경영전문석사학위 논문

지역인구 예측을 위한 딥러닝 모델 연구

- 부산지역 인구 기준 -

2022년 2월

서울과학종합대학원대학교

박 승 용

경영전문석사학위 논문

지역인구 예측을 위한 딥러닝 모델 연구

- 부산지역 인구 기준 -

2022년 2월

서울과학종합대학원대학교

박 승 용

지역인구 예측을 위한 딥러닝 모델 연구

지도교수 김 호 현

이 논문을 경영학 석사 학위논문으로 제출함
2022년 2월
서울과학종합대학원대학교
박 승 용

박승용의 석사 학위논문을 인준함
2022년 1월

위 원 장 신 호 상 (인)

위 원 문 달 주 (인)

위 원 김 호 현 (인)

초 록

현재 지역의 장래인구 추계에 사용하는 코호트 요인법은 미래의 가정을 전제로 작성한 시나리오에 기반하고 있다는 불확실성과 자유로운 지역 간 이동을 지역별로 고려해야 하는 추계의 복잡성을 가진 특징이 있다. 이에 따라 지역별 개발 계획 수립 시 지역 간 이동 요소에 대한 가정을 최대로 하여 계획 인구를 증가시키는 등 이해관계에 따라 장래 인구를 유리하게 조정하기도 한다. 그러나 국가 전체적인 인구 감소 추세로 실제 인구가 계획 인구를 충족하지 못하며, 과다한 장래 인구 예측에 따른 문제를 발생시키고 있다.

따라서 본 연구에서는 이러한 한계점을 극복할 방안으로, 데이터의 학습만으로 예측값을 도출할 수 있는 딥러닝 모델을 적용하여 인위적인 가정 요소와 통계 모델 선택의 문제를 배제하고 특정 지역의 인구 예측이 가능한지를 타진해보았다.

이를 위해 먼저 인구 예측에 필요한 딥러닝과 관련된 이론적 배경을 살펴보고, 선행연구를 분석하였다. 그리고 대한민국 부산지역의 인구 데이터를 사례로 하여 최적의 딥러닝 모델을 만들고, 통계청의 추계 인구와 비교해 보았다. 입력변수는 1997년부터 2020년까지 부산 지역의 출생아 수, 사망자 수, 전출자 수 및 전입자 수와 부산지역의 주민등록 인구수 등 5개 데이터를 사용하였다. 해당 데이터는 1997~2015년까지를 훈련 셋으로, 2016~2020년까지를 테스트 셋으로 구분하였다.

인구 예측을 위한 모델 생성을 위하여 인공신경망의 기본 구조인 MLP와 시계열 데이터에 적합한 RNN, LSTM, GRU를 적용하였다. 은닉층 수와 노드 수를 10가지로 구성하여 학습한 결과에 따르면, LSTM에서 가장 좋은 예측값을 보여준 모델이 생성되었다. 예측인구는 매년 감소되는 인구 추세를 따라갔으며, MAPE는 0.419로 평가되었다.

본 연구에서 사용한 데이터 범위가 짧았던 한계는 있었으나, 인위적인

가정이나 통계 모델 설정 없이도 딥러닝 모델을 적용하여 데이터 학습만으로 특정 지역의 인구를 예측하는 것이 가능함을 확인하였다. 향후 데이터의 보완 및 입력 변수 추가, 모델 개선을 통해 지역 단위의 인구를 보다 정확하게 예측하고, 정책 판단의 근거로 활용할 수 있을 것으로 기대된다.

목 차

표 목 차

그 림 목 차

제 Ⅰ 장 서 론

제1절 연구의 배경 및 목적

2020년말 기준 주민등록 인구수가 연간 출생자 수 30만명 선이 붕괴되면서, '인구 데드크로스[1]'를 기록하며 처음으로 감소한 것으로 나타났다(행정안전부, 2021). 폴 몰런드는 '인구의 힘'에서, 인구는 가장 중요한 생산요소이자 소비 수요이자, 나라를 유지할 수 있는 국방력의 근원임에 따라 인구가 한 나라에 끼치는 영향은 막대하다고 보았다. 이처럼 중요한 인구 규모가 우리나라에서는 뚜렷하게 감소하고 있다.

그러나 수도권은 비수도권에서 수도권으로 인력이 유출되는 수도권 인구집중 현상으로 오히려 인구 규모가 증가하고 있다. 전출입에 따른 사회적 증감의 영향을 크게 받는 지역별 인구 증감을 보면, 2011년 대비 2020년 기준으로 수도권(서울, 경기, 인천 등 3개 광역지방자치단체) 인구는 105만 명이 증가하였다. 그러나 비수도권(부산, 대구 등 14개 광역지방자치단체) 인구는 6.35만 명이 줄어든 것으로 나타났다(행정안전부, 2021). 수도권의 인구 집중도 1990년 42.7%에서 2020년 50.1%로 일본의 28.0%를 초과하는 수준을 보이고 있다(감사원, 2021). 이는 2014년에 '지방소멸'을 주장한 마스다 히로야의 국가인 일본보다 더 높은 수준의 집중도이다. 이에 따라 일자리와 정주 여건이 취약한 비수도권은 인구 감소에 따른 충격이 더 크게 진행될 수밖에 없는 실정이다. 지방소멸과 관련된 기사건수도 2015년도 30건에서 2021년도 8월말 현재, 1,853건으로 기하급수적으로 증가하는 등(빅카인즈(bigkinds.or.kr), 뉴스검색 분석) 지역의 인구 감소에 관한 관심도 높아지고 있다.

국가와 지방자치단체에서도 이처럼 심화하는 지역의 인구 문제에 대응하기 위해 다양한 정책을 추진하고 있으며, 이러한 정책의 근거가 되는 장래인구와 관련하여, 통계청에서는 5년 주기로 총인구와 시도 단위의

1) 데드크로스 : 사망자 수가 출생아 수 보다 많아지면서 인구가 자연 감소하는 현상

‘장래인구추계’를 공표하고 있다. 다만, 장래인구추계는 미래의 가정을 전제로 작성되는 시나리오에 기반하고 있음에 따라, 미래 가정의 변화가 발생할 경우 인구추계의 불확실성이 커지는 한계점도 가지고 있다(통계청, 장래인구특별추계(시도편):2017~2047년). 실제로 2021년이 다음 정기 장래인구추계 공표 예정 시기였으나, 이전 자료가 급변하는 인구변화를 제대로 반영하지 못하여 2019년에 ‘장래인구 특별 추계’를 공표하기도 하였다.

또한 시도 단위 인구추계의 시나리오는 5개[2]로, 총인구 추계 시나리오인 27개에 비해 다양한 가정에 기반하고 있지 않으며, 빈번한 지역 간 인구이동을 고려해야 하는 점, 지역별로 인구구성 요인의 예측값이 별도로 필요한 점 등 고려 요소가 많은 문제점을 기본적으로 가지고 있다.

이와 같은 가정에 기반한 불확실성과 추계의 복잡성을 가진 인구추계 모델의 특성은 이해관계에 따라 인구추계 데이터가 조정 가능한 편향성을 가지는 한계점을 보이게 된다. 실제 각 지방자치단체에서는 자유로운 지역 간 인구이동 요소를 반영하여, 사회적 유입인구 증가를 가정한 도시계획을 수립하고 추진하고 있다. 이에 따라 지역별 계획인구가 현실보다 과다하게 산정[3]되어, 계획만큼 증가하지 못한 인구로 결국은 다시 해당 지방자치단체의 재정을 압박하는 부정적 원인으로 작용하고 있다(경향신문, 2021.10.18.).

이에 본 연구에서는 기존의 인구추계 방법의 복잡성과 가정에 기반한 시나리오로 인해 발생할 수 있는 불확실성 등의 문제점을 보완하는 방안으로, 딥러닝을 활용하여 우리나라 제2의 도시로 비수도권 지역을 대표하는 상징성을 가지는 부산지역의 인구를 예측해보고 그 가능성을 타진하

---

2) 5개 시나리오(통계청, 장래인구특별추계, 2019) : 중위 시나리오를 기준으로 인구변동요인(출생, 사망, 인구이동)별 4개의 시나리오(고위・저위・무이동・출산율 현수준)를 추가로 설정, 국내이동은 중위가정만을 적용
- 고위(저위) 추계는 시도별 인구성장이 최대(최소)가 되는 시나리오로, 각 시도별 출산율 및 기대수명의 고위(저위)가정을 적용
- 무이동추계는 출산율과 기대수명은 중위수준이고 향후 국내 및 국제 이동은 미발생
- 출산율 현수준 추계는 기대수명과 국내 이동은 중위수준이고, 출산율은 ’18년 출산율이 지속

3) 2030년 기준 통계청 추계 인구는 5,193만명이나, 전국 계획인구 합산은 5,535만명임

고자 한다. 데이터 학습만으로 예측값을 도출할 수 있는 딥러닝으로 지역의 인구 예측이 가능하다면, 인위적인 가정 요소를 배제하고, 특정 지역 단위 인구 예측을 효과적으로 수행할 수 있는 등 인구 예측에 새로운 관점을 제시할 수 있을 것이다.

제2절 연구 방법 및 구성

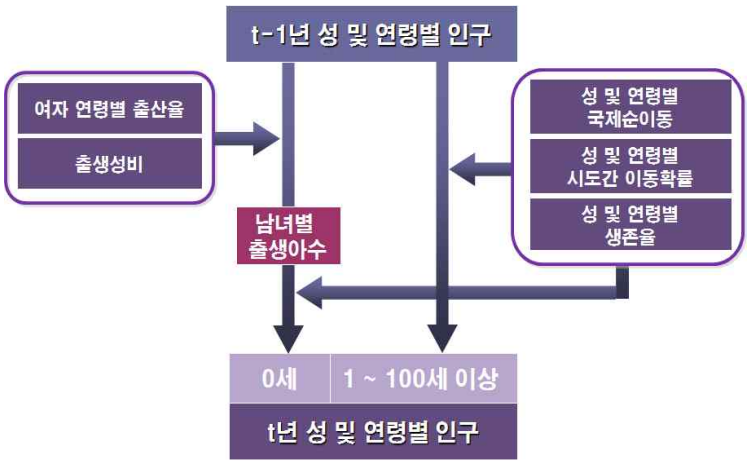
연구 방법은 지역 인구와 관련한 출생, 사망, 인구이동 등 인구균형방정식 요소를 데이터로 하여, 시계열 데이터에 적합한 것으로 평가받는 딥러닝 구조를 적용하여 예측 모델을 만들고 그 정확성을 평가한다.

연구 구성은 다음과 같다. 제1장 서론에서는 연구 배경과 목적, 연구 방법을 제시한다. 제2장에서는 인구 예측 및 딥러닝과 관련한 이론적 배경을 살펴보고, 관련 선행연구를 분석한다. 제3장에서는 부산지역 인구 예측을 위해 필요한 데이터 수집, 모델 설정 및 모델 평가 지표 등 연구 방법을 제시한다. 제4장에서는 하이퍼파라미터별로 모델의 성능을 비교한 뒤, 최적의 하이퍼파라미터를 도출하고, 모델별로 예측된 값을 실제 값과 비교하여 딥러닝을 통한 인구 예측 정확도를 확인한다. 마지막으로 결론에서는 연구 결과에 대한 요약 및 시사점, 연구의 한계를 제시한다.

제 Ⅱ 장 이론적 배경 및 선행연구 분석

제1절 통계청의 시도별 장래인구추계 방법

통계청에서는 시도별 장래인구추계에 코호트 요인법을 사용하고 있다. 코호트 요인법은 인구균형방정식을 적용해 <그림 1>과 같이 다음 해의 인구를 연속적으로 산출하는 방법이다. 인구균형방정식[4]은 출생·사망·이동 등 인구변동요인에 대한 미래수준을 예측하여 기준인구에 출생과 이동은 더하고, 사망은 제하는 방식이다. 즉, 장래인구의 규모와 구조는 인구추계가 이루어지는 초기의 인구 규모와 구조 그리고 설정된 출생, 사망, 인구이동 가정에 의해서 결정되며, 시도별 인구추계의 경우 전국단위와 달리 국내이동자 수가 고려되는 특성을 가지고 있다. 미래인구 예측 시나리오는 고위(인구성장 최대), 저위(인구성장 최소), 중위, 무이동, 현 수준 출산율 등 5개 시나리오를 기반으로 하고 있다.



출처 : 통계청 시도별 장래인구특별추계

<그림 1> 코호트 요인법에 의한 인구추계 과정

4) 인구균형방정식(Demographic Balancing Equation) : $P_t = P_{t-1} + B_{t-1} - D_{t-1} + NM_{t-1}$
(P_t: t년 인구, B_{t-1}: $t-1$년 출생아수, D_{t-1}: $t-1$년 사망자수, NM_{t-1}: $t-1$년 국내 및 국제 순이동자수)

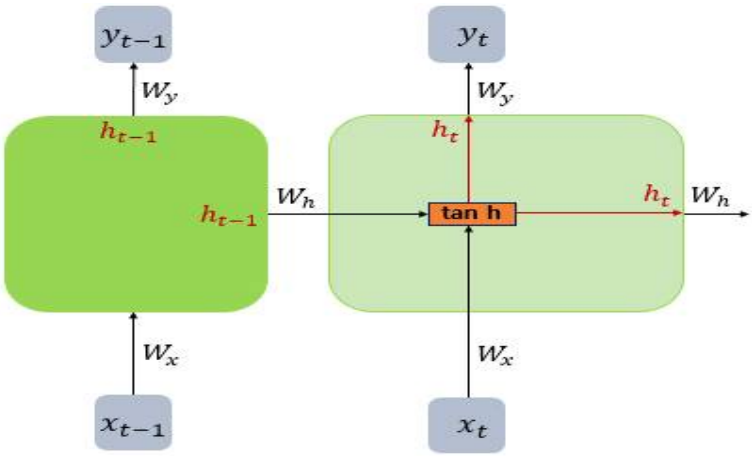
제2절 딥러닝

딥러닝은 여러 개의 은닉층을 가진 심층신경망에 기반한 학습 방법으로, 대량의 데이터로부터 데이터의 패턴인식이나 추론에 활용되고 있다. 심층신경망은 여러 개의 은닉층을 가지고 있는데, 상위 계층으로 올라갈수록 하위 계층을 통합하여 복잡한 특성을 추출하게 된다. 딥러닝은 훈련데이터에 대한 오류함수의 값을 줄이는 방향으로 가중치를 수정하는 오류역전파 알고리즘으로 학습을 수행한다. 다만, 층수가 깊은 심층신경망이나, 출력이 입력으로 되돌아오는 순환신경망에서는 가중치의 수정값이 0에 근접하는 '기울기의 상실' 문제가 발생하는데, 이를 극복하기 위해 활성화함수를 시그모이드함수가 아닌 미분값이 1 또는 0인 ReLU함수를 많이 사용하고 있다. 그 밖에도 학습속도와 안정성을 높이기 위해 입력 데이터 값을 정규화하고, 전이학습을 통해 학습에 성공한 신경망의 연결선 가중치를 새로 학습할 신경망 가중치에 이식하는 방법, 과적합 방지를 위해 연결선 수를 줄이는 등 딥러닝의 성능향상을 위한 다양한 방법들이 개발되고 있다.

현재 딥러닝은 데이터 특성 및 해결해야 할 문제에 따라 좋은 성과를 내는 방법론이 선택, 적용되고 있다. 공간의존성이 있는 영상 인식에서는 CNN, 시간적 행동 양태를 표현할 수 있는 RNN, RNN에서의 기울기 소멸 문제를 해결해 연속적인 음성 인식과 문장의 이해 등 시간정보 패턴에 좋은 성과를 내는 LSTM, 데이터를 분류하는 것이 아닌 생성의 목적으로 사용하여, 훈련 데이터와 유사한 새로운 데이터를 생성하는 GAN 등이 있으며, 새로운 심층 신경망도 계속적으로 제안되고 있다.

(1) 순환신경망 RNN (Recurrent Neural Network)

RNN은 <그림 2>와 같이 은닉층 노드에서 활성화 함수를 통해 나온 결괏값을 출력층 방향으로 보내며, 다시 은닉층 노드의 다음 계산의 입력으로 보내는 특징을 가지고 있다.



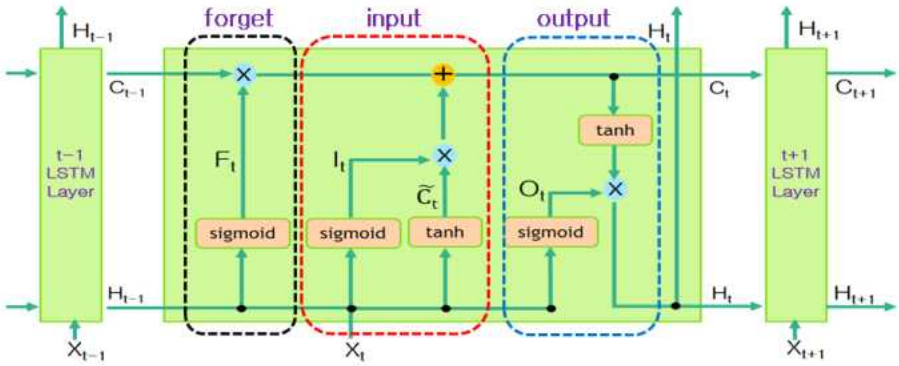
출처 : 딥러닝을 이용한 자연어 처리 입문

<그림 2> RNN(Recurrent Neural Network) 구조

이를 통해 이전의 값을 기억하는 일종의 메모리 역할 수행이 가능하게 되어 시간적 상관관계를 가지는 데이터(Sequential data) 입력을 처리할 수 있게 된다. 그러나, RNN은 각 출력 부분의 기울기가 이전 값에도 의존적이다 보니, 시점이 길어질수록 반복적으로 곱해지는 가중치에 의해 누적에러가 기하급수적으로 증가하거나, 빠르게 0으로 수렴하는 문제가 발생할 수 있는 한계점을 가지고 있다(이은주, 2017).

(2) 장단기기억신경망 LSTM (Long Short-Term Memory)

LSTM은 RNN의 한계점을 보완하기 위해서 등장한 것으로, RNN의 은닉층 셀에 3개의 gate(forget gate, input gate, output gate)를 추가하여 현재 시점의 정보를 바탕으로, 과거 내용을 얼마나 잊을지, 기억할지 등을 계산하고 그 결과에 현재 정보를 추가해서 다음 시점으로 정보를 전달한다. 이를 통해 RNN보다 긴 시퀀스의 입력을 처리하는데 좋은 성능을 보이고 있다.



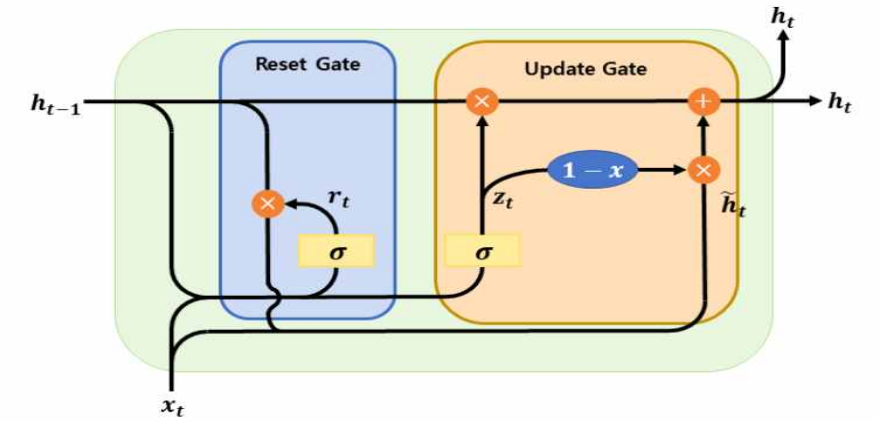
출처 : NeoWizard 블로그

<그림 3> LSTM(Long Short-Term Memory) 구조

<그림 3>의 LSTM의 기본구조를 살펴보면, forget gate는 과거 정보를 얼마나 잊거나 기억할지 결정하는 게이트로, 현 시점의 데이터와 과거 은닉층 값에 각각 가중치를 곱한 후 sigmoid 함수를 적용하여 과거 정보를 얼마나 활용할지 결정하게 된다. input gate는 현 시점의 데이터와 과거 은닉층 값에 각각 가중치를 곱한 후 sigmoid 함수를 적용하여, 어떤 정보를 업데이트 할지 결정(I_t)하고, 현재 시점의 데이터와 과거 은닉층 값에 각각의 가중치를 곱한 후 tanh함수를 적용해 현재 시점의 새로운 정보($\tilde{C}_t$)를 생성한다. ouput gate는 forget gate와 input gate에서 변경된 현재 시점의 메모리 셀 값(C_t)을 얼마나 빼서 다음 층으로 전달할지 결정한다.

(3) 게이트 순환 유닛 GRU (Gated Recurrent Unit)

LSTM에서는 3개의 게이트가 존재하는 복잡한 구조를 가지고 있다. 이를 업데이트 게이트와 리셋 게이트 두 가지 게이트만으로 구성하여, 구조를 간단화 시킨 것이 GRU이다.



출처 : GRU를 활용한 인공지능기반 구조물 시계열 응답 예측

<그림 4> GRU(Gated Recurrent Unit) 구조

<그림 4>의 GRU 기본구조를 보면, Reset Gate의 r_t는 과거의 은닉 상태를 얼마나 무시할 것인지를 결정한다. Update Gate는 은닉층을 갱신하는 게이트로, LSTM의 forget gate와 input gate를 합쳐놓은 역할을 수행하며, 과거와 현재 정보의 최신화 비율을 결정하게 된다.

제3절 선행연구 분석

지역 인구 예측과 관련하여 통계청의 장래인구추계 방법론의 한계를 극복하기 위한 다양한 시도가 이루어졌다. 통계청의 장래인구추계 방법론인 코호트 요인법은 외생적으로 주어지는 인구변동 요인들에 대한 가정에 기초해서 장래 인구를 계산하는 도구로, 가정의 불확실성에 따른 한계를 내포하고 있다.

우해봉(2009)은 우리나라 인구추계의 정확성과 시사점 연구에서, 코호트 요인법에 기반한 인구추계의 정확성이 기초 자료의 질적 향상에 따른 정확성 향상 외에는 뚜렷하게 향상되고 있다는 패턴을 찾기가 어렵다고 하였다. 또한 인구변동 요소와 관련하여 출산력은 과대 추정되고 기대 수

명은 과소 추정되는 문제가 지속적으로 나타났다. 이에 따라 1980년대 이후에는 확률적 인구추계의 활용도가 높아지고 있다고 하였다.

김형기, 문경종(2011)의 한국의 시도별 장래인구 예측에서도 코호트 요인법의 경우 세분된 코호트별 변동요인에 의한 추정에 어려움이 증가하므로, 출생, 사망, 외국계인구, 입지변화, GRDP 등 5개 요소에 의한 확률적 인구예측 방법(시계열 모형)을 제시하였다. 또한 송용호(2012)는 지역 방법론에 관한 연구에서 지역 단위 인구 예측에는 지역의 수만큼 인구구성 요인의 예측값이 필요함에 따른 비효율성을 지적하며, 각 지역의 인구변화 추세를 활용하는 방법을 제안하기도 하였다.

그 외에도 조대현, 이상일(2011)은 이지역 코호트-요인법을 이용한 부산광역시 장래 인구추계에서 지역 간의 상호작용을 고려하여 통계청의 순이동 코호트 요인법을 보완하기 위해 전체 지역을 두 지역으로 구분한 이지역 코호트 요인법에 대한 연구가 이루어지기도 하였다.

O. Folorunso1 el al.(2010)은 나이지리아의 인구 예측 연구에 인공신경망을 활용하였다. 그 이유는 인공신경망이 통계적으로 낙후된 나이지리아 인구 데이터의 결함과 같이 불리한 작동 조건에서도 성능이 크게 저하되지 않으며, 선형 데이터 및 비선형 데이터에 모두 좋은 성능을 보여주기 때문이다. 출산율과 사망률, 이주의 세 가지 인구 통계학적 변수를 입력 데이터로 하고, 두 개의 은닉층을 가진 인공신경망을 MATLAB(6.5버전) SW를 통해 구현한 연구 결과 인구 통계학적 모델인 코호트 요인법의 정확도인 64.55~86.43%보다 높은 81.02~99.15%의 정확도를 보여주었다.

시계열데이터 예측과 관련하여, 시장 상황이 일정한 추세를 보이는 경우 시계열 분석모형(ARIMA모형 등)과 머신러닝 방법(LSTM모형 등) 모두 유의미한 예측력을 보여주었으나, 시장이 비선형 형태로 급변하는 경우 선형모형을 가정하는 시계열 분석모형보다 비선형 모델링이 가능한 머신러닝 방법이 유의미하다고 보았다(배성완, 유정석, 2018). 또한 시계열 데이터 분석에 머신러닝을 적용해도 예측력이 우수하며, 시계열 분석의 경우 통계적 가정이 다른 여러 종류의 모델이 존재함에 따른 분석의

어려움이 있지만, 머신러닝은 해당 모델의 알고리즘에 대한 이해만 있다면 시계열 분석보다 용이하게 모델링이 가능한 장점을 시사한 바 있다. (지미경, 2019)

최근 주식, 부동산 등 자산 가격뿐만 아니라 물동량 예측, 수요량 예측 등 다양한 분야의 예측에 딥러닝을 활용한 연구가 활발하다. 그러나 딥러닝을 활용한 인구 예측 연구는 상대적으로 드문 편이다. 기지국을 활용한 유동인구나 서울시 생활인구 등 특정 지역 내 단기간 인구 유동성을 예측하기 위한 시도는 있으나, 지역 단위의 인구 예측에 딥러닝 등 머신러닝을 활용한 사례는 찾아보기 어려웠다.

하지만 해외 연구 사례에서도 인공신경망을 활용한 인구예측에 대한 가능성을 확인할 수 있었으며, 시계열 데이터 예측에 있어 머신러닝 방법에 의한 예측력의 우수성이 입증되고 있는 만큼 이번 연구에서도 딥러닝을 활용한 우리나라 부산지역의 장래인구 예측 모델의 가능성을 타진해 보고자 한다.

제 Ⅲ 장 연구방법

제1절 데이터 설명

본 연구에서 사용한 데이터는 <표 1>과 같이 인구균형방정식의 구성요소인 출생아 수, 사망자 수, 전출자 수 및 전입자 수와 부산지역의 주민등록 인구수 등 5개 변수와 관련한 데이터를 대상으로 하였다. 해당 데이터는 통계청의 국가통계포털(kosis.kr)을 통해 수집하였다. 기간은 1997년부터 2020년까지이며, 데이터 주기는 '연' 단위이다. 총 데이터 수는 120개(24×5) 이다.

<표 1> 데이터 개요

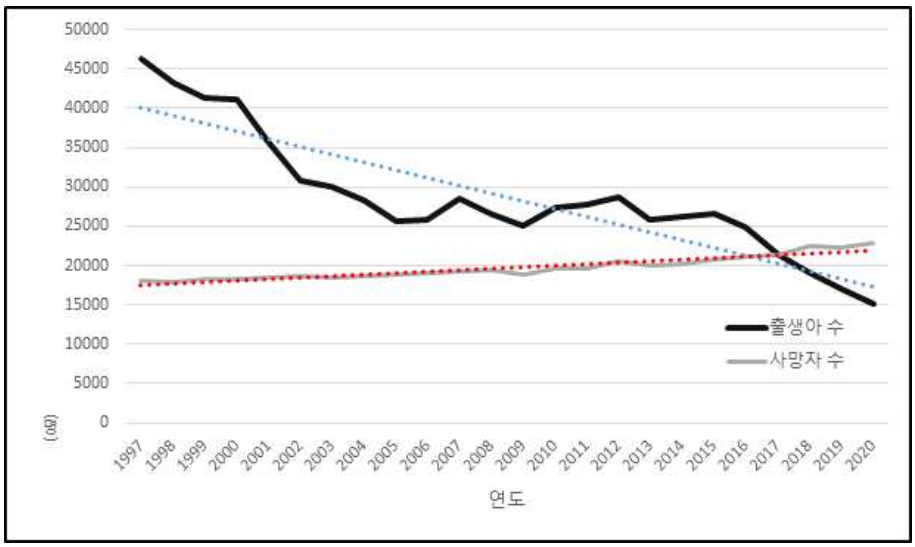
데이터명	데이터 정의	데이터수
출생아수	대한민국 국민에 의해 발생된 출생인구의 수	24개
사망자수	국내거주 사망인구의 수	24개
전출인구	행정구역 경계를 넘어 특정 지역에서 다른 지역으로 이동해 간 인구	24개
전입인구	행정구역 경계를 넘어 다른 지역에서 특정 지역으로 이동해 온 인구	24개
주민등록 인구수	지자체별 주민등록(거주자)이 되어있는 자의 인구	24개

<표 2>의 수집된 부산지역 인구 데이터를 보면, 부산지역의 자연 인구 증감 요소인 출생아 수 및 사망자 수의 경우, 출생아 수는 매년 감소하고 있으며 사망자 수는 증가하는 추세를 보이고 있다. 출생아 수는 1997년 46,284명이나, 2020년 현재는 32% 수준인 15,058명 수준이며, 사망자 수는 1997년 18,005명대비 125% 수준인 22,950명에 이르고 있다. 주민등록 인구수도 매년 감소하여, 2020년 현재 인구수는 1997년 대비 88% 수준인 339만명 수준이다.

<표 2> 부산지역 인구 데이터

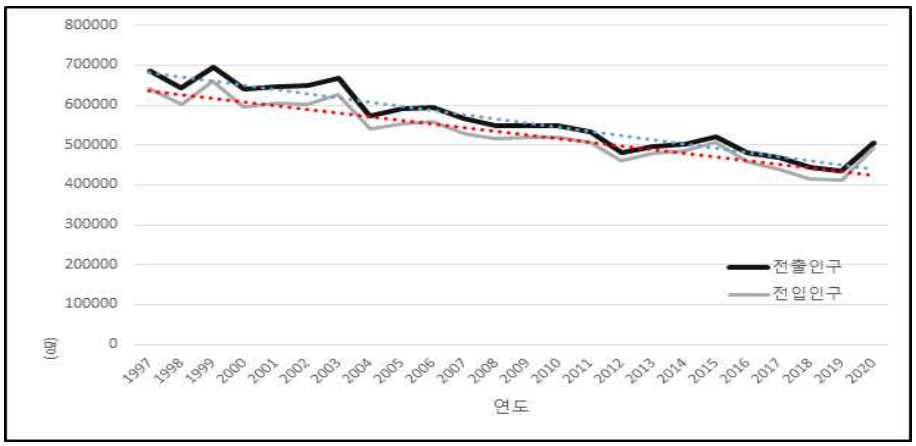
연도	출생아 수 (명)	사망자 수 (명)	전출인구 (명)	전입인구 (명)	주민등록 인구수(명)
1997	46,284	18,005	686,323	641,886	3,851,312
1998	43,200	17,928	643,339	602,418	3,829,098
1999	41,237	18,312	693,417	660,060	3,817,270
(중 략)					
2019	17,049	22,260	435,058	411,704	3,413,841
2020	15,058	22,950	506,176	491,829	3,391,946

1997년에서 2020년까지의 출생아 수와 사망자 수에 대해 Excel을 활용하여 선형 추세선을 그어보면 <그림 5>의 점선과 같이 출생아 수는 우하향하며 감소하고 있고, 사망자 수는 점진적으로 증가함을 알 수 있다. 이에 따라 2017년도에는 사망자 수가 출생아 수를 역전하는 '인구 데드크로스'가 발생하였다. 이는 우리나라 총인구를 대상으로 한 '인구 데드크로스' 시점인 2020년보다 3년 이른 시기이며, 그 차이는 매년 벌어지고 있다.



<그림 5> 출생아 수 및 사망자 수 (1997~2020년)

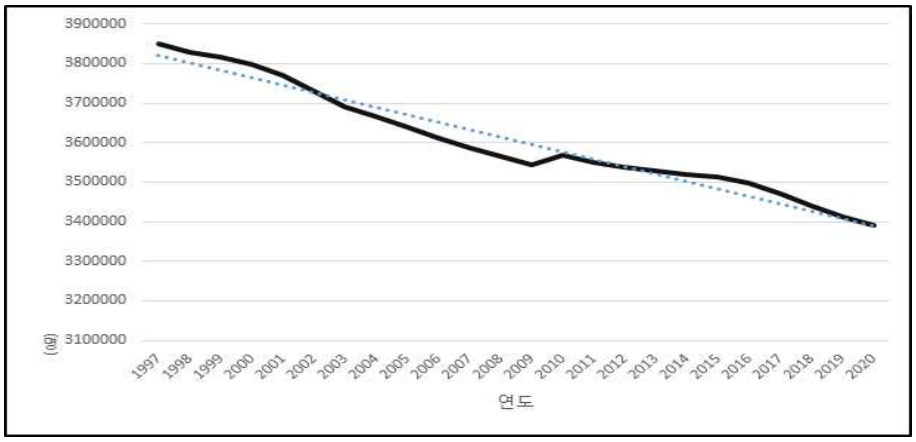
<그림 6>의 지역 간 이동인구를 나타내는 전입 및 전출인구를 보면, 전입인구가 많은 연도에는 전출인구도 많았으며, 전입인구가 적은 연도에는 전출인구도 적은 수를 나타내는 등 비슷한 움직임을 보인다.



<그림 6> 전입·전출인구 (1997~2020년)

매년 전출인구가 전입인구보다 많은 수가 발생하고 있어, 부산지역 인구는 지역간 이동에서도 지속적으로 유출되고 있음을 알 수 있다. 또한 전출 및 전입 인구 규모는 점선의 추세선과 같이 하향 추세에 있다. 2020년도는 예외적으로 전출 및 전입인구가 일시적으로 상승하였다.

이러한 결과 부산지역 인구는 <그림 7>과 같이 2010년도에 일시적 상승을 보인 것을 제외하고는 매년 감소하는 추세이다.



<그림 7> 부산지역 인구 (1997~2020년)

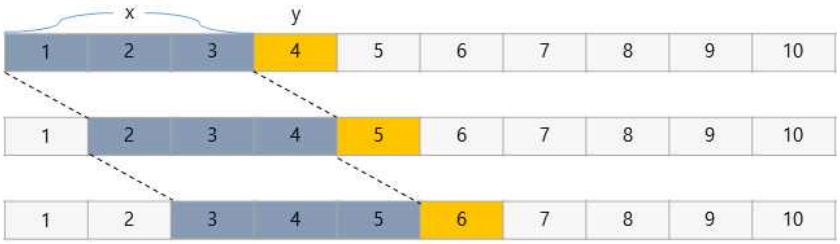
제2절 모델 설정

부산지역 인구예측을 위해 시계열 데이터에 적합한 RNN, LSTM, GRU를 적용하였으며, 신경망의 기본 구조인 MLP[5](Multi Layer Perceptron) 모델 결과와 비교하였다.

모델은 i5@1.60GHz, 8GB RAM, 238GB SSD HW에서 Jupyter note-book(Python 3.7.4 버전, tensorflow 2.6.0 버전, keras 2.6.0 버전)환경에서 설정하였다.

데이터의 처리 방법은 인공 신경망에서 시계열 예측을 수행하는 기법인 슬라이딩 윈도우(Sliding Window)를 적용하였다. 다른 시계열 관련 연구에서 슬라이딩 윈도우(Sliding Window)가 적용되었을 때의 예측력이 적용하지 않았을 때보다 더 좋게 나타났다.(전형진, 2021) 슬라이딩 윈도우(Sliding Window)는 <그림 8>과 같이 이전 일정 시간의 데이터를 입력(x)으로, 다음 시간 데이터를 목표(y)로 하고, 입력(x)와 목표(y)가 전체 훈련 셋에 걸쳐 순차적으로 슬라이딩 되는 기법이다. 이때 윈도우 사이즈를 어떻게 설정하느냐에 따라서도 모델 성능에 영향을 줄 수 있는데, 윈도우 사이즈를 크게 한다고 해서 모델 성능이 좋아지는 것은 아니다.(Frank et al., 2001) 이번 연구에서는 5개의 변수(출생아 수, 사망자 수, 전입인구, 전출인구, 주민등록 인구)를 입력으로 하였으며, 다음 윈도우의 주민등록 인구를 출력으로 하였다. 윈도우 크기는 '3년' 간격으로 설정하였으며 출력값은 '3년 + 1년' 시점의 주민등록 인구가 된다.

5) MLP는 입력층과 출력층 사이에 하나 이상의 은닉층을 두고, 역전파 학습 알고리즘을 통해 학습을 수행하는 구조를 가지고 있다. 이를 통해 이전 단층 신경망에서 해결하지 못한 XOR문제를 해결하였으며, 문자인식, 영상인식, 가격예측 등 다양한 분야에 신경망이 활용되도록 기여하였다.



<그림 8> 슬라이딩 윈도우(Sliding Window)

데이터는 학습을 위한 데이터 셋인 훈련 셋(Train Set)과 모델의 성능을 최종적으로 평가하기 위한 테스트 셋(Test Set)을 8:2로 분리하였다. 훈련 셋으로 학습한 모델의 성능을 측정하기 위하여 훈련 셋을 다시 검증 셋(Validation Set)으로 나누기도 하지만, 이번 연구에서는 훈련 데이터를 최대한 확보하기 위하여 별도로 구분하지 않았다.

다음으로 변수마다 값의 범위가 달라서 특정 데이터 크기에 영향을 받는 것을 방지하기 위하여, 데이터 전처리(Feature Scaling)를 통해 5개 변수의 값을 일정한 수준으로 맞춰주는 데이터 전처리 과정을 거쳤다. 데이터 전처리(Feature Scaling) 방법은 변수의 값을 0과 1 사이에 오도록 하는 정규화 방법을 사용하였으며, Sklearn에서 제공하는 MinMaxScaler를 통해 구현하였다. 정규화의 수식은 다음과 같다.

$$x_i = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

딥러닝 모델의 성능은 입력 뉴런 수, 은닉층 수 등 조정가능한 하이퍼파라미터(Hyper-parameter) 최적화에 따라 결정된다. 최적화된 하이퍼파라미터 조합을 찾기 위하여, 먼저 MLP에 Sklearn의 RandomizedSearchCV[6]를 적용해 층 수와 노드 수의 기준값을 설정하였다. 그리고 기준값보다 더

6) RandomizenSearchCV는 각 반복마다 하이퍼파라미터에 임의의 값을 대입해 지정한 횟수만큼 평가하는 방식으로, 하이퍼파라미터마다 각기 다른 값을 탐색하는 장점을 가지고 있다.

복잡하거나 단순한 신경망을 구성하여 그 성능을 비교하였다. MLP에 적용한 하이퍼파라미터 값을 다른 딥러닝 모델에도 적용하여, 성능을 비교하였다.

데이터 학습 횟수인 Epoch설정을 위해서 Keras의 EarlyStopping 함수를 적용하였다. 학습이 진행될수록 훈련 셋의 정확도는 올라가지만, 테스트 셋의 실험 결과는 점점 나빠지게 되는 과적합이 발생하게 되므로, EarlyStopping 함수를 통해 일정 횟수 동안 오차가 줄어들지 않으면 학습을 멈추도록 하였다. 최대 Epoch은 500회로 하였으며, EarlyStopping 모니터링 지표는 'MSE', 오차가 줄어들지 않는 횟수를 설정하는 매개지표인 'Patience'는 20회로 설정하였다.

오차를 비교하여 가장 작은 방향으로 이동시키며, 가중치를 업데이트하는 방법으로는 빠른 속도와 안정적 성능으로 평가 받는 아담(Adam)을 적용하였다.

제3절 모델 평가 지표

부산지역 인구 예측 모델 평가 지표로 예측 모델에서 주로 사용되고 있는 MAE, RMSE, MAPE 등 세 가지 평가 지표를 사용하였다.

MAE(Mean Absolute Error)는 실제값에서 예측값을 뺀 값의 절댓값 평균으로, 평균적인 오차가 어느 정도 인지를 직관적으로 보여 주는 장점이 있다.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y_i}\right|$$

RMSE(Root Mean Squared Error)는 MSE에 루트를 씌운 것으로, MSE가 실제값에서 예측값을 뺀 값을 제곱함에 따라 1 미만은 더 작아지고, 1 이상은 더 크게 측정되는 단점을 보완하였다.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$$

MAPE(Mean Absolute Percentage Error)는 실제값에서 예측값을 뺀 값을 실제값으로 나눈 것으로, 실제값에 대한 오차의 비율을 나타낸다. 이를 통해 MAE와 RMSE가 예측하려는 값에 영향을 받는 단점을 보완하고, 오차의 비율을 통해 모델을 비교 평가할 수 있다.

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{y_i}\right| \times 100$$

MAE나 RMSE는 직관적으로 예측값의 정확성을 파악할 수 있는 장점이 있는 반면, 0~무한대의 값을 가지게 되어 값이 클수록 오차값이 커지는 단점을 가지고 있다. MAPE는 0~100%의 값을 가지게 되므로 MAE나 RMSE와 비교하여 직관성은 떨어지더라도, 성능 비교에는 장점이 있다. 따라서 이번 연구에서는 세 가지 평가 지표를 가지고 예측 모델의 성능을 검증한다.

제 Ⅳ 장 결과분석

제1절 MLP를 통한 예측결과

1997년부터 2015년까지의 부산지역 인구 데이터를 훈련 셋으로 하여, MLP 모델을 구축하였다. RandomizedSearchCV를 통해 도출된 제일 적합한 층수와 노드 수를 기준으로, 해당 모델보다 복잡한 모델과 단순한 모델을 만들어 결괏값을 비교하였다.

<표 3>과 같이 기준 모델은 3개의 층과 24개의 노드 수로 구성된 구조로 되어 있으며, 이를 기준으로 층수는 2~4개, 노드 수는 층별로 동일한 노드 수를 설정하거나 층별 노드 수를 줄여나가는 구조로 하이퍼파라미터를 설정하여 10개의 모델을 만들었다.

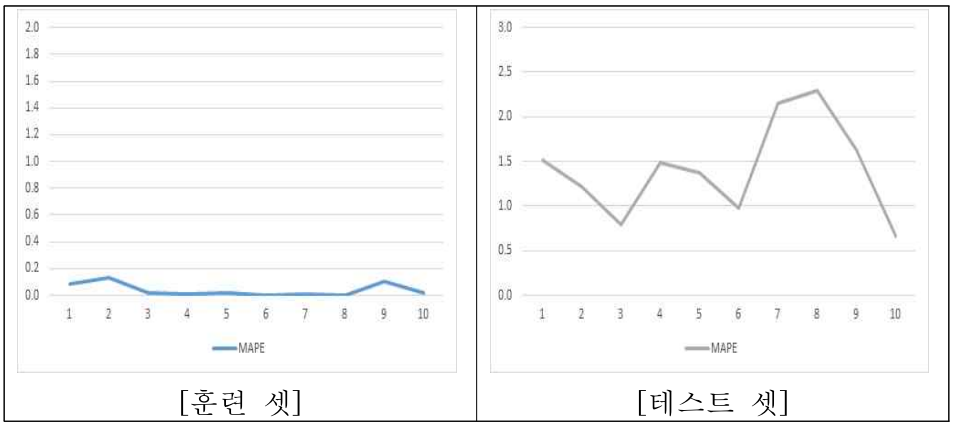
<표 3> 모델별 하이퍼파라미터 값

모델	기준	1	2	3	4	5	6	7	8	9	10
층	3	2	2	3	3	3	3	4	3	3	3
노드 수	24	12	48	12	17	20	48	12	24-12-6	12-6-3	17-17-9

모델별 평갓값은 <표 4>와 같으며, 훈련 셋과 테스트 셋의 MAPE는 <그림 9>와 같다. 전반적으로 훈련 셋에 과적합 되는 모습을 보였으며, 모델별로 비교 시 모델의 복잡성과 모델의 성능과는 관련성이 없었다. 층수를 4로 깊게 구성한 경우(7번모델)에도 좋지 않은 성능을 보였으며, 노드 수가 많거나 층별 노드 수가 순차적으로 줄어드는 모델에서도 MAE와 RMSE의 값이 크게 나타났다.

<표 4> MLP 모델별 평가 결과

모델	MAE		RMSE		MAPE	
	훈련셋	테스트셋	훈련셋	테스트셋	훈련셋	테스트셋
기준	700.3	28008.2	1057.9	29870.5	0.019	0.812
1	2983.5	51745.0	3837.4	58093.3	0.083	1.511
2	4813.0	41579.5	5917.0	50079.1	0.133	1.216
3	542.9	27096.9	648.3	29528.3	0.015	0.791
4	529.8	50881.9	649.0	72381.7	0.015	1.493
5	743.8	47296.5	1017.8	51739.1	0.021	1.380
6	0.0	33383.5	0.1	36516.0	0.000	0.974
7	401.3	73490.7	546.4	85225.0	0.011	2.148
8	0.5	78675.5	0.6	93261.6	0.000	2.301
9	3660.3	55713.5	5173.3	65938.8	0.102	1.629
10	687.6	22730.8	1039.3	23382.5	0.019	0.660

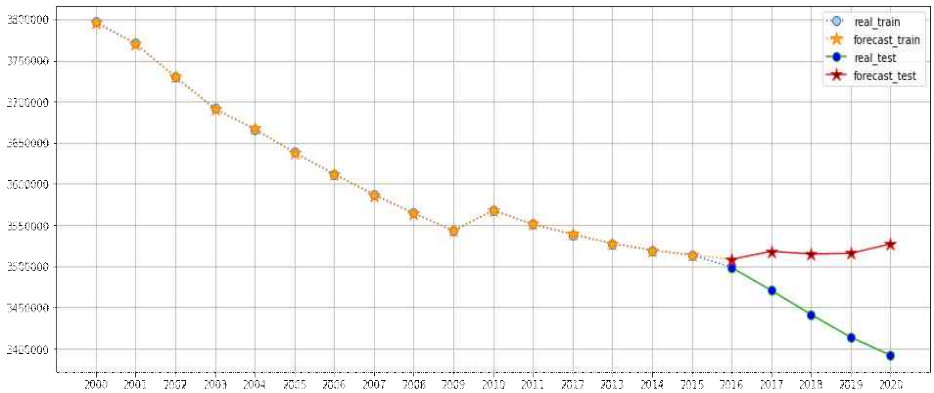


[훈련 셋] [테스트 셋]

<그림 9> MLP의 훈련 셋 및 테스트 셋에 대한 MAPE 결과

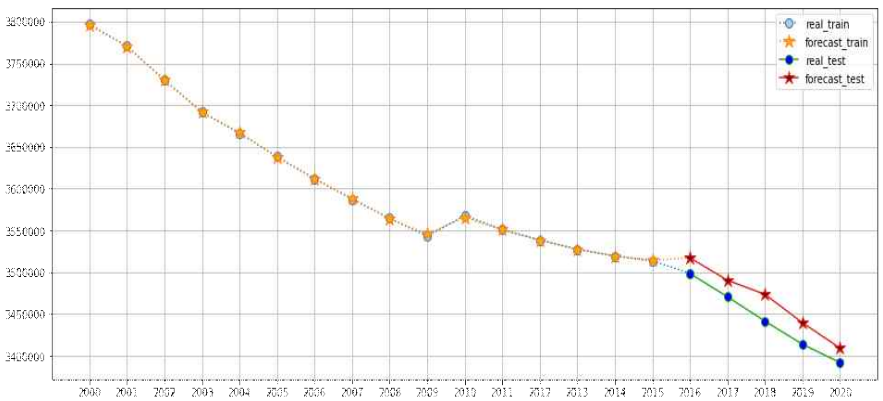
특히 4개 층을 가진 7번 모델의 경우, <그림 10>과 같이 테스트 셋에 대한 예측값이 정반대로 나타나기도 했다. 2000년부터 2015년까지의 훈련

셋의 실제 데이터값에는 과적합되며 점선 별표와 같이 거의 일치되는 경향을 나타내었고, 2016년부터 2020년까지의 테스트 셋의 실제 데이터는 계속 하향 추세이지만, 예측값인 실선 별표는 오히려 상승하는 추세를 보여 주었다.



<그림 10> MLP_7번 모델 실제값과 예측값 결과 비교

기준모델과 비교 시 가장 성능이 좋은 모델은 3개의 은닉층과 노드 수는 17개-17개-9개의 구조로 되어 있는 10번 모델이었다.



<그림 11> MLP_10번 모델 실제값과 예측값 결과 비교

<그림 11>에서와 같이 훈련 셋에 대한 점선 별표의 10번 모델의 값은

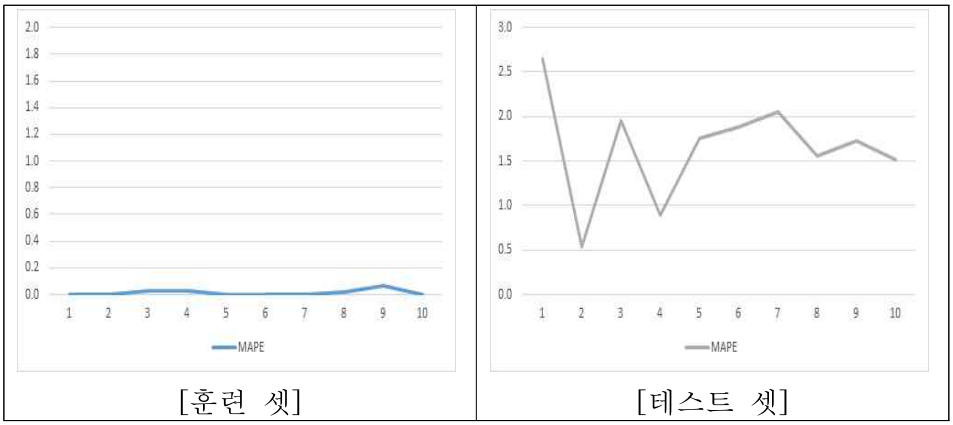
실제 데이터에 과적합 된 모습을 보이기는 하지만, 5개년(2016~2020년)에 걸쳐 감소 추세에 있는 테스트 셋의 실제 데이터와 비교해 보면, 실선 별표와 같이 인구수가 매년 감소하는 예측값을 보여 주었다. MAPE는 0.66을 기록하였다.

제2절 RNN, LSTM, GRU 모델별 결과 비교

제1절의 <표 3>의 모델 구조를 RNN, LSTM, GRU에 각각 적용하고, 해당 모델별 결과를 비교하여 어떤 모델이 가장 좋은 성능을 내는지 알아보았다. 먼저, RNN에 대한 모델별 평가 결과는 <표 5>와 같이, MLP에 비해 훈련 셋에 과적합 되는 모습을 보였다. <그림 12>의 테스트 셋에 대한 MAPE를 보면, 전반적으로 모델이 복잡할수록 MAPE값이 올라가는 경향이 나타났다. RNN의 최적 모델은 2번 모델로, 은닉층 2개와 48개씩의 노드로 구성된 모델이다. MLP_10번 모델의 MAPE 0.66과 비교 때에도 0.54의 MAPE를 보이는 등 좋은 평가 결괏값을 나타내었다.

<표 5> RNN 모델별 평가 결과

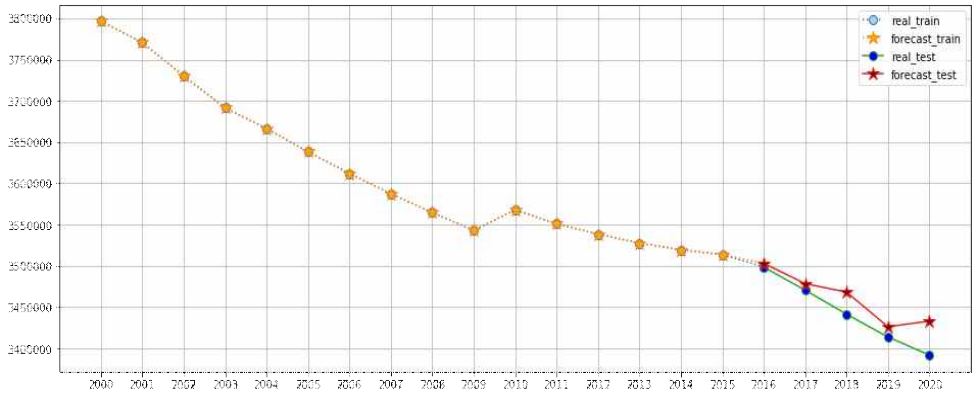
모델	MAE		RMSE		MAPE	
	훈련셋	테스트셋	훈련셋	테스트셋	훈련셋	테스트셋
MLP	687.6	22730.8	1039.3	23382.5	0.019	0.660
1	103.9	90701.8	145.6	110336.1	0.003	2.654
2	0.0	18558.7	0.1	23034.0	0.000	0.543
3	905.8	67004.4	1353.1	73505.5	0.025	1.956
4	1186.3	30788.4	1449.5	32691.4	0.032	0.897
5	0.3	60319.0	0.3	70176.2	0.000	1.763
6	2.9	64253.4	4.3	72928.8	0.000	1.877
7	25.9	70242.4	39.3	77436.7	0.001	2.051
8	711.3	53513.0	975.5	61023.7	0.020	1.564
9	2437.4	59161.1	2901.5	67249.3	0.067	1.729
10	106.3	51967.9	182.3	57294.6	0.003	1.517



[훈련 셋] [테스트 셋]

<그림 12> RNN의 훈련 셋 및 테스트 셋에 대한 MAPE 결과

그러나 <그림 13>과 같이 2번 모델의 실제값과 예측값 비교 그래프를 보면 2019년까지는 예측이 잘 이루어지다가 2020년은 인구수가 증가하는 것으로 예측하였다. 이는 2020년에 일시적으로 전출입 인구가 증가한 경향과 닮았으며, 전반적인 인구수 하향 추세를 잘 따라가고 있다고는 볼 수 없다.



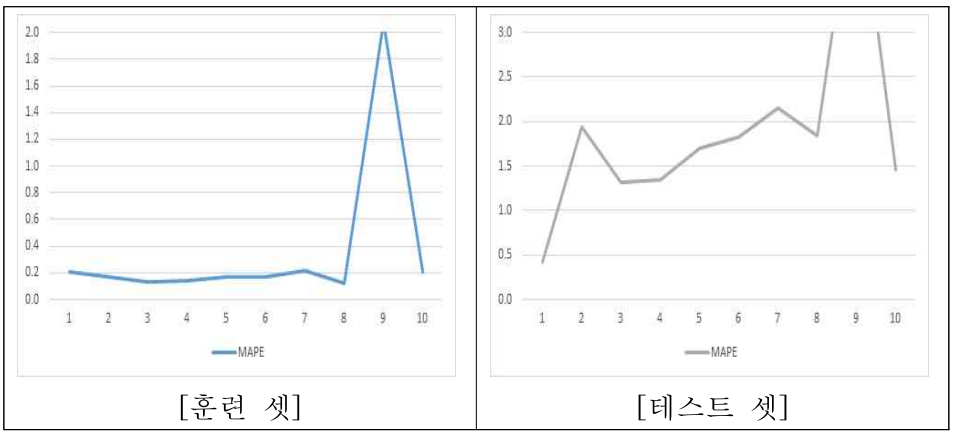
<그림 13> RNN_2번 모델 실제값과 예측값 결과 비교

LSTM 모델별 평가 결과는 <표 6>과 같으며, RNN과 비교 시 훈련 셋에 과적합이 크지는 않았다. <그림 14>의 MAPE 그래프를 보면, LSTM 모델도 복잡한 구조로 갈수록 MAPE값이 높아지는 경향을 보이

고 있다. LSTM의 최적 모델은 1번 모델로, 은닉층 2개와 12개 씩의 노드로 구성된 단순한 모델이다. MLP와 비교시 MAPE값은 0.419로 더 우수한 성능을 기록하였다.

<표 6> LSTM 모델별 평가 결과

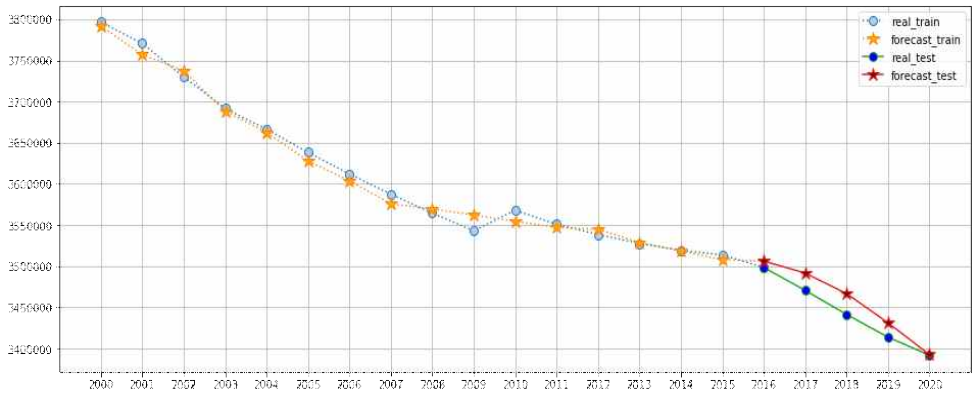
모델	MAE		RMSE		MAPE	
	훈련셋	테스트셋	훈련셋	테스트셋	훈련셋	테스트셋
MLP	687.6	22730.8	1039.3	23382.5	0.019	0.660
1	7406.2	14448.2	8855.7	17004.9	0.205	0.419
2	5997.7	66446.1	8073.3	74818.9	0.166	1.941
3	4633.2	45282.3	6532.4	49618.3	0.129	1.322
4	5272.7	46173.0	7321.6	50572.4	0.146	1.348
5	6152.8	58338.9	8471.5	65103.7	0.170	1.704
6	6317.0	62432.2	8810.6	70631.8	0.175	1.824
7	7955.7	73582.2	10048.7	81756.2	0.220	2.149
8	4578.7	63209.7	7047.3	70882.5	0.128	1.846
9	75795.0	167062.4	89176.8	171380.3	2.081	4.865
10	7473.6	50042.0	9654.6	53412.0	0.206	1.459



[훈련 셋] [테스트 셋]

<그림 14> LSTM의 훈련 셋 및 테스트 셋에 대한 MAPE 결과

다만, 1번 모델을 제외하고는 전반적으로 성능이 좋은 편은 아니었으며, 특히 9번 모델의 경우 훈련 셋과 테스트 셋 모두 제대로 학습하지 못하는 결과를 보여주었다.



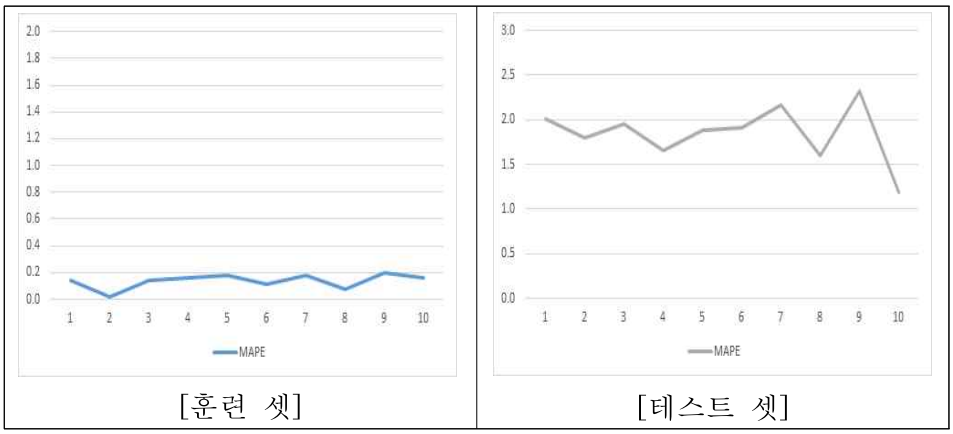
<그림 15> LSTM_1번 모델 실제값과 예측값 결과 비교

LSTM_1번 모델은 RNN과는 달리, 테스트 셋의 예측값을 나타낸 <그림 15>의 실선 별표에서 볼 수 있듯이, 매년 인구가 감소되는 추세를 잘 나타내고 있음은 물론, 예측 인구 수도 근사값으로 보여 주었다.

GRU의 경우 <표 7>에서와 같이 전반적으로 평가 결괏값이 좋지 않았으며, 이번 연구에서 적용한 하이퍼파라미터로는 좋은 성능을 보여 주지는 못하였다.

<표 7> GRU 모델별 평가 결과

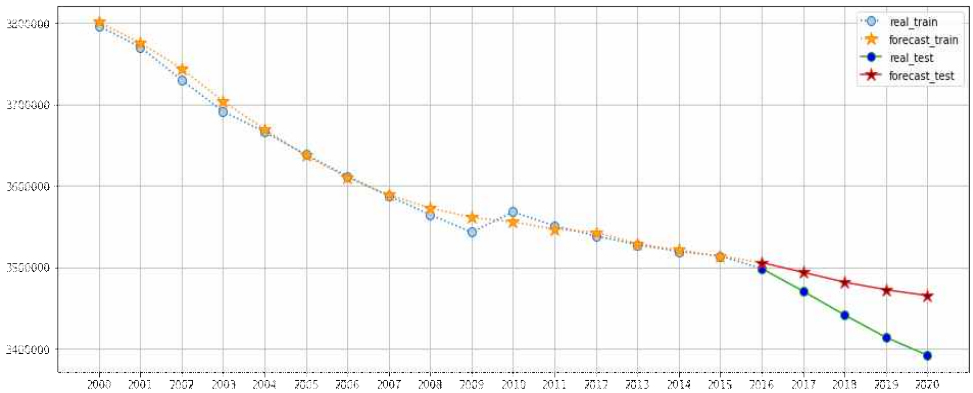
모델	MAE		RMSE		MAPE	
	훈련셋	테스트셋	훈련셋	테스트셋	훈련셋	테스트셋
MLP	687.6	22730.8	1039.3	23382.5	0.019	0.660
1	5250.3	68708.8	6832.9	78191.0	0.146	2.008
2	876.7	61556.2	1100.8	68954.6	0.024	1.798
3	5092.7	66823.7	7535.4	76471.1	0.142	1.953
4	5905.9	56603.1	7406.4	63351.6	0.164	1.653
5	6373.0	64328.4	8634.0	72657.3	0.176	1.879
6	3955.4	65224.1	5765.8	73692.7	0.110	1.906
7	6493.2	74434.8	8568.1	83371.0	0.181	2.174
8	2768.0	54602.8	3309.9	61194.6	0.076	1.595
9	7006.3	79641.6	8523.7	88341.2	0.195	2.326
10	5928.5	40478.3	7970.3	46903.4	0.164	1.183



[훈련 셋] [테스트 셋]

<그림 16> GRU의 훈련 셋 및 테스트 셋에 대한 MAPE 결과

<그림 16>에서와 같이 훈련 셋에 대한 학습은 비교적 이루어진 것으로 보이나, 테스트 셋의 결과를 보면 최저 MAPE가 1.18일 정도로, 좋은 성능을 보인 모델은 없었다. 특이한 사항은 GRU에서 테스트 셋의 MAPE가 높은 7번 모델과 9번 모델의 경우 LSTM에서도 높은 MAPE를 기록한 좋지 않은 모델이었다.



<그림 17> GRU_10번 모델 실제값과 예측값 결과 비교

GRU 모델 중 상대적으로 우수한 10번 모델을 보면, <그림 17>과 같이 연도가 갈수록 예측값이 실제값에서 멀어지는 모습을 보여 주었으며, MAPE는 1.183을 기록하였다.

<그림 18> 모델별 테스트셋에 대한 MAPE 결과

<그림 18>의 RNN, LSTM, GRU의 모델별 MAPE 결괏값을 보면, 가장 우수한 성능을 보여준 모델은 LSTM_1번 모델이며, RNN_2번 모델도 MLP에서 제일 우수한 모델인 MLP_10번 모델보다 좋은 성능평가 값을 보여주었다. 다만 RNN_2번 모델은 성능평가 값은 좋았으나 매년 감소하

고 있는 인구 추세와는 달리 2020년은 오히려 상승하는 인구수를 예측하는 문제를 드러내었다.

제3절 통계청 지역인구 추계와 결과 비교

마지막으로 통계청에서 공표한 지역인구 추계값과 이번 연구에서 성능이 가장 좋은 LSTM_1번 모델의 예측값을 비교하였다. <표 8>의 천명 단위 기준의 RMSE값을 보면, 통계청 추계 모델은 31.8, LSTM_1은 17.1이 나와 LSTM_1이 실제 주민등록 인구수를 더 근접하게 예측한 것으로 나타났다.

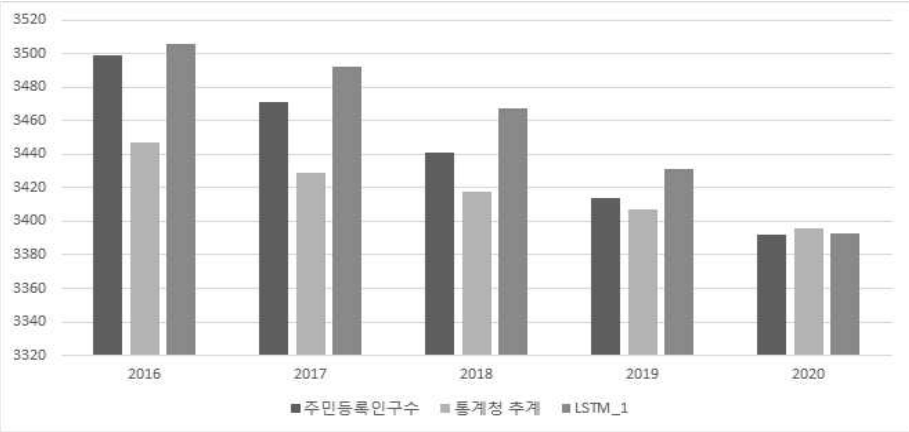
<표 8> 주민등록 인구수와 통계청 추계 및 LSTM 모델 비교

(단위:천명)

연도	주민등록 인구수 (y) / (성장률 %)	통계청 추계		LSTM_1	
		인구수 (y1)/ (성장률 %)	차이 (y-y1)	인구수 (y2)/ (성장률 %)	차이 (y-y2)
RMSE			31.8		17.1
2016	3499	3447	52	3506	-7
2017	3471(-0.80)	3429(-0.52)	42	3492(-0.40)	-21
2018	3441(-0.86)	3418(-0.32)	23	3467(-0.72)	-26
2019	3414(-0.78)	3407(-0.32)	7	3431(-1.04)	-17
2020	3392(-0.64)	3396(-0.32)	-4	3393(-1.11)	-1

실제 주민등록 인구수와 각 예측값을 비교해 보면, <그림 19>와 같이 통계청의 추계 차이도 매년 줄어드는 추세를 보인다. 다만 통계청의 인구추계는 인구총조사(등록센서스)를 기준[7]으로 추계하기 때문에, 주민등록 인구수와는 차이가 있음을 고려할 필요는 있다.

7) 기준인구는 2015년 7월 1일 시점의 인구로, 등록센서스 시도별 인구(2015년 11월 1일)에 2015년 7~10월 사이 발생한 인구변동요인과 국적변동을 가감해 작성



<그림 19> 주민등록 인구수와 통계청 추계 및 LSTM 모델값 비교

기준연도인 2015년도의 통계청 기준인구는 3,452천명으로, 주민등록 인구수 3,513천명보다 6만1천명이 적다. 이와 같은 기준 인구수의 차이로 인하여 통계청 추계 인구의 RMSE가 LSTM_1보다 크게 나타난 이유가 될 수 있다.

매년 추계 인구성장률을 <표 8>에서 보면, 통계청의 인구성장률은 2017년도에 -0.5% 이후 2018~2020년도는 매년 -0.3%로 일정하게 감소 추세를 보인다. LSTM_1 모델에서는 2017년도 0.4% 감소 이후 2020년도 -1.1% 성장률을 보이는 등 매년 인구성장률 감소비율이 커졌다. 실제 주민등록 인구수는 2017년도 0.8% 감소하였으며, 2020년도에는 -0.6%를 기록하는 등 인구성장률 감소 크기는 조금씩 줄어들었다. 인구성장률 기준에서 보면, 통계청의 추계 인구성장률 감소 크기는 전반적으로 낮은 수준이었으며, LSTM_1의 모델의 인구성장률은 좀 더 큰 감소 비율을 보여주었다.

제 V 장 결 론

제1절 요약 및 시사점

통계청에서는 '코호트 요인법'으로 5년 단위의 '시도별 장래인구 추계'를 공표하고 있다. 그러나 시나리오 기반의 추계 예측으로, 미래 가정의 변화에 따른 불확실성이 커지는 한계를 가지고 있으며, 선행연구에서도 코호트 요인법에 의한 인구예측의 정확성에 대한 문제 제기가 이루어져 왔다.

이러한 기존의 지역 인구 추계 방법의 한계성을 극복하는 방법으로 이번 연구에서, 데이터 학습을 통해 패턴을 추출할 수 있는 딥러닝을 적용한 지역인구 예측 모델의 가능성을 타진해보았다. 딥러닝은 영상인식, 분류, 가격 예측 등 다양한 분야에서 활용되고 있으며, 시계열 예측에서는 RNN, LSTM, GRU의 성능이 좋은 것으로 평가받고 있다.

연구에 사용된 데이터는 1997년에서 2020년까지의 연 단위 출생아 수, 사망자 수, 전입 인구, 전출 인구, 주민등록 인구이며, 총 데이터 수는 120개이다. 딥러닝의 성능은 은닉층 수, 노드 수 등 하이퍼파라미터에 의해 좌우되는데, 이번 연구에서는 서로 다른 하이퍼파라미터가 적용된 10개의 모델에 대한 예측 성능을 평가하였다. 평가 지표는 MAE, RMSE, MAPE를 적용하였다. 먼저 기본신경망 구조인 MLP를 적용한 모델에서는 가장 좋은 성능을 보인 모델의 MAPE는 0.66이며 5개년에 걸친 예측값이 실제값의 추이를 잘 따라가고 있음을 보여 주었다.

다음 3개 딥러닝 구조에 MLP에 적용한 하이퍼파라미터를 적용한 결과 LSTM모델에서 성능이 좋은 모델이 나왔다. RNN 모델의 경우 성능 평갓값은 좋았으나, 실제 인구 감소 추세를 잘 따라가지 못하는 문제점을 노출하였다. 특히 테스트 셋의 마지막 구간인 5년 차에는 정반대의 예측을 보여주었다. LSTM의 경우 모델 간 성능 편차가 크게 나타나긴 했으나 가장 좋은 성능을 보인 모델의 경우 인구 감소 추세를 잘 보여주었다.

특히 해당 모델의 경우 2개 층으로 이루어진 단순한 구조로, 모델의 복잡성보다는 하이퍼파라미터를 어떻게 조정하느냐가 예측 성능에 영향을 미친다고 볼 수 있다. 반면 GRU 모델의 경우 LSTM의 복잡한 구조를 단순화하여 파라미터 수가 적음에 따라, 학습 시간이 더 짧게 걸리고 보다 적은 데이터로도 학습이 가능하다는 특징으로 알려졌지만, 이번 연구에서 적용된 하이퍼파라미터에서는 성능 평가값도 좋지 못하였으며, 인구 감소 추세도 따라가지 못하였다. 다만, 이번 연구에서 사용된 데이터의 길이가 짧은 관계로 LSTM과 GRU의 고유한 특징이 충분히 활용되었다고는 볼 수 없다.

마지막으로 실제 주민등록 인구수에 대한 통계청 모델, LSTM 모델의 인구 예측값을 비교한 결과 LSTM 모델이 통계청 모델보다 오차값이 적었다. 다만, 통계청의 모델은 기준 인구를 주민등록 인구수가 아닌 인구총조사 기준인구를 쓴다는 점을 고려할 필요는 있다.

이상의 연구에서 변수별 가중치 설정, 인구변동 요인별 모델 선택의 사전 작업 없이도 적절한 하이퍼파라미터를 적용한 딥러닝을 활용할 경우, 데이터만으로도 지역 인구를 효과적으로 예측할 수 있는 모델을 만들어 낼 수 있음을 확인할 수 있었다. 즉 지역별로 필요한 데이터만 있다면, 사람의 가정에 기반한 시나리오를 설정하고 통계학적 모델을 적용하는 등의 복잡한 과정 없이도 그 지역의 인구 예측을 위한 모델을 개발하고 적용할 수 있다는 가능성을 확인할 수 있었다. 따라서 이번 연구 대상인 부산광역시 등 광역지방자치단체뿐만 아니라 규모가 작은 시군구 단위의 기초지방자치단체의 인구수도 데이터만 있다면 이해관계에 따른 편향성 없이 객관적이고 효과적으로 해당 지역 인구를 예측할 수 있을 것으로 기대된다.

제2절 연구의 한계 및 향후과제

본 연구의 한계는 예측 목표인 주민등록 인구수가 2010년도까지는 연단위로만 집계되어, 데이터의 크기가 충분하지 않았던 점이다. 2011년부

터는 주민등록 인구수가 월 단위로 집계되고 있어, 앞으로는 더 많은 데이터를 확보할 수 있을 것이다.

또한 인구 예측 모델의 가능성을 타진하기 위한 목적으로, 인구균형방정식 요소 데이터로만 예측 모델을 구성함에 따라, 자유로운 데이터 추가가 가능한 딥러닝의 장점을 적극적으로 활용하지 못한 부분이 있다. 이에 따라 이번 연구에서 개발된 모델을 인구정책에 직접 활용하기에는 한계를 가진다. 지역의 개발 계획의 기준이 되는 계획인구의 적정성, 인구정책에 따른 인구증감의 예측 등 정책의 효과성을 선행적으로 판단하기에는 사용된 변수가 제한적이기 때문이다.

지역 인구의 경우 자연 발생적인 출생아 수나 사망자 수 외에 지역 간 이동인 사회적 유입 인구도 중요한 요소인 만큼, 향후에는 지역의 개발(주택 공급, 일자리, 법인체 수, 의료시설 수 등)과 인구 유입 정책의 시행 등 사회경제적 요소를 입력변수로 적용한다면, 인구정책이나 개발계획이 인구 증감에 미치는 영향력을 더 잘 보여줄 수 있는 모델로 발전시켜 나갈 수 있을 것이다.

참고문헌

김진형(2020), AI최강의 수업, 매일경제신문사.

오렐리앙 제롱(2020), 핸즈온 머신러닝 2판, 한빛미디어.

윤영선(2020), 딥러닝으로 걷는 시계열 예측, 비제이퍼블릭.

조태호(2020), 모두의 딥러닝 2판, 길벗.

한국포스트휴먼연구소, 한국포스트휴먼학회(2019), 인공지능의 이론과 실제 편저, 아카넷.

감사원(2012), “인구구조변화 대응실태 Ⅰ(지역)”, 감사보고서.

통계청(2019), “장래인구특별추계 (시도편) : 2017~2047년”.

김형기, 문경종(2011), “한국의 시도별 장래인구 예측”, 국토계획, 46(6), 79-99.

배성완, 유정석(2018), “머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측”, 주택연구, 26(1), 107-133.

송용호(2012), 시군별 지역인구 추계방법론에 관한 연구, 석사학위논문, 고려대학교.

우해봉(2009), “우리나라 인구추계의 정확성과 시사점”, 조사연구, 10(2), 71-96.

이은주(2017), “CNN과 RNN의 기초 및 응용 연구”, 방송과 미디어, 22(1), 90-91.

전건열, 박재형, 정종원, 윤형철(2021), “GRU를 활용한 인공지능기반 구조물 시계열 응답 예측”, 한국방재학회논문집, 21(3), 171-179

전형진(2021), 하이브리드 모형과 주식 보조지표를 활용한 주가 예측 - 슬라이딩 윈도우 학습방법 활용, 석사학위논문, 한국외국어대학교.

조대현, 이상일(2011), “이지역 코호트-요인법을 이용한 부산광역시 장래인구 추계”, 대한지리학회지, 46(2), 212-232.

지미경(2019), 시계열 분석과 머신러닝을 이용한 양파 소비예측 모델링 - 정형, 비정형 데이터를 활용하여, 석사학위논문, 충북대학교.

경향신문, 2021년 10월 18일자, 2030년 한국 인구는 5535만?…지자체, 인구 감소 예상에도 계획인구 '뻥튀기'.

O. Folorunso1, A. T. Akinwale1, O. E. Asiribo2 and T. A. Adeyemo1(2010), "Population prediction using artificial neural network", *African Journal of Mathematics and Computer Science Research* 3(8), 155- 162.

R.J.Frank, N.Davey and S.P.Hunt(2001), Time Series Prediction and Neural Networks, Department of Computer Science, University of Hertfordshire, Hatfield, UK.

Abstract

A study on deep learning model for predicting the region population.

- Based on the population of Busan -

Park, Sung Yong
Seoul School of Integrated Sciences and Technologies

The Cohort-component method used to projections the future population of the current region is characterized by the uncertainty that it is based on a scenario prepared on the premise of future assumptions and the complexity of the projections that must consider free movement between regions by region.

Accordingly, when establishing regional development plans, the future population is adjusted in an advantageous manner according to stakeholders, such as increasing the planned population by maximizing the assumptions about interregional movement factors. However, the actual population does not meet the planned population due to the trend of population decline in the country as a whole, and there is a problem with the excessive forecasting of the future population.

Therefore, this study examined whether it is possible to predict the

population of a specific region by excluding artificial assumptions and statistical model selection by applying a deep learning model that can derive predictions only by learning data.

To this end, first, the theoretical background related to deep learning necessary for population prediction was examined, and prior research was analyzed. And, using the population data of Busan, Korea as an example, an optimal deep learning model was created and compared with the population projections by Statistics Korea. Five data were used as input variables from 1997 to 2020, including the number of births and deaths, the number of people moving out and moving-in in the Busan area, and the number of resident registration populations in the Busan area. The data were divided into a train-set from 1997 to 2015 and a test-set from 2016 to 2020.

To create a model for population prediction, MLP, which is the basic structure of an artificial neural network, and RNN, LSTM, and GRU suitable for time series data were applied. According to the training results by configuring the number of hidden layers and the number of nodes in 10 types, the model that showed the best prediction value in LSTM was generated. The prediction population followed the decreasing population trend every year. MAPE was evaluated as 0.419.

In conclusion, although there was a limit to the short range of data used in this study, but it was confirmed that it is possible to predict the population of a specific region only by data learning by applying a deep-learning model without artificial assumptions or statistical model setting. It is expected that future data will be supplemented, input variables added, and models will be improved to more accurately predict the local population and use it as a basis for policy decisions.

Key words: deep learning, population prediction, RNN, LSTM, GRU