붙임 4 **데이터분석 보고서 작성방법 및 서식(응모분야1)**

접수번호	※작성하지 않음

「2023년 통계데이터 활용대회」 데이터분석 보고서

제 목	건설 경기의 미래를 바라보다: 통계 모형과 자연어 모형을 활용한 건설 경기 예측 모델 개선

신 청 자 명	소속/직위	국민대학교	성 명	박성대
	휴대전화	010-7346-5508	전자우편	psdae62@gmail.com
제출일	2023.07.21			

건설 경기의 미래를 바라보다: 통계 모형과 자연어 모형을 활용한 건설 경기 예측 모델 개선

1. 배경

□ 주제 선정

건설 경기는 부동산 경기의 선행지표이자, 국내 경기변동에 상당한 영향을 미치는 부문 중 하나이다. 건설경기는 성장기여율이 연평균 1.1%p(15~17년)가량으로 경제성장률에 미치는 영향이 매우 높다. 또한 생산 및 고용 연계성도 타 부문에 비해 높아 경제적 파급효과가 큰 것으로 평가되어, 건설 경기에 대한 정확한 예측은 정부, 기업, 그리고 민간의 의사결정에 도움을 준다.

건설 경기는 금년 1분기 원자재 공급난이 해소되며 1.9%(전기대비)로 반등했으나, GDP에 대한 성장기여도는 여전히 0%p(실질GDP성장률 0.9%중 기여도)에 머무르고 있다. 한국은행 경제전망보고서에 따르면 건설투자는 24년 상반기까지 역성장이 예상되며, 향후 건설 경기는 높은 불확실성을 유지할 것으로 보인다.

본 연구는 최근의 부동산 시장 불안정성, 정책 불확실성과 원자재 공급난 등 다양한 요인을 반영하는 건설 경기 예측 모델을 구축해 전망 분석의 정확도를 높이고자 한다. 나아가, 부동산 경기 예측 및 GDP 예측 모형에 응용 가능한 방안을 제시해 중장기적 정책 수립 및 기업의 의사 결정에 도움을 주고자 한다.

□ 분석 필요성(문제점) 및 전략

기존의 건설 경기 예측 모델은 건설업 선행 및 동행 지표들을 활용하거나, 특정 거시변수의 움직임이 건설 경기에 미치는 영향을 파악하는데 집중해왔다(박상우&황나윤, 2022; 박철한, 2021; 진경호 외, 2015 등) 이에 따라 건설업과 거시경제변수를 종합적으로 고려한 연구가 부족한 상황이다. 또한 건설 경기의 결정요인 중 계량화가 어렵거나 신속히 계량변수로 포착되지 않는 경우가 있다. 이러한 한계로 선행연구 중 종합 건설경기에 대해 50% 이상 예측 정확도를 갖는 경우가 없었고, 이는 보다 정확한 건설 경기 예측 모델의 필요성을 시사한다.

따라서 본 연구는 건설경기에 영향을 미치는 다양한 거시변수와 MDIS에서 제공하는 건설 산업 및 심리 지수를 이용하여 통계 모델을 구성한다. 또한 계량변수에 반영되지 못한 정보를 포착하기 위해 뉴스 텍스트 데이터를 함께 이용한다. 이를 통해 종합적 관점에서 건설 경기 변동 요인을 알아보고, 비정형 텍스트 데이터를 활용해 예측 모델의 정확도를 개선하는 것이 본 연구의 목적이다.

2. 데이터 분석

□ 데이터 선정(사용한 데이터 및 이유 등)

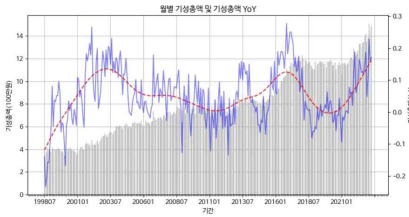
- 종속변수: 건설기성액

본 연구에서는 건설경기를 나타내는 지표로 예측의 신속성, 지표의 순환성을 고려해 건설기성액(계절조정)을 사용하였다. 변수의 안정성을 위해 전년동월대비 증감률을 이용하였다. 또한 건설경기의 지속성이 높은 점, 다수의 선행연구에서 종속변수의 시차변수를 독립변수로 이용한 점을 참고해 AR(1)의 시차변수도 포함했다. 건설기성액의 전년동월대비 증감률은 2012년 상반기까지 순환주기가 약 10년 이상으로 장기간 유지되었으나, 2012년 하반기부터는 5년 이하의 순환주기를 가지며 이전에 비해 변동성이 확대된 것을 확인할 수 있다. 현재 건설경기는 작년 하반기 고점에 도달한 뒤 수축국면에 진입할 것으로 보인다.

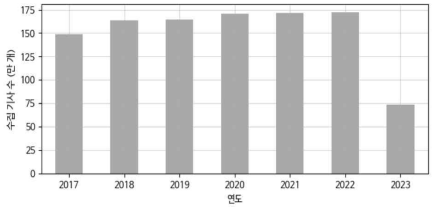
- 텍스트 데이터

분석에 사용한 데이터는 네이버 뉴스의 '경제' 섹션에 해당하는 기사를 연도별로 스크래핑하여 사용하였다. 2017년 1월부터 2023년 5월까지 월 평균 약 13만 8천개의 데이터를 수집하였으며, 각 기사별로 본문과 날짜 정보를 수집하여 저장하였다. 분석 과정에서는 키워드를 기준으로 필터링 과정을 거쳐 사용하였다.

<그림 1> 건설기성액 추이



<그림 2> 전체 뉴스데이터 수집 개수



주) 회색 막대 그래프는 기성총액 원자료,
빨간색 선은 기성총액 전년동기대비 증감률

- 독립변수: 공통 거시 변수 및 건설업 특수 변수

거시변수로는 건설 원자재 가격, 국제 공급망 압력, 국내경기를 반영할 수 있는 여러 변수들을 포함했다. 우선 건설 원자재 가격으로는 두바이유(현물, 달러당 배럴)와 수입물가지수(달러 단위) 중 철강과 유연탄을 활용했다. 이와 함께 2021년 팬데믹으로 초래된 국제 공급망 차질 사태를 반영하여 국제 공급망 압력지수[1]를 포함했다. 국내경기 변수로는 건설자금조달비용을 의미하는 신용스프레드[2]와 원/달러 환율을 이용했다. 또한 건설업 경기를 포착하기 위해 선행지표로서 건축 인허가를 반영하는 건축허가면적과, 동행지표로서 건설업 내 노동공급을 의미하는 건설 취업자수를 포함했다. 분양실적 등 부동산 시장 흐름에 따른 수요측 압력을 고려하기 위해 부동산 심리지수를 추가했다.

독립변수의 시차는 건설기성액 증감률과의 시차상관계수를 계산해 결정했다. 비용측 압력은 음의 상관관계를, 선행 및 동행지표는 양의 상관관계를 보였으며, 부동산 심리지수는 30개월 시차에서 0.6의 상관관계를 보였다. 건설 취업자수의 경우 동행지표로 0개월 시차에서 가장 높은 상관관계를 보였으나 예측을 위해 임의로 1기의 시차를 부여했다. 자세한 설명은 아래 표에서 확인할 수 있다.

<표 1> 분석 데이터

변수	기호	정의	기간	시 차 (개월)	출처
건설기성	*CRA*	전년동월대비 건설기성액 증감률	1997.07~2023.05	-	KOSIS
두바이유	*Oil*	전년동월대비 두바이유(달러/배럴) 증감률	1983.01~2023.05	15	ECOS
철강	*Steal*	전년동월대비 철강 수입물가(달러) 증감률	1971.01~2023.05	14	ECOS
유연탄	*Coal*	전년동월대비 유연탄 수입물가(달러) 증감률	1971.01~2023.05	24	ECOS
공급망압력지수	*GSCPI*	세계 공급망 압력지수	1997.10~2023.05	17	FED Newyork
신용스프레드	*Spread*	(회사채 3년, BBB- 금리) - (국채 3년 금리)	1976.01~2023.05	6	ECOS
원/달러 환율	*EXR*	전년동월대비 원/달러환율(매매기준) 증감률	1964.05~2023.05	24	ECOS
건축허가면적	*CPA*	전년동월대비 건축허가면적 증감률	2002.01~2023.05	30	KOSIS
건설 취업자수	*EMP*	전년동월대비 건설취업자수 증감률	2015.01~2023.05	1	MDIS, KOSIS
부동산 심리지수	*RES*	전년동월대비 부동산 심리지수 증감률	2011.07~2023.05	30	MDIS, KOSIS

□ 데이터 분석(분석 프로세스, 분석방법, 접근방법 등)

분석 단계는 크게 세 단계로 이루어진다. 우선 텍스트 데이터를 이용한 분석을 실시한다. 이후 텍스트 데이터 분석 결과를 지수화해, 계량 지표만을 사용한 다중회귀분석 모형과 비교하여 모형의 개선 여부를 점검한다. 이처럼 텍스트 데이터를 지수화할 경우 건설 경기 예측 뿐 아니라 GDP 예측모형, 부동산 경기 예측 등 다양한 분야에 쉽게 적용 가능하다는 장점이 있다. 마지막으로 분석 결과를 이용하여 향후 1년간의 건설경기를 예측하고 시사점을 제시한다.

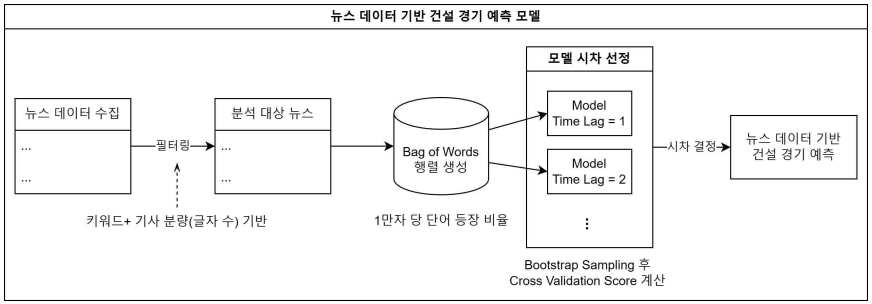
- 텍스트 데이터 분석 모형

텍스트 데이터의 분석은 아래 <그림 3>의 개념도와 같이 진행된다. 우선 수집된 전체 기사 중 건설 경기 예측에 있어 유의미한 기사만을 사용하기 위하여 필터링 과정을 실시하였다. '부동산' '건설' 단어가 기사의 본문에 등장한 횟수를 카운트하여 두 단어 중 하나 이상이 등장한 기사를 선별한다. 이후 수집된 기사 중 비교적 정보량이 많은 기사만을 선택하기 위하여 2000자 미만의 기사는 제외하였다. 위 필터링 과정을 거쳐 월평균 약 3300개의 기사가 분석 대상으로 추려

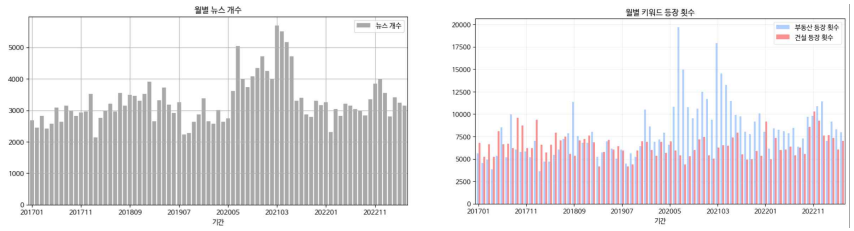
1) Global Supply Chain Pressure Index
2) 회사채 3년 BBB-와 국채 3년 간 금리 차

진다. 필터링 결과 부동산 키워드의 경우 2020년 중반부터 2021년 중반까지 등장한 횟수가 급격히 증가했다 감소하는 형태를 보였다. 건설 키워드의 경우 부동산 키워드에 비해 특정 기간에 대한 쏠림 현상이 적게 나타났다.

<그림 3> 텍스트 데이터 분석 모형 개념도



<그림 4> 뉴스 데이터 텍스트 필터링 결과



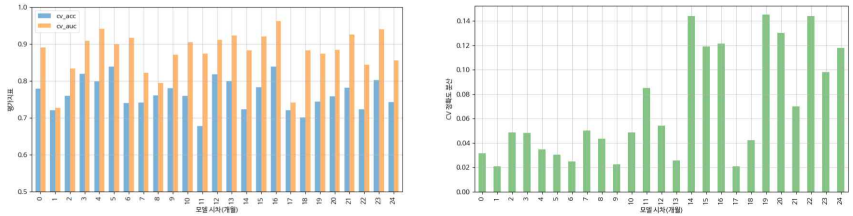
모델 훈련과 평가를 위해 데이터를 가공하는 과정은 다음과 같다. 먼저 필터링한 기사를 월별로 취합한 다음 Konlpy의 Okt 형태소 분석기를 사용하여 명사만을 추출한다. 추출한 명사들을 기반으로 Word Count를 실시하고, 열의 명칭이 각 단어이고 행의 값을 월별로 해당 단어가 등장한 횟수로 하는 Bag of Words 행렬을 생성한다. 다음으로 월별로 기사 수가 다른 부분을 보완하기 위하여 단어의 등장 횟수를 월별 단어 중 등장 비율로 변경한다. 이렇게 생성된 데이터에 시차를 적용하여 예측 모형을 생성한다.

이후 Bag of Words를 빈도로 정규화한 데이터에 Random Forest 모델을 적용하여 뉴스 데이터 기반의 건설 경기 변동 예측 모델을 제작하였다. 종속 변수는 건설기성액의 전년동월대비 상승(1), 하락(0) 여부로 설정하였다. 모든 Random Forest 모델에 대해 n_estimators를 100으로 설정하였고, 분석 과정에서 무작위성이 존재하는 부분에는 Random Seed를 1로 고정하였다.

예측 모델을 적합시키기 앞서, 종속변수와 독립변수 간 적절한 시차를 구하는 것이 필요하며 그 과정은 다음과 같이 이뤄진다. 가장 먼저 1~24개월 범위에서 월 단위 시차를 모델별로 설정한다. 이후 각 모델별로 시차를 적용한 데이터에 부트스트래핑을 적용하여 50개 크기의 학습 데이터를 무작위 복원추출로 생성한다. 해당 학습 데이터에 대해 K=3인 K-Fold Cross Validation을 실시한다. 이 과정에서 학습 데이터는 3개의 부분으로 분할되며, 2개 부분의 데이터로 모델 적합 후 나머지 1개 부분의 데이터로 평가 지표를 생성한다. 시차별 모델 평가의 결과로 3개의 평가 지표가 반환되며, 이를 평균한 값을 모델의 평가 지표로 한다. 또한 평가 지표 값 간 표준편차를 구하여 시차 선정의 보조 지표로 사용한다.

시차 분석 결과 정확도 기준 단기에는 3개월과 5개월, 중기에는 12개월의 시차를 선정하였다. 12개월 이후의 모델들은 정확도의 표준편차가 높아지는 형태를 보여 노이즈에 취약할 가능성이 높아 제외하였다.

<그림 5> 시차(lag)별 score / auc / std



- 다중회귀분석 모형

다중회귀분석은 이용 가능한 데이터의 길이, 시차의 길이에 따라 크게 세 모형으로 나누어 실시했다. 모형(1)은 1990년대 이전의 데이터가 존재하는 비교적 장기 데이터만으로 이루어진 모형이며, 분석기간은 $t = 2001.04 \sim 2023.05$이다. 모형(2)는 모든 변수를 활용했으며, 분석기간은 $t = 2015.01 \sim 2023.05$이다. 모형(3)은 중기 예측을 위해 12개월 이상의 시차를 갖는 변수만으로 이루어진 모형으로, 뉴스 텍스트 지수를 포함했다. 분석기간은 $t = 2018.01 \sim 2023.05$이다.

$$(1)\ CRA_t = \alpha_0 + \alpha_1 CRA_{t-1} + \alpha_2 Oil_{t-15} + \alpha_3 Steal_{t-14} + \alpha_4 Coal_{t-24} + \alpha_5 GSCPI_{t-17} + \alpha_6 Spread_{t-6} + \alpha_7 EXR_{t-24} + u_t$$

$$(2)\ CRA_t = \beta_0 + \beta_1 CRA_{t-1} + \beta_2 Oil_{t-15} + \beta_3 Steal_{t-14} + \beta_4 Coal_{t-24} + \beta_5 GSCPI_{t-17} + \beta_6 Spread_{t-6} + \beta_7 EXR_{t-24} + \beta_8 CPA_{t-30} + \beta_9 EMP_{t-1} + \beta_{10} RES_{t-30} + v_t$$

$$(3)\ CRA_t = \gamma_0 + \gamma_1 CRA_{t-1} + \gamma_2 Oil_{t-15} + \gamma_3 Steal_{t-14} + \gamma_4 Coal_{t-24} + \gamma_5 GSCPI_{t-17} + \gamma_6 EXR_{t-24} + \gamma_7 CPA_{t-30} + \gamma_8 RES_{t-30} + \gamma_9 INDEX_{t-12} + e_t$$

마지막으로 향후 1년간의 건설경기 예측은 다음을 이용했다(단 $1 \le h \le 12$).

$$\widehat{CRA}_{t+h} = \hat{\gamma}_0 + \hat{\gamma}_1 CRA_{t+h-1} + \hat{\gamma}_2 Oil_{t+h-15} + \hat{\gamma}_3 Steal_{t+h-14} + \hat{\gamma}_4 Coal_{t+h-24} + \hat{\gamma}_5 GSCPI_{t+h-17} + \hat{\gamma}_6 EXR_{t+h-24} + \hat{\gamma}_7 CPA_{t+h-30} + \hat{\gamma}_8 RES_{t+h-30} + \hat{\gamma}_9 INDEX_{t+h-12}$$

□ 분석 결과 및 해석

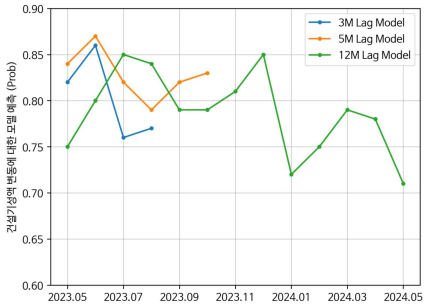
회귀분석 결과 장기시계인 모형(1)에서도 예측력을 의미하는 Adj. R-squared가 0.5를 넘었고, MDIS 제공데이터를 포함한 모형(2)의 경우 0.750의 우수한 예측력을 보였다. 비용 변수와 부동산 심리지수가 대체로 유의한 것으로 나타났다.

〈표 2〉 다중회귀분석 결과

변수	모형(1): 장기시계	모형(2): 단기시계	모형(3): 장기시차
$Const.$	0.0722*** (0.022)	0.0882 (0.105)	-0.0185 (0.012)
CRA_{t-1}	0.6065*** (0.052)	0.4020*** (0.103)	0.1322 (0.098)
Oil_{t-15}	-0.0122 (0.010)	-0.0461*** (0.013)	-0.0444*** (0.015)
$Steal_{t-14}$	-0.0297** (0.012)	-0.0352* (0.020)	0.0098 (0.017)
$Coal_{t-24}$	-0.0201* (0.012)	-0.0205 (0.013)	-0.0015 (0.012)
$GSCPI_{t-17}$	0.0057 (0.005)	0.0104** (0.005)	0.0138*** (0.005)
$Spread_{t-6}$	-0.0068** (0.003)	-0.0079 (0.016)	
EXR_{t-24}	-0.0310 (0.042)	0.0395 (0.108)	0.0416 (0.082)
CPA_{t-30}		0.0716* (0.040)	0.0225 (0.026)
EMP_{t-1}		0.0745 (0.261)	
RES_{t-30}		0.3074*** (0.079)	0.3642* (0.078)
$INDEX_{t-12}$			0.0744*** (0.013)
분석기간	$t = 2001.04 \sim 2023.05$	$t = 2015.01 \sim 2023.05$	$t = 2018.01 \sim 2023.05$
Adj. R-squared	0.516	0.750	0.815
F-statistic	37.66***	33.65***	24.59***
Durbin-Watson	2.228	2.121	1.933

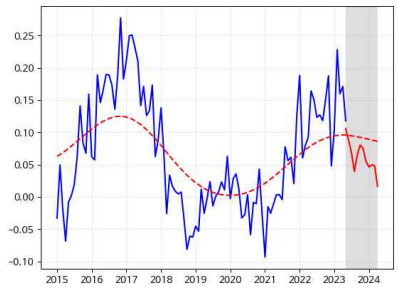
1) *** p<.01, ** p<.05, * p<.1, 2) ( ) 안의 값은 이분산에 견고한 표준오차(HC3)
3) Durbin-Watson 통계량은 오차항의 독립성을 의미하며 값이 2 근처일 경우 오차항의 독립성을 만족한다.

〈그림 6〉 텍스트 모델 건설기성 지표



주) 2023.05 이후 기간에 대한 모델 예측 기반

〈그림 7〉 건설기성 예측



주) 회색 음영은 2023.05 이후 예측치

텍스트 모델을 활용하여 2023년 5월 이후 기간에 대한 건설기성 변동 방향을 예측하였을 때, 시간이 지남에 따라 상승 확률이 떨어지는 형태를 보였다. 이를 기반으로 향후 단기적으로 건설기성액은 전년동월대비 양의 방향으로 변동할 것으로 예상되나 그 폭(증가폭)이 점진적으로 줄어들며 하방 추세로 전환될 가능성이 높을 것으로 추론할 수 있다.

뉴스 텍스트 데이터 분석을 통해 생성한 12개월 시차 지수($INDEX_{t-12}$)를 포함해 장기시차 변수로 구성된 모형(3)을 이용해 향후 1년간의 건설경기를 예측했다. 분석 결과 건설기성은 이번 하반기 수축 국면에 진입한 이후 2024년 상반기까지 하락세를 지속할 것으로 예상되며, 이는 상당 부분 건설 경기 사이클과 부동산 경기 위축에 따른 결과로 해석된다.

3. 분석 활용 전략

□ 기대효과

본 연구는 거시경제 변수와 건설업 특수 변수, 뉴스 텍스트 데이터와 같이 정형 및 비정형 데이터를 복합적으로 이용해 건설경기 예측 모델을 구축하고, 예측 모델의 정확도를 향상시켰다는 점에서 의의를 갖는다. 본 연구에서 구축한 건설경기 텍스트 지수는 부동산 경기 예측 및 GDP 예측 모형에도 활용 가능하며, 이들 모형에 있어서도 예측 정확도 향상에 기여해 유관 정책 수립과 민간의 의사결정에 도움을 줄 것으로 기대한다.

이와 함께 본 연구에서 제안한 뉴스 텍스트 데이터를 이용한 거시경제변수 예측 모형은 뉴스 텍스트 데이터의 지수화를 통해 기존 연구에 사용되는 계량변수와 비정형 데이터를 종합적으로 사용할 수 있도록 한다. 뉴스 텍스트 데이터는 신속성과 유연한 시계열 간격을 갖고 있어 분석 목적에 따라 다양하게 활용될 수 있으므로, 본 연구의 방법론을 응용해 반도체 경기 예측 모형 등 다양한 거시경제변수 예측 모형을 개선할 수 있을 것으로 기대한다.

□ 방향제시

본 연구는 시간적 제약으로 인해 5년 미만의 단기 시계에 대해서만 뉴스 텍스트 데이터 분석을 진행했다는 점에서 한계를 갖는다. 최근 건설 경기 사이클의 순환 주기가 짧아져 단기 및 중기 예측에는 높은 정확도를 보였으나, 장기 예측에 있어서는 한계를 보이는 모습이다. 또한 텍스트 분석 과정에서 단어의 등장 횟수만을 고려해 텍스트에 드러난 태도, 의견, 성향과 같은 주관적인 정보를 고려하지 못했다는 한계가 있다. 이는 향후 연구 기간이 장기로 주어질 경우 감성분석 등을 통해 보완할 수 있을 것으로 생각한다. 한편 MDIS 제공 데이터인 부동산심리지수, 건설업 취업자수는 10년 미만의 자료 길이를 갖고 있어, 추후에 이들 지표에 대한 장기적인 데이터가 확보된다면 긴 순환주기를 갖고 있는 건설업 경기의 예측 정확도를 더욱 높일 수 있을 것으로 기대한다.

참고문헌

박상우, 황나윤(2022). 최근 건설경기 상황에 대한 평가 및 시사점: 공급제약 요인을 중심으로. **BOK 이슈노트 제2022-20호.** 한국은행.

박철한(2021). **건설 경기종합지수를 활용한 공종별 건설경기 예측**. 한국건설산업연구원.

진경호, 이교선, 이수경, 양준석, 김현우, 김성종(2015). **건설경기 분석 및 예측**. 한국건설기술연구원.