2021년 수원시 빅데이터플랫폼 유지관리

수원시 인구예측 분석

시계열 분석과 코호트 모형을 결합한
수원시 행정동 장래 인구 추계

수원시
SUWON CITY

mobigen
Make The Best For Smart Business

2021. 11

목 차

1 분석 개요

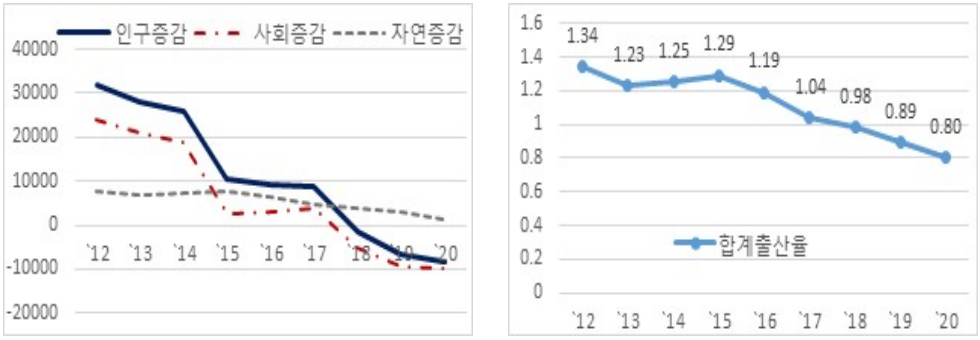
가. 분석 배경 및 문제정의

1) 수원시 인구구조 변화 상황

□ 2018년부터 인구성장률 마이너스 기록, 2020년 인구성장률은 -0.7%에 달함

□ 수원시 2020년 출산율은 0.80으로 전국(0.84), 경기도(0.88)보다 낮음

□ 65세 이상 고령인구 131,936명으로 전체 인구의 11.12%에 달함



[그림 1-1] 수원시 인구구조 변화 그래프

2) 문제 정의

□ 65세 이상 고령인구 급증과 맞물린 초저출산현상으로 지역소멸 위험 현실화

□ 30대 이하 연령층과 유소년 인구 급감으로 생산가능인구 지속적 감소 발생

□ 급변하는 인구구조가 야기한 성장능력 저하로 노인빈곤율 증가와 청·장년층 조세부담 증가 예상

□ 인구구조 변화 대응을 위한 수원형 인구정책 수립 필요성 대두

3) 분석 목표

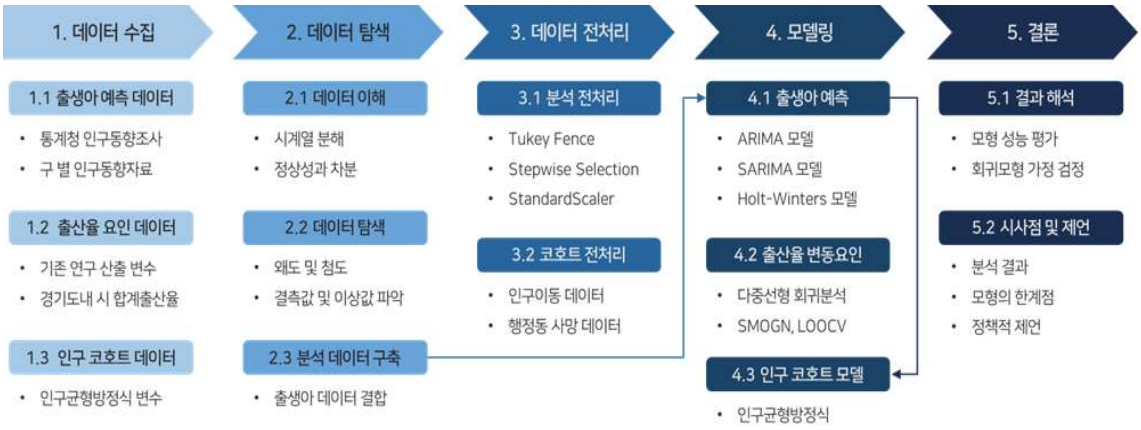
□ 인구정책 개별사업의 통합 · 체계화

□ 저출생 · 고령사회 대응체계 구축과 정책지원 활성화

□ 인구구조 변화에 대한 지역사회 인식개선 확산

나. 분석 프로세스 제시

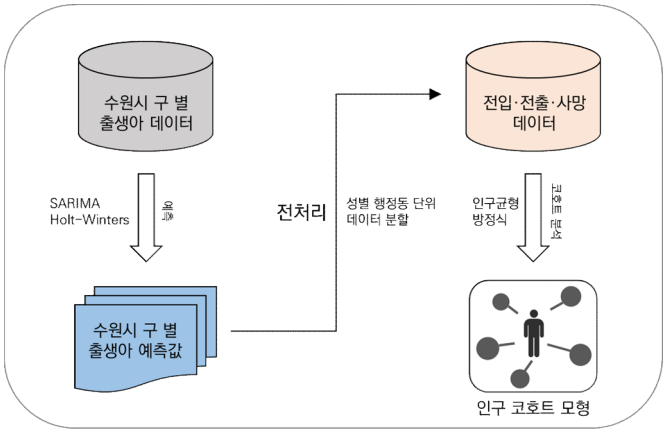
1) 수원시 인구예측 분석 프로세스



[그림 1-2] 인구예측 분석 프로세스

수원시 인구예측 분석은 시계열 분석 기법을 이용한 출생아 예측, 회귀분석을 이용한 출산율 변동 요인 분석, 코호트 요인법을 사용한 인구 코호트 모형으로 총 3가지 모델로 구성되어 있다.

2) 코호트 모형 구축



[그림 1-3] 인구 코호트 모형 프로세스

시계열 모형인 SARIMA, Holt-winters 지수평활법으로 수원시와 4개구 출생아 예측을 수행하였다. 이후 4개 구 출생아 예측값을 인구 코호트 모형에 사용할 수 있도록 전처리하여 성별 · 행정동 데이터로 분할하였다. 성별 · 행정동 출생아 데이터를 출생아 인구에 적용하여 2021~2025년의 수원시 행정동 단위 코호트 0~4세 인구를 추계해 모형을 구축하였다.

2 출생아 예측

가. 활용 데이터

데이터 명	설명	형태	출처
인구동향조사	시도/시/군/구 출생아 데이터	정형 데이터	통계청
장안구 인구동향 자료	장안구 행정동 출생신고 데이터	정형 데이터	장안구
권선구 인구동향 자료	권선구 행정동 출생신고 데이터	정형 데이터	권선구
팔달구 인구동향 자료	팔달구 행정동 출생신고 데이터	정형 데이터	팔달구
영통구 인구동향 자료	영통구 행정동 출생신고 데이터	정형 데이터	영통구

[표 2-1] 활용 데이터 목록

나. 데이터 이해와 탐색

1) 데이터 이해

□ 통계청 데이터와 수원시 구 별 데이터 결합

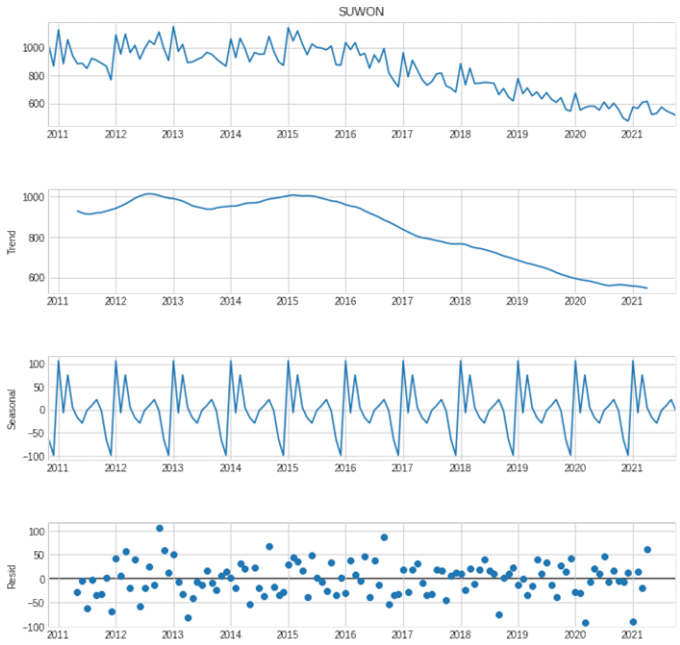
통계청 인구동향조사 데이터는 공표 주기가 연간이기 때문에 출생아 수 최신 동향을 파악하는데 어려움이 있다. 수원시 구 별 인구동향 자료의 경우 월 기준의 최신 데이터를 갖고 있지만 과거 데이터가 2018년부터 시작하는 한계점이 있다. 따라서 두 데이터를 결합하여 과거 추세와 최신 동향을 포함하는 데이터를 구축하였다.

□ 오차 측정

통계청과 구 별 출생아 데이터를 비교한 결과 두 데이터의 평균 오차율이 2019년 2.68, 2020년 0.66으로 큰 차이를 보이지 않는다는 것을 확인하였다. 2021년 1월부터 10월까지의 구 별 출생아 데이터를 통계청 데이터와 결합하여 최종적으로 2010년부터 약 11년의 월 단위 시계열 데이터 셋을 생성하였다.

2) 데이터 파악

□ 시계열 분해(Time Series Decomposition)



[그림 2-1] 수원시 출생아 데이터 시계열 분해

수원시 출생아 데이터의 경우 각 구성요인이 서로 독립적이라 가정하였고, 가법(additive)모형의 시계열 분해를 적용하여 추세(Trend) · 계절성(Seasonality) · 잔차(Residual)로 분해하였다. 그 결과 수원시 출생아는 감소 추세가 나타나며, 반복적인 계절성 패턴을 보이는 것을 확인할 수 있었다.

추세란 데이터가 장기적으로 변화해 가는 큰 흐름을 나타내며, 계절성은 일정한 주기를 갖고 반복적으로 같은 패턴으로 변화하는 형태를 나타낸다. 잔차는 불규칙 요인으로 추세·계절성·순환성이 제거된 후 남아있는 영향을 의미한다.

□ 정상성(Stationarity)과 차분(Differencing)

정상성이란 추세가 관측되지 않고, 변동이 시간의 흐름에 따라 일정한 시계열을 의미한다. 현재의 값을 기반으로 미래를 예측하는 시계열 분석에서 예측 효과를 증대시키기 위해서는 데이터가 정상성을 갖도록 변환해야 할 필요성이 있다.

수원시 출생아 데이터는 추세와 계절성이 나타나기 때문에 정상성을 가질 수 있도록 차분을 통해 비정상 요소의 제거를 수행하였음. 차분은 연속된 관측치의 차이를 계산하는 것이며, 이전 시점과 현재 시점의 차이를 의미한다.

3) 데이터 탐색

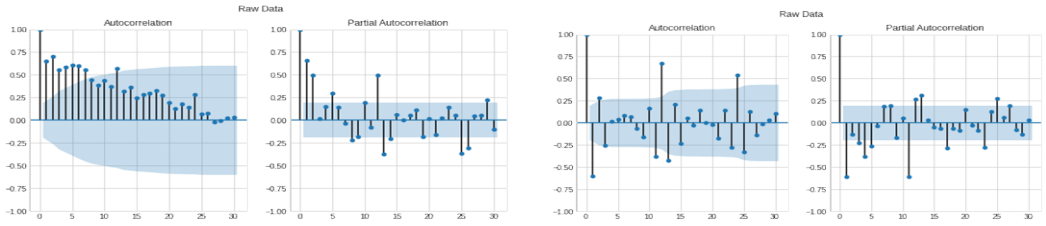
□ ADF (Augmented Dickey-Fuller) 검정
& KPSS (Kwiatkowski-Phillips-Schmidt-Shin) 검정

ADF와 KPSS 검정은 시계열이 정상 상태인지 확인하는 검정이며 ADF의 귀무가설은 시계열이 안정적이지 않음, 대립가설은 정상성을 가짐을 의미한다. KPSS 검정의 귀무가설과 대립가설은 ADF와 정반대로 해석됨에 유의한다.

ADF 검정을 시행한 결과 차분 전 p-value는 0.977에서 차분 후 0.081로 감소하였고, 이는 차분 후 시계열이 10% 유의수준에서 귀무가설을 기각하고 대립가설을 채택하여 단위근(unit root)이 없는 정상 상태임을 나타낸다.

KPSS 검정을 시행한 결과 차분 전 p-value는 0.01에서 차분 후 0.1로 증가하였고, 이는 차분 후 시계열이 5% 유의수준에서 귀무가설을 기각하지 못하고 채택하여 마찬가지로 단위근이 없는 정상 상태임을 나타낸다.

□ 자기상관함수 (ACF, Auto-Correlation Function)
& 편자기상관함수 (PACF, Partial Auto-Correlation Function)



[그림 2-2] 차분 전·후 ACF, PACF 그래프

자기상관함수(ACF)는 시계열의 시차에 따른 자기상관을 의미하며, 시차가 커질수록 ACF는 0에 가까워진다. 정상 시계열은 급격하게 0에 수렴하며, 비정상 시계열은 0을 향해 천천히 감소한다. 편자기상관함수(PACF)는 시간의 효과를 제거한 시차간의 부분상관계수이며, 시차가 다른 두 시계열 데이터 간의 순수한 상호연관성을 나타낸다.

두 함수를 시각화한 그래프를 이용하여 시계열의 정상성을 확인함과 동시에 ARIMA 모델의 최적 p, q 파라미터를 확인할 수 있다. 차분 전·후 그래프를 비교하면 ACF가 1차 차분 후에 급격히 0으로 수렴하는 것을 볼 수 있으나, 최적 파라미터를 판단하는 lag을 정하기에는 어려운 부분이 존재하였다.

다. 분석 모델링

1) ARIMA (Auto-regressive Integrated Moving Average)

□ ARIMA(p,d,q)

자기회귀누적이동평균(ARIMA) 모형은 p, d, q의 파라미터를 갖고 있다. d는 차분 횟수를 의미하며 데이터 탐색 과정을 통해 1차 차분의 유효함을 파악했기 때문에 d는 1로 설정하였다.

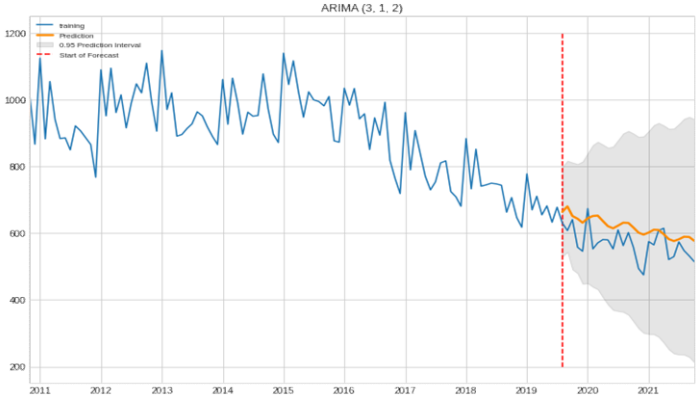
p는 AR(Auto Regressive) 모형의 lag을, q는 MA(Moving Average) 모형의 lag을 의미한다. p와 q의 경우에 ACF, PACF 그래프를 통해 명확히 파악하지 못했기 때문에 그리드 서치 기반의 auto.arima() 함수를 이용하여 AIC(Akaike Information Criterion) 값이 가장 작은 최적 파라미터를 탐색을 시도하였다.

□ ARIMA 모형 적합

shuffle을 허용하지 않고 8:2의 비율로 학습, 테스트 데이터 셋을 구분하였다. 학습 데이터에 대하여 auto.arima()를 적용한 결과 ARIMA(3,1,2) 모형이 AIC 값 1192.474를 기록하여 최적의 파라미터로 도출되었다.

□ 분석 결과

ARIMA(3,1,2) 모형을 적합시켰고, 예측 결과를 수원시 출생아 실제 값과 함께 아래의 그래프로 표현하였다.



[그림 2-3] ARIMA(3,1,2) 예측 결과 그래프

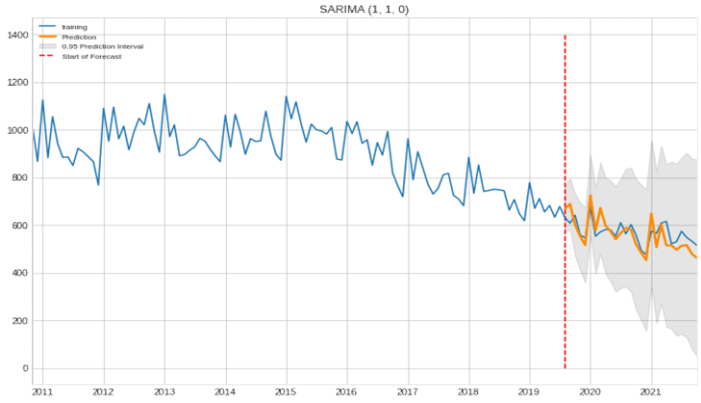
그래프의 빨간 점선은 예측이 시작된 시점을 의미하고, 파란 실선은 실제 수원시 출생아 수 그래프, 노란 실선은 ARIMA 예측 결과를 의미한다. 예측을 수행한 결과, 추세는 일정 부분 반영이 되었으나 예측의 정확도를 높이기 위해서는 계절성을 포함한 모델이 필요하다는 것을 파악하였다.

2) SARIMA 모델

□ SARIMA 모형 적합

auto.arima에 계절성을 반영하고 주기(m)를 12로 지정하여 새롭게 파라미터를 탐색하였다. 그 결과 SARIMA(1,1,0) 모형이 AIC 값 996.884를 기록하며 계절성이 반영되지 않은 기존 ARIMA 모형에 비해 더 낮게(좋게) 표출되었다.

□ 분석 결과



[그림 2-4] SARIMA(1,1,0) 예측 결과 그래프

계절성을 반영한 SARIMA 모델 결과, 예측값을 의미하는 노란색 실선과 95%의 신뢰수준에서 상한과 하한을 나타내는 회색 영역이 계절성이 없는 ARIMA 모델보다 실제 데이터를 더 잘 반영함을 확인할 수 있다.

3) Holt-Winters 지수평활 모델

□ 지수평활법 (Exponential Smoothing)

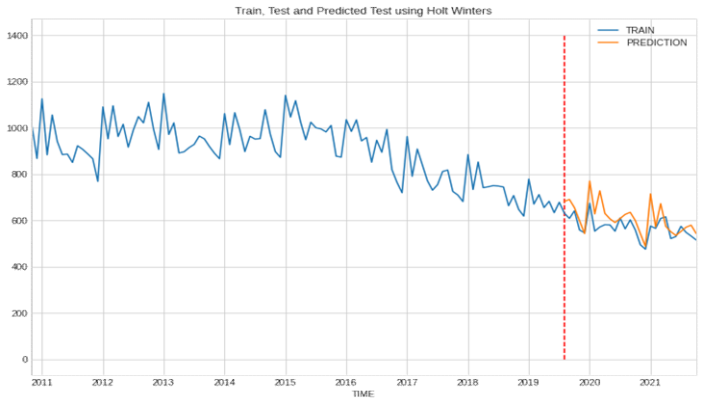
수원시 출생아 수의 추세를 확인하면, 급격한 감소 추세를 보인 2015~2020년과 달리 2021년부터는 비교적 완만한 형태를 보이고 있다. 이 경우 ARIMA, SARIMA 모델보다 최근 시계열에 더 많은 가중치를 부여하는 지수평활 모델의 예측 결과가 더 유효할 가능성이 있다.

□ Holt-Winters 모형 적합

기본적으로 지수평활법은 현재 값과 이전 예측치의 합산으로 다음 예측치를 계산하는데, 추세와 계절성이 확인된 출생아 데이터에 대해서는 적용하기 어려운 부분이 있다. 따라서 수준(level)과 추세, 계절성을 반영할 수 있는 Holt-Winters 지수평활 모델을 사용하였고, SARIMA와 마찬가지로 계절성의

주기(m)를 12로 설정하였다. 이전의 시계열 분해 과정을 통해 시계열 데이터가 가법 모형임을 확인했기 때문에 Holt-Winters 지수평활의 추세와 계절성에도 동일하게 가법 모형을 적용하여 예측을 수행하였다.

□ 분석 결과



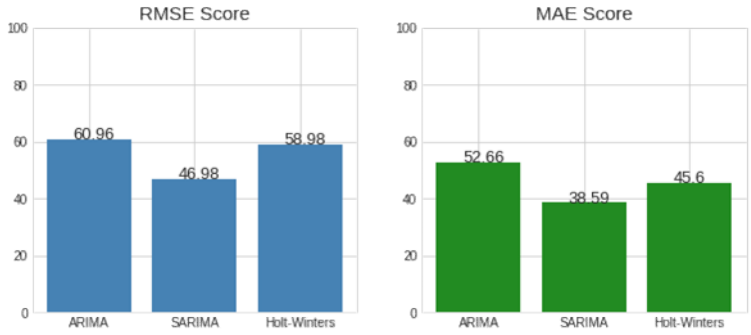
[그림 2-5] Holt-Winters 예측 결과 그래프

감소 추세가 완만해지는 최신 데이터에 더 많은 가중치를 반영한 Holt-Winters 모델 측성에 따라 출생아 수의 예측 결과가 SARIMA 모델보다 비교적 완만한 감소세를 나타낸다. 최종 모델 선정을 위한 모델 간 명확한 비교는 그래프가 아닌 RMSE와 MAE 평가지표를 사용해 진행하였다.

라. 분석 결과

1) 수원시 출생아 예측

□ 최종 예측모델 선정

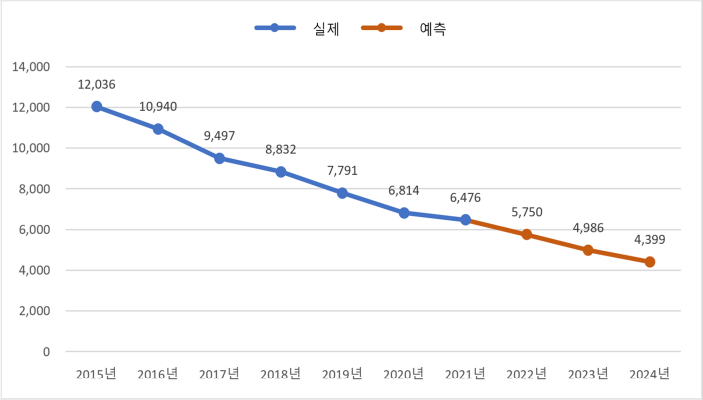


[그림 2-6] 예측모델 별 성능평가 점수

RMSE(Root Mean Squared Error)와 MAE(Mean Absolute Error) 점수를 이용하여 ARIMA, SARIMA, Holt-Winters 지수평활 모델의 성능을 평가했다.

최근 시계열의 추세에 가중치를 부여한 Holt-Winters 모형보다 ARIMA 모델에 계절성을 부여한 SARIMA 모델이 RMSE와 MAE 평가 결과에서 더 우수한 모델로 나타났다.

□ 결과 해석



[그림 2-7] 연도별 수원시 출생아 추이

모델 성능평가 점수에 따라 최종 모델을 SARIMA로 선정하였고, 2021년 11월부터 2024년 12월까지 약 3년의 출생아 수를 예측하였다. 그 결과 2022년 5,750명, 2023년 4,986명, 2024년 4,399명으로 지속적인 출생아 감소 추세를 보였다.

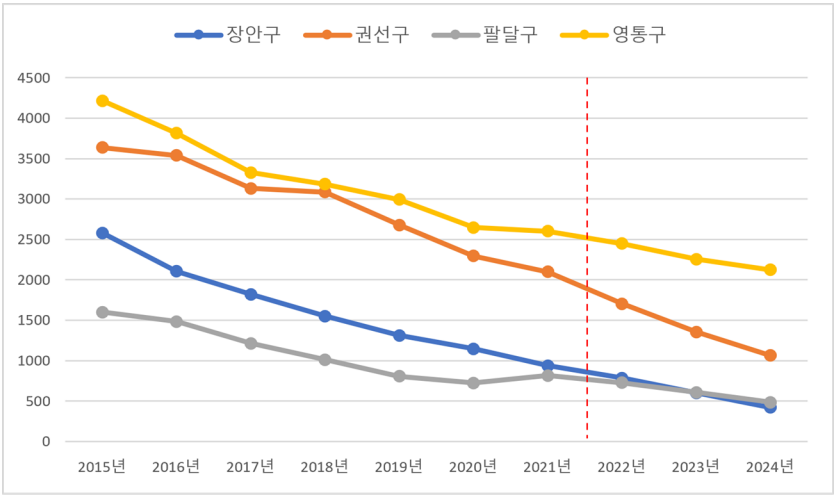
2) 수원시 구 별 출생아 예측

□ 최종 예측모델 선정

인구 코호트 모형에 필요한 2021년~2025년의 0~4세 인구를 예측하기 위해 수원시 출생아와 동일하게 구 별 출생아 예측을 수행하였다. 같은 출생아 수 데이터지만 각 구별로 추세가 다르게 나타나기 때문에 모델 성능평가 점수를 이용하여 최종 예측 모델을 각각 선정하였다.

RMSE와 MAE 점수에 따르면 장안구와 팔달구의 경우 Holt-Winters 모델이 SARIMA 모델보다 우수하게 평가되어 Holt-Winters 지수평활을 이용해 출생아 수를 예측하였다. 권선구와 영통구의 경우 SARIMA 모델이 Holt-Winters 모델보다 우수하게 평가되어 SARIMA 모델을 이용해 출생아 수를 예측하였다.

□ 결과 해석



[그림 2-8] 연도별 수원시 4개 구 출생아 추이

가장 높은 출생아 감소율을 보일 것으로 예측되는 구는 장안구로 나타났다. 장안구의 경우 출생아가 2015년 2,582명에서 2025년에 421명으로 83% 감소할 것으로 예측되었으며, 집계 시작 기간을 2020년 출생아 1,148명으로 변경해도 4개 구 중 가장 높은 수치인 약 63% 감소할 것으로 예측되었다.

가장 적은 출생아 감소율을 보인 구는 영통구이며, 2015년 4,215명에서 2025년 2,123명으로 약 49% 감소할 것으로 예측되었다. 마찬가지로 집계 시작 기간을 2020년 출생아 2,647명으로 변경한다고 해도 수원시 4개 구 중 가장 낮은 수치인 약 19% 감소할 것으로 예측되었다.

□ 시사점 및 제언

수원시와 각 구의 출생아 수 예측 결과, 대상 지역과 예측 기간에 관계없이 전부 감소 추세를 보이고 있음을 알 수 있다. 그러나 최근 시점의 감소 추세가 완만해지는 경향이 보이기 때문에, 모형의 예측 정확도 향상을 위해서는 투입 데이터인 출생아 수의 주기적인 갱신이 필요하다.

해당 모델은 월 별 출생아 데이터가 갱신되는 대로 추가할 수 있으며, 그에 따라 시계열이 변화하면서 각 모델의 성능평가 점수가 지금과 다를 수 있다. 따라서 최종 모델 선정을 상황에 따라 조정해야할 필요성이 있으며, 이를 위해 성능 점수에 따른 모델 선정 프로세스를 자동화하여 모델의 효용을 증대시켰다.

3 출산율 변동요인

가. 활용 데이터

데이터 명	설명	형태	출처
고령인구비율	전체 인구 대비 65세 이상 인구	정형 데이터	통계청
조혼인율	당해 언앙인구 대비 총 혼인 수	정형 데이터	통계청
지가변동률	연간 기준일 사이 해당 지역의 지가변화	정형 데이터	한국감정원
총 고용률	15-64세 생산가능인구 중 취업자 비율	정형 데이터	통계청
가임기 여성 고용률	15-49세 가임 여성 중 취업자 비율	정형 데이터	통계청
스트레스 인지율	평소 일상생활 중 스트레스를 “대단히 많이” 또는 “많이” 느끼는 사람의 비율	정형 데이터	통계청
외국인 비율	인구 천 명당 외국인 수	정형 데이터	통계청
초등학교 수	국공립/사립 초등학교 수	정형 데이터	통계청
유치원 수	국공립/사립 유치원 수	정형 데이터	통계청
재정자립도	자치단체 예산규모 중 해당 지역 자체 수입 비중	정형 데이터	통계청
합계출산율	15-49세의 가임 여성 1명이 평생 동안 낳을 것으로 예상되는 평균 출생아 수	정형 데이터	통계청

[표 3-1] 활용 데이터 목록

나. 데이터 이해와 탐색

1) 데이터 이해

□ 독립 변수 (Independent Variable)

`경기도 저출산 원인분석 및 출산동향예측` 연구를 바탕으로 고령인구비율을 포함한 총 10개의 변수 데이터를 수집하였다. 기존 연구에서는 약 20개의 변수가 사용되었으나, 그 중 유효한 것으로 판단된 10개의 변수에 대해서만 분석을 위하여 사용하였다.

□ 종속 변수 (Dependent Variable)

출생아 수 자체를 종속 변수로 사용하기에는 경기도 도시 간의 인구 차이가 크기 때문에 예측에 어려움이 따른다. 따라서 15~49세 가임여성 1명이 평생 동안 낳을 것으로 예상되는 평균 출생아 수인 합계출산율을 종속 변수로 이용하여 분석을 진행하였다.

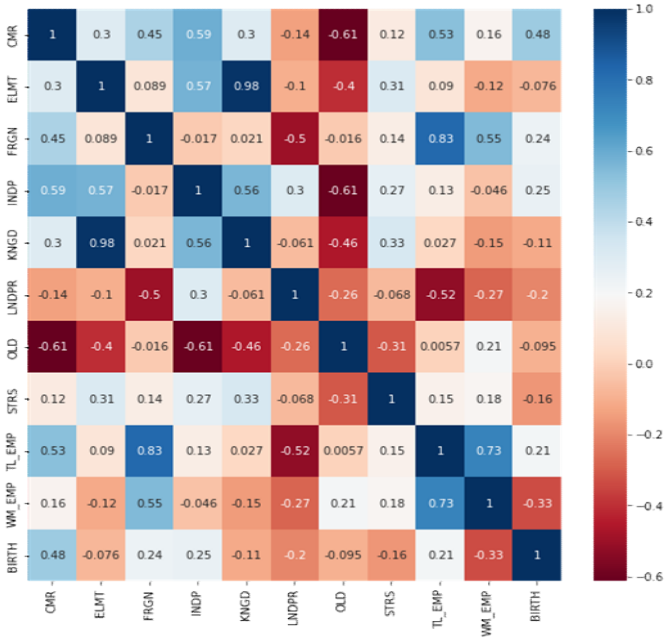
2) 데이터 탐색

□ 왜도(Skewness) 및 첨도(Kurtosis)

왜도는 분포가 정규분포에 비해서 얼마나 비대칭인지를 나타내는 통계량이다. 정규분포는 왜도 0을 갖으며, 절댓값 3을 넘지 않아야 왜도가 크지 않다고 판단할 수 있다. 첨도는 데이터 분포의 뾰족한 정도를 나타내는 통계량이다. 정규분포의 첨도는 0이며 집단이 이질적일수록 첨도가 낮고, 동질적일수록 첨도가 높아진다.

변수들이 정규 분포 형태일 때 더 신뢰할 수 있는 예측 결과를 얻을 수 있기 때문에 출산율 변동요인 변수에 대하여 왜도와 첨도를 파악하였다. 그 결과 왜도는 절댓값 기준 최대 1.01 첨도는 1.34를 나타내었고, 따라서 데이터 분포에 큰 이상이 없다고 판단하였다.

□ 변수 상관관계 파악



[그림 3-1] 변수 간 상관계수 히트맵

상관관계는 두 변수 간의 선형 관계에 초점을 맞추어 변수간의 밀접한 정도를 나타낸다. 변수 간 상관관계가 클수록 모델의 예측 정확도가 떨어지기 때문에 상관관계의 크기를 나타내는 상관계수(Correlation Coefficient) 파악 과정이 반드시 필요하다.

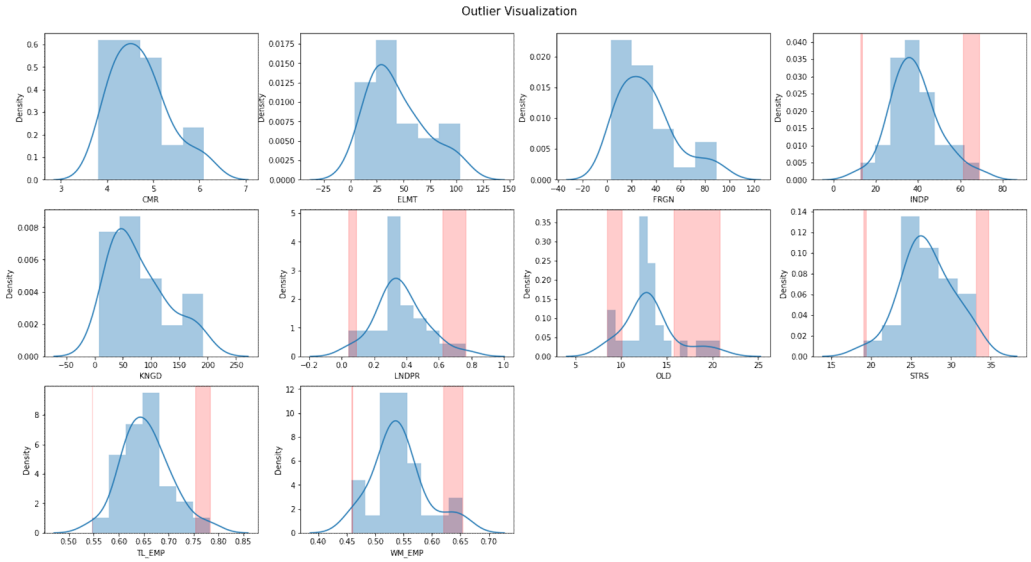
상관계수가 0일 때 변수가 서로 선형 관계가 없고, -1과 1에 가까울수록 각각 서로 음과 양의 상관관계가 있다고 볼 수 있다. 히트맵 결과, 강한 선형관계를 나타내는 기준인 상관계수 절댓값이 0.7이 넘는 변수가 검출되었기 때문에 추후 변수 선택(Feature Selection) 과정이 필요하다.

□ 결측값 및 이상값 파악

모든 변수에서 결측값은 존재하지 않았으며, 이상값은 6개 변수에서 나타났다. 이상값 검출에는 사분위수 범위(IQR, Interquartile Range)를 기반으로 하는 Tukey Fences 방식을 이용하였고, 상·하한 허용 범위를 넘어가는 값들에 대해 이상값으로 판별하였다.

3) 데이터 전처리

□ 이상값 처리

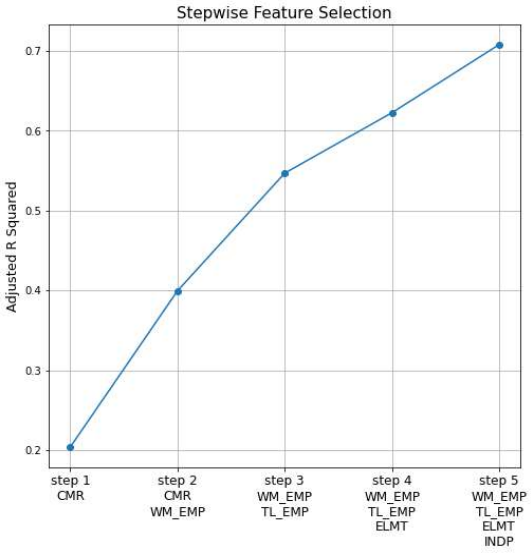


[그림 3-2] 변수 별 이상값 분포 그래프

세 번째 사분위(Q3)에서 첫 번째 사분위(Q1)을 뺀 값인 IQR을 이용해 상한값을 Q3+(1.5*IQR)로, 하한값을 Q1-(1.5*IQR)로 설정하였고, 해당 범위를 벗어나는 값들에 대해 상·하한 값으로 대체하는 것으로 이상값 처리를 수행하였다.

□ 변수 선택

	VIF	Features
0	229.722521	CMR
1	268.854691	ELMT
2	11.846034	FRGN
3	57.948136	INDP
4	241.238954	KNGD
5	14.890493	LNDPR
6	150.599531	OLD
7	88.278219	STRS
8	1118.762144	TL_EMP
9	475.888352	WM_EMP



[그림 3-3] 변수 별 VIF와 단계적 선택법

데이터 탐색 과정에서 변수 간 상관관계가 존재한다는 것이 확인되었기 때문에 다중공선성(Multicollinearity)의 발생 위험이 있다. 다중 회귀 모델에서 독립변수가 다중공선성 문제가 있는지 판별하는 척도인 분산팽창인자(VIF)을 구한 결과 기준값 10을 초과하는 변수들이 다수 발견되었다. 따라서 변수 선택 과정을 통해 다중공선성 문제를 해결하고자 한다.

모든 변수가 포함된 모델에서 출발하여 가장 도움이 되지 않는 변수를 삭제하고, 모델에서 빠져 있는 변수 중에서 모델을 가장 개선시키는 변수를 추가하는 방식을 반복하여 최적의 변수만 가져오는 단계적 선택법(Stepwise Selection)을 이용하여 변수 선택을 진행하였다.

p-value를 0.05로 지정하고 단계적 선택법을 수행한 결과 가임기 여성 고용률, 총 고용률, 초등학교 수, 재정 자립도 4가지 변수가 조정된 결정계수 값 0.7을 기록하면서 최종 변수로 선택되었다.

□ 변수 스케일링(Feature Scaling)

스케일링은 변수 간 비교가 가능하도록 각 변수들의 범위를 동일하게 만들어 주는 방법이다. 이번 분석에서는 변수들을 평균이 0, 분산이 1인 가우시안 정규분포를 가진 값으로 변환하는 표준화(Standardization) 방식을 사용하여 모든 변수에 대해 스케일링을 수행하였다.

다. 분석 모델링

1) 다중선형 회귀분석

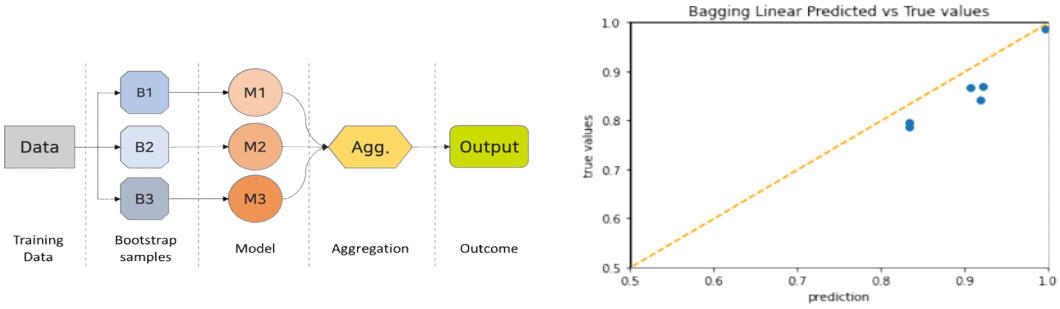
□ 교차 검증 (Cross Validation)

경기도 내 군을 제외한 28개의 시 데이터에 대해 교차검증을 위하여 8:2 비율로 학습, 테스트 데이터를 분리하였다. 출산율 영향요인 분석을 위한 데이터가 검증(Validation) 데이터 셋을 추가로 분할하는 Hold-out 교차 검증을 사용할 만큼 충분히 크지 않기 때문에 따로 구분하지 않았으며, 이 경우에는 과적합(Overfitting)의 위험성이 존재한다.

□ 다중 선형 회귀

다중 선형 회귀분석은 2개 이상의 독립 변수로 1개의 종속 변수를 예측하는 회귀 모형이다. 과적합 문제를 피할 수 있도록 L1 규제를 적용한 Lasso 회귀와 L2 규제를 적용한 Ridge 회귀를 일반적인 선형 회귀와 함께 수행하였다.

□ Bagging(Bootstrap Aggregating)



[그림 3-4] Bagging과 Bagging Regression 결과

앙상블의 일종인 배깅(Bagging)은 학습데이터에서 부트스트랩 샘플링을 이용해 여러 번 샘플을 추출하고 개별 모델을 학습시킨 후 결과를 집계하여 평균을 내서 최종 예측을 수행한다. 현재 학습 데이터의 수가 부족하여 회귀모델의 일반화 성능이 떨어질 수 있다고 판단했기 때문에 추가적으로 Bagging 회귀를 사용하여 예측을 수행하였다.

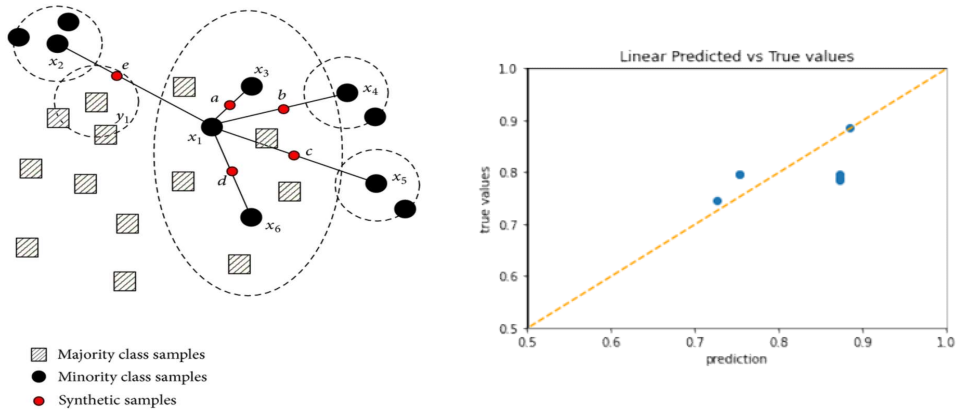
□ 분석 결과

선형, Ridge, Lasso, Bagging 회귀 모형으로 각각 예측한 후 테스트 데이터를 이용해 검증을 수행하였다. 검증 결과 성능 평가 점수로 사용한 RMSE와 MAE, 결정계수 모두 일반적인 선형 회귀를 가장 정확도가 높은 모델로 판단하였다.

그러나 학습 데이터가 부족한 상태에서 단순 선형회귀 모형은 여전히 과적합 위험과 일반화 성능에서의 약점을 갖고 있기 때문에 이를 개선하기 위해 데이터 자체에 대한 오버샘플링(Oversampling)을 시도하였다.

2) SMOGN (Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise)

□ SMOTE-GN



[그림 3-5] SMOTE 적용법, SMOGN 회귀분석 결과

SMOTE는 데이터의 개수가 비교적 적은 클래스의 표본을 가져온 뒤 임의의 값을 추가하여 새로운 샘플을 만드는 오버샘플링 방식이다. 단순 오버샘플링의 경우 과적합의 위험을 증대시킬 수 있기 때문에 최근접이웃(KNN)을 이용하여 소수 데이터에 대한 보정을 수행하는 SMOTE 기반의 SMOGN 오버샘플링을 수행하였다.

SMOTE는 기본적으로 분류 문제에 대해 작동하도록 구현되어있기 때문에 가우시안 잡음(Gaussian Noise)을 이용해 회귀 문제에 대해 적용할 수 있도록 변형한 SMOGN을 사용하였다.

□ 하이퍼파라미터(Hyperparameter)

SMOGN의 하이퍼파라미터 k는 고려할 이웃의 수를 의미하며, 표본 크기(n)보다 작은 양의 정수를 지정해야 한다. 오버·언더샘플링의 경계를 지정하는 희소성 임계점(Threshold of Rarity)을 지정하고, 데이터의 극단값과 희귀값을 지정하는 하이퍼파라미터인 박스플롯 계수(Coefficient)를 지정해 SMOGN의 샘플링 옵션을 설정할 수 있다.

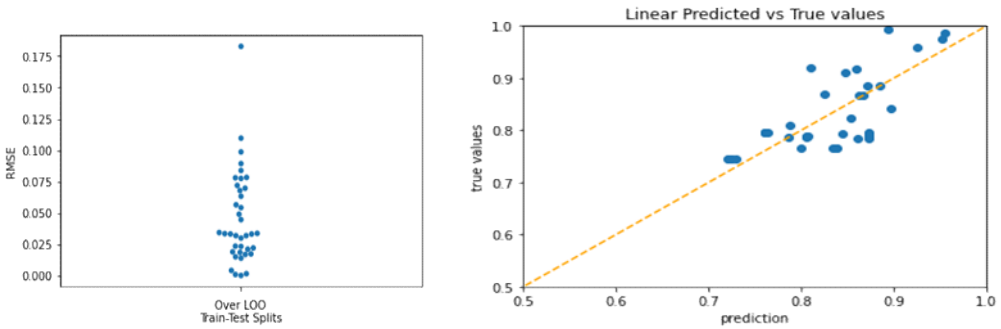
□ 분석 결과

k는 반복 시행을 통해 확인한 적절한 값 k=4를 지정하였고, 희소성 임계점은 기본값으로 설정된 0.8로 지정하였다. 계수의 경우 적용 가능한 최댓값을 지정하여 28개의 전체 데이터에 대해 SMOGN 함수를 적용하였다. 그 결과 38개의 오버샘플링 데이터를 얻을 수 있었다.

3) LOOCV(Leave-One-Out Cross Validation)

□ LOOCV 교차검증

학습, 테스트 데이터에 검증(Validation) 데이터를 추가로 분할하는 Hold-out 교차검증 방식은 과적합 문제를 피할 수 있는 방법이지만, 데이터 전체 개수가 크지 않다면 분할된 데이터가 전체를 대표하지 못하는 문제가 발생할 수 있다.



[그림 3-6] Loocv 활용 회귀분석 결과

LOOCV 방식의 경우 전체의 데이터에 대해 1개의 테스트 데이터와 n-1의 학습 데이터로 분할한다. 이 과정을 n번 진행하는 것으로 검증을 수행하고, 전체 수행 결과에 대해 평균을 계산하는 교차검증 방식이다. Hold-out 교차검증과 달리 LOOCV는 전체 데이터가 반드시 한 번씩은 테스트 데이터에 포함되기 때문에 무작위성(Randomness)이 존재하지 않고 안정적인 검증이 가능하다는 장점이 있다.

□ 모형 적합

SMOGN으로 오버샘플링한 38개의 전체 데이터에 대해 1개의 테스트 데이터와 37개의 학습데이터로 분할 후 반복 검증하는 것으로 LOOCV를 사용한 선형 회귀모형을 적합하였다. 그 결과 데이터 분할 시 시드로 지정한 random_state 값에 따라 예측 성능이 크게 변하던 불안정성 문제를 해결할 수 있었다.

라. 분석 결과

1) 모형 해석

□ 최종 예측모델 선정

[그림 3-7] Loocv 활용 회귀분석 결과

RMSE와 MAE 점수가 각각 0.059, 0.046을 기록하면서 처음 수행한 선형 회귀분석보다 SMOGN과 LOOCV를 적용한 선형 회귀분석이 오차가 적다는 것을 확인하였다. 또한 결정계수 0.458로 모델의 설명력이 이전보다 높아진 것을 확인하여 해당 모델을 최종 예측 모형으로 선정하였다.

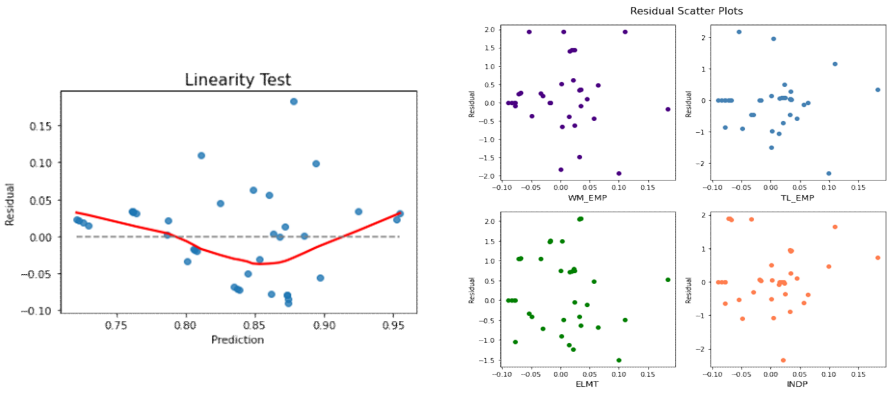
2) 모형 검정

□ 선형 회귀분석 4가지 가정

선형 회귀분석을 이용한 예측은 모형이 선형성 · 독립성 · 등분산성 · 정규성을 띈다는 4가지 가정을 만족해야한다.

□ 선형성 검정

예측값과 잔차를 비교 하는 것으로 선형성 검정을 수행하였다. 모형이 선형임을 나타내려면 잔차의 추세를 나타내는 빨간 실선이 예측값을 의미하는 가운데 점선에서 크게 벗어나지 않은 상태여야 한다. 그래프 상 잔차의 추세가 직선을 나타내기 때문에 해당 모형은 예측값에 따라 잔차가 크게 달라지지 않고 선형성을 띈다고 파악할 수 있다.

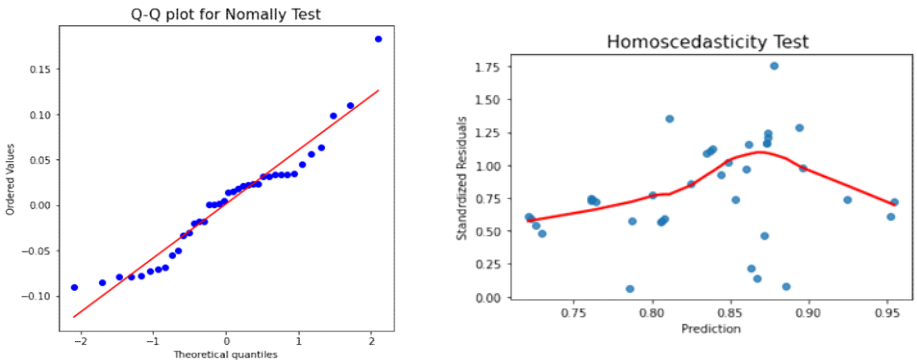


[그림 3-8] 선형성 검정 그래프

각 변수에 대해 표현한 잔차도를 보면 SMOGN 과정에서 발생한 값 외에는 각 독립변수의 잔차가 무작위로 분포되어있다는 것을 확인할 수 있다. 검정 결과를 종합한 결과 종속변수와 독립변수가 선형관계임을 의미하는 선형성 가정을 만족한다고 판단하였다.

□ 독립성 · 정규성 · 등분산성 검정

독립성 검정을 위해 데이터 탐색 과정에서 수행했던 VIF 값 확인을 다시 수행하였다. 변수 선택 과정을 거치면서 다중공선성 문제가 해결되어 4개의 독립변수가 VIF 값 10 미만으로 판별되었다. 따라서 독립변수 간에 상관관계가 없다는 독립성 가정을 만족한다고 판단하였다.



[그림 3-9] 정규성 Q-Q plot과 등분산성 검정 그래프

정규성 검정을 위해 x축은 현재 데이터 분포 상에서의 분위수(Quantile) 값을, y축은 정규분포에서 이론적 분위수 값을 나타내는 Q-Q plot 그래프를 그렸다. Q-Q plot 확인 결과 선형을 보이기 때문에 잔차가 정규분포임을 가정하는 정규성 가정을 만족한다고 할 수 있다.

등분산성 검정을 위해 예측값에 따른 잔차의 변화를 나타내는 그래프를 그렸다. 잔차의 절대값을 제곱근하여 표준화된 잔차를 구하고, 국소 회귀를 통해 잔차의 추세를 표현하였다. 추세선이 수평을 보이기 때문에 잔차가 특정한 패턴을 보이지 않고 분산이 같음을 의미하는 등분산성 가정을 만족한다고 판단하였다.

3) 시사점 및 제언

□ 회귀식 해석

회귀계수(coefficient)를 확인한 결과 가임기 여성 고용률이 -0.08, 총 고용률이 0.072, 초등학교 수가 -0.044, 재정자립도가 0.013으로 나타났다. 회귀계수는 다른 변수들이 일정할 때 해당 독립변수에 따라 종속변수가 얼마나 변동하는지를 나타낸다.

$y =$ 합계출산율, $x_1 =$ 가임기여성고용률, $x_2 =$ 총고용률, $x_3 =$ 초등학교수, $x_4 =$ 재정자립도

라고 정의했을 때, 다중회귀분석 결과를 오차를 제외한 회귀식으로 나타내면 $y = 0.870994 + -0.08x_1 + 0.072x_2 + -0.044x_3 + 0.013x_4$이며 회귀 모형은 y의 변동을 45.8% 설명할 수 있다.

□ 모형의 한계점

SMOGN과 LOOCV를 이용하여 학습 데이터의 수를 최대한 확보하였으나, 여전히 원활하게 회귀분석을 진행하기에는 데이터가 부족하였다. 초기 분석 계획대로 수원시 행정동 단위로 독립 변수들에 대한 데이터를 확보한다면, 수원시 출산율에 영향을 끼치는 변동 요인에 대해 정확도 높은 분석이 가능할 것으로 기대한다.

□ 시사점 및 제언

회귀분석 결과 수원시 2021년 예상 합계출산율은 0.768로 2020년 합계출산율 0.796보다 감소하는 것으로 확인되었다. 회귀모형을 통해 출산율 감소요인을 파악할 수 있었으며, 분석에 사용한 변수들이 1년 단위의 갱신 주기를 갖기 때문에 분석 시점에서는 사용하지 못했으나, 추후 독립변수를 2021년 데이터로 갱신한다면 2022년에 대한 출산율 예측이 가능하다.

4 인구 코호트 모형

가. 분석 배경

1) 모형 선정

□ 인구 코호트 모형

2019년 진행했던 '수원시 인구예측 빅데이터 분석'에 따르면 머신러닝 기법을 포함한 다양한 모델 별 인구 예측 결과 중 코호트 방식의 예측 정확도가 가장 높은 것으로 나타났다. 따라서 기존 연구와 통계청 장래인구추계를 바탕으로 코호트 분석을 수행하였으며, 수원시 행정동 기반 5세 단위 2021~2025년 인구 코호트 모형을 구축하였다.

2) 모델 개선

□ 출생아 예측 모형 결합

코호트의 출생아를 같은 값으로 가정한 기존 연구에서 예측 정확도를 높이기 위해 모형 개선을 수행하였다. 출생아 수 예측 모형을 결합하여 0~4세 코호트 인구에 예측 출생아 수를 적용하였다. 기존 코호트 모델과 성능 비교를 수행한 결과, 개선된 모형의 정확도가 더 높았기 때문에 최종 인구 코호트 모델로 선정하였다.

나. 활용 데이터

활용 데이터	설명	형태	출처
성연령별 인구현황	주민등록 인구 기준 2015 ~ 2020년 성연령별 행정동 인구 데이터	정형 데이터	행정안전부
성별 출생자 수	성별 2015 ~ 2020년 행정동 출생자 수 데이터	정형 데이터	통계청
성별 사망자수	읍면동별 성별 2015 ~ 2020년 사망자수 데이터	정형 데이터	통계청
성연령별 사망자수	시군구별 성연령별 2015 ~ 2020년 사망자수 데이터	정형 데이터	통계청
성연령 전입전출	수원시 2015 ~ 2020년 전입 전출 데이터	정형 데이터	통계청 MDIS

[표 4-1] 활용 데이터 목록

다. 데이터 이해와 탐색

1) 데이터 전처리

□ 인구 데이터

행정구역
경기도 수원시 영통구 영통1동(4111758000)
경기도 수원시 장안구 송죽동(411159100)
경기도 수원시 장안구 조원1동(411159700)
경기도 수원시 장안구 연무동(411160000)
경기도 수원시 권선구 세류1동(411135200)

→

구	동
영통구	영통1동
장안구	송죽동
장안구	조원1동
장안구	연무동
권선구	세류1동

[표 4-2] 인구 데이터 전처리

구, 행정동, 나이별 인구수를 예측하기 위해 인구 데이터 변수 '행정구역'에서 두 개의 파생변수 '구'와 '동'을 생성하였다.

□ 인구이동 데이터

전입자_만나이
27
91
10
3
34

→

전입자_만나이
25 - 29세
90세 이상
10 - 14세
0 - 4세
30 - 34세

[표 4-3] 인구이동 데이터 전처리

코호트 단위인 5세 단위로 인구수를 예측하기 위해 인구이동 데이터 변수 '전입자_만나이'를 구간 변수로 변형하였다.

□ 사망 인구 데이터

행정동 성·연령별 사망자 수를 구하기 위해 시군구 · 성별 · 연령별 데이터를 갖고 있는 시군구 사망자 데이터와 시군구 · 행정동 · 성별 데이터를 포함하는 읍면동 인구동태 데이터를 결합하였다.

■ 시군구 사망 데이터

시군구	성별	연령	사망_2019
장안구	남자	0 – 4세	7
장안구	남자	5 - 9세	30
장안구	남자	90세 이상	23

→

시군구	성별	전체_사망_2019
장안구	남자	359
팔달구	남자	210

[표 4-4] 시군구 사망 데이터 전처리

구의 전체 사망자 대비 특정 동의 사망자 비율을 구하기 위해 시군구 성별 사망 데이터 변수에서 구별로 모든 연령대의 사망자수를 합한 전체 사망 변수를 생성하였다.

■ 행정동, 성별 사망 비율 데이터

시군구	행정동	성별	사망_2019
장안구	파장동	남자	65
장안구	율천동	남자	43

+

시군구	성별	전체_사망_2019
장안구	남자	359
팔달구	남자	210

↓

시군구	행정동	성별	사망_2019_비율
장안구	파장동	남자	0.13
장안구	율천동	남자	0.086

[표 4-5] 행정동 사망 데이터 전처리

구의 전체 사망자에 대한 특정 동의 사망자 비율을 구하기 위해 읍면동 인구동태 데이터의 행정동별 사망자수 데이터 변수와 구별로 모든 연령대의 사망자수를 합한 전체 사망 변수를 나누어 '사망비율' 변수를 생성하였다.

■ 행정동 성·연령별 사망자 수

시군구	행정동	성별	사망_2019_비율
장안구	파장동	남자	0.13
장안구	율천동	남자	0.086

+

시군구	성별	연령	사망_2019
장안구	남자	0 - 4세	30
장안구	남자	5 - 9세	150

↓

시군구	행정동	성별	연령	사망_2019
장안구	파장동	남자	0 - 4세	3
장안구	파장동	남자	5 - 9세	12
장안구	파장동	남자	…	
장안구	파장동	남자	90세 이상	9

[표 4-6] 성·연령별 사망자 데이터 전처리

행정동, 연령별 예상 사망자수를 구하기 위해 시군구, 행정동별 사망자수 비율과 시군구, 연령별 사망자수를 곱하여 시군구, 행정동, 성별, 연령별 예상 사망자수 변수를 생성하였다.

□ 0-4세 인구 추정

출생아 예측 모델로 얻은 2021~2025년 수원시 4개 구 예측 출생아수 데이터를 사용하였다. 2020년 행정동 별 가임기 여성 인구 비율을 기반으로 4개 구 출생아 수에서 행정동 별 출생아 수를 도출하였다.

또한 최근 3개년도 출생아 남, 여 비율의 평균값을 이용하여 행정동 별 출생아 수에서 남, 여 출생아수를 구하였다. 최종적으로 5개년도의 출생아수를 합하여 코호트 모형에 사용될 행정동 단위의 성별 0~4세 인구를 도출하였다.

라. 분석 모델링

□ 코호트 분석

코호트(cohort)란 일정한 지리적 공간에 존재하는 인구 가운데 특정한 기간 동안 유의미한 생에 경험을 공유하는 사람들을 의미한다. 코호트 분석은 특정 기간 동안 공통된 특성이나 경험을 갖는 사용자 집단을 그룹화하여 시간 흐름에 따른 행동 패턴을 추적하는 분석 방법이다.

□ 인구균형 방정식

인구 변화에 직접적인 영향을 끼치는 출생, 사망, 전입, 전출 네 가지 요인을 인구변동 요인이라고 부른다. 코호트 모형을 이용한 인구 추계는 인구변동 요인별 미래 수준을 각각 예측한 후, 추계의 시작점이 되는 기준 인구에서 출생아 수와 국내 순이동자 수를 더하고, 사망자 수는 빼는 인구균형 방정식을 적용하여 다음 해 인구를 반복적으로 산출해 나가는 방법이다.

인구균형방정식 (Demographic Balancing Equation)

$PB_{05} = PC_{00} - B_{00} + D_{00} + (MI_{00} - MO_{00})$

- PB : 시군구 읍면동 남녀별 기준인구
- PC : 시군구 읍면동 남녀별 인구주택총조사 인구
- B : 출생아수
- D : 사망자수
- MI-MO : 순이동자수

□ 코호트 모형의 한계

인구균형 방정식에 사용하는 출생, 사망, 전입, 전출 외의 데이터는 모형에 포함되지 않기 때문에 코호트 모형이 사회 현상을 정확히 반영한다고 말할 수는 없다. 그러나 인구 추계 방식 중 기본적인 분석 모형이며, 과거 연구 결과 다른 모델에 비해 정확도가 가장 높게 나왔기 때문에 이번 인구 예측 분석 방법으로 코호트 모형을 선정하게 되었다.

시계열 분석 기반의 출생아 예측 모형과 코호트 모형을 결합한 이번 인구 코호트 모형과 같이 추후 연구에서 다양한 사회적 요인을 반영할 수 있는 분석 기법과 결합하여 앙상블 모형을 구축한다면 코호트 모형의 한계를 줄이고, 예측 정확도가 높은 모형을 구축할 수 있을 것이라 기대한다.

마. 분석 결과

1) 수원시 인구수 예측 검증

□ 최종 예측모델 선정

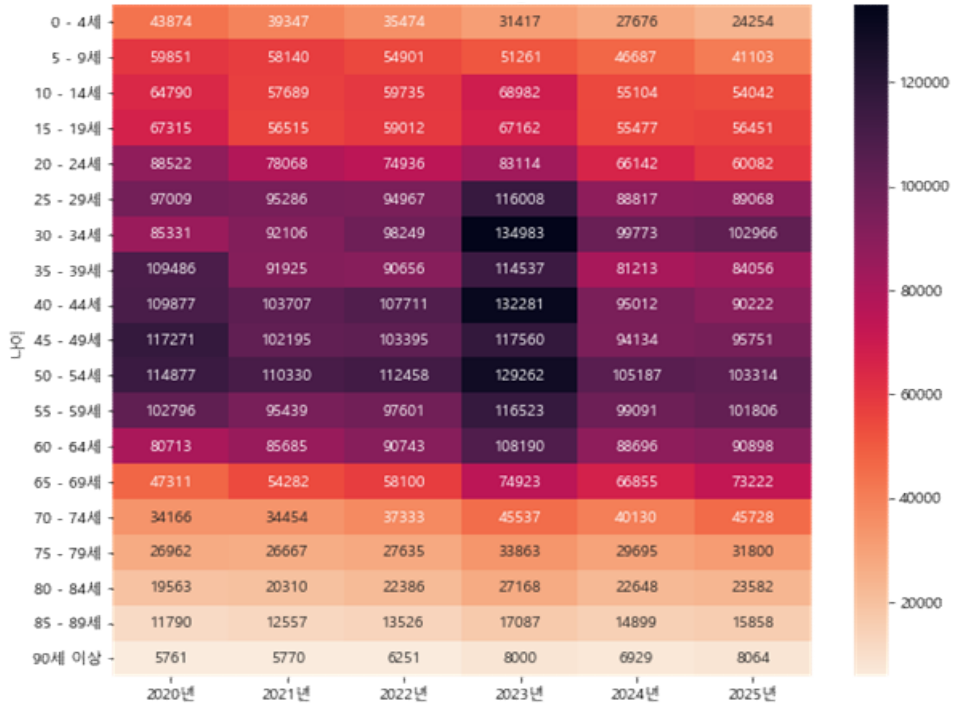
모델	연도	mape	smape	rmse	rmsle
기존 코호트	2020	32.32890	26.05991	238.44480	0.34337
	2021	27.49814	22.55524	205.90012	0.32328

⇩

모델	연도	mape	smape	rmse	rmsle
출생아 예측 코호트 모형	2020	28.20529	23.81684	220.54217	0.30617
	2021	22.89875	20.18530	183.18327	0.28102

[표 4-7] 코호트 모형 성능평가

MAPE(Mean Absolute Percentage Error), SMAPE(Symmetric Mean Absolute Error), RMSE, RMSLE(Root Mean Square Log Error)를 이용해 코호트 모델 성능을 평가했다. 2020년도 예측 값과 2020년 실제 값, 2021년도 예측 값과 분석 시점인 2021년 10월 기준 실제 값을 기반으로 검증을 진행하였다.

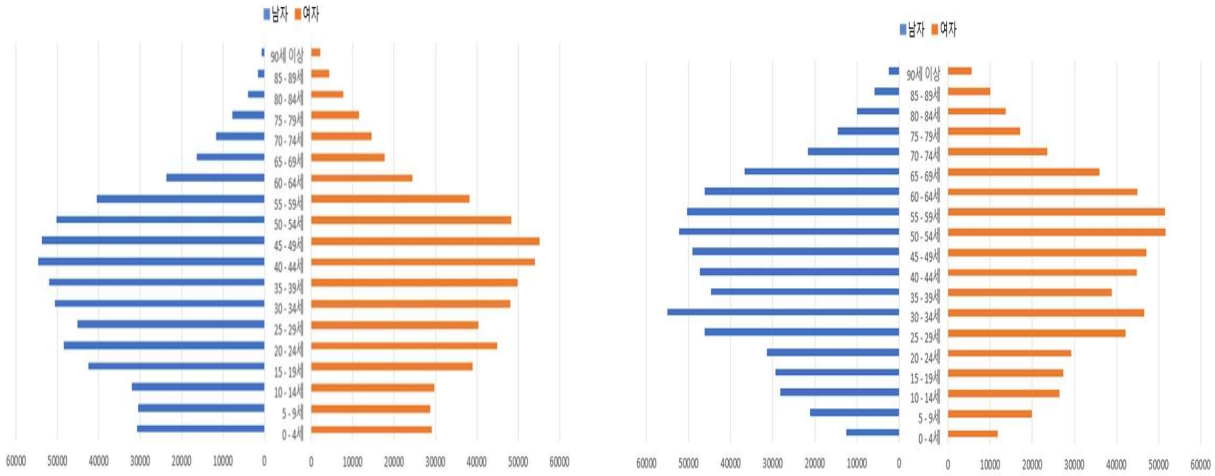


[그림 4-1] 인구 코호트 모형 히트맵

그 결과 출생아 예측모델을 결합한 코호트 모형이 0-4세 인구 부분에서 더 정확한 예측을 수행하였고, 모든 성능 평가 점수에서 더 적은 오차를 보여 해당 모델을 최종 모델로 선정하였다.

2) 결과 해석

□ 2015년, 2025년 5세 단위 인구 동향



[그림 4-2] 2015년 · 2025년 인구 그래프 비교

2015년과 2025년의 수원시 인구를 비교한 결과, 총 인구는 1,184,624명에서 1,193,266명으로 소폭 상승하였다. 총 인구 증가의 요인은 고령 인구의 증가와 전입 요인으로 볼 수 있다.

2015년부터 50세 이상의 모든 5세 단위 인구에서 증가 추세를 보이고 있고, 수원시의 50세부터 90세 이상 전체 인구는 2015년부터 10년 동안 약 65% 증가한 것으로 예상되었다.

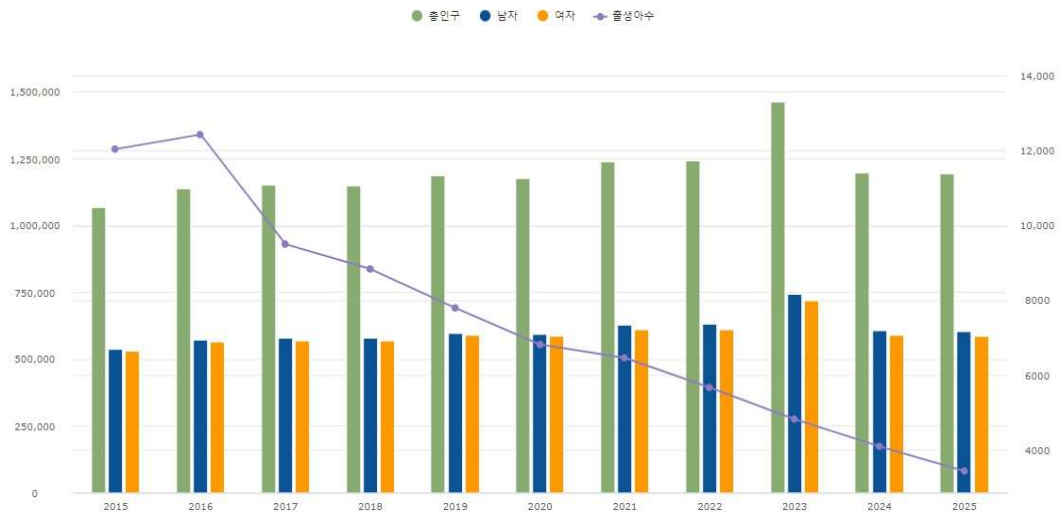
그러나 50대 미만 인구에서는 전입 요인으로 소폭의 증가를 보인 25-34세를 제외하고 모든 인구 단위에서 감소 추세가 나타났고, 50대 미만 전체 인구는 2015년 대비 11% 감소하는 것으로 나타났다.

특히 감소추세를 크게 보인 인구 단위는 출생아를 포함한 0-4세 단위로 10년 동안 약 60% 감소하였다. 이는 수원시 전체 출생아가 지속적으로 감소할 것이라 예측되기 때문에 코호트 모형에도 해당 추세가 반영된 결과이다.

□ 연도 별 인구 변동

분석 시점의 최신 데이터인 2020년과 최종 예측값인 2025년의 인구 변동을 비교한 결과, 총 인구는 2023년까지 증가 추세를 보이고 그 후부터 다시 원래 수준으로 감소하는 형태를 보인다.

2023년에 전체 인구수의 증가가 나타난 이유는 2018년 25-34세 연령대에서 평균적인 전입 인구수인 8,395명 대비 103,031명 많게 기록되었기 때문이다. 데이터 구조 상 인구 코호트가 5년 단위로 구성되어 있기 때문에 2018년의 전입 인구가 5년 뒤 코호트 모형에 반영되었고, 그 결과 2023년의 인구가 급증하는 현상이 나타나게 되었다.



[그림 4-3] 연도 별 인구 그래프

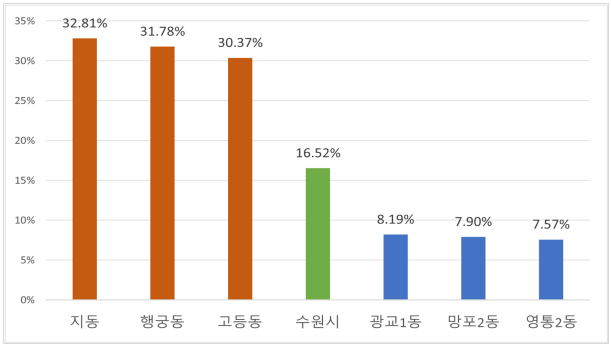
총 인구 증가에 가장 큰 영향을 끼친 것은 노년(65세 이상) 인구 단위의 증가이다. 노년 인구는 2020년 130,231명으로 전체 인구의 11%를 차지했는데, 2025년 197,185명으로 5년간 약 1.5배 증가하면서 전체 인구의 16%를 차지할 것으로 예상되었다.

15-64세 인구를 의미하는 생산가능 인구는 2020년 892,865명에서 2025년 876,094명으로 2.04% 감소하였다. 많은 수의 생산가능 인구가 노년 인구로 편입되고 있으나, 전입 인구의 증가로 15-64세의 총 인구수는 크게 변하지 않았다고 해석할 수 있다.

그러나 0-14세 인구를 의미하는 유소년 인구는 2020년 154,349명에서 2025년 119,987명으로 22.64% 가량 크게 감소하였다. 출생아 수가 꾸준히 감소하고 있는 것이 그 원인이며, 2020년 출생아수 6,814명에서 2025년 3435명으로 49.58% 가량 감소할 것으로 예측하였다.

□ 행정동 별 인구 변동

■ 노인 인구

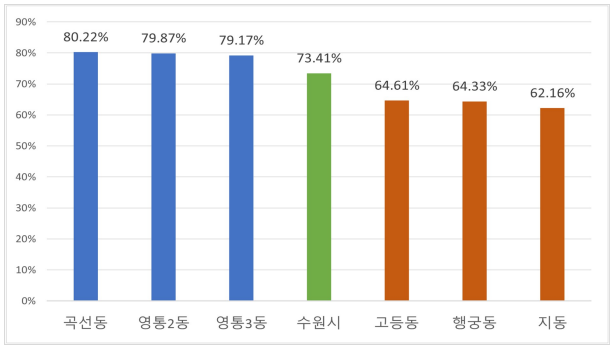


[그림 4-4] 노인 인구 그래프

인구 코호트 모형에 따라 수원시 행정동 인구 변동을 살펴보았다. 65세 이상 노년 인구가 전체의 30% 이상을 차지하는 행정동이 2020년에는 하나도 존재하지 않았으나, 2025년에는 고등동 · 지동 · 행궁동 3곳이 발생하였다.

특정 행정동의 노인 인구 비율은 2025년 수원시 전체의 노인 인구 비율로 예상되는 16%보다 크게 앞서고 있다는 것을 알 수 있다. 광교1동, 망포2동과 영통2동이 약 8%의 노인 인구 비율을 보이는 것을 보면, 수원시 내 행정동 간에서도 인구 구조에 큰 차이가 있다는 것을 확인할 수 있다.

■ 생산가능 인구

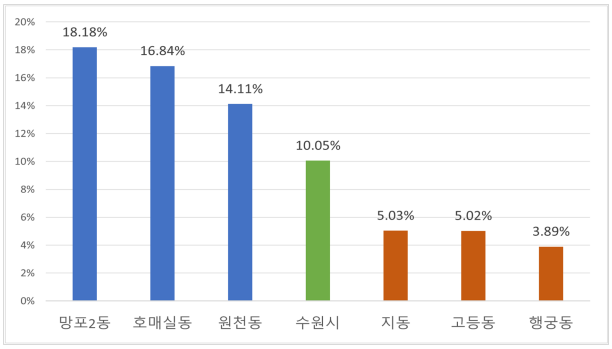


[그림 4-5] 생산가능 인구 그래프

생산가능 인구의 경우 전체 인구의 70% 이상인 행정동이 2020년 43개 동에서 2025년 33개 지역으로 감소할 것으로 나타났다. 또한 생산가능인구가 전체 인구의 65%에 이르지 못하는 행정동은 2020년에는 존재하지 않았으나, 2025년에 고등동 · 연무동 · 지동 · 행궁동 총 4곳이 발생할 것으로 나타났다.

노인 인구와 마찬가지로 특정 행정동의 경우 2025년 수원시 전체의 생산가능 인구 비율로 예측되는 73%에 크게 미치지 못하는 인구구조를 가질 것을 확인할 수 있다.

■ 유소년 인구



[그림 4-6] 유소년 인구 그래프

수원시 전체 인구에서 유소년 인구 비율은 10.05%로 노인 인구 비율인 16.52%보다 더 낮은 것을 확인할 수 있으며, 이는 2025년 수원시 인구구조는 노인 인구가 유소년 인구보다 더 많은 상태가 될 것임을 의미한다.

유소년 인구가 전체 인구의 15% 이상을 차지하는 행정동은 2020년 15개에서 2025년에는 망포2동과 호매실동을 제외하면 모두 사라질 것으로 예측된다. 반면 유소년 인구수가 10% 미만인 지역은 2020년 17개에서 2025년 전체 행정동 중 절반을 넘는 30개가 될 것으로 예측된다.

유소년 인구 비중은 노인 인구와 마찬가지로 상위 행정동과 하위 행정동 간에 큰 차이가 발생하고 있다. 지동, 고등동, 행궁동의 경우 노인 인구 비중은 매우 높게 나타나는 반면, 생산가능 인구와 유소년 인구 비율은 가장 낮은 불균형 인구 구조를 갖고 있다.

지역사회의 인구소멸 위험성을 판단하는 인구소멸 지수(노인인구 대비 20~39세 여성 인구)로 확인했을 때 수원시 전체의 인구소멸 지수는 약 0.8로 5단계 중 3단계에 해당하는 주의 단계에 해당하지만, 불균형 인구 구조를 갖고 있는 지동, 고등동, 행궁동의 경우 5단계 중 가장 위험한 인구 소멸지수 0.2 미만에 해당하는 소멸 고위험지역이 될 것으로 나타났다.

5 결론 및 정책적 제언

가. 결론

□ 인구 코호트 모형

출생아 예측 모형을 결합한 인구 코호트 결과에 따르면, 2025년까지 출생아 수는 지속적으로 감소하여 약 24,254명에 이를 것으로 예측하였다. 수원시 총 인구 자체는 1,193,266명으로 소폭 증가할 것으로 예측하였고, 그 원인은 전입 인구 증가와 노인 인구 사망률 감소로 볼 수 있다. 출생아라는 신규 코호트의 유입이 줄어들은 상황에서 전체 인구가 늘어난 것은 수원시 전체인구 중 노인 인구의 비중이 크게 상승했다는 것을 의미한다.

□ 출산율 변동 요인

출산율 변동 요인 분석에 따르면 가임기 여성의 고용률이 높을수록 출산율이 낮게 나타나는 현상이 있고, 생산가능 인구 전체에서 고용률이 높을수록 출산율은 높게 나타났다. 이는 출산율과 고용률이 밀접한 관계에 있다는 것을 의미하며, 가임기 여성 고용률은 음의 상관관계를 갖기 때문에 가임기 여성 취업자에 대한 정책적 지원방안이 필요하다는 것을 의미한다. 이외에 초등학교 수와 재정자립도와 같이 경제적 · 사회적 환경도 출산율에 영향을 끼치는 것으로 나타났다.

나. 정책적 제언

□ 시사점

수원시의 2025년 예상 노년 인구 비중은 16%로 고령사회 기준인 14% 보다도 높은 연령대의 인구구조를 갖게 될 것이다. 5년의 예측 기간 동안 수원시 노년 인구 비중이 크게 상승한 것을 보았을 때 지역사회의 고령화 추세가 계속 이어질 것이라 예상할 수 있고, 그 추세가 빠르기 때문에 신속한 인구정책의 체계화와 정책적 대응 체계 구축이 필요할 것으로 보인다.

대부분의 행정동에서 생산가능 인구와 유소년 인구가 감소할 것으로 예측되었고, 노년 인구가 행정동 인구 전체의 30% 이상을 차지하는 심각한 고령화 현상도 발견되었다. 불균형 상태의 인구구조를 개선하기 위해서는

출산율 변동요인으로 나타난 가임 여성 고용률 등의 사회적 요인을 충분히 반영할 수 있는 정책적 지원이 필요하며, 25~34세에서 높게 나타난 전입 인구를 유지할 수 있는 방안도 필요할 것으로 보인다.

특히 지동, 행궁동, 고색동은 유소년 · 생산 가능 인구에 비해 노인 인구 비율이 가장 높은 행정동으로 인구 소멸지수가 5단계 중 가장 심각한 단계의 소멸 고위험 지역으로 예상되기 때문에 해당 행정동에 대해 지역사회 단위에서의 노력과 더불어 수원시의 선제적인 정책 수립이 필요할 것으로 판단된다.

□ 제언

시계열 분석과 코호트 모형이 결합된 현재의 인구 코호트 모형을 사용할 시 기존 코호트 모형보다 비교적 정확한 0~4세 코호트 예측이 가능하여 관련 정책 수립에 이점이 존재한다. 특히 출산율 하락에 따른 인구 정책이 시민들의 큰 관심사이기 때문에 정책 방향 결정에 유용할 것이라 판단된다.

또한 코호트 모형이 2025년까지 5세 단위의 성별·행정동 장래 인구를 추계하기 때문에 인구 정책 뿐 아니라 복지 예산 결정, 지역단위 민원 분석 등 수원시의 다양한 정책 수립에도 활용 할 수 있을 것으로 기대한다.

□ 개선점

수원시 인구 코호트 모형의 분석 목표는 행정동 단위였으나 코호트 모형에 사용하는 인구균형 방정식의 투입 데이터가 행정동 단위까지 존재하지 않아, 이를 위한 관련 전처리가 필수적이었다. 전처리 과정이 많이 포함되어 있을수록 모형의 복잡도가 높아져 유지관리가 어렵고, 데이터 변형으로 인해 예측 결과의 정확도가 떨어질 위험이 있다.

행정동 단위의 데이터 수집이 이루어진다면 출생율 변동요인 분석과 인구 코호트 모형의 예측 정확도를 높일 수 있을 뿐 아니라, 수원시 데이터 기반 행정을 위한 분석에도 활용될 수 있기 때문에 데이터 구축 단계에서의 개선이 필요할 것으로 보인다.

참고 문헌

◈ 김기환·전새봄, 출산율 예측모형을 이용한 한국의 출산력 시나리오 분석, 2015

◈ 김태헌·김동회·정구현, 코호트 요인법을 이용한 시군구별 장래인구추계, 통계연구, 2006

◈ 김혁우·이필도, 계절 ARIMA 모형을 이용한 화장수요예측 - 수원시를 중심으로, 2017

◈ 서울대학교 보건대학원, 경기도 저출산 원인분석 및 출산동향예측 연구용역, 2016

◈ 수원시, 2021년 인구정책 시행계획, 2021

◈ Bruce H. Andrews 외 3명, Building ARIMA and ARIMAX Models for Predicting Long Term Disability Benefit Application Rates in the Public/Private Sectors, 2013

◈ Luís Torgo 외 3명, SMOTE for Regression, 2013

◈ Mick Smith·Rajeev Agrawal, A Comparison of Time Series Model Forecasting Methods on Patent Groups, 2015

◈ Paula Branco, Luis Torgo, Rita P. Riberio, SMOGN: a Pre-processing Approach for Imbalanced Regression, 2017