎞ × 1㎞ 격자

번호	구분	내용
1	통계명칭	총 인구 수(전체) - 1km
2	산출방식 (단위)	합계(명)
3	행정구역	대전광역시
4	전체 통계	1,466,776
5	보안처리	60 건

23,884 - 35,123 이하
17,868 - 23,884 이하
13,316 - 17,868 이하
9,098 - 13,316 이하
5,178 - 9,098 이하
1,550 - 5,178 이하
1,550 이하
빈격자

<그림 2-1> 인구 격자통계지도 검색 사례1(1㎞)

○ 두 번째 인구통계 검색 사례로 500m 이하의 격자를 선택하는 경우 시도와 시군구 단위를 지정해야 함을 확인하였다. 대전시 서구를 대상으로 250m 격자의 유아 인구 수를 검색해 보면 다음과 같다.

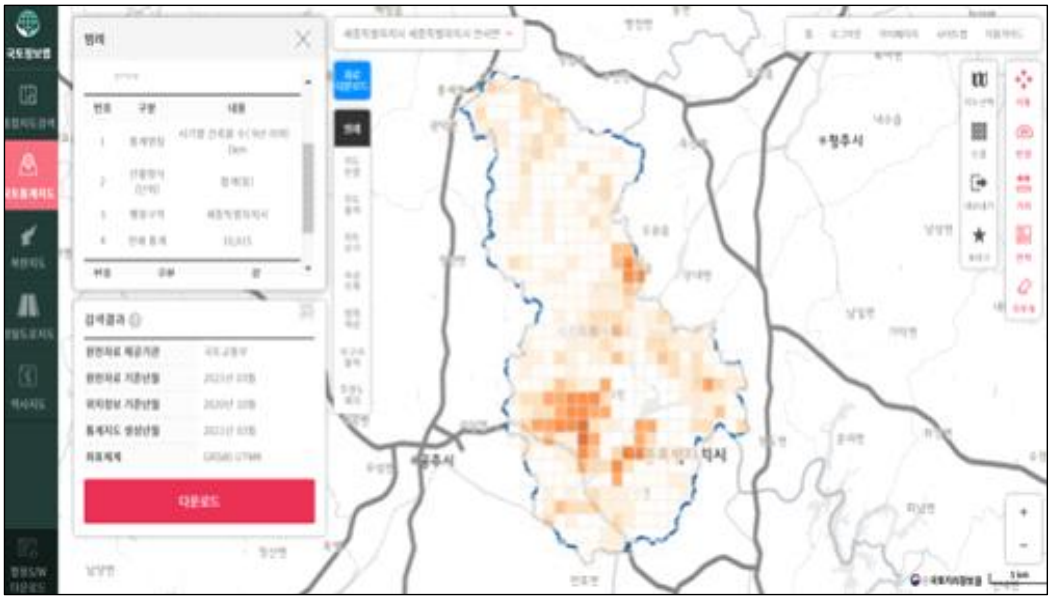
- 항목 : 유아 인구 수 (대전광역시 서구)
- 통계구역기준 : 250m × 250m 격자
- 참고 : 500m 이하의 격자를 선택하는 경우 시도와 시군구 단위를 지정해야 함

번호	구분	내용
1	통계명칭	유아 인구 수(전체) - 250m
2	산출방식 (단위)	합계(명)
3	행정구역	대전광역시 서구
4	전체 통계	24,072
5	보안처리	62 건

<그림 2-2> 인구 격자통계지도 검색 사례2(250m)

○ 건물 격자통계 검색 사례로, 건축연식이 9년 이하인 건축물 수에 대한 1㎞ 격자 통계 검색 결과는 다음과 같다.

- 항목 : 건축연식 9년 이하인 건축물 수 (세종시 일대)
- 통계구역기준 : 1㎞ × 1㎞ 격자



<그림 2-3> 건물 격자통계지도 검색 사례(1㎞)

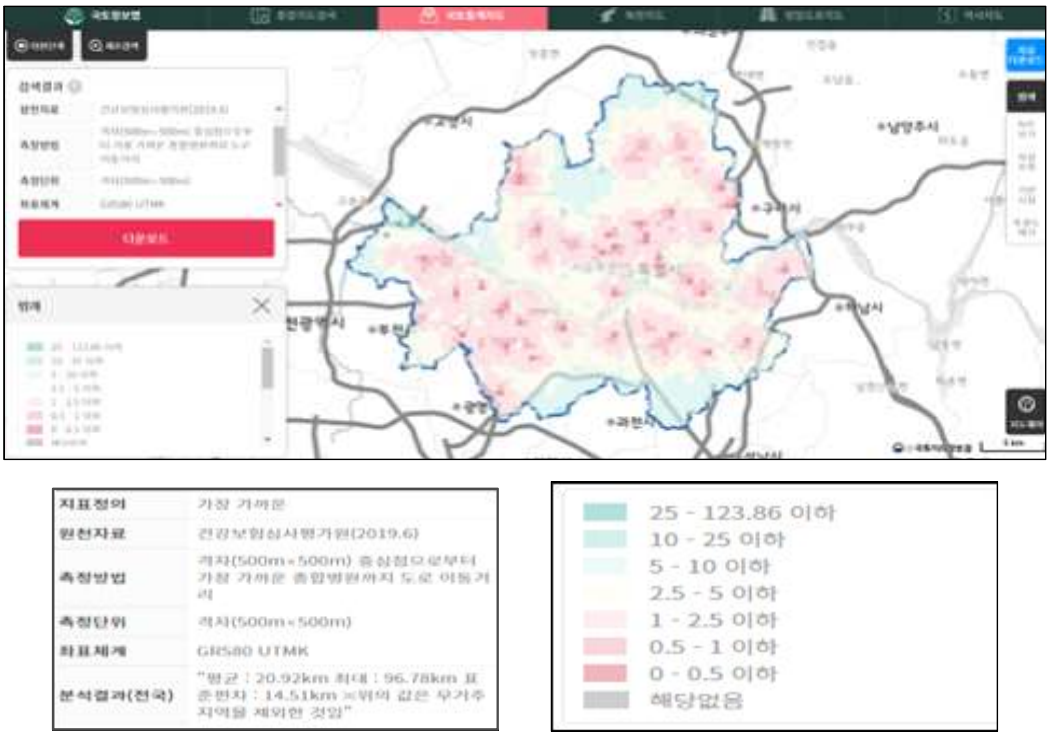
○ 토지 격자통계 검색 사례로, 공시지가에 대한 500m 격자통계 검색 결과는 다음과 같다. 단, 데이터 수집 및 산출시점, 지오코딩, 연속지적도의 면적 차이 등의 사유로 가격 오차가 발생할 수 있음을 표시해 준다.

- 항목 : 공시지가 (대전시 유성구)
- 통계구역기준 : 500m × 500m 격자

<그림 2-4> 토지 격자통계지도 검색 사례(500m)

○ 국토지표 격자통계 검색 사례로, 종합병원 접근성 대한 500m 격자통계 검색 결과는 다음과 같다.

- 항목 : 생활과 복지 항목 중 종합병원 접근성 (서울시)
- 통계구역기준 : 500m × 500m 격자



<그림 2-5> 국토지표 격자통계지도 검색 사례(500m)

1.2 한국 통계청 통계지리정보서비스(SGIS)

□ 통계지리정보서비스(sgis.kostat.go.kr) 중 대화형 통계지도에서 격자단위를 선택하면 국가지점번호체계에 따른 격자단위 통계지도를 생성할 수 있고, 자료의 출처는 통계청 조사 자료를 기반으로 한다.

○ 총조사 주요지표 항목에서는 주요지표를 선택하여 조건을 설정할 수 있도록 하는데, 제공되는 주요지표는 총인구, 가구, 주택, 사업체 4가지이다.

<표 2-7> 통계지리정보서비스 격자자료 서비스의 제공 자료

조사구분	대분류	소분류	세부내용
총조사 주요지표	총인구	-	총인구수
	가구	-	가구수
	주택	-	주택수
	사업체수	-	사업체수
인구주택 총조사	조사연도	-	2000년도, …, 2019년도
	인구조건	성별	전체, 남, 여
		연령	A 이상 B 미만(5세 단위)
	가구조건	세대구성	1세대가구, 2세대가구, 3세대가구, 4세대이상가구, 1인가구, 비혈연가구
	주택조건	주택유형	단독주택, 아파트, 연립주택, 다세대주택, 비주거용건물(상가, 공장, 여관 등) 내 주택, 주택 이외의 거처
		건축연도	1979년 이전,1980년~1989년, …, 2018년
		연면적	A 초과 B 이하(0㎡ ~ 230+㎡)
전국사업체 조사	조사연도	-	2000년도, …, 2019년도
	사업체수	-	-
	종사자수	-	-

○ 인구주택총조사 항목에서는 조사연도, 인구조건, 가구조건, 주택조건을 설정할 수 있도록 하고, 전국 사업체조사 항목에서는 조사연도와 대상(사업체수, 종사자수)을 선택하여 조회된 자료를 바탕으로 지도 위에 시각화 서비스를 제공한다.

○ 각 항목별로 100km, 10km, 1km, 100m 단위로 격자통계 서비스를 제공하며 격자 해당 통계 값이 속하는 범례의 구간으로 표현하고 있다. 지도의 레벨(1~13레벨)에 따라 표현되는 항목의 격자단위가 설정된다.

○ 통계청 자료의 좌표체계와 지점번호 체계는 국토지리정보원의 국토통계서비스와 동일하다.

□ 인구주택총조사 자료의 주택조건을 제외한 대부분의 항목에서 시계열 조회가 가능하다는 장점이 있다.

○ 인터렉티브 서비스로 화면에 표시된 격자자료의 다운로드는 불가하며, 필요한 경우 자료제공 신청과 별도의 심의 절차를 거친 후 제공받을 수 있다.

□ 데이터 정보보호를 위해 인구주택총조사 자료는 통계 값이 5 이하인 경우, 전국 사업체조사 자료는 통계 값이 3 이하인 경우 N/A 처리한다.

<표 2-8> 통계지리정보서비스 격자자료 서비스의 격자단위별 지도 레벨

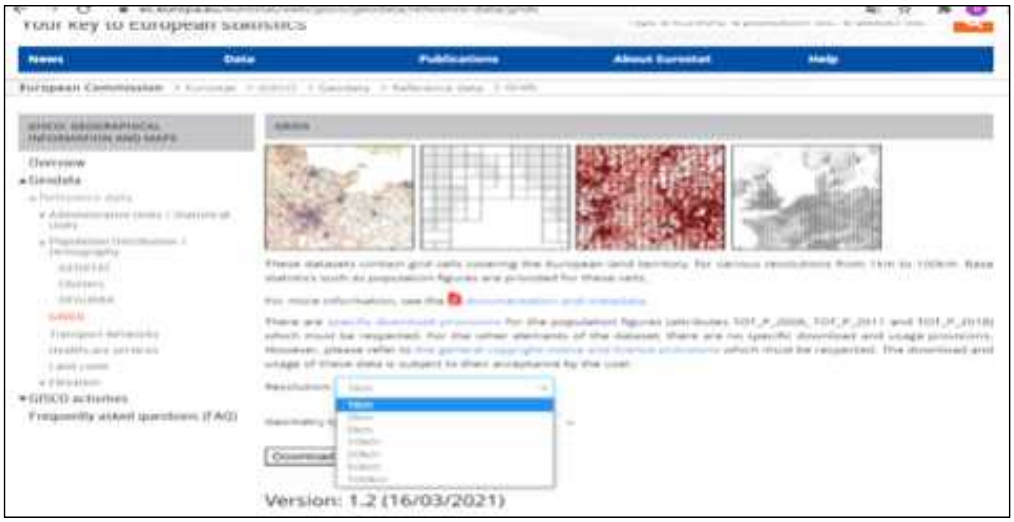
격자단위	지도레벨
100km	사용하지 않음
10km	1-5
1km	6-8
100m	9-13

1.3 유럽연합통계청(EUROSTAT)

□ 유럽연합통계청(Eurostat)에서는 행정구역 기반의 통계 제공을 위한 GISCO(Geographic Information System of the COmmission), 격자 기반 통계 제공을 위한 GEOSTAT (GeoStatistics) 프로젝트를 수행하고 있다.

○ 유럽연합통계청은 자료를 직접 수집하지 않고, 회원국의 통계청에서 수집한 자료를 취합하여 배포하는 역할을 수행한다.

○ 격자통계는 1km, 2km, 5km, 10km, 20km, 50km, 100km에 이르기까지 다양하게 제공되며, 격자id, 좌표(coordinate), 인구 수(2006년, 2011년, 2018년), 육지비율, 해안까지의 거리, 국경까지의 거리, 국가코드, 기준연도 등이 속성정보로 포함된다.



Attribute	Datatype	Description
GRD_ID	String	The grid cell identifier according to INSPIRE specification.
X_LLC	int	X coordinate of the lower left corner.
Y_LLC	int	Y coordinate of the lower left corner.
TOT_P_YYYY	int	Total population figure for year YYYY. The years 2006, 2011 and 2018 are provided. Source: GEOSTAT project - be aware of the specific license conditions described below.
LAND_PC	double	Estimate of the land share (in percentage) of the grid cell.
DIST_COAST	double	Estimate of the minimum distance to the coast, in meters.
DIST_BORD	double	Estimate of the minimum distance to a country land border, in meters.
CNTR_ID	String	List of country codes the grid cell intersects or is nearby to (within approximately 1.5 km).
NUTSYYYY_X	String	List of NUTS codes (for level X, from 0 to 3) the grid cell intersects or is nearby to (within approximately 1.5 km). YYYY is the NUTS version, 2016 or 2021.

<그림 2-6> 유럽연합통계청의 격자통계 제공 현황 및 속성정보

□ GEOSTAT 프로젝트를 통해 유럽연합 차원에서 유럽 전체 지역에 대한 격자통계를 제작한다.

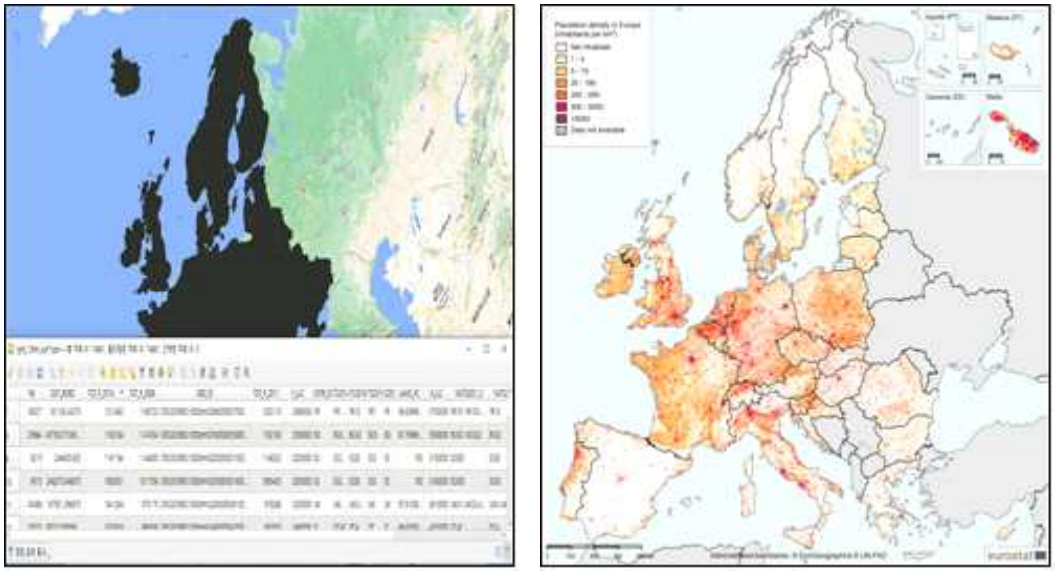
○ 유럽연합통계청에서의 격자통계는 EU 회원국인 29개 국가를 대상(발칸반도 국가, 동유럽 국가, 터키 등은 제외)으로 하며, 1km 단위를 최소 격자 크기로 지정하였다.

○ GEOSTAT 프로젝트는 유럽의 다양한 특성들을 나타내는 것을 목표로 하며, 격

자 데이터의 생산을 장려하기 위해 국가통계기관 및 국제단체로부터 사례, 방법론 및 경험을 수집했다.

○ 1km보다 상세한 수준의 인구 수 또는 기타 정보를 취득하기 위해서는 유럽 국가별 격자통계를 이용하여야 가능하다.

○ 개별 국가의 격자통계 생산과 이용은 각 국가의 정보화 수준에 따라 매우 다르게 이루어지고 있으며, 다수의 격자통계 데이터가 유료로 제공된다.



(a) 10km 격자통계 (b) 1km 격자통계

<그림 2-7> 유럽연합통계청의 격자통계 제공 예시

□ GEOSTAT(2011)에 따르면, 격자통계 사용자의 대다수는 공간 분석(spatial analysis)을 수행하는 것을 목적으로 격자통계를 사용한다.

○ 이 프로젝트에서는 유럽 전역에 대한 적합성, 기밀성, 이용성을 충족시키는 격자 크기로 1㎞를 권장하며, 보다 작은 크기로는 500m, 250m의 이용을 권장한다.

○ 관심 영역과 격자 크기 간에는 명확한 상관관계가 존재하였다. 국가 및 소규모 지역의 격자를 원하는 데이터 사용자는 최대 250m의 작은 격자 크기를 선호하고, 국가들 또는 대륙 단위의 데이터를 필요로 하는 사용자는 1km의 격자 크기를 선호하는 것으로 조사되었다.

○ 노출 제어의 관점에서 볼 때, 공개할 수 없는 값을 고정된 최소값으로 대체하는 것이 작은 격자를 더 큰 격자로 집계(aggregation)하는 것보다 더 적절하다고 판단하고 있다.

□ 지리통계유럽포럼(The European Forum for Geostatistics)은 유럽 전체를 대상으로 한 격자통계 공통 체계를 구축하기 위해 노력하고 있다.

○ 유럽연합은 격자의 좌표체계를 표준화하기 위해 INSPIRE(Infrastructure for Spatial Information in Europe)에서 정의한 ETRS(European Terrestrial Reference System) 89-LAEA(Lambert Azimuthal Equal Area) 좌표체계를 이용하여 격자 ID를 부여한다.

<표 2-9> INSPIRE의 격자 ID 부여 방법

구분	특징
투영원점	-10°E, 52°N
투영법	Lambert Azimuthal Equal Area
좌표기준점	투영원점에서 남쪽 3,210㎞, 서쪽 4,321㎞
데이텀	ETRS(European Terrestrial Reference System) 89
타원체	GRS80타원체
ERTS89-LAEA 격자 ID 부여방법 (격자 체계)	• 예:1kmN2599E4695 (좌하단 코너의 좌표가 4,695,000m, 2,599,000m인 1km 격자) • [격자크기][격자크기단위]+ N[격자 좌하단 코너의 Y좌표/1000]+E[격자 좌하단 코너의 X좌표/1000]
INSPIRE의 격자 ID 부여방법 (통계단위)	• 예1: CRS3035RES1000mN5400000E1200000 • 예2: CRS4326RES0-10-0dmsLON45-20-0LAT4-50-0 • CRS: Coord. Ref. System(좌표체계)를 의미함 • 3035: ERTS89/LAEA 투영법을 나타내는 EPSG(European Petroleum Survey Group) 코드로 전 세계에서 통용되는 좌표계를 정의함 • RESxxm: 격자의 크기가 XXm임을 나타냄 • Nxxx: 격자 좌하단의 Northing 직각좌표 • Exxx: 격자 좌하단의 Easting 직각좌표 • (예2) RESx-x-xdmsLONx-x-xLATx-x-x: 투영좌표 대신 경위도(도분초) 좌표를 사용할 경우(이 경우 EPSG코드는 WGS84 4236 등 타원체를 나타내는 코드를 사용함)

1.4 핀란드통계청(Statistics Finland)

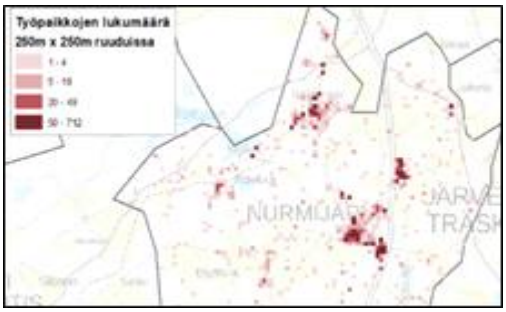
□ 핀란드통계청에서는 격자 기반의 DB를 생산하고 있으며, 격자통계는 250m, 1km, 5km 단위로 생산한다.

○ 제공되는 데이터 항목으로는 인구 구조, 교육, 주요 활동 및 소득 유형, 가구의 삶과 소득 단계, 건물 및 직장 등이 포함된다. 유료로 제공되며 라이센스 단위로 요금을 책정한다.

○ 격자의 좌표체계는 ETRS89-TM35FIN 투영법에 기반하고 있으며 데이터베이스는 최신 자료로 매년 업데이트된다.

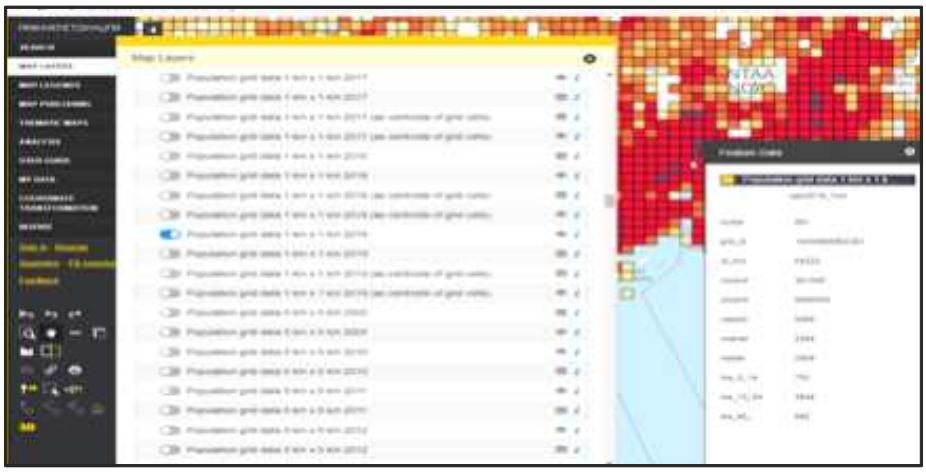
○ 인터넷을 통한 데이터 검색은 1km와 5km에 대한 인구자료만 확인이 가능하며 격자를 선택하면 INSPIRE 격자 ID와 통계 값을 확인할 수 있다.

Number of licences	Grid size		
	250 m x 250 m	1 km x 1 km	5 km x 5 km
1	EUR 4,800	EUR 3,440	EUR 1,720
2-4	EUR 8,300	EUR 5,950	EUR 2,970
5-9	EUR 12,700	EUR 9,100	EUR 4,550
10 or more	according to the agreement		



출처: www.tilastokeskus.fi/tup/ruututietokanta/hinnasto_en.html

<그림 2-8> 핀란드통계청의 격자통계 서비스 요금(좌)와 250m 격자 유료샘플(우)



출처 kartta.paikkatietoikkuna.fi/?lang=en

<그림 2-9> 핀란드통계청의 1km 격자 인구 데이터

1.5 오스트리아통계청(Statistics Austria)

□ 오스트리아통계청에서는 격자통계를 250m, 500m, 1km, 2km, 5km, 10km 단위로 생산한다.

○ 제공되는 데이터 항목은 인구, 건물, 주거, 노동 관련 데이터이며 2011년 기준 데이터를 제공한다.

○ 특정 항목은 100m 단위로도 제공되지만, 대부분은 최소 250m 격자 단위로 제공된다. 다운로드 서비스는 유료이며, 인터넷을 통한 서비스는 제공하지 않는 것으로 확인된다.

○ EU의 INSPIRE 지침에 따라 투영 원점을 10° E, 52° N으로 설정한 LAEA (Lambert Azimuthal Equal Area) 투영법에 의한 평면직각좌표를 기준으로 격자통계를 생산한다.

Fees for internal, non-commercial use in €

A fixed basic amount of €282 will be charged per order to cover consultancy, processing and billing costs. Selected data (packages) in grid size 10km are available for free.

Fees for one Object/Characteristics	100m	250m	500m	1.000m	2.000m	5km
Object	3.200,-	1.600,-	800,-	400,-	200,-	100,-
Characteristics	x	2.720,-	1.360,-	680,-	340,-	170,-

Packages derived from Registers	100m	250m	500m	1.000m	2.000m	5km
Package A – Register-Objects	8.000,-	4.000,-	2.000,-	1.000,-	500,-	250,-
Package Population Stock	x	5.440,-	2.720,-	1.360,-	680,-	340,-
Package Buildings and Dwellings	x	11.920,-	5.960,-	2.980,-	1.490,-	745,-

Packages derived from the Register-based Labour Market Statistics	100m	250m	500m	1.000m	2.000m	5km
Package RbLMS-Objects	9.600,-	4.800,-	2.400,-	1.200,-	600,-	300,-
Package RbLMS-Demography and Employment	x	9.920,-	4.960,-	2.480,-	1.240,-	620,-
Package RbLMS-Education	x	4.880,-	2.440,-	1.220,-	610,-	305,-
Package RbLMS-Households	x	9.240,-	4.620,-	2.310,-	1.155,-	578,-
Package RbLMS-Families	x	4.480,-	2.240,-	1.120,-	560,-	280,-
Package RbLMS-Commuters	x	8.640,-	4.320,-	2.160,-	1. 080,-	540,-
RbLMS-Grid based Commuting Matrix	x	8.640,-	4.320,-	2.160,-	x	x

Packages from the Census 2011	100m	250m	500m	1.000m	2.000m	5km
Package Census2011 - Objects	14.400,-	7.200,-	3.600,-	1.800,-	900,-	450,-
Package Census2011 - Workplace / Local Units of Employment	x	11.680,-	5.840,-	2.920,-	1.460,-	730,-
Package Census2011 - Buildings and Dwellings	x	14.640,-	7.320,-	3.660,-	1.830,-	915,-
Package Census2011 - Demography and Employment	x	9.920,-	4.960,-	2.480,-	1.240,-	620,-
Package Census2011 - Education	x	4.880,-	2.440,-	1.220,-	610,-	305,-
Package Census2011 - Households	x	9.240,-	4.620,-	2.310,-	1.155,-	578,-
Package Census2011 - Families	x	4.480,-	2.240,-	1.120,-	560,-	280,-

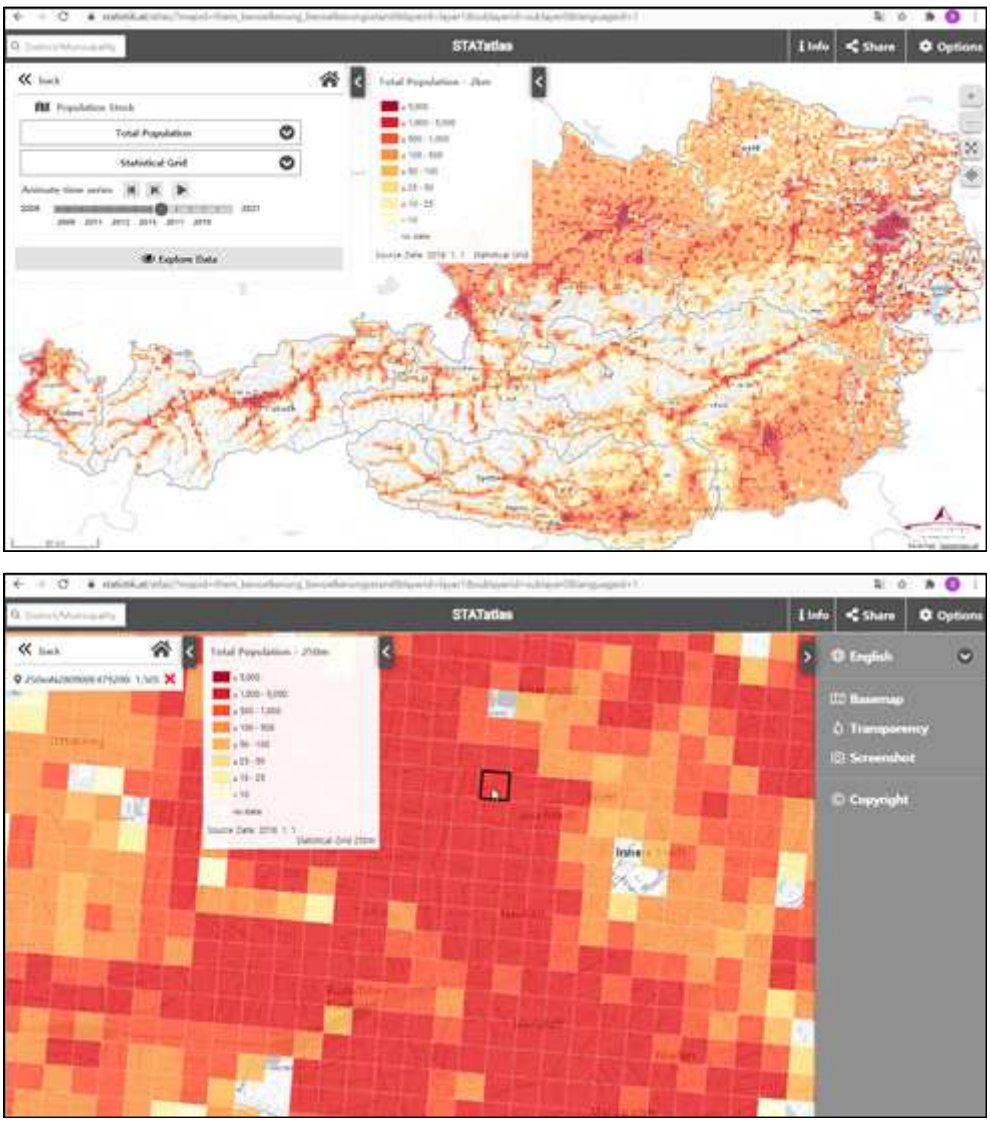
출처 www.statistik.at/web_en/classifications/regional_breakdown/grid/index.html

<그림 2-10> 오스트리아통계청의 격자통계 제공 현황 및 서비스 요금

□ 인터렉티브 서비스에서의 격자 단위는 250m, 500m, 1km, 2km로 서비스되고 있다.

○ 제공되는 데이터 항목은 총인구, 건물 수, 건축시기, 용도, 주거형태이며 인구는 시계열 서비스도 제공된다.

○ 격자를 클릭하면 위치정보와 통계 값이 제시되고, 격자의 위치정보는 EU의 INSPIRE의 격자 체계를 따른다.



출처: statistik.at/atlas

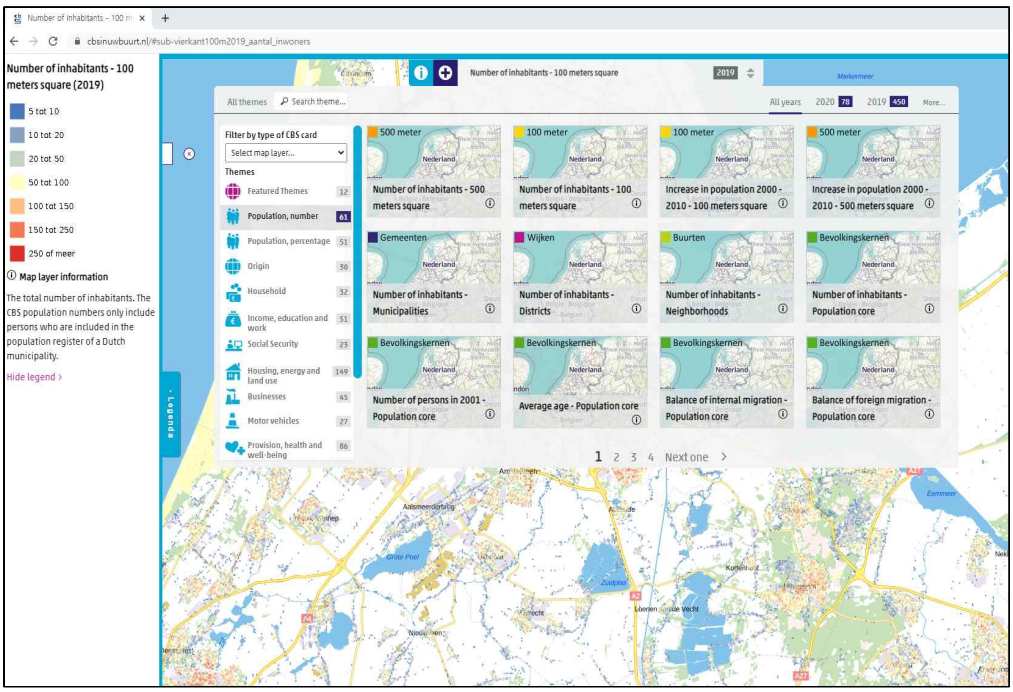
<그림 2-11> 오스트리아통계청의 격자통계 2km(위)와 250m(아래) 서비스 예시

1.6 네덜란드통계청(Central Bureau voor de Statistiek, CBS)

□ 네덜란드는 유럽 국가 중 격자통계가 가장 잘 구축되어 있으며, 인터렉티브 서비스와 다운로드 서비스 모두 제공한다.

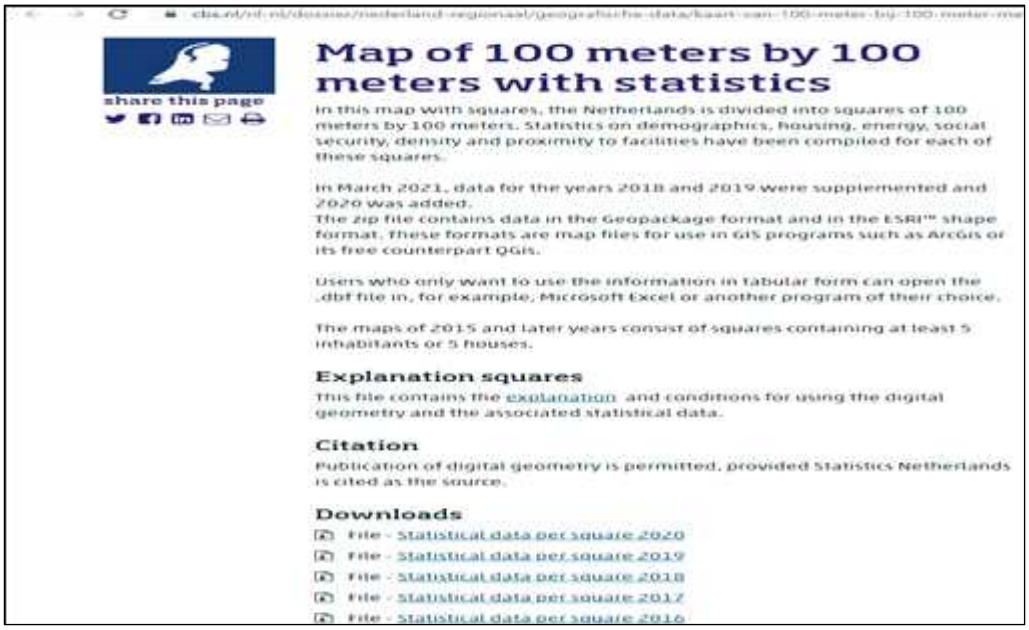
○ 제공되는 최소 격자는 100m로 인구, 주거, 에너지, 사회안전, 시설밀도 등 많은 자료를 제공하고 있으며, 매년 자료가 업데이트 되는 것으로 확인된다.

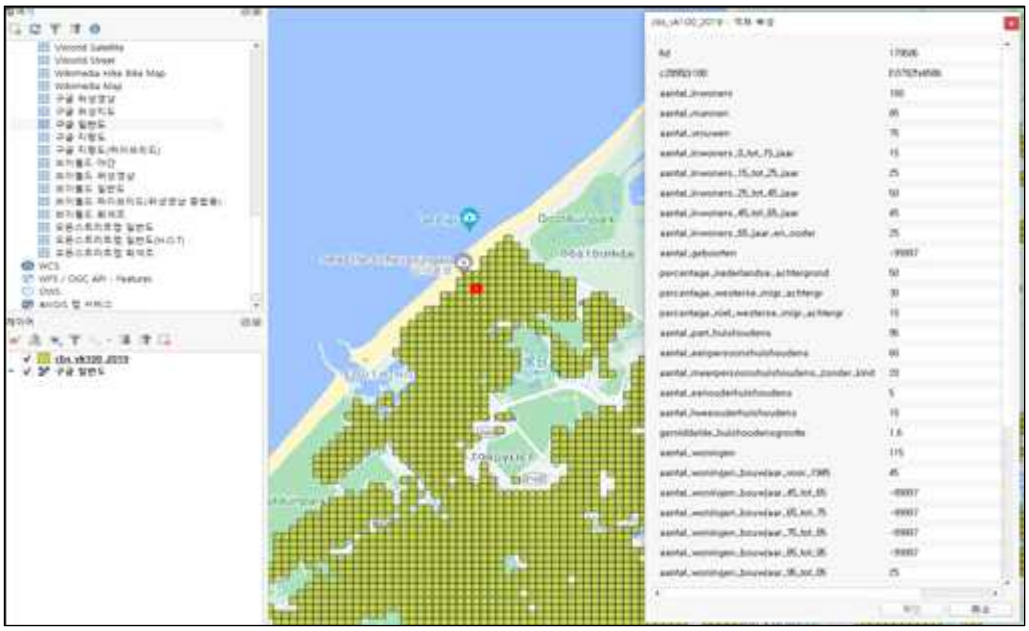
○ 별도의 인증과정 없이 데이터를 개방하며, 다운로드 서비스를 통해 자료를 검토한 결과 100m, 500m 격자에 대해 연도별 통계 값을 제공받을 수 있다. 따라서 격자통계의 시계열 변화를 관측할 수 있고, 격자 간 통계 값의 비교도 가능하다.



출처: cbsinuwbuurt.nl

<그림 2-12> 네덜란드통계청의 격자통계 인터렉티브 서비스 현황





출처: www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/kaart-van-100-meter-bij-100-meter-met-statistieken

<그림 2-13> 네덜란드통계청의 100m 격자통계 다운로드 서비스 현황(위)과 표시화면(아래)

1.7 노르웨이통계청(Statistics Norway)

□ 노르웨이에서는 격자통계에 대해 인터렉티브 및 다운로드 서비스를 모두 제공한다.

- ○ 서비스 격자의 크기는 두 서비스 모두 5km, 1km, 250m이며 인구, 가구, 주택에 대한 정보를 제공하며, 격자는 UTM 투영법을 기준으로 분할되어 있다.
- ○ 인터렉티브 서비스에서 격자를 선택하면 위치정보와 격자별 통계 값을 확인할 수 있고, 다운로드 서비스는 별도의 사이트에서 제공받을 수 있다.

출처: kart.ssb.no

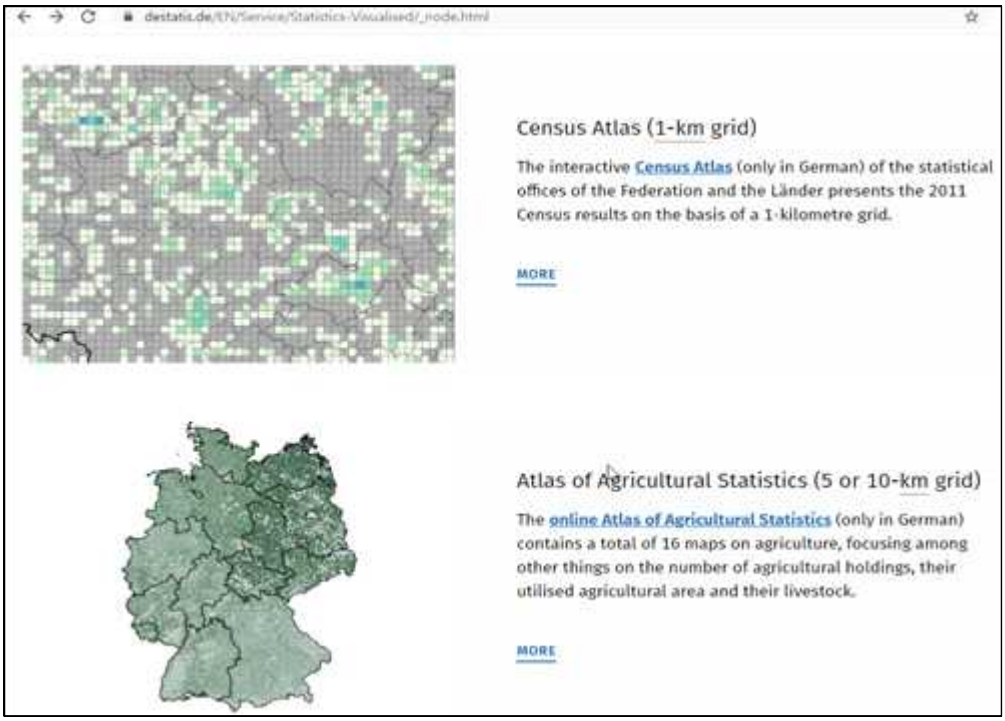
출처: www.ssb.no/natur-og-miljo/geodata#Nedlasting_av_rutenettsstatistikk

<그림 2-14> 노르웨이통계청의 격자통계 인터렉티브 서비스(좌)와 다운로드 서비스(우)

1.8 독일연방통계청(Germany Federal Statistical Office, DESTATIS)

□ 독일연방통계청은 2011년 센서스와 2010년 농업통계 지도집을 통해 인터렉티브 방식의 격자통계를 제공한다.

○ 자바스크립트 기반으로 제작된 인터넷 지도집은 격자통계와 더불어 행정경계, 지명, 자연지명, 고속도로 등도 표현되어 있어 격자의 위치를 파악할 수 있다.



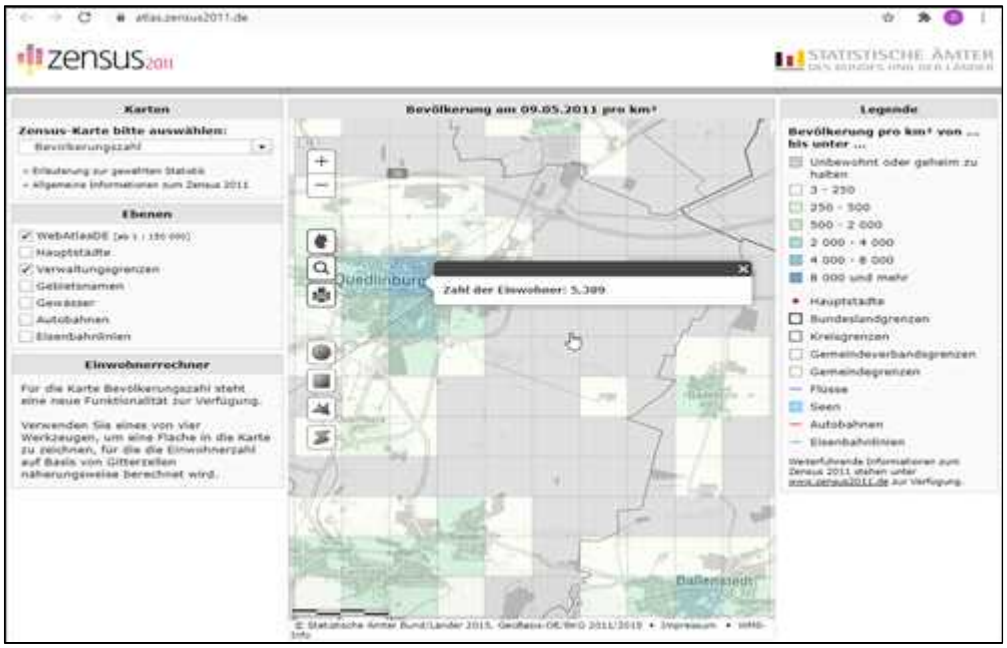
출처 : www.destatis.de/EN/Service/Statistics-Visualised/_node.html

<그림 2-15> 독일통계청의 격자통계 제공 현황

□ 2011년 센서스 지도집(https://atlas.zensus2011.de/)에서 1km 격자통계를 제공한다.

○ 제공항목으로는 인구(인구, 외국인비율, 성비, 평균연령 등), 주택(아파트 공실률, 평균생활공간면적 등)이며, 인구의 경우 2명 이하의 값은 제공하지 않는다.

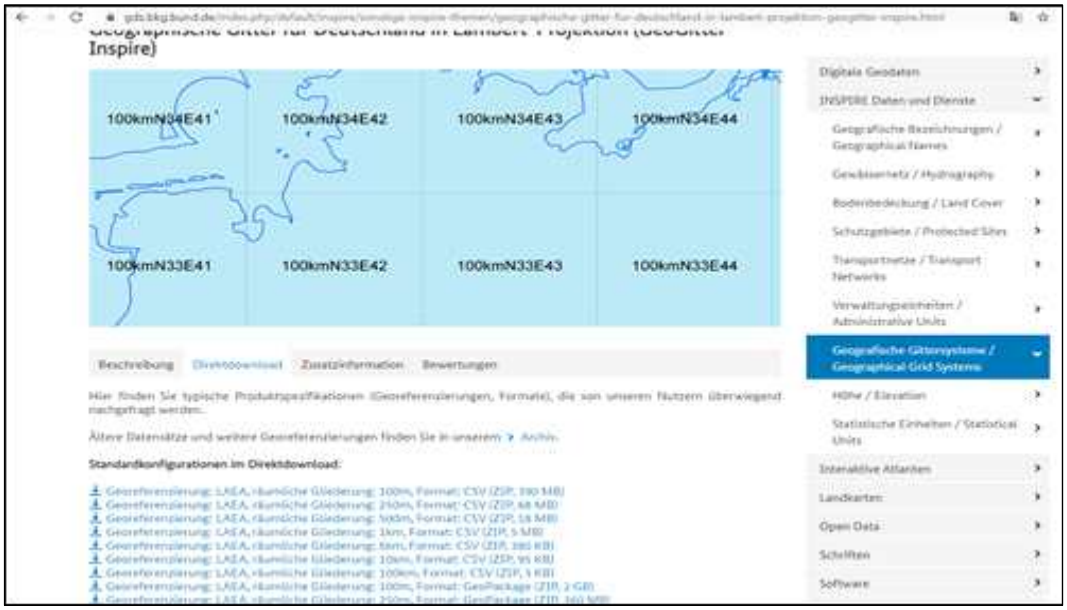
○ 다운로드 서비스는 별도의 사이트에서 100m와 1km에 대해 가능하며, 위치정보는 LAEA(Lambert Azimuthal Equal Area) 투영법을 따르는 INSPIRE 정의를 따른다.



<그림 2-16> 독일통계청의 인구 격자통계 제공 예시(2011년, 1km 격자)

출처 : www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html?nn=3065474#Gitter

<그림 2-17> 독일통계청의 2011년 센서스 격자통계 다운로드 서비스

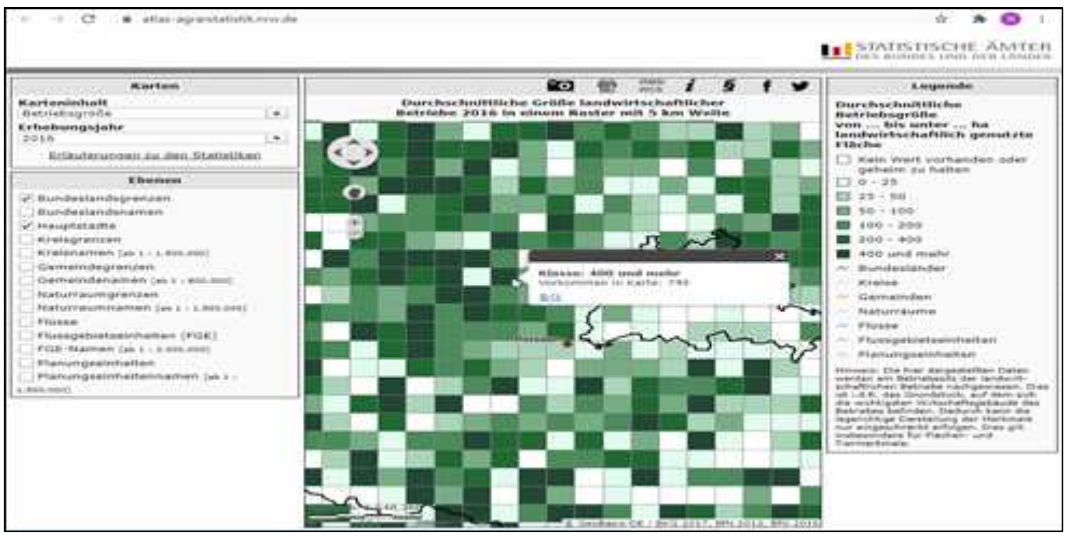


출처: gdz.bkg.bund.de/index.php/default/inspire/sonstige-inspire-themen/geographische-gitter-fur-deutschland-in-lambert-projektion-geogitter-inspire.html

<그림 2-18> 독일통계청의 격자 위치정보 다운로드(LAEA 투영법 기준)

□ 농업격자통계(www.atlas-agrarstatistik.nrw.de)는 2010년과 2016년 기준으로 5km, 10km 단위의 농업 통계를 제공한다.

○ 제공되는 항목으로는 격자별 농장의 평균크기, 가축(소, 돼지) 사육 두수, 식량(옥수수, 밀) 재배면적, 초지 면적 등 총 17종이다.



<그림 2-19> 독일통계청의 농업 통계 5km 격자 제공 예시(2016년, 농장 평균 크기)

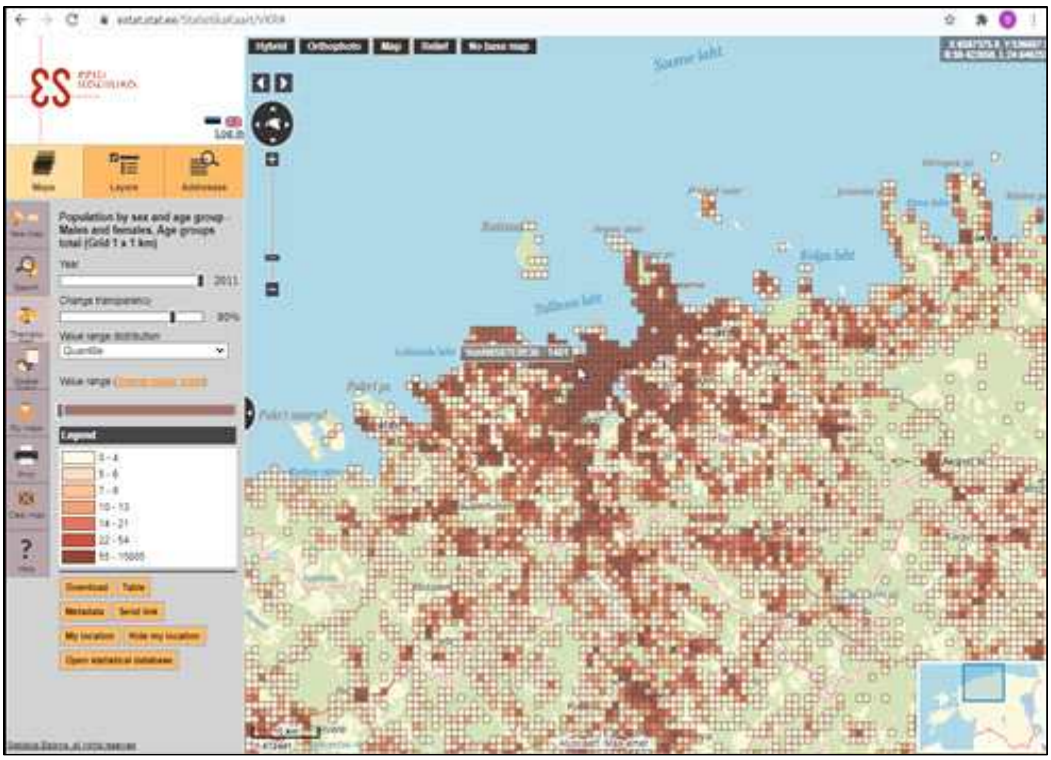
1.9 에스토니아통계청(Statistics Estonia)

□ 에스토니아통계청은 인터렉티브 서비스를 통해 1km 격자통계를 제공하고 있으며, ‘shp’ 또는 ‘kml’ 등의 파일 형태로 다운로드할 수 있으나 격자 ID는 조회만 가능하다.

○ 인터렉티브 서비스의 제공 자료는 1989년, 2000년, 2011년 센서스, 2015년, 2016년 인구 및 경제 자료이며, 자국어와 영어로 제공된다.

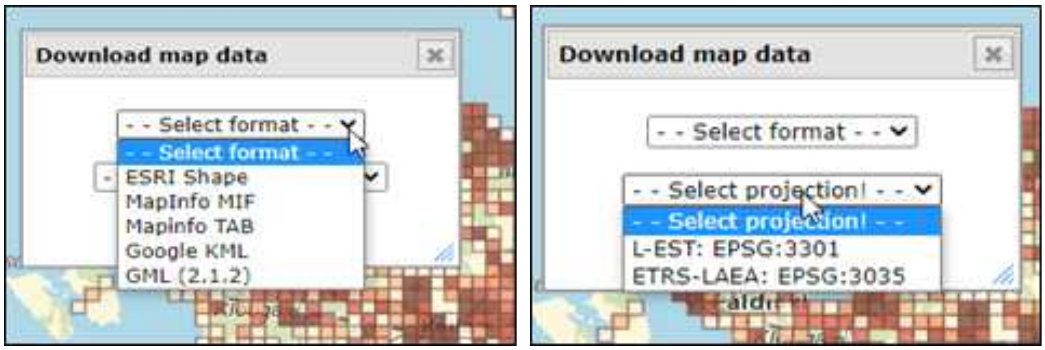
○ 격자통계에 대한 정보보호는 항목마다 수준이 다르며 인구 수가 3 이하인 경우에는 ‘<4’로 표시하고 구체적인 값은 공개하지 않고 있다.

○ 사용자가 요구하는 지역수준에 따라 격자의 단위는 달라지며, 전국단위는 1km, 도시지역은 500m, 대도시는 100m 단위이다(vana.stat.ee/dokumendid/296028).



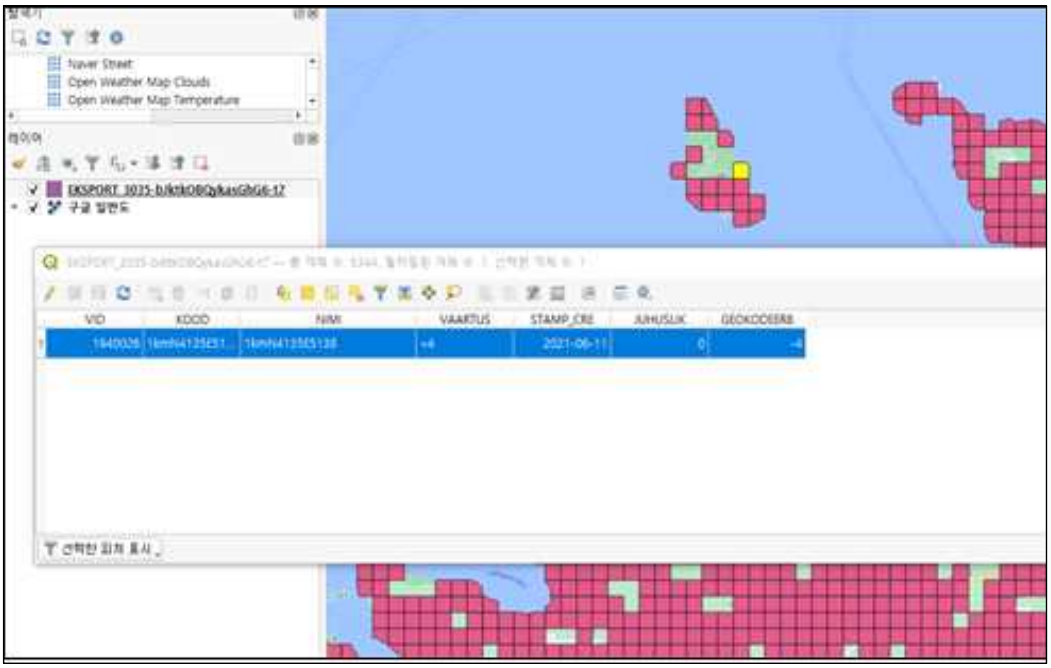
출처 : estat.stat.ee/StatistikaKaart/VKR

<그림 2-20> 에스토니아통계청의 격자통계 제공 예시(1km 격자, 인구)



출처: estat.stat.ee/StatistikaKaart/VKR

<그림 2-21> 에스토니아통계청 격자통계의 다운로드 파일형식(좌)과 좌표계(우) 설정



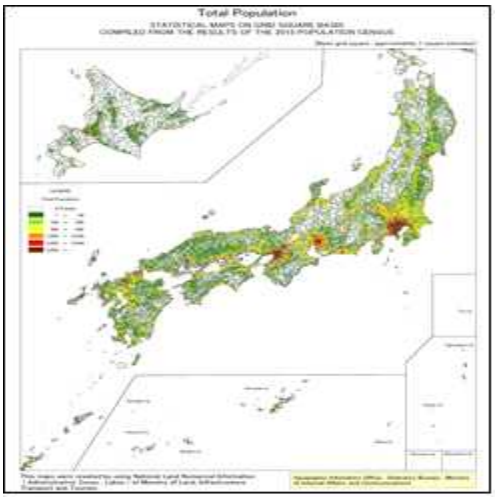
출처 : estat.stat.ee/StatistikaKaart/VKR

<그림 2-22> 에스토니아통계청 격자통계 정보보호 예시(1km 격자, 인구)

1.10 일본총무성통계국(Statistics Bureau of Japan)

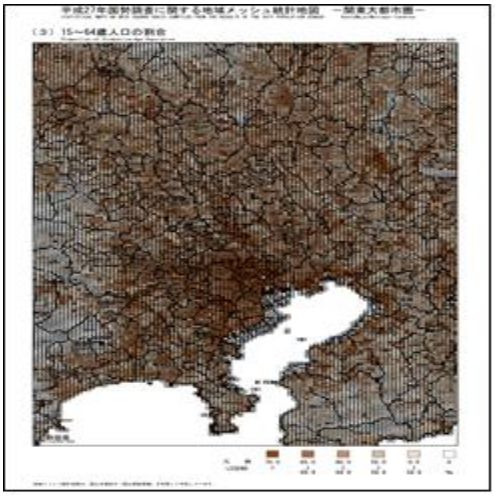
□ 일본총무성통계국에서는 인구센서스, 경제센서스, 사업체 센서스 데이터를 활용하여 격자통계를 생산한다.

○ 일본의 격자통계는 센서스의 기본 단위인 '기본 단위구'에 속하는 지역을 메쉬 구획에 대응하는 방식을 이용하고 있다. 일본의 격자 단위인 메쉬에 대한 자세한 사항은 부록3을 참조하면 된다.

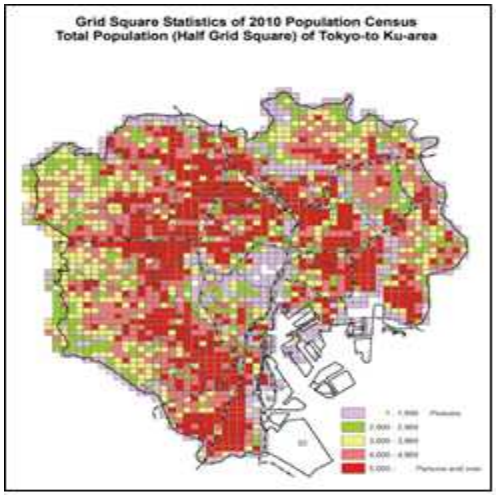


(a) 일본 전체 총 인구 수(1km 격자)

(b) 대도시 격자



(c) 간토대도시권 15-64세 인구비율(1km 격자)



(d) 도쿄지역 총인구 수(500m 격자)

<그림 2-23> 일본의 격자통계 제공 예시1

□ 일본어 홈페이지와 영문 홈페이지를 구분하여 제공하는 자료의 종류 및 형태를 달리 한다.

○ 영문 홈페이지(https://www.stat.go.jp/english)에서는 격자의 축척이나 레벨을 조정할 수 없으며, 가공 및 분석이 가능한 형태의 데이터 다운로드는 불가하고, pdf 버전의 다운로드만 허용된다.

○ 총인구 등 일부 자료는 일본의 모든 지역이 나타나기는 하지만, 영문 홈페이지에서는 대부분 대도시권에 대한 1km와 500m 격자통계를 제공한다.

□ 일본어 홈페이지(https://www.e-stat.go.jp/gis)를 통해서는 250m, 500m, 1km 크기의 격자통계에 대한 인터렉티브 및 다운로드 서비스가 이용 가능하다.

○ jSTAT MAP 인터렉티브 서비스는 행정구역과 격자 기준 모두 제공되는 우리나라의 통계지리정보서비스와 유사한 개념으로 격자 서비스를 통해 인구, 경제, 사업체, 농림업 관련 정보를 볼 수 있다.

○ 다운로드 서비스의 경우, 메쉬에 대한 경계 데이터(shp format)와 해당 격자통계 값(csv format)을 다운로드할 수 있으며, GIS 소프트웨어의 join 기능 등을 이용하면 벡터(폴리곤) 데이터로 가공이 가능하다.

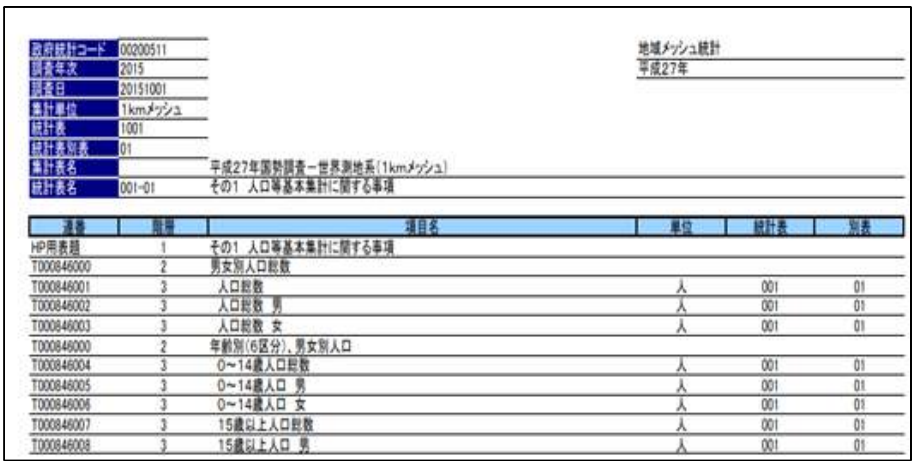
○ 또 데이터 정의서를 다운로드 받을 수 있는데, 이를 통해 데이터의 세부 항목명과 정보보호에 관한 사항을 확인할 수 있다.

○ 인구조사의 경우 250m, 사업체 조사 및 경제센서스는 500m, 농림업 조사는 1km를 최소 격자크기로 하여 제공한다.

○ 정보보호는 격자통계 값이 2 이하인 경우 인접 격자에 해당 격자의 값을 더하는 방식을 사용하고 있다. 그러나 총인구 수(합계, 남, 여), 가구 수, 가구원 수는 정보보호에서 제외된다(www.stat.go.jp/english/data/mesh/pdf/h27outline.pdf).

(a) 일본어 통계GIS 홈페이지

(b) 인터렉티브 격자통계(iSTAT MAP)



政府統計コード	00200511	
調査年次	2015	
調査日	20151001	
集計単位	1kmメッシュ	
統計表	1001	
統計表別表	01	
集計表名		平成27年国勢調査ー世界測地系(1kmメッシュ)
統計表名	001-01	その1 人口等基本集計に関する事項

地域メッシュ統計
平成27年

通番	階層	項目名	単位	統計表	別表
HP用表題	1	その1 人口等基本集計に関する事項			
T000846000	2	男女別人口総数			
T000846001	3	人口総数	人	001	01
T000846002	3	人口総数 男	人	001	01
T000846003	3	人口総数 女	人	001	01
T000846000	2	年齢別(6区分),男女別人口			
T000846004	3	0~14歳人口総数	人	001	01
T000846005	3	0~14歳人口 男	人	001	01
T000846006	3	0~14歳人口 女	人	001	01
T000846007	3	15歳以上人口総数	人	001	01
T000846008	3	15歳以上人口 男	人	001	01

(c) 3차 메쉬 데이터 정의서

출처 : jstatmap.e-stat.go.jp/jstatmap/main/trialstart.html

<그림 2-24> 일본의 격자통계 제공 예시2

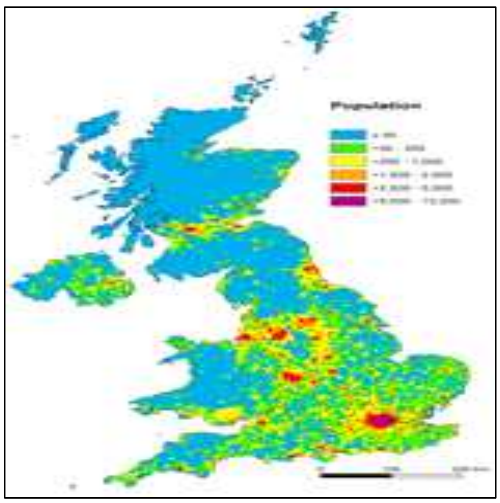
1.11 영국통계청(Office for National Statistics)

□ 공식적인 격자통계는 생산하지 않는 것으로 파악되며, 격자통계 생산을 위한 통계적 방법론에 대한 연구는 진행된 것으로 확인된다.

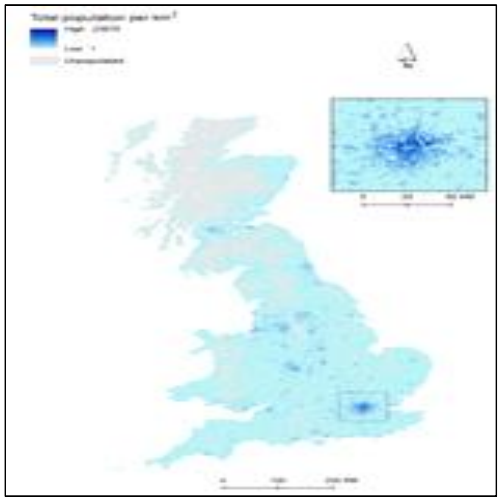
○ 과거 1971년 센서스를 바탕으로 1980년 출간된 통계지도집의 표지를 격자 기반으로 제작할 수준의 방법론 연구가 있었다.

○ 현재 대학의 연구 차원에서 영국 통계청의 자료를 바탕으로 한 격자 기반의 인구자료 만들고 있는 것이 확인된다.

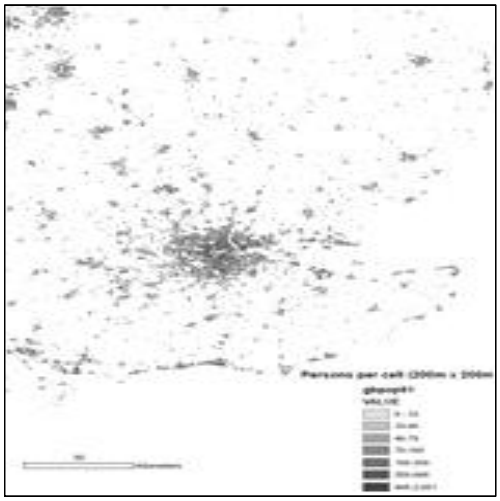
(a) 통계지도집 (1980년 발간)



(b) 인구자료의 격자(5㎞)



(c) 인구자료의 격자(1㎞)



(d) 인구자료의 격자(200m)

<그림 2-25> 영국의 격자통계 제공 예시

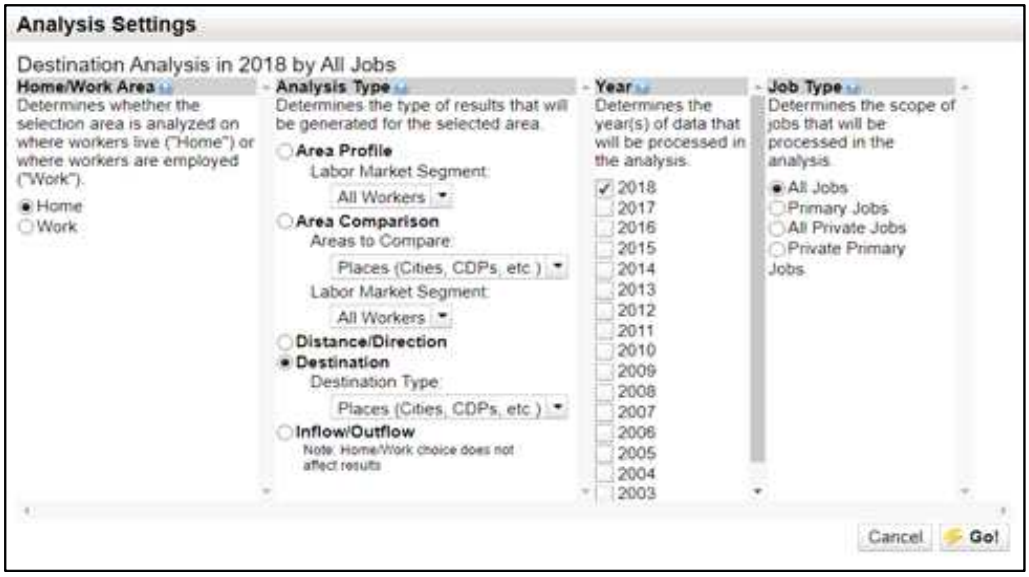
1.12 미국통계청(U.S. Census Bureau)

□ 미국통계청의 경제연구센터(Center for Economic Studies)에서는 종단연구모델(LEHD, Longitudinal Employer-Household Dynamics)에 기반을 둔 지도서비스를 제공한다(onthemap.ces.census.gov).

○ 온더맵(onthemap)은 특정 지역을 기준으로 근로자의 연령별, 소득별, 근로 부문별 상태를 파악하기 위해 마련되었다.

○ 지역 선택 후 분석을 실행하면 옵션 창이 나타나는데, 근로자 수 파악기준, 지역, 연도, 업종 등을 선택할 수 있다.

○ 첫 번째 옵션은 근로자 수의 산정기준을 거주 또는 근로 중에서 선택할 수 있다. 두 번째 옵션은 지역 프로파일(지역 근로자 수를 연령별/임금수준별/업종별로 선택), 지역 비교, 거리와 방향 비교, 근무지역별 근로자 수, 지역 내 유입 근로자와 유출근로자 등의 분석 옵션을 제공한다. 세 번째 옵션은 연도 선택이며, 네 번째 옵션은 공공과 민간 등으로 구분되는 업종기준을 선택하는 것이다.



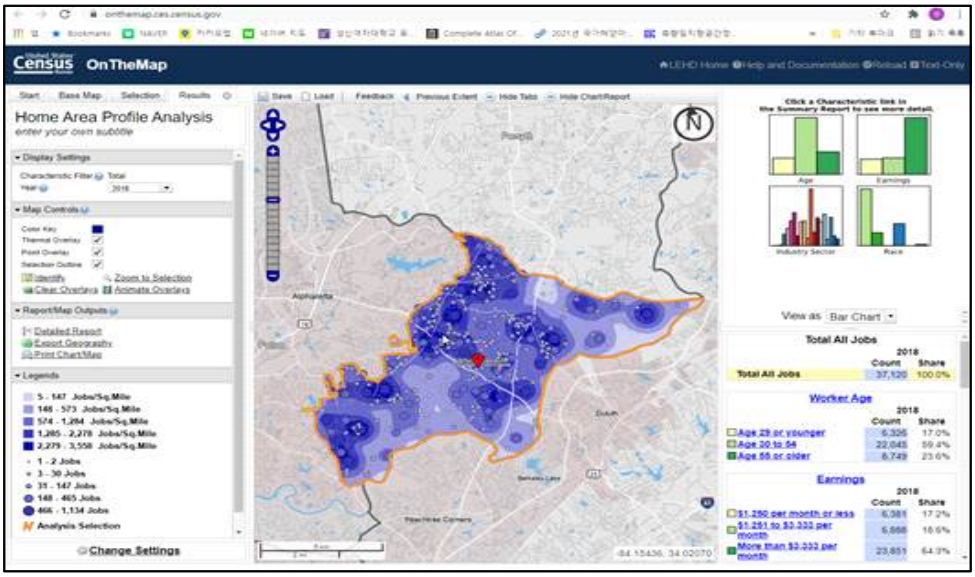
<그림 2-26> 온더맵(onthemap) 분석 옵션 예시

□ 온더맵(onthemap)의 원 자료는 센서스 블록 단위로 수집되며, 지역 내 근로자의 현황을 상세히 나타내고 있다.

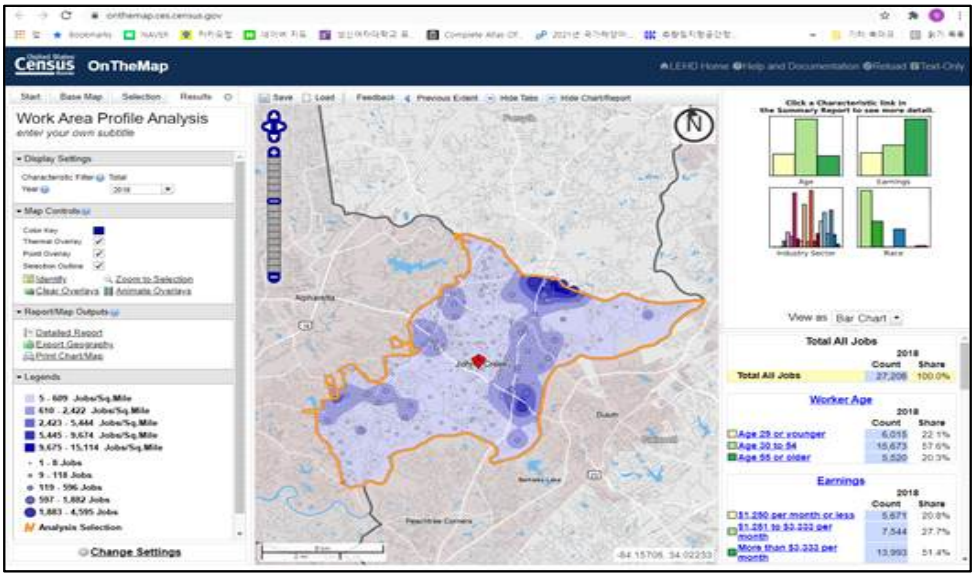
○ 미국 내에는 약 8백만 개의 센서스 블록이 있으며, 센서스 블록은 매년 변화할

수도 있다. 또, 정보보호를 위해 재현자료(synthetic data)를 제작하여 공개된다.

○ 예시로 미국 조지아주 Johns Creek 시의 거주지 기준 근로자 수와 근로지역 기준 근로자 수를 검색한 결과를 제시하였다. 이 지역은 대표적인 대도시 근처의 주거지역으로 주거지역 기준 근로자 수(37,120명)가 근로지역 기준 근로자 수(27,208명)보다 많다는 것을 알 수 있다.



(a) 주거지역 기준 근로자 수

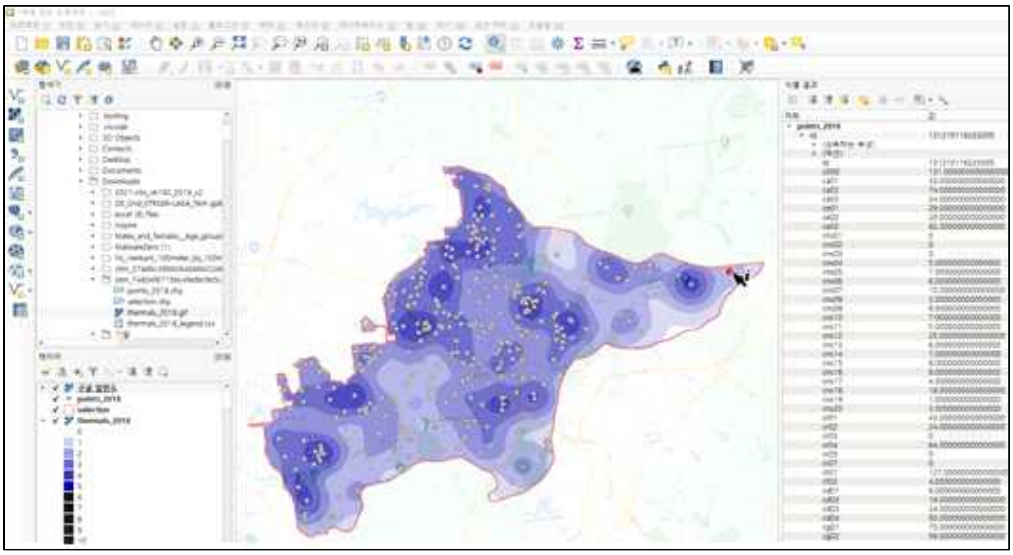


(b) 근로지역 기준 근로자 수

<그림 2-27> 미국 온더맵(onthemap) 격자통계 제공 예시

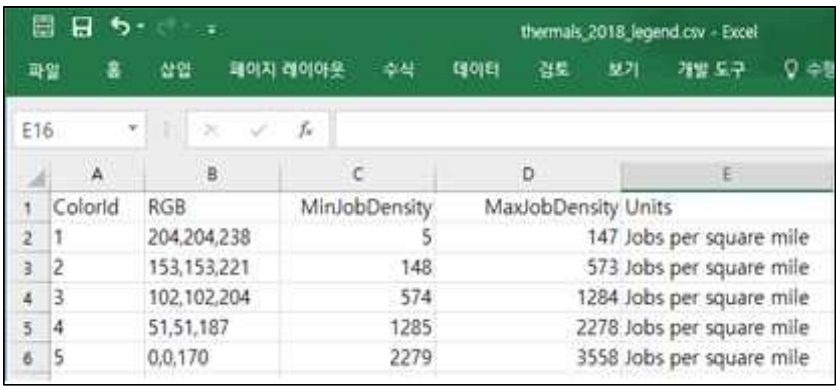
□ 온더맵(onthemap)에서는 데이터를 다운로드 받을 수 있으며 이를 구글지도에 표시할 수도 있다.

○ 점데이터(points 2018)는 각 센서스 블록의 통계 값이며, 붉은색으로 나타난 폴리곤(selection)은 해당 지역의 경계를 나타내며, 래스터 데이터(thermals 2018)는 점데이터를 내삽하여 표현한 것이다.



<그림 2-28> 온더맵(onthemap) 다운로드 데이터의 GIS 소프트웨어 적용

○ 점데이터는 다음과 같은 속성 필드를 가진 센서스 블록의 데이터이다. 격자자료인 래스터 자료는 앨버스 정적 원추도법으로 투영한 10m 격자간격으로 제공되는데 격자 색상의 RGB 값, 그에 해당하는 최소값, 최대값 정보 등을 확인할 수 있다.



	A	B	C	D	E
1	ColorId	RGB	MinJobDensity	MaxJobDensity	Units
2	1	204,204,238	5	147	Jobs per square mile
3	2	153,153,221	148	573	Jobs per square mile
4	3	102,102,204	574	1284	Jobs per square mile
5	4	51,51,187	1285	2278	Jobs per square mile
6	5	0,0,170	2279	3558	Jobs per square mile

<그림 2-29> 격자의 색상, 최솟값, 최댓값, 단위 정보

Residence Area Characteristics (RAC) File Structure				
Pos	Variable	Type	Length	Explanation
1	h_geocode	Char	15	Residence Census Block Code
2	C000	Num	8	Total number of jobs
3	CA01	Num	8	Number of jobs for workers age 29 or younger
4	CA02	Num	8	Number of jobs for workers age 30 to 54
5	CA03	Num	8	Number of jobs for workers age 55 or older
6	CE01	Num	8	Number of jobs with earnings $1250/month or less
7	CE02	Num	8	Number of jobs with earnings $1251/month to $3333/month
8	CE03	Num	8	Number of jobs with earnings greater than $3333/month
9	CNS01	Num	8	Number of jobs in NAICS sector 11 (Agriculture, Forestry, Fishing and Hunting)
10	CNS02	Num	8	Number of jobs in NAICS sector 21 (Mining, Quarrying, and Oil and Gas Extraction)
11	CNS03	Num	8	Number of jobs in NAICS sector 22 (Utilities)
12	CNS04	Num	8	Number of jobs in NAICS sector 23 (Construction)
13	CNS05	Num	8	Number of jobs in NAICS sector 31-33 (Manufacturing)
14	CNS06	Num	8	Number of jobs in NAICS sector 42 (Wholesale Trade)
15	CNS07	Num	8	Number of jobs in NAICS sector 44-45 (Retail Trade)
16	CNS08	Num	8	Number of jobs in NAICS sector 48-49 (Transportation and Warehousing)
17	CNS09	Num	8	Number of jobs in NAICS sector 51 (Information)
18	CNS10	Num	8	Number of jobs in NAICS sector 52 (Finance and Insurance)
19	CNS11	Num	8	Number of jobs in NAICS sector 53 (Real Estate and Rental and Leasing)
20	CNS12	Num	8	Number of jobs in NAICS sector 54 (Professional, Scientific, and Technical Services)
21	CNS13	Num	8	Number of jobs in NAICS sector 55 (Management of Companies and Enterprises)
22	CNS14	Num	8	Number of jobs in NAICS sector 56 (Administrative and Support and Waste Management and Remediation Services)
23	CNS15	Num	8	Number of jobs in NAICS sector 61 (Educational Services)
24	CNS16	Num	8	Number of jobs in NAICS sector 62 (Health Care and Social Assistance)
25	CNS17	Num	8	Number of jobs in NAICS sector 71 (Arts, Entertainment, and Recreation)
26	CNS18	Num	8	Number of jobs in NAICS sector 72 (Accommodation and Food Services)
27	CNS19	Num	8	Number of jobs in NAICS sector 81 (Other Services [except Public Administration])

<그림 2-30> 센서스 블록 점데이터의 속성 필드 정보

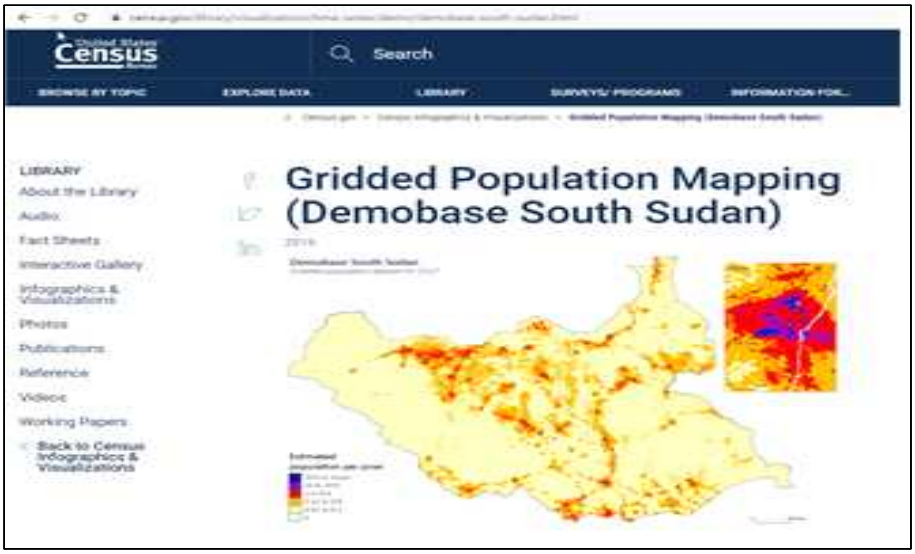
□ 온더맵(onthemap) 데이터는 근로자에 대한 정보를 센서스 블록 기준으로 계산하여 시각적인 효과를 나타내기 위한 것으로 격자통계로서 의미가 있는 것은 아니다.

○ 격자의 크기는 10m이지만 표시되는 통계 값은 1마일(약 1.6㎞)당 근로자 수를 나타내는 것으로, 격자통계라 할 수 없고 시각적 표현을 위한 래스터 데이터로 볼 수 있다.

○ 격자통계가 의미를 가지기 위해서는 해당 지역 내 격자에 해당하는 값을 모두 더하면 해당 지역 전제의 근로자 수가 산출되어야 한다. 즉 10m 격자를 집계하여 100m 격자의 통계 값이 산출되어야 격자통계로 이용될 수 있다.

○ 이와 같은 방식의 사용 이유에 대해 알려진 바는 없지만, 센서스 블록 데이터를 10m 격자로 내삽한 후 대외적으로 공개하지는 않았다. 대신 공개하는 격자의 크기는 1마일로 하여 근로자 수에 해당하는 구간(interval) 값으로 제공한다.

□ 온더맵(onthemap) 서비스 외에도 미국 통계청에서는 센서스 데이터와 위성영상을 이용하여 격자 단위의 상세한 인구지도를 생성하는 Demobase 프로젝트로 남수단, 르완다, 파키스탄, 아이티 등 저개발국에 대한 인구지도를 제작한 바 있다.



출처 : www.census.gov/library/visualizations/time-series/demo/demobase-south-sudan.html

<그림 2-31> 미국 통계청 제작 르완다 100m 격자 인구분포

□ 그 밖에 NASA 등에서 연구 차원으로 생성한 격자통계 사례를 찾아볼 수 있다.

○ NASA의 SEDAC(Socioeconomirc Data and Applications Center)에서는 불규칙한 센서스 블록 데이터를 이용하여 1990, 2000, 2010년 인구, 가구, 주택 등에 관해 레스터 형태의 격자 데이터를 제작하였다.

○ 격자간격은 30초(지상거리 1km 정도)로 하되, 1990년 인구 100만 이상 대도시 39개 지역과 2000년 인구 100만 이상 50개 대도시 지역에 대해서는 격자간격을 7.5초(지상거리 250m)로 설정하여 데이터를 제작하였다.

○ SEDAC은 격자통계를 별도의 벡터 파일로 제공하지는 않고 있으며, 다운로드 역시 불가능하다.

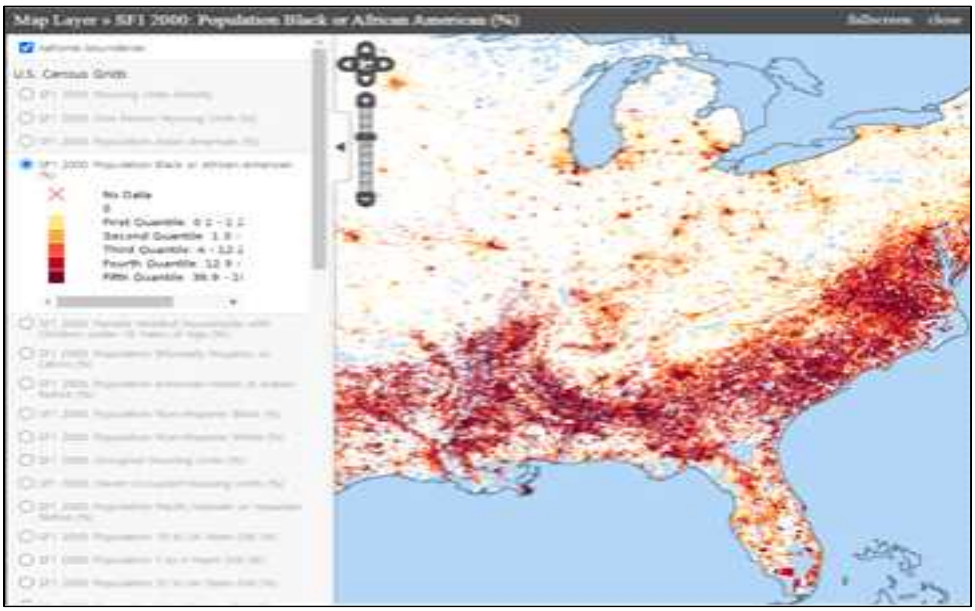
○ 또, 미국은 공간적 범위가 넓기 때문에 직각좌표에 의해 격자를 설정하기보다는 경위도를 기준으로 격자를 설정하여 격자통계를 제작하고 있다.

<표 2-10> SEDAC 격자통계 제공 항목

항목	세부 항목
인구	연령, 인종, 소득, 빈곤, 교육수준, 이민
가구	가구유형, 1인 가구, 18세 미만 여성 가구주, 언어고립 가수 등
주택	점유형태, 차량유무, 준공시기

출처: sedac.ciesin.columbia.edu/data/collection/usgrid

<그림 2-32> 2010년 흑인 비율 pdf 버전



출처: sedac.ciesin.columbia.edu/data/collection/usgrid

<그림 2-33> 2010년 흑인 비율 leaflet 이용 서비스 버전

2. 변동형 격자(varying size grid) 체계

□ 변동형 격자 체계는 격자 크기의 기저(base)를 먼저 결정한 후, 자료 제공을 위한 격자의 크기를 다양하게 바꾸는 것을 허용하는 방식이다.

○ 변동형 격자 체계에 해당하는 예시로 쿼드트리(quadtree) 방법론과 적용 사례를 소개하고자 한다.

□ 쿼드트리 방법론은 데이터를 1㎞ 격자를 기반으로 다양한 크기[2]의 격자로 생성함으로써 기밀성(confidentiality)을 보호하면서도 훨씬 더 정밀한 격자를 만들 수 있도록 유도한다.

○ 쿼드트리 방법론은 기본적으로 1㎞ 격자를 순차적으로 4개의 작은 격자로 분할하고, 필요에 따라 4개의 격자를 집계하여 보다 큰 격자로 만드는 방식이다.

○ 이 방법론에서 기저로 사용되는 격자의 크기는 62.5m이며, 이는 1km 격자를 보다 작은 크기의 격자로 4번 분할하였을 때 생기는 격자의 크기에 해당한다.

○ 따라서 쿼드트리 방법론 적용 시 생성 가능한 격자의 크기는 62.5m, 125m, 250m, 500m, 1㎞ 등으로 다양하게 소 격자를 만들어 볼 수 있다.

○ 이 방법론은 기저 격자인 62.5m 격자를 기반으로 k-익명성이 보장되거나 1km에 도달할 때까지 집계하여 제공하는 방식을 취한다. 이때 임계값 k의 기준은 통계 제공기관에서 결정하게 된다.

○ 따라서 쿼드트리 방법론은 임계값보다 작은 관측치를 가진 셀들은 상위 수준으로 집계되도록 하여 특정 격자에 포함되는 관측치의 수가 정해진 임계값 k보다 작지 않도록 하는 k-익명성 모델과 유사한 점이 있다.

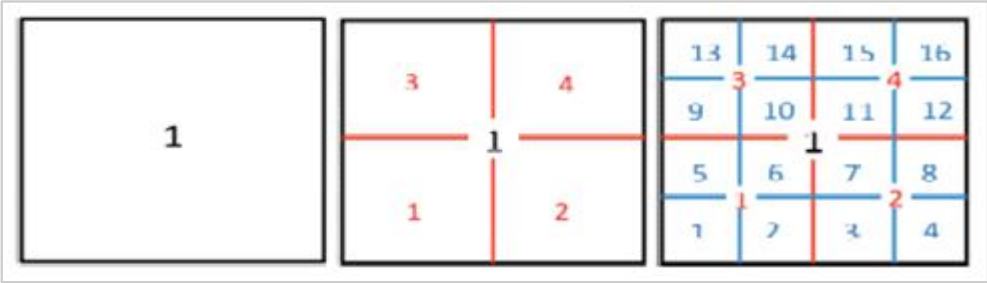
□ 쿼드트리 방법론에서 사용되는 셀 분할 방법의 예시는 <그림 2-34>와 같다.

○ 초기 셀이 4개의 더 작은 셀로 분할되면, 분할된 4개의 셀들은 왼쪽 아래에서 시작하여 왼쪽에서 오른쪽으로, 아래쪽에서 위쪽 순서로 숫자가 매겨진다.

○ 관측치가 많은 격자는 정확한 데이터를 제공하기 위해 보다 작은 격자로 나눠지고, 관측치가 작은 격자는 노출위험을 피하기 위해 보다 큰 격자로 합쳐진다.

---

2) 1㎞, 500m, 250m, 125m, 62.5m



<그림 2-34> 쿼드트리 분할 예시

□ 쿼드트리 방법론을 실제로 적용한 사례로 스페인 카탈로니아 지역에 있는 데이터 분석 및 모델링 연구기관(The Statistical Institute of Catalonia, Idescat)의 자료를 검토하였다.

○ Idescat에서는 기밀성 요건과 충동하지 않고 최대한 정확하게 지리 공간 데이터를 공개하고 시각화하는 기술에 대해 자동화 연구를 수행하고 있다.

○ 해당 방법론은 스페인 북동부 지방인 카탈로니아의 2014년 인구등록 데이터에 적용되었으며, 유럽연합통계청의 표준 격자 체계에도 활용될 수 있도록 하였다.

○ 쿼드트리 방법론의 셀 분할 적용을 위한 자세한 절차는 다음과 같다.

◦ (1단계) 전 지역에 걸쳐 1㎢ 격자를 생성한다.

◦ (2단계) 최소 셀 크기를 62.5㎡로 설정한다.

◦ (3단계) 격자 셀 크기가 최소 셀 크기보다 큰 경우 (a)-(d)를 반복한다.

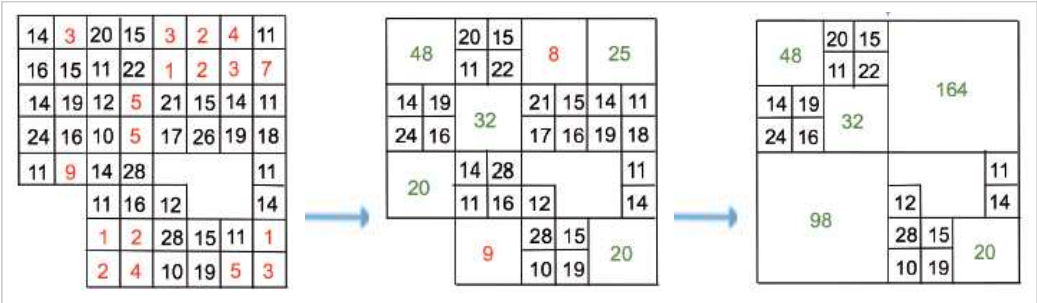
- (a단계) 각 셀에 대해 교차를 수행한다.
- (b단계) 각 대응 셀에 대해 개별 정보들을 집계한다.
- (c단계) 셀 내에 데이터가 없는 격자 셀을 버린다.
- (d단계) 새로운 격자를 만들기 위해 각 격자 셀을 4개의 셀들로 부분 분할한다.

□ 쿼드트리 방법론을 적용하여 다양한 크기를 지닌 격자 생성 절차의 예시는 다음과 같다.

○ 셀 안에 기재된 숫자는 해당 지역에서 주어진 기준을 충족시키는 관측치의 수를 나타낸다. 임계값의 기준 k는 10으로 가정하였다.

○ 왼쪽에 있는 첫 번째 격자는 임계값보다 작은 셀을 여러 개 가지며, 이 값들은 붉은색으로 표시되었고, 이 셀들은 쿼드트리 구조에서 상위 수준으로 그룹화된다. 이러한 절차는 ① 격자 내 모든 셀 값들이 임계값보다 크거나, ② 숨겨지는

셀의 크기가 1㎢에 도달할 때까지 반복된다.



<그림 2-35> 쿼드트리 방법론의 개념적 예시(익명 임계값=10 가정)

□ 쿼드트리 방법론은 집계와 익명성 보장을 위한 이중 임계값 체계를 활용하여 정보는 보호하되 정보손실을 최소화한다.

○ 집계 임계값(aggregation threshold)은 셀 집계를 멈추기 위한 기준 값을 의미하며, 각 셀의 관측치 수를 집계 기준으로 이용한다.

○ 익명 임계값(anonymity threshold)은 k-익명성을 보장하기 위한 기준 값을 의미하며, 익명 임계값에 도달하지 않을 때는 N/A(Not Available)로 표기된다.

○ GEOSTAT 1B(2013)에서는 익명성 임계값을 10, 30, 50, 100 등 여러 개로 설정하여 평가하는 것을 권장하였다.

○ 집계 임계값을 낮게 설정할 경우 크기가 작은 격자가 유지될 가능성이 높아 해상도(resolution)는 높을 수 있으나, 익명 임계값으로 인해 N/A 처리되는 격자가 많아져서 이용 가능한 정보는 줄어들게 된다.

□ 쿼드트리 방법론은 계층적 데이터 구조를 나타내는 기법의 하나로 발전해 왔다.

○ Hunter(1978)는 지도 제작 어플리케이션에 사용되는 데이터를 나타내기 위해 쿼드트리 방법론을 검토했다.

○ Strobl(2005)은 오스트리아의 인구 조사 데이터를 시각화하기 위한 방법으로 쿼드트리 방법론을 적용하여 제안했다.

○ Kaminger와 Meyer(2007)는 쿼드트리 방법론이 자동화하기 쉽고 정보의 노출제어가 가능하면서도 데이터의 전체적인 개요를 제공한다고 하였다.

○ Kardos et al.(2005)은 육각형 또는 마름모 쿼드트리 방법론을 제안하기도 했다.

3. 격자통계 제공의 시사점

□ 조사 결과를 토대로 국내 기관의 격자통계 서비스 현황을 정리하면 다음과 같다.

○ 국내 기관 중 공식통계로 격자단위 서비스를 제공하는 곳은 통계청과 국토정보원으로 파악되며, 두 기관 모두 인구, 주택 관련 자료를 서비스한다.

○ 국토지리정보원 국토통계지도에서는 인구, 건물, 토지 등 조사 자료와 이를 가공하여 정의한 국토지표를 격자통계로 제공하고, 격자의 크기는 국토지표를 제외하고 100m부터 100㎞까지로 하여 총 6종을 제공한다.

○ 통계청 통계지리정보서비스에서는 인구, 주택, 사업체 등 조사 자료를 격자통계로 제공하고, 격자의 크기는 100m부터 100㎞까지로 하여 총 4종을 제공한다.

○ 두 기관 모두 정보보호에 대한 기준 값을 인구 자료의 경우 5로 정하고 통계 값 5 이하인 경우 N/A 표시하여 제공한다.

<표 2-11> 국토지리정보원 국토통계지도 격자통계 제공 현황

<table>
<tr><th>분야</th><th>대분류</th><th>소분류(세부내용)</th><th>격자
크기</th><th>정보
보호</th></tr>
<tr><td rowspan="7">인구</td><td>총인구</td><td>성별(전체, 남자, 여자)</td><td rowspan="13">100km,
10km,
1km,
500m,
250m,
100m
(6종)</td><td rowspan="7">5 이하
N/A
처리</td></tr>
<tr><td>유소년인구</td><td>성별(전체, 남자, 여자)</td></tr>
<tr><td>생산가능인구</td><td>성별(전체, 남자, 여자)</td></tr>
<tr><td>고령인구</td><td>성별(전체, 남자, 여자)</td></tr>
<tr><td>유아인구</td><td>성별(전체, 남자, 여자)</td></tr>
<tr><td>학생인구</td><td>초등/중등/고등(전체, 남자, 여자)</td></tr>
<tr><td>연령별인구</td><td>20대/…/100세 이상(전체, 남자, 여자)</td></tr>
<tr><td rowspan="6">건물</td><td>일반</td><td>건축면적, 건폐율, 용적률, 대지면적,
연면적, 높이, 지상/지하 층수</td><td rowspan="13">처리
없음
(가공
데이터)</td></tr>
<tr><td>건축물</td><td>사용 승인일, 건축물 사용연한</td></tr>
<tr><td>주용도</td><td>건축물 수(공동, 단독), 주 용도 코드</td></tr>
<tr><td>구조</td><td>건축물 수(목구조, 철근구조),
주 용도 코드</td></tr>
<tr><td>개별주택</td><td>개별주택 가격, 주거용도 면적</td></tr>
<tr><td>토지</td><td>공시지가</td><td>-</td></tr>
<tr><td rowspan="4">국토
지표</td><td>토지와 주택</td><td>노후건물/주택, 토지이용 압축도,
토지이용 복합도</td><td rowspan="4">500m</td></tr>
<tr><td>생활과 복지</td><td>공원, 공공시설 등에 대한 도로 이동 거리</td></tr>
<tr><td>국토 인프라</td><td>고속도로IC, 철도, 주차장 이동 거리</td></tr>
<tr><td>환경과 안전</td><td>경찰서, 소방서, 지진대피소까지 이동 거리</td></tr>
</table>

주: 1km(또는 500m) 이하의 격자지도는 시도와 시군구를 선택하여야 조회 가능

<표 2-12> 통계청 통계지리정보서비스 격자통계 제공 현황

분야	대분류	소분류(세부내용)	격자 크기	정보 보호
총조사 주요 지표	총인구	총인구수	100km, 10km, 1km, 100m (4종)	처리 없음
	가구	가구수		
	주택	주택수		
	사업체	사업체수		
인구 주택 총조사	조사연도	2000년도, …, 2019년도		5 이하 N/A (not available) 처리
	인구	성별(전체, 남, 여)		
		연령(A 이상 B 미만, 5세 단위)		
	가구	세대구성(1세대가구, …, 4세대이상가구, 1인가구, 비혈연가구 등)		
	주택	주택유형(단독주택, 아파트, 연립주택, 다세대주택, 비주거용건물 등)		
		건축연도		
		연면적(A 초과 B 이하)		
전국 사업체 조사	조사연도	2000년도, …, 2019년도		3 이하 N/A 처리
	사업체수	-		
	종사자수	-		

□ 해외 국가별 격자통계 서비스 형태를 인터렉티브(interactive) 서비스와 다운로드 서비스로 분류하여 살펴보면 다음과 같다.

○ 사용자가 사이트에 접속하여 원하는 내용을 조회하면 바로 확인이 가능한 인터렉티브 서비스와 지리정보 프로그램에 연계해서 사용할 수 있도록 데이터를 직접 다운로드 받는 서비스로 구분된다.

○ 공통적으로 인구와 주택 자료는 대부분의 나라에서 서비스하고 있지만, 국가별로 구체적인 서비스 항목에는 차이가 있다.

○ 인터렉티브 서비스와 다운로드 서비스에서 최소 격자의 크기가 상이한 경우도 있으나, 동일한 경우가 더 많다.

○ 다운로드 서비스는 유료와 무료로 나뉜다. 다운로드 시 격자의 통계 값과 격자 위치를 별도의 파일로 제공하는 경우가 있으며, 양자를 통합하여 하나의 파일로 제공하는 경우도 있다.

<표 2-13> 해외 국가별 격자통계 서비스 형태

국가	제공자료	최소 격자	서비스 형태		
			인터렉티브 서비스 (최소격자)	다운로드	
				서비스 (최소격자)	파일 형식
일본	인구, 주택	250m	○ (250m)	○ (250m)	위치:shp 통계:csv
	경제 및 사업체	500m	○ (500m)	○ (500m)	
	농림업	1km	○ (1km)	○ (1km)	
유럽연합 통계청	인구	1km	-	○ (1km)	gpkg
네덜란드	인구, 주거, 에너지, 사회안전, 시설밀도	100m	○ (100m)	○ (100m)	shp, gpkg
노르웨이	인구, 가구, 주택, 산업, 농업	250m	○ (250m)	○ (250m)	위치:shp 통계:csv
핀란드	인구, 교육, 가구, 건물, 소득	250m	○ (1km)	○ (250m)	유료
오스트리아	인구, 건물, 주거, 노동	100m	○ (250m)	○ (100m)	유료
독일	인구, 건물	100m	○ (1km)	○ (100m)	위치:gpkg 통계:csv
	농업통계	5km	○ (5km)	-	-
에스토니아	인구, 주택, 경제	100m	○ (1km)	○ (1km)	shp, kml

□ 해외 국가별 격자통계 서비스 현황을 격자 크기와 주요 항목을 중심으로 정리하여 주목할 만한 사항을 살펴보았다.

○ 일본의 경우 통계조사 명칭을 기준으로 항목이 제공되고 있으며, 국세조사 자료를 통해 종사상 지위, 취업세대에 해당하는 인구 수 등을 공개하고 있다.

○ 유럽연합통계청에서는 인구 자료만 공개되고 최소 격자로 1㎞를 사용해서 개별 국가보다는 다소 큰 크기를 사용한다.

○ 네덜란드는 격자의 크기가 다양하지 않지만 가장 많은 자료를 격자통계로 제공하고 있다. 특히 다른 국가에서는 서비스하지 않는 소득, 에너지, 사회안전, 사회기반시설에 대한 서비스 수준의 자료까지 공개하고 있다.

○ 오스트리아의 경우는 격자의 크기를 100m, 250m, 500m, 1㎞, 2㎞, 5㎞ 등으로 하여 상당히 다양하게 제공하고 있다.

○ 노르웨이와 핀란드는 제공 격자의 크기가 250m, 1㎞, 5㎞ 등으로 동일하였다.

○ 독일의 경우 인구 관련 자료에서 출생국가, 종교, 국적, 시민권자 수 등을 제공한다는 점에서 다른 국가들과 차별점이 있다.

○ 아래 표에 제시된 기준연도는 최소 격자단위를 제공하는 데이터의 연도이다. 최근 자료를 즉시 반영하는 국가도 있지만, 국가적 차원의 센서스 시행 주기인 5~10년 단위로 격자통계를 작성하는 국가도 있다.

○ 거의 매년 격자통계를 생산하는 대표적인 국가는 네덜란드, 노르웨이, 핀란드를 들 수 있다. 반면 일본, 오스트리아, 독일, 에스토니아는 센서스 조사 결과를 바탕으로 격자통계를 생산하고 있다.

○ 유럽연합 통계청을 제외하고는 10㎞ 이상의 데이터를 제공하지 않고, 대부분의 나라에서 최소 250m 격자보다 세밀한 격자통계를 생산하며 1㎞ 격자를 가장 많이 사용하고 있다.

○ 제공 항목별 범주를 표시해 주는 국가도 있지만, 핀란드, 오스트리아, 에스토니아의 경우와 같이 항목에 대한 메타데이터가 표시되지 않는 나라는 세부적인 항목을 정확히 파악하는데 한계가 있다.

<표 2-14> 해외 국가별 격자통계 서비스 현황

국가	조사 분야	기준 연도	격자크기						주요 항목
			100m	250m	500m	1km	2km	5km	
일본	국세 조사	2015		o	o	o			총인구(합산, 남, 여), 성별/연령별 인구, 외국인, 세대수, 종사상 지위, 인구이동, 통근통합, 취업세대
	경제 센서스	2016			o	o			산업 대분류별 사업체, 종사자
	농림업 센서스	2015				o			업체, 고용, 주업/부업, 연령별종사자, 자급농가, 판매농가, 임업가
유럽 연합	인구	2018				o	o	o	인구, 육지비율, 바다(국경)까지 거리 * 10km, 100km 격자 제공
네덜 란드	인구	2020	o		o				총인구(합, 남, 여), 연령별 인구, 출생인구
	가구	2020	o		o				가구, 세대구성, 평균가구원
	주택	2020	o		o				주택, 무인주택, 건설연도, 가격, 자가/ 임대, 주택형태

국가	조사 분야	기준 연도	격자크기						주요 항목
			100m	250m	500m	1km	2km	5km	
네덜란드	소득	2020	o		o				평균소득, 저소득, 고소득
	에너지	2020	o		o				평균가스소비, 전력소비
	사회 보장	2020	o		o				실업수당
	생활 서비스	2020	o		o	o		o	병원, 소방서, 백화점 등 생활서비스 제공 기관/서비스 제공 * 3km 격자 제공
노르웨이	인구	2019		o		o		o	인구
	주택	2021		o		o		o	주택, 휴가용 건물
	건물	2021		o		o		o	병원, 요양원, 건물질량
	농업	2015		o		o		o	농업용토지, 농업기업
	산업	2021		o		o		o	사업체
핀란드	인구	2019		o		o		o	총인구(합, 남, 여), 연령별 인구, 평균연령
	교육	2019		o		o		o	18세 이상 인구, 교육수준
	소득	2018		o		o		o	평균소득, 소득유형, 구매력
	가구	2019		o		o		o	가구, 세대구성, 주거유형
	건물	2019		o		o		o	건물, 용도별 건물
	노동	2018		o		o		o	총종사자, 산업별 종사자
	취업	2018		o		o		o	근로유형
오스트리아	기초 데이터	2011	o	o	o	o	o	o	총인구, 가구, 주간인구, 통근인구
	노동	2011		o	o	o	o	o	종사자, 상업별 종사자
	건물	2011		o	o	o	o	o	건물, 주민, 건물유형, 건축시기, 연면적, 방(room)수
	인구 고용	2011		o	o	o	o	o	성별/연령별 인구, 출신국가, 성별/고용상태별 인구
	교육	2011		o	o	o	o	o	학력, 공교육 참여
	가구	2011		o	o	o	o	o	세대유형, 연령별 가구구성,
독일	인구	2011	o			o			인구, 연령별인구, 혼인상태, 출생국가, 종교, 국적, 시민권

국가	조사 분야	기준 연도	격자크기						주요 항목
			100m	250m	500m	1km	2km	5km	
독일	가구	2011	o			o			가구유형, 가구원
	주택	2011	o			o			소유현황, 면적, 방(room)수, 건물유형, 건축시기, 난방유형, 건물크기
에스토니아	인구	2011	o		o	o			총인구/성별, 성별/연령별 인구
	가구, 주택	2011	o		o	o			가구원, 주거면적
	건물	2011	o		o	o			건물유형
	생활	2011	o		o	o			교육수준, 혼인상태, 고용여부

□ 해외 국가별 격자통계 서비스의 특징과 정보보호와 관련한 사항을 요약하면 다음과 같다.

○ 국가마다 정보보호에 대한 기준은 대부분 마련되어 있으나, 해당 수준은 다양하다. 격자의 크기에 따라 정보보호 수준이 다른 경우도 있으며, 동일 국가라 하더라도 시기에 따라 정보보호의 기준이 바뀌기도 한다.

○ 제공 항목에 따른 정보보호 수준 역시 다양하다. 일본, 핀란드, 오스트리아에서는 총인구, 가구 수는 정보보호 대상에서 제외하였지만, 특성별 인구 수(예를 들면 5세 단위 연령대별 성별 인구 등)는 정보보호의 대상이다.

○ 정보보호 방법에서의 공통점으로는 총계 값의 경우 조사된 값에 대한 정보보호 없이 그대로 제공한다는 것이다. 개별 항목별 통계 값은 각 국가에서 정한 기준 값보다 작은 경우에 N/A로 제공함을 확인할 수 있다.

○ 정보보호 방법에서 N/A 처리하는 국가가 많기는 하나, 일본의 경우 인접 격자에 추가하는 방안을 사용하는 국가도 있다. 이때 미공개 사유가 정보보호 차원인지, 국가적 기밀에 해당하는 미공개 지역인지 등의 사유를 명확하게 밝히지 않는 경우가 대부분이다.

○ 각국의 통계자료는 주로 해당 국가의 자국어로만 제공되며, 특히 정보보호에 관한 사항은 별도의 문서 형태로 나타나는 경우가 많아 이를 모두 조사하는데 어려움이 있다.

<표 2-15> 해외 국가별 격자통계 서비스 특징 및 정보보호 방법

국가	서비스 특징	정보보호 방법
일본	국가적 차원의 메쉬 코드 활용	통계값 2 이하인 경우 인접 격자와 합산하여 제공
유럽 연합	속성정보로 인구, 육지비율, 해안(국경)까지 거리 등 제공	해당 없음
네덜 란드	가장 다양하고 방대한 자료 수록	인구(가구) 4 이하인 경우 N/A 처리
노르 웨이	가장 최근연도 자료 공개 (주택, 건물, 산업 등)	통계값 3 이하인 경우 N/A 처리
핀란드	다운로드 서비스는 라이센스 단위로 제공	가구 수 10 이하인 격자의 소득정보, 인구 수 3 이하인 경우 N/A 처리
오스 트리아	인터렉티브 서비스는 양호하나, 다운로드 서비스 관련 정보 없음	인구 4 이하, 건물 3 이하, 주거 3 이하인 경우 N/A 처리
독일	100m, 1km 격자만 제공	인구 2 이하인 경우 N/A 처리
에스 토니아	인터렉티브 뷰어에서 다운로드 서비스 제공	인구 3 이하인 경우 N/A 처리

□ 이상의 결과를 토대로 통계지리정보서비스의 격자통계 제공을 위한 시사점을 도출해 보면 다음과 같다.

○ 첫째, 현행 통계지리정보서비스의 격자통계 서비스는 우수한 편이나 제공 항목을 보다 다양화할 필요가 있다.

- 단순한 인구 구조와 함께 인구동태, 인구이동 등에 관한 항목이나, 건물의 물리적인 측면 이외의 항목에 대해서도 검토해 볼 필요가 있다.

<표 2-16> 일부 국가에서 공통적으로 제공하는 항목

제공항목	세부내용	비고
노동	종사상 지위, 실업수당, 근로유형(고용상태, 고용여부),	일본, 핀란드, 네덜란드, 오스트리아, 에스토니아
인구	통근인구	일본, 오스트리아
농업	농업용 토지면적, 농업기업	일본, 노르웨이
교육	교육수준, 공교육 참여	핀란드, 오스트리아, 에스토니아
소득	평균소득, 소득유형, 구매력	네덜란드, 핀란드

<표 2-17> 특정 국가에서 특징적으로 제공하는 항목

국가	항목(세부내용)
네덜란드	에너지(평균가스소비, 전력소비), 서비스(병원, 백화점 등 생활서비스 제공)
오스트리아	건물(방(room) 수)
독일	건물(방(room) 수, 난방유형), 인구(국적, 종교, 출생국가, 시민권자)
유럽연합	육지비율, 바다(국경)까지 거리

○ 둘째, 현행 SGIS에서 제공하는 4종류(100m, 1㎞, 10㎞, 100㎞)의 격자에서 그 크기를 보다 다양화할 필요가 있다.

- 최소 격자의 크기를 100m 단위로 하는 국가도 있지만, 250m 격자를 최소 크기로 하는 국가도 상당수 있다. 또, 1㎞와 10㎞의 중간 크기로 5㎞를 제공하는 국가도 상당수 있다.

- 통계지리정보서비스와 같이 격자의 크기가 10배씩 증가하는 것은 격자통계의 활용성이 낮아지는 주요한 원인 중 하나일 수 있다. 따라서 500m, 250m, 50m 등 다양한 수준의 격자통계에 대한 검토가 필요하다.

<표 2-18> 해외 국가별 제공 격자크기 현황

국가	제공 격자크기						비고
	100m	250m	500m	1km	2km	5km	
일본		o	o	o			
유럽연합				o	o	o	10㎞, 100㎞ 제공
네덜란드	o		o				3㎞ 제공
노르웨이		o		o		o	
핀란드		o		o		o	
오스트리아		o	o	o	o	o	100m 제공(기초데이터)
독일	o			o			
에스토니아	o		o	o			

○ 셋째, 현행 SGIS에서는 5(인구, 가구, 주택) 또는 3(사업체) 이하인 경우 N/A 처리하고 있으며, 이러한 정보보호 기준에 대한 면밀한 검토가 필요하다.

- 해외 국가에서는 익명성 보장의 기준 값으로 인구 자료는 2 또는 3을 사용하는 경우가 다수 있는 것으로 파악되므로, 현행 기준보다 완화된 기준 적용에 대한 검토가 필요하다.
- 정보보호를 위한 방법으로는 대부분의 국가에서 N/A 처리 방식을 선택하고 있어 현행 SGIS의 정보보호 방법과 차이가 없음을 확인하였다. 그러나 일본의 경우처럼 총량을 보전한다는 측면에서 인접 격자와 합산하여 제공하는 방식(예를 들면 쿼드트리 방법론 등)에 대한 검토도 의미 있을 것으로 판단된다.
- 정리하자면 정보보호를 적용할 항목과 하지 않을 항목을 구분하고, 격자 크기에 따라 정보보호 기준을 설정해야 한다.
- 또 정보보호가 필요한 경우 N/A로 표시할 것인지, 아니면 다른 격자와 합산하는 쿼드트리 방식을 취할 것인지 등에 대한 검토도 필요하다.

□ 마지막으로 격자통계 사용자의 요구(needs)를 정리해 보면 다음과 같다.

○ 국내에서는 이와 관련한 실태조사 또는 연구가 진행된 바가 없는 것으로 확인되어, 유럽연합 통계청의 GEOSTAT 프로젝트 결과를 인용하였다.

○ GEOSTAT(2011)에 따르면, 격자통계 사용자의 대다수는 공간 분석(spatial analysis)을 수행하는 것을 목적으로 격자통계를 사용하는 것으로 파악된다.

○ 사용자의 관심 영역과 격자 크기 간에는 명확한 상관관계가 존재하였다. 단일 국가 및 소규모 지역을 분석하는 경우에는 최대 250m 이하의 작은 격자 크기를 선호하고, 국가 간 또는 대륙 단위의 데이터를 필요로 하는 사용자는 1km의 격자 크기를 선호하는 것으로 조사되었다.

○ 노출 제어의 관점에서 공개할 수 없는 값은 고정된 최솟값으로 대체하는 방법을 작은 격자를 더 큰 격자로 집계(aggregation)하는 방법보다 더 선호하는 것으로 파악된다.

○ 향후 국내에서도 이와 관련한 연구를 진행해 보는 것은 의미가 있을 것으로 생각된다.

제 3 장

격자 체계별 노출제어와 정보보호

1. 지리 데이터(geographic data)의 정보보호

□ 공간 데이터(spatial data)를 공개할 때 정보보호의 측면에서 두 가지 문제 상황이 발생할 수 있다.

○ 개별 지리 좌표(individual geographic coordinate)는 그 자체가 기밀일 수 있으며, 개별 위치가 특정 시점에 식별되지 않도록 익명화해야 한다. 예를 들면 위치 인식 어플리케이션에서 어떤 장치를 사용하여 특정 목적을 가지고 사용자의 지리적 위치를 인식하고 처리하는 방식이다(Beresford & Stajano, 2003).

○ 반면 지리 좌표가 통계 DB의 많은 속성 값 중 하나로 공개되어도 문제 상황이 발생할 수 있다. 공간 좌표의 공개는 개별 개인 정보의 재식별을 초래하여 노출위험이 발생하므로 익명화해야 한다(Armstrong & Ruggles, 2005). 예를 들면 개별 위치에 대한 정보가 포함된 지도를 게시할 때 역 지오코딩(reverse geocoding)[3]을 통해 개별 주소를 표시할 수 있기 때문이다.

□ 개별 데이터 기록을 배포해야 할 경우 지오매스킹 기법(geomasking technique)을 적용해야 한다.

○ 지오매스킹 기법은 실제 위치를 찾을 수 없도록 각 지점의 위치를 개별적으로 변경한다(Armstrong et al., 1999; Hampton et al., 2010).

○ 서로 다른 유형의 임의 교란(random perturbation)을 모든 개별 지점에 적용할 수 있어 실제 관측된 지점과 지오매스킹된 지점이 일치하지 않게 된다.

○ 교란을 직접적으로 적용하면 재식별이 어려워 노출위험은 감소하지만, 동시에 공간 분석을 수행하기 위한 데이터의 유용성이 감소한다.

---

3) 지오코딩(geocoding)은 특정 지점에 관한 주소나 지명을 위도, 경도의 좌표로 변환시키는 작업이다.

○ 이러한 측면에서 개별 데이터의 공간 분포를 변경하면 추가적인 분석을 수행할 때 편향이 생길 수 있고 분석 결과의 왜곡이 발생하지 않도록 하는 것이 더욱 어려워질 수 있다(Duncan et al., 2001; Kwan et al, 2004; Reiter, 2012).

□ 통계지리정보서비스와 같은 지도의 시각화에 있어 지오매스킹 기법은 최선의 접근법은 아닌 것으로 보인다.

○ 지오매스킹 기법을 적용함으로써 시각적으로 역설적인 모순이 발생할 수 있기 때문이다. 예를 들어 사람이 거주할 수 없는 지역인 강, 바다, 산 등에 집계된 인구 수를 표시하게 되면 시스템 전체의 신뢰성을 감소시킬 수 있다.

□ 대안으로 생각할 수 있는 것은 집계하는 방법을 통해 재식별을 피할 수 있도록 개별 정보를 그룹화하는 것이다.

○ 이때 마이크로 데이터를 집계하여 정보보호를 제공하는 기본 모델로 정의된 것이 바로 k-익명성(anonymity)이다.

○ k-익명성 모델에서는 한 개인에 대한 정보가 동일한 DB 내에서는 적어도 (k-1)명의 개인과 구별될 수 없도록 한다. 즉, 동일한 정보를 가진 개인이 적어도 k명이 되도록 그룹을 만들어 집계하게 된다.

□ 격자 데이터의 정보보호 측면에서 k-익명성은 특정 격자에 속한 개인의 정보가 적어도 (k-1)명의 다른 개인과 구분될 수 없다는 것을 의미한다.

○ 따라서 지리적 집계 방법은 집계 과정에서 생성된 모든 구역이 최소 k명을 포함하도록 설계되어야 한다(Vu et al., 2012).

○ 이때 집계된 정보의 품질은 집계를 위한 격자의 크기와 격자에 의해 분할되는 경계에 따라 달라진다.

○ 특히, 편향의 크기를 비교해 보면 격자의 크기가 격자의 형태보다 훨씬 더 중요하다(Briant et al., 2010; Andersson et al., 2012).

○ 격자의 형태는 다양하게 존재할 수 있는데 이에 대한 자세한 내용은 부록을 참조할 수 있다.

○ 마지막으로, 자료 이용자 입장에서는 k-익명성 확보를 위해 보다 큰 격자로 집계하는 것보다 특정한 최소값으로 대체하는 것을 더 선호하는 것으로 조사되었다(GEOSTAT, 2011).

2. 고정형 격자의 노출제어와 정보보호

□ 고정형 격자의 노출위험을 검토하기 위해 2019년 기준 인구주택총조사, 전국사업체조사 자료를 입수하였다.

○ R 패키지에서 분석 가능하도록 대전광역시에 해당하는 국가지점번호 격자 '다바' 지역의 자료로 한정하여 제공받았다.

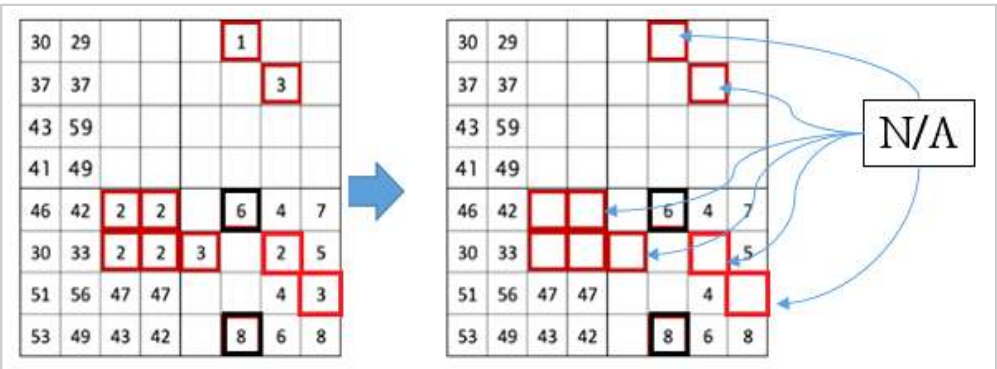
○ 추출된 격자의 크기는 100m, 500m, 1km, 10km, 100km로 하였으며, 250m 격자의 경우 기술적 문제로 제공받는 자료에서 제외되었다.

○ 노출위험을 제어하기 위해 k-익명성 모델을 활용하였으며 다양한 k 값을 적용하였다. 이때 익명성이 확보되지 않는 격자에 대해서는 N/A 처리함으로써 정보보호가 되도록 하였다.

○ 따라서 고정형 격자의 공표기준 산정을 위해 노출위험, 정보보호, 자료 유용성을 고려하여 N/A 표시 격자의 비율이 중요하다고 하겠다. 단, 격자의 통계 값이 0인 경우를 포함하여 비율을 계산하였다.

○ 만약 빈도가 0인 격자를 N/A 표시하지 않고 그대로 제공하게 되면 격자의 위계구조로 인해 상위 격자로부터 하위 격자의 노출위험이 크게 증가하게 되므로 이를 반드시 고려해야 한다.

○ 또 특정 제공 항목이 많은 범주로 세분화되어 있을수록, 격자의 크기가 작아질수록 통계 값이 0인 경우가 급격히 증가하게 됨을 고려해야 한다.



<그림 3-1> 고정형 격자의 노출제어와 정보보호(예시)

2.1. 인구주택총조사 자료

□ 2019년 기준 인구주택총조사 자료의 경우 인구 테이블, 가구 테이블, 주택 테이블 등 총 3가지로 구성된다.

○ 인구 자료의 경우 각 관측치에 해당하는 성별(남, 여) 및 연령에 대해 조사된 데이터이다. 연령의 경우 각 세단위로 조사된 것을 10세 단위로 변환하여 활용하였다.

○ 가구 자료의 경우 세대유형에 대해 조사한 것으로 1세대 가구, 2세대 가구, 3세대 가구, 4세대 이상 가구, 1인 가구, 비혈연 가구로 분류된다.

○ 주택 자료는 거처의 종류(단독주택, 아파트, 연립주택, 다세대주택 등), 주택면적(20㎡ 이하, 20~40㎡, …, 165~230㎡, 230㎡ 초과 등), 건축연도(2019년, 2018년, …, 2010년, 2009~2005년, …, 1989~1980년, 1979년 이전 등)로 구성되어 있다.

가. 인구(population) 자료 분석

□ 인구 자료에서 총 인구 수, 성별 인구 수, 연령별 인구 수에 대하여 격자 크기별로 k-익명성을 확인해 보았다.

○ 분석을 위한 격자 크기는 100m, 500m, 1km로 설정하였고, 익명성 확보를 위한 k 값은 1부터 10으로 하여 N/A 처리 비율을 구해 보았다.

○ 개념상 100km 격자 1개는 10km 격자 100개, 1km 격자 10,000개, 500m 격자 40,000개, 100m 격자 1,000,000개로 구성된다. 하지만 실제 인구가 관측되는 격자에서만 통계 값이 산출되므로 분석에 사용된 격자의 수와 개념적 격자의 수는 차이가 있음을 밝힌다. 개념상 100m 격자수(1백만 개)의 약 14.8%에 해당하는 격자에서만 인구가 관측되었다.

<표 3-1> 인구 자료의 크기별 격자 수('다바' 격자)

격자 크기	100m	500m	1km	10km	100km
개념적 격자 수	1,000,000	40,000	10,000	·100	1
통계 값이 있는 격자 수	148,714	25,108	8,274	98	1

□ 총 인구 수에 대하여 k-익명성 모델 적용 시 N/A 처리 비율을 격자 크기별로 살펴보았다.

○ 격자 크기가 100m인 경우, 현행 익명성 기준 5에서는 59.70%의 격자를 N/A 처리하게 된다. 만약 익명성 기준을 3 또는 2 로 완화할 경우 N/A로 표시 되는 격자 비율이 각 15.25%p, 26.94%p 감소한다.

○ 격자 크기가 500m인 경우에는 N/A 표시 격자의 비율이 현행 기준 21.46%에서 14.28%, 10.09%로 감소하게 되고, 격자 크기가 1km인 경우에도 현행 기준 7.77%에서 5.33%, 3.57%로 감소하게 된다.

○ 격자 크기와 무관하게 익명성 기준을 현행 5에서 2로 완화할 경우 N/A 표시 비율이 약 절반 수준으로 감소하게 됨을 확인하였다.

○ 익명성 기준 5 또는 3을 적용할 경우, 총 인구 수 제공에 적절한 격자의 크기는 1km로 판단된다. 사용자 입장에서 조금 더 수용적인 태도를 취할 수 있다면 500m 격자의 제공도 검토해 볼 수 있을 것이다.

<표 3-2> (총 인구 수) 익명성 기준에 의한 N/A 비율(%)

익명성 기준		1	2	3	4	5	10
격자 크기	100m	13.54	32.76	44.45	53.39	59.70	74.64
	500m	4.33	10.09	14.28	18.17	21.46	35.17
	1km	1.60	3.57	5.33	6.77	7.77	13.04

□ 성별 인구 수에 대하여 k-익명성 모델 적용 시 N/A 표시 비율을 격자 크기별로 살펴보았다.

○ 100m 격자의 경우 익명성 기준 5를 적용하면 N/A 표시 비율이 75% 이상을 차지하고, 익명성 기준 3을 적용하면 그 비율이 71% 수준으로 약 4%p 감소하는데 그친다.

○ 500m 격자가 되면 N/A 표시 비율이 100m 격자의 절반 수준으로 낮아지게 된다. 익명성 기준 5의 경우 41%이고, 익명성 기준 3인 경우 31% 정도가 된다.

○ 이러한 현상은 1km 격자에서도 유사하게 확인된다. 익명성 기준이 5인 경우 N/A 표시 비율이 15.37%, 익명성 기준 3인 경우 11.54%까지 감소하였다.

○ 이상과 같은 결과로부터 노출위험과 자료 유용성을 고려한 성별 인구 수 제공에 적절한 격자 크기는 1km로 판단할 수 있다. 이때 익명성 기준이 3이라고 가정하면 전체 격자 중 약 88.5%의 격자에 해당하는 통계 값을 제공할 수 있다.

<표 3-3> (성별 인구 수) 익명성 기준에 의한 N/A 비율(%)

익명성 기준		1	2	3	4	5	10
격자 크기	100m	49.97	63.93	71.23	75.41	75.41	84.42
	500m	17.48	**25.03**	**31.31**	36.51	41.22	58.29
	1km	6.56	9.29	11.54	13.55	15.37	23.54

□ 연령별 인구 수에 대하여 k-익명성 모델 적용 시 N/A 표시 비율을 격자 크기별로 살펴보았다.

○ 현행 서비스에서는 연령을 5세 단위(예, 0~5세, 5~10세 등)로 제공하고 있다. 이 경우 익명성 기준이 1~10인 경우 100m, 500m 등의 격자에서는 모두 N/A 표시됨을 확인하였다.

○ 원인으로는 연령을 5세 단위로 나눌 경우 범주의 수가 21개에 달하게 되어 1km 이내의 소단위 격자에서는 빈도가 모두 10 이하로 나타나기 때문으로 파악된다.

<표 3-4> (5세 단위 연령별 인구 수) 익명성 기준에 의한 N/A 비율(%)

익명성 기준		1	2	3	4	5	10
격자 크기	100m	100	100	100	100	100	100
	500m	99.89	99.98	100	100	100	100
	1km	99.26	99.82	99.94	99.99	100	100

○ 이에 연령을 10세 단위로 재코딩 하여 0~10세, 10~20세, …, 60~70세, 70세 이상 등 모두 8가지의 범주로 구분하고 익명성 모델을 적용 후 N/A 표시 비율을 조사하였다.

○ 5세 단위 연령별 인구 수와 비교해 보면, 100m 격자의 경우 N/A 표시 비율이 약 5~7%p 감소하였지만 여전히 높은 비율을 유지하고 있다. 500m 격자에서는 그 비율이 10%p 내외로 감소하고, 1km 격자에서는 약 20%p 내외로 감소한다.

○ 따라서 10세 단위 연령별 인구 수 제공에 있어서는 100m 또는 500m 격자보다는 1km 격자를 활용하는 것이 적당한 것으로 판단된다.

<표 3-5> (10세 단위 연령별 인구 수) 익명성 기준에 의한 N/A 비율(%)

익명성 기준		1	2	3	4	5	10
격자 크기	100m	90.05	91.56	92.60	93.44	94.16	96.54
	500m	83.35	86.86	88.74	89.86	90.60	92.41
	1km	60.39	69.29	74.47	78.14	80.84	86.79

□ 연령 항목에 대해 15세 단위, 20세 단위, 30세 단위로 전반적 재코딩하여 연령별 인구 수를 구하고 k-익명성 모델을 적용하여 보았다.

○ 15세 단위의 재코딩은 0~15세, …, 45세~60세, 60세 이상으로 총 5가지의 범주로 구분하고, 20세 단위의 재코딩은 0~20세, 20~40세, 40~60세, 60세 이상으로 총 4가지의 범주로 구분하였다. 30세 단위의 재코딩은 0~30세, 30~60세, 60세 이상으로 총 3가지의 범주로 구분하였다.

○ 전반적 재코딩을 통해 제공 항목의 범주를 5개에서 4개로 줄이면 N/A 표시 비율이 감소함을 확인할 수 있다. 100m 격자의 경우 익명성 기준 5 또는 3으로 적용 시 약 2%p 정도 감소하는데 그쳤으나, 500m 격자에서는 감소폭이 익명성 기준 5인 경우 3.91%p, 익명성 기준 3인 경우 6.14%p로 나타났다.

○ 이러한 현상은 1km 격자에서 가속화되는데, 익명성 기준 5에서는 10.09%p, 익명성 기준 3에서는 12.01%p 감소하여 500m 격자에 비해 약 2배 이상 크게 감소함을 확인할 수 있다.

○ 전반적 재코딩을 통해 제공 항목의 범주를 3개로 줄이고, 1km 격자에서 익명성 기준 3에서 N/A 표시 비율이 30% 수준으로 감소하므로 공표 기준으로 고려해 볼 수 있다.

<표 3-6> (재코딩 후 연령별 인구 수) 익명성 기준에 의한 N/A 비율(%)

재코딩 방법*		ⓐ			ⓑ			ⓒ		
익명성 기준		2	3	5	2	3	5	2	3	5
격자 크기	100m	88.40	89.58	91.11	86.36	87.89	89.69	81.34	84.16	87.12
	500m	79.85	83.35	87.06	72.01	77.21	83.15	54.87	61.78	71.78
	1km	53.23	60.90	70.09	42.02	48.89	60.00	**24.98**	**30.42**	39.57

* : ⓐ 방법(범주 5개) : 0~15세, 15~30세, 30~45세, 45세~60세, 60세 이상
ⓑ 방법(범주 4개) : 0~20세, 20~40세, 40~60세, 60세 이상
ⓒ 방법(범주 3개) : 0~30세, 30~60세, 60세 이상

나. 가구(household) 자료 분석

□ 가구 자료에서 세대구성 범주별 가구 수에 대하여 격자 크기별로 k-익명성을 확인해 보았다.

○ 세대구성에 관한 범주는 1세대 가구, 2세대 가구, 3세대 가구, 4세대 이상 가구, 1인 가구, 비혈연 가구 등 총 6가지로 분류된다.

○ 분석을 위한 격자 크기는 100m, 500m, 1km로 설정하였고, 익명성 확보를 위한 k 값은 1부터 10으로 하여 N/A 표시 비율을 구하고, 현행 기준과 비교해 보았다.

○ 인구 자료와 마찬가지 이유로 실제 가구가 관측되는 격자에서만 통계 값이 산출되므로 개념상 존재하는 격자의 수와 분석에 고려된 격자수에 다소 차이가 있다. 즉, 개념상 존재해야 하는 100m 격자수(1백만 개)의 약 14.7%에 해당하는 격자에서만 가구가 관측되었다.

<표 3-7> 가구 자료의 크기별 격자 수('다바' 격자)

격자 크기	100m	500m	1km	10km	100km
통계 값이 있는 격자 수	147,947	25,026	8,257	98	1

□ 총 가구 수에 대하여 k-익명성 모델 적용에 따른 N/A 표시 비율을 살펴보았다.

○ 익명성 기준 5를 적용할 경우 100m 격자에서 N/A 표시 비율은 77.42%에 달하지만, 500m 격자에서는 38.22%로 절반가량 낮아진다. 또 1km 격자에서는 13.6%로 더욱 급격히 낮아짐을 확인할 수 있다.

○ 익명성 기준 3을 적용할 경우 100m 격자에서 N/A 표시 비율은 69.49%이고, 500m 격자에서는 27.46%로 약 10%p 감소한다. 또 1km 격자에서는 9.65%로 급격히 감소한다.

<표 3-8> (총 가구 수) 익명성 기준에 의한 N/A 비율(%)

익명성 기준		1	2	3	4	5	10
격자 크기	100m	43.59	61.00	69.49	74.33	77.42	84.02
	500m	12.17	**20.71**	**27.46**	33.37	38.22	56.96
	1km	4.12	7.41	9.65	11.72	13.60	22.03

□ 세대구성 범주별 가구 수에 대하여 k-익명성 모델 적용 시 N/A 표시 비율을 살펴보았다.

○ 원 자료에 k-익명성 모델을 적용 할 경우 100% 격자에서 N/A 표시되는 것을 확인할 수 있어서 세대구성 범주에 대한 전반적 재코딩이 필요하다고 판단된다.

○ 이는 1km 격자에서 4세대 이상 가구에 해당하는 가구가 전체의 약 6.6%에 불과한 것에서 기인한다. 구체적으로 전체 8,257개의 1km 격자 중에서 542개의 격자에서만 해당 가구가 존재한다.

○ 다른 원인으로 비혈연 가구에 주목할 필요가 있다. 100m 격자 기준으로 비혈연 가구에 대한 빈도가 0인 것이 전체의 86%에 달하고, 빈도가 0 이상 10 이하인 경우도 전체의 99%를 차지한다. 다시 말해 비혈연 가구의 빈도가 10을 초과하는 100m 격자는 전체의 0.2%에 불과하다.

○ 따라서 세대구성 유형에 대한 재코딩 작업이 필요하고 N/A 표시 비율 계산 시 비혈연 가구는 제외하거나 다른 범주와 묶어 줄 필요가 있다. 그러나, 비혈연 가구의 경우 개념적으로 다른 범주와 묶는 것이 어려워 분석에서 제외하였다.

○ 재코딩 방법으로는 3세대 가구와 4세대 가구를 통합하여 3세대 이상 가구로 분류하는 방법과 2세대, 3세대, 4세대 가구를 통합하여 2세대 이상 가구로 분류하는 방법을 고려하였다. 두 가지 방법 모두 비혈연 가구는 제외하였다.

○ 이와 같이 특정 범주를 제외하는 경우에는 총계 값과 항목별로 집계한 값에서 차이가 발생함으로 자료 제공 시 노출위험에 관해 충분한 검토가 필요하다.

조사된 범주
1세대 가구
2세대 가구
3세대 가구
4세대 이상 가구
1인 가구
비혈연 가구

재코딩 방법 ⓐ	재코딩 방법 ⓑ*
1세대 가구	1세대 가구
2세대 가구	2세대 이상 가구
3세대 이상 가구	
1인 가구	1인 가구

* ⓐ, ⓑ 방법에서 비혈연 가구 제외함

<그림 3-2> 세대구성 항목의 재코딩

○ 재코딩 방법에 따른 차이를 살펴보면 동일한 익명성 기준으로 100m 격자에서의 N/A 표시 비율은 범주가 4개인 경우보다 범주가 3개인 경우에 약 4~7%p 내외

로 소폭 감소한다.

○ 500m 격자에서는 감소폭이 약 6%p~21%p 내외로 증가하지만, N/A 표시 비율이 전반적으로 70% 이상 나타난다.

○ N/A 표시 비율이 가장 큰 변화를 보이는 격자는 1km로 3개 범주로 재코딩 하였을 때 익명성 기준 5에서는 43.72%, 익명성 기준 3에서는 33.31%로 낮아진다. 범주 수가 4개인 경우에 비해, 범주 수가 3개인 경우 익명성 기준 3에서 N/A 표시 비율이 절반 이상 감소한다.

<표 3-9> (재코딩 후 세대구성 유형별 가구 수) 익명성 기준에 의한 N/A 비율(%)

재코딩 방법*		ⓐ			ⓑ		
익명성 기준		2	3	5	2	3	5
격자 크기	100m	94.46	95.95	97.79	87.29	89.02	91.25
	500m	85.57	88.94	92.00	64.30	71.58	80.06
	1km	62.98	70.76	80.62	**27.54**	**33.31**	43.72
	10km	0	0	0	0	0	0

* : ⓐ 방법(범주 4개) : 1세대 가구, 2세대 가구, 3세대 이상 가구, 1인 가구
ⓑ 방법(범주 3개) : 1세대 가구, 2세대 이상 가구, 1인 가구
ⓐ 방법, ⓑ 방법 모두 비혈연 가구는 제외함

다. 주택 자료 분석

□ 주택 자료에서 주택면적, 주택유형, 건축연도별 건축물 수에 대하여 격자 크기별로 k-익명성을 적용해 보았다.

○ 주택면적에 관한 범주는 20㎡, 20~40㎡, …, 130~165㎡, 165~230㎡, 230㎡ 초과 등 총 9가지로 분류된다.

○ 주택유형에 관한 범주는 단독주택, 아파트, 연립주택, 다세대주택, 비거주용건물 내 주택, 주택 이외의 거처 등 총 6가지로 분류된다.

○ 건축연도에 관한 범주는 1979년 이전, 1980~1989년, 1990~1999년, 2000~2004년, 2005~2009년, 2010년, …, 2019년 등 총 15가지로 분류된다.

○ 분석을 위한 격자 크기는 100m, 500m, 1km로 설정하였고, 익명성 확보를 위한 k 값은 1부터 10으로 하여 N/A 표시 비율을 구하고, 현행 기준과 비교해 보았다.

○ 인구, 가구 자료와 마찬가지 사유로 실제 주택이 관측되는 격자에서만 통계 값

이 산출되므로 개념상 존재하는 격자의 수와는 다소 차이가 있다. 즉, 개념상 100m 격자수(1백만 개)의 약 13.8%에 해당하는 격자에서만 주택이 관측되었다.

<표 3-10> 주택 자료의 크기별 격자 수('다바' 격자)

격자 크기	100m	500m	1km	10km	100km
통계 값이 있는 격자 수	138,165	24,036	8,127	98	1

□ 총 주택 수에 대하여 k-익명성 모델 적용 시 N/A 표시 비율을 살펴보았다.

○ 100m 격자의 경우 현행 익명성 기준 5에서는 78.93%가 제공되지 않는데, 익명성 기준을 3으로 조정하게 되면 그 비율이 70.88%로 감소하게 된다.

○ 500m 격자의 N/A 표시 비율은 100m 격자의 거의 절반 수준으로 낮아져서 익명성 기준 5에서는 39.92%, 익명성 기준 3에서는 29%이다.

○ 마찬가지로 1km 격자에서 N/A 표시 비율은 500m 격자의 1/3 수준인 14.64%와 10.32%이다.

<표 3-11> (총 주택 수) 익명성 기준에 의한 N/A 비율(%)

익명성 기준		1	2	3	4	5	10
격자 크기	100m	45.50	62.43	70.88	75.73	78.93	85.83
	500m	13.40	**22.24**	**29.00**	34.87	39.92	58.65
	1km	4.59	7.86	10.32	12.45	14.64	23.31

□ 주택면적별 주택 수에 대하여 k-익명성 모델 적용 시 N/A 표시 비율을 살펴보았다.

○ 조사된 주택면적에 관한 범주는 20㎡ 이하, 20~40㎡ 이하, …, 130~165㎡ 이하, 165~230㎡ 이하, 230㎡ 초과 등 총 9가지로 분류된다.

○ 조사된 주택면적에 관한 범주를 그대로 두고 익명성 기준에 의한 비율을 조사해 본 결과 100m, 500m, 1km 격자에서는 거의 모든 격자가 N/A로 표시된다. 그러나 10km 격자로 집계되면 익명성 기준 5에서 19.39%, 익명성 기준 3에서 11.22%의 격자에서 N/A 표시된다.

<표 3-12> (주택면적별 주택 수) 익명성 기준에 의한 N/A 비율(%)

익명성 기준		1	2	3	4	5	10
격자 크기	100m	99.99	100	100	100	100	100
	500m	98.49	98.88	99.08	99.25	99.38	99.71
	1km	96.28	96.95	97.49	97.72	97.91	98.58
	10km	4.08	7.14	11.22	16.33	19.39	34.69

○ 1km 이내의 소 격자에서 제공 가능한 크기를 찾아보기 위해 주택면적 항목을 재코딩하였다. 재코딩은 기존의 9가지 수준에서 5가지 수준(40㎡ 이하, 40~85㎡ 이하, 85~130㎡ 이하, 130~230㎡ 이하, 230㎡ 초과), 3가지 수준(60㎡ 이하, 60~130㎡ 이하, 130㎡ 초과), 2가지 수준(100㎡ 이하, 100㎡ 초과) 등으로 하여 N/A 표시 비율을 검토하였다.

조사된 범주
20㎡ 이하
20 ~ 40㎡ 이하
40 ~ 60㎡ 이하
60 ~ 85㎡ 이하
85 ~ 100㎡ 이하
100 ~ 130㎡ 이하
130 ~ 165㎡ 이하
165 ~ 230㎡ 이하
230㎡ 초과

재코딩 방법 ⓐ	재코딩 방법 ⓑ	재코딩 방법 ⓒ
40㎡ 이하	60㎡ 이하	100㎡ 이하
40~85㎡ 이하		
85~130㎡ 이하	60~130㎡ 이하	
130~230㎡ 이하	130㎡ 초과	100㎡ 초과
230㎡ 초과		

<그림 3-3> 주택면적 항목의 재코딩

○ 재코딩된 주택면적 항목의 수가 5가지인 경우 주택 수에 대한 익명성 기준 5 또는 3을 적용하면 100m, 500m, 1km에서 N/A 표시되는 격자의 비율이 92.78 ~ 99.99%로 나타나 재코딩 후에도 그 비율에 변화가 있다고 판단하기는 어려운 수준이다.

○ 주택면적 항목의 수를 3개한 경우에도 익명성 기준 5와 3에서의 N/A 표시비율은 100m 격자에서 99.02%에서 98.17%로 0.85%p 감소하여 거의 변화가 없다. 500m 격자에서는 그 비율이 93.23%에서 89.98%로 3.25%p 감소하여 100m 격자에서의 감소폭보다는 크지만 여전히 N/A 표시 비율이 크다. 1km 격자에서도 항목의 수를 2개로 줄이는 경우에만 N/A 표시 비율이 매우 높게 나타난다.

○ 주택면적 항목의 수를 줄여도 N/A 표시 비율이 낮아지지 않는 원인은 주택면적이 130㎡를 초과하는 주택의 수가 매우 적기 때문인 것으로 파악된다. 이에 주택면적의 범주를 2가지로 압축하되, 100㎡를 기준으로 하여 분류하였다.

<표 3-13> (재코딩 후 주택면적별 주택 수) 익명성 기준에 의한 N/A 비율(%)

재코딩 방법*		ⓐ			ⓑ			ⓒ		
익명성 기준		2	3	5	2	3	5	2	3	5
격자 크기	100m	99.87	99.94	99.99	97.44	98.17	99.02	92.65	94.37	96.07
	500m	96.38	96.91	97.54	86.51	89.98	93.23	65.42	73.87	82.90
	1km	91.42	92.78	94.33	59.25	68.25	78.84	**31.87**	39.63	52.06
	10km	1.02	4.08	5.10	0	0	0	0	0	0

* : ⓐ 방법(범주 5개) : 40㎡ 이하, 40~85㎡ 이하, 85~130㎡ 이하, 130~230㎡ 이하, 230㎡ 초과
ⓑ 방법(범주 3개) : 60㎡ 이하, 60~130㎡ 이하, 130㎡ 초과
ⓒ 방법(범주 2개) : 100㎡ 이하, 100㎡ 초과

○ 주택면적의 범주를 2가지로 하면 1km 격자에서의 N/A 표시 비율이 익명성 기준 5에서 약 52%, 익명성 기준 3에서 약 40% 수준으로 낮아짐을 확인할 수 있다.

□ 주택유형별 주택 수에 대하여 k-익명성 모델 적용 시 N/A 표시 비율을 살펴보았다.

○ 조사된 주택유형에 관한 범주는 단독주택, 아파트, 연립주택, 다세대주택, 비거주용건물 내 주택, 주택 이외의 거처로 모두 6가지로 분류된다.

○ 조사된 주택유형에 관한 범주에 대해 익명성 기준에 의한 비율을 조사해 본 결과 100m, 500m, 1km 격자에서는 98.73% 이상의 격자에서 N/A로 표시된다. 이를 집계한 10km 격자에서 익명성 기준 5를 적용하면 52.04%, 익명성 기준 3에서 51.02%의 격자가 N/A 표시된다.

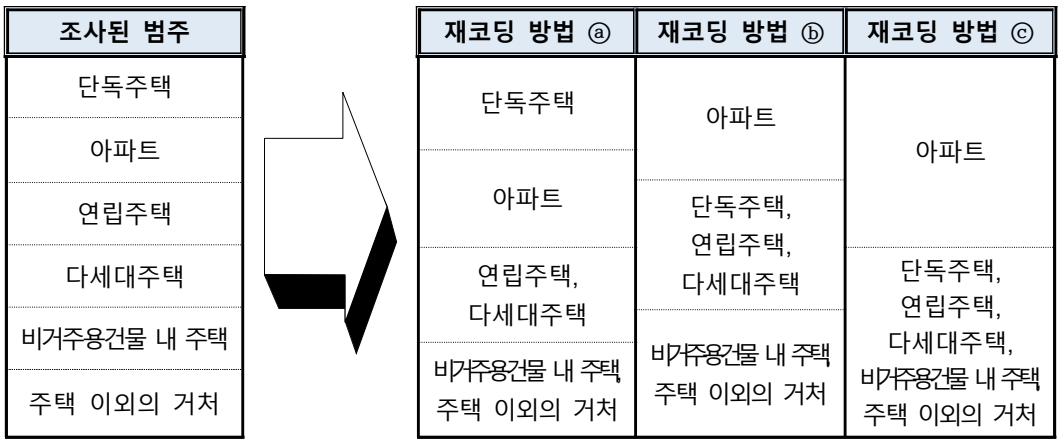
<표 3-14> (주택유형별 주택 수) 익명성 기준에 의한 N/A 비율(%)

익명성 기준		1	2	3	4	5	10
격자 크기	100m	100	100	100	100	100	100
	500m	99.59	99.65	99.67	99.70	99.73	99.80
	1km	98.73	98.83	98.90	98.92	98.98	99.18
	10km	48.98	50.00	51.02	52.04	52.04	59.18

○ 주택유형 항목에 대하여 여러 가지 방법으로 재코딩한 후 k-익명성 모델을 적용해 보았다. 어떠한 재코딩 방법을 적용하더라도 1km 이내의 격자에서 N/A 비율은 95% 전·후로 나타나며, 그 이하로 낮추는 것은 불가능함을 알 수 있다.

○ 그 원인으로 일부 격자에 아파트 또는 단독주택이 밀집해 있고, 연립주택, 다세대주택, 비거주용건물 내 주택, 주택 이외의 거처는 전반적으로 성긴(sparse) 빈도를 보인다. 이러한 현상은 주거지역과 비주거지역이 확연하게 나뉘는 도시의 특징일수 있다.

○ 따라서 주택유형별 주택 수 제공에 있어 k-익명성을 확보하면서 동시에 자료의 유용성을 재고하려면 주택유형을 2가지 범주로 재코딩하여 10km 격자로 하는 것이 적당함을 알 수 있다.



<그림 3-4> 주택유형 항목의 재코딩 방법

<표 3-15> (재코딩 후 주택유형별 주택 수) 익명성 기준에 의한 N/A 비율(%)

재코딩 방법*		ⓐ			ⓑ			ⓒ		
익명성 기준		2	3	5	2	3	5	2	3	5
격자 크기	100m	99.96	99.97	99.99	99.84	99.91	99.95	98.57	98.72	98.92
	500m	97.27	97.51	97.78	96.49	96.86	97.29	93.50	93.63	93.93
	1km	94.68	94.98	95.55	93.05	93.76	94.67	89.58	89.65	89.75
	10km	22.45	22.45	23.47	18.37	18.37	18.37	18.37	18.37	18.37

* : ⓐ 방법(범주 4개) : 단독주택, 아파트, (연립 + 다세대)주택, (비거주용건물 내 주택 + 주택 이외의 거처)
ⓑ 방법(범주 3개) : 아파트, (단독 + 연립 + 다세대)주택, (비거주용건물 내 주택 + 주택 이외의 거처)
ⓒ 방법(범주 2개) : 아파트, (단독 + 연립 + 다세대 + 비거주용건물 내 주택 + 주택 이외의 거처)

□ 건축연도별 주택 수에 대하여 k-익명성 모델 적용 시 N/A 표시 비율을 살펴보자.

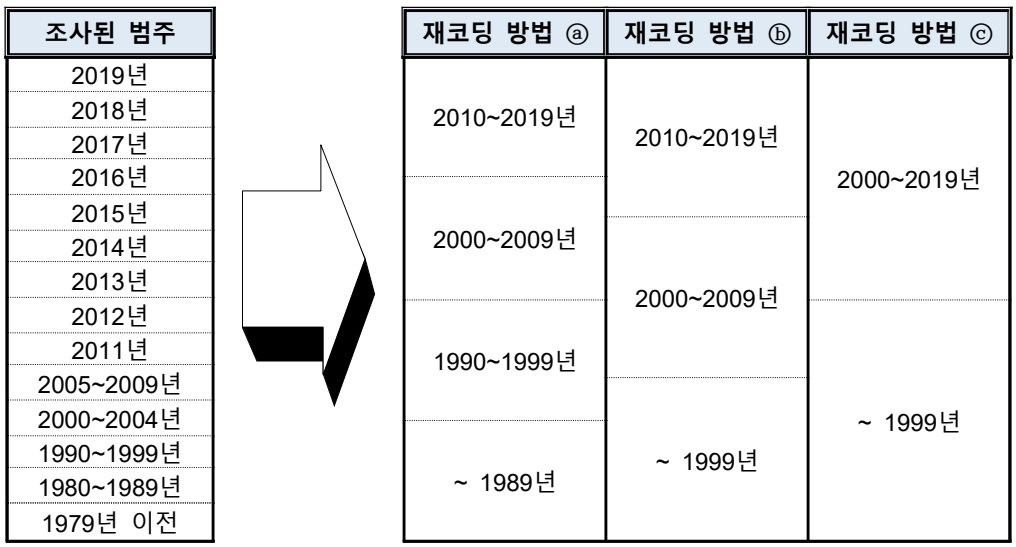
○ 건축연도에 관해 조사된 범주는 1979년 이전, 1980~1989년, 1990~1999년, 2000~2004년, 2005~2009년, 2010년, …, 2019년 등 총 15가지로 분류된다.

○ 건축연도에 관한 범주가 너무 많이 세분화되어 있어 k-익명성 모델 적용 시 거의 대부분의 격자에서 N/A로 표시된다. 이 경우에도 전반적 재코딩이 필요하다. 단, 1980~1989년에 건축된 주택의 수가 100m 격자에서 3개 이내인 경우가 전체의 95%를 차지한다. 1979년 이전에 건축된 주택이 수도 100m 격자에서 3개 이내인 경우가 전체의 94.2%를 차지하므로 이러한 수치를 고려하여 재코딩하였다.

○ 구체적인 재코딩 방법으로는 4가지 범주(2010~2019년, 2000~2009년, 1990~1999년, 1989년 이전)로 하는 방법, 3가지 범주(2010~2019년, 2000~2009년, 1999년 이전)으로 하는 방법, 2가지 범주(2000~2019년, 1999년 이전)로 하는 방법을 고려하였다.

<표 3-16> (건축연도별 주택 수) 익명성 기준에 의한 N/A 비율(%)

익명성 기준		1	2	3	4	5	10
격자 크기	100m	100	100	100	100	100	100
	500m	99.85	99.94	99.96	99.98	99.99	100
	1km	98.60	99.09	99.30	99.46	99.56	99.84
	10km	2.04	2.04	2.04	2.04	3.06	4.08



<그림 3-5> 건축연도 항목의 재코딩 방법

□ 건축연도 항목을 재코딩 방법을 적용하고 해당 격자별 주택 수에 대하여 k-익명성 모델 적용 시 N/A 표시 비율을 살펴보자.

○ 건축연도 항목을 재코딩하여 범주를 4개, 3개, 2개 수준으로 만든 경우 100m 격자에서의 N/A 표시 비율은 거의 100%에 가깝게 나타난다. 500m 격자의 경우에서도 범주가 4개 또는 3개로 하면 90% 이상의 격자가 N/A로 표시되게 된다.

○ 비교적 눈에 띄는 변화를 보이는 경우는 범주를 2가지로 한 경우에 1km 격자가 되면 익명성 기준 3 또는 5에서 N/A 표시 비율이 1/3 수준으로 감소하여 약 30~40%가 된다.

○ 이러한 현상에 대한 원인은 주택유형별 주택 수와 마찬가지로 일부 격자에만 주택이 밀집해 있기 때문으로 유추된다.

○ 건축연도별 주택 수 제공의 경우에는 건축연도를 2가지 범주로 재코딩하여 1km 격자로 하는 것이 적당함을 알 수 있다.

<표 3-17> (재코딩 후 건축연도별 주택 수) 익명성 기준에 의한 N/A 비율(%)

재코딩 방법*		ⓐ			ⓑ			ⓒ		
익명성 기준		2	3	5	2	3	5	2	3	5
격자 크기	100m	99.60	99.82	99.96	99.20	99.55	99.82	94.68	96.44	97.78
	500m	88.14	92.01	95.38	84.79	89.54	93.88	58.87	67.57	78.98
	1km	53.53	63.06	76.62	48.90	58.51	72.81	**23.99**	**29.45**	39.26
	10km	0	0	0	0	0	0	0	0	0

* : ⓐ 방법(범주 4개) : 2010~2019년, 2000~2009년, 1990~1999년, 1989년 이전
ⓑ 방법(범주 3개) : 2010~2019년, 2000~2009년, 1999년 이전
ⓒ 방법(범주 2개) : 2000~2019년, 1999년 이전

2.2. 전국사업체조사 자료

□ 2019년 기준 전국사업체조사 자료의 경우 조사대상 사업체의 종사자 수와 산업분류에 대한 항목으로 구성된다.

○ 총 사업체 수, 종사자 수에 따른 사업체 수, 산업분류별 사업체 수에 대하여 격자 크기별로 k-익명성을 검토하였다.

○ 분석을 위한 격자의 크기는 100m, 500m, 1km로 설정하였고, 익명성 확보를 위한 k값은 1에서 10까지로 하여 N/A 표시 비율을 조사하였다.

○ 인구주택총조사 자료에서와 마찬가지로 개념상 존재하는 격자의 수와 실제 사업체가 관측되는 격자의 수에 차이가 있다. 즉, 개념상 100m 격자수(1백만 개)의 약 6.7%에 해당하는 격자에서만 사업체가 관측되었다.

<표 3-18> 사업체 자료의 크기별 격자 수('다바' 격자)

격자 크기	100m	500m	1km	10km	100km
통계 값이 있는 격자 수	66,668	15,494	6,653	98	1

□ 총 사업체 수에 대하여 k-익명성 모델 적용 시 N/A 표시 비율을 살펴보았다.

○ 100m 격자의 경우 현행 익명성 기준 3에서는 68.48%가 제공되지 않는데, 익명성 기준을 2로 조정하게 되면 그 비율이 59.61%로 감소하게 된다.

○ 500m 격자의 N/A 표시 비율은 100m 격자에 비해 약 18%p 감소하여 익명성 기준 3에서는 50.6%, 익명성 기준 2에서는 41.95%이다.

○ 마찬가지로 1km 격자에서 N/A 표시 비율은 500m 격자에 비해 약 13%p 감소하여 익명성 기준 3에서는 36.80%, 익명성 기준 2에서는 28.95%이다.

○ 총 사업체 수 제공 시 익명성 기준 3에서 N/A 표시 비율을 고려하면 1km 격자가 적당하다고 할 수 있다.

<표 3-19> (총 사업체 수) 익명성 기준에 의한 N/A 비율(%)

익명성 기준		1	2	3	4	5	10
격자 크기	100m	41.94	59.61	68.48	73.74	77.38	86.69
	500m	28.07	41.95	50.60	56.82	61.15	74.04
	1km	17.83	**28.95**	36.80	42.52	47.24	60.92

가. 종사자 규모별 사업체 자료 분석

□ 종사자 규모별 사업체 수에 대하여 격자 크기별로 k-익명성을 적용해 보자.

○ '다바' 격자에 있는 사업체의 경우 종사자 수는 최소 1명인 경우부터 최대 18,625명인 경우로 범위가 매우 크다. 종사자 수가 10명 이하인 사업체가 '다바' 격자 전체의 91.7%를 차지한다.

<표 3-20> '다바' 격자의 종사자 규모별 사업체 비율

종사자 수	1	2	3	4	5	≤5	≤10	≤20	≤30	≤40	≤50
사업체 비율(%)	35.2	21.7	13.5	7.4	4.9	82.7	91.7	95.8	97.3	98.0	98.5

○ 이에 종사자 규모를 몇 가지 범주로 나누는 재코딩 작업을 하고 k-익명성 모델을 적용하여 N/A 표시 비율을 계산하는 방법으로 공표기준을 검토해 보았다.

○ 구체적인 재코딩 방법으로는 4가지 범주(1인, 2인, 3인, 4인 이상)로 하는 방법, 3가지 범주(1인, 2인, 3인 이상)로 하는 방법, 2가지 범주(1인, 2인 이상)로 하는 방법을 고려하였다.

<그림 3-6> 종사자 규모의 재코딩 방법

□ 종사자 수에 대하여 재코딩 방법을 적용하고 해당 격자별 사업체 수에 대하여 k-익명성 모델 적용 시 N/A 표시 비율을 살펴보자.

○ 종사자 수를 재코딩 하여 범주를 4개 또는 3개 수준으로 만든 경우 100m 격자에서의 N/A 표시 비율은 거의 100%에 가깝게 나타난다. 500m 격자의 경우에서도 범주가 4개 또는 3개로 하면 90% 정도의 격자가 N/A로 표시되게 된다.

○ 그나마 비교적 눈에 띄는 변화를 보이는 경우는 범주를 2가지로 한 경우에 1km 격자가 되면 익명성 기준 2 또는 3에서 N/A 표시 비율이 감소하여 약 61~68%가 된다.

○ 이러한 현상에 대한 원인은 '다바' 격자의 특성상 특정 격자에 사업체가 집중되어 있기 때문으로 유추된다.

○ 종사자 규모별 사업체 수를 제공하는 경우에는 종사자 규모를 2가지 범주로 재코딩하고, 1km 이상의 격자로 하는 것이 적당함을 알 수 있다.

<표 3-21> (재코딩 후 종사자 규모별 사업체 수) 익명성 기준에 의한 N/A 비율(%)

재코딩 방법*		ⓐ			ⓑ			ⓒ		
익명성 기준		1	2	3	1	2	3	1	2	3
격자 크기	100m	92.64	95.78	97.44	86.92	91.18	93.80	78.49	84.18	87.60
	500m	82.31	86.28	88.39	76.77	82.23	85.19	67.03	75.06	79.56
	1km	71.71	77.83	81.41	64.11	71.62	76.36	51.74	61.36	68.09
	10km	0	2.04	2.04	0	0	0	0	0	0

* : ⓐ 방법(범주 4개) : 1인, 2인, 3인, 4인 이상
ⓑ 방법(범주 3개) : 1인, 2인, 3인 이상
ⓒ 방법(범주 2개) : 1인, 2인 이상

나. 산업분류별 사업체 자료 분석

□ 산업분류별 사업체 수에 대하여 격자 크기별로 k-익명성을 적용해 보려 한다.

○ 전국사업체조사의 산업분류는 한국표준산업분류(10차 개정)을 따르고 있으며, 각 사업체별 세세분류에 의해 구분된다. 산업의 세세분류는 총 1,196개로 분류되고 있어 조사된 자료를 재코딩하지 않고 그대로 제공하는 것은 불가함을 예측할 수 있다.

○ 또 세분류는 495개, 소분류는 232개, 중분류는 77개로 분류되어 범주 수가 많음을 알 수 있다. 대분류 또한 21개의 범주로 분류되기 때문에 소단위 격자통계로 제공하기에는 적절하지 않음을 알 수 있다.

○ 이에 산업분류에 관한 범주의 수를 줄이기 위해 대분류를 기준으로 재코딩하여 일차적으로 1차 산업(농업, 어업, 임업 등), 2차 산업(광업, 제조업 등), 3차 산업(건설업, 도·소매업 등), 4차 산업(서비스업 등)으로 분류하여 현황을 살펴보았다.

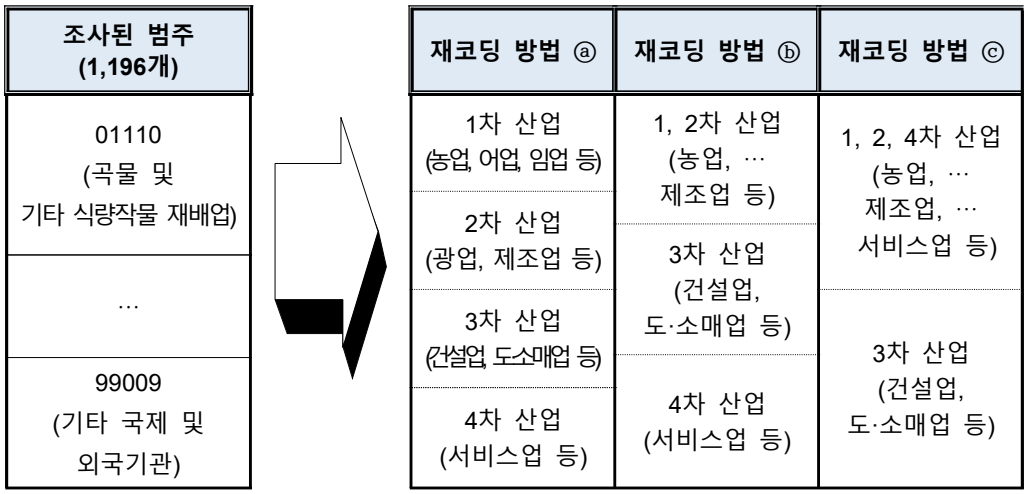
○ '다바' 격자의 특성으로 인해 1차 산업에 해당하는 사업체의 수는 극히 소수(0.2%)에 불과하고, 2차 산업도 전체의 11.2%에 불과하다. 따라서 1차 또는 2차 산업에 해당하는 사업체의 비율이 낮기 때문에 소 격자에서의 통계 제공이 어려울 것으로 예상되어 이차적인 재코딩을 하였다.

<표 3-22> '다바' 격자의 재코딩된 산업분류별 사업체 비율

재코딩된 산업분류	1차 산업 (농업, 임업 등)	2차 산업 (광업, 제조업 등)	3차 산업 (건설업, 도소매업 등)	4차 산업 (서비스업 등)
사업체 비율(%)	0.2	11.2	53.3	35.3

□ 산업분류에 대하여 재코딩 방법을 적용하고 해당 격자별 사업체 수에 대하여 k-익명성 모델 적용 시 N/A 표시 비율을 살펴보자.

○ 구체적인 재코딩 방법으로는 4가지 범주(1차 산업, 2차 산업, 3차 산업, 4차 산업)로 하는 방법, 3가지 범주(1, 2차 산업, 3차 산업, 4차 산업)로 하는 방법, 2가지 범주(1, 2, 4차 산업, 3차 산업)로 하는 방법을 고려하였다.



<그림 3-7> 산업분류의 재코딩 방법

○ 산업분류를 재코딩하여 범주를 4개 또는 3개 수준으로 만든 경우 100m 격자에서의 N/A 표시 비율은 거의 100%에 가깝게 나타난다. 500m 격자의 경우에서도 범주가 4개 또는 3개로 하면 90% 이상의 격자가 N/A로 표시되게 된다.

<표 3-23> (재코딩 후 산업분류별 사업체 수) 익명성 기준에 의한 N/A 비율(%)

재코딩 방법*		ⓐ			ⓑ			ⓒ		
익명성 기준		1	2	3	1	2	3	1	2	3
격자 크기	100m	100	100	100	96.56	98.70	99.42	77.50	83.67	87.19
	500m	99.90	99.98	99.99	83.06	88.05	91.05	62.82	71.00	75.95
	1km	99.34	99.85	99.97	69.28	76.28	81.06	48.71	57.75	63.49
	10km	11.22	26.53	35.71	1.02	1.02	1.02	0	0	0

* : ⓐ 방법(범주 4개) : 1차 산업, 2차 산업, 3차 산업, 4차 산업
ⓑ 방법(범주 3개) : (1, 2차) 산업, 3차 산업, 4차 산업
ⓒ 방법(범주 2개) : (1, 2, 4차) 산업, 3차 산업

○ 범주를 2가지로 줄여야 1km 격자가 되면 익명성 기준 2 또는 3에서 N/A 표시 비율이 감소하여 약 58~63% 정도가 된다.

○ 이러한 현상에 대한 원인은 '다바' 격자의 특성상 특정 산업이 집중되어 있기 때문으로 유추된다.

○ 산업분류별 사업체 수를 제공하는 경우에는 산업분류에 대한 범주를 2가지 수준으로 재코딩하고, 1km 이상의 격자로 하는 것이 적당함을 알 수 있다.

2.3. 소결

□ 2019년 인구주택총조사 자료와 전국사업체조사 자료에 대한 고정형 격자 체계에서의 k-익명성 모델 적용 결과를 종합해 보자.

○ k-익명성 모델을 적용하였으므로 격자에서 관측된 통계 값이 익명성 기준 값 이하인 경우 N/A로 표시하여 제공하게 된다.

○ 앞서 분석된 내용을 종합하여 보면 노출위험과 자료의 유용성을 고려했을 때 N/A 표시 비율을 30% 이하로 낮추기는 어려운 것으로 판단하였다.

○ 따라서 격자통계 제공 기준을 N/A 표시 비율이 30% 내외인 경우로 산정해 보았다.

□ 고정형 격자 체계에서 공표기준 산정을 위해 격자의 크기와 익명성 기준 값, 그리고 N/A 표시 비율의 상관관계를 검토해 보자.

<표 3-24> 고정형 격자 체계에서의 상관성

기준	조건	N/A 표시 비율
동일한 격자 크기	익명성 기준 값 ↓	↓
동일한 익명성 기준 값	격자 크기 ↑	↓

○ 고정형 격자 체계에서 자료 제공을 위해 고려해야 할 사항으로는 노출위험은 낮추면서 자료의 유용성 확보를 위해 N/A 표시 비율이 최소화될 수 있도록 하는 것이다.

○ 이때 동일한 격자 크기에서 익명성 기준 값을 작게 설정할수록 N/A 표시 비율

도 낮아진다. 따라서 자료의 유용성을 고려하여 익명성 기준 값을 가급적이면 작게 할 필요가 있다고 하겠으나, 노출위험을 고려하여 익명성 기준의 최솟값은 2 또는 3으로 정할 수 있다.

○ 반면 동일한 익명성 기준 값에 대해 격자의 크기가 클수록 N/A 표시 비율도 낮아져서, 격자의 크기는 가급적이면 크게 할 필요가 있다. 그러나 격자의 크기가 클수록 자료의 유용성을 낮아지게 된다.

○ 따라서 익명성 기준이 2 또는 3이면서 N/A 표시 비율이 30% 수준을 만족하는 최소 크기의 격자를 공표 기준으로 산정할 수 있다.

<표 3-25> 고정형 격자 체계의 제공 항목별 공표 기준

분류	제공 항목*	제공 가능한 최소 격자 크기	익명성 기준 값	N/A 비율 (%)
인구	총계	100m	2	32.76
	성별	500m	2	25.03
			3	31.31
	연령①	1km	2	24.98
			3	30.42
가구	총계	500m	2	20.71
			3	27.46
	세대구성②	1km	2	27.54
			3	33.31
주택	총계	500m	2	22.24
			3	29.00
	주택면적③	1km	2	31.87
	주택유형④	1km	2	89.58**
	건축연도⑤	1km	2	23.99
			3	29.45
사업체	총계	1km	2	28.95
	종사자 규모⑥	1km	2	51.74**
	산업분류⑦	1km	2	57.75**

* 전반적 재코딩 항목 ① 연령(30세 단위), ② 세대구성 유형(1세대, 2세대 이상, 1인 가구/ 비혈연 가구 제외)
③ 주택면적 (100㎡ 이하, 100㎡ 초과), ④ 주택유형 (아파트, 아파트 이외)
⑤ 건축연도 (1999년 이전, 2000년 이후)
⑥ 종사자 규모 (1인, 2인 이상), ⑦ 산업분류 ((1, 2, 4차) 산업, 3차 산업)

** 제공 불가 항목에 해당함

□ 이상의 사항을 고려하여 인구주택총조사 자료와 전국사업체조사 자료에 대하여 ‘다바’ 격자에서 제공 가능한 고정형 격자의 크기, 익명성 기준 값, N/A 표시 비율을 고려하여 공표기준을 산정해 보면 다음과 같다.

○ 인구, 가구, 주택 자료에 대하여 현행 익명성 기준보다는 완화하여 제공이 가능한 것으로 판단된다.

○ 단, 주택유형별 주택 수, 종사자 규모별 사업체 수, 산업분류별 사업체 수의 경우는 N/A 표시 비율이 높아 제공이 불가한 것으로 보인다.

□ 구체적으로 인구 자료의 경우 현행 익명성 기준인 5보다 완화된 2 또는 3으로 제공이 가능한 것으로 판단되며, 제공 항목에 따라 최소 격자의 크기는 다소 차이가 있다.

○ 총 인구 수의 경우 최소 격자의 크기를 100m로 하고 익명성 기준을 2로 설정하게 되면, 전체 격자 중 약 32%의 격자가 N/A 표시되고 약 68%의 격자는 조사된 통계 값을 제공할 수 있다.

○ 성별에 따른 인구 수의 경우는 제공 가능한 최소 격자의 크기가 500m로 판단되며 익명성 기준을 3으로 하게 되면 N/A 표시되는 격자가 전체의 약 31% 수준이 된다.

○ 연령별 인구 수의 경우는 각 세 단위로 조사된 것을 30세 단위로 전반적 재코딩하여 제공이 가능한 것으로 판단된다. 이때 제공 가능한 최소 격자의 크기는 1km이며 익명성 기준을 3으로 설정하게 되면 N/A 표시 비율이 약 30%가 된다.

○ 인구 자료의 경우 현행 적용되고 있는 익명성 기준 값 5보다는 완화된 기준인 2 또는 3으로 격자통계의 제공이 가능한 것으로 판단된다. 또 총 인구 수의 경우 제공 가능한 최소 격자의 크기가 100m이며, 성별 인구 수는 500m 격자에서, 연령별 인구 수는 30세 단위로 전반적 재코딩한 후에 1km 격자로 제공 가능하다.

□ 가구 자료의 경우도 현행 익명성 기준인 5보다 완화된 3을 적용하여 제공이 가능한 것으로 판단되며, 제공 항목에 따라 최소 격자의 크기는 다소 차이가 있다.

○ 총 가구 수의 경우 최소 격자의 크기를 500m로 하고 익명성 기준을 3으로 설정하게 되면, 전체 격자 중 약 27%의 격자가 N/A 표시되고 약 73%의 격자는 조사된 통계 값을 그대로 제공하여 유용성을 확보할 수 있다.

○ 세대구성 유형별 가구 수의 경우는 1세대, 2세대 이상, 1인 가구로 전반적 재코딩하여 제공 가능한 최소 격자의 크기가 1km로 판단된다. 익명성 기준을 3으로 하게 되면 N/A 표시되는 격자가 전체의 약 33% 수준이 되어 약 67% 정도의 자료 유용성을 확보하게 된다.

○ 가구 자료의 경우 현행 적용되고 있는 익명성 기준 값 5보다 완화된 기준인 3을 적용하여 격자통계의 제공이 가능한 것으로 판단된다. 단, 총 가구 수는 제공 가능한 최소 격자의 크기가 500m이며, 세대구성 유형별 가구 수는 전반적 재코딩 후 1km 격자로 제공 가능하다.

□ 주택 자료의 경우 총 주택 수, 주택면적별 주택 수, 건축연도별 주택 수는 격자통계 제공이 가능하나, 주택유형별 주택 수는 제공이 불가한 것으로 판단된다.

○ 총 주택 수의 경우 최소 격자의 크기를 500m로 하고 익명성 기준을 3으로 설정하게 되면, 전체 격자 중 약 29%의 격자가 N/A 표시되고 약 71%의 격자는 자료의 유용성을 확보할 수 있다.

○ 주택면적별 주택 수의 경우는 주택면적을 100㎡ 이하, 100㎡ 초과로 전반적 재코딩하게 되면 제공 가능한 최소 격자의 크기가 1km로 판단된다. 이때 익명성 기준을 3으로 하게 되면 N/A 표시되는 격자가 전체의 약 32% 수준이 되어 약 68% 정도의 자료 유용성 확보가 가능하다.

○ 주택유형별 주택 수의 경우는 주택유형에 대하여 전반적 재코딩을 수행하여도 N/A 표시 비율이 약 90% 수준에서 더 이상 낮아지지 않아 격자통계의 제공이 불가한 것으로 판단된다.

○ 건축연도별 주택 수의 경우는 건축연도를 1999년 이전, 2000년 이후로 전반적 재코딩하게 되면 제공 가능한 최소 격자의 크기가 1km로 판단된다. 이때 익명성 기준을 3으로 하게 되면 N/A 표시되는 격자가 전체의 약 30% 수준이 되어 약 70% 정도의 자료 유용성 확보가 가능하다.

○ 주택 자료의 경우 주택유형별 주택 수를 제외하고 현행 익명성 기준보다 완화된 기준인 3을 적용하여 격자통계의 제공이 가능한 것으로 판단된다. 단, 제공 가능한 최소 격자의 크기는 총 주택 수는 500m 단위이며, 주택면적별 주택 수와 건축연도별 주택 수는 전반적 재코딩 후 1km 격자로 제공 가능하다.

□ 사업체 자료의 경우 총 사업체 수는 격자통계 제공이 가능하나, 종사자 규모별 사업체 수와 산업분류별 사업체 수는 제공이 불가한 것으로 판단된다.

○ 총 사업체 수의 경우 최소 격자의 크기를 1km로 하고 익명성 기준을 2로 설정하게 되면, 전체 격자 중 약 29%의 격자가 N/A 표시되고 약 71%의 격자는 자료의 유용성을 확보할 수 있다.

○ 종사자 규모별 사업체 수의 경우는 종사자 규모에 대하여 전반적 재코딩을 수행하여도 N/A 표시 비율이 약 52% 수준에서 더 이상 낮아지지 않아 격자통계의 제공이 불가한 것으로 판단된다.

○ 산업분류별 사업체 수의 경우도 산업분류에 대한 전반적 재코딩 수행 후에도 N/A 표시 비율이 약 58% 수준에서 더 이상 감소하지 않아 격자통계의 제공이 불가한 것으로 판단된다.

○ 사업체 자료의 경우 총 사업체 수는 1km 격자에서 익명성 기준 2를 적용하여 제공이 가능하다. 그러나, 종사자 규모별 사업체 수와 산업분류별 사업체 수는 전반적 재코딩 후에도 자료의 유용성을 확보할 수 없어 제공이 불가하다.

3. 변동형 격자의 노출제어와 정보보호

□ 본 절에서는 변동형 격자 체계를 2019년 인구주택총조사 자료와 전국사업체조사 자료에 적용하고 공표기준을 산정해 보았다.

○ 변동형 격자 체계는 격자 크기의 기저(base)를 먼저 결정한 후, 자료 제공을 위해 격자의 크기를 다양하게 바꾸는 것을 허용하는 방식이다. 본 절에서는 변동형 격자 체계에 해당하는 쿼드트리(quadtree) 방법론을 적용하여 분석한다.

○ 쿼드트리 방법론은 일차적으로 1㎞ 격자를 재귀적으로 4개의 작은 격자로 분할하고, 사용자가 정한 기준에 따라 상위 수준으로 집계하거나 작은 격자를 보존하는 방법론이다.

○ 본 연구에서는 250m, 125m를 기저 격자 크기로 사용하고, 임계값보다 작은 관측치를 가진 격자는 보다 큰 격자로 집계하는 방법론을 적용한다. 여기서, 250m, 125m는 1km 격자를 보다 작은 크기의 격자로 각각 2번, 3번 분할했을 때 생기는 격자 크기에 해당한다.

○ 쿼드트리 방법론은 임계값보다 작은 관측치를 가진 격자를 상위 수준으로 집계하므로, 특정 격자에 포함되는 관측치의 수가 정해진 임계값 k보다 작지 않도록 하는 k-익명성 모델과 유사한 점이 있다.

○ 이러한 방법론의 특성을 반영하여 변동형 격자의 노출위험을 평가하기 위해 익명성 기준 k에 따른 기저 격자의 비율을 살펴본다.

□ 변동형 격자 체계를 기반으로 각 조사 자료에 대응하는 건물 좌표값을 결합하여 다양한 크기의 격자를 새롭게 생성하였다.

○ 건물의 좌표값을 기반으로 '다바 격자' 전체에 걸쳐 1km 격자를 생성한 후 기저 격자 크기를 각각 125m, 250m로 설정하여 2개의 계층적 공간 데이터 구조를 만들었다.

○ 노출제어를 위해 k-익명성 모델을 적용하고, 전반적 재코딩 방법을 적용하였다.

- 다양한 익명성 임계값 k에 대하여 익명성 임계값 이하인 격자는 쿼드트리 방식에 의해 상위 수준으로 집계된다.

- 따라서 각 격자에 해당하는 통계 값은 k보다 작지 않도록 구성되어 익명성 모델을 만족하게 된다.

○ 또 정보보호를 위해 변동형 격자에서는 격자 내 통계값이 임계값 k보다 크거나, 격자의 크기가 1㎢에 도달할 때까지 반복적으로 집계하게 된다.

- 반면 고정형 격자에서는 특정 격자에 해당하는 통계 값이 k보다 작은 경우 해당 격자를 N/A 처리함으로써 정보보호를 수행하였다.

○ '다바' 격자 건물의 UTM-K(GRS80) 좌표계를 기준으로 격자를 생성하였다. 단산, 강 등 건물이 없는 구역은 분석 대상에서 제외되었고, 최종적으로 생성된 격자의 수는 표와 같다.

○ 변동형 격자 체계는 좌표계를 기준으로 새롭게 격자를 생성하였으므로, 각 테이블마다 격자 크기별로 할당된 격자 코드를 활용한 고정형 격자 체계 분석에 활용된 격자 수와 다소 차이가 있음을 밝힌다.

○ 변동형 격자의 자료 제공에 대한 기준은 기저 격자의 비율로 설정하였다. 고정형 격자의 경우 k-익명성을 만족하지 못하는 격자는 N/A 표시하므로 그 비율을 기준으로 하였다. 그러나 변동형 격자의 경우 k-익명성을 만족하는 격자는 그 형태를 유지하므로 기저 격자의 비율을 자료 제공 기준으로 삼고자 한다.

○ 고정형 격자에서 N/A 표시 비율이 30% 정도이면 제공 가능한 것으로 판단하였으므로, 변동형 격자에서는 기저 격자의 비율이 70% 수준이면 제공 가능한 것으로 판단하였다.

<표 3-26> ' 다바 ' 격자의 크기별 격자 수(변동형 격자 체계)

격자 크기	125m	250m	500m	1km	10km	100km
생성 격자 수	138,385	66,448	26,122	8,414	98	1

□ 변동형 격자 체계에서 격자 크기를 최소화하면서 정보손실을 최소화하기 위해 잔차(residual) 격자의 개념이 활용된다.

○ 변동형 격자 체계에서 격자 내 값들이 k 값보다 작은 경우 쿼드트리 방식에 의해 상위 수준으로 집계된다. 이때 가능한 한 작은 격자를 보존하여 보다 정밀한 자료를 제공하기 위해 k 값보다 작은 격자 중 일부를 감추는 방법이 고안되었다. 그러나 격자를 감추는 것은 필수적으로 정보손실을 유발하게 된다.

○ 따라서 격자 크기 최소화와 정보손실 최소화 간 균형을 유지하기 위해 불균등 지수와 정보손실 비율을 계산하여 이용하였다. 불균등 지수와 정보손실 비율의 기준이 되는 값들은 0과 1 사이에서 정해질 수 있으며, 본 연구에서는 R 소프트웨어 내 쿼드트리 알고리즘에서의 기본값인 0.25와 0.4를 각각 이용한다.

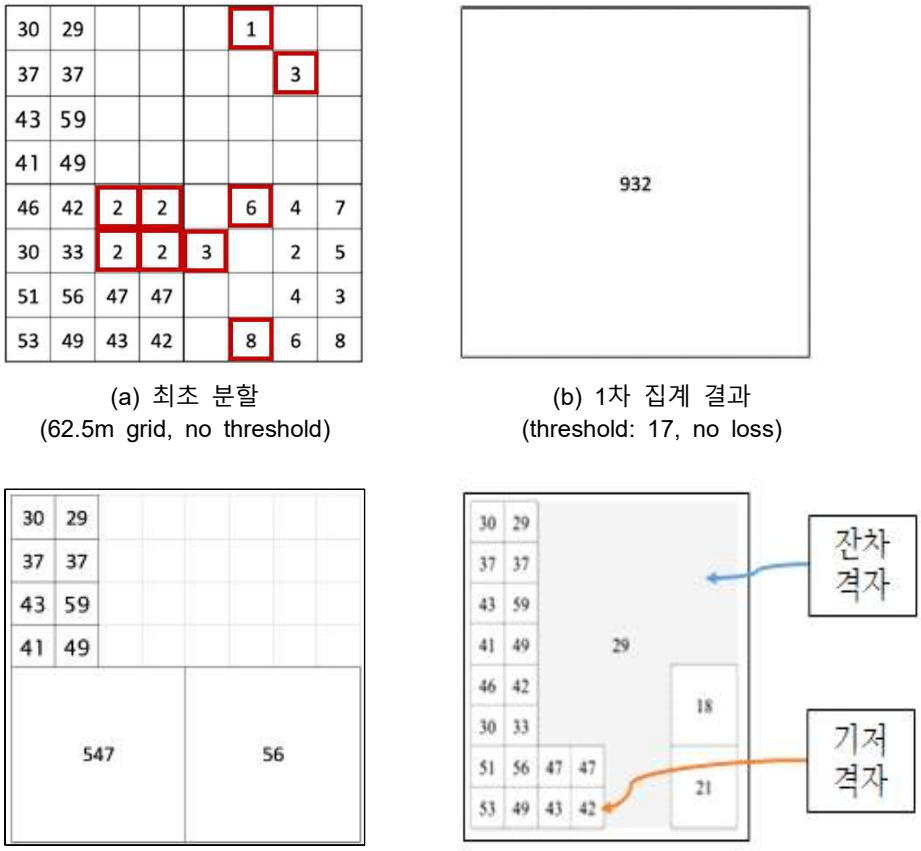
○ 불균등 정도를 계산하는 Theil 지수(the Theil index of inequality, 1972)는 다음과 같다.

$$\text{Theil index} = \sum_{i=1}^{4}(x_i \bullet \ln(x_i/\overline{x}))/\sum_{i=1}^{4}x_i$$

여기서, x_i는 i사분면 내 자료의 개수를 의미한다.

○ 정보손실 비율은 전체 격자 중 감춰진 격자에 해당하는 비율을 의미한다.

○ Theil 지수가 0.25보다 큰 경우 정보손실 비율이 0.4보다 작게 유지되는 한 k 값보다 작은 격자를 제외하고 계속해서 분할을 하게 된다. 분할의 과정이 끝나고 나면 k 값보다 작은 격자는 집계되어 잔차 격자로 남게 된다.



(a) 최초 분할
(62.5m grid, no threshold)

(b) 1차 집계 결과
(threshold: 17, no loss)

(c) 재분할 과정
(threshold: 17, loss: 4)

(d) 재분할 결과
(threshold: 17, loss: 29)

출처 : Lagonigro, R. et al.(2020)

<그림 3-8> 기저 격자와 잔차 격자(예시)

3.1 인구주택총조사자료

가. 인구(population) 자료 분석

□ 인구 자료에서의 노출위험 검토를 위해, 총 인구 수, 성별 및 연령별 인구 수에 대하여 익명성 기준 k를 기반으로 기저 격자의 비율을 살펴보았다.

○ 격자통계 제공 기준으로 기저 격자의 비율이 70% 수준인 경우로 설정하고 검토하였다.

□ 총 인구 수에 대하여 익명성 기준 k에 따른 쿼드트리 방법론을 적용하여 생성된

격자는 표와 같다.

○ 현행 익명성 기준 5에서 기저 격자 크기를 125m로 설정하는 경우, 62.3%의 기저 격자가 제공되며, 대략 81.5%의 격자를 250m보다 작은 크기의 격자로 제공할 수 있다.

○ 동일한 기준하에 기저 격자를 250m로 설정하는 경우 74.4%의 기저 격자가 제공될 수 있다.

○ 익명성 기준을 3으로 완화하여 기저 격자 크기를 125m로 설정한다면, 72.7%가 기저 격자로 유지되며, 250m로 설정하면 84.0%까지 유지될 수 있다.

○ 결론적으로 익명성 기준을 현행대로 유지한다면, 250m 이상의 격자를 제공할 수 있을 것으로 판단된다. 만약 익명성 기준을 2 또는 3으로 완화하는 경우에는 125m의 격자의 제공도 가능할 것으로 판단된다.

<표 3-27> (총 인구 수) 익명성 기준에 의한 격자 생성 결과

익명성 기준	기저 격자 크기	생성격자 비율					기저격자 비율(%)**
		125m	250m	500m	1km	잔차격자*	
1	125m	92.9	4.6	2.0	0.4	0.0	92.9
	250m	-	95.3	3.9	0.8	0.0	95.3
	500m	-	-	98.1	1.9	0.0	98.1
2	125m	85.1	9.1	3.5	2.2	1.6	85.1
	250m	-	91.5	6.4	2.1	0.9	91.5
	500m	-	-	97.2	2.8	0.2	97.2
3	125m	72.7	15.7	6.8	4.8	3.6	72.7
	250m	-	84.0	11.3	4.7	2.8	84.0
	500m	-	-	95.4	4.6	0.7	95.4
4	125m	66.4	18.2	9.2	6.2	4.6	66.4
	250m	-	78.8	14.8	6.5	3.9	78.8
	500m	-	-	93.9	6.1	1.0	93.9
5	125m	62.3	19.2	11.1	7.5	5.3	62.3
	250m	-	74.4	17.3	8.2	4.8	74.4
	500m	-	-	92.3	7.7	1.4	92.3
10	125m	55.6	17.8	14.8	11.8	6.7	55.6
	250m	-	62.8	22.3	14.9	7.2	62.8
	500m	-	-	84.6	15.4	3.0	84.6

* 잔차격자 비율 = (잔차격자 개수 / 전체격자 개수) ×100

** 기저격자 비율 = (기저격자 개수 / 전체격자 개수) ×100

□ 성별 인구 수에 대하여 익명성 기준 k에 따른 쿼드트리 방법론을 적용하여 생성된 격자는 <표 3-28>과 같다.

○ 현행 익명성 기준 5에서 기저 격자 크기를 125m로 설정하는 경우, 56.0%의 격자가 기저 격자로 유지되며, 대략 73.2%의 격자를 250m보다 작은 크기의 격자로 제공할 수 있다.

○ 또 동일한 기준하에 기저 격자 크기를 250m로 설정하는 경우 62.7%의 기저 격자가 유지되고 약 84%를 소 격자통계로 제공할 수 있다.

○ 익명성 기준을 3으로 완화하여 기저 격자 크기를 125m로 설정한다면, 58.4%의 격자가 기저 격자로 유지되며, 250m로 설정하면 69.2%까지 유지될 수 있다.

○ 결론적으로 현행 익명성 기준을 적용하는 경우에는 500m 격자로 제공해야 기저 격자 비율이 84% 정도가 된다. 익명성 기준을 2 또는 3으로 완화하는 경우 250m 격자의 제공이 가능할 것으로 판단된다.

<표 3-28> (성별 인구 수) 익명성 기준에 의한 격자 생성 결과

익명성 기준	기저 격자 크기	생성격자 비율					기저격자 비율(%)**
		125m	250m	500m	1km	잔차격자*	
1	125m	81.5	11.9	4.9	1.7	0.7	81.5
	250m	-	89.2	8.7	2.1	0.4	89.2
	500m	-	-	96.3	3.7	0.1	96.3
2	125m	63.7	19.0	10.0	7.2	5.3	63.7
	250m	-	76.1	16.0	7.9	4.9	**76.1**
	500m	-	-	93.0	7.0	1.3	93.0
3	125m	58.4	19.2	12.8	9.6	6.7	58.4
	250m	-	69.2	19.6	11.2	6.7	**69.2**
	500m	-	-	89.8	10.2	2.3	89.8
4	125m	56.5	18.3	13.8	11.3	7.3	56.5
	250m	-	65.2	21.0	13.8	7.7	65.2
	500m	-	-	86.7	13.3	3.1	86.7
5	125m	56.0	17.2	14.1	12.6	7.6	56.0
	250m	-	62.7	21.4	15.9	8.2	62.7
	500m	-	-	84.1	15.9	3.5	84.1
10	125m	56.6	14.0	12.4	17.0	7.1	56.6
	250m	-	57.1	19.3	23.6	8.2	57.1
	500m	-	-	72.2	27.8	4.2	72.2

* 잔차격자 비율 = (잔차격자 개수 / 전체격자 개수) ×100

** 기저격자 비율 = (기저격자 개수 / 전체격자 개수) ×100

□ 연령별 인구 수에 대하여 익명성 기준 k에 따른 쿼드트리 방법론을 적용하여 생성된 격자는 표와 같다.

○ 고정형 격자에서, 연령 항목에 대한 전반적 재코딩의 필요성이 언급되었다. 이에 더하여, R package에서 제공하는 쿼드트리 함수 또한 5개 이상의 범주를 지닌 데이터 분석을 지원하지 않기 때문에 고정형 격자에서와 동일한 방식으로 연령 항목을 재코딩하되, 범주가 5개 미만인 ⓑ, ⓒ방법만을 이용하여 항목을 재구성하였다.

○ ⓑ방법은 총 4가지 범주(0~20세, 20~40세, 40~60세, 60세 이상)로 구성되며, ⓒ방법은 총 3가지 범주(0~30세, 30~60세, 60세 이상)로 구성된다.

<표 3-29> (재코딩 후 연령별 인구 수) 익명성 기준에 의한 격자 생성 결과

재코딩 방법*	익명성 기준	기저 격자 크기	생성격자 비율					기저격자 비율(%)***
			125m	250m	500m	1km	잔차격자**	
ⓑ	2	125m	61.8	13.1	10.8	14.3	6.9	61.8
		250m	-	62.8	17.5	19.7	7.6	62.8
		500m	-	-	76.4	23.6	4.4	76.4
	3	125m	63.5	11.5	9.6	15.5	6.8	63.5
		250m	-	62.2	15.8	21.9	7.6	62.2
		500m	-	-	72.4	27.6	4.7	72.4
	5	125m	65.4	11.1	7.8	15.7	6.4	65.4
		250m	-	64.0	13.2	22.8	7.1	64.0
		500m	-	-	69.9	30.1	4.0	69.9
ⓒ	2	125m	58.4	16.5	12.7	12.3	6.5	58.4
		250m	-	64.3	19.7	15.9	6.8	64.3
		500m	-	-	82.0	18.0	3.1	**82.0**
	3	125m	58.7	14.7	11.9	14.7	7.3	58.7
		250m	-	61.6	18.7	19.8	8.2	61.6
		500m	-	-	77.6	22.4	4.1	**77.6**
	5	125m	59.8	12.7	10.8	16.7	7.4	59.8
		250m	-	59.5	17.2	23.4	8.5	59.5
		500m	-	-	72.1	27.9	4.5	72.1

* ⓑ 방법(범주 4개) : 0~20세, 20~40세, 40~60세, 60세 이상
ⓒ 방법(범주 3개) : 0~30세, 30~60세, 60세 이상
** 잔차격자 비율 = (잔차격자 개수 / 전체격자 개수) ×100
*** 기저격자 비율 = (기저격자 개수 / 전체격자 개수) ×100

○ 현행 익명성 기준 5에서 기저 격자 크기를 125m로 설정하는 경우, 각각 65.4%,

59.8%의 격자가 기저 격자로 유지되며, 대략 84.3%, 83.3%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ 또한 동일한 기준 하에 기저 격자 크기를 250m로 설정하는 경우 64.0%, 59.5%의 격자가 기저 격자로 유지되며, 대략 77.2%, 76.6% 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ ⓑ, ⓒ방법을 이용하여 항목을 재구성한 이후, 익명성 기준을 현행대로 유지하여 기저 격자 크기를 500m로 설정한다면 기저격자 비율이 각각 69.9%, 72.1%로 격자를 제공할 수 있을 것으로 판단된다.

○ 그러므로 ⓑ방법을 이용하여 연령 항목을 재코딩하고, 익명성 기준을 현행대로 유지하여 500m 격자를 제공하는 것이 적절한 방법으로 판단된다.

나. 가구(household) 자료 분석

□ 가구 자료에서의 노출위험 검토를 위해 총 가구 수, 세대구성 유형별 가구 수에 대하여 익명성 기준 k를 기반으로 기저 격자에 해당되는 비율을 살펴보았다.

○ 전체적인 자료 구성은 고정형 격자와 동일하며, 세대구성 유형별 가구 수는 고정형 격자와 동일한 방식으로 재코딩하여 분석을 수행하였다.

□ 총 가구 수에 대하여 익명성 기준 k에 따른 쿼드트리 방법론을 적용하여 생성된 격자는 표와 같다.

○ 현행 익명성 기준 5에서 기저 격자 크기를 125m로 설정하는 경우, 53.8%의 격자가 기저 격자로 유지되며, 대략 71.6%의 격자를 250m보다 작은 크기의 격자로 제공할 수 있다. 또한 동일한 기준하에 기저 격자 크기를 250m로 설정하는 경우 60.0%의 격자가 기저 격자로 유지되며, 대략 83.7%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ 익명성 기준을 3으로 완화하여 기저 격자 크기를 125m로 설정한다면, 55.8%의 격자가 기저 격자로 유지되며, 250m로 설정하면 67.6%까지 유지될 수 있다.

○ 익명성 기준을 5에서 3으로 완화하는 경우, 기저격자 비율이 각각 2.0%p, 7.6%p만 증가하여 제공 가능한 격자 크기에 큰 차이가 나지 않는다고 판단된다.

○ 익명성 기준을 현행대로 유지하는 경우 500m 격자를 제공하는 것이 적합할 것으로 판단된다.

<표 3-30> (총 가구 수) 익명성 기준에 의한 격자 생성 결과

익명성 기준	기저 격자 크기	생성격자 비율					기저격자 비율(%)**
		125m	250m	500m	1km	잔차격자*	
1	125m	92.5	4.8	2.2	0.5	0.0	92.5
	250m	-	94.8	4.3	0.9	0.0	94.8
	500m	-	-	97.9	2.1	0.0	97.9
2	125m	63.4	20.5	10.2	5.9	4.1	63.4
	250m	-	77.2	16.2	6.6	3.7	77.2
	500m	-	-	93.4	6.6	1.1	93.4
3	125m	55.8	21.0	14.4	8.8	5.5	55.8
	250m	-	67.6	22.0	10.4	5.4	67.6
	500m	-	-	89.4	10.6	2.0	89.4
4	125m	53.8	19.4	15.7	11.0	6.3	53.8
	250m	-	62.6	23.7	13.7	6.6	62.6
	500m	-	-	85.7	14.3	2.7	85.7
5	125m	53.8	17.8	15.8	12.7	6.6	53.8
	250m	-	60.0	23.7	16.2	7.1	60.0
	500m	-	-	82.3	17.7	3.2	82.3
10	125m	55.6	13.7	12.7	18.0	6.7	55.6
	250m	-	54.9	19.8	25.3	7.7	54.9
	500m	-	-	69.3	30.7	4.1	69.3

* 잔차격자 비율 = (잔차격자 개수 / 전체격자 개수) ×100
** 기저격자 비율 = (기저격자 개수 / 전체격자 개수) ×100

□ 세대구성 유형별 가구 수에 대하여 익명성 기준 k에 따른 쿼드트리 방법론을 적용하여 생성된 격자는 표와 같다.

○ 고정형 격자와 동일한 방법으로 세대구성 항목을 재코딩하여 분석을 수행하였다. ⓐ방법은 총 4가지 범주(1세대 가구, 2세대 가구, 3세대 이상 가구, 1인 가구)로 구성되며, ⓑ방법은 총 3가지 범주(1세대 가구, 2세대 이상 가구, 1인 가구)로 구성된다. 단, 두 방법 모두에서 비혈연 가구는 극소수인 것으로 파악되어 분석에서 제외하였다.

○ 현행 익명성 기준 5에서 기저 격자 크기를 125m로 설정하는 경우, 각각 40.2%, 54.6%의 격자가 기저 격자로 유지되며, 대략 77.3%, 77.5%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다. 또한 동일한 기준하에 기저 격자 크기를 250m로 설정하는 경우 61.7%, 53.0%의 격자가 기저 격자로 유지되며, 대략 75.4%, 68.8%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ 익명성 기준을 3으로 완화하여 기저 격자 크기를 125m로 설정한다면, 각각 50.1%, 54.4%의 격자가 기저 격자로 유지되며, 250m로 설정하면 59.4%, 51.9%의 격자가 기저 격자로 유지될 수 있다.

○ 이러한 분석 결과에 따라, ⓐ방법으로 항목을 재코딩하고 익명성 기준을 5에서 3으로 완화하는 경우, 기저격자 비율이 각각 9.9%p, -2.3%p 변화하는 데 그쳐 제공 가능한 격자 크기에 큰 차이가 나지 않는다고 판단되었다. 따라서 익명성 기준을 현행대로 유지하여 500m 격자를 제공하는 것이 적합할 것으로 고려되며, 이는 ⓑ방법으로 항목을 재코딩한 경우에도 동일한 양상을 보였다.

○ 그러므로 ⓐ방법을 이용하여 세대구성 유형을 재코딩하고, 익명성 기준을 현행대로 유지하여 500m 격자를 제공하는 것이 정밀하게 정보를 제공하면서도 정보를 보호할 수 있는 방안으로 판단된다.

<표 3-31> (재코딩 후 세대구성 유형별 가구 수) 익명성 기준에 의한 격자 생성 결과

재코딩 방법*	익명성 기준	기저 격자 크기	생성격자 비율					기저격자 비율(%)***
			125m	250m	500m	1km	잔차격자**	
ⓐ	2	125m	52.1	15.5	11.5	20.9	6.8	52.1
		250m	-	56.7	16.6	26.6	6.3	56.7
		500m	-	-	65.4	34.6	3.6	65.4
	3	125m	50.1	18.0	10.4	21.4	6.9	50.1
		250m	-	59.4	14.5	26.1	5.9	59.4
		500m	-	-	65.0	35.0	3.1	65.0
	5	125m	40.2	26.2	10.9	22.6	7.9	40.2
		250m	-	61.7	13.7	24.6	6.2	61.7
		500m	-	-	67.5	32.5	1.9	67.5
ⓑ	2	125m	52.5	14.0	14.8	18.7	7.8	52.5
		250m	-	52.0	22.4	25.7	9.2	52.0
		500m	-	-	69.5	30.5	5.8	**69.5**
	3	125m	54.4	12.5	12.5	20.6	7.7	54.4
		250m	-	51.9	19.1	28.9	9.1	51.9
		500m	-	-	64.4	35.6	6.0	**64.4**
	5	125m	54.6	12.5	10.4	22.5	6.8	54.6
		250m	-	53.0	15.8	31.1	7.2	53.0
		500m	-	-	59.3	40.7	4.5	59.3

* ⓐ 방법(범주 4개) : 1세대 가구, 2세대 가구, 3세대 이상 가구, 1인 가구
ⓑ 방법(범주 3개) : 1세대 가구, 2세대 이상 가구, 1인 가구
ⓐ 방법, ⓑ 방법 모두 비혈연 가구는 제외함
** 잔차격자 비율 = (잔차격자 개수 / 전체격자 개수) ×100
*** 기저격자 비율 = (기저격자 개수 / 전체격자 개수) ×100

다. 주택 자료 분석

□ 주택 자료에서의 노출위험을 살펴보기 위해, 총 주택 수, 주택면적별 주택 수, 주택유형별 주택 수, 건축유형별 주택 수에 대하여 익명성 기준 k를 기반으로 기저 격자에 해당되는 비율을 살펴보았다.

○ 전체적인 자료 구성은 고정형 격자와 동일하고, 주택면적별 주택 수, 주택유형별 주택 수, 건축연도별 주택 수는 고정형 격자에서와 같은 방식으로 재코딩하여 분석을 수행하였다.

□ 총 주택 수에 대하여 익명성 기준 k에 따른 쿼드트리 방법론을 적용하여 생성된 격자는 표와 같다.

<표 3-32> (총 주택 수) 익명성 기준에 의한 격자 생성 결과

익명성 기준	기저 격자 크기	생성격자 비율					기저격자 비율(%)**
		125m	250m	500m	1km	잔차격자*	
1	125m	89.6	6.0	3.5	0.9	0.0	89.6
	250m	-	91.7	6.5	1.8	0.0	91.7
	500m	-	-	96.1	3.9	0.0	96.1
2	125m	60.3	21.3	11.6	6.9	4.4	60.3
	250m	-	73.9	18.1	8.0	4.0	73.9
	500m	-	-	91.2	8.8	1.4	91.2
3	125m	53.3	21.2	15.6	9.9	5.7	53.3
	250m	-	64.6	23.3	12.0	5.7	64.6
	500m	-	-	87.2	12.8	2.1	87.2
4	125m	51.4	19.5	16.8	12.3	6.4	51.4
	250m	-	59.8	24.9	15.3	6.7	59.8
	500m	-	-	83.4	16.6	2.8	83.4
5	125m	51.0	18.2	16.7	14.1	6.8	51.0
	250m	-	57.6	24.6	17.8	7.0	57.6
	500m	-	-	80.0	20.0	3.3	80.0
10	125m	51.6	15.0	13.7	19.7	6.7	51.6
	250m	-	52.9	20.5	26.7	7.1	52.9
	500m	-	-	67.4	32.6	3.7	67.4

* 잔차격자 비율 = (잔차격자 개수 / 전체격자 개수) ×100
** 기저격자 비율 = (기저격자 개수 / 전체격자 개수) ×100

○ 현행 익명성 기준 5에서 기저 격자 크기를 125m로 설정하는 경우, 51.0%의 격자

가 기저 격자로 유지되며, 대략 85.9%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다. 또한 동일한 기준하에 기저 격자 크기를 250m로 설정하는 경우 57.6%의 격자가 기저 격자로 유지되며, 대략 82.2%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ 익명성 기준을 3으로 완화하여 기저 격자 크기를 125m로 설정한다면, 53.3%의 격자가 기저 격자로 유지되며, 250m로 설정하면 64.6%까지 유지될 수 있다.

○ 따라서 익명성 기준을 현행대로 유지한다면, 500m 격자를 제공하여야 격자의 70% 이상이 소 격자 크기를 유지할 수 있다. 익명성 기준을 2로 완화하는 경우에는 250m 격자로 제공이 가능할 것으로 판단된다.

□ 주택면적별 주택 수에 대하여 익명성 기준 k에 따른 쿼드트리 방법론을 적용하여 생성된 격자는 표와 같다.

○ 본 자료에는 주택 면적이 '99'로 코딩된 응답이 26,850개 포함되어 있어, 이를 N/A 처리하여 분석을 수행하였다.

○ 고정형 격자에서 언급된 전반적 재코딩의 필요성과 함께 R 패키지에서 제공하는 쿼드트리 함수 또한 5개 이상의 범주를 지닌 데이터 분석을 지원하지 않는다. 따라서 고정형 격자에서와 동일한 방식으로 주택면적 항목을 재코딩하되, 범주가 5개 미만인 ⓑ, ⓒ방법만을 이용하여 항목을 재구성하였다.

○ ⓑ방법은 총 3가지 범주(60㎡ 이하, 60~130㎡ 이하, 130㎡ 초과)로 구성되며, ⓒ방법은 총 2가지 범주(100㎡ 이하, 100㎡ 초과)로 구성된다.

○ 현행 익명성 기준 5에서 기저 격자 크기를 125m로 설정하는 경우, 각각 30.3%, 35.7%의 격자가 기저 격자로 유지되며, 대략 68.1%, 70.2%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ 또한 동일한 기준하에 기저 격자 크기를 250m로 설정하는 경우 47.2%, 43.1%의 격자가 기저 격자로 유지되며, 대략 64.2%, 64.2% 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ ⓒ방법을 이용하여 항목을 재구성한 이후, 익명성 기준을 3으로 완화하여 기저 격자 크기를 125m로 설정한다면, 36.6%의 격자가 기저 격자로 유지되며, 대략 73.8%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다. 또한 기저 격자 크기를 250m로 설정하면 44.3%의 격자가 기저 격자로 유지되며, 대략 68.3%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ 따라서 익명성 기준을 3으로 완화하고, ⓒ방법을 이용하여 항목을 재코딩한다면 500m 격자로 제공할 수 있을 것으로 판단된다.

<표 3-33> (재코딩후 주택면적별 주택 수) 익명성 기준에 의한 격자 생성 결과

재코딩 방법*	익명성 기준	기저 격자 크기	생성격자 비율					기저격자 비율(%)***
			125m	250m	500m	1km	잔차격자**	
ⓑ	2	125m	35.7	17.0	16.5	30.8	9.2	35.7
		250m	-	42.5	20.7	36.8	9.5	42.5
		500m	-	-	57.3	42.7	5.7	57.3
	3	125m	34.1	18.6	14.8	32.5	9.1	34.1
		250m	-	43.8	18.2	38.0	9.2	43.8
		500m	-	-	54.8	45.2	5.3	54.8
	5	125m	30.3	23.3	14.5	31.8	9.3	30.3
		250m	-	47.2	17.0	35.7	9.3	47.2
		500m	-	-	56.5	43.5	5.2	56.5
ⓒ	2	125m	37.3	20.0	20.4	22.2	7.9	37.3
		250m	-	46.5	26.6	26.9	8.2	46.5
		500m	-	-	69.7	30.3	4.2	**69.7**
	3	125m	36.6	18.7	18.5	26.2	8.5	36.6
		250m	-	44.3	24.0	31.7	8.7	44.3
		500m	-	-	63.6	36.4	4.7	**63.4**
	5	125m	35.7	18.0	16.5	29.7	8.0	35.7
		250m	-	43.1	21.1	35.8	8.1	43.1
		500m	-	-	57.6	42.4	4.2	57.6

* ⓑ 방법(범주 3개) : 60㎡ 이하, 60~130㎡ 이하, 130㎡ 초과
ⓒ 방법(범주 2개) : 100㎡ 이하, 100㎡ 초과
** 잔차격자 비율 = (잔차격자 개수 / 전체격자 개수) ×100
*** 기저격자 비율 = (기저격자 개수 / 전체격자 개수) ×100

□ 주택유형별 주택 수에 대하여 익명성 기준 k에 따른 쿼드트리 방법론을 적용하여 생성된 격자는 표와 같다.

○ 고정형 격자에서와 동일한 방법으로 주택유형 항목을 재코딩하여 분석을 수행하였다. ⓐ방법은 총 4가지 범주(단독주택, 아파트, 연립주택·다세대주택, 비거주용건물 내 주택·주택 이외의 거처)로 구성되며, ⓑ방법은 총 3가지 범주(아파트, 단독주택·연립주택·다세대주택, 비거주용건물 내 주택·주택 이외의 거처)로, ⓒ방법은 총 2가지 범주(아파트, 아파트 외 항목)로 구성된다.

<표 3-34> (재코딩 후 주택유형별 주택 수) 익명성 기준에 의한 격자 생성 결과

재코딩 방법*	익명성 기준	기저 격자 크기	생성격자 비율					기저격자 비율(%)***
			125m	250m	500m	1km	잔차격자**	
ⓐ	2	125m	8.0	37.5	27.1	27.4	13.8	8.0
		250m	-	45.0	27.7	27.3	13.4	45.0
		500m	-	-	76.3	23.7	4.4	76.3
	3	125m	4.8	36.5	30.4	28.3	14.1	4.8
		250m	-	41.2	30.7	28.1	13.7	41.2
		500m	-	-	76.1	23.9	4.5	76.1
	5	125m	3.1	31.6	36.3	29.0	13.6	3.1
		250m	-	34.6	36.5	28.9	13.3	34.6
		500m	-	-	74.4	25.6	6.3	74.4
ⓑ	2	125m	11.9	35.2	26.3	26.5	12.5	11.9
		250m	-	46.3	27.3	26.4	11.8	46.3
		500m	-	-	74.2	25.8	5.5	74.2
	3	125m	9.5	33.6	29.0	27.9	13.5	9.5
		250m	-	42.7	29.7	27.7	12.9	42.7
		500m	-	-	74.4	25.6	5.5	74.4
	5	125m	6.2	31.3	34.3	28.2	13.8	6.2
		250m	-	37.2	34.7	28.1	13.6	37.2
		500m	-	-	75.2	24.8	6.1	75.2
ⓒ	2	125m	23.2	32.7	25.2	18.9	8.2	23.2
		250m	-	53.2	28.3	18.5	6.4	53.2
		500m	-	-	80.1	19.9	2.1	**80.1**
	3	125m	21.3	32.7	25.5	20.5	8.9	21.3
		250m	-	51.8	28.3	19.9	7.1	51.8
		500m	-	-	79.1	20.9	2.2	**79.1**
	5	125m	19.2	32.4	26.2	22.2	9.7	19.2
		250m	-	49.6	28.5	21.9	8.3	49.6
		500m	-	-	77.5	22.5	3.0	77.5

* ⓐ 방법(범주 4개) : 단독주택, 아파트, (연립 + 다세대)주택, (비거주용건물 내 주택 + 주택 이외의 거처)
ⓑ 방법(범주 3개) : 아파트, (단독 + 연립 + 다세대)주택, (비거주용건물 내 주택 + 주택 이외의 거처)
ⓒ 방법(범주 2개) : 아파트, (단독 + 연립 + 다세대 + 비거주용건물 내 주택 + 주택 이외의 거처)
** 잔차격자 비율 = (잔차격자 개수 / 전체격자 개수) ×100
*** 기저격자 비율 = (기저격자 개수 / 전체격자 개수) ×100

○ 재코딩 방법에 따라 현행 익명성 기준 5에서 기저 격자 크기를 125m로 설정하는 경우 각각 3.1%, 6.2%, 19.2%의 격자가 기저 격자로 유지되며, 약 71.0%, 71.8%, 77.8%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다. 또한 동

일한 기준하에 기저 격자 크기를 250m로 설정하는 경우 34.6%, 37.2%, 49.6%의 격자가 기저 격자로 유지되며, 대략 71.1%, 71.9%, 78.1% 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ 마찬가지로 익명성 기준을 3으로 완화하여 기저 격자 크기를 125m로 설정한다면, 각각 4.8%, 9.5%, 21.3%의 격자가 기저 격자로 유지되며, 250m로 설정하면 41.2%, 42.7%, 51.8%까지 유지될 수 있다.

○ 또, ⓐ방법으로 항목을 재코딩하고 익명성 기준을 5에서 3으로 완화하는 경우, 기저 격자 비율이 각각 1.7%p, 6.6%p 증가하는 데 그쳐 제공 가능한 격자 크기에 큰 차이가 나지 않는다고 판단된다. 따라서 익명성 기준을 현행대로 유지하여 500m 격자를 제공하는 것이 적합할 것으로 고려되며, 이는 ⓑ, ⓒ 방법으로 항목을 재코딩한 경우도 동일한 양상을 보였다.

○ 결과적으로 ⓐ방법을 이용하여 세대구성 항목을 재코딩하고, 익명성 기준을 현행대로 유지하여 500m 격자를 제공하는 것이 적절한 것으로 판단된다.

□ 건축연도별 주택 수에 대하여 익명성 기준 k에 따른 쿼드트리 방법론을 적용하여 생성된 격자는 표와 같다.

○ 고정형 격자에서와 마찬가지로 관측된 원래 범주에 대하여 k-익명성 모델을 적용할 경우 대부분의 격자에서 N/A 표시되는 것이 확인된다. 따라서 건축연도 항목에 대한 전반적 재코딩을 수행하였다.

○ 고정형 격자와의 비교를 위해 고정형 격자에서와 동일한 기준으로 건축연도 항목을 재코딩하였다. ⓐ방법은 총 4가지 범주(1989년 이전, 1990~1999년, 2000~2009년, 2010~2019년)로 구성되며, ⓑ방법은 총 3가지 범주(1999년 이전, 2000~2009년, 2010~2019년)로, ⓒ방법은 총 2가지 범주(1999년 이전, 2000~2019년)로 구성된다.

○ 현행 익명성 기준 5에서 기저 격자 크기를 125m로 설정하는 경우, 각각 6.3%, 9.9%, 23.4%의 격자가 기저 격자로 유지되며, 대략 45.3%, 49.3%, 62.4%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다. 또한 동일한 기준하에 기저 격자 크기를 250m로 설정하는 경우 25.6%, 28.2%, 35.2%의 격자가 기저 격자로 유지되며, 대략 44.8%, 48.3%, 58.9% 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

<표 3-35> (재코딩 후 건축연도별 주택 수) 익명성 기준에 의한 격자 생성 결과

재코딩 방법*	익명성 기준	기저 격자 크기	생성격자 비율					기저격자 비율(%)***
			125m	250m	500m	1km	잔차격자**	
ⓐ	2	125m	15.7	17.5	22.8	44.1	11.7	15.7
		250m	-	29.3	24.3	46.4	11.8	29.3
		500m	-	-	50.3	49.7	7.6	50.3
	3	125m	11.1	18.3	20.0	50.5	10.3	11.1
		250m	-	27.1	20.9	52.0	10.1	27.1
		500m	-	-	43.4	56.6	6.2	43.4
	5	125m	6.3	20.2	18.8	54.8	9.6	6.3
		250m	-	25.6	19.2	55.3	9.2	25.6
		500m	-	-	40.3	59.7	4.9	40.3
ⓑ	2	125m	18.0	18.4	22.5	41.0	11.4	18.0
		250m	-	32.0	24.5	43.5	11.2	32.0
		500m	-	-	52.9	47.1	7.2	52.9
	3	125m	13.7	18.9	20.5	47.0	10.0	13.7
		250m	-	29.5	21.7	48.9	9.7	29.5
		500m	-	-	46.4	53.6	5.9	46.4
	5	125m	9.9	19.9	19.5	50.7	9.6	9.9
		250m	-	28.2	20.1	51.7	9.3	28.2
		500m	-	-	44.0	56.0	5.0	44.0
ⓒ	2	125m	29.5	21.5	25.1	23.9	9.7	29.5
		250m	-	42.8	30.3	26.9	9.7	42.8
		500m	-	-	71.2	28.8	5.4	**71.2**
	3	125m	26.9	19.7	23.7	29.7	10.1	26.9
		250m	-	38.9	27.9	33.2	10.2	38.9
		500m	-	-	63.8	36.2	6.0	**63.8**
	5	125m	23.4	18.1	20.9	37.6	9.8	23.4
		250m	-	35.2	23.7	41.1	9.5	35.2
		500m	-	-	54.2	45.8	5.4	54.2

* ⓐ 방법(범주 4개) : 2010~2019년, 2000~2009년, 1990~1999년, 1989년 이전
ⓑ 방법(범주 3개) : 2010~2019년, 2000~2009년, 1999년 이전
ⓒ 방법(범주 2개) : 2000~2019년, 1999년 이전
** 잔차격자 비율 = (잔차격자 개수 / 전체격자 개수) ×100
*** 기저격자 비율 = (기저격자 개수 / 전체격자 개수) ×100

○ 익명성 기준을 3으로 완화하여 기저 격자 크기를 125m로 설정한다면, 각각 11.1%, 13.7%, 26.9%의 격자가 기저 격자로 유지되며, 250m로 설정하면 27.1%, 29.5%, 38.9%까지 유지될 수 있다.

○ ⓒ방법을 이용하여 항목을 재구성한 이후, 익명성 기준을 3으로 완화하여 기저 격자 크기를 125m로 설정한다면, 26.9%의 격자가 기저 격자로 유지되며, 대략 70.3%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다. 또한 기저 격자 크기를 250m로 설정하면 38.9%의 격자가 기저 격자로 유지되며, 대략 66.8%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ 따라서 익명성 기준을 3으로 완화하고, ⓒ방법을 이용하여 항목을 재코딩하면 500m 격자의 제공도 가능한 것으로 판단된다.

3.2 전국사업체조사자료

□ 사업체 자료에서의 노출위험을 살펴보기 위해, 총 사업체 수, 종사자 규모별 사업체 수, 산업분류별 사업체 수에 대하여 익명성 기준 k를 기반으로 기저 격자에 해당되는 비율을 살펴보았다.

○ 전체적인 자료 구성은 고정형 격자와 같으며, 종사자 규모별 사업체 수, 산업분류별 사업체 수는 고정형 격자에서와 같이 자료 내 관련 항목에 대해 재코딩하여 분석을 수행하였다.

□ 총 사업체 수에 대하여 익명성 기준 k에 따른 쿼드트리 방법론을 적용하여 생성된 격자는 표와 같다.

○ 현행 익명성 기준 3에서 기저 격자 크기를 125m로 설정하는 경우, 54.5%,의 격자가 기저 격자로 유지되며, 대략 83.3%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ 또한 동일한 기준하에 기저 격자 크기를 250m로 설정하는 경우 58.1%의 격자가 기저 격자로 유지되며, 대략 76.9%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ 익명성 기준을 2로 완화하여 기저 격자 크기를 125m로 설정한다면, 58.0%의 격자가 기저 격자로 유지되며, 250m로 설정하면 60.2%까지 유지될 수 있다.

○ 익명성 기준을 3에서 2로 완화하여도 기저격자 비율이 각각 3.5%p, 2.1%p 증가하는 데 그쳐 제공 가능한 격자 크기에 큰 차이가 나지 않는다고 판단되었다. 따라서 익명성 기준을 현행대로 유지하여 500m 격자를 제공하는 것이 적합할 것으로 고려된다.

<표 3-36> (총 사업체 수) 익명성 기준에 의한 격자 생성 결과

익명성 기준	기저 격자 크기	생성격자 비율					기저격자 비율(%)**
		125m	250m	500m	1km	잔차격자*	
1	125m	67.9	12.4	10.3	9.4	0.0	67.9
	250m	-	66.8	17.3	15.8	0.0	66.8
	500m	-	-	72.4	27.6	0.0	72.4
2	125m	58.0	15.7	11.6	14.7	3.5	58.0
	250m	-	60.2	18.7	21.1	2.9	60.2
	500m	-	-	67.9	32.1	1.4	67.9
3	125m	54.5	16.8	12.0	16.7	4.7	54.5
	250m	-	58.1	18.8	23.1	4.3	58.1
	500m	-	-	66.5	33.5	2.3	66.5
4	125m	53.5	16.6	12.3	17.6	5.3	53.5
	250m	-	57.0	18.9	24.1	5.1	57.0
	500m	-	-	66.3	33.7	2.5	66.3
5	125m	52.9	16.6	12.5	18.0	5.7	52.9
	250m	-	56.7	19.1	24.2	5.5	56.7
	500m	-	-	66.7	33.3	2.8	66.7
10	125m	48.1	19.2	13.1	19.6	6.7	48.1
	250m	-	56.4	18.7	24.9	6.3	56.4
	500m	-	-	66.8	33.2	3.1	66.8

* 잔차격자 비율 = (잔차격자 개수 / 전체격자 개수) ×100
** 기저격자 비율 = (기저격자 개수 / 전체격자 개수) ×100

□ 종사자 규모별 사업체 수에 대하여 익명성 기준 k에 따른 쿼드트리 방법론을 적용하여 생성된 격자는 표와 같다.

○ 앞서 언급하였듯이 '다바' 격자에 있는 사업체 내 종사자 수는 최소 1명에서 최대 18,625명으로 매우 큰 범위를 가지는 것이 확인되어 재코딩 작업이 필요하다.

○ 고정형 격자에서와 동일한 방법으로 종사자 규모를 재코딩하여 분석을 수행하였다. ⓐ방법은 총 4가지 범주(1인, 2인, 3인, 4인 이상)로 구성되며, ⓑ방법은 총 3가지 범주(1인, 2인, 3인 이상)로, ⓒ방법은 총 2가지 범주(1인, 2인 이상)로 구성된다.

○ 현행 익명성 기준 3에서 기저 격자 크기를 125m로 설정하는 경우, 각각 37.4%, 46.6%, 43.3%의 격자가 기저 격자로 유지되며, 대략 78.1%, 80.4%, 78.2%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다. 또한 동일한 기준하에 기저 격자 크기를 250m로 설정하는 경우 59.0%, 60.4%, 56.5%의 격자가 기저 격자로 유지되며, 대략 75.6, 76.4%, 73.2% 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ 결과적으로 ⓐ방법을 이용하여 종사자 규모 항목을 정밀하게 재코딩하고, 익명성 기준을 현행대로 유지하여 500m 격자를 제공하는 것이 가능하다고 판단된다.

<표 3-37> (재코딩 후 종사자 규모별 사업체 수) 익명성 기준에 의한 격자 생성 결과

재코딩 방법*	익명성 기준	기저 격자 크기	생성격자 비율					기저격자 비율(%)***
			125m	250m	500m	1km	잔차격자**	
ⓐ	1	125m	48.4	19.4	13.0	19.2	5.9	48.4
		250m	-	56.3	18.9	24.9	5.4	56.3
		500m	-	-	65.5	34.5	3.1	65.5
	2	125m	44.1	22.6	12.5	20.9	8.5	44.1
		250m	-	58.7	16.6	24.7	8.1	58.7
		500m	-	-	68.8	31.2	3.8	68.8
	3	125m	37.4	27.3	13.4	21.9	9.7	37.4
		250m	-	59.0	16.6	24.4	9.3	59.0
		500m	-	-	70.4	29.6	4.6	70.4
ⓑ	1	125m	51.5	17.7	12.0	18.8	4.0	51.5
		250m	-	55.7	18.3	26.0	3.5	55.7
		500m	-	-	62.0	38.0	2.1	62.0
	2	125m	50.1	18.8	11.5	19.6	7.1	50.1
		250m	-	58.2	16.7	25.1	6.8	58.2
		500m	-	-	66.1	33.9	3.7	**66.1**
	3	125m	46.6	22.2	11.6	19.6	7.8	46.6
		250m	-	60.4	16.0	23.7	7.4	60.4
		500m	-	-	69.1	30.9	3.5	**69.1**
ⓒ	1	125m	47.8	17.9	12.2	22.2	2.2	47.8
		250m	-	50.2	18.2	31.6	1.8	50.2
		500m	-	-	53.3	46.7	1.1	53.3
	2	125m	46.5	19.4	11.6	22.5	5.9	46.5
		250m	-	53.7	16.8	29.5	5.6	53.7
		500m	-	-	58.7	41.3	2.9	58.7
	3	125m	43.3	22.7	12.2	21.8	7.1	43.3
		250m	-	56.5	16.7	26.8	6.6	56.5
		500m	-	-	63.7	36.3	3.0	63.7

* ⓐ 방법(범주 4개) : 1인, 2인, 3인, 4인 이상
ⓑ 방법(범주 3개) : 1인, 2인, 3인 이상
ⓒ 방법(범주 2개) : 1인, 2인 이상
** 잔차격자 비율 = (잔차격자 개수 / 전체격자 개수) ×100
*** 기저격자 비율 = (기저격자 개수 / 전체격자 개수) ×100

□ 산업분류별 사업체 수에 대하여 익명성 기준 k에 따른 쿼드트리 방법론을 적용하여 생성된 격자는 표와 같다.

<표 3-38> (재코딩 후 산업분류별 사업체 수) 익명성 기준에 의한 격자 생성 결과

재코딩 방법*	익명성 기준	기저 격자 크기	생성격자 비율					기저격자 비율(%)***
			125m	250m	500m	1km	잔차격자**	
ⓐ	1	125m	3.3	9.5	24.3	63.0	1.6	3.3
		250m	-	12.8	24.3	62.8	1.3	12.8
		500m	-	-	37.1	62.9	1.0	37.1
	2	125m	2.3	0.0	18.2	79.5	0.0	2.3
		250m	-	2.3	18.2	79.5	0.0	2.3
		500m	-	-	20.5	79.5	0.0	20.5
	3	125m	0.0	10.0	10.0	80.0	0.0	0.0
		250m	-	10.0	10.0	80.0	0.0	10.0
		500m	-	-	20.0	80.0	0.0	20.0
ⓑ	1	125m	31.3	22.6	18.6	27.5	5.2	31.3
		250m	-	46.4	22.1	31.4	4.9	46.4
		500m	-	-	59.9	40.1	3.1	59.9
	2	125m	18.2	26.6	21.4	33.7	10.1	18.2
		250m	-	42.1	23.2	34.7	9.0	42.1
		500m	-	-	60.4	39.6	5.0	60.4
	3	125m	11.9	24.9	26.7	36.5	10.3	11.9
		250m	-	35.1	27.9	37.0	9.7	35.1
		500m	-	-	60.2	39.8	5.2	60.2
ⓒ	1	125m	52.2	17.6	13.4	16.9	2.6	52.2
		250m	-	55.5	20.5	24.0	2.2	55.5
		500m	-	-	63.9	36.1	1.2	63.9
	2	125m	51.4	17.0	12.4	19.1	6.1	51.4
		250m	-	55.1	18.8	26.1	6.3	55.1
		500m	-	-	64.5	35.5	3.4	**64.5**
	3	125m	50.6	17.8	12.3	19.2	6.9	50.6
		250m	-	56.3	18.3	25.4	7.1	56.3
		500m	-	-	66.1	33.9	4.0	**66.1**

* ⓐ 방법(범주 4개) : 1차 산업, 2차 산업, 3차 산업, 4차 산업
ⓑ 방법(범주 3개) : (1,2차) 산업, 3차 산업, 4차 산업
ⓒ 방법(범주 2개) : (1,2,4차) 산업, 3차 산업
** 잔차격자 비율 = (잔차격자 개수 / 전체격자 개수) ×100
*** 기저격자 비율 = (기저격자 개수 / 전체격자 개수) ×100

○ 앞서 산업 세세분류가 총 1,196개로 분류되고 있어 조사된 자료를 그대로 제공하는 것은 불가할 것으로 판단되어 전반적 재코딩 작업을 수행하였다.

○ 고정형 격자에서와 동일한 방법으로 종사자 규모 항목을 재코딩하여 분석을 수행하였다. ⓐ방법은 총 4가지 범주(1차 산업, 2차 산업, 3차 산업, 4차 산업)로 구성되며, ⓑ방법은 총 3가지 범주(1·2차 산업, 3차 산업, 4차 산업)로, ⓒ방법은 총 2가지 범주(1·2·4차 산업, 3차 산업)로 구성된다.

○ 현행 익명성 기준 3에서 기저 격자 크기를 125m로 설정하는 경우, 각각 0.0%, 11.9%, 50.6%의 격자가 기저 격자로 유지되며, 대략 20.0%, 63.5%, 80.7%의 격자를 500m보다 작은 크기의 격자로 제공할 수 있다. 또한 동일한 기준하에 기저 격자 크기를 250m로 설정하는 경우 10.0%, 35.1%, 56.3%의 격자가 기저 격자로 유지되며, 대략 20.0, 63.0%, 74.6% 격자를 500m보다 작은 크기의 격자로 제공할 수 있다.

○ ⓐ방법으로 산업분류 항목을 재코딩하고 익명성 기준을 현행대로 유지하는 경우, 1km 격자는 전체 격자의 80%를 차지하게 된다. 이는 단지 익명성 기준보다 작은 격자가 최상위 수준으로 집계된 결과일 수 있어 익명성 기준을 충족시켰다고 보기는 어렵다. 또, 익명성 기준을 10으로 강화하게 되면 대부분의 격자가 조건을 충족시키지 못하게 되므로 쿼드트리 격자를 생성할 수 없게 된다.

○ 따라서 ⓒ방법을 이용하여 산업분류 항목을 재코딩하고, 익명성 기준을 현행대로 유지하여 500m 격자를 제공하는 것이 가능한 것으로 판단된다.

3.3. 소결

□ 2019년 인구주택총조사 자료와 전국사업체조사 자료에 대한 변동형 격자 체계에서의 분석 결과를 종합해 보자.

○ 전체적인 자료의 구성은 고정형 격자 체계와 동일하며, ‘다바’ 격자에서 건물이 존재하는 포인트에 UTM-K(GRS80) 좌표계를 기준으로 격자를 새로이 생성하였다.

○ 노출위험이 낮고 자료의 활용 가능성을 고려하였을 때 자료 제공이 가능하다고 판단할 수 있는 기준으로 기저 격자 비율이 70% 수준인 경우로 설정해 보았다. 이때 기저 격자로는 125m, 250m, 500m 등을 고려하였다.

□ 변동형 격자 체계에서 공표기준 산정을 위해 격자의 크기와 익명성 기준 값, 그리고 기저 격자 비율의 상관관계를 검토해 보자.

<표 3-39> 변동형 격자 체계에서의 상관성

기준	조건	기저 격자 비율
동일한 기저 격자	익명성 기준 값 ↓	↑
동일한 익명성 기준 값	기저 격자 크기 ↑	↑

○ 변동형 격자 체계에서 자료 제공을 위해 고려해야 할 사항으로는 노출위험은 낮추면서 자료의 유용성 확보를 위해 기저 격자 비율이 최대화될 수 있도록 하는 것이다.

○ 이때 동일한 기저 격자 크기에서 익명성 기준 값을 작게 설정할수록 기저 격자 비율도 높아진다. 따라서 자료의 유용성을 고려하여 익명성 기준 값을 가급적이면 작게 할 필요가 있다고 하겠으나, 노출위험을 고려하여 익명성 기준의 최솟값은 2 또는 3으로 정할 수 있다.

○ 반면 동일한 익명성 기준 값에 대해 기저 격자의 크기가 클수록 기저 격자 비율도 높아져서, 격자의 크기는 가급적이면 크게 할 필요가 있다. 그러나 격자의 크기가 클수록 자료의 유용성을 낮아지게 된다.

○ 따라서 익명성 기준이 2 또는 3이면서 기저 격자 비율이 70% 수준을 만족하는 최소 크기의 격자를 공표 기준으로 산정할 수 있다.

□ 인구주택총조사 자료와 전국사업체조사 자료에 대하여 '다바' 격자에서 제공 가능한 변동형 격자의 크기, 익명성 기준 값, 기저격자 비율을 정리하면 표와 같다.

○ 인구, 가구, 주택, 사업체 자료에 대하여 현행 익명성 기준보다는 완화하여 제공이 가능한 것으로 판단된다.

<표 3-40> 변동형 격자 체계의 제공 항목별 공표 기준

분류	제공 항목*	제공 가능한 최소 격자 크기	익명성 기준 값	기저 격자 비율 (%)
인구	총계	125m	2	85.1
			3	72.7
	성별	500m	2	76.1
			3	69.2
	연령①	500m	2	82.0
			3	77.6
가구	총계	500m	2	77.2
			3	67.6
	세대구성②	500m	2	69.5
			3	64.4
주택	총계	250m	2	73.9
			3	64.6
	주택면적③	500m	2	69.7
			3	63.4
	주택유형④	500m	2	80.1
			3	79.1
	건축연도⑤	500m	2	71.2
			3	63.8
사업체	총계	500m	2	67.9
			3	66.5
	종사자 규모⑥	500m	2	66.1
			3	69.1
	산업분류⑦	500m	2	64.5
			3	66.1

* 전반적 재코딩 항목 ① 연령(30세 단위), ② 세대구성 유형(1세대, 2세대 이상, 1인 가구/ 비혈연 가구 제외)
③ 주택면적 (100㎡ 이하, 100㎡ 초과), ④ 주택유형 (아파트, 아파트 이외)
⑤ 건축연도 (1999년 이전, 2000년 이후)
⑥ 종사자 규모 (1인, 2인, 3인 이상), ⑦ 산업분류 ((1, 2, 4차) 산업, 3차 산업)

□ 인구 자료의 경우 현행 익명성 기준인 5보다 완화된 2 또는 3으로 제공이 가능한 것으로 판단되며, 제공 항목에 따라 최소 격자의 크기는 다소 차이가 있다.

○ 총 인구 수의 경우 제공 가능한 기저 격자의 크기는 125m이다. 익명성 기준을 2로 하면 기저 격자의 비율이 약 85%가 되고, 익명성 기준을 3으로 하면 그 비율은 약 73%가 된다.

○ 성별에 따른 인구 수의 경우는 제공 가능한 최소 격자의 크기가 500m로 판단되며 익명성 기준을 2로 하게 되면 기저 격자가 전체의 약 76% 수준이 된다.

○ 연령별 인구 수의 경우는 각 세 단위로 조사된 것을 30세 단위로 전반적 재코딩하여 제공이 가능한 것으로 판단된다. 이때 제공 가능한 최소 격자의 크기는 500m이며 익명성 기준을 2로 설정하게 되면 기저 격자 비율이 82%가 된다.

○ 결론적으로 인구 자료의 경우 총 인구 수는 125m 격자, 성별, 연령별 인구 수는 500m 격자로 제공 가능하다.

□ 가구 자료의 경우도 현행 익명성 기준인 5보다 완화된 2 또는 3을 적용하여 제공이 가능한 것으로 판단되며, 제공 항목에 따라 최소 격자의 크기는 다소 차이가 있다.

○ 총 가구 수의 경우 최소 격자의 크기를 500m로 하고 익명성 기준을 2로 설정하게 되면, 전체 격자 중 약 77%의 격자가 기저 격자로 표시되어 통계 값을 그대로 제공하여 유용성을 확보할 수 있다.

○ 세대구성 유형별 가구 수의 경우는 1세대, 2세대 이상, 1인 가구로 전반적 재코딩하고 비혈연 가구는 제외하면 제공 가능한 최소 격자의 크기가 500m로 판단된다. 익명성 기준을 2로 하게 되면 기저 격자가 전체의 약 70% 수준이 되어 자료 유용성을 확보하게 된다.

○ 결론적으로 가구 자료의 경우 총 가구 수와 세대구성 유형별 가구 수 모두 500m 격자로 제공 가능하다.

□ 주택 자료의 경우 주택면적별 주택 수, 주택유형별 주택 수, 건축연도별 주택 수를 격자통계로 제공하기 위해서는 전반적 재코딩이 필요하다.

○ 총 주택 수의 경우 최소 격자의 크기를 250m로 하고 익명성 기준을 2로 설정하게 되면, 전체 격자 중 약 74%의 격자가 기저 격자로 표시되어 자료의 유용성을 확보할 수 있다.

○ 주택면적별 주택 수의 경우는 주택면적을 100㎡ 이하, 100㎡ 초과로 전반적 재코딩하게 되면 제공 가능한 최소 격자의 크기가 500m로 판단된다. 이때 익명성 기준을 2로 하게 되면 기저 격자를 전체의 약 70% 수준으로 제공이 가능하다.

○ 주택유형별 주택 수의 경우는 주택유형을 아파트와 아파트 외(단독 주택, 연립 주택, 다세대 주택, 비거주용 건물 내 주택, 주택 이외의 거처 등)으로 전반적 재코딩을 수행할 필요가 있다. 이 경우 최소 격자의 크기를 500m로 하면 익명성 기준 2에서 기저 격자 비율을 약 80% 수준으로 제공 가능하다.

○ 건축연도별 주택 수의 경우는 건축연도를 1999년 이전, 2000년 이후로 전반적 재코딩 하게 되면 제공 가능한 최소 격자의 크기가 500m로 판단된다. 이때 익명성 기준을 2로 하게 되면 기저 격자가 전체의 약 71% 수준이 된다.

○ 결론적으로 주택 자료의 경우 총 주택 수는 250m 격자, 주택면적, 주택유형, 건축연도별 주택 수는 500m 격자로 제공 가능하다.

□ 사업체 자료의 경우 종사자 규모별 사업체 수와 산업분류별 사업체 수를 격자통계로 제공하기 위해서는 전반적 재코딩이 필요하다.

○ 총 사업체 수의 경우 기저 격자의 크기를 500m로 하고 익명성 기준을 2로 설정하게 되면, 전체 격자 중 약 68%가 기저 격자로 표현되므로 자료의 유용성을 확보할 수 있다.

○ 종사자 규모별 사업체 수의 경우는 종사자 규모를 1인, 2인, 3인 이상으로 전반적 재코딩을 수행하면 기저 격자 비율을 약 70% 수준으로 하여 격자통계의 제공이 가능하다.

○ 산업분류별 사업체 수의 경우도 산업분류에 대해 (1, 2, 4)차 산업과 3차 산업으로 전반적 재코딩 수행하면 기저 격자 비율을 약 65% 수준으로 하여 격자통계의 제공이 가능하다.

○ 결론적으로 사업체 자료의 경우 총 사업체 수, 종사자 규모 및 산업분류별 사업체 수 모두 500m 격자로 제공 가능하다.

4. 격자 체계에 따른 비교

□ 격자 체계에 따른 변화를 살펴보기 위해 ‘다바’ 격자 내에서도 인구가 밀집된 격자와 사업체가 밀집된 격자를 중심으로 비교해 보았다.

○ 인구주택총조사 자료와 전국사업체조사 자료의 각 항목별 총계 값에 대해서 고정형 격자와 변동형 격자 체계를 각각 적용하여 비교해 보았다.

○ 앞서 격자 체계별 공표 기준에서 정리된 바를 바탕으로 익명성 기준 값으로 2 또는 3을 고려하였다. 이때 제공 항목별로 익명성 기준 값은 동일하게 설정하여 각 격자 체계별로 기저 격자 크기가 달라지므로 격자 지도에서의 변화 양상을 비교해 보고자 한다.

○ 단, 다음에 소개되는 격자 서비스의 예시는 그래프 작성을 위해 선택된 특정 격자나 활용된 자료의 특성에 따라 달라질 수 있음을 밝힌다.

<표 3-41> 고정형 격자 체계와 변동형 격자 체계의 비교

분류	제공 항목	익명성 기준 값	(기저)격자 크기	
			고정형	변동형
인구	총계	2	100m	125m
가구	총계	3	500m	500m
주택	총계	3	500m	250m
사업체	총계	2	1km	500m

□ 총 인구 수에 대하여 익명성 기준 2를 적용한 격자 체계별 서비스 형태를 비교해 보자.

○ 다음에 제시된 격자 서비스 예시 자료에서 왼쪽 그래프는 해당 지역을 기저 격자 크기로 생성한 후 해당 격자의 통계 값이 익명성 기준 값 이하인 경우는 붉은 색으로 표시한 것이다. 오른쪽 그래프는 격자별 통계 값의 밀도(density)에 따라 색을 다르게 하여 표시한 것으로, 밝은 색일수록 밀도가 높은 격자로 이해할 수 있다.

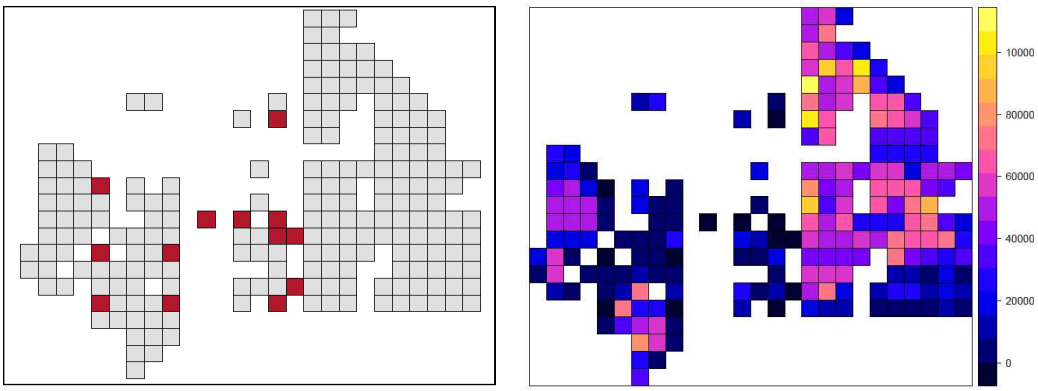
○ 고정형 격자의 왼쪽 그래프에서 익명성 기준 2 이하인 격자가 존재하나, 그 비율이 30% 이하임을 확인할 수 있다. 붉은 색으로 표시된 격자의 통계 값은 N/A 표시하여 제공되고, 대부분의 격자는 관측된 통계 값 그대로 제공이 가능하다.

○ 반면 변동형 격자의 경우 모두 기저 격자인 125m 크기로 제공이 가능함을 확인할 수 있다.

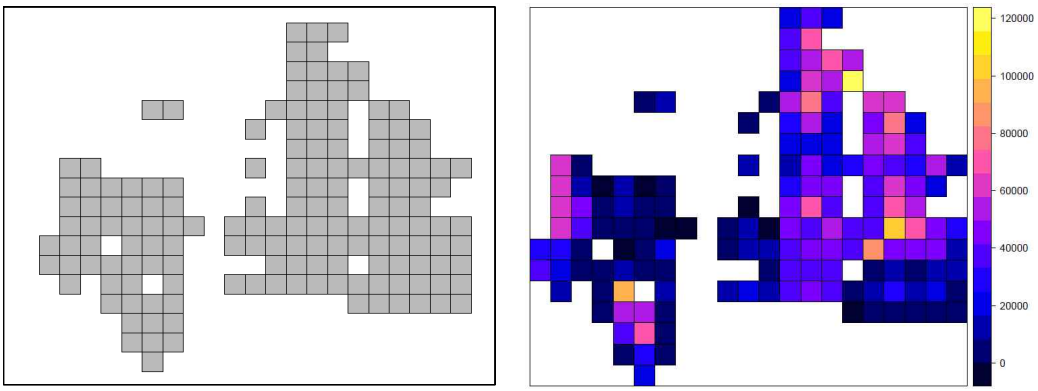
○ 고정형 격자와 변동형 격자를 비교해보면 일차적으로 기저 격자의 크기가 달라 표시된 격자의 위치도 바뀌고, 격자별 밀도도 달라짐을 알 수 있다.

○ 또, 이 경우 변동형 격자보다 고정형 격자의 크기가 작기 때문에 비슷한 위치의 격자 밀도가 고정형 격자에서 더 높게 표시된다.

○ 참고로 여백의 형태로 비어있는 부분은 해당 포인트에 건물이 없어 격자가 생성되지 않는 영역(예, 도로, 산, 강 등)에 해당한다.



(a) 고정형 격자(100m 단위)



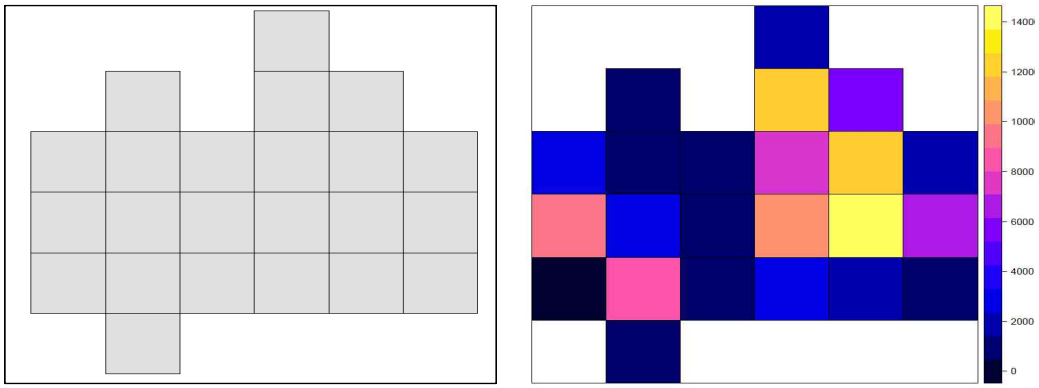
(b) 변동형 격자(125m 단위)

<그림 3-9> 총 인구 수에 대한 격자 서비스 예시(익명성 기준 2인 경우)

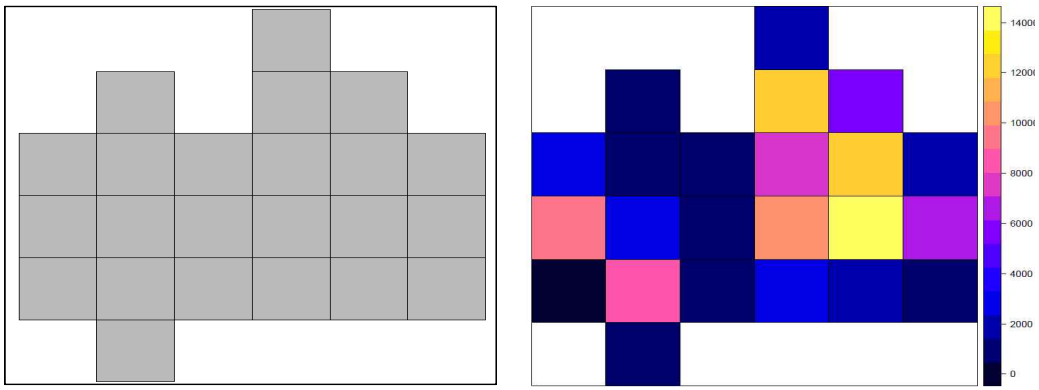
□ 총 가구 수에 대하여 익명성 3을 기준으로 고정형 격자와 변동형 격자 서비스의 형태를 비교해 보자.

○ 총 가구 수에 대한 기저 격자의 크기가 고정형과 변동형에서 동일하게 500m로 반영되었고, 두 격자 서비스의 형태가 동일하게 나타남을 확인 할 수 있다.

○ 변동형 격자의 경우, 해당 격자에서 잔차 격자가 관측되지 않아 고정형 격자와 동일한 형태로 표시되었다.



(a) 고정형 격자(500m 단위)



(b) 변동형 격자(500m 단위)

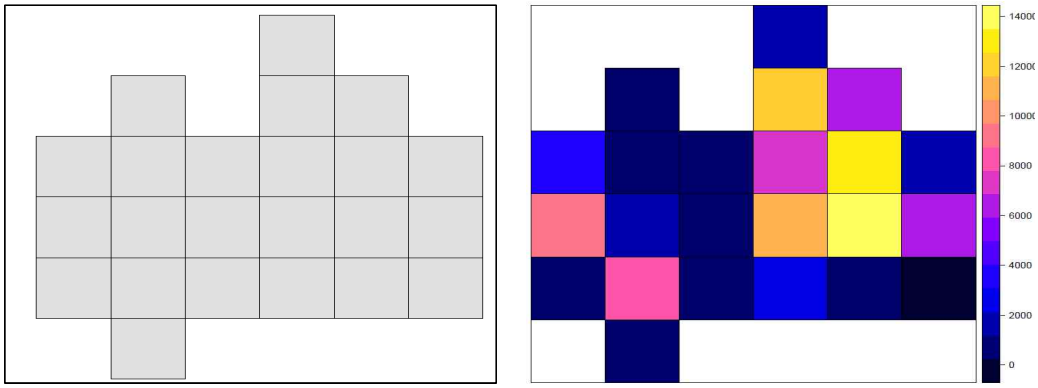
<그림 3-10> 총 가구 수에 대한 격자 서비스 예시(익명성 기준 3인 경우)

□ 총 주택 수에 대하여 익명성 3을 기준으로 격자 체계별 서비스 형태를 검토해 보자.

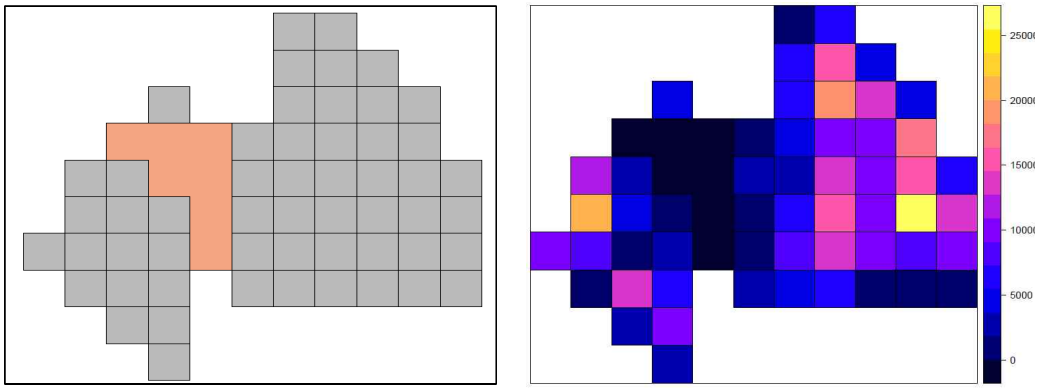
○ 총 주택 수에 대한 기저 격자는 고정형 격자에서는 500m 단위이고, 변동형 격자에서는 250m 단위로 2배의 차이를 보인다.

○ 이 경우 기저 격자의 크기가 더 작은 변동형 격자를 제공하게 되면 밀도가 낮은 부분의 정보는 그대로 전달하면서 보다 세분화하여 밀도를 표시할 수 있으므로 자료의 유용성을 높일 수 있다.

○ 이 격자의 경우 변동형 격자 체계를 적용하면 잔차 격자가 발생하게 되고, 해당 영역의 밀도는 매우 낮은 것을 확인 할 수 있다.



(a) 고정형 격자(500m 단위)



(b) 변동형 격자(250m 단위)

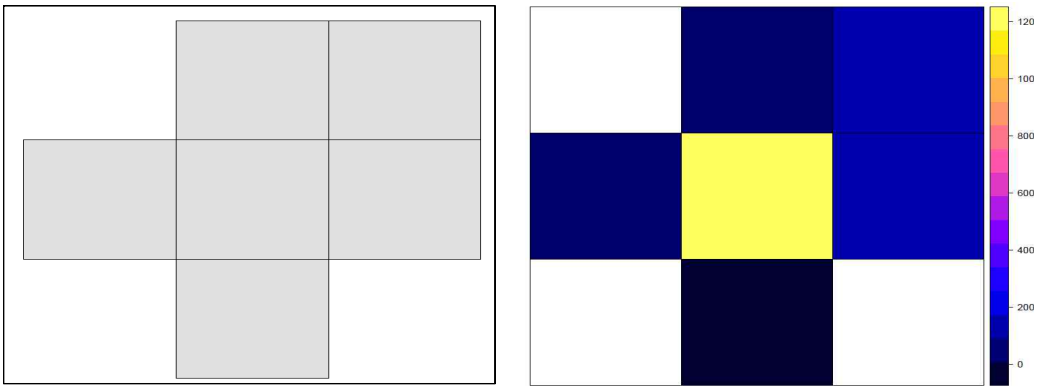
<그림 3-11> 총 주택 수에 대한 격자 서비스 예시(익명성 기준 3인 경우)

□ 총 사업체 수에 대하여 익명성 2를 기준으로 고정형 격자와 변동형 격자의 서비스 형태를 비교해 보자.

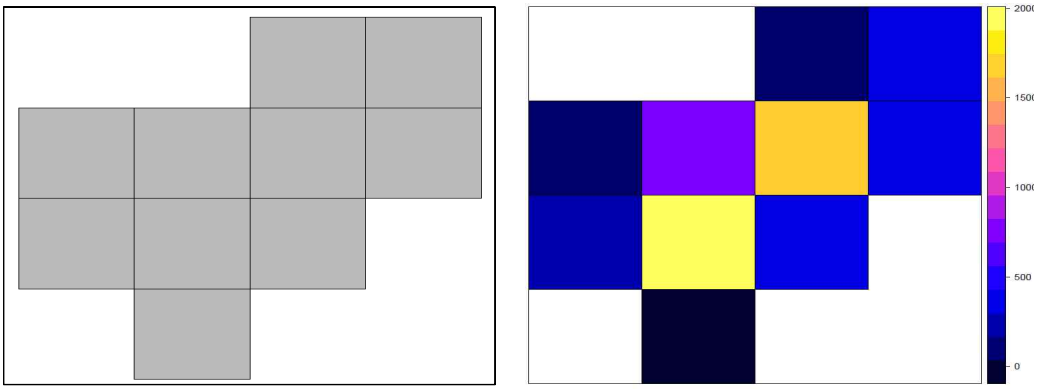
○ 총 사업체 수에 대한 기저 격자는 고정형 격자에서는 1km 단위인데 비해, 변동형 격자에서는 500m 단위로 1/2배 작다.

○ 이 경우 기저 격자의 크기가 더 작은 변동형 격자를 제공하게 되면 밀도가 낮은 부분의 정보는 그대로 전달하면서 보다 세분화하여 밀도를 표시할 수 있으므로 자료의 유용성을 높일 수 있다.

○ 또 해당 격자에서는 변동형 격자 체계를 적용하여도 잔차 격자가 없어, 고정형 격자와 매우 유사한 형태를 보인다.



(a) 고정형 격자(1km 단위)



(b) 변동형 격자(500m 단위)

<그림 3-12> 총 사업체 수에 대한 격자 서비스 예시(익명성 기준 2인 경우)

5. 격자통계 공표기준에 대한 제언

□ 고정형 격자 체계의 경우 1km 이하의 소 격자에서는 항목별 총계 값 제공이 가능한 것으로 판단된다.

○ 구체적으로 총 인구 수는 100m 격자 단위, 총 가구 수 및 총 주택 수는 500m 격자 단위, 총 사업체 수는 1km 격자 단위에서 제공 가능함이 확인된다.

○ 그러나, 항목별로 세분화된 통계는 성별 인구수를 제외하고는 1km 이상의 격자 단위에서 제공이 가능할 것으로 판단된다.

○ 이 경우 제공 항목별 전반적 재코딩 작업이 필수적으로 동반되는데, 자료 제공이 가능한 수준은 재코딩 후 범주의 수가 2~3개 정도로 대폭 축소되어야 한다.

○ 고정형 격자 체계의 경우 전반적 재코딩의 내용에 따라 N/A 표시 비율이 비교적 큰 변화를 보인다. 따라서 제공 가능한 격자통계를 생산하기 위해서는 제공 항목의 특성을 반영한 전반적 재코딩이 되도록 노력해야 한다.

□ 변동형 격자 체계의 경우 대부분의 항목에서 500m 격자 단위로 제공 가능한 것으로 판단된다.

○ 특히 총 인구 수, 총 주택 수의 경우에는 각각 125m와 250m 격자 단위에서 제공이 가능할 것으로 예상된다.

○ 결과적으로 동일한 제공 항목을 고려했을 때 고정형 격자에 비해 더 작은 격자 단위로 제공이 가능하다.

○ 변동형 격자를 적용하는 경우에도 제공 항목별로 전반적 재코딩 작업은 필요하며, 재코딩 후 범주의 수는 4개 이하이면 적용이 가능하다.

○ 그러나 고정형 격자에 비해 전반적 재코딩의 내용에 비교적 강건한(robust) 변화를 보인다. 다시 말해 전반적 재코딩으로 인해 제공 항목의 범주 수를 줄인다고 해서 기저 격자의 비율이 크게 변화하지는 않는 것으로 판단된다.

□ 고정형 격자 체계와 변동형 격자 체계에서 공표기준으로 산정한 전반적 재코딩의 내용은 동일하다.

○ 각 격자 체계별 공표기준에서 언급된 바와 같이, 연령, 세대구성, 주택면적, 주택유형, 건축연도, 산업분류에 대한 전반적 재코딩은 동일하게 적용 가능하다.

○ 그러나 종사자 규모에 대한 전반적 재코딩은 고정형 격자에서는 범주 수가 2개(1인, 2인 이상)인데 비해, 변동형 격자에서는 범주 수가 3개(1인, 2인, 3인 이상)로 다소 차이를 보였다.

□ 마지막으로 앞서 언급된 결과는 현행 익명성 기준을 보다 완화하기 위한 근거로 활용이 가능할 것이다.

○ 인구주택총조사 자료에 대한 현행 익명성 기준 5보다 완화된 2 또는 3으로 적용하여 공표기준을 산정할 수 있음을 앞서 언급하였다.

○ 익명성 기준에 대한 절대적인 기준은 없으며, 제공 자료의 특성이나 제공 기관의 입장, 또는 사회적 합의에 의해 결정하는 것이 통상의 관례로 알려져 있다.

○ 익명성 기준을 2 또는 3으로 설정하여 분석하고 결과를 제시하였으므로, 향후에라도 익명성 기준 설정에 대한 논의가 필요한 시점에 기초자료로 활용이 가능할 것으로 기대한다.

제 4 장

결론 및 시사점

□ 본 연구에서는 향후 SGIS 격자 서비스의 다양화를 위해 현행 제공되고 있는 DB에 대한 소단위 격자 제공 가능성을 검토하였다.

○ 이를 위해 국내·외의 격자통계 제공 현황 및 방법을 조사하고, 각 국가별로 제공하는 격자의 크기와 정보보호 방안을 검토하였다.

○ 또 현재 서비스되고 있는 DB 자료를 입수하여 소단위 격자통계를 산출하고 노출위험 분석을 통해 공표기준을 산정하였다.

□ 제2장에서는 국내·외 격자통계 제공 현황과 정보보호 방안을 검토하여 시사점을 도출하였다.

○ 격자통계 제공 사례는 고정형 격자(fixed size grid) 체계와 변동형 격자(varying size grid) 체계로 구분하였다.

- 고정형 격자 체계는 격자의 크기를 100m, 500m 등으로 사전에 규격화하여 제공하는 것을 의미한다.

- 변동형 격자 체계는 기저(base) 격자의 크기만 사전에 결정하고, 자료 제공 시 다양한 크기의 격자를 허용하는 방식을 의미한다.

○ 대부분의 국가 통계 기관에서는 고정형 격자 체계를 채택하고 있으며, 변동형 격자 체계는 일부 국가에서 연구 차원으로 접근한 것으로 확인된다.

○ 통계청 SGIS의 격자통계 서비스는 제공 항목과 격자의 크기를 다양화할 필요성이 제기된다.

- 일본과 유럽 일부 국가에서 공통적으로 많이 제공하는 항목으로 노동(종사상 지위, 실업수당 등), 농업(토지면적, 농업기업 등), 교육(교육수준, 공교육 참여 등), 소득(평균소득, 구매력) 등이 있고, 네덜란드에서만 제공하는 에너지(평균 가스소비, 전력소비 등) 항목도 눈여겨볼 만하다.

- 격자의 크기와 관련해서는 최소 격자의 크기를 100m, 250m로 설정한 경우가 많다. 또 1km와 10km의 중간 크기로 5km를 제공하는 국가도 상당수 있다.

○ 데이터 정보보호를 위한 방법으로 조사된 사례 모두에서 익명성 기준 이하에 해당하는 격자에 대해 N/A 처리 방식을 택하는 것으로 확인되었다.

- 그러나 익명성 보장의 기준 값으로 인구 자료의 경우 2 또는 3을 사용하는 경우가 다수 있는 것으로 파악되어, 현행 기준인 5보다 완화된 기준 적용에 대한 검토가 필요하다.

○ 마지막 절에서는 유럽연합통계청의 GEOSTAT(2011) 프로젝트에서 확인된 격자통계 사용자의 요구(needs)를 정리하였다.

- 사용자의 관심 영역과 격자 크기 간에는 상관관계가 존재하는데, 소 격자 분석을 하는 경우에는 최대 250m 이하의 격자 크기를 선호하는 것으로 조사된다.

- 노출 제어의 관점에서 제공할 수 없는 값은 특정한 최솟값으로 대체하는 방법을 작은 격자를 더 큰 격자로 집계(aggregation)하는 방법보다 선호하는 것으로 조사되었다.

□ 제3장에서는 사례 검토를 통해 도출된 시사점을 기반으로 통계 주제별 격자통계 공표기준을 제안해 보았다.

○ 분석에 활용한 것은 '다바' 격자의 2019년 인구주택총조사 자료와 전국사업체조사 자료이다. 고정형 격자와 변동형 격자에 대해 k-익명성 모델을 적용하되, 제공 항목별로 전반적 재코딩을 수행하였다.

- 항목별 전반적 재코딩의 구체적인 방법을 나열하면 다음과 같다. 자료 제공이 가능한 수준이 되려면 재코딩 후 범주의 수가 2~3개 정도로 대폭 축소되어야 한다.

<표 4-1> 제공 항목별 전반적 재코딩 방법

분류	제공 항목	재코딩 방법	분류	제공 항목	재코딩 방법
인구	성별	(남, 여)	가구	세대구성	1세대 가구, 2세대 이상 가구, 1인 가구 (비혈연 가구 제외)
	연령	0~30세, 30~60세, 60세 이상	주택	주택면적	100㎡ 이하, 100㎡ 초과
사업체	종사자규모	1인, 2인 이상		주택유형	아파트, 그 외
	산업분류	(1, 2, 4차)산업, 3차 산업		건축연도	1999년 이전, 2000년 이후

○ **(고정형 격자 체계)** 격자통계 제공 기준으로 N/A 표시 비율이 30% 내외인 경우로 설정하여 요약하면 다음과 같다.

- 일부 항목을 제외하고 인구, 가구, 주택, 사업체 관련 항목에서는 익명성 기준 2로 적용이 가능하다.
- 제공 격자의 크기는 총 인구 수에 대해서만 100m 단위가 가능하고, 총 가구 수 및 총 주택 수는 500m, 총 사업체 수는 1km 단위가 가능하다. 총계 값을 제외하고 세부 항목별로 제공 가능한 격자의 크기는 대체로 500m 또는 1km 단위임을 알 수 있다.
- 그러나 주택유형별 주택 수, 종사자 규모별 사업체 수, 산업분류별 사업체 수의 경우는 N/A 표시 비율이 높아 제공이 불가한 것으로 판정된다.
- 고정형 격자의 경우 전반적 재코딩 방법에 따라 N/A 표시 비율이 비교적 큰 변화를 보이므로, 제공 가능한 격자통계 생산을 위해서는 주의가 필요하다.

<표 4-2> 고정형 격자 체계의 제공 항목별 공표 기준 요약

분류	제공 항목	익명성 기준	격자 크기	분류	제공 항목	익명성 기준	격자 크기
인구	총계	2	100m	가구	총계	2 or 3	500m
	성별	2 or 3	500m		세대구성	2 or 3	1km
	연령	2 or 3	1km	주택	총계	2 or 3	500m
사업체	총계	2	1km		주택면적	2	1km
	종사자규모	-	-		주택유형	-	-
	산업분류	-	-		건축연도	2 or 3	1km

주: "-"는 소단위 격자에서 제공 불가

○ **(변동형 격자 체계)** 격자통계 제공 기준으로 기저 격자 비율이 70% 수준인 경우로 설정하여 요약하면 다음과 같다.

- 전체적인 자료의 구성은 고정형 격자 체계와 동일하며, '다바' 격자에서 건물이 존재하는 포인트에 UTM-K(GRS80) 좌표계를 기준으로 격자를 생성하였다.
- 전반적 재코딩 방법은 고정형 격자 체계와 동일하나, 종사자 규모 항목은 1인, 2인, 3인 이상으로 분류되었다.

- 또, 변동형 격자의 경우 전반적 재코딩 방법에 따라 기저 격자의 비율이 크게 변화하지는 않는 것으로 관찰되면, 재코딩 후 범주의 수가 4개 이하이면 적용이 가능한 것으로 판단된다.
- 인구, 가구, 주택, 사업체 관련 항목에서 익명성 기준 2 또는 3으로 적용이 가능함을 알 수 있다.
- 제공 격자의 크기는 대부분의 항목에서 500m 단위로 가능한 것으로 판단되며, 특히 총 인구 수는 125m 단위, 총 가구 수는 250m 단위도 가능하다.

<표 4-3> 변동형 격자 체계의 제공 항목별 공표 기준 요약

분류	제공 항목	익명성 기준	격자 크기	분류	제공 항목	익명성 기준	격자 크기
인구	총계	2 or 3	125m	가구	총계	2 or 3	250m
	성별	2 or 3	500m		세대구성	2 or 3	500m
	연령	2 or 3	500m	주택	총계	2 or 3	500m
사업체	총계	2 or 3	500m		주택면적	2 or 3	500m
	종사자규모	2 or 3	500m		주택유형	2 or 3	500m
	산업분류	2 or 3	500m		건축연도	2 or 3	500m

□ 추가적으로 향후 SGIS 격자통계 서비스와 연계가 가능한 행정자료 기반의 데이터베이스에 대해 검토해 보았다.

○ 먼저 행정자료를 격자통계로 제공하기 위한 전제조건으로 통계를 생산하기 위해 만들어진 DB에 개체별 식별번호와 주소 정보가 존재해야 한다.

- 해당 DB에서 직접적으로 주소를 알 수 없는 경우에는 주소가 있는 다른 DB (예를 들면 인구통계등록부, 기업통계등록부 등)과 연계가 가능해야 한다.
- 만약 행정자료의 주소가 정확하지 않은 경우가 많고, DB 간 연계가 원활하지 않다면 그 문제를 해결하는 것이 관건이 될 수 있다.

○ 국세청 과세자료 중에서 근로소득지급명세서를 통해 소득자의 성별, 근무처, 근무시작일자, 근무종료일자 등을 파악할 수 있으므로 이와 관련한 격자통계 생산이 가능할 것이다.

- 근로소득지급명세서는 근로소득이 있는 사람의 자료만 생성되어 사업소득, 기타소득, 퇴직소득 등과 구분이 가능하므로 이를 활용해 볼 수도 있다.

○ 4대 사회보험 중에서 고용보험 자료를 통해 고용보험 가입자 수, 상실자 수, 자격취득자 수 등의 격자통계 생산이 가능할 것이다.

- 최근 정부에서 발표한 '전 국민 고용보험 가입' 정책이 가시화된다면 향후에 더욱 격자통계가 필요할 수 있다.

- 그러나 일본, 네덜란드 등에서 제공하고 있는 종사상 지위, 근로 유형, 실업 수당 등의 항목은 아직 수집되고 있지 않아 제공이 어려운 것으로 파악된다.

○ 일본, 노르웨이 등 에서 제공하고 있는 농업 분야 통계와 유사한 항목은 격자통계 생산이 가능할 것으로 기대된다.

- 농지원부에서 농업인 개인대체식별번호, 주소, 농지소재지, 농지면적 등을 파악할 수 있으므로 제공이 가능할 수 있다.

○ '20년에 구축된 18세 미만 아동가구 DB를 활용하면 가구유형, 가구원 수, 부모의 근무형태(상용, 임시/일용, 자영업, 특수고용 등), 부모가 임금 근로자인 경우 월평균 급여액 등에 관한 격자통계 생산을 시도해 볼 수 있다.

○ 재산세(주택) 자료를 통해 제공이 가능한 항목이 있는지 검토해 보았으나, 현행 제공 항목(주택 면적, 건축연도 등) 외에 추가로 제공할 수 있는 항목은 없는 것으로 파악 된다.

- 만약 재산세(주택) 부과 금액 등을 제공 받을 수 있다면 격자통계 제공이 가능하고, 많은 관심을 받을 것으로 예상된다.

□ 마지막으로 본 연구의 성과와 한계를 정리하고자 한다.

○ (성과) 소 단위 격자통계 제공에 있어 필요한 기준을 구체적인 수치로 제시하여, 객관적 판단이 가능하도록 하였다.

- 고정형 격자에서는 100m, 500m 단위, 변동형 격자에서는 125m, 250m, 500m 단위의 통계 제공을 위해 필요한 전반적 재코딩 방법을 제안하였다.

- 익명성 기준 값에 대해 현행 수준과 완화된 기준을 적용한 경우 수치의 변화를 모두 제시함으로써 판단의 기준이 될 수 있도록 하였다.

○ (한계) 격자통계를 제공하는 해외 사례의 정보보호 수준에 대하여 보다 면밀한 검토가 필요하다.

- 본 연구에서는 해외 사례의 전반적인 현황을 검토 하였지만, 정보보호와 관련한 보다 세부적인 내용을 확인하기 위해서는 국가별 제공 항목의 데이터 정의

서에 포함된 처리 규정을 모두 검토해야 한다.

- 그러나 국가별 데이터 정의서는 대부분 자국의 언어로만 작성되어 있어, 이를 모두 검토하기에는 한계가 있었다.

참고문헌

Andersson, M., Klaesson, J. & Larsson, J. P. (2012). *How local are spatial density externalities? Evidence from square grid data*. Technical report (No. 2012/10). Lund University, CIRCLE-Center for Innovation, Research and Competences in the Learning Economy.

Armstrong, M. P. & Ruggles, A. J. (2005). Geographic information technologies and personal privacy. *Cartographica : The International Journal for Geographic Information and Geovisualization, 40*, 63-73.

Armstrong, M. P., Rushton, G. & Zimmerman, D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in medicine, 18*, 497-525.

Behnisch, M., Meinei, G., Tramsen, S. & Diesselmann, M. (2013). Using Quadtree Representations in Building Stock Visualization and Analysis. *ERDKUNDE 67(2)*, 151-166.

Beresford, A.R. & Stajano, F. (2003). Location privacy in pervasive computing. *IEEE Pervasive computing, 2*, 46-55.

Duncan, G. T., Elliot, M. & Salazar-Gonzalez, J. J. (2011). *Statistical confidentiality", Statistics for social and behavioral sciences*. New York, Springer.

GEOSTAT 1A. (2011). *ESSnet porject GEOSTAT 1A-representing census data in a European population grid*. Technical report. The European Forum for GeoStatistics.

GEOSTAT 1B. (2013). *ESSnet porject GEOSTAT 1B-representing 2011 census data on grid*. Technical report. The European Forum for GeoStatistics.

GEOSTAT 1B. (2014). *ESSnet porject GEOSTAT 1B-representing census data in a European population grid*. Technical report. The European Forum for GeoStatistics.

Hampton, K. H., Fitch, M. K., Allshouse, W. B., Doherty, I. A., Gesink, D. C., Leone, P. A., Serre, M. L. & Miller, W. C. (2010). Mapping health data : improved privacy protection with donut method geomasking. *American Journal of Epidemiology, 172*, 1062-1069.

Hunter, G. M. (1978). *Efficient computation and data structures for graphics*. PH. D. thesis, Princeton, NJ, USA.

KaMinGer, I. & Meyer, W. (2007). *Neue raster-orientierte statistik in Europa*. AGIT-Symposium Salzburg, 303-308.

Kardos, J., BenWell, G. I. & Moore, A. (2005). The visualisation of uncertainty for spatially referenced census data using hierarchical tessellations. In : *Transactions in GIS, 9(1)*, 19-34.

Kwan, M. P., Casas, I. & Schmitz, B, C. (2004). Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks?. *Cartographica: The International Journal for Geographic Information and Geovisualization, 39*, 15-28.

Lagonigro, R., Oller, R. & Martori, J. C. (2017), A quadtree approach based on European geographic grids: reconciling data privacy and accuracy. *SORT 41(1)*, 136-158.

Lukan, J., Smukavec, A. (2017). *Statistical disclosure control considerations of publishing data on*

grid squares and territorial units in the application STAGE. Proceedings 2017 Statistical IDC.

Szibalski, M. (2007), Kleinräumige Bevölkerungs- und Wirtschaftsdaten in der amtlichen Statistik Europas. In: *Wirtschaft und Statistik 2*, 137−143.

Statistics Austria. (2021). *Statistical data based on statistical grids - Data and Fees*.

Statistics Bureau of Japan. (2015). *Outline of the Grid Square Statistics, Compiled from the 2015 Population Census Results.*

Strobl, J. (2005). *Hierarchishe aggregation: Detailinformation versus Datenschutz am Beispiel adressbezogen georeferenzierter datensatze*. Salzburger Geographische Arbeiten 38, Salzburg, 163-171.

Vu, K., Zheng, R. & Cao, J. (2012). *Efficient algorithms for k-anonymous location privacy in participatory sensing*. In INFOCOM, 2012 Proceedings IEEE, 2399-2407.

Younghee, H. & Min-Jeong, P. (2020). *A study on the application of confidentiality to frequency tables provided by SGIS*. SRI.

국토지리정보원 https://www.ngii.go.kr/kor/main.do

미국 인구조사 그리드 https://sedac.ciesin.columbia.edu/data/collection/usgrid

미국 항공우주국(NASA) https://earthdata.nasa.gov/

스웨덴 통계청 https://www.scb.se/eng/index.asp

영국 통계청 https://www.ons.gov.uk/

오스트리아통계청 http://www.statistik.at/web_en/classifications/regional_breakdown/grid/index.html

유럽연합통계청 https://ec.europa.eu/eurostat/web/main/home

일본 통계청 https://www.stat.go.jp/english/index.html

핀란드 통계청 https://www.stat.fi/index_en.html

통계청 SGIS 통계지리정보서비스 https://sgis.kostat.go.kr/view/index

부 록

1. 국토지리정보원의 국토조사 관련 근거 법령

□ 국토지리정보원은 『국토기본법』 제25조 및 동법 시행령 제10조에 의해 국토조사를 실시하고 있다(<부표 1-1>).

○ 국토조사는 국토에 관한 계획 정책의 수립 등에 활용될 수 있는 자료를 조사하고, 이를 공표함으로써 효율적인 국토의 이용 및 관리를 지원하기 위해 시행되어 진다.

○ 국토조사의 목적은 첫째, 국토의 현황과 변화상을 조사 · 취합 · 관리하고, 둘째, 국토와 관련된 시대상 및 사회적 요구사항이 반영된 정보를 수집 · 가공하며, 셋째, 국토계획 및 평가를 위한 기초자료로 활용될 수 있도록 지원하고 공공 및 민간에게 양질의 국토정보를 제공함에 있다.

<부표 1-1> 국토조사 관련 근거 법령

관련 법령	내용
국토기본법 제25조	제25조(국토 조사) ① 국토교통부장관은 국토에 관한 계획 또는 정책의 수립, 「국가공간정보 기본법」 제32조제2항에 따른 공간정보의 제작, 연차보고서의 작성 등을 위하여 필요할 때에는 미리 인구, 경제, 사회, 문화, 교통, 환경, 토지이용, 그 밖에 대통령령으로 정하는 사항에 대하여 조사할 수 있다. ③ 국토교통부장관은 효율적인 국토조사를 위하여 필요하면 제1항에 따른 조사를 전문기관에 의뢰할 수 있다.
국토기본법 시행령 제10조	제10조(국토조사의 실시) ②국토조사는 다음 각호의 구분에 따라 실시하며, 국토교통부장관은 국토조사를 효율적으로 실시하기 위하여 국토조사 항목 및 조사주체 등 필요한 사항에 대하여 관계 중앙행정기관의 장 및 시・도지사와 사전협의를 거쳐 국토조사계획을 수립할 수 있다. ③ 국토조사는 행정구역 또는 일정한 격자(格子) 형태의 구역 단위로 할 수 있다. 제19조(권한의 위임) 국토교통부장관은 법 제33조제1항에 따라 다음 각 호의 권한을 국토지리정보원장에게 위임한다.

2. 국가지점번호

□ 국가지점번호는 도로명주소가 없는 지역에서 응급상황 발생 시 위치참조를 하기 위해 마련된 위치참조체계로서 국토 및 이와 인접한 해양을 격자형(grid)으로 일정하게 구획한 지점마다 부여한 번호(도로명주소법 제2조제10호)이다.

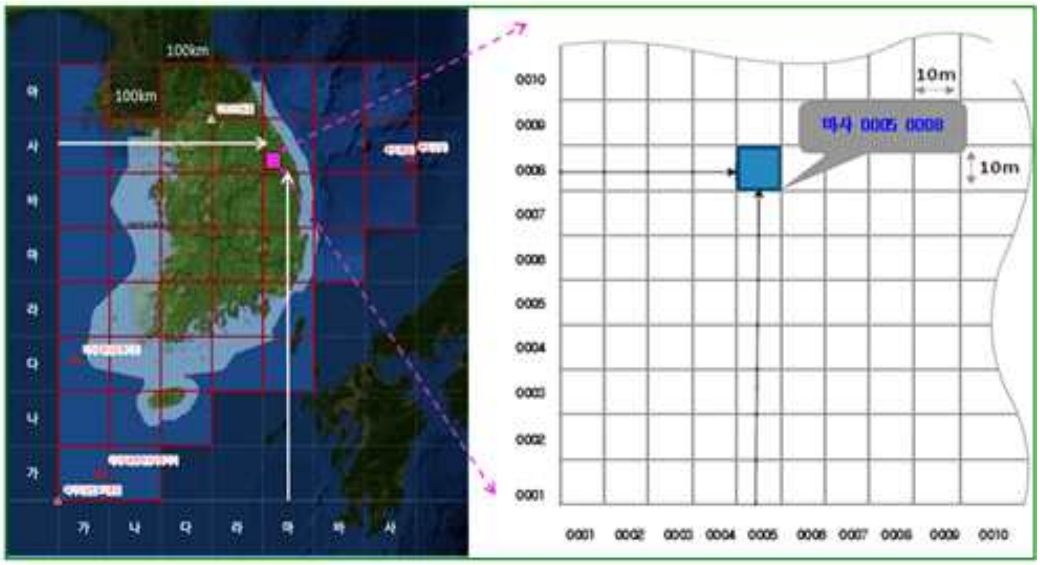
○ 부여 기준은 100km 단위는 문자를 사용하고 이하부터는 가로와 세로를 10,000으로 나눈 정수의 숫자를 연결하여 사용하고 있다(도로명주소법 시행령 제11조의13).

○ 국가지점번호 기준점 : 단일평면직각좌표계(UTM-K)의 서쪽 300km, 남쪽 700km

○ 타원체: 세계측지계(GRS80)

○ 표기방법: 첫 번째 문자 동쪽으로 가나다순, 두 번째 문자 북쪽으로 가나다순으로 부여

○ 표현 최소 단위:10m

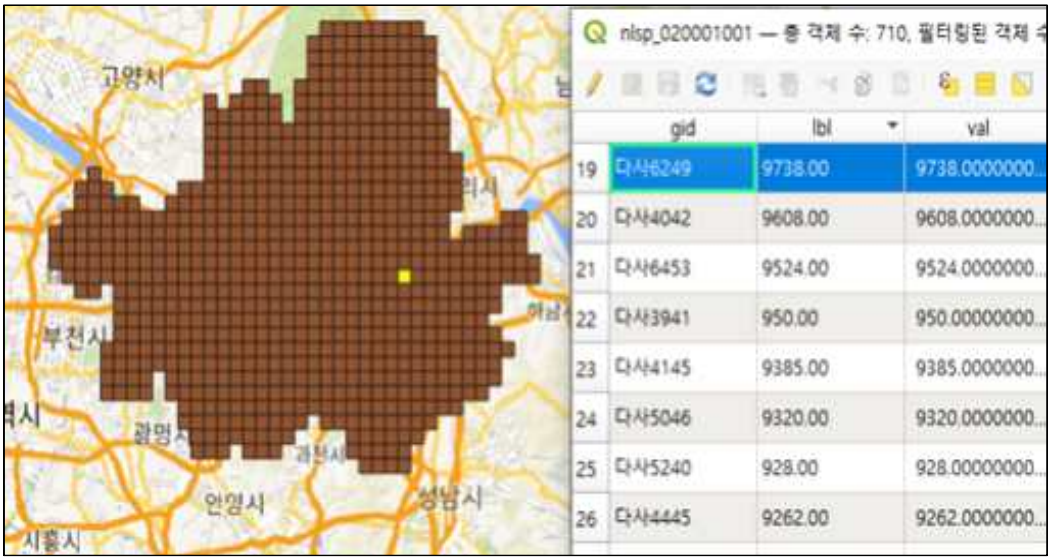


<부그림 2-1> 국가지점번호 예시(마0005(가로), 사0008(세로) → 마사 00050008)

3. 국토통계지도의 다운로드 서비스

□ 국토통계지도의 다운로드 서비스에서 제공하는 자료의 형태는 래스터 데이터가 아닌 벡터(폴리곤) 데이터이다.

○ 벡터의 속성정보에는 gid(국자지점번호) 필드가 포함되어 격자의 위치를 알 수 있으며, 나머지 필드에 격자가 가지는 속성(예를 들면, 인구 등)을 나타내고 있다.



<부그림 3-1> 1km 격자의 총인구 데이터 도형정보(좌)와 속성정보(우)

□ 자료를 조회한 후 다운로드 한 파일은 <부그림 3-2>와 같이 도형정보를 나타내는 .shp 파일, 속성정보를 나타내는 .dbf 파일, 도형정보와 속성정보의 연결정보를 가진 .shx 파일, 그리고 투영정보를 포함한 .prj 파일로 구성되어 있다.

파일명	압축크기	원본크기	압축률	종류
nlsp_020001001.dbf	6,508	342,673	98%	DBF 파일
nlsp_020001001.prj	378	761	50%	PRJ 파일
nlsp_020001001.shp	12,535	86,052	85%	SHP 파일
nlsp_020001001.shx	1,511	5,156	71%	SHX 파일

<부그림 3-2> 국토통계지도의 다운로드 파일 구성

□ 국토지리정보원의 국토통계지도 서비스는 격자 형태로만 제공되는 것은 아니다.

○ 시도, 시군구, 읍면동, 동리 등의 법정경계, 우편번호인 국가기초구역, 도시지역, 관리지역, 농림지역, 자연환경보전지역으로 구분되는 국토의 계획 및 이용에 관한 법률에 의한 용도지역 단위로도 통계지도를 제공하고 있다(<부표 3-1>).

○ 따라서 국토통계지도는 격자단위, 법정경계, 국가기초구역, 용도지역 등으로 구분하여 제공되며, 모든 국토통계지도는 모두 벡터(폴리곤) 형식 공간정보로 제공된다.

<부표 3-1> 국토지리정보원의 국토통계지도 서비스 유형 및 서비스 단위

서비스 유형	서비스 단위
격자	100km, 10km, 1km, 500m, 250m, 100m (50m, 10m 격자는 통계 서비스 제외)
법정경계	시도, 시군구, 읍면동, 리
국가기초구역	국가기조구역(우편번호)
용도지역	도시지역, 관리지역, 농림지역, 자연환경보전지역

4. 일본의 격자단위 메쉬(Mesh)

□ 일본의 격자통계는 센서스의 기본 단위인 '기본 단위구'에 속하는 지역을 메쉬(Mesh) 구획에 대응하는 방식을 이용하고 있다. 따라서 일본의 격자통계를 이해하기 위해서는 일본의 격자 단위인 메쉬에 대한 이해가 필요하다.

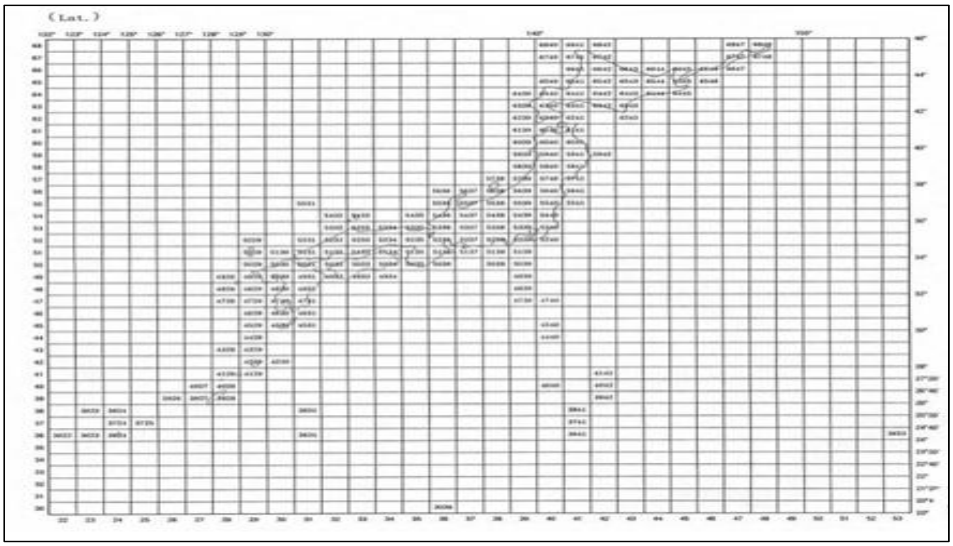
○ 일본의 격자 체계는 1963년 표준 지역 메쉬 및 표준 지역 메쉬법, 1976년 일본 공업 규격 코드(JIS C 6304, JIS X 0410) 등을 거치면서 마련되었다.

○ 일본은 국가, 지방 자치 단체는 물론 민간 기관 등 각 방면에서 메시(격자)법에 의해 지역 정보를 수집하여 활용하고 있다.

□ 중위도에 위치한 일본은 위도 1도의 간격이 경도 1도의 간격보다 길기 때문에 격자가 정사각형 형태를 갖추기 위해 위도(동서) 방향과 경도(남북) 방향의 간격을 2:3으로 설정하고 있다.

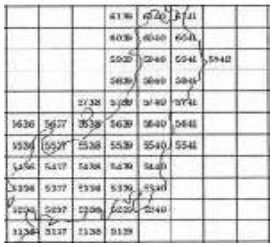
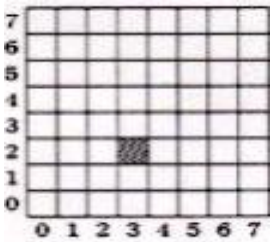
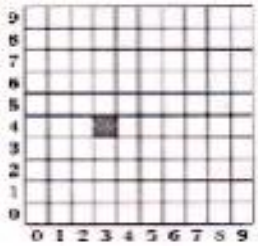
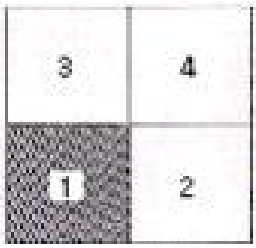
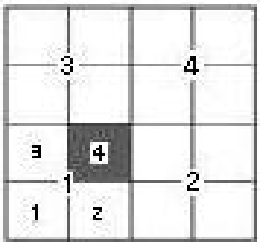
○ 격자는 1차 메쉬로부터 분할되며, 1차 메쉬는 북위 20도에서 46도 사이, 동경 122도에서 154도 사이에 해당한다. 1차 메쉬는 위도 40분 간격, 경도 1도 간격으로 나뉘며, 4자리 숫자로 구성되어 있는데 처음 2자리는 30에서 68 사이로 위도 방향을 마지막 2 자리가 22에서 53 사이로 경도 방향을 나타낸다(<부그림 3-1>).

○ 이후 2차 메쉬, 3차 메쉬, 4차 메쉬, 5차 메쉬 등은 계층적으로 구분되는 방법을 사용하고 있다(<부표 4-1>).



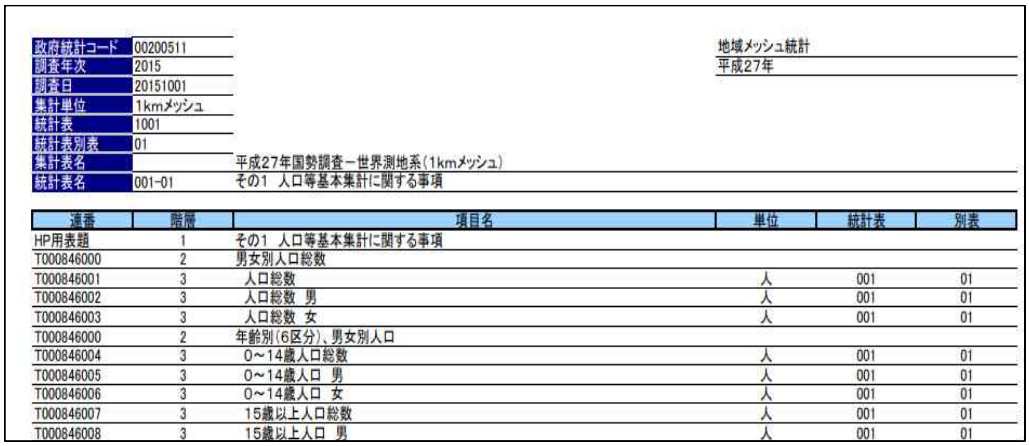
<부그림 4-1> 일본의 1차 격자

<부표 4-1> 일본의 메쉬 코드

분할	구분방법	위도간격	경도간격	지상거리	메쉬코드 예
1차 메쉬	위도 2/3도와 경도 1도	40분	1도	약 80km	5339 
2차 메쉬	1차영역을 8×8 영역으로 64분할	5분	7분 30초	약 10km	5339-23 
기본격자 (3차 메쉬)	2차영역을 10×10 영역으로 100분할	30초	45초	약 1km	5339-23-43 
1/2 격자 (4차 메쉬)	기본격자를 2×2로 4분할	15초	22.5초	약 500m	5339-23-43-1 
1/4 격자 (5차 메쉬)	기본격자를 4×4로 16분할	7.5초	11.25초	약 250m	5339-23-43-1-4 

(a) 3차 메쉬 경계 파일(shp 포맷)

(b) 3차 메쉬 인구통계 자료(csv 포맷)



政府統計コード	00200511
調査年次	2015
調査日	20151001
集計単位	1kmメッシュ
統計表	1001
統計表別表	01
集計表名	平成27年国勢調査－世界測地系（1kmメッシュ）
統計表名	001-01　その1　人口等基本集計に関する事項

地域メッシュ統計
平成27年

連番	階層	項目名	単位	統計表	別表
HP用表題	1	その1　人口等基本集計に関する事項			
T000846000	2	男女別人口総数			
T000846001	3	人口総数	人	001	01
T000846002	3	人口総数　男	人	001	01
T000846003	3	人口総数　女	人	001	01
T000846000	2	年齢別(6区分)、男女別人口			
T000846004	3	0～14歳人口総数	人	001	01
T000846005	3	0～14歳人口　男	人	001	01
T000846006	3	0～14歳人口　女	人	001	01
T000846007	3	15歳以上人口総数	人	001	01
T000846008	3	15歳以上人口　男	人	001	01

(c) 3차 메쉬 데이터 정의서

<부그림 4-2> 일본의 메쉬 자료 제공 예시

5. 해외 국가별 격자통계 제공 웹사이트

□ 제2장의 고정형 격자 체계에서 소개된 해외 국가별 격자통계 서비스 제공 웹사이트를 정리하였다. 단, 미국, 영국은 공식적인 격자통계 서비스가 없어 제외하였다.

○ 해외 국가의 웹사이트는 대부분 자바스크립트 기반의 인터렉티브 서비스와 다운로드 서비스를 제공하고 있다. 일부 국가의 다운로드 서비스는 유료로 제공된다.

○ 국가마다 자국어가 달라 웹사이트를 통해 조사를 완벽히 하기는 쉽지 않으며, 일부 국가의 경우 구글 번역 활용하여 조사하였다.

<부표 5-1 > 해외 국가별 격자통계 제공 웹사이트

국가	해당 웹사이트
일본	영문: www.stat.go.jp/english/data/mesh/index.html 일본어: www.e-stat.go.jp/gis - 인터렉티브: jstatmap.e-stat.go.jp/jstatmap/main/trialstart.html - 다운로드: www.e-stat.go.jp/gis/statmap-search?type=1
유럽연합 통계청	다운로드:ec.europa.eu/eurostat/web/gisco/geodata/reference-data/grids
네덜란드	인터렉티브: cbsinuwbuurt.nl/ 다운로드 :www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/kaart-van-100-meter-bij-100-meter-met-statistieken
노르웨이	인터렉티브:kart.ssb.no 다운로드:www.ssb.no/natur-og-miljo/geodata#Nedlasting_av_rutenettsstatistikk
핀란드	인터렉티브:kartta.paikkatietoikkuna.fi/?lang=en 다운로드(유료): 이메일
오스트리아	인터렉티브:statistik.at/atlas 다운로드(유료):www.statistik.at/atlas/reg-datenkatalog/
독일	인터렉티브(센서스):atlas.zensus2011.de/ 인터렉티브(농업):www.atlas-agrarstatistik.nrw.de/ 다운로드(통계) :www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html?nn=3065474#Gitter 다운로드(위치) :gdz.bkg.bund.de/index.php/default/inspire/sonstige-inspire-themen/geographische-gitter-fur-deutschland-in-lambert-projektion-geogitter-inspire.html
에스토니아	인터렉티브:estat.stat.ee/StatistikaKaart/VKR 다운로드:estat.stat.ee/StatistikaKaart/VKR

6. 격자의 다양한 형태

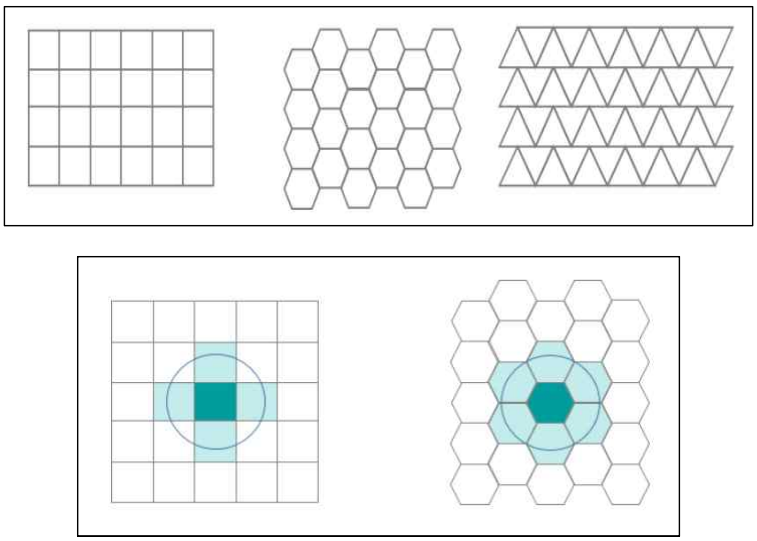
□ 격자의 형태는 정삼각형, 사각형, 육각형 등이 존재할 수 있으나, 실제로 사용되는 대부분의 격자는 사각형 형태이다.

○ 이론적으로는 육각형 격자가 가장 안정적인데, 특정 격자에 인접한 6개의 격자까지의 거리가 $\sqrt{3}$으로 모두 동일하기 때문이다. 또, 원형에 가장 가까운 형태로 표현이 가능하기 때문에 활용도가 높을 수 있다. 하지만 실제 구현이 쉽지 않다는 단점이 있다.

○ 정삼각형 격자의 경우 3개의 점으로 한 개의 면을 만들면 하나의 평면이 되므로 내삽이 용이하다는 장점이 있다.

○ 사각형 격자의 경우 인접하는 격자가 8개 존재하지만 격자 간의 거리가 달라지는 단점이 있다. 동서남북으로 인접하는 4개의 격자까지의 거리가 1이라고 가정하면, 모서리 부분(북동, 남동, 남서, 북서 방향)에 인접하는 4개의 격자까지의 거리는 $\sqrt{2}$가 되어, 인접한 격자를 정의하기에 명확하지 않은 부분이 있다.

○ 사각형 격자의 장점은 행렬로 표시할 수 있어 가장 편리하다는 것이고, 이것은 실제 구현에 있어 매우 중요한 이점이 된다. 사각형 격자는 위치를 표시할 때 행과 열이 모두 정수로 표현이 가능하지만, 삼각형, 육각형 격자에서는 정수로 표현이 불가한데, 이것은 PC에서 색상을 행렬로 표시하는 것과 관련이 있다.



출처 https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-whyhexagons.htm

<부그림 6-1> 다양한 격자의 형태 예시

Abstract

A Study on the establishment of publication standards for SGIS grid services by statistical topic

Younghee Hong, Yeongju Oh, Jaejoon Jeong

In this study, to diversify grid statistics of the Statistical Geographic Information Service(SGIS), domestic and foreign cases were investigated and benchmarked to review the possibility of providing smaller unit grid statistics.

The grid statistics cases were divided into a fixed size grid system and a varying size grid system. The size of the fixed size grid used by most national statistical institutions is standardized, and in many cases, the minimum grid size is set to 100m and 250m. Varying size grids have been used at the research level in some countries in a way that allows grids of various sizes, and the minimum grid sizes vary at 125m, 250m, and 500m. As a method of data privacy, it is marked as N/A in the fixed size grid if it falls below the anonymity standard, and in the varying size grid, it is aggregated into a larger grid. At this time, because many cases use 2 or 3 as the anonymity standard value, it is judged that the current standard of Statistics Korea needs to be relaxed.

In the actual data obtained from the 2019 Population and Housing Census and the National Business Survey of the 'DABA' grid, a fixed size grid and a varying size grid system based on k-anonymity were applied and analyzed. Overall recoding for each provided item was required to reduce the risk of exposure in both grid systems, and the publication standard was calculated in consideration of information loss. As a result, the N/A ratio was determined to be 30% in the fixed size grid, and the base grid ratio was determined to be 70% in the varying size grid. The small grid that can be provided in anonymity standard 2 or 3 is somewhat different depending on the items, and when reviewed under the same conditions, it was found that a varying system can provide a smaller grid than the fixed system.

Finally, we hope that this study can be used as basic reference to provide various unit grid statistics in the future.

Keywords : fixed size grid, varying size grid, global recoding, disclosure

집필진

- 홍영희 (여성가족부 사무관)
- 오영주 (통계청 통계개발원 통계방법연구실 주무관)
- 정재준 (성신여자대학교 지리학과 교수)

연구보고서 2021-02

통계주제별 통계지리정보 격자 서비스의 공표기준 산정 연구

발 행	2022년 4월 13일
발 행 인	통계개발원장 직무대리 김진
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 공공누리 공공저작물 자유이용허락

통계청
통계개발원