부동산학연구 제30집 제1호, 2024. 3, pp. 7~19 http://dx.doi.org/10.19172/KREAA.30.1.1

Journal of the Korea Real Estate Analysts Association
Vol.30, No.1 2024. 3, pp. 7~19

건축물대장 데이터 품질 개선을 위한 용적률 결측오류 분석 및 결측치 처리방안

Analysis of Missing Data of Floor Area Ratio in Building Register and Proposed Solutions to Improve Data Quality

이 지 은 (Lee, Jieun)*

< Abstract >

This study examines the issue of missing floor area ratio (FAR) data in the building register of South Korea. The analysis reveals that the missing FAR data is not randomly missing, but rather it is found much more frequently among the deteriorated buildings built before 1990 and among the single-family homes or Class-1 neighborhood living facilities. This suggests that simply excluding the missing data would distort the facts. To address this problem, this study proposes practical solutions that researchers and data users can easily apply to minimize the bias caused by the missing data and improve data quality. By filling the missing 'site area(㎡)' values using continuous cadastral map spatial data, 60.4% of the missing FAR can be filled. Additionally, by applying the ratio of 'total floor area for FAR calculation' to 'total floor area' from a homogeneous group of non-missing buildings in terms of use-type and built year, it is possible to estimate and impute 35.7% of the missing FAR. In the long term, it is necessary to update missing values for key variables such as 'site area(㎡)' and 'FAR', open other related real-estate databases to the public, and establish a quality control strategy for newly added buildings in the building register.

Keyword : Building Register, Floor Area Ratio, Missing Values, Real Estate Big Data, Data Quality

I. 서론

1. 연구 배경

부동산 및 도시계획 분야의 실무자와 연구자에게 있어 도시의 개발밀도를 나타내는 '용적률' 데이터의 중요성은 아무리 강조해도 지나침이 없다. 용적률은 건축물이 위치한 대지면적 대비 건축물 바닥면적 합계(연면적)의 비율로, 도시계획 및 개발에 있어 중요한 역할을 한다. 용적률은 토지이용밀도를 나타내는 대표적인 지표로서, 효율적 토지이용을 위한 도시밀도 관리 수단으로 기능하고 있다. 이에 정확하고 완전한 용적률 자료는 건축물을 기반으로 계획 및 관리 방안을 도출하고자 하는 모든 분석에 필수적이다.

시민, 개발 주체, 공공의 관점에서도, 효과적인 용적률 계획 및 관리의 중요성은 아무리 강조해도 지나치지 않다. 용적률 계획은 그것이 실현되었을 때 도시 활력, 교통, 기반시설 등 도시환경에 직접적인 영향을 미치고, 개발 사업자는 용적률을 이해함으로써 개발사업 완료 이후 활용 가능 면적을 예측하고, 경제적·재무적 타당성을 고려한 계획 마련이 가능하다(이인성 외, 2009). 또한, 세계적으로 복합적인 토지이용이 대두되면서 용도지역체계가 재편되고 있고(이주일 외,

* 본 학회 정회원, The University of Hong Kong, Postdoc Fellow, jieun0441@naver.com

2022), 이에 따라 새로운 도시공간구조 마련에 있어 용적률 자료는 앞으로 더욱 그 활용 가치가 높아질 것으로 전망된다.

이러한 중요성에도 불구하고, 현재 용적률 자료 데이터의 신뢰성에는 우려의 목소리가 높다. 건축물대장을 다룬 연구에 따르면, 건축물대장에는 용적률 값이 0으로 기입되어 있는 등 유효하지 않은 데이터의 비중이 높다(이성화, 2010). 또한 '개별주택특성조사' 수행 시 별도로 조사하는 실제값과 건축물대장상의 자료값을 비교한 결과, 용적률 자료의 일치도는 13.5% 수준에 그쳐 활용 상 한계가 지적된 바 있다(강영옥 · 이주일, 2005).

낮은 용적률 데이터 품질 문제는 분석의 신뢰성을 떨어뜨림으로써 예측 기반 밀도계획을 무력화하고, 이에 따라 부적절한 도시개발구역 설정, 인프라 부족이나 과잉 등 비효율적 토지이용을 초래할 수 있다. 또한, 용적률 데이터의 불일치는 신규 개발이 대상지에 미칠 영향을 정확하게 평가할 수 없게 함으로써, 부적절한 밀도 수준으로 인한 외부효과로 교통혼잡, 녹지공간 부족, 주민 삶의 질 저하 등의 문제를 초래할 수 있다.

특히, 용적률의 결측치 문제는 더욱 신중한 검토가 필요하다. 현재 용적률 데이터는 유효하지 않은 데이터의 존재, 즉, 완전성의 결함뿐만 아니라, 단순히 값이 비어있는 상태가 아닌, 0으로 기입된 상태로 민간에 개방 · 제공되고 있는 데에 있다. 이로 인해 만약 데이터 이용자들이 분석 이전에 일부 자료의 용적률이 0으로 기재되어 있음을 인지하지 못한다면, 분석 대상지의 평균 용적률을 구함에 있어, 상당수 건물의 용적률을 0으로 산입하여 계산함으로써 현상을 왜곡하여 파악하게 될 위험이 있다.

이러한 문제를 인식하고 국토교통부에서는 건축물대장 자료의 정확성을 높이기 위해 2006년 시범사업을 시작으로 2007년부터 2009년까지 「건축물대장 기초자료 정비사업」을 3차에 걸쳐 실시한 바 있다(국토교통부, 2009). 이러한 과정을 거쳐 상당한 개선이 이루어졌으나, 여전히 중요 변수 누락이나 논리 오류가 발견되어 신뢰할 만한 부동산 통계로 활용하기에 무리가 있는 상황이다(이성화, 2010).

수많은 연구에서 용적률 변수에 초점을 두고 공동주택 가격에 미치는 영향이나 주거지 밀도분포, 실현용적률에 관한 결정요인 분석 등의 연구를 진행해왔으나, 건축물대장 내 용적률 데이터 결측치의 문제는 그간 관심의 대상에서 배제되었다. 이에 본 연구는 건축물대장 내 핵심 변수인 용적률 결측치에 초점을 두고, 용적률 결측치의 특성과 데이터 품질 개선방안을 고찰하였다.

2. 연구의 목적

만약 건축물대장의 용적률 변수 결측 건축물이 비결측 건축물과 동일한 특성 분포를 갖는다면, 용적률 결측값을 제외해도 평균적인 개발밀도를 구하는 등 연구나 계획의 분석 기초자료로 활용하는 데 무리가 없을 것이다. 그러나 용적률이 누락된 건축물들이 특정한 지역에 집중되어 있다거나, 용적률이 결측인 건물과 그렇지 않은 건물들이 구조적 차이를 갖는다면, 이들 결측치를 단순히 제외하여 활용하는 것은 문제가 된다. 비결측 데이터만을 가지고 추론하였을 때 전체 양상의 이해에 오류가 생기기 때문이다. 이 때문에 우리는 건축물대장 데이터를 활용하기에 앞서 용적률 결측치가 결측치의 세 가지 유형 중 어떤 유형에 해당하는지 먼저 진단하여야 한다.

결측치는 크게 완전 무작위 결측, 무작위 결측, 비무작위 결측 세 가지 범주로 분류된다(Rubin, 1976). 먼저 '완전 무작위 결측(MCAR, Missing Completely at Random)'은 어떤 변수의 누락이 다른 데이터와 관련 없이 무작위로 발생한 경우를 말한다. 이 경우는 결측을 제외하고 분석해도 문제가 없다.

한편, '무작위 결측(MAR, Missing at Random)'은 결측값이 결측치 변수와는 상관성이 없지만 다른 변수와는 관련이 있는 경우로, 만약 결측치의 분포가 실제값과 관련 없이 다른 변수와 연관되어 발생하는 경우, 무작위 결측이라 할 수 있다. 이 경우, 관측할 수 있는 자료로부터 결측치 추정이 가능하므로 다양한 결측치 대체 방법을 활용할 수 있다.

'비무작위 결측(MNAR, Missing Not at Random)'은 어떤 변수의 결측값이 해당 변수와 관련이 있는 경우를 말한다. 이 경우 결측치의 발생은 관측된 값과 결측된 값 모두에 영향을 받는 상태이므로 결코 무시할 수 없다. 예를 들어, 만약 용적률 결측치가 용적률이 낮은 건축물에서 더 빈번하게 발견된다면, 비무작위 결측에 해당한다. 이러한 비무작위 결측을 처리하려면 다양한 가정(what-if) 분석을 통해 결측 원인에 대한 단서를 찾고, 세세하게 추가 조사를 하는 등 유효한

통계적 추론이 될 수 있도록 접근해야 한다(Burren, 2018).

비록 용적률 결측이 '완전 무작위 결측'에 해당하지 않더라도 용적률 결측값과 다른 변수들과의 관계를 찾아 이로부터 결측치를 적절한 값으로 대체할 수 있다면, 결측 용적률의 편향을 일부 줄일 수 있을 것이다. 또한 만약 용적률 결측이 '비무작위 결측'에 해당한다면, 본 연구에서 이러한 결측치의 문제를 인식함으로써 추후 건축물대장 기초자료 정비 시 가장 시급히 바로잡아야 할 변수로 용적률을 선정하거나, 건축물 자료 정비 대상 지역 선정에 도움이 될 수 있을 것이다.

이에 본 연구에서는 용적률 결측이 지역, 노후도, 용도 등 타 변수와의 연관성을 갖는지 살펴보고, 용적률 결측과 높은 연관성을 갖는 변수들을 밝히며, 정확한 용적률 자료 활용을 위해 건축물대장 용적률 결측치 처리방안을 도출하는 것을 목적으로 한다.

II. 선행 연구 검토

그동안 용적률 자료의 활용은 용도지역과 관련한 개발현황 분석이나, 실현용적률을 종속변수로 하여 특정 지역의 실현용적률 결정요인을 분석하거나, 용적률이 공동주택 가격에 미치는 영향을 규명하는 연구 위주로 이루어져 왔으며, 출처로는 대부분 건축물대장이 활용되었다(김수현 · 최창규, 2019; 김수현 · 최창규, 2021; 김형보, 1998; 백태경 · 김영훈 · 최정미, 2004; 윤상복 외, 2004; 이주일 외, 2022; 이희정 · 김기호, 2001).

이에 건축물대장의 용적률 변수를 활용한 선행연구를 대상으로 용적률 데이터 결측치의 처리 방식이나, 용적률 자료의 대표성 및 신뢰성에 관한 검토 여부를 검토한 결과, 대부분의 연구에서는 용적률 결측치를 어떻게 처리하였는지에 관한 언급이 없거나(김수현 · 최창규, 2019; 박재빈 외, 2017; 윤병훈 · 남진, 2013; 윤병훈 · 남진, 2014; 윤혜림 · 남진, 2013), 건축물대장 상의 용적률 결측치 존재를 인지한 후 해당 데이터는 제외 또는 분리하여 분석에 활용했음을 알 수 있었다(김수현 · 최창규, 2019; 김수현 · 최창규, 2021; 이윤상 · 남진, 2014; 이희정 · 김기호, 2001). 즉, 용적률(개발밀도)이 공동주택 가격에 미치는 영향 등 용적률의 역할이나, 용적률을 결정하는 다양한 공간 특성 요인들이 다양하게 규명되었지만, 그러한 분석을 수행하기에 앞서 용적률 결측 자료를 비결측 자료와 비교 · 검토하고 그 결과를 포함하여 보고한 연구는 부재한 실정이다.

그러나 용적률 결측을 포함한 건축물대장 데이터 품질 문제를 인식한 연구자들은 이전부터 건축물대장의 오류 유형을 밝히고 대장의 정비와 활용 가능성 개선 방안을 모색해 왔다(강영옥 · 이주일, 2005; 이성화, 2010; 신상영 · 장영희, 2006; 김정옥 외, 2008; 김승범, 2015). 이성화(2010)는 건축물대장 내 존재하는 105개의 논리오류 유형을 도출하여 건축물대장 내 각종 변수 항목들에 존재하는 오류의 유형을 소상히 밝혔으며, 특히 용적률과 관련된 오류로는 (용적률산정연면적/대지면적)*100의 값과 용적률 변수의 기재값이 다르게 기입되어 있는 경우, 용적률에 기입된 값이 0이거나 값 자체가 기입되어 있지 않은 경우로 2개의 논리오류 유형이 있음을 밝혀냈다. 또한, 건축물대장 전반에 걸쳐 이렇게 수많은 논리 오류가 발생하게 된 원인을 건축물대장 관련 제도의 변천사와 관리항목의 변화, 최초 원시 자료의 부실 등록, 대장기록 사항의 의무화 제정 지연으로 밝히고 있다.

또한 강영옥 · 이주일(2005)은 건축물대장이 건축물 정보의 근간이지만 실제 사용상에서 여러 한계를 가지며, 자료의 신뢰도가 매우 낮은 실정임을 밝히고, 개선방안을 제시하였다. 제시된 개선방안으로는 건축물 과세대장, 위법건축물대장, 수치지도 정보를 통합한 건축물 정보 통합 DB를 만들 것, 이러한 복수의 대장들 간 연계를 가능케 하는 고유값(key)을 부여하고 분류체계를 일치시킬 것, 각 대장별 정확도를 높이기 위해 행정업무과정 상 문제 발생 소지를 없앨 것, 총체적인 현장 조사를 통해 실제 상황과 기재 내용 및 각 대장들 간 기재 내용의 일치도를 높일 것이 제시되었다.

신상영 · 장영희(2006)는 특히 건축물대장 자료의 누락과 정확성 문제를 다루면서 건축물대장과 과세대장 간 주택용도 불일치가 단독주택에서 특히 심각함을 밝혀, 건축물 유형에 따라 대장 자료의 품질 차이가 두드러짐을 시사하였다.

위와 같은 선행연구들을 통해 건축물대장의 문제들이 밝혀졌으나, 용적률 변수를 분석하는 연구자의 관점에서 결측치의 문제를 다루거나, 실제 용적률을 주요 관심 변수로 하여 분석한 연구 중 건축물대장의 결측치를 단순 제외하는 것 외에 대표성 확보를 위해 결

측치의 특성을 검토하거나 적절한 처리를 병행한 연구는 찾을 수 없었다. 이에 본 연구는 도시 내 개발밀도 및 정비계획과 관련하여 용적률을 다루는 도시계획 및 부동산 분야 연구자의 관점에서 건축물대장 내 용적률 데이터의 결측치에 관한 탐색적 분석을 진행하여 용적률 결측치의 특성을 밝히고, 건축물대장의 용적률 데이터 품질 개선방안을 제시함으로써 향후 용적률 통계를 보다 유용하게 활용할 수 있도록 돕고자 한다. 구체적으로는 먼저 건축물대장 통계 이용자들이 건축물대장의 용적률 결측 오류의 존재를 정확히 인지하고, 이로 인한 편향(Bias)을 최소화하여 활용할 수 있게 한다. 이를 통해 개발밀도 현황 분석의 신뢰도를 높이고, 정확한 현황 파악을 기반으로 한 효율적인 토지이용계획 및 실현에 이바지할 수 있을 것으로 기대된다.

Ⅲ. 분석자료 및 방법

1. 분석자료

1) 건축물대장 개요

본 연구는 서울시 내 전수 건축물 데이터 자료 분석을 위해 '건축데이터 민간개방시스템'에서 제공하는 건축물대장 총괄표제부와 표제부 자료를 활용하였다(국토교통부, 2023). 건축물대장은 시장·군수·구청장이 관리하는 건축물 및 그 부지에 관한 현황을 관리하는 대장으로, 건축물의 소유·이용 상태를 확인하거나 건축정책의 기초자료로 활용하는 대표적인 우리나라의 건축물 데이터이다(국토교통부, 2023).

건축물대장은 일반건축물대장과 집합건축물 대장이 있다. 일반건축물이란, 건축물의 소유권이 하나의 주체에 귀속되는 경우를 말한다. 이러한 일반건축물에 해당하는 건축물 및 대지 현황은 '일반건축물대장'에 기재 및 관리된다. 기본적으로 건축물대장은 건축물 1개 동을 단위로 하여 건축물마다 작성하며, 부속건축물이 있는 경우는 주된 건축물대장에 포함하여 작성된다. 한편, 아파트나 집합 상가와 같이 여러 소유자가 1개 동의 건축물에 존재하는 경우, 집합건축물로 분류하며 집합건축물에 해당하는 건축물 및 대지 관련 정보는 '집합건축물대장'에 기재·관리된다.

집합건축물대장은 표제부와 전유부로 나누어 작성하는데, 표제부는 집합건축물의 총괄적인 현황을, 전유부는 건축물 내 호별 소유 현황을 기재한다. 또한 하나의 대지에 2개 이상의 건축물이 있는 경우, 총괄표제부를 작성한다(국토교통부, 2023).

국토교통부 녹색건축과 건축행정시스템(세움터)에서는 건축인허가 현황, 건축물폐쇄·말소대장과 함께 건축물대장을 민간에 개방하고 있다. 건축물대장은 원시데이터 자료 전체가 다운로드가 가능하도록 공개되어 있어, 그 투명성이 가장 높고, 데이터 품질 개선의 여지도 가장 높다. 본 연구는 이처럼 활용 가치가 높은 건축물대장의 정확하고 효과적인 활용 방안을 모색하는 차원에서 진행되었다.

2) 분석 자료의 범위

분석 자료는 일반건축물과 집합건축물 총괄표제부와 표제부를 모두 포괄하는 2023년 12월 기준 서울시 내 전수 데이터이다. 서울시 건축물대장 내 총괄표제부 자료는 19,406개, 표제부 자료는 592,374개 행으로 이루어져 있으며, 주소, 대지면적, 주용도코드, 연면적, 용적률 산정 연면적, 사용승인연도, 용적률 등의 변수를 포함하고 있다. 분석 단위는 필지 단위를 기준으로 하였다. 통상 건축행위는 필지단위로 발생하므로, 개발밀도 분석은 필지단위의 세밀한 분석을 수행함이 바람직하기 때문이다(이윤상·남진, 2014). 일반건축물대장은 1개 필지에 1개 동이 건축된 경우가 기재되지만, 공동주택이나 오피스 등과 같이 한 필지에 여러 동의 건축물이 건축되는 경우, 개별 동의 용적률을 확인할 필요성은 매우 부족하다. 따라서 본 연구는 총괄표제부와 표제부 자료를 단순 병합하지 않고, 1개의 대지에 여러 동의 건축물이 있는 경우는 해당 단지 전체에 관한 용적률을 기재하고 있는 총괄표제부 자료만 분석에 포함되도록 하여 분석 자료를 구성하였다. 이렇게 준비한 서울시 내 건축물 빅데이터 자료는 520,522개 행으로 이루어져 있다.

본 연구는 "용적률 결측 오류(0값 기입 케이스)가 건축물대장 내에 얼마나 존재하는가?"라는 질문으로 시작한다. 대지면적 대비 연면적[1]의 비율을 나타내는 용적률의 개념적 정의상, 건축물이 존재한다면 용적률

[1] 연면적은 건축물 바닥면적의 합계이다. 실제 건축물대장에 용적률로 기재되는 연면적 값은 건물의 '연면적' 변수값이 아닌, '용적률 산정 연면적'을 대지면적으로 나눈 값이라는 점도 밝혀둔다.

값은 논리적으로 0이 될 수 없다. 그렇기에 용적률이 0으로 기입된 경우는 명백히 논리오류[2]이다. 본래 결측치란 값이 아예 없는 경우를 가리키지만, 본 연구에서 활용한 건축물대장 자료에서 용적률 값이 기입되지 않은 케이스는 없었다. 따라서 본 연구는 용적률이 0 값으로 기재된 경우를 용적률 결측치로 정의하였다.

2. 분석방법

1) 결측치와 타 변수 간 탐색적 데이터 분석

본 연구는 전통적 통계 도구를 사용하여 가설을 검증하는 확증적 데이터 분석(Confirmatory Data Analysis, CDA) 방식이 아닌, 기술통계에 해당하는 탐색적 데이터 분석(Exploratory Data Analysis, EDA) 방법을 활용하였다. EDA는 데이터의 추세, 분포, 관계를 시각화하여 살펴봄으로써 가설 검증 방식으로는 미처 발견하기 어려운 패턴을 발견할 수 있다(Tukey, 1977). 존 튜키는 그동안 가정을 세우고 그 가정을 확증하는 방식의 연구는 지나치게 강조되어 온 반면, 적절한 연구 설계 및 가설 검증을 담보하기 위한 과정으로서 필수적인 탐색적 데이터 분석은 지나치게 간과되어 왔음을 지적하면서 EDA와 CDA간 상호보완적 관계가 있음을 주장하였다(Tukey, 1980). 통계적 관점에서 보면, EDA를 통해 추론통계적 방법론이 근간으로 하는 가정의 유효성 평가와 적절한 모형 수립이 가능하며, 이론적 측면에서는 더 뚜렷하고 타당한 연구질문과 가설을 수립할 수 있다(Jebb et al., 2017). 즉, 탐색적 데이터 분석을 수행하면 기본 추세 및 패턴에 대한 통찰력을 확보할 수 있으며, 데이터셋과 관련된 다양한 요소에 대한 이해를 크게 향상시킬 수 있을 뿐 아니라, 데이터셋에서 중요한 오류나 이상치, 결측치 값을 식별할 수 있어 데이터의 가치를 최대화할 수 있다(Jebb et al., 2017). 이러한 탐색적 접근은 작은 표본에 의존할 수밖에 없었던 이전과는 달리, 전수 빅데이터 분석이 가능해진 최근 더욱 널리 활용되고 있다(김승범, 2015). 이러한 탐색적 분석과정은 어떤 연구에서든 본격적인 모델링 이전에 거쳐야 하는 과정이지만, 그동안 용적률 결측치를 대상으로 한 탐색적 데이터 분석은 많은 경우 생략되거나, 그 결과가 공유되지 못하였다.

본 연구는 특정 표본이 아닌 서울시 내 건축물 전수 자료를 대상으로 하고, 기존 용적률 자료 결측치의 추세 및 패턴을 찾아내는 데 주요 목적이 있으므로 Jebb et al.(2017)이 언급한 바와 같이 데이터의 추세 및 패턴 파악, 데이터셋의 주요 결측값 식별 등에 적합한 탐색적 분석(EDA)을 통해 분석하였다.

본 연구는 다음과 같은 순서로 분석을 진행하였다. 먼저 서울시 내 건축물대장 데이터 중 용적률 결측치의 비중을 파악한 후, 용적률과 관련성이 있을 수 있는 다른 변수들과의 관계를 탐색하여 용적률 결측치의 무작위성을 검토하였다. 만약 결측치 비율이 미미하다면, 해당 데이터를 단순히 제거 후 계산하더라도 그에 따른 부작용 또한 작겠지만, 혹 그 비중이 높다면, 결측 데이터와 비결측 데이터 간에 체계적 차이의 존재 여부는 검증이 필요한 중요한 문제가 된다.

이 과정에서 먼저 서울시 내 자치구별 용적률 결측 비율을 파악하였다. 또한, 용적률이 결측인 건축물과 그렇지 않은 건축물의 노후도(건물 연한) 분포를 각각 박스플롯(Box-plot)으로 시각화·대조함으로써 건축물 노후도와 용적률 결측치 간 관계성을 검증하였다. 이후 건축물 주용도와 결측치 발생 확률 간 연관성 여부를 밝히기 위해 용적률 결측 데이터의 용도별 분포와 전체 데이터의 용도별 분포를 비교하고, 전체적인 분포 양상에 비해 용적률이 결측인 데이터에서 그 비중이 더욱 높게 발견되는 주용도 유형 건축물이 있는지 탐색하였다.

2) 용적률 결측 발생 원인 분석 및 보완방안 도출

이후 분석에서는 건축물대장 데이터를 분석에 활용하는 연구자의 입장에서 용적률 결측치 자료들을 유형별로 분류하고, 결측치에 내재된 편향을 최소화하기 위한 실제적인 방안을 도출하였다. 구체적 방법으로는 건축물대장 테이블에 존재하는 '대지면적' 과 '용적률 산정 연면적' 변수의 현황에 따라 4개 범주로 분류하고, 각 케이스별로 용적률 결측치로 인해 발생하는 오차를 최소화하여 분석에 활용할 수 있는 방안을 제시하였다. 이어 결론에서는 건축물대장 자료의 생성·관리 주체의 관점에서 근본적·장기적으로 결측치 문제를 개선하고, 데이터 품질 및 신뢰도를 높일 방안을 제언하였다.

[2] 이는 이성화(2010)의 연구에서도 지적된 논리오류 105종 중 하나에 해당한다.

Ⅳ. 분석 결과

1. 탐색적 데이터 분석

1) 용적률 결측의 비중 및 지역별 분포

서울시 내 건축물대장의 용적률 결측율은 36.7%로,[3] 상당히 높은 비중으로 결측이 발견된다(<표 1> 참조). 이러한 결측의 무작위성을 밝히기 위해, 먼저 용적률 결측치의 자치구별 분포를 탐색하였다.

자치구별 용적률 결측 비율을 분석한 결과, 결측의 비중이 가장 높은 중구와 종로구는 각각 60.1%, 57.1%, 가장 낮은 강서구와 강남구는 각각 14.5%, 13.2%로, 자치구별 편차가 상당히 크게 나타났다. <표 1>과 같이 각 자치구 내 건축물 데이터 수 대비 결측치 비율을 기준으로 내림차순으로 정렬했을 때, 유서 깊은 건축물이 다수 분포하는 중구와 종로구는 결측율이 약 60%에 달하는 반면, 비교적 최근에 개발된 강서구, 강남구와 같은 지역은 결측율이 낮아, 이러한 용적률 결측율의 지역 간 차이는 상당 부분 개발 시기의 차이에서 기인한 것으로 추측되었다.

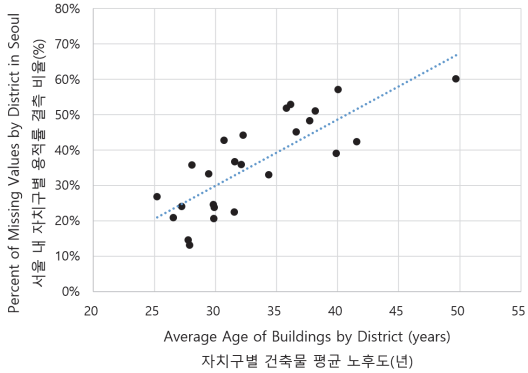
2) 용적률 결측과 노후도 간 관계 탐색

용적률 결측과 노후도 간 관계를 탐색하기 위해 자치구별 건축물의 평균 노후도[4]를 구하고 시각화한 결과, 실제로 각 지역 건축물의 평균 노후도가 높을수록 용적률 결측 비율도 높아지는 경향이 발견되었다(<표 1>, <그림 1> 참조). 이어 용적률 기재 데이터와 결측 데이터를 분리하여 건축물 노후도를 Box-Plot으로 분석한 결과, <그림 2>와 같이 뚜렷한 차이가 나타났다. 용적률이 기입되어 있는 건축물은 노후도 평균이 24.1년인데 반해, 용적률 결측 건축물은 44.3년으로 무려 20년의 차이를 보였으며, 제1사분위수(Q1) ~ 제3사분위수(Q3) 박스 구간도 겹치지 않는 양상을 보였다(<그림 2>, <표 2>). 최근 승인된 건축물들에서도 계속 용적률 결측이 발생하고 있긴 하지만, 용적률 결측치의 절대다수(75%)는 35년 이상 노후 건축물(1990년 이전 승인)이 차지하고 있다. 이를 통해 건축물대장의 용적률 결측은 건축 시기와 매우 높은 관련성이 있으며, 세 가지 결측치 유형 중 완전 무작위 결측에 해당하지 못함을 확인하였다.

<표 1> 각 지역별 건축물 내에서 용적률 결측 비율

자치구	건축물 데이터 수		결측율 (%)	노후도 평균(년)
	결측치	전체		
중구	10,048	16,717	60.1%	49.7
종로구	13,663	23,929	57.1%	40.0
영등포구	12,262	23,162	52.9%	36.2
서대문구	10,523	20,315	51.8%	35.8
성북구	15,205	29,788	51.0%	38.2
동대문구	11,403	23,593	48.3%	37.7
성동구	6,988	15,466	45.2%	36.6
구로구	8,600	19,446	44.2%	32.3
은평구	11,710	27,358	42.8%	30.7
용산구	9,224	21,758	42.4%	41.6
강북구	9,714	24,885	39.0%	39.9
마포구	8,255	22,476	36.7%	31.6
노원구	3,568	9,943	35.9%	32.1
관악구	10,656	29,836	35.7%	28.1
양천구	5,214	15,667	33.3%	29.5
동작구	7,385	22,391	33.0%	34.4
송파구	5,608	20,925	26.8%	25.2
금천구	3,339	13,575	24.6%	29.8
강동구	4,235	17,617	24.0%	27.3
광진구	5,568	23,351	23.8%	29.9
도봉구	3,123	13,890	22.5%	31.6
서초구	3,244	15,552	20.9%	26.6
중랑구	5,235	25,432	20.6%	29.9
강서구	3,217	22,134	14.5%	27.8
강남구	2,814	21,316	13.2%	27.9
합계	190,801	520,522	36.7%	33.1

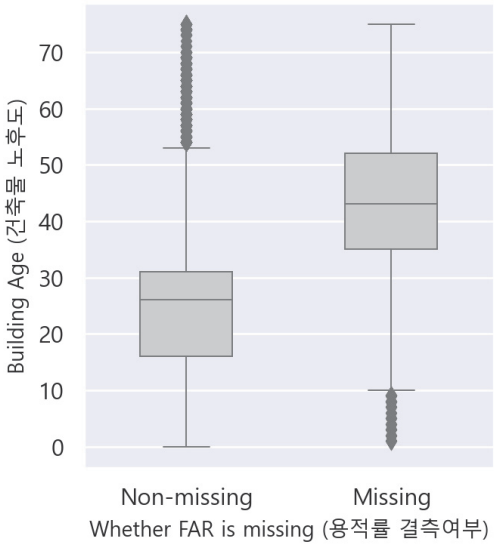
<그림 1> 구별 용적률 결측 비율과 노후도간 관계



[3] 1개 필지에 여러 동이 있는 경우, 개별 동 자료는 제외한 본 연구의 분석 자료 기준 수치이며, 원자료인 총괄표제부 결측율은 53.1%(10,298건/19,406건), 표제부 원자료의 결측율은 44.1%(261,477건/592,374건)로 더욱 높은 실정이다(2023년 12월 자료 기준).

[4] 노후도는 (2023년 - 승인연도)로 계산하였으며, 승인연도 변수 오류값(ex: 19, 0, NaN 등)의 영향을 최소화하기 위해, 승인연도가 0으로 기재된 경우, 결측인 경우, 1948 (대한민국 정부 수립연도)보다 작은 수치로 기재된 경우는 제외 후 승인연도 분포를 도식하였다.

<그림 2> 용적률 결측여부에 따른 건축물 노후도



<표 2> 용적률 결측에 따른 건축물 노후도 분포

(단위: 년; Unit: year)

구분	용적률 비결측 Non-Missing	용적률 결측 Missing
평균(Mean)	24.1	44.3
표준편차(Std.)	11.8	10.3
최소값(Min)	0.0	1.0
1사분위(Q1)	16.0	35.0
2사분위(Q2)	26.0	43.0
3사분위수(Q3)	31.0	52.0
최대값(Max)	75.0	75.0

3) 용적률 결측치와 주용도와의 관계 탐색

용적률 결측치가 노후 건축물에 집중되어 있다는 점에 더하여, 용적률 결측 건축물들이 특정 용도에 집중되어 있지는 않은지 탐색하기 위해, 용적률 결측 데이터 내 각 용도의 비율과 전체 건물 중 각 용도가 차지하는 비율 간 차이를 조사하였다(<표 3> 참조).

분석 결과, 건축물의 용도별 비중 차이를 고려하더라도 결측 빈도수와 결측 비율 모두에서 단독주택과 제1종근린생활시설 두 용도에 결측치가 집중되어 있음을 알 수 있었다. 전체 용적률 결측치 대비 각 용도의 해당 비율(<표 3>의 E열)을 보면, 단독주택이 68.0%로 가장 큰 비율을 차지하고, 2위는 제1종근린생활시설 13.9%로, 두 용도는 총 81.9%를 차지한다. 그러나 전체 건축물 중에서 이들 용도의 비중(<표 3>의 F열)은 단독주택 52.8%, 제1종근린생활시설 10.3%로, 단독주택과 제1종근린생활시설 용도가 전체 건축물에서 차지하는 비중 63.1%에 비해 결측치 건축물에서 차지하는 비중이 현저히 높다. <표 3>의 G열에서 볼 수 있는 바와 같이 단독주택의 경우 결측치 데이터셋 내 비율이 15.1%p 더 높았으며, 제1종근린생활시설의 경우 3.6%p 더 높았다. 이를 통해 서울시 내 건축물 중 단독주택과 제1종근린생활시설 용도 건축물에서 용적률 결측 오류가 더 빈번하게 발생함을 알 수 있다.

4) 소결

이상의 분석을 통해 건축물대장의 용적률 결측치는 비무작위 결측(MNAR)에 해당한다는 결론을 내릴 수 있다. 건축물대장의 용적률 결측치는 건축물 노후도, 건축물 용도(단독주택 및 제1종근린생활시설), 지역적 분포와 독립적이지 않은 관계를 보였으며, 단독주택과 제1종근린생활시설 두 용도는 전체 건축물 중에서는 약 63%를 차지하지만, 용적률 결측치 중에서는 80% 이상을 차지하고 있다. 특히 단독주택은 층수 제한이 있어 타용도에 비해 낮은 용적률로 개발되는 경향이 뚜렷하기에, 용적률 변수와 결측치 발생 간 상관성을 배제할 수 없어 비무작위 결측에 해당한다. 이러한 비무작위 결측(MNAR)의 보완을 위해서는 현장의 실제 값에 관한 추가적인 조사를 하여야만 정확한 추론이 가능하다. 그러나 몇 가지 가정을 도입한다면(예를 들어, 같은 지역 내 동일 용도 건축물 내에서는 결측데이터와 비결측데이터 간 차이가 없다고 가정시), 그러한 가정이 참이라는 전제 하에 결측치를 대체할 수 있다. 이어지는 장에서는 이러한 한계 하에서 연구자가 활용할 수 있는 결측 데이터 보완 방안을 모색한다.

2. 용적률 결측치로 인한 데이터 품질 보완방안

1) 결측치 유형 분류

건축물대장의 용적률은 '용적률 산정 연면적'을 '대지면적'으로 나누어 계산된다. 이 두 개 변수의 결측 여부에 따라 용적률 결측 데이터를 아래 표와 같이 네 가지 경우로 범주화하면(<표 4> 참조), 용적률 결측이 오로지 대지면적 변수의 부재로 인해 발생한 경우(case 1)가 60.4%, 대지면적과 용적률 산정 연면적이 모두 부재한 경우(case 2)가 35.7%로 두 개 유형이

<표 3> 건축물 주용도와 용적률 결측 탐색 결과

용도 Use	A. 용적률 기입값 개수 no. of filled FAR for each use	B. 용적률 결측치 개수 no. of missing FAR for each use	C. 각 용도별 건축물 데이터 수 no. of building records for each use C=A+B	D. 각 용도 용적률 결측율 (%) percent of missing out of no. of building records D=B/C	E. 결측치 내 각 용도의 구성비율 (%) percent of missing for each use out of total no. of missing E=B/sum(B)	F. 전체 데이터 대비 각 용도의 구성비율 (%) percent of records for each use out of total no. of building records F=C/sum(C)	G. 비율 차이 (%p) percent point diff. G = E-F
전체 Total	329,721	190,801	520,522	36.7%	100.0%	100.0%	0.0%
단독주택 Detached Houses	145,326	129,694	275,020	47.2%	68.0%	52.8%	15.1%
제1종근린생활시설 Class 1 neighborhood living facilities	26,858	26,524	53,382	49.7%	13.9%	10.3%	3.6%
제2종근린생활시설 Class 2 neighborhood living facilities	42,629	15,271	57,900	26.4%	8.0%	11.1%	-3.1%
공동주택 Multi-family housing	92,238	13,561	105,799	12.8%	7.1%	20.3%	-13.2%
공장 Factories	1,027	913	1,940	47.1%	0.5%	0.4%	0.1%
업무시설 Business facilities	8,361	863	9,224	9.4%	0.5%	1.8%	-1.3%
근린생활시설 Neighbourhood living facilities	379	527	906	58.2%	0.3%	0.2%	0.1%
숙박시설 Lodging facilities	1,237	519	1,756	29.6%	0.3%	0.3%	-0.1%
창고시설 Warehouses	536	505	1,041	48.5%	0.3%	0.2%	0.1%
노유자시설 Facilities for older persons and children	2,240	497	2,737	18.2%	0.3%	0.5%	-0.3%
판매시설 Sales facilities	401	393	794	49.5%	0.2%	0.2%	0.1%
교육연구시설 Education and research facilities	3,041	388	3,429	11.3%	0.2%	0.7%	-0.5%
종교시설 Religious facilities	1,882	337	2,219	15.2%	0.2%	0.4%	-0.2%
자동차관련시설 Facilities for motor vehicles	1,044	180	1,224	14.7%	0.1%	0.2%	-0.1%
문화및집회시설 Facilities for cultural activities and assembly	592	91	683	13.3%	0.0%	0.1%	-0.1%

※ 지면 한계상 전체 데이터 대비 해당 건축물 개수가 0.1% 미만에 해당하는 용도는 생략하였다.

전체 결측치의 96.1%를 차지하고 있다. case 3은 '용적률 산정 연면적' 변수가 결측인 경우, case 4는 '대지면적'과 '용적률 산정 연면적' 모두 기재되어 있으나 용적률 변수만 결측인 경우로 각각 3.7%, 0.2%를 차지한다.

<표 4> 건축물대장 용적률 결측 데이터 유형

용적률 결측치 유형	행수	비율(%)
case 1: '용적률 산정 연면적'은 기재되어 있지만 '대지면적'이 0으로 기재되어 있는 경우	115,291	60.4
case 2: '대지면적'과 '용적률 산정 연면적' 모두 0으로 기재되어 있는 경우	68,208	35.7
case 3: '대지면적'은 기재되어 있지만 '용적률 산정 연면적'이 0으로 기재되어 있는 경우	7,012	3.7
case 4: '대지면적'과 '용적률 산정 연면적' 모두 기재되어 있으나 용적률은 0으로 기재된 경우	290	0.2
총 용적률 결측치 수	190,801	100.0%

2) 결측치 보완 방안

비무작위 결측(MNAR) 유형의 결측 문제가 있을 때, 가장 이상적인 해결 방안은 참값이 기재되어 있는 다른 자료와의 연계를 통해 해당 참값을 채워 넣고, 불가한 케이스는 추가 조사를 통해 데이터 품질을 높이는 것이다. 건축물 과세대장, 부동산등기부등본 등과 같은 다른 자료를 통합·연계하여 관리하는 방안은 그동안 필요성이 꾸준히 제기되어 왔다(강영옥·이주일, 2005; 강혜진·정종호, 2019). 그러나 수많은 건축물의 통합·관리 작업이 단기간 내 이루어지기 어려운 상황에서, 본 연구는 당장 임박한 분석을 진행해야 하는 연구자가 취할 수 있는 데이터 품질 보완방안을 용적률과 관련된 다른 변수와의 관계를 통해 고찰·도출하였다.

먼저 case 1은 용적률 계산을 위한 분모 값이 되는 대지면적 변수가 0으로 기재되어[5] 발생한 결측치이다. 이 경우, '대지면적'의 값만 확보되면 '용적률 산정 연면적'을 '대지면적'으로 나누어 용적률 참값을 계산해낼 수 있다. 대지면적 값의 확보는 건축물대장에 구체적인 건축물의 주소가 기입되어 있기에, 이 주소 위치를 매개로 연속지적도[6]에서 필지별 면적 값을 가져와 대지면적 0 값을 대체할 수 있다. 만약 연속지적도상 주소와 면적 데이터셋이 구비되어 있지 않더라도, GIS 상에서 면적을 자동으로 계산하고, spatial join을 수행한다면 최소한의 자원 투입으로 대지면적 결측을 모두 채울 수 있다. 이렇게 대지면적 결측이 전부 채워지면, 가장 난제인 결측 case 2의 예측도 case 3과 동일하게 접근하여 결측을 보완할 수 있다.

case 3의 경우는 대지면적이 이미 확보되어 있으므로, 같은 지역 내 용적률이 기재되어 있는 동일 용도 건축물 자료의 '연면적' 변수와 '용적률 산정 연면적' 변수 간 비율(P)을 구한 후, 존재하는 '연면적' 변수에 이 비율(P)을 곱하여 결측인 '용적률 산정 연면적'을 추정하여 산입해 넣음으로써 용적률을 계산할 수 있다. 용적률 결측 자료들의 '연면적' 변수가 결측이 아닐 경우, 해당 지역의 건축물 용도별로 '연면적'과 '용적률 산정 연면적 변수' 간 비율이 일정하다는 가정하에, '연면적' 변수에 해당 지역 비결측 건축물의 해당 비율을 적용하여 '용적률 산정 연면적' 변수를 채우는 방식으로 일정 부분 해결할 수 있다. 만약 데이터의 '연면적' 변수가 결측인 경우라 하더라도, '대지면적', '건폐율', '층수' 세 변수의 곱을 통해 '연면적' 값을 계산 후 위 방법을 활용할 수 있다.

case 4는 '대지면적' 변수와 '용적률 산정 연면적' 변수가 모두 기재되어 있는 경우이므로, 단순히 '용적률 산정 연면적' 변수값을 '대지면적' 변수로 나누어서 용적률을 계산해 넣을 수 있다. 이 경우에 해당하는 용적률 결측치는 0.2%에 불과하지만, 가장 간단히 해결할 수 있는 경우이다.

위와 같은 방법을 활용하면 건축물대장의 용적률 결측치를 단순히 제외하고 분석하는 것보다 데이터 품질이 훨씬 개선된 현황 분석이 가능할 것이다. 이는 건축물데이터의 통합관리 체계가 구축되거나, 건축물대장 결측치의 정비가 이루어지기 전까지 연구자/분석자가 용적률 결측치 제거시 발생하는 편향의 최소화를 위해 단기에 쉽게 적용할 수 있는 방법일 것으로 판단된다.

그럼에도 불구하고 남을 수 있는 한계점은 다음과 같다. 첫째, 하나의 필지에 하나의 건축물이 있지 않고 건축물 하나 아래에 두 개 이상의 필지가 존재하거나,

5) 현재 서울시 내 건축물대장에는 대지면적이 0으로 표기되어 있는 데이터가 42.36%를 차지(592,440개 중 250,985개)한다.

6) 국가공간정보포털에 공개되어 있다.

하나의 건축물이 다른 건축물과 더불어 하나의 큰 필지에 위치하는 경우, 지적 공간자료를 연계하여 대지면적 결측값을 업데이트하는 과정에서 오차가 발생할 수 있다. 따라서, 핵심 기초자료인 '대지면적' 변수는 건축물대장 데이터의 정비 시 가장 시급히 수정이 이루어질 필요가 있다.

둘째, '연면적' 변수와 '용적률 산정 연면적' 변수 간 비율을 각 건축물의 용도나 노후도 등에 따라 세분된 가장 동질적인 건축물 집단 중 두 변수 모두 결측이 아닌 데이터로부터 구한 후 적용하더라도, 각 건축물만의 고유한 설계상 특징이 있을 수 있으므로 오차가 발생할 여지가 있음을 염두에 둘 필요가 있다. 따라서 데이터 이용자들은 이러한 오차를 최대한 줄일 수 있는 동질적인 건축물 그룹을 찾아서 활용할 필요가 있다. 건축구조, 사용승인연도, 지역과 같은 변수들을 함께 활용해서 필터링 후 해당 건축물들의 '연면적' 변수와 '용적률 산정 연면적' 변수 간 비율 수치를 활용한다면 오차를 최소화할 수 있을 것이다.

V. 결론

그동안 건축물대장 자료의 오류와 누락으로 인한 데이터 활용 상 어려움이 꾸준히 제기되어 왔음에도 불구하고, 이와 관련된 연구는 건축물 데이터를 구축하고 관리하는 주체 위주로 이루어져 왔다. 이에 본 연구는 건축물대장이 보유한 변수 중 가장 그 중요성이 높음에도 불구하고 현재 그 활용에 어려움을 초래하는 용적률 결측치에 초점을 두고, 용적률과 다른 변수 간 관계를 탐색하고, 데이터 이용자의 관점에서 데이터 품질을 개선하여 분석에 활용할 수 있는 방안을 고찰하였다.

분석 결과, 건축물대장의 용적률 결측은 결측의 세 가지 유형 중 비무작위 결측(NMAR)에 해당하며, 이로 인해 단순히 결측치를 제외한 분석으로는 현상의 심각한 왜곡을 피할 수 없음을 밝혔다. 구체적으로 서울시 내 건축물대장에서 최소 36.7%는 용적률 정보가 0으로 기입된 채 누락되어 있으며, 이러한 누락 비율은 1990년 이전 승인된 노후 건축물에서 현저히 높게 나타난다. 또한 용적률 결측치의 81.9% 이상은 그 용도적 특성상 저층·저밀도로 개발되는 단독주택과 제1종근린생활시설을 주용도로 하는 건축물에서 발견되어, 용적률 결측 여부가 용도 및 용적률 변수와 결코 독립적이지 못함을 드러내었다. 단독주택과 제1종근린생활시설은 통상 100~200% 이하의, 다른 용도의 건축물에 비해 비교적 낮은 용적률로 개발되기 때문이다. 따라서 결측치를 제외하고 분석하기보다는 결측치의 75% 이상을 차지하는 1972년~1990년 기간 승인된 노후 건축물의 각 용도별 용적률 평균값[7]으로 결측치를 대체하여 활용한다면, 편향을 대폭 줄일 수 있을 것이다.

본 연구는 이어서 이러한 건축물대장 자료의 한계에도 불구하고, 연구자가 데이터 품질을 높이기 위해 직접 다른 변수와의 관계 속에서 용적률 수치의 편향을 최소화할 방법을 고안, 제시하였다. 건축물대장 자료에서 용적률 계산의 재료가 되는 '대지면적' 변수와 '용적률 산정 연면적' 변수를 살피고, 이미 공개되어 있는 연속지적 공간 자료와 연계함으로써 60% 이상, 용적률이 올바로 기재되어 있는 다른 동질적 건축물 자료로부터 동일 용도의 '연면적' 변수와 '용적률 산정 연면적' 변수 간 비율을 구한 후, 존재하는 '연면적' 변수에 이 비율을 곱하여 결측인 '용적률 산정 연면적' 근사값을 생성함으로써 35% 이상의 용적률 결측을 대체할 수 있음을 밝혔다.

이상 밝혀낸 건축물대장 내 용적률 결측치 분포의 패턴과 개선방안에도 불구하고 본 연구는 서울시 자료만을 분석 대상으로 한 점, 실제 누락 건축물의 용적률 참값은 알 수 없어 실제 데이터 참값에 얼마나 수렴했는지를 구체적 수치로 제시하지는 못하였다는 점에서 다른 지역에서의 적용 및 일반화에는 유의할 필요가 있다. 이에 향후 용적률 결측 오류 건축물이 다수 집중된 일부 표본 지역을 대상으로 해당 용적률 참값을 조사하고, 본 연구에서 밝힌 패턴을 적용한 개선안을 적용 시 편향이 개선되는 수준을 계량화하는 연구가 진행된다면, 더욱 면밀한 검증이 가능할 것이다. 또한, 향후 노후 건축물의 재건축과 해당 지역들의 정비사업이 진행됨에 따라 건축물대장 자료 오류의 개선 추이에 관하여 연구자들이 관심을 두고 추적 관찰하는 것도 필요하다.

한편 건축물대장 관리 주체는 동일 문제가 반복되지

7) 1972년~1990년 기간 사용 승인된 건축물의 용적률 평균은 128.28%이다(승인연도가 0으로 기재되어 있는 경우 및 1948년(대한민국 정부 수립일)보다 더 이전 연도로 기재된 경우는 제외).

않도록 ‘대지면적’, ‘용적률 산정 연면적’, ‘용적률’과 같은 핵심 변수에 대해서는 자료 갱신과 함께 꾸준히 오류 및 결측을 수정·정비해나갈 필요가 있다. 특히, 건축물대장 내 ‘대지면적’ 변수는 건축물 위치를 기준으로 다른 지적 자료와 연계작업만 진행하면 간단히 해결할 수 있고, 이는 무려 60%의 결측치를 해결할 수 있다. 따라서 건축물대장 내 0으로 기재된 ‘대지면적’ 변수의 결측치 해결이 ‘용적률’ 변수 결측 오류 문제 해결의 첫걸음이 될 것이다.

과거 승인 및 등록된 건축물 정보를 개선하는 한편, 앞으로 신규로 승인·등록된 건축물대장에서 용적률과 같은 핵심 변수의 오류·누락이 발생하지 않도록 관련 부서의 구체적인 전략 수립과 실행이 이루어져야 꾸준한 데이터 품질 개선을 기대할 수 있을 것이다. 또한, 건축물대장 외에 해당 변수들의 참값을 알 수 있는 다른 자료(건축물 과세대장)들을 함께 민간에 전면 공개하여 연계 활용할 수 있게 한다면, 큰 비용이 소요되는 대규모의 조사 사업이 없더라도 단기간에 민간 연구자와 통계 이용자들이 더욱 신뢰성 있는 분석과 토지이용계획을 수행하기 위한 기초자료로 활용할 수 있을 것이다.

논문접수일 : 2023년 11월 14일
논문심사일 : 2024년 1월 31일
게재확정일 : 2024년 3월 13일

참고문헌

1. 강영옥 · 이주일, “건축물 정보 정비방안”, 서울시정개발연구원, 2005
2. 강혜진 · 정종호, “서울시 건축물데이터 통합관리 방안”, 서울기술연구원, 2019
3. 국토교통부, 건축물대장 기초자료 정비사업(3차) 감리. 국토교통부 입찰안내, 2009. https://www.molit.go.kr/USR/tender/m_83/mng.jsp?ID=12794
4. 국토교통부, 국토교통부 건축물대장 표제부(2023년 9월). 건축데이터 민간개방 시스템, 2023. https://open.eais.go.kr
5. 김수현 · 최창규, “용적실현비(A-FAR)에 영향을 미치는 용도지역별 대지특성에 대한 분석”, 「국토계획」 제54권 제2호, 대한국토 · 도시계획학회, 2019, pp. 33-45
6. 김수현 · 최창규, “주택유형별 용적률 실현에 영향을 미치는 대지특성에 대한 분석”, 「국토계획」 제56권 제4호, 대한국토 · 도시계획학회, 2021, pp. 1-14
7. 김승범, “건축물대장 원시데이터의 관계 구조와 탐색적 분석을 통한 데이터 활용”, 「한국문화공간건축학회 논문집」 통권 제50호, 2015, pp. 110-120
8. 김정옥 · 김지영 · 배영은 · 유기윤, “건축물대장을 이용한 수치지도 속성정보의 효율적 갱신방안: 새주소상업의 건물번호 이용을 중심으로”, 「한국측량학회지」 제26권 제3호, 2008, pp. 275-284
9. 김형보, “상업지역 건축물의 용적률 실현정도에 관한 실증분석”, 「국토계획」 제33권 제3호, 대한국토 · 도시계획학회, 1998, pp. 89-104
10. 박재빈 · 임하나 · 김수현 · 최창규, “다세대 · 다가구 우세지역과 아파트 우세지역의 건폐율과 용적률이 열섬효과에 미치는 영향분석”, 「국토계획」 제52권 제7호, 대한국토 · 도시계획학회, 2017, pp. 159-176
11. 백태경 · 김영훈 · 최정미, “지적도와 건축물대장 연계를 통한 토지이용 DB구축에 관한 연구”, 「한국지리정보학회지」 제7권 제4호, 2004, pp. 133-142
12. 신상영 · 장영희, 「서울시 주택재고 산정을 위한 데이터기반 연구」, 서울시정개발연구원, 2006
13. 윤병훈 · 남진, “서울시 개발밀도 실현율에 영향을 미치는 요인에 관한 연구”, 「국토계획」 제48권 제5호, 대한국토 · 도시계획학회, 2013, pp. 177-196
14. 윤병훈 · 남진, “구조방정식을 활용한 개발밀도 영향요인간 상호작용 분석”, 「국토계획」 제49권 제7호, 대한국토 · 도시계획학회, 2014, pp. 81-96
15. 윤상복 · 김형보 · 채성주, “신, 구도심부 용적률 실현의 특성과 영향요인에 관한 비교연구”, 「국토연구」 제40권, 2004, pp. 19-34
16. 윤혜림 · 남진, “서울시 개발밀도에 영향을 미치는 요소의 변화에 관한 연구 - 일반주거지역 종세분화 전, 후(2002-2011) 비교를 중심으로”, 「국토계획」 제48권 제3호, 2013, pp. 165-180
17. 이성화, “건축물대장 등록정보의 논리오류 유형 연구”, 「한국지적학회지」 제26권 제1호, 2010, pp. 145-161
18. 이윤상 · 남진, “서울시 상업지역의 개발밀도에 미치는 영향요인 연구”, 「국토계획」 제48권 제8호, 대한국토 · 도시계획학회, 2014, pp. 63-77
19. 이인성 · 임상준 · 김충식, “필지형상이 개발밀도에 미치는 영향 분석 - 서울시 강동구 천호 · 암사 지구단위계획구역을 대상으로 -”, 「한국도시설계학회지」 제10권 제4호, 2009, pp. 151-162
20. 이주일 외, 「서울형 용도지역 체계재편 실행계획 수립」, 서울연구원, 2022
21. 이지은, “서울시 지역특성이 실현용적률에 미치는 영향에 관한 연구”, 한양대학교 도시대학원 박사학위논문, 2011
22. 이희정 · 김기호, “서울시 일반주거지역 세분을 위한 주거지 밀도분포 특성 연구(1) - 일반주거지역 밀도결정요인 분석을 중심으로 -”, 「국토계획」 제36권 제5호, 대한국토 · 도시계획학회, 2001, pp. 73-88
23. Buuren, Stef van. Flexible Imputation of Missing Data. Second edition. Boca Raton, FL: CRC Press, 2018. Print
24. Jebb, Andrew T., Scott Parrigon, and Sang Eun Woo, “Exploratory Data Analysis as a Foundation of Inductive Research,” Human Resource Management Review, Vol. 27 No. 2, 2017, pp. 265-276
25. Rubin, D. B. “Inference and Missing Data,” Biometrika, 63, 1976, pp. 581-592
26. Tukey, John W., “*Exploratory data analysis*,” Reading, Massachusetts: Addison Wesley, 1977
27. Tukey, John W., “We Need Both Exploratory and Confirmatory,” The American Statistician, vol. 34, no. 1, 1980, pp. 23-25. JSTOR, https://doi.org/10.2307/2682991

<국문요약>

건축물대장 데이터 품질 개선을 위한 용적률 결측오류 분석 및 결측치 처리방안

이 지 은 (Lee, Jieun)

이 연구는 대표적인 건축물 자료인 건축물대장에 포함된 용적률 결측 오류 문제를 고찰하였다. 분석 결과, 서울시 내 건축물의 36.7%는 용적률이 0으로 기입되어 있으며, 노후도가 현저히 높고 단독주택 및 제1종 근린생활시설 용도의 건축물에서 발생 빈도가 높은 비무작위 결측인 것으로 나타났다. 이는 단순히 결측치를 제외한 분석은 사실을 왜곡시킬 수 있음을 시사한다. 이에 본 연구는 결측 데이터로 인한 편향을 최소화하는 방안을 제시하였다. 첫 번째로, 연속지적도 공간자료와 연계하여 '대지면적(㎡)' 변수의 결측을 해결함으로써 60.4%의 용적률 결측값을 계산된 용적률로 대체할 수 있다. 또한, 용도 및 사용승인 연도에 따라 지역 내 동질적인 비결측 건축물 집단의 '용적률 산정 연면적'의 '연면적'에 대한 비율을 활용하여 용적률 결측 건축물 집단의 '용적률 산정 연면적' 계산에 활용함으로써, 35.7%의 용적률 결측을 추정 및 대체할 수 있다. 장기적으로는 '대지면적', '용적률'과 같은 핵심 변수의 결측치를 중심으로 데이터를 정비하고, 관련 부동산 데이터베이스를 대중에 공개하며, 건축물대장에 신규로 추가되는 건축물에 대한 데이터품질 관리 전략 수립이 필요할 것이다.

주 제 어 : 용적률, 건축물대장, 결측치, 부동산빅데이터, 데이터품질