Journal of the Korean Association of Geographic Information Studies, 26(4) 2023, pp.97-115 ISSN 1226-9719(Print)
https://doi.org/10.11108/kagis.2023.26.4.097 ISSN 2287-6952(Online)

해석가능한 기계학습을 적용한 소지역 인구 추정에 관한 연구: 부산광역시를 대상으로*

김유현[1] · 김동현[2※]

A Study on the Population Estimation of Small Areas using Explainable Machine Learning: Focused on the Busan Metropolitan City*

Yu-Hyun KIM[1] · Donghyun KIM[2※]

요 약

최근 저출산, 고령화 등 인구의 구조가 급격히 변화하고 있고 인구 분포의 불균등성이 확대되고 있는 시점에서 인구 추정 방식의 변화가 요구되고 있으며 소지역 단위에서 보다 정확한 추정이 요구되고 있다. 본 연구는 이러한 인구 추정 방식 변화 요구에 대응하기 위해 부산광역시를 대상으로 해석가능한 기계학습 방법을 적용하여 500m 격자 단위에서 2040년 인구를 추정하는 것을 목적으로 하고 있다. 해석가능한 기계학습의 방법과 코호트 요인법을 각각 적용하여 격자별 인구추정 결과를 비교해본 결과, 기계학습 방법이 인구 구조 변동에 영향을 미칠 가능성이 있는 여러 변수의 조합 반영이 가능하여 보다 낮은 오차를 도출함으로써 소지역과 같이 인구 변화폭이 큰 지역의 추정에 있어 적용력이 높음을 확인하였다. 인구감소시대에 과대추정된 인구 값은 도시계획에서 투자의 비효율성과 특정 부문에 대한 과잉 투자에 따른 타 부문에서의 질적 저하와 같은 문제를 일으킬 가능성이 높으며, 과소추정된 인구 값 역시 도시의 축소를 가속화시켜 삶의 질을 저하시키는 문제를 초래하므로 적절한 인구 추정 방법과 대안을 마련해야 할 필요가 있을 것으로 판단된다.

주요어 : 인구 추정, 소지역, 코호트 요인법, 기계학습, 해석가능한 기계학습

2023년 11월 02일 접수 Received on November 02, 2023 / 2023년 11월 08일 수정 Revised on November 26, 2023 / 2023년 11월 16일 심사완료 Accepted on November 16, 2023

* 이 논문은 2023년 대한민국 교육부와 한국연구재단의 지원(NRF-2020S1A3A2A01095064)과 국토교통부의 스마트시티 혁신인재육성사업으로 지원을 받아 수행되었습니다.

1 부산대학교 공과대학 도시공학과 석사과정 / Graduate Student, Department of Urban Planning and Engineering, Pusan National University

2 부산대학교 공과대학 도시공학과 부교수 / Associate Professor, Department of Urban Planning and Engineering, Pusan National University

※ Corresponding Author E-mail: donghyun-kim@pusan.ac.kr

ABSTRACT

In recent years, the structure of the population has been changing rapidly, with a declining birthrate and aging population, and the inequality of population distribution is expanding. At this point, changes in population estimation methods are required, and more accurate estimates are needed at the subregional level. This study aims to estimate the population in 2040 at the 500m grid level by applying an explainable machine learning to Busan in order to respond to this need for a change in population estimation method. Comparing the results of population estimation by applying the explainable machine learning and the cohort component method, we found that the machine learning produces lower errors and is more applicable to estimating areas with large population changes. This is because machine learning can account for a combination of variables that are likely to affect demographic change. Overestimated population values in a declining population period are likely to cause problems in urban planning, such as inefficiency of investment and overinvestment in certain sectors, resulting in a decrease in quality in other sectors. Underestimated population values can also accelerate the shrinkage of cities and reduce the quality of life, so there is a need to develop appropriate population estimation methods and alternatives.

KEYWORDS : Population Estimation, Small Area, Cohort Component Method, Machine Learning, Explainable Machine Learning

서 론

인구 자료는 도시계획 및 각종 정책 자료의 근간이자 필수적인 선행지표이다. 인구 구조의 변화는 주택, 경제, 복지, 환경 등 모든 분야에 영향을 줄 뿐만 아니라 예측된 인구 자료를 바탕으로 도시계획시설의 규모나 서비스 공급 기준이 결정되기 때문이다. 부정확한 인구 예측은 유휴자본 또는 혼잡비용을 발생시켜 사회적 비효율을 야기하고 도시 전체의 질을 떨어뜨린다.

한 지역의 인구 변화는 국가 전체의 영향요인 뿐만 아니라 지역의 인구 구조와 다양한 원인이 복합적으로 상호작용하여 나타난다. 미래 인구의 변화와 구조에 대한 논의는 도시계획 뿐만 아니라 정책의 바탕이 되는 가장 기본적인 자료이다. 성장의 시기에 추정된 미래 인구는 각종 인프라와 도시의 확장을 위한 택지개발, 수요추정 등의 근거자료가 되어 왔으며 인구의 감소보다는 성장에 초점을 두고 추정되어 왔다. 최근 저출산, 고령화 등 인구의 구조가 급격히 변화하고 있고 인구 분포의 불균등성이 확대되고 있는 시점에서 인구 추정 방식의 변화가 요구되고 있으며 소지역 단위에서 보다 정확한 추정이 요구되고 있다.

전 세계 많은 국가의 주요 도시들은 인구 구조적인 측면에서 출산율 감소와 기대수명 증가로 인해 고령화와 인구감소 현상을 경험하고 있다(Baker *et al.*, 2014; Choi *et al.*, 2019; Lee and Cho, 2020). 하지만 동일 지자체 내에서도 인구 구조 변화가 다르게 나타나고 있다(Lee, 2019). 일부 신규 택지개발 지역들에서는 지속적인 인구 유입과 새로운 인프라 투자 등으로 고밀도의 사회활동이 이루어지는 반면 구도심이나 농어촌지역 등은 인구감소와 고령화 현상이 동시에 나타나고 있다. 미시적 공간 단위에서의 인구집계에 대한 MAUP(modifiable area unit problem)의 이슈는 꾸준히 제기되어 왔으며, 이는 보다 세밀한 지역에 대한 인구 연구의 필요성으로 이어진다(Goodchild *et al.*,

1993; Kim et al., 2022).

이러한 관심사를 반영하여 최근 소지역 인구 추정이 하나의 연구 주제로 논의되고 있으며, 소지역 인구 예측의 정확도를 높일 수 있는 연구 방법 또한 계속해서 발전하고 있는 추세이다(Davis, 1995; Swanson *et al.*, 2010; Wilson, 2011; Baker *et al.*, 2013; 2014; Inoue, 2017; Lee and Lee, 2006; Choi *et al.*, 2019; Lee and Cho, 2020; Lee, 2019). 기존 인구추정과정에서 사용하던 통계 기법들은 모형의 한계로 인하여 소지역 인구 추정 시 과소추정 하거나 과대추정하는 문제가 발생한다(Choi *et al.*, 2019; Lee, 2019). 이를 해결하기 위해서 기계학습 기법을 적용하는 방법이 최근 대안적으로 논의되고 있다.

본 연구의 목적은 부산광역시를 대상으로 해석가능한 기계학습의 방법을 적용하여 500m 격자 단위에서 2040년 인구를 추정하고자 하는 것이다. 부산광역시는 우리나라에서 두 번째로 큰 대도시이지만 고령화 정도가 심화되며 인구감소가 지속되고 있는 대표적인 지역이다. 특히 일부 소규모 지역의 경우는 근시일 내 지역소멸이 전망되고 있다. 하지만 도시기본계획에서 설정하고 있는 인구는 여전히 과대추정되고 있으며 소지역 단위에서는 더욱 그 차이가 크게 나타나고 있다. 본 연구는 기존의 인구추정 방법인 코호트 요인법과 대안적 방법인 해석가능한 기계학습의 방법을 비교하여 시사점을 도출한다. 연구는 다음과 같은 순서로 진행된다. 먼저 소지역 인구추정 및 해석가능한 기계학습과 관련한 연구동향을 살펴보고 해석가능한 기계학습의 방법과 코호트 요인법을 적용한 격자별 인구추정의 결과를 비교한다. 마지막으로는 연구의 시사점과 향후 연구방향을 제시한다.

선행연구 고찰

1. 소지역 단위 인구 전망

인구추정과 관련된 주요 논의 주제 중 하나는 소지역 인구 추정에 대한 것이다(Cai, 2007). 소지역 인구 추정은 정부 및 기업이 인프라 투자 및 지역 서비스 제공 시 보다 정확한 투자금액을 제시하거나 서비스가 집중적으로 제공되어야 할 특정 지역을 가리킬 수 있다는 점에서 유용하다. 하지만 공간 해상도를 높임에 따라 반비례하는 예측 정확성과 급속한 변화를 겪고 있는 인구구조적 환경 등으로 인해 소지역 인구추정 연구의 유용성과 필요성에도 불구하고 관련 연구는 아직 미비한 상태이다(Wilson *et al.*, 2023). 소지역 인구 추정과 관련된 연구가 어려운 이유는 두 가지로 논의된다. 먼저 코호트 요인법과 같은 대부분의 인구 예측 모델은 거시지리적 범위 내의 인구 구조를 다룰 수 있도록 개발되었다. 이러한 예측 모델을 소지역에 적용 시 공간 해상도가 맞지 않아 연구 결과의 신뢰성을 보장하기 어렵다. 다음으로 소지역의 경우 전통적 인구 예측 모델이 고려하지 못하는 사회환경 맥락과 토지이용과 같은 비인구학적 요인이 변수로서 더 크게 작용하지만 자료의 한계가 존재한다(Chi *et al.*, 2011). 하지만 최근 빅데이터 분야의 발전과 기계학습의 새로운 방법으로 인해 소지역 인구 추정 분야의 잠재력이 제고되고 있다.

인구추정에 있어 소지역에 대한 보편적 정의는 없으나, 일반적으로 소지역이란 데이터 확보가 가능한 지역 분류 중 가장 세밀한 공간 단위로 간주된다(Smith and Morison, 2005). 국외 선행연구 내 대표적인 소지역의 예로는 미국의 센서스 트랙(census tracts), 호주의 Statistical Areas Level 2 지역, 영국의 워드(ward)와 집계구(census super output areas) 등이 있으며, 해당 지역들의 거주 인구는 대부분 20,000명 미만이다(Wilson, 2015). 국내 소지역 관련 연구는 대부분 시 · 군 · 구 단위에서 이루어졌으며, 해당 연구들은 인구이동으로 인한 인구 예측 정밀성을 높이기 위한 과정에 집중하고 있다(Cho and Lee, 2011). 하지만 보다 세밀한 지역에 대한 인구 예측 수요가 증가하면서 소지역 단위는 점차 읍 · 면 · 동, 집계구, 격자 등 점점 더 좁은 범위로 좁혀지고 있다. 최근 들어서는 특히 격자단위 인구 예측 방법에 대한 연구가

국내외적으로 진행되고 있으며, 이에 따라 국내에서는 국토지리정보원, 이동통신사와 같은 공공 및 민간부문에서 격자단위 통계 데이터를 생산하여 해당 연구 분야를 지원하고 있는 추세이다(Lee, 2019).

2. 소지역 단위 인구 전망 예측 방법론

소지역 인구추정 연구는 예측의 정확성을 높이기 위해 다양한 방법론이 사용되고 있다. 관련 선행연구에서 사용한 방법론들은 크게 기초모델, 코호트 요인법, 공간통계모델, 시나리오 기반 모델, 인공지능 기법으로 분류된다(Lee, 2019). 기초모델은 경향외삽법과 비율법으로 다시 구분되며, 해당 모형들은 최소한의 데이터를 요구하고 계산과정이 단순하다는 장점이 있지만 그만큼 예측 기간이 증가할수록, 지역이 작아질수록 예측 오차가 증가하고 인구 구조의 변화를 설명하지 못한다는 한계가 존재한다. 코호트 요인법은 가장 보편적으로 활용되는 인구추정 방법론으로, 출생 · 사망 · 이동이라는 3가지 인구변동요인에 대한 인구균형방정식을 만들어 장래인구를 추정하는 방식이다. 코호트 요인법의 경우 과거에는 단순히 기준인구에 출생자, 사망자, 전입 · 전출자 수를 고려한 회계방식에 머물렀으나, 최근에는 확률모형 및 회귀모형을 적용하여 고도화되고 있다(Lee, 2019). 코호트 요인법은 인구 변화 패턴을 잘 포착하여 그 신뢰성을 전세계적으로 인정받아 보편적으로 사용되는 인구 전망법이지만, 소지역단위에 적용하기에는 한계가 존재한다. 코호트 요인법이 고려하는 인구변화요인을 미시적 단위에서 계산하기에 무리가 있기 때문이다. 공간통계모델 및 시공간 회귀모형의 경우 기존의 인구추정 방법론들이 간과하였던 인접 공간 간의 상호관계성을 반영할 수 있다는 장점이 있지만, 선행연구 결과에 따르면 단순회귀와 비교하였을 경우 예측오차를 크게 줄이지 못한다는 점에서 향후 방법론 개발이 더욱 이루어져야 함을 알 수 있다(Chi and Voss, 2011). 시나리오 기반 모델은 앞서 언급한 방법론들과는 다르게 인구변화와 관련한 이론적 배경 및 사회경제 · 물리환경적 맥락을 기반으로 장래 변화를 시뮬레이션하거나 미래 인구를 예측하는 기법이다. 주로 토지이용 및 미래 개발계획, 각 지자체가 공표한 도시기본계획 등을 따라 시나리오를 구성한 뒤, 시나리오별 장래인구를 추정하고 있음을 알 수 있다(Triantakonstantis and Mountrakis, 2012; Ford *et al.*, 2019; Chen *et al.*, 2020). 이러한 시나리오 기반 인구추정 방식은 토지이용계획 및 다양한 도시인구 변수 반영이 가능하여 예측값의 오차를 줄일 수 있다는 장점이 있으나 코호트 요인법과 같은 기존 인구통계모델과 마찬가지로 소지역 단위 자료 구축 단계 상 한계가 존재한다. 뿐만 아니라 인구 전망에 있어 인구학적 요인 및 자연적 인구 증감요인을 상대적으로 등한시하고 있다는 한계가 있다(Lee, 2019).

기계학습 기법이 인구추정에 적용되기 시작한 것은 최근 3~4년 전부터이다. 기계학습 기법을 이용한 인구추정의 연구는 대부분 기존 인구추정기법에 비해 기계학습 모델의 인구추정 성능이 유의미한지를 비교하는 내용이다(Riiman *et al.*, 2019; Chen *et al.*, 2020; Grossman *et al.*, 2022; Grossman, Wilson, and Temple, 2022). Riiman *et al.*(2019)은 미국 알리바마주 카운티 지역에 대한 ANN(arificial neural network) LSTM(long-short term memory) 결과값과 코호트 요인법 결과 값을 비교하였다. 이 연구는 모든 카운티의 데이터를 일괄 학습한 모델과 각 카운티별로 구분하여 학습한 모델의 두 가지 유형을 제시하였다. 가장 예측력이 높았던 경우는 각 카운티별로 ANN LSTM 모델을 학습한 경우였다. Chen *et al.*(2020)은 XGBoost, Random forest, Neural network, Support vector regression의 4가지 인공지능 알고리즘을 이용하여 2015년부터 2050년까지 5년 간격으로 인구분포를 추정하였다. 이 연구는 장래 인구 추정에 있어 토지이용, 도시 중심지까지의 거리 등 인구 분포 관련 공간적 변수 반영을 통해 현실성 있는 결과를 도출하고자 하였다. Grossman *et al.*(2022)은 인구 예측 시

여러 모델을 병합하여 사용하는 기법인 기계학습의 앙상블 기법의 유용성을 확인하고자 하였다. CSP(constant share of population), MEX(modified exponential model), VSG(variable share of growth model), LIN/EXP(linear/exponential), THETA, LGBM(light gradient boosting model)의 6가지 모델을 호주와 뉴질랜드의 SA2 지역에 적용하였다. 모델을 병합한 방식은 전체 모델 결과 값의 평균값을 도출하거나 각 모델의 최댓값 · 최솟값을 제외한 평균값을 도출하는 방식 두 가지를 이용하였으며 LGBM 모델을 포함하는 것이 인구 예측 모델의 정확도를 향상시키는 것으로 확인하였다. 이외 시계열적 인구 변화 추세 학습이 가능한 LSTM 알고리즘을 이용한 연구나 인구 변화 관련 과거 데이터 학습을 통해 장래 인구를 추정하는 기법인 기계학습 기법 등을 통한 연구들도 등장하고 있다(Weber, 2020; Grossman *et al.*, 2022).

3. 해석가능한 기계학습에 대한 논의

기계학습은 인간이 직접 논리를 구축하는 것이 아니라 학습 방식을 먼저 입력하고 기계가 스스로 논리를 만들어가게 제작하는 과정으로 정의할 수 있다. 조성된 알고리즘을 통해 의사결정을 할 수 있게 된 기계학습 산출물을 모델이라고 부른다(Ahn, 2020). 기계학습은 학습된 데이터들에 대한 전통적 통계학이 분석하기 힘든 비선형적 관계 도출을 통해 다양한 상황에 대한 유용한 적용이 가능하여 이미지 처리, 음성 인식, 텍스트 분석 등 다양한 분야에서 사용되고 있다(Caraviello *et al.*, 2006; Yan *et al.*, 2020). 하지만 기계학습의 연산 과정이 많을수록, 절차가 깊어질수록 학습 과정이 복잡해지며 오늘날 기계학습 모델 중 다수가 복잡한 매개변수로 이루어져 있다. 따라서 현대 기계학습의 의사결정 과정 중 다수는 학습 모델의 의사 결정 과정을 직접 이해할 수 없는 블랙박스(Black Box)의 성질을 가지고 있다(Ahn, 2020).

최근 해석가능한 기계학습 분야의 발전으로 기계학습의 블랙박스를 해석할 수 있는 방법론이 고안되기 시작하였다. 해석가능한 기계학습은 컴퓨터 시스템이나 인공지능 시스템은 복잡해지는 반면에 그것들의 자기 설명 기능에는 발전이 없었다는 점을 지적한 데서 발전하기 시작하였다(Van Lent *et al.*, 2004). 해석가능한 기계학습은 기계학습 모델의 블랙박스 성향을 인간이 이해할 수 있는 수준까지 분해하는 기술이다. 사회과학에서는 어떠한 이론이 대상이 이해할 수 있는 수준까지 변이되는 과정을 해석가능성(interpretability)이라고 부른다(Miller and Tim, 2018). 즉, 해석가능한 기계학습은 인공지능 모델이 특정 결론을 내리기까지 어떤 근거로 의사 결정을 내렸는지를 알 수 있게 설명 가능성을 추가하는 기법이다(Ahn, 2020). 이러한 해석가능한 기계학습의 대표적 방법론으로는 Skater, Local Interpretable Model-Agnostic Explanation(LIME), SHapley Additive exPlanations(SHAP), Partial Dependence Plot(PDP), 피처 중요도(Feature Importance) 등이 있다.

분석자료 및 방법

1. 대상 지역 및 분석의 공간 단위

본 연구는 대한민국 부산광역시를 연구의 공간적 범위로 설정하였다. 부산광역시는 수도권으로의 지속적인 인구 유출과 심각한 저출산으로 인해 총인구 규모, 합계출산율, 고령화율, 가구 구조, 지방소멸위험지수 등 주요 인구지표가 모두 부정적인 추세를 보이고 있다. 부산의 가파른 평균연령 증가 추세와 청년층 유출, 도심 내부가 공동화되는 현상 등은 지방소멸의 위기로 확대되고 있는 실정이다.

본 연구는 부산광역시를 500m 격자단위로 구획한 공간을 소지역으로 이용하였다. 국내의 격자 단위 인구 추정의 경우 1km 격자가 다수 이용되고 있다(Lee, 2019; Byeon *et al.*, 2023). 한 도시 내에서 인구추정을 위한 소지역을 정의할 때 1㎞ 격자는 비교적 큰 단위이다. 특히 구도심의 경우 면적이 좁아 소수의 격

자만이 해당될 가능성이 있다. 또한 산악지역이 많은 특성을 고려한다면 1㎞ 보다 세분화된 공간 단위가 적합하다.

2. 분석 자료, 변수 및 시간적 범위

본 연구는 국토지리정보원에서 제공하는 격자단위 인구자료를 이용하였다. 해당 자료는 행정안전부 주민등록시스템의 생년, 성별, 주소를 활용하여 개개인의 위치 좌표를 지오코딩한 뒤 격자단위로 집계하여 생산된 자료이다. 국토지리정보원은 2014년부터 매년 2회(6월, 10월) 행정안전부 주민등록시스템에 등록된 인구를 토대로 100m, 250m, 500m, 1㎞, 10㎞, 100㎞의 격자 통계를 공표하고 있다. 통계청 자료와 비교하였을 때 국토지리정보원이 다양한 격자단위 통계를 제공하고, 개인정보보호를 위한 데이터 처리 방식 상 값의 왜곡이 적다는 점에서 장점이 존재한다(Lee, 2019).

연구의 시간적 범위는 2020년~2040년이다. 이는 도시기본계획이 기준시점으로부터 20년을 기준으로 수립하도록 하고 있으며, 본 연구의 대상인 부산광역시 역시 2040년을 목표로 하는 도시기본계획을 수립하였기 때문이다. 본 연구는 각 방법론을 이용하여 인구추정을 진행하기에 앞서 2000~2020년 자료를 이용하여 2020년 인구를 예측하고, 실제 2020년 인구와 비교하여 오차를 계산한 다음 미래 추정을 진행하였다.

해당 연구 범위 내에서 부산광역시 2040년 인구를 추정하기 위해 기계학습 모델이 고려한 변수는 표 1과 같다. 변수는 크게 인구학적 특성의 변수와 사회경제적 변수로 구분되며, 인구학적 변수 중 생산가능 여자 인구 수와 고령 인구 수는 저출산·고령화로 인한 인구 구조 변화를 보기 위한 변수이다. 개별 주택 수와 노후 주택 비율은 개발 정도 및 질 좋은 인프라 환경이 인구 구조 변화에 미치는 영향을 보기 위한 변수이다. 추가적으로 인구 예측 모형은 이전 시점 인구를 기반으로 다음 시점의 인구를 예측하기에 전 시점 인구 수에 크게 의존할 것이라 가정하여 해당 인구 자료를 변수로 반영하였다.

인구 구조의 변화에는 무엇보다도 출산력이 미치는 영향력이 크다. 최근의 인구 감소 및 고령화현상에 대응하고, 적정 인구 규모를 유지하기 위해서는 일정 출산 수준을 유지하는 것이 필수적이다. 이러한 출산 수준은 근본적으로 가임여성인구 규모, 연령 구조, 혼인 수준 등에 의해 결정된다(Lee, 2017). 따라서 정확한 인구 규모 및 구조 추정에는 생산가능 여자 인구 규모 및 인구의 연령 구조 파악 과정이 선행되어야 하며, 이에 본 연구는 인구학적 변수로 생산가능 여자 인구 수와 고령 인구 수를 포함하였다.

인구 구조는 1차적으로는 출산 수준의 영향을 받으나, 이러한 출산율은 미시·거시적으로 사회경제적 현상에 영향을 받는다(Lee, 2017). 출산율 회복을 위한 사회경제적 대안들은 대부분 청년층의 고용 및 주거 안정에 중점을 두고

TABLE 1. Variables and Data

Variable		Definition	Data Sources
Dependent variable		Total Population	National Geographic Information Institute
Independent Variable	Demographical factors	Fertile Women	National Geographic Information Institute
		Aging Population	National Geographic Information Institute
	Housing Characteristics factors	Number of Individual Housing	National Geographic Information Institute
		Percentage of Old Housing	National Geographic Information Institute
	Model-based factors	Total Population at Previous Time	National Geographic Information Institute

있다. 이러한 정책 현황 및 방향은 사회 · 경제적 인프라 환경 개선이 출산 수요를 충족시켜 새로운 인구 구조 변화를 야기할 수 있을 것으로 사료된다. 출산율 회복을 위한 대표적인 사회경제적 인프라 환경은 주택 인프라이며(Wilson, 2015; Choi *et al.*, 2019; Wilson *et al.*, 2021), 이에 본 연구는 사회경제적 변수로 격자별 개별 주택 수와 노후 주택 비율을 반영하였다. 주택 수 자체는 해당 지역의 활성화 및 개발 정도를 가리키며, 노후 주택 비율은 해당 지역이 사회경제적으로 노후화된 정도를 의미한다(Wilson, 2015).

추가적으로 본 연구는 기계학습 알고리즘을 이용하여 인구 추계를 진행하는 연구로, 개별 변수의 영향력 및 변수 간 상호작용을 통해 기준 연도 대비 목표 연도의 인구를 추계해야 한다. 이러한 인구 추계 과정은 전 시점 인구 대비 다음 시점의 인구를 순차적으로 추계해 야하기 때문에 전 시점 인구를 고려해야 한다. 이에 본 연구는 미래 인구를 추계하기 위한 설명변수 중 하나로 전 시점 인구 수를 변수로서 구성하였다.

3. 분석 방법

본 연구의 분석은 검증과 적용의 두 단계로 구성된다. 검증은 연구 모형의 신뢰도를 판단하기 위한 과정으로, 각 방법론을 통해 2020년 인구를 예측한 뒤, 2020년 예측값과 실제값의 차이를 예측 오차 표현법 중 하나로 제시한 것이다. 검증 단계 이후 각 방법론을 통해 2020년~2040년에 대한 2040년 인구를 추정하는 적용 단계를 진행하였다.

인구추정 결과값의 정확성을 나타내는 지표인 예측 오차(FE: Forecast Error)는 추정 목표연도 시점에서의 인구 예측 값과 실제 인구 값의 차이로 정의된다. 이는 오차의 절댓값을 의미하는데, 실제로는 오차의 절대치를 그대로 사용하기보다 백분율로 표기된 오차값(PE: Percent Error)을 더 많이 사용한다. 오차를 백분율로 표기하는 것은 결과값을 상대적으로 비교하고자 하는 경우 유용하다. 어떤 경우에는 한 시점 및 특정 연도에 대해 복수의 예측치들이 존재하기도 한다. 여러 지역별 장래인구추정을 동시에 실시하거나 특정 시계열 구간에 여러 예측치들이 존재하는 경우가 이에 해당한다. 이때는 ME(Mean Error)와 MAE(Mean Absolute Error) 등과 같은 오차의 평균값을 제시하는 것이 유용하다. ME는 계산과정상 양의 값과 음의 값이 상쇄되어 정확한 오차를 제시하지 못할 수도 있는 반면, MAE는 오차들의 절댓값 평균을 의미하기에 누적 오차 계산이 가능하다. 하지만 ME와 MAE 또한 백분율로 그 값을 표기할 때보다 해석이 용이하고 적용성이 높다. ME와 MAE 또한 백분율로 그 값을 표기할 때 보다 해석이 용이하고 적용성이 높다. ME의 백분율 치환값인 MPE(Mean Percentage Error)와 MAE의 백분율 치환값인 MAPE(Mean Absolute Percentage Error) 각각 오차의 상쇄가 이루어지지 않거나 극단치의 영향으로 인해 높은 부정확성을 도출할 수 있다(Rayer, 2007; Swanson *et al.*, 2000). 예측값과 실제값 차이를 제곱하여 MSE(Mean Squared Error)로 제시하기도 하는데, 오차를 제곱할 경우 극단치의 영향을 과도하게 반영할 수 있기에 MSE의 제곱근인 RMSE(Root Mean Squared Error)를 사용하기도 한다.

본 연구에서 코호트 요인법은 PE값을 통해 연구모형의 정확성을 판단하였고, 기계학습 모델의 정확성은 PE 값에 RMSE 값을 추가적으로 적용하여 판단하였다. RMSE 값의 경우, 회귀모형과 같이 특정 예측모형과 실제값 간의 차이에 대한 복수 값이 존재하는 경우 도출할 수 있는 지표인데 코호트 요인법은 특정 예측모형을 가정하지 않고 각 시점에 대한 예측값만을 제시하기에 PE 값만 이용하였다.

1) 코호트 요인법

코호트 요인법은 자료 구득이나 인구 추정 과정에 있어 큰 제약이 없으면서도 인구 변화에 대한 설명을 제공할 수 있어 가장 보편적으로 사용되고 있는 인구 추정방법이다. 코호트 요인

법은 출생, 사망, 인구이동을 인구 변동의 핵심적 요인으로 고려하는 요인법이다(Smith *et al.*, 2001). 코호트 요인법은 특정 지역에서 발생하는 총 인구 변화를 출생, 사망, 인구이동에 의한 것으로 추정하는데, 이러한 인구 변화가 코호트라는 인구 집단별로 이루어짐을 나타낸다. 코호트란 인생사에서 동일한 경로를 걷는 집단이라는 의미로 인구 추정 분석 내에서는 n세 단위로 태어난 시점이 일치하는 인구집단을 의미하며, 본 연구에서는 10세 간격을 사용하였다.

코호트 요인법에 의한 인구 추정 과정을 요약하면 다음과 같다. 어느 한 지역에 대해 l 시점으로부터 z 기간 후 t 시점에서의 인구를 추정하고자 한다면, 목표 시점의 총인구는 시작 시점으로부터 살아남은 생존 인구, 순이동 인구, 출생 인구의 합으로 계산된다. 이때 l 시점에 x세인 인구는 사망 및 인구이동을 거쳐 t의 $x+z$세의 인구로 전환되며, 이 기간 동안 출생한 인구는 사망을 거쳐 t 시점에 0~z세의 인구를 구성하게 된다(Cho and Lee, 2011).

이를 수식으로 표현하면 식 1과 같다. P_{t-1}은 $t-1$ 시점의 인구를 의미하고, $B_{(t-1,t)}$은 한 시점 동안 태어난 출생 인구를 의미하며, $D_{(t-1,t)}$은 한 시점 동안 사망한 인구, $M_{(t-1,t)}$은 한 시점 동안 전입 · 전출하고 대상지에 남은 순이동 인구 수를 의미한다. 각 코호트에 대해 생존 인구 및 출생 인구, 순이동 인구를 더한 결과를 모두 합산하면 목표 시점의 인구를 구할 수 있다.

$$P_t = P_{t-1} + B_{(t-1,t)} - D_{(t-1,t)} + M_{(t-1,t)} \quad (1)$$

(P_t : t 시점의 인구, $P_{(t-1,t)}$: $t-1$ 시점의 인구, $B_{(t-1,t)}$: 출생 인구, $D_{(t-1,t)}$: 사망 인구, $M_{(t-1,t)}$: 순이동인구)

2) 해석가능한 기계학습

본 연구는 인구학적 특성 외에 공간적 특성 및 사회경제적 특성을 복합적으로 고려할 수 있도록 XGBoost 기계학습 기법과 해석가능한 수치를 도출하기 위해 SHAP(Shapley Additive exPlanations)을 적용하였다. XGBoost(Extreme Gradient Boost)는 성능이 약한 분류기(트리)를 여러 개 쌓아서 보다 강력한 분류 성능을 제공하는 방법인 부스팅 알고리즘 중 하나이다(Chen and Guestrin, 2016). 부스팅 알고리즘은 하나의 트리에 의해 분류를 수행한 뒤 오분류된 사례에 대해 가중치를 부여함으로써 분류기를 수정하고 성능을 개선해나가는 알고리즘이다. 이러한 부스팅 알고리즘은 예측 정확도가 높은 장점이 있지만, 과적합의 문제와 느린 수행시간으로 연산 효율성이 낮다는 단점이 있다. XGBoost는 규제항(regularization term)을 도입함으로써 과적합 문제를 해결하고, 병렬 CPU를 통한 학습이 가능케 함으로써 수행시간도 단축시켰다(Yang and Chun, 2022). 규제화된 손실함수를 목적함수로 이용하는 XGBoost의 목적함수는 식 2와 같다. 식 2의 $L(\Theta)$는 실제값 y_i와 예측값 $(\widehat{y_i})$의 차이, 즉 예측오류에 의한 손실을 의미하므로 $L(\Theta)$를 최소화함은 예측력이 우수한 모형을 선택함을 의미한다. $\Omega(\Theta)$는 모형이 복잡해질 시 패널티를 부여하여 복잡도를 조절하는 규제항으로서 $\Omega(\Theta)$를 최소화함은 단순한 모형을 선택함을 의미한다. 규제항은 개별 트리 f_k의 노드 함수를 입력변수로 사용하여 모델을 구성하는 모수의 크기를 줄이는 방향으로 진행된다(Kim, 2022).

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta) = \sum l(\widehat{y_i}, y_i) + \sum \Omega(f_k) \quad (2)$$

SHAP(Shapley Additive exPlanations) 기법은 전체 성과를 창출하는 데 각 변수가 얼마나 기여했는지를 수치로 표현하는 섀플리 값(Shapley Value)와 변수 간 독립성을 이용하는 기법이다. 각 변수의 기여도는 모든 변수를 조합하였을 때 나오는 성과와 해당 변수를 제외했을 때 나오는 성과의 차이를 각각 계산한 뒤 해당 값들을 가중평균하여 측정한다(Lee *et al.*, 2021). 이러한 논리는 게임 이론을 바탕으로

두고 있으며, 다수가 게임을 진행할 때 i 인물이 전체 성과에 기여하는 정도는 전체 기여도에서 i번째 인물이 제외된 기여도의 합을 뺀 값을 근거로 한다(Ahn, 2020). 섀플리 값을 구하는 방식은 식 3과 같다. Φ_i는 변수 i의 섀플리 값이며, n은 전체 변수 수, $f_x(z')$는 모든 변수의 기여도, $f_x(z' \setminus i)$는 변수 i를 제외하고 나머지 변수들을 이용해 구한 기여도를 의미한다(Ahn, 2020). 따라서 SHAP 기법은 각 변수의 기여도를 통해 예측 모델의 출력 결과 해석을 지원한다.

$$\Phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(n-|z'|-1)!}{n!} \cdot [f_x(z') - f_x(z' \setminus i)] \quad (3)$$

SHAP 기법은 회귀분석이 각 독립변수 간의 독립성을 가정하는 바와 달리 종속변수를 예측하는 변수들끼리 서로 의존적인 경우에도 적용 가능하며, 예측에 대해 개별 변수들이 갖는 영향력과 개별 변수들이 갖는 영향력의 근거를 제시할 수 있다는 점에서 장점이 있다. 또한 섀플리 값은 음수일 수 있으며, 섀플리 값이 음수로 나타남은 해당 변수가 예측 결과에 음(−)의 영향을 미친다고 해석할 수 있다(Lee *et al.*, 2021). 섀플리 값은 크게 전역적 섀플리 값(Global Shaple Value)과 국지적 섀플리 값(Local Shapley Value)으로 구분된다. 전역적 섀플리 값은 예측 결과에 전반적으로 영향을 미치는 변수들의 평균 영향력을 제공한다. 이 때 SHAP 기법은 변수 간 의존도 및 음의 영향력을 모두 반영하여 음의 영향력은 반영하지 못하는 피처 중요도 기법 등 타 해석가능한 기계학습 기법보다 정확한 해석값을 제시해준다(Lee *et al.*, 2021). 전역적 섀플리 값은 영향력의 크기가 큰 순서대로 위쪽에서부터 나열되며, 변수의 영향력이 양(+)일 경우 붉은색, 음(−)일 경우 푸른색으로 나타난다(Park *et al.*, 2022). 국지적 섀플리 값은 데이터 내 각 표본에 대한 해석을 제공할 수 있는 부분을 의미하며, 이는 예측 결과에 대한 개별 표본들의 설명이 가능하다는 점에서 장점이 있다. 하지만 국지적 섀플리 값을 구하고자 할 경우 대용량의 데이터를 다뤄야 하므로 연산량이 크고, 변수가 자주 바뀌는 경우에는 적용하기 어렵다는 단점이 있다(Lee *et al.*, 2021).

분석결과

1. 코호트 요인법 적용의 추정 결과

2000년~2020년에 대해 코호트 요인법을 이용하여 인구 추정를 실시한 결과와 PE 값은 표 2와 같다. 현재 부산광역시 개발 가능성이 가장 높은 지역인 강서구와 기장군 지역을 제외하고는 추정인구와 실제인구의 오차가 평균 23%인 것으로 나타났다. 또한, 코호트 요인법을 적용하였을 경우 실제 인구 값보다 부산시 인구가 56만 명 정도 증가할 것으로 예측되었다.

코호트 요인법의 오차는 대부분 인구 이동을 추정하는 과정과 20~40대 연령 인구를 추정하는 과정으로부터 기인하는 것으로 확인되었다. 두 과정 모두 시계열적 변이가 크고, 추정기간 및 전망기간에 대해 인구 변동률을 예측하기 어려운 부분이다. 이는 코호트 요인법이 단순히 이전 시점 인구와 출산율, 생존율 등의 증감분만 고려하여 사회 · 경제적 요인으로 인한 인구 변동률을 예측하는 데는 한계가 있음을 의미한다.

다음으로 2020년~2040년에 대해 코호트 요인법을 이용하여 인구 추정을 실시한 결과는 표 3과 같다. 앞서 2020년 인구 추정 결과값과 같이 강서구와 기장군을 제외하고는 모든 지역에서 급격한 인구 감소가 진행될 것으로 예상되었다. 예측한 2035년 인구와 2040년 인구를 비교해 보았을 때, 부산광역시 인구가 지속적으로 줄어들 것을 확인할 수 있고, 특히 영도구, 사상구, 중구, 서구, 부산진구 순으로 급격한 인구 감소가 예측되었다. 또한 기준연도 대비 추정연도 인구의 증감분을 확인할 수 있는 PC값을 계산한 결과, 2040년 부산광역시 전체 추정인구는 2020년 대비 평균 26.7% 감소할 것으로 나

TABLE 2. Forecasting Error of Cohort-Component Method

Region	Sex	Population Projection at 2020	Actual Population at 2020	Percent Error (%)
Busan	Male	1,317,651	1,638,751	19.59
	Female	1,472,815	1,710,265	13.88
Junggu	Male	14,039	20,416	31.24
	Female	16,039	21,019	23.69
Seogu	Male	34,940	51,127	31.66
	Female	38,883	54,172	28.22
Donggu	Male	17,726	42,453	58.25
	Female	20,478	44,791	54.28
Yeongdogu	Male	37,734	56,613	33.35
	Female	45,004	56,610	20.50
Busanjingu	Male	141,579	168,641	16.05
	Female	165,111	182,762	9.66
Dongnaegu	Male	70,711	127,743	44.65
	Female	83,048	135,600	38.76
Namgu	Male	130,011	131,683	1.27
	Female	143,259	137,426	4.24
Bukgu	Male	195,608	138,110	41.63
	Female	208,643	142,066	46.86
Haeundaegu	Male	165,812	187,806	11.71
	Female	189,946	201,729	5.84
Sahagu	Male	135,163	155,310	12.97
	Female	148,028	156,746	5.56
Geumjeonggu	Male	74,961	115,296	34.98
	Female	85,107	121,919	30.19
Gangseogu	Male	8,728	72,093	87.89
	Female	10,043	64,639	84.46
Yeonjegu	Male	84,903	98,184	13.53
	Female	93,996	106,760	11.96
Suyeonggu	Male	68,568	80,698	15.03
	Female	80,313	90,760	11.51
Sasanggu	Male	95,824	108,759	11.89
	Female	103,264	107,589	4.02
Gijanggun	Male	41,344	83,787	50.66
	Female	41,654	85,677	51.38

TABLE 3. Forecasting 2040 Population of Cohort-Component Method

Region	Sex	Population Projection at 2035	Population Projection at 2040	2035 Percent Change (%)	2040 Percent Change (%)
Busan	Male	1,284,881	1,139,738	-21.6%	-30.5%
	Female	1,442,910	1,321,865	-15.6%	-22.7%
Junggu	Male	12,330	9,450	-39.6%	-53.7%
	Female	13,747	11,035	-34.6%	-47.5%
Seogu	Male	33,415	26,873	-34.6%	-47.4%
	Female	38,202	31,924	-29.5%	-41.1%

TABLE 3. Continued

Region	Sex	Population Projection at 2035	Population Projection at 2040	2035 Percent Change (%)	2040 Percent Change (%)
Donggu	Male	33,239	29,125	−21.7%	−31.4%
	Female	37,098	33,379	−17.2%	−25.5%
Yeongdogu	Male	25,715	16,371	−54.6%	−71.1%
	Female	27,516	17,794	−51.4%	−68.6%
Busanjingu	Male	108,132	86,217	−35.9%	−48.9%
	Female	130,647	110,587	−28.5%	−39.5%
Dongnaegu	Male	110,834	102,321	−13.2%	−19.9%
	Female	127,878	121,927	−5.7%	−10.1%
Namgu	Male	91,174	76,253	−30.8%	−42.1%
	Female	101,864	87,657	−25.9%	−36.2%
Bukgu	Male	97,200	81,182	−29.6%	−41.2%
	Female	108,471	95,044	−23.6%	−33.1%
Haeundaegu	Male	140,309	121,735	−25.3%	−35.2%
	Female	162,733	146,425	−19.3%	−27.4%
Sahagu	Male	108,474	90,369	−30.2%	−41.8%
	Female	118,047	102,698	−24.7%	−34.5%
Geumjeonggu	Male	83,041	70,516	−28.0%	−38.8%
	Female	96,528	85,846	−20.8%	−29.6%
Gangseogu	Male	123,466	136,528	71.3%	89.4%
	Female	114,165	127,502	76.6%	97.3%
Yeonjegu	Male	84,606	77,879	−13.8%	−20.7%
	Female	104,846	101,301	−1.8%	−5.1%
Suyeonggu	Male	66,141	59,666	−18.0%	−26.1%
	Female	84,387	79,883	−7.0%	−12.0%
Sasanggu	Male	61,023	45,135	−43.9%	−58.5%
	Female	67,095	53,736	−37.6%	−50.1%
Gijanggun	Male	105,781	110,116	26.2%	31.4%
	Female	109,689	115,125	28.0%	34.4%

타났다.

2. 기계학습을 적용한 인구추정 결과

기계학습의 방법 역시 코호트 요인법과 마찬가지로 2005~2020년에 대해 검증의 과정을 진행하여 본 방법론의 예측 오차 및 신뢰도를 우선적으로 판단하였다. 기계학습 알고리즘을 이용하여 2020년의 부산시 인구추정을 수행한 결과는 표 4와 같다. 코호트 요인법이 평균 23%의 예측 오차(PE)를 도출했던 바와 달리 기계학습을 이용하여 인구 추정을 수행할 시 예측 오차(PE)가 2.36%로 낮아지는 것을 확인할 수 있다. 코호트 요인법 적용 시 오차 범위가 가장 컸던 강서구와 기장군에 대해서도 기계학습 기법을 통해 인구 추정을 진행할 시 높은 정확성을 가지는 결과를 도출할 수 있음을 확인하였다. 추가적으로 본 연구에서 사용한 XGBoost 모델의 정확성을 나타내는 지표인 RMSE 값은 48.27로 나타났다.

XGboost 알고리즘을 이용하여 부산광역시 각 소지역 인구를 추정한 뒤 합산한 결과, 부산시에 대한 2040년 전체 인구는 2020년 대비 1.75% 감소할 예정인 것으로 나타났다(표 5). 기계학습 분석 결과, PC값 기준 2020년 대비 가장 급격한 인구 변화를 보일 예정인 상위 5개 지역은 영도구, 금정구, 중구, 동구, 서구 순으로

TABLE 4. Forecasting Error of Machine Learning Method

Region	Population Projection at 2020	Actual Population at 2020	Percent Error (%)
Busan	3,428,035	3,349,016	2.36
Junggu	41,941	41,439	1.21
Seogu	108,214	105,303	2.76
Donggu	87,733	87,246	0.56
Yeongdogu	117,224	113,224	3.53
Busanjingu	358,557	351,403	2.04
Dongnaegu	260,425	263,345	1.11
Namgu	282,192	269,111	4.86
Bukgu	309,385	280,177	10.42
Haeundaegu	400,241	389,535	2.75
Sahagu	316,903	312,057	1.55
Geumjeonggu	236,116	237,219	0.46
Gangseogu	131,037	136,734	4.17
Yeonjegu	213,129	204,947	3.99
Suyeonggu	176,373	171,461	2.86
Sasanggu	221,165	216,350	2.23
Gijanggun	167,400	169,465	1.22

나타났으며, 해운대구와 강서구를 제외한 부산광역시 15개 시군구 지역 모두 인구가 감소할 것으로 나타났다.

기계학습 모형에 있어 피처 중요도를 이용하여 XGBoost 알고리즘을 해석한 결과는 그림 1과 같다. 피처 중요도 그래프(그림 1)에 따르면 XGBoost 알고리즘이 생산가능 여자 인구, 전 시점 인구, 주택 수, 고령 인구, 노후 주택 비율 순서로 부산광역시 인구 추정에 영향을 미친다고 측정함을 알 수 있다. 섀플리 값 분석 결과,

TABLE 5. Forecasting 2040 Population of Machine Learning Method

Region	Population at 2020	Population Projection at 2040	Percent Change(%)
Busan	3,349,016	3,290,566	−1.75%
Junggu	41,439	38,977	−5.94%
Seogu	105,303	101,423	−3.68%
Donggu	87,246	83,249	−4.58%
Yeongdogu	113,224	103,888	−8.25%
Busanjingu	351,403	346,271	−1.46%
Dongnaegu	263,345	260,335	−1.14%
Namgu	269,111	263,791	−1.98%
Bukgu	280,177	277,818	−0.84%
Haeundaegu	389,535	398,235	2.23%
Sahagu	312,057	305,201	−2.20%
Geumjeonggu	237,219	218,156	−8.04%
Gangseogu	136,734	143,420	4.89%
Yeonjegu	204,947	202,069	−1.40%
Suyeonggu	171,461	170,791	−0.39%
Sasanggu	216,350	208,427	−3.66%
Gijanggun	169,465	168,515	−0.56%

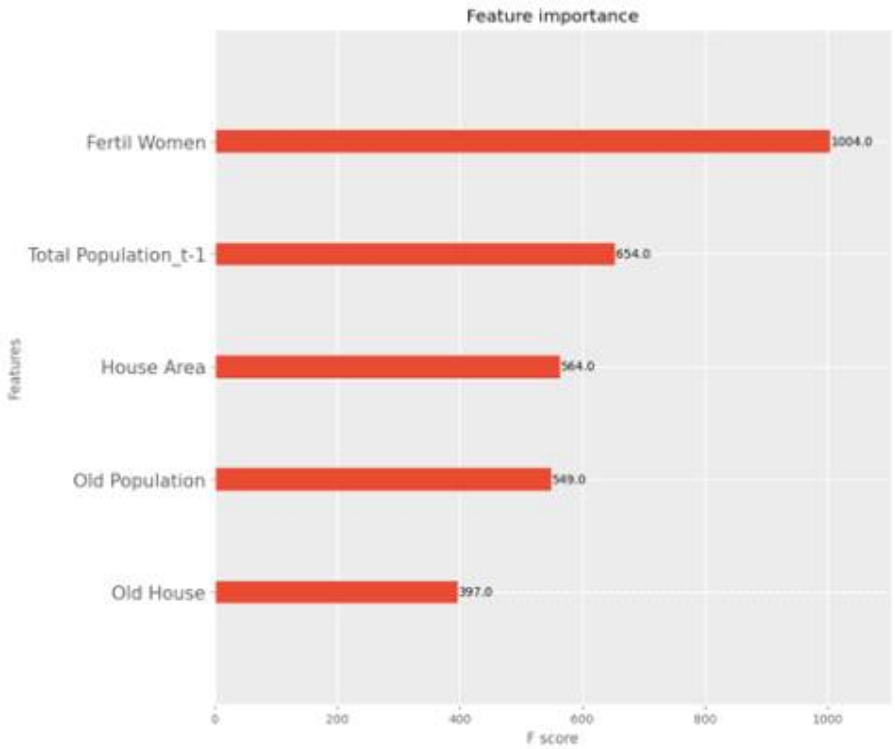

FIGURE 1. Results of Feature Importance Analysis

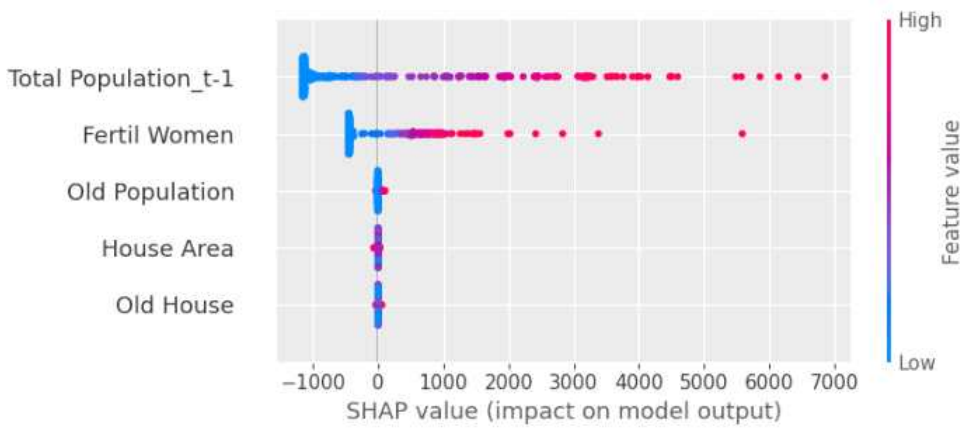

FIGURE 2. Results of SHAP

피처 중요도 기법 결과와 유사하게 본 모형을 해석함을 알 수 있다. 그림 2의 붉은색 점은 해당 변수가 인구 추정에 큰 영향을 미쳤음을 의미하는 반면, 파란색 점은 해당 행 변수가 인구 추정을 결정하는 데 적은 영향을 미치는 것으로 해석할 수 있다. 새플리 값은 전 시점 인구, 생

산가능 여자 인구, 고령 인구, 개별주택, 노후 주택 비율 순서로 인구 추정에 영향을 준다고 해석하고 있으며, 특히 전 시점 인구 변수와 생산가능 여자 인구 변수가 큰 역할을 한다. 두 변수의 분산 또한 크다. 그림 2에 의하면 인구 추정의 정확성 및 판단 근거는 전 시점 인구 수와 생산가능 여자 인구 수에 크게 의존하는 것을 알 수 있으며, 이 외의 변수들은 인구 추정에 있어 큰 영향을 주지 않음을 알 수 있다.

3. 인구추정 결과의 비교 분석

코호트 요인법과 기계학습 기법을 적용하여 부산광역시 인구를 추정한 결과를 종합하면 다음과 같다. 대한민국 부산광역시 500m 소지역 격자를 바탕으로 한 시군구별 인구 추정을 수행함에 있어 기계학습 분석 기법이 상대적으로 훨씬 높은 정확성을 도출하였으며, 두 방법론 모두 영도구, 중구, 서구 지역의 인구가 가장 급감할 것으로 나타났다. 두 방법론 모두 부산시 인구가 지속적으로 감소할 것으로 추정되었다. 코호트 요인법을 통해 연령별 인구를 파악한 결과 부산시 인구의 감소분은 계속해서 줄어드는 청년인구의 몫으로 해석할 수 있다. 인구 추정 과정 상 여러 요인을 반영하지 못하고 출생, 사망, 이주인구만을 변수로 사용하는 코호트 요인법은 인구 유출입이 가장 활발히 일어날 것으로 예상되는 강서구, 기장군 지역 등에 대해서는 부정확한 결과값을 도출하였다. 반면, 기계학습 기법은 인구 구조 변동에 영향을 미칠 가능성이 있는 여러 변수의 조합 반영이 가능하여 보다 낮은 오차를 도출함으로써 인구 변화폭이 큰 지역의 추정에 있어 적용력이 높음을 알 수 있다.

부산광역시 500m 격자에 대해 인구를 추정한 결과는 그림 3 및 그림 4와 같다. 코호트 요인법이 기계학습 기법에 비해 인구가 급감할 것으로 나타났음을 확인할 수 있다. 또한 코호트 요인법은 2040년 부산광역시 개별 시군구에 대해 기장군과 강서구 등 일부 지역에 인구가 집중적으로 분포할 것이라 나타난 반면, 기계학습

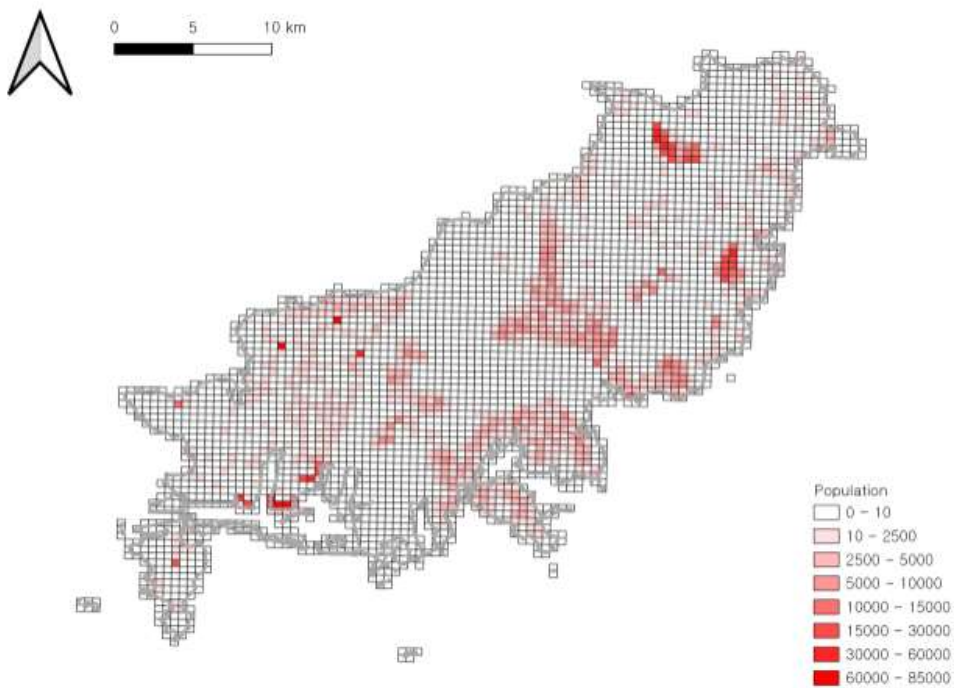

FIGURE 3. Population Projection Using Cohort-Component Method

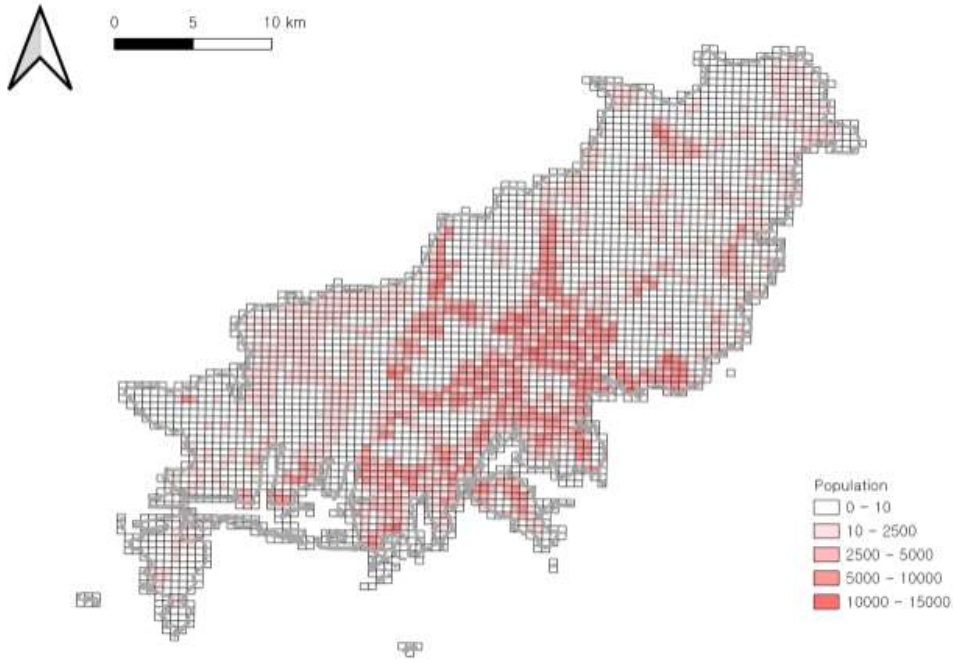

FIGURE 4. Population Projection using AI

기법은 사하구, 영도구, 중구 일대를 제외하고 전역적인 인구 분포를 갖게 될 것이라 추정하고 있다.

결 론

본 연구는 부산광역시를 대상으로 500m 격자를 이용한 소지역을 대상으로 기존의 코호트 요인법과 기계학습 방법을 이용하여 2040년 인구를 추정하였다. 부산광역시 2040년 인구는 코호트 요인법으로 추정한 결과가 2,461,603명이었으며 기계학습으로 추정한 결과가 3,290,566명으로 나타났다. 통계청에서 부산광역시 전체를 대상으로 추정한 2040년의 인구는 2,826,940명인데 이는 소지역으로 추정한 경우보다 365,337명 더 많게 추정되었다. 소지역에 대해 코호트 요인법을 적용하였을 때 지역별로 차이가 나는 값이 반영되었기 때문인 것으로 여겨진다. 하지만 기계학습 방법으로 추정하였을 경우 코호트 요인법으로 추정한 결과보다 828,963명 더 높게 나타났다. 이러한 인구 추정 값 차이의 발생은 인구의 이동, 출생, 사망 외 인구변동에 영향을 미칠 수 있는 예측 불가능한 요인으로부터 기인하는 것으로 예측해볼 수 있다.

최근 발표된 2040년 부산도시기본계획에서는 생잔모형에 기반한 자연적 인구변화의 값을 3,151,524명으로 전출을 -124,800명으로 사회적 인구증가를 510,313명으로 추정하여 계획인구를 약 350만명으로 설정하고 있다. 그리고 이 계획인구를 바탕으로 주택, 상수도, 하수도, 병원, 복지시설, 공원 등 다양한 계획지표를 설정하고 있다. 2030 부산도시기본계획(변경)에서 제시된 계획인구인 410만명 보다는 과대추정되지 않았지만 사회적 인구증가를 여전히 높은 수준인 50만명 이상으로 설정하고 있다. 본 연구의 결과 값과 비교할 경우 코호트 요인법 보다는 약 104만명을 기계학습으로 추정한 결과보다는 약 21만명을 과대추정하고 있다.

예측과 전망은 현 시점에서의 정보를 기준으로 추정한 값이기 때문에 그 특성 상 미래가 도

래하기 전에는 무엇이 맞는지 판단하기 어렵다. 또한 현재부터 미래까지 이어지는 경로 상에 어떠한 가정을 하고 있는가에 따라 그 값의 차이가 크게 달라진다. 특히 인구의 추정은 인구의 자연적인 출생과 사망 그리고 사회적인 측면의 이동이 결합되어 있기 때문에 더욱이 어려운 과정이다. 성장의 시기에 있어서는 예측한 것 보다 더 많은 인구가 도시에 거주할 수 있기 때문에 선투자적인 대비의 측면에서 과대추정된 값을 이용한다 하더라도 문제가 나타날 여지는 크지 않으나 정체와 감소의 시기에 있어서는 과대추정된 값은 도시계획에서 투자의 비효율성과 특정 부문에 대한 과잉 투자에 따른 타 부문에서의 질적 저하와 같은 문제가 나타날 가능성이 높다. 뿐만 아니라 정체와 감소의 시기에 지나치게 과소추정된 값 역시 도시의 축소를 가속화시키며 삶의 질을 저하시키는 문제가 나타난다. 따라서 추정된 인구 값을 기준으로 하여 미래에 대한 투자와 의사결정을 하는데 있어서는 현재의 엄격한 기준에서 합리적으로 추정된 값을 이용하여야 한다.

본 연구는 인구 추정에 있어 소지역의 범위를 500m 격자단위 자료를 이용하여 보다 세분화된 결과값을 제시하고자 하였고, 기계학습 기법을 통해 정확도 높은 인구 추정 결과값과 설명력을 제공하고자 하였다는 점에서 타 연구와 차별점을 갖는다. 또한, 높은 예측 성능을 갖는 XGBoost 모델을 사용하면서도 기계학습 블랙박스의 설명을 제공해주고자 한 점에서 의의를 갖는다. 하지만 XGBoost 알고리즘 이외 예측 성능이 높은 보다 많은 알고리즘의 적용을 통해 다양한 소지역 인구 추정 방법론을 제시한다면 향후 수행될 소지역 인구 추정 연구에 도움이 될 것으로 사료된다. 추가적으로 기계학습 모델 구축 시 장래 인구에 영향을 줄 수 있는 변인으로 표고, 경사도와 같은 지질학적 특성과 보다 넓은 사회·경제적 맥락을 다룰 수 있는 변수를 추가한다면 기계학습 모델의 오차를 더 줄일 수 있을 것으로 기대된다. KAGIS

REFERENCES

Baker, J., A. Alcantara., X. Ruan., K. Watkins., and S. Vasan. 2013. A comparative evaluation of error and bias in census tract-level age/sex-specific population estimates: component I (net-migration) vs component III (Hamilton-Perry). Population Research and Policy Review 32:919-942.

Baker, J., A. Alcántara., X. Ruan., K. Watkins., and S. Vasan, 2014. Spatial weighting improves accuracy in small-area demographic forecasts of urban census tract populations. Journal of Population Research 31:345-359.

Baker, J., D. Swanson., J. and Tayman. 2021. The accuracy of Hamilton-Perry population projections for census tracts in the United States. Population Research and Policy Review 40:1341-1354.

Breidenbach, P., M. Kaeding., and S. Schaffner. 2019. Population projection for Germany 2015-2050 on grid level (RWI-GEO-GRID-POP-Forecast). Jahrbücher für Nationalökonomie und Statistik 239(4):733-745.

Cai, Q. 2007. New techniques in small area population estimates by demographic characteristics. Population Research and Policy Review 26:203-218.

Caraviello, D. Z., K. A. Weigel., M. Craven., D. Gianola., N. B. Cook., K. V. Nordlund,, ... and M. C. Wiltbank. 2006. Analysis of reproductive performance of lactating cows on large dairy farms using machine learning algorithms. Journal of dairy science 89(12):4703-4722.

Chen, T., and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. In

Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 785-794.

Chen, Y., F. Guo., J. Wang., W. Cai., C. Wang., K. and Wang. 2020. Provincial and gridded population projection for China under shared socioeconomic pathways from 2010 to 2100. Scientific Data 7(1): 83.

Chen, Y., X. Li., K. Huang., M. Luo., and M. Gao. 2020. High-resolution gridded population projections for China under the shared socioeconomic pathways. Earth's Future 8(6):e2020EF001491.

Chi, G., and P. R. Voss. 2011. Small-area population forecasting: Borrowing strength across space and time. Population, Space and Place 17(5):505-520.

Davis, H. C. 1995. Demographic projection techniques for regions and smaller areas: a primer. UBC Press.

Ford, A., S. Barr., R. Dawson., J. Virgo., M. Batty., and J. Hall. 2019. A multi-scale urban integrated assessment framework for climate change studies: A flooding application. Computers, Environment and Urban Systems 75:229-243.

Goodchild, M. F., L. Anselin., and U. Deichmann. 1993. A framework for the areal interpolation of socioeconomic data. Environment and planning A. 25(3):383-397.

Lundberg, P., and P. Frodin. 1998. Ecosystem resilience and productivity: are predictions possible?. Oikos 603-606.

McKee, J. J., A. N. Rose., E. A. Bright., T. Huynh., and B. L. Bhaduri. 2015. Locally adaptive, spatially explicit projection of US population for 2030 and 2050. Proceedings of the National Academy of Sciences 112(5):1344-1349.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence 267:1-38.

Perry, C. 2013. The neighborhood unit. In The urban design reader. 98-109. Routledge.

Rayer, S. 2007. Population forecast accuracy: does the choice of summary measure of error matter?. Population Research and Policy Review 26:163-184.

Samuel, A. L. 1959. Some studies in machine learning using the game of checkers. IBM Journal of research and development 3(3):210-229.

Shapley, L. S. 1997. A value for n-person games. Classics in game theory 69.

Smith, S. K., and P. A. Morrison. 2005. Small-area and business demography. 761-785. Springer US.

Smith, S. K., J. Tayman., and D. A. Swanson. 2005. State and local population projections: Methodology and analysis.

Smith, S. K., J. Tayman., and D. A. Swanson. 2013. A practitioner's guide to state and local population projections. Springer Netherlands.

Stillwell, J., and M. Clarke. (Eds.). 2011. Population dynamics and projection methods (Vol. 4). Springer Science & Business Media.

Swanson, D. A. (Ed.). 2017. The frontiers of applied demography. Springer International Publishing.

Swanson, D. A., A. Schlottmann., and B. Schmidt.

2010. Forecasting the population of census tracts by age and sex: An example of the Hamilton-Perry method in action. Population Research and Policy Review 29:47-63.

Swanson, D. A., J. Tayman., and C. F. Barr. 2000. A note on the measurement of accuracy for subnational demographic estimates. Demography 37(2):193-201.

Triantakonstantis, D., and G. Mountrakis. 2012. Urban growth prediction: a review of computational models and human perceptions.

Van Lent, M., W. Fisher., and M. Mancuso. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the national conference on artificial intelligence. 900-907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Weber, H. 2020. How well can the migration component of regional population change be predicted? A machine learning approach applied to German municipalities. Comparative Population Studies-Zeitschrift für Bevölkerungswissenschaft. 45:143-178.

Wilson, T. 2015. New evaluations of simple models for small area population forecasts. Population. Space and Place 21(4):335-353.

Wilson, T., I. Grossman., and J. Temple. 2021. Evaluation of the best M4 competition methods for small area population forecasting. International Journal of Forecasting.

Yan, X., X. Liu., and X. Zhao. 2020. Using machine learning for direct demand modeling of ridesourcing services in Chicago. Journal of Transport Geography 83:102661.

Ahn, J.H. 2020. Explainable AI, Dissecting AI. (안재현. 2020. XAI 설명 가능한 인공지능, 인공지능을 해부하다. 위키북스, 파주).

Byeon, S.Y., D.C. Lee., and K.H. Kim. 2023. Estimating Gridded Population using Day-Time Satellite Image by CNN. Journal of The Korean Data Analysis Society 25(2):467-479 (변상영, 이동찬, 김기환. 2023. CNN 기반 주간 위성 이미지를 활용한 격자 단위 인구추정. 한국자료분석학회 25(2):467-479).

Cho, D.H., and S.I. Lee. 2011. Population Projections for Busan Using a Biregional Cohort-Component Method. Journal of the Korean Geographical Society 46(2): 212-232 (조대헌, 이상일. 2011. 이지역 코호트-요인법을 이용한 부산광역시 장래 인구추계. 대한지리학회지 46(2):212-232).

Choi, H.J., S.H. Choi., and S.J. Hong. 2019. Development of Estimation Method for Future Population by Eup-Myeon-Dong Unit. Journal of Real Estate Analysis 5(3):67-87 (최현정, 최석환, 홍성조. 2019. 읍면동 단위 장래인구 추정모형 개발에 관한 연구. 부동산분석 5(3):67-87).

Kim, B.S., H.C. Jeon., and D.B. Shin. 2022. Development and Accuracy Analysis of the Population Estimation Models based on Building Volume and Landuse. Journal of Korean Society for Geospatial Information Science 30(4):25-34 (김병선, 전해찬, 신동빈. 2022. 건물 부피와 토지이용 정보를 이용한 인구추정 모델 개발 및 정확도 분석. 대한공간정보학회지 30(4):25-34).

Kim, S.H. 2022. Empirical studies on the stock price prediction performance of Xgboost models. Journal of Social Science.

39(1):29-55 (김상환. 2022. Xgboost 모형의 주가예측성과에 대한 실증연구. 사회과학연구 39(1):29-55).

Lee, B.K. 2019. 2040 Future Population Projections Study. Korea Research Institute for Human Settlements (이보경. 2019. 2040년 장래인구 분포 전망 연구. 국토연구원).

Lee, C.H., and S.I. Lee. 2006. Generation of Synthetic Population for Buildings. Journal of Korea Planning Association 41(6):37-50 (이창효, 이승일. 2006. 건축물 단위의 인구분포 추정. 국토계획 41(6):37-50).

Lee, J.J., Y.R. Lee., D.H. Lim., and H.C. Ahn. 2021. A Study on the Employee Turnover Prediction using XGBoost and SHAP. The Journal of Information Systems 30(4):21-42 (이재준, 이유린, 임도현, 안현철. 2021. XGBoost 와 SHAP 기법을 활용한 근로자 이직 예측에 관한 연구. 정보시스템연구 30(4):21-42).

Lee, S.I., and D.H. Cho. 2020. Trend Extrapolation Methods for Small Area Population Projections in Korea. Journal of Geography Education 64:1-19 (이상일, 조대헌. 2020. 우리나라 소지역 인구 추계를 위한 방법론 연구: 추세외삽법을 중심으로. 지리교육논집 64:1-19).

Lee, S.S. 2017. Demographic Dynamics and Policy Response. Health and Welfare Policy Forum 2017(1):29-46 (이삼식. 2017. 인구 및 출산 동향과 대응 방향. 보건복지포럼 2017(1):29-46).

Park, J.S., S.J. Kim., and S.G. Lee. 2022. Analysis of Determining Factors of Urban Vitality with Mobile Phone Location-Based Origin-Destination Bigdata by Travel Purpose : Using the PageRank Algorithm and SHAP Machine Learning. Journal of Korea Planning Association 57(5):72-89 (박준상, 김선재, 이수기. 2022. 모바일폰 위치기반 생활이동 빅데이터를 활용한 통행목적별 도시활력 영향요인 분석: PageRank 알고리즘과 SHAP 기계학습을 활용하여. 국토계획 57(5):72-89).

Woo, H.B., J.Y. Yang., S.H. Cho., and H.S. Ahn. 2016. Demographic forecasting: analytic review and assessment. (우해봉, 양지윤, 조성호, 안형석. 2016. 인구추계 방법론의 현황과 평가).

Yang, G.P., and H.J. Chun. 2022. Analysis of the Relationship between Housing Prices and Regional Characteristics Using Machine Learning and XAI. Korea Real Estate Review 32(3):7-24 (양건필, 전해정. 2022. 기계학습과 XAI 를 활용한 아파트 가격과 지역특성과의 관계 분석. 부동산연구 32(3):7-24). KAGIS