

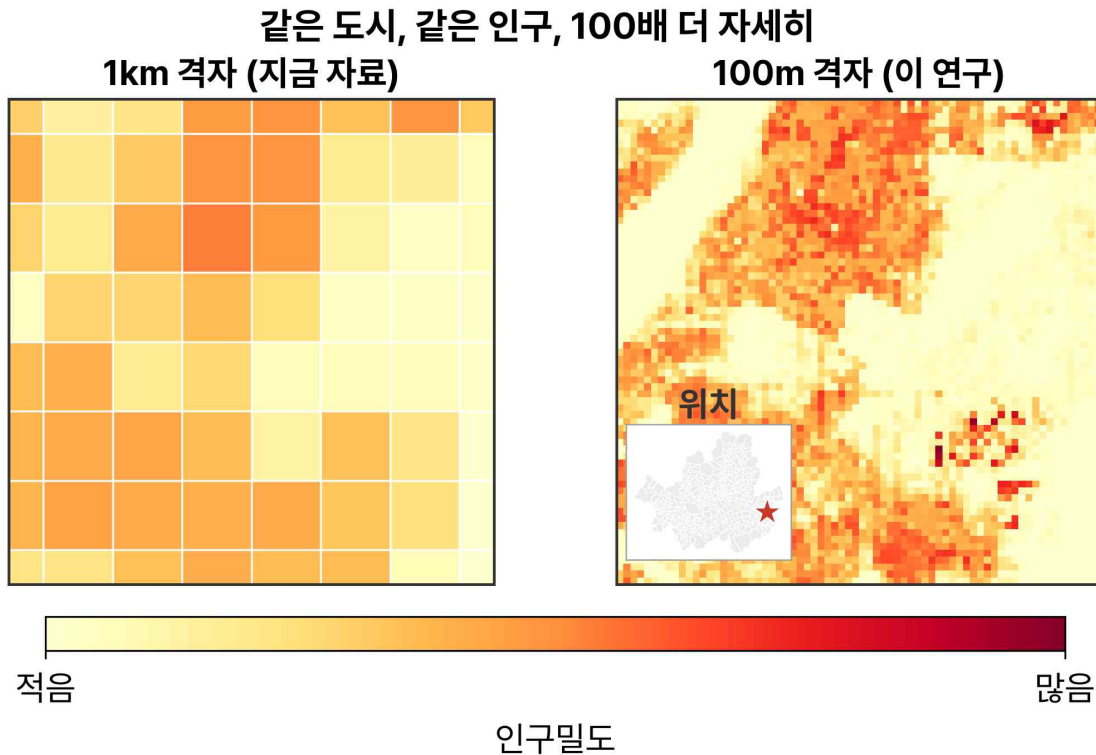
□ 추진배경 및 필요성

가로세로 1km 네모 하나에 아파트 단지와 산이 함께 담긴다. SGIS 1km 인구격자는 그 안 어디에 사람이 몰려 있는지 보여주지 못한다.

본 사례는 SGIS 1km 격자를 구속조건으로 전국 100m 인구격자와 행정동 신뢰등급 명단을 만들었다. WorldPop 전국 합계는 SGIS 총조사 인구와 0.17%밖에 차이 나지 않지만, 칸 단위 오차(RMSE)는 743.2명에 달한다. 총량이 맞는다고 칸 안의 분포까지 맞는 것은 아니다.

더 세밀한 실측값을 누구나 바로 쓸 수 없는 데는 이유가 있다. SGIS는 통계법 제33조에 따라 K-익명성 기준으로 정보손실을 통제해 공표하며(인구 1명 이상 격자의 14.7%가 5의 배수, 별첨), 개발지원센터 100m 격자경계는 면적 10km² 이하 읍면동에서만 열린다. 100m 격자 통계는 로그인·신청 절차를 거쳐야 해, 본 사례는 이를 쓰지 않고 공개 자료만으로 배분했다. 본 사례는 이 원칙을 우회하지 않는다. 100m 값은 실측값의 공개가 아니라 SGIS 1km 합계를 보존하도록 나눈 모형 추정치다.

같은 WorldPop 값도 1km 합을 SGIS 값에 묶으면 MAPE가 14.7%에서 9.83%로 떨어진다 - SGIS 1km 통계 없이는 나오지 않는 결과다.



〈그림 1〉 SGIS 1km 인구격자와 이 연구가 만든 100m 인구격자 비교 - 전국에서 1km 격자 간 대비가 가장 큰 지점(서울 송파·강동 일대, 7km×7km) (자료: 국가데이터처 SGIS 격자통계, WorldPop 100m, OpenStreetMap / 기준:2024년)

□ 추진과정

■ 활용분야 및 주요내용:

SGIS 자료제공에서 1km 인구격자 통계와 격자경계 SHP 전수(107,680칸), 행정동 경계·인구 통계를 받았다. 좌표계 통일·질량보존 빈닝·공변량 산출 절차는 별첨 「1. 방법 상세」에 실었다.

대화형 통계지도(총조사 시각화지도·통계주제도)로 D등급 1위인 강동구 둔촌1동과 인접 동 성내1동을 실측 화면과 대조해 검증했다(별첨 「4. SGIS 자료 활용·검증 기록」).

기법은 분석 전에 미리 정해 둔 XGBoost를 다른 트리 계열 5종·선형회귀와 시군구 GroupKFold(5)로 비교한 뒤 채택했다(모형 비교 전체는 별첨).

검증은 두 층이다. GroupKFold로 모형을 고르고, 튜닝에 한 번도 쓰지 않은 대전·경기·충남을 최종 홀드아웃으로 뒀다. 100m 값을 행정동으로 재합산해 SGIS 행정동 인구와 맞춰 보는 것이 최종 판정이다.

□ 추진내용

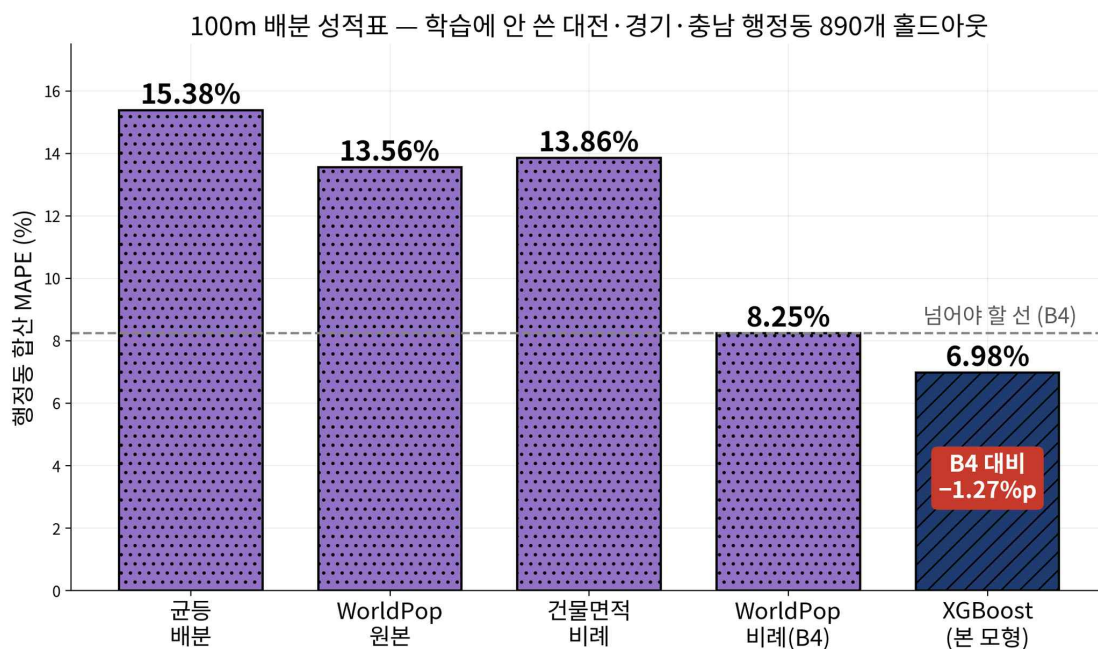
베이스라인 4종 성적

1km 인구를 100m로 나누는 방식 넷을 실제로 채점했다. 균등배분(B1)·WorldPop 원본(B2)·건물면적 비례(B3)·WorldPop 밀도 비례(B4)

를, 인구가 0이라 오차율을 계산할 수 없는 6곳을 뺀 전국 행정동 기준으로 비교했다.

방식(전국 행정동 기준)	MAPE	R ²
B1 균등배분	28.8%	0.9426
B2 WorldPop 원본	14.7%	0.9249
B3 건물면적 비례	36.4%	0.9239
B4 WorldPop 비례	9.83%	0.9786

베이스라인 중 최선인 B4는 균등배분보다 19.0%p 낮다. B2·B4는 같은 WorldPop 값에 SGIS 1km 구속 여부만 다르므로, 14.7%→9.83%의 차이가 곧 SGIS 1km 통계의 기여다.



자료: 국가데이터처 SGIS 격자·행정구역 통계(2024), WorldPop 100m(2024), OpenStreetMap(2026-08) · n=890개 행정동(대전·경기·충남, 모형 선택에 미사용)

〈그림 2〉 100m 배분 성적표 — 학습·튜닝에 쓰지 않은 대전·경기·충남 행정동 890개 홀드아웃(위 표의 전국 기준과 대상이 다름) (자료: 국가데이터처 SGIS 격자·행정구역 통계, WorldPop 100m, OpenStreetMap / 기준:2024년)

검증 결과 요약(한계 포함 상세는 별첨)

학습·튜닝에 한 번도 쓰지 않은 대전·경기·충남 행정동 890개에서 모형 MAPE는 6.98%로, SGIS 구속 베이스라인 B4(8.25%)보다 1.27%p 낮다. 개선이 가장 큰 곳은 도심 면적 최소 4분위(2.10%p)

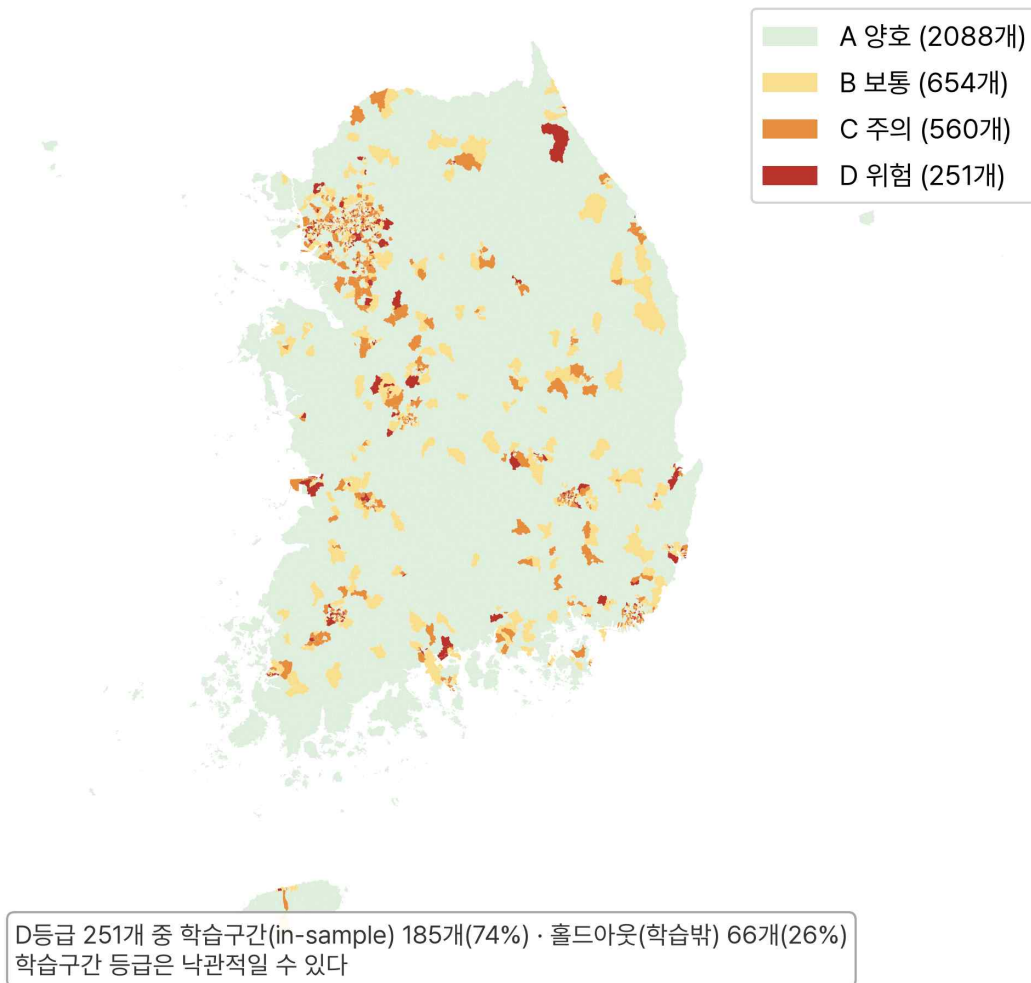
였다. 다만 개선폭은 홀드아웃 구성과 WorldPop 입력 여부에 따라 달라지고 일부 시도에서는 B4가 더 나왔다. 그래서 모형 결과를 일괄 적용하지 않고, 행정동마다 신뢰등급을 매겨 어디를 믿고 어디를 조심할지 함께 제공한다(홀드아웃 재추첨·시도별 결과는 별첨에 모두 공개). 아울러 100m 값을 1km로 다시 합치면 SGIS 원값과 일치(최대 오차 0.0064명)하도록 설계해, 원자료 총량을 훼손하지 않음을 자체 점검으로 확인했다.

□ 추진성과

산출물

100m 인구격자를 전국 단위로 만들었다. 인구 1명 이상인 칸은 1,546,957칸이고 전국 합계는 51,826,577명으로 SGIS 1km 원자료 합계와 사실상 같다. 행정동 3,553개 전수에 대한 신뢰등급 명단은 「1km 격자값을 그대로 써도 되는지」를 A~D 4단계로 판정한다. D등급(오차 20% 초과)은 251개다. 전체의 75%(2,663개)는 학습에 쓰인 구간이라 등급이 낙관적일 수 있어, 명단에 홀드아웃 여부 열을 함께 두어 이용자가 구분할 수 있게 했다. 이 명단으로 SGIS 1km 값을 그대로 쓸지 재계산할지 행정동 단위로 바로 판단할 수 있어 자료 활용의 신뢰성이 높아진다.

배분 신뢰등급 전국 지도 — 1km 격자값을 그대로 쓰면 안 되는 곳 (D등급 251개 동)



자료: 본 연구 산출 신뢰등급 명단(행정동 3,553개 전수, 모형 배분 기준) · 2024년 기준
전체의 75%(2,663개)는 학습구간이라 낙관적일 수 있음 · 홀드아웃 890개는 학습 밖에서 낸 값

〈그림 3〉 격자신뢰등급 전국지도 - 행정동 전수 판정 (자료: 본 연구 산출 · 검증 기준 대전·경기·충남 홀드아웃 / 기준: 2024년)

SGIS 화면 대조와 웹지도

이 명단을 SGIS 대화형 통계지도로 직접 대조했다. D등급 1위인 둔촌1동은 인접 동보다 뚜렷하게 낮은 값으로 나타나 D등급 판정과 일치했다. 둔촌주공 재건축으로 단기간에 인구가 크게 바뀐 지역이라는 정황과도 맞는다. 인접 동 성내1동은 실측 인구 18,416명과 모형 배분값 18,398명의 오차가 0.1%로 A등급 판정과 일치했다(별첨).

신뢰등급 명단을 동 이름으로 직접 찾는 웹지도 「동네돋보기」를 만들었다. 같은 자리를 1km와 100m로 나란히 보여주고 신뢰등급을 함께 띄우며, 전국 수치 대조와 100m→1km 되합산 자체검사를 통과해야 배포되도록 만들었다(별첨). 동 이름 검색만으로 오차 수준이 바로 확인돼 별도 분석 입력 없이도 활용할 수 있다. 정적 페이지로 구현을 마쳐 자체검사를 전부 통과한 상태다.

공개와 라이선스

OSM 건물·도로를 공변량으로 썼으므로 파생물에 ODbL Share-Alike 의무가 걸린다. SGIS 1km 원자료는 재배포하지 않고 우리가 만든 파생 격자만 공개해 그 의무를 지킨다(공개 범위 표는 별첨).

분석 코드와 파생 격자는 공개 저장소에 게시할 준비를 마쳤으며, 게시 이후 개발지원센터 이용사례 등록과 연구실·지자체 활용 문의를 이어서 진행한다.