

# 공공데이터 분석을 통한 변동성 요인 분석과 예측 모델 생성에 대한 연구

윤철희<sup>1</sup>, 한기형<sup>2\*</sup>

<sup>1</sup>경찰대학교 치안정책연구소 치안자율주행센터, <sup>2</sup>스타프리카

## Research on Analyzing Volatility Factors and Creating Predictive Models through Public Data Analysis

Cheol-Hee Yoon<sup>1</sup>, Ki-Hyoung Han<sup>2\*</sup>

<sup>1</sup>Police Science Institute Autonomous Driving Center, Korean National Police University  
<sup>2</sup>StarPrima

**요약** 공공데이터는 정보의 투명성, 시민참여 촉진, 효율적인 정책 결정을 위해 사용되고 있으며 또한, 데이터 접근이 편리해져서 누구나 쉽게 사용하고 있다. 이런 공공데이터는 과학적 연구기법과 다양한 목적 등을 통해 통찰력을 얻을 수가 있기 때문에, 새로운 분야 혹은 새로운 연구에서 데이터 기반 의사결정에 사용하고 있다. 본 논문 역시 주택가격 변동성 분석과 예측 모델 생성을 위해 공공데이터를 수집, 분석, 활용하였으며, 관련해 세종시 주택거래 가격변동성에 영향을 미치는 요인 분석을 하였다. 아파트 실거래내역과 거시경제 요인을 통합 하여 실거래 내역과 거시경제자료를 비교분석 후 상관성을 파악하였다. 또한, 트렌드 검색지수와 뉴스 보도지수를 활용하여 상관분석도 하였다. 최종적으로 주택가격 변동은 주거지역, 교통 편의성, 교육 시설, 경제적 상황 등 다양한 요인에 의해 영향을 받기 때문에, 본 논문에서는 공공데이터 기반으로 주택가격 예측 모델 생성을 하였으며, 합리적인 의사결정을 내릴 수가 있는 변동성 원인 상관관계를 제시하였다.

**Abstract** Open data is being used for information transparency, civic engagement, and efficient policy-making, and it is becoming more accessible and easy to use. These public data are being used for data-driven decision-making in new fields and new research because they can provide insights through scientific research techniques and for various purposes. This paper also collected, analyzed, and utilized public data to analyze house price volatility and create a prediction model, and analyzed the factors affecting the price volatility of housing transactions in Sejong City. By integrating actual apartment transaction history and macroeconomic factors, we compared and analyzed the actual transaction history and macroeconomic data to identify correlations. We also conducted a correlation analysis using the trend search index and news coverage index. Finally, since house price fluctuations are influenced by various factors such as residential areas, transportation convenience, educational facilities, and economic conditions, this paper generated a house price prediction model based on public data and presented the correlation of volatility causes for rational decision-making.

**Keywords** : Interest Rates, Money-supply, Data Crawling, OPEN API, Visualization

본 연구는 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-00901, 가상자산 부정거래 등 사이버범죄 활동 정보 추적 기술)로 수행되었음.

\*Corresponding Author : Ki-Hyoung Han(StarPrima.)

email: hankdn1@naver.com

Received November 13, 2023

Revised December 5, 2023

Accepted December 8, 2023

Published December 31, 2023

# 1. 서론

## 1.1 연구목적

코로나19 이후 금융과 부동산 정책 등의 변화에 맞추어 여러 기관 및 관계자들은 공공 데이터를 바탕으로 '주택 가격 변동성 요인 분석'에 대한 연구를 수행하고 있다. 분석에서 자주 거론된 변동성 요인은 지역별 교육 정보, 교통, 생활 편의 시설 위치, 거주 환경 특성, 투자 자산요소 등 공공 데이터이며, 분석을 통해 효율적인 의사결정 모델을 생성하고 있다. 본 논문은 주택 가격 변동에 대한 원인을 찾기 위해 '공공데이터'를 기반으로 '웹 크롤링', 'Open API' 방법을 이용해 데이터 수집을 수행하였으며, 전처리 과정에서의 데이터 추출 방법, 분석에서의 지수 분석 및 가격 변동 요인 결과도출 등 공공데이터를 활용하였다. 또한, 주택 가격 변동 원인을 파악하는 방법과 치안을 목적으로 공공데이터를 분석하는 추가적인 기여를 제시하는 것을 본 논문의 연구 목적으로 하였다.

## 1.2 연구의 범위 및 방법

본 연구의 데이터는 2014년부터 2022년까지의 세종시 데이터를 사용하였다. 세종시 아파트 부동산 거래가격에 영향을 미치는 변수로는 매매량, 소비 및 사회심리, 경제지표, 인구 이동량, 여론 지수 등 약 21여 종의 데이터를 사용하였다.

수집 데이터를 전처리하여, 이상값에 대한 결측치를 제외처리 하였으며, 부동산 지수를 기반으로 알고리즘화 하였다. 데이터 전처리 시 추출된 데이터는 OPEN API를 이용하여 변수로 사용된 데이터의 방향성 산출과 기준 데이터를 산정하였다. 전체적인 연구 프로세스 과정은 아래 Fig. 1과 같다.

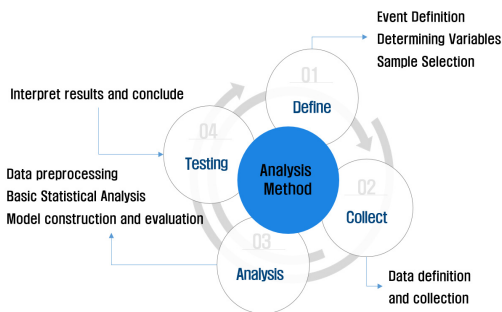


Fig. 1. Process of Research

추가로 경기순환과 경제지수를 파악하기 위해 다양한 공공데이터를 수집 후 전처리하였으며, 이상값에 대한 결측치를 제외처리 하였다. 그 결과 시각화를 통해 변동성 요인 분석과 예측 모델 생성에 대한 연구 수행이 가능하였다.

# 2. 본론

## 2.1 경기순환과 경제지수

본 연구는 세종시 지역 아파트 값 가격의 상승 및 하락에 영향을 주는 다양한 요인 산출을 시도하는 것으로 시작된다. 첫째, 변동성 원인 분석을 위해 한국 부동산원, 한국은행, KB부동산 등의 주택 거래 정보에서 주택 시세 전망 정보 등을 수집하였다. 둘째, 주택 가격 상승 요인으로 추정되고 있는 도시 유입인구 증가에 대한 데이터를 수집하였다. 셋째, 주택공급 부족과 팬데믹 후 양적완화로 통화량 증가, 그리고 저금리와 유동성 분석을 추가 시도하였다. 마지막으로 사회심리적 요인(주택가격 상승 시 실수요자보다 투기를 목적으로 불법 탈법적인 가수요가 증가하여 집값 상승) 등을 파악하였다[1].

또한, 주택 공급량, 부동산 불법 투기 방지, 다주택자 매도유도, 주택가격 상승에 대한 기대심리 억제 정책 등을 추가로 분석해 세종시의 아파트 가격 변동에 영향을 미치는 요인 선정을 하였다.

## 2.2 변동분석

### 2.2.1 데이터 수집

아파트 거래량 변동성에 영향을 끼치는 요인에 대한 선행연구 사례 분석을 통해 주요 요인을 선정하였다. 부동산 거래 변동성 관련 아파트 실거래 내역, 경제지수, 여론 동향 데이터를 수집한 후 매매거래, 소비 및 사회심리, 부동산, 경제지표, 이동량, 여론지수 등 약 21여 종의 자료를 대상으로 주택 유형별 매매, 주택유형별 전세, 시군구 법정동 코드, 사회심리 등의 경제지수와 여론동향을 분석하였다. 기간, 시군구 단위 법정동코드 등 주택유형별 거래 데이터는 Open API URL과 조합하여 호출할 수 있었다[2].

경제지수는 한국은행 ECOS 에서 제공하는 Open API를 사용하였으며, XML 형식 BeautifulSoup을 통해서 데이터프레임으로 변환하여 저장하였다.

Table 1. Data List

Item		Collected Data
TrainingData	Open Government Data Portal	Actual Price of an Apartment Sale
		The Actual Price of an Apartment Rental
Consumption Psychological	Open Government Data Portal	Consumer Price Index
		Housing Market Consumer Sentiment Index
		Residential Rental Market Consumer Sentiment Index
		Housing Market Consumer Sentiment Index
Social Psychology	Bank of Korea Economic Statistics	News Sentiment Index
		Economic Sentiment Index
		House Price Index
Real Estate Metrics	Open Government Data Portal	Apartment Sales Volume
		Average Price for Apartment Rentals
		Average Price of an Apartment for Sale
		Actual Price Index for Apartment Sales
		Apartment Transaction Price Index
Currency/Interest Rate	Bank of Korea Economic Statistics	Mortgage Rates
		Mortgage Amount
		Currency Rates(M2)
Movement	Open Government Data Portal	People Moved (by city and country)
		Number of City Moves (by municipality)
Public Opinion Trends	Google	Sejong Apartment Search Index
	Korean Media Promotion Foundation	Sejong Apartment News Coverage Index

통계청에서 제공하는 데이터는 Open API를 제공하는 라이선스키 + URL를 조합하여 데이터를 수집하였다. 여론동향은 구글 트렌드에서 키워드를 입력해 검색지수를 통합하여 저장하였고, 아파트 매매거래내역은 공공데이터포털에서 수집한 자료 3종을 통합하였다. 데이터 주기는 일별로 분류했으며, 거래유형(매매, 전세), 지역, 거래시기, 면적(평형), 아파트 단지명 등 31개 항목을 선별하였다. 그리고 ECOS, KOSIS, 빅카인즈, 구글트렌드 등에서 수집한 15종의 데이터를 통합하여, 월별로 소비

자 물가지수, 소비심리지수, 아파트 거래량, 사람 이동량, 검색지수 등을 수집하였다.

### 2.2.2 데이터 전처리

지역명에서 특정문자열 검색하여 시/읍/면으로 지역구분 항목으로 생성하여 2013년 이전 자료 등 활용성이 낮은 항목은 Fig. 3 과 같이 삭제 후 데이터를 분석하였다.

```

## 항목명 소문자로 치환
cols = df.columns.str.lower()

for i in range(len(cols)):
    df.rename(columns={df.columns[i]:cols[i]}, inplace=True)

df.columns

Index(['price', 'trade_type', 'build_year', 'year', 'road_nm', 'road_bon',
       'road_bu', 'road_sgg_cd', 'road_l_cd', 'load_ud_cd', 'load_cd',
       'si_gun', 'emd_nm', 'bon', 'bu', 'sgg_cd', 'emd_cd', 'bon1', 'dg_nm',
       'month', 'day', 'code', 'usemj', 'real_estate_type', 'bunji', 'law_cd',
       'floor', 'release_date', 'release_cause', 'price_by_py', 'std_date',
       ...])

## 지역항목 생성
ddf['si_gun_gu'] = ddf['emd_nm'].str.split(" ").str[0]

## 지역항목 값 확인
ddf['si_gun_gu'].value_counts()

초치원읍      8591
중촌동        3856
고운동        3833
도담동        3423
한솔동        3215
아름동        3179
새롬동        1545
    
```

Fig. 3. Create a regional item

거래구분항목에서 빈 값으로 나타난 Null 값은 “기타”로 처리하였으며, 결측치가 있는 10 여개 항목에 대해 지역별 2014년 전세/매매 최저가를 산출하였다. 2013년 빈 값 보정 후 시각화 추이 분석을 위해 날짜 항목 생성하였으며, 일부 항목은 Data Type을 변환하였으며, 추가로 특수문자를 제거하였다. 또한, 아파트 실거래 내역 및 경제지수 데이터를 확인 후 경제지수별 분석 및 시각화를 실시 하였다. SQL Query 방식의 데이터 함수를 이용하여 경제지수 컬럼명, 지역명, 경제지수명을 호출하여 조회하였다, 특히, 시각화함수를 사용하여 거래종류 및 경제지표를 실거래내역과 경제지표 데이터 중심으로 증감율을 표현하였다[3].

### 2.2.3 변동성 분석 및 시각화

#### 1) 경제지수 분석

소비자 물가지수는 가계 일상소비생활을 영위하기 위해 구입하는 상품(460여개)과 서비스의 가격변동을 대상으로 측정하였으며, 기준년도를 중심으로 소비자물가지

수는 매년 가파르게 상승하는 것을 그래프로 표현하였다. 물가지수 변동폭이 가장 높은 곳은 충청북도로 나타났다. 2021년부터 물가지수 변동폭을 추이하면, 코로나 팬데믹의 영향으로 물가지수가 가파르게 상승한 것으로 분석되었다.

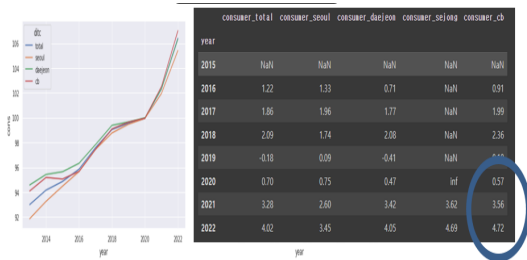


Fig. 4. Consumer Price Index and Percentage Change

또한, 국토연구원의 데이터 중 가구주, 중개업소를 대상으로 매월 제공하는 부동산시장 소비자 심리지수를 통해 추가분석을 하였다. 소비자심리지수는 0~200사이의 값으로 지수가 100을 넘으면 가격상승이나 거래증가로 해석된다. 분석된 지수에 따르면 2020년 기준 세종지역 주택매매시장 소비자심리지수가 전국에서 가장 높고 실제로 매매가가 상승하였음을 파악할 수 있었다[4].

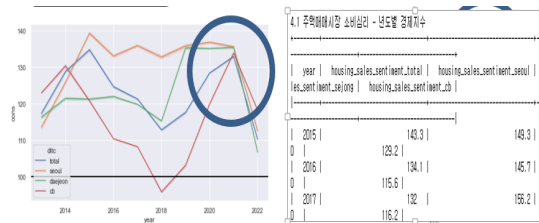


Fig. 5. Housing Market Consumer Sentiment

### 2) 금융지수 분석

한국은행에서 매월 조사하는 주택담보대출금리, 주택담보대출금액, M2 통화량, 아파트 매매실거래지수, 주택담보대출금리 등을 이용해 상관관계를 파악하였다. 통화량과 대출금액 관계는 양의 상관의 특징을 보였으며, 통화량과 주택담보대출금리 관계는 음의 상관의 특징을 보였다. 특히, 아파트대출금액, 아파트 매매실거래지수는 양의 상관으로 통화량과 대출금액, 아파트 매매 실거래지수가 증가함을 파악하였다. 반대로, 2021년 중반에는 대출금리가 상승하면서 아파트 매매거래량은 급격히 하락하였음을 상관관계 분석으로 파악하였다[5].



Fig. 6. Correlation analysis of Rates, Currency, Live Trading

### 3) 뉴스보도 및 온라인 동향 분석

구글 트렌드 검색 키워드를 통해 아파트 매매, 아파트 전세로 검색 지수를 추출 후 분석을 수행하였다. 결과로 구글 트렌드 매매검색지수가 높으면 전세검색지수가 상대적으로 낮음을 파악했으며 특히, 2022년 경우에는 아파트매매 검색지수가 급격히 떨어졌으며, 검색지수를 분석한바 아파트 매매와 아파트 매매거래 지수가 낮아지고 있음을 추가 파악할 수 있었다[6-10].

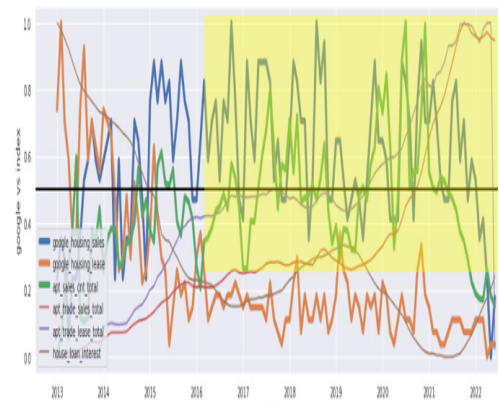


Fig. 7. Google Search Index and Trading Volume

한국언론진흥재단 뉴스보도 키워드 분석은 세종시 아파트, 세종시 아파트 분양, 세종시 아파트 매매, 세종시 아파트 전세 키워드로 수행하였다. 세종시 부동산 아파트 분양, 매매, 전세 관련 뉴스 보도가 줄어드는 추세를 파악하였으며, 2021년 이후 뉴스 보도 비교 시 실제 거래량은 낮아졌음을 알 수 있다.

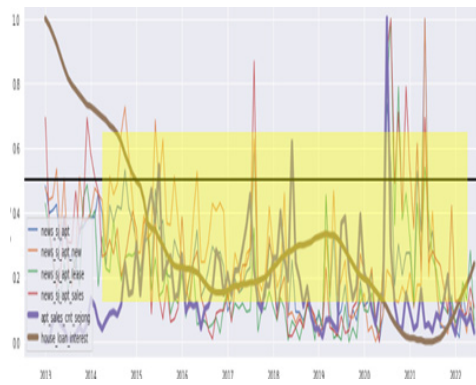


Fig. 8. News Coverage Index and Trading Volume

### 3. 결론

#### 3.1 데이터 분석 결과

세종시 아파트 매매거래와 금리, 통화량 증감률 비교 분석 결과 세종시 부동산의 매매량은 2019년 집계된 금리와 통화량 그리고 매매가의 변동폭의 변화 기율기와 정방향으로 비례하며, 금리와 통화량은 반비례하는 변동폭을 보였다. 2021년부터 금리가 오르고 통화량이 줄자 매매량 건수가 급격히 하락하는 결과가 도출되었다.

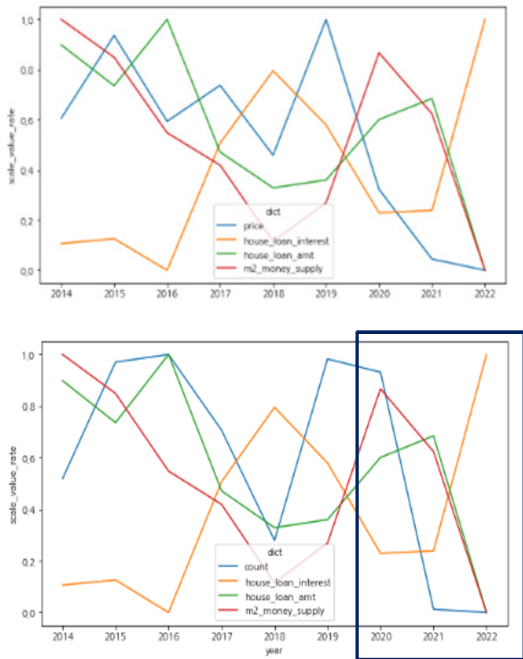


Fig. 9. Price Index of Price and Volume

세종시 아파트 매매거래와 사회심리지수의 시계열 증감률 비교 분석 결과로는 사회심리지수와 매매가가 정의 상관관을 보이고 있음을 알 수 있다. 2021년 심리지수가 하락하면서 거래량도 크게 떨어진 것으로 나타남을 볼 수 있다. 그리고, 세종시의 아파트 매매거래와 여론지수 증감률 비교분석 결과로는 세종시 매매가와 트렌드 검색지수는 뉴스보도량과 검색 및 거래량으로 정의 상관으로 약하게 비례함을 알 수 있었다. 세종시 아파트 매매 거래와 물가지수 증감률 비교분석 결과는 매매가 대비 물가지수가 매매가와 전국 단위 주택, 전세시장 소비심리지수가 거의 정비례하며, 물가지수와 주택시장 소비심리지수는 반비례함을 보였다. 특히, 거래량 대비 물가지수는 2020년 부동산지수와 전국단위 부동산지수 증감률이 일치하였다.

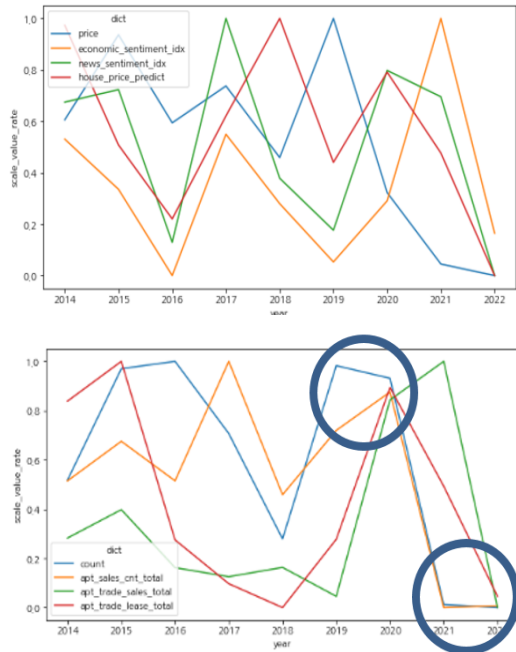


Fig. 10. Price and Volume Relative to the Sentiment Index

#### 3.2 연구의 공헌

공공데이터는 정보의 투명성, 시민참여 촉진, 효율적인 정책 결정을 위해 사용되고 있으며 또한, 데이터 접근이 편리해져서 누구나 쉽게 사용하고 있다. 본 논문은 세종시의 공공데이터를 활용하여 과학적 연구기법과 분석을 통해 데이터 기반 의사결정에 사용하는 방법을 제시하고 있다. 세종시 매매내역과 경제지수를 시계열 추이를 분석하였으며, 경제 지표 관점에서 전국 부동산 매매

량을 실거래지수, 금리, 통화량, 심리지수, 검색지수에 따른 상관성을 분석하였다. 또한, 여론지수를 위해 포털 검색과 뉴스 보도 검색을 수행하여 매매거래량 대비 변동폭도 분석하였다. 이를 통해 경제 지표 외에도 구글, 뉴스 등 SNS여론도 부동산 변동성에 영향을 미치고 있음을 확인도 하였다.

본 연구는 첫째, 세종시 주택거래 가격변동성에 영향을 미치는 요인 분석을 하였다. 이를 통해 아파트 실거래 내역과 거시경제 요인을 데이터기반으로 분석 후 영향력 높은 결정요인 확인하였다. 둘째, 세종시 주변지역과의 아파트 실거래에 대한 시계열 추이 분석을 통해 가격변동성의 미래 예측가능성과 효과성 역시 확인하였다. 셋째, 포털 트렌드 검색지수와 한국언론 진흥재단의 뉴스 보도 지수를 활용하여 미시적 의사결정 활용가능성 제시하였다. 넷째, 범죄예방 분야에서 부동산 변동성과 범죄율을 통한 범죄율 평가를 확인할 가능성을 확인하였다. 특히, 주택가격은 주거지역, 교통 편의성, 교육 시설, 경제적 상황 등이 다양한 요인에 의해 영향을 받기 때문에, 주택가격은 일반적으로 안전한 지역, 교육 기관이 뛰어난 곳 등에서 높게 형성될 가능성이 높기 때문이다. 마지막으로, 도시의 경제적 성장과 직접적으로 연관되어지는 도시 범죄율 연구에 대한 필요성이다. 부동산 가격은 사회 경제적인 요인, 교육 수준, 실업률, 경찰 치안 등 다양한 분야로부터 영향을 복합적으로 받고 있지만, 특히, 범죄율이 높은 지역에서는 주택가격이 낮게 형성될 가능성이 높기 때문이다.

향후 그 연관성 및 분석기법을 고도화하여 연구할 계획이며, 본 논문에서 사용한 공공데이터 변동성을 예측했던 기법을 기반으로 치안관련 분석을 추가 수행할 예정이다. 현재, 가상자산을 악용한 범죄 발생 분석과 범죄 추적을 위한 데이터분석 기법을 접목하고 있으며, 112 신고량과 범죄율을 융합 비교 분석 연구를 고대하고 있다.

## References

[1] Y. Kim, J.C Park, and H.Y Oh. "Real Estate Price Outlier Analysis Using Autoencoder Technique." Journal of the Korea Communications Society 25. pp.1739-1748. Dec. 2021.

[2] R. Y. Kim, "Linkage analysis of house price changes in the Seoul Metropolitan Area and Chungcheong Province after the launch of Sejong City". Real Estate Analysis, 7(3), 79-98. 2021.

[3] J. Hong, S.R. Ahn. "Are Seoul apartment prices a

bubble? Real Estate Analysis", 8(1), 1-21. 2022.

[4] H, Song, J. Jung, "A Study of Determinants of Apartment Prices by Living Area in Sejong Administrative Center Complex." Journal of Real Estate 86, pp.41-65. 2022.

[5] B. G. Cho, K.B.Park, S.H Ha. "Comparison of apartment real transaction price index prediction models by region using machine learning algorithms: Verification of LIME interpretability." Information Systems Research 29, pp.119-144, March. 2020.

[6] <https://www.hapt.co.kr/news/articleView.html?idxno=153607>

[7] <https://wisdomagora.com/Korea-House-Price-Rise-Causes>

[8] [http://www.ohmynews.com/NWS\\_Web/View/at\\_pg.aspx?CNTN\\_CD=A0002806708](http://www.ohmynews.com/NWS_Web/View/at_pg.aspx?CNTN_CD=A0002806708)

[9] <https://m.post.naver.com/viewer/postView.naver?volumeNo=33937974&memberNo=38946978>

[10] <https://www.r114.com/?c=information&m=noticeInfo&a=noticeInfoDetail&bno=30&num=884>

윤 철 희(Cheolhee Yoon)

[정회원]



- 2016년 9월 : 고려대학교 정보보호대학원 디지털포렌식학과 (공학석사)
- 2023년 2월 : 연세대학교 일반대학원 기술정책과정 (공학박사)
- 2005년 2월 ~ 현재 : 경찰대학 치안정책연구소 연구관

<관심분야>

데이터분석, 정보통신, 인공지능

한 기 형(Ki-Hyoung Han)

[정회원]



- 2010년 2월 : 건국대학교 정보통신대학원 프로젝트기술경영학(경영석사)
- 1995년 7월 ~ 2019년 12월 : 한전KDN ERP 컨설턴트
- 2014년 1월 ~ 2014년 12월 : 한수원 사이버보안 관제센터PM
- 2022년 1월 ~ 현재 : 스타프리카 기술상무

<관심분야>

정보보안, 빅데이터, 인공지능