기본연구보고서 2025-8

인공지능을 활용한 건축물 데이터 품질 고도화 방향 연구

Enhancing Building Data Quality Using Artificial Intelligence

안의순 Ahn, Euisoon
허한결 Heo, Hankyul
남기천 Nam, Kicheon
강범준 Kang, Bumjoon
이선재 Lee, Sunjae
박동준 Park, Dongjoon

기본연구보고서 2025-8

인공지능을 활용한 건축물 데이터 품질 고도화 방향 연구

Enhancing Building Data Quality Using Artificial Intelligence

지은이 안의순, 허한결, 남기천, 강범준, 이선재, 박동준
펴낸곳 건축공간연구원
출판등록 제2015-41호 (등록일 '08. 02. 18.)
인쇄 2025년 10월 26일, 발행: 2025년 10월 31일
주소 세종특별자치시 가름로 143, 8층
전화 044-417-9600
팩스 044-417-9608

http://www.auri.re.kr
가격: 25,000원, ISBN: 979-11-5659-516-8

연구진

연구책임	안의순 부연구위원
연구진	허한결 부연구위원 남기천 연구원
외부연구진	강범준 서울대학교 건축학과 교수 이선재 서울대학교 건설환경종합연구소 연구조교수 박동준 서울대학교 건축학과

연구심의위원	염철호 선임연구위원 오성훈 선임연구위원 조영진 선임연구위원 김영현 연구위원 이여경 연구위원 김태현 서울연구원 선임연구위원 노희용 한국기술교육대학교 산업경영학부 교수
연구자문위원	강준경 금오공과대학교 건축학부 교수 김문현 한국행정연구원 부연구위원 김민석 부경대학교 건축학과 교수 김수영 경북대학교 건축학부 교수 노병준 순천향대학교 AI·빅데이터학과 교수 문봉주 해안종합건축사사무소 소장 박성호 한국교육개발원 선임연구위원 박수현 이화여자대학교 융합전자반도체공학부 교수 박영섭 희림종합건축사사무소 수석 박종기 순천향대학교 건축학과 교수 박찬영 국립경국대학교 건축공학과 교수 변나향 충북대학교 건축학부 교수 서영선 정보통신정책연구원 부연구위원 서유진 광운대학교 건축학과 교수 손동화 충북대학교 건축학과 교수

연구요약

서론

건축물은 국가 정책의 기반이 되는 핵심 공간단위로서, 건축물 현황을 정확히 파악하는 것은 핵심 과제이다. 그럼에도 정부에서 생산·관리하고 있는 건축물과 관련한 데이터는 그 품질과 상호 연계성에 있어 제약이 있다. 건축물대장으로 대표되는 건축행정 데이터는 건축물의 현황 정보를 담고 있는 공적장부로서 건축물 관련 범정부 데이터 연계의 핵심이나, 선행연구 등에서 오류, 누락, 불일치 등 품질 문제가 지속적으로 제기되고 있는 현실이다.

건축물 현황을 정확하게 파악하기 위해서는 건축물에 대한 전수 현장조사가 유일한 방법이며, 데이터 자체의 논리적 오류도 관리 주체가 전수 검토를 할 필요가 있다. 그러나 이러한 방법은 매우 큰 비용이 발생하여 실행가능성이 낮다. 이에 따라 본 연구에서는 사람의 개입을 최소화하면서 데이터 전수에 대한 검증을 수행하기 위한 방법으로 인공지능 방법론을 활용하여 건축물 데이터의 오류를 검증하고, 누락, 이상값, 불일치 등 데이터 품질 문제를 도출하는 과정을 통한 데이터 연계 및 품질확보 방안을 마련하고자 하였다. 또한 건축물 데이터의 개방과 연계를 고도화하기 위한 기술적, 제도적 기반 마련을 위하여, 먼저 세움터의 건축행정 데이터에서 도입된 건축물 단위 고유식별자인 건물ID를 바탕으로 건축물 데이터 연계 방안을 모색하였다.

따라서 본 연구에서는 인공지능을 활용한 건축물 데이터의 품질 개선방향을 모색하고, 이에 기반하여 건축물 데이터 품질 고도화 적용 방안을 제안하였다. 품질 제고 방법론을 적용한 건축물 데이터를 유통하기 위한 방안을 마련하고, 건축 인허가 단계에서 신규 데이터의 품질 제고를 위한 방안도 함께 제안하고자 하였다.

건축물 데이터 품질 고도화 방향

■ 건축물 데이터 관련 제도 분석

「건축법」은 건축허가 업무 등을 전자정보처리 시스템으로 처리하도록 하고(제32조), 건축물의 유지와 관리를 위하여 허가권자가 건축물대장에 건축물 현황 정보를 보관하고 정비하도록 한다(제38조).

이에 따라 건축행정시스템 세움터가 운영되고 있다. 건축물대장 기재내용 변경의 경우 소유자 신청에 의하거나 지자체장이 직권으로 변경할 수 있는데, 이 때는 건축물대장 기재내용의 오류 또는 누락사항을 정정 또는 기재한 경우 지체 없이 건축물의 소유자에게 통지하여야 한다.

「공공데이터의 제공 및 이용 활성화에 관한 법률」(약칭 공공데이터법)은 공공기관이 보유·관리하는 데이터(공공데이터)를 국민에게 제공하도록 규정하며, 건축행정시스템(세움터)에서 관리하는 건축행정 전산자료도 건축허브를 통하여 가공하지 않은 형태 그대로 제공하고 있다. 이렇게 제공되고 있는 건축물 데이터는 연속성과 일관성이 보장되지는 않는다. 특히 2024년까지 건축데이터 민간개방시스템을 통하여 제공되던 건축행정 데이터가 2025년 초부터 건축허브로 이관되면서 고유키(PK)가 변경되어 과년도 데이터와 연계 활용이 어려운 상황이다.

■ 건축물 데이터 품질 및 활용 현황 분석

「건축법」 제38조는 건축물대장에 대한 지속적 정비 의무를 규정하고 있으며, 「공공데이터법」 제22조는 공공데이터의 품질관리를 규정하고 있다. 이에 따라 국토교통부 및 지자체는 건축물대장을 지속적으로 정비하고 있다. 2022년 세움터에서는 건축물대장, 건축인허가, 주택인허가 데이터에 대하여 정비 대상 항목을 86개 업무규칙으로 제시하고, 오류 현황을 검토하였다. 전체 점검 대상 건수 6억 2천만 건 중 약 855만 건에서 오류가 나타나 전체 오류율은 1.37%이었으나, 특정 업무규칙에서는 10~30%대의 높은 오류율이 나타났다. 건축물대장에서는 (순번 13) 대지면적보다 건축면적이 큰 경우가 25.02%로 가장 높게 나타났으며, 이어 (순번 16) 표제부 건축면적 합계의 불일치도 16.78%으로 높게 나타났다.

건축물 데이터는 건축통계 작성, 건축물 관련 연구 등에 폭넓게 활용되고 있다. 현재 건축통계는 건축행정시스템에 기반한 자동 집계를 통한 보고통계로 작성되고 있다. 건축물 데이터 중 건축물통계에서는 시군구코드, 연면적, 소유구분, 용도코드, 층수 등이 활용되며, 건축허가·착공·준공 통계에서는 시군구코드, 용도코드, 구조코드, 연면적, 허가구분, 동수, 허가일, 착공일, 사용승인일 등이 활용된다.

다음으로 건축물 데이터 활용 선행연구를 고찰하여 주요 활용 데이터와 주요 오류 사례를 검토하였다. 주요 오류 사례는 주 용도 코드에 정상 코드가 아닌 값이 기재된 경우, 동 명칭, 주 건축물 수 등이 누락된 경우, 연면적 등 값이 지나치게 큰 이상값인 경우 등이 나타났다.

■ 건축물 데이터 특성을 고려한 품질 고도화 방향

건축물 데이터 품질 및 활용 현황을 고려하여, 면적 및 용도 데이터의 품질과 데이터 연계 품질의 고도화 방향을 제시하였다. 면적 데이터는 수치 데이터로, 건축통계와 각종 연구 분석에서 핵심적으로 활용되는 변수이다. 특히 건축물대장 품질 점검에서 가장 높은 오류율이 확인되어 우선적으로 검토하고자 하였다. 건물동 단위 면적 데이터에 관한 2022년 정비규칙 4개(대지면적, 건축면적, 건폐율, 용적률)를 기존 규칙에 대한 검증대상으로 설정하고, 지역별, 준공년도별 경향을 분석하였다. 또한 기존 건축물대장 정비에서 다루지 않은 신규 검증규칙을 개발하였다. 다음으로 규칙 기반 진단의 한계를

넘어 오류 가능성을 탐색하기 위하여 기계학습 기반 이상값 탐지를 병행하고자 하였다. Isolation Forest와 One-Class SVM 등 대표적인 기계학습 기반 이상탐지 알고리즘을 적용하고, 규칙 기반 진단과 동일한 방법론을 적용하여 이상값 발생 경향을 검토하였다.

용도 데이터도 건축통계와 연구에서 활용도가 높은 항목으로 나타났다. 건축물대장에 표출되지 않는 용도 분류 코드가 실제 통계작성 등에 활용되고 있어, 기재내용과 분류 코드 간의 일관성을 확보할 수 있는 방안을 제시하고자 하였다. 건축물대장 기재내용에 해당하는 자유 형식 텍스트인 '기타 용도' 데이터와 분류코드인 '주 용도 코드'를 입력값으로 하여, 기계학습 알고리즘인 나이브 베이즈 모델을 적용하여 정합성을 검증하였다.

마지막으로 건물ID에 기반하여 인허가 및 건축물대장 데이터를 연계, 그 품질 현황을 분석하고 품질 고도화 방안을 도출하고자 하였다. 건축물대장과 건축인허가, 주택인허가 데이터의 건물ID 부여율을 검토하고, 건물ID 기반으로 데이터를 연계하여 연계 성공률, 기재 데이터 일관성 등을 검토하고자 하였다.

건축물 데이터 품질 고도화 시범적용

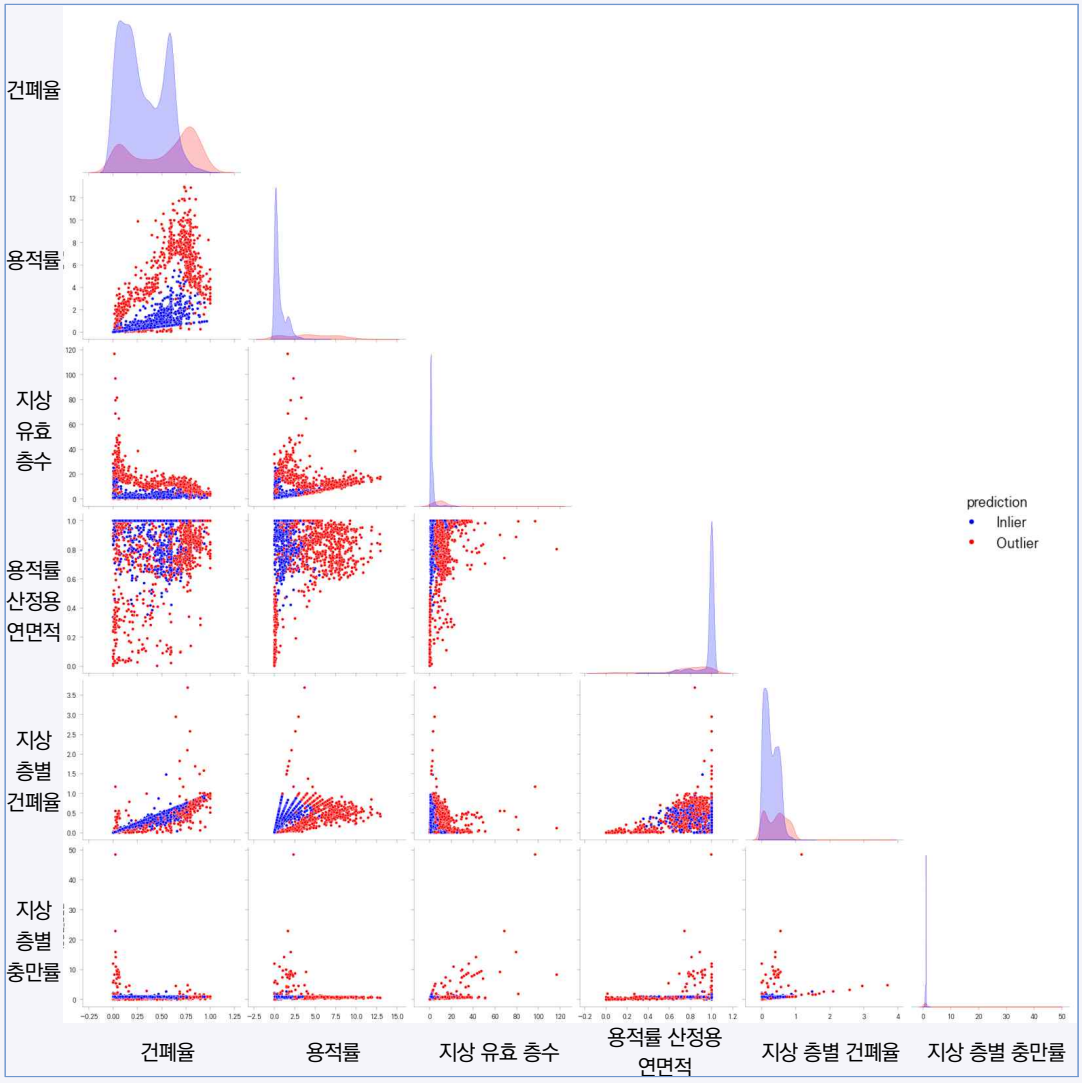
■ 건축물대장 면적 데이터 품질 고도화

건축물대장 면적 데이터의 오류 및 이상 현황을 기존 업무규칙 기반 검증, 신규 검증규칙 개발, 그리고 기계학습을 통한 이상값 탐지를 종합하여 다각적으로 검토하였다. 기존 업무규칙 기반 검증 결과 0.13%~5.01%의 오류율이 나타났으며, 건축면적이 연면적보다 큰 경우가 가장 흔한 오류로 나타났다. 또한 신규 검증규칙으로 용적률 산정 연면적이 (일반) 연면적보다 큰 경우(오류율 0.26%), 연면적이 본 연구에서 도출한 상한을 초과하는 경우(0.86%) 등을 개발하여 검증하였다.

기계학습 기반 이상값 탐지를 위하여 먼저 건축물 규모에 영향받지 않는 무차원 지표로 대지면적 대비 건축면적 비율, 연면적 대비 용적률 산정용 연면적 비율, 층별 충만률 등을 도출하였다. 이 데이터에 대하여 Isolation Forest와 One-Class SVM 등 방법론을 상호 보완적으로 활용할 수 있는 가능성을 검증하였다. 이상값의 시도별·연도별 발생 경향을 분석한 결과 시기별로 일정한 오류율을 보이고 서울, 부산의 오류율이 높게 나타나는 등 규칙 기반 오류와 다른 패턴을 보여, 기존 규칙 기반 검증으로 포착하기 어려운 오류를 탐지하였음을 시사하였다.

■ 건축물대장 용도 데이터 품질 고도화

건축물대장 용도 데이터를 대상으로 체계적인 품질 진단과 오류 탐지를 수행하였다. 먼저 건축물대장 용도 데이터의 전반적 현황 파악을 위하여 용도별 단어 출현 빈도 및 분포를 분석하였다. 그 과정에서 상위 단어가 주 용도 코드와 일치하지 않는 경우를 발견하였고, 복합용도 건축물의 경우 다양한 용도가 혼합 기재되는 양상도 관찰되었다.



기계학습 기반 이상값 산점도
출처: 연구진 작성

다음으로 나이브 베이즈 분류 알고리즘을 적용하여 기재내용에서 분류 코드를 예측하였다. 동별 용도 예측은 84.78%, 층별 예측은 78.97% 정확도를 달성하였다. 다만 세부 예측 검토를 통해 건축물 용도의 불균형으로 인한 오분류를 확인하였다. 또한 여러 용도가 복합된 경우 주 용도를 제대로 분류하지 못함을 보였다. 층별개요의 용도 기재내용은 해당 층의 주 용도에 대한 내용을 포함하고 있지 않은 경우가 있다. 따라서 건축물 모든 층의 용도를 고려한 검증이 필요함을 확인하였다.

■ 건축물대장-인허가 연계 품질 고도화

건축물대장과 인허가 데이터의 연계 품질 고도화에는 건물ID 기반 건축물 데이터 연계 데이터를 구축하고, 건축물대장과 건축인허가, 주택인허가 데이터의 교차검증을 통하여 연계 가능성을 검토하였다. 건축물대장 기준 2024년 사용승인 건축물 55,784동을 기준으로 건물ID 부여율은 97.0%, 건축인허가는 99.4%, 주택인허가는 99.4% 부여로 나타났다. 대다수의 경우 건축물과 건물ID가 일대일로 대

응되었으나, 건축인허가 데이터의 경우 여러 건축물에 동일한 건물ID가 부여된 사례가 상당수 확인되었다. 대지 내 여러 건축물의 경우, 동일 건축물에 대한 중복 레코드 존재 등이 원인으로 판단된다.

건물ID 기반 연계 성공률은 건축인허가의 경우 93.6%, 주택인허가는 34.9%로 나타났다. 연계 성공한 데이터의 정합성을 검토한 결과, 면적 변수의 경우 건축인허가는 면적 변수 90% 이상 일치하였으나, 주택인허가의 일치율은 60%대 후반~70%대 초반으로 나타났다. 용도 변수의 경우 건축인허가와 주택인허가 모두 90%대 후반 일치율을 보였다.

■ 건축물 데이터 오류 특성에 기반한 맞춤형 정비 방안 필요

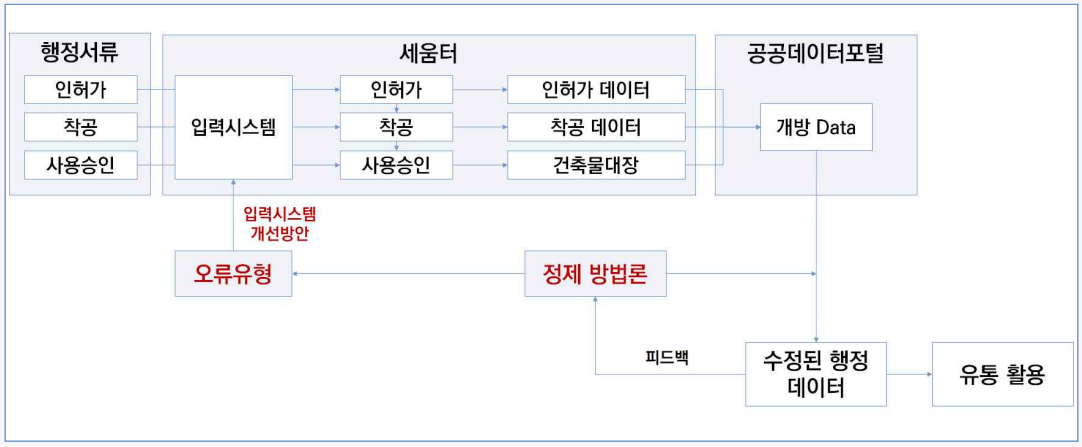
건축물 데이터 품질 고도화 방법론 시범 적용 결과 건축물대장 데이터 오류는 시기적 및 지역적 행정체계와 제도적 요인에 기인한 구조적 특성을 지님을 확인하였다. 데이터 유형별 고려 외에도 실제 데이터 오류 발생 맥락을 고려하여야 한다. 정기적·획일적 정비 방식은 한계가 있다. 오류 발생 맥락을 반영한 맞춤형 정비 체계로 전환하여야 한다. 기존 규칙 검증으로는 오류 포착이 어렵다. 인공지능 기반 이상탐지를 병행하는 다층적 품질검증 체계가 필요하다. 건물ID 중복 불일치 사례는 연계 품질 관리가 중요함을 부각한다. 건물ID의 고유성 정합성 확보를 통한 범용 연계키 활용 기반을 강화하여야 한다.

건축물 데이터 품질 제고를 위한 방안 제안

■ 품질 제고 건축물 데이터 유통

본 연구는 방법론 적용을 통하여 품질이 제고된 건축물 데이터 유통 방안을 제안하였다. 공적장부 기반 건축행정 데이터는 정부 및 지자체에서 직권 정정하기 어려운 현실이다. 품질 개선 데이터를 기존 원본 데이터에 반영 정정하는 대신 품질 개선 데이터를 세움터 외부에 저장하고, 오픈소스로 유통하는 방식을 제안하였다.

다음으로 본 연구에서 도출된 고도화 방법론 자체를 오픈소스로 유통하는 방안을 제안하였다. 오류 탐지 규칙 활용 방식, 인공지능 기반 이상치 도출 방식, 건물ID 기반 교차검증 세 가지 방법론이 있다. 이 방법론들에 포함된 개별 규칙 프로세스를 문서 및 코드 형식으로 공개한다. 이는 재현성을 확보하고 구체적인 방법론 활용을 촉진할 수 있다. 공개 방식은 단계적 추진을 통하여 개선사항을 반영한다. 오픈소스 기반 유통에는 공개뿐만 아니라 이용자들의 피드백을 보고받아 반영하고 개선해나가는 과정이 포함된다. 최종적으로 공공데이터 포털 등을 통하여 배포하는 것을 목표로 한다. 건축물 데이터 품질 개선 알고리즘의 유통 운영 지속가능성을 위해 운영 주체를 명확히 하여야 한다. 버전 관리를 통하여 지속적인 품질 제고가 이루어질 수 있다.



건축물 데이터 품질 제고 방안
출처: 연구진 작성

■ 신규 데이터 품질 제고

신축 건축물에 대한 신규 데이터의 품질 제고를 위하여, 인허가 단계에서 동별 및 층별개요 입력 시스템에서 건축물 용도 및 면적 입력 내용을 검증하는 방안을 제시하였다. 면적 항목은 숫자를 직접 입력하는 구조로 되어있어 오류 발생 가능성이 높을 것으로 판단된다. 구조, 용도 등 항목은 검색을 통하여 정해진 목록 안에서 고르는 구조로 오타로 인한 오류는 없으나, 텍스트로 입력하는 기재내용(기타 구조, 기타 용도 등)에서 오류 가능성이 있다.

면적 항목의 경우, 면적 간 관계에 기반한 자동 입력을 도입하거나, 면적 합계 등 입력 내용이 계산된 값과 비교하여 불일치하는 경우 경고창을 띄우는 방식을 도입할 수 있다. 용도 등 항목에서 분류 내용과 기타 기재내용 사이에 오타나 불일치 등 오류가 발생하는 경우 마찬가지로 경고하고 사용자가 수정할 수 있도록 할 필요가 있다.

사용승인 단계에서는 인허가 단계와 동일한 동별 및 층별개요 입력 검증과 함께, 건물ID를 적극적으로 활용하여 세움터 데이터 검증 기능에서 연계키로 사용하는 방안도 제안하였다. 사용승인 단계 인허가 데이터와 건축물대장 생성 시 데이터가 일치하지 않는 경우, 입력자에게 경고하여 다시 한번 확인할 수 있는 기회를 제공하고, 건축허가 현황을 존중하는 범위 안에서 데이터 무결성을 높이는 방안을 제안하였다.

주제어
건축물대장, 기계학습, 건물ID, 공공데이터, 오픈소스

차례

표 차례

그림 차례

List of Figures

제1장

서론

1. 연구의 배경 및 목적
2. 연구의 범위 및 방법
3. 연구의 차별성

1. 연구의 배경 및 목적

1) 연구의 배경 및 필요성

■ 건축물 관련 범정부 데이터의 연계 및 품질확보 방안 마련 시급

건축물은 부동산, 지역, 산업 등 국가 정책의 기반이 되는 핵심 공간단위로서, 건축물 현황을 정확히 파악하는 것은 국가 행정 전반을 좌우하는 핵심 과제이다. 그럼에도 정부에서 생산·관리하고 있는 건축물과 관련한 데이터는 다양한 데이터가 여러 주체에 의해 관리되고, 데이터의 품질과 상호 연계성에 있어 제약이 있는 것이 현실이다.

건축물대장으로 대표되는 건축행정 데이터는 건축물의 현황 정보를 담고 있는 공적장부로서 건축물과 관련한 범정부 데이터 연계의 핵심이 되는 중요한 데이터이다. 그러나 선행연구를 통하여 건축행정 데이터의 오류, 누락, 불일치 등 품질 문제가 지속적으로 제기되고 있다. 또한, 다른 건축물 관련 공공데이터와 비교할 때 전국 건축물 현황부터 크게 차이가 나는 등, 데이터 연계를 위한 기반 마련이 미흡한 것이 현실이다. 국가통계지도의 경우, 건축물대장 기준 전국 건축물 수는 7,951,324동이나, 도로명주소대장의 '건물등' 수는 전국 10,752,596동이며, 약 70%에서만 건물정보가 일치한다(국토지리정보원, 2024; 2024년 1월 기준).

건축물 현황을 정확하게 파악하기 위해서는 개별 건축물별로 현장조사를 수행하는 것이 유일한 방법이며, 데이터 자체의 논리적 오류를 제거하기 위해서도 데이터 관리 주체가 모든 내용을 빠짐없이 검토하는 것이 가장 확실한 방법일 것이다. 그러나 전국 모든 건축물에 대하여 주기적으로 현장조사를 수행하거나 다양한 주체에 의해 생산된 범정부 데이터를 모두 확인하는 것은 것은 매우 큰 비용이 들어 실행가능성이 낮다. 사람의 개입을 최소화하면서 데이터 전수에 대한 검증을 수행하기 위하여 인공지능 방법론을 활용하여 건축물 데이터의 오류를 검증하고, 누락, 이상값, 불일치 등 데이터 품질 문제를 도출하는 과정을 통한 데이터 연계 및 품질확보 방안 마련이 요구된다. 이러한 과정을 통하여 문제가 있는 건축물 데이터의 범위를 확정하고, 합리적인 수준의 비용으로 현장조사 등을 통한 건축물 데이터 품질 고도화를 달성할 수 있을 것이다.

[표 1-1] 건축물 데이터 오류 관련 선행연구 현황

선행연구	건축물 데이터 오류 검토 결과
안의순·박종훈(2023), 재해 취약 지하층 주택 현황 분석	- '주_용도_코드'의 영문자 Z와 숫자 4자리로 코드체계 변경 전 코드를 표현하거나, "창고", "사찰", "축사", "학교" 등 전혀 다른 값이 포함
조영진 외(2022), 건축물 화재 발생 예측 모델	- 화재 데이터와 건축물대장 데이터 연계과정에서 주소정보의 불일치 사례 발견 및 연계 프로세스 방안 도출 - '동_명칭' 컬럼 오류 검토
조영진 외(2023), 건축물 화재 및 홍수 리스크 분석 모델 개발	- 데이터 연계과정에서 주소정보의 불일치를 해결하기 위해 조영진 외(2022, p.63-64)의 데이터 연계 프로세싱 방안을 준용 - 총괄표제부 '주_건축물_수', 표제부 '연면적', '건축_면적', '대지_면적', '건폐율', '용적률', '용적률_산정_연면적', '높이', '지상_층_수', '지하_층_수', '주_용도_코드', '구조_코드' 등 오류 검토
조영진 외(2024), 범죄예방 환경설계 고도화 및 인증제도 개선	- 표제부 '경과연수', '높이', '건폐율', '용적율' 컬럼 오류 및 이상치 검토 - 이상치 처리 방안은 z-score를 사용하여 이상치를 탐지하였으며, z-score가 ±3 시그마 범위를 벗어난 데이터를 제거
조영진 외(2024), 빈 건축물 추정	- 필지 단위의 전기에너지 사용량 정보를 건축물 동 단위로 배분하는 과정에서 '연면적' 컬럼에 정보가 없는 경우 미확정 정보로 분류
송유미 외(2024), 건축물 연령지표 개발	- 건축물대장 표제부 테이블의 '사용승인_일' 컬럼을 추출하여 검토한 결과, 건축행정 전산화 이후 데이터 누락은 발생되지 않았으며, 문화재 또는 고(古)건축물과 같이 과거에 축조되거나 건축행정 이전 시기에 지어진 건축물의 경우 사용승인일이 기록되지 않은 사례가 일부 확인

출처: 연구진 작성

[국가통계지도 미매칭 현황]

[반지하 현황 전수 현장조사]

[그림 1-1] 건축물 데이터의 품질 확보 필요성

출처: (왼쪽) 국토지리정보원. (2024). 국토정보플랫폼 국토정보맵.
https://map.ngii.go.kr/ms/map/NlipMap.do?tabGb=statsMap (검색일: 2024.6.5.)
(오른쪽) 김보미. (2023). 5200여 반지하 주택 '전수조사' 성동구…장마 전 예방 위해 3개월 만에 끝내. 경향신문. 2월 17일 기사. https://www.khan.co.kr/national/national-general/article/202302171820001 (검색일: 2023.3.2.)

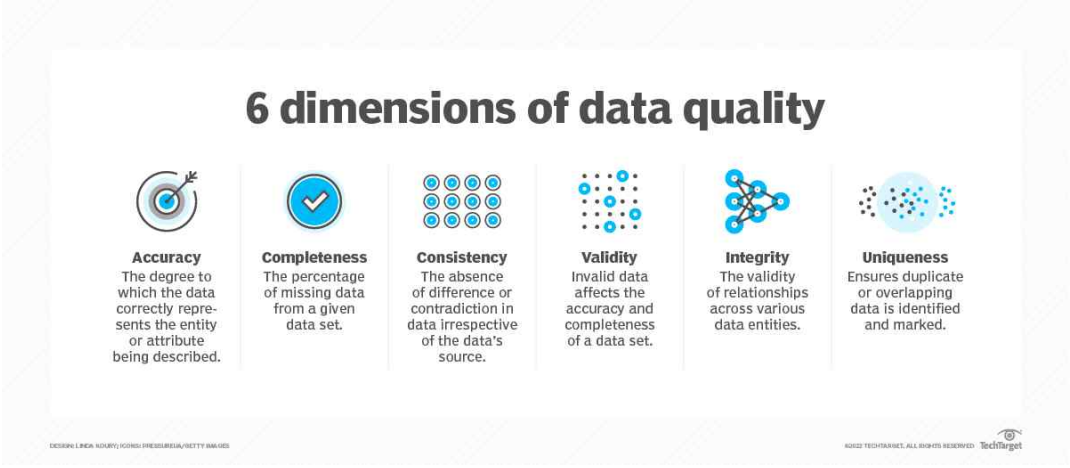
■ 건축물 데이터 연계 및 품질 확보를 위한 기반 마련 방안 필요

건축물 데이터의 연계와 품질 확보를 위해서는 건축물 데이터의 개방과 연계를 고도화하기 위한 기술적, 제도적 기반 마련도 필요하다. 인공지능을 활용하여 건축물 데이터 품질을 확보하는 방향을 검토할 필요가 있다. 통계, 기계학습, 딥러닝을 포함한 인공지능 방법론을 활용하여 건축물 데이터의 오류,

누락, 불일치 등 품질 문제를 찾고, 최종적으로는 허가권자의 현장조사 및 현황 파악이 이루어질 수 있도록 충분한 근거를 제공할 수 있다.

또한 최근 세움터의 건축행정 데이터에서 도입된 건축물 단위 고유식별자인 건물ID는 건축물 데이터를 연계하는 공통식별자로서 높은 활용 가치를 지닐 것으로 기대된다. 현재는 건축물대장과 건축 및 주택 인허가대장의 연계에 활용되고 있는데, 이를 바탕으로 연계 데이터를 활용한 정부 서비스 제공 등 디지털 혁신 방안을 모색할 필요가 있다.

마지막으로 건축물 데이터의 연계, 품질 확보 등을 지속적으로 수행하기 위한 방안을 마련할 필요가 있다. 특히, 건축물 관련 공공데이터의 민간 활용을 위해서는 개방 데이터의 품질 고도화가 중요하며, 행정 목적으로 작성된 건축행정 데이터와 독립적으로 활용을 위한 데이터를 관리하고 품질 고도화를 적용한 데이터를 개방 및 유통하는 방안을 검토할 필요가 있다.



[그림 1-2] 데이터 품질의 차원

출처: Craig, L. (2023). How data quality shapes machine learning and AI outcomes. Enterprise AI. https://www.techtarget.com/searchenterpriseai/feature/How-data-quality-shapes-machine-learning-and-AI-outcomes (검색일: 2025.2.14.)

2) 연구목적

본 연구의 목적은 먼저 인공지능을 활용한 건축물 데이터의 품질 개선 방향을 모색하는 것이다. 이를 위하여 먼저 건축물 데이터의 품질 현황과 오류 양상을 검토하여 현재 지니는 한계를 파악한다. 이후 인공지능 기반 오류 검출 기법을 적용함으로써 데이터 품질을 개선할 구체적인 방향을 제안한다.

다음 목적은 건축물 데이터 품질 고도화의 적용 방안을 마련하는 것이다. 제안된 방법론을 적용하여 품질이 제고된 건축물 데이터의 유통 체계를 구축할 방안을 모색하고, 건축 인허가 단계에서 신규로 생산되는 데이터의 품질을 향상시킬 방안을 아울러 제시한다.

2. 연구의 범위 및 방법

1) 연구범위

■ 건축물 데이터의 범위

본 연구의 대상은 넓게는 건축물 관련 공공 및 민간의 모든 정보를 포괄하지만, 접근성, 개방성, 활용도 등을 고려하여 그 범위를 건축행정 자료로 한정하였다. 특히, 새롭게 도입된 건물ID가 적용된 건축물대장과 건축인허가, 주택인허가 자료를 주요 연구 대상으로 선정하였다.

건축행정시스템 세움터의 데이터 구조에서 각 대장은 데이터베이스의 여러 테이블로 구성된다. 테이블은 다시 컬럼의 집합으로 구성된다. 본 연구에서는 대장-테이블-컬럼 위계를 기준으로 건축행정 데이터의 구조를 파악하고 데이터 품질 고도화 방법론을 도출하고자 하였다.

■ 데이터 품질 고도화의 범위

본 연구에서는 활용도가 높은 건축행정 데이터를 중심으로 데이터 품질 고도화 방법론을 도출하고, 이를 일부 데이터에 시범적으로 적용하여 도출된 방법론을 검증하고자 하였다. 특히 개방 공공데이터의 품질 제고와 정확한 통계 추정치 산출의 두 가지 측면을 고려하여 데이터 품질 고도화 시범적용 대상 범위를 설정하고자 하였다. 구체적으로는 건축행정 데이터의 활용 사례를 검토하고, 여러 선행 사례에서 활용된 컬럼을 활용도 높은 컬럼으로 정의하여 건축물대장의 면적 및 용도 데이터를 품질 고도화 시범적용 대상으로 설정하였다. 건물ID의 경우 대장 간 교차검증을 위한 연계키로 건물ID 자체의 품질을 검증하면서 건물ID 기반 건축물대장과 인허가 데이터의 연계 품질을 검토하였다.

[표 1-2] 건축물대장 주요 테이블 및 컬럼 (예시)

테이블	컬럼
건축물대장 기본개요	'시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지'
건축물대장 표제부	'연면적', '건축_면적', '대지_면적', '건폐율', '용적률', '용적률_산정_연면적', '높이', '지상_층_수', '지하_층_수', '주_용도_코드', '구조_코드'

출처: 조영진 외(2023), 건축물 화재 및 홍수 리스크 분석 모델 개발. 바탕으로 연구진 작성

■ 연구의 내용적 범위

건축물 데이터 품질은 정부 정책 결정과 민간 활용에 있어서 매우 중요한 요소이다. 인공지능을 활용한 전수 검증을 통하여 여러 차례 반복된 대장 정비를 통해서도 정정되지 않고 있는 오류를 찾아낼 수 있는 방법론을 도출하고자 하였다. 또한 이를 건축행정 데이터 생애주기의 각 단계에 적용하기 위한 방향을 제시하고자 하였다.

건축물 데이터의 오류는 행정 간 연계 문제, 시스템 간 연계 문제, 데이터 입력 실수 등 다양한 원인으로 발생할 수 있다. 본 연구에서는 건축물 데이터 검증을 위하여 규칙 기반 검증, 통계 기반 접근, 기계학습 등 인공지능 기법을 활용한 다양한 전략을 시범적으로 적용하고, 그 중 효과가 확인된 전략을 최종 정책 방향에 포함시켜 제시하고자 하였다. 또한, 정책화 방향에 있어서 기축 건축물과 신축 건축물 대상, 인허가대장과 건축물대장을 분리하여 접근함으로써, 건축행정과 데이터의 생애주기를 고려한 세분화된 품질 고도화 전략을 도출하고자 하였다.

2) 연구방법

- 문헌조사 및 법·제도 분석
  - 건축물 데이터 구조, 특성, 품질 현황 관련 문헌 고찰
  - 건축물 데이터 활용 선행연구 검토
- 논리적 검증
  - 건축물의 특성을 반영한 건축물 데이터 내 논리적 오류(음수, 100% 초과 등) 검출
  - 건축물 관련 데이터 연계·교차검증 통한 논리적 오류 검출
- 통계·기계학습
  - 정형 데이터(예: 범주형, 수치형)의 오류(통계적 이상치 등) 검출
  - 텍스트 데이터(예: 수기 입력)의 오류(오기, 오분류 등) 검출
- 전문가 자문
  - 건축 분야 전문가 자문을 통한 건축물 데이터 관리방안 마련
  - 통계, 빅데이터, 기계학습, 인공지능 분야 전문가 자문을 통한 연구 수행

3. 연구의 차별성

본 연구는 인공지능을 활용하여 건축물 데이터의 품질을 고도화하는 방향을 제시하는 데 초점을 둔다. 기존 연구들이 건축물 데이터의 오류 및 품질 문제를 다루었지만, 본 연구는 통계, 기계학습 등 인공지능 기술을 활용하여 데이터 오류를 검출하고 품질을 개선하는 구체적인 방법론을 제시한다는 점에서 차별성이 있다.

연구는 문헌연구와 사례조사를 바탕으로 통계 및 기계학습을 활용한 정형 데이터 오류 검출과 자연어 데이터 오류 검출 방법을 분석한다. 또한, 현재 건축물 데이터의 현황을 검토한 후, 건축물 데이터의 특성에 맞는 인공지능을 적용한 오류 검출 및 품질 확보 방안을 도출하고, 이를 실제 건축물 데이터 품질 고도화에 적용할 수 있는 방향을 제시한다. 특히, 인공지능을 활용하여 데이터의 오류를 자동으로 식별하고 보완할 수 있는 체계를 마련함으로써 기존의 수작업 검증 방식과 달리, 데이터 품질 관리의 효율성과 투명성을높이는 데 기여한다는 점에 연구의 차별성이 있다.

[표 1-3] 선행연구 현황

구분	연구과제명	연구목적	연구방법	주요연구내용
건축행정정보 및 통계	건축행정정보의 정책적 활용 및 건축통계 개선방안 연구 (조상규, 성은영, 2012)	건축행정정보와 건축통계의 문제점 분석, 개선방안 도출	- 국내 건축행정정보 및 건축통계 현황 검토 - 해외 건축통계 구축 현황 사례 조사 (영국, 노르웨이, 핀란드, 미국, 일본)	- 단기 개선방안: 건축물 용도분류 개선, 지역단위 세분화, 변동 현황 통계추가 등 - 중장기 개선방안: 이력 추적 가능한 DB 구조 개선, 유관 데이터 연계프로세스 개선, 건축물 정기조사 도입 검토 등
	건축행정 통계 개선 및 공간정보 융합 방안 연구 (조영진 외, 2022)	건축물 행정통계 고도화, 건축물 통계정보와 타 통계를 연계 하는 융·복합 통계 생산 가능성 을 모색	- 문헌연구 - 사례조사 - 건축행정 데이터 분석 - 건축행정 통계, 융복합 통계 시범생산	- 기존 건축물 정보 구축현황 및 통계 제공 현황 분석 - 해외 건축통계 제공 현황 사례 분석 및 시사점 도출 - 건축행정통계의 고도화를 위 한 추가 통계표 제안 및 생산 - 공간 단위의 타 데이터와의 융 · 복합 신규 통계 방향모색 - 신규 건축통계 작성 및 활용을 위한 제도적 장치 마련

구분	연구과제명	연구목적	연구방법	주요연구내용
건축물 지리정보 연계	건축물 정위치 등록에 관한 연구 (정동훈 외, 2014)	지적도 상에 건축물을 정위치 등록하기 위한 법·제도적 개선방안 제시	– 문헌연구 – 사례조사 – 지적, 건축 분야 공무원, 교수, 건축사 등 전문가 인터뷰	– 지적전산파일을 이용한 건축설계 현황 분석 – 지적현황측량을 통한 건물위치 등록현황 파악 – 건축행정업무 개선을 통한 건축물 정위치 등록방안 – 건축물 정위치 등록정보 활용방안
통계·인공지능 활용	기계학습(Machine Learning) 기반 이상 탐지 (Anomaly Detection) 기법 연구 – 보건사회 분야를 중심으로. (오미애 외, 2018)	이상탐지 연구의 체계적이고 포괄적인 개요, 관련 이슈 분석, 보건복지 분야에 활용성을 높일 수 있는 방안 모색	– 문헌연구 – 사례조사 – 데이터 탐색적 분석 – 기계학습 기반 이상탐지 방법론	– 이상 탐지 개념 정의 – 국내·외 사례 연구 분석 – 기계학습, 딥러닝 기반 이상 탐지 기법 조사 – 치매 조기 진단을 위한 이미지 자료(FDG-PET) 활용성 탐색적 분석 – 노인 학대 노출에 대한 이상(anomaly) 재정의와 특성 탐색적 분석 – 이상 탐지 기법 관련 이슈, 정책 제언 도출
	과학기술 행정 혁신을 위한 인공지능 활용 방안 (양현채 외, 2020)	과학기술 행정에서 인공지능 활용 현황 파악, 활용 가능 영역 도출 과학기술 행정 혁신을 위한 인공지능의 활용 방안 모색	– 문헌연구 – 사례조사 – 시스템 구축, 과학기술 정책 등 전문가 인터뷰	– 공공부문과 과학기술 행정에서 인공지능 활용 필요성 타진 – 인공지능 기술의 활용 대상인 과학기술 행정의 개념과 범위 구체화 – 과학기술 행정의 인공지능 활용 사례 유형화 – 향후 예상 장애요인 도출 – 인공지능의 활용 방향성 제시
	대규모 언어모델(LLM)을 활용한 건축민원 대응 효율화 방안 연구 (조상규, 김신성, 2023)	인공지능 기술을 활용한 건축 법규 해석과 관련된 민원 처리 자동화 프로세스 제안	– 법·제도 검토 – LLM 활용 사례 분석 – 프로토타입 시스템 개발 – LLM 기반 자동 평가 시스템 개발	– 대규모 언어모델 특성 검토 – 건축 관련 민원 특성 파악 – 건축 법규해석 질의 처리 프로세스 구축 – 건축법 관련 업무 자동화 시스템 구축 방안 제시

4. 연구흐름도

[그림 1-3] 연구흐름도
출처: 연구진 작성

제2장

건축물 데이터 품질 고도화 방향

1. 건축물 데이터 관련 제도 분석
2. 건축물 데이터 품질 및 활용 현황 분석
3. 건축물 데이터 특성을 고려한 품질 고도화 방향

1. 건축물 데이터 관련 제도 분석

1) 건축법과 건축물 데이터

■ 건축법 개요

「건축법」은 건축물의 대지, 구조, 설비 및 용도에 관한 기준을 규정하여 공공복리를 증진하는 것을 그 목적으로 하는 법률이다(제1조). 건축행위는 사유재산인 대지 안의 행위이다. 그러나 「건축법」은 이를 개인의 재산권 행사에 국한하지 않는다. 공공의 안전과 계획적 도시 개발 등 공적 차원에서 규율한다는 점에서 의의가 있다. 「건축법」의 내용은 '건축' 행위에 대한 내용과 건축물의 특성에 대한 내용으로 나뉜다. 건축물의 '건축'은 신축, 증축, 개축, 재축, 이전과 대수선, 용도변경, 리모델링 등 건축물을 변화시키는 행위를 다룬다.

■ 건축행정 전산자료

- 건축행정 전산화 근거

「건축법」은 "건축허가 업무 등의 효율적인 처리를 위하여"(제32조 제1항), 전자정보처리 시스템을 이용하여 업무를 처리하도록 한다. 건축허가 업무의 전산처리는 「건축법」 제32조 제1항은 "허가권자는 … 이 법에 규정된 업무를 처리할 수 있다"고 규정하고 있으나, 같은 항에서 세부 규정을 위임받은 같은 법 시행규칙에서는 "허가권자는 … 불가피한 경우를 제외하고는 전자정보시스템을 이용하여 건축허가 등의 업무를 처리하여야 한다"(제22조의2)고 규정하고 있어, 실질적으로 의무규정이며, 현재 건축허가 업무는 건축행정시스템 세움터를 통하여 처리되고 있다.

건축행정 전산화를 위하여 허가권자는 「건축법」 제31조에 따라 건축행정 관련 신청서, 신고서, 첨부서류, 통지, 보고 등을 전산적으로 제출하게 할 수 있다. 건축행정에 해당하는 건축법 조항은 「건축법」 제10조(건축 관련 입지와 규모의 사전결정), 제11조(건축허가), 제14조(건축신고), 제16조(허가와 신고사항의 변경), 제19조(용도변경), 제19조의2(복수 용도의 인정), 제20조(가설건축물), 제21조(착공신고 등), 제22조(건축물의 사용승인) 등 건축행위를 포괄한다.

제31조(건축행정 전산화)

① 국토교통부장관은 이 법에 따른 건축행정 관련 업무를 전산처리하기 위하여 종합적인 계획을 수립 · 시행할 수 있다. 〈개정 2013. 3. 23.〉
② 허가권자는 제10조, 제11조, 제14조, 제16조, 제19조부터 제22조까지, 제25조, 제29조, 제30조, 제38조, 제83조 및 제92조에 따른 신청서, 신고서, 첨부서류, 통지, 보고 등을 디스켓, 디스크 또는 정보통신망 등으로 제출하게 할 수 있다. 〈개정 2019. 4. 30.〉

제32조(건축허가 업무 등의 전산처리 등)

① 허가권자는 건축허가 업무 등의 효율적인 처리를 위하여 국토교통부령으로 정하는 바에 따라 전자정보처리 시스템을 이용하여 이 법에 규정된 업무를 처리할 수 있다. 〈개정 2013. 3. 23.〉

출처: 「건축법」. 법률 제20424호, 2024. 3. 26., 일부개정

제22조의2(전자정보처리시스템의 이용)

① 법 제32조제1항에 따라 허가권자는 정보통신망 이용환경의 미비, 전산장애 등 불가피한 경우를 제외하고는 전자정보시스템을 이용하여 건축허가 등의 업무를 처리하여야 한다.
② 제1항에 따른 전자정보처리시스템의 구축, 운영 및 관리에 관한 세부적인 사항은 국토교통부장관이 정한다. 〈개정 2013. 3. 23.〉

출처: 「건축법 시행규칙」. 국토교통부령 제1416호, 2024. 12. 16., 일부개정

- 건축행정시스템과 건축행정 전산자료

「건축법」제32조와 「건축법 시행규칙」제22조의2, 「건축행정시스템 운영규정」제5조에 따라 건축행정업무를 수행할 수 있는 건축행정시스템 세움터를 구축 및 운영하고 있다. 「건축행정시스템 운영규정」제20조에 따르면 건축행정시스템을 이용하여 신고·신청할 수 있는 전자민원의 종류는 총 49종이다. 또한, 건축행정시스템 세움터에 기재된 민원서비스 종류는 총 150종으로, 건축인허가 관련 민원 34종, 주택인허가 관련 민원 33종, 건축물대장 관련 민원 19종, 정비사업 관련 민원 31종, 사업자 관련 민원 25종, 건축위원회심의 관련 민원 5종, 녹색건축 관련 민원 3종으로 구성되어있다[1]. 즉, 해당 민원 중 「건축법」제32조에 따른 "건축허가 업무 등"에 해당하는 경우 전산자료의 형태로 건축행정시스템에 저장되며 건축행정 전산자료가 된다. 대상업무 중 건축업무, 주택업무, 건축물대장업무에 속하는 상세업무의 전산처리 과정에서 발생한 정보가 본 연구에서 다루고자하는 건축물 데이터의 범위에 해당할 것이다.

제5조 (구축 및 운영 원칙)

① 운영기관의 장은 건축행정시스템을 구축하는 경우 운영자지침서의 내용에 따라 적합하게 구축하여야 한다.
② 운영기관의 장은 건축행정시스템의 안정적인 운영을 위하여 시스템 관리자를 지정하고 유지보수 대책을 수립하여야 한다.
③ 운영기관의 장은 표준프로그램을 변형시킬 수 없으며, 표준프로그램과 연계한 개별프로그램을 개발 · 운영하고자 하는 경우에는 국토교통부장관에게 관련 내용을 사전에 보고하여야 한다.

제20조(전자민원종류)

① 건축행정시스템을 이용하여 신고 · 신청할 수 있는 민원의 종류는 별표 3에 열거한 민원으로 한다.
② 업무담당자 등이 건축행정시스템을 이용하여 전자 발급할 수 있는 증명민원의 종류는 별표 4와 같다.
③ 운영기관의 장은 처리단계를 공개할 필요가 있는 민원에 대하여는 그 처리 진행상황을 해당 민원인에게 공개할 수 있다.

출처: 「건축행정시스템 운영규정」. 국토교통부훈령 제1369호, 2021. 2. 18., 일부개정

1) 세움터. https://www.eais.go.kr/moct/awp/ada08/AWPADA08L02. 2024.02.26. 접속

[표 2-1] 건축행정시스템 세움터 대상업무

대상업무	상세업무
건축업무	건축허가, 착공, 사용승인, 건축신고, 공작물, 가설건축물, 위반건축물
주택업무	주택건설사업계획승인, 착공, 사용검사, 행위허가
건축물대장업무	대장작성, 기재사항변경, 열람, 발급, 말소
정비사업	행정계획, 조합, 사업시행인가, 관리처분, 착공, 준공인가, 정비사업전문관리업
건축관련업자	건축사, 건축사사무소, 임대사업자, 주택건설사업자, 주택관리사
건축정보집계	건축허가현황, 건축착공현황, 건축물현황, 주택건설실적

출처: 건축행정시스템 세움터. https://www.eais.go.kr/moct/awp/agd01/AWPAGD01V01. 2025.02.06. 접속

[표 2-2] 건축행정시스템 이용 신고 신청 대상민원

민원사무명	민원사무명	민원사무명
[임시]사용승인신청서	건축물대장합병신청서	행위허가신청서
가설건축물존치기간연장신고서	건축물부존재증명발급신청서	건축사사무소[·업무신고사항변경·휴업·폐업]신고서
가설건축물축조신고서	건축물소유자 ·변경·정정신청서	건축사업무신고서
건축·대수선·용도변경신고서	건축물지번 ·변경·정정신청서	건축사업무신고필증재교부신청서
건축·대수선·용도변경허가신청서	건축물표시 ·변경·정정신청서	임대사업자등록사항변경신고서
건축관계자변경신고서	집합건축물대장 전유부 변경[분할·합병]신청서	임대사업자등록신청서
건축물철거·멸실신고서	관리사무소장배치등변경신고서	임대조건변경신고서
공작물축조신고서	관리사무소장배치등신고서	임대조건신고서
도로폐지·변경신청서	사업계획[변경]승인신청서	임대주택매각계획서
사전결정신청서	사용검사[임시사용승인]신청서	임대주택분양전환신고서
착공신고서	사용검사신청서	임대주택분양전환허가신청서
착공연기신청서	임대주택조합[설립·변경·해산]인가신청서	주택관리사[보]자격증재교부신청서
건축물대장말소신청서	주택조합[설립·변경·해산]인가신청서	주택관리사자격증교부신청서
건축물대장생성신청서	직장주택조합설립신고서	주택관리업등록사항변경신고서
건축물대장의분리·결합신청서	착공신고서	주택관리업등록신청서
건축물대장재작성신청서	착공연기신청서	
건축물대장전환신청서	행위신고서	

출처: 「건축행정시스템 운영규정」 별표 3. 국토교통부훈령 제1369호, 2021. 2. 18., 일부개정

■ 건축물대장

「건축법」은 건축물의 유지와 관리를 위하여 허가권자가 건축물대장에 건축물과 그 대지의 현황, 구조내력에 대한 정보를 보관하고 정비하도록 한다(제38조). 건축물대장 생성 대상은 「건축물대장의 기재 및 관리 등에 관한 규칙」제12조에서 규정하고 있다. 이에 따르면 크게 세 가지로 구분할 수 있다. 첫째, 사용승인을 하는 경우로, 「건축법」제22조 제2항에 따른 사용승인서를 내준 경우이다. 둘째, 사용승인이 없는 경우이다. 건축물의 건축·대수선·용도변경과 가설건축물의 건축 및 공작물의 축조 등이 이에 해당한다. 셋째로 「주한미군기지 이전에 따른 평택시 등의 지원 등에 관한 특별법」제8조에 따라 반환되는 공여구역의 건축물이 이에 해당한다. 이 세 가지 모두에 해당하지 않는 건축물의 경우 공사를 완료한 후 「건축물대장의 기재 및 관리 등에 관한 규칙」제12조 제2항에 따라 건축물대장 생성을 신청하여야 하며 이 경우 건축물대장이 생성된다.

제12조(건축물대장의 생성)

① 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 다음 각 호의 구분에 따라 건축물대장을 생성하여야 한다. 다만, 법 제20조에 따른 가설건축물은 제외한다. 〈개정 2011. 9. 16., 2017. 1. 20.〉

1. 법 제22조제2항에 따라 사용승인(다른 법령에 따라 사용승인으로 의제되는 준공검사 · 준공인가 등을 포함한다. 이하 같다)을 하는 경우: 사용승인된 내용에 따라 생성
2. 법 제29조에 따른 공용건축물의 공사완료를 통보받은 경우: 영 제22조제3항 및 「건축법 시행규칙」 제22조제2항에 따라 제출된 서류에 따라 생성
3. 「주한미군기지 이전에 따른 평택시 등의 지원 등에 관한 특별법」 제8조에 따라 반환되는 공여구역의 건축물에 대하여 국방부장관의 요청이 있는 경우: 해당 건축물에 대한 국유재산대장부본 및 건물배치도에 따라 생성

②제1항 외의 건축물의 공사를 완료한 자는 별지 제10호서식의 건축물대장 생성 · 재작성 신청서에 다음 각 호의 서류를 첨부하여 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 신청하여야 한다. 〈개정 2009. 1. 20., 2017. 1. 20., 2018. 12. 4.〉

1. 대지의 범위와 그 대지의 사용에 관한 권리를 증명하는 서류
2. 건축물현황도
3. 현황측량성과도(경계복원측량도로 갈음할 수 있다)

③제2항에 따라 건축물대장생성 신청을 받은 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 신청내용이 건축물 및 대지의 실제 현황과 합치되고 건축법령이 정한 건축기준 및 관계 법령 등의 규정에 적합한 건축물에 대하여 건축물대장을 생성하여야 한다. 〈개정 2009. 1. 20., 2017. 1. 20.〉

④특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 제2항에 따른 건축물대장생성 신청이 없는 경우에는 건축주 또는 소유자에게 건축물대장생성의 신청을 권고하거나 제3항의 기준에 따라 직권으로 해당 건축물에 대한 건축물대장을 생성할 수 있다. 직권으로 건축물대장을 생성하는 경우에 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 해당 건축물의 건축주 또는 소유자에게 그 사실을 통지하여야 한다. 〈개정 2009. 1. 20., 2017. 1. 20.〉

⑤제4항 후단에 따른 통지를 하려는 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장이 통지받는 자의 주소 또는 거소를 알 수 없는 때에는 건축물대장생성 사실을 해당 특별자치시 · 특별자치도 또는 시 · 군 · 구의 게시판에 게시하거나 특별자치시 · 특별자치도 또는 시 · 군 · 구의 공보나 일간신문에 게재함으로써 건축주 또는 소유자에게 통지된 것으로 본다. 〈개정 2009. 1. 20., 2017. 1. 20.〉

출처: 「건축물대장의 기재 및 관리 등에 관한 규칙」. 국토교통부령 제1235호, 2023. 8. 1., 일부개정

또한, 「건축물대장의 기재 및 관리 등에 관한 규칙」제29조에 따라 건축물대장정보가 전산화된 경우에도 전산화된 자료를 건축물대장으로 보도록 규정하고 있다.

제29조(전자정보처리시스템에 의한 건축물대장 사무처리 등)

①법 제32조제1항에 따른 전자정보처리시스템(이하 "전자정보처리시스템"이라 한다)을 이용하여 건축허가업무 등의 사무를 처리하여 보조기억장치(자기디스크, 자기테이프, 그 밖에 이와 유사한 방법에 의하여 일정한 건축물정보사항을 확실하게 기록 · 보관할 수 있는 전자적 정보저장매체를 포함한다. 이하 같다)에 건축물대장정보를 전자적 형태로 기록한 경우 그 전산기록을 건축물대장으로 본다. 〈개정 2009. 1. 20.〉

출처: 「건축물대장의 기재 및 관리 등에 관한 규칙」. 국토교통부령 제1235호, 2023. 8. 1., 일부개정

「건축법」제38조에 따르면 「건축법」제22조 제2항에 따른 사용승인서를 내준 경우, 「건축법」제11조에 따른 건축허가 대상 건축물(제14조에 따른 신고 대상 건축물을 포함) 외의 건축물의 공사를 끝낸 후 기재를 요청한 경우, 그 밖에 대통령령(「건축법 시행령」제25조)으로 정하는 경우 건축물대장에 건축물과 대지의 현황 및 구조내력 관련 정보를 적어 보관하고 지속적으로 정비하도록 하고 있다. 또한, 건축물대장의 서식, 기재 내용, 기재 절차 등에 대해서는 「건축물대장의 기재 및 관리 등에 관한 규칙」을 통해 규정하고 있다.

제38조(건축물대장)

① 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 건축물의 소유 · 이용 및 유지 · 관리 상태를 확인하거나 건축정책의 기초 자료로 활용하기 위하여 다음 각 호의 어느 하나에 해당하면 건축물대장에 건축물과 그 대지의 현황 및 국토교통부령으로 정하는 건축물의 구조내력(構造耐力)에 관한 정보를 적어서 보관하고 이를 지속적으로 정비하여야 한다. 〈개정 2012. 1. 17., 2014. 1. 14., 2015. 1. 6., 2017. 10. 24.〉

1. 제22조제2항에 따라 사용승인서를 내준 경우
2. 제11조에 따른 건축허가 대상 건축물(제14조에 따른 신고 대상 건축물을 포함한다) 외의 건축물의 공사를 끝낸 후 기재를 요청한 경우
3. 삭제 〈2019. 4. 30.〉
4. 그 밖에 대통령령으로 정하는 경우

② 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 건축물대장의 작성 · 보관 및 정비를 위하여 필요한 자료나 정보의 제공을 중앙행정기관의 장 또는 지방자치단체의 장에게 요청할 수 있다. 이 경우 자료나 정보의 제공을 요청받은 기관의 장은 특별한 사유가 없으면 그 요청에 따라야 한다. 〈신설 2017. 10. 24.〉

③ 제1항 및 제2항에 따른 건축물대장의 서식, 기재 내용, 기재 절차, 그 밖에 필요한 사항은 국토교통부령으로 정한다. 〈개정 2013. 3. 23., 2017. 10. 24.〉

출처: 「건축법」. 법률 제20424호, 2024. 3. 26., 일부개정

- 건축물대장의 구조

「건축물대장의 기재 및 관리 등에 관한 규칙」제5조에 건축물대장의 작성방법을 규정하고 있다. 건축물대장은 일반건축물대장, 집합건축물대장의 표제부, 집합건축물대장의 전유부, 건축물대장의 총괄표제부, 다가구주택의 호(가구)별 면적대장으로 구분할 수 있으며 「건축물대장의 기재 및 관리 등에 관한 규칙」제7조에 각 대장의 기재사항이 정리된 서식을 규정하고 있다. 또한, 「건축물대장의 기재 및 관리 등에 관한 규칙」제5조 제1항의 별표에서 각 서식의 작성항목과 작성요령에 대해 설명하고 있다.

[표 2-3] 건축물대장 작성항목

항목명	항목명	항목명
일반건축물대장		
대지위치	주용도	사용승인일
지번	층수	건축물 인증 현황
명칭	높이	내진설계 적용 여부
호수/가구수/세대수	지붕	내진능력
도로명 주소	부속건축물	특수구조 건축물
대지면적	조경면적, 공개 공지 · 공간 면적	지하수위
건축면적	건축선 후퇴면적, 건축선 후퇴거리	기초형식
건폐율	건축물 현황	구조설계 해석법
연면적	소유자 현황	관리계획 수립 여부
용적률 산정용 연면적	건축주 · 설계자 · 공사감리자 · 공사시공자(현장관리인)	건축물관리점검 현황
용적률	주차장 · 승강기 · 하수처리시설 · 급수설비(저수조)	변동사항 (변동일, 변동내용 및 원인)
지역, 지구, 구역	허가일	그 밖의 기재사항
주구조	착공일	
집합건축물대장		
건축물 현황	전유부분 · 공용부분	그 밖의 항목
호명칭	공동주택(아파트) 가격	
건축물대장 총괄표제부		
연면적 · 용적률 산정용연면적	총 호수/가구수/세대수	특이사항
건축물 수	총 주차대수	그 밖의 항목
주용도	건축물 현황	
다가구주택의 호(가구)별 면적대장		
호(가구)별 전용면적	그 밖의 항목	

출처: 「건축물대장의 기재 및 관리 등에 관한 규칙」 별표. 국토교통부령 제1235호, 2023. 8. 1., 일부개정

- 건축물대장의 변경

건축물대장이 변경되는 경우는 크게 두 가지로 구분할 수 있다. 건축물의 소유자가 변경 신청하여 변경되는 경우와 건축물 소유자의 신청이 없음에도 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장이 변경하는 경우이다.

건축물 소유자가 변경신청하여 변경되는 경우는 '건축물이 있는 대지의 변동에 따른 건축물대장 변경', '건축물대장의 재작성', '건축물대장의 합병', '집합건축물대장의 전유부의 변경', '건축물대장의 표시사항 변경', '건축물대장의 소유자 변경', '건축물대장의 지번 변경', '건축물대장의 도로명주소 변경', '건축물대장 기재내용 정정', '건축물의 해체·멸실 등에 따른 건축물대장의 말소'로 구분 가능하며 각각 「건축물대장의 기재 및 관리 등에 관한 규칙」제13조부터 제22조에 해당한다. 건축물의 소

유자는 모든 경우 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 신청함으로써 건축물대장을 변경할 수 있다.

건축물 소유자의 신청이 없음에도 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장이 변경하는 경우는 그 성격과 소유자에 대한 통지 여부에 따라 크게 두 가지로 구분 가능하다. 첫째는 건축주 또는 건축물의 소유자가 건축물대장의 변경을 신청한 것은 아니나 다른 종류의 신청에 따라 건축물대장도 변경되는 경우이다. 건축물의 건축공사가 완료되어 사용승인을 받을 때 그 내용에 따라 건축물대장의 내용도 허가권자의 직권으로 변경하는 경우(「건축물대장의 기재 및 관리 등에 관한 규칙」제18조 제1항 단서)와 등기관서로부터 소유권 변동이 통지되어 건축물대장 기재사항을 정리하는 경우(같은 규칙 제19조 제1항) 등이 이에 해당한다.

둘째는 건축물대장의 내용이 건축물 소유자의 신청 없이 허가권자의 직권으로 변경되고, 그 내용을 건축물의 소유자에게 통지하여야 하는 경우이다.「건축물대장의 기재 및 관리 등에 관한 규칙」제20조, 제20조의2, 제21조, 제22조에 해당하는 다음 네 가지 경우가 있다. "건축물 소유자의 신청에 의하는 경우 외에 지번이 변경되었음을 증명하는 서류(토지이동정리결의서와 토지 · 임야 분할 · 합병신청서 사본 등을 말한다)가 첨부된 지적공부 소관청의 지적정리통지에 의하는 경우", "건축물 소유자가 신청하는 경우 외에 도로명주소개별대장이 첨부된 도로명주소 소관청의 도로명주소정리통지에 따르는 경우", "건축물대장 기초자료 등을 통해 건축물대장의 기재내용에 잘못이 있거나 기재내용이 누락되어 있음을 발견한 경우에는 그 사실을 확인한 경우", "건축물이 해체 · 멸실되었음에도 소유자나 관리자가 건축물대장의 말소 신청을 하지 아니하거나「건축물관리법」제30조에 따라 해체허가를 받지 않거나 해체신고를 하지 않은 경우 또는 같은 법 제34조에 따라 멸실신고를 하지 않은 경우"이다.

또한, 건축물대장의 기재 내용 중 소유권과 같이 등기사항에 해당하는 부분이 변경되는 경우에는「건축법」제39조에 따라 관할 등기소에 그 등기를 촉탁하여야 한다. 등기촉탁의 절차에 관련한 사항은「건축물대장의 기재 및 관리 등에 관한 규칙」제26조에 규정하고 있다.

제13조(건축물이 있는 대지의 변동에 따른 건축물대장 변경)

①건축물의 소유자는 건축물이 있는 대지의 분할이나 합병에 따라 건축물대장을 나누거나 합치려는 경우 또는 대지가「공간정보의 구축 및 관리 등에 관한 법률」에 따라 이미 분할이나 합병된 경우에는 별지 제11호서식의 건축물대장의 분리 · 결합신청서에 다음 각 호의 서류를 첨부하여 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 신청하여야 한다. 〈개정 2009. 1. 20., 2009. 12. 14., 2012. 11. 16., 2015. 6. 4., 2017. 1. 20., 2018. 12. 4.〉

1. 건축물현황도(건축물현황도의 내용이 변경된 경우에 한한다)
2. 현황측량성과도(「공간정보의 구축 및 관리 등에 관한 법률」 제23조제1항에 따라 지적측량을 실시하여야 하는 경우에 한정하며, 경계복원측량도로 갈음할 수 있다)
3. 토지 등기사항증명서(등기필증의 제시로 갈음할 수 있다)

②특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은「건축법」,「국토의 계획 및 이용에 관한 법률」,「공간정보의 구축 및 관리 등에 관한 법률」,「집합건물의 소유 및 관리에 관한 법률」 등 관계 법령에 적합한 때에만 제1항에 따라 건축물대장을 나누거나 합칠 수 있다. 〈개정 2009. 1. 20., 2009. 12. 14., 2015. 6. 4., 2017. 1. 20.〉

③특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 제1항의 신청에 의하여 건축물대장을 나누거나 합친 때에는 기존 건축물대장을 폐쇄하여야 한다. 다만, 분할이나 합병된 대지에 있는 건축물에 대한 건축물대장 중 하나는 기존 건축물대장의 기재내용을 변경하는 방법으로 작성할 수 있으며, 이 경우에는 기재내용이 변경된 건축물대장을 폐쇄하지 아니한다. 〈개정 2009. 1. 20., 2017. 1. 20.〉

제14조(건축물대장의 재작성)

①건축물의 소유자는 제12조제1항의 사용승인 신청서류 또는 동조제2항의 건축물대장생성 신청서류와 다르게 일반건축물에 대하여 집합건축물대장이 작성되었거나 집합건축물에 대하여 일반건축물대장이 작성되었음을 발견한 때에는 별지 제10호서식의 건축물대장 생성 · 재작성 신청서에 건축물대장이 잘못 작성되었음을 증명하는 서류를 첨부하여 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 건축물대장의 재작성을 신청하여야 한다. 〈개정 2009. 1. 20., 2017. 1. 20., 2018. 12. 4.〉

②특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 제1항의 신청에 의하여 건축물대장이 잘못 작성된 것으로 확인되면 건축물대장을 재작성하고 기존 건축물대장을 폐쇄하여야 한다. 〈개정 2009. 1. 20., 2017. 1. 20.〉

제15조(건축물대장의 전환)

①건축물의 소유자는 건축물대장의 전환을 하려는 경우에는 별지 제12호서식의 건축물대장전환신청서에 다음 각 호의 서류를 첨부하여 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 해당 건축물대장의 전환을 신청하여야 한다. 〈개정 2009. 1. 20., 2012. 11. 16., 2017. 1. 20.〉

1. 전환하려는 건축물의 건축물현황도(건축물현황도의 내용이 변경된 경우에 한한다)
2. 전환하려는 건축물의 등기사항증명서(등기필증의 제시로 갈음할 수 있다)
3. 해당 건축물에 거주하는 임차인에게 그 건축물의 용도변경으로 인하여 동 번호 및 호수 등이 변경된다는 사실을 통지하였음을 증명하는 서류

②특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 건축물대장을 전환하는 경우에는 「집합건물의 소유 및 관리에 관한 법률」 등 관계 법령에 적합한지를 검토하여야 한다. 〈개정 2009. 1. 20., 2017. 1. 20.〉

③특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 제1항의 신청에 의하여 건축물대장을 전환한 때에는 기존 건축물대장을 폐쇄하여야 한다. 〈개정 2009. 1. 20., 2017. 1. 20.〉

제16조(건축물대장의 합병)

①건축물의 소유자는 건축물대장의 합병을 하려는 경우에는 별지 제13호서식의 건축물대장 합병신청서에 다음 각 호의 서류를 첨부하여 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 해당 건축물대장의 합병을 신청하여야 한다. 〈개정 2009. 1. 20., 2012. 11. 16., 2017. 1. 20.〉

1. 합병하려는 건축물의 건축물현황도(건축물현황도의 내용이 변경된 경우에 한한다)
2. 합병하려는 건축물의 등기사항증명서(등기필증의 제시로 갈음할 수 있다)

②제1항에 불구하고 「부동산등기법」 제42조제1항에 따라 합병의 등기를 할 수 없는 건축물의 건축물대장은 합병할 수 없다. 〈개정 2012. 11. 16.〉

③제15조제2항은 제1항에 따른 건축물대장의 합병에 관하여 준용한다

④특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 제1항의 신청에 의하여 건축물대장의 합병을 한 때에는 기존 건축물대장을 폐쇄하여야 한다. 〈개정 2009. 1. 20., 2017. 1. 20.〉

제17조(집합건축물대장의 전유부의 변경)

①건축물의 소유자는 집합건축물의 전유부분을 두 개 이상으로 분할하거나 두 개 이상의 전유부분을 합병하려는 경우에는 별지 제14호서식의 집합건축물대장 전유부변경(분할 · 합병)신청서에 다음 각 호의 서류를 첨부하여 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 신청하여야 한다. 〈개정 2009. 1. 20., 2012. 11. 16., 2017. 1. 20.〉

1. 건축물현황도 중 해당 층의 평면도 및 단위세대평면도
2. 건물 등기사항증명서(등기필증의 제시로 갈음할 수 있다)

②특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 신청내용을 확인한 후 집합건축물대장의 전유부를 변경하는 경우에는 건축물대장 중 전유부의 해당 부분을 폐쇄하고 변경된 내용에 따라 새로이 작성하는 방법에 의하여야 한다. 다만, 분할하거나 합병하는 전유부분의 호 명칭이 기존의 호 명칭과 동일한 경우에는 건축물대장의 기재내용을 변경하는 방법에 따를 수 있다. 〈개정 2009. 1. 20., 2017. 1. 20.〉

③제15조제2항 · 제16조제2항은 제1항에 따른 집합건축물대장의 전유부 변경에 관하여 준용한다.

제18조(건축물대장의 표시사항 변경)

①건축물의 소유자는 건축물대장의 기재내용 중 건축물 표시사항을 변경(지번의 변경은 제20조에 따르고, 도로명주소의 변경은 제

20조의2에 따른다)하려는 때에는 별지 제15호서식의 건축물표시 변경신청서에 다음 각 호의 서류를 첨부하여 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 신청하여야 한다. 다만, 법 제22조제2항에 따라 사용승인된 경우에는 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장이 직권으로 사용승인서에 따라 변경한다. 〈개정 2009. 1. 20., 2011. 9. 16., 2017. 1. 20.〉

1. 건축물현황도(건축물현황도의 내용이 변경된 경우에 한한다)
2. 건축물의 표시에 관한 사항이 변경되었음을 증명하는 서류

②특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 제1항에 따른 건축물표시 변경신청에 의하여 건축물의 표시에 관한 사항을 변경하려는 때에는 신청내용이 건축물 및 대지의 실제현황과 합치되는지 여부를 대조 · 확인하여야 한다. 〈개정 2009. 1. 20., 2017. 1. 20.〉

제19조(건축물대장의 소유자 변경)

①특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 등기관서로부터 소유권 변동자료가 통지된 때에는 건축물대장의 소유자에 관한 사항을 정리하여야 한다. 〈개정 2009. 1. 20., 2017. 1. 20.〉

②건축물의 소유자는 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 별지 제16호서식의 건축물소유자 변경신청서에 등기사항증명서를 첨부(등기필증 제시를 포함한다)하여 건축물대장의 변경을 신청할 수 있다. 〈개정 2009. 1. 20., 2012. 11. 16., 2017. 1. 20.〉

③건축물의 소유자는 건축물대장이 건축법령에 적합하게 생성되었으나 「부동산등기법」 제29조제2호에 해당하여 소유권보존등기를 할 수 없는 건축물의 경우에는 제2항에도 불구하고 별지 제16호서식의 건축물소유자 변경 · 정정 신청서에 소유자가 변경되었음을 증명하는 다음 각 호의 서류를 첨부하여 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 그 건축물과 관련된 행정행위를 위한 기재내용의 변경을 신청할 수 있다. 〈개정 2009. 1. 20., 2017. 1. 20., 2018. 12. 4.〉

1. 「부동산등기법」에 따라 등기가 되지 않는 것을 증명하는 서류(특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장이 「부동산등기규칙」 제166조에 따른 대법원예규로 확인할 수 있는 경우에는 그 확인으로 갈음할 수 있다)
2. 건축물의 소유권에 관한 사항이 변경되었음을 증명하는 서류
3. 「공증인법」 제50조에 따른 증서의 등본 또는 같은 법 제57조에 따른 인증된 사서증서의 등본(「부동산 거래신고 등에 관한 법률」에 따른 부동산거래의 신고를 하고, 매매대금이 완납된 경우에는 같은 법에 따른 신고필증으로 갈음할 수 있다)

제20조(건축물대장의 지번 변경)

①건축물의 소유자는 건축물대장의 기재내용 중 지번에 관한 사항을 변경하려는 때에는 별지 제17호서식의 건축물지번 변경신청서에 다음 각 호의 서류를 첨부하여 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 신청하여야 하며, 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 이를 확인한 후 그 지번을 변경하여야 한다. 〈개정 2009. 1. 20., 2012. 11. 16., 2015. 6. 4., 2017. 1. 20.〉

1. 토지대장 또는 임야대장
2. 현황측량성과도(「공간정보의 구축 및 관리 등에 관한 법률」 제23조제1항에 따라 지적측량을 실시하는 경우에 한정하며, 경계복원측량도로 갈음할 수 있다)
3. 대지 소유자의 동의서(변경되는 대지의 소유권자와 건축물의 소유권자가 다른 경우에 한한다)

②특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 건축물 소유자의 신청에 의하는 경우 외에 지번이 변경되었음을 증명하는 서류(토지이동정리결의서와 토지 · 임야 분할 · 합병신청서 사본 등을 말한다)가 첨부된 지적공부 소관청의 지적정리통지에 의하여 건축물대장의 지번을 직권으로 변경할 수 있다. 〈개정 2009. 1. 20., 2017. 1. 20.〉

③ 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 제2항에 따라 건축물대장의 지번을 직권으로 변경한 경우에는 해당 건축물의 소유자에게 그 사실을 알려야 한다. 이 경우 제12조제5항을 준용한다. 〈신설 2009. 1. 20., 2017. 1. 20.〉

제20조의2(건축물대장의 도로명주소 변경)

① 건축물의 소유자는 건축물대장의 기재내용 중 도로명주소에 관한 사항을 변경하려는 때에는 별지 제17호의2서식의 건축물대장 도로명주소 변경신청서에 다음 각 호의 서류를 첨부하여 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 신청하여야 하며, 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 이를 확인한 후 그 도로명주소를 변경하여야 한다. 〈개정 2017. 1. 20.〉

1. 도로명주소개별대장
2. 건축물 소유자의 동의서(변경되는 건축물의 소유자와 거주자가 다르면서 거주자가 신청하는 경우에 한정한다)

② 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 제1항에 따라 건축물 소유자가 신청하는 경우 외에 도로명주소개별대장이 첨부된 도로명주소 소관청의 도로명주소정리통지에 따라 건축물대장의 도로명주소를 직권으로 변경할 수 있다. 〈개정 2017. 1. 20.〉
③ 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 제2항에 따라 건축물대장의 도로명주소를 직권으로 변경한 경우에는 해당 건축물의 소유자에게 그 사실을 알려야 한다. 이 경우 제12조제5항을 준용한다. 〈개정 2017. 1. 20.〉
[본조신설 2011. 9. 16.]

제21조(건축물대장 기초자료의 관리 및 건축물대장의 기재내용 정정)

① 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 건축물대장의 기재누락이나 소유권 불일치와 같은 오류사항 등을 조사하여 건축물대장 기초자료를 작성 · 관리할 수 있으며, 국토교통부장관은 이에 필요한 세부기준을 정할 수 있다. 〈신설 2009. 1. 20., 2013. 3. 23., 2017. 1. 20.〉
②특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 제1항의 건축물대장 기초자료 등을 통해 건축물대장의 기재내용에 잘못이 있거나 기재내용이 누락되어 있음을 발견한 경우에는 그 사실을 확인한 후 직권으로 이를 정정하거나 기재할 수 있다. 이 경우 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 지체 없이 그 내용을 건축물의 소유자에게 통지하여야 한다. 〈개정 2009. 1. 20., 2017. 1. 20.〉
③건축물의 소유자는 건축물대장의 기재내용에 잘못이 있음을 발견한 경우에는 별지 제15호서식의 건축물표시 정정신청서, 별지 제16호서식의 건축물소유자 정정신청서, 별지 제17호서식의 건축물지번 정정신청서 또는 별지 제17호의2서식의 건축물도로명주소 정정신청서에 다음 각 호의 서류를 첨부하여 그 잘못된 부분의 정정을 신청할 수 있다. 〈개정 2009. 1. 20., 2011. 9. 16., 2012. 11. 16., 2015. 6. 4.〉
1. 건축물대장의 표시사항을 정정하려는 경우에는 잘못이 있는 부분의 건축물현황도면과 이를 증명하는 서류
2. 건축물대장의 소유자에 관한 사항을 정정하려는 경우에는 건물 등기사항증명서(등기필증의 제시로 갈음할 수 있다)
3. 건축물대장의 지번에 관한 사항을 정정하려는 경우에는 토지대장 또는 임야대장. 이 경우 건축물의 대지위치에 관한 사항일 경우에는 현황측량성과도(「공간정보의 구축 및 관리 등에 관한 법률」 제23조제1항에 따라 지적측량을 실시하는 경우에 한정하며, 경계복원측량도로 갈음할 수 있다)를 포함한다.
4. 건축물대장에 도로명주소에 관한 사항을 정정하려는 경우에는 도로명주소개별대장
④특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 제3항에 따라 건축물의 기재내용을 정정하려는 때에는 신청내용이 건축물 및 대지의 실제 현황과 합치되는지 여부를 대조 · 확인하여야 한다. 〈개정 2009. 1. 20., 2017. 1. 20.〉
⑤제12조제5항은 제2항 후단의 통지에 관하여 준용한다. 〈개정 2009. 1. 20.〉
[제목개정 2009. 1. 20.]

제22조(건축물의 해체 · 멸실 등에 따른 건축물대장의 말소)

①건축물의 소유자나 관리자는 건축물의 전부 또는 일부가 해체 · 멸실 등으로 없어진 경우에는 별지 제18호서식의 건축물대장 말소신청서를 작성하여 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 건축물대장의 말소를 신청하여야 한다. 다만, 「건축물관리법」 제30조에 따라 해체허가를 받거나 해체신고를 한 경우 또는 같은 법 제34조에 따라 멸실신고를 한 경우에는 그러하지 아니하다. 〈개정 2009. 1. 20., 2011. 9. 16., 2012. 11. 16., 2017. 1. 20., 2020. 5. 1.〉
②제1항에 불구하고 법 제29조에 따른 공용건축물의 경우에는 해당 기관의 장이 해체 · 멸실 등으로 없어진 건축물의 개요, 해체 · 멸실 등의 사유 및 해체 · 멸실 등 전 · 후 사진(영 제22조제1항 단서에 따라 설계도서의 제출을 생략할 수 있는 건축물의 경우에는 해당 기관장의 확인서로 사진을 갈음할 수 있다)을 첨부하여 문서로 요청하여야 한다. 〈개정 2009. 1. 20., 2020. 5. 1.〉
③특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 건축물이 해체 · 멸실되었음에도 소유자나 관리자가 건축물대장의 말소 신청을 하지 아니하거나 「건축물관리법」 제30조에 따라 해체허가를 받지 않거나 해체신고를 하지 않은 경우 또는 같은 법 제34조에 따라 멸실신고를 하지 않은 경우에는 직권으로 해당 건축물대장을 말소할 수 있다. 이 경우 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 지체 없이 그 내용을 건축물의 소유자에게 통지하여야 한다. 〈개정 2009. 1. 20., 2011. 9. 16., 2017. 1. 20., 2020. 5. 1.〉
④제12조제5항은 제3항 후단의 통지에 관하여 준용한다.
[제목개정 2023. 8. 1.]

출처: 「건축물대장의 기재 및 관리 등에 관한 규칙」. 국토교통부령 제1235호, 2023. 8. 1., 일부개정

제39조(등기촉탁)

① 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 다음 각 호의 어느 하나에 해당하는 사유로 건축물대장의 기재 내용이 변경되는 경우(제2호의 경우 신규 등록은 제외한다) 관할 등기소에 그 등기를 촉탁하여야 한다. 이 경우 제1호와 제4호의 등기촉탁은 지방자치단체가 자기를 위하여 하는 등기로 본다.〈개정 2014. 1. 14., 2017. 1. 17., 2019. 4. 30.〉

1. 지번이나 행정구역의 명칭이 변경된 경우
2. 제22조에 따른 사용승인을 받은 건축물로서 사용승인 내용 중 건축물의 면적 · 구조 · 용도 및 층수가 변경된 경우
3. 「건축물관리법」 제30조에 따라 건축물을 해체한 경우
4. 「건축물관리법」 제34조에 따른 건축물의 멸실 후 멸실신고를 한 경우

② 제1항에 따른 등기촉탁의 절차에 관하여 필요한 사항은 국토교통부령으로 정한다.〈개정 2013. 3. 23.〉

출처: 「건축법」. 법률 제20424호, 2024. 3. 26., 일부개정

제26조(등기촉탁의 절차 등)

① 법 제39조제1항에 따른 등기촉탁 대상 건축물의 소유자 또는 건축주는 사용승인서를 교부받은 날 또는 건축물대장이 말소된 날부터 1개월 이내에 다음 각 호의 서류를 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장에게 제출하여야 한다. 다만, 해당 건축물에 대한 등기가 없는 경우에는 그러하지 아니하다.〈개정 2009. 1. 20., 2011. 9. 16., 2012. 11. 16., 2017. 1. 20., 2017. 7. 18., 2023. 8. 1.〉

1. 등록면허세영수필 확인서 및 통지서(「지방세법」 제26조제1항 본문에 따라 등록면허세 부과가 면제되는 경우는 제외한다)
2. 건물 등기사항증명서 등 등기에 필요한 서류

② 제1항에 따른 서류를 제출받은 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 건축물대장을 변경한 날 또는 건축물대장을 말소한 날부터 1개월 이내에 별지 제25호서식의 건물표시변경 등기촉탁서에 다음 각 호의 서류를 첨부하여 관할 등기소에 등기를 촉탁하여야 한다.〈개정 2009. 1. 20., 2011. 9. 16., 2012. 11. 16., 2017. 1. 20., 2017. 7. 18., 2023. 8. 1.〉

1. 건축물대장 등본
2. 등기촉탁서 부본
3. 등록면허세영수필 확인서 및 통지서(「지방세법」 제26조제1항 본문에 따라 등록면허세 부과가 면제되는 경우는 제외한다)
4. 삭제〈2017. 7. 18.〉

③ 특별자치시장 · 특별자치도지사 또는 시장 · 군수 · 구청장은 제2항에 따른 등기촉탁 내용을 별지 제26호서식의 건물표시변경등기촉탁대장에 기재하여야 한다.〈개정 2009. 1. 20., 2017. 1. 20., 2017. 7. 18.〉

출처: 「건축물대장의 기재 및 관리 등에 관한 규칙」. 국토교통부령 제1235호, 2023. 8. 1., 일부개정

2) 공공데이터법과 건축물 데이터 개방

■ 공공데이터법 개요

「공공데이터의 제공 및 이용 활성화에 관한 법률」(약칭 공공데이터법)은 공공기관이 보유·관리하는 데이터(공공데이터)를 국민이 이용할 수 있도록 제공하고, 민간 활용 활성화를 위한 제도적 기반을 마련하는 것을 그 목적으로 한다(제1조, 제2조). 이를 달성하기 위하여 공공데이터법은 공공기관이 보유·관리하는 데이터 중 제공 가능한 데이터를 사전에 목록화하여 등록하고(제18조), 공공데이터 포털 등을 통하여 제공하도록 규정하고 있다(제21조, 제26조).

공공데이터법 제17조 제1항은 공공기관이 보유·관리하는 모든 데이터를 원칙적으로 국민에게 제공하여야 한다고 제공 대상 공공데이터의 범위를 규정하고 있다. 다만, 정보공개법에서 정하는 국가안전보장 등 비공개 대상 정보와, 「저작권법」 등에서 보호하는 제3자의 권리에 대한 정당한 이용허락을 받지 않은 정보 등은 제공 대상에서 제외하고 있다. 그럼에도 불구하고 비공개 대상 정보를 기술적으로 분리하는 것이 가능한 경우에는 그 부분을 제외한 데이터를 제공하도록 규정하고 있어, 공개할 수 없는 부분을 제외한 모든 데이터에 대한 적극적인 제공이 공공기관의 의무임을 명확히 하고 있다.

■ 공공데이터의 정의

「공공데이터법」에 따른 공공데이터는 공공기관이 보유·관리하는 데이터를 의미하며, 「공공데이터법」 제2조에서는 공공기관과 공공데이터를 정의하고 있다. 이 법에서 공공기관의 정의는 국가기관, 지방자치단체, 공공기관, 지방공사 및 지방공단, 특수법인, 학교 등이며, 공공데이터의 정의는 행정정보, 공공기관이 생산한 정보, 웹기록물 및 행정정보 데이터세트 등의 기록정보자료 등이다.

따라서 건축행정업무를 수행하는 건축행정시스템(세움터)에서 생성·관리되고 있는 건축행정 전산화 자료는 공공데이터에 해당한다.

제2조(정의)

1. “공공기관”이란 국가기관, 지방자치단체 및 「지능정보화 기본법」 제2조제16호에 따른 공공기관을 말한다.
2. “공공데이터”란 데이터베이스, 전자화된 파일 등 공공기관이 법령 등에서 정하는 목적을 위하여 생성 또는 취득하여 관리하고 있는 광(光) 또는 전자적 방식으로 처리된 자료 또는 정보로서 다음 각 목의 어느 하나에 해당하는 것을 말한다.
   가. 「전자정부법」 제2조제6호에 따른 행정정보
   나. 「지능정보화 기본법」 제2조제1호에 따른 정보 중 공공기관이 생산한 정보
   다. 「공공기록물 관리에 관한 법률」 제20조제1항에 따른 전자기록물 중 대통령령으로 정하는 전자기록물
   라. 그 밖에 대통령령으로 정하는 자료 또는 정보

출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률」. 법률 제19408호, 2023. 5. 16., 타법개정

제2조(정의)

16. "공공기관"이란 다음 각 목의 어느 하나에 해당하는 기관을 말한다.
  가. 「공공기관의 운영에 관한 법률」에 따른 공공기관
  나. 「지방공기업법」에 따른 지방공사 및 지방공단
  다. 특별법에 따라 설립된 특수법인
  라. 「초 · 중등교육법」, 「고등교육법」 및 그 밖의 다른 법률에 따라 설치된 각급 학교
  마. 그 밖에 대통령령으로 정하는 법인 · 기관 및 단체

출처: 「지능정보화 기본법」. 법률 제20410호, 2024. 3. 26., 일부개정

제1조의2(공공데이터에 해당하는 전자기록물의 범위)

「공공데이터의 제공 및 이용 활성화에 관한 법률」(이하 "법"이라 한다) 제2조제2호다목에서 "대통령령으로 정하는 전자기록물"이란 웹기록물 및 행정정보 데이터세트 등의 기록정보자료를 말한다.

출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률 시행령」. 대통령령 제33842호, 2023. 11. 7., 일부개정

■ 건축물 데이터 개방

국토교통부는 건축행정시스템 세움터에서 생산·보유하고 있는 건축행정 전산자료 중 개인정보(소유자 정보) 등 일부 데이터를 제외하고 공공데이터법에 따라 개방 및 제공하고 있다. 2024년 말까지 건축데이터 민간개방시스템(https://open.eais.go.kr)을 통하여 건축행정 전산자료를 제공하였고, 2025년 초부터 건축허브(https://www.hub.go.kr)로 이관되어 건축물대장, 인허가 등 건축·주택 업무 유형별 현황 데이터를 가공하지 않은 형태 그대로 제공하고 있다. 특히 건축허브에서는 세움터에서 생산된 데이터뿐만 아니라 건물에너지(전기, 가스), 건축물유지관리 등 건축물 관련 데이터를 종합적으로 제공하고 있다.

제공 방식은 검색 결과에 대한 파일 다운로드(csv, 엑셀 등 형식) 제공, OpenAPI 제공, 대용량 텍스트 파일 제공 등 3가지 방식이다. 파일 다운로드의 경우 주소 검색을 통하여 일정 지역에 대한 전체 데이터를 테이블 별로 조회 및 다운로드할 수 있다. OpenAPI의 경우 개발자나 시스템 간 자동 연동을 위한 RESTful API 제공. JSON, XML 포맷을 지원한다. 대용량 제공의 경우 전국 건축물에 대한 정보를 가공하지 않은 텍스트 파일 형태로 제공하고 있으며, 매월 갱신 데이터를 주기적으로 제공하고 있다.

[그림 2-1] 건축HUB 데이터 제공 현황
출처: 건축HUB, https://www.hub.go.kr/portal/main.do(검색일: 2025.09.16.)

건축물 데이터는 가공 없이 제공되고 있기 때문에, 데이터의 연속성과 일관성이 보장되지는 않는다. 2025년 초 건축허브로의 이관 이후 2024년까지 건축데이터 민간개방시스템을 통하여 제공된 과거 데이터와 차이점이 발생하였다. 가장 중요하게는 기존 민간개방시스템의 건축데이터에서 사용되던 PK(Primary Key)가 변경되었다. 클라우드 세움터 전환 과정에서 내부적으로는 이미 PK가 변경되었으나, 기존 민간개방시스템을 통하여 제공되던 공공데이터에는 자체 변환을 통하여 예전과 동일한 PK로 제공하고 있었다. 그러나 건축허브로 이관되면서 이러한 가공이 이루어지지 않고, 이미 내부적으로 사용되고 있었던 신규 PK를 그대로 제공하게 변경되었다. 신규 PK는 기존 PK를 일정 규칙에 따라 변환한 것이다. 건축허브는 기존 PK에서 신규 PK로 변환하는 규칙을 제공하고 있으나, 그 역방향으로의 연계에는 어려움이 있어 과년도 데이터와 연계 활용에 한계가 있는 현황이다.

■ 공공데이터 품질관리

- 공공데이터의 제공 운영실태 평가

「공공데이터법」 제9조제1항에 따라 매년 공공기관을 대상으로 공공데이터의 제공기반조성, 제공현황 등 제공 운영실태를 평가하고 있다. 평가기준에는 공공데이터 제공 책임관 및 실무 담당자의 임명·운영 실태, 공공데이터 목록의 등록 및 유지관리 실태, 공공데이터 목록정보의 공표 실태, 공공데이터 포털에의 등록 실태, 공공데이터의 제공기반 구축 실태 등이 포함된다.

> 제9조(공공데이터의 제공 운영실태 평가)
>
> ① 행정안전부장관은 매년 공공기관(국회 · 법원 · 헌법재판소 및 중앙선거관리위원회는 제외한다. 이하 이 조에서 같다)을 대상으로 공공데이터의 제공기반조성, 제공현황 등 제공 운영실태를 대통령령으로 정하는 바에 따라 평가하여야 한다.〈개정 2014. 11. 19., 2017. 7. 26.〉
>
> 출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률」. 법률 제19408호, 2023. 5. 16., 타법개정

제10조(공공데이터의 제공 운영실태 평가의 기준 등)

① 행정안전부장관은 법 제9조제1항에 따라 공공데이터의 제공 운영실태를 평가하려는 경우에는 그 평가대상 · 평가기준 및 평가방법 등을 미리 공표하여야 한다.〈개정 2014. 11. 19., 2017. 7. 26.〉

② 제1항에 따른 평가기준에는 다음 각 호의 사항이 포함되어야 한다.〈개정 2014. 11. 19., 2017. 7. 26.〉

1. 법 제12조제1항에 따른 공공데이터제공책임관 및 실무담당자의 임명 · 운영 실태
2. 법 제18조제1항에 따른 공공데이터 목록의 등록 및 유지관리 실태
3. 법 제19조제3항에 따른 공공데이터 목록정보의 공표 실태
4. 법 제19조제4항 및 제21조제2항에 따른 제공대상 공공데이터의 법 제21조제1항에 따른 공공데이터 포털(이하 "공공데이터 포털"이라 한다)에의 등록 실태
5. 법 제24조제1항 및 제2항에 따른 공공데이터의 제공기반 구축 실태
6. 그 밖에 공공데이터의 제공 운영실태 평가를 위하여 필요하다고 행정안전부장관이 인정하는 사항

출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률 시행령」. 대통령령 제33842호, 2023. 11. 7., 일부개정

동법 제9조제2항에서는 평가 결과 개선이 필요하다고 권고한 사항에 대해서 공공기관에 시정요구 등 조치를 위할 것을 명시하고 있다. 시정조치를 요구받은 공공기관의 장은 그 시정요구를 받은 날부터 30일 이내에 필요한 조치를 하고, 그 조치 결과를 행정안전부장관에게 통보해야 한다.

제9조(공공데이터의 제공 운영실태 평가)

② 행정안전부장관은 제1항에 따른 평가결과를 전략위원회와 국무회의에 보고한 후 이를 공공기관의 장에게 통보하고 공표하여야 하며, 전략위원회가 개선이 필요하다고 권고한 사항에 대하여는 해당 공공기관에 시정요구 등의 조치를 취하여야 한다.〈개정 2014. 11. 19., 2017. 7. 26.〉

출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률」. 법률 제19408호, 2023. 5. 16., 타법개정

제11조(공공데이터의 제공 운영실태 평가의 시정조치 및 조치결과의 통보)

법 제9조제2항에 따른 시정조치를 요구받은 공공기관의 장은 그 시정요구를 받은 날부터 30일 이내에 필요한 조치를 하고, 그 조치결과를 행정안전부장관에게 통보하여야 한다. 다만, 조치가 완료되지 아니한 경우에는 그 사유와 향후 조치계획을 제출하여야 한다.〈개정 2014. 11. 19., 2017. 7. 26.〉

출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률 시행령」. 대통령령 제33842호, 2023. 11. 7., 일부개정

- 공공데이터의 품질관리

「공공데이터법」 제22조에 따라 공공데이터의 품질관리 및 품질수준 확보를 위한 조치로 품질 진단·평가, 개선지원 등 필요한 시책을 수립·추진하여야 하며, 이를 위해 공공데이터에 대한 품질을 진단·평가하고 그 결과를 공표할 수 있다.

제22조(공공데이터의 품질관리)

① 공공기관의 장은 해당 기관이 생성 또는 취득하여 관리하는 공공데이터의 안정적 품질관리 및 적정한 품질수준의 확보를 위하여 필요한 조치를 취하여야 한다.
② 과학기술정보통신부장관과 행정안전부장관은 공공데이터의 적정한 품질수준의 확보와 제공 촉진을 위하여 품질 진단 · 평가, 개선지원 등 필요한 시책을 수립 · 추진하여야 한다.〈개정 2014. 11. 19., 2017. 7. 26.〉
③ 행정안전부장관은 과학기술정보통신부장관과 협의하여 제2항에 따라 정기적으로 사회적 · 경제적 파급효과가 큰 제공대상 공공데이터에 대한 품질 진단 · 평가를 실시하고 그 결과를 공표할 수 있다.〈개정 2014. 11. 19., 2017. 7. 26.〉
④ 그 밖에 공공데이터 품질 진단 · 평가 등에 필요한 사항은 대통령령으로 정한다.

출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률」. 법률 제19408호, 2023. 5. 16., 타법개정

공공데이터 품질의 진단 및 평가 등에 필요한 사항은 「공공데이터법 시행령」 제17조에 따른다. 품질 진단·평가의 기준은 공공데이터의 구조 및 성능, 품질관리체계, 표준화 준수, 오류 등이다. 또한 품질 진단·평가를 실시하려는 경우 품질 진단·평가 계획을 수립하여야 하며, 품질 진단·평가 대상 공공데이터, 추진체계, 절차 및 시간, 결과에 따른 시정조치가 포함되어야 한다. 품질 진단·평가 결과에 따라 개선이 필요한 공공데이터에 대해 해당 공공기관의 장에게 시정을 요구할 수 있다.

제17조(공공데이터의 품질 진단 및 개선)

① 법 제22조제3항에 따른 공공데이터의 품질 진단 · 평가의 기준은 다음 각 호와 같다.〈개정 2014. 11. 19., 2017. 7. 26.〉
1. 공공데이터 구조 및 성능
2. 공공데이터 품질관리체계
3. 공공데이터 표준화 준수
4. 공공데이터값 오류
5. 그 밖에 행정안전부장관이 필요하다고 인정하는 사항

② 행정안전부장관은 법 제22조제3항에 따라 공공데이터의 품질 진단 · 평가를 실시하려는 경우에는 과학기술정보통신부장관과 협의하여 다음 각 호의 사항이 포함된 품질 진단 · 평가 계획을 수립하여야 한다.〈신설 2016. 4. 5., 2017. 7. 26.〉
1. 품질 진단 · 평가 대상 공공데이터
2. 품질 진단 · 평가의 추진체계
3. 품질 진단 · 평가의 절차 및 기간
4. 품질 진단 · 평가 결과에 따른 시정조치
5. 그 밖에 품질 진단 · 평가 시행에 필요한 사항

④ 행정안전부장관은 법 제22조제3항에 따른 품질 진단 · 평가 결과에 따라 개선이 필요한 제공대상 공공데이터에 대하여 과학기술정보통신부장관과 협의하여 해당 공공기관의 장에게 시정을 요구할 수 있다.〈개정 2014. 11. 19., 2016. 4. 5., 2017. 7. 26.〉

출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률 시행령」. 대통령령 제33842호, 2023. 11. 7., 일부개정

2. 건축물 데이터 품질 및 활용 현황 분석

1) 건축물 데이터 구조 및 특성

■ 건축물 데이터 구조

건축물 데이터 품질 및 활용 현황을 고려하여 선별된 건축물 데이터를 대상으로, 데이터 구조 및 특성을 파악하였다. 분석 범위는 건축허브에서 제공하는 건축물 관련 데이터 중 본 연구에서 다루고자 하는 건축물대장, 건축인허가, 주택인허가 데이터를 대상으로, 품질 고도화 필요성 및 활용도가 높은 데이터를 중심으로 그 구조를 검토하고자 한다. 이들 데이터는 건축행정시스템 세움터에서 건축허브와의 연계를 통하여 제공하는 것으로, 가공 없이 원데이터를 제공하고 있어 세움터 데이터 구조와 일치한다고 볼 수 있다. 다만, 개인정보 등 비공개 정보를 제외하였기 때문에 일부 테이블, 컬럼 등이 제외되었다.

- 건축물대장 데이터 구조

건축물대장은 기본적으로 독립된 건축물(동) 단위로 생성되나, 한 대지에 여러 건물동이 있는 경우, 한 건물동을 여러 소유자가 구분소유하는 경우 등에는 대지 단위, 호 단위 건축물대장이 함께 생성된다. 건물동 단위 작성되는 건축물대장은 한 소유자가 구분없이 소유하는 경우 일반건축물대장이 작성되며, 여러 소유자가 구분소유하는 집합건물의 경우에는 집합건축물대장 중 표제부가 작성된다. 집합건물의 경우 표제부와 함께 구분소유의 단위인 호별로 전유부가 함께 작성된다. 집합건물 해당 여부와 별개로, 한 대지에 여러 건물동이 있는 경우 대지별로 총괄표제부가 작성된다. 본 연구에서는 기본적으로 동 단위 데이터 활용도가 높은 점을 반영하여 일반건축물대장 또는 집합건축물대장(표제부)를 중심으로 데이터 구조 및 특성을 파악하고자 한다. 다만, 건축물대장 구조 상 총괄표제부에 기재한 내용은 동 단위 건축물대장에 기재하지 않도록 하고 있어, 총괄표제부가 존재하는 경우 동 단위 건축물대장과 함께 검토할 필요가 있다.

[그림 2-2] 건축물대장 종류
출처: 연구진 작성

동 단위 건축물대장에 해당하는 일반건축물대장과 집합건축물대장(표제부)에 기재되는 내용은 크게 다르지 않으며, 건축허브에서 제공하는 건축물대장 데이터에서 두 대장은 동일한 테이블(표제부)로 관리되고 있다. 표제부 테이블(일반건축물대장 포함)에는 건물ID, 고유번호, 대지 위치, 명칭 등 건축물을 식별할 수 있는 기본정보와, 대지면적, 연면적, 건축면적, 건폐율, 용적률 등 면적 관련 정보, 구조, 용도, 지붕, 높이, 층수 등 건축물 관련 개요, 층별 건축물 현황(구조, 용도, 면적 등), 소유자 현황 등 건축물 관련 정보가 종합되어있다.

발급확인번호 : MAML-AWBK-GBIG-RGXQ-VPYK

■ 건축물대장의 기재 및 관리 등에 관한 규칙 [별지 제1호서식](개정 2023. 8. 1.)

일반건축물대장(갑)

(2쪽 중 제1쪽)

건물ID		고유번호		명칭	KT&G세종타워B	호수/가구수/세대수	0호/0가구/0세대
대지위치	세종특별자치시 어진동	지번	524	도로명주소	세종특별자치시 가름로 143 (어진동)		
※대지면적	6,001.4 ㎡	연면적	48,074.7 ㎡	※지역		※지구	경관지구	※구역	지구단위계획구역
건축면적	4,194.94 ㎡	용적률 산정용 연면적	28,945.22 ㎡	주구조	철근콘크리트구조	주용도	업무시설, 판매시설, 교육연구시설(타워A 6층)	층수	지하: 5층, 지상: 13층
※건폐율	69.8994 %	※용적률	482.3078 %	높이	50.8 m	지붕	(철근)콘크리트	부속건축물	동 ㎡
※조경면적	945.91 ㎡	※공개 공지·공간 면적	650.2 ㎡	※건축선 후퇴면적	622.36 ㎡	※건축선후퇴 거리	3 m		

건축물 현황					소유자 현황			
구분	층별	구조	용도	면적(㎡)	성명(명칭) 주민(법인)등록번호 (부동산등기용등록번호)	주소	소유권 지분	변동일 변동원인
주1	지5층	철근콘크리트구조	기계실,전기실	1,580.7			1/1	2020.6.29.
주1	지4층	철근콘크리트구조	지하주차장	2,405.65				소유권보존
주1	지4층	철근콘크리트구조	부대시설(창고 등)	771.29		- 이하여백 -		
주1	지3층	철근콘크리트구조	지하주차장	3,995.86		※이 건축물대장은 현소유자만 표시한 것입니다.		

이 등(초)본은 건축물대장의 원본내용과 틀림없음을 증명합니다.

세종특별자치시장

세종특별자치시장인

발급일:
담당자:
전 화:

※ 표시 항목은 총괄표제부가 있는 경우에는 적지 않을 수 있습니다.

297mm×210mm[백상지 80g/㎡]

[그림 2-3] 일반건축물대장 예시
출처: 세움터(https://www.eais.go.kr/)에서 연구진 발급

일반건축물대장 또는 표제부의 기재내용은 건축물대장 데이터 중 여러 테이블이 연계되어 있다. 기본개요 테이블은 고유번호, 대지위치 등 모든 건축물대장을 식별할 수 있는 정보가 저장되어 있다. 표제부 테이블은 건물동마다 유일하게 부여되는 정보를 포괄하고 있으며, 일대다(1:n) 관계로 연계되는 층별 건축물 현황은 층별개요 테이블에서 별도로 관리하고 있다. 반대로 다대일(n:1) 관계로 여러 건물동이 하나의 대지에 연계되는 총괄표제부의 경우, 대지별로 고유한 정보는 일반건축물대장 또는 표제부에 기재하지 않고 이러한 내용을 별도로 표기하고 있다.

[그림 2-4] 건축물대장과 건축물대장 데이터 테이블 연계
출처: 연구진 작성

- 건축물대장 데이터 구성

건축물대장 데이터는 건축서비스산업 정보체계 건축HUB를 통해 세움터 현황 데이터를 가공 없이, 업무별 대용량 개방 데이터 형태로 매월 주기적으로 제공되고 있다. 2025년 02월 기준 자료는 총 10개의 테이블로 구성되며, 각 테이블은 여러 개의 컬럼으로 이루어져 있다. 모든 테이블에는 첫 번째 컬럼으로 '관리_건축물대장_PK'가 포함되어 있어, 이를 기준으로 테이블 간 연결이 가능하다.

건축물대장 데이터는 테이블별 정보와 건수에 따라 다음 표와 같이 구분할 수 있다. 각 건물동 단위의 기록(예: 101동, 102동)을 담은 표제부가 있으며, 이 여러 표제부를 묶는 상위 단위(예: 아파트 단지 전체)는 총괄표제부로 구분된다. 또한, 표제부 안에는 개별 세대나 점포 등 구분소유된 부분(예: 101동 101호, 상가 1층 101호)이 기록되어 있으며, 이를 전유부라고 한다. 따라서 전유부는 표제부 안에 속하는 하위 요소이고, 총괄표제부는 표제부를 아우르는 상위 요소로 이해할 수 있다.

[표 2-4] 건축물대장 데이터 종류와 수

구분	내용
기본개요	데이터에는 모든 관리대장 PK 값이 포함되며, 총괄표제부, 표제부, 일반건축물 세 유형으로 구분된다. 각 데이터는 해당 대지의 위치 정보를 포함한다.
총괄표제부	대표 필지에 여러 동의 건물이 존재하는 경우, 각 동의 면적은 모두 합산하여 제공한다.
표제부	일반건축물대장과 집합건축물대장 표제부는 개별적으로 독립된 건물들에 대한 데이터를 포함한다.
지역지구구역	지역 · 지구 · 구역 정보는 대표 필지 중심으로 기록된다. 만약 하나의 필지가 3개 구역에 걸쳐 있다면, 동일 PK로 3개의 레코드를 생성한다.
부속지번	건물이 등록된 대표 필지 외에도 여러 지번에 걸쳐 있는 경우 그 내역을 포함한다. 가령 한 대표 필지에 속한 건물이 추가로 3개의 필지에 걸쳐 있는 경우, 3개의 레코드가 동일한 PK로 생성된다.
전유부	모든 전유부는 O동 O호 형태로 목록화되어 있다. 이 정보는 전유공용면적 테이블에도 모두 포함되므로 중복이 될 수 있다. 다만, 이 테이블은 중복 없이 동 · 호 번호가 기재되어 있고, 전유공용면적 테이블에 비하여 상대적으로 용량이 작다는 점에서 활용 가치가 있다.
전유공용면적	집합건축물 전유부에 대해서는 모든 전유면적과 공용면적이 표기되어 있다.
층별개요	건물의 각 층에 있는 모든 용도 및 시설의 면적은 분할된 형태로 기록되어 있으며, 각각 하나의 레코드로 존재한다.
공동주택가격	공동주택의 각 전유부에는 연도별 공시지가 정보가 기록되어 있으며, 예를 들어 3개 연도의 공시지가를 포함할 경우 해당 전유부는 3개의 레코드를 가진다.
오수정화시설	오수정화시설의 개요 정보가 포함되어 있다.

출처: 김승범. (2015). 건축물대장 원시데이터의 관계 구조와 탐색적 분석을 통한 데이터 활용. 한국문화공간건축학회, 50, 110-120. 바탕으로 재구성

- 건축인허가 데이터 구성

건축인허가 데이터는 「건축법」 등에 따라 국토교통부가 관리하는 지방자치단체의 건축 인허가 정보를 바탕으로 구성되며, 건축HUB를 통해 매월 주기적으로 제공된다. 2025년 2월 기준으로 제공되는 데이터는 총 17개의 테이블로 구성되어 있으며, '호별전유공용면적'을 제외한 나머지 테이블에는 공통적으로 '관리_허가대장_PK'가 포함되어 있어, 이 일련번호를 기준으로 테이블 간의 연계가 가능하다. 그 외에도 '관리_동별_개요_PK', '관리_층별_개요_PK', '관리_호별_전유_공용_면적_PK', '관리_호별_명세_PK', '관리_대지_위치_PK', '관리_전유_공용_면적_PK' 등의 고유 식별자가 포함되어 있다. 본 연구에서는 2024년에 사용승인이 발생한 건축인허가 데이터만을 분석 대상으로 삼고 있으며, 각 테이블별 정보와 건수는 다음 표와 같이 정리할 수 있다.

[표 2-5] 건축인허가 데이터 종류와 수

구분	내용
기본개요	인허가 대상 건축물의 기본 정보
주택유형	해당 건축물이 속한 주택 유형(준주택, 도시형생활주택 등)
동별개요	동일 인허가 건에 포함된 각 동의 규모, 층수, 용도 등
층별개요	각 동의 층별 면적, 층고, 층수, 용도 등 층 단위 상세정보
공작물관리대장	옹벽, 광고판 등 건축물이 아닌 공작물에 대한 관리 정보
가설건축물	임시 건축물(컨테이너 등)에 대한 허가 정보
호별전유공용면적	각 호(세대)의 전유면적과 공용면적에 대한 세부정보
부설주차장	건축물에 부설된 주차장의 위치, 지목 등
대지위치	건축물의 지번, 도로명 주소 등 공간적 위치 정보

구분	내용
호별개요	각 호의 동, 층, 평형 등 세대 단위 정보
대수선	기존 건축물의 주요 구조 변경 등 대수선 허가 관련 정보
철거멸실관리대장	건축물 철거 또는 멸실에 대한 허가 및 관리 기록
주차장	독립된 주차장 시설의 위치, 규모, 용도 등의 정보
지역지구구역	해당 건축물의 위치가 속한 용도구역, 용도지역 등
도로대장	인접 도로 지정번호, 구분 코드, 폭, 길이, 면적 등
오수정화시설	정화조 설치 여부, 용량, 처리 방식 등 하수처리 관련 정보
전유공용면적	건물 층별 전유 및 공용면적 정보

출처: 연구진 작성

- 주택인허가 데이터 구성

주택인허가 데이터는 주택건설사업계획승인 정보에서 관리하는 주택 인허가 정보를 바탕으로 구성되며, 건축HUB를 통해 매월 주기적으로 제공된다. 2025년 2월 기준으로 제공되는 데이터는 총 16개의 테이블로 구성되어 있으며, '관리_주택대장_PK'는 건축인허가 데이터와는 달리, 6개의 테이블에 존재하지 않으며 총 10개의 테이블에만 존재하고 있다. 그 외에도 '관리_동별_개요_PK', '관리_층별_개요_PK', '관리_호별_명세_PK', '관리_행위_호_전유_공용_PK', '관리_전유_공용_PK', '관리_형벌_개요_PK', '관리_공동주택대장_PK' 등의 고유 식별자가 포함되어 있다. 본 연구의 분석 대상인 2024년에 사용승인이 발생한 주택인허가 데이터에 대한 각 테이블별 정보와 건수는 다음 표와 같이 정리할 수 있다.

[표 2-6] 주택인허가 데이터 종류와 수

구분	내용
기본개요	인허가 대상 주택의 기본 정보
행위개요	건축행위의 종류, 규모, 용도 등 사업 행위에 대한 개요
대지위치	해당 주택사업의 위치, 지목, 대지 형상 등 공간정보
동별개요	사업 대상지에 포함된 각 동의 용도, 층수, 세대수 등 정보
층별개요	각 동의 층별, 연면적, 용도, 층수 등 층 단위 정보
호별개요	각 세대(호)의 평형, 층수 등 세대 정보
부대시설	주차장, 관리사무소, 보안등 등 부대시설 현황
주차장	주차장에 대한 위치, 규모, 차량 수용 가능 수 등
부설주차장	건축물에 부설된 주차장의 위치, 지목 등
행위호전유공용면적	세대별 전유/공용면적 정보 (행위별 기준으로 관리됨)
전유공용면적	사업 전체 또는 동별 전유/공용면적 정보
관리공동부대복리시설	주차장, 노인정, 경비실, 유치원 등 부대복리시설 정보
지역지구구역	해당 사업 대상지가 속한 용도구역, 용도지역 등
복리분양시설	어린이놀이터, 회의장, 배드민턴 등 비주거 시설 정보
관리공동형별개요	승강기, 수도, 난방 등 주택관리에 대한 개요
오수정화시설	정화조, 하수처리시설 등 오수처리 관련 계획 및 설치 정보

출처: 연구진 작성

2) 건축물데이터 품질 현황

건축물 데이터의 품질 고도화 방향을 설정하기 위하여 건축물 데이터 품질 및 활용 현황을 살펴보았다. 첫째로 국토교통부와 허가권자인 지자체의 건축물대장 정비 추진 과정을 통하여 건축물 데이터 품질 현황을 파악하였다. 건축물대장 정비 근거와 정비 추진 현황을 검토하고, 정부와 지자체가 수행한 건축물대장 품질 점검 업무 규칙 내용을 통하여 건축물 데이터 오류 현황을 고찰하였다. 둘째로는 건축물 데이터 중 활용도 높은 데이터를 파악하기 위하여 건축물 데이터 활용 현황을 분석하였다. 먼저 건축물 데이터에 기반하여 생산되는 건축물 통계, 건축허가·착공·준공통계의 작성 방식을 검토하여 통계 생산에 활용되는 건축물 데이터 성격을 분석하였다. 다음으로 건축물 데이터를 활용한 연구 사례 조사를 통하여 건축·도시 연구에 활용되고 있는 건축물 데이터 현황을 파악하였다. 이를 통하여 활용도가 높고 품질 고도화가 요구되는 건축물 데이터를 선별하고 품질 고도화 방향 설정 대상으로 설정하고자 하였다.

- 건축물대장 정비 관련 규정

「건축법」 제38조 제1항은 건축물대장에 대한 지속적 정비 의무를 규정하고 있다. 제3항에서 세부 사항을 국토교통부령에 위임하고 있으며, 이에 따라 건축물대장의 기재 및 관리 등에 관한 규칙 제21조에서 건축물대장 기재내용의 오류 및 누락을 정정할 수 있도록 하고 있다. 허가권자는 건축물대장 기초자료를 작성·관리하고(제1항), 이를 통해 기재내용의 오류 또는 누락을 발견한 경우 이를 직권으로 정정할 수 있다(제2항 전단). 건축물의 소유자도 기재내용의 정정을 신청할 수 있으며(제3항), 허가권자가 직권 정정한 경우에도 건축물의 소유자에 그 내용을 통지하여야 한다(제2항 후단).

- 건축물대장 정비 추진

국토교통부는 과거부터 꾸준히 건축물대장 정리 및 정비를 추진하여왔으며, 공공데이터 개방 이후로는 2018년부터 「공공데이터의 제공 및 이용 활성화에 관한 법률」 제9조(공공데이터의 제공 운영실태 평가), 제22조(공공데이터의 품질관리)에 따라 데이터 품질관리 체계를 구축·운영하면서 건축물대장 오류 데이터 정비를 위해 지자체 협조를 요청하였다. 협조 요청 내용은 2018년 통계 항목 중심의 정비 항목 12개 업무규칙(주용도코드, 구조코드, 연면적, 층별면적, 소유자, 사용승인일 등), 2019년 활용도 높은 항목 중심의 7개(높이, 허가일, 착공일, 사용승인일, 소유자, 시군구코드, 법정동코드 등), 2020년 소유권 오류 점검 업무규칙 1개 등이다. 2022년 3월 11일 기준, 각 연도별 정비 완료율은 58%('18년), 27%('19년), 27%('20년) 수준으로 미비하였다.

[표 2-7] 지자체 건축물대장 품질 점검 업무 규칙

구분	점검 업무규칙	주요 점검사항	비고
2018년 (통계 항목 중심)	(1) 일반건축물(표제부) 주용도코드	건축물의 주용도코드 누락, 과거 코드 등록 자료에 대한 검증	구조, 용도별 건축물 통계 활용
	(2) 층별 용도코드	건축물의 층별 용도 누락, 과거 코드 등록 자료에 대한 검증	
	(3) 일반건축물(표제부) 구조코드	건축물의 주구조코드 누락, 과거 코드 등록 자료에 대한 검증	
	(4) 층별 구조코드	건축물의 층별 구코드 누락, 과거 코드 등록 자료에 대한 검증	
	(5) 전유부 세부용도 코드	전유공용면적 중 세부용도코드가 누락되거나 과거코드가 등록된 자료에 대한 검증	
	(6) 전유부 세부구조 코드	전유공용면적 중 세부구조코드가 누락되거나 과거코드가 등록된 자료에 대한 검증	
	(7) 일반건축물(표제부) 연면적과 층별 면적	건축물의 연면적과 각층의 면적 합계가 상이한 자료에 대한 검증	면적 구분별 건축물 통계 활용
	(8) 소유자 중복	일반건축물 및 전유부에 대하여 소유자 정보가 중복 기재된 자료에 대한 검증	소유조회 등 재산권 확인 활용
	(9) 소유자 지분	건축물의 소유권 지분 합이 100%가 아닌 자료에 대한 검증	
	(10) 건축물대장 사용승인일 누락	건축물대장 상에 사용승인일이 누락된 자료에 대한 검증	노후 건축물 통계 추출 활용
	(11) 일반건축물(표제부) 필로티 건축물	필로티 건축물로 추정되나, 1층 정보에 면적 제외가 설정되지 않은 자료에 대한 검증	필로티 건축물 현황 파악 활용
	(12) 건축물대장의 변동원인 중복	건축물대장의 변동사항이 중복 기재된 자료에 대한 검증	불필요한 중복 자료 제거
2019년 (활용도 높은 항목)	(13) 건축물의 높이	2층 이상, 높이 50m 이상 건축물 중 층간 높이가 5m가 넘는 주택용도 건축물	
	(14) 허가/착공/사용승인일의 오류	허가, 착공, 사용승인일에 미래 일자가 입력된 자료 검증	
	(15) 허가/착공/사용승인일의 순서	허가, 착공, 사용승인일의 순서가 변경된 자료 검증	
	(16) 층별현황 등록여부	건축물의 층별현황이 미등록된 자료 검증	
	(17) 소유자 등록여부	건축물의 소유자가 미등록된 자료 검증	
	(18) 건축물의 초고층	건축물의 지상층 수와, 층별개요의 최고층 자료 검증	
	(19) 시군구, 법정동 코드	건축물의 주소(시군구, 법정동) 코드가 잘못 부여된 데이터 검증	
2020년	(20) 소유권 오류	건축물대장과 건물등기사항증명서의 소유권이 상이한 자료 검증(부동산고유번호 매핑자료에 한함)	

출처: 세움터 내부자료

- 86개 정비 대상 업무 규칙

2022년 세움터에서는 건축물대장 오류 데이터 분석 및 정비 대상 항목 개발을 통해 86개 업무 규칙에 대한 유형을 제시하였다. 업무규칙 86개 중 건축물대장의 정비대상 항목은 39개였고, 건축인허가대장은 21개, 주택인허가대장은 26개이다.

39개 건축물대장 업무규칙은 건축면적, 연면적, 용적률산정연면적, 건폐율, 용적률, 주부속구분, 주용도코드, 허가일, 착공일, 사용승인일, 대지면적, 지상층수, 지하층수, 건폐율, 용적률, 세대수, 호수, 호명칭, 소유자 등을 정비한다. 21개 건축인허가대장에서는 대지면적, 건축면적, 연면적, 주차대수, 허가일, 건폐율,

용적률, 주용도코드, 구조코드, 용적률산정연면적, 지상층수, 지하층수, 가구수, 세대수, 호수 등이며, 26개 주택인허가대장에서는 대지면적, 건축면적, 연면적, 주차대수, 허가일, 건폐율, 용적률, 주용도코드, 구조코드, 용적률산정연면적, 지상층수, 지하층수, 가구수, 세대수, 호수, 소유자 지분, 높이, 소유자 등이다.

[표 2-8] 2022년 86개 업무규칙 및 정비대상 항목

순번	구분	업무규칙명
1	건축물대장	(건축면적) 총괄표제부 건축면적과 표제부 건축면적 합계의 일관성
2		(연면적) 총괄표제부 연면적과 표제부 연면적 합계의 일관성
3		(용적률산정연면적) 총괄표제부 용적률산정연면적과 표제부 용적률산정연면적 합계의 일관성
4		(건폐율) 총괄표제부 건폐율 계산의 정확성
5		(용적률) 총괄표제부 용적율 계산의 정확성
6		(주건축물수) 총괄표제부 주건축물 수와 표제부 주건축물수 합과의 일관성
7		(부속건축물수) 총괄표제부 부속건축물 수와 표제부 부속건축물수 합과의 일관성
8		(부속건축물 면적) 총괄표제부 부속건축물 면적과 표제부 부속건축물수 연면적 합과의 일관성
9		(세대수) 총괄표제부 세대수와 표제부 세대수 합의 일관성
10		(주용도코드) 총괄표제부 주용도코드의 정확성
11		(사용승인일, 허가일) 총괄표제부 내 사용승인일 보다 큰 허가일 검증(정확성)
12		(착공일, 허가일) 총괄표제부 내 착공일 보다 큰 허가일 검증(정확성)
13		(대지면적) 총괄표제부 내 대지면적의 값이 건축면적보다 작은 경우의 데이터 검증(정확성)
14		(대지면적) 일반건축물대장 및 표제부 내 대지면적의 값이 건축면적보다 작은 경우의 데이터 검증(정확성)
15		(주용도코드) 일반건축물대장 및 표제부의 주용도코드 입력의 정확성
16		(건축면적) 일반건축물대장 및 표제부의 바닥면적의 합이 가장 큰 면적과 표제부의 건축면적이 다른 데이터 검증(정확성)
17		(지상층수) 일반건축물 및 표제부의 지상층수와 층별 개요 지상층 번호 최대값이 상이한 경우 데이터 검증(일관성)
18		(지하층수) 일반건축물 및 표제부의 지하층수와 층별 개요 지하층 번호 최대값이 상이한 경우 데이터 검증(일관성
19		(건폐율) 일반건축물대장 및 표제부 건폐율 계산의 정확성 → 소수점 이하 처리에 대한 명확한 법규 처리 없어 반올림 자리수가 불분명함(숫점 이하 절사처리)
20		(용적률) 일반건축물대장 및 표제부 용적율 계산의 정확성
21		(연면적) 일반건축물대장 및 표제부의 연면적 층별 개요 면적의 합이 상이한 데이터 검증(일관성)
22		(사용승인일, 허가일) 일반건축물대장 및 표제부 내 사용승인일 보다 큰 허가일 검증(정확성)
23		(허가일, 착공일) 일반건축물대장 및 표제부 내 착공일 보다 큰 허가일 검증(정확성)
24		(세부용도코드) 층별 개요의 세부용도코드 정확성
25		(구조코드) 일반건축물대장 및 표제부의 구조코드 입력 정확성
26		(구조코드) 층별 개요의 구조코드 입력 정확성
27		(주부속구분코드) 일반건축물대장 및 표제부 내 주부속구분코드 입력 누락(완전성)
28		(세대수+호수) 표제부 내 세대수와 호수의 합이 전유부의 수와 같아야 함(일관성)
29		(전유부면적) 전유 부분 면적의 미입력(정확성)
30		(세부용도코드) 전유부의 세부용도코드 오류(정확성)
31		(구조코드) 전유부의 구조코드 오류(정확성)
32		(호명칭) 동일 표제부 내 호명칭 중복 오류(정확성)
33		(총괄표제부 누락) 동일 대지위치에 2개 이상 주건축물 존재 시 총괄표제부 생성되지 않은 데이터 오류 검증(정확성)
34		(소유자 중복) 소유자정보 일반건축물 및 전유부의 소유권변동에 대한 동일 소유자 등록정보 검증(정확성)
35		(표제부 연결 누락) 건축물대장 기본개요에 전유부 중 상위 표제부와 연결되지 않은 데이터 오류 검증(정확성)
36		(표제부 누락) 전유부는 존재하나 표제부가 생성되어 있지 않는 데이터 오류 검증(정확성)
37		(표제부 누락) 전유부가 존재하지 않은 표제부 데이터 오류 검증(정확성)
38		(표제부 누락) 전유부가 존재하는 일반건축물데이터 오류 검증(정확성)
39		(변동원인 중복) 각 대장별 동일한 변동원인이 있는 경우 데이터 오류 검증(정확성)
40	건축인허가	(대지면적) 대지면적이 건축면적 보다 작은 경우의 데이터 오류 점증
41		(연면적) 건축허가대장의 연면적과 동별개요의 연면적 합이 다른 데이터 오류 검증(일관성)
42		(연면적) 동별 연면적이 층별 바닦면적의 합과 다른 데이터 오류 검증
43		(주차대수) 건축허가대장의 총주차대수와 주차장 테이블의 주차대수 합이 다른 데이터 오류 검증
44		(허가일) 건축허가대장 내 허가일자의 유효성 검증

순번	구분	업무규칙명
45	대장	(건폐율) 건축허가대장 건폐율 계산을 정확성
46		(용적률) 건축허가대장 용적률 계산의 정확성
47		(주용도코드) 건축허가대장 주용도코드 입력의 정확성
48		(구조코드) 건축허가대장 동별 구조코드 입력의 정확성
49		(건축면적) 건축허가대장의 건축면적과 동별 건축면적 합계가 상이한 데이터 오류 검증
50		(용적률산정연면적) 건축허가대장의 용적률산정연면적과 동별 용적률산정연면적의 합계가 상이한 데이터 오류 검증(일관성)
51		(용적률산정연면적) 동별 용적률산정연면적과 층별 면적의 합계가 상이한 데이터 오류 검증(일관성)
52		(연면적) 건축허가대장의 연면적과 동별 연면적의 합계가 상이한 데이터 오류 검증(일관성)
53		(층수) 동별개요의 지상층수와 층별 개요 지상층 번호 최대값이 상이한 데이터 오류 검증(일관성)
54		(지하층수) 동별개요의 지하층수와 층별 개요 지하층 번호 최대값이 상이한 데이터 오류 검증(일관성)
55		(가구수) 허가대장 기본개요의 가구수와 동별 개요 가구수 합이 상이한 데이터 오류 검증(일관성)
56		(세대수) 허가대장 기본개요의 세대수와 동별 개요 세대수 합이 상이한 데이터 오류 검증(일관성)
57		(호수) 허가대장 기본개요의 호수와 동별 개요 호수 합이 상이한 데이터 오류 검증(일관성)
58		(부속건축물) 건축허가대장 기본개요의 부속건축물 면적과 동별 개요의 부속건축물 연면적의 합이 상이한 데이터 오류 검증(일관성)
59		(가설건축물) 가설건축물 구조코드 입력 오류 검증(정확성)
60		(가설건축물) 가설건축물 용도코드 입력 오류 검증(정확성)
61	주택인허가대장	(대지면적) 대지면적이 건축면적 보다 작은 경우의 데이터 오류 검증
62		(연면적) 주택허가대장의 연면적과 동별개요의 연면적 합이 다른 데이터 오류 검증(일관성)
63		(연면적) 동별 연면적이 층별 연면적의 합과 다른 데이터 오류 검증
64		(주차대수) 주택허가대장의 총주차대수와 주차장 테이블의 주차대수 합이 다른 데이터 오류 검증(일관성)
65		(허가일) 주택허가대장 내 허가일자의 유효성 검증
66		(건폐율) 주택허가대장 건폐율 계산의 정확성
67		(용적률) 주택허가대장 용적률 계산의 정확성
68		(주용도코드) 주택허가대장 주용도코드 입력의 정확성
69		(구조) 구조코드 입력의 정확성
70		(건축면적) 주택허가대장의 건축면적과 동별 건축면적 합계가 상이한 데이터 오류 검증(일관성)
71		(용적률산정연면적) 주택허가대장의 용적률산정연면적과 동별 용적률산정연면적 합계가 상이한 데이터 오류 검증(일관성)
72		(용적률산정면적) 동별 용적률산정연면적과 층별 면적의 합계가 상이한 데이터 오류 검증(일관성)
73		(층수) 동별개요 지상층수와 층별 개요 지상층 번호 최대값이 상이한 데이터 오류 검증(일관성)
74		(지하층수) 동별개요 지하층수와 층별 개요 지하층 번호 최대값이 상이한 데이터 오류 검증(일관성)
75		(가구수) 허가대장 기본개요의 가구수와 동별 개요 가구수 합이 상이한 데이터 오류 검증(일관성)
76		(세대수) 허가대장 기본개요의 세대수와 동별 개요 세대수 합이 상이한 데이터 오류 검증(일관성)
77		(호수) 허가대장 기본개요의 호수와 동별 개요 호수 합이 상이한 데이터 오류 검증(일관성)
78		(부속건축물) 주택허가대장 기본개요의 부속건축물 면적과 동별 개요의 부속건축물 연면적의 합이 상이한 데이터 오류 검증(일관성)
79		소유자의 지분 점검
80		건축물대장의 사용승인일 누락 점검
81		일반건축물(표제부)의 필로티 건축물 점검
82		주택용도 건축물의 층간 높이 점검
83		일반건축물(표제부)의 층별현황 누락 점검
84		소유자 누락 점검
85		건축물대장 미조회 대상 주소 점검
86		건축물대장과 등기사항증명서간 소유권 점검

출처: 세움터 내부자료

주: 지자체 품질정비 대상은 음영처리

- 건축물 데이터 오류 현황

2022년 당시 세움터에서 산출한 86개 업무규칙에 대한 오류 검토 결과, 전체 점검 대상 건수 6억 2천만 건 중 약 855만 건에서 오류가 발견된 것으로 나타났다. 전체 오류율은 1.37%로 낮지만, 특정 업무규칙에서는 10~30%대의 높은 오류율이 나타났다. 건축물대장에서는 (순번 13) 대지면적보다 건축면적이 큰 경우가 25.02%로 가장 높게 나타났다. 이후 (순번 16) 표제부 건축면적 합계의 불일치가

16.78%였고, (순번 1~5, 20, 21) 표제부와 총괄표제부 간 건축면적, 연면적, 용적률 등의 불일치가 13~17% 수준으로 높게 나타났다. 면적·용적률 산출 과정에서 합계 불일치가 빈번한데, 이는 현행 시스템의 입력 및 산정 방식에 구조적 문제가 존재할 가능성이 있는 것으로 판단된다.

건축인허가대장에서는 (순번 43) 건축허가대장과 주차장 테이블 간의 주차대수 불일치가 21.50%로 높게 나타났으며, 이루 면적과 관련된 합계의 불일치(순번 41, 42, 49, 50, 51)가 0.5~3.6% 수준으로 나타났다. 주택인허가대장에서는 (순번 76) 기본개요와 동별개요 세대수가 상이한 경우가 30.44%로 높게 나타났다. 또한 (순번 62, 63, 70~72) 건축면적, 연면적, 용적률산정연면적 합계의 불일치가 1~1.5% 수준으로 나타났으며, (순번 86) 건축물대장과 등기사항증명서간 소유권 불일치가 7.29%, 소유권 검증(순번 78) 부속건축물 면적 물일치가 4.47%로 높게 나타났다.

[표 2-9] 2022년 86개 업무규칙 및 정비대상 항목의 오류율

순번	구분	업무규칙명	오류건수	점검대상 건수	오류율
1	건축물대장	(건축면적) 총괄표제부 건축면적과 표제부 건축면적 합계의 일관성	103,141	593,350	17.38%
2		(연면적) 총괄표제부 연면적과 표제부 연면적 합계의 일관성	80,147	593,350	13.51%
3		(용적률산정연면적) 총괄표제부 용적률산정연면적과 표제부 용적률산정연면적 합계의 일관성	78,756	593,350	13.27%
4		(건폐율) 총괄표제부 건폐율 계산의 정확성	6,006	599,790	1.00%
5		(용적률) 총괄표제부 용적율 계산의 정확성	81,647	593,341	13.76%
6		(주건축물수) 총괄표제부 주건축물 수와 표제부 주건축물수 합과의 일관성	53,987	593,350	9.10%
7		(부속건축물수) 총괄표제부 부속건축물 수와 표제부 부속건축물수 합과의 일관성	8,670	77,685	11.16%
8		(부속건축물 면적) 총괄표제부 부속건축물 면적과 표제부 부속건축물수 연면적 합과의 일관성	12,033	77,685	15.49%
9		(세대수) 총괄표제부 세대수와 표제부 세대수 합의 일관성	6,221	45,837	13.57%
10		(주용도코드) 총괄표제부 주용도코드의 정확성	42,664	599,790	7.11%
11		(사용승인일, 허가일) 총괄표제부 내 사용승인일 보다 큰 허가일 검증(정확성)	78	599,790	0.01%
12		(착공일, 허가일) 총괄표제부 내 착공일 보다 큰 허가일 검증(정확성)	1,253	599,790	0.21%
13		(대지면적) 총괄표제부 내 대지면적의 값이 건축면적보다 작은 경우의 데이터 검증(정확성)	150,276	599,790	25.05%
14		(대지면적) 일반건축물대장 및 표제부 내 대지면적의 값이 건축면적보다 작은 경우의 데이터 검증(정확성)	1,136	612,719	0.19%
15		(주용도코드) 일반건축물대장 및 표제부의 주용도코드 입력의 정확성	214,276	7,952,243	2.69%
16		(건축면적) 일반건축물대장 및 표제부의 바닥면적의 합이 가장 큰 면적과 표제부의 건축면적이 다른 데이터 검증(정확성)	1,334,538	7,952,243	16.78%
17		(지상층수) 일반건축물 및 표제부의 지상층수와 층별 개요 지상층 번호 최대값이 상이한 경우 데이터 검증(일관성)	77,890	7,952,243	0.98%
18		(지하층수) 일반건축물 및 표제부의 지하층수와 층별 개요 지하층 번호 최대값이 상이한 경우 데이터 검증(일관성	10,515	7,952,243	0.13%
19		(건폐율) 일반건축물대장 및 표제부 건폐율 계산의 정확성 →소수점 이하 처리에 대한 명확한 법규 처리 없어 반올림 자리수가 불분명함(소점 이하 절사처리)	54,486	7,324,326	0.74%
20		(용적률) 일반건축물대장 및 표제부 용적율 계산의 정확성	723,474	5,212,430	13.88%
21		(연면적) 일반건축물대장 및 표제부의 연면적 층별 개요 면적의 합이 상이한 데이터 검증(일관성)	290,813	7,952,243	3.66%
22		(사용승인일, 허가일) 일반건축물대장 및 표제부 내 사용승인일 보다 큰 허가일 검증(정확성)	24,958	7,952,243	0.31%
23		(허가일, 착공일) 일반건축물대장 및 표제부 내 착공일 보다 큰 허가일 검증(정확성)	63,393	7,952,243	0.80%
24		(세부용도코드) 층별 개요의 세부용도코드 정확성	566,758	20,632,081	2.75%
25		(구조코드) 일반건축물대장 및 표제부의 구조코드 입력 정확성	53,350	7,952,243	0.67%
26		(구조코드) 층별 개요의 구조코드 입력 정확성	24,918	20,632,081	0.12%
27		(주부속구분코드) 일반건축물대장 및 표제부 내 주부속구분코드 입력 누락(완전성)	794	7,952,243	0.01%
28		(세대수+호수) 표제부 내 세대수와 호수의 합이 전유부의 수와 같아야 함(일관성)	50,381	527,287	9.55%
29		(전유부면적) 전유 부분 면적의 미입력(정확성)	6,338	17,868,797	0.04%
30		(세부용도코드) 전유부의 세부용도코드 오류(정확성)	62,279	101,131,720	0.06%
31		(구조코드) 전유부의 구조코드 오류(정확성)	14,089	17,868,797	0.08%
32		(호명칭) 동일 표제부 내 호명칭 중복 오류(정확성)	8,343	17,847,974	0.05%

순번	구분	업무규칙명	오류건수	점검대상 건수	오류율
33		(총괄표제부 누락) 동일 대지위치에 2개 이상 주건축물 존재 시 총괄표제부 생성되지 않은 데이터 오류 검증(정확성)	326,740	2,110,903	15.48%
34		(소유자 중복) 소유자정보 일반건축물 및 전유부의 소유권변동에 대한 동일 소유자 등록정보 검증(정확성)	679,099	28,445,882	2.39%
35		(표제부 연결 누락) 건축물대장 기본개요에 전유부 중 상위 표제부와 연결되지 않은 데이터 오류 검증(정확성)	2,501	17,868,797	0.01%
36		(표제부 누락) 전유부는 존재하나 표제부가 생성되어 있지 않는 데이터 오류 검증(정확성)	6,410	17,869,673	0.04%
37		(표제부 누락) 전유부가 존재하지 않은 표제부 데이터 오류 검증(정확성)	3,743	527,357	0.71%
38		(표제부 누락) 전유부가 존재하는 일반건축물데이터 오류 검증(정확성)	357	17,869,673	0.00%
39		(변동원인 중복) 각 대장별 동일한 변동원인이 있는 경우 데이터 오류 검증(정확성)	94,473	41,434,613	0.23%
40	건축인허가대장	(대지면적) 대지면적이 건축면적 보다 작은 경우의 데이터 오류 점증	362	2,564,065	0.01%
41		(연면적) 건축허가대장의 연면적과 동별개요의 연면적 합이 다른 데이터 오류 검증(일관성)	13,753	2,564,065	0.54%
42		(연면적) 동별 연면적이 층별 바닥면적의 합과 다른 데이터 오류 검증	33,038	3,233,609	1.02%
43		(주차대수) 건축허가대장의 총주차대수와 주차장 테이블의 주차대수 합이 다른 데이터 오류 검증	551,217	2,564,065	21.50%
44		(허가일) 건축허가대장 내 허가일자의 유효성 검증	501	3,907,109	0.01%
45		(건폐율) 건축허가대장 건폐율 계산을 정확성	7,638	2,564,065	0.30%
46		(용적률) 건축허가대장 용적률 계산의 정확성	8,329	2,564,065	0.32%
47		(주용도코드) 건축허가대장 주용도코드 입력의 정확성	23,921	2,564,065	0.93%
48		(구조코드) 건축허가대장 동별 구조코드 입력의 정확성	824	3,233,609	0.03%
49		(건축면적) 건축허가대장의 건축면적과 동별 건축면적 합계가 상이한 데이터 오류 검증	17,459	2,564,065	0.68%
50		(용적률산정연면적) 건축허가대장의 용적률산정연면적과 동별 용적률산정연면적의 합계가 상이한 데이터 오류 검증(일관성)	19,753	2,564,065	0.77%
51		(용적률산정연면적) 동별 용적률산정연면적과 층별 면적의 합계가 상이한 데이터 오류 검증(일관성)	117,094	3,233,609	3.62%
52		(연면적) 건축허가대장의 연면적과 동별 연면적의 합계가 상이한 데이터 오류 검증(일관성)	13,753	2,564,065	0.54%
53		(층수) 동별개요의 지상층수와 층별 개요 지상층 번호 최대값이 상이한 데이터 오류 검증(일관성)	3,878	3,233,609	0.12%
54		(지하층수) 동별개요의 지하층수와 층별 개요 지하층 번호 최대값이 상이한 데이터 오류 검증(일관성)	808	3,233,609	0.02%
55		(가구수) 허가대장 기본개요의 가구수와 동별 개요 가구수 합이 상이한 데이터 오류 검증(일관성)	18,687	2,564,065	0.73%
56		(세대수) 허가대장 기본개요의 세대수와 동별 개요 세대수 합이 상이한 데이터 오류 검증(일관성)	14,941	2,564,065	0.58%
57		(호수) 허가대장 기본개요의 호수와 동별 개요 호수 합이 상이한 데이터 오류 검증(일관성)	16,705	2,564,065	0.65%
58		(부속건축물) 건축허가대장 기본개요의 부속건축물 면적과 동별 개요의 부속건축물 연면적의 합이 상이한 데이터 오류 검증(일관성)	7,354	2,564,065	0.29%
59		(가설건축물) 가설건축물 구조코드 입력 오류 검증(정확성)	400	645,247	0.06%
60		(가설건축물) 가설건축물 용도코드 입력 오류 검증(정확성)	108	450,523	0.02%
61	주택인허가대장	(대지면적) 대지면적이 건축면적 보다 작은 경우의 데이터 오류 검증	254	19,252	1.32%
62		(연면적) 주택허가대장의 연면적과 동별개요의 연면적 합이 다른 데이터 오류 검증(일관성)	230	19,252	1.19%
63		(연면적) 동별 연면적이 층별 연면적의 합과 다른 데이터 오류 검증	1,123	264,712	0.42%
64		(주차대수) 주택허가대장의 총주차대수와 주차장 테이블의 주차대수 합이 다른 데이터 오류 검증(일관성)	252	19,252	1.31%
65		(허가일) 주택허가대장 내 허가일자의 유효성 검증	52	19,252	0.27%
66		(건폐율) 주택허가대장 건폐율 계산의 정확성	88	19,252	0.46%
67		(용적률) 주택허가대장 용적률 계산의 정확성	96	19,252	0.50%
68		(주용도코드) 주택허가대장 주용도코드 입력의 정확성	30	19,252	0.16%
69		(구조) 구조코드 입력의 정확성	76	264,712	0.03%
70		(건축면적) 주택허가대장의 건축면적과 동별 건축면적 합계가 상이한 데이터 오류 검증(일관성)	290	19,252	1.51%
71		(용적률산정연면적) 주택허가대장의 용적률산정연면적과 동별 용적률산정연면적 합계가 상이한 데이터 오류 검증(일관성)	205	19,252	1.06%
72		(용적률산정면적) 동별 용적률산정연면적과 층별 면적의 합계가 상이한 데이터 오류 검증(일관성)	2,720	264,712	1.03%
73		(층수) 동별개요 지상층수와 층별 개요 지상층 번호 최대값이 상이한 데이터 오류 검증(일관성)	1,461	264,712	0.55%
74		(지하층수) 동별개요 지하층수와 층별 개요 지하층 번호 최대값이 상이한 데이터 오류 검증(일관성)	2,436	264,712	0.92%
75		(가구수) 허가대장 기본개요의 가구수와 동별 개요 가구수 합이 상이한 데이터 오류 검증(일관성)	322	19,252	1.67%
76		(세대수) 허가대장 기본개요의 세대수와 동별 개요 세대수 합이 상이한 데이터 오류 검증(일관성)	5,861	19,252	30.44%
77		(호수) 허가대장 기본개요의 호수와 동별 개요 호수 합이 상이한 데이터 오류 검증(일관성)	57	19,252	0.30%

순번	구분	업무규칙명	오류건수	점검대상 건수	오류율
78		(부속건축물) 주택허가대장 기본개요의 부속건축물 면적과 동별 개요의 부속건축물 연면적의 합이 상이한 데이터 오류 검증(일관성)	861	19,252	4.47%
79		소유자의 지분 점검	391,988	24,089,589	1.6%
80		건축물대장의 사용승인일 누락 점검	129,650	7,306,781	1.8%
81		일반건축물(표제부)의 필로티 건축물 점검	56,715	7,306,781	0.8%
82		주택용도 건축물의 층간 높이 점검	388	33,599,870	0.0%
83		일반건축물(표제부)의 층별현황 누락 점검	1,383	7,306,781	0.0%
84		소유자 누락 점검	7,133	24,089,589	0.0%
85		건축물대장 미조회 대상 주소 점검	52	26,293,089	0.0%
86		건축물대장과 등기사항증면서간 소유권 점검	1,755,609	24,089,589	7.29%
총합계			8,550,786	623,681,869	1.37%

출처: 세움터 내부자료
주: 지자체 품질정비 대상은 음영처리

3) 건축물 데이터 활용 현황

■ 건축통계 산출

- 건축통계 개요 및 작성 방법[2)]

건축통계는「건축법」제30조에 근거하여 작성되며, 건축허가·착공·준공통계와 건축물통계로 구분된다. 건축허가·착공·준공통계는 최초 건축허가통계로 1962년부터 작성되었으며, 1975년 12월 통계작성 승인을 받아 공식 지정통계로 운영되기 시작하였다(국토교통부, 2024b, p. 3). 현재는 건축허가, 착공, 준공 등 현황을 모두 아우르는 통계로 작성되고 있다. 건축물통계는 1985년 2월 통계작성 승인을 받은 후 생산되기 시작하였으며(국토교통부, 2024c, p. 2),건축물대장에 등재된 건축물 현황을 집계한다.

「건축법」제30조는 건축통계 작성과 관련하여 허가권자가 국토교통부령으로 정하는 바에 따라 국토교통부 장관이나 시·도지사에 건축허가 현황 등을 보고하도록 하고 있다(법 제1항). 이와 관련하여 건축물착공통계조사시행규칙은 건축물의 착공통계조사 시행에 관한 사항을 규정하고 있다. 이에 따르면 시장·군수와 서울특별시장·직할시장은 착공신고서를 접수한 현황을 착공조사표에 취합하여 도지사에 송부하고, 시·도지사는 이를 취합하여 국토교통부 장관에 송부하여야 한다(규칙 제9조).

그러나 건축허가·착공·준공통계는 2001년 12월부터, 건축물통계는 2002년 1월부터 기존 수기로 작성되었던 건축통계 서식을 통한 조사를 건축행정정보시스템(AIS)에 기반한 자동 집계를 통한 보고통계로 변경 시행하면서 수기 조사는 이루어지지 않고 있으며, 현재 건축통계는 건축행정 데이터에 기반하여 작성되고 있다.

- 건축통계 작성 활용 건축물 데이터

건축물통계 및 건축허가·착공·준공 통계의 공표주기는 '건축물 통계' 8종과 '건축허가·착공·준공 통계' 3종의 경우 연도별 공표되며, '건축허가·착공·준공 통계' 6종은 월별 공표된다. 건축물 통계는 면

2) 국토교통부(2024b)와 국토교통부(2024c)를 참고하여 연구진 작성

적별, 소유구분별, 용도별, 층수별로 17개 시도 및 인구 50만 이상[3] 도시를 구분하여 제공한다. 건축허가·착공·준공 통계는 연도별, 동수별로 건축물 구조 및 용도를 구분하여 제공하며, 시도별 허가·착공·준공 현황에서는 17개 시도 및 용도별로 해당 용도의 구조 및 허가구분을 적용하여 제공한다.

건축물통계에서는 시군구코드, 연면적, 소유구분, 용도코드, 층수 등이 활용되며, 건축허가·착공·준공 통계에서는 시군구코드, 용도코드, 구조코드, 연면적, 허가구분, 동수, 허가일, 착공일, 사용승인일 등이 활용된다.

[표 2-10] 건축물 관련 통계의 공표주기 및 통계

통계명	공표주기	대분류	소분류	단위
건축물 통계 (8)	연도별	면적별 건축물현황(17개시도)/(50만 이상 도시)	1백㎡미만, 1백~2백㎡미만, 2백~3백㎡미만, 3백~5백㎡미만, 5백~1천㎡미만, 1천~3천㎡미만, 3천~1만㎡미만, 1만㎡이상	동
		소유구분별 건축물현황(17개시도)/(50만 이상 도시)	국공유, 개인, 법인, 그 외 기타	동
		용도별 건축물 현황(17개시도)/(50만 이상 도시)	주거용[1], 상업용[2], 공업용[3], 교육·사회용[4], 기타[5]	동
		층수별 건축물 현황(17개시도)/(50만 이상 도시)	1층, 2~4층, 5층, 6~10층, 11~20층, 21~30층, 31층 이상, 기타	동
건축허가·착공·준공 통계 (9)		연도별 건축착공현황	구조별(콘크리트, 철골, 철골콘크리트, 조적, 목조, 기타) 용도별(주거용, 상업용, 공업용, 교육·사회용, 기타)	동, 연면적
		연도별 건축허가현황		
		연도별 건축준공현황		
	월별	동수별 연면적별 건축착공현황		
		동수별 연면적별 건축허가현황		
		동수별 연면적별 건축준공현황		
		시도별 건축착공현황	허가구분(신축, 증축/개축/이전/대수선, 용도변경) 구조별(콘크리트, 철골, 철골콘크리트, 조적, 목조, 기타) ※ 시도별 건축물 용도별(29종/단독 및 공동주택은 상세 11종 세분)	동, 연면적
		시도별 건축허가현황		
		시도별 건축준공현황		

출처: 국토교통 통계누리, https://stat.molit.go.kr/portal/main/portalMain.do(검색일: 2025.09.16.)
주:
1) 단독주택(단독, 다중, 다가구주택 등), 공동주택(아파트, 연립, 다세대주택 등)
2) 근린생활, 판매시설, 운수시설, 업무시설, 숙박시설, 위험물저장 및 처리시설, 자동차관련시설
3) 공장
4) 문화 및 집회시설, 종교시설, 의료시설, 교육연구시설, 노유자시설, 수련시설, 운동시설, 묘지관련시설, 관광휴게시설
5) 동식물관련시설, 창고시설, 분뇨 및 쓰레기시설, 교정 및 군사시설, 방송통신시설, 발전시설

3) 2024년 말 기준 서울특별시 강서구, 강남구, 송파구, 대구광역시 달서구, 인천광역시 서구, 경기도 수원시, 성남시, 안양시, 부천시, 평택시, 안산시, 고양시, 남양주시, 시흥시, 용인시, 화성시, 충청북도 청주시, 충청남도 천안시, 전라북도 전주시, 포항시, 경상남도 창원시, 김해시

건축 허가 및 착공 통계 작성 방식

구분	작성 방법
용도별 연면적 및 동수	(연면적) 건축물 각 층 정보에 입력된 주용도 코드별 면적 합산 * 1층~2층 주용도코드가 제2종근린생활시설이고, 3층~5층 주용도코드가 다세대주택인 경우, 1층~2층 면적은 제2종근린생활시설로, 3층~5층 면적은 다세대 주택 면적으로 집계 (동수) 건축허가신청서 및 착공신청서에 입력된 동별개요 1건을 1개동으로 집계하며, 건축물 각 층별 주용도코드별 면적 합 중 가장 넓은 면적의 용도를 동 대표용도로 산정하여 용도별 동수 집계 * 1층~2층 주용도 코드가 제2종근린생활시설이고, 3층~5층 주용도 코드가 다세대주택인 경우, 가장 넓은 면적을 차지하는 다세대주택 1동으로 집계
구조별, 사업주체별, 건축구분별 연면적 및 동수	(연면적) 건축허가신청서 및 착공신고서의 건축구분, 건축주, 동 주구조코드별 층면적 합산 (동수) 건축허가신청서 및 착공신고서의 건축구분, 건축주, 동 주구조코드별 동수 집계
허가 및 착공 및 준공 기간	(허가) 건축허가일자를 기준으로 작성 (착공) 건축 관리진행상황 중 업무단계 구분이 '착공'인 경우의 건축착공일자를 기준으로 작성 (준공) 공사가 실제로 완공된 기준으로 작성
기타 세부 작성 기준	전국 자치단체 행정자료를 집계하여 동수 및 연면적 산출 옥탑 면적 제외 부속건축물 면적 및 동수 포함

출처 : 국토교통부(2022b).『건축허가 및 착공통계』통계정보보고서. pp.15-16; 국토교통 통계누리, https://stat.molit.go.kr/ (검색일: 2025.09.17.); 조영진 외(2023), 데이터 기반 정책을 위한 건축물 생산량 지수 개발 연구, 건축공간연구원, p.19 재인용. 바탕으로 연구진 작성

■ 건축물 데이터 활용 연구

- 개요

건축물 데이터는 건축통계 작성에 그치지 않고, 공공 및 민간 영역에서 폭넓게 활용되고 있다. 데이터가 가진 세부성과 공간성을 바탕으로 재해·재난 관리, 범죄예방, 위법 건축물 관리, 빈 건축물 추정, 도시 환경 분석 등 다양한 분야에서 응용 가능성이 확인되고 있다. 본 연구에서는 건축물 데이터를 활용한 연구 사례를 검토하여 다양한 주제의 연구에서 건축물 데이터가 어떻게 활용되고 있는지 파악하고, 활용도가 높은 건축물 데이터를 도출하고자 하였다. 이를 통하여 건축물 데이터가 공공데이터로서 갖는 가치를 확인하고, 정책 수립, 연구 개발, 민간 서비스 등 다방면에서 활용될 수 있는 건축물 데이터에 대한 품질 고도화 방향 설정이 이루어질 수 있도록 하였다.

- 재해 취약 지하층 주택 현황 분석[4)]
  - 활용 데이터

안의순·박종훈(2023, p54-58)은 건축물 현황에 대한 정보를 기재한 건축물대장을 통하여 재해 취약 지하층 주택 판별에 필요한 데이터를 수집하였다. 분석 과정의 활용 데이터는 다음과 같다. 첫째, 지하층 주택 포함 여부 판단을 위해 건축물대장 표제부와 건축물대장 층별개요 데이터의 '주_용도_코드', '주_용도_코드명', '기타_용도'를 활용하였다. 둘째, 사용승인일 기준 노후도 판별을 위해 건축물대장 표제부 데이터의 '사용승인_일'을 활용하였다. 셋째, 재해 취약 구조 여부를 판단하기 위해 표제부의

4) 안의순, 박종훈. 2023. 건축행정 데이터 기반 재해 취약 주택 현황 분석. 건축공간연구원

'구조_코드'를 활용하였다. 넷째, 필로티 주차장이 포함된 경우 제외를 위해 층별 용도 데이터의 '기타_용도' 기재 내용을 확인하였다. 다섯째, 전국 재해 취약 지하층 주택 목록을 추출하고, 이를 시·도별, 시·군·구별, 읍·면·동별 동수로 집계하기위해 건축물대장 기본개요의 '시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지'를 활용하였다.

- 데이터 오류

안의순·박종훈(2023, p.57)은 분석과정에서 문제가 발생한 데이터 오류를 제시하였다. 건축물대장 대용량 데이터에 포함된 '주_용도_코드'는 행정표준코드로 정의되어 있다. 주 용도 코드는 건축물 용도를 5자리 숫자로 표현한다. 그러나 일부 대장에 대한 데이터에서는 영문자 Z와 숫자 4자리로 기존 4자리 코드를 포함하고 있거나, "창고", "사찰", "축사", "학교" 등 다른 값이 포함된 것을 확인하였다.

[표 2-11] 지하층 주택 현황 분석을 위해 사용된 건축물대장 테이블 및 컬럼 정보

테이블 정보	컬럼 정보
건축물대장 기본개요	'시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지'
건축물대장 표제부	**'주_용도_코드'**, '주_용도_코드명', '기타_용도', '사용승인_일', '구조_코드'
건축물대장 층별개요	'주_용도_코드', '주_용도_코드명', '기타_용도',

출처: 안의순·박종훈(2023), 건축행정 데이터 기반 재해 취약 지하층 주택 현황 분석, p.54~58.
주: 데이터 오류를 제시한 컬럼은 볼드처리

- 건축물 화재 발생 예측 모델[5)]

  - 활용 데이터

조영진 외(2022, p.58)는 건축물 화재 예측 모델 개발을 위해, 서울시 건축물을 대상으로 화재 발생 데이터와 범죄 발생 데이터를 연계·결합하였다. 화재 발생 데이터와 건축행정 데이터 연계를 위해 주소정보로 PNU 코드[6)]를 생성하여 연계키로 활용하였다. 또한, 대지 내 건물동의 구분을 위하여 화재 발생 건축물의 동 명칭을 확인하였다. 활용된 컬럼은 '시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지', '동_명칭'이다. 이후 지오코딩을 통한 공간정보화 이후 화재 관련 변수 3종, 범죄 관련 변수 6종, 건축물 관련 변수 6종을 독립변수로 구축하였다. 건축물 관련 변수 6종의 컬럼은 '주_용도_코드', '구조_코드', '연면적', '용적률', '위반_건축물_여부', '사용승인_일'이다.

- 데이터 오류

조영진 외(2022, p.63-64)는 화재 데이터와 건축물대장 데이터 연계과정에서 주소정보의 불일치를 해결하기 위해 매칭 프로세스를 제시하였다. 첫째, 단일건물인 경우 화재데이터에 건축물대장 데이터 결합, 둘째, 집합건물에서 해당 주소의 건물이 한 동인 경우 데이터 결합, 셋째, 집합건물/건축물대장의DONG_NM(동이름)이 화재 세부주소와 일치하는 경우 해당 데이터 결합, 넷째, 부수적으로 건축물대장 상 동 이름이 '제00동' 형식, 'A동'을'에이동'으로 기재한 경우 등 의미에 부합하는 해당 동을 찾아 결합, 다섯째, 화재데이터 세부주소에서 동 이름을 숫자만 기재한 경우 해당 동이 있는지 찾아서 결

5) 조영진, 허한결, 안의순, 류수연, 현태환. 2022. 빅데이터 기반 건축물 화재 예측 모델 개발 연구. 건축공간연구원

6) 19자리 숫자로 구성된 코드로 00(시·도)000(시·군·구)000(읍·면·동)00(리)0(대장구분)0000(본번)0000(부번)으로 이루어졌다.

합, 여섯째, 연결되지 않을시 화재발생 당해연도 건축물 연결, 일곱째, 비매칭결과 에러 처리코드의 순이다(조영진 외, 2022, p.63-64). 프로세스 진행 결과, 주소가 일치하는 건축물들이 있지만 세부주소 비일치 경우 직접 검토 후 수정하는 과정을 거쳤다.

연구의 결론에서는 건축행정데이터와 재난·재해데이터의 연계 시 높은 수준의 무결성을 달성하는 것이 시급함을 제시하였다. 현행 건축행정 데이터는 건축물 고유번호(PK번호)를 중심으로 구성되어 있으나, 향후 재난·안전 관련 원시데이터와의 연계를 위해서는 건물ID 기반 체계로 전환할 필요가 있다. 특히 공공데이터를 건축물 단위로 제공함으로써, 데이터 활용의 효율성과 정책적 효과를 동시에 높일 수 있음을 제안하였다.

[표 2-12] 건축물 화재 예측 모델 개발을 위해 사용된 건축물대장 테이블 및 컬럼 정보

테이블 정보	컬럼 정보
건축물대장 기본개요	'시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지'
건축물대장 표제부	**'동_명칭'**, '주_용도_코드', '구조_코드', 연면적', '용적률', '위반_건축물_여부', '사용승인_일'

출처: 조영진 외(2022), 빅데이터 기반 건축물화재 예측모델 개발 연구, p.52~74.
주:
1) 분석과정에서 GIS건물통합정보를 활용, GIS건물통합정보는 건축행정 정보를 기반으로 가공된 정보임으로 원자료로 볼 수 있는 건축행정 정보의 건축물대장을 기준으로 테이블과 컬럼 정보를 기재
2) 데이터 오류를 제시한 컬럼은 볼드처리

- 건축물 화재 및 홍수 리스크 분석 모델 개발[7)]

  - 활용 데이터

조영진 외(2023, p.56-79)는 건축물과 화재 데이터 연계를 위해 조영진 외(2022, p.58)에서 활용한 방법을 준용하였다. 사용된 건축물대장의 컬럼은 '시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지', '동_명칭'이다. 홍수데이터와 건축물 연계 과정에서는 biz-gis의 지오코딩툴을 이용하였으며, 건축물대장의 지번주소를 사용하여, 화재 데이터 연계 과정에서 활용한 컬럼과 동일하다.

건축물 화재 리스크 분석 모델 개발 과정에서는 입력 데이터 선정을 위해 건축물대장 표제부 테이블의 '주_용도_코드'(주거용도 수 및 상업용도 수) '사용승인_일'(노후도), '지상_층_수', '지하_층_수', '용적률', '승용_승강기_수', '비상용_승강기_수' 컬럼을 활용하였고, 건축물 홍수 리스크 분석보델 개발 과정에서는 건축물대장 총괄표제부 테이블의 '주_건축물_수', 건축물대장 표제부 테이블의 '연면적', '건축_면적', '대지_면적', '건폐율', '용적률', '용적률_산정_연면적', '높이', '지상_층_수', '지하_층_수', '주_용도_코드', '구조_코드'를 활용하였다.

  - 데이터 오류

조영진 외(2023, pp.58, 87)는 화재 데이터와 건축물대장 데이터 연계과정에서 주소정보의 불일치를 해결하기 위해 조영진 외(2022, p.63-64)의 데이터 연계 프로세싱 방안을 준용하였다. 집합건물/건축물대장의 DONG_NM(동 이름)이 화재 세부주소와 일치하는 경우 해당 데이터를 연계·결합하고, 건

7) 조영진, 허한결, 송유미, 현태환. 2023. 빅데이터 기반 건축물 화재 및 홍수 리스크 분석 모델 개발 연구. 건축공간연구원

축물대장 상 동 이름이 의미상 일치하는 경우('A동'을 '에이동'으로 기재한 경우 등) 수작업을 통하여 결합하였다.

입력 데이터의 전처리를 위해 건축물대장 총괄표제부 테이블의 '주_건축물_수' 컬럼과, 건축물대장 표제부 테이블의 '연면적', '건축_면적', '대지_면적', '건폐율', '용적률', '용적률_산정_연면적', '높이', '지상_층_수', '지하_층_수' 컬럼은 Min-max scaling을 활용하였고, '주_용도_코드', '구조_코드'는 One-hot encoding을 활용하였다.

[표 2-13] 화재 및 홍수 리스크 분석 모델 개발을 위해 사용된 건축물대장 테이블 및 컬럼 정보

테이블 정보	컬럼 정보
건축물대장 기본개요	'시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지'
건축물대장 총괄표제부	**'주_건축물_수'**
건축물대장 표제부	**'동_명칭'**, '사용승인_일', '승용_승강기_수', '비상용_승강기_수', **'연면적', '건축_면적', '대지_면적', '건폐율', '용적률', '용적률_산정_연면적', '높이', '지상_층_수', '지하_층_수', '주_용도_코드', '구조_코드'**

출처: 조영진 외(2023). 빅데이터 기반 건축물 화재 및 홍수 리스크 분석 모델 개발 연구. 건축공간연구원, p.56~79.
주: 데이터 오류를 제시한 컬럼은 볼드처리

- 범죄예방 환경설계 고도화 및 인증제도 개선[8)]
  - 활용 데이터

조영진 외(2024, p.92)는 범죄 예방과 건축·도시 물리환경 특성 사이의 관계 분석을 위해 필요한 물리환경 변수를 구축하였고, 그 중 건축물과 관련한 변수는 GIS건물통합정보를 활용하였다. 활용된 정보는 건축물대장 표제부의 '경과연수', '높이', '건폐율', '용적율', '주_용도_코드' 컬럼이다.

  - 데이터 오류

조영진 외(2024, p.103-105)는 통계 및 기계학습 분석을 위한 데이터셋 처리를 위해 종속변수 프로세싱, 변수 스케일 조정, 이상치 처리[9)], 결측치 제거, 다중공선성과 상관계수 낮은 변수 선택의 과정을 거쳤다. 그중 이상치 처리 방안은 z-score를 사용하여 이상치를 탐지하였으며, z-score가 ±3 시그마 범위[10)]를 벗어난 데이터를 제거하였다. 결측치 제거 방안은 대표적으로 평균값, 중앙값, 최빈값 등을 활용한 대체가 있음을 제시하며, 연구에서는 총 3456개의 그리드 중 인구가 0인 그리드와, 중복된 그리드, 건물이 없는 그리드를 제거한 후 2713개의 그리드를 사용하였다.

8) 조영진, 안의순. 박성남, 고영호, 권오규, 임보영, 임리사, 김유진, 이정현, 2024. 범죄예방 환경설계(CPTED) 고도화 및 인증제도 개선 방향. 건축공간연구원

9) 이상치의 발생에는 여러 요소들이 있으며, 데이터 수집과정에서 생긴 오류는 그중 하나이다(James et al., 2013, pp.96-97; 조영진 외, 2024, p.104 재인용)

10) 이상치 처리에 주로 사용되는 방안은 z-score 혹은 IQR이며, z-score의 경우 데이터셋의 z-score 절대값 3 시그마 범위를 벗어나는 데이터는 이상치로 간주하는것이 보편적인 방법이다(Iglewicz, B., & Hoaglin, D. C., 1993,pp.10; 조영진 외, 2024, p.104 재인용)

[표 2-14] 범죄예방 환경설계 고도화 및 인증제도 개선을 위해 사용된 건축물대장 테이블 및 컬럼 정보

테이블 정보	컬럼 정보
건축물대장 표제부	**'경과연수', '높이', '건폐율', '용적율'**, 주_용도_코드'

출처: 조영진 외(2024), 범죄예방 환경설계(CPTED) 고도화 및 인증제도 개선 방향, p.92.
주:
1) 분석과정에서 GIS건물통합정보를 활용, GIS건물통합정보는 건축행정 정보를 기반으로 가공된 정보임으로 원자료로 볼 수 있는 건축행정 정보의 건축물대장을 기준으로 테이블과 컬럼 정보를 기재
2) 데이터 오류를 제시한 컬럼은 볼드처리

- 위반 건축물 통계분석 및 모니터링[11)]
  - 활용 데이터

조영진 외(2024, p.231-246)는 위반 건축물모니터링을 위해 건축물대장 기본개요 테이블의 '위반_건축물_여부' 컬럼을 활용해 위반 건축물 정보를 구축하였다. 이후 통계분석을 위해 건축물대장 기본개요 '시군구_코드'와 표제부 테이블의 '주_용도_코드', '연면적', '사용승인_일', '지상_층_수', '구조_코드' 컬럼을 활용하였다.

[표 2-15] 위반 건축물 통계분석 및 모니터링을 위해 사용된 건축물대장 테이블 및 컬럼 정보

테이블 정보	컬럼 정보
건축물대장 기본개요	'위반_건축물_여부', '시군구_코드'
건축물대장 표제부	'주_용도_코드', '연면적', '사용승인_일', '지상_층_수', '구조_코드'

출처: 조영진 외(2024), 2024년 건축물관리지원센터 업무 위탁, 국토교통부·건축공간연구원, p.233-246.

- 빈 건축물 추정[12)]
  - 활용 데이터

조영진 외(2024, p.191)는 빈 건축물 추정과정에서 사업자등록 정보와 매칭을 위해 건축물대장 기본개요 테이블의 '시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지' 컬럼을 활용하였고, 전기에너지 정보와 매칭과정에서는 '건축물대장_PK' 컬럼을 활용하였다. 이후 제외 Filter 적용을 위해 건축물대장 표제부 테이블의 '주_부속_구분_코드', '사용승인_일', '주_용도_코드'를 활용하여 주거용도 및 사용승인 5년 이내 및 안보시설을 제외하였다.

추정 이후 모수 저감을 위해 폐쇄말소 대장과 데이터셋을 결합하였으며, 폐쇄말소대장 기본개요 테이블의 '폐쇄말소_구분_코드', '폐쇄말소_구분_코드_명', '대장_종류_코드', '폐쇄말소_구분_코드_명'을 활용하였다. 그리고 동식물관련시설의 축사, 재배사 등의 제외를 위해 건축물대장 표제부 테이블의 '대표용도' 컬럼을 활용하였다.

11) 조영진, 유광흠, 박종훈, 안의순, 허한결, 현태환, 송유미, 김효정, 남기천, 김가해, 박미래, 2024, 2024년 건축물관리지원센터 업무 위탁, 국토교통부·건축공간연구원

12) 조영진, 유광흠, 박종훈, 안의순, 허한결, 현태환, 송유미, 김효정, 남기천, 김가해, 박미래, 2024, 2024년 건축물관리지원센터 업무 위탁, 국토교통부·건축공간연구원

- 데이터 오류

조영진 외(2024, p.196)은 건축물대장과 전기에너지 정보 매칭과정에서 건축물대장 표제부 테이블의 '연면적' 컬럼에 정보가 없는 오류를 제시하였다. 필지 단위의 전기에너지 사용량 정보를 건축물 동 단위로 배분하는 과정에서 '연면적' 컬럼에 정보가 없는 경우 미확정 정보로 분류하였다.

[표 2-16] 빈 건축물 추정을 위해 사용된 건축물대장 테이블 및 컬럼 정보

테이블 정보	컬럼 정보
건축물대장 기본개요	'시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지', '안보_시설_구분_코드'
건축물대장 표제부	'건축물대장_PK', '주_부속_구분_코드', '사용승인_일', '주_용도_코드', '대표용도', **'연면적'**
폐쇄말소대장 기본개요	'폐쇄말소_구분_코드', '폐쇄말소_구분_코드_명', '대장_종류_코드', '폐쇄말소_구분_코드_명'

출처: 조영진 외(2024), 2024년 건축물관리지원센터 업무 위탁, 국토교통부·건축공간연구원, p.191.
주: 데이터 오류를 제시한 컬럼은 볼드처리

- 건축물 생산량 지수 개발[13)]

- 활용 데이터

조영진 외(2023, p.75-77)는 건축물 생산량 지수는 개발을 위해 건축행정시스템을 통하여 수집되는 '건축 허가', '건축 착공', '건축 준공'의 세 가지 항목으로 작성하였다. 각 작성단위는 전국 총계와 함께 구조, 용도, 지역 등 세분류 지수가 함께 산출된다. '건축 허가'와 '건축 착공' 지수는 모든 세분류를 반영하여 작성하였으며, '건축 준공' 지수는 구조 분류를 제외한 23종으로 산출하였다. 이 연구에서는 연면적 데이터를 수집하여 연면적 단위로 집계하여 활용하였다.

[표 2-17] 건축물 생산량 지수 작성 구분

작성단위 구분	작성단위 세부구분		작성 세부 항목
건축 허가(27)	전국(10)	총계(1)	전국 총계
		구조(4)	철골철근콘크리트, 조적, 목조, 기타
		용도(5)	주거, 상업, 공업, 교육 및 사회, 기타
	시도(17)	시(8)	서울, 부산, 대구, 인천, 광주, 대전, 울산, 세종
		도(9)	경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주
건축 착공(27)	전국(10)	총계(1)	전국 총계
		구조(4)	철골철근콘크리트, 조적, 목조, 기타
		용도(5)	주거, 상업, 공업, 교육 및 사회, 기타
	시도(17)	시(8)	서울, 부산, 대구, 인천, 광주, 대전, 울산, 세종
		도(9)	경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주
건축 준공(23)	전국(6)	총계(1)	전국 총계
		용도(5)	주거, 상업, 공업, 교육 및 사회, 기타
	시도(17)	시(8)	서울, 부산, 대구, 인천, 광주, 대전, 울산, 세종
		도(9)	경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주

출처: 조영진 외(2023). 데이터 기반 정책을 위한 건축물 생산량 지수 개발 연구, 건축공간연구원, p.76.

13) 조영진, 허한결, 안의순, 송유미, 2023. 데이터 기반 정책을 위한 건축물 생산량 지수 개발 연구. 건축공간연구원

생산량 지수 개발에 활용된 건축물 데이터는 건축인허가 기본개요 테이블의 '건축_허가_일', '실제_착공_일', '사용승인_일' 컬럼, 건축인허가 동별개요 테이블의 '시군구_코드', 건축인허가 층별개요 테이블의 '주_용도_코드', '구조_코드', '연면적' 컬럼이 활용되었다.

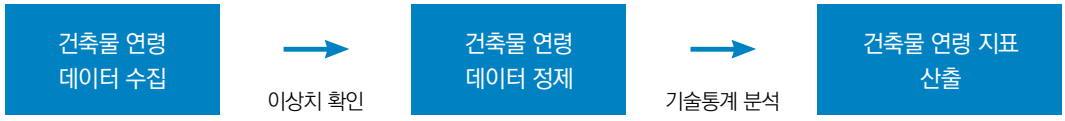
[표 2-18] 건축물 생산량 지수 개발을 위해 사용된 건축물대장 테이블 및 컬럼 정보

테이블 정보	컬럼 정보
건축인허가 기본개요	'건축_허가_일', '실제_착공_일', '사용승인_일'
건축인허가 동별개요	'시군구_코드'
건축인허가 층별개요	'주_용도_코드', '구조_코드', '연면적'

출처: 조영진 외(2023), 데이터 기반 정책을 위한 건축물 생산량 지수 개발 연구, 건축공간연구원, p.19, 76-77.

- 건축물 연령지표 개발[14)]

송유미 외(2024, p.43)는 건축물 연령 지표 개발을 위해, 건축물의 물리적 경과연수를 확인할 수 있는 데이터를 수집, 정제, 분석하였다.



[그림 2-5] 건축물 연령 지표 개발 프로세스
출처 : 송유미 외(2024, p.43)

- 활용 데이터

건축물의 연령 확인을 위해 건축물대장 표제부 테이블의 '허가_일', '착공_일', '사용승인_일' 컬럼을 검토하였고, '사용승인_일'을 건축물의 연령을 확인하는 기준일로 확정하였다. 이후 건축물 연령 데이터 추출 과정에서 건축물대장 표제부 테이블의 '주_부속_구분_코드'와 건축물대장 층별개요 테이블의 '주_부속_구분_코드'를 활용하여 부속 건축물을 제외하였고, 건축물현황 집계는 동수가 아닌 면적을 기준으로 하기 위해 건축물대장 표제부 테이블의 '주_용도_코드', '연면적' 컬럼과, 건축물대장 층별개요 테이블의 '주_용도_코드', '면적' 컬럼을 활용하였다.

- 데이터 오류

송유미 외(2024, p.46)는 건축물대장 표제부 테이블의 '사용승인_일' 컬럼을 추출하여 검토하였다. 검토 결과, 건축행정 전산화 이후 데이터 누락은 발생되지 않았으며, 문화재 등 과거에 축조되거나 건축행정 이전 시기에 지어진 건축물의 경우 사용승인일이 기록되지 않은 사례가 일부 확인되었다. 또한 사용승인일 기재 형식이 연월일 8자리 숫자가 아닌 경우도 확인되었다.

14) 송유미, 조영진, 안의순. 2024. 건축행정 데이터를 활용한 건축물 연령 지표 개발 연구. 건축공간연구원

[표 2-19] 사용승인일 기입 유형

구분		분석 및 조치
사용승인일 1자리		- 의미를 파악할 수 없는 숫자(0, 1 등)와 문자(`)
사용승인일 2자리		- '19' 가 대다수 - 1900년 ~1999년 중 사용승인일이라는 의미로 추정 - 정확한 사용승인 연도 확인 곤란
사용승인일 3자리		- 읍면지역의 노후건축물(단독, 창고, 판매시설)로 확인되며, 일부 사찰(대구 동화사 등)로 확인 - 전산화에 따른 오기이나 정확한 승인연도 추정이 곤란
사용승인일 4자리	0001 ~ 2023	- 연도로 추정 - 월일을 파악하기 곤란하여 가옥대장 등 기존 장부를 참고하거나 담당공무원이 청문 등을 통하여 사용승인연도를 추정하여 표기한 경우로 판단
	0101 ~ 1231	- 일자로 추정 - 대부분 노후 소규모 건축물(단독, 1종근생)이나 장부상 연도파악 곤란

출처 : 송유미 외(2024), 건축행정 데이터를 활용한 건축물 연령 지표 개발 연구. 건축공간연구원, p.46.

또한 건축물 연령 데이터 추출 과정에서 건축물 현황 집계에 연면적을 활용하였는데, 연면적 데이터의 오류 현황도 제시하였다. '연면적' 컬럼에는 주용도의 연면적과 부속용도의 연면적이 혼재된 경우가 존재하였다. 부속건축물 관련 정보도 총괄표제부와 각 표제부의 수치의 합이 상이한 경우가 있었다. 그 외에도 수치 입력 오류 등이 존재함을 제시하였다.

층별 면적에서도 오류가 발견되었다. 한 예로 단일 층 면적이 약 1억 9천만 ㎡인 경우가 있었다. 많은 경우 소수점 오류로 인한 것으로 판단된다. 이와 유사한 오류를 삭제하기 위하여 층별 면적이 연면적보다 큰 경우는 오기로 판단, 해당 데이터는 제외하였다.

[표 2-20] 건축물 연령지표 개발을 위해 사용된 건축물대장 테이블 및 컬럼 정보

테이블 정보	컬럼 정보
건축물대장 표제부	'허가_일', '착공_일', **'사용승인_일'**, '주_부속_구분_코드', '주_용도_코드', **'연면적'**
건축물대장 층별개요	'주_부속_구분_코드', '주_용도_코드', **'면적'**

출처: 송유미 외(2024). 건축행정 데이터를 활용한 건축물 연령 지표 개발 연구. 건축공간연구원, p.45~51.
주: 데이터 오류를 제시한 컬럼은 볼드처리

- 건축물 데이터 활용 현황 종합

건축물 데이터를 활용한 연구들은 건축물대장과 건축인허가 데이터를 중심으로 화재, 홍수, 범죄, 위반 건축물, 빈 건축물, 건축물 생산량 및 연령 분석 등 다양한 목적으로 활용되었다. 주요 활용 데이터는 건축물대장 기본개요에서 시군구 코드, 법정동 코드, 대지 구분 코드, 번·지, 위반 건축물 여부, 안보시설 구분 등, 표제부에서 주 용도 코드 및 명칭, 구조 코드, 사용승인일, 연면적, 건폐율·용적률, 층수, 대표용도 등, 건축물대장 층별개요에서 주_용도_코드, 구조_코드, 연면적, 면적 등, 총괄표제부에서 주 건축물 수, 건축인하가 자료에서 건축 허가일, 착공일, 사용승인일, 연면적 등이다.

화재 및 침수, 사업자등록 정보 등 건축물 유관 데이터와 건축물대장을 결합하기 위해 기본개요의 주소정보가 많이 활용되었고 이에 대한 매칭 프로세스가 제시되었다, 재해 취약 지하층 주택 파악 연구에서는 용도코드에 대한 오류가 제시되었고, 지수·지표 개발, 빈 건축물 추정 등에서는 면적 관련 정보에 대한 오류가 두드러졌다.

[표 2-21] 건축물 데이터 활용 현황

구분	데이터 활용 및 오류	활용 데이터
안의순·박종훈(2023), 재해 취약 지하층 주택 현황 분석	재해 취약 지하층 주택 판별 - '주_용도_코드'의 영문자 Z와 숫자 4자리로 코드 길이 변경 전 코드를 표현하거나, "창고", "사찰", "축사", "학교" 등 전혀 다른 값이 포함	- 건축물대장 기본개요: '시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지' - 건축물대장 표제부: **'주_용도_코드',** '주_용도_코드명', '기타_용도', '사용승인_일', '구조_코드' - 건축물대장 층별개요: '주_용도_코드', '주_용도_코드명', '기타_용도'
조영진 외(2022), 건축물 화재 발생 예측 모델	건축물 화재 예측 모델 개발을 위해, 서울시 건축물을 대상으로 화재 발생 데이터와 범죄 발생 데이터를 연계 및 결합 - 화재 데이터와 건축물대장 데이터 연계과정에서 주소정보의 불일치를 해결하기 위해 매칭 프로세스를 제시	- 건축물대장 기본개요: '시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지' - 건축물대장 표제부: **'동_명칭',** '주_용도_코드', '구조_코드', 연면적', '용적률', '위반_건축물_여부', '사용승인_일'
조영진 외(2023), 건축물 화재 및 홍수 리스크 분석 모델 개발	건축물과 화재 데이터 연계를 위해 조영진 외(2022, p.58)에서 활용한 방법을 준용, 홍수데이터 연계 과정에서 biz-gis의 지오코딩툴을 이용 - 데이터 연계과정에서 주소정보의 불일치를 해결하기 위해 조영진 외(2022, p.63-64)의 데이터 연계 프로세싱 방안을 준용, 데이터 전처리를 위해 Min-max scaling과 One-hot encoding을 활용	- 건축물대장 기본개요: '시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지' - 건축물대장 총괄표제부: **'주_건축물_수'** - 건축물대장 표제부: '동_명칭', '사용승인_일', '승용_승강기_수', '비상용_승강기_수', **'연면적', '건축_면적', '대지_면적', '건폐율', '용적률', '용적률_산정_연면적', '높이', '지상_층_수', '지하_층_수', '주_용도_코드', '구조_코드'**
조영진 외(2024), 범죄예방 환경설계 고도화 및 인증제도 개선	범죄 예방과 건축·도시 물리환경 특성 사이의 관계 분석을 위해 필요한 물리환경 변수를 구축 - 이상치 처리 방안은 z-score를 사용하여 이상치를 탐지하였으며, z-score가 ±3 시그마 범위를 벗어난 데이터를 제거	- 건축물대장 표제부: **'경과연수', '높이', '건폐율', '용적율',** 주_용도_코드'
조영진 외(2024), 위반 건축물 통계분석	위반 건축물모니터링을 위해 건축물대장 기본개요 테이블의 '위반_건축물_여부' 컬럼을 활용해 위반 건축물 정보를 구축	- 건축물대장 기본개요: '위반_건축물_여부', '시군구_코드' - 건축물대장 표제부: '주_용도_코드', '연면적', '사용승인_일', '지상_층_수', '구조_코드'
조영진 외(2024), 빈 건축물 추정	사업자등록 정보와 매칭을 위해 건축물대장 기본개요 테이블의 주소정보를 활용, 이후 제외 Filter 적용을 위해 건축물대장 표제부 테이블의 '주_부속_구분_코드', '사용승인_일', '주_용도_코드'를 활용하여 주거용도 및 사용승인 5년 이내 및 안보시설을 제외, 모수저감을 위해 폐쇄말소대장 및 표제부 테이블의 '대표용도'컬럼 활용 - 필지 단위의 전기에너지 사용량 정보를 건축물 동 단위로 배분하는 과정에서 '연면적' 컬럼에 정보가 없는 경우 미확정 정보로 분류	- 건축물대장 기본개요:'시군구_코드', '법정동_코드', '대지_구분_코드', '번', '지', '안보_시설_구분_코드' - 건축물대장 표제부: '건축물대장_PK', '주_부속_구분_코드', '사용승인_일', '주_용도_코드', '대표용도', **'연면적'** - 폐쇄말소대장 기본개요: '폐쇄말소_구분_코드', '폐쇄말소_구분_코드_명', '대장_종류_코드', '폐쇄말소_구분_코드_명'

구분	데이터 활용 및 오류	활용 데이터
조영진 외 (2024), 건축물 생산량 지수 개발	건축물 생산량 지수는 개발을 위해 건축행정시스템을 통하여 수집되는 '건축 허가', '건축 착공', '건축 준공'의 세 가지 항목으로 작성, 각 분류 안에는 전국 총계 지수가 포함되며, 그 외에 전국단위와 시도단위에 따라 구조, 용도, 시도 등 세분류에 따른 지수가 포함	- 건축인허가 기본개요: '건축_허가_일', '실제_착공_일', '사용승인_일' - 건축인허가 동별개요: '시군구_코드' - 건축인허가 층별개요: '주_용도_코드', '구조_코드', '연면적'
송유미 외 (2024), 건축물 연령지표 개발	건축물 연령 지표 개발을 위해, 건축물의 물리적 경과연수를 확인할 수 있는 데이터를 수집, 정제, 분석 - 건축물대장 표제부 테이블의 '사용승인_일' 컬럼을 추출하여 검토한 결과, 건축행정 전산화 이후 데이터 누락은 발생되지 않았으며, 문화재 또는 고(古)건축물과 같이 과거에 축조되거나 건축행정 이전 시기에 지어진 건축물의 경우 사용승인일이 기록되지 않은 사례가 일부 확인	- 건축물대장 표제부: '허가_일', '착공_일', **'사용승인_일'**, '주_부속_구분_코드', '주_용도_코드', **'연면적'** - 건축물대장 층별개요: '주_부속_구분_코드', '주_용도_코드', **'면적'**

출처: 연구진 작성
주: 데이터 오류를 제시한 컬럼은 볼드처리

3. 건축물 데이터 특성을 고려한 품질 고도화 방향

1) 개요

앞서 확인한 건축물 데이터 품질 및 이용 현황에 근거하여, 면적, 용도, 연계 데이터의 품질 고도화 방안을 제시하고자 한다. 면적 데이터는 건축물대장 품질 점검에서 가장 높은 오류율을 기록하였다. 이는 일회성 입력 실수를 넘어선 구조적 결함이 있음을 나타낸다. 아울러 연면적, 건폐율, 용적률 등 면적 관련 지표는 건축 통계 및 연구 분석에서 핵심적으로 이용되는 변수이다. 그러므로 면적 데이터는 품질 고도화의 최우선 검토 대상으로 선정되어야 한다. 용도 데이터도 통계 생산, 재해 취약 주택 분석, 위반 건축물 분석 등 연구에 유용성이 높다. 건축물대장에서 용도별 분류 코드가 표출되지 않는다는 문제가 있다. 통계 생산과 연구는 분류 코드를 기준으로 이루어지기에, 기재 내용과 분류 코드 간 일관성을 확보하고 용도 데이터 품질 현황을 파악하여야 한다. 이것이 품질 고도화의 핵심 과제이다.

마지막으로 건축행정데이터와 재난·재해데이터의 연계 연구 사례에서 나타난 것처럼, 건축물 데이터는 건축물과 관련된 다른 분야 데이터와 연계되었을 때 그 가치가 더 높아질 수 있다. 실제 선행연구에서도 화재·침수, 사업자등록 정보 등 다른 행정데이터와의 연계 활용이 활발히 이루어지고 있다. 건물ID는 건축행정 데이터 내에서 인허가 데이터와 건축물대장 데이터의 연계키로 활용되고 있으나, 그 품질 현황에 대한 연구가 이루어지지 않은 실정이다. 따라서 건물ID 기반 인허가 및 건축물대장 연계 품질 현황을 분석하고 품질 고도화 방안을 도출하고자 한다.

2) 건축물 면적 데이터 품질 고도화

■ 면적 데이터 검증 대상 설정

건축물 면적 데이터는 대지면적, 건축면적, 바닥면적, 연면적, 용적률 산정 연면적 등 면적 항목과, 건폐율, 용적률 등 연관 항목을 포괄한다. 오류율과 활용도를 고려하여 검증 대상을 설정하기 위하여, 2022년 건축물대장 등 정비 규칙 86개 중 건물동 단위 면적 관련 규칙을 검토하고 검증 대상 항목을 설정하였다. 2022년 정비 규칙은 건축물대장과 건축인허가, 주택인허가 등 건축행정 데이터 전반을 포괄하는데, 전국 모든 건축물의 면적 데이터 품질 검증을 위하여 건축물 현황에 대한 데이터인 건축물대장 대상 규칙에 한정하여 검증 대상 항목을 검토하였다. 마찬가지로, 대지 단위 데이터인 총괄표

제부, 구분소유되는 호 단위 데이터인 전유부 등도 제외하였다. 일반건축물대장 및 집합건축물대장 표제부를 포괄하여 건물동 단위 데이터를 기록한 '표제부' 테이블을 대상으로, 면적 관련 데이터 품질를 검증하고 있는 규칙은 14번 대지면적, 16번 건축면적, 19번 건폐율, 20번 용적률 등이다. 해당 규칙에서는 대지면적, 건축면적, 바닥면적 등을 포괄하고 있으며, 건폐율 및 용적률 계산의 정확성 검증에서 대지면적, 건축면적, 용적률 산정 연면적 등의 비율 산출을 통하여 이러한 항목의 품질을 검증하고 있다. 본 연구에서도 정비 규칙에서 언급된 면적 관련 항목을 면적 데이터 검증 대상으로 설정하며, 추가로 신규 검증규칙 개발, 기계학습 분석을 위한 변수 도출 등을 통하여 면적 데이터를 다각적으로 검증하고자 한다.

[표 2-22] 2022년 건축물대장 정비규칙 중 건물동 단위 면적 관련 규칙

순번	업무규칙명	오류건수	점검대상건수	오류율
14	(대지면적) 일반건축물대장 및 표제부 내 대지면적의 값이 건축면적보다 작은 경우의 데이터 검증(정확성)	1,136	612,719	0.19%
16	(건축면적) 일반건축물대장 및 표제부의 바닥면적의 합이 가장 큰 면적과 표제부의 건축면적이 다른 데이터 검증(정확성)	1,334,538	7,952,243	16.78%
19	(건폐율) 일반건축물대장 및 표제부 건폐율 계산의 정확성 → 소수점 이하 처리에 대한 명확한 법규 처리 없어 반올림 자리수가 불분명함(소숫점 이하 절사처리)	54,486	7,324,326	0.74%
20	(용적률) 일반건축물대장 및 표제부 용적율 계산의 정확성	723,474	5,212,430	13.88%

출처: 세움터 내부자료 바탕으로 연구진 작성
주: 지자체 품질정비 대상은 음영처리

■ 면적 데이터 분석 방법론

건축물대장 면적 데이터 품질 고도화는 먼저 기초통계 분석 및 오류 현황 검토를 통해 데이터의 전반적 분포 특성을 파악하고, 극단값, 결측, 음수 표기 등 오류 유형을 확인한다. 2022년 정비규칙 중 건물동 단위 면적 관력 규칙을 현재 시점 데이터에 적용하여 오류를 검증한다. 정비규칙의 유효성과 오류 잔존 여부 등을 검토하고, 항목별 오류율과 시기·지역별 분포를 분석한다.

다음으로 신규 검증규칙을 개발하고자 한다. 총 86개 검증규칙 중 건물동 단위 면적 관련 규칙은 4개 뿐이다. 이는 충분한 검증이 이루어지지 못하였을 가능성을 시사한다. 신규 검증규칙 도출 및 검증을 통하여 규칙 기반 검증의 범위를 확장한다.

마지막으로, 데이터에 내재된 패턴 탐지를 위해 기계학습 기반 이상값 탐지를 병행한다. 규칙 기반 검증만으로는 포착 어려운 잠재적 오류 비정형적 패턴을 식별한다. 면적 데이터는 건축물 규모에 비례하는 변수를 다수 포함한다. 이러한 데이터에 이상값 탐지를 적용하면 크기 차이가 큰 건축물이 이상값으로 도출되는 문제가 있다. 건축물 규모의 절대적 크기 차이에 영향받지 않는 비교 가능 지표를 구성한다.

Isolation Forest나 One-Class SVM 등 대표 기계학습 기반 이상탐지 알고리즘을 적용한다. 두 알고리즘 모두 비지도 학습 기반의 이상값 탐지 기법이다. Isolation Forest은 의사결정나무 알고리즘 바

탕으로 분리의 용이성으로 이상값을 판별한다. 분포 형태나 밀도에 대한 가정에 의존하지 않기 때문에 비모수 데이터나 수많은 변수를 포함하는 고차원 데이터에 효과적으로 적용된다.

One-Class SVM은 주어진 데이터를 둘러싸는 경계를 학습하고 그 내부에 속하지 않는 데이터를 이상값으로 판별하는 방식이다. One-Class SVM은 중심성 경향이 나타나는 데이터에 적합한 방법론이다. Isolation Forest와 One-Class SVM은 서로 다른 특성으로 이상값을 정의하기 때문에 상호 보완적 방법론으로 복합 사용 가능하다.

이와 같이 데이터 구축, 기존 규칙 적용, 신규 규칙 개발, 기계학습 기법의 적용을 단계적으로 병행한 분석 방법론은 건축물대장 면적 데이터의 다양한 오류 유형과 잠재적 이상 패턴을 포괄적으로 식별할 수 있는 체계적 접근이다.

3) 건축물 용도 데이터 품질 고도화

■ 용도 데이터 검증 대상 설정

건축물 데이터 중 텍스트 데이터의 검증 및 품질 고도화 시범 적용 범위로 건축물 용도를 선정하였다. 건축물 용도는 건축물 현황 통계에서 지역과 더불어 주된 분류 기준이다. 또한, 세움터 데이터 정비 업무 역시 건축물 용도를 중점적으로 포함하였다.

2022년 세움터 정비대상 업무규칙 중 건축물 용도와 관련된 내용은 건축물대장 주용도코드의 정확성(10번, 15번), 세부용도코드 정확성(24번, 30번), 건축/주택인허가대장 주용도코드의 정확성(47번, 68번) 등이다. 여기서 말하는 주용도코드와 세부용도코드는 각각 건축법 시행령의 건축물 용도 구분에 따른 용도 코드 대분류(000으로 끝나는 숫자 5자리 코드), 소분류(대분류와 첫 2자리가 같고 용도별 숫자 3자리가 이어지는 5자리 코드)에 해당한다. 다만 건축허브에서 제공하는 건축물대장 데이터 기준으로는 컬럼명이 '주 용도 코드'로 동일하다.

건축물 용도 코드의 정확성을 판단하는 기준은 여러 가지가 있을 수 있으나, 건축물대장에 포함된 용도 관련 데이터에 한정한다면 '기타 용도' 데이터가 기준이 될 수 있다. 기타 용도 데이터는 건축물 용도에 대한 자유 형식 텍스트를 포함하는데, 실제로 건축물대장을 발급하는 경우 문서에 기재되는 내용은 기타 용도 데이터이며, 주 용도 코드는 건축물대장 문서에 표출되지 않는다. 따라서 건축물대장 기재내용에 해당하는 기타 용도 데이터를 기준으로 주 용도 코드가 정확하게 부여되었는지를 판단할 수 있다.

■ 용도 데이터 분석 방법론

- 나이브 베이즈

나이브 베이즈 모델은 통계학의 베이즈 정리를 토대로 가장 확률이 높은 클래스로 분류 문제를 수행하는 지도학습 머신러닝 알고리즘이다(scikit-learn developers, 2025). 베이즈 정리는 새로운 증거가 관측되었을 때 이에 기초하여 어떤 사건의 확률을 갱신하여 사후확률을 얻는 수학적 원리를 표현한다

(Murty & Devi, 2011). 분류 문제의 예측 관점에서 이를 다시 표현하면, 사후확률은 특성 $x_1, \ldots, x_n$이 있는 대상이 특정한 분류 클래스 C_k일 확률, 즉 해당 클래스로 분류되는 것이 정답일 확률이 된다(scikit-learn developers, 2025).

$$P(C_k|x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n|C_k)P(C_k)}{P(x_1, \ldots, x_n)}$$

나이브 베이즈 모델에서 사후확률은 어떤 대상이 특정 클래스 C_k에 속할 때, 그 대상이 특성 $x_1, \ldots, x_n$을 가지고 있을 조건부 확률(우도)와, 그러한 특성이 있는지 여부가 관찰되기 전에 어떤 대상이 해당 클래스에 속할 사전확률, 그리고 어떤 대상이 특성 $x_1, \ldots, x_n$을 가지고 있을 확률(정규화 상수)로 정의된다(scikit-learn developers, 2025).

특정 분류 C_k에 속한 어떤 대상이 특성 $x_1, \ldots, x_n$을 가지고 있을 확률은 구하기 어려운 다변량 확률 문제이다. 나이브 베이즈는 이를 단순화하기 위하여, 각 특성 x_i들이 동일한 분류 클래스 C_k 안에서 나타날 확률이 서로 조건부 독립이라고 가정한다(나이브 가정). 따라서 특정 클래스에서 여러 특성이 나타날 조건부 확률(우도)은 단순히 해당 클래스에서 각 특성이 나타날 조건부 확률의 곱으로 표현될 수 있다(scikit-learn developers, 2025). 이를 식으로 나타내면 다음과 같다.

$$P(x_1, x_2, \ldots, x_n|C_k) = \prod_{i=1}^{n} P(x_i|C_k)$$

한편 분모에 있는 정규화 상수는 분류 클래스와 무관하게 특성 $x_1, \ldots, x_n$을 가지고 있을 확률을 의미하므로, 각 클래스에 속하면서 해당 특성을 가질 확률을 모두 더하여 구할 수 있다(scikit-learn developers, 2025). 따라서 최종적으로 나이브 베이즈 모델의 사후확률은 다음 식과 같이 나타낼 수 있다.

$$P(C_k|x_1, \ldots, x_n) = \frac{P(C_k)\prod_{i=1}^{n} P(x_i|C_k)}{\sum_{j=1}^{K} P(C_j)\prod_{i=1}^{n} P(x_i|C_j)}$$

분류 문제에 나이브 베이즈 모델을 적용하는 경우, 분모의 정규화 상수는 클래스 C_k에 따라 달라지지 않으므로, 관심사가 되지 않는다. 분자를 최대로 하는 C_k를 찾는 것이 곧 특성 $x_1, \ldots, x_n$을 가지고 있는 대상이 속할 확률이 가장 높은 분류 클래스를 찾는 것이 된다. 이를 위하여 나이브 베이즈 모델은 분자에 있는 전체 대상 중 C_k에 속할 확률, C_k에 속한 대상이 각 특성 x_i를 가질 확률 등을 데이터에서 추출하여 모델을 구축한다(scikit-learn developers, 2025).

- 나이브 베이즈 기반 분류

건축물 용도 텍스트와 건축법상 용도 구분의 정합성을 검증하기 위하여, 건축물 용도 텍스트에서 건축법상 용도 구분을 예측하는 모델을 구축하고 예측 결과가 건축물대장의 분류와 일치하는지 확인하고자 하였다. 모델의 예측 결과가 데이터상 분류와 일치하지 않는 경우 용도 분류가 잘못된 오류를 탐

지한 것으로 해석할 수 있다. 다만, 단순히 예측 모델이 잘못 탐지하는 경우가 있을 수 있으므로 시범 적용을 통하여 적용 가능성을 검토하고자 하였다.

나이브 베이즈 모델에서 특성 $x_1, \ldots, x_n$을 지닌 대상이 분류 클래스 C_k에 속할 확률(사후확률)은 전체 대상 중 C_k에 속할 확률($P(C_k)$)과 C_k에 속한 대상이 각 특성 x_i를 가질 확률의 곱($\prod_{i=1}^{n} P(x_i|C_k)$)에 비례한다. 나이브 베이즈 기반 분류 모델은 이들 특성을 가진 대상을 사후확률이 가장 높은 C_k로 분류하는 모델이다.

$$P(C_k|x_1,\cdots,x_n) \propto P(C_k)\prod_{i=1}^{n} P(x_i|C_k)$$

건축물 용도 텍스트를 건축법상 용도 구분에 맞추어 분류하는 문제에서, C_k는 각 용도 구분에 해당하며, 특성 $x_1, \ldots, x_n$은 텍스트를 구성하는 여러 특성이 된다. 건축물대장의 건축물 용도 텍스트는 주로 다양한 용도와 관련된 단어가 나열되는 구조로 이루어져 있다. 여러 단어가 조합되어 하나의 건축물 용도를 지칭하기보다는, 단어 각각이 건축물이 사용되는 용도 (중 하나)를 구체적으로 지칭하고 있다. 건축물 용도 텍스트는 각 단어의 집합으로, 특성 x_i는 건축물 용도 텍스트에서 나타나는 단어 중 하나로 생각할 수 있다. 각 단어가 나타날 확률은 서로 조건부 독립이라는 나이브 베이즈의 가정(나이브 가정)에 대해서는 건축물이 여러 용도로 이용되는 경우에도 각 단어의 등장 여부가 다른 단어가 지칭하는 건축법상 용도 구분에 영향을 주지 않으므로 무리가 없다고 판단된다.

$P(C_k)$은 전체 집단 분류 문제에 나이브 베이즈 모델을 적용하는 경우, 분모의 정규화 상수는 클래스 C_k에 따라 달라지지 않으므로, 관심사가 되지 않는다. 분자를 최대로 하는 C_k를 찾는 것이 곧 특성 $x_1, \ldots, x_n$을 가지고 있는 대상이 속할 확률이 가장 높은 분류 클래스를 찾는 것이 된다. 이를 위하여 나이브 베이즈 모델은 분자에 있는 전체 대상 중 C_k에 속할 확률, C_k에 속한 대상이 각 특성 x_i를 가질 확률 등을 데이터에서 추출하여 모델을 구축한다.

4) 인허가-건축물대장 데이터 연계 품질 고도화

■ 건축물대장, 건축인허가, 주택인허가 데이터 연계

- 인허가-건축물대장 데이터 연계의 한계

건축물대장은 '관리_건축물대장_PK', 건축인허가는 '관리_허가대장_PK', 주택인허가는 '관리_주택대장_PK' 등 각각 다른 고유키를 사용하고 있어, 현재 데이터 구조에서 유일한 대안은 주소를 기준으로 테이블을 연결하는 것이다. 모든 테이블에는 대지_위치, 도로명_대지_위치 칼럼이 존재하는데, 이 중 도로명_대지_위치 칼럼은 약 6%(5,509건)의 결측치를 가지고 있어 결측치가 없는 대지_위치를 활용하는 편이 더욱 용이하다. 또한 공간정보 등 다른 데이터베이스와의 연결에서는 주로 지번을 사용하는데, 이를 위해 시군구_코드, 법정동_코드, 대지_구분, 번, 지를 결합한 19자리의 PNU 코드가 활용된다. 대지_위치와 PNU 코드는 표기 방식은 다르지만, 모두 실제로는 건축물이 위치한 주소를 나

타낸다고 볼 수 있다. 즉, 이 두 가지 정보를 활용하면 같은 주소에 위치한 건축물의 정보를 상호 연결하는 것이 가능하다.

이러한 방식으로 보완한다면, 하나의 필지에 단일 건축물이 있을 경우 연결은 효율적으로 가능하다. 그러나 필지-건축물 관계가 복잡해지면 문제는 여전히 해소되지 않는다. 즉, 하나의 필지에 다수 건축물이 있거나, 다수 필지에 건축물이 집합적으로 존재하는 상황에서는 그 연결 관계를 명확히 파악하기 어렵다.

- 건물ID

건물ID는 건축물에 적용하는 새로운 연계키이다. 기존 건축물대장 고유번호와 데이터베이스 PK 활용 등의 한계점을 보완하기 위해 도입되었다. 기존 등기정보 매칭용 공통고유번호를 건축물 인허가 단계로 확대 적용한 것이다. 건물ID는 총 16자리 숫자로 종류(2자리), 허가년도(4자리), 시군구(3자리), 일련번호(7자리)로 구성된다. 종류 코드는 건축물대장 종류를 나타내며, 총괄(20), 동(21: 일반/표제), 호(22: 전유)로 구분된다. 허가년도는 건축물이 허가된 연도를 의미하며, 시군구 코드는 행정구역 변경 여부와 관계없이 동일하게 유지되는 고정값이다. 마지막으로 일련번호는 종류, 허가년도, 시군구별로 순차적으로 부여된다.

건물ID의 관리 원칙은 다음과 같다. 첫째 건축물 동 단위로 부여하며 부속건축물도 포함한다. 둘째 인허가 처리 과정에서 시스템이 자동으로 부여한다. 셋째 한번 부여된 건물ID는 변경하거나 삭제할 수 없고 고유 식별키로 영구히 유지하여야 한다. 넷째 건축물 단위 정보를 유통할 때 핵심 식별키로 활용하고 건물ID를 중심으로 이력을 추적 관리할 수 있도록 한다.

■ 건물ID 기반 연계 가능성 평가

건물ID는 인허가와 건축물대장 테이블 간 정보 보완과 검증에 유용하게 활용될 수 있다. 동일 건물ID으로 테이블 정보를 비교 참조하여 누락 데이터를 보완하거나 서로 다른 입력값을 검증할 수 있다. 세움터에 따르면 인허가 데이터와 건축물대장 데이터는 건물ID 기반으로 자동 연계되고 있다. 따라서 최근 데이터는 누락 등 연계 품질 문제가 원천적으로 예방되었다.

본 연구는 건물ID 기반 데이터 연계 품질을 기존 주소 기반 연계와 비교 검증하고자 한다. 세움터 데이터가 건물ID 기반으로 완전하게 연계 가능한지 검증한다. 각 데이터셋과 테이블별로 건물ID가 올바른 형태로 결합된 비율을 평가한다. 각 테이블 간 속성 매칭 방식, 중복 데이터 처리, 다대일·일대다 관계 처리 방식 등을 검토한다. 이 과정에서 발생 가능 문제점을 검토하고 이를 해결할 논리적 연계 구조와 설계 방안을 구체적으로 검토하는 것을 목적으로 한다.

제3장

건축물 데이터 품질 고도화 시범적용

1. 건축물대장 면적 데이터 품질 고도화

1) 개요

본 절에서는 건축물대장 면적 데이터의 품질 고도화 방법론을 개발 및 시범적용하였다. 먼저 품질 진단을 위한 분석 데이터를 구축하고 데이터의 전반적 분포와 특성을 파악하였다. 이어 기존 정비규칙을 적용한 오류 검증을 실시하여 규칙 기반 진단의 유효성을 재확인하였다. 또한 신규 검증규칙을 개발하여 적용하였다. 이를 통하여 기존 정비규칙으로 포착하지 못한 오류 유형을 보완하였다. 마지막으로 기계학습 기법을 활용한 이상값 탐지를 병행하였다. 이러한 과정을 거쳐 면적 데이터의 오류 특성을 탐색하였다.

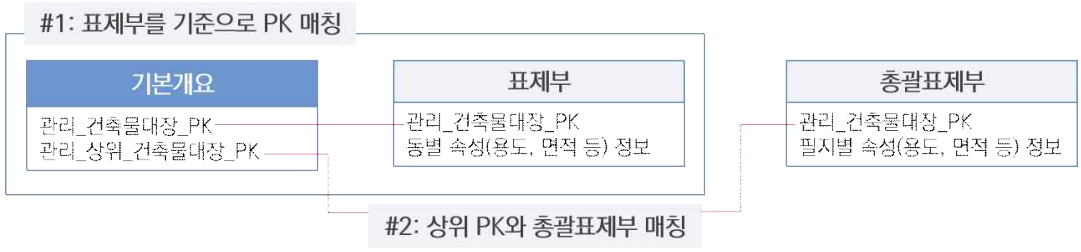
이와 같이 규칙 기반 검증과 인공지능을 활용한 오류 탐지를 병행하는 다층적 접근을 통해, 건축물대장 면적 데이터의 오류 양상을 입체적으로 파악하고, 향후 데이터 품질 관리 체계 고도화를 위한 근거를 마련하고자 하였다.

2) 분석 데이터 구축 및 오류 현황 검토

- 분석 데이터 구축

건축물대장 면적 데이터 품질 고도화 시범적용에 활용한 분석 데이터는 건축허브에서 제공하는 대용량 제공 서비스에서 2025년 2월 기준 건축물대장 데이터를 바탕으로 구축하였다. 건축물대장에는 대지별, 건물동별, 호별 대장이 모두 포함되어 있으나, 본 분석에서는 건물동별 대장의 정비를 기본적인 범위로 설정하였기 때문에 건물동 단위 데이터를 기재하고 있는 표제부 데이터를 기본으로 활용하였다. 다만, 총괄표제부가 존재하는 경우 건축물대장에 기재되지 않는 사항이 있어 해당 내용에 대한 파악을 위하여 총괄표제부를 표제부에 연계하여 분석 데이터를 구축하였다.

[그림 3-1] 건축물대장 면적 데이터 품질 고도화 흐름도
출처: 연구진 작성



[그림 3-2] 건축물대장 테이블 매칭 구조 모식도
출처: 연구진 작성

건축물대장의 연계는 표제부의 관리_건축물대장_PK 컬럼의 값을 기준으로 기본개요 테이블에서 동일한 컬럼인 관리_건축물대장_PK와 매칭시키고, 기본개요 테이블에 기재된 관리_상위_건축물대장_PK를 연계키로 활용한다. 기본개요 테이블은 대지별 총괄표제부, 건물동별 일반건축물대장 및 표제부, 호별 전유부 등 모든 건축물대장에 대한 기본개요가 포함된 테이블로, 관리_상위_건축물대장_PK의 값은 대장 종류에 따라 그 의미가 달라진다. 건물동별 건축물대장인 일반건축물대장 또는 집합건축물대장 표제부의 경우, 상위 건축물대장은 대지별 총괄표제부가 해당된다. 총괄표제부는 대지 내 건축물이 두 개 동 이상 존재해야 생성되므로, 관리_상위_건축물대장_PK에는 대지 내 모든 건축물에 대한 총괄표제부가 존재하는 경우에만 그 고유키(PK)가 기재된다. 이 PK를 총괄표제부 테이블의 관리_건축물대장_PK와 연계할 수 있고, 이를 통하여 건물동 단위 표제부 데이터에 (대지 내 여러 건축물이 있는 경우) 대지 총괄 정보를 연계할 수 있다. 만약 대지 내 건축물이 1개 동만 있으면 총괄표제부

가 존재하지 않고, 분석 데이터에서는 해당 내용이 빈 값(null)으로 나타나게 된다. 이상과 같은 분석 데이터 매칭 구조는 그림 3-2와 같다.

[표 3-1] 활용한 건축물대장 테이블의 레코드 수 및 컬럼 정보

테이블	레코드 수	컬럼 수	컬럼 명
기본개요	27,994,157	30	관리_건축물대장_PK, 관리_상위_건축물대장_PK, 대장_구분_코드, 대장_구분_코드_명, 대장_종류_코드, 대장_종류_코드_명, 대지_위치, 도로명_대지_위치, 건물_명, 시군구_코드, 법정동_코드, 대지_구분_코드, 번, 지, 특수지_명, 블록, 로트, 외필지_수, 새주소_도로_코드, 새주소_법정동_코드, 새주소_지상지하_코드, 새주소_본_번, 새주소_부_번, 지역_코드, 지구_코드, 구역_코드, 지역_코드_명, 지구_코드_명, 구역_코드_명, 생성_일자
총괄표제부	615,541	64	관리_건축물대장_PK, 대장_구분_코드, 대장_구분_코드_명, 대장_종류_코드, 대장_종류_코드_명, 신_구_대장_구분_코드, 신_구_대장_구분_코드_명, 대지_위치, 도로명_대지_위치, 건물_명, 시군구_코드, 법정동_코드, 대지_구분_코드, 번, 지, 특수지_명, 블록, 로트, 외필지_수, 새주소_도로_코드, 새주소_법정동_코드, 새주소_지상지하_코드, 새주소_본_번, 새주소_부_번, 대지_면적(㎡), 건축_면적(㎡), 건폐_율(%), 연면적(㎡), 용적_률_산정_연면적(㎡), 용적_률(%), 주_용도_코드, 주_용도_코드_명, 기타_용도, 세대_수(세대), 가구_수(가구), 주_건축물_수, 부속_건축물_수, 부속_건축물_면적(㎡), 총_주차_수, 옥내_기계식_대수(대), 옥내_기계식_면적(㎡), 옥외_기계식_대수(대), 옥외_기계식_면적(㎡), 옥내_자주식_대수(대), 옥내_자주식_면적(㎡), 옥외_자주식_대수(대), 옥외_자주식_면적(㎡), 허가_일, 착공_일, 사용승인_일, 허가번호_년, 허가번호_기관_코드, 허가번호_기관_코드_명, 허가번호_구분_코드, 허가번호_구분_코드_명, 호_수(호), 에너지효율_등급, 에너지절감_율, 에너지_EPI점수, 친환경_건축물_등급, 친환경_건축물_인증점수, 지능형_건축물_등급, 지능형_건축물_인증점수, 생성_일자
표제부	8,027,067	77	관리_건축물대장_PK, 대장_구분_코드, 대장_구분_코드_명, 대장_종류_코드, 대장_종류_코드_명, 대지_위치, 도로명_대지_위치, 건물_명, 시군구_코드, 법정동_코드, 대지_구분_코드, 번, 지, 특수지_명, 블록, 로트, 외필지_수, 새주소_도로_코드, 새주소_법정동_코드, 새주소_지상지하_코드, 새주소_본_번, 새주소_부_번, 동_명, 주_부속_구분_코드, 주_부속_구분_코드_명, 대지_면적(㎡), 건축_면적(㎡), 건폐_율(%), 연면적(㎡), 용적_률_산정_연면적(㎡), 용적_률(%), 구조_코드, 구조_코드_명, 기타_구조, 주_용도_코드, 주_용도_코드_명, 기타_용도, 지붕_코드, 지붕_코드_명, 기타_지붕, 세대_수(세대), 가구_수(가구), 높이(m), 지상_층_수, 지하_층_수, 승용_승강기_수, 비상용_승강기_수, 부속_건축물_수, 부속_건축물_면적(㎡), 총_동_연면적(㎡), 옥내_기계식_대수(대), 옥내_기계식_면적(㎡), 옥외_기계식_대수(대), 옥외_기계식_면적(㎡), 옥내_자주식_대수(대), 옥내_자주식_면적(㎡), 옥외_자주식_대수(대), 옥외_자주식_면적(㎡), 허가_일, 착공_일, 사용승인_일, 허가번호_년, 허가번호_기관_코드, 허가번호_기관_코드_명, 허가번호_구분_코드, 허가번호_구분_코드_명, 호_수(호), 에너지효율_등급, 에너지절감_율, 에너지_EPI점수, 친환경_건축물_등급, 친환경_건축물_인증점수, 지능형_건축물_등급, 지능형_건축물_인증점수, 생성_일자, 내진_설계_적용_여부, 내진_능력

출처: 연구진 작성

- 분석 데이터 기초통계

표제부 테이블은 8,027,067행 데이터를 포함하고 있으며, 건물동별로 1개 행의 데이터가 기재되므로, 이는 약 800만 동에 대한 데이터에 해당한다. 데이터 전처리에 앞서 표제부 테이블의 주요 컬럼의 기초통계를 도출하였다. 먼저 미기재열의 경우, 수치 데이터에서는 미기재율이 0으로 나타났다. 이는 건축물대장 데이터 중 수치 데이터의 경우 대장상 미기재 데이터를 '0'으로 기재하고 있기 때문인데, 이에 대하여는 뒤에서 더 자세하게 검토하였다. 텍스트 데이터의 경우 미기재율이 0인 경우(시군구코드, 법정동코드 등)도 있었고, 대체로 미기재율(NULL 비율)이 1% 미만으로 나타났으나, 건축허

가 등 일자 관련 데이터의 미기재율은 높게 나타났다. 사용승인_일의 경우 미기재율이 7.8%에 달하였고, 허가_일과 착공_일의 미기재율은 각각 33.67%, 48.29%로 매우 높게 나타났다.

면적 관련 데이터의 분포를 검토하기 위하여 최소값, 최대값, 평균, 중앙값 등을 검토하였다. 대지_면적(㎡), 건축_면적(㎡), 연면적(㎡), 용적_률_산정_연면적(㎡) 등 면적 데이터와, 면적과 관련된 건폐_율(%)과 용적_률(%) 등 컬럼을 확인하였다. 모든 컬럼에서 최소값이 0 또는 음수로 나타났는데, 건축물 관련 면적이 음수일 수 없다는 점을 고려하면 오류로 판단된다. 대지_면적(㎡)의 경우 최소값이 0인 데이터가 있었으며, 건축_면적(㎡)은 -93.66, 연면적(㎡)과 용적_률_산정_연면적(㎡)은 각 -132.16으로 음수값이 나타났다. 건폐_율(%)과 용적_률(%)도 최소값이 각 -31.01, -43.76이 나타났다. 최대값의 경우, 면적 데이터에서 대지면적은 4,333,333,333 (단위 ㎡, 이하 동일), 건축면적은 56,237,885, 연면적은 83,374,375 등 매우 큰 값이 나타났는데, 여의도 면적(윤중로 제방 안쪽 면적 기준) 2.9 ㎢ (= 2,900,000 ㎡)의 수십~수백 배에 달하는 값으로, 정상적인 건축물 1개 동의 규모로 보기 어렵다. 건폐율과 용적률(단위 %)의 경우도 1백~1천 정도가 정상 범위이나 각각 1,000,000,000, 11,314,584이 최대값으로 나타나 오류로 판단된다.

[표 3-2] 표제부의 주요 컬럼 기초통계

컬럼명	최소값	최대값	고유값의 수	평균	표준편차	1사분위수 (25%)	중앙값 (50%)	3사분위수 (75%)	NULL 비율
시군구_코드	0	99999	252						0
법정동_코드	0	99999	881						0
주_부속_구분_코드	0	1	2						0
주_부속_구분_코드_명	부속건축물	주건축물	2						0.01
대지_면적(㎡)	0	4333333333	139670	6725.65	1606405.60	0	114	424	0
건축_면적(㎡)	-93.66	56237885	319695	244.79	35605.07	50	90	158	0
건폐_율(%)	-31.01	1000000000	63811	8699.78	2282410.88	0	4	35	0
연면적(㎡)	-132.16	83374375	492711	638.73	56449.29	60	117	303	0
용적_률_산정_연면적(㎡)	-132.16	98585340	426509	561.99	68400.84	53	104	262	0
용적_률(%)	-43.76	11314584	156252	41.70	4021.47	0	4	41	0
구조_코드		`	43						0.25
구조_코드_명	강파이프구조	흙벽돌조	43						0.25
기타_구조	\t 경량철골구조	? 조	114205						0.72
주_용도_코드	0	b8990	108						0.36
주_용도_코드_명	가설건축물	휴양콘도미니엄	92						0.36
기타_용도	\t 공동주택	힐링센터	244470						0.65
높이(m)	-13.6	88055	26694	5.72	123.98	0	4	8	0
지상_층_수	0	131	71	1.95	2.64	1	1	2	0
지하_층_수	-1	93	24	0.16	0.42	0	0	0	0
부속_건축물_수	0	1666	115	0.21	1.25	0	0	0	0
부속_건축물_면적(㎡)	0	320514349	74878	163.94	118975.29	0	0	0	0
총_동_연면적(㎡)	-132.16	568282882	506857	892.33	241540.52	47	101	275	0
허가_일		ㅋ	43392						33.67
착공_일		99911111	29239						48.29
사용승인_일		9990408	56445						7.8

출처: 연구진 작성

텍스트 데이터는 시군구_코드와 법정동_코드에서 최소값과 최대값이 0과 99999로 나타났는데 정상적인 시군구 및 법정동 코드값이 아니어서 오류로 판단된다. 주_용도_코드, 허가_일, 착공_일, 사용승인_일 등 텍스트로 기재된 코드 및 일자 데이터도 최소값과 최대값은 정상적인 코드값 및 일자가 아닌 것으로 나타났으며, 모두 오류로 판단된다.

총괄표제부 테이블은 615,541개 대지에 대한 데이터로 구성되어있다. 주요컬럼의 기초통계를 검토한 결과, 미기재율(NULL 비율)은 표제부와 유사한 분포를 보이나 다소 높은 값을 보였다. 주_용도_코드와, 주_용도_코드_명의 미기재율은 각 3.27%, 기타_용도는 9.18%로 표제부에 비해 상대적으로 높게 나타났다. 또한 허가_일과 착공_일, 사용승인_일도 미기재율이 59.84%, 62.44%, 56.35%로 표제부보다 높게 나타났다.

[표 3-3] 총괄표제부의 주요 컬럼 기초통계

컬럼명	최소값	최대값	고유값의 수	평균	표준편차	1사분위수 (25%)	중앙값 (50%)	3사분위수 (75%)	NULL 비율
시군구_코드	11110	99999	252						0
법정동_코드	0	99999	872						0
대지_면적(㎡)	0	2434234234	74247	11171.48	3113159.74	0	875	2730	0
건축_면적(㎡)	0	104787000	222777	2033.10	246797.36	135	287	805	0
건폐_율(%)	0	355805.62	36463	25.04	483.77	0	21	38	0
연면적(㎡)	0	3986783503	259206	19454.93	5390585.68	147	354	1117	0
용적_률_산정_연면적(㎡)	0	2844507236	268646	11538.37	3700734.52	142	341	1061	0
용적_률(%)	0	93289180	50446	197.44	118919.36	0	25	45	0
주_용도_코드	1000	n0300	71						3.27
주_용도_코드_명	골프연습장	휴양콘도미니엄	55						3.27
기타_용도	\t 공장	? 린생활시설, 위락시설,주택	35739						9.18
세대_수(세대)	0	14308	2131	19.42	135.09	0	0	0	0
가구_수(가구)	0	1730	210	0.85	5.55	0	0	1	0
주_건축물_수	0	98792	298	5.16	333.64	2	2	3	0
부속_건축물_수	0	20000	103	0.66	34.85	0	0	0	0
부속_건축물_면적(㎡)	0	629634307	48689	2765.29	894312.88	0	0	0	0
허가_일		98 1231	15206						59.84
착공_일		99 1115	12350						62.44
사용승인_일	1960505	99 1129	14017						56.35

출처: 연구진 작성

면적 관련 데이터에서는 대지_면적(㎡)과 건축_면적(㎡), 연면적(㎡)의 최소값이 0이었으나, 표제부와 달리 음수인 경우는 없었다. 최대값을 검토한 결과 표제부와 유사하게 매우 큰 값이 나타났고, 건폐_율(%)과 용적_률(%)은 최대값이 각 355,805과 93,289,180으로 오류로 판단되는 값이 관측되었다. 텍스트 데이터 중 시군구_코드의 경우 최소값이 11110으로 정상값이었으나, 최대값은 99999로 오류로 나타났다. 법정동_코드는 최소값은 0, 최대값 역시 99999로 오류가 관측되었다. 허가_일, 착공_일, 사용승인_일도 최소값 및 최대값이 오류로 나타났다. 사용승인_일의 최소값은 '1960505'으로 컬

럼 값 정의에 따른 8자리 연월일 데이터가 아니며, 1960년 또는 1996년에 대한 오기일 수 있는 값이 기재되었다.

기초통계 검토 결과를 종합하면 주요 컬럼에서 최소값과 최대값에서 오류로 판단되는 값이 나타났으며, 이러한 오류는 단순 통계 조회를 통하여 아주 쉽게 발견되었다. 이렇게 명확하고 쉽게 찾을 수 있는 오류가 건축물대장 데이터 상에서 정정되지 않고 있음을 확인하였다.

- 면적 관련 데이터 이상치 사례 검토

기초통계 분석에서 확인된 이상치 사례를 검토하여 실제 오류 여부를 확인하였다. 분석 데이터 검토 및 세움터 건축물대장 조회를 통하여 건축물대장 현황을 파악하고, 독립적인 확인을 위하여 토지이용계획, 민간 지도 및 거리뷰 서비스 등도 확인하여 건축물대장 데이터의 오류 여부를 판단하였다.

면적 관련 수치 데이터의 경우 분석 데이터 상 빈 값(NULL)인 경우는 없고, '0'으로 기재된 경우만 존재하였다. 그러나 건축물대장에서 0일 가능성이 없는 면적 관련 데이터가 0인 경우가 존재하여, 이상치로 판단되었다. 1984년 준공된 공동주택 사례의 경우, 대지 내 여러 건축물이 존재하고 구분소유되는 집합건축물로서 건축물대장에 총괄표제부와 집합건축물대장(표제부)가 존재하였다. 일반건축물대장과 표제부의 경우 총괄표제부에 대지면적, 건폐율, 용적률 등이 기재된 경우 건물동 단위 대장에 그 내용을 적지 않을 수 있고, 그러한 내용을 대장에 명기하고 있다. 그러나 이 사례의 경우 총괄표제부에 대지면적, 건폐율, 용적률이 0으로 기재되어 있어 오류로 나타났다. 토지이용계획 상 실제 필지의 면적은 36,959.8㎡이다. 총괄표제부에 해당 값이 정상적으로 기재된 경우, 일반건축물대장과 표제부의 해당 값은 0으로 기재되어있더라도 건축물대장에서 해당 란은 공란으로 나타난다. 그러나 이 건축물의 경우 총괄표제부에 해당 값이 정상적으로 기재되지 않았으므로, 표제부에도 대지면적, 건폐율, 용적률이 '0' 기재되는 오류가 발생하였다.

[표 3-4] 사용승인일 1984년 공동주택 사례

관리_건축물대장_PK	동명	표제부						총괄표제부					
		대지면적	건축면적	건폐율	연면적	용적률산정연면적	용적률	대지면적	건축면적	건폐율	연면적	용적률산정연면적	용적률
10241100181447	상가	0	2040.33	0	9482.35	6084.08	0	0	0	0	72027.55	0	0
10241100181665	15	0	849.36	0	13194	12362.4	0	0	0	0	72027.55	0	0

출처: 연구진 작성

건축물대장 총괄표제부(갑)

(2쪽 중 제1쪽)

건물ID	고유번호	건축물 명칭	특이사항
대지위치	지번	도로명주소	

대지면적	연면적	지역	지구	구역
0 ㎡	72,027.55 ㎡	제3종일반주거지역		
건축면적	용적률 산정용 연면적	건축물 수	주용도	
㎡	㎡	6	공동주택	
건폐율	용적률	총 호수/가구수/세대수	총 주차 대 수	부속 건축물
0 %	0 %	61호/0가구/450세대	474	동 ㎡
조경면적	공개 공지/공간 면적	건축선 후퇴면적	건축선 후퇴거리	
㎡	㎡	㎡	m	

건축물 현황

구분	명칭	도로명주소	건축물 주구조	건축물 지붕	층수	용도	연면적(㎡)	변동일	변동원인
주1			철근콘크리트조	슬라브	1/15	주거시설	13,194	2011.10.4.	표시변경(직권)
주2			철근콘크리트조	슬라브	1/15	주거시설	12,130.8	2011.10.4.	표시변경(직권)
주3			철근콘크리트조	슬라브	1/15	주거시설	13,194	2011.10.4.	표시변경(직권)
주4			철근콘크리트조	슬라브	1/15	주거시설	10,832.4	2011.10.4.	표시변경(직권)

총괄표제부

집합건축물대장(표제부, 갑)

(2쪽 중 제1쪽)

건물ID	고유번호	명칭	호수/가구수/세대수
			60호/0가구/0세대
대지위치	지번	도로명주소	

※대지면적	연면적	※지역	※지구	※구역
0 ㎡	9,482.35 ㎡	일반주거지역 외 2		지구단위계획구역
건축면적	용적률 산정용 연면적	주구조	주용도	층수
2,040.33 ㎡	6,084.08 ㎡	철근콘크리트조	판매시설	지하: 1층, 지상: 3층
※건폐율	※용적률	높이	지붕	부속건축물
0 %	0 %	m	슬라브	동 ㎡
※조경면적	※공개 공지/공간 면적	※건축선 후퇴면적	※건축선 후퇴거리	
㎡	㎡	㎡	m	

표제부(상가동)

[그림 3-3] 사용승인일 1984년 공동주택의 총괄표제부(PK: 10241100198307)
출처: 세움터(https://www.eais.go.kr/)에서 연구진 발급 후 강조 내용 작성

2011년 준공된 다른 공동주택 사례의 경우, 총괄표제부에는 모든 면적 관련 수치 데이터가 정상적으로 기재되어 있다. 따라서 표제부에서 대지면적, 건폐율, 용적률이 기재되지 않을 수 있고, 실제로 데이터에서도 0으로 기재되어있다. 그러나 실제 건축물대장을 확인한 결과, 대지면적, 건폐율, 용적률 값이 미기재가 아닌 0으로 기재되는 오류가 존재하였다. 이는 미기재와 '0' 기재를 데이터 상에서 별도로 구분하지 않고 '0' 값 데이터를 상황에 따라 미기재 또는 '0' 기재로 판단하는 과정에서 발생한 오류일 것으로 추정된다. 한편 같은 단지 상가동의 경우 총괄표제부가 있는 경우에도 기재가 생략될 수 없는 건축면적이 미기재 처리되어 있어, 0 값의 미기재 판단이 건축물대장 상 미기재 가능 항목에 한정되지 않고 모든 수치 데이터에서 '0' 기재를 미기재 의미로 사용하고 있음을 추정할 수 있다.

[표 3-5] 사용승인일 2011년 공동주택 사례

관리_건축물대장_PK	동명	대지면적	건축면적	건폐율	연면적	용적률산정연면적	용적률
10821100199335	101동	표제부					
		0	7235.08	0	23199.06	22728.84	0
		총괄표제부					
		69432.00	20105.66	28.96	169706.83	113575.15	163.58

출처: 연구진 작성

집합건축물대장(표제부, 갑)

(2쪽 중 제1쪽)

건물ID	고유번호	명칭	호수/가구수/세대수
			0호/0가구/209세대

대지위치	지번	도로명주소	

※대지면적	연면적	※지역	※지구	※구역
0 ㎡	23,199.06 ㎡			

건축면적	용적률 산정용 연면적	주구조	주용도	층수
7,235.08 ㎡	22,728.84 ㎡	철근콘크리트구조	공동주택(아파트)	지하: 1층, 지상: 26층

※건폐율	※용적률	높이	지붕	부속건축물	
0 %	0 %	78.5 m	(철근)콘크리트평지붕		동 ㎡

※조경면적	※공개 공지/공간 면적	※건축선 후퇴면적	※건축선 후퇴거리
㎡	㎡	㎡	m

건축물 현황

구분	층별	구조	용도	면적(㎡)
주1	지1	철근콘크리트구조	설비배관공간	470.22
주1	1층	철근콘크리트구조	계단실,승강기,홀	492.26
주1	2층	철근콘크리트구조	아파트	1,260.35
주1	3층	철근콘크리트구조	아파트	1,823.16

건축물 현황

구분	층별	구조	용도	면적(㎡)
주1	4층	철근콘크리트구조	아파트	1,823.16
주1	5층	철근콘크리트구조	아파트	1,823.16
주1	6층	철근콘크리트구조	아파트	1,428.79
주1	7층	철근콘크리트구조	아파트	1,428.79

표제부(101동)

집합건축물대장(표제부, 갑)

(2쪽 중 제1쪽)

건물ID	고유번호	명칭	호수/가구수/세대수
			11호/0가구/0세대

대지위치	지번	도로명주소	

※대지면적	연면적	※지역	※지구	※구역
0 ㎡	1,457.58 ㎡			

건축면적	용적률 산정용 연면적	주구조	주용도	층수
㎡	949.67 ㎡	철근콘크리트구조	제1,2종근린생활시설	지하: 2층, 지상: 1층

※건폐율	※용적률	높이	지붕	부속건축물	
0 %	0 %	5.5 m	(철근)콘크리트평지붕		동 ㎡

※조경면적	※공개 공지/공간 면적	※건축선 후퇴면적	※건축선 후퇴거리
㎡	㎡	㎡	m

건축물 현황

구분	층별	구조	용도	면적(㎡)
주12	지2	철근콘크리트구조	주차장	427.14
주12	지2	철근콘크리트구조	전기실,발전기실	44.67
주12	지2	철근콘크리트구조	계단실	36.1
주12	1층	철근콘크리트구조	제1,2종근린생활시설	949.67

건축물 현황

구분	층별	구조	용도	면적(㎡)
		- 이하여백 -		

표제부(상가동)

[그림 3-4] 사용승인일 2011년 공동주택(총괄표제부 PK: 10821100198694)의 표제부

출처: 세움터(https://www.eais.go.kr/)에서 연구진 발급 후 강조 내용 작성

앞선 사례에 대한 검토를 통하여 수치 데이터의 경우 0 값의 해석이 총괄표제부 여부에 따라 달라지는 경우가 있음을 확인할 수 있다. 총괄표제부가 없어서 이의 영향이 없는 경우에도 유사한 오류가 발생하는지를 검토하기 위하여, 총괄표제부 유무에 따른 면적 관련 데이터의 차이를 검토하였다. 이외에도 주건축물과 부속건축물의 경우에 면적 관련 데이터의 기재 여부 및 오류가 다르게 나타나는지 검토하였다. 이를 위하여 총괄표제부의 유무 및 주·부속 건축물별로 면적 관련 데이터의 평균을 비교하여 차이 여부를 검토하였다. 그 결과, 총괄표제부가 없는 주건축물에서 평균 건폐율이 12,797.1%로 나타나는 등 이상값의 영향이 큰 것으로 나타났다. 앞서 건폐율의 중앙값은 4%에 불과하나 최대값이 1,000,000,000%에 달한 점을 고려하면 이러한 현상은 극소수 오류에서 기인하는 것으로 추정된다.

[표 3-6] 총괄표제부 유무 및 주·부속 건축물(표제부)별 면적 관련 데이터의 평균

총괄표제부 유무	주·부속 구분	건축면적	건폐율	연면적	용적률산정연면적	용적률	지상층수	지하층수	부속건축물 수	부속건축물 면적	총 동 연면적
무	주	154.6	12797.1	405.2	328.1	58.0	1.9	0.2	0.2	9.6	602.5
	부속	53.5	1.1	70.7	61.1	1.2	0.9	0.0	0.1	11.7	75.4
유	주	542.4	4.4	1285.8	1383.5	9.3	2.6	0.1	0.2	504.5	1796.7
	부속	264.0	0.3	1789.2	226.6	0.7	0.9	0.3	0.5	1224.4	1715.1

출처: 연구진 작성

이어 주건축물이 건축물 현황 통계의 집계 대상이며, 건축물 데이터 품질에서의 중요도도 큰 점을 고려하여 주건축물만을 대상으로 데이터를 검토하였다. 분석 데이터에서 주/부속 구분 코드가 0(주건축물)으로 기재된 레코드 수는 7,364,967건으로, 2024년 건축물 현황 통계 7,421,603동과 비교할 때 다소 작은 수치이다. 이러한 차이는 여러 가지 오류에서 기인할 수 도 있으나, 단순히 일부 비공개 대상 데이터(보안시설 등)가 개방데이터에서 제외된 것이 원인일 수 있다.

개별 사례를 확인하기 위하여 주건축물 중 총괄표제부가 있는 건축물 5동과 총괄표제부가 없는 대지 내 단독 건축물 5동에 대하여 면적 관련 데이터, 건축물대장 현황과 실제 건축물 현황을 비교하였다. 총괄표제부가 있는 건축물의 5동 중 4개동에서 대지면적, 건폐율, 용적률이 '0'으로 나타나, 총괄표제부가 있는 경우 대지면적, 건폐율, 용적률을 기재하지 않는 경우가 대다수인 한편, 총괄표제부가 있는 경우에도 해당 항목이 기재되는 경우도 있는 것으로 나타났다. 한편 총괄표제부가 없는 대지 내 단독 건축물의 경우 면적 관련 정보가 모두 기입되어 있었다.

[표 3-7] 총괄표제부 유무에 따른 면적관련 데이터 현황

총괄표제부 유무	대지면적	건축면적	건폐율	연면적	용적률산정연면적	용적률
유	0	187.6	0	367.6	367.6	0
	644	134.28	20.85	133.38	133.38	20.71
	0	342.41	0	399.61	342.41	0
	0	324.04	0	532	478.7	0
	0	102.24	0	196.32	196.32	0
무	333	83.1	24.95	83.1	83.1	24.95
	786	134.33	17.09	324.94	324.94	41.341
	181	89.93	49.69	179.86	179.86	99.37
	1350	96	7.11	96	96	7.11
	264	93.96	35.59	149.64	149.64	56.68

출처: 연구진 작성

총괄표제부가 없는 대지 내 단독 주건축물의 오류 현황을 자세히 검토하기 위하여 추가로 20동 사례를 추출하여 검토한 결과, 대지면적, 건폐율, 용적률 '0' 기재 사례를 확인할 수 있었다. 대지 내 단독

건축물의 경우 해당 항목이 미기재될 수 없으므로, 해당 항목이 0으로 기재될 수 있으며 이 경우 대지면적 등 항목 오류로 판단할 수 있다.

부속건축물 수와 면적이 건축물대장에 기재되며, 해당 주건축물의 연면적에 부속건축물의 면적을 합산한 총 동 연면적을 기재하고 있다. 그러나 많은 사례에서 부속건축물 면적이 기재되어 있음에도 해당 주건축물 연면적과 총 동 연면적이 동일한 수치로 기재되거나, 총 동 연면적이 아예 0으로 기재되는 등 오류가 다수 발견되었다. 또한 건축면적, 연면적이 0으로 기재되고, 용적률 산정 연면적만 정상적으로 기재된 경우도 발견되었다.

[표 3-8] 총괄표제부 없는 단독건축물의 면적 및 층수 관련 데이터 현황

관리_건축물대장_PK	대지면적	건축면적	건폐율	연면적	용적률산정연면적	용적률	지상층수	지하층수	부속건축물 수	부속건축물 면적	총 동 연면적
119315983	0	49.58	0	49.58	49.58	0	1	0	0	0	49.58
1081113335	0	27.04	0	27.04	27.04	0	1	0	0	0	27.04
10000000000000489111	460	77.04	16.75	77.04	77.04	16.75	1	0	0	0	77.04
1230139935	0	33.2	0	33.2	33.2	0	1	0	0	0	33.2
1089162437	0	57.52	0	195.92	114.04	0	2	1	1	1	0
1142111879	4569	25	0.547	25	25	0.547	1	0	1	10.9	25
10311100177445	127.6	83.64	65.55	126.28	126.28	98.97	2	0	0	0	126.28
125212000022070	189	96.2	50.96	192.4	192.4	101.8	2	0	0	0	192.4
106818883	0	80.98	0	96.18	96.18	0	2	0	2	10.57	96.18
1008122597	136.9	65.91	48.14	199.89	131.82	96.29	2	1	0	0	199.89
12321100210411	420.6	205.15	48.78	197.68	197.68	47	1	0	0	0	197.68
10431100216597	2158	430.48	19.95	429.54	429.54	19.9	1	0	1	119.04	429.54
121512066	248	54.76	22.08	54.76	54.76	22.08	1	0	0	0	54.76
1021112562	89	50.22	56.42	99.44	99.44	111.73	2	0	0	0	99.44
1207125697	278	59.4	21.37	59.4	59.4	21.37	1	0	0	0	0
1261156296	302	166.98	55.29	659.52	659.52	218.38	5	0	0	0	659.52
1120149958	292	112	38.36	112	112	38.36	1	0	0	0	112
12761100396101	448.6	268.28	59.8	836.23	836.23	186.41	4	0	0	0	836.23
1067112692	0	0	0	135.9	124.38	0	2	1	2	2.61	135.9
112514656	1310	523.05	39.93	1046.1	1046.1	79.85	2	0	0	0	1046.1

출처: 연구진 작성

다음으로 비교적 최근 건축된 건축물 중에서 연면적이 음수인 사례를 검토하여 그 원인을 분석하였다. 사용승인일 기준 최근 10년 건축한 건축물의 경우에도 연면적 등이 음수인 오류 사례가 존재하였다. 충청북도 괴산군 장연면 광진리에 위치한 건축물의 경우 건축물대장에서 건축면적, 연면적, 용적률 산정용 연면적, 건폐율, 용적률이 모두 음수로 나타나고 있다. 이는 2008년 개축하면서 동일한 건축물에 대하여 별개의 대장이 생성되어 한 건축물에 두 개의 건축물대장이 존재하게 된 사례로, 로드뷰를 통해 현재 해당 대지에 건축물이 1동만 존재함을 확인할 수 있다. 면적 관련 항목이 음수인 건축물대장은 2020년 개축 이전 현황인 시멘트벽돌구조, 콘크리트 평슬래브 지붕 등이 기재되어 있고, 현존하는 건축물에 해당하는 내용은 별도 건축물대장에 기재되어 있다.

2018.05. 기준 로드뷰

2024.09. 기준 로드뷰

[그림 3-5] 충청북도 괴산군 장연면 광진리 건축물대장 오류 사례 로드뷰
출처: 카카오맵 로드뷰(https://map.kakao.com/)(접속일: 2025.09.04.)

일반건축물대장(갑)

(2쪽 중 제1쪽)

건물ID	고유번호	명칭	호수/가구수/세대수
		주1	0호/0가구/0세대

대지위치	지번	도로명주소	

※대지면적	연면적	※지역	※지구	※구역
302 ㎡	-132.16 ㎡			

건축면적	용적률 산정용 연면적	주구조	주용도	층수
-93.66 ㎡	-132.16 ㎡	시멘트벽돌구조	제1종근린생활시설	지하: 층, 지상: 2층

※건폐율	※용적률	높이	지붕	부속건축물
-31.01 %	-43.76 %	6.9 m	콘크리트평슬래브	동 ㎡

※조경면적	※공개 공지·공간 면적	※건축선 후퇴면적	※건축선후퇴 거리
㎡	㎡	㎡	m

변동사항				그 밖의 기재사항
변동일	변동내용 및 원인	변동일	변동내용 및 원인	
2020.8.3.	[개축]-132.16㎡ 개축 - 이하여백 -			

광진리(개축) 음수 표기 사례

일반건축물대장(갑)

(2쪽 중 제1쪽)

건물ID	고유번호	명칭	호수/가구수/세대수
		주1동	0호/0가구/0세대

대지위치	지번	도로명주소	

※대지면적	연면적	※지역	※지구	※구역
302 ㎡	114.67 ㎡			

건축면적	용적률 산정용 연면적	주구조	주용도	층수
114.67 ㎡	114.67 ㎡	경량철골구조	제1종근린생활시설	지하: 층, 지상 1층

※건폐율	※용적률	높이	지붕	부속건축물
37.97 %	37.97 %	5.2 m	경량철골트러스위샌드위치철판넬	동 ㎡

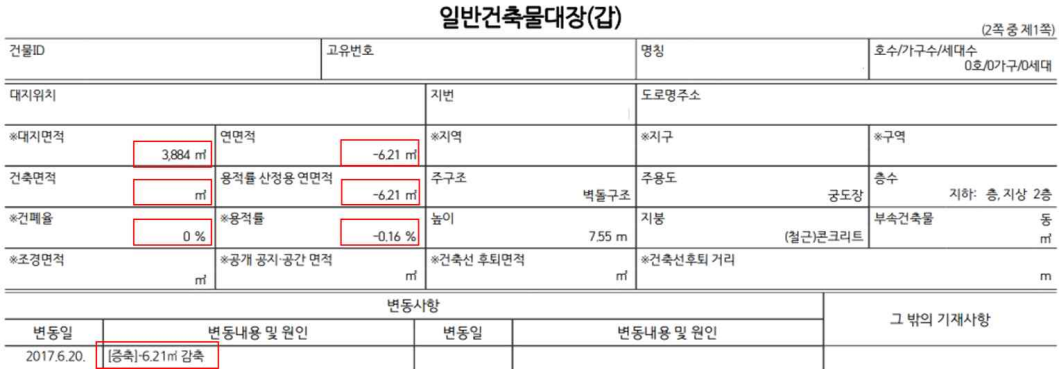
※조경면적	※공개 공지·공간 면적	※건축선 후퇴면적	※건축선후퇴 거리
㎡	㎡	㎡	m

변동사항				그 밖의 기재사항
변동일	변동내용 및 원인	변동일	변동내용 및 원인	
2020.8.3.	사용승인일자(2020.8.3.),[개축]114.67㎡ 개축, 소유자등록(균형개발과 개축신고번호: 2019-1호) - 이하여백 -			

광진리 정상 건축물

[그림 3-6] 충청북도 괴산군 장연면 광진리 건축물대장 오류 사례
출처: 세움터(https://www.eais.go.kr/)에서 연구진 발급 후 강조 내용 작성

2017년 증축 이후 건축물대장과 별개로 감축만 등재하여 연면적이 음수로 기재되고, 기존 대장에는 면적이 음수로 기재된 사례 또한 존재하였다. 충청북도 괴산군 괴산읍 서부리에 위치한 건축물의 경우 2001년 완공 이후 2004년에 증축하였고, 2011년 건축물대장 기초자료 정비에의거 건폐율을 직권변경하였다. 이후 2017년 증축 과정에서 별개의 감축만 표기된 건축물대장이 생성되었고, 기존 대장에는 면적에 음수 표기가 기재된 사례이다.



일반건축물대장(갑)

(2쪽 중 제1쪽)

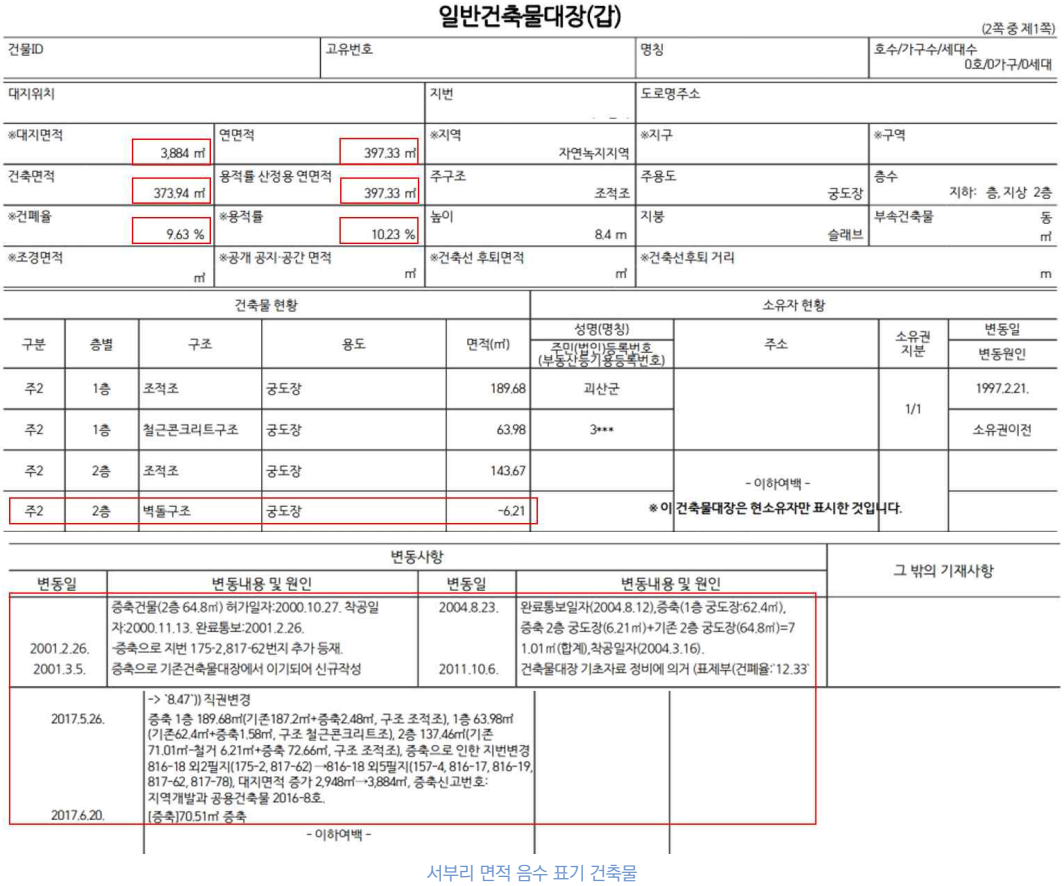
건물ID	고유번호	명칭	호수/가구수/세대수 0호/0가구/0세대
대지위치	지번	도로명주소	

※대지면적	연면적	※지역	※지구	※구역
3,884 ㎡	-6.21 ㎡			
건축면적	용적률 산정용 연면적	주구조	주용도	층수
㎡	-6.21 ㎡	벽돌구조	궁도장	지하: 층, 지상 2층
※건폐율	※용적률	높이	지붕	부속건축물 동
0 %	-0.16 %	7.55 m	(철근)콘크리트	㎡
※조경면적	※공개 공지·공간 면적	※건축선 후퇴면적	※건축선후퇴 거리	
㎡	㎡	㎡	m	

변동사항

변동일	변동내용 및 원인	변동일	변동내용 및 원인	그 밖의 기재사항
2017.6.20.	[증축]-6.21㎡ 감축			

서부리(증축) 음수 표기 대장 생성 사례



일반건축물대장(갑)

(2쪽 중 제1쪽)

건물ID	고유번호	명칭	호수/가구수/세대수 0호/0가구/0세대
대지위치	지번	도로명주소	

※대지면적	연면적	※지역	※지구	※구역
3,884 ㎡	397.33 ㎡	자연녹지지역		
건축면적	용적률 산정용 연면적	주구조	주용도	층수
373.94 ㎡	397.33 ㎡	조적조	궁도장	지하: 층, 지상 2층
※건폐율	※용적률	높이	지붕	부속건축물 동
9.63 %	10.23 %	8.4 m	슬래브	㎡
※조경면적	※공개 공지·공간 면적	※건축선 후퇴면적	※건축선후퇴 거리	
㎡	㎡	㎡	m	

건축물 현황 / 소유자 현황

구분	층별	구조	용도	면적(㎡)	성명(명칭) 주민(법인)등록번호 (부동산등기용등록번호)	주소	소유권 지분	변동일 변동원인
주2	1층	조적조	궁도장	189.68	괴산군		1/1	1997.2.21.
주2	1층	철근콘크리트구조	궁도장	63.98	3***			소유권이전
주2	2층	조적조	궁도장	143.67		- 이하여백 -		
주2	2층	벽돌구조	궁도장	-6.21	※ 이 건축물대장은 현소유자만 표시한 것입니다.			

변동사항

변동일	변동내용 및 원인	변동일	변동내용 및 원인	그 밖의 기재사항
	증축건물(2층 64.8㎡) 허가일자:2000.10.27. 착공일자:2000.11.13. 완료통보:2001.2.26.	2004.8.23.	완료통보일자(2004.8.12),증축(1층 궁도장:62.4㎡), 증축 2층 궁도장(6.21㎡)+기존 2층 궁도장(64.8㎡)=71.01㎡(합계),착공일자(2004.3.16).	
2001.2.26.	-증축으로 지번 175-2,817-62번지 추가 등재.			
2001.3.5.	증축으로 기존건축물대장에서 이기되어 신규작성	2011.10.6.	건축물대장 기초자료 정비에 의거 (표제부(건폐율:`12.33`->`8.47`)) 직권변경	
2017.5.26.	증축 1층 189.68㎡(기존187.2㎡+증축2.48㎡, 구조 조적조), 1층 63.98㎡(기존62.4㎡+증축1.58㎡, 구조 철근콘크리트조), 2층 137.46㎡(기존 71.01㎡-철거 6.21㎡+증축 72.66㎡, 구조 조적조), 증축으로 인한 지번변경 816-18 외2필지(175-2, 817-62) →816-18 외5필지(157-4, 816-17, 816-19, 817-62, 817-78), 대지면적 증가 2,948㎡→3,884㎡, 증축신고번호: 지역개발과 공용건축물 2016-8호.			
2017.6.20.	[증축]70.51㎡ 증축			
	- 이하여백 -			

서부리 면적 음수 표기 건축물

[그림 3-7] 충청북도 괴산군 괴산읍 서부리 건축물대장 오류 사례

출처: 세움터(https://www.eais.go.kr/)에서 연구진 발급 후 강조 내용 작성

2004년 신축한 대구광역시 달성군 화원읍 본리리의 건축물은 사용승인 후 2009년 용도변경, 2011년 대장 정비 등 있었으나, 2012년 용도변경이 있었지만, 최종적으로 건축면적, 연면적 미기재, 용적률과 용적률 산정용 연면적에 음수가 기재되었다. 다만 검토 결과 층별개요 표기는 정상으로 판단된다.

일반건축물대장(갑)

(2쪽 중 제1쪽)

건물ID	고유번호	명칭	호수/가구수/세대수
			0호/6가구/0세대

대지위치	지번	도로명주소	

※대지면적	연면적	※지역	※지구	※구역
330 ㎡	0 ㎡	제2종일반주거지역		
건축면적	**용적률 산정용 연면적**	**주구조**	**주용도**	**층수**
㎡	-9.24 ㎡	철근콘크리트구조	제2종근린생활시설	지하: 1층, 지상: 3층
※건폐율	※용적률	높이	지붕	부속건축물
0 %	-2.8 %	11.5 m	슬래브	동 ㎡
※조경면적	※공개 공지·공간 면적	※건축선 후퇴면적	※건축선후퇴 거리	
㎡	㎡	㎡		m

건축물 현황

구분	층별	구조	용도	면적(㎡)
주1	지1층	철근콘크리트구조	일반음식점	141.29
주1	1층	철근콘크리트구조	사무소	122.01
주1	1층	철근콘크리트구조	다가구주택(1가구)	57.15
주1	1층	철근콘크리트구조	주차장	9.24

소유자 현황

성명(명칭) / 주민(법인)등록번호 (부동산등기용등록번호)	주소	소유권 지분	변동일 / 변동원인
강한철			2009.6.23.
			등기명의인표시변경
	- 이하여백 -		
	※이 건축물대장은 현소유자만 표시한 것입니다.		

변동사항

변동일	변동내용 및 원인	변동일	변동내용 및 원인
2004.3.4.	2004.02.25 신규작성(신축)		-> '지1')) 직권변경
2009.6.15.	2009.06.15 용도변경(1층 사무소 121.63㎡ →다가	2012.12.27.	2012.12.27 용도변경(지상1층 단독주택(1가구) 188.
	구주택) : 2009-종합민원과-용도변경신고-18		40㎡ →단독주택(1가구) 57.15㎡, 주차장 9.24㎡, 사
2011.4.13.	건축물대장 기초자료 정비에 의거 (층별개요(층번호명:'지하1층'		무소 122.01㎡:2012-종합민원과-용도변경허가-13)

그 밖의 기재사항

본리리 음수 및 '0' 표기 건축물

[그림 3-8] 대구광역시 달성군 화원읍 본리리 건축물대장 오류 사례
출처: 세움터(https://www.eais.go.kr/)에서 연구진 발급 후 강조 내용 작성

검토 결과를 종합하면, 대지면적, 건축면적, 연면적, 용적률 산정 연면적, 건폐율, 용적률 등 면적 관련 데이터를 포함한 모든 수치 데이터에 대하여 '0' 기재와 미기재를 구분하지 않고 있다. 미기재 여부를 상황에 따라 구분하여 건축물대장에 표출하나 그 구분에서도 오류가 발생하고 있다. 음수 기재 사례는 개축 등 정보 변동사항 발생 시 건축물대장 생성 등 절차 오류로 인하여 발생하는 것으로 추정된다. 수치 데이터에 0으로 기재된 경우는 모두 미기재(null)로 처리하고, 미기재 및 오류 사례를 고려하여 건축물 데이터 검증 규칙을 엄밀하게 정의할 필요가 있을 것으로 판단된다.

3) 기존 정비규칙 적용한 오류 검증

■ 기존 정비규칙 기반 검증 방법론

- 면적 데이터 관련 기존 건축물대장 정비규칙

건축물대장 면적 데이터 품질 고도화를 위하여 관련된 기존 정비규칙을 검토하고, 분석 데이터에 기존 정비규칙을 적용하여 현재 오류 현황을 파악하였다. 또한 지속적으로 자동화된 오류 탐지가 가능하도록 검증 방법론을 구축하였다. 정비규칙이 가장 상세하게 도출된 2022년 정비규칙을 기준으로, 건물동 단위 일반건축물대장 및 표제부의 대지면적, 건축면적, 건폐율, 용적률 등 면적 관련 데이터 대상 규칙을 검토하였다.

건축물대장 정비규칙 중 동 단위 면적 관련 데이터와 관련된 정비규칙은 총 4개로, 각각 대지면적, 건축면적, 건폐율, 용적률 데이터에 대한 검증 내용을 담고 있다. 건물동 단위 검증 규칙이지만 점검 대상 건수가 동일하지 않고 각 규칙마다 다른데, 따라서 오류 건수를 직접 비교하기는 어려우며, 2024년 말 현재 기준 분석 데이터에 대한 검증 결과와도 직접 비교가 어렵다. 따라서 해당 정비규칙 오류 현황은 오류 건수를 점검 대상 건수로 나눈 오류율을 기반으로 판단하고, 현재 기준 분석 데이터에서 나타난 오류율을 이와 비교하여 2022년 이후 건축물대장 정비에 따른 품질 개선 현황을 검토하였다.

[표 3-9] 면적 관련 데이터 대상 정비규칙 현황

순번	업무규칙명	오류건수	점검대상건수	오류율
14	(대지면적) 일반건축물대장 및 표제부 내 대지면적의 값이 건축면적보다 작은 경우의 데이터 검증(정확성)	1,136	612,719	0.19%
16	(건축면적) 일반건축물대장 및 표제부의 바닥면적의 합이 가장 큰 면적과 표제부의 건축면적이 다른 데이터 검증(정확성)	1,334,538	7,952,243	16.78%
19	(건폐율) 일반건축물대장 및 표제부 건폐율 계산의 정확성 → 소수점 이하 처리에 대한 명확한 법규 처리 없어 반올림 자리수가 불분명함(숫점 이하 절사처리)	54,486	7,324,326	0.74%
20	(용적률) 일반건축물대장 및 표제부 용적율 계산의 정확성	723,474	5,212,430	13.88%

출처: 세움터 내부자료

규칙 14는 일반건축물대장 및 표제부에 기재된 대지면적과 건축면적의 관계를 점검하여, 건축면적이 대지면적을 초과하는 경우를 오류로 식별하는 규칙이다. 건축면적은 건축물 지상부의 수평투영면적으로, 일반적으로 대지 경계선 내에서 건축이 이루어지고, 추가로 건폐율 규제에 따라 건축면적이 대지면적의 일정 비율 이하로 제한되므로, 건축면적이 대지면적을 초과하는 경우 오류 가능성이 높다고 판단할 수 있다. 이러한 검증은 모든 일반건축물대장 및 표제부에 대해 가능하나, 2022년 당시 점검 대상 건수는 전체 건축물이 아닌 약 60만 동으로 계획되어, 이 중 약 1천 건의 오류를 탐지하였다. 오류율로 따지면 0.19%로 매우 적은 오류율을 보였다.

규칙 16은 일반건축물대장 및 표제부에 기재된 "바닥면적의 합이 가장 큰 면적"이 건축면적과 다른 경우를 검증하고 있다. 규칙 16의 업무규칙명 자체로는 해석상 모호한 점이 있다. 건축물 바닥면적의 합

은 연면적인데, “연면적이 가장 큰 면적”이라는 표현은 연면적의 정의상 성립하지 않는다. 관계 전문가 자문 결과 “바닥면적의 합” 앞에 ‘층별’이 생략된 것으로 추정된다. 「건축물대장 기재 및 관리 등에 관한 규칙」 별표는 건축물 현황의 작성에 관하여 구분별(주/부속건축물), 층별, 구조별, 용도별로 면적을 달리 작성하도록 규정하고 있다. 즉 층별개요 테이블의 데이터는 실제로는 구조별, 용도별로도 구별된 개요이며, 이에 따라 각 층별로 구분·구조·용도별 바닥면적의 합산을 통하여 각 층의 실제 면적을 구하고, 이를 건축면적과 비교할 필요가 발생한다. 규칙 16 정비 결과인 오류율 16.78%은 건축면적과 바닥면적에 대한 산정 기준이 상이한 점을 고려하지 않고, 대소 관계가 아니라 일치 여부를 검증하였기 때문에 높게 나타난 것으로 판단된다. 이러한 배경을 폭넓게 고려하여, 본 연구에서는 동별 면적 데이터의 검토가 될 수 있도록 규칙 16번을 일부 변경하여 연면적이 건축면적보다 작은 경우를 검증 대상으로 설정하였다. 다만 이 경우 규칙 16 정비 결과와 직접 비교는 어렵다.

규칙 19는 건폐율 계산의 정확성을 검증한다. 건폐율은 대지면적에 대한 건축면적의 비율로 정의된다. 데이터 품질 관점에서 모호한 지점은 소수점 이하 반올림 처리 규칙이 엄밀하게 정의되지 않아, 일관되지 않게 적용되고 있는 점이다. 대지면적의 예를 들면, 건축물대장의 기재 및 관리 등에 관한 규칙 별표 1 ‘건축물대장의 작성방법’에서 “산정한 후 소수점 둘째 자리에서 반올림해 소수점 첫째 자리까지 적을 것”을 규정하고 있다. 건폐율에 대해서는 그러한 규칙이 제시되지 않았으나, 임의로 소수점 이하 자리에서 반올림 처리를 하는 경우가 있다. 해당 규칙은 이를 검증하기 위하여 대지면적과 건축면적의 비율과 기재된 건폐율 값을 비교하여 계산의 정확성을 검증한다. 건축물대장 정비 당시 오류율은 0.74%로 낮은 편이었다.

규칙 20은 용적률 계산의 정확성을 검증한다. 규칙 19의 건폐율 검증과 유사하게 대지면적 대비 (용적률 계산용) 연면적의 비율과 기재된 용적률 값을 비교하여 계산의 정확성을 검증한다. 다만 점검 대상 건수가 전체 건축물(약 7백만 동)이 아닌 일부로 추정되며, 오류 건수 및 오류율이 건폐율과 달리 높은 편으로, 직접 비교는 어려우나 전국 건축물 대상 검증의 중요성이 높은 편으로 판단된다.

- 오류 판별 기준 정의

건축물대장은 수치 데이터를 제한된 자리수로 표출하고 있다. 건축물대장 데이터에서는 매우 작은 오차가 존재하더라도, 대장 기준으로는 오류가 발생하지 않을 수 있다. 따라서 정상적으로 관리되고 있는 데이터에서도 작은 오차가 존재할 수 있으며, 이러한 경우 오류로 판별하지 않도록 주의가 필요하다. 또한, 분석 데이터는 수치 데이터를 부동소수점 형식으로 저장하고 있다. 부동소수점 형식은 null 값 처리 및 다양한 라이브러리에서 기본 지원하는 장점이 있으나, 이진법을 사용하고 있어 십진법의 정수 및 소수를 정확하게 표현하지 못하고 매우 작은 오차가 발생하는 부동소수점 오류가 존재한다. 이러한 점을 고려할 때, 단순히 정확한 일치 여부를 오류 판별에 사용하기보다는 일정 정도 오차범위를 고려한 오류 판별 기준을 설정할 필요가 있다.

이를 위하여 총괄표제부의 연면적과, 표제부의 연면적 합, 그리고 층별개요의 면적 합의 3개 면적 쌍 607,144건을 대상으로 오차 범위를 분석하고, 이를 통하여 오류로 판단하지 않는 정상적인 오차 범위를 설정하고자 하였다. 총괄표제부 연면적에서 표제부 연면적을 뺀 차이를 구간으로 나누어 정리하였

다. 이때 구간을 로그 스케일로 0에 가까울수록 작은 구간을 설정하여, 매우 작은 차이의 전반적인 경향을 파악하고자 하였다. 구간 범위가 각 구간마다 다르기 때문에, 비교를 위하여 구간 범위 내 오차율을 산정하였다. '오차구간'은 총괄표제부와 표제부 두 테이블간의 연면적 차이를 나타내는 구간이며, '빈도'는 607,144건의 데이터를 모수로 연면적의 차이가 해당 구간에 얼마나 속해있는지를 나타낸다. '백분율/구간범위는' 연면적 차 데이터가 얼마나 해당 구간에 밀집되어 있는지에 대한 밀도이다. 즉, 같은 백분율이라도 구간의 폭이 좁으면 밀도가 높게 나타나며, 넓으면 밀도가 낮게 나타난다.

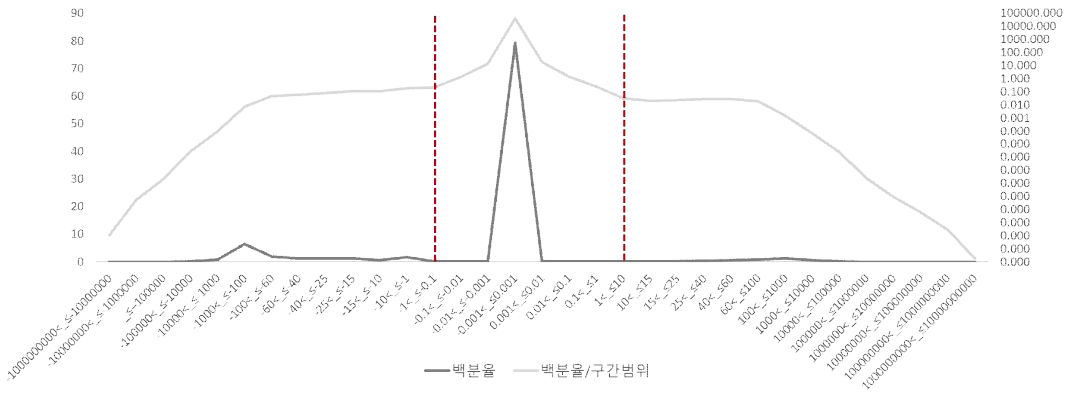
[표 3-10] 총괄표제부와 표제부 연면적 일치 구간

오차구간	구간범위	빈도	백분율	백분율/구간범위
_≤-10000000000	∞	0	0.000	
-10000000000<_≤-1000000000	9000000000	0	0.000	
-1000000000<_≤-10000000	990000000	7	0.001	0.000
-10000000<_≤-1000000	9000000	31	0.005	0.000
_≤-100000	900000	114	0.019	0.000
-100000<_≤-10000	90000	1,407	0.232	0.000
-10000<_≤-1000	9000	5,540	0.912	0.000
-1000<_≤-100	900	38,687	6.372	0.007
-100<_≤-60	40	12,075	1.989	0.050
-60<_≤-40	20	7,112	1.171	0.059
-40<_≤-25	15	7,271	1.198	0.080
-25<_≤-15	10	7,002	1.153	0.115
-15<_≤-10	5	3,419	0.563	0.113
-10<_≤-1	9	9,749	1.606	0.178
-1<_≤-0.1	0.9	1,079	0.178	0.197
-0.1<_≤-0.01	0.09	742	0.122	1.358
-0.01<_≤-0.001	0.009	780	0.128	14.274
-0.001<_≤0.001	0.002	481,793	79.354	39,676.996
0.001<_≤0.01	0.009	1,010	0.166	18.484
0.01<_≤0.1	0.09	739	0.122	1.352
0.1<_≤1	0.9	1,364	0.225	0.250
1<_≤10	9	1,731	0.285	0.032
10<_≤15	5	616	0.101	0.020
15<_≤25	10	1,421	0.234	0.023
25<_≤40	15	2,551	0.420	0.028
40<_≤60	20	3,310	0.545	0.027
60<_≤100	40	4,603	0.758	0.019
100<_≤1000	900	8,032	1.323	0.001
1000<_≤10000	9000	3,388	0.558	0.000
10000<_≤100000	90000	1,338	0.220	0.000
100000<_≤1000000	900000	132	0.022	0.000
1000000<_≤10000000	9000000	53	0.009	0.000
10000000<_≤100000000	90000000	32	0.005	0.000
100000000<_≤1000000000	900000000	15	0.002	0.000
1000000000<_≤10000000000	9000000000	1	0.000	0.000

출처: 연구진 작성

검토 결과, -0.001 초과 0.001 이하의 0에 수렴하는 구간에서의 빈도가 481,793건(79.354%)으로 가장 높게 나타났다. 이후 마이너스 오차의 폭이 넓어질수록 보합의 추세를 보이다가, -1 이하 구간에서 증가 추세를 보이며, 플러스 오차에서는 완만하게 증가하는 추세를 보인다.

다음으로 '백분율'과 '백분율/구간범위'를 히스토그램으로 나타내어 분석하였다. 히스토그램에서 주축은 백분율이며, 보조축은 백문율/구간범위를 로그 눈금 단위(log scale)로 표현한 것이다. 백분율/구간범위의 값은 극단으로 작은 값부터 아주 큰 값까지 분포되어 있으므로, 데이터의 상대적인 밀도(=백분율/구간범위)의 크기 차이를 효과적으로 시각화하기 위함이다. 일반 선형 눈금에서는 작은 값이 시각적으로 거의 표현되지 않거나, 큰 값만 두드려져 전체 분포 해석이 어렵다. 따라서 로그 눈금 단위를 활용하여 상대적인 비율 차이를 명확히 보고자 하였다.



[그림 3-9] 총괄표제부와 표제부 연면적 차 히스토그램
출처: 연구진 작성

검토 결과, 대다수는 0을 포함하는 -0.001 초과 0.001 이하 구간에서 밀집이 두드러지며, 백분율/구간범위로 봤을 때, -0.1 이하와 1초과 구간에서 단위구간당 백분율(밀도)가 점진적으로 감소하는 경향이 확인되었다. 이는 데이터의 -0.1~1 사이의 구간에서 집중적으로 분포하고 있음을 나타내며, 그 외 구간에서는 분포 밀도가 희박해짐을 의미한다. 점선은 -0.1에서 1 사이의 구간을 나타내고 있는데, 해당 구간 안에 높은 밀도 영역이 모두 포함되고 있는 것을 확인할 수 있다.

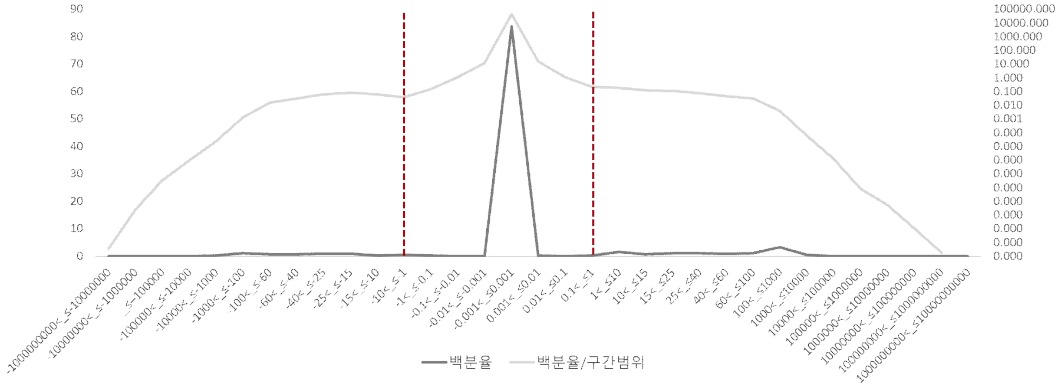
다음으로 표제부 연면적 합과 층별개요 면적 합을 비교한 결과, -0.001 초과 0.001 이하의 0에 수렴하는 구간에서의 빈도는 508,248건(83.711%)로 총괄표제부와 표제부 연면적의 차(79.354%)보다도 높게 나타났다. -1 이하 구간과 1 초과 구간에서 급격히 증가하는 추세를 보인다. 히스토그램에서도 비슷한 추이가 나타났다. 0에 수렴하는 -0.001 초과 0.001 이하 구간에서 밀집이 두드러지며, 백분율/구간범위로 봤을 때, -1 이하와 0.1 초과의 구간에서 단위구간당 밀도가 점진적으로 감소하는 경향이 확인되었다. 이는 데이터의 -1~0.1 사이의 구간에서 집중적으로 분포하고 있음을 나타낸다.

[표 3-11] 표제부와 층별개요 연면적 일치 구간

오차구간	구간범위	빈도	백분율	백분율/구간범위
_≤-10000000000	∞	0	0.000	
-10000000000<_≤-1000000000	9000000000	0	0.000	
-1000000000<_≤-10000000	990000000	2	0.000	0.000
-10000000<_≤-1000000	9000000	14	0.002	0.000
_≤-100000	900000	183	0.030	0.000
-100000<_≤-10000	90000	491	0.081	0.000

오차구간	구간범위	빈도	백분율	백분율/구간범위
-10000<_≤-1000	9000	1,212	0.200	0.000
-1000<_≤-100	900	7,212	1.188	0.001
-100<_≤-60	40	3,609	0.594	0.015
-60<_≤-40	20	3,533	0.582	0.029
-40<_≤-25	15	5,441	0.896	0.060
-25<_≤-15	10	5,015	0.826	0.083
-15<_≤-10	5	1,838	0.303	0.061
-10<_≤-1	9	2,086	0.344	0.038
-1<_≤-0.1	0.9	781	0.129	0.143
-0.1<_≤-0.01	0.09	579	0.095	1.060
-0.01<_≤-0.001	0.009	633	0.104	11.584
-0.001<_≤0.001	0.002	508,248	83.711	41,855.639
0.001<_≤0.01	0.009	817	0.135	14.952
0.01<_≤0.1	0.09	608	0.100	1.113
0.1<_≤1	0.9	1,183	0.195	0.216
1<_≤10	9	9,820	1.617	0.180
10<_≤15	5	3,514	0.579	0.116
15<_≤25	10	6,543	1.078	0.108
25<_≤40	15	6,823	1.124	0.075
40<_≤60	20	5,558	0.915	0.046
60<_≤100	40	7,156	1.179	0.029
100<_≤1000	900	20,175	3.323	0.004
1000<_≤10000	9000	3,300	0.544	0.000
10000<_≤100000	90000	691	0.114	0.000
100000<_≤1000000	900000	43	0.007	0.000
1000000<_≤10000000	9000000	29	0.005	0.000
10000000<_≤100000000	90000000	6	0.001	0.000
100000000<_≤1000000000	900000000	1	0.000	0.000
1000000000<_≤10000000000	9000000000	0	0.000	0.000

출처: 연구진 작성



[그림 3-10] 표제부와 층별개요 연면적 차 히스토그램
출처: 연구진 작성

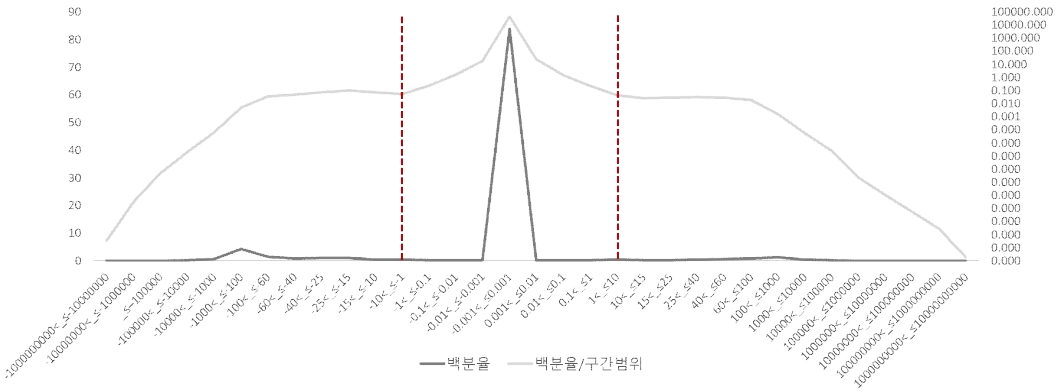
마지막으로 총괄표제부 연면적에서 층별개요 면적 합을 뺀 차이 구간은 표 3-12와 같다. -0.001 초과 0.001 이하의 0에 수렴하는 구간에서의 빈도는 508,720건(83.789%)로 총괄표제부와 표제부, 표제

부와 층별개요의 연면적의 차(각 79.354%, 83.711%) 대비 높게 나타났다. -1 이하 구간에서 급격히 증가 추세를 보이며, 플러스 오차에서는 완만하게 증가하는 추세를 보인다. 히스토그램에서는 0에 수렴하는 -0.001 초과 0.001 이하 구간에서 밀집이 두드러지며, 백분율/구간범위로 봤을 때, -1 이하와 1초과 구간에서 단위구간당 백분율(밀도)가 점진적으로 감소하는 경향이 확인되었다. 이는 데이터의 -1~1 사이의 구간에서 집중적으로 분포하고 있음을 나타내며, 그 외 구간에서는 분포 밀도가 희박해짐을 의미한다.

[표 3-12] 총괄표제부와 층별개요 연면적 일치 구간

오차구간	구간범위	빈도	백분율	백분율/구간범위
_≤-10000000000	∞	0	0	
-10000000000<_≤-1000000000	9000000000	0	0	
-1000000000<_≤-10000000	990000000	2	0.000	0.000
-10000000<_≤-1000000	9000000	18	0.003	0.000
_≤-100000	900000	255	0.042	0.000
-100000<_≤-10000	90000	1,104	0.182	0.000
-10000<_≤-1000	9000	3,298	0.543	0.000
-1000<_≤-100	900	26,286	4.329	0.005
-100<_≤-60	40	8,945	1.473	0.037
-60<_≤-40	20	5,567	0.917	0.046
-40<_≤-25	15	6,701	1.104	0.074
-25<_≤-15	10	5,964	0.982	0.098
-15<_≤-10	5	2,146	0.353	0.071
-10<_≤-1	9	3,022	0.498	0.055
-1<_≤-0.1	0.9	1,202	0.198	0.220
-0.1<_≤-0.01	0.09	876	0.144	1.603
-0.01<_≤-0.001	0.009	956	0.157	17.495
-0.001<_≤0.001	0.002	508,720	83.789	41,894.509
0.001<_≤0.01	0.009	1,403	0.231	25.676
0.01<_≤0.1	0.09	833	0.137	1.524
0.1<_≤1	0.9	1,248	0.206	0.228
1<_≤10	9	2,338	0.385	0.043
10<_≤15	5	793	0.131	0.026
15<_≤25	10	1,687	0.278	0.028
25<_≤40	15	2,805	0.462	0.031
40<_≤60	20	3,411	0.562	0.028
60<_≤100	40	4,651	0.766	0.019
100<_≤1000	900	8,197	1.350	0.002
1000<_≤10000	9000	3,172	0.522	0.000
10000<_≤100000	90000	1,318	0.217	0.000
100000<_≤1000000	900000	124	0.020	0.000
1000000<_≤10000000	9000000	54	0.009	0.000
10000000<_≤100000000	90000000	32	0.005	0.000
100000000<_≤1000000000	900000000	15	0.002	0.000
1000000000<_≤10000000000	9000000000	1	0.000	0.000

출처: 연구진 작성



[그림 3-11] 총괄표제부와 층별개요 연면적 차 히스토그램
출처: 연구진 작성

검토 결과, 소수점 이하의 근소한 차이와 라운딩 에러(반올림 오차) 등으로 오차범위 설정이 필요하다고 판단되었다. 3개 면적 쌍 검토를 통해 각 면적의 차가 1 ㎡ 이하인 경우 오류가 아니고 일치하는 것으로 설정하였다. 따라서 오류 판별은 1 ㎡ 오차범위를 설정하고 이를 벗어나는 경우에만 오류로 판단하였다. 건폐율과 용적률은 단위가 ㎡가 아니라 %이며, 1% 차이는 오차범위로서는 지나치게 크다고 판단되어 0.1%을 기준으로 사용하였다. 최종적으로 종합하면 오류의 판정은 면적의 경우 1 ㎡ 이상 크거나 작은 경우, 비율의 경우 0.1% 이상 크거나 작은 경우 오류로 판정하였다.

■ 기존 업무규칙 적용 검증 결과

- 검증 결과 통계
  - (규칙 14) 대지면적

대지면적의 판정은 표제부의 대지면적 또는 총괄표제부의 대지면적 보다 건축면적이 1㎡ 이상 클 경우 오류로 판정하였다. 즉 1㎡ 미만인 경우 유효하며, 건축면적이 존재하더라도 표제부의 대지면적 또는 총괄표제부의 대지면적에 값이 없거나, 반대로 표제부의 대지면적 또는 총괄표제부의 대지면적에 값이 있어도 건축면적에 값이 없으면 NULL로 처리된다.

대지면적의 오류율은 2022년 0.19%에 비해 0.13%로 적어졌다. 하지만 2022년 점검 대상 건수는 612,719건이며, 2024년 말에 해당하는 2025년 점검 대상 건수는 7,364,967건으로 모수가 달라 대소 관계에 대한 해석에는 주의가 필요하다. 큰 틀에서는 2022년 대비 오류율이 유사한 수준인 것으로 판단할 수 있다. 이러한 오류율은 대지면적 또는 건축면적 미기재를 의미하는 NULL의 건수가 2,164,467건이고, 이를 오류로 산정하지 않았음에도 2022년과 유사한 수준으로 나타난 것이다.

[표 3-13] 대지면적의 오류율

순번	2022년			2025년				
	오류건수	점검대상건수	오류율	NULL	유효건수	오류건수	점검대상건수	오류율
14	1,136	612,719	0.19%	2,164,467	5,190,727	9,773	7,364,967	0.13%

출처: 세움터 내부자료

- (규칙 16) 건축면적

건축면적의 오류 판정은 표제부의 건축면적이 연면적보다 1㎡ 이상 클 경우 오류로 판정하였다. 연면적은 각 층의 바닥면적 합계로 복층 이상 건축물에서는 건축면적보다 크고, 단층 건물에서도 거나 같아야 한다. 따라서 소수점 이하 근소한 차이 및 반올림 오차 등을 고려하여 건축면적이 연면적보다 1㎡ 이상 클 경우 오류로 판정하였다. 또한 표제부에 건축면적 또는 연면적에 값이 없는 경우 판정이 불가능해 NULL 처리되었다.

건축면적의 오류율은 2022년 16.78%에 비해 5.01%로 적어졌다. 하지만 2022년 점검대상건수는 7,952,243건이며, 2025년의 점검대상건수는 부속건축물을 제외한 7,364,967건으로 모수가 달라 해석에는 주의가 필요하다. 또한 표제부의 건축면적 또는 연면적의 값이 없는 NULL의 건수가 343,908건 나타났다.

[표 3-14] 건축면적의 오류율

순번	2022년			2025년				
	오류건수	점검대상건수	오류율	NULL	유효건수	오류건수	점검대상건수	오류율
16	1,334,538	7,952,243	16.78%	343,908	6,647,554	373,505	7,364,967	5.01%

출처: 세움터 내부자료

- (규칙 19) 건폐율

건폐율의 오류판정은 건축물대장에 기입된 건폐율을 기준으로 건축면적과 대지면적을 활용하여 건폐율을 재계산하고 기재된 값과의 차이가 0.1% 이상 차이나는 건축물을 오류로 규정하였다. 건축면적과 대지면적, 건폐율 중 값이 하나라도 없다면 NULL로 처리된다.

건폐율의 오류율은 2022년 0.74%에 비해 1.16%로 늘어났다. 점검대상건수에 일부 차이가 있지만 7백3십만여건으로 유사하고, 오류 건수도 늘어난 점을 고려할 때, 실제로 오류율이 더 높게 나타난 것으로 판단된다. 이러한 차이는 본 연구의 오류 판별 규칙에서 건폐율 및 용적률의 오류 기준을 0.1%로 엄격하게 설정한 것이 원인일 수 있다.

[표 3-15] 건폐율의 오류율

순번	2022년			2025년				
	오류건수	점검대상건수	오류율	NULL	유효건수	오류건수	점검대상건수	오류율
19	54,486	7,324,326	0.74%	3,336,827	3,942,588	85,541	7,364,967	1.16%

출처: 세움터 내부자료

- (규칙 20) 용적률

용적률의 오류판정은 건축물대장에 기입된 용적률을 기준으로 용적률 산정 연면적과 대지면적을 활용하여 용적률을 재계산하여 값의 차이가 0.1% 이상 차이나는 건축물을 오류로 규정하였다. 용적율 산정 연면적과 대지면적, 용적률 중 값이 하나라도 없다면 NULL로 처리된다.

용적률의 오류율은 2022년 13.88%에 비해 1.12%로 크게 감소한 것으로 보인다. 2022년 점검 대상 건수는 5,212,430건으로 적은 편이나, 오류 건수로 비교하여도 2022년 약 72만 건에서 2025년 약 8만 건으로 크게 감소하였다. 용적률 검증은 건폐율과 동일한 기준을 적용하여 수행되었으므로 실제로 용적률 오류가 2022년 대비 낮게 나타난 것으로 판단된다.

[표 3-16] 용적률의 오류율

순번	2022년			2025년				
	오류건수	점검대상건수	오류율	NULL	유효건수	오류건수	점검대상건수	오류율
19	723,474	5,212,430	13.88%	3,396,161	3,886,770	82,136	7,364,967	1.12%

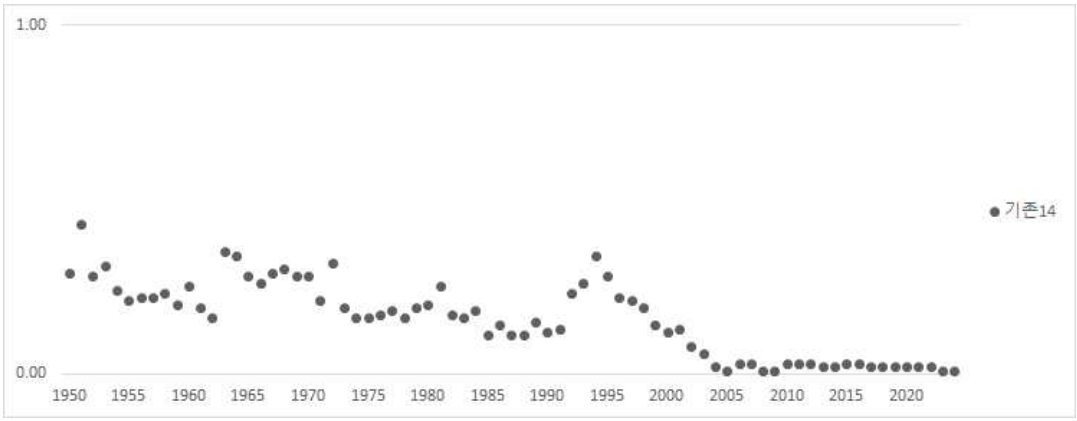
출처: 세움터 내부자료

• 준공 시기별 비교

검증 규칙별 오류율 도출 결과를 건축물 사용승인 연도별로 분석하여 준공 시기에 따른 오류 발생 경향을 분석하였다. 1% 간격 세로축을 기준으로 연도별 오류율을 점으로 나타내어 연도별 오류 발생 추세를 검토하였다. 전국 건축물 준공 시기 분포, 건축물대장 데이터의 정확성, 최근 오류 추세 검토 필요성 등을 고려하여 1950년부터 2024년까지 75년간의 데이터를 검토하였다.

- (규칙 14) 대지면적

대지면적 오류는 전체 기간 동안 1% 미만으로 나타났으며, 1950년대부터 1990년대 중반까지 오류율에 큰 차이를 보이지 않았으나, 1994년부터 오류율이 지속적으로 감소하여 2000년대 중반부터 오류율이 0에 가깝게 나타나고 있다. 최근에는 건축면적이 대지면적보다 크게 기재된 건축물이 거의 없음을 확인할 수 있다.

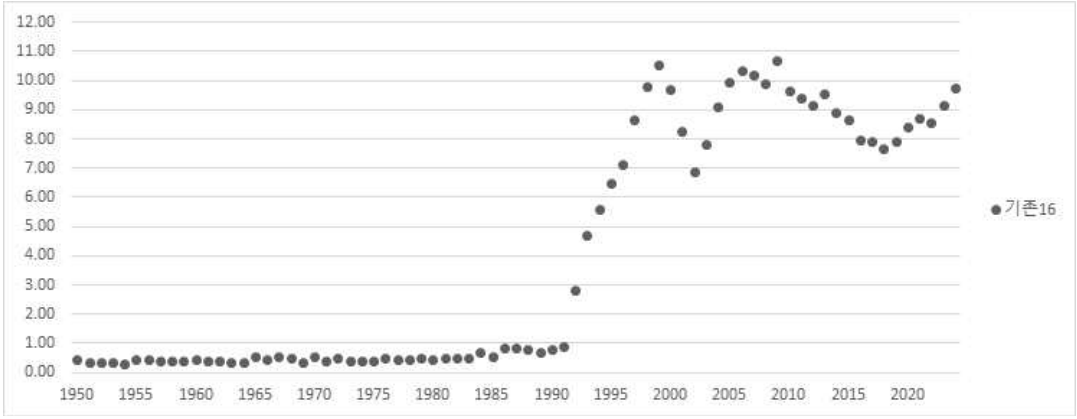


[그림 3-12] 사용승인 연도별 대지면적 오류
출처: 연구진 작성

- (규칙 16) 건축면적

건축면적이 연면적보다 큰 경우는 1980년대까지 거의 나타나지 않았으나, 1990년대 급격히 증가하기 시작하여 1999년 10%를 상회하였고, 이후 2000년대 초반까지 8% 이하로 감소하다 이후 8~10% 이내로 등락을 반복하고 있다. 2022년 건축물대장 정비 당시 산출된 오류율 16.78%에 비하면 낮기는 하지만 다른 검증규칙에 비하면 높은 수준이었다.

건축면적이 연면적보다 큰 건축물 사례를 검토한 결과, 대부분 단층 건축물이었다. 앞서 기존 규칙 검토에서 논의한 바와 같이 건축면적과 바닥면적의 산정 기준이 상이하며, 건축면적이 단층의 바닥면적보다 큰 경우가 발생할 수 있다. 다만 이러한 건축물이 1990년대 이후에 집중적으로 나타나는 것은 「건축법」 및 하위 법령의 면적 산정방법의 변화로 인한 것으로 추정된다.

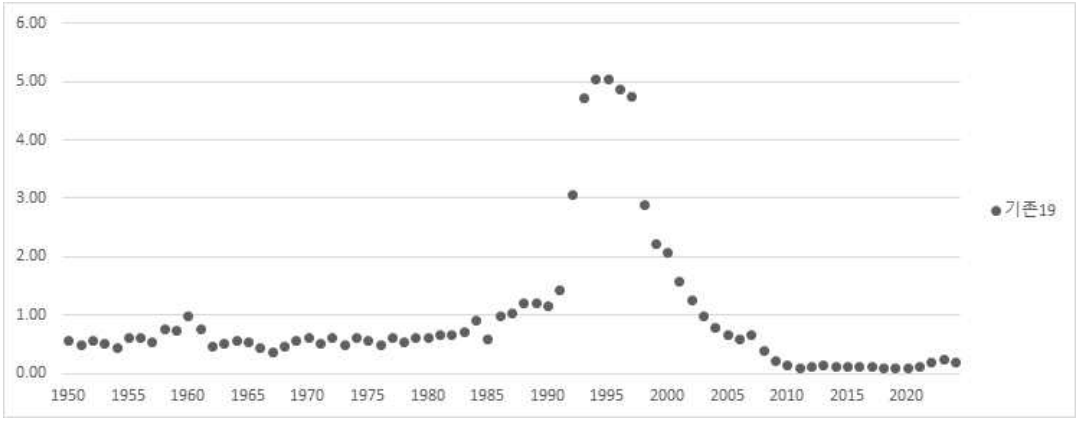


[그림 3-13] 사용승인 연도별 건축면적 오류
출처: 연구진 작성

- (규칙 19) 건폐율

건폐율은 1990년대를 제외하면 오류율이 1% 미만으로 나타나고 있으나, 1990년대에 오류율이 집중적으로 높게 나타나 최대 약 5%에 달하였다. 특정 시기에 집중된 오류 발생은 건축물대장 전산화와 관

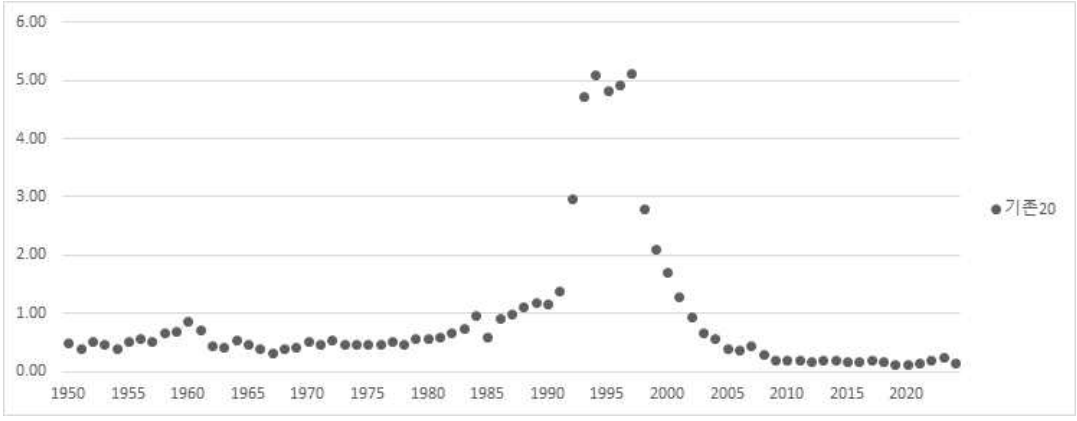
련된 행정·기술적 요인, 전산 입력체계의 미비, 전산화 전환기 혼란 등과 연관된 것으로 해석할 수 있다. 따라서 이러한 시기적 특성을 고려하여 오류 발생의 배경을 규명하는 한편, 해당 시기 건축물대장에 대한 다각적인 검증이 필요하다.



[그림 3-14] 사용승인 연도별 건폐율 오류
출처: 연구진 작성

- (규칙 20) 용적률

용적률의 오류율도 건폐율과 유사하게 1990년대에 집중하여 나타나고 있다. 1993년~1997년 동안 용적률 오류율은 약 5%에 달하다, 1990년대 이후로 급격히 감소하여 최근에는 오류율이 1% 미만으로 나타나고 있다.



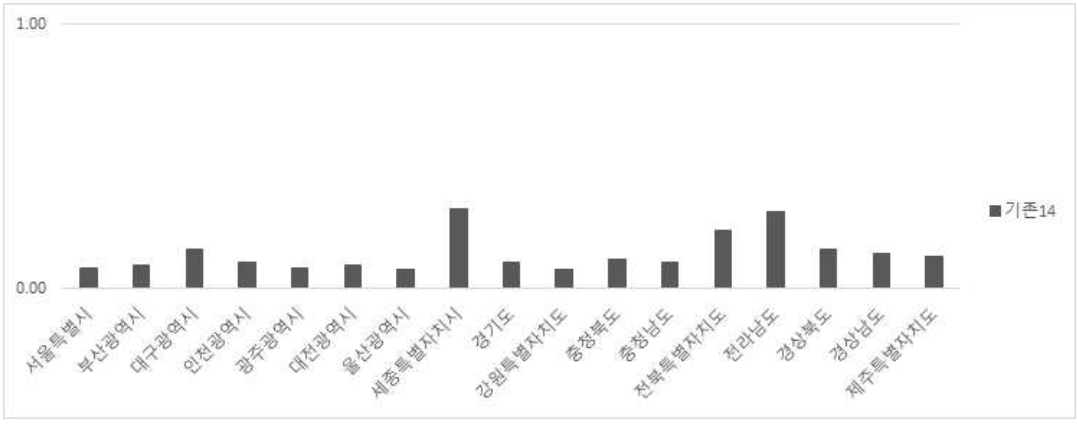
[그림 3-15] 사용승인 연도별 용적률 오류
출처: 연구진 작성

- 지역별 비교

시기별 비교에 이어, 검증 규칙별 오류율을 시도별로 분석하여 지역별 오류 발생 분포를 분석하였다. 사용승인 연도별 비교와 동일하게 1% 간격 세로축을 기준으로 시각화하였으며, 각 시도별 오류율을 막대로 나타내어 전국 17개 시도의 오류율 분포를 검토하였다.

- (규칙 14) 대지면적

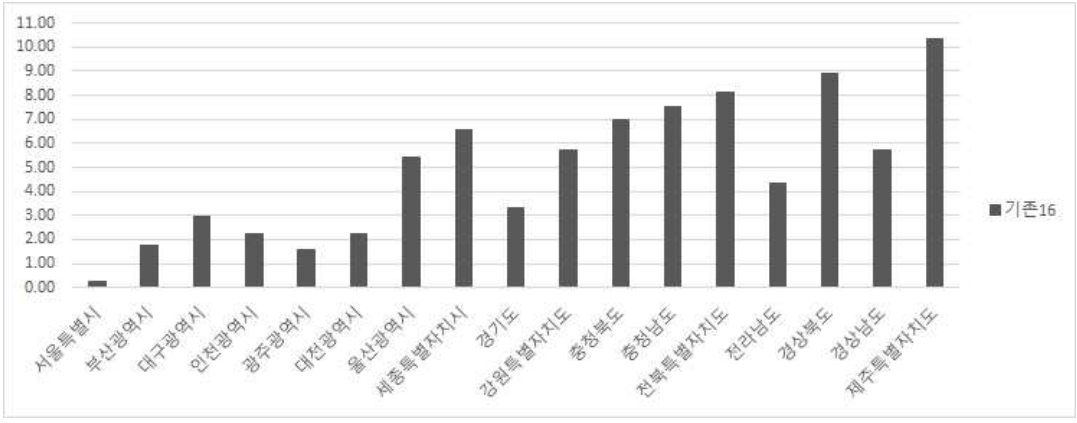
대지면적의 경우 모든 시도에서 1% 미만 오류율을 보였으나, 세종특별자치시(0.30%), 전라남도(0.29%), 전북특별자치도(0.22%) 등 지역에서 비교적 높게 나타났다. 울산광역시(0.07%), 서울특별시(0.08%), 광주광역시(0.08%) 등에서는 매우 낮은 오류율을 보였다.



[그림 3-16] 시도별 대지면적 오류
출처: 연구진 작성

- (규칙 16) 건축면적

건축면적의 경우 대지면적과 달리 오류율이 높았고, 시도별로도 편차가 크게 나타났다. 제주특별자치도(10.34%)가 가장 높은 오류율을 보였으며, 경상북도(8.92%), 전북특별자치도(8.12%), 충청남도(7.52%), 충청북도(6.98%) 등도 두드러지게 높았다. 반면 서울특별시(0.25%), 광주광역시(1.59%) 등은 상대적으로 낮은 편이었다. 1990년대 이후 오류가 집중되어 발생하였다는 시기별 결과와 비교할 때, 특정 시기, 일부 지역에서 오류가 집중적으로 나타나고 있다고 판단된다.

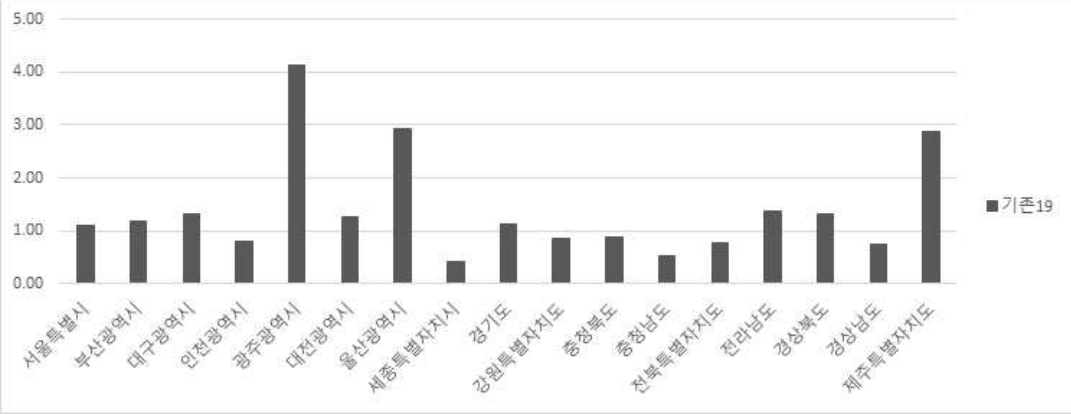


[그림 3-17] 시도별 건축면적 오류
출처: 연구진 작성

- (규칙 19) 건폐율

건폐율 오류율은 대부분 시도에서 1% 전후로 나타나고 있어 다른 오류에 비하면 낮은 편은 아니었으나, 특히 특정 시도에서 오류가 집중되어 나타나는 분포를 보였다. 광주광역시(4.14%), 울산광역시

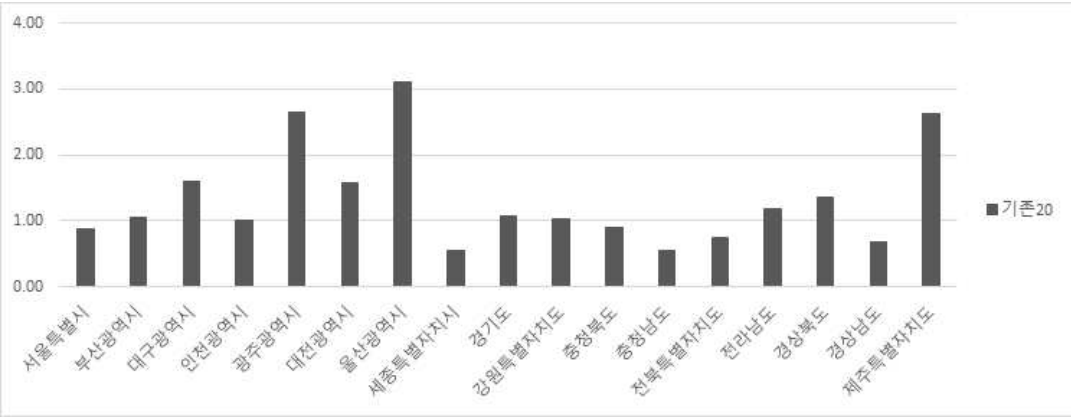
(2.92%), 제주특별자치도(2.88%) 등 지역에서는 오류율이 다른 시도에 비해 2배 이상 높게 나타났다. 건폐율 오류가 1990년대에 집중되어 발생하였다는 시기별 결과와 비교할 때, 특정 시기, 일부 지역에서 오류가 집중적으로 나타났다가 현재는 감소한 것으로 판단된다.



[그림 3-18] 시도별 건폐율 오류
출처: 연구진 작성

- (규칙 20) 용적률

용적률 오류율도 대부분 시도에서는 1% 전후로 나타나나 건폐율 오류와 동일하게 울산광역시(3.11%), 광주광역시(2.65%), 제주특별자치도(2.62%)에서 높은 오류율을 보였다. 건폐율과 마찬가지로 오류가 1990년대에 집중되어 발생하였다는 시기별 결과와 비교할 때, 특정 시기, 일부 지역에서 오류가 집중적으로 나타났다가 현재는 감소한 것으로 판단되며, 건폐율 오류와 용적률 오류가 유사한 패턴을 보이는 점을 고려하면 유사한 원인으로 발생한 것으로 추정된다.



[그림 3-19] 시도별 용적률 오류
출처: 연구진 작성

- 검증결과 종합

건축물대장의 주요 기재 항목별 오류율은 항목별 특성과 검증 방식에 따라 상이한 경향을 보였다. 대지면적(규칙 14) 전체 오류율은 0.1%대에 머물렀다. 이는 대지면적과 건축면적 간 불일치 사례가 드물다는 점을 의미한다. 특히 최근에는 오류가 거의 발생하지 않고 있다.

건축면적(규칙 16)은 다른 항목에 비해 높은 오류율을 나타내었다. 오류율 자체는 2022년 16.78%에서 2024년 말 기준 5.01%로 감소하였는데, 감소 이유는 오차범위에 따른 차이 등으로 추정된다. 2024년 기준 오류율도 여전히 다른 규칙 대비 높은 수준이고 준공년도에 따른 추세도 1990년대 이후 높은 오류율을 보였다.

건폐율(규칙 19)과 용적률(규칙 20)은 오류율 추세가 유사하게 나타났다. 두 지표 모두 시기별로는 1990년대, 지역별로는 광주광역시, 울산광역시, 제주특별자치도에 한하여 높은 오류율을 보였다. 이후 급격히 감소하여 최근에는 1% 내외 수준으로 안정된 상태다. 이는 특정 전환기에 일부 지역에서 발생한 오류가 현재까지 누적·잔존한 결과로 해석된다. 또한 두 지표는 대지면적·건축면적 값을 기반으로 산출된 파생 항목으로 그 성격이 유사하다. 따라서 동일 원인에 의한 오류로 추정된다.

여전히 높은 오류율을 보이는 항목, 특정 시기·지역에 집중된 오류 발생 등은 건축물대장 품질이 행정적·기술적 요인에 크게 영향을 받아왔음을 보여준다. 따라서 건축물대장 정비 등 검증 규칙을 도출한 이후에는 동일한 규칙을 통한 검증을 지속적으로 수행하여 조치 경과 및 현황을 모니터링하고 오류 발생 유형별 맞춤형 품질 고도화 방안을 마련할 필요가 있다.

4) 신규 검증규칙 개발

- 개요

기존 건축물대장 정비에서 도출된 검증 규칙 적용에 더하여, 기존 정비에서 도출되지 않았던 새로운 검증 규칙을 개발하고 현재 기준 오류 현황을 분석하였다. 건물동 단위 일반건축물대장 및 표제부의 면적 관련 데이터에 대하여 새롭게 비교 및 검증할 수 있는 규칙 총 2개를 개발하여 검증하였다.

- 신규 검증규칙 정의 및 결과
  - (신규 01) 용적률 산정 연면적

용적률 산정 연면적은 용적률 산정에 사용되는 연면적을 산출한 항목으로, 연면적 중 용적률 산정 시 제외되는 면적을 연면적에서 빼 용적률을 산정할 수 있도록 산출한 값이다. 연면적의 정의인 건축물의 각 층 바닥면적의 합계에서 지하층, 지상층 주차장(부속용도인 경우), 초고층·준초고층 피난안전구역, 경사지붕 아래 대피공간 등 일부 면적을 제외한 값을 의미한다. 따라서 용적률 산정 연면적은 용적률 산정에서 제외되는 면적을 빼기 전인 (일반) 연면적에 대비하여 면적이 작을 것으로 판단할 수 있다. 따라서 용적률 산정 연면적의 검증은 용적률 산정 연면적이 (일반) 연면적보다 클 경우 오류로 판별하는 방식이다. 구체적으로는 일반건축물대장 또는 표제부의 용적률 산정 연면적이 연면적 보다 1㎡ 이상 클 경우 오류로 판정하였다. 또한 표제부에 용적률산정연면적 또는 연면적에 값이 없는 경우 판정이 불가능해 NULL 처리되었다.

용적률 산정 연면적의 오류율은 0.26%로 나타났다. 유효건수는 6,999,231건(95.03%)로 높게 나타났고, 용적률 산정 연면적 또는 연면적에 값이 없는 NULL의 건수는 346,441건으로 나타났다.

[표 3-17] 용적률 산정 연면적의 오류율

신규	NULL	유효건수	오류건수	점검대상건수	오류율
01	346,441	6,999,231	19,295	7,364,967	0.26%

출처: 연구진 작성

- (신규 02) 연면적 상한

연면적은 건축물의 각 층 바닥면적의 합계로 산출할 수 있는데, 각 층 바닥면적은 각각 공간적으로 가능한 상한이 존재한다. 지상층의 경우 바닥면적이 건축면적을 초과할 수 없으며, 지하층은 대지면적을 초과할 수 없다. 따라서 건축면적과 지상층수 곱을 통하여 지상층 바닥면적의 상한을, 대지면적과 지하층수의 곱을 통하여 지하층 바닥면적의 상한을 산출할 수 있으며, 두 값의 합을 통하여 건축물 연면적의 상한을 산출할 수 있다. 정리하면 연면적이 (건축면적×지상층수)+(대지면적×지하층수)보다 큰 경우 오류로 판별하는 방식이다. 구체적으로 연면적 상한 오류 판정은 연면적이 연면적 상한값((건축면적×지상층수)+(대지면적(표제부 또는 총괄표제부)×지하층수))보다 1㎡ 이상 클 경우 오류로 판정하였다.

연면적 상한의 오류율은 0.86%로 나타났다. 유효건수는 5,134,625건(69.72%)로 나타났고, NULL의 건수는 2,167,029건으로 나타났다.

[표 3-18] 연면적 상한의 오류율

신규	NULL	유효건수	오류건수	점검대상건수	오류율
01	2,167,029	5,134,625	63,313	7,364,967	0.86%

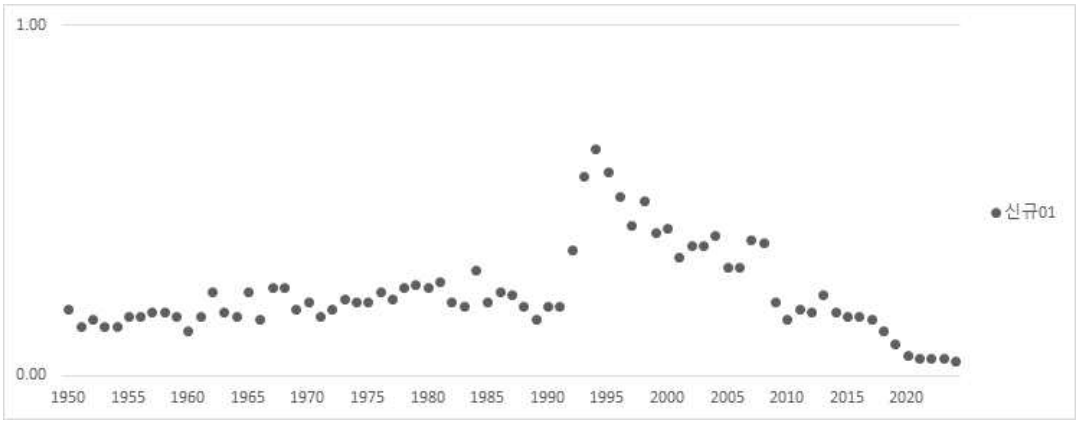
출처: 연구진 작성

• 준공 시기별 비교

기존 검증규칙과 동일한 방식으로 신규 검증규칙의 준공 시기별 오류율을 검토하였다. 기존 검증규칙과 마찬가지로 1% 간격 세로축을 기준으로 연도별 오류율을 점으로 나타내어 연도별 오류 발생 추세를 검토하였다. 1950년부터 2024년까지 75년간의 데이터를 검토하였다.

- (신규 01) 용적률 산정 연면적

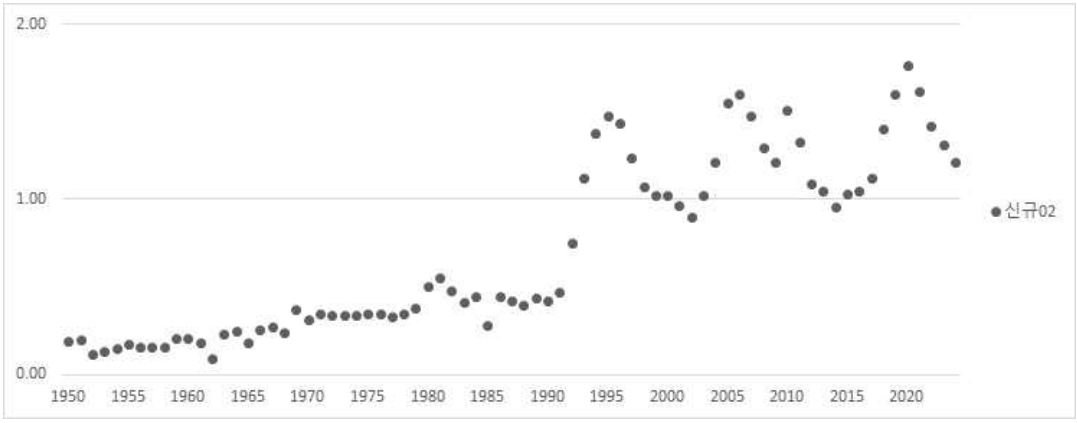
용적률 산정 연면적은 모든 기간 동안 오류율이 1% 미만으로 낮은 편이었으나, 1994년을 정점으로 오류율이 소폭 상승한 뒤 점차 낮아지는 경향을 보이고 있다. 해당 시기 건축면적, 건폐율, 용적률이 오류율 증가를 보인 것과 비교하면, 건축면적과 건폐율처럼 용적률 산정 연면적과 용적률도 해당 시기에 전산화 전환 등 원인으로 오류율이 다소 증가하였던 것으로 추정할 수 있다.



[그림 3-20] 사용승인 연도별 용적률 산정 연면적 오류
출처: 연구진 작성

- (신규 02) 연면적 상한

연면적 상한 오류는 1980년대까지 1% 미만의 오류율을 보이다가, 1995년을 정점으로 1차 상승을 보였다가 다시 감소하였다. 그러나 용적률 산정 연면적과 다르게, 2006년, 2020년 등 주기적으로 오류 발생 증가가 관측되고, 현재도 1% 이상의 오류율을 보이는 것으로 나타나고 있다.



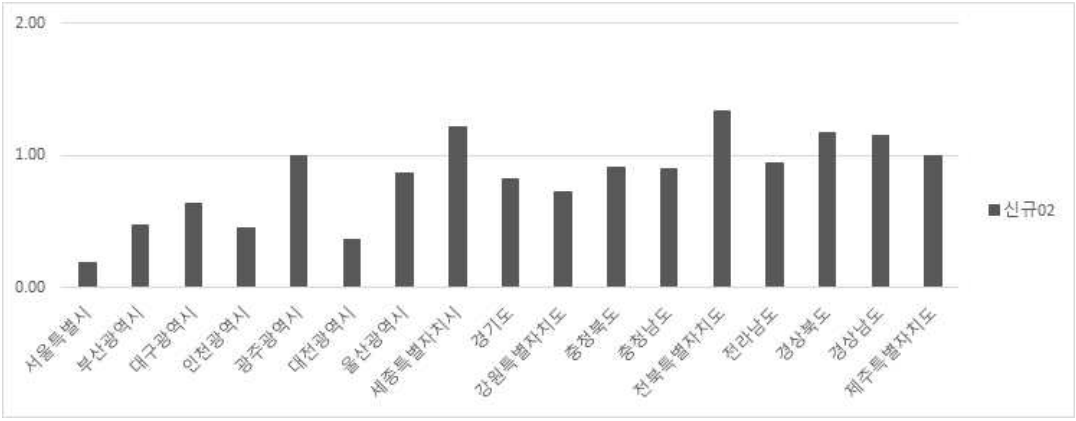
[그림 3-21] 사용승인 연도별 연면적 상한 오류
출처: 연구진 작성

- 지역별 비교

신규 검증규칙별 오류율을 시도별로 분석하여 지역별 오류 발생 분포를 분석하였다. 기존 규칙과 동일하게 1% 간격 세로축을 기준으로 시각화하였으며, 각 시도별 오류율을 막대로 나타내어 전국 17개 시도의 오류율 분포를 검토하였다.

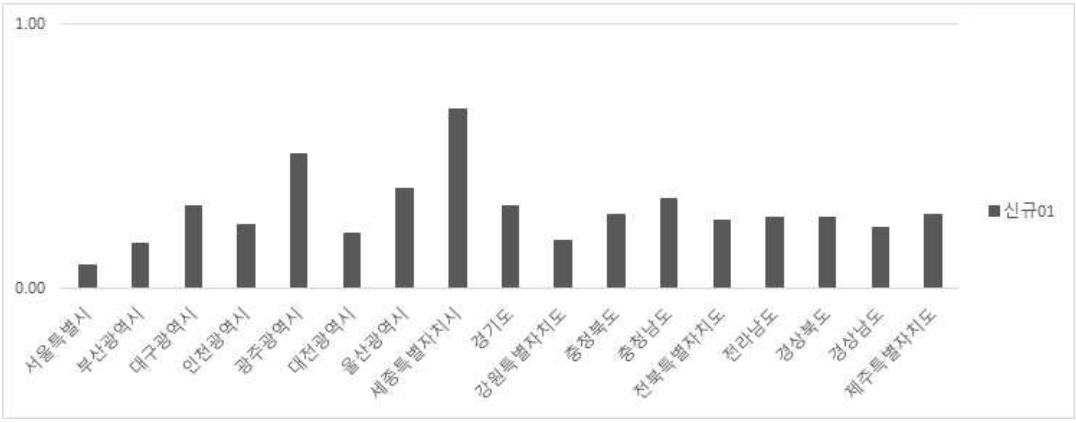
- (신규 01) 용적률 산정 연면적

용적률 산정 연면적의 경우 세종특별자치시(0.68%)에서 가장 높은 오류율을 보였다. 다음으로는 광주광역시(0.51%)가 높은 편이었고, 나머지 시도는 대체로 비슷한 수준인 가운데, 서울특별시가 0.09%로 가장 낮게 나타났다. 앞서 시기별 비교에서 용적률 산정 연면적 오류는 1990년대~2000년대에 집중하여 나타났는데, 세종시 행정도시 건설은 2010년대 들어서야 본격화된 점을 고려하면 두 경



[그림 3-23] 시도별 연면적 상한 오류
출처: 연구진 작성

향이 하나의 현상을 나타낸다고 해석하기보다는, 1990년대부터 발생한 오류에 대하여 건축물대장 정비 현황이 시도별로 다르게 나타나고 있다고 추정할 수 있다.



[그림 3-22] 시도별 용적률 산정 연면적 오류
출처: 연구진 작성

- (신규 02) 연면적 상한

연면적 상한의 경우 특광역시보다 도 지역의 오류율이 비교적 높은 편으로 나타났다. 세종특별자치시(1.21%)와 광주광역시(1.00%)는 높은 오류율을 보였으나, 나머지 특광역시는 1% 미만 오류율을 보였고, 전북특별자치도(1.33%), 경상북도(1.17%), 경상남도(1.15%) 등 도 지역은 다소 높은 오류율을 보였다. 연면적 상한 오류가 최근까지도 주기적으로 증가하고, 과거 대비 높은 수준을 유지하고 있는 것을 볼 때, 도 지역에서 연면적 관련 오류가 지속적으로 발생하고 있다고 판단할 수 있다.

- 검증결과 종합

신규 검증규칙은 기존 건축물대장 정비 과정에서 고려되지 않았던 검증 항목을 추가로 도출하여, 면적 관련 데이터의 품질 고도화 방법론을 보완하고자 한 것이다. 본 연구에서는 건물동 단위 일반건축물대장과 표제부 데이터를 활용하여 용적률 산정 연면적, 연면적 상한 등 총 2개의 신규 검증규칙을 정의하고 오류 현황을 분석하였다.

용적률 산정 연면적 검증은 연면적 대비 용적률 산정 연면적이 더 큰 경우를 오류로 판정하였다. 전체 오류율은 0.26%로 낮았다. 시기별 분석 결과 1990년대 중반에 오류율이 일시적으로 상승하였다가 이후 점차 감소하는 경향이었다. 이는 건축면적, 건폐율, 용적률 등 같은 시기 다른 항목의 오류율이 높아진 양상과 같았다. 전산화 전환과 같은 제도적 변화가 오류 증가의 원인일 가능성을 시사한다. 세종시의 경우 2010년대 이후 도시 건설이 본격 진행되었다. 따라서 준공연도별 경향보다는 지역별 대장 정비 수준 차이에 기인하였다고 추정한다.

둘째, 연면적 상한 검증은 건축물 연면적이 가능한 상한값(건축면적×지상층수 + 대지면적×지하층수)을 초과하는지를 판정하였다. 분석 결과 전체 오류율은 0.86%이었다. 이 오류는 최근까지도 반복적으로 발생함을 확인하였다. 지역별로는 세종특별자치시(1.21%), 전북특별자치도(1.33%), 경상북도(1.17%), 경상남도(1.15%) 등에서 높았다. 전반적으로 도 지역의 오류율이 특광역시에 비해 높은 경향을 보였다. 연면적 데이터 관리에서 지역 간 차이가 여전히 존재한다. 특히 도 지역의 품질 관리 체계 강화가 필요함을 시사한다.

신규 검증규칙 분석은 기존 정비 규칙만으로 포착하지 못했던 면적 데이터의 구조적 오류를 발견하는 가능성을 확인하였다. 오류 분포가 지역별로 집중되었고, 시기별로는 지속적인 오류 발생을 관측하였다. 이는 제도적, 지역적 관리 미비에 대한 추정을 뒷받침하였다. 오류 발생이 집중되는 특정 시기나 지역을 우선 정비하는 정책적 접근이 필요하다.

5) 기계학습을 통한 이상값 탐지

- 개요

건축물 면적 데이터 품질 고도화를 위해 기계학습 기반 이상값 탐지를 수행하였다. 기존 업무규칙 기반 검증으로 포착하기 어려운 잠재적 오류나 비정상 패턴을 식별하기 위한 것이다. 건축물 규모에 따라 변하지 않는 특성을 반영하고자 무차원 변수(dimensionless variables)를 정의하였다. 대지면적, 건축면적, 연면적, 용적률 산정용 연면적, 층수 등 기본 변수를 조합하여 건폐율, 용적률, 유효층수, 충만률 등 다양한 비율 지표를 산출하였다. 이들 간 상관성을 검토하여 독립적인 무차원 변수 집합을 최종 선정하였다.

이와 같이 도출된 변수를 입력값으로 기계학습에 적용하였다. 기계학습 모델은 대표적인 이상탐지 기법인 Isolation Forest와 One-Class SVM을 적용하였다. 두 알고리즘은 전체 분포 속에서 현저히 다른 패턴을 보이는 특정 건축물을 판정한다. 두 방법론을 조합하여 판정 차이를 보완하고 확실한 이상값을 도출하였다.

이상값 도출 결과를 시기 및 지역 단위로 집계하였다. 특정 지역이나 시기에 높은 비율의 이상값이 나타나는지를 확인하고 그 성격을 검증하였다. 기계학습 기법이 기존 규칙 기반 검증을 보완하며 데이터 내 오류나 패턴을 식별할 수 있는 가능성을 확인하였다.

• 분석 데이터 전처리

분석 전 데이터 전처리 과정을 거쳤다. 데이터 정합성을 확보하고 결측값, 규칙 기반 검증 대상 데이터를 제거하였다. 규칙 기반 검증과 동일한 데이터에서 기본 제약을 충족하는 사례만 추출하였다. 대지면적은 건축면적 이상, 연면적은 건축면적 이상, 연면적은 용적률 산정용 연면적 이상이어야 한다. 면적 변수가 결측이면 총괄표제부 값을 사용하고, 다시 결측이면 0으로 대치하였다. 이 과정을 통해 명백한 오류가 제거된 데이터셋이 구성되었다. 모델 투입 전 계산 과정에서 발생할 수 있는 무한대나 결측값을 모두 정제하였다.

• 무차원 변수 도출

기계학습 기반 이상값 탐지 방법론은 정상적인 분포에서 나타나기 힘든 극단적인 값을 찾아낸다. 그러나 건축물 면적 관련 데이터 대상으로 이상값 탐지 방법론을 그대로 적용할 경우, 단순히 매우 크거나 매우 작은 건축물이 이상값으로 도출될 수 있다. 단순히 규모 면에서 극단적인 건축물이 아니라 오류 가능성이 높은 이상값을 도출하기 위하여, 면적이나 층수 등 절대 규모의 차이를 제거하고 건축물의 형태와 이용 특성을 상대적으로 파악할 수 있는 무차원 변수를 설계 및 산출하였다.

이를 위하여 먼저 기존 면적 관련 변수의 차원분석을 실시하였다. 무차원 변수의 도출은 물리학적 차원분석의 원리에 기반하였다. 물리학의 차원분석은 길이·질량·시간과 같은 기본 차원을 단위로 하여 변수들 간의 관계를 해석하고, 단위가 서로 다른 변수들을 결합하여 크기에 무관한 지표를 만드는 방법론이다. 본 연구에서는 물리학적 차원분석을 그대로 적용하기보다는, 건축법과 건축물대장 체계에서 핵심적인 통제 변수인 면적과 층수를 건축학적 기본 차원으로 설정하고, 이를 조합하여 변수의 차원을 표현하는 변형 방법론을 적용하였다. 먼저 면적 관련 변수의 차원을 분석한 결과, [면적], [층수], [면적]과 [층수]의 곱 등의 차원을 지닌 것으로 나타났다. 이에 기반하여 같은 차원을 가진 두 변수의 비율을 구하거나, 경우에 따라 3개 변수를 조합하여 [비율] 차원 변수를 도출하는 것을 무차원 변수 정의의 목표로 하였다. 다만 용적률은 [비율]과 [층수]의 곱이지만 건축물의 특성을 나타내는 대표적인 변수임을 고려하여, 용적률과 동일한 [비율][층수] 차원의 변수도 무차원 변수로 도출하였다.

[표 3-19] 면적 관련 변수 차원분석 결과

이름	차원
대지면적	[면적]
건축면적	[면적]
연면적	[면적][층수]
용적률 산정용 연면적	[면적][층수]
지상 층수	[층수]

출처: 연구진 작성

[표 3-20] 무차원 변수 정의

이름	정의	차원
건폐율 (재산출)	건축면적/대지면적	[면적]/[면적] = [비율]
용적률 (재산출)	용적률 산정용 연면적/대지면적	[면적][층수]/[면적] = [비율][층수]
지상 유효층수	용적률 산정용 연면적/건축면적	[면적][층수]/[면적] = [비율][층수]
용적 산정률	용적률 산정용 연면적/연면적	[면적][층수]/[면적][층수] = [비율]
지상층별 건폐율	용적률 산정용 연면적/지상층수/대지면적	[면적][층수]/층수/[면적] = [비율]
지상층별 충만률	용적률 산정용 연면적/지상층수/건축면적	[면적][층수]/층수/[면적] = [비율]

출처: 연구진 작성

이러한 과정을 거쳐 총 6개의 무차원 변수를 도출하였다. 우선 건폐율은 건축면적을 대지면적으로 나눈 값으로, 건물이 대지 위에서 얼마나 많은 면적을 점유하는지를 보여주는 가장 기본적인 지표이다. 이는 대지 규모와 무관하게 건축물이 수평적으로 얼마나 채워져 있는지 상대적으로 평가할 수 있다.

다음으로 용적률은 용적률 산정용 연면적을 대지면적으로 나눈 값으로 정의하였다. 연면적 전체를 쓰지 않고 법적으로 산정되는 연면적을 사용함으로써, 제도적 기준에 부합하는 집적 강도를 측정할 수 있도록 하였다. 이 지표는 단순히 건물 크기만을 반영하는 것이 아니라, 용도별·공간별로 산정에서 제외되는 면적을 고려함으로써 법적 규제에 따른 건축 양상을 포착할 수 있도록 하였다.

세 번째로 지상 유효층수는 용적률 산정용 연면적을 건축면적으로 나눈 값이다. 이는 일정한 바닥면적이 몇 개 층으로 중첩되어 있는지를 나타내는 지표로, 동일한 건폐율이라 하더라도 저층 건축물인지, 혹은 고층으로 적층된 건축물인지를 구분할 수 있다. 따라서 지상 유효층수는 수직적 집적의 정도를 드러내는 지표이다.

네 번째로 용적 산정률은 용적률 산정용 연면적을 연면적으로 나눈 값이다. 이는 전체 연면적 가운데 실제 용적률 산정에 포함되는 면적의 비율을 보여주며, 지하층이나 부속용도와 같이 법적 산정에서 제외되는 면적이 차지하는 비중을 간접적으로 드러낸다. 따라서 용적 산정률이 낮게 나타난 건축물은 전체 규모 대비 용적률 산정 면적이 상대적으로 작아, 특정한 형태의 건축물을 나타낼 가능성을 시사한다.

다섯 번째로 지상층별 건폐율은 용적률 산정용 연면적을 지상층수로 나누고 다시 대지면적으로 나눈 값이다. 다르게는 용적률을 지상층수로 나눈 것으로 생각할 수 있다. 이는 지상 1개 층이 평균적으로 대지에서 차지하는 면적 비율을 보여주며, 층수가 다른 건축물 간 수평 밀도를 비교하는 데 유용하다. 저층 건축물과 고층 건축물을 단순 건폐율로 비교하기 어려운 점을 보완해주는 지표로 기능한다.

마지막으로 지상층별 충만률은 용적률 산정용 연면적을 지상층수로 나누고 이를 다시 건축면적으로 나눈 값으로, 지상 각 층이 건축면적 대비 얼마나 충실하게 채워져 있는지를 보여준다. 지상 유효층수를 실제 지상층수로 나눈 것으로 생각할 수도 있다. 이는 건물의 평면적 규모와 실제 층별 이용의 밀도를 동시에 고려한 지표로, 동일한 건폐율과 용적률을 가진 건축물이라도 내부 공간 활용의 차이를 드러낼 수 있다.

이 여섯 가지 무차원 변수는 모두 면적과 층수라는 기본 차원을 조합하여 도출되었으며, 계산 과정에서 단위가 상쇄되거나 간단한 비율로 환원되어 규모와 무관하게 비교 가능한 형태를 띤다. 각 변수는 상호 간에 일정 부분의 상관관계를 가지지만, 건축물의 수평 점유(건폐율, 층별 건폐율), 수직 적층(지상 유효층수, 층별 충만률), 용적률 산정 구조(용적률, 용적 산정률)라는 세 가지 측면을 각각 강조한다는 점에서 보완적 관계에 있다. 따라서 이 변수들은 건축물의 형태적·구조적 특성을 다차원적으로 반영하면서도, 절대적 규모에 좌우되지 않는 비교 가능성을 제공한다.

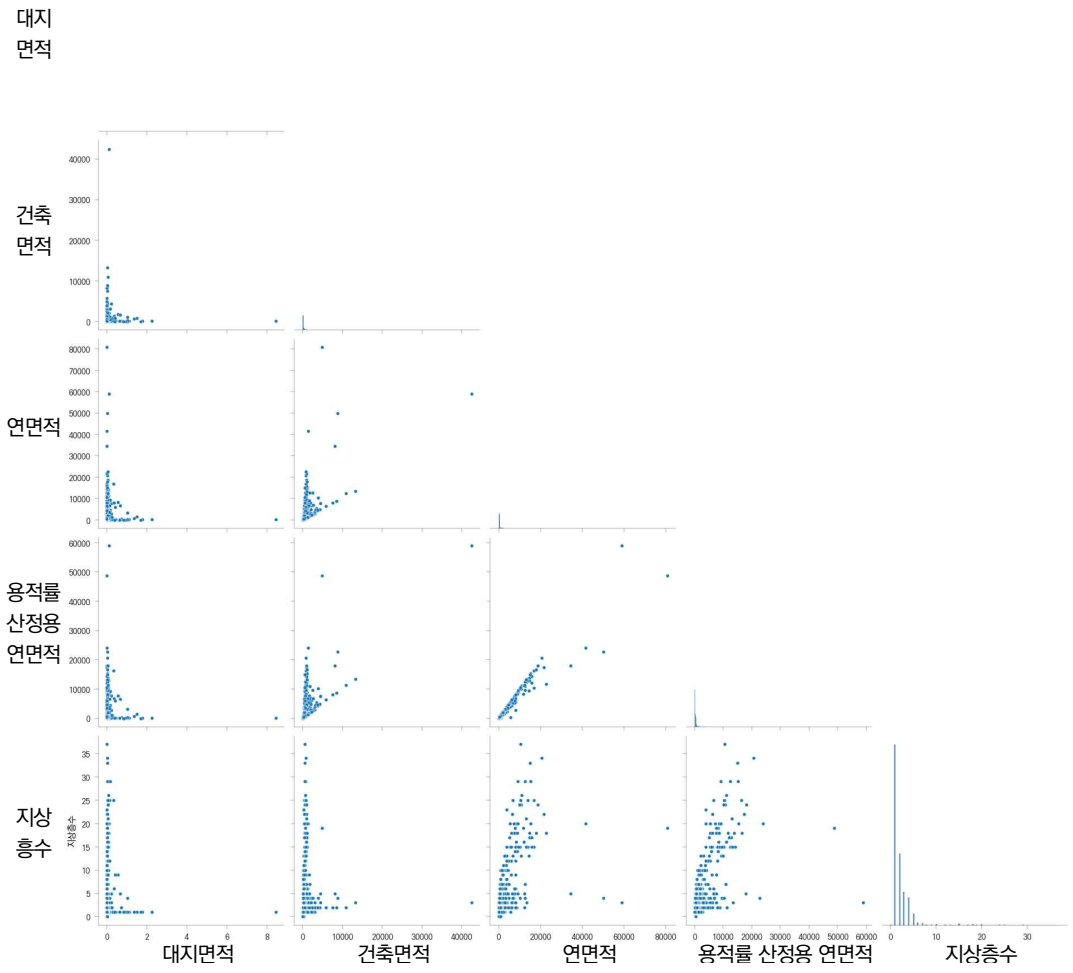
이와 같이 차원분석을 적용하여 산출된 지표들은 건축물의 절대적 크기 차이를 배제하면서도 본질적인 형상과 구조적 특성을 포착할 수 있도록 설계되었다. 따라서 이후 단계에서 적용된 이상값 탐지 알고리즘은 데이터의 스케일이나 단위에 의존하지 않고, 건축물 간 비교 가능한 상대적 패턴을 바탕으로 잠재적 이상을 식별할 수 있다.

- 무차원 변수 기초통계

무차원 변수 산출 후 기초통계를 검토한 결과, 원시 변수와 도출된 무차원 변수 모두에서 극단적인 분포와 이상치가 확인되었다. 전처리 과정에서 일부 오류 유형을 원천적으로 제외하였으나, 그렇지 않은 부분에서는 여전히 극단적 분포가 확인된다.

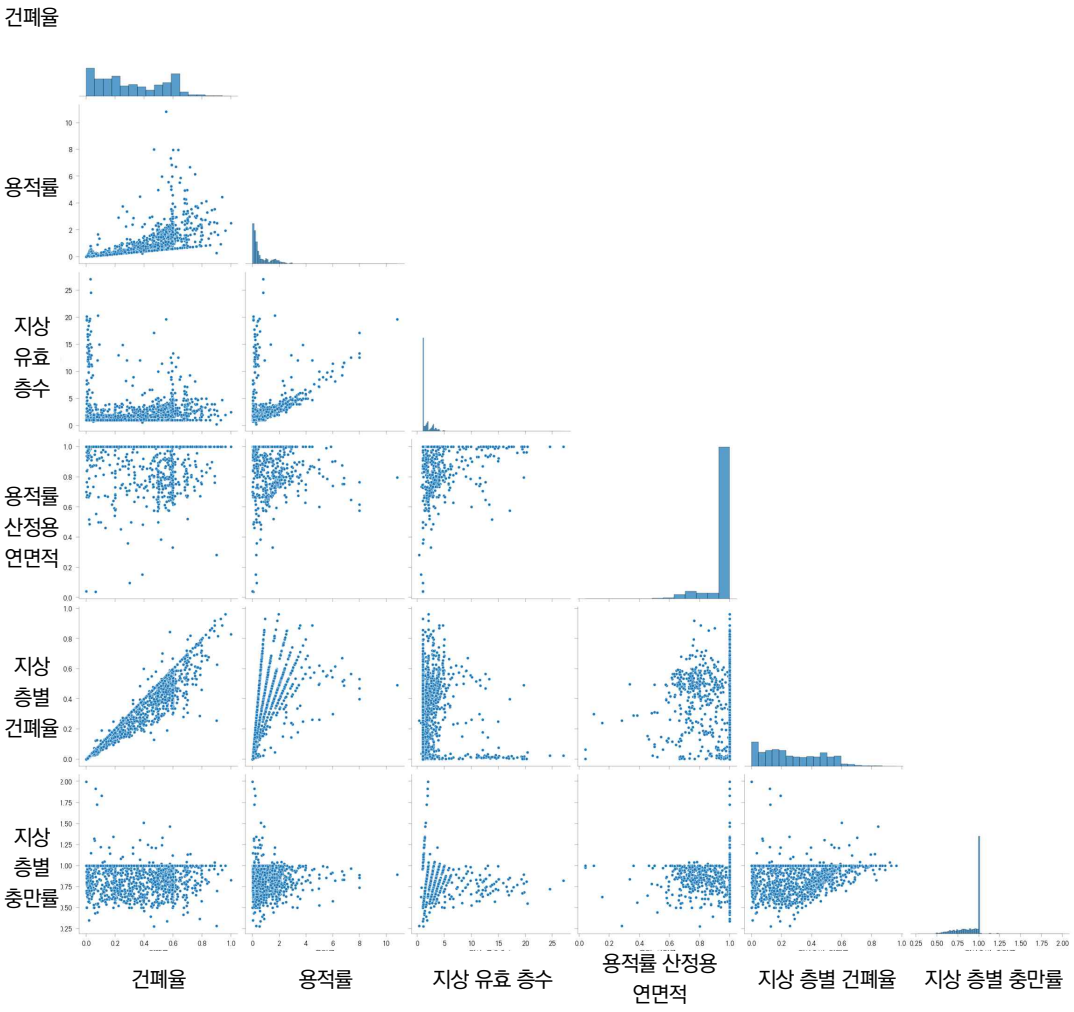
무차원 변수 중 첫 번째인 건폐율은 0에서 1(100%) 사이 값으로 정상 범위 내에서 나타나고 있다. 건폐율의 평균은 0.31로, 대체로 대지의 약 30% 정도가 건축면적으로 사용됨을 의미한다. 용적률은 평균이 0.65로 비교적 낮지만, 최대 4,947(배)까지 나타나 명백한 이상치가 포함되어 있음을 보여준다. 지상 유효층수의 중앙값은 1로 대부분 저층 건축물임을 시사하나, 최대값이 40만 층을 넘는 비현실적 기록이 존재해 데이터 오류가 뚜렷하다. 용적 산정률은 평균이 0.96으로 대부분의 연면적이 용적률 산정에 포함됨을 의미하며, 사분위수 범위가 모두 1.0에 집중되어 있어 안정적이다. 지상층별 건폐율은 평균이 0.27로 나타났으며, 극단적으로 549 이상으로 산출된 사례가 있어 이상값이 존재한다. 마지막으로 지상층별 충만률의 중앙값은 1.0으로, 대부분 건축물에서 층별로 건축면적이 충실하게 활용되는 것으로 나타나지만, 최대값이 66,603에 달해 역시 비현실적인 수치가 확인되었다.

종합하면, 기초통계 분석은 건축물대장 면적 데이터가 전반적으로 정상적인 분포를 보이면서도, 일부 항목에서 음수값·비현실적 극단값·비정상적 분산이 존재함을 보여준다. 이는 데이터 입력 오류나 예외적 특수 건축물에서 기인할 수 있으며, 후속 단계의 이상값 탐지 알고리즘을 통해 정제와 보완이 필요함을 시사한다.



[그림 3-24] 면적 관련 원시 변수 산점도

출처: 연구진 작성



[그림 3-25] 면적 관련 무차원 변수 산점도

출처: 연구진 작성

[표 3-21] 면적 관련 변수 기초통계

	대지면적	건축면적	연면적	용적률_산정용_연면적	지상층수
count	4827834.0000	4777956.0000	5081126.0000	4791078.0000	5085645.0000
mean	17015.6430	256.3988	815.2429	691.9291	2.3504
std	2346984.4250	1364.1303	60153.1923	38027.5603	3.1328
min	0.1200	0.0010	0.0010	-9.2400	0.0000
25%	244.4000	75.8500	85.9500	87.3800	1.0000
50%	495.0000	116.1300	178.8000	173.1000	1.0000
75%	1307.0000	199.5000	425.3100	404.4900	3.0000
max	4333333333.0000	2218284.2800	83374375.0000	58461732.0000	123.0000

	건폐율	용적률	지상_유효층수	용적_산정률	지상층별_건폐율	지상층별_충만률
count	4777956.0000	4724847.0000	4707543.0000	4791077.0000	4720271.0000	4703097.0000
mean	0.3080	0.6543	2.3107	0.9649	0.2739	0.9818
std	0.2223	3.1704	277.0266	0.0969	0.4277	56.2283
min	0.0000	-0.0280	0.0009	0.0001	-0.0093	0.0001
25%	0.1188	0.1371	1.0000	1.0000	0.1078	0.8404
50%	0.2521	0.3043	1.0000	1.0000	0.2278	1.0000
75%	0.5275	0.9516	2.0000	1.0000	0.4394	1.0000
max	1.0000	4946.7970	408459.3333	1.0000	549.6441	66603.0000

출처: 연구진 작성

산점도 검토 결과, 면적 관련 원시 변수의 경우 값의 분포가 극단적으로 집중되어 있어 히스토그램 상 분포 확인이 어려운 수준이었다. 무차원 변수의 경우 다수 변수가 비교적 고른 분포를 보였다.

- 기계학습 기반 이상탐지

기계학습 기반 이상탐지는 기존 규칙 검증만으로 포착하기 어려운 잠재적 오류를 식별하기 위해 수행되었다. 입력 데이터는 앞서 도출된 여섯 가지 무차원 변수를 중심으로 구성되었으며, 이를 통해 건축물의 절대적 규모 차이를 배제한 상대적 특성이 모델에 반영되도록 하였다. 학습 과정에서는 계산 효율성을 고려하여 최대 100만 건을 표본 추출하여 모델을 학습시키고, 학습된 모델을 사용하여 전체 데이터를 대상으로 예측을 수행하도록 하였다.

분석에는 두 가지 대표적인 비지도 이상탐지 알고리즘을 적용하였다. 첫째, Isolation Forest는 랜덤 분할을 통해 개별 관측치가 다른 집단과 얼마나 쉽게 분리되는지를 측정하는 방식으로 작동한다. 이러한 과정을 거쳐 비모수 분포를 보이는 다차원 데이터에서 이상값을 탐지할 수 있다. 둘째, One-Class SVM은 데이터의 외곽을 감싸는 초평면을 학습하여 내부와 외부를 구분하는 알고리즘으로, 대규모 데이터 적합을 위해 Nystroem 커널 근사와 SGD 기반 최적화를 결합하였다.

비지도 기계학습 알고리즘은 정답(label, 참값 true value)이 주어지지 않은 상태에서 데이터의 숨겨진 구조나 패턴을 탐색하는 데 활용된다. 이 때문에 비지도 학습은 지도 학습처럼 정확도나 재현율과 같은 명확한 성능 지표를 직접 계산하기는 어렵다. 대신 결과 검증은 주로 간접적인 방식이나 상대적인 척도를 통해 이루어진다. 차원 축소나 이상탐지와 같은 영역에서는 결과 해석의 직관성, 시각화를 통한 구조적 일관성, 혹은 도메인 지식에 근거한 전문가 검토가 중요한 검증 방법으로 사용된다. 본 연

구에서는 기존의 업무규칙 기반 검증 결과와 비교하여 공통적으로 식별되는 사례가 얼마나 되는지를 살펴봄으로써, 새로운 탐지 방식의 타당성을 간접적으로 점검하였다. 기존 규칙 및 신규 규칙의 시기별, 지역별 분포과 비교하여 기계학습 기반 이상탐지에서 도출된 결과의 타당성을 평가하고, 특정 시기나 지역에 편중되어 나타나는 이상치가 행정 데이터 입력 오류나 제도 변화와 연관될 가능성을 건축 분야 도메인 지식을 바탕으로 해석하였다.

SVM은 Isolation Forest 에 비해서는 중심이 있는 분포를 가정하므로, 서로 다른 특성을 보이는 두 알고리즘을 상호보완적으로 사용하고자 하였다. 구체적으로 판정 결과는 두 알고리즘을 독립적으로 적용한 뒤 합의 규칙을 통해 최종 확정하였다. 즉, Isolation Forest와 One-Class SVM이 동시에 이상치로 판정한 경우만을 최종 이상값으로 처리하여 보수적인 기준을 채택하였다. 이를 통해 한쪽 알고리즘에서만 나타나는 불안정한 결과는 제외하고, 신뢰도 높은 이상 패턴만을 도출할 수 있었다. 두 알고리즘 모두 재현성을 확보하기 위해 동일한 시드(seed)를 설정하였으며, 계산 시간과 안정성을 검증하였다. 두 알고리즘 모두 이상치 비율(contamination)을 사전에 설정하도록 하고 있는데, 기존 검증규칙 및 신규 검증규칙을 통하여 확인한 오류율 수준을 고려하여 이상값 비율은 보수적으로 1%로 설정하였다.

최종적으로 산출된 이상값은 건축물 단위에 라벨링된 뒤, 시도 및 준공연도별로 집계되어 이상값 비율이 계산되었다. 이 과정은 기계학습 기반으로 도출된 이상값의 특성을 파악하고, 특정 지역이나 시기에 비정상적 패턴이 상대적으로 높은 빈도로 발생하는지를 검토할 수 있게 하였다.

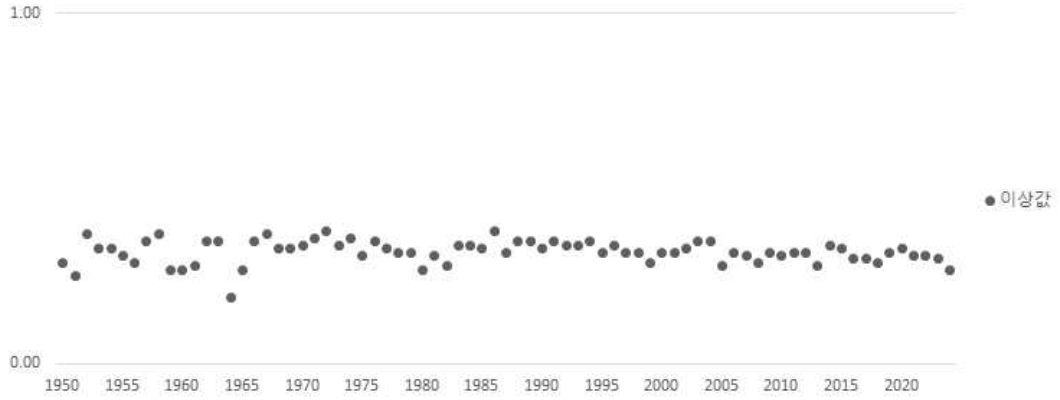
- 이상값 탐지 결과

기계학습 기반 이상탐지를 통해 도출된 이상값의 특성을 검토하기 위하여 먼저 산점도를 분석하였다. 이상값의 비율이 1%로 매우 적어, 이상값의 분포를 확인하기 위하여 이상값(빨간색)과 정상값(파란색)의 비율이 1:3이 되도록 과장하여 표본 추출 후 산점도 및 밀도 분포를 작성하였다. 분석 결과 일부 산점도에서 이상값 분포가 명확하게 분리되어 보이는 경우가 있었다. 예를 들어, 건폐율과 용적률, 지상 유효층수 3개 변수 사이의 산점도 3개에서는 이상값과 정상값 분포가 명확하게 분리되어 나타났다. 다른 변수에서는 두 집단의 구분이 조금 더 어려웠으나, 다차원 공간에서 이상값이 확실하게 분리되고 있다고 판단할 수 있었다.

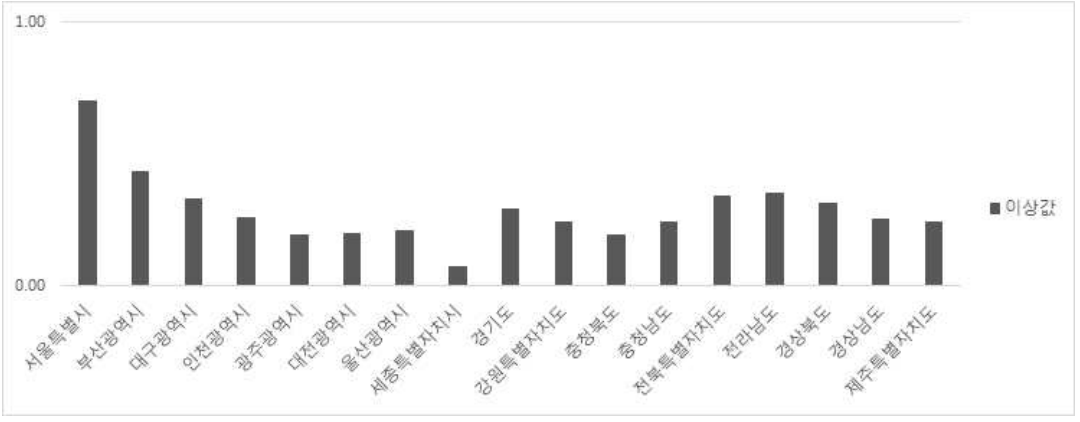
준공연도별로 이상값 분포를 분석한 결과, 이상값은 특정 시기에 집중되지 않고 고르게 나타나는 모습을 보여, 규칙 기반 오류와 다른 모습을 보였다. 시도별 분포에서도 서울특별시가 가장 높고, 그 다음으로 부산광역시가 높으나 나머지 시도에서는 큰 차이를 보이지 않았는데, 기존 검증규칙과 신규 규칙 모두에서 서울시는 오류율이 낮은 편에 속했던 것과 비교하면 기계학습 기반 이상값은 다른 오류와 확연히 다른 분포를 보이고 있음을 확인할 수 있다.

이처럼 기계학습 기반 이상탐지 결과는 일반적인 건축물의 특성에서 벗어나는 건축물을 이상값으로 탐지하여 제시하고 있으며, 이러한 사례는 규칙 기반 검증을 통해서는 확인하기 어려운 오류를 드러내고 있다. 다만 전문가가 설정한 규칙을 통하여 검증하였기 때문에 도출된 이상값을 바로 오류로 확정할 수 있는 규칙 기반 검증과 달리, 기계학습 기반 이상값은 오류 여부 확인에 사람의 개입이 추가적

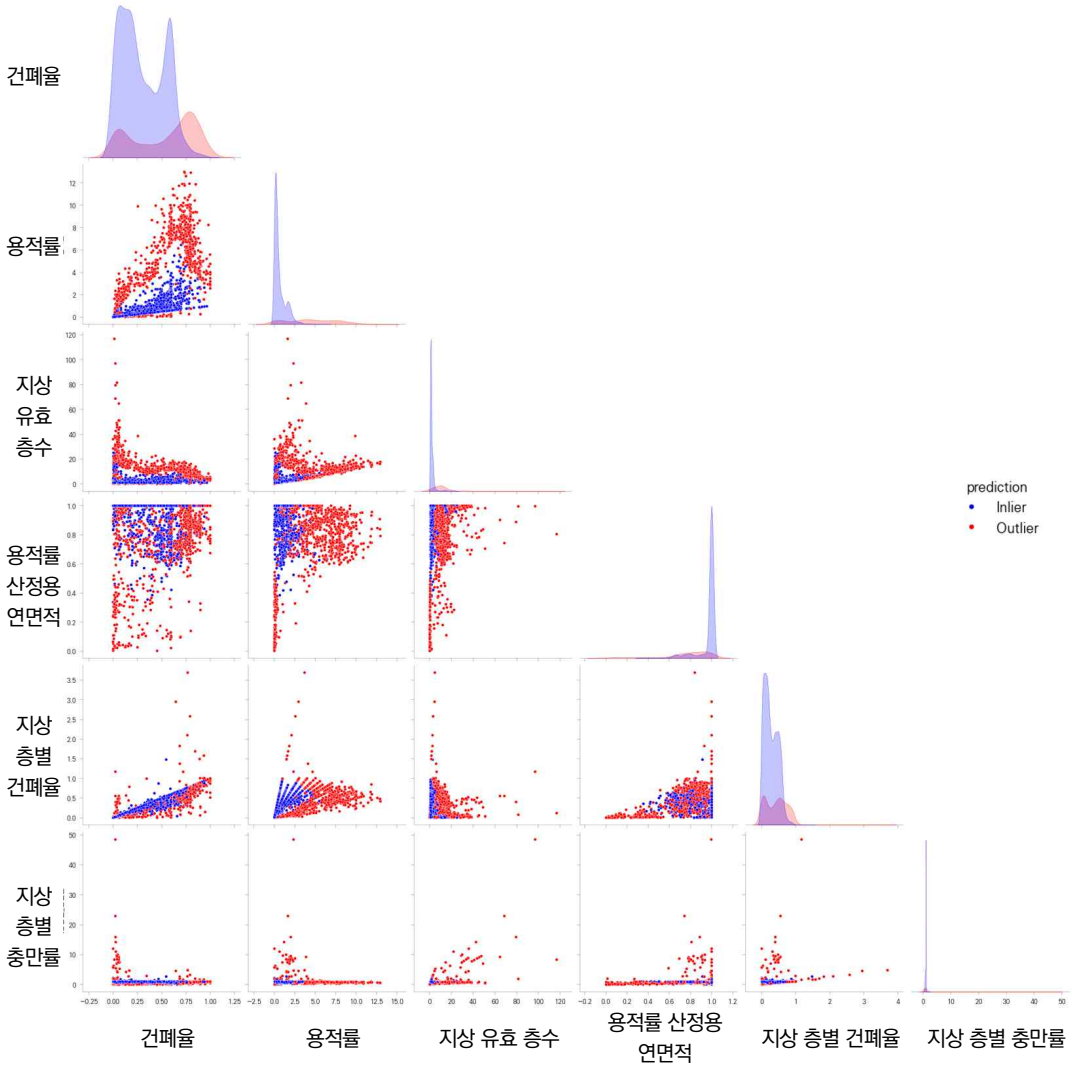
으로 필요하다. 다만 특정 시기나 지역에 한정되지 않는 오류 가능성이 높은 후보를 도출할 수 있다는 점에서 규칙 기반 검증과 함께 보완적으로 활용될 수 있는 방법론임을 확인할 수 있다.



[그림 3-26] 사용승인 연도별 이상값 비율
출처: 연구진 작성



[그림 3-27] 시도별 이상값 비율
출처: 연구진 작성



[그림 3-28] 기계학습 기반 이상값 산점도
출처: 연구진 작성

6) 결과 종합

본 절에서는 건축물대장 면적 데이터의 오류 및 이상 현황을 기존 업무규칙 기반 검증, 신규 검증규칙 개발, 그리고 기계학습을 통한 이상값 탐지를 종합하여 다각적으로 검토하였다. 건축물대장은 8,027,067동에 대한 데이터를 포함하고 있으나, 데이터상 0으로 기재된 결측값, 면적이 음수로 기재된 오류 등이 포함된 것으로 확인되었다. 기존 규칙 및 신규 규칙 기반 검증은 결측값을 포함한 정상 데이터 7,364,967동을 대상으로 수행하였고, 기계학습 기반 이상값 탐지는 모든 결측값과 오류를 제외한 4,703,097동을 대상으로 하였다.

먼저 기존 업무규칙 기반 검증의 경우, 대지면적·건축면적·연면적 등 기초 변수 간의 불일치를 식별하는 규칙을 중심으로 오류 현황을 집계하였다. 건축물대장 내 수치 데이터는 소수점 이하 근소한 차이가 다소 존재하여, 면적 데이터의 경우 1 ㎡, 퍼센트 단위 비율 데이터의 경우 0.1%의 오차범위를 설정하고, 이를 벗어나는 차이만 오류로 판정하였다.

대지면적이 건축면적보다 작은 오류는 9,773건 발생하여 0.13%의 오류율을 보였다. 2022년 건축물대장 정비 당시 오류율 0.19%와 유사한 수준으로, 오류가 잔존하는 것으로 판단되었다. 2005년 이후 준공 건축물에서는 오류율이 0에 가까운 수준으로 감소하여 현재는 오류가 거의 발생하지 않는 것으로 나타났다.

건축면적이 연면적보다 큰 오류는 373,505건 발생하여 5.01%의 오류율을 보였다. 2022년 당시 오류율 16.78%보다 낮지만 다른 오류 유형에 비해 높은 수준이었다. 시기별로는 1990년대 이후에 집중되어 있으며, 최근 준공 건축물에서도 여전히 8~10% 수준의 오류율을 보이고 있다. 지역별로는 제주특별자치도(10.34%) 등 도 지역에서 높게 나타나고 있다.

건폐율 산출 오류는 85,541건 발생하여 1.16%의 오류율을 보였다. 시기별로는 1990년대 준공 건축물에만 집중되어 있었고, 2000년대부터 꾸준히 감소하여 현재는 오류율이 0에 가까운 수준이었다.

용적률 산출 오류는 82,136건 발생하여 1.12%의 오류율을 보였다. 건폐율 오류와 마찬가지로 1990년대에만 집중되어 나타났으며, 지역별로도 두 오류가 같은 패턴을 보였다.

다음으로 기존 건축물대장 정비에서 확인된 적 없는 신규 검증규칙을 개발하고, 오류 현황을 집계하였다. 기존 검증규칙의 경우 연면적은 건축면적 오류 판단을 위한 비교 수치로만 사용되고 있어, 연면적 관련 신규 검증규칙을 2개 개발하였다.

용적률 산정 연면적이 (일반) 연면적보다 큰 오류는 19,295건 발생하여 0.26%의 오류율을 보였다. 시기별로는 1990년대 준공 건축물에서 오류율이 급격히 증가 후 점차 감소하여 현재는 오류율이 0에 가까운 수준으로 감소하였다. 지역별로는 세종특별자치시(0.68%)에서 오류율이 높게 나타났다.

지상층 바닥면적의 상한을 건축면적으로, 지하층 상한을 대지면적으로 설정하여 산출한 연면적 상한을 초과하는 오류는 63,313건 발생하여 0.86%의 오류율을 보였다. 시기별로는 1990년대 이후 기간에서 1% 이상의 오류율이 주기적으로 발생하는 것으로 나타나고 있다. 지역별로는 차이가 크지는 않으나 도 지역이 특광역시 지역보다 오류율이 높게 나타났다.

마지막으로 기계학습 기반 이상값 탐지를 통해서는 규칙 기반 검증으로는 포착하기 어려운 비정형적 오류가 추가적으로 드러났다. 대지면적 대비 건축면적 비율, 연면적 대비 용적률 산정용 연면적 비율, 층별 충만률 등 무차원 지표 기반으로, Isolation Forest와 One-Class SVM 등 방법론을 병행 적용한 결과, 알고리즘별 차이를 확인하고, 상호 보완적으로 활용할 수 있는 가능성을 검증하였다. 기계학습 기반 이상값 판정 결과를 시도·연도 단위로 집계한 결과, 시기별로 일정한 오류율이 나타나 다른 규칙 기반 검증 결과와 차이를 보였다. 지역별로는 서울, 부산 순으로 높은 오류율을 보였는데, 기존 및 신규 규칙 기반 검증에서는 서울의 오류율이 가장 낮은 편에 속하여, 지역별 분포에서도 차이를 보였다. 이는 단순히 규칙 기반으로 탐지할 수 없는 오류가 정정되지 않고 있다는 점을 시사한다.

종합하면, 건축물대장 면적 데이터 대상 시범적용 결과는 규칙 기반 검증과 기계학습 기반 탐지의 상호보완적 활용이 데이터 품질 고도화에 필수적임을 확인시켜준다. 규칙 기반 검증은 법적·제도적 기준에 부합하지 않는 명백한 오류를 신속하게 걸러내는 데 강점을 가지며, 기계학습 기반 탐지는 데이터 분포 상의 비정상적 패턴을 탐지하여 잠재적 오류를 보완적으로 식별할 수 있으므로, 향후 데이터 품질 고도화 방향은 이러한 방법론을 복합적으로 적용할 필요가 있다.

2. 건축물대장 용도 데이터 품질 고도화

1) 개요

본 절에서는 건축물대장 내 용도 데이터를 대상으로 체계적인 품질 진단과 오류 탐지를 수행하였다. 분석 데이터 구축을 통해 용도 분류 체계 및 기타 기재 항목을 정비하였다. 용도별 단어 출현 빈도 및 분포를 분석함으로써 데이터의 전반적 현황과 특성을 파악하였다. 이를 바탕으로 용도 분류 코드의 불일치, 비표준 용어 사용 등 주요 오류 가능성을 검토하였다.

다음으로, 기계학습 기반 접근법을 적용하여 데이터의 잠재적 오류를 탐지하였다. 나이브 베이즈(Naive Bayes) 분류 알고리즘을 활용하여 건축물의 구조적·물리적 속성(예: 연면적, 층수, 주요구조, 주용도 키워드 등)에 기초한 예측 용도 분류를 산출하였다. 그 결과를 실제 기재된 용도 분류와 비교·검증하여 자동분류 결과와 실제 기재값 간의 불일치 사례를 도출하였다.

이와 같이 용도 데이터의 현황 검토와 통계적 탐지, 기계학습 기반 진단을 병행함으로써 규칙 기반 점검을 넘어 데이터의 내재적 오류를 실증적으로 파악하였다.

[그림 3-29] 건축물대장 용도 데이터 품질 고도화 흐름도
출처: 연구진 작성

2) 건축물대장 용도 데이터 현황

표제부에는 건축물 동의 주 용도에 대한 건축법에 따른 용도 구분(현행 29개 분류)이 기재된다. 기재 내용은 주 용도 구분에 대한 코드와 코드명으로 구성된다. 예를 들어 '01000'은 단독주택(대분류), '02000'은 공동주택 등이다. 다만 건축물 동의 용도가 기재되지 않은 경우도 약 3만 건에 달한다.

건축법에 따른 용도 구분은 건축법 및 하위법령 개정에 따라 변화하는데, 표제부에 기재된 용도 코드 및 코드명은 건축물대장 작성 당시 구분을 아직 따르고 있는 경우가 존재한다. 예를 들어 현행 건축법에서 근린생활시설은 제1종근린생활시설과 제2종근린생활시설로 구분되고 있는데, 이에 해당하는 '03000', '04000' 코드 외에도 'Z3000'과 '근린생활시설'로 기재된 건축물 동 데이터도 약 1만 건에 달한다. 그 외에도 숫자 5자리로 정의된 주 용도 코드에 'JY' 등 정의에서 벗어난 데이터가 입력된 경우도 소수이지만 존재한다.

층별개요에는 건축물 층의 주 용도에 대한 건축법에 따른 세부 용도 구분이 기재된다. 아파트, 다세대주택, 연립주택 등이 표제부의 주 용도에는 동일하게 '02000' 공동주택으로 기재되지만, 층별개요에서는 각 층의 주 용도가 '02001' 아파트, '02002' 연립주택, '02003' 다세대주택 등으로 구분된다. 층별개요에서도 건축물 층의 용도가 기재되지 않은 경우도 약 2만 5천 건에 달하고 있으며, Z로 시작하는 과거 용도 구분이 기재된 경우도 존재한다. '창', '홀' 등 정의에서 벗어난 데이터도 발견된다.

건축물대장 표제부와 층별개요에는 주 용도 코드 및 코드명 외에도 '기타 용도'라는 데이터가 존재한다. 기타 용도는 건축물의 용도를 자유 텍스트로 표현하고 있는데, 건축물대장을 발급하였을 때 실제로 문서에 표기되는 내용은 이 기타 용도의 내용이 된다. 기타 용도에 기재되는 내용은 용도 구분에 따른 용도명과 동일한 경우가 가장 많지만, 그 외에도 실제 건축물의 용도를 반영하는 다양한 용도명 또는 여러 용도명의 집합으로 구성되는 경우가 많다. 예를 들어, 다가구주택(대분류 단독주택)의 경우 '단독주택(다가구주택)', '단독주택(5가구)', '다가구주택(14가구)' 등 다양한 형태로 기재되어 있으며 일관된 규칙이 존재하지 않는다.

층별개요의 기타 용도 데이터는 건축물 동 단위로 발급되는 건축물대장 표제부의 층별개요 부분에 기재되는 각 층별 용도를 표현한다. 표제부의 기타 용도 데이터와 차이점은 건축물 동 단위로는 건축물의 용도가 항상 존재하나, 층별개요에는 반드시 층별 용도를 지칭하지 않는 데이터가 존재하는 경우가 있다. 예를 들어, 연면적에서 제외되는 면적은 물리적으로는 별도 층이 아니지만, 층별개요에서 그 용도가 '연면적 제외'인 별도 항목으로 기재되는 경우가 있다. 따라서 층별개요 데이터가 건축물의 층 정보를 정확하게 반영하고 있다고는 할 수 없으며, 층별개요의 용도 관련 데이터도 동일 한계가 있다.

표제부의 건축물 동별 주 용도 코드, 용도명과 기타 용도 데이터의 관계를 많이 나타나는 사례 순으로 살펴보면, 용도명과 기타 용도 데이터가 일치하는 경우가 가장 많다. 그러나 용도 구분과 정확히 일치하지 않는 경우, 여러 용도가 기재된 경우도 상당수 있다. 단독주택 용도 건축물에 기타 용도가 '주택'이라고 기재되어 있거나, 제1종근린생활시설 및 제2종근린생활시설에 기타 용도가 단순히 '근린생활

[표 3-22] 표제부 용도별 건축물 동 데이터 현황 (상위)

코드	용도명	건수
01000	단독주택	4,344,979
04000	제2종근린생활시설	719,292
03000	제1종근린생활시설	595,408
02000	공동주택	578,529
21000	동물및식물관련시설	440,924
18000	창고시설	413,645
17000	공장	404,723
10000	교육연구시설	89,291
11000	노유자시설	48,284
15000	숙박시설	47,399
06000	종교시설	40,446
19000	위험물저장및처리시설	37,588
14000	업무시설	37,176
20000	자동차관련시설	37,005
-	(미기재)	29,129

코드	용도명	건수
23000	교정및군사시설	21,555
05000	문화및집회시설	16,114
13000	운동시설	14,176
30000	자원순환관련시설	13,221
22000	분뇨.쓰레기처리시설	12,880
07000	판매시설	12,551
Z3000	근린생활시설	10,771
27000	관광휴게시설	9,466
09000	의료시설	7,607
31000	야영장시설	7,233
16000	위락시설	5,494
12000	수련시설	5,191
33000	국방,군사시설	5,060
25000	발전시설	4,602
08000	운수시설	3,974

출처: 연구진 작성

시설'로 기재된 경우가 상위권에 든다. 또한, 동별 용도는 대분류에 따라 코드가 부여되지만, 기타 용도에는 '아파트', '다세대주택' 등 소분류가 기재된 경우도 존재한다.

층별개요의 기타 용도 데이터는 표제부의 경우와 반대로 주 용도 코드, 용도명은 건축물 전체의 세부 용도로 부여되지만, 기타 용도에서 해당 층의 실제 용도를 기재하고 있는 경우가 많다. 예를 들어 단독주택에 딸린 부속건축물을 단순히 '창고', '부속사' 등으로 층별개요에 기재하는 경우 등이다. 따라서 층별개요의 경우 해당 층의 기타 용도 데이터가 그 층의 정확한 용도 구분을 나타낸다고 보기는 어렵다.

기타 용도 데이터를 단어별로 분리하여 등장 빈도를 분석해보면, 표제부에서는 '단독주택' 및 '주택'을 합쳐 약 420만 건으로 가장 많이 나타난다. 그 다음으로 '근린생활시설'이 약 51만 건으로 많은데, 이는 '제1종근린생활시설' 및 '제2종근린생활시설'과는 별개의 단어다. 예를 들어, 제1종근린생활시설 건축물의 건축물대장 표제부에 용도가 '제1종근린생활시설'이 아니라 '근린생활시설'로 기재된 경우 등이 더 많음을 의미한다. 유사하게 근린생활시설은 단순히 용도만이 아니라 규모 등 다른 조건에 따라 1, 2종이 구분되는데, 기타 용도 데이터에서 단순히 세부 용도(점포, 소매점, 사무실 등)만을 기재하고 있는 경우가 많았다.

층별개요의 기타 용도 데이터를 단어별로 분리하여 등장 빈도를 분석하면 단순히 '주택'이 가장 많이 나타나 약 386만 건에 달한다. 그 다음으로는 아파트, 단독주택, 창고, 다가구주택, 다세대주택 등 세부 용도를 나타내는 단어가 뒤를 잇는다. 층별개요는 주 용도 코드와 기타 용도 데이터 모두 세부 용도를 기재하고 있어 두 기재내용이 일치하는 경우가 많다. 그러나 공동주택, (제1종, 제2종)근린생활시

[표 3-23] 층별개요 용도별 건축물 층 데이터 현황 (상위)

코드	용도명	건수	코드	용도명	건수
01001	단독주택	6,118,573	17301	물품 제조공장	77,765
02001	아파트	2,877,653	10101	초등학교	67,484
01003	다가구주택	1,923,975	11201	노인복지시설	65,775
02003	다세대주택	1,261,033	14299	기타일반업무시설	63,723
03001	소매점	903,494	20001	주차장	59,401
18001	창고	655,616	06101	교회	58,931
04402	사무소	600,119	03021	마을회관	54,167
04001	일반음식점	563,381	04006	수리점	49,329
21101	축사	493,776	03007	마을공동시설	43,710
17100	일반공장	432,244	03011	대피소	41,623
03999	기타제1종근린생활시설	330,306	03199	기타공공시설	40,845
17999	기타공장	284,297	10103	고등학교	38,953
02005	부대시설	234,621	10999	기타교육연구시설	36,370
04999	기타제2종근린생활시설	221,739	10104	대학교	35,347
14202	오피스텔	213,301	04002	휴게음식점	31,598
04005	제조업소	196,149	02006	복리시설	30,457
21999	기타동식물관련시설	164,832	10003	학원	30,258
02002	연립주택	157,461	19001	주유소	29,711
18999	기타창고시설	135,702	10102	중학교	29,482
03002	휴게음식점	116,898	23003	국방 · 군사시설	29,010
03005	의원	112,282	15199	기타일반숙박시설	28,669
04010	학원	108,025	04201	교회	28,198
04499	기타사무소	90,362	17200	유해공장	26,926
18101	일반창고	90,040	15002	생활숙박시설	25,828
15102	여관	87,025	04035	고시원	25,587
14204	사무소	82,929	-	(미기재)	25,487
01002	다중주택	80,811	09107	병원	22,981

출처: 연구진 작성

설, 업무시설 등 대분류에 해당하는 단어도 층별개요 기타 용도 데이터 등장 빈도 상위권에 위치하고 있다. 이는 '공동주택(아파트)', '제2종근린생활시설(사무소)' 등 대분류를 먼저 기재하고 괄호 안에 세부 용도를 기재하는 경우가 많기 때문으로 판단된다. 마찬가지로 세부 용도 이후에 층별 이용 현황에 대한 상세내역을 부기하는 경우가 있다. '다가구주택(1가구)', '다세대주택(2세대)', '계단실(연면적제외)' 등 자주 등장하는 부기 내용으로 인하여 '1가구'(약 34만 건), '연면적제외'(약 23만 건) 등 단어의 등장 빈도도 높은 편에 속하고 있다.

다음으로 표제부, 층별개요에서 나타나는 주 용도(코드, 용도명)와 기타 용도 데이터의 연관관계를 분석하였다. 가장 등장 빈도가 높은 '단독주택' 및 '주택'의 경우 주 용도가 단독주택으로 분류된 경우가

[표 3-24] 표제부 동별 용도 관련 데이터 예시

코드	용도명	'기타 용도' 데이터
01000	단독주택	단독주택(다가구주택),제1종근린생활시설((소매점)
		다가구주택(14가구)
		공동주택(다세대), 근린생활시설
		단독주택(5가구)
		법당, 주택
		근린생활시설및 주택
		단독주택(다가구주택 및 근생)
		근린생활시설, 공동주택(다세대주택)
		주택, 변소및창고
		주택.근린생활시설
03000	제1종근린생활시설	의료시설(소매점, 제1종근린생활시설)
		제1종근린생활시설(소매점),창고시설
		제1종,제2종 근린생활시설, 위락시설
		마을회관
		점포, 주택
		수위실
		근린생활시설, 교육연구시설, 단독주택
		펌프장
		제1종근린생활시설(휴게음식점, 마을공동시설)
		탁구장
04000	제2종근린생활시설	위락시설 및 제1,2종근린생활시설
		제1,2종근린생활시설, 주택
		2종근린생활시설, 주택
		공장, 사무실
		휴게음식점
		교회당
		근린생활시설, 다가구주택(4가구)
		근린생활시설, 교육연구시설
		제2종근린생활시설(단독주택)
		숙소

출처: 연구진 작성

많았으나, (제1종, 제2종)근린생활시설, 창고시설 등 다른 용도로 분류된 경우도 약 26만 건에 달하였다. 마찬가지로 '근린생활시설', '(제1종, 제2종)근린생활시설' 등 단어가 기타 용도에 포함되었으나 주 용도가 근린생활시설이 아닌 단독주택, 공동주택, 업무시설 등인 건축물이 약 32만 동이었다. 기타 용도에 다른 용도를 지칭하는 단어가 포함된 것이 그 자체로 데이터 오류의 근거가 될 수는 없다.

[표 3-25] 층별개요 층별 용도 관련 데이터 예시

코드	용도명	'기타 용도' 데이터
01001	단독주택	주택,소매점
		주택
		ㄴ
		차고
		창고
01003	다가구주택	다가구용주택(2가구)
		보일러실
		단독주택(다가구주택,5가구)
		계단실
		단독주택(다가구주택2가구)
02003	다세대주택	도시형생활주택(단지형다세대)-2세대
		단지형다세대주택(1세대)
		단지형다세대주택(계단실)
		공동주택(계단실)
		다세대(1세대)
03999	기타제1종근린생활시설	창고
		(연면적 제외)
		물치
		점포,목욕장
		연면적제외
04999	기타제2종근린생활시설	숙직실
		화장실
		기타제2종근린생활시설
		제2종근린생활시설(영업용)
		체력단련장
14299	기타일반업무시설	주차장
		램프
		복도
		(연면적제외)
		기계실,물탱크실

출처: 연구진 작성

[표 3-26] 표제부 동별 용도 관련 데이터 중 최다 항목

코드	용도명	'기타 용도' 데이터	건수
01000	단독주택	단독주택	1,722,659
01000	단독주택	주택	1,592,478
04000	제2종근린생활시설	제2종근린생활시설	327,131
17000	공장	공장	304,174
18000	창고시설	창고시설	230,618
03000	제1종근린생활시설	제1종근린생활시설	185,441
21000	동물및식물관련시설	동.식물관련시설	158,005
02000	공동주택	공동주택	112,514
02000	공동주택	다세대주택	86,467
03000	제1종근린생활시설	근린생활시설	81,974
04000	제2종근린생활시설	근린생활시설	80,714
18000	창고시설	창고	74,362
02000	공동주택	공동주택(아파트)	72,079
21000	동물및식물관련시설	동물및식물관련시설	71,888
01000	단독주택	농가주택	63,702
02000	공동주택	아파트	52,504
21000	동물및식물관련시설	축사	46,340
10000	교육연구시설	교육연구시설	44,507
01000	단독주택	다가구주택	42,196
01000	단독주택	주택, 창고	40,403
21000	동물및식물관련시설	동물관련시설	35,544
03000	제1종근린생활시설	점포	26,767
03000	제1종근린생활시설	근린생활시설, 주택	25,912
01000	단독주택	근린생활시설, 주택	25,833
02000	공동주택	공동주택(다세대주택)	25,447
01000	단독주택	단독주택(다가구주택)	24,853
15000	숙박시설	숙박시설	24,087
01000	단독주택	주택, 근린생활시설	24,063
01000	단독주택	주택, 점포	22,333

출처: 연구진 작성

[표 3-27] 층별개요 층별 용도 관련 데이터 중 최다 항목

코드	용도명	'기타 용도' 데이터	건수
01001	단독주택	주택	3,192,276
02001	아파트	아파트	1,749,831
01001	단독주택	단독주택	1,568,151
02001	아파트	공동주택(아파트)	703,628
18001	창고	창고	462,082
02003	다세대주택	다세대주택	319,380
03001	소매점	소매점	283,287
04001	일반음식점	일반음식점	221,384
21101	축사	축사	219,071
01003	다가구주택	주택	211,141
01003	다가구주택	다가구주택	203,381
04402	사무소	제2종근린생활시설(사무소)	201,476
04402	사무소	사무소	194,720
03001	소매점	제1종근린생활시설(소매점)	190,472
01001	단독주택	창고	182,852
02003	다세대주택	다세대주택(2세대)	176,864
04001	일반음식점	제2종근린생활시설(일반음식점)	172,643
17100	일반공장	공장	151,485
03001	소매점	점포	151,353
03999	기타제1종근린생활시설	근린생활시설	145,800
01003	다가구주택	다가구주택(1가구)	137,020
01001	단독주택	변소	118,406
01003	다가구주택	다가구주택(2가구)	118,249
01001	단독주택	농가주택	114,026
01003	다가구주택	단독주택(다가구주택)	98,522
04402	사무소	사무실	81,901
01003	다가구주택	주택(1가구)	81,416
01001	단독주택	단독주택(단독주택)	80,157
04005	제조업소	제2종근린생활시설(제조업소)	73,132
01001	단독주택	부속사	72,147

출처: 연구진 작성

[표 3-28] 기타 용도 단어 빈도 분석 (중복 포함)

동별 기타 용도 단어	건수
단독주택	2,114,809
주택	2,069,361
근린생활시설	513,863
제2종근린생활시설	504,077
공장	362,588
제1종근린생활시설	290,552
창고시설	280,232
창고	272,363
공동주택	262,082
식물관련시설	187,056
동	176,831
다세대주택	165,076
다가구주택	150,273
아파트	133,031
점포	114,963
축사	99,921
동물및식물관련시설	81,241
및	78,442
농가주택	78,145
소매점	70,131
2종근린생활시설	67,233
교육연구시설	66,596
제1	60,711
사무실	57,564
동물관련시설	45,354
다가구용단독주택	45,153
변소	44,164
사무소	42,023
업무시설	40,072
숙박시설	38,638

층별 기타 용도 단어	건수
주택	3,857,882
아파트	2,584,109
단독주택	2,217,159
창고	1,067,948
공동주택	1,005,215
다가구주택	982,559
다세대주택	850,405
제2종근린생활시설	803,173
소매점	611,069
계단실	592,979
사무소	569,395
공장	473,228
근린생활시설	472,425
일반음식점	467,760
제1종근린생활시설	428,788
축사	357,873
사무실	344,953
1가구	341,216
2세대	338,449
주차장	304,072
2가구	277,684
화장실	234,731
연면적제외	234,214
점포	233,634
변소	185,847
업무시설	176,578
도시형생활주택	169,266
오피스텔	161,115
제조업소	157,698
기계실	127,458

출처: 연구진 작성

[표 3-29] 표제부 주 용도별 기타 용도 상위 단어 빈도 분석 (중복 포함)

코드	용도명	단독주택	주택	근린생활시설	제2종 근린생활시설	공장	제1종 근린생활시설	창고시설	창고	공동주택	식물 관련시설
01000	단독주택	2,013,416	1,905,896	149,834	59,280	1,812	37,238	18,215	123,186	1,552	1,035
02000	공동주택	381	7,592	17,905	6,709	197	4,613	119	848	257,505	2
03000	제1종 근린생활시설	33,428	78,241	161,711	14,378	628	229,181	3,758	7,722	519	37
04000	제2종 근린생활시설	46,781	44,144	150,633	410,247	2,671	11,261	4,150	7,960	451	68
05000	문화 및 집회시설	153	101	457	266	13	250	38	141	5	9
06000	종교시설	622	1,080	789	990	4	221	44	520	4	1
07000	판매시설	84	146	1,163	509	18	177	112	153	78	0
09000	의료시설	53	114	568	224	1	477	17	129	4	0
10000	교육연구시설	483	513	2,004	750	146	621	82	2,014	61	5
11000	노유자시설	793	537	958	580	15	1,111	52	445	24	0
13000	운동시설	98	55	441	343	9	259	36	207	5	1
14000	업무시설	658	439	7,343	1,356	214	1,434	181	480	1,689	1
15000	숙박시설	1,324	1,991	4,827	1,592	15	642	114	411	8	1
16000	위락시설	406	646	1,287	527	9	189	52	77	0	1
17000	공장	698	2,308	1,999	1,108	353,329	410	996	17,227	68	16
18000	창고시설	10,184	11,151	2,032	2,927	2,366	1,116	250,220	97,398	13	337
19000	위험물 저장 및 처리 시설	199	285	824	386	341	338	43	386	2	0
20000	자동차 관련시설	277	175	1,138	954	261	397	97	491	38	4
21000	동물 및 식물 관련시설	3,331	9,767	220	215	124	70	1,459	10,254	3	185,441

출처: 연구진 작성

건축물이 여러 용도로 사용되는 복합건축물의 경우, 기타 용도에 모든 용도가 기재되지만 주 용도는 그 중 하나만 기재되므로 특정 용도를 지칭하는 단어가 다른 용도 건축물의 기타 용도 데이터에 포함될 수 있다. 그러나 단독주택과 공장처럼 복합용도 건축물로 볼 수 없는 사례도 상당수 존재해, 용도 관련 데이터의 교차검증 필요성을 확인할 수 있다.

[표 3-30] 층별개요 주 용도별 기타 용도 상위 단어 빈도 분석 (중복 포함)

코드	용도명	주택	아파트	단독주택	창고	공동주택	다가구 주택	다세대 주택	제2종 근린생활 시설	소매점	계단실
01000	단독주택	5,598	0	787	1,240	0	111	8	2	4	13
01001	단독주택	3,304,862	14	1,850,839	248,983	90	16,365	1,283	718	3,795	23,325
01002	다중주택	2,868	55	24,639	154	49	32	16	15	34	9,101
01003	다가구 주택	443,844	300	335,083	13,907	883	964,535	28,517	377	888	154,179
02000	공동주택	286	376	9	3	324	0	112	1	1	16
02001	아파트	3,278	2,578,652	7	1,549	802,980	0	617	301	339	51,640
02002	연립주택	10,103	965	68	749	26,645	18	743	82	162	7,325
02003	다세대 주택	57,668	162	719	1,836	164,745	322	818,436	301	408	154,203
02004	생활편익 시설	22	296	1	49	240	2	2	1,415	179	236
02005	부대시설	77	1,079	14	2,636	3,294	12	203	68	42	25,295
02006	복리시설	0	112	0	129	661	0	6	102	45	444
02007	기숙사	46	66	1	141	3,592	0	1	8	10	354
03001	소매점	13,134	83	824	8,933	66	236	91	4,408	590,818	9,704
03002	휴게 음식점	95	0	49	662	0	17	1	522	515	1,157
03005	의원	194	9	25	369	8	6	8	1,663	437	3,210
03007	마을 공동시설	18	0	24	22,775	0	0	0	5	16	141
03021	마을회관	29	0	7	1,765	5	2	2	18	17	252
03199	기타 공공시설	44	2	20	745	1	8	0	80	62	547
03202	사무소	9	0	9	154	2	5	2	551	34	375
03999	기타 제1종 근린생활 시설	3,904	60	266	4,891	112	67	50	3,114	3,119	14,249
04001	일반 음식점	877	10	392	5,989	16	126	16	201,845	2,511	7,051
04002	휴게 음식점	37	0	5	225	1	3	0	11,272	187	551
					…						

출처: 연구진 작성

층별개요에서도 주 용도와 기타 용도 데이터의 교차검증 필요성이 확인된다. 먼저 주 용도 코드가 세부 용도가 아닌 대분류로 잘못 기재된 경우가 있다. 단독주택의 경우, 표제부는 단독주택(대분류)에 해당하는 '01000'으로 기재되고, 층별개요에서 단독주택(세부 용도)에 해당하는 '01001'로 기재되어야 하며, 대부분의 경우 정확히 기재되어있다. 그러나 층별개요 주 용도 코드와 용도명에 대분류에 해

당하는 '01000 단독주택' 또는 '02000 공동주택'으로 기재된 경우가 적지 않게 존재한다. 특히 이들 오류 사례 중에는 기타 용도에 '다가구주택', '다세대주택' 등 단독주택(세부 용도)이 아닌 다른 용도를 지칭하는 단어가 포함된 경우도 있다. 따라서 주 용도와 기타 용도의 교차검증을 통하여 용도 관련 데이터가 정확히 기재되지 않은 경우를 탐지할 필요가 있다.

종합하면, 표제부와 층별개요에 기재된 건축물 동별, 층별 주 용도 및 기타 용도 데이터는 상당수 동일하거나 대체로 일치하였으나, 일부 경우에는 주 용도와 기타 용도의 불일치, 나아가 오류 가능성이 확인되는 경우가 있었다. 그러나 표제부의 주 용도는 건축법에 따른 용도 대분류를 기재하고 기타 용도에는 세부 용도에 대한 내용을 기재하고 있는 등 단순히 1:1로 비교할 수 없는 경우가 많았다. 따라서 건축물대장 용도 관련 데이터의 교차검증을 위해서는 표제부와 층별개요의 주 용도(코드, 용도명), 기타 용도 데이터를 종합적으로 검토할 필요가 있다.

3) 기계학습 기반 용도 데이터 오류 탐지

나이브 베이즈 예측 모델을 활용한 분석의 첫 단계로 분석 데이터를 구축하였다. 표제부와 층별개요에서 주 용도 코드 및 용도명, 기타 용도 텍스트가 모두 정상적으로 존재하는 경우만 선별하였으며, 충분한 훈련 데이터를 확보하기 위하여 주 용도 및 기타 용도에 최소 일정 건수(동 기준 100건, 층 기준 200건) 이상 데이터가 존재하는 경우만 포함하였다. 이러한 과정을 거쳐 구축된 데이터는 동 기준 7,951,068건, 층 기준 20,943,324건이었다. 동, 층 모두 단독주택이 가장 많았으며, 제2종, 제1종 근린생활시설이 동 기준 각각 2, 3위로 나타났다. 아파트는 동 단위로는 근린생활시설보다 적었으나 층 단위로는 2위였다.

나이브 베이즈 모델을 기타 용도 단어에서 주 용도 코드를 예측하는 나이브 베이즈 모델을 학습시킨 결과, 동별 용도 예측은 84.78%, 층별 예측은 78.97% 정확도를 달성하였다. 그러나 전반적으로 단순한 모델로 높은 정확도를 달성하였음에도, 세부 예측 결과를 검토한 결과 여러 한계가 발견되었다. 먼저 대다수 건축물이 단독주택, 제2종 근린생활시설 용도인 불균형한 현황을 반영하여 예측 결과도 단독주택 등으로 잘못 분류하는 경우가 많았다. 또한, 여러 용도가 복합된 경우('주택, 근린생활시설' 등)가 과적합되어 텍스트의 의미와 전혀 다른 용도로 분류하는 경우가 발생하였다.

층별 예측 결과를 검토한 결과, 각 층의 기타 용도 텍스트만으로는 해당 층의 실제 용도를 예측하기 어려운 경우가 있음을 확인하였다. 예를 들어 단독주택에 딸린 창고의 경우 층별개요에서 '창고'로 기재하고, 이에 따라 건축물대장 문서에도 창고로 표출되더라도, 해당 층의 주 용도는 단독주택(세부 용도)으로 기재된다. 해당 층의 기타 용도만으로는 이러한 점을 파악할 수 없으므로, 층별 용도 데이터의 정확성을 판단하기 위해서는 건축물 모든 층의 용도를 고려할 필요가 있음을 확인하였다.

[표 3-31] 분석 데이터 내 주 용도별 건축물 동, 층 수

코드	용도명	동 수
01000	단독주택	4,326,700
04000	제2종근린생활시설	718,151
03000	제1종근린생활시설	594,056
02000	공동주택	578,273
21000	동물및식물관련시설	439,989
18000	창고시설	409,787
17000	공장	404,164
10000	교육연구시설	88,934
11000	노유자시설	48,184
15000	숙박시설	47,324
06000	종교시설	40,251
19000	위험물저장및처리시설	37,497
14000	업무시설	37,108
20000	자동차관련시설	36,908
23000	교정및군사시설	20,999
05000	문화및집회시설	15,969
13000	운동시설	14,124
30000	자원순환관련시설	13,204
22000	분뇨.쓰레기처리시설	12,842
07000	판매시설	12,500
27000	관광휴게시설	9,419
09000	의료시설	7,592
31000	야영장시설	7,232
16000	위락시설	5,482
12000	수련시설	5,174
33000	국방,군사시설	4,993
25000	발전시설	4,602
08000	운수시설	3,966
24000	방송통신시설	3,343
26000	묘지관련시설	1,783
29000	장례시설	518
합계		7,951,068

코드	용도명	층 수
01001	단독주택	6,107,851
02001	아파트	2,866,578
01003	다가구주택	1,922,416
02003	다세대주택	1,260,327
03001	소매점	902,487
18001	창고	655,182
04402	사무소	599,795
04001	일반음식점	563,247
21101	축사	493,734
17100	일반공장	432,017
03999	기타제1종근린생활시설	327,167
17999	기타공장	284,157
02005	부대시설	234,099
04999	기타제2종근린생활시설	220,235
14202	오피스텔	213,283
04005	제조업소	196,058
21999	기타동식물관련시설	164,800
02002	연립주택	156,682
18999	기타창고시설	135,639
03002	휴게음식점	116,871
03005	의원	112,237
04010	학원	107,982
04499	기타사무소	90,263
18101	일반창고	90,025
15102	여관	86,848
14204	사무소	82,760
01002	다중주택	80,799
17301	물품 제조공장	77,761
10101	초등학교	67,422
	…	
04039	복합유통게임제공업의시설	204
합계		20,943,324

출처: 연구진 작성

[표 3-32] 표제부 나이브 베이즈 예측 결과

코드	용도명	기타 용도 텍스트	건수	예측 코드	예측 용도명	1: 일치 0: 불일치
01000	단독주택	단독주택	1,722,659	01000	단독주택	1
01000	단독주택	주택	1,592,478	01000	단독주택	1
04000	제2종근린생활시설	제2종근린생활시설	327,131	04000	제2종근린생활시설	1
17000	공장	공장	304,174	17000	공장	1
18000	창고시설	창고시설	230,618	18000	창고시설	1
03000	제1종근린생활시설	제1종근린생활시설	185,441	04000	제2종근린생활시설	0
21000	동물및식물관련시설	동.식물관련시설	158,005	21000	동물및식물관련시설	1
02000	공동주택	공동주택	112,514	02000	공동주택	1
02000	공동주택	다세대주택	86,467	02000	공동주택	1
03000	제1종근린생활시설	근린생활시설	81,974	01000	단독주택	0
04000	제2종근린생활시설	근린생활시설	80,714	01000	단독주택	0
18000	창고시설	창고	74,362	01000	단독주택	0
02000	공동주택	공동주택(아파트)	72,079	02000	공동주택	1
21000	동물및식물관련시설	동물및식물관련시설	71,888	21000	동물및식물관련시설	1
01000	단독주택	농가주택	63,702	01000	단독주택	1
02000	공동주택	아파트	52,504	02000	공동주택	1
21000	동물및식물관련시설	축사	46,340	21000	동물및식물관련시설	1
10000	교육연구시설	교육연구시설	44,507	10000	교육연구시설	1
01000	단독주택	다가구주택	42,196	01000	단독주택	1
01000	단독주택	주택, 창고	40,403	01000	단독주택	1
21000	동물및식물관련시설	동물관련시설	35,544	21000	동물및식물관련시설	1
03000	제1종근린생활시설	점포	26,767	01000	단독주택	0
03000	제1종근린생활시설	근린생활시설, 주택	25,912	01000	단독주택	0
01000	단독주택	근린생활시설, 주택	25,833	01000	단독주택	1
02000	공동주택	공동주택(다세대주택)	25,447	02000	공동주택	1
01000	단독주택	단독주택(다가구주택)	24,853	01000	단독주택	1
15000	숙박시설	숙박시설	24,087	15000	숙박시설	1
01000	단독주택	주택, 근린생활시설	24,063	01000	단독주택	1
01000	단독주택	주택, 점포	22,333	01000	단독주택	1
01000	단독주택	창고	22,231	01000	단독주택	1
01000	단독주택	변소	22,188	01000	단독주택	1
11000	노유자시설	노유자시설	22,081	11000	노유자시설	1
20000	자동차관련시설	자동차관련시설	21,966	20000	자동차관련시설	1

출처: 연구진 작성

[표 3-33] 층별개요 나이브 베이즈 예측 결과

코드	용도명	기타 용도 텍스트	건수	예측 코드	예측 용도명	1: 일치 0: 불일치
01001	단독주택	주택	3,192,276	01001	단독주택	1
02001	아파트	아파트	1,749,831	02001	아파트	1
01001	단독주택	단독주택	1,568,151	01001	단독주택	1
02001	아파트	공동주택(아파트)	703,628	02001	아파트	1
18001	창고	창고	462,082	18001	창고	1
02003	다세대주택	다세대주택	319,380	02003	다세대주택	1
03001	소매점	소매점	283,287	03001	소매점	1
04001	일반음식점	일반음식점	221,384	04001	일반음식점	1
21101	축사	축사	219,071	21101	축사	1
01003	다가구주택	주택	211,141	01001	단독주택	0
01003	다가구주택	다가구주택	203,381	01003	다가구주택	1
04402	사무소	제2종근린생활시설(사무소)	201,476	04402	사무소	1
04402	사무소	사무소	194,720	04402	사무소	1
03001	소매점	제1종근린생활시설(소매점)	190,472	03001	소매점	1
01001	단독주택	창고	182,852	18001	창고	0
02003	다세대주택	다세대주택(2세대)	176,864	02003	다세대주택	1
04001	일반음식점	제2종근린생활시설(일반음식점)	172,643	04001	일반음식점	1
17100	일반공장	공장	151,485	17100	일반공장	1
03001	소매점	점포	151,353	03001	소매점	1
03999	기타제1종근린생활시설	근린생활시설	145,800	03999	기타제1종근린생활시설	1
01003	다가구주택	다가구주택(1가구)	137,020	01003	다가구주택	1
01001	단독주택	변소	118,406	01001	단독주택	1
01003	다가구주택	다가구주택(2가구)	118,249	01003	다가구주택	1
01001	단독주택	농가주택	114,026	01001	단독주택	1
01003	다가구주택	단독주택(다가구주택)	98,522	01003	다가구주택	1
04402	사무소	사무실	81,901	04402	사무소	1
01003	다가구주택	주택(1가구)	81,416	01003	다가구주택	1
01001	단독주택	단독주택(단독주택)	80,157	01001	단독주택	1
04005	제조업소	제2종근린생활시설(제조업소)	73,132	04005	제조업소	1
01001	단독주택	부속사	72,147	01001	단독주택	1
01003	다가구주택	주택(2가구)	72,064	01003	다가구주택	1
02002	연립주택	연립주택	67,176	02002	연립주택	1
14202	오피스텔	업무시설(오피스텔)	64,755	14202	오피스텔	1

출처: 연구진 작성

또한, 단순히 기타 용도에 기재된 내용이 충분하지 않은 경우도 문제가 되었다. 다가구주택 기타 용도가 '주택', '단독주택'이거나, 일반 공장이 아닌 기타공장, 일반 창고가 아닌 기타창고시설이 단순히 '공장', '창고'로 기재된 경우 등, 해당 내용만으로는 세부 용도까지 일치 여부를 파악하기 어려운 경우였다.

건축물 용도별로 예측 정확도를 파악한 결과, 용도에 따른 예측 정확도가 크게 다른 것으로 나타났다. 동 단위 예측에서, 사례 수가 가장 많은 단독주택의 경우 예측 정확도가 96.83%에 달하였으나, 제1종 근린생활시설은 정확도가 16.59%에 불과하였다. 이는 개별 사례 검토에서 드러난 바와 같이 이름이 유사하고 대상 수가 더 많은 제2종 근린생활시설로 분류되는 경우가 많기 때문인 것으로 판단된다. 전반적으로 사례 수가 많은 용도에서 정확도가 높고, 수가 적은 용도에서 정확도가 낮았기 때문에, 전체 정확도는 높았으나 편향이 존재함을 확인하였다.

건축물 용도 예측 결과에 따른 동일한 데이터의 중복 사례 수를 검토한 결과, 이러한 경향이 모든 용도 건축물에서 전반적으로 나타나는 것을 알 수 있었다. 단독주택(대분류)의 경우, 나이브 베이즈 모델이 예측에 성공한 사례의 경우 동일한 주 용도, 기타 용도로 기재된 건축물 동 데이터가 평균 약 116건인 반면, 실패한 경우에는 약 10건에 그쳤다. 이러한 경향은 앞에서 논의한 제1종 근린생활시설을 제외하고 모든 용도에서 동일하게 나타났다. 이는 사례 수가 많은 용도 건축물의 예측 정확도가 더 높은 경향이 단순히 용도 구분에 따라서 일어나는 것이 아니라, 모든 사례에서 데이터가 충분치 않은 경우 잘못 분류될 확률이 높아진다는 것을 의미한다.

층 단위 예측 결과에서도 동일한 경향이 나타났다. 가장 수가 많은 단독주택의 정확도는 93.29%이었으나 '기타 제2종 근린생활시설'의 경우 4.01%에 그쳤다. 또한 대다수 용도에서 성공한 사례의 중복 건수가 실패 사례보다 크게 높게 나타났다.

동 단위 데이터와 층 단위 데이터, 건축물 용도 등을 가리지 않고 중복 사례 수가 많은 데이터에 대한 예측 정확도가 일관되게 높게 나타나는 것은 전형적인 과적합 현상에 해당한다. 이는 텍스트 데이터의 언어적 의미를 고려하지 않고 등장 빈도만으로 훈련하는 기계학습 방법론의 특성 때문으로 판단된다. 사례가 적은 데이터에 대한 정확도를 높이기 위해서는 건축법령 등 외부 텍스트를 훈련에 활용하거나, 사전 훈련된 LLM 모형 등을 활용하는 방안이 있다.

[표 3-34] 건축물 용도별 용도 예측 정확도 (동 단위)

코드	용도명	총 건수	성공 건수	정확도
01000	단독주택	4,326,700	4,189,356	96.83%
02000	공동주택	578,273	541,520	93.64%
03000	제1종근린생활시설	594,056	98,580	16.59%
04000	제2종근린생활시설	718,151	499,771	69.59%
05000	문화및집회시설	15,969	11,384	71.29%
06000	종교시설	40,251	28,507	70.82%
07000	판매시설	12,500	7,224	57.79%
08000	운수시설	3,966	2,674	67.42%
09000	의료시설	7,592	6,278	82.69%
10000	교육연구시설	88,934	76,632	86.17%
11000	노유자시설	48,184	32,812	68.10%
12000	수련시설	5,174	3,737	72.23%
13000	운동시설	14,124	10,700	75.76%
14000	업무시설	37,108	23,389	63.03%
15000	숙박시설	47,324	41,344	87.36%
16000	위락시설	5,482	2,346	42.79%
17000	공장	404,164	371,022	91.80%
18000	창고시설	409,787	277,607	67.74%
19000	위험물저장및처리시설	37,497	31,203	83.21%
20000	자동차관련시설	36,908	29,089	78.81%
21000	동물및식물관련시설	439,989	408,871	92.93%
22000	분뇨.쓰레기처리시설	12,842	9,477	73.80%
23000	교정및군사시설	20,999	16,717	79.61%
24000	방송통신시설	3,343	1,568	46.90%
25000	발전시설	4,602	3,893	84.59%
26000	묘지관련시설	1,783	1,344	75.38%
27000	관광휴게시설	9,419	7,420	78.78%
29000	장례시설	518	271	52.32%
30000	자원순환관련시설	13,204	11,358	86.02%
31000	야영장시설	7,232	6,320	87.39%
33000	국방,군사시설	4,993	3,715	74.40%

출처: 연구진 작성

[표 3-35] 건축물 용도 예측 결과별 평균 사례 수 (동 단위)

코드	용도명	성공 평균 사례 수	실패 평균 사례 수
01000	단독주택	115.62	10.47
02000	공동주택	22.75	4.06
03000	제1종근린생활시설	8.74	20.79
04000	제2종근린생활시설	30.10	9.40
05000	문화및집회시설	18.13	2.28
06000	종교시설	17.60	4.90
07000	판매시설	12.08	3.11
08000	운수시설	8.83	2.05
09000	의료시설	9.14	1.97
10000	교육연구시설	23.63	2.92
11000	노유자시설	21.52	6.42
12000	수련시설	21.35	3.73
13000	운동시설	17.12	2.43
14000	업무시설	11.03	2.97
15000	숙박시설	16.48	2.67
16000	위락시설	17.38	2.56
17000	공장	27.92	3.53
18000	창고시설	72.60	14.55
19000	위험물저장및처리시설	15.68	2.95
20000	자동차관련시설	21.79	3.39
21000	동물및식물관련시설	53.82	5.78
22000	분뇨.쓰레기처리시설	20.87	3.02
23000	교정및군사시설	103.19	7.55
24000	방송통신시설	11.53	3.51
25000	발전시설	31.91	3.17
26000	묘지관련시설	12.44	2.14
27000	관광휴게시설	19.48	2.57
29000	장례시설	15.06	2.55
30000	자원순환관련시설	39.99	3.13
31000	야영장시설	52.23	4.24
33000	국방,군사시설	86.40	21.66

출처: 연구진 작성

[표 3-36] 건축물 용도별 용도 예측 정확도 (층 단위, 일부)

코드	용도명	총 건수	성공 건수	정확도
01001	단독주택	6,107,851	5,698,090	93.29%
01003	다가구주택	1,922,416	1,445,495	75.19%
02001	아파트	2,866,578	2,673,013	93.25%
02002	연립주택	156,682	111,829	71.37%
02003	다세대주택	1,260,327	1,074,819	85.28%
02005	부대시설	234,099	114,137	48.76%
03001	소매점	902,487	800,650	88.72%
03002	휴게음식점	116,871	100,708	86.17%
03999	기타제1종근린생활시설	327,167	165,524	50.59%
04001	일반음식점	563,247	495,787	88.02%
04005	제조업소	196,058	166,074	84.71%
04402	사무소	599,795	537,567	89.63%
04999	기타제2종근린생활시설	220,235	8,822	4.01%
14202	오피스텔	213,283	165,329	77.52%
17100	일반공장	432,017	297,976	68.97%
17999	기타공장	284,157	64,747	22.79%
18001	창고	655,182	567,636	86.64%
18999	기타창고시설	135,639	41,216	30.39%
21101	축사	493,734	445,586	90.25%
21999	기타동식물관련시설	164,800	21,194	12.86%

출처: 연구진 작성

[표 3-37] 건축물 용도 예측 결과별 평균 사례 수 (층 단위, 일부)

코드	용도명	성공 평균 사례 수	실패 평균 사례 수
01001	단독주택	216.30	31.33
01003	다가구주택	93.28	47.43
02001	아파트	399.61	9.10
02002	연립주택	55.39	8.87
02003	다세대주택	54.51	17.22
02005	부대시설	13.97	11.53
03001	소매점	63.01	6.97
03002	휴게음식점	87.04	5.90
03005	의원	45.11	3.83
03999	기타제1종근린생활시설	61.86	6.70
04001	일반음식점	88.79	5.70

04005	제조업소	41.19	4.70
04010	학원	81.43	5.05
04402	사무소	128.76	5.97
04999	기타제2종근린생활시설	3.29	9.58
14202	오피스텔	30.03	4.63
17100	일반공장	16.82	6.38
17999	기타공장	11.05	8.53
18001	창고	106.84	9.27
18999	기타창고시설	28.64	11.91
21101	축사	49.00	9.78
21999	기타동식물관련시설	14.18	15.02

출처: 연구진 작성

4) 결과 종합

건축물대장 용도 데이터 분석 결과, 표제부와 층별개요의 주 용도와 기타 용도 데이터는 대체로 일관성을 보였으나, 불일치 사례도 상당수 확인되었다. 표제부에서는 건축법에 따른 대분류 중심으로 기재되는 반면, 기타 용도에는 세부 용도나 실제 이용 현황이 자유롭게 기재되어 동일 건축물임에도 상이한 정보가 공존하는 경우가 많았다. 층별개요 또한 주 용도와 기타 용도가 일치하지 않는 사례가 나타났으며, 일부는 대분류 코드가 잘못 입력되거나 정의에서 벗어난 텍스트가 기록되기도 했다.

용도 코드와 코드명은 법령 개정 이력과 데이터 입력 관행의 영향을 받아 혼재된 상태로 존재하였다. 예를 들어 '근린생활시설'은 제1종과 제2종으로 세분화되었음에도 여전히 단일 용어로 기재된 데이터가 다수 발견되었고, Z코드나 정의 외 표현('창', '홀' 등)도 일부 남아 있었다. 이러한 요인으로 인해 동일 개념이 여러 방식으로 표기되고 코드와 텍스트 간 불일치가 누적되면서 데이터 품질이 저하되는 현상이 확인되었다.

기타 용도 데이터는 건축물의 실제 활용 현황을 반영한다는 장점이 있으나, 자연어로 기재되어 있고 표준화되지 않아 관리가 어렵다는 문제가 드러났다. 특정 세부 용도나 가구 수, 층 이용 현황 등이 괄호 안에 부기되는 경우가 많아 텍스트 처리 과정에서 불규칙성이 발생하였다. 단어 빈도 분석 결과, '주택', '근린생활시설' 등이 상위권을 차지했으나 이는 주 용도 코드와 반드시 일치하지 않았고, 복합용도 건축물의 경우 다양한 용도가 혼합 기재되는 양상도 관찰되었다.

주 용도와 기타 용도가 불일치하는 경우 그 내용이 오류일 가능성이 높았으며, 교차검증 필요성을 다시 한번 확인하였다. 단독주택임에도 기타 용도에 '공장'이 포함되거나, 근린생활시설임에도 기타 용도에 '주택'이 기재된 사례처럼 복합용도의 범위를 넘어서는 불일치도 다수 발견되었다. 층별개요에서는 주 용도 코드가 대분류로 잘못 기재된 경우가 많았으며, 이 과정에서 기타 용도에 세부 용도가 따로 기록되기도 하였다. 따라서 표제부와 층별개요의 데이터를 종합적으로 비교·검증하는 체계 마련이 필요하다.

기계학습 기반의 예측 실험에서는 나이브 베이즈 모델을 적용하여 비교적 높은 정확도(동별 84.78%, 층별 78.97%)를 달성하였다. 그러나 사례 수가 많은 단독주택이나 제2종 근린생활시설에서는 높은 정확도를 보였으나, 사례가 적거나 명칭이 유사한 제1종 근린생활시설에서는 정확도가 16.59%에 불과했다. 이는 데이터 불균형과 텍스트 빈도 기반 학습의 한계 때문으로, 중복 사례가 많을수록 예측이 성공하는 과적합 현상이 나타났다. 층별 데이터에서는 창고 등 부속 용도가 본래 주 용도와 괴리되는 경우가 많아, 개별 층만이 아니라 건축물 전체 맥락을 고려해야 하는 필요성이 확인되었다. 단순 빈도 기반 기계학습 접근만으로는 이러한 복잡성을 충분히 반영하기 어렵다는 점을 확인하였다. 따라서 향후 데이터 품질 고도화를 위해서는 표제부와 층별개요 간 교차검증 체계를 마련하고, 용어와 코드의 표준화를 추진할 필요가 있다.

3. 건축물대장-인허가 연계 품질 고도화

1) 개요

최근 도입된 건물 단위 연계키인 건물ID는 건축물 데이터를 서로 연계하고 데이터 간 일관성과 추적성을 확보하는 데 필수적인 역할을 할 것으로 기대된다. 본 연구에서는 건물 ID 부여의 현황을 진단하고, 이를 기반으로 연계 데이터의 정합성을 검증하고자 한다. 본 절에서는 건축물대장, 건축인허가, 주택인허가 데이터를 통합하여 분석용 데이터베이스를 구축하고, 건물 ID의 부여 현황과 연계 효율을 정량적으로 평가하였다.

먼저, 건물 ID의 부여율, 고유성, 일관성을 지표로 설정하여 적용 실태를 점검하였다. 이를 통해 행정데이터별 부여 패턴의 차이, 다중 ID 부여 사례, 누락 발생 구간 등을 검토하였다.

다음으로, 건물 ID를 기준으로 한 데이터 연계 교차 검증을 수행하였다. 연계 성공률을 산출하여 건축물대장과 인허가 데이터 간의 매칭 정확도를 평가하고, 연계 결과를 바탕으로 면적 데이터와 용도 데이터의 상호 정합성을 교차 검증하였다. 이를 통해 동일 건물에 대한 주요 속성 정보의 불일치 유형을 식별하고, 연계 기준 및 식별체계의 보완 방향을 도출하였다.

[그림 3-30] 건물ID 기반 데이터 연계 품질 고도화 흐름도
출처: 연구진 작성

2) 건물ID 도입현황 파악 및 검증

■ 건물ID 데이터 검토

건물ID는 건축물대장에 부여되는 고유 식별번호로, 인허가 시 건축행정시스템 세움터에서 건물동 단위로 부여되어 등기부등본과 매칭키로 활용된다. 본 연구에서는 이를 중심으로 건축 관련 데이터를 연계 · 검증하였다. 연구에 활용된 데이터는 건축물대장 기본개요, 건축인허가 동별개요, 주택인허가 동별개요 등이다. 세 데이터 모두에서 건물ID는 '공통_건물_고유_번호'라는 칼럼명으로 제공되며, 건물 단위 연계의 주요 키 필드로 사용된다.

본 연구는 건물ID 기반 연계 가능성 분석을 목적으로 하므로, 건축HUB에서 제공되는 개방 데이터 중 연구 범위에 해당하는 건물을 선별하였다. 분석 단위는 개별 건물동 단위로, 그 중 사용승인 시점이 2024년인 건물로 한정하였다. 건축물대장은 표제부 데이터를 활용하여 '사용승인_일'을 2024년으로 제한하고, '허가번호_구분_코드_명'에서 신축, '주_부속_구분_코드_명'에서 주건축물을 필터링하였다. 그 결과 55,784건이 도출되었다. 건축인허가는 기본개요와 동별개요를 연계하여 동일 기준으로 필터링한 결과, 71,367건이 추출되었다. 주택인허가는 기본개요와 동별개요를 연계하여 최종적으로 6,165건을 확보하였다.

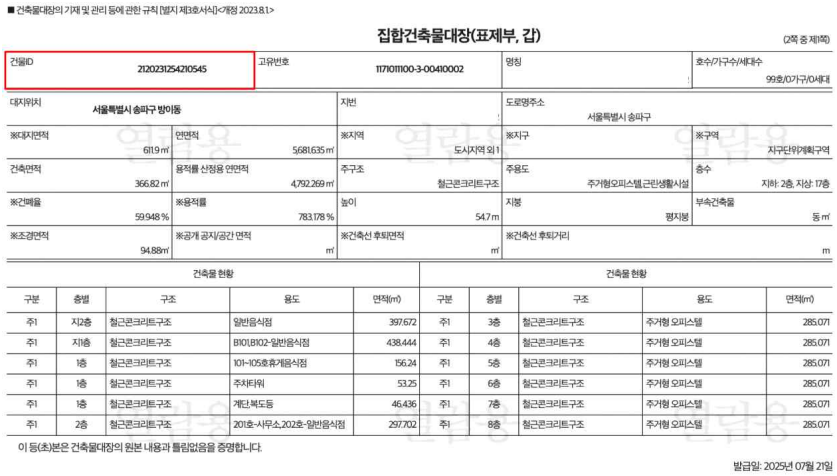
연계 분석을 위해, 건축HUB 개방 데이터와 별도로 제공받은 건물ID 데이터를 상호 결합하였다. 결합 기준은 건축물대장의 경우 건축HUB 데이터의 '관리_건축물대장_PK'와 건물ID 데이터의 '건축물대장일련번호_PK'를 매칭하여 결합하였고, 건축인허가 · 주택인허가는 두 데이터 모두 건축HUB의 '관리_동별_개요_PK'와 건물ID 데이터의 '동별_개요_일련번호'를 매칭하여 연계하였다. 이와 같은 방식으로 서로 다른 출처의 데이터를 건물 단위에서 결합하였다.

■ 건축물대장 검토

- 건물ID의 부여율 검토

본 연구에서는 건물ID의 부여율을, 앞서 정의한 결합 방식에 따라 건축HUB 개방 건축물대장 데이터에 실제로 건물ID가 연계된 비율로 정의하였다. 즉, 분석 대상 총 55,784건의 레코드 중 건물ID가 성공적으로 결합된 건의 비율을 의미한다.

결합 결과, 전체 중 54,123건(97.0%)에서 건물ID가 부여되었으며, 나머지 1,661건(3.0%)은 '건축물대장일련번호_PK'가 일치하지 않아 ID가 부여되지 않았다. 데이터 생성 시점 차이나 정비 지연 등에서 비롯된 것으로 보인다. 실제로, 일부 미결합 건의 '건축물대장일련번호_PK'에 대해 건축물대장을 직접 발급하여 확인한 결과, 건물ID가 존재함을 확인하였다. 이는 제공받은 건물ID 데이터에 아직 반영되지 않았던 경우로, 데이터 제공 시점의 차이나 연계 체계 간 동기화 문제로 인한 것으로 추정된다. 따라서 실제 부여율은 97%보다 더 높을 수 있으며, 일부 누락은 데이터 갱신 주기와 수집 경로에 기인했을 가능성이 있다. 그럼에도 불구하고, 전체의 약 97%에 달하는 높은 부여율은 건물ID를 활용한 데이터 연계 및 분석 기반이 충분히 마련되어 있음을 보여준다.



■ 건축물대장의 기재 및 관리 등에 관한 규칙 [별지 제3호서식] <개정 2023.8.1>

집합건축물대장(표제부, 갑)

(2쪽 중 제1쪽)

건물ID	2120231254210545	고유번호	1171011100-3-00410002	명칭		호수/가구수/세대수	99호/07가구/0세대
대지위치	서울특별시 송파구 방이동	지번		도로명주소	서울특별시 송파구		
※대지면적	611.9 ㎡	연면적	5,681.635 ㎡	※지역	도시지역 외 1	※지구	※구역 자구단위계획구역
건축면적	366.82 ㎡	용적률 산정용 연면적	4,792.269 ㎡	주구조	철근콘크리트구조	주용도	주거형오피스텔,근린생활시설
※건폐율	59.948 %	※용적률	783.178 %	높이	54.7 m	지붕	평지붕
※조경면적	94.88㎡	※공개 공지/공간 면적	㎡	※건축선 후퇴면적	㎡	※건축선 후퇴거리	m

건축물 현황					건축물 현황				
구분	층별	구조	용도	면적(㎡)	구분	층별	구조	용도	면적(㎡)
주1	지2층	철근콘크리트구조	일반음식점	397.672	주1	3층	철근콘크리트구조	주거형 오피스텔	285.071
주1	지1층	철근콘크리트구조	B101,B102-일반음식점	438.444	주1	4층	철근콘크리트구조	주거형 오피스텔	285.071
주1	1층	철근콘크리트구조	101~105호휴게음식점	156.24	주1	5층	철근콘크리트구조	주거형 오피스텔	285.071
주1	1층	철근콘크리트구조	주차타워	53.25	주1	6층	철근콘크리트구조	주거형 오피스텔	285.071
주1	1층	철근콘크리트구조	계단,복도등	46.436	주1	7층	철근콘크리트구조	주거형 오피스텔	285.071
주1	2층	철근콘크리트구조	201호~사무소,202호-일반음식점	297.702	주1	8층	철근콘크리트구조	주거형 오피스텔	285.071

이 등(초)본은 건축물대장의 원본 내용과 틀림없음을 증명합니다.

발급일: 2025년 07월 21일

[그림 3-31] 건물ID 미결합 건축물대장 표제부의 건물ID 표기 현황
출처: 연구진 작성

- 건물ID의 고유성 검토

건물ID의 고유성은 하나의 건물ID가 단일 건물에만 부여되는지, 혹은 다수 건물에 중복 부여되는지를 기준으로 평가하였다. 즉, 앞서 건물ID가 성공적으로 결합된 54,123건을 대상으로, 각 건물ID가 일대일 대응(one-to-one) 관계인지, 일대다 대응(one-to-many) 관계를 가지는지 검토하였다.

분석 결과, 총 54,123건 중 54,103건(99.96%)은 단일 건물과 일대일 대응을 보였다. 반면 7건의 건물ID는 두 개 이상의 건물과 일대다 대응하여, 고유성이 확보되지 않은 사례에 해당하였다. 예를 들어, 건물ID '2020242254085195'는 5개 건물에 동일하게 부여되어 있었으며, 실제 건축물대장 발급 결과 개별 동별 대장에 동일한 ID가 등재된 것으로 확인되었다. 이는 단지 내 여러 건물동에 동일한 건물ID가 잘못 부여된 사례로 해석할 수 있다. 전문가 자문 결과 이 건축물의 경우 인허가 때가 아닌 사용승인 때 건물동이 추가된 경우로, 이러한 경우 (건물ID는 인허가 때 부여되므로) 기존 건물동의 건물

일반건축물대장(갑)

건물ID	2020242254085195	고유번호	4165034029-1-0038000	명칭	제2동

일반건축물대장(갑)

건물ID	2020242254085195	고유번호	4165034029-1-0038000	명칭	제3동

일반건축물대장(갑)

건물ID	2020242254085195	고유번호	4165034029-1-0038000	명칭	제4동

일반건축물대장(갑)

건물ID	2020242254085195	고유번호	4165034029-1-0038000	명칭	제5동

일반건축물대장(갑)

건물ID	2020242254085195	고유번호	4165034029-1-0038000	명칭	제6동

[그림 3-32] 고유성에 위배되어 건물ID가 부여된 건축물대장 표제부 사례
출처: 연구진 작성

ID가 복제되는 오류가 발생한 것으로 추정된다. 따라서 건물ID는 대체로 고유성이 확보되지만, 일부 복수 건물에 동일 ID가 부여된 오류 사례도 일부 존재하며, 이는 향후 데이터 연계 시 건물 단위 식별 정확성에 영향을 미칠 수 있는 요소로 별도 검토가 필요하다.

- 건물ID의 일관성 검토

건물ID의 일관성은, 앞서 살펴본 고유성과는 반대로 하나의 건물에 단일 건물ID만 부여되었는지 여부를 기준으로 평가하였다. 즉, 동일 건물에 복수의 건물ID가 중복 부여된 사례가 있는지를 검토함으로써 체계 내부의 정합성을 확인하였다.

분석 결과, 건물ID가 성공적으로 결합된 54,123건 중 하나의 건물에 복수 건물ID가 부여된 사례는 전혀 없었다. 이는 건물ID가 동일 건물 내에서 중복 없이 일관성을 확보하고 있으며, 시스템 전반에서 식별자로서의 정합성이 유지되고 있음을 보여준다. 즉, 고유성 측면에서는 일부 예외 사례가 존재하지만, 일관성 측면에서는 100% 양호한 관리가 이루어지고 있음을 확인하였다.

[표 3-38] 건축물대장 건물ID 부여 현황 요약

평가 항목	정의	결과 요약	주요 수치/비율
부여율	대상 레코드 중 건물ID가 실제로 부여된 비율	대부분의 건물에 ID 부여	54,123 / 55,784건 (97.0%)
고유성	하나의 건물ID가 여러 건물에 중복 부여되지 않았는지 여부	일부 예외 존재 (복수 건물 동일 ID)	7건의 건물ID가 다건물에 대응
일관성	하나의 건물에 복수의 건물ID가 부여되지 않았는지 여부	모든 건물에 ID 일관 부여	중복 없음 (100% 일관성)

출처: 연구진 작성

■ 건축인허가 검토

- 건물ID의 부여율 검토

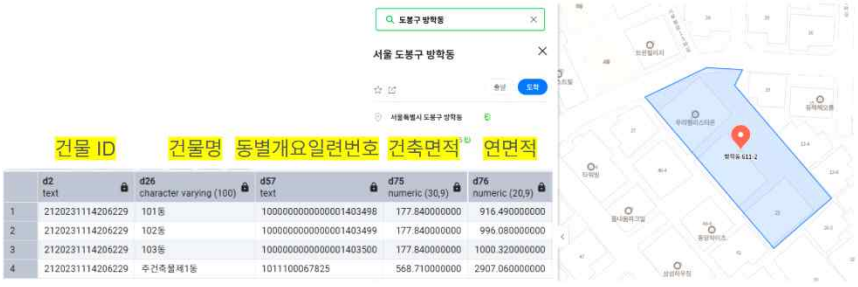
건축인허가 데이터의 건물ID 부여율은 총 71,367건의 분석 대상 중 70,910건(99.4%)에서 건물ID가 성공적으로 부여되어 매우 높은 수준을 보였다. 반면 457건(0.6%)은 부여되지 않은 것으로 나타났다.

미부여 사례는 두 가지 유형으로 구분된다. 첫째, 연계 자체가 불가능한 경우(4건)이다. '건축물대장일련번호_PK' 값이 일치하지 않아 건물ID 데이터와 매칭되지 않는 경우이다. 둘째, 연계는 가능하나 ID 누락(453건): '건축물대장일련번호_PK' 값은 존재하지만, 해당하는 '공통_건물_고유_번호'가 건물ID 데이터에 반영되지 않는 경우이다. 실제 건축물대장을 발급해 확인한 결과, 해당 건물에는 건물ID가 정상적으로 부여되어 있었다. 이러한 사례는 건축물대장에서는 건물ID가 부여되었으나 건축인허가 데이터에는 부여되지 않은 것으로, 인허가 시점이 아닌 사용승인 시점에 건물ID가 부여된 경우로 추정된다.

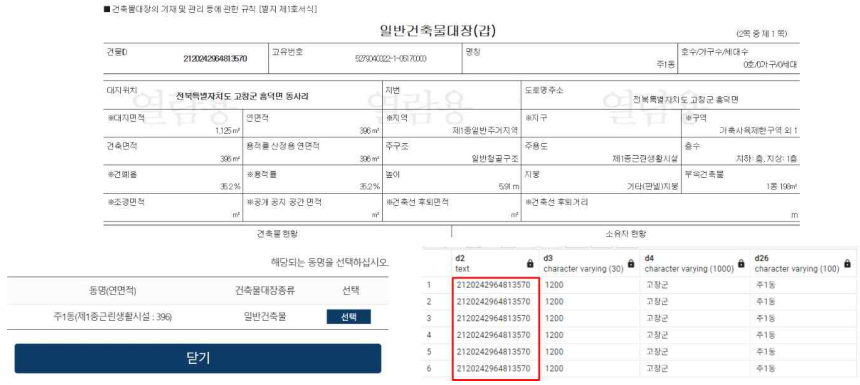
- 건물ID의 고유성 검토

건축인허가 데이터에서 건물ID의 고유성을 분석한 결과, 결합된 70,910건 중 48,060건은 건물ID가 단일 건물과 일대일 대응하였다. 그러나 11,408건은 하나의 건물ID가 다수의 건축인허가 레코드와 대응하여 고유성이 확보되지 않은 사례에 해당하였다.

고유성 미확보 사례는 크게 두 가지 유형으로 구분된다. 총괄표제부로 추정되는 정보가 포함된 경우이다. 한 사례에서는 대지 내 모든 건축물 정보를 총괄표제부에도 개별 건축물과 동일한 건물ID가 부여되어 일대다 대응하는 사례가 발견되었다. 이 사례에서는 대지 내 3개 건물이 존재하나, 종합 정보를 담은 총괄표제부 레코드가 추가되어 총 4개 레코드가 존재하며, 동일 건물ID가 부여되었다. 다른 경우는 동별개요의 기재 내용이 중복 기록된 경우이다. 하나의 건물이 건축물대장상 단일 건물로 등재되어 있음에도, 건축인허가 동별개요에는 여러 개의 레코드로 기재된 사례가 존재하였는데, 대부분 속성 값(대지위치, 연면적 등)이 동일하여 동일 건물을 중복 기록한 것으로 추정된다.



[그림 3-33] 총괄표제부 레코드 포함으로 인한 고유성 위배 사례
출처: 연구진 작성



[그림 3-34] 단일 건물에 다수 동별개요 레코드가 기록된 고유성 위배 사례
출처: 연구진 작성

- 건물ID의 일관성 검토

건축인허가 데이터에서는 건물ID의 일관성에 문제가 없었다. 즉, 동일 건물에 복수의 건물ID가 중복 부여된 사례는 전혀 존재하지 않았다. 총 70,910건의 결합 레코드 중 어느 경우에도 하나의 건물에는 단일 건물ID만 부여되어 있었으며, 이는 건물ID가 식별자로서의 역할을 충실히 수행하고 있음을 의

미한다. 따라서 고유성 측면에서는 일부 구조적 한계가 발견되었지만, 일관성 측면에서는 전면적으로 안정적인 관리가 이루어지고 있음을 확인하였다.

[표 3-39] 건축인허가 데이터에 대한 건물ID 부여 현황 요약

평가 항목	정의	결과 요약	주요 수치/비율
부여율	분석 대상 중 건물ID가 실제로 부여된 비율	대부분 부여됨	70,910 / 71,367건 (99.4%)
고유성	하나의 건물ID가 하나의 건물에만 대응하는지 여부	다수 건물에 동일 ID 부여 사례 존재 (동별개요 중복 등 구조적 요인)	48,060건(일대일), 11,408건(일대다)
일관성	하나의 건물에 복수 건물ID가 부여되지 않았는지 여부	전면적으로 양호, 중복 없음	100% 일관성 유지

출처: 연구진 작성

■ 주택인허가 검토

- 건물ID의 부여율 검토

주택인허가 데이터의 건물ID 부여율은 총 6,165건 중 6,125건(99.4%)에서 건물ID가 성공적으로 부여되어 매우 높은 수준을 보였다. 반면 40건(0.6%)은 부여되지 않은 것으로 나타났다.

이 미부여 사례는 대부분 특정 아파트 단지에서 발생하였다. 이들 건은 '건축물대장일련번호_PK' 값이 존재하여 건축물대장과의 연계는 가능했으나, 해당하는 '공통_건물_고유_번호' 값이 건물ID 데이터에서 누락되어 있었다. 전문가 자문 결과, 기존에는 건축물대장 생성 때 건물ID가 부여되었으나, 2023년 초 클라우드세움터 개시 이후에 인허가 때 건물ID를 부여하는 것으로 부여 시점이 변경된 바 있다. 주택인허가의 경우 사용검사 전 임시 사용승인 단계이거나, 정비사업의 경우 사용검사 이후 소유권 이전고시 전 단계인 경우 사용승인은 마쳤으나 건축물대장이 아직 생성되지 않은 상태일 수 있다. 따라서 2022년 말 이전 인허가된 건축물의 경우 인허가 당시 건물ID 부여가 이루어지지 않고, 사용승인일이 2024년으로 기재된 경우에도 건축물대장 생성 및 건물ID 부여가 아직 이루어지지 않은 상태일 수 있다. 이러한 점을 고려할 때, 주택인허가 데이터의 건물ID 부여율은 2023년 이후 인허가 건축물에서는 매우 높은 수준을 달성할 것으로 예상된다.

- 건물ID의 고유성 검토

고유성 분석 결과, 성공적으로 결합된 6,125건 중 5,717건(93.3%)은 건물ID가 단일 건물 레코드와 일대일 대응하였다. 그러나 178건(2.9%)은 하나의 건물ID가 여러 개의 주택인허가 레코드와 일대다 대응하여 고유성이 확보되지 않았다.

이러한 사례는 건축인허가 데이터에서 확인된 고유성 위반과 동일한 구조를 보인다. 즉, 실제로는 하나의 건물이지만, 주택인허가 데이터에서는 여러 개의 '관리_동별_개요_PK' 값으로 분리 기록된 경우이다. 이로 인해 동일 건물에 하나의 ID가 부여되어 있음에도 데이터 상에서는 중복된 레코드가 생성되며, 결과적으로는 하나의 건물ID가 다수의 레코드와 연결되는 문제가 발생한다. 이는 건물ID 자

체의 오류라기보다는, 주택인허가 데이터 작성 단위 및 입력 관행의 특성에서 비롯된 현상으로 해석된다.

- 건물ID의 일관성 검토

주택인허가 데이터에서는 건물ID의 일관성에 문제가 없었다. 즉, 하나의 건물에 복수의 건물ID가 중복 부여된 사례는 단 한 건도 존재하지 않았다. 분석 대상인 6,125건의 결합 레코드에서 모든 건물은 단일 건물ID와 대응하였으며, 이는 건물ID 체계가 유일 식별자로서 기능하고 있음을 보여준다.

따라서 고유성 측면에서는 일부 중복 문제가 나타났으나, 일관성 측면에서는 건축물대장 및 건축인허가 데이터와 마찬가지로 매우 안정적이었다.

[표 3-40] 주택인허가 데이터 건물ID 부여 현황 요약

평가 항목	정의	결과 요약	주요 수치/비율
부여율	분석 대상 중 건물ID가 실제로 부여된 비율	대부분 부여됨	6,125 / 6,165건 (99.4%)
고유성	하나의 건물ID가 하나의 건물에만 대응하는지 여부	일부 위반 사례 존재 (중복 입력 등)	5,717건(일대일), 178건(일대다)
일관성	하나의 건물에 복수 건물ID가 부여되지 않았는지 여부	중복 없음, 100% 일관성 유지	6,125건 모두 일관

출처: 연구진 작성

3) 건물ID 기반 연계 데이터의 교차검증

■ 건물ID 기반 연계 분석

- 주소 및 건물명 기반 연계

건물ID가 도입되기 이전의 전통적 연계 방식과 비교하여 그 효용을 검증하기 위해, 먼저 '대지_위치'와 '건물_명'을 기준으로 건축물대장과 건축인허가 데이터를 연계하였다. 이는 건물명·주소 등 고유성을 일부 반영하는 항목을 활용하여 건물ID 없이도 연계가 가능한지, 그 한계를 평가하기 위함이다. 분석 대상은 건축HUB 건축물대장 표제부(55,784건), 건축인허가 동별개요(71,367건)이다. 기재 현황은 두 데이터 모두 '대지_위치'는 전부 존재하였으나, '건물_명'은 건축물대장에서 9,174건(16.4%), 건축인허가에서 54,940건(77.0%)만 존재하여 누락 비중이 매우 컸다. 연계 결과 주소와 건물명을 활용해 조인한 결과 총 9,599건의 연계되었으며, 이는 전체 건축인허가 레코드(71,367건) 대비 약 13.5%의 성공률이다. 이 결과는 건물명 정보가 불완전한 상황에, 주소·건물명만으로는 연계 성공률이 낮아 고유 식별자인 건물ID가 필수적임을 시사한다.

다음으로 건축물대장과 주택인허가 데이터 간 연계 가능성을 검토하였다. 주택인허가 데이터에는 건축인허가와 달리 '동명' 칼럼이 존재하므로 이를 함께 고려하였다. 분석 대상은 건축물대장 표제부(55,784건), 주택인허가 동별개요(6,165건)이다. 기재 현황은 '대지_위치'는 전부 기재되어 있었으며, '건물_명'은 5,533건(89.7%), '동명'은 6,152건(99.8%)에 기록되어 있었다. 다만 '동명'은 공동주

택의 동 번호(예: 101동, 102동)와 같은 내부 단지 구분 값이 많아 건물 식별자로 활용에 한계가 있었다. 연계 결과 '대지_위치' + '건물_명'은 2,568건 연계, '대지_위치' + '동명'은 13건 연계, 두 방식을 통합 시 총 2,573건 연계(전체 6,165건 대비 약 41.7% 성공률) 결과를 얻었다. 이는 건축인허가의 연계 성공률(13.5%)보다는 높지만, 여전히 절반 이상 연계되지 않는다는 점에서 건물명 기반 연계의 구조적 한계를 보여준다. 특히 공동주택처럼 건물명이 모호하거나 누락되기 쉬운 유형에서는 건물ID와 같은 고유 식별자 없이는 정밀한 연계가 어렵다는 점이 확인되었다.

[표 3-41] 주소 및 건물명 기반 연계 결과 요약

비교 항목	건축물대장 ↔ 건축인허가	건축물대장 ↔ 주택인허가
분석 대상 수 (오른쪽 기준)	71,367건	6,165건
'대지_위치' 기재율	100%	100%
'건물_명' 기재율	77.0%	89.7%
'동명' 기재율	없음	99.8%
연계 방식 ①: '대지_위치' + '건물_명'	9,599건	2,568건
연계 방식 ②: '대지_위치' + '동명'	해당 없음	13건
통합 연계 건수	9,599건	2,573건
연계 성공률	약 13.5%	약 41.7%
주요 한계점	건물명 누락 및 불일치 다수	'동명'은 식별자로 부적합함

출처: 연구진 작성

- 건물ID 기반 연계

주소 및 건물명 기반 연계와 비교하여 건물ID 도입의 효용을 검증하기 위해, 건물ID를 기준으로 건축물대장과 건축인허가 데이터 간 연계 성공률을 측정하였다. 분석 대상은 건물ID가 부여된 건축인허가 70,910건이었으며, 건축물대장 데이터는 사용승인일을 기준으로 필터링하지 않고 전체 데이터(7,967,416건)를 기준으로 연계하였다.

그 결과, 건축인허가와 연결되는 건수는 66,343건(93.6%)으로 나타났다. 즉, 필터링 과정에서 약 3.9%가 누락된 것이며, 이는 사용승인 연도 · 신축 여부 · 주건축물 여부 등 조건이 일부 불일치했음을 의미한다. 또한, 남은 6.4%는 전체 건축물대장 목록에서도 매칭되지 않았는데, 이는 사용승인일 기재 이후에도 건축물대장이 생성되지 않은 상태로 해석된다. 한 예로, 서울특별시 동대문구 용두동 모 건축물의 경우 건축인허가 동별개요에는 총 23건이 존재하나, 건축물대장에는 해당 건물이 아직 등재되지 않아 연계되지 않았다.

주택인허가 데이터에 대해서도 동일한 방식으로 검토하였다. 분석 대상은 건물ID가 부여된 주택인허가 6,125건이었다. 연계 결과 2,136건(34.9%)이 매칭되었다. 65.1%의 건축물이 전체 건축물대장 대상으로도 매칭되지 않았는데, 이는 주택인허가 데이터에서 특히 사용승인일 기재 이후 건축물대장 생성이 이루어지지 않은 비율이 높음을 시사한다. 앞서 논의한 바와 같이 임시 사용승인을 받은 경우나 정비사업에서 사용검사 이후 소유권 이전고시 절차가 진행 중인 경우 등이 주택법에 따른 인허가에서

더 빈번함을 시사한다. 한 예로 서울특별시 강동구 성내동 건축물 사례에서, 주택인허가 동별개요에는 총 6건이 등록되어 있으나, 건축물대장에는 해당 정보가 존재하지 않았다.

[표 3-42] 건물ID 기반 연계 결과 요약

비교 항목	건축물대장 ↔ 건축인허가	건축물대장 ↔ 주택인허가
분석 대상 (ID 부여됨)	70,910건	6,125건
연계 성공 건수(성공률)	66,363건 (93.6%)	2,136건 (34.9%)

출처: 연구진 작성

■ 연계데이터 정합성 검증

- 개요

앞서 건물ID를 기준으로 연계한 결과, 건축물대장 전체 데이터와 건축인허가 데이터는 66,343건(93.6%), 건축물대장과 주택인허가 데이터는 2,136건(34.9%)이 결합되는 것으로 확인되었다. 그러나 단순히 건물ID를 매개로 연계가 이루어졌다고 해서, 해당 레코드들이 실제로 동일 건물임을 보장하지는 않는다. 동일한 건물ID를 공유하는 레코드라 하더라도, 건축면적 · 용도 등 핵심 속성값이 일치하지 않을 수 있기 때문이다. 따라서 건물ID 기반 연계 체계의 신뢰성과 유효성을 확보하기 위해서는, 연계된 데이터 간 속성값 정합성을 교차검증할 필요가 있다.

앞서 확인한 바와 같이, 현재 데이터베이스에는 건물ID 고유성 문제가 일부 존재한다. 건축인허가 데이터에서는 11,408건이 일대다 대응 관계를 보였으며, 주택인허가 데이터에서는 178건이 일대다 대응 관계를 보였다. 이러한 대응 구조는 서로 다른 데이터셋 간 속성 비교의 정확성을 저해하여, 직접적인 정합성 검증을 어렵게 만든다. 따라서 교차검증을 수행하기 위해서는, 중복 레코드를 사전에 처리하여 건물ID-레코드 간 일대일 대응 관계를 확보하는 전처리 과정이 필수적이다.

이를 위해 우선 중복 유형을 분류한 뒤, 유형별 처리 방안을 적용하여 비교 가능한 데이터셋을 구축하였다. 이후 모든 데이터셋에서 공통적으로 존재하는 속성을 통해 데이터 간 정합성을 검증하고자 하였다. 교차검증은 건축물대장, 건축인허가, 주택인허가 데이터에 공통적으로 존재하는 속성을 기준으로 수행하였다. 주요 속성은 위치정보, 구조 · 위계정보, 면적정보, 용도정보 등이며, 특히 면적과 용도를 핵심 비교 지표로 활용하였다. 비교 단위는 필터링된 건축 · 주택 인허가 동별개요 자료로, 대조 기준은 건축물대장 표제부 전체 약 800만 건이다. 이를 통해 건물ID 기반 연계의 정확성과 속성 정합성을 실증적으로 검증하고자 하였다.

- 중복 레코드 처리

건물ID 기반 교차검증을 수행하기 위해서는, 동일한 건물ID를 공유하는 중복 레코드 처리가 선행되어야 한다. 본 연구에서는 이러한 중복 레코드를 다음 두 가지 유형으로 분류하였다. 유형 1은 내부 속성 완전 일치형이다. 동일한 건물ID를 가진 레코드들 내부 속성(건물명, 주소, 면적 등)이 모두 동일한 경우이다. 사실상 동일한 레코드가 중복 입력된 사례로 판단하였다. 유형 2는 내부 속성 불일치형이다. 동일한 건물ID를 가진 레코드들이 서로 다른 속성값을 포함하는 경우이다. 동일 건물임에도 불구

[표 3-43] 각 데이터셋 공통 컬럼 항목

변수명	데이터 타입	카테고리
대지위치	VARCHAR(500)	위치정보
시군구코드	VARCHAR(5)	위치정보
법정동코드	VARCHAR(5)	위치정보
대지구분코드	VARCHAR(1)	위치정보
번	VARCHAR(4)	위치정보
지	VARCHAR(4)	위치정보
블록	VARCHAR(20)	위치정보
로트	VARCHAR(20)	위치정보
주부속구분코드	VARCHAR(1)	위계정보
구조코드	VARCHAR(2)	구조정보
생성일자	VARCHAR(8)	기타
건물명	VARCHAR(100)	기타
특수지명	VARCHAR(200)	기타
건축면적(m^2)	NUMERIC(19,9)	면적정보
연면적(m^2)	NUMERIC(19,9)	면적정보
용적률 산정 연면적(m^2)	NUMERIC(19,9)	면적정보
주용도코드	VARCHAR(5)	용도정보

출처: 연구진 작성

하고 속성 정보가 상이하게 기록된 사례로, 데이터 정합성에 직접적 영향을 미친다. 이러한 중복 현상은 건물ID가 동일함에도 불구하고 하나의 ID에 여러 데이터 행이 매칭되어, 데이터 간 일대일 대응을 불가능하게 만드는 요인으로 작용한다. 따라서 교차검증에 앞서 적절한 처리 절차가 필요하다.

유형 1의 처리는 동일한 건물ID를 가진 레코드들 가운데 날짜 · PK 정보를 제외한 모든 속성이 일치하는 경우 하나의 레코드만 남기고 나머지는 제거하였다.이는 '생성일자'와 같이 데이터 갱신 과정에서 발생하는 단순 중복을 제거하기 위함이다.

유형 2의 처리는 동일한 건물ID를 가진 레코드들이 상이한 속성값을 가질 경우, 속성별 비교 목적에 따라 차등적으로 처리하였다. 먼저 면적 정보를 비교하여 중복 레코드 간 면적값이 동일하면 하나로 압축, 상이하면 해당 건물ID 전체를 비교 대상에서 제외하였다. 다음으로 용도 정보를 비교하여 면적과 동일한 방식을 적용하여 처리하였다. 이와 같은 전처리를 통해 중복 레코드로 인한 왜곡을 최소화하고 건물ID 기반 연계 검증에 필요한 일대일 대응 데이터셋을 확보하였다.

유형 1(속성 완전 일치형)의 중복 처리를 수행한 결과, 건축인허가는 전체 70,910건 중 16,173건에서 PK와 날짜를 제외한 모든 속성이 동일하게 나타났으며, 중복 그룹당 하나의 행만 남기는 방식으로 8,074건으로 압축하였다. 최종적으로 62,811건의 레코드가 확보되었다. 주택인허가는 전체 6,125건 중 363건에서 동일 유형의 중복이 확인되었으며, 이를 157건으로 압축하여 최종 5,919건의 레코드가 남았다. 건축물대장은 전체 7,967,416건 중 약 3.3만 건에서 속성값이 모두 동일한 현상이 발견되었으나, 건물ID 값 자체는 중복되지 않아 일대일 매칭에는 영향을 미치지 않았다.

[표 3-44] 중복 유형 1 처리 결과

건축물대장		건축인허가		주택인허가	
처리 이전	처리 이후	처리 이전	처리 이후	처리 이전	처리 이후
7,967,416건	이전과 동일	70,910건	62,811건	6,125건	5,919건

출처: 연구진 작성

유형 1 데이터를 제외한 후, 유형 2(속성 불일치형)에 대한 분석을 수행하였다. 건축인허가 데이터(62,811건 중 6,681건) 중 면적 정보가 동일한 중복(568건)은 284건으로 압축되어 최종 56,414건의 면적 비교용 테이블이 구축되었다. 또한 용도 정보가 동일한 중복 6,557건을 3,276건으로 압축하여, 최종 59,406건의 용도 비교용 테이블을 구성하였다. 주택인허가(5,919건 중 48건)는 면적 정보가 동일한 중복(42건)은 21건으로 압축되어 최종 5,892건의 면적 비교용 테이블이 구축되었다. 용도 정보가 동일한 중복(32건)은 16건으로 압축되어 최족 5,887건의 용도 비교용 테이블이 마련되었다. 건축물대장(7,967,416건 중 87건)은 면적 정보가 동일한 중복(29건)은 7건으로 압축되어 최종 7,967,336건의 면적 비교용 테이블이 구축되었다. 용도 정보가 동일한 중복(74건)은 22건으로 압축되어 최종 7,967,351건의 용도 비교용 테이블이 마련되었다. 중복 레코드 처리 과정을 통해, 모든 데이터셋에서 하나의 건물ID가 하나의 데이터 행과만 대응하는 일대일 연계 가능 데이터셋이 마련되었다.

[표 3-45] 건축인허가 중복 유형 2 현황

모두 일치	면적그룹만 일치	용도그룹만 일치	모두 불일치	총계
498건	70건	6,059건	54건	6,681건

출처: 연구진 작성

[표 3-46] 주택인허가 중복 유형 2 현황

모두 일치	면적그룹만 일치	용도그룹만 일치	모두 불일치	총계
28건	14건	4건	2건	48건

출처: 연구진 작성

[표 3-47] 건축물대장 중복 유형 2 현황

모두 일치	면적그룹만 일치	용도그룹만 일치	모두 불일치	총계
29건	0건	45건	13건	87건

출처: 연구진 작성

[표 3-48] 최종 일대일 연계 데이터셋 구축 결과

건축물대장		건축인허가		주택인허가	
처리 이전	처리 이후 (용도/면적)	처리 이전	처리 이후 (용도/면적)	처리 이전	처리 이후 (용도/면적)
7,967,416건	7,967,336건 (-80건)	70,910건	56,414건 (-14,496건)	6,125건	5,892건 (-233건)
	7,967,351건 (-65건)		59,406건 (-11,504건)		5,887건 (-238건)

출처: 연구진 작성

■ 면적 정보 정합성 검증

건물ID를 기준으로 연계된 데이터의 정합성을 검증하기 위해, 건축면적, 연면적, 용적률 산정 연면적의 세 가지 변수를 비교 대상으로 선정하였다. 이들 변수는 건축물의 물리적 규모를 나타내는 불변 속성으로서, 동일 건물에 대해서는 데이터 간 일치가 보장되어야 한다. 또한, 세 데이터셋이 공유하는 공통 속성 중 유일한 수치형 변수로 정량적 비교가 가능하며, 오차 범위 설정을 통한 실질적 일치 여부 판단에도 적합하다. 정확한 비교를 위해, 해당 면적 변수들과 건물ID만을 포함하는 비교 전용 테이블을 새로 생성하고, 불일치 데이터가 발견될 경우 원본 테이블로 추적하는 방식을 채택하였다.

면적 변수 검토 결과, 일부 값은 단순히 누락되지 않고 '0' 또는 음수값으로 기입되어 있었다. 이는 실제 면적이 아닌 미입력값 혹은 오류값으로 판단되어 모두 결측으로 처리하였다. 결측률은 데이터셋별로 상이하였으나, 이는 비교 대상에서의 자동 제외를 의미하지 않는다. 예를 들어, 건축물대장에서 결측이지만 인허가 데이터에는 값이 존재하는 경우, 해당 정보를 활용하여 결측을 보완할 수 있다. 반대로 두 데이터 모두에서 결측인 경우에는 내부 보완이 불가능하다.

[표 3-49] 면적 변수 결측 현황

테이블명	총 행수	컬럼명	NULL(건)	0(건)	음수(건)	총 결측	결측률(%)
건축물대장 (면적자료)	7,967,336건	건축면적	0	454,329	10	454,339	5.70
		연면적	0	60,821	10	60,831	0.76
		용적률산정연면적	0	439,418	11	439,429	5.52
건축인허가 (면적자료)	56,414건	건축면적	0	137	3	140	0.25
		연면적	0	13	3	16	0.03
		용적률산정연면적	0	153	3	156	0.28
주택인허가 (면적자료)	5,892건	건축면적	0	528	2	530	9.00
		연면적	0	783	2	785	13.32
		용적률산정연면적	0	716	2	718	12.19

출처: 연구진 작성

위와 같은 결측 현황을 반영하여, 건물ID 단위의 면적 정보 비교 시 다음 네 가지 유형을 정의하였다. 첫째 일치 유형은 두 데이터베이스 모두 동일한 면적값을 가지는 경우로, 정합성이 확보된 상태이므로 별도 조치가 불필요하다. 둘째 불일치 유형은 두 데이터베이스 모두에 값이 존재하나 서로 다른 수치가 기록된 경우이다. 이는 입력 오류, 측정 기준 차이, 시점 차이 등의 가능성이 있어 추가 검증이 필요하다. 셋째 보완 가능 유형은 한쪽 데이터는 결측이나 다른 한쪽에는 정상 값이 존재하는 경우로, 비결측 데이터의 값으로 보완이 가능하여 데이터 상호 보완 효과를 기대할 수 있다. 넷째 완전 결측 유형은 두 데이터베이스 모두 결측값인 경우로, 내부 보완이 불가능하여 현장 실측 또는 제3의 데이터 소스가 필요하다.

건축물대장과 건축인허가 연계 비교를 수행하였다. 공통 건물ID를 기준으로 비교 가능한 건수는 총 52,469건으로, 이는 필터링된 건축인허가 데이터(56,414건) 중 93.01%에 해당한다. 이 데이터를 대상으로 앞서 정의한 네 가지 유형(일치, 불일치, 보완가능, 완전결측)에 따라 분포를 분석하였다. 일치율은 세 변수 모두 90% 내외의 안정적인 비율을 보여 데이터 간 정합성이 전반적으로 양호한 것으로 확인되었다. 불일치율은 8~10% 수준으로 존재하며, “서울특별시 강서구 개화동” 건물의 경우 건축면적이 건축물대장에서는 100.39㎡, 건축인허가에서는 107.96㎡로 상이하게 기록되어 있다. 보완가능한 건은 0.1~0.2% 수준의 극히 일부에 불과하며, “충청북도 충주시 호암동” 건물은 인허가 데이터에는 누락되었으나 건축물대장에 증축 이력으로 기록된 값과 일치하는 사례를 보였다. 완전결측은 0.1~0.2% 수준의 극히 일부에 불과하며, “경기도 남양주시 조안면 시우리” 건물은 두 데이터 모두에서 건축면적과 용적률 산정 연면적이 결측된 상태로 나타났다.

[표 3-50] 불일치 유형 현황(건축물대장 ↔ 건축인허가)

변수명	불일치건수	평균차이(m^2)	평균차이율(%)	최대차이(m^2)	최대차이율(%)
건축면적	4,394건	110.40	29.99	49346.55	9900
연면적	5,122건	138.24	27.02	49561.52	3045
용적률산정연면적	4,666건	135.50	27.26	27648.89	3045

출처: 연구진 작성

불일치 유형의 평균 차이율은 27~30% 수준으로 상당한 격차를 보였으며, 건축면적은 평균 110.40㎡, 연면적은 138.24㎡, 용적률 산정 연면적은 135.50㎡의 차이를 나타났다. 극단적인 경우 건축면적에서 49,346.55㎡, 9,900%까지의 차이가 발생하기도 했다. 불일치 데이터를 검토한 결과, 주된 원인은 두 데이터베이스 간 등재 시점 차이로 인해 한쪽에는 변경 이전, 다른 쪽에는 변경 이후 정보가 기재되어 발생하거나 측정 기준의 차이로 인해 불일치 현상이 발생하는 것으로 확인되었다. 이러한 불일치가 실제 오류인지 아니면 시점 차이에 따른 정상적인 변화인지를 구별하기 위해서는 허용 오차 범위 설정을 통한 추가 분석이 필요하다.

불일치 사례라도 실제로는 동일한 건축물일 가능성을 고려해, 두 값의 평균 ±2.5% 범위를 허용 오차로 설정하였다. 허용오차 이내 데이터는 약 10% 수준을 차지했으며, 평균 차이율은 0.9% 내외로 매우 낮게 나타났다. 건축면적 11.22㎡, 연면적 16.53㎡, 용적률 산정 연면적 17.77㎡의 미세한 차이는 측정 방식이나 반올림 처리 등의 기술적 차이로 해석할 수 있다. 반면 허용오차를 초과하는 데이터는 약 90%를 차지하며, 평균 차이율이 35% 내외로 상당한 격차를 보였다. 최대 차이율은 190% 수준까지 나타나 극단적인 불일치 사례가 다수 존재함을 확인했다. 특히 불일치 건물 4,854건 중 3,734건(76.9%)이 세 변수 모두에서 허용오차를 초과하는 동반 오류 현상을 보였다.

건축물대장과 건축인허가 간 면적 정보의 정합성은 90% 내외 수준을 보여 전반적으로 양호한 편이나, 발생하는 불일치는 단순한 입력 오류보다는 데이터 등재 시점의 차이나 제도적 차이에서 기인하는 것으로 분석된다. 허용 오차 ±2.5% 기준을 적용할 경우 불일치 건 중 약 10%는 실질적으로 동일한 값으로 간주할 수 있어, 향후 데이터 품질 평가 시 이러한 허용 범위를 고려한 보정 기준 마련이 필요하다. 다만 불일치 사례의 나머지 90%가 여전히 허용오차를 초과하고 있어, 건축행정 데이터의 관리 및 갱신 체계에 대한 근본적인 개선이 요구되는 상황이다.

[표 3-51] 불일치 유형의 허용오차 내 · 외 구분(건축물대장 ↔ 건축인허가)

변수명	분류(±2.5%)	건수	비율(%)	평균차이(m^2)	평균차이율(%)	최대차이율(%)
건축면적	허용오차 이내	446	10.15	11.22	0.89	2.50
	허용오차 초과	3,948	89.85	121.60	36.24	196.04
연면적	허용오차 이내	477	9.31	16.53	0.83	2.48
	허용오차 초과	4,645	90.69	150.74	35.29	187.67
용적률 산정연면적	허용오차 이내	454	9.73	17.77	0.84	2.50
	허용오차 초과	4,212	90.27	148.19	35.67	190.28

출처: 연구진 작성

다음으로 건축물대장과 주택인허가 연계 비교를 수행하였다. 공통 건물ID를 가진 데이터 중 비교 가능한 건수는 1,946건으로, 이는 필터링된 주택인허가 데이터(5,892건) 중 33.03%에 해당한다. 동일한 방식으로 네 가지 유형(일치, 불일치, 보완가능, 완전결측)에 따라 검토하였다. 일치율은 약 70% 수준으로 나타나 건축인허가와의 비교(90% 수준)보다 상대적으로 낮은 정합성을 보였다. 불일치율은 2% 내외로 건축인허가(8~10%)에 비해 낮은 수준이었으나, “부산광역시 부산진구 개금동” 경우 건축물대장에는 건축면적 313.26㎡, 용적률 산정 연면적 4,242.56㎡로 기록되어 있으나 주택인허가에는 각각 10.81㎡, 125.85㎡로 기재되어 상당한 차이를 보였다. 보완가능은 약 25%로 건축인허가(0.1% 미만)에 비해 현저히 높은 비율을 나타냈으며, “서울특별시 중구 신당동”은 건축물대장에는 상가동의

면적 정보가 존재하나 주택인허가에서는 동일 건물의 데이터 행이 3개 존재함에도 모두 결측으로 나타나는 사례를 보였다. 완전결측은 건축면적과 용적률 산정 연면적에서 4% 내외 수준을 보였으며, "경상남도 진주시 가좌동"은 연면적과 용적률 산정 연면적은 일치하나 건축면적은 두 데이터 모두에서 결측된 상태로 확인되었다.

[표 3-52] 건축물대장 ↔ 주택인허가 면적 변수별 유형 분포

유형	건축면적		연면적		용적률산정연면적	
	건수	비율	건수	비율	건수	비율
일치	1,331	68.40%	1,401	71.99%	1,344	69.06%
불일치	28	1.44%	35	1.80%	33	1.70%
보완가능	502	25.80%	493	25.33%	478	24.56%
완전결측	85	4.37%	17	0.87%	91	4.68%

출처: 연구진 작성

불일치 유형의 평균 차이율은 59~72% 수준으로 건축인허가 연계(27~30%)보다 2배 이상 높은 격차를 보였으며, 건축면적은 평균 571.43㎡, 연면적은 3,858.86㎡, 용적률 산정 연면적은 3,855.83㎡의 차이를 나타났다. 극단적인 경우 모든 변수에서 99%에 달하는 차이가 발생하기도 했다. 불일치 데이터를 검토한 결과, 주된 원인은 개별 동 건물과 집합건축물 전체 면적의 혼재, 혹은 일부 층이나 동의 면적만 입력된 오류에서 기인하는 것으로 확인되었다. 이러한 불일치가 실제 오류인지 아니면 데이터 입력 범위 차이에 따른 것인지를 구별하기 위해서 건축인허가 연계와 마찬가지의 허용 오차 범위를 설정하여 추가 분석하였다.

[표 3-53] 불일치 유형 현황(건축물대장 ↔ 주택인허가)

변수명	불일치건수	평균차이(m^2)	평균차이율(%)	최대차이(m^2)	최대차이율(%)
건축면적	28건	571.43	71.94	2334.53	98.97
연면적	35건	3858.86	71.22	17192.13	99.93
용적률산정연면적	33건	3855.83	59.13	12152.98	99.92

출처: 연구진 작성

허용오차 이내 데이터는 건축면적 1건, 연면적 7건, 용적률 산정 연면적 3건, 그 평균 차이율은 0.1% 내외로 매우 낮게 나타났다. 이러한 미세한 차이는 데이터 갱신이나 반올림 처리 등의 원인으로 해석할 수 있다. 반면 허용오차를 초과하는 데이터는 변수별 80~97%를 차지하며, 평균 차이율은 114~162%, 최대 차이율은 200% 근접 수준까지 나타났다. 불일치 건물 중 15건(34.9%)은 세 변수 모두에서 허용오차를 초과하는 동반 오류 현상을 보였다.

[표 3-54] 불일치 유형의 허용오차 내 · 외 구분(건축물대장 ↔ 주택인허가)

변수명	분류(±2.5%)	건수	비율(%)	평균차이(m^2)	평균차이율(%)	최대차이율(%)
건축면적	허용오차 이내	1	3.57	0.01	0.01	0.01
	허용오차 초과	27	96.43	592.59	129.95	195.92
연면적	허용오차 이내	7	20.00	7.61	0.82	1.87
	허용오차 초과	28	80.00	4821.68	162.02	199.72
용적률 산정연면적	허용오차 이내	3	9.09	3.50	0.08	0.22
	허용오차 초과	30	90.91	4241.07	114.24	199.70

출처: 연구진 작성

건축물대장과 주택인허가 간 면적 정보의 정합성은 70% 수준으로 건축인허가 연계(90%)보다 상당히 낮은 편이나, 보완가능 비율이 25%로 높아 데이터 간 상호 보완 효과를 기대할 수 있다. 또한, 허용 오차 ±2.5% 기준을 적용하더라도 불일치 건의 대부분(80% 이상)이 여전히 허용오차를 초과하고 있어, 데이터 간 기록값의 차이를 넘어서는 내재적 문제가 존재함을 알 수있다. 따라서 주택인허가는 결측이나 이상값의 처리 등 데이터의 품질 관리 측면에서 개선이 우선적으로 필요한 상황이다.

면적 정보 정합성 검증 결과를 종합하면, 각 데이터베이스 간 면적 변수의 연계 상태는 다음 세 가지 최종 유형으로 분류할 수 있다. “정합성 높음” 유형은 두 데이터베이스의 면적 변수가 실질적으로 일치하는 경우이다. 실제 값이 완전히 일치하거나 평균값 대비 ±2.5%의 허용오차 범위 내에 두 값이 모두 위치하는 경우가 이에 해당한다. 즉, 허용오차 내에서 미세한 차이가 존재하는 경우가 포함되나, 이는 데이터 구득 시점의 차이, 미세한 기입 오류 등에서 기인한 것으로 판단되어 실무적으로 동일한 값으로 간주할 수 있는 수준이다.

“연계 보정 필요” 유형은 두 데이터베이스의 면적 변수에 불일치가 존재하나 상호 보완을 통해 처리 가능한 경우이다. 허용오차 2.5%를 초과하는 불일치 데이터 행과 하나의 데이터베이스에만 값이 존재하는 보완가능 데이터 행이 모두 포함된다. 이들은 데이터 간 교차검증이나 우선순위 설정을 통해 정확한 값의 정정 가능성이 있는 것으로 파악되었다.

“연계 불가 유형”은 두 데이터베이스 모두에서 면적 정보가 결측되어 데이터 내에서는 처리가 불가능한 경우이다. 이러한 완전결측의 경우 현장 실측이나 제3의 데이터 소스를 통해서만 면적 정보를 획득할 수 있는 상태를 의미한다.

이러한 분류 체계에 따라 건축물대장과 건축인허가, 건축물대장과 주택인허가 연계의 면적 변수의 정합성을 최종 정리한 결과는, 건축물대장-건축인허가 연계에서는 세 면적 변수 모두 90% 이상의 높은 정합성을 보였다. 건축면적은 48,404건(92.25%), 연면적은 47,818건(91.14%), 용적률 산정 연면적

은 48,105건(91.70%)이 정합성 높음으로 분류되었다. 연계 보정이 필요한 케이스는 전체의 약 8% 수준으로 나타났으며, 연계 불가 케이스는 0.2% 이하로 극히 제한적인 것으로 확인되었다. 건축물대장-주택인허가 연계에서는 건축인허가 대비 상대적으로 낮은 정합성을 보였다. 정합성 높음으로 분류된 케이스는 건축면적 1,332건(68.45%), 연면적 1,408건(72.35%), 용적률 산정 연면적 1,347건(69.22%)으로 약 70% 수준에 머물렀다. 연계 보정이 필요한 케이스는 전체의 약 27%로 건축인허가보다 3배 이상 높게 나타났으며, 연계 불가 케이스 또한 건축면적 4.37%, 용적률 산정 연면적 4.68%로 건축인허가보다 높은 비율을 차지하였다.

따라서 건축인허가 데이터는 비교적 안정적인 정합성을 확보한 반면, 주택인허가 데이터는 보완 및 개선이 시급한 영역임이 드러났다. 특히 주택인허가 연계에서 보완가능 사례가 전체의 4분의 1 이상을 차지한다는 점에서, 체계적인 데이터 정비 · 관리 시스템의 구축이 필요하다.

[표 3-55] 용도정보 정합성 검증 결과(건축물대장 ↔ 건축인허가)

변수명	분류	건수	비율
건축면적	정합성 높음	48,404건	92.25%
	연계 보정 필요	3,959건	7.55%
	연계 불가	106건	0.20%
	합계	52,469건	100.00%
연면적	정합성 높음	47,818건	91.14%
	연계 보정 필요	4,649건	8.86%
	연계 불가	2건	0.01%
	합계	52,469건	100.00%
용적률산정연면적	정합성 높음	48,105건	91.70%
	연계 보정 필요	4,269건	8.14%
	연계 불가	95건	0.18%
	합계	52,469건	100.00%

출처: 연구진 작성

[표 3-56] 용도정보 정합성 검증 결과(건축물대장 ↔ 주택인허가)

변수명	분류	건수	비율
건축면적	정합성 높음	1,332건	68.45%
	연계 보정 필요	529건	27.18%
	연계 불가	85건	4.37%
	합계	1,946건	100.00%
연면적	정합성 높음	1,408건	72.35%
	연계 보정 필요	521건	26.77%
	연계 불가	17건	0.87%
	합계	1,946건	100.00%

변수명	분류	건수	비율
용적률산정연면적	정합성 높음	1,347건	69.22%
	연계 보정 필요	508건	26.10%
	연계 불가	91건	4.68%
	합계	1,946건	100.00%

출처: 연구진 작성

■ 용도 정보 정합성 검증

건물ID를 기준으로 연계된 데이터의 용도 정보 정합성을 검증하기 위해, 본 연구에서는 '주용도코드'를 비교 대상으로 선정하였다. 주용도코드는 건축물의 주된 사용 목적을 나타내는 범주형 변수로서, 앞서 검증한 면적 정보(수치형 변수)와는 성격이 상이하다. 수치형 변수의 경우 값의 차이를 정량화하여 허용 오차 범위 내에서 실질적 일치 여부를 판단할 수 있었다. 반면, 범주형 변수인 주용도코드는 수치적 차이의 개념이 존재하지 않으므로, 값의 동일성 여부에 따른 이분법적 판단만이 가능하다.

용도 정보의 비교를 위해, 면적 변수 검증과 동일하게 다음 네 가지 유형으로 구분하였다. 일치 유형은 두 데이터베이스 모두 동일한 주용도코드를 보유한 경우이다. 불일치 유형은 두 데이터베이스 모두 주용도코드가 존재하나 서로 다른 값을 가진 경우이다. 보완가능 유형은 한쪽 데이터베이스에서는 결측이나, 다른 쪽에서는 정상적인 주용도코드가 존재하는 경우이다. 이 경우, 데이터 간 상호 보완을 통해 결측 보정이 가능한 상태이다. 완전결측 유형은 두 데이터베이스 모두 주용도코드가 결측인 경우이다. 이 경우, 데이터 내부적으로는 보완 불가하며 외부 검증 자료가 필요하다.

용도 정보 정합성 검증은 단순 일치 여부를 판단할 수 있다는 점에서 면적 변수 검증과 본질적인 차이가 있다. 즉, 차이의 정도를 정량화할 수 없기에 실질적으로 두 데이터가 같은 건물을 가리키고 있는지 알 수 없다. 따라서 불일치 사례에 대해 어느 데이터베이스의 값이 '정확한지' 판단하는 것은 본 연구 범위 내에서는 불가능하다. 불일치 검증을 보완하기 위해서는 행정자료, 현장조사, 또는 제3의 데이터 소스가 추가적으로 필요하다.

건축물대장과 건축인허가 연계 비교를 수행하였다. 공통 건물ID를 가진 데이터 중 비교 가능한 건수는 55,352건으로, 이는 필터링된 건축인허가 데이터(59,406건) 기준 93.18%에 해당한다. 일치 유형은 주용도코드와 주용도코드명 모두 54,558건(98.57%)이 일치하는 것으로 나타났다. 코드값이 일치하는 경우 코드명도 함께 일치하는 패턴을 보였으며, 이는 두 변수가 상호 연동되어 정확하게 관리되고 있음을 의미한다. 불일치 유형은 794건(1.43%)에서 주용도코드와 주용도코드명이 동시에 상이한 값을 가지는 것으로 확인되었다. 코드가 불일치하는 경우 코드명도 함께 불일치하는 일관된 패턴을 보였다. 보완가능 · 완전결측 유형은 용도 변수에서는 전혀 존재하지 않았다. 이는 면적 변수와 달리 용도 정보가 두 데이터베이스 모두에서 완전하게 입력되어 있음을 나타낸다. 면적 정보에서 90% 수준의 일치율을 보인 것과 비교하여, 용도 정보는 98.57%라는 높은 일치율을 기록하였다. 또한 보완가능 유형이나 완전결측 유형이 존재하지 않는다는 점에서 용도 정보의 데이터 품질이 면적 정보보다 상대적으로 양호한 상태인 것으로 판단된다.

[표 3-57] 각 변수별 유형 분포 비교(건축물대장 ↔ 건축인허가)

유형	주용도코드		주용도코드명	
	건수	비율	건수	비율
일치	54,558	98.57%	54,558	98.57%
불일치	794	1.43%	794	1.43%
보완가능	0	0%	0	0%
완전결측	0	0%	0	0%

출처: 연구진 작성

불일치 유형(794건)을 세부적으로 분류한 결과, 상위 10개 패턴이 전체의 79.8%를 차지하였다. 주요한 불일치는 제1종근린생활시설, 제2종근린생활시설, 단독주택 간의 혼동이 대부분을 차지하는 것으로 나타났다. 창고시설, 숙박시설, 노유자시설, 공장 등도 불일치 유형에 포함되었으나, 이들 역시 대부분 근린생활시설 및 단독주택과의 혼재된 형태로 불일치가 발생하였다.

한 예로 서울특별시 동대문구 전농동 건물의 경우, 건축물대장 주용도는 '제2종근린생활시설', 건축인허가는 '제1종근린생활시설'로 기재되어 있었다. 이는 근린생활시설 세부 분류 기준이 데이터베이스 간 다르게 적용된 결과로 해석할 수 있다. 다른 사례인 서울특별시 관악구 신림동 건물의 경우, 건축물대장은 '제2종근린생활시설 외 1(단독주택 포함)'으로 복합 기재되어 있으나 건축인허가는 '단독주택' 단일 용도로만 기록되어 있었다. 이는 두 데이터 간의 복합용도 입력 방식 차이에서 기인한 불일치로 해석할 수 있다.

[표 3-58] 불일치 그룹별 상위 10개 건수(건축물대장 ↔ 건축인허가)

대장_주용도코드	인허가_주용도코드	대장_주용도코드명	인허가_주용도코드명	불일치건수
04000	03000	제2종근린생활시설	제1종근린생활시설	255건
04000	01000	제2종근린생활시설	단독주택	93건
03000	04000	제1종근린생활시설	제2종근린생활시설	85건
01000	04000	단독주택	제2종근린생활시설	56건
03000	01000	제1종근린생활시설	단독주택	38건
15000	01000	숙박시설	단독주택	36건
04000	18000	제2종근린생활시설	창고시설	25건
01000	03000	단독주택	제1종근린생활시설	24건
11000	03000	노유자시설	제1종근린생활시설	13건
17000	04000	공장	제2종근린생활시설	9건

출처: 연구진 작성

다음으로 건축물대장과 주택인허가를 연계 비교하였다. 공통 건물ID를 가진 데이터 중 비교 가능한 건수는 1,941건으로, 필터링된 주택인허가 데이터(5,887건) 기준 32.96%가 연계되었다. 일치는 주용도코드와 주용도코드명 모두 1,884건(97.06%) 이 일치하는 것으로 나타났다. 불일치는 57건(2.94%)에서 주용도코드와 주용도코드명이 동시에 상이한 값을 가졌으며, 코드와 코드명 모두 동일한 데이터 행에서 불일치하는 패턴을 확인했다. 보완가능 · 완전결측은 건축인허가 연계와 마찬가지

로 전혀 존재하지 않았다. 즉, 건축인허가와 마찬가지로 코드와 코드명 간의 매핑이 정확하게 관리됨을 확인했다.

불일치 57건 중 상위 10개 패턴이 30건(52.6%)을 차지하였으며 주요한 불일치는 근린생활시설 계열 혼동, 복합용도/세부 용도 혼동 등으로 나타났다. 이처럼 건축인허가 연계에서 집중적으로 나타났던 “제1종/제2종 근린생활시설과 단독주택 혼동” 패턴과 달리, 주택인허가 연계에서는 보다 다양한 세부 용도에서 불일치가 분포하는 특징을 보였다.

불일치 사례로는 인천광역시 계양구 병방동 건물의 경우, 건축물대장 주용도는 '제1,2종 근린생활시설'로 기재되어 있으나, 건축인허가는 '소매점'으로 기록되어 있었다. 건축물대장을 확인한 결과 1층 층별 용도에 '소매점'이 기재되어 있어, 이는 층별 세부용도가 주용도로 입력된 사례로 해석할 수 있다. 또다른 사례로 서울특별시 노원구 하계동 건물의 경우, 건축물대장은 '제2종근린생활시설'로 기재되어 있으나, 건축인허가는 '학원'으로 기록되어 있었다. 건축물대장 확인 결과 1-2층 층별 용도에 '학원'이 기재되어 있어, 이는 세부 용도가 주용도로 전환된 사례로 해석할 수 있다.

[표 3-59] 불일치 그룹별 상위 10개 건수(건축물대장 ↔ 주택인허가)

대장_주용도코드	인허가_주용도코드	대장_주용도코드명	인허가_주용도코드명	불일치건수
04000	03001	제2종근린생활시설	소매점	6건
04000	04010	제2종근린생활시설	학원	4건
04000	04402	제2종근린생활시설	사무소	4건
04000	10000	제2종근린생활시설	교육연구시설	3건
03000	03005	제1종근린생활시설	의원	3건
02000	04402	공동주택	사무소	2건
11000	11201	노유자시설	노인복지시설	2건
03000	03001	제1종근린생활시설	소매점	2건
03000	02000	제1종근린생활시설	공동주택	2건
07000	04001	판매시설	일반음식점	2건

출처: 연구진 작성

용도 정보 정합성 검증 결과, 건축물대장-건축인허가 연계에서는 55,352건 중 54,558건(98.57%), 건축물대장-주택인허가 연계에서는 1,941건 중 1,884건(97.06%)의 용도 정보가 일치하는 것으로 나타났다. 이는 면적 정보에서 확인된 정합성 수준(건축인허가 90%, 주택인허가 70%)보다 높은 수치로, 용도 정보의 데이터 품질이 상대적으로 우수한 상태임을 보여준다.

용도 정보의 데이터 품질 특성을 살펴보면, 면적 정보와 달리 보완가능 유형이나 완전결측 유형이 전혀 존재하지 않는 것으로 확인되었다. 이는 필터링 조건에 따른 인허가 자료에서 용도 정보가 모두 입력되어 있어 결측 문제가 없음을 의미한다. 또한 주용도코드와 주용도코드명의 일치 및 불일치 건수가 완전히 동일하게 나타났으며, 이는 두 변수가 항상 동일하게 동작하여 코드와 코드명 간의 매핑 관리 체계가 양호함을 시사한다.

불일치 유형의 주요 원인을 분석한 결과, 세 가지 주요 패턴이 확인되었다. 첫째, 근린생활시설의 세부 분류 기준 차이로, 제1종과 제2종 간의 구분이나 소매점, 학원, 사무소 등 세부 용도 기재 방식의 차이

가 불일치를 야기하였다. 둘째, 복합용도 건축물에서 대표 용도를 선정하는 기준이 데이터베이스 간 상이하여 불일치가 발생하였다. 셋째, 건축물대장의 층별 세부용도가 인허가 자료의 주용도로 기재되는 사례가 확인되었으며, 이는 각 데이터베이스의 용도 기재 방식과 분류 체계의 구조적 차이에 기인한 것으로 판단된다.

[표 3-60] 용도정보 정합성 검증 결과

연계 구분	총 연계건수	일치		불일치	
		건수	비율	건수	비율
건축물대장 ↔ 건축인허가	55,352	54,558	98.57%	794	1.43%
건축물대장 ↔ 주택인허가	1,941	1,884	97.06%	57	2.94%

출처: 연구진 작성

4. 소결

1) 분석 종합

이 장에서는 건축물대장 내 면적 데이터, 용도 데이터와 함께, 건축물대장과 인허가 데이터의 연계키 등 주요 건축물 데이터를 대상으로 데이터 유형별 품질 고도화 방법론을 시범적으로 적용하였다.

건축물대장 면적 데이터 품질 고도화에는 건축물대장 정비 당시 기존 업무규칙 기반 검증, 신규 검증 규칙 개발 및 적용, 기계학습 기반 이상값 탐지 등을 종합적으로 적용하였다. 먼저 건축물대장 정비에서 사용된 기존 업무규칙에 근거하여 대지면적, 건축면적, 건폐율, 용적률 등 데이터의 오류율을 분석하고, 시기별, 지역별 분포를 검토하였다. 면적 데이터는 1 ㎡ 등 오차범위를 설정하고 이를 벗어나는 차이를 오류로 판정하였다. 또한 연면적과 관련하여 건축물대장 정비에서 검토되지 않은 2개 규칙을 신규 검증규칙으로 개발하여 오류 발생 현황을 분석하였다.

기존 규칙 및 신규 규칙을 적용한 검증 결과 현재는 오류가 거의 발생하지 않는 경우도 있었으나, 모든 규칙에서 0.13%~5.01%의 오류율이 나타나, 건축물대장 정비를 통한 오류 정정이 신속하게 이루어지고 있지 못한 현황을 확인할 수 있었다. 시기별 및 지역별 분석 결과 오류 유형별로 분포가 다양하게 나타났으며, 특히 건축면적이 연면적보다 큰 오류의 경우 최근 준공 건축물에서 8~10%의 오류율이 지속적으로 발생하고 있는 것으로 나타났다.

기계학습 기반 이상값 탐지에서는 비지도 기계학습 이상탐지 방법론을 사용하여 규칙 기반 검증으로 걸러지지 않는 비정형적 오류를 탐지하였다. 건축물 규모에 영향받지 않는 무차원 지표를 도출하고, 규모와 독립적으로 면적 데이터의 이상값을 탐지한 결과, 건폐율, 용적률 등에서 일반적인 건축물과 크게 다른 이상값을 도출하였다. 시기별, 지역별 분포에서도 다른 규칙 기반 검증에서 도출된 오류와 달리 시기별로 일정한 오류율이 나타나, 단순히 규칙 기반으로 탐지할 수 없는 비정형적 오류를 탐지할 수 있는 가능성을 확인하였다.

건축물대장 용도 데이터 품질 고도화에는 건축물대장 용도 데이터에 기계학습 기반 분류 방법론을 적용하였다. 건축물대장 건물동 단위 표제부 테이블과 층 단위 층별개요 테이블에는 용도 분류가 코드로 기재되어 있으나, 오류 사례 및 상위 단어 빈도 분석 결과 코드 분류와 자연어로 기재된 '기타 용도' 컬럼의 내용과 일치하지 않는 경우를 확인하였다. 기계학습 분류 모델인 나이브 베이즈 모델을 활용

하여 기타 용도 단어에서 코드 분류를 예측하는 모델을 구축한 결과, 동별 용도 예측은 84.78%, 층별 예측은 78.97% 정확도를 달성하였다.

분류 모델이 실제 코드 분류와 다른 용도로 잘못 분류하는 경우를 검토한 결과, 실제 건축물대장 기재 내용이 불일치하는 오류 사례도 있었으나, 예측 모델이 단독주택 등 흔한 용도로 잘못 분류하거나, 여러 용도가 복합된 경우에 잘못 분류하는 등 과적합으로 인한 예측 실패 사례도 나타났다. 층별 데이터 분석에서는 개별 층의 분류를 위하여 건축물의 전체 맥락을 고려할 필요성이 확인되었다. 향후 데이터 품질 고도화를 위해서는 건물동 단위 표제부와 층 단위 층별개요 교차검증 체계를 마련하고, 딥러닝 모델이나 사전 훈련된 LLM 모델 등을 활용하여 복잡한 예측 문제에 접근할 필요가 있다.

마지막으로 건축물대장과 인허가 데이터의 연계 품질 고도화에는 건물ID 기반 건축물 데이터 연계 데이터를 구축하고, 건축물대장과 건축인허가, 주택인허가 데이터의 교차검증을 통하여 연계 가능성을 검토하였다. 건축물대장 기재 기준 2024년 사용승인 건축물 55,784동을 기준으로 건물ID는 54,123동(97.0%)에 부여되었으며, 그 중 54,103건은 건축물과 건물ID가 일대일로 부여되었다. 그러나 극소수 사례에서 여러 건축물에 동일한 건물ID(7건)가 부여된 경우가 발견되었다. 건축인허가(99.4% 부여), 주택인허가(99.4% 부여)에서도 건물ID가 대부분 부여되었으나, 건축인허가 데이터의 경우 11,408건에서 여러 건축물에 동일한 건물ID가 부여된 것이 확인되었다. 대지 내 여러 건축물의 경우, 동일 건축물에 대한 중복 레코드 존재 등이 원인으로 판단된다.

건물ID 기반 연계 성공률은 건축인허가의 경우 93.6%, 주택인허가는 34.9%로 나타났다. 연계 성공한 데이터의 정합성을 검토한 결과, 면적 변수의 경우 건축인허가는 90% 이상 일치하였으나, 주택인허가의 일치율은 60%대 후반~70%대 초반으로 나타났다. 용도 변수의 경우 건축인허가와 주택인허가 모두 90%대 후반 일치율을 보였다.

2) 시사점

본 연구에서는 시기별·지역별 오류율을 검토함으로써 건축물대장 데이터의 오류 발생 경향과 원인을 일정 부분 추정할 수 있었다. 특정 시기에 특정 오류가 집중되거나, 지역별로 오류율의 편차가 나타나는 현상은 단순한 입력상의 문제를 넘어 행정체계의 특성이나 제도적 요인과 연관되어 있음을 시사한다. 이는 기존 건축물대장 정비와 같이 정기적·획일적 정비 방식으로는 오류를 근본적으로 줄이기 어렵다는 점을 보여주며, 오류 발생의 맥락을 반영한 맞춤형 정비 체계, 즉 데이터 품질 정비의 고도화가 필요함을 의미한다.

또한 기존 규칙 기반 검증만으로는 포착되지 않는 오류가 여전히 존재함을 확인하였다. 이에 따라 연면적 관련 신규 규칙을 추가하여 검증 범위를 확장하고, 기계학습을 활용한 이상탐지 기법을 도입함으로써 규칙 기반 검증의 한계를 보완할 수 있음을 보였다. 이러한 결과는 향후 건축물대장 품질 고도화가 단일한 규칙 중심 접근을 넘어, 신규 규칙 개발과 기계학습 기반 탐지를 병행하는 다층적 검증 체계로 전환되어야 함을 시사한다.

아울러 건물ID를 중심으로 한 데이터 연계 분석은 건축물대장과 인허가 데이터 간의 품질 관리 필요성을 드러냈다. 건물ID는 건축물 정보를 다른 행정 데이터와 연결하는 핵심 키로 기능할 수 있으나, 실제 분석 결과 일부에서는 동일 건물ID가 복수 건축물에 부여되거나, 주택인허가 데이터에서 연계 성공률이 낮게 나타나는 등 관리상의 문제점이 확인되었다. 이는 건물ID 부여·관리 기준의 정교화와 데이터 현행화 노력이 동반되어야 함을 보여준다. 건물ID를 건축물 데이터 전반에 걸쳐 범용적인 건축물 단위 연계키로 활용하기 위해서는 건물ID의 고유성, 일관성, 정합성 등 품질 고도화가 우선될 필요가 있다.

제4장

건축물 데이터 품질 제고를 위한 방안 제안

1. 개요

■ 건축물 데이터 품질 제고 방안

본 연구는 건축물 데이터 관련 제도, 품질 및 활용 현황을 검토하고, 품질 고도화 방법론을 시범적으로 적용한 결과를 토대로, 데이터 특성을 반영한 품질 향상 방안을 제시하였다. 이를 바탕으로 건축물 데이터의 생산부터 유통까지 전 단계에 품질 고도화 방법론을 적용함으로써, 최종적으로 활용되는 건축물 데이터의 품질을 체계적으로 향상시키는 것을 기본 방향으로 설정하였다.

현재 건축행정 데이터는 세움터에서 생산 및 관리되고 있으며, 개방 데이터의 유통에 한하여 건축허브를 거쳐 이루어지고 있다. 건축허브는 세움터 데이터를 최소한의 가공만 거쳐 원데이터 그대로 제공하고 있으며, 따라서 현재 유통 및 활용되는 건축물 데이터에는 세움터 데이터에 내재된 오류가 그대로 전파되고 있다. 본 연구에서는 먼저 데이터 생산 및 유통 과정에서 오류 검증 및 정제가 이루어질 수 있는 방안을 종합하고, 이어 각 방안에 대한 내용을 구체적으로 제시하고자 한다.

본 연구에서 제안하고자 하는 건축물 데이터의 품질 제고 방안은 크게 두 가지로 구분할 수 있다.

첫째는 이미 생산된 건축물 데이터의 품질을 개선하여 유통하는 것이고, 둘째는 앞으로 새롭게 생산될 데이터의 품질을 개선하는 것이다. 먼저, 이미 생산된 기존 건축물 데이터의 품질을 개선하는 방안은 현재 공공데이터포털이나 건축HUB 등을 통해 제공되고 있는 건축물대장 데이터를 대상으로 한다. 즉, 이미 유통되고 있는 데이터의 품질을 향상시키기 위해 체계적인 데이터 정제 방법론을 마련하고, 이를 적용·보완해 나가는 절차를 제안하고자 한다.

다른 하나는 향후 생산되는 건축물 데이터의 품질을 개선하는 것이다. 즉 건축물의 허가 및 사용승인을 위한 입력시스템을 개선하여 입력 자료의 오류를 최소화할 수 있는 방안을 제안하고자 한다. 이를 위해 본 연구에서 건축물 데이터 품질 개선 과정에서 발견한 오류 유형을 검토하고, 이를 시스템에 적용할 수 있는 방안을 제안하고자 한다.

[그림 4-1] 건축물 데이터 품질 제고 방안
출처: 연구진 작성

■ 품질 제고 건축물 데이터 유통

품질 제고 건축물 데이터를 유통하기 위한 방안을 제안하고자 한다. 기존에는 연구 및 정책수립을 위해 세움터, 공공데이터포털, 건축HUB 등에서 생산 및 유통하는 건축물 데이터를 활용하였다. 기존에 제공하고있는 건축물 데이터는 공적장부에 기반하고 있는 경우가 많아 국가기관에서 직권정정할 수 없는 경우가 많다. 즉, 기존 시스템에 저장된 원본 데이터를 정정하여 제공하기는 어렵다. 품질이 개선된 데이터를 기존 시스템에 반영하여 제공하기보다는 새로운 시스템에 저장하고 유통하는 방식을 검토할 필요가 있다.

[그림 4-2] 건축물 데이터 품질 제고 방안
출처: 연구진 작성

본 연구에서는 두 가지 방향으로 품질 제고 건축물 데이터 유통 방안을 모색하고자 하였다. 첫째는 공공데이터포털과 같은 유통 시스템을 통한 정제된 건축물 데이터 유통 방안이다. 공공기관은 생산하는 공공데이터를 관리하고 유통할 의무가 있으며, 공공데이터포털과 같은 시스템은 이를 지원하는 플랫폼이다. 이에 본 연구에서 제시하는 품질 제고 건축물 데이터의 공공데이터포털을 통한 유통 가능성과 유통 방법을 검토하고자 한다.

둘째로 건축물 데이터 품질 개선을 위해 개발한 분석 프로세스나 분석 알고리즘을 오픈소스 형식으로 유통하는 방법을 검토하고자 하였다. 오픈소스는 누구나 자유롭게 접근하여 활용·수정·배포할 수 있고, 이에 따라 다양한 이용자의 피드백이 이루어져 지속적인 개선이 가능하다. 특히 현재 건축물 데이터 유통방식과 비교할 때, 최종 활용 단계에서 발견된 오류를 유통 데이터에서 정정할 수 있는 방안이 필요하다. 본 연구는 이 같은 오픈소스의 순기능을 활용하기 위해 건축물 데이터 품질 개선 프로세스 및 알고리즘을 오픈소스 형태로 유통할 수 있는 방법을 검토하고자 한다.

■ 신규 데이터 품질 제고

본 연구에서 도출한 건축물 데이터 오류 검출 방법을 건축행정시스템 세움터에 도입할 수 있는 방안을 제시하고자 한다. 이를 위해 세움터 내 건축허가 및 사용승인 시스템의 구조와 운영 방식을 조사하고, 시스템 이용자와 관리자의 의견을 수렴하였다. 특히 건축행정정보 중에서 건축물 데이터의 핵심 요소로 빈번히 활용되는 면적 및 용도 관련 항목이 시스템상에서 어떻게 기입되는지를 검토하였다. 본 연구에서 검토한 건축물 데이터의 오류 유형과 세움터 시스템 내 항목별 입력 절차를 비교하여 오류 발생 가능성과 원인을 도출하고, 이를 기반으로 오류를 저감할 수 있는 시스템 개선 방향을 제안하였다.

2. 품질 제고 건축물 데이터 유통

1) 공공데이터포털을 통한 품질 제고 건축물 데이터의 유통

■ 품질 제고 건축물 데이터의 정의

본 연구에서 정의하는 '품질 제고 건축물 데이터'란 본 연구에서 제시하는 건축물 데이터 품질 제고 방법론을 적용하여 데이터에 내재된 오류 및 결측값을 제거하거나 다른 값으로 대체한 데이터를 통칭한다. 이는 건축법 등에 따른 건축물대장 정비, 허가권자의 직권정정 등 과정을 거쳐 원데이터 자체가 정정되는 경우와는 구분된다. 본 연구는 건축물대장 등 공적 장부 자체를 변경하지 않고, 별도의 품질 제고 건축물 데이터를 연구 및 통계작성 등 용도로 유통하는 방안을 제안하고자 한다.

■ 공공데이터포털을 통한 데이터 유통 방안

본 연구에서는 품질 제고 건축물 데이터의 유통을 위한 구체적 방안으로 공공데이터포털 활용을 제안하고자 한다. 공공데이터포털은「공공데이터의 제공 및 이용활성화에 관한 법률」제21조(공공데이터 포털의 운영)에 근거하여 운영되며, 공공데이터의 효율적 제공을 위하여 운영되는 공공데이터의 통합제공시스템이다. 공공데이터 개방의 목적은 "중앙정부·지방자치단체 및 공공기관이 보유·관리하는 공공데이터를 일반 국민이 자유롭게 이용할 수 있도록 다양한 형태로 개방·제공"하는 것이며[15], 본 연구에서 제안하는 품질 제고 건축물 데이터의 경우도 일반 국민이 별도 데이터의 존재를 인식하고 활용하도록 하기 위하여 공공데이터포털을 통하여 유통할 필요가 있다.

> 제21조(공공데이터 포털의 운영)
>
> ① 행정안전부장관은 공공데이터의 효율적 제공을 위하여 통합제공시스템(이하 "공공데이터 포털"이라 한다)을 구축 · 관리하고 활용을 촉진하여야 한다. 〈개정 2014. 11. 19., 2017. 7. 26.〉
>
> ② 행정안전부장관은 공공기관의 장에게 공공데이터 포털의 구축과 운영에 필요한 공공데이터의 연계, 제공 등의 협력을 요청할 수 있다. 이 경우 요청을 받은 공공기관의 장은 특별한 사유가 없는 한 이에 따라야 한다. 〈개정 2014. 11. 19., 2017. 7. 26.〉
>
> ③ 그 밖에 공공데이터 포털의 구축 · 관리 및 활용촉진 등 필요한 사항은 대통령령으로 정한다.
>
> 출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률」. 법률 제19408호, 2023. 5. 16., 타법개정

15) 행정안전부. https://www.mois.go.kr/frt/sub/a06/b02/openData/screen.do. 2025.05.07. 접속

- 공공데이터 해당 여부 검토

품질 제고 건축물 데이터는 「공공데이터의 제공 및 이용활성화에 관한 법률」제2조(정의) 제2호에 따른 공공데이터에 해당한다. 품질 제고 건축물 데이터가 본 연구를 수행한 건축공간연구원에 의해 본 연구에서 제안한 품질 제고 방법론을 적용하여 생산된다고 가정할 때, 건축공간연구원은 「공공기관의 운영에 관한 법률」에 따른 공공기관에 해당하며, 품질 제고 건축물 데이터는 「지능정보화 기본법」 제2조(정의)제1호에서 정의하는 "광(光) 또는 전자적 방식으로 처리되는 부호, 문자, 음성, 음향 및 영상 등으로 표현된 모든 종류의 자료 또는 지식"으로서의 정보에 해당한다. 따라서 품질 제고 건축물 데이터는 「공공데이터의 제공 및 이용활성화에 관한 법률」제2호나목에서 정한 「지능정보화 기본법」제2조(정의)제1호에 따른 정보 중 공공기관이 생산한 정보에 해당하므로, 공공데이터에 해당한다.

> 제2조(정의)
>
> 2. "공공데이터"란 데이터베이스, 전자화된 파일 등 공공기관이 법령 등에서 정하는 목적을 위하여 생성 또는 취득하여 관리하고 있는 광(光) 또는 전자적 방식으로 처리된 자료 또는 정보로서 다음 각 목의 어느 하나에 해당하는 것을 말한다.
> 가. 「전자정부법」 제2조제6호에 따른 행정정보
> 나. 「지능정보화 기본법」 제2조제1호에 따른 정보 중 공공기관이 생산한 정보
> 다. 「공공기록물 관리에 관한 법률」 제20조제1항에 따른 전자기록물 중 대통령령으로 정하는 전자기록물
> 라. 그 밖에 대통령령으로 정하는 자료 또는 정보
>
> 출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률」. 법률 제19408호, 2023. 5. 16., 타법개정

- 민간에 개방하여야 하는 공공데이터 해당 여부 검토

다음으로 품질 제고 건축물 데이터는 민간에 개방하여야 하는 공공데이터에 해당한다. 민간에 개방하여야 하는 제공 대상 공공데이터에 해당하려면 「공공데이터의 제공 및 이용활성화에 관한 법률」제17조(제공대상 공공데이터의 범위)를 충족해야 한다. 공공기관이 보유 및 관리하는 공공데이터는 원칙적으로는 모두 국민에게 제공되어야 하나 법 제9조에 따른 비공개대상정보나 「저작권법」 및 다른 법령에서 명시하는 제3자 권리가 포함된 것 중 이용허락을 받지 않은 정보는 제외하여야 한다.

「공공데이터 관리지침」 또한 공공데이터 제공을 위해 필요한 절차와 요건을 제시하고 있다. 첫째로 지침 제7조(제공대상 여부의 확인)에 따라 품질 제고 건축물 데이터가 제공대상 공공데이터인지 여부를 확인하여야 한다. 지침 또한 상위 법령과 동일하게 「공공기관의 정보공개에 관한 법률」제9조에 따른 비공개대상정보와 저작권 관련 제한이 없는 한 공공데이터를 제공해야 하는 것으로 규정하고 있다. 본 연구에서 제시한 품질 제고 건축물 데이터는 이미 개방 유통되고 있는 공공데이터를 가공한 것으로, 법령 및 지침에 따른 비개방 데이터가 포함되어있지 않으므로 품질 제고 건축물 데이터 또한 제공 대상 공공데이터에 해당한다.

제7조(제공대상 공공데이터의 범위)

① 공공기관의 장은 해당 공공기관이 보유 · 관리하는 공공데이터를 국민에게 제공하여야 한다. 다만, 다음 각 호의 어느 하나에 해당하는 정보를 포함하고 있는 경우에는 그러하지 아니한다.
1. 「공공기관의 정보공개에 관한 법률」 제9조에 따른 비공개대상정보
2. 「저작권법」 및 그 밖의 다른 법령에서 보호하고 있는 제3자의 권리가 포함된 것으로 해당 법령에 따른 정당한 이용허락을 받지 아니한 정보
② 공공기관의 장은 제1항에도 불구하고 제1항 각 호에 해당하는 내용을 기술적으로 분리할 수 있는 때에는 제1항 각 호에 해당하는 부분을 제외한 공공데이터를 제공하여야 한다.
③ 행정안전부장관은 제1항제2호의 제3자의 권리를 포함하는 것으로 분류되어 제공대상에서 제외된 공공데이터에 대한 정당한 이용허락 확보를 위한 방안을 제시할 수 있으며, 공공기관의 장은 그 방안에 따라 필요한 조치를 취하여야 한다. 〈개정 2014. 11. 19., 2017. 7. 26.〉

출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률」. 법률 제19408호, 2023. 5. 16., 타법개정

제7조(제공대상 여부의 확인)

① 공공기관의 장은 다음 각 호의 어느 하나에 해당하는 정보를 포함하는 경우를 제외하고 해당 기관이 보유 · 관리하는 모든 공공데이터를 제공해야 한다.
1. 「공공기관의 정보공개에 관한 법률」 제9조에 따른 비공개 대상 정보
2. 저작권 등 다른 법령에서 보호하고 있는 제3자의 권리가 포함된 정보. 단, 권리자의 정당한 이용허락을 받은 경우에는 제공대상에 해당한다.
② 공공기관의 장은 제1항 각 호에 따라 제공대상에 해당하지 않는 데이터라 하더라도, 해당 부분을 기술적으로 분리할 수 있는 경우에는 이를 제외한 공공데이터를 제공해야 한다.

출처: 「공공데이터 관리지침」. 행정안전부고시 제2021-70호, 2021. 10. 26., 전부개정

- 연구기관이 생산한 공공데이터 유통 사례 검토

공공기관이 보유한 데이터가 제공대상 공공데이터에 해당하는 경우 공공기관의 장은 「공공데이터의 제공 및 이용활성화에 관한 법률」제18조(공공데이터 목록의 등록) 제1항에 따라 공공데이터 목록을 등록하여야 한다. 실제 공공데이터포털의 목록개방현황 자료를 보면 2025년 5월[16]을 기준으로 총 87,716개의 공공데이터가 등록되어 있으며, 이 중 보건환경연구원, 농어촌연구원, 한국건설기술연구원 등 연구기관의 공공데이터도 포함되어있음을 확인할 수 있다.

제18조(공공데이터 목록의 등록)

① 공공기관의 장은 해당 공공기관의 소관 공공데이터 목록을 대통령령으로 정하는 바에 따라 행정안전부장관에게 등록하여야 한다. 〈개정 2014. 11. 19., 2017. 7. 26.〉
② 행정안전부장관은 제1항에 따른 등록의 누락이 있는지를 조사하여 누락된 공공데이터 목록의 등록을 요청할 수 있다. 〈개정 2014. 11. 19., 2017. 7. 26.〉
③ 행정안전부장관은 제1항 및 제2항에 따라 등록된 공공데이터 목록에 관한 정보를 그 내용별, 형태별, 이용대상별 등 이용에 용이하게 분류하여 관리 · 제공하여야 한다. 〈개정 2014. 11. 19., 2017. 7. 26.〉
④ 행정안전부장관은 공공데이터의 체계적 관리와 제공 및 이용 활성화 정책의 효율적 집행을 위하여 제21조에 따른 공공데이터 포털에 공공데이터목록등록관리시스템을 구축 · 운영하여야 한다. 〈개정 2014. 11. 19., 2017. 7. 26.〉

출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률」. 법률 제19408호, 2023. 5. 16., 타법개정

16) 공공데이터포털. https://www.data.go.kr/data/15062804/fileData.do. 2025.05.08. 접속

• 공공데이터 유통방식 결정 및 처리

「공공데이터 관리지침」제3조 제2항에 따라 공공데이터는 원천데이터 형태로 유통할 수 있도록 해야 한다. 각 기관은 제공대상 공공데이터의 유형 및 특성과 함께 데이터 이용자의 편의성을 고려하여 개방할 필요가 있다. 공공데이터 제공·관리 실무 매뉴얼에서는 갱신이 빈번한 데이터의 경우 오픈 API 방식을, 인물, 지명, 도서데이터 등 정적데이터 형태로 연관성 높은 데이터가 많은 경우 LOD 등 방식을 권고하고 있다. 또한, 데이터 분석·학습 등에 활용되는 데이터, 오픈API 제공을 위한 개방DB, API 서비스 환경 등 제공기반이 마련되지 않은 경우, 데이터가 빈번하게 생성 및 갱신되지 않는 경우 등은 파일데이터를 제공하도록 권고하고 있다[17].

품질 제고 건축물 데이터는 파일데이터 형식이 합리적인 데이터 제공방식일 것으로 판단된다. 품질제고 건축물 데이터는 건축행정정보를 바탕으로 생산되는 데이터이며, 건축행정정보와 달리 비정기적으로 생산 및 갱신될 것으로 추정된다. 또한, 과거 활용 형태를 볼 때 통계생산이나 데이터 분석 및 학습에 활용될 가능성이 높은 데이터이다. 이에 본 연구에서는 품질 제고 건축물 데이터를 파일데이터 형식으로 제공하는 방안을 제시한다.

• 공공데이터 목록 등록

「공공데이터의 제공 및 이용 활성화에 관한 법률」제18조제1항에 따라 기관은 생성하거나 수집한 날로부터 15일 이내에 행정안전부장관에게 공공데이터 목록을 등록하도록 되어있다. 행정안전부장관에게 공공데이터 목록을 등록하는 행위는 공공데이터목록등록관리시스템[18]을 통해 공공데이터 목록을 등록하는 것으로 대체할 수 있다[19].

공공데이터 목록 등록을 위한 절차는 크게 세 단계로 구분된다. 첫째, 기관 내에서 제공 대상 공공데이터로 식별된 데이터 목록을 구성하여야 한다. 둘째로 구성한 공공데이터 목록에 기반하여「공공데이터의 제공 및 이용 활성화에 관한 법률 시행규칙」별지 제2호서식에 있는 공공데이터 목록 등록서를 작성하고, 등록하여야 한다. 마지막으로 기관이 제공하고 있는 공공데이터와 공공데이터 목록의 일치여부를 확인하고 관리하여야 한다.

• 공공데이터 등록

품질 제고 건축물 데이터가 공공데이터 목록에 등록될 경우「공공데이터의 관리지침」별표1 공공데이터 목록·데이터 등록기준에 따라 등록되어야 한다. 데이터의 등록은 크게 파일데이터 형태와 오픈API의 두 가지로 구분할 수 있다. 파일데이터 형태로 데이터를 제공할 경우 1년 이하의 갱신주기를 유지하여야 한다. 또한 데이터는 CSV, XML, JSON과 같이 구조화된 형태의 개방 파일형식으로 등록하는 것을 원칙으로 하고 있다. 오픈API의 경우 실제 데이터와 데이터에 대한 기술문서가 일치하도록 관리하여야 한다. 본 연구에서 제시한 품질 제고 건축물 데이터는 “데이터가 빈번하게 생성 및 갱신되지 않는 경우”에 해당하여 파일데이터로 제공할 필요가 있다.

17) 행정안전부. 2021. 공공데이터 제공·관리 실무 매뉴얼. p.25

18) all.data.go.kr

19) 행정안전부. 2021. 공공데이터 제공·관리 실무 매뉴얼. p.26

「공공데이터의 제공 및 이용 활성화에 관한 법률」 별지 제4호 서식에 있는 제공대상 공공데이터의 포털 등록 서식을 작성하여야 품질 제고 건축물 데이터를 공공데이터포털을 통해 제공할 수 있다.

■ 공공데이터의 제공 및 이용 활성화에 관한 법률 시행규칙 [별지 제2호서식] <개정 2016. 11. 22.>

(앞쪽)

공공데이터 목록 등록서

「공공데이터의 제공 및 이용 활성화에 관한 법률」 제18조제1항·제27조제5항 및 같은 법 시행령 제15조제1항에 따라 공공데이터 목록을 아래와 같이 등록합니다.

① 공공데이터 목록 명칭(국문)		
② 공공데이터 목록 명칭(영문)		
③ 정부기능 분류체계(BRM)		
④ 보유 근거(법령)		
⑤ 키워드(3개)		
⑥ 공공데이터 설명		
⑦ 제공대상 여부	제공 [ ]	부분제공 [ ]
	미제공 [ ] 미제공 사유: ( ) 미제공 근거 법률: ( )	
⑧ 향후 제공	향후 제공 [ ] 향후 제공 연도: ( ) 향후 제공 사유: ( )	
⑨ 제공신청에 의한 등록	예 [ ] 아니오 [ ]	

210㎜×297㎜(백상지 80g/㎡)

■ 공공데이터의 제공 및 이용 활성화에 관한 법률 시행규칙 [별지 제4호서식] <개정 2017. 7. 26.>

(앞쪽)

제공대상 공공데이터 등록서

「공공데이터의 제공 및 이용 활성화에 관한 법률」 제19조제4항에 따라 제공대상 공공데이터를 같은 법 제21조에 따른 공공데이터 포털에 등록합니다.

① 제공대상 공공데이터 명칭				
② 제공대상 공공데이터 설명				
③ 공공저작물 여부	예 [ ] 아니오 [ ]			
④ 제공대상 공공데이터 제3자 권리 포함 유무	제3자 권리(저작권, 소유권 등) 미포함 [ ]	제3자 권리(저작권, 소유권 등) 포함 [ ]		
		이용허락 등 권리 확보 [ ]	이용허락 등 권리 미확보 [ ]	
⑤ 이용허락범위	저작자표시 [ ]	저작자표시-비영리 [ ]		
	저작자표시-변경금지 [ ]	저작자표시-비영리-변경금지 [ ]		
	저작자표시-동일조건 변경허락 [ ]	저작자표시-비영리-동일조건 변경허락 [ ]		
	사유 : ( )			
⑥ 제공대상 공공데이터 업데이트 주기				
⑦ 제공대상 공공데이터의 차기 등록 예정일	년 월 일			
⑧ 제공대상 공공데이터 비용부과 유무	무료 [ ]	유료 [ ] (원)		
⑨ 제공대상 공공데이터 비용 부과기준 및 단위	건 [ ] (부과단위:) 금액( )	이용량 [ ] (부과단위:) 금액( )	이용기간 [ ] (부과단위:) 금액( )	기타 [ ] * 부과기준 단위 등이 포함된 자료 별도제출 금액( )
⑩ 제공대상 공공데이터 제공 형태	공공데이터 포털에서 다운로드 [ ]	기관자체에서 다운로드 [ ]		
	전자기록매체 저장 제공 [ ]	링크드 오픈데이터(LOD) 연계파일등록 [ ]		
	링크드 오픈데이터(LOD) 연계URL등록 [ ]			
⑪ 제공대상 공공데이터 / 위치(URL)				
⑫ 제공대상 공공데이터 파일명				
⑬ 제공대상 공공데이터 매체 유형	텍스트 [ ]	동영상 [ ]	이미지 [ ]	소리 [ ] 기타 [ ] (유형:)
⑭ 매체 유형별 건수	유형: ( )	건수: ( )	대표확장자명: ( )	
⑮ 언어				
⑯ 제공신청에 의한 등록	예 [ ] 아니오 [ ]			
⑰ 공공데이터 개방 표준	「공공데이터 개방 표준」(행정안전부고시)을 따른 데이터입니까? 예[ ] 아니오[ ]			
⑱ 기타 이용 유의사항				

210㎜×297㎜(백상지 80g/㎡)

[그림 4-3] 공공데이터 목록 등록서 및 제공대상 공공데이터 등록서

출처: 「공공데이터의 제공 및 이용 활성화에 관한 법률 시행규칙」. 행정안전부령 제436호, 2023. 11. 7., 일부개정. 별지 제2호서식, 별지 제4호서식

2) 품질 고도화 방법론의 오픈소스 기반 유통

■ 방법론의 오픈소스 기반 유통 필요성

건축물 데이터 품질 고도화 방법론을 오픈소스로 유통함으로써 단발성으로 그칠 수 있는 연구 성과를 반복 가능한 프로세스로 전환하여 지속 가능한 개선이 이루어지도록 할 수 있다. 즉, 본 연구가 제시한 품질 개선 절차를 현장에 안정적으로 확산시키기 위한 효율적이 방법이다. 건축물 데이터 품질 고도화 방법론을 오픈소스로 유통할 경우 네 가지 장점이 있다.

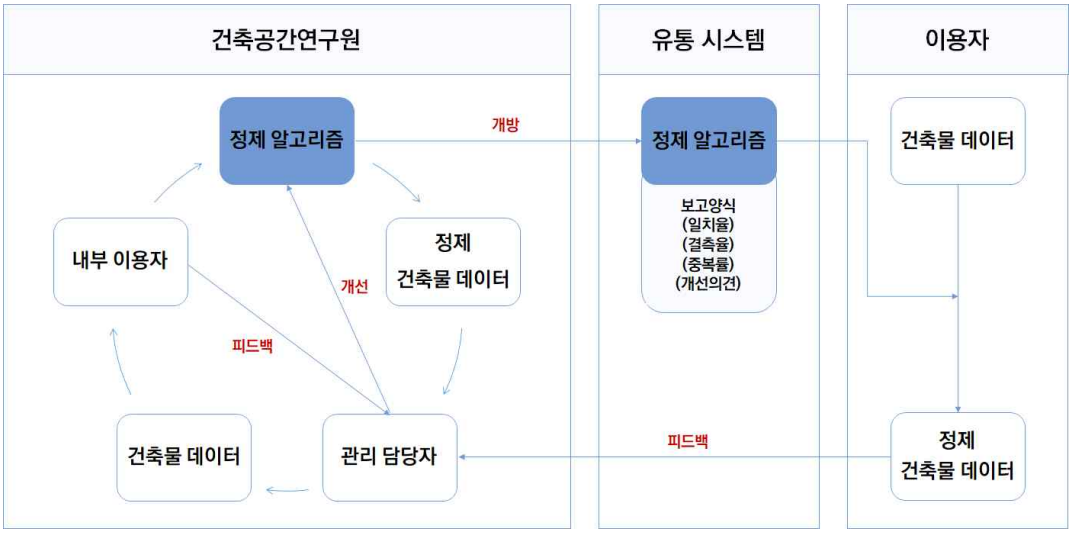
첫째, 검증 규칙과 임계값, 지자체별 예외 처리 등 핵심 로직이 공개될 때 결과의 재현성이 향상되고, 반복 사용에 따른 외부 검증과 이에 따른 문제 제기가 가능해져 모델의 신뢰도를 높일 수 있다. 둘째, 동일한 기능을 하는 방법론의 중복 개발을 방지할 수 있고, 기존 방법론을 기점으로 개선된 모듈 간 상호운용성을 확보해 (사실상의) 표준 방법론 형성에 기여한다. 셋째, 커뮤니티 기반의 이슈 추적과 취약점 신고 체계가 구축되면 오류와 결함의 발견·수정 속도가 빨라지고 보안 대응의 선제성이 강화된다. 넷째, 중앙정부와 지자체, 민간이 동일한 공개 규칙과 데이터 사전을 참조함으로써 정책–기술 간 피드백 루프가 촘촘해지고, 현장 요구가 곧바로 방법론의 개정과 버전 업데이트로 환류되는 선순환 구조가 정착된다.

공간정보 분야와 데이터 품질 분야에서 이미 검증된 오픈소스 사례는 본 연구의 유통 전략을 구체화하는 데 중요한 시사점을 제공한다. QGIS와 PostGIS, GeoPandas 등은 플러그인 구조와 관련 문서를 공개하는 체계를 통해 다양한 실무자들의 요구를 반영할 수 있었고, 오픈소스 방식을 적용함으로써 국가별, 지역별로 상이한 데이터 관행을 표준화할 수 있는 바탕이 되었다. OpenStreetMap은 사용자 참여 기반의 검수 체계, 그리고 지역 사용자의 모니터링을 통해 신속하게 오류를 수정하는 방식을 보여주었다. 정부와 공공기관이 일부 데이터 플랫폼을 오픈소스를 위한 플랫폼으로 전환하고, 활용 가이드와 보안 정책을 함께 배포하여 오픈소스 방식의 장점을 활용하려는 흐름이 점차 확산되고 있다. 본 연구의 건축물 데이터 품질 제고 방법론 역시 오픈소스 방식으로 유통할 때 가장 적은 비용으로 효율적으로 개선될 것으로 기대할 수 있다.

■ 오픈소스 기반 유통 방안

본 연구에서 제작한 건축물 데이터 품질 제고 방법론은 크게 세 가지로 구분할 수 있다. 첫째로 오류 탐지 규칙을 활용하는 방식, 둘째로 인공지능 및 기계학습을 활용한 이상치 도출 방식, 셋째로 인허가–건축물대장을 건물ID를 축으로 교차검증하는 방식, 총 세 가지로 구분할 수 있다. 이러한 품질 제고 방법론에 포함된 개별 규칙 및 프로세스를 문서와 코드 형식으로 공개하면 재현성을 확보하고 구체적인 방법론의 활용을 촉진할 수 있다. 이러한 오픈소스 유통의 장점을 활용하기 위해서는 알고리즘의 소스코드, 규칙, 관련 설명문 등 핵심 로직을 함께 공개할 필요가 있다. 세부적으로는 입력자료의 구득 방법 뿐 아니라 행정구역별 예외 규칙을 묶은 규칙 팩, 스키마 표준화 도구 등 분석을 위한 전처리 방법과 분석 결과를 함께 공개해야 한다. 인공지능 모듈의 경우 학습에 사용한 방법, 변수, 파라미터를 함께 제공하여 다른 이용자가 동일한 성능을 재현할 수 있도록 해야한다.

알고리즘의 공개 방식은 단계적으로 추진할 필요가 있다. 초기에는 연구원 내부 공개를 통해 최초 버전을 게시하고, 신규 건축물 데이터에 대한 테스트 구동 및 시범 적용을 통해 확인된 오류와 개선 사항을 반영한 이후 공공데이터 포털 등을 통해 배포할 필요가 있다. 공개 목록에는 설치 안내 데이터 사전, 튜토리얼, 변동 이력, 보안 정책 등을 포함하여 모든 이용자가 알고리즘을 적용해볼 수 있도록 해야한다. 알고리즘 자체는 가장 보편적으로 사용되는 파이썬 패키지 형태로 제공하는 것이 효율적일 것으로 판단된다. 높은 범용성, 방대한 라이브러리, 빠른 개발 속도가 장점인 파이썬 패키지는 다양한 분야의 사용자가 활용하기 간편하며, 피드백을 반영하기에 적절하다.



[그림 4-4] 오픈소스 공개 및 피드백 흐름도
출처: 연구진 작성

오픈소스 기반 유통에는 공개뿐만 아니라 이용자들의 실행 결과에서 발견된 오류 유형과 임계값의 적정성을 확인하여 보고받아 반영하고 개선해나가는 과정이 포함된다. 보고에는 일차적으로 분석 결과인 일치율·결측률·중복률 등 핵심 지표와 함께, 오류의 행정구역별 분포와 유형별 발생 추세 등이 포함될 필요가 있다. 또한, 검증 실패 사유나 개선 오류 건축물 데이터 추적 및 알고리즘 개선을 위한 알고리즘 개선 방안을 제시할 수 있도록 하여 건축물 데이터 품질 개선 알고리즘에 반영할 수 있도록 해야한다.

건축물 데이터 품질 개선 알고리즘의 유통과 운영의 지속가능성을 위해 운영 주체를 명확히 할 필요가 있다. 유지관리자는 규칙과 코드의 최종 승인 권한을 갖고, 외부 아이디어를 반영할 수 있는 프로세스를 구축할 필요가 있다. 제안된 아이디어 중 반영이 필요한 것과 그렇지 않은 것을 구분하여 반영할 수 있도록 심의할 수 있도록 운영해야한다. 알고리즘 내 규칙 변경은 기존 버전과 동기화하여 추적 가능하도록 관리할 필요가 있다. 이러한 방식은 품질이 개선된 건축물 데이터의 배포만으로는 달성하기 어려운 재현성, 투명성 확보를 가능하게 하며, 실무자 피드백에 의한 지속적 고도화를 동시에 확보할 수 있다는 점에서, 지속적인 건축물 데이터 품질 제고가 이루어질 수 있을 것으로 기대한다.

3. 신규 데이터 품질 제고

1) 건축허가·신고-신축 허가 단계

■ 입력시스템 개요

건축허가·신고-신축 허가 단계에서는 크게 기본개요와 동별 및 층별개요를 작성하게 되어있다. 기본개요에서는 '건축주', '설계자', '대지조건' 등 내용을 입력한다. 동별 및 층별개요의 주요 작성 항목은 동별개요 부분과 층별개요 부분으로 구분할 수 있다.

- 동별개요 입력 시스템

동별개요 부분에서는 직접 값을 입력하는 항목과 층별개요 정보 입력 후 자동으로 입력값이 계산되는 네 개의 항목이 구분된다. 직접 값을 입력하는 항목으로는 '주/부구분', '동 명칭 및 번호', '세대/호/가구', '지붕구조', '지붕 마감 재료', '건축면적', '연면적', '용적률산정용 연면적', '높이', '승용/비상용 승강기', '특수구조 건축물'이 있다. 층별개요 정보 입력 후 자동으로 입력값이 계산되는 항목은 '주용도', '주구조', '세부구조', '지하층수/지상층수'이다. 동별개요를 모두 입력한 후 층별개요 항목이 입력 가능하기 때문에 직접 값을 입력하는 항목만 우선적으로 입력하도록 되어있다.

동별개요에서 직접 값을 입력하는 항목은 '주/부구분'과 '동 명칭 및 번호'를 제외하고 모두 직접 숫자를 입력하도록 되어있다. 이 중 '건축면적'은 층별개요에서 포함되지 않는 사항으로 동별개요에서 작성함이 타당하다. 다만, '연면적', '용적률산정용 연면적'은 모두 층별개요에서 작성하도록 되어있는 항목임에도 불구하고 개별적으로 기입하도록 되어있다.

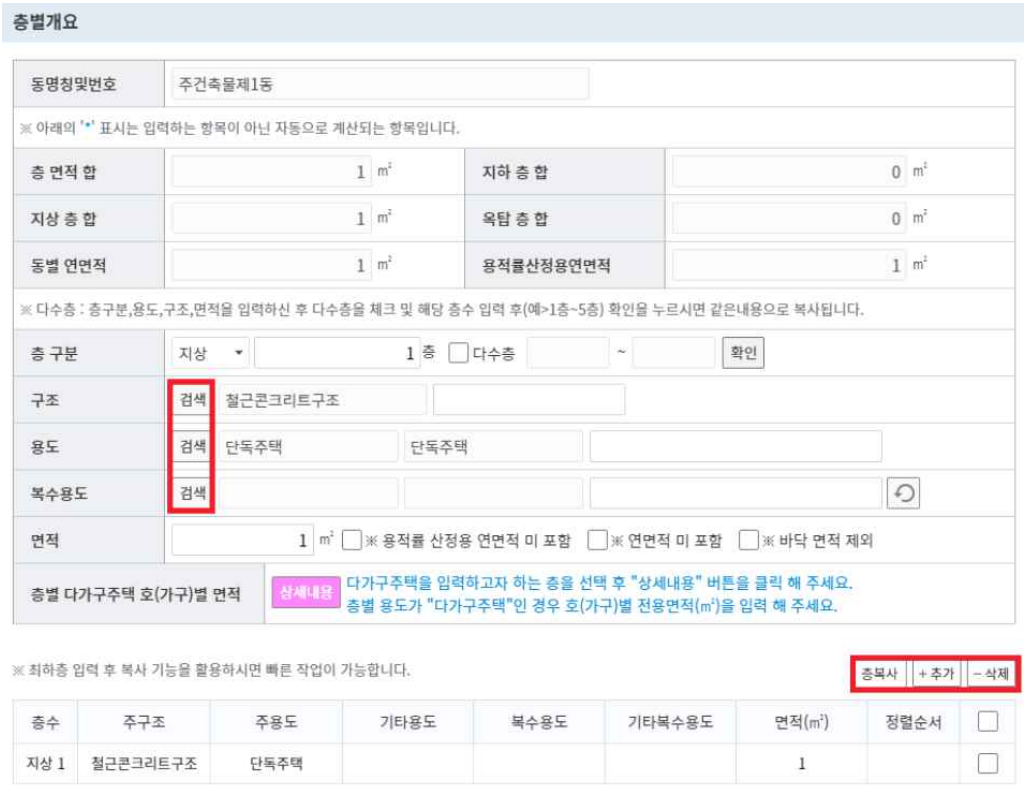
본 연구에서 건축물 대장 중 총괄표제부, 표제부(동), 층별개요를 다루었으나 해당 시스템에는 총괄표제부를 작성하는 창이 없다. 즉, 건축물이 여러 동이 있어 동별개요를 여러 개 작성할 경우 총괄표제부가 생성될 것으로 판단된다. 이 때 총괄표제부에 기입되는 면적 관련 정보들은 동별개요의 값들을 이용해서 자동 계산될 것이나, 본 연구에서 도출한 결과에 따르면 일정부분 차이가 발생하고 있는 것을 확인하였다.

[그림 4-5] 건축허가·신고-신축의 동별개요 작성 탭
출처: 국토교통부. (2022c). 클라우드 기반 건축행정시스템 민원인 매뉴얼(건축인허가 민원). p. 31

- 층별개요 입력 시스템

층별개요 입력 창에서는 '층 구분', '구조', '용도', '복수용도', '면적', '층별 다가구주택 호(가구)별 면적' 항목을 입력하도록 되어있다. 이 중 '구조', '용도', '복수용도' 항목은 검색을 통해 해당하는 구조, 용도를 선택하여 입력할 수 있는 방식을 취하고 있어 오타에 의한 오류는 발생하기 어려울 것으로 판단된다. 다만, 기타용도의 경우 직접 입력할 수 있도록 되어있어 일부 문제가 발생할 여지가 있다. '층구분'과 '면적'의 경우 직접 숫자를 입력하는 형태로 되어있어 오타에 의한 오류가 발생할 가능성이 상대적으로 높다.

본 연구에서는 건축물 데이터 품질 개선을 위해 용도 관련 정보와 면적 관련 정보를 검토하였다. 층별개요에서 '면적' 관련 항목에는 숫자 기입 이외에 용적률 산정용 연면적 미 포함 여부 체크박스, 연면적 미 포함 여부 체크박스, 바닥 면적 제외 여부 체스박스가 있다. 각각은 해당 면적이 용적률 산정, 연면적 산정, 바닥 면적 산정에 포함되는지 여부를 확인하기 위함이다.



[그림 4-6] 건축허가·신고-신축의 층별개요 작성 탭
출처: 국토교통부. (2022c). 클라우드 기반 건축행정시스템 민원인 매뉴얼(건축인허가 민원). p. 32

■ 시스템 개선 방안

면적과 관련된 값 중 연면적과 용적률산정용 연면적은 현재 동별개요와 층별개요에서 모두 직접 입력하도록 되어있다. 다만, 한 동 안에서의 층별개요 면적의 합계는 동별개요의 면적과 일치하여야 한다. 그리고 총괄표제부의 면적은 여러 건축물 동의 동별개요 면적의 합계와 일치하여야 한다. 뿐만아니라 모든 동의 층별개요의 면적 합계는 총괄표제부의 면적과 일치하여야 한다. 그러나 현행 건축물 데이터를 보면 이 같은 면적 값의 불일치가 나타나고 있으며, 이는 향후 시스템 상에서 해결할 수 있는 문제로 판단된다.

첫째로 동별개요 및 총괄표제부 면적의 자동계산 방법을 도입할 수 있다. 즉, 층별개요의 면적값만 잘 입력된다면 동별개요의 면적이 자동으로 계산되고, 자동으로 계산된 동별개요의 면적을 이용하여 총괄표제부의 면적이 자동으로 계산되는 시스템으로 개선하는 방법이 있다. 다만, 이 같은 방법은 수학적인 면적 및 면적 합계는 일치하나, 층별개요의 면적값이 잘못 들어간 경우 동별개요의 면적과 총괄표제부의 면적이 모두 잘못 적용되는 오류를 발생시킬 가능성이 있다.

둘째로 면적값의 합계가 일치하지 않는 경우 시스템이 경고 메세지를 띄우고, 다음 프로세스로의 진행을 막는 방법을 도입할 수 있다. 즉, 항목별 값의 입력 방식은 현 상태를 유지하나 동별개요의 면적과 층별개요의 면적 합계가 일치하지 않는다면 동별개요 창에서 해당 항목을 붉은색으로 표시하고,

에러메세지를 띄우는 방식이다. 이 경우 정보 입력자는 잘못 기입된 항목을 확인하고, 동별개요와 층별개요의 개별 면적 값을 다시 확인하여 합계가 일치할때까지 수정을 거듭하게 된다. 다만 이 같은 방법은 건축물의 동 수나 층 수가 과도한 경우 입력이 잘못된 부분을 확인하는데 많은 시간이 소요될 수 있다는 단점이 있다.

또한 용도와 관련된 항목 입력 방법도 개선이 필요하다. 현행 시스템을 보면 용도 카테고리를 검색해서 입력하는 부분과 기타용도를 직접 타이핑하여 입력하는 방식이 혼용되고 있다. 검색을 통해 입력되는 용도와 직접 타이핑하여 입력하는 기타용도가 다를 수는 있다. 그러나 불가능한 용도가 기입되는 오류가 발생하는 경우가 발생하고 있음을 확인하였다. 예를 들어 검색을 통해 기입되는 값은 다가구주택이나 직접 타이핑하여 기입된 값은 단독주택(10가구)로 기입된 경우이다. 이 같은 오류를 체크하여 전혀 다른 용도가 기입된 경우 경고 메시지를 띄우고 다음 절차로 진행되지 않도록 하는 시스템을 개발할 필요가 있다.

2) 건축허가·신고-사용승인 단계

■ 입력시스템 개요

건축허가·신고-사용승인 단계에서는 크게 기본개요와 동별 및 층별개요를 작성하게 되어있다. 기본개요에서는 '신청구분' 항목에서 사용승인과 임시사용승인 중 하나, 전체와 일부 중 하나, 일반과 집합 중 하나를 선택하도록 되어있다. 동별 및 층별개요의 주요 작성 항목은 동별개요 부분과 층별개요 부분으로 구분할 수 있다.

- 동별개요 입력 시스템

사용승인 단계의 동별개요 입력 시스템은 신축 단계의 동별개요 입력 시스템과 달리 모든 값을 직접 입력하도록 되어있다. 이 중 '주/부구분', '내진설계 적용 여부', '건축협정, 결합건축' 항목은 카테고리를 선택하는 형태로 되어있고, '주용도', '주구조', '지붕구조' 항목은 검색을 통해 해당하는 구조, 용도를 선택하여 입력할 수 있는 방식을 취하고 있어 상대적으로 오류가 발생할 가능성이 낮다. 다만 '세대/호/가구', '건축면적', '연면적', '용적률산정용연면적', '지하층수/지상층수', '높이', '승용 승강기/비상용 승강기', '내진능력' 항목은 직접 값을 입력하는 형식을 취하고 있다.

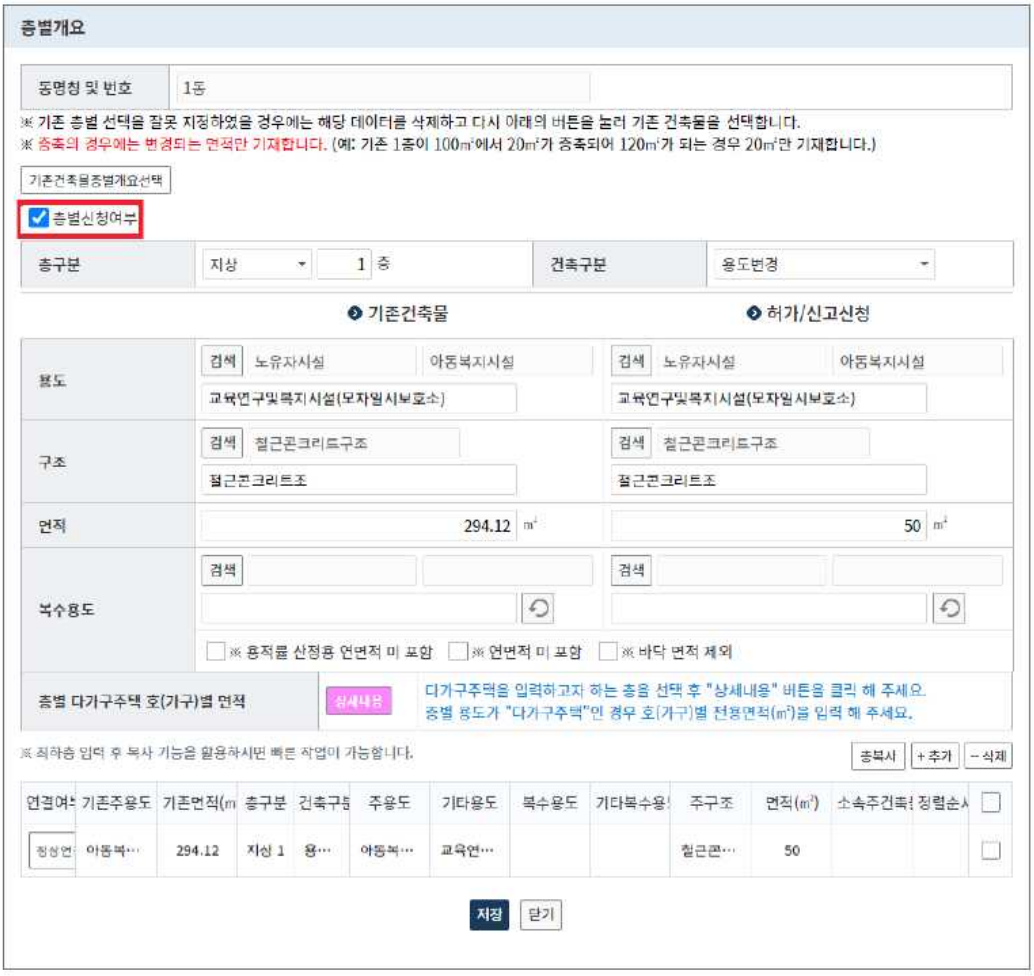
본 연구에서 건축물 데이터 품질 고도화를 위해 주로 검토하였던 '건축면적', '연면적', '용적률산정용연면적' 등 면적관련 항목은 숫자를 직접 입력하는 형식으로 되어있다. 동별개요의 면적 관련 자료 또한 마찬가지로, 개별 동의 층별개요에서 면적 관련 정보를 모두 기입하고 있다. 즉, 층별개요의 면적 관련 값의 합계를 통해 동별개요의 면적 값을 계산할 수 있을 것으로 판단되나 따로 기입하는 방식을 활용하고 있다.

[그림 4-7] 건축허가·신고-사용승인의 동별개요 작성 탭
출처: 국토교통부. (2022c). 클라우드 기반 건축행정시스템 민원인 매뉴얼(건축인허가 민원). p. 54

- 층별개요 입력 시스템

층별개요 입력 시스템에는 '층구분', '건축구분', '용도', '구조', '면적', '복수용도', '층별 다가구주택 호(가구)별 면적' 항목을 입력하도록 되어있다. 이 중 '층구분'의 지상 및 지하 여부와 '건축구분'은 카테고리를 선택하는 항목으로 설정되어있다. '용도', '구조', '복수용도' 항목은 검색을 통해 값을 선택할 수 있도록 하였다. 다만 값을 선택하는 것 이외에도 기타용도에 해당하는 부분은 직접 타이핑하여 입력할 수 있도록 해 놓았다. '면적'과 '층구분'의 층수 항목은 직접 수치를 입력하도록 되어있다.

즉 층별개요 입력 시스템에서 상대적으로 오류 발생 가능성이 적은 항목은 '층구분'의 지상 및 지하 여부와 '건축구분', '용도', '구조', '복수용도'으로 볼 수 있다. 반면 '면적'과 '층구분'의 층수 항목은 숫자를 직접 입력하기 때문에 상대적으로 오류 발생 가능성이 높다고 볼 수 있다.



[그림 4-8] 건축허가·신고-사용승인의 층별개요 작성 탭
출처: 국토교통부. (2022c). 클라우드 기반 건축행정시스템 민원인 매뉴얼(건축인허가 민원). p. 55

■ 시스템 개선 방안

건축 허가 단계와 마찬가지로 사용승인 단계에서도 시스템 개선으로 향후 생산될 건축물 데이터의 오류를 저감시킬 수 있다. 연면적과 용적률 산정용 연면적의 경우, 층별개요와 동별개요 모두 값을 직접 입력하도록 되어 있어 불일치가 발생할 가능성이 높으며, 실제로도 불일치가 발생하는 것을 확인하였다. 정의상 각 동 층별 면적의 총합은 해당 동의 동별개요 내 면적과 일치해야 한다. 또한, 전체 건축물 동의 면적 총합은 총괄표제부 내 면적과 일치해야 한다. 그러나 현행 시스템에서는 이러한 관계가 자동으로 검증되지 않고, 이를 검증할 수 있는 방법을 시스템에 포함시킬 필요가 있다.

첫 번째 방안으로, 동별 및 층별 면적 값을 자동으로 산출하여 입력하는 방법이 있다. 층별 면적을 입력하

면 여러 층의 값을 자동으로 합산하여 동별개요 내 면적에 반영하고, 여러 동의 면적이 자동 집계되어 총괄표제부의 면적으로 연결되도록 설계할 수 있다. 이는 데이터의 일관성을 유지할 수 있도록 도와주며, 행정정보로서의 온전성을 개선하는데에도 도움을 준다. 다만, 이 방식은 값의 비교를 통해 입력된 값에 오류가 있음을 확인할 수는 있지만, 층별개요의 입력값이 부정확할 경우 연쇄적으로 동별 및 총괄표제부의 면적 데이터가 잘못 반영될 수 있다.

두 번째로는, 면적 간 불일치 발생 시 검증 기능을 도입하는 방법이 있다. 예를 들어, 층별 면적 합계가 동별개요 면적과 일치하지 않을 경우, 입력 화면 상에 시각적인 경고와 함께 명확한 오류 메시지를 제공하고, 다음 단계로의 진행을 제한하는 방식이다. 이로 인해 입력자는 각 항목을 재점검하여 오류를 수정할 수 있게 된다. 마지막으로 건축 허가 단계에서 제시한 것처럼 용도와 관련된 항목의 입력 방법 개선이 필요하다.

3) 건물ID를 활용한 건축물 사용승인 시 데이터 검증

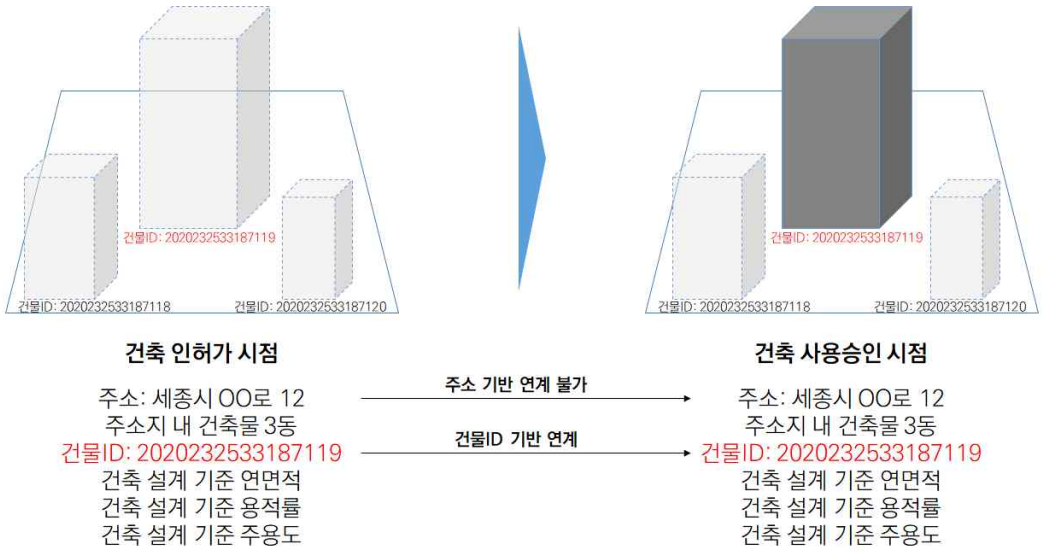
■ 데이터 검증을 위한 건물ID 활용가능성

건축물 데이터 생산 과정에서 건물ID는 건축물 인허가 단계에서 기입된 정보와 건축물 사용승인 단계에서 기입되는 정보가 서로 일치하는지 여부를 확인할 수 있는 연결고리 역할을 할 수 있다. 동일한 건축물은 건축물의 인허가 단계와 사용승인 단계에서 기입하는 정보가 일치할 가능성이 높다. 이 같은 점을 활용하여 건물 ID를 기반으로 건축인허가 정보와 건축물대장의 값이 일치하는지 비교하여 건축물대장 작성 시 잘못 기입되는 정보를 체크할 수 있다. 이를 위한 건물 ID 활용가능성은 건축인허가 데이터와 건축물대장의 건물ID 부여율, 고유성, 일관성 측면을 3장에서 검토하였다.

건축인허가 데이터의 건물ID 부여율은 99.4%로 매우 높았으며, 고유성은 48,060건이 일대일 대응되나 11,408건이 일대다 대응되어 다소 문제가 있는 것으로 나타났다. 일관성 측면에서는 모든 건축물에 대해 하나의 건물에 하나의 건물 ID만 부여되어 일관성 100%로 매우 높게 나타났다.

건축물대장의 건물ID 부여율은 97%로 매우 높았으며, 고유성은 단 7건이 일대다 대응되어 활용가능성이 높게 나타났다. 또한 일관성 측면에서 모든 건축물에 대해 하나의 건물에 하나의 건물 ID만 부여되어 일관성 100%로 매우 높게 나타났다.

전체적으로 볼 때 건물ID를 이용하여 동일 건축물의 건축인허가 데이터와 건축물대장을 연계할 수 있을 것이다. 건축인허가 데이터의 고유성 문제는 향후 개선이 될 것으로 예상되며, 고유성 문제가 건물ID를 이용하여 동일 건축물을 1:1로 연계하는데 문제가 되지 않을 것으로 예상할 수 있다. 특히, 기존의 방식인 주소나 건물명 기반의 건축인허가 데이터 및 건축물대장 연계율이 20% 미만으로 매우 떨어지는 상황에서 건물ID는 건축물 단위 데이터 연계에서 매우 활용성이 높은 고유키로 활용될 수 있다.

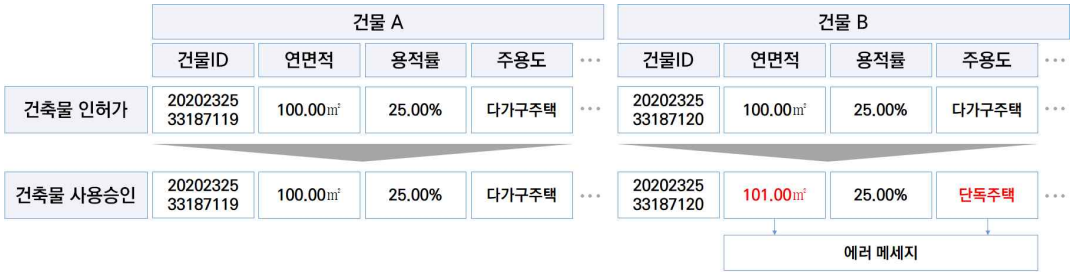


[그림 4-9] 건물ID 기반의 건축인허가 데이터와 건축물대장의 연계
출처: 연구진 작성

■ 건물ID 활용 방안

본 연구에서는 건물ID를 활용하여 건축물 인허가데이터와 건축물 사용승인 시 기입하는 정보가 다른 경우 에러 메세지를 제공하는 방식을 제안하고자 한다. 건물ID를 활용하여 건축물 사용승인시 건축물 인허가 데이터와 일치성 여부를 확인할 수 있으며 이를 건축물 사용승인 시스템에 적용하여 향후 건축물 사용승인을 통해 생산되는 건축물대장의 오류를 저감할 수 있다. 이를 통해 세움터 시스템을 통해 건축물 사용승인 정보를 입력하는 입력자는 건축물 인허가데이터와 다른 값을 기입하였음을 확인하고, 다시 한 번 확인할 수 있는 기회를 제공할 수 있다.

건축물 사용승인시 기입하는 정보가 건축물 인허가 데이터와 다른 경우에도 다음 프로세스로의 진행을 제한하는 방식은 피하고자 한다. 전문가 자문 결과 건축 인허가 중 사용승인 단계에서 건축물 정보가 변경되는 경우가 발생하는 것으로 나타났으며, 이러한 현황을 존중하는 범위 안에서 데이터 무결성을 높이는 방안을 제안하고자 한다.



[그림 4-10] 건물ID 기반 건축물 사용승인 단계 데이터 검증 예시
출처: 연구진 작성

제5장

결론

1. 연구 요약
2. 연구의 정책적 시사점
3. 연구의 한계 및 향후 연구 추진 방향

1. 연구 요약

건축행정 데이터는 범정부 데이터 연계의 핵심이 되는 데이터로, 공공과 민간의 다양한 주체가 생성·관리하는 건축물 관련 데이터의 중심 역할을 한다. 그러나 건축행정 데이터의 오류, 누락, 불일치 등 품질 문제가 지속적으로 제기되고 있는 현실이다. 건축물 현황을 정확하게 파악하기 위해서는 건축물 전수에 대한 현장조사가 필요하나, 이에 앞서 오류 가능성이 높은 조사 대상을 추출하기 위한 방안이 요구된다.

또한, 건축물 데이터의 개방과 연계를 고도화하기 위한 기술적, 제도적 기반 마련이 필요하다. 최근 건축행정 데이터에 도입된 건물ID는 건축물 데이터를 연계하는 공통식별자로서 높은 활용 가치가 기대된다. 건물ID의 도입 현황, 건물ID 도입 후 데이터 품질 등에 대한 검토를 통하여 향후 활용을 위한 기반을 마련하고자 하였다.

본 연구의 목적은 인공지능을 활용한 건축물 데이터의 품질 개선방향을 모색하고, 건축물 데이터 품질 고도화를 위한 정책화 등 적용 방안을 마련하는 것이다. 연구 대상인 건축물 데이터는 건축행정 데이터 중 건물ID가 적용된 건축물대장, 건축인허가, 주택인허가 데이터로 설정하였다. 각 데이터는 대장-테이블-컬럼 위계에 따라 여러 테이블로 구성되는데, 건물동 단위 데이터 분석을 기본으로 하여 각 분석에서 필요한 데이터를 연계하여 분석 대상 데이터를 구축하였다. 품질 고도화를 위한 구체적인 건축물 데이터의 범위는 건축물 데이터의 품질 현황과 활용도를 고려하여 주요 테이블 및 컬럼 항목을 도출하였다. 방법론적으로는 문헌고찰, 법·제도 분석, 논리적 오류 검증, 통계 및 기계학습을 통한 이상값 검출, 건축 및 인공지능 분야 전문가 자문 등을 수행하였다.

본 연구는 크게 세 단계로 진행되었다. 2장에서는 건축물 데이터와 관련된 건축법, 공공데이터법 등 제도를 분석하고, 건축통계 및 관련 연구에서 건축물 데이터의 활용 현황을 분석하였다. 이를 통하여 건축물 데이터 중 활용도가 높은 항목을 도출하고 동시에 선행연구에서 보고한 오류 현황 등을 함께 고려하여 면적 관련 수치 데이터와 용도 관련 자연어 데이터의 품질 고도화 방향을 제시하였다. 또한 건물ID 기반 건축물대장과 인허가 데이터의 연계 시범적용을 통한 품질 평가 및 향후 활용 제고 방향도 제시하였다.

3장에서는 앞에서 제시된 품질 고도화 방향의 시범적용을 통하여 품질 현황을 파악하고 방법론을 심화하였다. 건축물대장 면적 데이터 분석 결과, 기존 건축물대장 정비 규칙과 본 연구에서 개발한 신규

규칙 모두에서 0.13%~5.01%의 오류율이 나타나 건축물대장 정비를 통한 오류 정정이 신속하게 이루어지고 있지 못한 현황을 확인할 수 있었다. 시기별 및 지역별 분석 결과 건축면적 오류의 경우 최근까지 8~10%의 오류율이 나타나는 등, 신규 건축물 데이터에 대한 품질 고도화 필요성이 확인되었다. Isolation Forest와 One-Class SVM 등 기계학습 기반 이상값 탐지 분석에서는 규칙 기반 검증과 다른 패턴을 보이는 비정형적 오류가 도출되었다. 시기별로 특정 시기에 집중되지 않고 고른 분포를 보이며, 지역별로는 서울, 부산 순으로 높은 오류율을 보였다. 이러한 결과는 규칙 기반 검증과 기계학습 기반 탐지를 병행하여 다양한 오류를 검출하는 상호보완적 활용이 필요함을 보여주었다.

건축물대장 용도 데이터 분석 결과, 용도 분류 코드('주 용도 코드')와 대장 기재내용('기타 용도') 사이 불일치 사례가 상당수 확인 되었다. 단어 빈도 분석과 나이브 베이즈 모델을 적용한 기계학습 기반 예측 분석 결과 모두에서 기재내용에 따른 분류 결과와 데이터에 별도 항목으로 기재된 분류 코드가 일치하지 않는 경우가 나타났다. 기재내용에서 분류 코드를 예측하는 모델 구축 결과, 동별 용도 예측은 84.78%, 층별 용도 예측은 78.97% 정확도를 달성하였다. 다만 예측 실패 중 일부는 실제 오류로 인한 것이었으나, 사례가 적은 용도에서는 정확도가 낮게 나타나는 등 빈도 기반 분석의 한계도 드러났다. 향후 품질 고도화를 위하여 건축물 전체 맥락을 고려한 복잡성을 반영할 수 있는 방법론 도입이 필요함을 확인하였다.

건물ID 기반 건축물대장과 인허가 데이터를 연계하여 분석한 결과, 건물ID는 2024년 사용승인 건축물의 97%에 부여되고, 건축인허가 데이터와 93.6%, 주택인허가와는 34.9% 연계 성공률을 보였다. 주택인허가는 연계 성공한 경우에도 면적 변수값의 일치율이 60%대 후반~70%대 초반으로 낮게 나타났다. 건물ID를 중심으로 한 건축물대장과 인허가 데이터 간의 품질 관리 필요성을 확인하였다. 특히 건물ID 부여의 고유성, 일관성, 정합성 등 건물ID 데이터 자체 품질 고도화가 우선될 필요가 있다.

건축물 데이터는 세움터에서 생산·관리하고 있으며, 건축물 데이터의 품질 제고를 위한 프로세스도 세움터를 기준으로 두 가지로 구분할 수 있다. 첫째는 이미 생산된 건축물 데이터의 품질을 개선하여 유통하는 것이며, 이때 기존 원본 데이터를 정정하는 대신 품질 개선 데이터를 세움터 외부에 저장하고 오픈소스로 유통하는 방식을 제안하였다.

둘째는 향후 생산될 데이터의 품질을 개선하는 것이다. 이를 위해서 세움터 시스템 내 건축물 허가 및 사용승인 위한 입력 시스템의 개선을 통하여 오류 발생을 최소화하는 방안을 제안하였다. 면적 및 용도 데이터 입력 시스템을 검토하여 시각적 경고를 동반하는 검증 기능 도입을 제시하였다. 특히 건물ID를 적극적으로 활용하여 세움터 데이터 검증 기능에서 연계키로 사용하는 방안도 제안하였다.

2. 연구의 정책적 시사점

■ 건축물대장 품질정비 현황 파악

세운터에서는 자체적으로 86개 업무규칙을 선정하여 건축물대장 상 오류를 검출하고 있으며, 그 중 주요 항목에 오류가 발생한 경우 오류가 있는 건축물대장 목록을 지자체로 보내고, 지자체에서 건축물대장의 오류를 정비하도록 지시하고 있다. 본 연구에서는 전국단위에서 건축물대장의 품질정비가 얼마나 진척되었고 어느 정도의 품질정비 필요 항목이 남아있는지 확인하였다. 뿐만 아니라 오류가 있는 건축물대장의 효율적인 정비 방향을 확인하기 위해 광역지자체별, 건축물의 사용승인 연도별 오류율을 확인하였다.

이를 통해 품질정비가 시급한 광역지자체를 도출하였고, 오류율이 높은 건축물대장 상 사용승인일을 확인함으로써 오류항목별 오류가 발생하는 경향을 파악할 수 있었다. 이는 단순히 오류발생 경향에 멈추지 않고 오류가 발생하는 원인을 추정할 수 있는 근거가 될 수 있음을 확인하였다. 예를 들어 건축물의 준공 및 사용승인이 집중적으로 발생한 1990년대 중반에 건폐율, 용적률, 용적률 산정 연면적 오류가 증가함을 확인할 수 있었다. 즉, 과도한 행정처리 업무량을 오류 발생의 원인 중 하나로 볼 수 있으며, 이 같은 오류가 많이 발생하는 지자체에 건축물 인허가 및 사용승인 담당 공무원 추가가 필요함을 시사한다.

■ 건축물대장 품질정비 대상 추가 제안

본 연구에서는 세움터에서 제안한 기존 4가지 면적 관련 오류 탐지 업무규칙에 더하여 3가지 오류 탐지 방법을 제안하였다. 세부적으로는 두 개의 신규규칙을 통해 용적률 및 연면적과 관련된 오류를 검출할 수 있으며, 기계학습을 통한 이상값 탐지 알고리즘을 통해 면적과 관련된 변수의 오류가 있을 가능성이 높은 건축물대장을 도출할 수 있다. 즉, 신규 업무규칙 두 개를 제안함으로써 기존에 오류 탐지 업무규칙에서 도출된 문제가 있는 건축물 대장에 더하여, 품질정비가 필요한 약 8만개의 건축물 대장을 제시하였다.

뿐만 아니라, 기계학습 기반의 이상값 탐지 알고리즘을 통해 기존 및 신규 업무규칙을 통해 검출되지 않을 수 있는 새로운 형태의 오류를 검출할 수 있는 기반을 마련하였다. 기계학습을 통해 도출되는 오

류 유형이 규칙 기반 검증에서 검출될 수 없는 비정형 오류를 제시하기 때문에, 기존에는 전혀 발견할 수 없었던 오류가 있는 건축물 대장을 도출할 수 있는 방법론을 본 연구를 통해 제시하였다.

■ 건축물 데이터 품질제고를 위한 방향성 도출

기존 세움터에서 제시하고 있는 건축물 대장 품질정비 방안은 법적·물리적으로 발생이 불가능한 오류를 도출하여 정비하는 수준이었다. 그러나 건축물 관련 데이터에는 법적·물리적으로 발생 가능하나 현실의 건축물과는 다른 값이 기입되어있는 건축물이 있다. 본 연구에서는 건축물 데이터의 활용성 제고를 위해 전방위적인 건축물 데이터의 품질 제고가 필요하다고 판단하였으며, 이를 위해 기계학습 모형 활용 방안을 제안하고 시범 적용해 보았다. 이를 통해 기존에는 발견할 수 없었던 오류의 추가적인 검출이 가능해질 것으로 확인하였으며, 기계학습과 같은 인공지능 활용이 건축물 데이터 품질 제고를 위해 향후 나아갈 방향 중 하나임을 확인하였다.

기존 건축물 대장 품질 제고는 건축물 대장 자체의 품질 제고를 목적으로 하여 직권정정 관련 문제, 오류 정정을 위해 들어가는 시간상의 문제가 발생하고 있다. 이는 건축물 대장에 기인한 데이터의 활용 시 다양한 문제점을 야기함을 선행연구를 통해 확인할 수 있었다. 그러나 본 연구에서는 건축물 대장이 아닌 통계나 정책연구를 위한 건축물 데이터의 품질 제고 방안을 제안하였기 때문에 데이터의 활용성 측면에서 더 좋은 성과를 도출할 것으로 기대된다. 특히, 건축물 데이터 품질 제고 및 유통 뿐 아니라 오픈소스를 활용한 건축물 데이터 품질 제고 모델 및 알고리즘의 개선까지 제시함으로써 향후 건축물 데이터의 품질제고를 위한 더 효율적인 방향을 제시하였다는 점에서 시사점이 있다.

본 연구에서 제시한 건축물 데이터 품질제고 방안을 통해 향후 건축물 데이터를 활용한 통계나 정책연구의 품질 또한 개선될 것으로 판단된다. 특히, 기존 건축물 데이터를 활용한 통계나 정책연구의 대부분이 면적이나 용도 데이터를 활용하고 있으며, 본 연구에서 주로 다룬 항목이 면적이나 용도 관련 내용임을 고려하면 연구의 파급효과 또한 높은 것으로 판단된다.

■ 건물ID 기반 건축물 데이터 정책의 효용 확인

건물ID는 건축행정정보의 연계를 위하여 최근 도입되었다. 본 연구에서는 건물ID를 건축물 데이터 품질제고를 위해서도 활용할 수 있음을 제안하였다. 건물ID가 건물의 인허가시점부터 사용승인, 건축물 대장의 폐말소 시점까지 고유한 값으로 연계되기 때문에 다양한 건축물 항목의 추적이 가능하기 때문이다.

그러나 건물ID가 데이터 연계를 위한 고유한 키로 활용되기 위한 다양한 조건들을 충족시킬 필요가 있음을 확인하였다. 연구를 통해 아직까지도 건물ID가 모든 건축물 대장에 기입되지 않았을 가능성 또한 확인하였으며, 일부 건축물의 경우 건물 ID의 연계가 완벽하지는 않을 수 있다는 것 또한 확인하였다. 즉, 본 연구에서 확인한 건물ID 활용방안을 도입하는 것도 중요하지만, 건물ID의 활용성 제고를 위해 개선할 사항은 여전히 남아있다.

3. 연구의 한계 및 향후 연구 추진 방향

본 연구에서 제시한 건축물 데이터 품질 제고 방법론은 기계학습 등 인공지능 방법론을 활용하여 기존 규칙 기반 검증으로 탐지하기 어려운 비정형 오류를 탐지할 수 있다는 의의가 있다. 그러나 한편으로는 탐지된 이상값이 실제 오류인지 확인하는 과정이 필요한 한계가 있음이 확인되었다. 특히 통계적 이상치이지만 정상인 값을 오류로 잘못 분류할 가능성이 존재하였다.

본 연구에서 제안한 인공지능 기반 방법론의 단점을 보완하기 위하여 후속 과제를 다음 세 가지 방향으로 제시한다. 첫째, 규칙 기반 검증과 인공지능 기반 방법론을 상호 보완적으로 활용할 필요가 있다. 1차적으로 규칙 기반 검증을 적용하여 정제된 데이터에 2차로 기계학습을 적용해 숨겨진 이상값을 탐지하는 단계별 접근을 검토할 수 있다. 규칙 기반 검증은 법적, 명시적 오류를 인간이 이해하기 쉬운 방식으로 관리하는 데 활용하고, 새로운 오류 사례와 그에 따른 새로운 규칙 발굴을 위하여 인공지능 기반 방법론을 적용하는 것이다. 인공지능 기반 방법론이 탐지한 오류는 오탐지 가능성을 배제할 수 없으므로 허가권자의 검증이 필요하며, 이는 본 연구에서 제시한 바와 같이 신규 데이터 생성 또는 기존 데이터의 정비 단계에서 이루어질 수 있다. 새로운 오류 사례 검토와 이를 규칙 기반으로 검증하기 위한 새로운 규칙 개발은 연구자 등 제3자가 수행할 수 있는 영역이며, 특정 데이터의 오류 특성에 집중하는 후속 연구를 통하여 규칙 기반 검증 방법론도 고도화할 수 있을 것으로 기대한다.

둘째, 학습 데이터 구축을 통하여 인공지능 기반 품질 제고 방법론을 지도학습 방식으로 전환하는 것이다. 본 연구에서 제시한 비지도 기계학습 기반 방법론은 전문가에 의하여 오류 여부가 분류된(라벨링된) 학습용 데이터가 없기 때문에 그러한 라벨링 없이도 수행할 수 있는 이상탐지 기법에 한정되었다는 한계가 있다. 오류 여부가 라벨링된 학습 데이터를 구축하고 이를 기반으로 건축물 데이터의 품질 제고 방법론을 지도학습 기반 방식으로 전환하여 오류 탐지 정확성을 높일 수 있다. 학습 데이터 구축을 위한 한 방법으로는 본 연구에서 제안한 비지도 학습 기반 모델의 분석 결과를 전문가가 검토 후 라벨링하여 학습 데이터로 축적하는 연구가 이루어질 수 있다.

셋째, 다양한 인공지능 모델을 접목하여 인공지능 방법론 자체를 고도화하는 연구이다. 설명가능 AI(SHAP, LIME 등)의 도입, 새로운 검증 규칙 후보를 자동으로 제안하는 지능형 규칙 추천, LLM을 활용하여 법제 텍스트에서 검증 규칙을 자동으로 제안하는 시스템 등 다양한 연구 방향이 전문가 자문

결과 도출되었다. 또한, 누락 및 오류 데이터를 추정하여 대체하기 위한 정상값 예측 모델의 개발에 대한 관심도 높게 나타났다.

한편, 이러한 후속 연구가 수행되기 위해서는 원데이터에 해당하는 건축행정 데이터의 개방 수준도 높아져야 할 것이다. 현재 건물ID는 세움터 내부 데이터에 포함되어 있고, 건축물대장 발급 시 표출되고 있으나, 건축허브를 통한 대용량 원시 데이터 제공 방식으로는 제공되지 않고 있다. 건축물대장의 건물ID 부여 현황 데이터가 민간에 개방되어야 공공 및 민간에서 생산되는 다양한 건축물 관련 데이터에서도 건물ID를 표준 연계키로 활용할 수 있을 것이다.

참고문헌

강필성. (2022). 설명 가능한 멀티모달 이상치 탐지 방법론 개발 및 산업 데이터 응용. DSBA. 1월 13일. http://dsba.korea.ac.kr/%ec%84%a4%eb%aa%85-%ea%b0%80%eb%8a%a5%ed%95%9c-%eb%a9%80%ed%8b%b0%eb%aa%a8%eb%8b%ac-%ec%9d%b4%ec%83%81%ec%b9%98-%ed%83%90%ec%a7%80-%eb%b0%a9%eb%b2%95%eb%a1%a0-%ea%b0%9c%eb%b0%9c-%eb%b0%8f-%ec%82%b0/ (검색일: 2024.6.4.)

강황식. (2006). 건교부, 건축물대장 1668만건 일제 정비. 한국경제. 4월 2일 기사. https://www.hankyung.com/article/2004112548771 (검색일: 2024.8.2.)

건축법. 법률 제20424호, 2024. 3. 26., 일부개정.

건축법 시행규칙. 국토교통부령 제1416호, 2024. 12. 16., 일부개정.

건축물대장의 기재 및 관리 등에 관한 규칙. 국토교통부령 제1235호. 2023. 8. 1., 일부개정.

건축행정시스템 세움터, https://www.eais.go.kr/.

건축행정시스템 운영규정. 국토교통부훈령 제1369호, 2021. 2. 18., 일부개정

건축HUB, https://www.hub.go.kr/portal/main.do.

공공데이터포털, https://www.data.go.kr/.

공공데이터 관리지침. 행정안전부고시 제2021-70호, 2021. 10. 26., 전부개정

공공데이터의 제공 및 이용 활성화에 관한 법률. 법률 제19408호, 2023. 5. 16., 타법개정.

공공데이터의 제공 및 이용 활성화에 관한 법률 시행령. 대통령령 제33842호, 2023. 11. 7., 일부개정.

국토교통부. (2022a). '22년 시·도별 건축물대장 정비계획 수립 및 정비 요청. https://www.open.go.kr/othicInfo/infoList/infoListDetl2.do (검색일: 2025.1.23.)

국토교통부. (2022b). 건축허가 및 착공통계 통계정보보고서.

국토교통부, (2022c). 클라우드 기반 건축행정시스템 민원인 매뉴얼.

국토교통부. (2023a). 건축물대장의 기재 및 관리 등에 관한 규칙 일부개정안 입법예고. 법제처. https://www.moleg.go.kr (검색일: 2025.1.16.)

국토교통부. (2023b). 전체민원. 건축행정시스템 세움터. https://www.eais.go.kr/moct/awp/ada08/AWPADA08L02 (검색일: 2023.10.4.)

국토교통부. (2024a). 건축데이터 개방서비스 관련 PK전환 규칙 안내. 건축HUB. https://www.hub.go.kr/portal/bbs/ntc/idx-ntc-detail.do (검색일: 2025.1.13.)

국토교통부. (2024b). 건축물통계 통계정보보고서. https://www.k-stat.go.kr/metasvc/msba100/statsdcdta?statsConfmNo=116011&kosisYn=Y

국토교통부. (2024c). 건축허가·착공·준공통계 통계정보보고서.
https://www.k-stat.go.kr/metasvc/msba100/statsdcdta?statsConfmNo=116005&kosisYn=Y

국토교통부. (2024d). 국토교통부_건축HUB_건축물대장정보 서비스. 공공데이터포털.
https://www.data.go.kr/data/15134735/openapi.do#/API%20%EB%AA%A9%EB%A1%9D/getBrBasisOulnInfo (검색일: 2025.1.14.)

국토교통부. (2025a). 대용량 제공 서비스. 건축HUB.
https://www.hub.go.kr/portal/opn/lps/idx-lgcpt-pvsn-srvc-list.do (검색일: 2025.2.7.)

국토교통부. (2025b). 건축통계 정확도 향상 위해 집계방식 개선. 1월 24일 보도자료.

국토교통 통계누리, https://stat.molit.go.kr/portal/main/portalMain.do.

국토지리정보원. (2024). 국토정보플랫폼 국토정보맵.
https://map.ngii.go.kr/ms/map/NlipMap.do?tabGb=statsMap (검색일: 2024.6.5.)

김보미. (2023). 5200여 반지하 주택 '전수조사' 성동구…장마 전 예방 위해 3개월 만에 끝내. 경향신문. 2월 17일 기사. https://www.khan.co.kr/national/national-general/article/202302171820001 (검색일: 2023.3.2.)

김승범. (2015). 건축물대장 원시데이터의 관계 구조와 탐색적 분석을 통한 데이터 활용.
한국문화공간건축학회, 50, 110–120.

나라경제 2022년 7월호. (2022). KDI 경제정보센터.

박진우. (2024). 마포 공덕자이, 9년 만에 이전고시…주담대·전세대출 길 열렸다. 한국경제. 12월 19일 기사.
https://www.hankyung.com/article/202412196833i (검색일: 2025.1.14.)

송유미, 조영진, 안의순. (2024). 건축행정 데이터를 활용한 건축물 연령 지표 개발 연구. 건축공간연구원.

안의순, 박종훈. (2023). 건축행정 데이터 기반 재해 취약 지하층 주택 현황 분석. 건축공간연구원.

양현채, 박기범, 엄미정, 목은지. (2020). 과학기술 행정 혁신을 위한 인공지능 활용 방안.
https://www.stepi.re.kr/site/stepiko/report/View.do?cbIdx=1292&reIdx=973&cateCont=A0201 (검색일: 2024.6.3.)

여인규. (2021). 국토부, 전국 3차원 건물모형 서비스 개시. 칸. 7월 4일 기사.
https://www.kharn.kr/news/article.html?no=16649 (검색일: 2024.11.21.)

오미애, 박아연, 김용대, 진재현. (2018). 기계학습(Machine Learning) 기반 이상 탐지(Anomaly Detection) 기법 연구 - 보건사회 분야를 중심으로. 한국보건사회연구원.
https://repository.kihasa.re.kr/handle/201002/32608 (검색일: 2024.6.26.)

이민철. (2024). Kiwi: 통계적 언어 모델과 skip-bigram을 이용한 한국어 형태소 분석기 구현. 디지털인문학, 1(1), 109–136. https://doi.org/10.23287/KJDH.2024.1.1.6

이선재. (2024). 필지별 건축물 이력 작성: 건축물대장/폐쇄말소대장 결합 [발표자료]. 서울대학교 건축도시공간연구실 내부자료.

이선재. (2025). 건축물데이터 품질개선 프로젝트 [계획]. 서울대학교 건축도시공간연구실 내부자료.

이영주. (2022). 디지털플랫폼정부 시대, 데이터로 국토 변화를 읽고 지속가능한 미래를 준비. 국토, 55–59.

장요한, 이영민. (2022). 국토조사를 위한 국가기본도 도로망도 기반 접근성 보정방안에 관한 연구 [발표자료]. 대한교통학회 제87회 학술발표회. 9월 30일.

장요한, 이영민. (2022). 국토조사를 위한 국가기본도 도로망도 기반 접근성 보정방안에 관한 연구.
대한교통학회 학술대회지, 87, 211–212. (2022.9.28–30.).

장요한, 이영민, 임시영, 백정한, 김현경. (2022). 국가기본도 도로망도 기반 국토조사 접근성 지표 생산방안 연구. 국토지리정보원.
https://www.ngii.go.kr/kor/contents/view.do?sq=1345&board_code=contents_data (검색일:

2025.1.22.)

전라남도 해남군. (2022). 2022년 건축물대장 정비 추진계획(안). https://www.open.go.kr/othicInfo/infoList/infoListDetl.do?prdnNstRgstNo=DCT2B6206AA1831BFB03D4FFFFBA3E3C5CC&prdnDt=20220908174212&nstSeCd=B&title=%EC%9B%90%EB%AC%B8%EC%A0%95%EB%B3%B4&kwd=%EA%B1%B4%EC%B6%95%EB%AC%BC%EB%8C%80%EC%9E%A5+%EC%A0%95%EB%B9%84+%EA%B3%84%ED%9A%8D&searchInsttCdNmPop=&preKwds=%EA%B1%B4%EC%B6%95%EB%AC%BC%EB%8C%80%EC%9E%A5+%EC%A0%95%EB%B9%84+%EA%B3%84%ED%9A%8D&reSrchFlag=off&insttSeCd=&eduYn=N&startDate=20220601&endDate=20230531&pSelt=&insttCdNm=&insttCd=&searchMainYn=&rowPage=10&viewPage=1&sort=s&offSet=4&prevUrl=%2FothicInfo%2FinfoList%2ForginlInfoList.do&othbcSeCd=&hash=true (검색일: 2025.1.23.)

전창해. (2021). "일제잔재 청산" 충북도 공적 장부의 일본식 이름 지운다. 연합뉴스. 3월 1일 기사. https://www.yna.co.kr/view/AKR20210226133900064 (검색일: 2025.1.14.)

정동훈, 김진, 배상근, 이길재. (2014a). 건축물 정위치 등록에 관한 연구: 지적도상 건축물 등록에 관한 연구. 국가건축정책위원회. https://www.riss.kr/link?id=G3808455&ssoSkipYN=Y (검색일: 2024.6.3.)

정동훈, 김진, 배상근, 이길재. (2014b). 건축물 정위치 등록에 관한 연구: 지적도상 건축물 등록에 관한 연구. 국가건축정책위원회. https://www.codil.or.kr/viewDtlConRpt.do?gubun=rpt&pMetaCode=OTKCRK180224 (검색일: 2024.6.3.)

정용찬, 노희용, 김윤화, 오윤석, 박민규, 이동희, 정경오. (2023). 국가통계 관리체계 효율화 방안 연구. 통계청. https://www.kisdi.re.kr/report/view.do?key=m2101113024770&masterId=3934580&arrMasterId=3934580&artId=1779956 (검색일: 2025.1.22.)

조상규, 성은영. (2012). 건축행정정보의 정책적 활용 및 건축통계 개선방안 연구. 건축공간연구원.

조상규, 조영진, 송유미. (2019). 건축행정 빅데이터의 효율적 활용을 위한 정보체계 개선 연구. 건축공간연구원.

조상규, 김신성. (2023). 대규모 언어모델(LLM)을 활용한 건축민원 대응 효율화 방안 연구. 건축공간연구원.

조영진, 류수연, 현태환. (2022). 건축행정 통계 개선 및 공간정보 융합 방안 연구. 건축공간연구원.

조영진, 허한결, 안의순, 류수연, 송유미, 현태환. (2022). 빅데이터 기반 건축물 화재 예측 모델 개발 연구. 건축공간연구원.

조영진, 허한결, 안의순, 류수연, 현태환. (2022). 빅데이터 기반 건축물 화재 예측 모델 개발 연구. 건축공간연구원

조영진, 허한결, 안의순, 송유미. (2023). 데이터 기반 정책을 위한 건축물 생산량 지수 개발 연구. 건축공간연구원.

조영진, 허한결, 송유미, 현태환. (2023). 빅데이터 기반 건축물 화재 및 홍수 리스크 분석 모델 개발 연구. 건축공간연구원.

조영진, 유광흠, 박종훈, 안의순, 허한결, 현태환, 송유미, 김효정, 남기천, 김가해, 박미래. (2024). 2024년 건축물관리지원센터 업무 위탁. 국토교통부·건축공간연구원.

조영진, 안의순. 박성남, 고영호, 권오규, 임보영, 임리사, 김유진, 이정현, (2024). 범죄예방 환경설계(CPTED) 고도화 및 인증제도 개선 방향. 건축공간연구원.

지능정보화 기본법. 법률 제20410호, 2024. 3. 26., 일부개정.

최병남. (2023). 디지털플랫폼정부의 플랫폼, 디지털국토. 국토, 6-14.

최준호, 김인한. (2014). 개방형 BIM기반의 건축인허가 적법성검토체계 구축을 위한 사전프로세스 적용 방안에 관한 연구. 대한건축학회 논문집 - 계획계, 30(9), 3-12.

최훈. (2022). 스마트 주소정보 플랫폼 구축해 AI 로봇 배송 등 혁신서비스 창출한다. 나라경제, 33(7). https://eiec.kdi.re.kr/publish/naraView.do?fcode=000020000400001000 05&cidx=13892&sel_year=2022&sel_month=11&pp=20&pg=1 (검색일: 2024.6.5.)

행정안전부. (2021). 공공데이터 제공·관리 실무 매뉴얼.

Chatbot Arena LLM Leaderboard. (2025). Chatbot Arena (formerly LMSYS). https://lmarena.ai/?leaderboard (검색일: 2025.2.7.)

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H.,Zhu, B., Jordan, M., Gonzalez, J. E., Stoica, I. (2024). Chatbot arena: An open platform for evaluating llms by human preference.

Craig, L. (2023). How data quality shapes machine learning and AI outcomes. Enterprise AI. https://www.techtarget.com/searchenterpriseai/feature/How-data-quality-shapes-machine-learning-and-AI-outcomes (검색일: 2024.6.5.)

Geng, Y., Chen, J., Ye, Z., Yuan, Z., Zhang, W., Chen, H. (2021). Explainable zero-shot learning via attentive graph convolutional network and knowledge graphs. Semantic Web, 12(5), 741–765. https://doi.org/10.3233/SW-210435

Swalin, A. (2018). How to Make Your Machine Learning Models Robust to Outliers. KDNuggets. 8월 28일. https://www.kdnuggets.com/how-to-make-your-machine-learning-models-robust-to-outliers (검색일: 2024.7.2.)

Summary

Enhancing Building Data Quality Using Artificial Intelligence

Ahn, Euisoon Heo, Hankyul Nam, Kicheon Kang, Bumjoon Lee, Sunjae Park, Dongjoon

Introduction

Buildings are fundamental spatial units that form the basis of policymaking, so having an accurate picture of the nation's building stock is a critical task. However, data related to buildings produced and managed by the government are limited in both quality and interoperability. The building register—the representative administrative dataset for buildings—is an official ledger that records current building information and serves as the core foundation for inter-ministerial data integration on buildings. Nevertheless, previous studies have repeatedly pointed out issues of errors, omissions, and inconsistencies in these datasets.

A complete on-site survey of all buildings would be the only way to fully ascertain the actual building status, but such an approach would be prohibitively costly and impractical. Therefore, this study aimed to verify building data errors and identify data quality issues such as omissions, outliers, and inconsistencies by applying artificial intelligence (AI) methodologies that minimize human intervention. Furthermore, to strengthen the technical and institutional foundation for the open access of building data, the study explored data linkage methods based on "Building ID"—a unique building-level identifier introduced in the "Seumter," or the e-AIS, the electronic Architectural Information System.

Accordingly, this study explored AI-based strategies to improve the quality of building data and proposed practical applications for quality enhancement. It also suggested measures to distribute quality-enhanced datasets and improve the quality of newly generated data during the building permit process.

Directions for Enhancing Building Data Quality

■ Current state of legislation and regulations regarding building data

Under the Building Act, building permit processes must be handled via an electronic information processing system (Article 32), and permit authorities must maintain and update the Building Register to ensure accurate building information (Article 38). Accordingly, the Seumter system has been in operation.

Also under the Act on Promotion of the Provision and Use of Public Data, public institutions are required to provide the data they manage so that they may be utilized by the public (Article 3). Unprocessed administrative building data managed through Seumter are provided accordingly via the Building Hub platform. However, this data, along with the original register, lacks continuity and consistency.

In addition, the primary key (PK) structure has changed since the transition from the previous Open Building Data System to the new Building Hub in early 2025, making it difficult to link data across the time series.

■ Quality and Utilization of Building Data

The Ministry of Land, Infrastructure and Transport (MOLIT) and local governments have been continuously refining the building register. In 2022, Seumter defined 86 rules for inspecting the building register along with building/housing permit data, and reviewed their error rates. Out of 620 million total records, about 8.55 million contained errors, corresponding to an overall error rate of 1.37%. However, some rules showed much higher error rates between 10% and 30%.

Building data are widely used for statistical and research purposes. Current construction statistics are automatically generated through the administrative system. In these datasets, key variables include district codes, floor area, ownership type, usage codes, and floor counts, while permit-related datasets use district codes, usage and structure codes, floor area, type of permit, number of buildings, and key approval dates.

A review of prior research revealed common data errors, including invalid main-use codes, missing building names or main-building counts, and unrealistic outliers in total floor area.

■ Directions for Building Data Quality Enhancement

Based on the review of current quality and utilization, this study proposes strategies to enhance data on building area and usage, as well as linkage quality between datasets.

The building area data is a numerical variable central to statistics and research. Since the building register shows the highest error rate in rules on area data, it was prioritized for review. Four rules from the Seumter inspection (site area, building area, building coverage ratio, and floor area ratio) were verified again as of now and analyzed by region and year of completion. In addition, new verification rules were developed to supplement the existing framework.

To go beyond rule-based checks, AI-based anomaly detection techniques such as Isolation Forest and One-Class SVM were employed. These were applied complementarily to detect potential errors that cannot be captured through the rule-based method.

The building usage data were also analyzed, as they are widely used in both statistics and research. To ensure consistency between free-text "other use" entries and coded "main use" classifications, a Naive Bayes classifier was trained to predict main use codes from textual descriptions.

Finally, the study examined data linkage quality using Building ID by connecting building registers with building and housing permit datasets, and analyzing linkage rates, data consistency, and matching success.

Pilot Application of Building Data Quality Enhancement

■ Building Register Area Data

A comprehensive diagnostic process combining existing rule-based validation, newly developed rules, and AI-based anomaly detection revealed an error rate ranging from 0.13% to 5.01%.

New validation rules were added for cases where the floor area used for floor area ratio (FAR) calculation exceeded the total floor area (0.26%), and where the total floor area exceeded the derived upper threshold (0.86%).

For anomaly detection, dimensionless indicators were created—such as the ratio of building area to site area, the ratio of FAR floor area to total floor area, and floor occupancy ratios. Results showed temporal stability but different regional variation pattern from the existing and new rules: higher anomaly rates were found in Seoul and Busan, suggesting that the AI approach successfully detected previously unidentifiable error patterns.

Building Register Usage Data

The usage data were systematically analyzed for quality issues. Frequency analysis revealed mismatches between frequently occurring words and official use codes, particularly in mixed-use buildings. Using a Naive Bayes classifier, prediction accuracy reached 84.78% at the building level and 78.97% at the floor level. However, class imbalance and multiple-use entries led to misclassification, highlighting the need for layer-specific verification across all floors.

Linking between Building Register and Permit Data

A Building ID–based linked dataset was established to assess linkage quality between the building register and building/housing permit data. Among 55,784 buildings approved for occupancy in 2024, Building ID assignment rates were 97.0% for the building register, 99.4% for building permits, and 99.4% for housing permits.

Linkage success rates were 93.6% for building permits and 34.9% for housing permits. Among successfully linked records, area consistency exceeded 90% for building permits but ranged from 67–73% for housing permits; usage data consistency exceeded 95% for both datasets.

A shift toward a tailored maintenance framework that accounts for error patterns in building data

The pilot application confirmed that building data errors are not random but structured by temporal, regional, and institutional factors. Hence, uniform periodic maintenance alone is insufficient. A context-aware and customized maintenance framework integrating AI-based anomaly detection with rule-based validation is required.

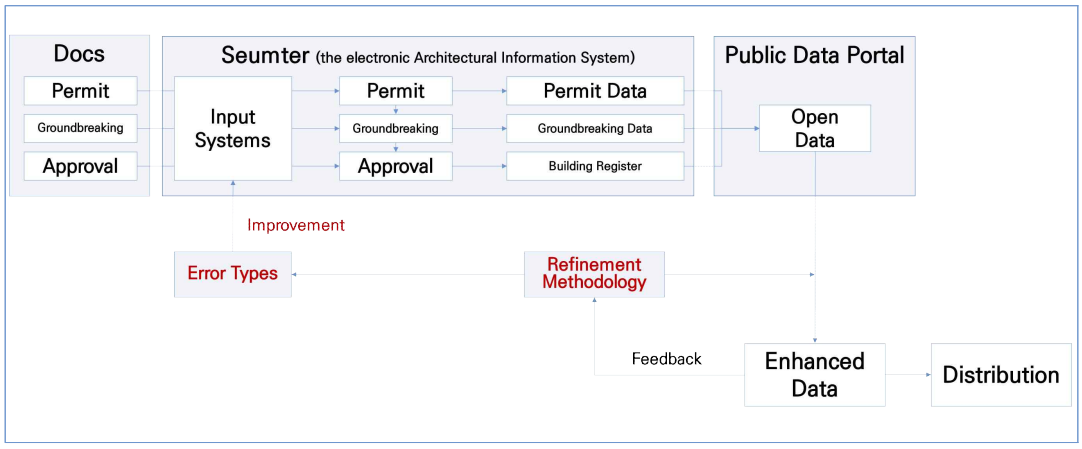
Furthermore, Building ID needs stricter management of linkage quality, ensuring the uniqueness and reliability of Building IDs as universal linkage keys across datasets.

Proposals for Building Data Quality Improvement

Independent Distribution of Enhanced Building Data

Given the administrative difficulty of directly amending official registers, this study proposes maintaining enhanced data externally from Seumter and distributing them as open-source datasets. Also the methodologies themselves—(1) rule-based error detection, (2) AI-based anomaly detection, and (3) cross-verification using Building IDs—should be openly published as reproducible documentation and code.

A phased release through the Public Data Portal is recommended, incorporating user feedback and version control to ensure continuous improvement and sustainability of the enhancement.



Proposals for Building Data Quality Improvement
Source: Written by the research team

Improving the Quality of New Data

For newly constructed buildings, validation should be integrated into the building and floor summary input systems during the permit process. Area values, currently entered manually, are prone to errors; therefore, automatic calculations and alert mechanisms should be introduced when discrepancies arise. For usage-related fields, inconsistencies or typos between free-text entries ("other use") and classified codes should trigger warnings for user correction.

At the approval for use stage, Building IDs should be actively used as linkage keys to verify data consistency between permit data and the building register data. When inconsistencies occur, the system should alert users, providing an opportunity to correct the data while maintaining the authority of the permitting agency.

Keywords
Building Register, Building ID, Machine Learning, Public Data, Open Source

부록

분석 코드

■ 면적 데이터 대상 기계학습 기반 이상탐지 코드

• 분석 코드

```
# %%
import time
from pathlib import Path

import duckdb
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
from sklearn.ensemble import IsolationForest
from sklearn.kernel_approximation- import Nystroem
from sklearn.linear_model import SGDOneClassSVM
from sklearn.pipeline import make_pipeline
from tabulate import tabulate

# from sklearn.neighbors import LocalOutlierFactor


plt.style.use("../auri.mplstyle")

pd.options.display.unicode.east_asian_width = True

# %%
parquet_path = Path("../data/processed/주건축물_검증규칙.parquet")
```

```
df_orig = pd.read_parquet(parquet_path)
print(f"Loaded shape: {df_orig.shape}")
df_orig.head()

# %%
con = duckdb.connect(database=":memory:")
query = """
SELECT
 *,
FROM
 df_orig
WHERE
 COALESCE("대지_면적(㎡)", "총괄_대지_면적(㎡)", 0)
 >= COALESCE("건축_면적(㎡)", 0) -- not 기존14
 AND COALESCE("연면적(㎡)", 0) >= COALESCE("건축_면적(㎡)", 0) --
not 기존16
 AND COALESCE("연면적(㎡)", 0) -- not 신규01
 >= COALESCE("용적_률_산정_연면적(㎡)", 0)
 AND COALESCE("대지_면적(㎡)", "총괄_대지_면적(㎡)", 0) >= 0 -- not
신규03
 AND COALESCE("건축_면적(㎡)", "총괄_건축_면적(㎡)", 0) >= 0
 AND COALESCE("연면적(㎡)", 0) >= 0
 AND COALESCE("지상_층_수", 0) >= 0
 AND COALESCE("지하_층_수", 0) >= 0
;
"""
rel = con.sql(query)
con.sql("SELECT COUNT(*) FROM rel").show()

# 검증규칙 확실히 통과한 건축물만 선택
df_filtered = rel.df()
with pd.option_context("display.max_columns", None):
print(tabulate(df_filtered.head(), headers="keys", tablefmt="psql"))

# %%
df_filtered.shape
```

```
59
60  # %%
61  df_filtered.dtypes
62
63  # %% [markdown]
64  # ### 무차원 변수 산출
65  #
66  # 건축물의 규모에 비례하지 않는(무차원) 개별 건축물의 특성을 반영하는 변수를 산출하기 위하여 면적 관련 변수의 차원 분석을 시행.
67  #
68  # | 이름 | 차원 |
69  # | ------------------- | ------------ |
70  # | 대지면적 | [면적] |
71  # | 건축면적 | [면적] |
72  # | 연면적 | [면적][층수] |
73  # | 용적률 산정용 연면적 | [면적][층수] |
74  # | 지상 층수 | [층수] |
75  #
76  # 〈!-- | 건폐율 | [면적]/[면적] = [비율] |
77  # | 용적률 | [면적][층수]/[면적] = [비율][층수] | --〉
78  # 〈!-- | 지하 층수 | [층수] | --〉
79  #
80  # 물리학의 차원분석은 길이, 질량 등을 다루며, 개수, 비율 등은 무차원수로 보나, 건축물 대상 차원분석에서는 건축법에 따라 건축물의 특성을 통제하는 변수인 면적, 층수, 비율 등을 별도의 차원으로 두었음.
81  #
82  # 1차적으로 비율 변수를 산출한 후, 각 비율 변수의 상관관계를 검토한 후 비례 관계가 나타나지 않는 무차원 변수를 최종 산출하고, 클러스터링 및 기계학습에 활용.
83  #
84  # | 이름 | 정의 | 차원 |
85  # | --------------- | ----------------------------- | ------------------------------ |
86  # | 건폐율 (재산출) | 건축면적/대지면적 | [면적]/[면적] = [비율] |
87  # | 용적률 (재산출) | 용적률 산정용 연면적/대지면적 | [면적][층수]/[면적] = [비율][층수] |
```

```
# | 지상 유효층수 | 용적률 산정용 연면적/건축면적 | [면적][층수]/[면적] = [비율][층수] |
# | 용적 산정률 | 용적률 산정용 연면적/연면적 | [면적][층수]/[면적][층수] = [비율] |
#
# 〈!-- | 실질 용적률 | 연면적/대지면적 | [면적][층수]/[면적] = [비율][층수] |
# | 유효층수 | 연면적/건축면적 | [면적][층수]/[면적] = [비율][층수] | --〉
#
# 용적률, 유효층수 등 연면적이 적용된 비율 변수를 층수로 나누어 층별 비율 변수를 도출.
#
# | 이름 | 정의 | 차원 |
# | --------------- | ------------------------------------ | ------------------------------- |
# | 지상층별 건폐율 | 용적률 산정용 연면적/지상층수/대지면적 | [면적][층수]/층수/[면적] = [비율] |
# | 지상층별 충만률 | 용적률 산정용 연면적/지상층수/건축면적 | [면적][층수]/층수/[면적] = [비율] |
#


# %%
# 우선 COALESCE 적용
대지면적 = df_filtered["대지_면적(㎡)"].fillna(df_filtered["총괄_대지_면적(㎡)"])
# 건축면적 = df_filtered["건축_면적(㎡)"].fillna(df_filtered["총괄_건축_면적(㎡)"])
지상층수 = df_filtered["지상_층_수"].fillna(0)
지하층수 = df_filtered["지하_층_수"].fillna(0)
총층수 = 지상층수 + 지하층수

df_dimensionless = pd.DataFrame(
 {
"관리_건축물대장_PK": df_filtered["관리_건축물대장_PK"],
"대지면적": 대지면적,
"건축면적": df_filtered["건축_면적(㎡)"],
```

```
115 "연면적": df_filtered["연면적(㎡)"],
116 "용적률_산정용_연면적": df_filtered["용적_률_산정_연면적(㎡)"],
117 "지상층수": 지상층수,
118 "건폐율": df_filtered["건축_면적(㎡)"] / 대지면적,
119 "용적률": df_filtered["용적_률_산정_연면적(㎡)"] / 대지면적,
120 # "실질_용적률": df_filtered["연면적(㎡)"] / 대지면적,
121 # "유효층수": df_filtered["연면적(㎡)"] / df_filtered["건축_면적(㎡)"],
122 "지상_유효층수": df_filtered["용적_률_산정_연면적(㎡)"]
123  / df_filtered["건축_면적(㎡)"],
124 "용적_산정률": df_filtered["용적_률_산정_연면적(㎡)"]
125  / df_filtered["연면적(㎡)"],
126 # "층별_건폐율": df_filtered["연면적(㎡)"] / 총층수 / 대지면적,
127 "지상층별_건폐율": df_filtered["용적_률_산정_연면적(㎡)"] / 지상층수 / 대지
    면적,
128 # "층별_충만률": df_filtered["연면적(㎡)"]
129 # / 총층수
130 # / df_filtered["건축_면적(㎡)"],
131 "지상층별_충만률": df_filtered["용적_률_산정_연면적(㎡)"]
132  / 지상층수
133  / df_filtered["건축_면적(㎡)"],
134  }
135 )
136
137 # 무한대/NaN 정리
138 df_dimensionless = df_dimensionless.replace(
139  [pd.NA, float("inf"), -float("inf")], np.nan
140 )
141
142 # %%
143 df_dimensionless.dtypes
144
145 # %%
146 df_dimensionless.head()
147
148 # %%
149 with pd.option_context("display.float_format", "{:.4f}".format):
```

```
 display(df_dimensionless.describe())

# %%
sampled = df_dimensionless.sample(n=3000, random_state=1106).copy()


# %%
sampled.head()

# %%
sns.pairplot(sampled.iloc[:, 1:6], corner=True)

# %%
sns.pairplot(sampled.iloc[:, 6:], corner=True)

# %% [markdown]
# ### Anormaly Detection
#
# Isolation Forest, One-Class SVM using SGD
#

# %%
df = (
 df_dimensionless.iloc[:, 6:12] # select only the dimensionless columns
 .dropna()
# .sample(n=1_000_000, random_state=1106)
 .copy()
)

df.head()

# %%
df.shape

```

```
# %%
import time
import numpy as np
import pandas as pd
from sklearn.ensemble import IsolationForest
from sklearn.kernel_approximation- import Nystroem
from sklearn.linear_model import SGDOneClassSVM
from sklearn.neighbors import LocalOutlierFactor
from sklearn.pipeline import make_pipeline

# settings
outliers_fraction = 0.01

anomaly_algorithms = [
 (
"One-Class SVM (SGD)",
 make_pipeline(
 Nystroem(gamma=0.1, random_state=1106, n_components=1_000),
 SGDOneClassSVM(
 nu=outliers_fraction,
 fit_intercept=True,
 random_state=1106,
 ),
 ),
 ),
 (
"Isolation Forest",
 IsolationForest(contamination=outliers_fraction, random_state=1106),
 ),
]

# Choose sizes to test
sizes = [10_000, 100_000, 1_000_000] # 4_703_097 failed on 64GB
RAM machine
results = []

```

```
for algo_name, clf in anomaly_algorithms:
for size in sizes:
 X = df.sample(n=size, random_state=1106).to_numpy()

 start = time.perf_counter()
 clf.fit(X)
 elapsed = time.perf_counter() - start

 results.append((size, elapsed))
print(f"clf={algo_name} n={size:<6} time={elapsed:.4f} sec")


# %%
X = df.to_numpy()
X_sampled = df.sample(n=1_000_000, random_state=1106).to_numpy
()

# settings
outliers_fraction = 0.01


# %%
clf = make_pipeline(
 Nystroem(gamma=0.1, random_state=1106, n_components=1_000),
 SGDOneClassSVM(
 nu=outliers_fraction,
 fit_intercept=True,
 random_state=1106,
 ),
)
clf.fit(X_sampled)
y_pred = clf.predict(X)
# align by index when assigning back
df_dimensionless["pred_one-class_svm"] = pd.Series(y_pred, index=d
f.index)

```

```
253 # runtime ~6 min.
254
255 # %%
256 clf = IsolationForest(contamination=outliers_fraction, random_state=11
    06)
257 clf.fit(X_sampled)
258 y_pred = clf.predict(X)
259
260 # align by index when assigning back
261 df_dimensionless["pred_isolation_fores-t"] = pd.Series(y_pred, index=d
    f.index)
262
263 # runtime <1 min.
264
265 # %%
266 # 1: inlier, -1: outlier
267 df_dimensionless.head()
268
269 # %%
270 df_orig.loc[1]
271
272 # %% [markdown]
273 # 숭인동아파트(숭인모범) 아파트 한 동에 대해, SVM은 inlier로 IF는 outlier
    로 판정함.
274 #
275
276 # %%
277 with pd.option_context("display.float_format", "{:.4f}".format):
278  display(df_dimensionless.describe())
279
280 # %%
281 stat_func_names = ["count", "mean", "std", "min", "max"]
282
283 stats = df_dimensionless.groupby("pred_one-class_svm")[
284  [
285 "대지면적",
```

```
286 "건축면적",
287 "연면적",
288 "용적률_산정용_연면적",
289 "지상층수",
290 "건폐율",
291 "용적률",
292 "지상_유효층수",
293 "용적_산정률",
294 "지상층별_건폐율",
295 "지상층별_충만률",
296  ]
297 ].agg(stat_func_names) # type: ignore
298 stacked = stats.stack(level=1, future_stack=True) # columns → index
299 stacked = stacked.swaplevel(0, 1) # flip the first two index levels
300 stacked.index = stacked.index.set_names(["통계", *stats.index.names])
301 stacked = stacked.reindex(index=stat_func_names, level=0)
302
303 with pd.option_context(
304 "display.float_format", "{:.4f}".format, "display.max_columns", None
305 ):
306  display(stacked)
307
308
309 # %% [markdown]
310 # ---
311 #
312
313 # %%
314 stat_func_names = ["count", "mean", "std", "min", "max"]
315
316 stats = df_dimensionless.groupby("pred_isolation_fores-t")[
317  [
318 "대지면적",
319 "건축면적",
320 "연면적",
321 "용적률_산정용_연면적",
```

```
    "지상층수",
    "건폐율",
    "용적률",
    "지상_유효층수",
    "용적_산정률",
    "지상층별_건폐율",
    "지상층별_충만률",
  ]
].agg(stat_func_names) # type: ignore
stacked = stats.stack(level=1, future_stack=True) # columns → index
stacked = stacked.swaplevel(0, 1) # flip the first two index levels
stacked.index = stacked.index.set_names(["통계", *stats.index.names])
stacked = stacked.reindex(index=stat_func_names, level=0)

with pd.option_context(
"display.float_format", "{:.4f}".format, "display.max_columns", None
):
  display(stacked)


# %%
cols = ["pred_one-class_svm", "pred_isolation_fores-t"]
pred_sum = df_dimensionless[cols].sum(axis=1, min_count=2)
# -2 for outlier by both, 0 for outlier by one

pred_both = pd.Series(pred_sum == -2, index=df_dimensionless.inde
x)
df_dimensionless["pred_both"] = pred_both.mask(pred_sum.isna())

df_dimensionless.head()

# %%
outliers = df_dimensionless[df_dimensionless["pred_both"] == True] #
noqa: E712
inliers = df_dimensionless[df_dimensionless["pred_both"] == False] #
noqa: E712
```

```
355
356 # Choose how many to keep
357 n_out = min(len(outliers), 1000) # keep up to 1k outliers
358 n_in = min(len(inliers), 3000) # keep up to 3k inliers
359
360 out_sample = outliers.sample(n=n_out, random_state=1106)
361 in_sample = inliers.sample(n=n_in, random_state=1106)
362
363 plot_df = pd.concat([out_sample, in_sample]).sort_index()
364 plot_df["prediction"] = plot_df["pred_both"].map({True: "Outlier", False:
    "Inlier"})
365 # Attach only the features + predictions
366 plot_df = plot_df.iloc[:, [*range(6, 12), -1]]
367 plot_df.head()
368
369 # %%
370 plot_df["prediction"].value_counts()
371
372 # %%
373 sns.pairplot(
374  plot_df, hue="prediction", corner=True, palette={"Outlier": "red", "Inlie
    r": "blue"}
375 )
376
377 # %% [markdown]
378 # ---
379 #
380 # 시도별, 연도별 표를 만들어보자.
381 #
382
383 # %%
384 df_results = df_orig.copy()
385 df_results["이상값"] = df_dimensionless["pred_both"]
386 # 무한대/NaN 정리
387 df_results = df_results.replace([np.nan, pd.NA, float("inf"), -float("inf"
    )], pd.NA)
```

```
df_results["시도_코드"] = df_results["시군구_코드"].str[:2]
df_results["사용승인_년"] = df_results["사용승인_일"].str[:4]
df_results = df_results[
 [
"관리_건축물대장_PK",
"시도_코드",
"사용승인_년",
"주_용도_코드",
"기존14",
"기존16",
"기존19",
"기존20",
"신규01",
"신규02",
"이상값",
 ]
]
with pd.option_context("display.max_columns", None):
print(tabulate(df_results.head(30), headers="keys", tablefmt="psql"))

# %%
df_results.to_parquet("../results/area_data_anomaly.parquet", index=Fal
se)

```

- 시각화 코드

```
# %%
import time
from pathlib import Path

import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
from tabulate import tabulate
```

```
10
11  plt.style.use("../auri.mplstyle")
12  pd.options.display.unicode.east_asian_width = True
13
14  # %%
15  parquet_path = Path("../data/processed/주건축물_검증규칙.parquet")
16  df_orig = pd.read_parquet(parquet_path)
17  print(f"Loaded shape: {df_orig.shape}")
18  with pd.option_context("display.max_columns", None):
19  print(tabulate(df_orig.head(30), headers="keys", tablefmt="psql"))
20
21  # %%
22  df_results = pd.read_parquet("../results/area_data_anomaly.parquet")
23  df_results = df_results.convert_dtypes()
24  print(f"Loaded shape: {df_results.shape}")
25  with pd.option_context("display.max_columns", None):
26  print(tabulate(df_results.head(30), headers="keys", tablefmt="psql"))
27
28  # %%
29  df_results.dtypes
30
31  # %%
32  df_results.head()
33
34  # %% [markdown]
35  # ---
36  #
37  # 시도별, 연도별 표를 만들어보자.
38  #
39
40  # %%
41  df_year = pd.DataFrame(
```

```
 {
"사용승인_년": [f"{year:04d}" for year in range(1950, 2025)],
 }
)
df_year

# %%
df_usage = df_orig[["주_용도_코드", "주_용도_코드_명"]].drop_duplicates()
df_usage = df_usage.set_index("주_용도_코드").sort_index()
df_usage = df_usage.reindex(
 index=[f"{i:02d}000" for i in range(1, 34)]
) # 01000, 02000, ···
df_usage

# %%
df_kcad = pd.read_csv(
"../data/processed/code_kcad_sgg_2024.csv", dtype="string"
).sort_values("행정구역분류")
df_kcad["시도_코드"] = df_kcad["시군구코드"].str[:2]
df_sido = df_kcad[["시도_코드", "시도"]].drop_duplicates()
df_sido

# %%
df_results.shape

# %%
df_results["기존16"].value_counts(dropna=False)

# %%
# mean()은 NA 제외하고 계산

print(df_results["기존16"].mean())
```

```
print(373505 / (7364967 - 343908))

# %%
# NA → False로 처리

print(df_results["기존16"].astype("boolean").fillna(False).mean())
print(373505 / (7364967))

# %%
# NA → False로 처리한 df 준비
# Convert columns to the best possible dtypes using dtypes supporting pd.NA
df_gb = df_results.iloc[:, 1:].copy().convert_dtypes()
df_gb.iloc[:, 3:] = df_gb.iloc[:, 3:].fillna(False) # NA → False
df_gb

# %%
df_gb.dtypes

# %%
gb = (
 df_gb.iloc[:, [0, *range(3, df_gb.shape[1])]]
 .groupby("시도_코드")
 .mean()
 .astype(float)
 .mul(100)
 .round(2)
)
df_ratio = df_sido.set_index("시도_코드").join(gb)
df_ratio

# %%
```

```
105 df_ratio.to_csv(
106 "../results/area_data_anomaly_ra-tio_by_sido.csv", index=True, encodin
    g="utf-8-sig"
107 )
108
109 # %%
110 df_orig["기존14"].value_counts(dropna=False)
111
112 # %%
113 df_orig[df_orig["시군구_코드"].str.startswith("38")]
114
115 # %% [markdown]
116 # ---
117 #
118 # 연도
119 #
120
121 # %%
122 # Convert columns to the best possible dtypes using dtypes suppor
    ting pd.NA
123 df_gb = df_results.iloc[:, 1:].copy().convert_dtypes()
124 df_gb.iloc[:, 3:] = df_gb.iloc[:, 3:].fillna(False) # NA → False
125
126 key_col = "사용승인_년"
127
128 gb = (
129  df_gb.iloc[:, [1, *range(3, df_gb.shape[1])]]
130  .groupby(key_col)
131  .mean()
132  .astype(float)
133  .mul(100)
134  .round(2)
```

```
)

df_ratio_year = df_year.set_index(key_col).join(gb).fillna("-")
with pd.option_context("display.max_rows", None):
 display(df_ratio_year)

# %%
df_ratio_year.to_csv(
"../results/area_data_anomaly_ra-tio_by_year.csv", index=True, encodin
g="utf-8-sig"
)

# %% [markdown]
# ---
#
# 용도
#

# %%
# Convert columns to the best possible dtypes using dtypes suppor
ting pd.NA
df_gb = df_results.iloc[:, 1:].copy().convert_dtypes()
df_gb.iloc[:, 3:] = df_gb.iloc[:, 3:].fillna(False) # NA → False

key_col = "주_용도_코드"

gb = (
 df_gb.iloc[:, [2, *range(3, df_gb.shape[1])]]
 .groupby(key_col)
 .mean()
 .astype(float)
 .mul(100)
```

```
  .round(2)
)

df_ratio_usage = df_usage.join(gb).fillna("-")
with pd.option_context("display.max_rows", None):
  display(df_ratio_usage)

# %%
df_ratio_usage.to_csv(
"../results/area_data_anomaly_ra-tio_by_usage.csv", index=True, encod
ing="utf-8-sig"
)

```

■ 용도 데이터 대상 기계학습 기반 오류 검증 코드

• 분석 코드

```
# %%
from pathlib import Path
import pandas as pd
import duckdb

data_root = Path("D:\\데이터\\건축데이터 건축허브 개방데이터")
path_DB = data_root / "건축물대장_2025년_02월.db"

results_dir = Path("../results")

# %%
# Open a DuckDB connection
con = duckdb.connect(database=path_DB, read_only=True)

# print the list of tables in the database
tables = con.execute("SHOW TABLES").fetchall()
print("Tables in the database:")
for table in tables:
print(table[0])
print()

# show heads of the tables
for table in tables:
 table_name = table[0]
print(f"Head of {table_name}:")
# print(con.execute(f"SELECT * FROM {table_name} LIMIT 5").fetchdf
())
print(con.sql(f"SELECT * FROM {table_name} LIMIT 5"))
print()

# %%
# Create a list to store the table names and their record counts
table_counts = []

```

```
# Iterate through the tables and count the records
for table in tables:
 table_name = table[0]
 count = con.execute(f"SELECT COUNT(*) FROM {table_name}").fetchone()[0]
 table_counts.append({"Table Name": table_name, "Record Count": count})

# Convert the list to a DataFrame
record_counts_df = pd.DataFrame(table_counts)

# Display the DataFrame
display(record_counts_df)

# %%
# Create a temporary view with the selected columns
층별개요_필터링 = con.sql("""
 SELECT
 "주_용도_코드",
 "주_용도_코드_명",
 "기타_용도"
 FROM 층별개요
 WHERE "기타_용도" IS NOT NULL AND "기타_용도" != ''
 AND "주_용도_코드_명" IS NOT NULL AND "주_용도_코드_명" != ''
""")
# Count the records in the filtered view
con.sql("SELECT COUNT(*) FROM 층별개요_필터링").fetchone()[0]

# %%
# Perform value count using SQL
value_counts_sql = con.sql("""
 SELECT
 "주_용도_코드",
 "주_용도_코드_명",
 "기타_용도",
 COUNT(*) AS "Count"
```

```
 FROM 층별개요_필터링
 GROUP BY "주_용도_코드", "주_용도_코드_명", "기타_용도"
 ORDER BY "Count" DESC
""")

# Fetch and display the result
value_counts_df = value_counts_sql.fetchdf()
display(value_counts_df.head(5))
# Save the top 30 rows to a CSV file
# value_counts_df.head(30).to_csv(
# results_dir / "표제부_용도_기재내용_top30.csv", index=False, encoding="utf-8-sig"
# )

# %%
# Create a temporary view to split the '기타_용도' column into arrays using a regular expression
value_counts_split = con.sql("""
 SELECT
 *,
 regexp_split_to_array(lower("기타_용도"), '[^\\p{L}\\d]+') AS "기타_용도_분리"
 FROM value_counts_sql
""")

# Fetch and display the result
value_counts_split.fetchdf()

# %%
# Create a temporary view to split the '기타_용도' column into rows using a regular expression
value_counts_split_table = con.sql("""
 SELECT
 *,
 regexp_split_to_table(lower("기타_용도"), '[^\\p{L}\\d]+') AS "기타_용도_분리"
```

```
 FROM value_counts_sql
""")

# Fetch and display the result
value_counts_split_t-able.fetchdf()

# %%
# Group by valid (5 digit) '주_용도_코드',
# sum the 'Count' column, and filter where the sum is >= 100
class_counts_sql = con.sql("""
 SELECT
 "주_용도_코드",
 "주_용도_코드_명",
 SUM("Count") AS "Total_Count"
 FROM value_counts_sql
 WHERE LENGTH("주_용도_코드") = 5 AND "주_용도_코드" ~ '^\\d+$'
 GROUP BY "주_용도_코드", "주_용도_코드_명"
 HAVING SUM("Count") >= 200
 ORDER BY "Total_Count" DESC
""")

# Fetch and display the result
class_counts_df = class_counts_sql.fetchdf().astype({"Total_Count": in
t})
display(class_counts_df)
# Save to a CSV file
class_counts_df.to_csv(
 results_dir / "층_naive_bayes_클래스별_문서빈도.csv",
 index=False,
 encoding="utf-8-sig",
)

# %%
# Group by '기타_용도_분리',
# sum the 'Count' column, and filter where the sum is >= 100
word_counts_sql = con.sql("""
```

```
 SELECT
 "기타_용도_분리",
 SUM("Count") AS "Total_Count"
 FROM value_counts_split_t-able
 WHERE "기타_용도_분리" IS NOT NULL AND "기타_용도_분리" != ''
 GROUP BY "기타_용도_분리"
 HAVING SUM("Count") >= 200
 ORDER BY "Total_Count" DESC
""")

# Fetch and display the result
word_counts_df = word_counts_sql.fetchdf().astype({"Total_Count": int})
display(word_counts_df)
# Save to a CSV file
word_counts_df.to_csv(
 results_dir / "층_naive_bayes_단어별_문서빈도.csv",
 index=False,
 encoding="utf-8-sig",
)

# %%
# Construct training data with '주_용도_코드' and '기타_용도_분리'
# Filter with class_counts_sql and word_counts_sql
train_data_sql = con.sql("""
 SELECT
 t."주_용도_코드" AS class,
 t."기타_용도_분리" AS word,
 t."Count" AS weight,
 FROM value_counts_split_t-able t
 JOIN class_counts_sql c
 ON t."주_용도_코드" = c."주_용도_코드"
 JOIN word_counts_sql w
 ON t."기타_용도_분리" = w."기타_용도_분리"
 WHERE t.Count >= 200
 ORDER BY t."Count" DESC
```

```
""")
train_df = train_data_sql.fetchdf()
train_df

# %%
import duckdb
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB


# 2. Vectorize & train

# The multinomial Naive Bayes classifier is suitable for classification with
# discrete features (e.g., word counts for text classification). The multinomial
# distribution normally requires integer feature counts. However, in practice,
# fractional counts such as tf-idf may also work.

# word is already a word, but no harm from count vectorization
# regex does not support unicode, so we need to customize the pattern
vectorizer = CountVectorizer(token_pattern=r"[\\w\\d가-힣]+")
X = vectorizer.fit_transform(train_df["word"])
y = train_df["class"]
weights = train_df["weight"]
clf = MultinomialNB(alpha=1.0)
clf.fit(X, y, sample_weight=weights)

# %%
vectorizer.vocabulary_

# %%
# test the classifier
```

```
test_text = "물탱크,EV기계실(연면적 제외)"
X_test = vectorizer.transform([test_text])
X_test.toarray()

# %%
clf.predict(X_test)[0]


# %%
# 3. Register a prediction UDF
def nb_predict(text: str) -> str:
 x = vectorizer.transform([text])
return clf.predict(x)[0]


# Remove the function if it already exists
try:
 con.remove_function("nb_predict")
except Exception as e:
if "No function by the name of" not in str(e):
raise

# Register the function
con.create_function("nb_predict", nb_predict)

# %%
preds = con.sql("""
 SELECT
 *,
 nb_predict("기타_용도") AS predicted_label,
 CASE
 WHEN "주_용도_코드" = predicted_label THEN 1
 ELSE 0
 END AS is_correct
 FROM value_counts_sql
""")
```

```
preds_label = con.sql("""
 SELECT
 preds.*,
 class_counts_sql."주_용도_코드_명" AS "predicted_label_name-",
 FROM preds
 LEFT JOIN class_counts_sql
 ON preds."predicted_label" = class_counts_sql."주_용도_코드"
""")
preds_df = preds_label.fetchdf()
preds_df

# %%
# Save preds to a CSV file
preds_df.to_csv(
 results_dir / "층_naive_bayes_예측결과.csv",
 index=False,
 encoding="utf-8-sig",
)

```

- 시각화 코드

```
# %%
import duckdb

# Load the CSV file into DuckDB
con = duckdb.connect()
results = con.sql("SELECT * FROM '../results/층_naive_bayes_예측결과.csv'")
# Display the first 5 rows of the results
display(results.df().head())

# %%
# Calculate the overall accuracy
overall_accuracy = con.sql("""
 SELECT
```

```
 SUM(Count * is_correct) / SUM(Count) AS overall_accuracy
 FROM results
""")
print("Overall Accuracy: {:.2f}%".format(overall_accuracy.df().iloc[0, 0] * 100))

# calculate the accuracy by 주_용도_코드
accuracy = con.sql("""
 SELECT
 주_용도_코드,
 주_용도_코드_명,
 SUM(Count) AS total_count,
 SUM(Count * is_correct) AS correct_count,
 SUM(Count * is_correct) / SUM(Count) AS accuracy
 FROM results
 WHERE LENGTH("주_용도_코드") = 5 AND "주_용도_코드" ~ '^\\d+$'
 GROUP BY 주_용도_코드, 주_용도_코드_명
 HAVING SUM(Count) >= 100
 ORDER BY SUM(Count) DESC
""")
accuracy_df = (
 accuracy.df()
 .astype({"total_count": "int", "correct_count": "int"})
 .head(20)
 .sort_values(by="주_용도_코드", ascending=True)
)
accuracy_df["accuracy"] = accuracy_df["accuracy"].apply(
lambda x: "{:.2f}%".format(x * 100)
)
display(accuracy_df)
# Save the accuracy results to a CSV file
accuracy_df.to_csv(
"../results/층_naive_bayes_accurac-y.csv", index=False, encoding="utf-8-sig"
```

```
)
accuracy.df().astype({"total_count": "int", "correct_count": "int"}).to_csv(
"../results/층_naive_bayes_accurac-y_orig.csv", index=False, encoding=
"utf-8-sig"
)

# %%
# # top and bottom 5 by accuracy
# top_5_accuracy = accuracy.df().nlargest(5, "accuracy")
# bottom_5_accuracy = accuracy.df().query("accuracy > 0").nsmallest
(5, "accuracy")
# print("Top 5 Accuracy:")
# display(top_5_accuracy)
# print("Bottom 5 Accuracy:")
# display(bottom_5_accuracy)

# %%
# Calculate the average Count for is_correct values 0 and 1 groupe
d by 주_용도_코드 and 주_용도_코드_명
average_count = con.sql("""
 SELECT
 주_용도_코드,
 주_용도_코드_명,
 AVG(CASE WHEN is_correct = 1 THEN Count ELSE NULL END) AS
avg_count_correct,
 AVG(CASE WHEN is_correct = 0 THEN Count ELSE NULL END) AS
avg_count_incorrect,
 FROM results
 WHERE LENGTH("주_용도_코드") = 5 AND "주_용도_코드" ~ '^\\d+$'
 GROUP BY 주_용도_코드, 주_용도_코드_명
 HAVING SUM(Count) >= 100000
 ORDER BY 주_용도_코드
""")
average_count_df = average_count.df()
```

```
# format the average counts to 2 decimal places
average_count_df["avg_count_correct"] = average_count_df["avg_count_correct"].apply(
lambda x: "{:.2f}".format(x)
)
average_count_df["avg_count_incorrect"] = average_count_df["avg_count_incorrect"].apply(
lambda x: "{:.2f}".format(x)
)
display(average_count_df)
# Save the average counts to a CSV file
average_count_df.to_csv(
"../results/층_naive_bayes_average-_count.csv", index=False, encoding="utf-8-sig"
)

```