데이터로 설계하는
대한민국의 내일

제15회
국가통계·데이터발전포럼

2026. 5. 14.(목), 10:00~17:00 | 더플라자호텔

주관 : 국가데이터처, 한국통계학회, 한국경제학회, 한국조사연구학회

데이터로 설계하는
대한민국의 내일

제15회
국가통계·데이터발전포럼

2026. 5. 14.(목), 10:00~17:00 | 더플라자호텔

주관 : 국가데이터처, 한국통계학회, 한국경제학회, 한국조사연구학회

국가데이터처

목차

기조연설

국가데이터처 : 빅데이터와 AI시대의 국가데이터 통할본부 008
류근관 명예교수(서울대학교 경제학부)

세션1

국가데이터처 격상과 데이터 거버넌스

발표1 국가데이터기본법 제정(안) 및 의의 028
김지은 과장(국가데이터처 국가데이터기획협력과)

발표2 국가데이터 허브 기능 강화 로드맵 034
김수영 과장(국가데이터처 국가데이터허브정책과)

발표3 AI 친화적 국가 메타데이터 구축 방안 044
이선희 과장(국가데이터처 인공지능통계혁신과)

세션2

경제를 읽고, 내일을 설계하다: 경제총조사

발표1 경제총조사의 의의 및 추진방향 060
김혜련 과장(국가데이터처 경제총조사과)

발표2 국가통계의 변천과 경제총조사 데이터의 학술적 정책적 가치 074
홍석철 교수(서울대학교 경제학부)

발표3 국가데이터처 미시경제데이터의 정책의사결정 활용도 제고 방향 084
장우현 소장(한국조세재정연구원 국가회계재정통계센터)

발표4 일본의 경제총조사 추진전략 및 방향 094
Nanami suzuki(일본 총무성 통계국)

세션3-1

국가데이터 시대의 공식통계: 데이터 통합, 활용, 그리고 안전한 연계

발표1 빅데이터 결합을 통한 개인사업자 금융지원 정책 효과 분석 110
김미루 팀장(한국개발연구원 국채연구팀)

발표2 AI 시대의 조사데이터 통합: 가중치 보정의 역할과 가능성 144
권용현 교수(육군사관학교 수학과)

발표3 안전한 데이터 연계와 활용을 위한 플랫폼 운영모델 제안 156
박민정 과장(국가데이터처 통계기준과)

발표4 행정자료 기반 주거비지수 작성방안 168
홍현정 서기관(국가데이터처 물가동향과)

세션3-2

AI 에이전트와 사회조사

발표1 대형언어모델이 조사연구에 가져오는 새로운 가능성과 위협 188
강동현 교수(연세대학교 사회학과)

발표2 AI Agent와 사회현상 연구 210
한경식 교수(한양대학교 데이터사이언스학부)

발표3 LLM을 이용한 텍스트 분석 236
한상원 교수(성균관대학교 사회학과)

기조연설

국가데이터처 : 빅데이터와 AI시대의 국가데이터 통할본부

류근관 명예교수(서울대학교 경제학부)

국가데이터처: 빅데이터와 AI 시대의 국가데이터 통할본부

류근관
서울대학교 경제학부 명예교수
ryu@snu.ac.kr

Rapidly Changing Economy

- AI revolution and data
- pragmatism: from ideology to data
- “No data, no evidence-based policy”
- information asymmetry (lemon problem) and role of data
- open vs. proprietary data

Three BIGs

Big Data + Big Model ⇨ Something Big !

⇧

Big Computing

Three Big’s: Big Data, Big Model, Big Computing

- Complementarity among features
- Complementarity between data and model
  - (Big Data, Big Model)
  - (Small Data, Small Model)
- Big computing backs up (Big Data, Big Model)

No Data, No AI

- Data
  - Existing AI models such as foundational LLMs have used up public domain data
  - Not much left to be utilized
  - Need to explore high quality domain specific data

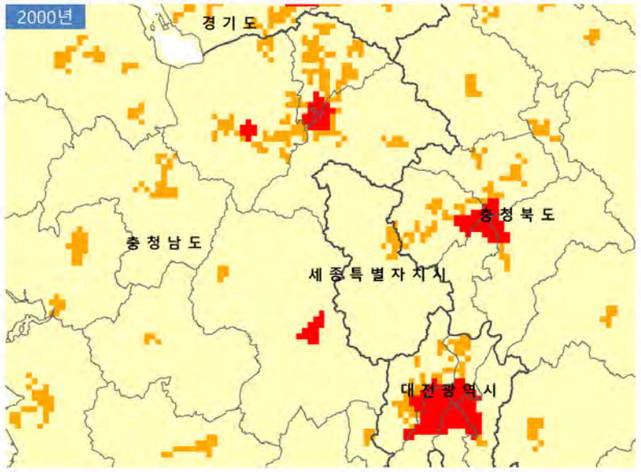
No Data, No AI

- Issues
  - Privacy protection in urgent need
  - risk of re-identification increases as more pieces of information are linked
  - pseudonymization is not an answer (c.f. curse of dimensionality)
  - an answer is privacy protection based on PQC (post quantum cryptography)

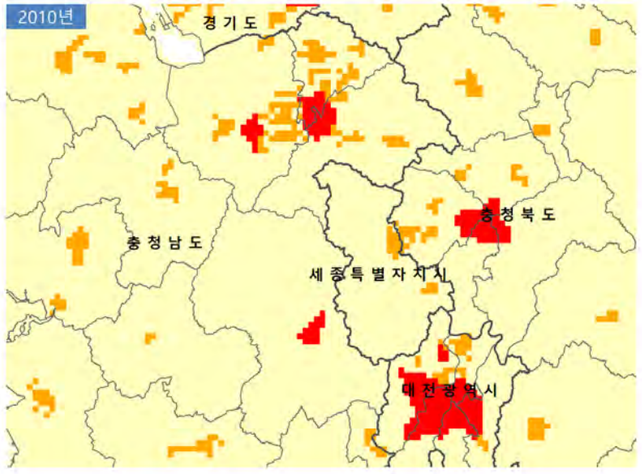
Geo-spatial panel data

- Link diverse data using (i, j, t) as observation units
  - (i, j)=(latitude, longitude)
  - t=time
  - Thus, y_{ijt} is a geo-spatial panel data on y
  - When y = population density, useful to investigate growth and decline of regions
  - Useful to study why regions grow or decline as a function of several $x's$ such as governmental policy, industry, job availability, education, hospital, age composition, etc.
  - So far, geo-spatial panel data has been under-utilized by economists.

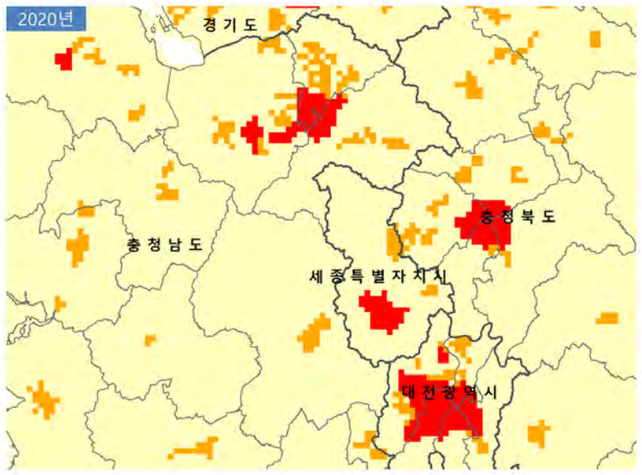
대전·충청권 도시 변화



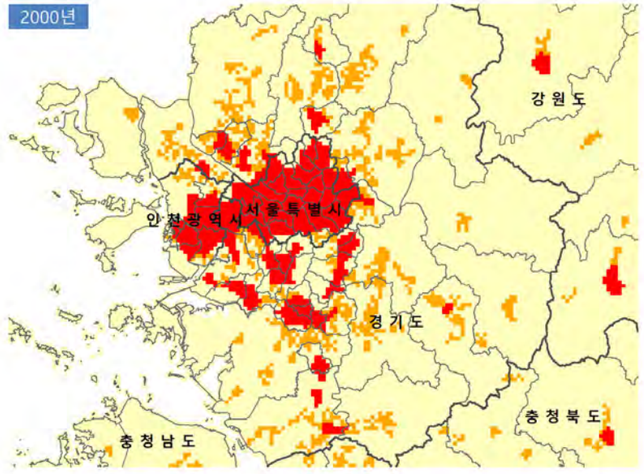
대전·충청권 도시 변화



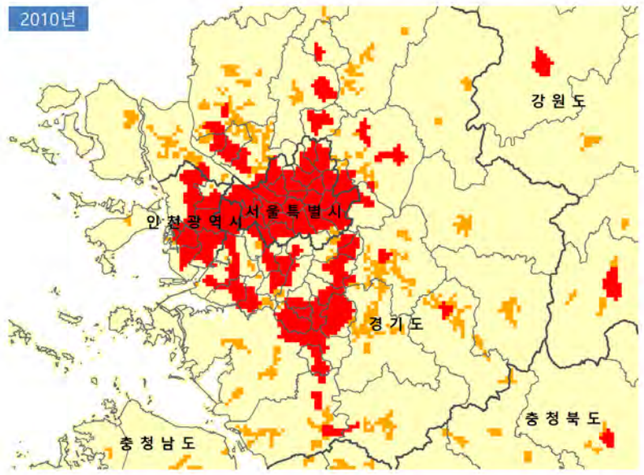
대전·충청권 도시 변화



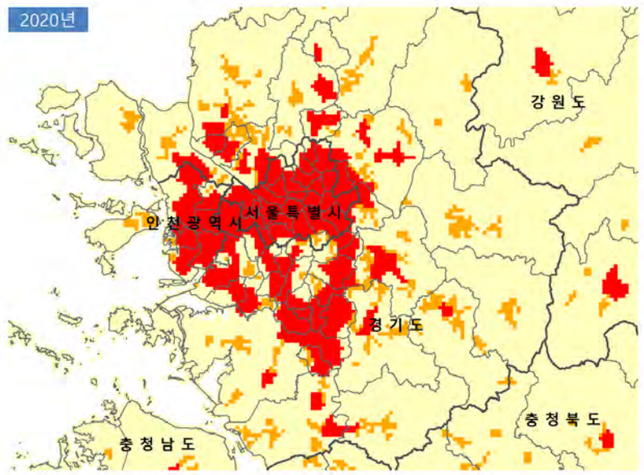
수도권 도시 변화



수도권 도시 변화



수도권 도시 변화



Statistical Registers

- Two foundational statistical registers
  - Population (individuals & households)
  - Establishments

  → The key axes for integrated data linkage

Data Linkage: Need to link income and asset

- Household Asset & Income by Age of Household Head

  2021 Household Financing and Welfare Survey, KOREA

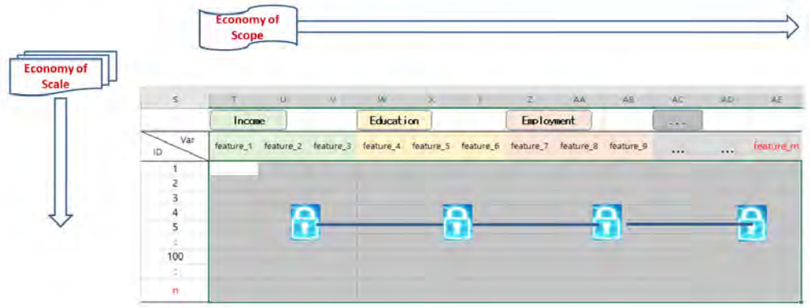
Age of Household Head	P(high asset)	P(low income)	P(low income | high asset)	P(high asset | low income)
All	0.400	0.400	0.187	0.187
65 or over	0.355	**0.704**	**0.452**	**0.228**
below 65	0.417	0.286	0.102	0.148

High asset=top 40% in terms of asset
Low income=bottom 40% in terms of income

- To better understand the old age poverty problem in Korea, data linkage is an absolute necessity!

Vision for Data Platform

Diverse data can be safely linked and analyzed



Vision for Data Platform

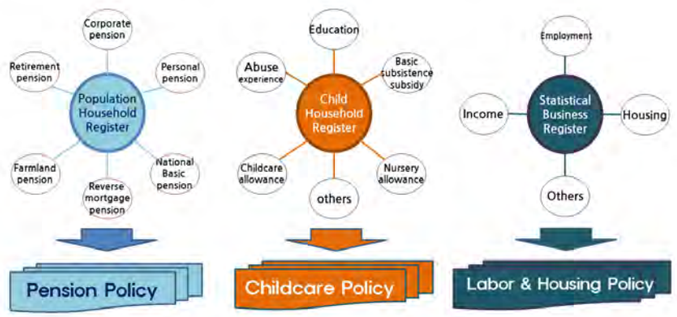
(As-Is)

- **Data scattered across diverse public and private institutions**
  - lack of cooperation and trust among the data holders
  - lack of standardization
  - differences in data storage

(To-Be)

- **Data Linkage using statistical registers**
- **Data Security using HE, DP, and synthetic data**

Data Linkage Using Statistical Registers



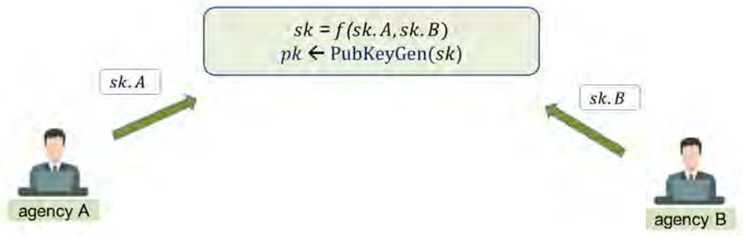
Data Protection Using HE, DP, and SD

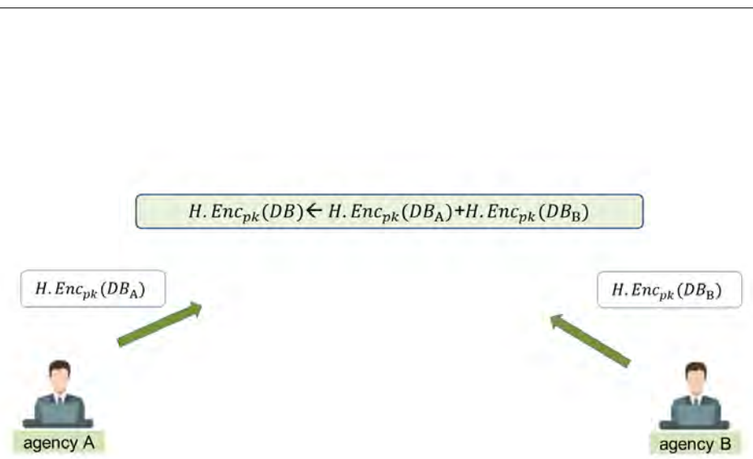
- Homomorphic Encryption (HE)
- Differential Privacy (DP)
- Synthetic Data (SD)

Homomorphic Encryption (HE)

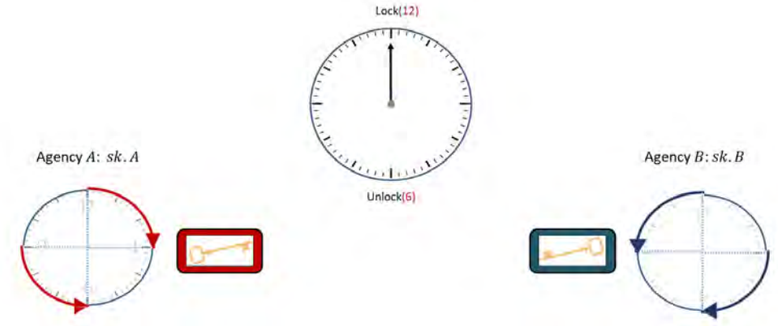
- Encryption and computation are commutative
- No decoding required for computation
- Need to set up decryption key management system

A Key Management Scheme for Encryption & Decryption
(based on personal communication with Junghee Chun and Jungdae Hong)

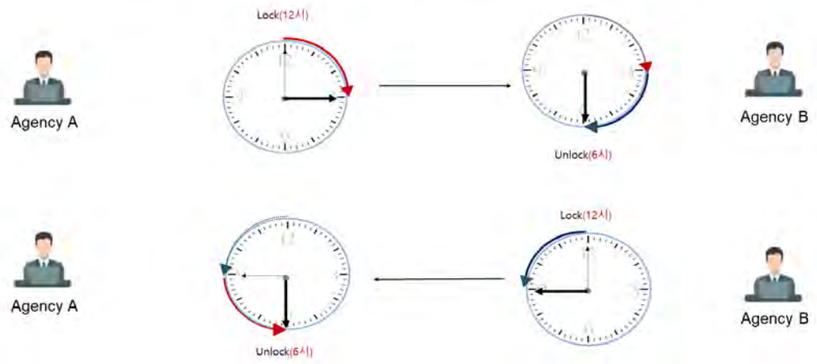




No single agency can decrypt the results



Only mutual agreement between A and B can decrypt the results



Differential Privacy (DP)

"released data (statistics) = observed data (statistics) + random noise"

Global Differential Privacy

Two pieces of aggregate (global) information reveal a disaggregate (local) information

$n_a = 100\ \bar{Y}_a = \$30{,}000$
$n_b = 99\ \bar{Y}_b = \$29{,}000$
$\Rightarrow$ "excluded person's income" is recovered as
$y_{100} = \mathbf{\$129{,}000}$

To protect the individual information, one may report the aggregate (global) information with a small, negligible error. Then individual information is protected by much inflated error (n times big !)

Local Differential Privacy

Have you ever used any illegal drug last year? (response based on coin toss)

$$y=\begin{cases}1, Tail\ or\ 'Head and\ Used' \\ 0, \quad 'Head\ and\ Not\ used'\end{cases}$$

→ can still identify the true proportion, p, of illegal drug users while protecting privacy of those whose y values are equal to 1.
c.f. Let $q = p(y = 1)$. $q = 1/2 + p/2$. Thus, $p = 2q - 1$.

Synthetic Data:

Synthetic data are randomly generated data according to a data generating process such that

- the resulting data do not reveal individual privacy
- and that the resulting data "look like" the original data.

Role of HE, DP, and SD within Statistics Korea's Data Platform

- HE: safe linkage and usage of data across diverse data owners
- Local DP and SD: providing data analysts with sufficient "practice opportunities" such as data exploration and model development in plain text which is absolutely necessary as encrypted data are "heavy" and inconvenient
- Global DP: Release computed statistics in an error-ridden form if privacy is extractable from the statistics by repeated queries

Healthcare Information

- World-class quality and quantity of Korean medical data
- Difficulty in sharing individual patient data due to privacy issues
- If individual medical data are further linked to family lineage, the Korean data would be ideal for research on genetic disease

Comprehensive Household Debt Statistics

- Household debt is increasing in Korea. Yet, no comprehensive debt statistics is available. By linking a private credit bureau's micro financial information with the population and household registers of Statistics Korea, we are currently in the process of making the comprehensive household debt statistics in Korea.
- Further, we also link other registers such as economic activities registers, business registers, and residence ownership registers. Not for sample. Rather, for the entire population. At the household level as well as at the individual level.

Comprehensive Pension Statistics

- Korea ranked 1st in terms of old-age poverty rate out of 36 OECD countries in 2019
- Korea has become a super-aging society : the percentage of population above 65 will reach around 40% by 2050
- Based on a key agreement/decision reached on April 14, 2022, Statistics Korea is currently compiling Comprehensive Pension Statistics
- Various pension data have been safely linked based on Statistical Registers
- Comprehensive pension statistics compiled Oct. 2023 for the 1st time ever in Korea (based on April 2022 agreement)

Worldwide recognition of Statistics Korea's Data Vision

1. UN has included Statistics Korea's Data Platform Project as Case Study 11 in "United Nations Guide on Privacy-Enhancing Technologies for Official Statistics," 2023, United Nations Committee of Experts on Big Data and Data Science for Official Statistics
https://unstats.un.org/bigdata/task-teams/privacy/guide/2023_UN%20PET%20Guide.pdf

2. RWI-Essen, a research institute in Germany, has decided to include "Statistics Korea's Governmental Data Platform" as a best-practice example for innovative methodology and data use worldwide in its final report for the German Labor Ministry

빅데이터 AI 시대의 국가데이터처

1. 조직 체계
   – 현재는 부처 간 데이터가 칸막이에 막혀 국정 운영의 효율성이 저하됨
   – **개선 방향**: 집행부처로부터 독립된 국가데이터처를 중심으로 연계된 데이터 기반의 **국정 기록·평가 체계를 실질적으로 구축해야 함**

2. 국가데이터 연계의 통할 권한
   – 정부 부처와 공공기관 자체가 **핵심적인 국가 데이터 생산자**
   – 성장률 둔화, 연금개혁, 청년, 인구, 지역 문제는 **상호 연결된 구조적 문제 (연립방정식)**
   – 따라서 데이터를 개별적으로 관리할 것이 아니라 **연계하여 통할할 수 있는 실질적 권한 필요**
   – 그 실질적 통할 권한이 국가데이터처에 부여되어야 함

빅데이터 AI 시대의 국가데이터처

- **집행과 측정의 분리**
  -국정이든 금융이든 집행과 측정 간 방화벽(firewall) 없으면 위험. 도덕적 해이 봉착
  -국토교통부가 부동산 정책도 집행하고 아파트 가격 상승률 측정도 해서 일었던 통계 불신
  -233년 된 영국 Barings 은행이 head of trading과 head of audit 겸한 Leeson에 의해 파산
  -조선왕조실록은 왕권으로부터 독립적으로 기록되고 관리됨
- 집행부처의 일부였던 **통계청의 국가데이터처로의 독립이 갖는 역사적 의미**
  -집행과 측정의 분리 통해 국가데이터 및 국가통계의 신뢰성 및 국정 운영의 효율성 제고
  -AI, 빅데이터 시대에 꼭 맞는 정부 조직 개편

국가데이터처: 빅데이터와 AI 시대의 국가데이터 통할본부

Session 01

국가데이터처 격상과 데이터 거버넌스

좌장

류근관 명예교수(서울대학교 경제학부)

1. **국가데이터기본법 제정(안) 및 의의**

   김지은 과장(국가데이터처 국가데이터기획협력과)

2. **국가데이터 허브 기능 강화 로드맵**

   김수영 과장(국가데이터처 국가데이터허브정책과)

3. **AI 친화적 국가 메타데이터 구축 방안**

   이선희 과장(국가데이터처 인공지능통계혁신과)

토론

권순일 부사장(업스테이지), **신신애 본부장**(한국지능정보사회진흥원)

국가데이터기본법 제정(안) 및 의의

2026. 05. 14.

국가데이터처

목 차

01 국가데이터기본법 제정(안) 개요
02 국가데이터 지정
03 국가데이터 품질관리
04 국가데이터 플랫폼
05 국가데이터 이용센터

1 국가데이터기본법 제정(안) 개요

국가데이터처

국가데이터 지정
국가 차원에서 관리 및 연계·활용의 필요성이 인정되는 데이터

국가데이터플랫폼
국가데이터를 효율적으로 관리·활용하기 위한 플랫폼 (메타데이터, 데이터관계도 등에 기반한 데이터 탐색, 분석 신청 등)

국가데이터 연계·결합
국가데이터 간 데이터 분석, 분석결과 제공 등

국가데이터 품질관리
국가데이터 보유·관리기관 소관 국가데이터의 안정적 품질관리 및 적정한 품질수준의 확보·유지 목적

국가데이터 이용센터
국가데이터의 연계·활용을 지원하기 위하여 전문인력·시설 등을 갖춘 안전한 환경

관련 특례규정
개인정보, 과세정보 관련 특례 규정 (국가데이터 연계·결합 관련)

2 국가데이터 지정

국가데이터처

1) 국가데이터 정의

데이터*

국가데이터

국가 차원에서 관리 및 연계 · 활용의 필요성이 인정되는 데이터로
- ✓ 인구·산업·고용·노동·재난·안전·국토·교통·보건·복지·환경 등의 분야에서 공공안전, 재난대응 또는 사회 문제 해결 등 공공가치의 창출 가능성이 있거나 사회·경제적 파급효과가 큰 데이터
- ✓ 국민의 의사결정 지원 또는 정책 수립을 위하여 국민 수요 또는 정책 수요가 높다고 평가되는 데이터
- ✓ 그 밖에 이에 준하는 데이터로서 국가데이터처장이 국가데이터로 지정하는 것이 필요하다고 인정하는 데이터

공공데이터 | 민간데이터

* 데이터기반행정법, 데이터산업법, 지능정보화 기본법 등에서 기 규정

2 국가데이터 지정

국가데이터처

2) 국가데이터 지정 및 지정 제외 대상

지정 대상

▪ **국가 차원에서 관리 및 연계 · 활용의 필요성이 인정되는 데이터**

- 인구·산업·고용·노동·재난·안전·국토·교통·보건·복지·환경 등의 분야에서 공공안전, 재난대응 또는 사회 문제 해결 등 공공가치의 창출 가능성이 있거나 사회·경제적 파급효과가 큰 데이터
- 국민의 의사결정 지원 또는 정책 수립을 위하여 국민 수요 또는 정책 수요가 높다고 평가되는 데이터
- 그 밖에 이에 준하는 데이터로서 국가데이터처장이 국가데이터로 지정하는 것이 필요하다고 인정하는 데이터

지정 제외

▪ **(원칙) 국가데이터 지정 제외**

- 국가안전보장 또는 국방·통일·외교관계 등에 관한 데이터
- 국가핵심기술 등 대통령령으로 정하는 기술과 관련된 데이터
- 영업비밀에 해당하는 데이터
- 그 밖에 국가의 중대한 이익을 크게 해칠 우려가 있다고 국가데이터처장이 인정하는 데이터

▪ **(단서) 국가기관등의 장과 협의를 거쳐 관리·연계 및 활용이 필요하다고 인정하는 경우 위원회 심의·의결을 거쳐 지정 가능**

2 국가데이터 지정

국가데이터처

3) 국가데이터 지정 절차

지정 대상

가목	인구·산업·고용·재난 등 공공가치 창출 또는 파급효과
나목	국민 수요·정책 수요가 높다고 평가되는 데이터
다목	그 밖에 처장이 국가데이터로 지정이 필요하다고 인정

지정 프로세스*

수요조사
현황·실태 파악 → 지정에 반영

직권 지정 (공공)	지정 신청(공공+민간)
협의 국가기관등과 사전 협의	지정 신청** 기관이 처장에 신청
위원회 심의 심의·의결	협의 국가기관등과 협의
	위원회 심의 심의·의결

공고·통보
지정 공고 및 보유·관리기관 통보

* 국가데이터 관리 지원, 컨설팅, 품질인증 등
** (민간데이터) 지원 사업 우선 지원 등 조치 및 보상

지정 제외

1호	국가안전보장 또는 국방·통일·외교관계 등에 관한 데이터
2호	국가핵심기술 등 대통령령으로 정하는 기술
3호	영업비밀에 해당하는 데이터
4호	그 밖에 처장이 인정하는 데이터

단서: 사전 협의로 필요 시 지정 가능

3 국가데이터 품질관리

국가데이터처

국가데이터처장	보유·관리기관의 장	품질인증기관
품질 기준 수립·진단·평가 시정 요구·포상	품질 확보·유지 조치 분류체계 준수	인증 기준에 따라 품질인증 실시

품질관리 전체 사이클

+ 국가데이터 지정 전 품질진단	01 분류체계 구축·준수	02 메타데이터·관계도 관리	03 품질 진단·평가	04 품질 인증
지정 전 사전 품질진단·평가	처장이 분류체계 구축·보완	처장이 작성·관리 기준 마련	처장이 정기 품질진단 실시	처장이 인증기관 지정
전문가 자문단 구성·의견 청취	통계법§22 표준분류 연계 검토	보유기관별 메타데이터 작성	결과 공표	대통령령 기준에 따라 인증
결과 반영하여 지정 여부 결정	보유기관 준수 의무	데이터관계도 작성·관리	현저 미흡 → 지정 해제 검토	인증자 지원사업 우선 지원
세부사항 대통령령 위임	미준수 시 시정 요구	미흡 시 처장이 시정 요구	우수 기관·개인 포상	인증 취소 규정

⚑ 제재: 분류체계 미준수·메타데이터 미흡 → 시정 요구 / 품질 현저 미흡 → 국가데이터 지정 해제

✦ 인센티브: 품질인증 취득 시 지원사업 우선 지원 / 우수 기관·개인 포상 / 공공기관 경영평가 반영 협조 요청

4 국가데이터플랫폼

국가데이터처

DATA FABRIC

메타데이터 관리
국가데이터·메타데이터 체계적 관리 및 최신성 유지

데이터 탐색 지원
데이터관계도 기반 탐색 및 상호운용성 유지

저장 체계 마련
데이터 유형별 저장 체계 구축

정책 의사결정 지원
AI·데이터 분석 통한 정책 수립 지원 및 이력 관리

국가데이터 플랫폼

연계·결합 지원
국가데이터 연계·결합 분석 업무 수행 지원

탐색용 국가데이터 제공
익명처리된 국가데이터 제공

등록관리시스템
수요조사 결과 기반 지정대상 목록 관리

플랫폼 연계
다른 법령상 데이터 플랫폼 상호 연계·활용

운영주체: 국가데이터처장 (데이터 보유·관리기관의 장에게 소관 데이터·메타데이터 제공 요청 가능)

5 국가데이터 이용센터

국가데이터처

1) 국가데이터 이용센터 주요 기능

① 데이터 처리
국가데이터 연계·결합·가공 등 데이터 처리

② 분석 결과 제공
국가데이터 분석 자료 및 결과 제공

③ 민간데이터* 연계
국가데이터 + 민간데이터* 연계·결합·가공 등 처리

* 계약에 의한 구매 및 업무협약 등을 통해 제공요청한 민간데이터

④ 보존 대상 데이터세트 관리
개인정보 보호법상 파기 의무 특례 부여

⑤ 처장 지정 업무
국가데이터처장이 필요하다고 인정하는 사항

5 국가데이터 이용센터

국가데이터처

3) 국가데이터 이용센터 내 특례 규정

개인정보 처리 특례

▪ 개인정보보호법상 민감정보, 고유식별정보 처리 특례
- 국가기관등의 정책수립을 위한 국가데이터 연계·결합 시 불가피한 경우 민감정보, 주민등록번호 등 고유식별정보 처리

▪ 개인정보보호법상 파기 의무 특례
- 반복적인 정책 수립·평가 또는 국가안전보장, 재난대응 또는 공공안전을 위하여 계속 보존할 필요성이 있는 경우 등
- 보존대상 데이터세트로 등록하여 저장·관리

과세정보 처리 특례

▪ 이용센터로 지정받은 국세통계센터 내 이용
- 국가데이터 이용센터로 지정된 국세통계센터
- 국가기관등의 장 및 연구기관의 장
- 국가데이터로 지정된 과세정보와 다른 국가데이터(예: 통계등록부) 연계·활용 가능

5 국가데이터 이용센터

국가데이터처

2) 국가데이터 연계·결합 절차

국가데이터
- 공공데이터 (직권) 국가기관등과 사전협의, 위원회 심의의결 + (지정신청)
- 민간데이터 (지정신청)

민간데이터 계약·구매·업무협약

국가데이터플랫폼
① 메타데이터·관계도 기반 데이터 탐색
② 데이터 유형별 저장 체계 운영
③ 비식별처리 및 연계·결합 지원

국가데이터 이용센터
① 국가데이터 연계·결합·가공 등 처리
② 국가데이터 + 민간데이터 연계
③ 분석 자료 및 결과 제공

연계·결합 관련 세부 절차

01 수요조사	02 지정	03 데이터 수집	04 신청·검토	05 연계·결합	06 결과 제공
국가데이터 지정 수요 확인 조사	국가기관등과 협의, 위원회 심의·의결 국가데이터 지정	보유기관에 데이터 및 메타데이터 요청 (플랫폼)	이용센터에 분석 신청	데이터 연계·결합 수행, 비식별 조치 후 분석	분석 결과 제공 비밀 보호

⚑ 개인정보 특례: 민감정보·고유식별정보 처리 가능 / 주민등록번호등 처리 가능 (국가데이터 연계·결합을 위하여 불가피한 경우 한정)

감사합니다

국가데이터 허브 기능 강화를 위한
추진전략 로드맵(안)

2026. 5. 14.(목) / 국가통계·데이터발전포럼

김수영 국가데이터허브정책과장

국가데이터처

목 차

Ⅰ. 추진배경
Ⅱ. 대내외 환경분석
Ⅲ. 현황 및 문제점
Ⅳ. 추진전략 로드맵(안)
Ⅴ. 실행기반
Ⅵ. 기대효과

Ⅰ. 추진배경

환경 변화
AI 대전환 시대 진입

- 데이터 중요성 증대, 데이터 관리 역량이 국가 경쟁력과 직결
- 사회·경제 복잡화로 데이터 규모·유형 확대, 데이터 활용 수요 다변화

핵심 변화: AI 활용 등 데이터 수요 급증 및 다변화

위상 변화
범정부 데이터 총괄기관

- 국가데이터처 승격에 따라 범정부 데이터 총괄기관으로서의 위상 강화
- 다양한 데이터의 융합·활용을 통해 가치를 창출할 수 있는 핵심 허브 역할 대두

핵심 역할: 데이터 수집, 연계, 분석, 제공의 전주기 관리

데이터 수집, 연계·결합, 활용까지 데이터 全 주기에 걸쳐
국가데이터 허브 기능 강화를 위한 체계적 추진전략 및 실행과제 수립 필요

3

Ⅱ. 대내외 환경 분석(SWOT)

강점 (Strength)
내부 역량 분석

- 통계생산 및 데이터 연계·분석 경험과 역량 보유
- 기존 데이터 관리 인프라 보유 (SDC, KOSIS, MDIS 등)

→ ST 전략: 신기술 도입으로 안전한 데이터 연계활용 지원

약점 (Weakness)
내부 문제점 분석

- 데이터 공백 현황 파악 미비
- 데이터 제공 시의성 한계

→ WO 전략: 데이터맵 구축을 통한 수요기반 데이터 확충

기회 (Opportunity)
외부 환경 분석

- AI·디지털 전환으로 데이터 수요 급증
- 범정부 데이터 총괄기관으로서 위상 강화

→ SO 전략: 적극적 데이터 연계·결합을 통한 데이터 가치 제고

위협 (Threat)
외부 위험요소 분석

- 부처 간 데이터 칸막이(Silo) 현상
- 개인정보 보호 강화로 데이터 활용 제약

→ WT 전략: 법 제정으로 데이터 칸막이 해소

4

III. 현황 및 문제점

	데이터 입수	데이터 연계	데이터 활용
현황	행정자료 입수 327종 327종 행정자료 및 67종 민간자료 입수 198종 행정DB 구축 및 제공 ('25.12. 기준)	융합데이터 구축 7종 아동가구, 청년, 취업활동, 연금, 소득이동, 생활인구, 중소기업	통계데이터센터(SDC) 운영 16개소 행정자료·통계등록부로 통계 생산 지원 SDC·KOSIS 등 데이터 서비스 플랫폼 운영
문제점	입수 체계 분산 및 데이터 공백 발생 • 입수 체계 분산 • 수요 대비 데이터 공백 발생 • 서식 및 데이터 형식 불일치	데이터 제공 시의성 부족 • 융합데이터 제공 시의성 한계 • 신규 데이터 요청시 건별 신속한 대응에 한계	물리적·시간적 제약, 데이터 환류 부족 • 물리적 제약: 직접 방문 필요 • 시간적 제약: 연계, 반출 시간 소요 • 자료제공기관의 데이터 활용 한계

IV. 추진전략 로드맵(안)

비전: 국가데이터 가치 창출을 위한 핵심 허브로 도약 — 목표 3개 (전략 9개) / 실행기반 4개 (2026~2030)

목표 1: 데이터 공백 발굴 및 해소
- 추진전략 1-1. 수집체계 일원화
- 추진전략 1-2. 데이터 공백 해소
- 추진전략 1-3. 데이터 표준 수립

목표 2: 데이터 연계·결합 기반 강화
- 추진전략 2-1. 융합데이터 개선·개발
- 추진전략 2-2. 범용 모듈형 데이터셋
- 추진전략 2-3. 데이터뱅크 도입

목표 3: 데이터 활용 확산 및 가치 제고
- 추진전략 3-1. 하이브리드 방식 전환
- 추진전략 3-2. 온라인 분석환경 구축
- 추진전략 3-3. 범정부 활용 촉진

실행기반 1: 법·제도 | 실행기반 2: 시스템 구축 | 실행기반 3: AI 활용 | 실행기반 4: 조직·인력

IV. 추진전략 로드맵(안) 목표 1: 데이터 공백 발굴 및 해소

1-1. 수집체계 일원화

추진전략
- (필요성) 데이터 입수체계 분산에 따른 기능 중복 및 관리 어려움
- (추진방안) 여러 부서로 분산되어 있는 데이터 수집 채널 일원화 및 통계등록부 구축 등 기능별 전담체계 확립
  - ✓ 現 행정자료관리과 개편(데이터입수관리과) 및 소속기관(특수법인) 일부 업무 이관 등 업무 재설계 검토
- (기대효과) 입수 데이터의 체계적 관리 및 대외 협상력 강화

추진일정

구분	내용	추진일정
단기	부서간 기능조정, 특수법인 설립 등 조직·기능 재편	~'27년
중장기	데이터 수집 절차 및 형식 표준화, 품질 관리 체계 구축	~'30년

데이터 수집체계 일원화(안)

	AS-IS	TO-BE
관련 부서	• 행정자료관리과(행정자료) • 인구총조사과(행정자료) • 빅데이터통계과(민간자료) 등	• 데이터관리과(가칭) (행정자료, 민간자료)
담당업무 (기능)	• 행정자료관리과 -자료입수(행정자료) -품질점검 등 자료정제1차 -품질진단·관리 -행정자료DB 구축, 제공	• 데이터관리과(가칭) -자료입수(행정자료↑, 민간자료✦) -품질점검 등 자료정제1차 -품질진단·관리 -DB 구축, 제공 ※ 특수법인에 일부 기능 이관
	• 자료입수·활용부서 -자료입수(행정자료, 민간자료) -품질점검 등 자료정제1차 -작성통계 기준으로 자료정제2차 -자료 이용 통계생산	• 자료활용부서 <통계생산부서> -작성통계 기준으로 자료정제2차 -자료 이용 통계생산 * 업무필요에 따라 개별입수有 <융합데이터 구축 부서> - 융합데이터 구축 - 범용 모듈형 데이터셋 구축

IV. 추진전략 로드맵(안) 목표 1: 데이터 공백 발굴 및 해소

1-2. 데이터 공백 해소

추진전략
- (필요성) 미보유 데이터 발굴 및 단계별 입수 전략 마련 필요
- (추진방안) 정책·연구수요 대비 처 내 보유데이터 현황을 보여주는 '데이터맵'을 작성, 데이터 공백을 식별하여 필요 데이터 확보
  - ✓ 보유데이터 매핑 → 우선순위별 데이터 공백 해소
- (기대효과) 전략적 공백 발굴을 통한 수요 기반 데이터 확충

추진일정

구분	내용	추진일정
단기	정책 수요 카탈로그 구축, 데이터맵 작성, 핵심공백 선정	~'27년
중장기	민간데이터 협력체계 확보, 자동 매핑시스템 구축	~'30년

데이터맵 작성(예시)

외부 수요			매칭	내부 보유	
부문	분야	분석변수		기 확보 데이터	입수 필요 데이터
정책수요 (국정과제)	가계부채	부채(대출)	→	NICE, KCB 자료(일부)	신용정보원 자료(전수)
		자산(금융,실물)	→	주택소유통계 DB	금융자산 자료
		소득	→	일자리행정통계 DB, 연금통계 DB	국세청 소득자료 (일자리 외)
		소비(카드사용 등)	→	카드사 자료(일부)	카드사 자료(전수)
		고용	→	취업활동통계등록부	-
	자살예방	인구 특성	→	인구통계등록부	-
		사망원인	→	인구동향(사망원인)	-
		상담자료	→	-	상담자료
	...	...	→	...	...
연구수요	고령자	고용	→	취업활동통계등록부	-
		주거	→	주택소유통계 DB	-
		소득	→	취업활동통계등록부 연금통계DB	국세청 소득자료 (일자리 외)
		보건·복지	→	장애인등록자료	기초생활보장자료
	소득이동	인구·가구 특성	→	인구·가구통계등록부	-
		소득·자산·부채	→	일자리행정통계 DB, 주택소유통계 DB 등	국세청 소득자료, 금융자산 자료
		교육	→	교육부 NEIS 자료	해외대학 자료
		취업	→	취업활동통계등록부	-
	...	...	→	...	...

IV. 추진전략 로드맵(안) 목표 1: 데이터 공백 발굴 및 해소

1-3. 데이터 표준 수립

추진전략

- (필요성) 내부 시스템별 항목명이 상이하고, 부처별 표준화된 서식이 없어 관리가 어려움
- (추진방안) 내부 시스템 및 범부처 서식·데이터 표준 수립 및 권고
  - ✓ (내부) 시스템별 상이한 데이터 항목명 표준화 및 메타정보 구조화
  - ✓ (외부) 부처별 표준행정서식을 권고, 신규데이터 설계 시 사전협의제도 도입
- (기대효과) 항목 추출, 자료정제, 자료연계 등 업무 효율성 제고

추진일정

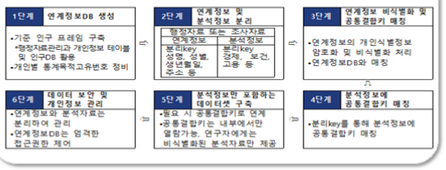
구분	내용	추진일정
단기	행정서식 항목 추출 및 메타정보 구조화, 내부시스템 항목명 표준화, 표준서식 수립, 공통연계키 설계	~'27년
중장기	표준서식 권고, 서식별 메타정보 자동생성시스템 구축, 데이터별 공통연계키 확보	~'30년

메타정보 구조화(예시)

① 서식, 데이터, 통계 정보 그룹을 구성하여 내용 작성, 이를 연결한 정보 구조 구성
② 각 그룹과 연결된 항목 정보 그룹 구성하여 내용 작성, 이를 연결한 정보 구조 구성
③ "①과 ②"를 연결하여 AI 활용에 적합한 항목 정보의 메타데이터 DB기획/설계

구분	항목명
서식정보 그룹	서식명, 소관부처, 법령명, 적용업무, 시행일 등
서식 항목정보 그룹	서식명, 항목명, 레벨 등
데이터 정보 그룹	연결된 서식명, 데이터명, 생성주체, 생성시점, 갱신주기, 갱신여부 등
데이터 항목정보 그룹	데이터명, 항목명, 레벨 등
통계 정보 그룹	연결된 데이터명, 통계명, 작성주체, 작성주기 등
통계 항목정보 그룹	통계명, 통계표명, 항목명, 활용 데이터명, 데이터 연결 여부 등

공통연계키 설계방안(안)



IV. 추진전략 로드맵(안) 목표 2: 데이터 연계·결합 기반 강화

2-1. 융합데이터 개선·개발

추진전략

- (필요성) 각종 현안에 신속히 대응하기 위해 다양한 데이터를 연계한 '융합데이터' 개선·개발로 데이터 가치 제고
- (추진방안) 융합데이터 시의성 개선 및 신규 융합데이터 구축
  - ✓ (시의성 개선) 자료 입수 시기, 정제·분석기간 단축
  - ✓ (신규 융합데이터) '26년 3종 융합데이터(주택소유자, 고령자, 사망자)를 추가 구축하고, 향후 신규 융합데이터 발굴·구축
- (기대효과) 데이터 기반 정책수립 및 평가 지원

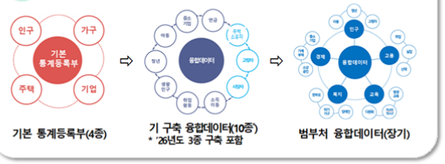
추진일정

구분	내용	추진일정
단기	신규 융합데이터 3종 구축, 융합데이터 시의성 개선	~'27년
중장기	국정과제를 뒷받침할 신규 융합데이터 발굴 및 구축으로 범부처를 아우르는 분야별 융합데이터 제공	~'30년

융합데이터 시의성 개선(예시)

구분	AS-IS	TO-BE
기업통계등록부	확정 자료 제공 16개월 소요	잠정 자료 확대 제공 12개월 소요(△5개월)
인구·가구·주택 통계등록부	확정 자료 이용 13개월 소요	내부DB 이용 11개월 소요(△2개월)
취업활동 통계등록부	연간 자료만 제공 14개월 소요	월간, 분기 확대 제공 3개월 소요(△11개월)

범부처 융합데이터 구축(안)



IV. 추진전략 로드맵(안) 목표 2: 데이터 연계·결합 기반 강화

2-2. 범용 모듈형 데이터셋 구축

추진전략

- (필요성) 자료 연계·결합에 많은 시간과 노력 소요, 기존 융합데이터는 한정된 항목 및 변수만 제공, 횡단 분석에 초점
- (추진방안) 개인, 기업 등 레벨에서 연계가 가능한 다양한 '범용 모듈형 데이터셋'을 미리 구축, 필요시 공통 결합키로 즉시 연계
  - ✓ 변수 확대, 데이터셋과 연계키 분리, 종단분석 체계 확장, 수시 업데이트
- (기대효과) 이용자의 대기시간 최소화 및 만족도 향상

추진일정

구분	내용	추진일정
단기	개별 데이터간 공통결합키 정비, 1~2개 데이터셋 시범구축, 결합키와 데이터셋 간 분리방안 마련	~'27년
중장기	범용 모듈형 데이터셋 구축, 종단패널 데이터 연계 지원	~'30년

융합데이터 - 범용 모듈형 데이터셋 비교

구분	융합데이터	범용 모듈형 데이터셋
개념	특정 주제 중심으로 여러 데이터를 미리 연계·구축해둔 완결형 데이터셋	다양한 변수 및 데이터를 필요에 따라 연계할 수 있도록 모아둔 조합형 데이터셋
연계 시점	사전 연계	필요시 연계
활용 방식	정해진 데이터 활용	원하는 변수를 직접 조합하여 분석
주제 범위	제한적 (공급자 중심)	확장 가능 (수요자 중심)
예시	아동가구, 청년, 취업활동 통계등록부	호주 통계청 PLIDA & BLADE

IV. 추진전략 로드맵(안) 목표 2: 데이터 연계·결합 기반 강화

2-3. 데이터뱅크 도입

추진전략

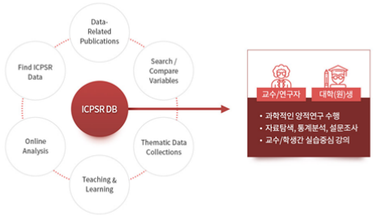
- (필요성) 가명결합 등 민간 데이터 활용 시, 결합 노력 및 소요 기간에 비해 활용도 저조(일회성 이용 후 파기 원칙)
- (추진방안) 민간 데이터와 연계된 결합데이터를 재활용할 수 있도록 데이터뱅크에 안전하게 보관하여 자산화
  - ✓ 5중 안전장치(Five Safes) 마련을 통해 보관에 대한 공감대 형성 필요
- (기대효과) 데이터 간 결합 절차 효율화 및 데이터 가치 제고

추진일정

구분	내용	추진일정
단기	결합데이터 등록제 도입, 데이터뱅크 시범 구축	~'27년
중장기	'데이터 자산 포트폴리오' 구축 및 관리, 범정부 공유·확장	~'30년

해외사례: 미국 미시건대 ICPSR

▶ 세계 최대 사회과학 조사·통계자료 아카이브
- 전세계 800여개 학술연구기관 컨소시엄으로 운영
- 논문에서 활용된 Raw data를 저장하여 재사용할 수 있도록 제공 → 검증, 통계분석, 새로운 논문 작성 등 2차 활용 가능
- 메타데이터, API까지 제공하여 재사용 촉진



IV. 추진전략 로드맵(안) 목표 3: 데이터 활용 확산 및 가치 제고

3-1. 하이브리드 방식 전환

추진전략

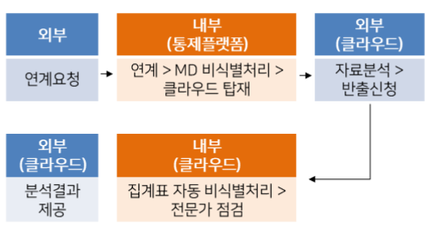
- (필요성) 기존 데이터센터 이용 시 물리적·시간적 제약 존재하여 이용자 불편
- (추진방안) 연계는 안전한 내부 통제환경에서, 분석은 클라우드에서 신속하게 수행할 수 있는 "하이브리드" 방식으로 전환
  - ✓ 5중 안전장치(Five Safes) 마련을 통해 보관에 대한 공감대 형성 필요
- (기대효과) 데이터 연계·분석 편의성 확대로 이용자 만족도 제고

추진일정

구분	내용	추진일정
단기	집계자료 비식별화 프로그램 개발, 연계·분석 기능 분리 원칙 수립, 제한적 클라우드 환경 시범 구축	~'27년
중장기	클라우드 분석 시스템 고도화, 주요 데이터셋 하이브리드 방식 활용체계 전환, 연구자 접근·허용 확대	~'30년

하이브리드 방식 데이터 활용(안)

내부 통제환경에서 연계 →
비식별처리된 MD를 클라우드에 제공 →
이용자가 원하는 대로 표두/표측 집계할 수 있도록 지원
※ MD 및 집계표를 자동으로 비식별화 처리



IV. 추진전략 로드맵(안) 목표 3: 데이터 활용 확산 및 가치 제고

3-2. 온라인 분석환경 구축

추진전략

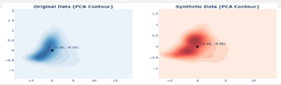
- (필요성) 활용 수요에 비해 폐쇄환경 의존으로 연계·분석 제약
- (추진방안) 동형암호 등 신기술을 통한 온라인 연계·분석환경 제공
  - ✓ **(재현자료제공)** 데이터센터 이용전, 데이터를 사전 분석할 수 있는 합성(재현)자료를 제공, 효율적인 데이터 활용 환경 구축
  - ✓ **(신기술 도입)** "원자료는 노출되지 않는(Invisible)" 연계 체계 확립
- (기대효과) 민감데이터도 안전하고 신속하게 활용하는 체계 마련

추진일정

구분	내용	추진일정
단기	재현자료 활용 시스템 구축, 신기술 방법론 연구 및 실증, 동형암호·다자간연산 시범시스템 구축	~'27년
중장기	동형암호 외부활용 플랫폼 구축을 통한 온라인 분석체계 전환, 범정부 확산	~'30년

재현자료, 동형암호 등 정보보호 신기술

▶ **재현자료:** 원자료와 합계나 평균, 분포 등 통계적 추론값은 유사하지만 가상으로 합성된 자료



▶ **동형암호:** 암호화된 데이터를 복호화하지 않은 상태에서 연산을 수행할 수 있는 암호 기술



▶ **다자간 연산:** 데이터를 조각으로 나누어 여러 서버에 분산, 분석하는 기술

IV. 추진전략 로드맵(안) 목표 3: 데이터 활용 확산 및 가치 제고

3-3. 범정부 활용 촉진

추진전략

- (필요성) 자료제공기관 대상 인센티브 부재로 협력 확보 한계
- (추진방안) 자료제공기관의 데이터 활용 지원 및 재현자료 제공
  - ✓ **(데이터 활용 지원)** 자료제공기관 요청 시, 해당 기관 데이터와 필요 데이터를 연계하여 비식별화 자료 또는 분석보고서 등 제공
  - ✓ **(AI학습데이터제공)** 대용량 데이터의 자동생성 재현자료 제공
- (기대효과) 데이터 환류를 통한 협력체계 강화 및 정책활용 확대

추진일정

구분	내용	추진일정
단기	2~3개 정책분야 대상 시범 서비스, 통계데이터센터(SDC)·내부 클라우드 내 정책분석 전용 트랙 신설	~'27년
중장기	주요 정책평가 보고서 정례화 및 확산, AI 학습을 위한 대용량 재현자료 제공	~'30년

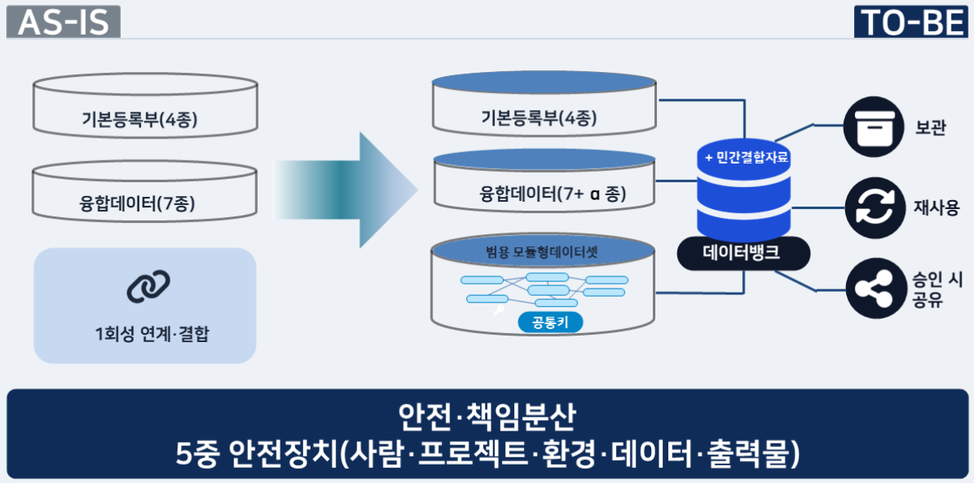
범정부 데이터 활용 촉진



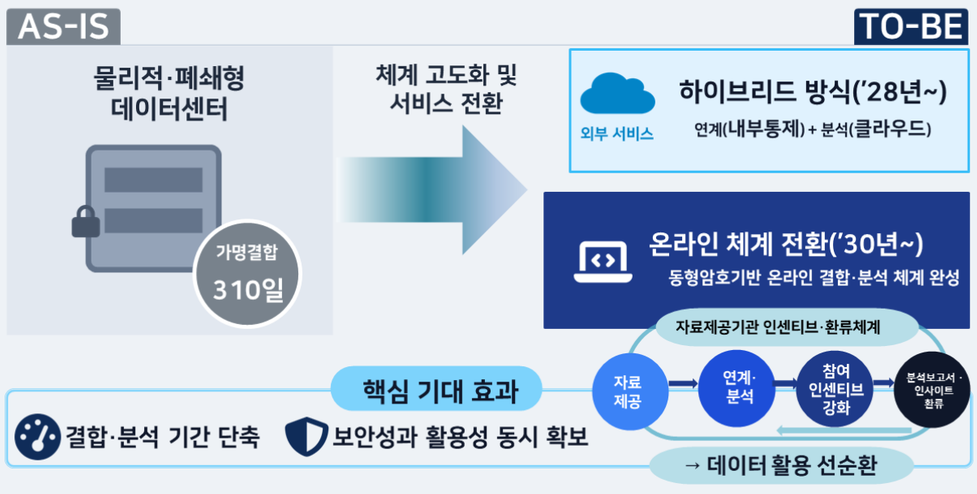
데이터 수집 관리 고도화



데이터 연계·결합 체계 고도화



데이터 활용·서비스 체계 고도화



V. 실행기반

법·제도
- 데이터 수집·연계·보관·재사용 법적 근거 마련
- 민관 협력체계 확립
- 데이터 표준 권고
- 국가데이터위원회
- 5중 안전장치 도입

시스템 구축
- 통합데이터수집포털 구축
- 융합데이터 및 범용 모듈형 데이터셋 구축
- 연계키 분리 보관
- 하이브리드 체계 도입
- 온라인 분석환경 마련

AI 활용
- 메타데이터 관리
- 데이터맵 항목 매핑
- 포트폴리오 구축
- 비식별화 프로그램, 재현자료 생성 지원
- 정책분석 지원

조직·인력
- 부서별 기능재편
- 표준화, 연계키 관리, 연계 등 전담부서 신설
- 정책지원 강화를 위한 전담인력 확보

VI. 기대효과

국가데이터 연계·활용 활성화
- 실시간 데이터 분석 및 활용
- 데이터 기반 의사결정 지원
- 국가데이터 활용 가치 제고

AI 고속도로 구축 기여
- 데이터 전 주기에 걸친 체계적 관리
- 재현자료 등 AI 학습데이터 제공 체계 구축
- 고품질 데이터의 신속한 제공

데이터 全주기에 걸친 체계적 관리로 국가 데이터 경쟁력 강화

국가데이터처
Ministry of Data and Statistics

2026 국가통계 · 데이터발전포럼

AI 시대,
통계가 인공지능과 대화하는 법

AI 친화적 메타데이터 구축을 통한 진정한 AI 전환(AX)의 비전과 전략

2026. 5. 14.

발표자 이선희
국가데이터처 국가데이터관리본부 인공지능통계혁신과

오늘 나눌 이야기

AI 시대 통계의 위기와 해결책, 그리고 미래 비전

1. 위기와 도전

- **통계 소비 패러다임의 변화**: AI 인터페이스로의 접점 이동
- **LLM 챗봇의 환각 현상**: 가짜 통계 생성 문제
- **추론의 역설**: 정확성과 똑똑함의 균형
- **API 개방의 한계**: 맥락 없는 원시 데이터
- **SDMX 표준의 구조적 한계**

2. AI 친화적 메타데이터 핵심 개념

- **온톨로지(Ontology)**: 개념 관계도 구축
- **GSIM(일반통계정보모델)**: 국제 공통 언어
- **AI 기반 자동 구축**: LLM 활용 효율화
- **메타데이터 구축 4대 원칙**

3. 4가지 혁신 활용 시나리오

- **시나리오 1**: LLM 챗봇 - 계절조정/원계열 혼란 해결
- **시나리오 2**: 가짜뉴스 검증 - AI 실시간 팩트체크
- **시나리오 3**: 데이터 자기 설명 - 자율형 데이터
- **시나리오 4**: 분석 지원 - 의미 기반 자동 매핑

4. TDO-KR와 미래 로드맵

- **신뢰할 수 있는 데이터 관측소**(TDO-KR) 구축
- **국가 데이터 전체로 확산** 계획
- **2026-2030+ 연차별 추진 로드맵**
- **글로벌 데이터망 연계** 비전

통계 소비 패러다임의 변화

과거의 공식 포털 접속에서 AI 인터페이스로의 '접점 이동'

과거
공식 포털 접속
- 국가데이터처 포털 접속
- 원하는 표 검색
- 엑셀 다운로드
- 수동 분석

현재
AI 인터페이스
- 챗봇 질의
- 일상어로 질문
- 즉시 답변
- 자동 분석

→ 국민이 통계를 만나는 **'첫 접점'**이 공식 포털에서 **AI 인터페이스**로 완전히 이동

통계 소비 패러다임의 3단계 변화

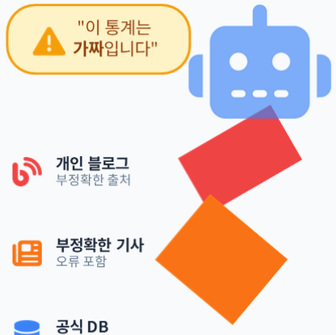
1. 정보 접근 방식의 변화
- **과거**: 포털 → 검색 → 다운로드 → 분석 (4단계)
- **현재**: 챗봇 → 질문 → 즉시 답변 (3단계)

2. 사용자 경험의 혁신
- **일상어 질의**: "요즘 경제 흐름 어때?"
- **맥락 이해**: 질문 의도 파악 및 관련 통계 추천
- **시각화**: 자동 차트/그래프 생성

3. 데이터 활용의 민주화
- **전문가 의존 → 일반인도 쉽게**
- **통계 문해력** 요구 감소
- **실시간 데이터** 접근 가능

경험을 통한 시사점: LLM 통계 챗봇과 환각

AI 통계 챗봇 도입 시도에서 발견한 환각(Hallucination) 문제와 공식 통계 기관의 신뢰성 위협



환각(Hallucination) 문제의 심각성

1. 공식 DB 대신 잘못된 출처 참조
AI가 공식 DB의 정확한 수치 대신 **개인 블로그나 부정확한 기사**를 인용하여 가짜 통계를 만들어냄

2. 환각 현상의 빈번한 발생
- **가짜 통계**: AI가 존재하지 않는 수치 생성
- **오류 인용**: 잘못된 출처에서 데이터 인용
- **맥락 왜곡**: 정확한 수치의 해석 오류

3. 공식 통계 기관의 신뢰성 위협
이는 공식 통계 기관의 **신뢰성을 위협**하는 치명적인 문제로, 정부 정책 결정에도 영향을 미칠 수 있음

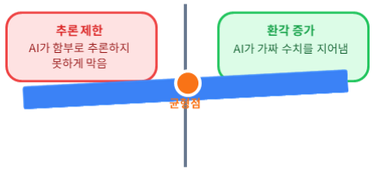
추론의 역설 (The Reasoning Paradox)

AI의 똑똑함과 정확성 사이의 균형을 찾는 딜레마

추론 제한의 문제

안전한 AI
- 환각 현상 방지
- 정확한 데이터만 사용
- 신뢰성 확보

⚠ 단점
- 똑똑함을 잃음
- 단순 검색기로 전락
- 추론 능력 제한



환각 증가의 문제

똑똑한 AI
- 자유로운 추론 가능
- 창의적 분석
- 복잡한 문제 해결

⚠ 단점
- 가짜 수치 생성
- 신뢰성 저하
- 공식 통계 왜곡

해결책 AI 친화적 메타데이터 구축을 통해 **똑똑하면서도 정확한 AI**를 만드는 것이 목표입니다.

API 개방의 한계와 AI의 '맥락맹'

숫자뿐인 원시 데이터(Raw Data)만으로는 부족한 이유

Raw Data vs Context

Raw Data
- 숫자만 존재
- 단위 정보 없음
- 분모 기준 불명
- 조사 대상 미표기
- 맥락 정보 부재

VS

Context
- 단위/기준 명시
- 분모 정의 포함
- 조사 대상 표기
- 맥락 정보 완비
- AI 이해 가능

AI는 숫자 이면의 맥락을 읽지 못함

AI의 '맥락맹' (Context Blindness)

문제: AI가 숫자의 의미를 읽지 못함

AI는 **숫자 이면에 숨겨진 단위, 분모의 기준, 조사 대상** 등 '맥락'을 읽지 못하는 일종의 '맥락맹'에 걸려 있습니다. 예를 들어 "실업률 5%"라는 숫자만으로는 **15세 이상 기준인지, 15-64세 기준인지** 알 수 없습니다.

원인: API 개방의 한계
- **원시 데이터만 제공**: 숫자만 있는 API는 AI에게 의미 없음
- **메타데이터 부재**: 변수의 의미와 관계가 명시되지 않음
- **비교 불가능**: 서로 다른 기준의 데이터를 구분 못함

해결책: AI 친화적 메타데이터
- **의미의 지도**를 구축하여 AI가 숫자의 맥락을 이해
- **온톨로지**를 통해 개념 관계를 명시
- **GSIM 표준**으로 통계 생산부터 서비스까지 일관성 확보

기존 통계 표준(SDMX)의 구조적 한계

데이터셋 중심 교환 vs AI가 요구하는 변수 단위 의미 해석

SDMX의 강점

데이터셋 중심 구조
- **표준화된 데이터 교환**: XML/JSON 형식
- **구조적 메타데이터**: DSD, 코드 리스트
- **국제 기관 협력**: BIS, ECB, IMF, OECD 등

데이터셋 단위 교환

대규모 데이터셋을 효율적으로 교환하는 데 최적화되어 있으며, 국제 통계 기관 간의 데이터 공유를 표준화했습니다.

구조적 정합성

데이터 구조 정의(DSD), 코드 리스트, 개념 체계를 명확히 정의하여 데이터 일관성을 확보합니다.

광범위한 채택

전 세계 통계 기관에서 널리 사용되며, 국제 통계 교환의 사실상 표준으로 자리잡았습니다.

AI 관점의 한계

변수 단위 의미 부재
- **데이터셋 단위**로 포장되어 **개별 변수**의 의미 파악 불가
- AI가 상자 안의 내용을 꺼내어 읽고 비교/추론하기 어려움
- **맥락 정보**가 부족하여 정확한 해석 불가

✕ 구조적 한계

데이터를 '데이터셋'이라는 큰 상자 단위로 포장하여, AI가 **개별 변수**들의 의미를 꺼내어 읽기 어려움

✕ 의미 해석 불가

비교 가능성을 판단하기 위한 분모 정의, 조사 대상 등 **맥락 정보** 부족

✕ 자동 추론 불가

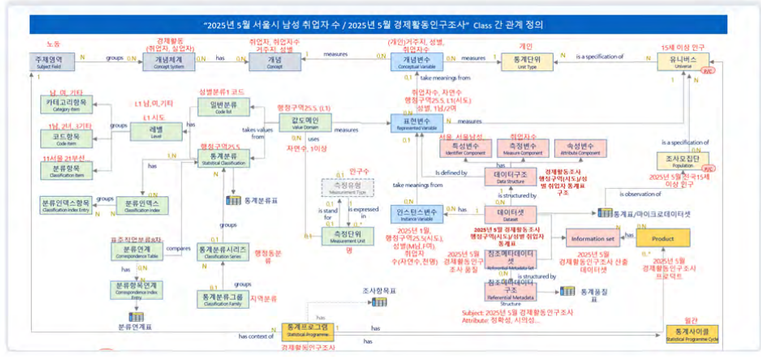
AI가 **자동으로** 변수 간 관계를 판단하고 변환하는 **추론 능력** 부족

핵심 인사이트: SDMX는 데이터 교환의 효율성을 높였지만, AI가 개별 변수의 의미를 이해하고 추론하기에는 **구조적인 한계**가 존재합니다.

AI 친화적 통계 메타데이터의 탄생

숫자 이면에 담긴 의미를 기계가 스스로 읽고 이해할 수 있는 '의미의 지도'

기존
- 데이터셋
- 파일 단위
- 수동 검색
- 제한된 의미
- 개별 접근



AI
- 의미망
- 변수 단위
- 자동 연결
- 맥락 이해
- 지능형 검색

의미의 지도 (Meaning Map)

숫자 이면에 담긴 의미, 비교 가능성, 변환 규칙을 기계가 스스로 읽고 이해하도록 명시

의미 연결 | 비교 가능 | 변환 규칙

핵심 개념 1: 온톨로지(Ontology)란?

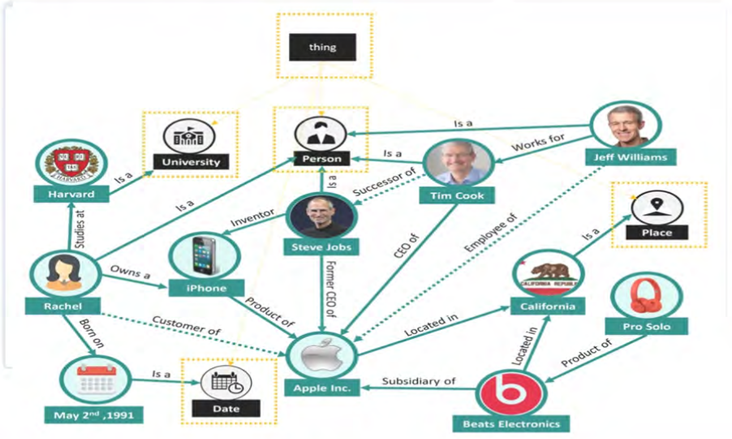
개념들의 관계도: '청년실업률'의 의미를 거미줄처럼 엮어서 이해하는 구조

온톨로지란?

정의

개념들의 관계도.

AI가 단어를 이해하도록 개념 간의 연결을 명시적으로 정의한 구조



온톨로지 구축의 딜레마와 AI 기반 해결책

LLM을 활용한 자동 구축으로 수만 개 통계 용어의 개념 관계도를 효율적으로 구축

구축 딜레마

수작업의 한계

- 수만 개의 통계 용어
- 일일이 관계도 작성
- 엄청난 시간/비용 소요

AI 해결책

- LLM이 초안 자동 생성
- 전문가 검토 및 승인
- 구축 속도/정확성 향상

AI 기반 온톨로지 구축 프로세스

1. 문서 읽기 — 조사 지침서 용어집 분석
2. 관계 추출 — 개념 관계도 초안 자동 생성
3. 전문가 검토 — 관계도 검증 정확성 확인
4. 승인/적용 — 최종 승인 시스템 적용
5. 지속 관리 — 지속적 업데이트 버전 관리
6. 최적화 — 성능 개선 정확도 향상
7. 완료 — 온톨로지 구축 완료

구축 속도 기존 대비 10배 향상

정확성 전문가 검증 후 승인

비용 절감 인력 소요 70% 감소

지속 관리 자동 업데이트

핵심 개념 2: GSIM(일반통계정보모델)

통계 생산부터 서비스까지의 모든 정보를 규격화한 국제 공통 언어

설계/생산
통계 설계
데이터 수집
가공/분석

GSIM (일반통계정보모델)
통계 생산 → 서비스까지
모든 정보를 규격화한 국제 공통 언어

서비스/배포
데이터 저장
API 제공
챗봇/포털

GSIM의 장점

생산 단계
- 통계 설계 및 데이터 수집
- 가공/분석 프로세스

GSIM 표준
- 통일된 규격화
- 공통 언어 정의

서비스 단계
- 데이터 저장/배포
- API/챗봇 제공

통일된 조립 설명서
포털/챗봇/보고서에서 동일한 기준으로 통계 조립

일관성 확보
생산부터 서비스까지 전 과정의 정보 규격화

국제 호환성
OECD, UN 등 국제 기관과의 연계 가능

메타데이터 구축 4대 원칙

AI 친화적 통계 메타데이터를 위한 체계적 구축 기준

01 공통 레지스트리

부처 칸막이를 넘는 공통 레지스트리
데이터셋별 메타데이터에서 **기관 전체 공통 레지스트리**로 전환하여, AI가 관련 변수를 쉽게 찾을 수 있도록 합니다.

공유 정의로 AI가 변수 탐색 가능
"실업률"과 같은 개념을 **기관 전체**에서 일관되게 정의하여 AI가 자동 탐색합니다.

02 변수 단위 표준화

데이터셋이 아닌 개별 변수 단위
데이터셋 레벨이 아닌 **변수 레벨**에서 의미를 정의하여 AI가 세부 내용을 파악합니다.

분모 정의, 조사 대상 등 명확화
단위, 분모, 조사 대상 등 세부 정보를 명시하여 AI가 정확히 해석합니다.

03 온톨로지 개념망

다차원 개념망 구축
SKOS/XKOS를 활용하여 **다차원 개념 관계망**을 구축하여 AI가 복잡한 관계를 추론합니다.

개념 간 관계 명시
"청년실업률" = "청년"(계층) + "실업"(상태) + "경제활동인구"(분모) 관계를 명시합니다.

04 GSIM 정합성

GSIM 표준 철저 준수
일반통계정보모델(GSIM)을 따르는 메타데이터 구조로 통일성을 확보합니다.

국제 표준과의 호환성
국제 공통 언어로 통계 생산부터 서비스까지 일관성을 유지합니다.

[시나리오 1] 계절조정 vs 원계열의 혼란

AI가 원계열(-2.1%)과 계절조정계열(+0.5%)을 구분하지 못해 상반된 해석이 나오는 문제

문제 상황

원계열	계절조정계열
-2.1%	+0.5%
전년 동월 대비	전월 대비

AI가 **원계열**을 보고 "경제 침체"라고 답할지, **계절조정계열**을 보고 "경제 회복"이라고 답할지 몰라 **혼란에 빠집니다.**

"요즘 경제 흐름 어때?"
사용자가 **일상어로 질문**하면, AI는 **어떤 지표를 선택**해야 할지 몰라 **상반된 해석**을 내놓습니다.

해결 필요
AI가 **질문 의도**를 파악하고 **적절한 지표**를 선택하도록 **메타데이터 규칙**을 명시해야 합니다.

메타데이터 기반 해결

1. 명시적인 메타데이터 규칙
메타데이터 안에 **"단기 흐름 분석에는 계절조정계열 사용을 권장함"**이라는 규칙을 명시합니다.

2. AI 선택 로직
AI는 사용자의 **질문 의도**를 파악하고, 메타데이터 규칙(GSIM의 Concept/Structure 그룹)을 읽어 **정확한 지표**를 선택합니다.

3. 정확한 답변
"최근 경제는 회복세"라고 답변하며, **계절조정계열** 수치를 바탕으로 **정확한 분석**을 제공합니다.

[시나리오 1] 메타데이터 기반의 똑똑한 답변

AI가 메타데이터 규칙을 읽어 정확한 계열을 선택하는 해결 과정

메타데이터 규칙

1. 계열 선택 규칙
"단기 흐름 분석에는 계절조정계열 사용을 권장함"
- 원계열: 전년 동월 대비 (YoY) 분석에 적합
- 계절조정계열: 전월 대비 (MoM) 분석에 적합

2. AI 선택 로직
- 사용자 질문 의도 파악
- 메타데이터 규칙(GSIM) 적용
- 적절한 계열 자동 선택

3. 결과 검증
- 선택된 계열의 적절성 확인
- 결과의 일관성 검증

AI 답변 프로세스

1. 사용자 질문
"요즘 경제 흐름 어때?"

2. 의도 파악
단기 흐름 분석 필요성 인식

3. 계열 선택
계절조정계열 적용

최종 답변
"최근 경제는 회복세"라고 답변하며, **계절조정계열** 수치를 바탕으로 **정확한 분석**을 제공합니다.

[시나리오 2] 끝없는 수동 모니터링의 한계

가짜뉴스 대응을 위한 담당자의 고된 수작업과 그 한계

현재의 수동 모니터링 현황

평일 모니터링 (6:00 ~ 20:00)
대변인실: **매 1시간 간격** 실시 (14회/일)
실사과: 하루 **3회 이상** (9시, 13시, 18시) 키워드 검색

주말/공휴일 모니터링
대변인실 및 실사과: 하루 **3회 이상** (8시, 13시, 18시)
3대 동향 및 가계수지동향과, 복지과, 인구동향과는 필수

늦은 밤/주말 대응
최근 **늦은 밤(9시 이후)** 및 **주말**에도 대응하는 추세
담당자가 **밤낮없이** 모니터링해야 하는 상황

핵심 문제: 끝없는 수작업으로 인한 **업무 부담**과 **지속 가능성** 문제

담당자의 고충

끝없는 키워드 검색
부서 관련 **키워드로 촘촘하게 검색**하고, **수치가 맞는지 직접 확인**해야 하는 **수작업**

수치 확인의 번거로움
기사 속 **통계 수치의 오류, 통계 해석 왜곡 여부**를 **일일이 검증**해야 함

주말/야간 대응 부담
주말이나 늦은 밤에도 대응해야 하는, **그야말로 끝이 없는** 고된 수작업

피로 누적: 지속적인 모니터링으로 인한 **업무 효율성 저하**

시나리오 2 해결책: AI 실시간 자동 팩트체크 시스템

기사를 실시간으로 모니터링하고 통계적 주장을 자동으로 검증하는 지능형 시스템

자동화된 검증

실시간 모니터링
- 24/7 기사 감시
- 통계 주장 추출
- 자동 대조

즉각 대응
- 왜곡 즉각 판별
- 반박 논리 초안
- 담당자 알림

AI 자동 팩트체크 프로세스

1. 기사 실시간 모니터링 — 24/7 기사 감시
2. 통계 주장 추출 — AI가 수치 추출
3. 공식 메타데이터 대조 — 온톨로지망 연결
4. 왜곡 즉각 판별 — 부정확성 감지
5. 반박 논리 초안 작성 — 자동 반박자료 생성
6. 담당자 알림 — 실시간 알림 전송
7. 공식 대응 — 해명자료 배포

실시간 감시	정확도 99%+	비용 절감	자동화
24/7 자동 모니터링	AI 정확한 판별	인력 소요 80% 감소	완전 자동화 프로세스

3. AI 친화적 국가 메타데이터 구축 방안









[시나리오 4] 이 데이터들, 마음대로 섞어도 될까요?

서로 다른 기관의 데이터셋을 합치려 할 때 발생하는 치명적 분석 오류 위험

데이터셋 합치기 시도

고용부 데이터	통계청 데이터
15세 이상	15-64세
모집단 기준	모집단 기준

모집단 기준 불일치: 연구자가 두 데이터를 마음대로 섞어 분석하려 할 때 발생

모집단 기준 불일치

A 지표는 **15세 이상** 기준이고, B 지표는 **15~64세** 기준이라 **직접 비교하면 안 됩니다**.

시점 차이

데이터 수집 **시점이 다르거나, 조사 주기가 일치하지 않아** 분석 오류가 발생합니다.

해결 필요

AI가 **의미 기반 자동 매핑**을 통해 코드와 단위를 자동으로 통일해줘야 합니다.

AI 데이터 사이언티스트

1. 의미 기반 자동 매핑

분석가가 두 데이터를 올리면 시스템이 **메타데이터를 즉각 분석**하여, 서로 다른 기관에서 쓰던 **코드나 단위를 자동으로 통일**해 줍니다.

2. 유의사항 분석

AI는 **"A 지표는 15세 이상 기준이고, B 지표는 15~64세 기준이라 직접 비교하면 안 됩니다"**라고 경고합니다.

3. 전문가적 대안 제시

단순히 안 된다고 끝나는 것이 아니라, **"대신 X 비율의 가중치를 주어 보정하거나, Y라는 대체 지표를 사용하십시오"**라며 데이터 전문가 수준의 깊이 있는 분석 대안을 제시합니다.

[시나리오 4] 의미 기반 자동 매핑

서로 다른 기관의 데이터셋을 자동으로 통합하는 전처리 자동화 프로세스

통합 과제

데이터 불일치

- 서로 다른 기관 코드
- 단위/시점 불일치
- 모집단 기준 차이

자동화 해결

- 의미 기반 매핑
- 코드/단위 통일
- 전처리 자동화

의미 기반 자동 매핑 프로세스

1. 데이터셋 업로드
고용부/국가데이터처 데이터 업로드

→

2. 메타데이터 분석
구조/단위/시점 즉각 분석

→

3. 코드/단위 통일
의미 기반 매핑 자동 변환

4. 통합 데이터셋 생성
의미 기반 매핑 완료 → 코드/단위 자동 통일 → 통합 데이터셋 생성 → 분석 준비 완료

처리 속도	정확성	비용절감	자동화
기존 대비 90% 감소	자동 매핑 99.5%	전처리 시간 95% 감소	수동 작업 불필요

[시나리오 4] 전문가 수준의 분석 가이던스

AI가 데이터 분석가들에게 전문가 수준의 분석 대안과 가이던스를 제공합니다

문제 상황: 직접 비교 불가

A 지표		B 지표
15세+	→ 직접 비교 불가	15-64세
전국민 기준		경제활동인구

모집단 기준이 달라 직접 비교하면 **치명적인 분석 오류**가 발생합니다.

"이 데이터들, 마음대로 섞어도 될까요?"

연구자들이 **고용부 데이터**와 **통계청 데이터**를 마음대로 섞어 분석하려 할 때, **모집단의 기준이나 시점이 달라** 치명적인 분석 오류가 발생할 수 있습니다.

해결 필요

AI가 **유의사항을 분석**하고 **전문가적 대안**을 제시해야 합니다.

전문가 수준의 분석 가이던스

1. 유의사항 분석

AI는 **"A 지표는 15세 이상 기준이고, B 지표는 15~64세 기준이라 직접 비교하면 안 됩니다"**라고 **즉각 경고**합니다.

2. 가중치 보정 방법

"대신 X 비율의 가중치를 주어 보정하거나"라며 **구체적인 보정 방법**을 제시합니다.

3. 대체 지표 추천

"Y라는 대체 지표를 사용하십시오"라며 **데이터 전문가 수준의 깊이 있는 분석 대안**을 제시합니다.

인프라의 완성: 신뢰할 수 있는 데이터 관측소 (TDO-KR)

TDO-KR 인프라 구조: 데이터 수집 → 처리 → 검증 → 공급의 4단계 파이프라인

데이터 수집	데이터 처리	AI 통합	데이터 공급
LOD (Linked Open Data)	AI 분석 엔진	RAG Adapter	Trusted Portal
SPARQL Endpoint	메타데이터 추출	LLM 연동	API Gateway
RDF/TTL 형식	품질 검증	추론 엔진	데이터 검증

데이터 흐름: 수집 → 처리 → AI 분석 → 검증 → 공급 (실시간)

수집 (원시 데이터) → 처리 (메타데이터 추출) → AI 분석 (의미 추론) → 공급 (신뢰 데이터)

확산 계획: 국가 통계를 넘어 국가 데이터 전체로

국가데이터처를 넘어 범정부 행정 데이터, 나아가 핵심 민간 데이터까지 지식 그래프로 연결



국가 데이터 확산 비전

1. 범정부 행정 데이터 연결
- **국가데이터처**를 중심으로 한 **행정 데이터** 통합
- **30개 부처**의 데이터를 지식 그래프로 연결
- **실시간 데이터** 교환 및 공유

2. 핵심 민간 데이터 확장
- **금융, 의료, 교육** 등 핵심 민간 데이터
- **기업** 및 **기관** 데이터 연계
- **데이터 품질** 표준화

3. 국가 데이터 인프라 구축
- **TDO-KR**을 통한 신뢰성 있는 데이터 제공
- **AI 친화적 메타데이터** 표준화
- **글로벌 데이터** 연계

로드맵: 2026-2030+ 연차별 추진 계획

AI 친화적 통계 메타데이터 구축을 위한 3단계 타임라인

3단계 로드맵: 기반 구축 → 전국 확산 → 글로벌 연계 — 2026-2030+

단계 1 2026-2028: 기반 구축
- **핵심 지표 메타데이터 시범 구축**
- 고용, 물가(CPI) 등 핵심 지표
- TDO-KR 베타 서비스 개시
- GSIM 정합성 확보

단계 2 2029: 전 기관 확산
- **전 부처 확대 적용**
- 전 부처 메타데이터 구축
- 자동 비교 가능성 판단
- 챗봇/AI 시스템 연동

단계 3 2030+: 글로벌 연계
- **글로벌 데이터망 연계**
- OECD, UN 등 국제 기관 연계
- 크로스보더 지식망 완성
- 글로벌 표준화 참여

주요 목표 및 성과

2026-2030+

메타데이터 구축 (핵심 지표 100% 커버리지) → 전국 확산 (전 부처 연계 완료) → 글로벌 연계 (국제 표준화 참여) → 완성 (글로벌 리더십)

요약: 데이터 개방을 넘어 '의미 개방'의 시대로

Open Data → Open Meaning으로의 패러다임 전환

"파일을 엑셀로 다운로드하고 API를 열어주는 1차원적 '데이터 개방(Open Data)'의 시대는 저물고 있습니다"

Open Data
- **1차원적 개방**
- 파일 다운로드
- API 개방
- 원시 데이터 제공

Open Information
- **2차원적 개방**
- 메타데이터 제공
- 의미 정보 포함
- 맥락 설명

Open Meaning
- **3차원적 개방**
- 의미 개방
- AI 추론 가능
- 자율 조합

변수 단위 의미 표준화
데이터셋이 아닌 개별 '변수' 단위로 의미를 쪼개어 AI가 정확히 해석

온톨로지 기반 개념망
다차원 개념 관계망을 구축하여 AI가 복잡한 관계를 추론

AI 친화적 메타데이터
기계가 스스로 읽고 이해할 수 있는 '의미의 지도'

Q&A 및 마무리

질문이 있으신 분들은 편하게 말씀해 주십시오

AI 친화적 통계 메타데이터는 단순한 기술적 도구가 아닙니다

국민과 국가 데이터가 가장 **안전하고, 정확하고, 똑똑하게** 만날 수 있도록 든든하게 받쳐주는 **미래의 다리**입니다

신뢰성 · 정확성 · 지능성

진정한 AX의 시대를 열어갈 국가데이터처의 여정에 많은 관심 부탁드립니다

Session 02

경제를 읽고, 내일을 설계하다 : 경제총조사

좌장

전현배 교수(서강대학교 경제학과)

1. **경제총조사의 의의 및 추진방향**
   김혜련 과장(국가데이터처 경제총조사과)
2. **국가통계의 변천과 경제총조사 데이터의 학술적 정책적 가치**
   홍석철 교수(서울대학교 경제학부)
3. **국가데이터처 미시경제데이터의 정책의사결정 활용도 제고 방향**
   장우현 소장(한국조세재정연구원 국가회계재정통계센터)
4. **일본의 경제총조사 추진전략 및 방향**
   Nanami Suzuki **연구관**(일본 총무성 통계국)

토론

김덕파 교수(고려대학교 경제학과), **정현준 연구위원**(정보통신정책연구원)

강영관 팀장(한국은행 국민계정기획팀), **유민희 연구위원**(한국경제인협회 미래전략팀)

국가데이터처

2025년 기준

경제총조사
의의 및 추진방향

- 제15회 국가 통계·데이터 발전 포럼 -

Census 2025 경제총조사
https://ecensus.go.kr/

DATE. 2026.5.14.(목)
국가데이터처 경제총조사과

CONTENTS

목차안내

1 **경제총조사 개요**
경제총조사의 목적과 조사 대상, 국가 통계로서의 역할과 주요 활동 분야 등 조사 전반의 기본 내용

2 **추진 경과 및 방안**
준비 과정과 시험·시범조사 결과를 바탕으로 조사 환경 변화에 대응하기 위한 추진 전략과 개선방향

3 **중점 추진 과제**
신산업 구조 반영, 조사 방식 개선, 행정자료 활용 확대 등 2025년 조사에서 중점적으로 추진할 핵심 과제

4 **실시개요 및 주요 추진 일정**
조사 내상과 방법, 조사 기간 및 추진 체계 등 실제 조사 수행을 위한 운영 계획과 향후 추진 일정 안내

Chapter 01

경제총조사 개요

경제총조사의 목적과 조사 대상, 국가 통계로서의 역할과 주요 활동 분야 등 조사 전반의 기본 내용

개념 의의

03

Census 2025 경제총조사

경제총조사 개념 및 의의

✔ **개념**

대한민국 전체 산업의 고용, 생산 등 구조 파악을 위해 **모든 경제단위(사업체)**를 대상으로 **5년마다** 실시하는 대규모 조사

✔ **의의**

1. **국가 산업구조**의 이해
2. **고품질 경제통계** 기반 확보
3. **효율적 정책수립** 자료 제공
4. **지역간 균형발전** 정책 기여
5. 경제변화 **추이 파악** 및 **미래예측** 가능

04

Census 경제총조사

경제총조사 활용 분야

1. 정부정책 수립·평가
   - 대·중소기업 및 소상공인 지원 정책
   - 여성기업 육성·지원, 지식재산 집약 산업 육성 등
2. 주요 가공통계 작성
   - 국민소득, 지역소득 및 산업연관표 작성
   - 시장집중도 분석, 사업다각화지수 등
3. 표본조사의 모집단
   - 사업체노동력조사, 환경산업통계조사 등의 모집단
   - 광공업생산 등의 기준년도 품목 및 가중치 산정
4. 연구 및 분석 자료
   - 대학·연구기관의 연구 시 기초자료
   - 민간 기업의 경영실태 파악 등

05

Census 경제총조사

대표적 활용 사례

❖ 2020년 기준 경제총조사 결과를 활용한 GDP 개편

"기준연도 개편만으로 경제지표 일제히 개선... 1인당 소득 日 추월"

'경제총조사의 경우 과거에는 조사 기반이었지만 등록 기반으로 바뀌어 소규모 사업자들이 많이 포착됐다. 이에 따라 기준연도의 GDP 레벨이 올라갔고' (연합뉴스, '24.6.5)

연합뉴스

"GDP 기준년 바꾸자 가계부채비율 100.4%→93.5%로 하락"

'모수인 경제규모가 확대되면서 기계와 정부의 부채비율 등은 하락' (파이낸셜, ' 24.6.5)

파이낸셜뉴스

06

Census 경제총조사

산업생태계 변화 측정 (2020년 ~2024년)

❖ 우리나라 사업체수, 종사자수 및 매출액 추이



* 2021년~2024년은 전국사업체조사 결과

07

Census 경제총조사

산업생태계 변화 측정 (2020년 ~2024년)

❖ 디지털 플랫폼 거래 추이



* 2021년~2024년은 전국사업체조사 결과

08

Census 경제총조사

산업생태계 변화 측정 (2020년 ~2024년)

❖ 무인 결제 기기 도입 추이

⑯ 무인 결제 기기(키오스크) 도입 여부	• 무인 결제 기기 : 키오스크(무인 정보 단말기) 등 해사로 자동 정산 및 결제 서비스를 제공
① 도입하고 있음 →	ⓘ 무인 결제 기기(키오스크) 대수 ☐☐☐ 대
② 도입하지 않음	



* 2021년~2024년은 전국사업체조사 결과

09

Census 경제총조사

그간 추진 사항 ('23~'25년)

의견 수렴	연구용역 실시	시스템 구축	사전조사 실시
✓ 국가데이터처 및 관련기관 의견수렴	✓ 경제총조사 개선을 위한 연구	✓ 경제총조사 통합시스템 구축 정보전략계획(ISP) 실시	✓ 1차 시험조사 (인천 미추홀구2개동, 전북 전주시2개동)
✓ 경제총조사 및 전국사업체조사 이용자그룹협의체 회의	✓ 경제총조사 조사효율성 증진을 위한 심층 연구	✓ 경제총조사 통합시스템 구축	✓ 2차 시험조사 (경남 밀양시, 함안군)
✓ 경제총조사 및 전국사업체조사 업무발전토론회			✓ 시범예행조사 (부산 북구, 경기 안성시, 경북 경산시)

011

Chapter 02

추진경과 및 방안

준비 과정과 시험·시범조사 결과를 바탕으로 조사 환경
변화에 대응하기 위한 추진 전략과 개선방향

추진 경과 | 환경 분석 | 추진 전략

010

Census 경제총조사

환경분석

새로운 산업 형태를 파악하기 위한 정책 부처 등의 **세부 자료 제공 요청 증가**	✓ 시의성 있는 조사항목 개발
조사대상 사업체의 **응답부담 경감** 및 **통계 정확도 향상 필요**	✓ 행정자료 활용 확대
응답거부, 인구감소 및 고령화로 인한 조사원 확보 곤란 등 **대면 조사환경 악화**	
조사대상 사업체수의 지속적 증가 에도 불구하고, 예산당국의 **조사비용 절감 요구 증대**	✓ 온라인조사 확대

12

Census 경제총조사

추진 전략 및 과제

목표 국민이 필요로 하는 경제총조사의 성공적 수행

추진방향 경제 대전환 시대 반영 및 조사 방식의 혁신

추진 전략

산업생태계 변화 반영	조사 방법 효율화	국민 응답 부담 경감	고품질 통계 제공
❶ 신산업 구조 반영 신규 항목 발굴	❸ 온라인 조사 확대	❻ 등록센서스율 제고	❽ 차세대 통합시스템 구축
❷ 최신 산업 분류 적용	❹ 응답자 친화적 조사표 설계	❼ 현장조사 규모 축소	❾ 인공지능 기술 도입
	❺ 현장조사 환경 개선		

13

Census 경제총조사

조사방식의 획기적 전환 및 온라인조사 확대

❶ **온라인조사 확대**

- 온라인조사 비율을 **기존 8.1%에서 '26년 10%로 확대**
- 온라인조사 대상처를 **대면조사 대상에서제외**하고 조사원 업무에 **미배정**

* 온라인조사율: ('20년기준) 8.1% → ('25년기준) 10.0% → ('30년기준) 30.0%

❷ **신규인력 도입**

- 온라인조사를 전담하는 **스마트조사 관리자(257명)**
- **프로파일링 전담인력**

❸ **온라인 조사 홍보 강화**

- **全 현장조사 사업체 대상 협조공문 확대 발송**
- 기업 CEO 대상 협의체 구성 등 온라인 조사 독려를 위한 대국민 **홍보 강화**

추진 성과 조사원이 방문하는 **현장조사 대상 사업체를 축소**하여 예산 절감

15

Chapter 03

경제총조사 중점추진과제

신산업 구조 반영, 조사 방식 개선, 행정자료 활용 확대 등
2025년 조사에서 중점적으로 추진할 핵심 과제

014

Census 경제총조사

신 산업 구조 반영 및 정부정책 지원을 위한 신규 항목 발굴

❶ **산업생태계 변화 반영**

- **인공지능(AI), 스마트공장** 등 신기술 활용에 대한 실태 및 **외국인 종사자**를 파악하여 **산업생태계 변화 반영** 및 **혁신 경제정책 지원**

❷ **국정과제 추진 기초자료 제공**

- 인공지능 대전환, 중소기업 경쟁력 강화, 상생의 기업환경 조성, 농업육성, 노동 대전환 등 **국정과제 추진**에 필요한 **기초자료 제공**

2025년 기준 신규 항목 (6개)

❶ 외국인 종사자 ❷ 스마트농장 운영 ❸ 로봇 활용
❹ 인공지능(AI) 활용 ❺ 스마트공장 운영 ❻ 무인매장 운영

추진 성과 국민이 필요로 하는 **수요자 맞춤형 통계** 자료 제공

16

Census 2025 경제총조사

[참고] 2025년 기준 경제총조사를 통한 국정과제 지원

조사항목	지원내용
AI 활용 사업체 현황(사업체 수, 종사자 수, 산업분야 등) 및 활용 분야	**과제 21: 인공지능 대전환(AX)** (과학기술정보통신부) - 인공지능 전환 및 혁신 정책의 기초자료
스마트공장 운영 현황 (사업체 수, 종사자 수 등)	**과제 35: 중소기업 경쟁력 강화** (중소벤처기업부) - 중소기업의 스마트공장 육성 정책의 기초자료
사업체의 디지털 플랫폼 거래 여부	**과제 67: 상생의 기업환경 조성** (중소벤처기업부) - 온라인 플랫폼 상생협력 제도화를 위한 실태 파악
스마트농장 운영 현황 (사업체 수, 종사자 수 등)	**과제 68: 국가전략 산업으로 농업 육성** (농림축산식품부) - 스마트농업 고도화를 위한 기초자료
외국인 종사자 규모 및 분포	**과제 92: 인구변동 등에 대응하는 노동 대전환** (고용노동부) - 외국인력(이주노동자 등) 수급 설계 등의 기초자료

17

Census 2025 경제총조사

AI 기술 도입 및 활용

1 콜센터 도입

- 보이스봇(챗봇) 등 **AI 기반* 콜센터**를 도입하여 응답 정확성 제고

* 음성인식, 자연어 처리 등 AI를 활용하여 24시간 상담 및 상담내용 자동 TEXT 저장

2 산업분류 자동코딩

- **AI 기반 산업분류 자동코딩 적용**을 통해 업무량 감소 및 분류 오류 예방

3 홍보방식 변화

- **AI 기술 활용 홍보방식을 실시간 분석**하여 홍보 우선순위 결정, 일정 조정 등 효율적 홍보 추진

추진 성과 최신 기술을 활용한 **응답 편의성 및 자료 정확성 제고**

19

Census 2025 경제총조사

행정자료 활용 확대 및 현장조사 규모 축소

1 조사표에서 삭제

- **기업체 유·무형자산 항목**을 법인세 자료국세청로 전면 대체하여 조사표에서 삭제

2 행정자료 대체 확대

- 고용허가제 자료고용부 활용 (외국인 종사자) 등 **행정자료 항목 대체 확대**
- 공동주택관리정보시스템한국부동산원 등 **신규 행정자료 발굴**

3 현장조사 규모 축소

- 1인 유튜버 등 개인사업체와 같은 **행정자료 대체 사업체를 확대**하여 **현장조사 규모 축소**

추진 성과 사업체의 직접 응답 **조사 항목을 축소**하여 응답부담 대폭 감소

18

Census 2025 경제총조사

조사환경 개선

1 맞춤형 교육 실시

- 다양한 사례와 온라인조사 입력 시스템을 활용한 **실습 위주의 조사원 맞춤형 교육 실시**

2 조사원 처우 개선

- 조사원1일 6천원, 조사관리자50,000원 교통보조비 신설 등 **조사요원 처우 개선**

3 안전대책 강화

- 안전 교육 강화, 안전용품 지급(녹음 기능 탑재 신분증 케이스, 투넘버 서비스 가입 지원 등) 등 **조사원 안전대책 강화**

추진 성과 **조사원들의 만족도를 높여** 조사 결과의 품질 제고

20

Chapter 04

실시개요 및 주요 추진 일정

조사 대상과 방법, 조사 기간 및 추진 체계 등 실제 조사 수행을 위한 운영 계획과 향후 추진 일정 안내

21

Census 2025 경제총조사

2025년 기준 경제총조사 개요

✔ 조사대상	조사 기준일 현재('25. 12. 31.), **산업활동을 수행**하고 있는 모든 사업체 * 약 750만개 사업체 중 현장조사 약 330만개
✔ 조사표/조사항목	조사표 **7종**업종별 6종, 본사 1종 / 조사항목 **38개**공통항목 12개, 특성항목 26개
✔ 조사방법	**면접조사** 원칙, **스마트조사**온라인(PC, 모바일), 전화, 이메일, 팩스 **병행**
✔ 조사기간	**'26. 6. 12. ~ 7. 22.** * 온라인조사 6. 1. ~ 6. 30.
✔ 조사체계	**국가데이터처**(주관), **지방자치단체**(실시)
✔ 조사인력/예산	조사요원 **약 1만 6천명** / 예산 **약 659억원**
✔ 결과공표	**잠정결과** 공표('26.12월) / **확정결과** 공표('27.6월)

22

Census 2025 경제총조사

2025년 기준 경제총조사 조사항목 (공통항목, 12개)

1 사업체명	2 사업체 대표자 (성명, 성별, 연령)	3 소재지
4 창설연월	5 사업자등록번호	6 조직형태
7 사업의 종류	8 영업기간(조업기간)	9 인공지능(AI) 활용신규
10 총 종사자 수 및 연간 급여액	11 외국인 종사자신규	12 사업실적 (매출액, 영업비용)

23

Census 2025 경제총조사

2025년 기준 경제총조사 조사항목 (특성항목, 26개)

1 스마트농장(양식장) 운영신규	2 스마트공장 운영신규	3 연간 제품별 출하액	4 연간 품목별 임가공 수입액
5 재고액	6 유형자산(사업체)	7 온라인 거래변경	8 매출형태별 수입액 (55101, 55103)
9 로봇 활용신규	10 보조금(751)	11 일일평균 영업시간	12 사업체 건물(사업장) 연면적
13 디지털 플랫폼 거래	14 상품 매입처별 구입액 구성비	15 상품 판매처별 매출액 구성비	16 상품 판매유형별 매출액 구성비
17 배달(택배) 판매	18 무인 결제 기기 활용변경	19 무인매장 운영신규	20 객실 수
21 월평균 객실 이용 건수(55101)	22 편의시설 개수(55101, 55103)	23 객석 수	24 판매 유형별 매출액 구성비변경
25 직능별 종사자 수변경 (582, 62, 63)	26 이용인원(고객)		

24


Census 경제총조사
주요 추진 일정
계획 수립
- 종합시행계획(안) 수립
- 분야별 세부시행계획(안) 수립
26년 2월
26년 3월
출범식 실시
- 실시본부 구성
- 출범식(통계책임관 회의) 실시
권역별 설명회
- 지자체 및 지방청 대상
권역별 설명회 실시
26년 4월
26년 5월
교육 실시
- 교관단, 지자체 공무원 및
조사요원 교육 실시
'25 년 기준
경제총조사 실시
- 2025년 기준 경제총조사 실시
26년 6월
* 온라인조사 6.1.~30.
면접조사 6.12.~7.22.
검토·내검 및 재종 7.23.~31.
26년 12월
조사결과 분석 및 공표
- 잠정 공표 2026. 12.
- 확정 공표 2027. 6.
25


Census 경제총조사
[참고] 로고, 슬로건 및 캐치플레이즈
✔ 로고
2025 Census 경제총조사
✔ 슬로건
Amazing Impact
내 일을 설계하는
오늘의 데이터
✔ 캐치플레이즈
사업데이터를 UP하면
내 경쟁력이 UP됩니다
26


국가데이터처
Census 2025 경제총조사
Amazing Impact
내 일을 설계하는 오늘의 데이터
THANK YOU
https://ecensus.go.kr/

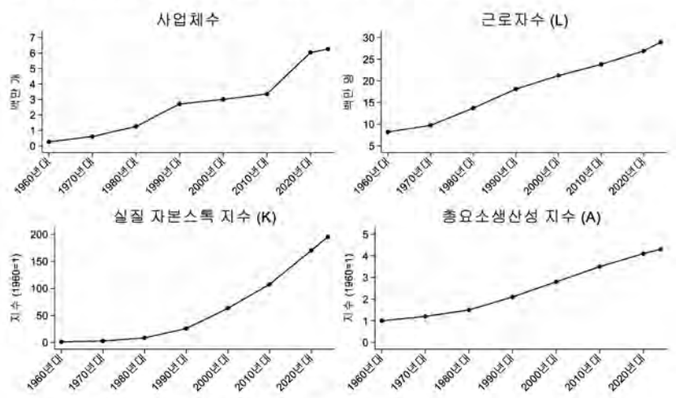
국가통계의 변천과 경제총조사의 학술적·정책적 가치

홍석철
서울대 경제학부
2026년 5월 14일

제15회 국가통계·데이터발전포럼

국가통계로 본 경제 발전

- 한국전쟁 이후 노동, 자본, 기술의 비약적인 증대는 고도 성장의 토대
  - 1960년대 이후 노동력은 약 3.5배, 실질자본스톡은 약 200배, 총요소생산성은 약 5배 증가하고, 사업체수는 약 25배 증가



- 사업체 및 근로자수: 통계청, 『경제총조사』, 『전국사업체조사(SBR)』, 『경제활동인구조사』
- 실질 자본스톡 지수: 한국은행, 『국민대차대조표』 및 Penn World Table (PWT 10.1)
- 총요소생산성 지수: Asian Productivity Organization (APO), 『Productivity Database』
- (주: 장기 시계열 분석을 위해 일부 데이터는 상기 공식 통계 및 국제 DB를 재구성하여 지수화)

2

국가통계는 경제발전을 진단하는 렌즈 역할



출처: 제미나이

3

경제발전과 국가통계의 변천

- ~1950년대: 국가 기틀 마련
  - 전쟁의 폐허 속에서 경제분석 보다는 국가재건과 행정유지를 위한 인적자원 파악에 집중
  - 인구조사의 변천: 1925년 (간이)국세조사 → 1960년 인구주택 국세조사 → 1990년 인구주택총조사
- 1960-70년대: 한강의 기적을 뒷받침한 국가통계의 체계화
  1. 광공업 및 제조업 통계
  - 경제개발 5개년 계획의 성공적인 추진을 위해 중화학 공업 중심의 산업 구조를 수치화하고 모니터링 시작
  - 산업총조사의 변천
    - 1955년: 광공업센서스(초기 산업 기반 파악)
    - 1968년: 광업 및 제조업 센서스
      - 1970년: 광공업동태조사 시작 (월별 생산, 출하, 재고량의 변동을 지수화 → 월간 경기동향분석, GDP 추계, 노동생산성 측정의 핵심 지표로 활용)
    - 1973년: 통계청(당시 경제기획원 조사통계국)이 조사를 인수, 5년 주기(기준년도 끝자리 3, 8)로 정례화하여 전문성 강화
    - 1998년: 전기·가스·수도사업까지 조사 범위를 확장하여 에너지 인프라 전반 포괄

4

경제발전과 국가통계의 변천

- 1960-70년대: 한강의 기적을 뒷받침한 국가통계의 체계화
  2. 상업 및 서비스업 통계
  - 제조업 성장에 발맞춰 재화의 유통과 서비스 산업 구조 변화를 체계적으로 모니터링 시작
  - 상업센서스 변천
    - 1968년: 상업센서스 도입
    - 1971년: 도소매업센서스로 세분화
  3. 노동 통계
  - 경제활동 가능 인구에 대한 정밀한 조사 시행
  - 경제활동인구조사의 변천
    - 1957년: 노동력조사 시작
    - 1962년: 경제활동인구조사로 개편
    - 1982년: 급변하는 고용 시장에 적극적으로 대응하기 위해 조사 주기를 분기에서 월별로 변경 (고용 정책의 시의성 확보)

5

경제발전과 국가통계의 변천

- 1980-90년대: 소비의 성장과 서비스 경제로의 전환
  1. 서비스업 통계의 고도화
  - 국민소득 증대에 따른 3차 산업의 팽창을 포착하기 위해 상업 중심의 조사가 서비스업 전반을 아우르는 체계로 확장
  - 도소매 및 서비스업 총조사의 변천
    - 1968년 상업센서스 → 1971년 도소매업센서스
    - 1986년: 기존의 도소매업 센서스에 서비스업을 통합하여 '도소매업 및 서비스업 총조사'로 확대 개편 (산업 구조 고도화 반영)
    - 2006년: 서비스 경제의 비대화에 따라 '서비스업총조사'로 독립 및 전문화
  2. 소비자물가지수(CPI)의 현대화
  - 소비 패턴의 다양화와 양적 성장에 발맞춰 물가 조사 체계가 전국 단위의 과학적 시스템으로 변모
  - 물가 조사의 변천
    - 1936년: 경성상공회의소 물가 조사 시초
    - 1965년: 경제기획원 조사통계국(현 국가데이터처)에서 소비자물가조사 시작
    - 1981년: 지수 산정 방식 고도화, 조사 품목 확대와 세분화, 조사 지역 전국 단위 확대, 전산화 시스템 도입

6

경제발전과 국가통계의 변천

- 1980-90년대: 소비의 성장과 서비스 경제로의 전환
  3. 데이터 거버넌스의 초석: 사업체기초통계조사 도입
  - 급변하는 경제 구조를 선명하게 포착하기 위해 모든 사업체를 망라하는 통계 모집단을 구축
  - 1994년: 사업체기초통계조사 실시
  - 2008년: '전국사업체조사'로 명칭 변경 및 체계 정립
    - 2011년부터는 기준연도(0과 5인 해)에는 '경제총조사'를 실시하고, 나머지 해에는 '전국사업체조사'를 실시하는 방식으로 주기를 조정하여 운영
- 2010년대 이후: 디지털 대전환과 데이터 거버넌스
  - 산업 간 경계가 허물어지면서, 과거 분절되어 있던 각종 산업·서비스 조사를 하나로 통합
  - 2011년: 산업총조사와 서비스업총조사를 통합하여 경제총조사 실시
    - 기존 산업총조사와 서비스업 총조사의 문제점 개선
      - 산업총조사와 서비스업총조사의 조사 시점 불일치에 따른 데이터의 동시성 결여
      - 포괄 범위의 사각지대에 따른 전체 산업의 조망 불가
      - 중복 조사 등에 따른 행정 자원의 낭비
  - 2016년: 등록기반 통계 도입
    - 행정자료를 연계하여 통계 생산 → 포괄 범위 확대, 비용 절감, 적시성 향상

7

경제총조사: 국가 경제의 정밀 진단과 미래 나침반
- 경제총조사의 주요 내용 -

- 경제총조사의 주요 활용 목표
  - 정부의 경제정책 수립 및 성과 평가
  - 사업체 대상 표본조사의 모집단 및 기준점 제공
  - 가공통계 작성을 위한 기초 자료 제공
  - 학계의 연구 자료 및 기업의 시장 분석·경영 계획 수립의 근거
- 주요 조사 항목: 생산의 3요소(L, K, A) 진단
  - 노동: 종사자수, 급여 현황 (인적자원의 배분과 비용 진단)
  - 자본: 자산 및 부채 현황, 자본금, 연간 매출 및 영업 비용 (자본 축적 및 수익성 진단)
  - 기술 및 혁신: 산업별 특성 항목, 디지털 전환 지표 (기술 확산 및 혁신 성과 진단)
  - 인프라: 사업체 일반 현황, 기업 조직 형태 (경제 생태계의 뼈대 파악)
- 최근 조사의 특징: 변화하는 경제 패러다임 반영
  - 2016년: 기업 조직의 형태와 프랜차이즈 관련 항목 심층 조사
  - 2021년: 디지털·플랫폼 경제로의 전환을 반영하기 위해 무점포 사업체 및 전자상거래 관련 항목 강화

8

경제총조사: 국가 경제의 정밀 진단과 미래 나침반
- 2026 경제총조사 주요 특징 -

- AX(AI 전환) 및 자동화 실태 조사 강화
  - 대한민국 경제의 미래를 좌우할 혁신 기술의 확산 정도를 측정
  - 지능형 기술 도입: AI·로봇 활용 여부 조사
  - 스마트 인프라: 스마트공장·스마트농장 현황을 통한 제조·농업 혁신 진단
  - 비대면 경제 고도화: 무인매장 운영 등 고용 구조와 소비 패턴의 변화를 상징하는 신규 항목 포함
- 제11차 한국표준산업분류(KSIC)가 적용
  - 변화된 산업 생태계 반영
  - 미래 모빌리티: 수소·전기차 등 친환경 이동 수단 및 관련 산업 세분화
  - 신산업 체계화: 기존 분류로 포착하기 어려웠던 첨단 기술 서비스 및 융복합 산업군을 재정립
- 데이터 기반의 지역 맞춤형 정책 수립
  - 지방 소멸 위기와 지역 경제 활성화를 위한 전략적 자산으로 활용
  - 산업 생태계 분석: 지역별 특화 산업의 건강도를 정밀 분석하여 자생적 성장 동력 확보 지원
  - 파편화된 행정 자료와 결합하여, 지역별·산업별로 최적화된 '맞춤형 경제 정책' 수립의 과학적 근거 제공

9

경제학 관점에서 바라본 국가통계의 변천

- 국가통계의 변천사는 노동과 자본을 넘어 혁신 기술까지, 시대별 핵심 생산 동력을 발굴하고 정량화해 온 과정
  - 노동: 양적 팽창(인구)에서 질적 분화(고용 형태)로
  - 자본: 물적 축적(설비)에서 운영 효율로
  - 기술: 보이지 않는 혁신의 가시화와 정량화
- 경제발전 단계별 국가통계의 변천

	초기	성장기	현재
노동(L)	**양적 확보가 최우선** 생산가능인구 파악 중심 (인구센서스 기반)	**노동의 이동과 배분** 산업 인력 공급 확대 중심 (경활조사 기반)	**고용의 질적 분화 진단** 단순한 인력 규모 파악을 넘어 초단시간 근로, 외국인력, 고령층 노동력 등 노동 공급의 구조적 변화의 정밀 추적
자본(K)	**물적 자본의 투입 측정** 공장(사업체)의 수, 기계 설비 규모 등 양적 투입량을 측정하는 것이 통계의 주 목적	**자본 축적의 가시화** 산업별 실질 자본스톡 추계 및 자본 형성 데이터 기반 마련	**자본 효율성 진단** 자본 투입 대비 매출과 이익률을 전수 조사함으로써, 경제 전반의 자본 효율성을 진단
기술(A)	**잔차(residual)의 시대** 노동과 자본을 제외한 나머지 성장분으로 기술을 추정	**기술 도입의 가속화** 연구개발 투자 규모 등 투입 위주의 기술 지표 관리	**혁신의 측정** 디지털 전환(DX) 및 AI·로봇 활용 여부 조사 등 혁신 진단

10

정책적·학술적 활용 가치
- 사업체 대상 표본조사의 모집단 및 기준점 제공 -

- 사례 1: 거시경제 지표 산출 및 국가 경제 진단의 근간
  - 산업구조의 벤치마크 제공: 경제총조사는 광업·제조업 등 산업별 구조를 확정하여, 매월/매년 실시하는 동태조사의 기준점 제공
  - GDP 추계 고도화 활용: 월별 출하·재고 지수 등 단기 지표와 총조사 데이터를 결합하여 국가 GDP 산출의 정확성 확보
  - 경제 진단의 나침반 역할: 총조사를 통해 보정된 데이터는 중장기 경기 국면 진단 및 성장 잠재력 분석의 기초 자산
- 사례 2: 미시 정책 설계를 위한 마스터 프레임 제공
  - 다양한 표본조사의 모집단 역할: 경제총조사 결과는 사업체노동실태현황조사, 전국소상공인실태조사, 산업연관표, 환경산업통계조사 등 국가가 실시하는 약 60여 종의 주요 표본조사에 정확한 모집단과 기준점 제공
  - '중소기업실태조사' 사례: 2022년 제56차 실태조사부터 모집단 구성 시 표본추출틀을 통계청 전국사업체조사에서 기업통계등록부(경제총조사 기반 행정데이터 통합 시스템)로 변경하여 정확성을 높였으며, 표본규모는 시계열항목 17,000개, 일반항목 5,000개 수준으로 운영

11

정책적·학술적 활용 가치
- 가공통계 작성을 위한 기초 자료 -

- 사례 1: 광공업생산지수 기준년도 품목·가중치 산정
  - 광공업생산지수는 2015년 기준으로 광업, 제조업, 전기·가스업 부문 총출하액의 1/5,000 이상이 되는 485개 품목을 대상으로 2015년 업종별·품목별 부가가치를 가중치로 사용하여 산출
  - 이 가중치의 원천이 바로 경제총조사이며, 5년마다 경제총조사가 실시될 때 품목과 가중치 갱신
- 사례 2: 국민소득(GDP) 및 지역소득(GRDP) 추계 기초자료
  - 경제총조사의 모든 자료는 국민소득, 지역소득 추계, 경제통계모집단 제공, 읍면동 통계 작성 등을 위한 자료로 활용
  - 지역소득 통계는 단순 참고 자료가 아니라, 지역내 산업구조의 실태를 파악하여 산업정책의 구체적 방향 결정, 조세부담역량의 평가척도인 주민의 조세부담률로부터 지방세 수입을 합리적으로 계획하고 지방세제 개정의 기초자료로도 활용

12

정책적·학술적 활용 가치
- 연구 및 경영 자료 제공 -

- 사례 1: 학술 연구 및 정책 평가의 실증적 근거
  - 경제총조사는 개별 사업체 단위의 미시 데이터를 제공함으로써, 단순한 현상 파악을 넘어 경제적 인과관계를 규명하는 고도화된 연구 가능
  - 강지수, 전현배, 조장희 (2017): 진입퇴출과 서비스업 생산성 성장: 과밀창업 억제 대상자영업을 중심으로, 경제학연구.
    - 2005년 서비스업총조사와 2010년 경제총조사를 연계 분석
    - 생계형 자영업의 과밀 창업이 전체 서비스업 생산성 성장을 저해하는 메커니즘을 실증적으로 규명
- 사례 2: 비즈니스 의사결정 및 창업 지원 인프라
  - 전수조사 데이터가 공간 정보(GIS)와 결합하여, 민간의 창업 성공률을 높이고 기업 생태계를 시각화하는 실용적 자산으로 변모
  - SGIS(통계지리정보서비스)를 통한 입지 분석 (https://sgis.mods.go.kr/)
    - 전국사업체조사와 경제총조사 데이터를 지도 위에 구현하여, 업종별 밀집도, 상권의 유동적 변화, 배후지 특성을 분석
    - 예비 창업자의 위험을 줄이는 데이터 기반 의사결정 도구
  - 기업생태 분석지도
    - 매년 실시되는 전국사업체조사와 5년 주기의 경제총조사를 보완적 활용
    - 기업의 생성-성장-소멸 과정을 추적

13

정책적·학술적 활용 가치
- 정부 정책 수립 및 평가 -

- 경제총조사는 개별 산업의 표면적 성장 뒤에 숨겨진 구조적 한계를 가시화하여, 정부가 최우선적으로 해결해야 할 정책 과제를 도출하는 데 기여
- 사례 1: 소상공인 과밀화 진단과 정책 연계
  - 과밀화 진단과 영세성 확인
    - 경제총조사의 업종별·지역별 사업체 수와 매출액 데이터를 분석 → 특정 상권 내 동일 업종의 '과밀 지수' 산출 → 과당 경쟁으로 인한 영세성 규명
  - 정책연계: 희망리턴패키지 사례
    - 이러한 진단 결과는 정부가 창업 지원 중심의 정책에서 벗어나, '업종 전환'을 지원하는 방향으로 소상공인 정책의 패러다임을 전환하는 근거 제공
    - 경제총조사에서 포착된 '매출 급감 사업체' 및 '영업이익 적자 구간 사업체'의 비중은 폐업 위기 소상공인의 규모를 예측하는 기초 자료
    - 폐업 지원: 폐업 과정에서의 정보 부족과 비용 부담을 덜어주기 위한 컨설팅 및 점포 철거비 지원의 규모를 산정
    - 재기 지원: 폐업 이후 '재취업' 또는 '유망 업종으로의 재창업'을 돕는 교육 프로그램을 설계할 때, 경제총조사상에서 성장성이 확인된 유망 산업 데이터를 활용하여 교육 커리큘럼 구성

14

거대한 변화에 직면한 대한민국
- 인구구조변화와 AI 기술혁신 -

- 지속된 저출생과 급격한 인구고령화 → 저성장의 고착화 위기
  - 인구구조변화는 노동력부족과 소비·투자 감소를 초래 → 저성장으로 귀결
  - 생산연령인구 전망
    - '24년 70.2% → '33년 이후 OECD 평균인 65% 보다 낮아지고 → '39년 이후에는 일본(59%) 보다 낮아지며 → '70년에는 46% 전망 (출처: 장래인구추계)
  - 잠재성장률 전망(KDI)
    - 2025-30년 1.5% → 2030년대 0.7% → 2040년대 0.1% (역성장 일상화)
- AI와 로봇 기술 → 경제·사회 위기 극복을 위한 혁신의 돌파구
  - 인구구조변화에 따른 노동력 부족과 생산성 저하를 해소할 수 있는 근본적인 해법
  - 성장을 견인해 온 제조업에 AI를 결합한 '지능형 제조 공정'은 글로벌 공급망 재편 속에서 국가 성장 동력을 유지하는 핵심 열쇠
  - 초고령사회에서 급증하는 복지 수요와 돌봄 공백을 빅데이터, 로봇, AI 기반의 Aging Tech로 대응 → 사회보장의 비용 효율성을 높이고 국가 재정의 지속가능성을 확보

15

지속가능한 미래 구축을 위한 경제총조사의 활용



16

경제총조사 활용 가치 증대를 위한 제언
- 연구 기획 및 학술 연구 연계 강화 -

- 주요 부처 주도의 연구 기획 및 정책 개발 활용 강화
  - 지속가능한 국가 경제와 사회 구축을 위해 필요한 정책 연구를 기획하고 학술 연구와 연계될 수 있도록 지원을 확대
  - 단순한 현황과 추이 파악을 넘어, 주요 요인과 결과의 인과성 분석이 뒷받침 되어야 보다 효과적이고 효율적 정책 수립 가능
- 미국의 Economic Census를 활용한 최신 연구 동향
  - Acemoglu et al. (2022), "AI and Jobs: Evidence from Online Vacancies," *Journal of Labor Economics*.
    - 2010년대 중반부터 경제총조사에 도입된 AI 관련 설문 항목을 활용하여, 어떤 특성을 가진 기업들이 AI를 도입하는지 분석
    - AI 도입 기업들은 특정 업무(Task)의 노동 수요를 줄이는 대신, 기술 보완적 직무에서의 생산성 향상 도모 규명
  - Autor et al. (2020), "The Fall of the Labor Share and the Rise of Superstar Firms", *Quarterly Journal of Economics*.
    - 약 30년간의 미국 경제총조사 전수 데이터를 분석
    - 각 산업군에서 소수의 '슈퍼스타 기업(Superstar Firms)'이 시장 점유율을 독식하는 현상을 규명 → 이들 기업은 생산성은 매우 높지만, 매출액 대비 노동소득 분배율은 낮아 전체 국가 경제의 노동 소득 분배율 하락을 주도

17

경제총조사 활용 가치 증대를 위한 제언
- 자료 연계 확대 -

- 행정자료 및 외부자료와의 연계 강화
  - 기업 성장의 동태적 변화와 요인을 규명하기 위해서는 경제총조사 간, 그리고 타 행정자료 및 외부자료와의 결합 필요
  - 이는 기업과 산업 분석 뿐만 아니라 사회 문제 해결을 위한 연구에도 기여
    - 예시: 경제총조사와 건강보험공단자료 결합 → 산업구조의 변화가 장기적으로 국민건강과 건강보험재정에 미치는 영향 예측 및 대응 전략 수립
- 자료 연계 연구 사례
  - Hong et al. (2026), "Foreign Aid and Economic Development: A Firm-level Study of Korean Experience in the 1950-70s"
    - 1950년대 미국 USAID의 국내 기업에 대한 원조기록과 1955-59년 『광업 및 제조업 사업체명부』, 1973년 『기업조사총록』 결합 분석
      - 원조기록: Small Business Circular, Weekly Report, Daily Report from US National Archives
    - 기존의 통념과는 달리 원조자금은 경쟁적이고 공정하게 배분되었으며, 1950년대 원조는 한국 산업과 대기업 성장의 토대가 되었음을 실증적으로 규명

18

과거-현재-미래를 연결하는 경제총조사



19

국가데이터처 미시경제데이터의 정책의사결정 활용도 제고 방향

제15회 국가통계·데이터 발전포럼 | 세션 3-1

장우현 (한국조세재정연구원 국가회계재정통계센터 소장/선임연구위원)

2026. 5. 14.

한국조세재정연구원

발표 순서

PART 1 국가데이터처 자료는 정책의 상식을 바꿀 힘이 있다
국가데이터처 미시경제데이터가 정책 의사결정에 결정적으로 작용했던 개인 경험 사례들

PART 2 개인적 활용의 정체와 세 가지 아쉬움
적시성 · 자료 연계 · 정책변수의 한계

PART 3 국가데이터처 자료의 잠재력 극대화를 기대하며
정책 제언 — 더 큰 기대를 담아

PART 1

국가데이터처 자료는 정책의 상식을 바꿀 힘이 있다

네 가지 사례로 살펴보는 국가 미시경제데이터의 위력

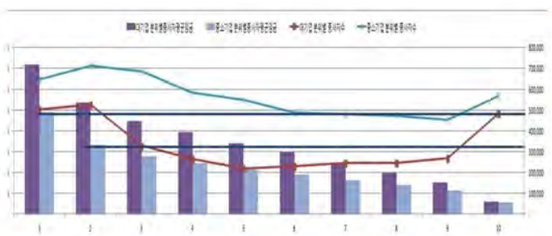
왜 국가 미시경제데이터가 결정적인가

국가 전수조사의 신인도(信認度)는 표본조사·민간자료가 대체할 수 없는 정책 설득력을 만든다

- **정책 의사결정의 최종 라운드에서는 「누구도 부정하기 어려운 보편적 근거」가 요구됨**
  - 표본조사는 일반화의 한계, 민간자료는 신인도와 공공성의 한계를 가짐
- **국가 미시경제데이터의 두 가지 결정적 무기**
  - 전수성(全數性) — 표본 추론을 넘어선 보편적 결론 도출 가능
  - 행정자료 식별자 연계 — 사업자등록번호·법인등록번호 기반 정책이력·재무정보 결합
- **대표 자산 — 광업·제조업조사(연 1회), 경제총조사(5년 주기 전수)**
  - 두 자료는 지난 십여 년간 한국의 중소기업·산업·재정 정책의 상식을 여러 차례 바꿔왔음

사례 ① VIP 주재 간담회의 핵심 슬라이드 — 경제총조사 기반 임금·고용 분위 격차

"소상공인을 제외하면 중소기업 상위 1분위가 대기업 상위 3분위의 평균 임금·고용 규모를 뛰어넘는다"



2011년 경제총조사 기준. 전산업 막대그래프 좌축(백만원), 선그래프 우축(명). 숫자가 낮은 분위가 더 높은 성과.

- **주요 회의에서 인용된 슬라이드**
  - 2025년 「나라재정 절약 간담회」 발제 핵심 자료
  - 경제총조사 기반 분석이 결정적 근거
- **경제총조사가 만든 정책적 함의**
  - 「대기업 vs 중소기업」 이분법의 위험성
  - 소상공인 분리 시 중소기업의 잠재력이 드러남
  - 우수 중소기업 = 정책 조준 재설계의 근거
- **전수조사의 신인도가 만든 설득력**
  - 최상 보고 라운드까지 도달한 데이터의 힘

출처: 장우한(2017), 2011년 공표 2010 기준 경제총조사 / 장우한(2025), 「효율적 재정지출을 위한 중소기업정책 재구조화 제안」, 사라새정 협약·감담회 발제자료, p.1

사례 ② 「99–88」을 「80–62」 / 「9.6–37」로 — 경제총조사와 법인등록번호의 만남

소상공인과 중소기업을 분리해 「진짜 중소기업」의 모습을 드러내다

기존 통념 — 「99–88」

사업체 99.9%, 고용 88%가 중소기업이라는 광범위한 통계

대기업 소속 사업장이 중소기업으로 분류되고, 소상공인이 중소기업에 묻여 통계가 「울타기」

재계산 — 2010 경제총조사 + 법인등록번호

「80–62」 모수 소상공인 제외 시 소상공인 외 중소기업 사업체수·고용 비중

「9.6–37」 모수 소상공인 포함 시 소상공인 외 중소기업 사업체수·고용 비중

종사자 인당 평균연봉 (2010 경제총조사, 만원)

구분	대기업	중소기업 (소상공인 제외)	소상공인
평균연봉	3,647	2,446	659

- **소상공인 포함 시 중소기업 평균연봉이 1,500만원으로 왜곡 — 정책 진단의 출발점이 흐려짐**
- **정책 파급력 — 약 5년 후 정책 상식이 됨**
  - 현재 중기부는 소상공인을 분리한 통계를 산출·제공
  - 산업·고용·재정 정책의 진단 단위가 바뀜

출처: 장우한·양용현·주석진(2013), 「중소기업지원정책의 개선방안에 관한 연구(II)」, KDI, pp.25–28 (표 1–8)

사례 ③ 광업·제조업조사 × 정책금융 지원이력 — 동태적 효율성 훼손의 발견

「효과가 없을 수 있다」에서 「동태적 효율성을 훼손할 수 있다」로 — 정책 상식의 전환

- **자료 구축 — 전수 × 행정자료 연계**
  - 광업·제조업조사 (10인 이상 전수, 연 1회)
  - 신보·기보·중진공 정책금융 지원이력
  - 사업자등록번호 기준 결합 → 19.6조원 지원이력 확보
- **분석 방법**
  - 성향점수 짝짓기(PSM) 추정 — 지원 vs 비지원 비교
  - 총요소생산성(TFP)을 정책평가지표로 활용

−2.73	2.5조원	+5.32%p
TFP 격차 (p<0.01)	잠재 부가가치 손실 (GDP의 0.2%)	잔존율 증가 (p<0.01)

정책 상식의 전환

본 연구 이전: 「중소기업 지원의 효과가 미흡할 수 있다」
본 연구 이후:
「지원이 동태적 효율성을 훼손해 마이너스를 만들 수 있다」
→ 시장 효율화 메커니즘에 정부 개입이 역행할 위험을 정량화

출처: 장우한·양용현·주석진(2013, 2014), 「중소기업지원정책의 개선방안에 관한 연구」, KDI / 장우한(2016), KDI Focus 제83호

사례 ④ 광업·제조업조사 — 진입·철수의 동학(動學)과 생산성 기여

전수성이 있어야만 가능한 분석 — 한국 산업의 동태적 활력을 정량화하다

2008년 vs 2011년 광업·제조업 사업체 동학 (10인 이상, 2011년 화폐가치 기준)

구분	사업체 수 증분	매출액 증분 (백만원)	영업이익 증분 (백만원)	고용 증분 (명)	인당 지급액 증분 (백만원)
지속사업체	0 (지속사업체 수 32,964)	351,945,908	38,592,663	183,350	4.49 (평균 39.87)
철수사업체 (2008)	-25,236	-416,703,202	-48,099,603	-839,056	-30.79
진입사업체 (2011)	25,419	296,745,516	23,083,206	766,676	25.82
합계	183	231,988,223	13,576,235	110,970	1.82

함의

- **3년간 32,964개만 지속, 25,236개 철수, 25,419개 진입 — 한국 산업의 활발한 진입·퇴출이 정량 확인됨**
- **지속사업체 인당 지급액 35.38 → 39.87(생산성 향상), 진입사업체는 25.82에서 출발 — 효율화 동학 가시화**
- **표본조사로는 도출 불가능 — 진입·철수의 정확한 식별은 전수조사의 산물**

출처: 장우한·양용현·주석진(2014), 「중소기업지원정책의 개선방안에 관한 연구(III)」, KDI, p.12 (표 4-2)

PART 1 소결 — 네 사례가 공통으로 보여주는 것

국가 전수조사 + 행정자료 식별자 연계 = 정책 의사결정의 「게임 체인저」

사례 ①
경제총조사 임금·고용 분위
중요 건담회 핵심 슬라이드

사례 ②
99–88 → 80–62 / 9.6–37
중기부 통계 산출 기준 변경

사례 ③
정책금융의 동태적 효율성 훼손
중소기업 자원 정책 상식의 전환

사례 ④
진입·철수의 생산성 동학
한국 신업 활력의 정량 측정

공통 분모

- 네 사례 모두 **광업·제조업조사 또는 경제총조사**를 활용
- 모두 **사업자등록번호 또는 법인등록번호 기반의 행정자료 연계**로 보편성 확보
- 발표 시점에서 **약 5년의 시차를 두고 정책 상식으로 정착** — 표본·민간자료로는 대체 불가능한 파급력

그런데
적어도 저의 경우에는 …

PART 2
개인적 활용의 정체와 세 가지 아쉬움

적시성 · 자료 연계 · 정책변수의 한계

활용의 정체 — 현황 진단

정책평가 연구자 현장체감으로는 결정적 기반으로서의 미시경제데이터 활용은 줄어들고, 「배경 설명용」으로 활용되는 경향

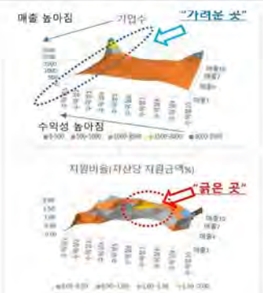
■ 최근의 개인적 활용 패턴
- 경제 보고서·발표의 첫 슬라이드 거시 배경 설명용
- 주로 경제총조사 요약치, 아쉽게도 2026년에는 5년 전 자료를 인용
- 정책 의사결정의 결정적 증거로서의 활용은 점차 감소

■ 연구 현장의 변화
- 연구자들이 점차 한국평가데이터(KoDATA) 등 민간 자료로 우회
- 국가데이터처 자료의 잠재력 대비 실제 활용도 격차 확대

■ 남는 의문 — 좋은 자료가 있음에도 왜 PART 1의 사례를 재생산하기 힘들어졌는가?
- 다음 페이지부터 그 원인을 세 가지로 진단

우회 사례 — KoDATA × SIMS「가려운 곳 따로, 긁은 곳 따로」

의미 있는 결과지만, 같은 분석을 국가데이터처 전수자료로 병행했다면 신인도 확보에 도움이 되었을 것으로 판단함



2021년 KoDATA 중기업 82,825개를 수익성·매출 10분위로 분할, SIMS 자금분포와 비교

출처: 장우현 외(2024) / 장우현(2025), 「효율적 재정지출을 위한 중소기업정책 재구조화 제안」, 나라재정 정책 감담회 발제자료, p.5

- **분석 개요**
  - KoDATA 회계자료 + SIMS 정책지원이력 병합
  - 수익성·규모 우수「가려운 곳」≠ 지원 집중「긁은 곳」
- **결정적 차이**
  - KoDATA는 사업자등록번호로 SIMS와 자유롭게 병합 가능
  - 국가데이터처 자료라면 추가 행정 절차·승인 필요
- **남은 아쉬움**
  - 전수성·공공성·신인도에서 국가 자료의 강점이 있음.
  - 「만약 광업·제조업조사로도 같은 분석이 가능했다면…」

진단 ① 적시성 (Timeliness)

정책 사이클은 분기·반기 단위로 빠르지만, 자료의 갱신 주기는 아쉬움

자료 갱신 주기



현장의 실제 모습

2025년 **나라재정 절약 간담회**의 핵심 슬라이드는
2011년 경제총조사 — 14년 전 자료

자료가 주는 통찰은 강력했지만, 「가장 최신 전수자료를 인용한다」고 자신 있게 말하기는 어려운 상황

정책 사이클(분기·반기)과 자료 사이클(1년·5년)의 미스매치 — 결정타가 되어야 할 시점에 자료가 도착하지 않는다

진단 ② 자료 연계 활용의 어려움 (Linkability)

민간 회계자료는 자유롭게 결합되는데, 국가 자료는 추가 절차의 비용이 존재

민간 회계자료 (KoDATA 등)

- **사업자등록번호 = 자유로운 연결 키**
  - SIMS 정책지원이력과 자유 병합
  - 세무·재무·고용 등 다양한 행정자료와 결합
- **빠른 연구·분석 사이클**
  - 가설 → 데이터 결합 → 분석 → 정책 제언
  - 한 사이클이 수개월 단위로 가능

국가데이터처 자료

- **연계 활용에 추가 행정 절차**
  - 승인 절차의 복잡성과 시간 소요
  - 사후 활용 제한(재이용·확장 어려움)
- **연구의 지속성·반복성 제약**
  - 한 분석이 다음 분석으로 이어지기 어려움
  - 아이러니 — 2009년 광업·제조업조사 × 정책금융 연계가 가능했던 모델이 지금은 더 어려움

연구자가 자연스럽게 KoDATA 같은 민간 자료로 옮겨가게 되는 구조적 이유

진단 ③ 정책 변수의 미흡 (Policy Variables)

국민의 세금이 투입된 정책의 효과를, 국민의 세금으로 만든 통계로 검증하기 어려운 구조

현재 결합·확보가 어려운 정책 변수

- 기업 보조금 수혜 이력
- 중소기업 정책 지원 여부 및 금액
- R&D·고용·수출 지원 등 분야별 수혜
- 세금 감면·공제 내역
- 지자체 단위 보조금

해외에서는 공개정보

- **EU State Aid Register**
  국가별 기업 보조금 수혜 내역 공개 검색
- **USA 연방 지원 공시 (USAspending.gov)**
  연방정부의 기업·기관 지원 내역 공개

정책 변수 없이는 정책 효과 평가의 출발점이 봉쇄 — 국민경제적 평가의 사각지대

PART 3

국가데이터처 자료의 잠재력 극대화를 기대하며

정책 제언 — 더 큰 기대를 담아

기대 효과 — PART 1의 사례들이 다시, 더 자주

국가데이터처 자료가 갖는 정책 의사결정의 핵심요소로서의 역할이 회복·확장됨

현재	보강 후
5년 전 자료를 한국경제 요약으로 활용하는 아쉬움	▶ 최신 전수자료가 의사결정의 핵심요소로
KoDATA 등 민간자료로 우회 분석	▶ 국가 전수자료로 보편적 주른 도출
정책 효과 평가가 사각지대에	▶ 정책변수 결합으로 즉시 평가 가능

국가데이터처 자료의 본래 잠재력 — 정책의 상식을 바꾸는 힘 — 의 극대화

정책 제언 — 세 가지 보강이 만들 수 있는 변화

세 가지가 함께 보강될 때 PART 1의 사례들이 다시, 그리고 더 자주 재생산될 수 있다

01	02	03
적시성 개선	**공개가능 DB 확보**	**정책 변수 보강**
경제총조사·광업·제조업조사 등의 조사 및 발표 주기 단축 및 잠정치 조기 공개	민감정보를 제외한 연계 가능 미시 DB를 별도 구축하여 활용 가능하도록 정비	기업 보조금·중소기업 정책 지원·세제 혜택 등 정책 결합 변수의 체계적 정비
정책 사이클과의 미스매치 해소	*연구·분석 사이클의 가속*	*국민경제적 정책 평가의 출발점 확보*

맺음

국가데이터처의 출범에 깊이 축하드립니다

국가 미시경제데이터는 **정책의 상식을 재정립할 힘**을 가지고 있습니다.
적시성·연계성·정책변수가 함께 보강된다면 그 잠재력은
극대화될 것입니다.

향후 충분한 예산 배정과 제도적 뒷받침이 그 길을 열어주리라
큰 기대를 가지고 지켜보겠습니다.

감사합니다.

장우현 · 한국조세재정연구원 선임연구위원

Japan's Approach to the Economic Census

Statistics Bureau,
Ministry of Internal Affairs
and Communications,
May 2026

総務省統計局
Statistics Bureau,
Ministry of Internal Affairs
and Communications

Contents

- Introduction
- Overview of Japan's Economic Census 2026
- Transition to Online Survey
- Implementation of Business Profiling
- Summary

Introduction

Objective of the Economic Census

- **To identify the economic structure** comprised of the activities of establishments and enterprises in all industrial sectors on a national and regional level

1. Identify the economic activities of establishments and enterprises
   – Understand the economic activities in all industrial sectors at the same point in time
   – Provide information on economic activities not only on a national but also regional level
2. Prepare population information of various statistical surveys
   – In order to improve the accuracy of sample surveys of establishments and enterprises, highly accurate population information is necessary

Introduction

Structure of the Economic Census

Economic Census for Business Frame

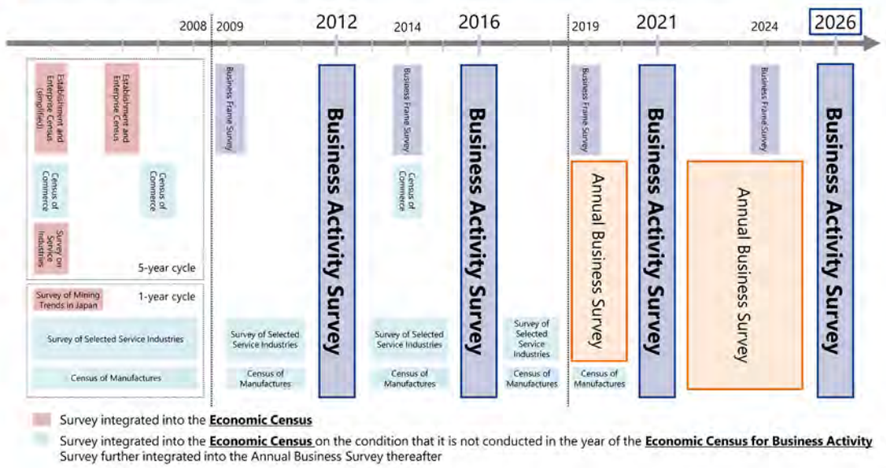
Focuses on preparing population information (lines of business, number of employees, etc.)

Economic Census for Business Activity

Focuses on understanding the status of economic activities of establishments and enterprises (sales, costs, etc.)

Introduction

Progress of the Systematic Development of Economic Structure Statistics



Contents

- Introduction
- **Overview of Japan's Economic Census 2026**
- Transition to Online Survey
- Implementation of Business Profiling
- Summary

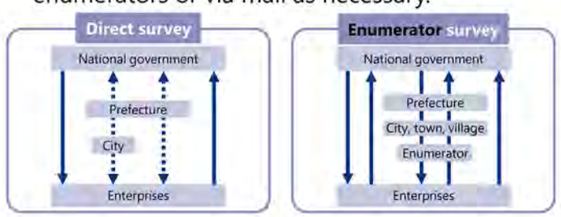
Overview of Japan's Economic Census 2026

- Date of the Survey: **June 1, 2026**
  - Reference date on "whether an organization is to be surveyed or not," "the name and location," "the number of employees," etc.
  - The economic activities (sales, expenses, etc.) over one year in 2025 are ascertained
- Structure of the Questionnaire
  - **Common survey items** to be ascertained for all industries
    - 16 items in total: sales, expenses, etc.
    - Some of these items are collected at the enterprise level instead of the establishment level, taking into consideration the burden on businesses.
  - **Specific survey items** to be ascertained for each industry
    - Example: The value of product shipments in the manufacturing industry -> Used to promote local industries

Overview of Japan's Economic Census 2026

Main Changes for the Economic Census 2026

- Application and review of the product classification:
  - Establish the product classification in the goods sector
  - Review the product classification in the service sector, which was introduced in 2021
- Revision of survey methods to **encourage online responses**:
  - At first, only the documents required for the online response are sent to respondents. Subsequently, paper questionnaires are provided by enumerators or via mail as necessary.



* These figures illustrate Survey A, which targets non-governmental establishments; Survey B targets national and local public bodies (omitted here).

Contents

- Introduction
- Overview of Japan's Economic Census 2026
- **Transition to Online Survey**
- Implementation of Business Profiling
- Summary

Transition to Online Survey

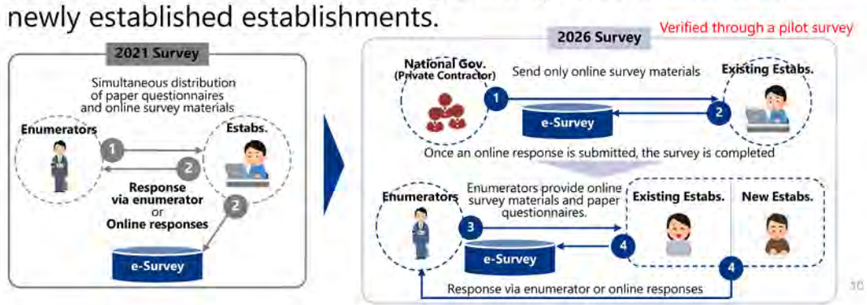
Background: Issues Identified in the 2021 Survey

- As the survey period of the previous census overlapped with the fourth wave of COVID-19 pandemic, local governments expressed the following views:
  - Detailed item information was often left blank, and the increase in WFH due to COVID-19 made it extremely difficult to fill in missing information through follow-up inquiries with the responsible staff. This was also affected reminder activities.
  - With enumerator-based surveys being discontinued and mail-in surveys being adopted in other business surveys, securing enumerators will become increasingly difficult unless survey methods are improved.
- **Promoting online responses is an effective way** to address the issues in both survey implementation and data review perspectives.

Transition to Online Survey

New Survey Method [Enumerator Survey]

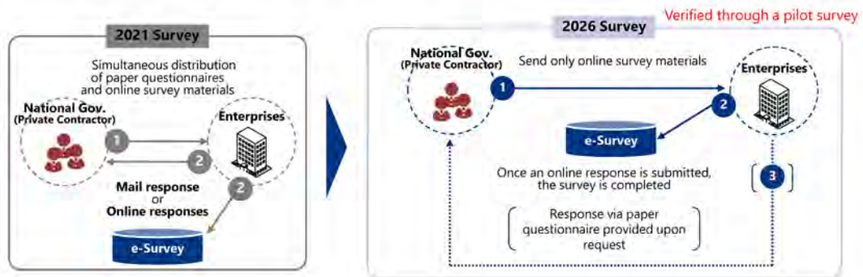
- Only online survey materials (e.g., IDs and passwords) are sent to existing establishments by a private contractor commissioned by the national government.
- Enumerators then provide online survey materials and paper questionnaires to non-responding existing establishments and newly established establishments.



Transition to Online Survey

New Survey Method [Direct Survey]

- Only online survey materials (e.g., IDs and passwords) are sent to target companies by a private contractor commissioned by the national government.
- Paper questionnaires are provided only upon request.



Transition to Online Survey

Results of the Pilot Survey in 2024

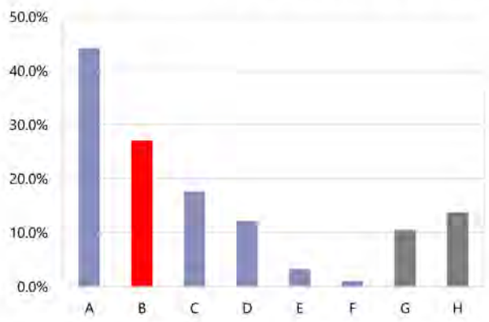
- A significant increase in the online response rate was verified in the pilot survey.
  -> The same survey methodology is adopted for the main survey.

Survey System		Online Response Rate	
		2024 Pilot Survey	2021 Survey
Enumerator Survey (Establishments)		60.7% (+20.8pt)	39.9%
	Individually Operated	46.2% (+15.4pt)	30.8%
	Corporations	71.2% (+23.0pt)	48.2%
Direct Survey (Enterprises)		87.4% (+27.0pt)	60.4%

Transition to Online Survey

Feedback from Respondents Using Paper (Enumerator Survey)

- Q1. What were the reasons for not responding online?
  (Excluding cases where no online-capable device was available)
  -> Did not consider online response at all. **[55.5%]**
  *Detailed reasons are shown below (multiple selections allowed).

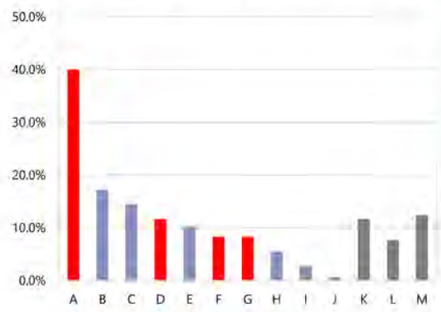


A. Usually respond using paper questionnaires
B. It is easier to grasp the overall content
C. It was difficult to operate in previous surveys
D. Security concerns
E. It is easier to respond while checking information on a screen
F. Paper questionnaires are easier to share internally
G. Other (not comfortable using a PC)
H. Other (not listed above)

Transition to Online Survey

Feedback from Respondents Using Paper (Enumerator Survey)

- Q1. What were the reasons for not responding online?
  (Excluding cases where no online-capable device was available)
  -> Considered online response but chose a paper-based. **[44.5%]**
  *Detailed reasons are shown below (multiple selections allowed).

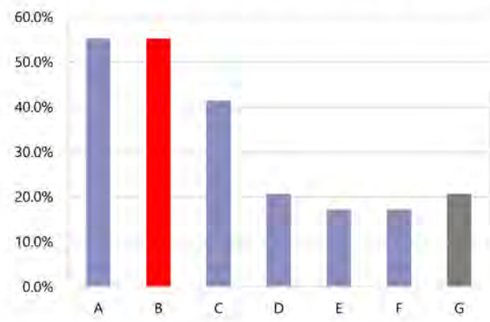


A. The initial ID and related information were unclear
B. Reluctance to register an email address
C. Security concerns
D. The survey materials were difficult to understand
E. The password was forgotten
F. Data could not be submitted
G. On-screen explanations were difficult to understand
H. Operational errors occurred, making it cumbersome to complete the response
I. The system was slow to respond
J. Error messages for missing required entries made the response process cumbersome
K. Other (paper-based response is faster, etc.)
L. Other (logging in is cumbersome, etc.)
M. Other (not listed above)

Transition to Online Survey

Feedback from Respondents Using Paper (Direct Survey)

- Q1. What were the reasons for not responding online?
  (Excluding cases where no online-capable device was available)
  -> Did not consider online response at all. **[45.3%]**
  *Detailed reasons are shown below (multiple selections allowed).

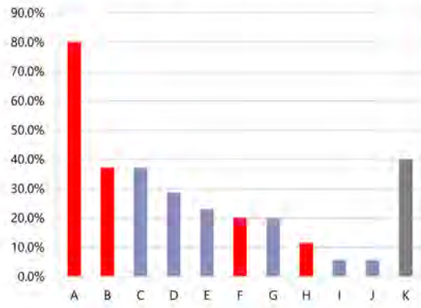


A. Usually respond using paper questionnaires
B. It is easier to grasp the overall content
C. It was difficult to operate in previous surveys
D. Security concerns
E. It is easier to respond while checking information on a screen
F. Paper questionnaires are easier to share internally
G. Other

Transition to Online Survey

Feedback from Respondents Using Paper (Direct Survey)

- Q1. What were the reasons for not responding online?
  (Excluding cases where no online-capable device was available)
  -> Considered online response but chose a paper-based. **[54.7%]**
  *Detailed reasons are shown below (multiple selections allowed).

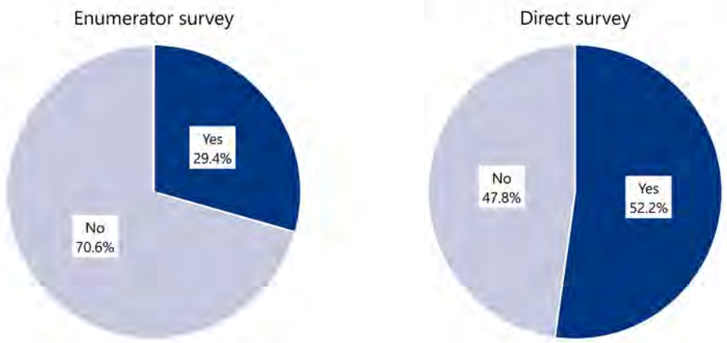


A. Data could not be submitted
B. On-screen explanations were difficult to understand
C. Operational errors occurred, making it cumbersome to complete the response
D. Error messages for missing required entries made the response process cumbersome
E. The system was slow to respond
F. The initial ID and related information were unclear
G. Security concerns
H. The survey materials were difficult to understand
I. The password was forgotten
J. Reluctance to register an email address
K. Other

Transition to Online Survey

Feedback from Respondents Using Paper

- Q2. If the online response method is improved, would you be inclined to respond online in the next survey?



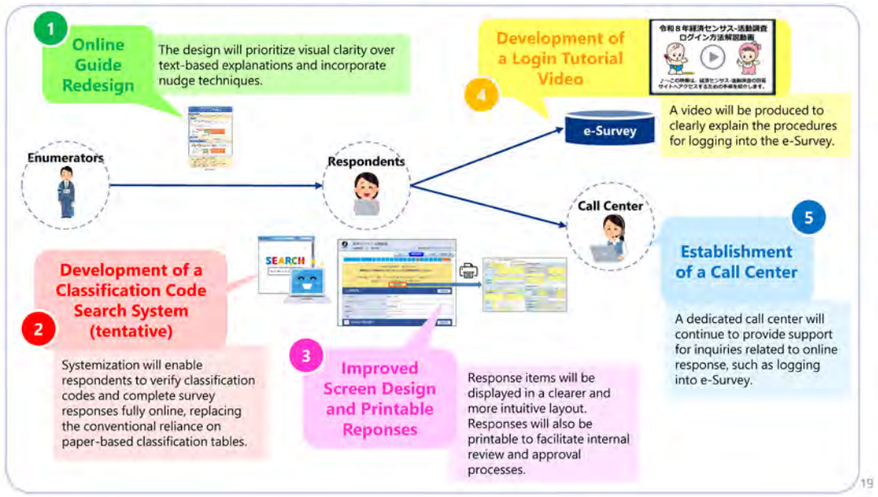
Transition to Online Survey

Approaches Based on the Results of the Pilot Survey

Reasons for choosing paper questionnaire	Enumerator Survey	Direct Survey
1. Preference for paper questionnaires that allow respondents **to grasp the overall content**	Paper-based materials explaining the survey item will be provided when survey documents are mailed.	
2. **Respondents have willingness** to respond online, but difficulties during the process leads to choosing paper questionnaires	In addition to the measures described on the next slide, clearer explanation in online guides and enhanced FAQs will be provided.	
3. A certain number of companies **do not support Excel macros** (Direct Survey only)		HTML-based electronic questionnaires will be newly introduced.

Transition to Online Survey

Major Approaches in 2026 Economic Census



Contents

- Introduction
- Overview of Japan's Economic Census 2026
- Transition to Online Survey
- **Implementation of Business Profiling**
- Summary

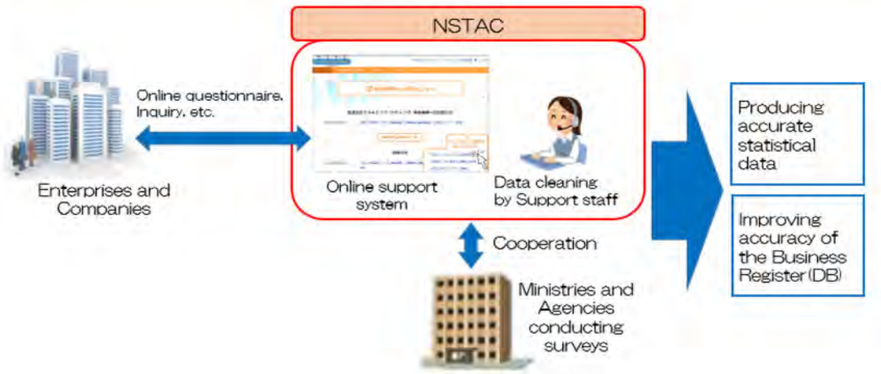
Implementation of Business Profiling

Overview of Business Profiling in Business Surveys

- The National Statistics Center (NSTAC) conducts the **"Enterprise Survey Support Program"** as business profiling activity.
  * NSTAC is an Agency Engaged in Administrative Execution, whose staff are national public servants.
- This program aims to **support enterprises in responding to official statistical surveys** and thereby **contribute to the production of accurate statistics**.
- As part of the program, ongoing developments such as corporate mergers and divisions are routinely captured, and survey response support is provided through a dedicated information system, which enables two-way communication between enterprises and NSTAC.

Implementation of Business Profiling

Illustration of Enterprise Survey Support Program



Implementation of Business Profiling

Target Enterprises and Staffing Framework

- Target: Enterprises frequently subject to complete enumeration in surveys and therefore, bear a relatively heavy responding burden
  – Submit securities reports and related documents pursuant to the Financial Instruments and Exchange Act (Act No. 25 of 1948)
  – Have annual sales of 100 billion yen or more (and in the case of incorporated companies, paid-in capital of 200 million yen or more)
  – Are mutual companies
  -> Approximately 5,000 enterprises
- Support Staff Allocation: **35 staff members**
  – Support staff are organized into five teams, with target enterprises assigned to each team based on attributes such as industry classification and corporate group affiliation.

Implementation of Business Profiling

Past Achievements (Annual Surveys)

Year	Number of Target Enterprises	Statistical Surveys Subject to Support	Response Rate	Online Response Rate
2018	Preparation		-	-
2019	3,000 in the service industry	Annual Business Survey	98.5%	58.1%
2020	3,000 in the service industry	Annual Business Survey	99.1%	67.9%
2021	**5,000 in all industry**	**Economic Census for Business Activity**	**99.0%**	**76.5%**
2022	5,000 in all industry	Annual Business Survey Survey of Research and Development Basic Survey of Japanese Business Structure and Activities	99.4%	82.2%
2023	5,000 in all industry	Annual Business Survey Survey of Research and Development Basic Survey of Japanese Business Structure and Activities	100.0%	84.8%
2024	5,000 in all industry	Annual Business Survey Survey of Research and Development Basic Survey of Japanese Business Structure and Activities	100.0%	87.9%
2025	5,000 in all industry	Annual Business Survey Survey of Research and Development Basic Survey of Japanese Business Structure and Activities	99.9%	89.5%
2026	**5,000 in all industry**	**Economic Census for Business Activity**	-	-

24

Summary

- **Preparations are currently underway** for the upcoming survey administration.
- In this survey, **materials for online responses are distributed in advance** in order to promote online participation.
- At the same time, **the user interface of the online response tools will be improved**, and **support measures – including enhanced guidance, instructional videos, and call center services – will be strengthened**.
- **Profiling activities for large enterprises** will also continue.

->These efforts are expected to reduce respondent burden and contribute to improved accuracy of survey results.

Contents

- Introduction
- Overview of Japan's Economic Census 2026
- Transition to Online Survey
- Implementation of Business Profiling
- **Summary**

25

Thank you for your attention!

経済センサス
活動調査

総務省 Ministry of Internal Affairs and Communications
経済産業省 Ministry of Economy, Trade and Industry

Session 03-1

국가데이터 시대의 공식통계 : 데이터 통합, 활용, 그리고 안전한 연계

좌장

정성규 교수(서울대학교 통계학과)

1. 빅데이터 결합을 통한 개인사업자 금융지원 정책 효과 분석

   김미루 팀장(한국개발연구원 국채연구팀)

2. AI 시대의 조사데이터 통합: 가중치 보정의 역할과 가능성

   권용현 교수(육군사관학교 수학과)

3. 안전한 데이터 연계와 활용을 위한 플랫폼 운영모델 제안

   박민정 과장(국가데이터처 통계기준과)

4. 행정자료 기반 주거비지수 작성방안

   홍현정 서기관(국가데이터처 물가동향과)

토론

박승환 교수(강원대학교 정보통계전공)

빅데이터 결합을 통한 개인사업자 금융지원 정책 효과 분석

김미루(KDI)

제15회 국가통계 · 데이터 발전포럼

May. 14, 2026

Table of Contents

1 서론
소상공인 지원정책 소개
연구 모형 및 자료의 소개
실증분석 결과
요약 및 시사점
부록1. 소상공인 대상 보증, 융자 지원사업 내용
부록2. 신종 코로나바이러스(CV) 관련 소상공인 지원
부록3. 성향점수 매칭 전후 성향점수 분포
부록4. 강건성 검증 실증분석 결과

연구 배경 및 목적

연구 배경

- 본 발표자료는 개인사업자 금융지원 정책의 효과 분석(2020; 오윤해, 김미루, 배준형)에 기반함.
- 소상공인 금융애로 해결을 위한 정부의 다양한 정책자금 지원.
- 2020년「소상공인 기본법」제정: 소상공인을 독립적 법적 지원 대상으로 명시.
- 코로나19로 인한 금융지원 확대:
  - 초저금리 특별융자: 26.4조 원
  - 긴급자금 전액 보증: 3조 원

연구 목적

- 소상공인 개인사업자 대상 금융지원 효과 분석.
- 기존 방식 개선 및 지속 가능한 정책금융 지원방안 구축.

연구 방법 및 자료

연구 대상 및 한계

- 대상: 소상공인이면서 개인사업자인 영세 자영업자.
- 한계 극복:
  - 기존 설문조사 기반 연구의 대표성 부족 보완.
  - 소상공인 개인사업자의 신용상태 및 지역 경쟁환경 고려.

분석 자료

- 통계청-KCB 개인사업자 DB:
  - 개인사업체 전수 데이터.
  - 대표 사업주의 신용정보 포함.
  - 지역별 업종 경쟁도 정보 제공.

연구 기대 및 구성

주요 분석 항목

- 정책 효과:
  - 매출 및 고용 증대.
  - 폐업 방지 효과.
- 부정적 영향 여부:
  - 신용도 저하.
  - 부채상환 부담 증가.
- 폐업 사업체 분석:
  - 폐업 전후 대표의 신용도 변화.

연구 구성

- **제2절**: 선행연구와 본 연구 차별점.
- **제3절**: 소상공인 대상 정책지원의 종류 및 성격.
- **제4절**: 연구 가설, 자료, 계량 분석 모형.
- **제5절**: 실증분석 결과.
- **제6절**: 요약 및 정책적 시사점.

Table of Contents

1 서론
2 소상공인 지원정책 소개
3 연구 모형 및 자료의 소개
4 실증분석 결과
5 요약 및 시사점
부록1. 소상공인 대상 보증, 융자 지원사업 내용
부록2. 신종 코로나바이러스(CV) 관련 소상공인 지원
부록3. 성향점수 매칭 전후 성향점수 분포
부록4. 강건성 검증 실증분석 결과

소상공인 지원 예산의 추이

<표 4-1> 중소벤처기업부 소상공인 지원정책 지출의 연도별 추이

(단위: 백만원)

항목 \ 연도	2013	2014	2015	2016	2017	2018	2019	2020(안)
소상공인시장진흥기금 지출 총계	**1,746,957**	**2,174,342**	**2,472,360**	**2,435,519**	**2,917,318**	**2,813,371**	**3,211,079**	**3,308,706**
지역신용보증	**110,000**	**54,800**	**96,000**	**60,000**	**53,550**	**35,620**	**46,850**	**101,114**
소상공인지원융자	**1,050,000**	**916,500**	**1,819,600**	**1,764,800**	**2,253,960**	**2,068,000**	**2,212,500**	**2,300,000**
소상공인지원융자	1,050,000	916,500	**1,809,500**	**1,755,000**	2,245,000	2,057,539	2,212,500	
소상공인이차보전			10,100	9,800	8,960			
소상공인지원	**87,492**	**74,168**	**130,159**	**119,998**	**122,747**	**135,876**	**207,884**	**228,440**
소상공인창업지원	10,752	10,252	**30,539**	**19,000**	13,130	17,189	22,617	20,891
소상공인성장지원	62,599	49,675	48,584	44,871	52,790	56,690	68,967	87,541
소공인특화지원	2,100	2,800	34,750	34,750	35,520	34,005	41,894	49,344
소상공인재기지원	3,000	3,000	6,100	14,070	13,500	14,218	56,460	44,188
소상공인지원인프라	9,041	8,441	10,186	7,307	7,807	14,204	17,946	21,361
소상공인자영업자를 위한 생활혁신	-	-	-	-	-	-	-	3,465
소상공인스마트상점기술보급	-	-	-	-	-	-	-	1,650
전통시장활성화지원(기금)	**74,662**	**76,336**	**255,796**	**280,087**	**291,984**	**332,339**	**421,107**	**470,485**
시장경영혁신지원	74,662	76,336	**159,312**	**180,025**	192,176	214,951	274,234	328,112
전통시장주차환경개선	-	-	96,484	100,062	99,808	117,388	146,873	142,373

주: 1) 2020년은 기존 예산안을 의미하며 2020년을 제외하고는 추경을 반영한 수치임.
2) 전통시장활성화지원(기금)의 예산명은 2018년 이전에는 소상공인경영개선지원(기금)으로 분류되었음.
자료: 중소벤처기업부, 「예산 및 기금운용계획 사업설명 자료」, 2015-2020을 취합하여 저자 작성.

소상공인 대상 금융지원: 지역신용보증재단(cnt.)

<표 4-2> 신용보증재단중앙회의 재보증 지원 실적

연 도	재보증 지원잔액 실적
2013	675,841개 업체에 6조 8,668억원 지원
2014	622,714개 업체에 6조 9,620억원 지원
2015	725,034개 업체에 8조 533억원 지원
2016	790,606개 업체에 8조 9,113억원 지원
2017	866,811개 업체에 9조 3,663억원 지원
2018	919,737개 업체에 10조 697억원 지원
2019	1,022,478개 업체에 11조 3,684억원 지원

자료: 중소벤처기업부, 「예산 및 기금운용계획 사업설명 자료」, 2017, 2020을 취합하여 정리.

소상공인 대상 금융지원: 지역신용보증재단(cnt.)

<표 4-3> 신용보증재단중앙회의 기업규모별 보증 실적

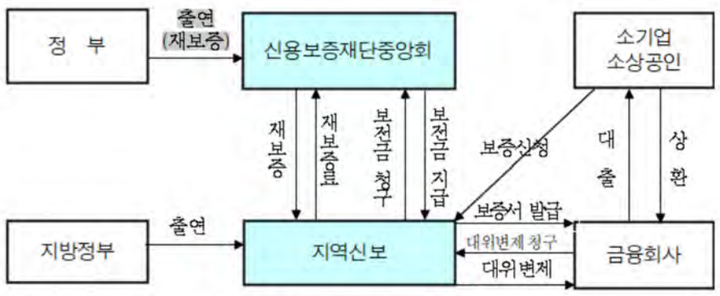
(단위: 억원, 개)

구 분		계	소상공인	소기업	중기업
보증공급	업체 수	464,329	456,504	6,555	1,270
	금액	111,849	100,564	8,934	2,351
보증잔액	업체 수	910,813	896,596	12,069	2,148
	금액	179,713	167,220	9,901	2,592

자료: 신용보증재단중앙회, 『2017년 신용보증제도 안내책자』, 2017.

소상공인 대상 금융지원: 지역신용보증재단

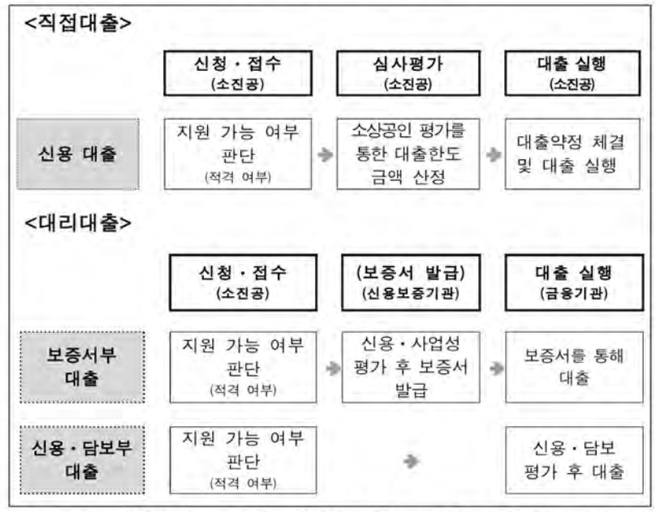
[그림 4-1] 신용보증재단의 보증 및 재보증 절차



자료: 중소벤처기업부, 「2020년 예산 및 기금운용계획」, p.1645.

소상공인 대상 금융지원: 소상공인시장진흥공단

[그림 4-2] 소상공인시장진흥공단 정책자금 융자사업의 절차



자료: 중소벤처기업부, 「2018년 중소벤처기업부 소관 소상공인 정책자금 융자계획」, 2018. 7. 3.

소상공인 지원 경상사업

<표 4-4> 소상공인 지원 경상사업의 세부 사업

항목	세부 사업 목록
소상공인 창업 지원	신사업창업사관학교, 소상공인사관학교, 소상공인창업교육, 소자본해외창업지원, 생활혁신형창업지원, 소상공인사이버평생교육원,신사업육성지원, 상권정보시스템 운영
소상공인 성장 지원	소상공인경영교육, 소상공인 역량강화, 소상공인 협업 활성화, 중소슈퍼 지원, 소상공인컨설팅, 소상공인협동조합 활성화, 나들가게 성과확산, 중소유통 통합구매, 상생협력프랜차이즈 육성, 업종별 경쟁력 강화, 소상공인 온라인 판로지원, 혁신형 소상공인 육성
소상공인 재기 지원	희망리턴패키지, 재창업패키지, 소기업・소상공인공제, 1인 자영업자 고용보험료 지원
소공인특화 (기술) 지원	소공인특화지원센터 설치・운영, 공공기반시설 구축・운영, 성장희망사다리, 제품・기술 가치 향상
소상공인 지원 인프라	소상공인 정보인프라 구축, 소상공인 네트워크 구축, 소상공인 정책・조사연구, 정책자금지원성과 향상, 소상공인연합회 지원, 공예산업 지원

자료: 중소벤처기업부, 「예산 및 기금운용계획 사업설명 자료」, 2015~2020을 취합하여 저자 작성.

Table of Contents

1. 서론
2. 소상공인 지원정책 소개
3. 연구 모형 및 자료의 소개
4. 실증분석 결과
5. 요약 및 시사점
6. 부록1. 소상공인 대상 보증, 융자 지원사업 내용
7. 부록2. 신종 코로나바이러스(CV) 관련 소상공인 지원
8. 부록3. 성향점수 매칭 전후 성향점수 분포
9. 부록4. 강건성 검증 실증분석 결과

연구 가설 설정(cnt.)

가설 1: 정책금융이 사업성과에 미치는 영향

- 정책금융 지원은 금리부담과 자금애로 해소로 긍정적 사업성과를 유도.
- 분석 대상: 매출액 증가, 고용인원 증가, 폐업 방지.

가설 1: 정책금융을 지원받은 소상공인 개인사업자는 비수혜업체에 비해 긍정적인 사업성과를 보인다.

가설 2: 경쟁압력이 높은 지역에서의 정책금융 효과 감소

- 동종업체 비율이 높은 지역에서는 경쟁도가 높아 정책금융 효과가 감소할 가능성.

가설 2: 지역 내 동종업체 비율이 높을수록 정책금융 지원 효과는 감소한다.

연구 가설 설정

가설 3: 사업주 신용 상태에 따른 정책금융 효과 차이

- 낮은 신용등급 및 높은 부채상환비율 사업체:
  - 효과 감소: 정책금융이 부채 청산에 주로 사용.
  - 효과 증가: 신용제약 해소로 생산성 증가.
- 이론적 근거: Banerjee and Duflo(2014).

가설 3: 정책금융 효과는 신용도가 낮거나 부채상환비율이 높을수록 증가한다.

정책금융과 대표자 신용

가설 4: 정책금융이 대표자 개인 신용에 미치는 영향

- 긍정적: 금리 부담 완화 → 신용도 향상.
- 부정적: 부채 부담 증가 → 신용도 악화.

가설 4: 정책금융 수혜 사업주는 비수혜 사업주에 비해 신용도 및 부채상환비율이 악화된다.

가설 5: 폐업 사업체에 대한 정책금융 효과

- 정책금융 지원이 폐업 시점을 지연시키고, 부채 규모를 증가시킬 가능성.
- 분석 대상: 폐업 당시 대표자의 신용 악화 및 부채 수준 변화.

가설 5: 폐업 사업체 중 정책금융 수혜 사업주는 비수혜 사업주보다 신용 악화 및 부채 증가가 더 크다.

연구자료의 소개(cnt.)

데이터 구성 및 출처

- **자료 결합**:
  - 통계청 비임금근로자(개인사업자) DB
  - 중소벤처기업부 중소기업지원이력 DB (SIMS)

특이점

- 기존 연구에서 활용되지 않은 **새로운 데이터베이스**
- 통계청 빅데이터 통계과와 협업하여 상세 분석 가능
- 중소기업 지원 및 개인사업자 경영 분석 가능

연구자료의 소개(cnt.)

통계청 비임금근로자(개인사업자) 부채 DB

- 통계청 개인사업자 DB는 새로운 데이터베이스로 통계청 빅데이터 통계과에서 2018년도부터 기업등록부와 KCB 개인신용정보를 결합하여 개인사업자의 경영과 신용 상황을 결합하여 구축
- 통계청에서는 해당 자료를 기반으로 2020년 10월 비임금근로자의 부채 수준을 발표
- 해당 데이터베이스에는 기업등록부상에서 제공되는 개인사업자의 경영정보와 대표자의 신용정보가 포함됨.
  - **경영정보:** 업종, 매출액, 고용상태, 폐업 여부 등
  - **대표자의 신용정보**: 대출부문별 대출 금액, 연체 여부, 신용등급 등

연구자료의 소개(cnt.)

중소기업통합관리시스템(SIMS) DB

- 중소벤처기업부를 통해 중소기업통합관리시스템(SIMS)의 정책지원이력 데이터베이스를 구득하여 개인사업자 DB와 사업자번호를 기준으로 결합
- 해당 결합자료를 통해 지역신용보증재단, 소상공인시장진흥공단 등에서 정책금융 및 인력, 수출, 경영 지원 등 다양한 지원을 받은 사업체의 지원 이력을 구분
- 본 연구는 다양한 소상공인 지원정책 중 비율이 가장 높은 정책금융 지원 수혜업체를 구분
- 2018년 통계청 기업등록부와 중소기업지원이력 DB(SIMS)를 결합하여 소상공인 중 정책지원을 받은 사업체의 비중을 계산
  - 정책지원 방식에는 경영, 금융, 인력 지원 외에도 기술(개발)/수출/내수(판로)/동반성장을 지원하는 경우도 포함
  - 그러나 해당 지원의 수혜 사업 중에는 영세 사업체의 비중이 작고 상대적으로 규모가 큰 사업체의 비중이 커서 기술/수출/내수/동반성장 지원 수혜 사업체는 분석에서 제외
  - 또한 창업 관련 정책지원을 수혜받은 사업체의 비율도 높았지만, 창업 사업체의 경우에는 다른 성격을 보일 것으로 예상되어 분석에서 제외함.

연구자료의 소개

<표 4-5> 2018년 소상공인·개인사업자 중 분야별 정책 수혜업체의 비중

(단위: %)

분 야	지원횟수		
	0회	1회	2회 +
경영	99.82	0.17	0.01
금융	95.03	2.80	2.16
기술	99.97	0.03	0.01
내수	99.96	0.02	0.02
동반성장	100	0.00	0.00
수출	99.94	0.04	0.02
인력	99.51	0.26	0.23
창업	99.81	0.18	0.01
전 분야	94.34	3.15	2.52

주: 2018년 통계청 기업등록부에 포함된 개인사업자 중 소상공인으로 분류되는 전체 사업체 537만 7,260개 중 중소기업통합관리시스템상에서 지원이력이 관찰되는 사업체 수의 비중을 특정하였음.
자료: 통계청의 2018년 기업등록부, 중소벤처기업부의 SIMS 결합 자료를 통해 저자 작성.

연구의 대상 및 시기(cnt.)

연구 대상

- 영세 자영업자에 해당하는 업체를 구분하기 위해, 대표사업주가 1개의 사업체만을 운영할 경우로 한정
  - 2015~17년에 생존하였던 연매출 5억원 미만의 사업체로 구성
  - 지원금액이 극단적으로 큰 관측치가 존재함에 따라 대출금액이 6천만원 이상인 경우도 표본에서 제외

정책자금 수혜업체

- 2012년 이후 정책자금을 수혜받기 전까지 여타 정책지원 이력이 전무한 사업체 중 2016년~17년에 정책금융을 이용한 업체를 선정
  - 신용보증재단의 보증지원을 통해 대출을 이용한 비중이 67.16%로 수혜업체의 대다수를 차지
  - 소상공인시장진흥공단을 통한 소상공인지원(융자)이 13.7%로 많은 비중을 차지
  - 기술보증기금(0.5%), 신용보증기금(1.8%)과 같이 중소기업 대상의 정책자금 지원을 이용한 업체도 발견
  - 서울시, 경기도와 같이 지자체에서 지원된 금융지원 사업도 표본에 포함되어 있음을 확인

정책자금 비수혜업

- 2012~18년에 어떠한 형태의 정책지원도 받지 못한 업체로 제한
  - 비수혜업체 중 정책지원을 받기 이전인 2015~16년 정보를 기준으로 성향점수 매칭기법을 통해 수혜업체와 유사한 대조군 비수혜업체를 구성
  - 따라서 연구의 표본은 2016년~17년에 직접융자, 이차보전, 신용보증 등의 금융지원을 수혜받은 업체와 정책 수혜 이력이 없지만 2015년과 2016년 기준에서 수혜 사업체와 가장 유사한 비수혜업체로 구성

연구의 대상 및 시기(cnt.)

<표 4-6> 2017년 수혜업체의 정책금융 세부 사업내역

분야	세부 사업명	사업체 수	비중
금융	경기도일자리창출 특별협약보증(자체/직접)	8	0.08
금융	경기취약소상공인 자금지원(자체/직접)	31	0.31
금융	경영안정화긴급자금지원	7	0.07
금융	기술보증기금출연	55	0.54
금융	매출채권보험계정출연	8	0.08
금융	소상공인경영안정지원	32	0.32
금융	소상공인육성자금이차보전	53	0.52
금융	소상공인 육성자금 지원업체 이차보전	52	0.51
금융	소상공인 정책자금 이차보전금 지원	25	0.25
금융	소상공인지원	4	0.04
금융	소상공인지원(융자)	1,385	13.69
금융	소상공인 창업 및 경영안정융자금 이차보전	97	0.96
금융	시중은행 협력자금 지원	484	4.78
금융	신용보증기금출연	183	1.81
금융	이차보전 및 융자지원	103	1.02
금융	중소기업경영안정자금지원	143	1.41
금융	중소기업금융지원	45	0.44
금융	중소기업운전자금지원	44	0.43
금융	중소기업육성자금이차보전지원	124	1.23
금융	중소기업자금지원이자보전	145	1.43
금융	중소기업직접융자	115	1.14
금융	지역신용보증재단재보증	6,796	67.16
금융	창업기업자금(융자)	22	0.22

자료: 통계청의 2018년 기업등록부, 중소벤처기업부의 SIMS 결합자료를 통해 저자 작성.

연구의 대상 및 시기

<표 4-7> 2017년 수혜업체의 정책금융 세부 사업내역(3건 이하 사업)

분 야	세부 사업명
금융	가스안전관리(융자)
금융	서울형마이크로크레딧지원
금융	에너지절약시설설치
금융	위생시설개선자금융자사업
금융	이자차액보전금지원
금융	재도약지원자금(융자)
금융	운전자금 이차보전 지원사업
금융	중소기업 경영안정자금 및 창업자금지원
금융	중소기업 경영안정사업 자금지원
금융	중소기업 육성자금 이차보전금지원

주: 통계청 빅데이터센터의 비식별 조치에 따라 관측치 3개 이하 사업장이 존재하는 경우 지원정책 수혜 사업체의 실제 비율은 나타내지 못하였음.
자료: 통계청의 2018년 기업등록부, 중소벤처기업부의 SIMS 결합 자료를 통해 저자 작성.

매칭을 통한 대조군의 선정(cnt.)

- 소상공인에 대한 정책자금 공급의 인과적 효과를 분석하기 위해 실험군(정책자금 수혜업체)과 대조군(비수혜업체)의 전·후 성과를 비교하는 이중차분기법을 적용
- 대조군을 선정하는 방식으로는 성향점수 매칭기법(Propensity Score Matching Estimation)을 적용
  - 먼저 로짓(logit) 모형 등으로 사업체가 정책자금 지원대상에 선정될 확률을 관찰된 특성을 통해 설명하는 모형을 산출
  - 이후 각 사업체의 성향점수를 계산한 후 정책자금 수혜업체와 가장 근접한 성향점수를 가진 사업체를 대조군으로 짝짓게 됨.
- 성향점수를 결정하는 특성변수로 2016년 사업체의 특성정보, 사업체 대표자의 특성 및 신용정보를 활용
  - **사업체의 특성정보(2016년 기준):** 연 매출액, 총 종사자 수, 업종 세세분류, 사업체 소재지의 시군구 더미, 복식부기 작성 여부
  - **대표자 특성:** 연령, 성별
  - **신용정보 및 부채구조:** 신용등급, 신용점수, 소득 대비 부채상환비율(DSR), 가계대출잔액 합, 기업대출잔액 합, KCB 연소득, 신용체크카드 사용금액(연간), 주담대/가계대출 비율, 신용대출/가계대출 비율, 비은행권 가계대출의 비중
- 자료 활용 및 매칭 결과
  - 대표자의 신용 및 부채구조 포함: 선택편의 감소
  - 관찰 가능한 변수로 매칭, PSME 분석 적합성 확보

매칭을 통한 대조군의 선정(cnt.)

<표 4-8> 대조군과 실험군의 평균 차이 t-test

주요 변수	수혜업체 평균	비수혜업체 수혜업체 평균 차이	T-test P-value
매출액(만원)	6710.28	0.15	0.91
연령	44.81	-0.20	0.17
고용인원(명)	0.16	0.00	0.96
복식부기 작성 여부 (복식=1, 간편=0)	0.11	0.00	0.32
성별(남성=1)	0.46	0.01	0.48
신용등급(1~10, 0)	4.23	0.03	0.33
신용점수	773.97	-2.87	0.15
DSR(소득 대비 상환 비율)	0.21	0.01	0.38
가계부채량(천원)	49442.67	104.64	0.94

자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

매칭을 통한 대조군의 선정(cnt.)

<표 4-9> 특성변수 선정 방식별 대조군과 실험군의 평균 차이 t-test

특성 정보	매칭방식 1 (주담대 비율, 신용대출 비율, 비은행 가계대출 비중, 주소 정보 제외)		매칭방식 2 (주담대 비율, 신용대출 비율, 대표자 성별, 신용점수, KCB 연소득, 비은행 가계대출 비중, 주소 정보 제외)		매칭방식 3 (주담대 비율, 신용대출 비율, 비은행 가계대출 비중 제외)	
변수	평균 차이	T-test P-value	평균 차이	T-test P-value	평균 차이	T-test P-value
매출액	-3.29	0.01	-3.58	0.00	-2.34	0.06
연령	-0.05	0.73	0.01	0.94	-0.23	0.14
고용인원	-0.02	0.03	-0.02	0.07	-0.02	0.04
복식부기 더미	0.00	0.42	0.00	0.20	0.00	0.73
성별	0.00	0.61	0.00	0.77	0.00	0.92
신용등급	0.04	0.19	0.06	0.06	0.02	0.46
신용점수	-2.84	0.15	-3.61	0.07	-1.41	0.48
DSR	0.01	0.36	0.00	0.62	0.00	0.89
가계부채량	418.71	0.76	969.15	0.47	-520.98	0.68
기업부채량	-210.58	0.80	605.44	0.47	679.72	0.44
KCB 소득	193.59	0.31	-89.98	0.64	-76.28	0.67
Pseudo R^2	0.050		0.057		0.065	

자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

매칭을 통한 대조군의 선정

<표 4-9> 특성변수 선정 방식별 대조군과 실험군의 평균 차이 t-test

특성정보	매칭방식 4 (비은행 가계대출 비중 제외, 주소 시도 더미 활용)		매칭방식 5 (비은행 가계대출 비중 제외)		매칭방식 6 (본 모형)	
변수	평균 차이	T-test P-value	평균 차이	T-test P-value	평균 차이	T-test P-value
매출	-1.79	0.16	-1.23	0.34	0.15	0.91
연령	-0.09	0.54	-0.22	0.14	-0.20	0.17
고용인원	-0.01	0.23	0.00	0.84	0.00	0.96
복식부기더미	0.00	0.92	0.00	0.65	0.00	0.32
성별	-0.01	0.10	0.01	0.48	0.01	0.48
신용등급	0.02	0.47	0.01	0.80	0.03	0.33
신용점수	-1.70	0.39	-0.85	0.67	-2.87	0.15
DSR	0.01	0.14	0.01	0.20	0.01	0.38
가계부채량	161.00	0.90	1260.56	0.33	104.64	0.94
기업부채량	1021.27	0.24	629.18	0.46	-638.16	0.41
KCB 소득	-5.94	0.97	-196.22	0.27	15.50	0.93
Pseudo R^2	0.067		0.069		0.089	

자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

분석 표본: 지역 및 업종 분포

지역별 분포 (<표 4-10>)

- 수도권: **37.6%** (서울 15.5%, 경기도 22.1%)
- 경상도: 경남 8.7%, 경북 6.4%

매출액 분포 ([그림 4-3])

- 대부분의 사업체가 연매출 **1억 원 이하** 영세 사업체

업종별 분포 (<표 4-11>)

- 도매 및 소매업: **24.18%**
- 숙박 및 음식점업: **22.96%**
- 협회 및 단체, 수리 및 기타 개인서비스업: **12.78%**
- 부동산업: **10.61%**

종속변수 및 설명변수 요약 통계

종속변수 요약 (<표 4-12>)

- **매출액 증감**:
  - 2017년 수혜업체: **5,629만 원** (비수혜업체: 3,227만 원)
- **고용인원 증감**:
  - 수혜업체: **0.1명**, 비수혜업체: 0.065명
- **폐업률**:
  - 수혜업체: **28%**, 비수혜업체: 30%
- 신용점수, DSR 변화 및 연체 여부도 종속변수로 활용

설명변수 요약 (<표 4-13>)

- **2017년 정책자금 수혜업체 기준**:
  - 사업주 평균 연령: **45세**, 사업체 업력: **2.6년**
  - 남성 사업주 비율: **46% (실험군)**, **47% (대조군)**
- 간편장부 대상자 비율(소상공인 중 75%) 및 복식부기 작성업체 비율은 두 집단에서 유사

분석 표본의 특징(cnt.)

<표 4-10> 표본 사업체의 광역시도 분포

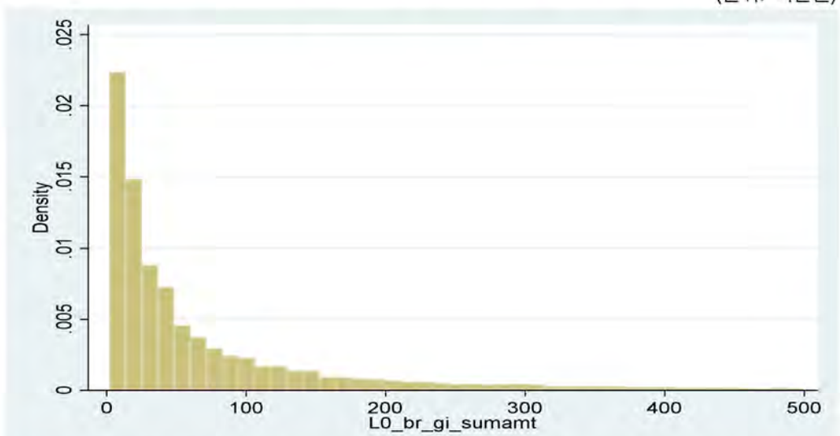
지 역	비중(%)
서울특별시	15.46
부산광역시	6.99
대구광역시	4.78
인천광역시	3.43
광주광역시	2.79
대전광역시	3.33
울산광역시	2.63
세종특별자치시	0.34
경기도	22.07
강원도	4.14
충청북도	3.59
충청남도	4.83
전라북도	3.75
전라남도	4.72
경상북도	6.36
경상남도	8.72
제주특별자치도	2.07

자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

분석 표본의 특징(cnt.)

[그림 4-3] 2017년 정책금융 수혜업체 및 대조군의 2016년 매출액 분포

(단위: 백만원)



자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

분석 표본의 특징

<표 4-11> 업종별(대분류) 사업체 분포

산업 대분류	비중
농업, 임업 및 어업	0.70
제조업	6.69
전기, 가스, 증기 및 공기 조절 공급업	0.03
수도, 하수 및 폐기물 처리, 원료재생업	0.07
건설업	4.14
도매 및 소매업	24.18
운수 및 창고업	3.15
숙박 및 음식점업	22.96
정보통신업	0.72
금융 및 보험업	0.31
부동산업	10.61
전문, 과학 및 기술 서비스업	2.67
사업시설 관리, 사업 지원 및 임대서비스업	1.63
교육서비스업	3.60
보건업 및 사회복지 서비스업	0.39
예술, 스포츠 및 여가 관련 서비스업	5.35
협회 및 단체, 수리 및 기타 개인서비스업	12.78

자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

종속변수 요약통계량

A. 2016년 정책금융 수혜업체 및 대조군

분류		변수	관측치	평균	표준편차
전체 (실험군, 대조군)	사업 성과	2년 후 매출 변화(만원), $\Delta_{t-1,t+2}$	9,810	5,711.6	17,633.3
		2년 후 고용 변화(만원), $\Delta_{t-1,t+2}$	10,425	0.10026	0.8417
		1년 후 매출 변화(만원), $\Delta_{t-1,t+1}$	10,410	5,618.4	19,594.6
		1년 후 고용 변화(명), $\Delta_{t-1,t+1}$	10,710	0.100	0.815
		1년 후 폐업 여부(더미)	10,738	0.085	0.278
	신용 상태	1년 후 신용점수 변화, $\Delta_{t-1,t+1}$	10,728	-2.484	120.99
		1년 후 DSR 변화, $\Delta_{t-1,t+1}$	10,728	0.112	0.772
	연체 여부	1년 후 신용카드 90일 이상 연체(더미)	10,597	0.004	0.061
		1년 후 가계부채 90일 이상 연체(더미)	10,738	0.017	0.128
		1년 후 기업부채 90일 이상 연체(더미)	10,738	0.002	0.046
2016년 대조군	사업 성과	1년 후 매출 변화(만원), $\Delta_{t-1,t+1}$	5,179	4,937.9	15,441.6
		1년 후 고용 변화(명), $\Delta_{t-1,t+1}$	5,358	0.010	0.806
		1년 후 폐업 여부(더미)	5,369	0.078	0.269
	신용 상태	1년 후 신용점수 변화, $\Delta_{t-1,t+1}$	5,369	-0.175	49.445
		1년 후 DSR 변화, $\Delta_{t-1,t+1}$	5,369	0.054	0.645
	연체 여부	1년 후 신용카드 90일 이상 연체(더미)	5,257	0.006	0.074
		1년 후 가계부채 90일 이상 연체(더미)	5,369	0.013	0.114
		1년 후 기업부채 90일 이상 연체(더미)	5,369	0.001	0.031
2016년 실험군	사업 성과	1년 후 매출 변화(만원), $\Delta_{t-1,t+1}$	5,231	6,292.1	22,960.1
		1년 후 고용 변화(명), $\Delta_{t-1,t+1}$	5,352	0.101	0.824
		1년 후 폐업 여부(더미)	5,369	0.091	0.287
	신용 상태	1년 후 신용점수 변화, $\Delta_{t-1,t+1}$	5,369	-9.187	72.863
		1년 후 DSR 변화, $\Delta_{t-1,t+1}$	5,359	0.171	0.877
	연체 여부	1년 후 신용카드 90일 이상 연체(더미)	5,340	0.002	0.045
		1년 후 가계부채 90일 이상 연체(더미)	5,369	0.020	0.141
		1년 후 기업부채 90일 이상 연체(더미)	5,369	0.003	0.058

B. 2017년 정책금융 수혜업체 및 대조군

분류		변수	관측치	평균	표준편차
전체	사업 성과	2년 후 매출 변화(만원), $\Delta_{t-1,t+2}$	18,036	4,680.3	17,216.6
		2년 후 고용 변화(만원), $\Delta_{t-1,t+2}$	19,205	0.0826	0.8234
		1년 후 매출 변화(만원), $\Delta_{t-1,t+1}$	18,036	4,680.3	17,216.6
		1년 후 고용 변화(명), $\Delta_{t-1,t+1}$	19,205	0.083	0.823
		1년 후 폐업 여부(더미)	19,588	0.097	0.300
	신용 상태	1년 후 신용점수 변화, $\Delta_{t-1,t+1}$	19,259	-5.165	126.77
		1년 후 DSR 변화, $\Delta_{t-1,t+1}$	19,259	0.116	0.894
	연체 여부	1년 후 신용카드 90일 이상 연체(더미)	19,406	0.006	0.080
		1년 후 가계부채 90일 이상 연체(더미)	19,588	0.019	0.137
		1년 후 기업부채 90일 이상 연체(더미)	19,588	0.005	0.071
2017년 대조군	사업 성과	1년 후 매출 변화(만원), $\Delta_{t-1,t+1}$	9,001	3,727.7	16,390.8
		1년 후 고용 변화(명), $\Delta_{t-1,t+1}$	9,771	0.065	0.729
		1년 후 폐업 여부(더미)	9,794	0.105	0.307
	신용 상태	1년 후 신용점수 변화, $\Delta_{t-1,t+1}$	9,794	0.547	48.246
		1년 후 DSR 변화, $\Delta_{t-1,t+1}$	9,791	0.0547	0.780
	연체 여부	1년 후 신용카드 90일 이상 연체(더미)	9,640	0.007	0.085
		1년 후 가계부채 90일 이상 연체(더미)	9,794	0.017	0.129
		1년 후 기업부채 90일 이상 연체(더미)	9,794	0.002	0.044
2017년 실험군	사업 성과	1년 후 매출 변화(만원), $\Delta_{t-1,t+1}$	9,035	5,629.3	17,952.3
		1년 후 고용 변화(명), $\Delta_{t-1,t+1}$	9,434	0.101	0.910
		1년 후 폐업 여부(더미)	9,794	0.088	0.284
	신용 상태	1년 후 신용점수 변화, $\Delta_{t-1,t+1}$	9,794	-7.283	62.480
		1년 후 DSR 변화, $\Delta_{t-1,t+1}$	9,468	0.179	0.995
	연체 여부	1년 후 신용카드 90일 이상 연체(더미)	9,766	0.016	0.125
		1년 후 가계부채 90일 이상 연체(더미)	9,794	0.022	0.145
		1년 후 기업부채 90일 이상 연체(더미)	9,794	0.008	0.090

자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

설명변수 요약통계량(cnt.)

A. 2016년도 정책금융 수혜업체 및 대조군

변 수	실험군(5,369)		대조군(5,369)	
	평균	표준편차	평균	표준편차
정책수혜 여부(수혜=1)	1.00	0.00	0.00	0.00
사업주 연령	46.20	10.13	45.89	10.51
사업체 업력	4.33	4.80	4.19	4.16
사업주 남성 더미	0.43	0.49	0.42	0.49
복식부기 기입 여부	0.08	0.28	0.10	0.30
2015년 매출액	5,766.57	7,550.27	6,031.92	8,286.76
2015년 고용인원	0.13	0.51	0.13	0.54
주민등록 인구 대비 동종업종(세세분류) 개수(동별)	0.95	3.34	1.08	3.48
2015년 사업주 신용점수	766.82	133.18	767.46	150.24
2015년 사업주 신용등급	4.33	1.97	4.33	2.18
2015년 사업주 DSR	0.21	0.35	0.20	0.50
2015년 사업주 가계대출잔액	43,841.29	75,794.71	44,814.94	88,803.52
이차보전 지원 더미	0.10	0.30		
지신보보증 지원 더미	0.74	0.44		

설명변수 요약통계량

B. 2017년도 정책금융 수혜업체 및 대조군

변수	실험군(9,794)		대조군(9,794)	
	평균	표준편차	평균	표준편차
정책수혜 여부(수혜=1)	1.00	0.00	0.00	0.00
사업주 연령	44.81	10.25	44.61	10.71
사업체 업력	2.64	3.91	2.56	3.42
사업주 남성 더미	0.46	0.50	0.47	0.50
복식부기 기입 여부	0.11	0.32	0.13	0.33
2016년 매출액	6,710.28	8,450.61	6,994.28	9,060.81
2016년 고용인원	0.16	0.57	0.18	0.62
주민등록 인구 대비 동종업종(세세분류) 개수(동별)	0.83	2.68	0.84	2.74
2016년 사업주 신용점수	773.97	129.94	773.32	150.11
2016년 사업주 신용등급	4.23	1.97	4.23	2.18
2016년 사업주 DSR	0.21	0.32	0.22	0.54
2016년 사업주 가계대출잔액	49,442.67	82,568.73	50,480.56	99,200.97
이차보전 지원 더미	0.12	0.32		
지신보보증 지원 더미	0.68	0.47		

자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

Table of Contents

1 서론
2 소상공인 지원정책 소개
3 연구 모형 및 자료의 소개
4 실증분석 결과
5 요약 및 시사점
부록1. 소상공인 대상 보증, 융자 지원사업 내용
부록2. 신종 코로나바이러스(CV) 관련 소상공인 지원
부록3. 성향점수 매칭 전후 성향점수 분포
부록4. 강건성 검증 실증분석 결과

정책자금의 효과 분석 결과(cnt.)

정책자금의 긍정적 사업성과 (<표 4-14>)

- **매출 증가 효과**:
  - 수혜업체는 비수혜업체에 비해
    - **1년 후**: 매출액 약 **1,880만 원 증가** (28.8%)
    - **2년 후**: 매출액 약 **1,078만 원 증가** (16.5%)
- **고용 증가 효과**:
  - **1년 후**: 고용인원 약 **0.035명 증가** (22.5%)
- **폐업률 감소 효과**:
  - **1년 후**: 폐업 확률 약 **0.9% 감소**

기존 연구와의 비교

- **이진국(2017)**: 정책금융 수혜업체의 매출 증가 보고
- **Banerjee and Duflo(2014)**: 신용제약 상황에서의 정부 대출 효과 확인
- **Adhikary et al.(2019)**: 인도네시아 신용보증 정책성과와 유사한 경향

시간에 따른 효과 변화

- 정책금융의 매출 및 고용 효과는 시간이 지남에 따라 약화

정책자금 효과에 영향을 미치는 요인

설명변수의 영향 (1년 후 기준)

- **사업주 연령 증가 (1세)**:
  - 매출액: **-44만 원**, 고용원: **-0.002명**, 폐업률: **-0.1% 감소**
- **사업체 업력 증가 (1년)**:
  - 매출액: **-424만 원**, 고용원: **-0.01명**, 폐업률: **-0.4% 감소**
- **대표자 성별 (남성)**:
  - 매출액과 고용인원 증가 효과가 더 유의

사업체 특성에 따른 효과

- **복식부기 작성 사업체**:
  - 매출액: **2,955만 원 증가**, 고용원: **0.27명 증가**, 폐업률: **-1.6% 감소**
- **전년도 고용인원 많을수록**:
  - 매출 증가 효과 더 큼 (1년 후: **2,698만 원**, 2년 후: **3,572만 원**)
  - 고용 증가 효과는 더 작음
- **전년도 매출액 클수록**:
  - 매출, 고용 증가 및 폐업 방지 효과 더 큼

정책금융 지원이 사업성과에 미치는 영향: 기본분석

<표 4-14> 정책금융이 영세 자영업자 사업성과에 미친 영향(기본분석)

종속변수	매출액 전·후 변화		고용인원 전·후 변화		폐업 여부
	1년 후 $\Delta_{t-1,t+1}$	2년 후 $\Delta_{t-1,t+2}$	1년 후 $\Delta_{t-1,t+1}$	2년 후 $\Delta_{t-1,t+2}$	1년 후 폐업 (폐업=1)
	(1)	(2)	(3)	(4)	(5)
정책금융 수혜 더미	1,880.15***	1,078.14***	0.035***	0.018	-0.009***
	(211.66)	(345.35)	(0.009)	(0.017)	(0.003)
2017년 정책금융 이용 더미	-1,981.83***		-0.041***		0.006
	(220.432)		(0.010)		(0.004)
사업주 연령	-44.223***	-75.040***	-0.002***	-0.002***	-0.001***
	(10.863)	(18.220)	(0.000)	(0.001)	(0.000)
사업체 업력	-424.475***	-448.578***	-0.010***	-0.008***	-0.004***
	(27.483)	(40.314)	(0.001)	(0.002)	(0.000)
남성 더미	1,303.63***	1,209.803***	0.018*	0.028	-0.006*
	(232.955)	(385.761)	(0.010)	(0.018)	(0.004)
복식부기 더미	2,955.07***	3,351.742***	0.269***	0.209***	-0.016**
	(453.62)	(775.468)	(0.020)	(0.038)	(0.007)
정책금융 이용 전년도 매출액	0.035**	0.058*	0.000***	0.000***	-0.000**
	(0.018)	(0.030)	(0.000)	(0.000)	(0.000)
정책금융 이용 전년도 고용	2,698.28***	3,572.01***	-0.299***	-0.238***	0.004
	(199.46)	(348.692)	(0.009)	(0.017)	(0.003)
상수항	9,115.31***	19,167.31***	0.020	0.098	0.115***
	(2,172.11)	(3,995.44)	(0.090)	(0.168)	(0.032)
관측치	28,427	9,806	29,895	10,420	30,285
R-squared	0.076	0.113	0.054	0.052	0.016

주: 시·도 더미와 소분류 업종 더미를 포함하였으나 추정계수는 생략하였음.
자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

정책금융 지원이 사업성과에 미치는 영향: 지원방식별

<표 4-16> 2017년 정책금융 지원방식별 사업체의 특징

변 수	이차보전		지역신용보증	
	평균	표준편차	평균	표준편차
매출액(만원)	7070.5	8463.9	6471	8246
고용인원	0.19	0.59	0.15	0.56
사업체 업력	2.64	4.14	2.65	3.89
사업주 신용등급	3.89	1.94	4.37	1.98
사업주 DSR	0.19	0.36	0.22	0.31

자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

정책금융 지원이 사업성과에 미치는 영향: 지역 동종업종 경쟁도별

<표 4-17> 지역 동종업권 경쟁도별 정책금융의 영향

종속변수	매출액 전·후 변화		고용인원 전·후 변화		폐업 여부
	1년 후 $\triangle_{t-1,t+1}$	2년 후 $\triangle_{t-1,t+2}$	1년 후 $\triangle_{t-1,t+1}$	2년 후 $\triangle_{t-1,t+2}$	1년 후 폐업 (폐업=1)
	(1)	(2)	(3)	(4)	(5)
정책금융 수혜	2,037.69***	1,232.40***	0.038***	0.022	-0.012***
	(221.01)	(359.86)	(0.010)	(0.017)	(0.004)
정책금융 X 동종업체 비율	-178.95**	-158.25	-0.003	-0.005	0.003***
	(71.38)	(101.03)	(0.003)	(0.005)	(0.001)
인구 대비 동종업체비율 (세세분류)	71.818	46.797	0.002	0.001	-0.002**
	(51.436)	(72.642)	(0.002)	(0.004)	(0.001)
관측치	28,427	9,806	29,895	10,420	30,285
R-squared	0.076	0.114	0.054	0.052	0.017

주: 2017년 표본 더미, 시·도 더미, 소분류 업종 더미, 연령, 업력, 남성, 기존 매출액 및 고용, 복식부기 더미를 포함하였으나 추정계수는 생략하였음.
자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

정책금융 지원이 사업성과에 미치는 영향: 대표자의 신용도별

<표 4-18> 대표자의 신용도별 정책금융 효과

종속변수	매출액 전·후 변화		고용인원 전·후 변화		폐업 여부
	1년 후 $\triangle_{t-1,t+1}$	2년 후 $\triangle_{t-1,t+2}$	1년 후 $\triangle_{t-1,t+1}$	2년 후 $\triangle_{t-1,t+2}$	1년 후 폐업 (폐업=1)
	(1)	(2)	(3)	(4)	(5)
정책금융 수혜	2,721.41***	2,602.23***	0.070***	0.074*	-0.007
	(477.47)	(786.30)	(0.021)	(0.038)	(0.008)
정책금융 X 기존 신용등급	-198.56**	-353.31**	-0.008*	-0.013*	-0.000
	(101.14)	(164.50)	(0.004)	(0.008)	(0.002)
전년도 대표 신용등급	-217.87***	-389.88***	-0.009***	-0.012**	0.006***
	(69.01)	(111.28)	(0.003)	(0.005)	(0.001)
관측치	28,427	28,427	29,895	29,895	30,285
R-squared	0.077	0.077	0.055	0.055	0.018

주: 2017년 표본 더미, 시·도 더미, 소분류 업종 더미, 연령, 업력, 남성, 기존 매출액 및 고용, 복식부기 더미를 포함하였으나 추정계수는 생략하였음.
자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

정책금융 지원이 사업성과에 미치는 영향: 대표자의 부채상환비율(DSR)별

<표 4-19> 대표자의 부채상환비율(DSR)별 정책금융 효과

종속변수	매출액 전·후 변화		고용인원 전·후 변화		폐업 여부
	1년 후 $\triangle_{t-1,t+1}$	2년 후 $\triangle_{t-1,t+2}$	1년 후 $\triangle_{t-1,t+1}$	2년 후 $\triangle_{t-1,t+2}$	1년 후 폐업 (폐업=1)
	(1)	(2)	(3)	(4)	(5)
정책금융 수혜	1,787.711***	1,302.798***	0.030***	0.023	-0.008**
	(239.529)	(384.698)	(0.011)	(0.018)	(0.004)
정책금융 X DSR	428.549	-1,174.744	0.024	-0.034	-0.002
	(537.400)	(845.270)	(0.023)	(0.040)	(0.008)
전년도 DSR	172.716	1,423.404***	0.014	0.072***	-0.006
	(311.977)	(499.079)	(0.013)	(0.023)	(0.005)
관측치	28,427	9,806	29,895	10,420	30,285
R-squared	0.077	0.118	0.055	0.054	0.018

주: 2017년 표본 더미, 시·도 더미, 소분류 업종 더미, 연령, 업력, 남성, 기존 매출액 및 고용, 복식부기 더미를 포함하였으나 추정계수는 생략하였음.
자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

정책금융 지원이 사업주 개인 신용에 미친 영향: 전체 사업체

<표 4-20> 정책금융이 대표자 신용점수·소득 대비 상환비율에 미친 영향 분석

변수	신용점수 변화, $\triangle_{t-1,t+1}$			DSR(소득 대비 상환비율) $\triangle_{t-1,t+1}$		
	(1)	(2)	(3)	(1)	(2)	(3)
정책금융	-24.203***	-17.557***	-25.006***	0.132***	-0.089***	0.097***
	(1.435)	(3.265)	(1.622)	(0.013)	(0.024)	(0.013)
정책금융 X 신용등급		-1.594**			0.050***	
		(0.686)			(0.005)	
전년도 신용등급	-9.480***	11.719***	-9.582***	0.109***	-0.016***	0.066***
	(1.418)	(0.462)	(1.424)	(0.010)	(0.003)	(0.010)
전년도 신용점수	-0.310***		-0.311***	0.002***		0.000**
	(0.021)		(0.021)	(0.000)		(0.000)
정책금융 X DSR			3.782			0.173***
			(3.565)			(0.023)
전년 DSR			-0.826			-0.728***
			(1.943)			(0.012)
관측치	29,948	29,948	29,948	30,077	30,077	30,077
R-squared	0.056	0.049	0.056	0.016	0.016	0.136

주: 2017년 표본 더미, 시·도 더미, 소분류 업종 더미, 연령, 업력, 남성, 기존 매출액 및 고용, 복식부기 더미를 포함하였으나 추정계수는 생략하였음.
자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

정책금융 지원이 사업주 개인 신용에 미친 영향: 전체 사업체

<표 4-21> 정책금융이 1년 후 가계 및 기업부채 연체에 미친 영향 분석

변수	1년 후 신용카드 90일 이상 연체		1년 후 가계부채 90일 이상 연체		1년 후 기업부채 90일 이상 연체	
	(1)	(2)	(3)	(4)	(5)	(6)
정책금융	-0.002***	0.008***	0.007***	0.006	0.003***	-0.007***
	(0.001)	(0.002)	(0.002)	(0.004)	(0.001)	(0.002)
정책금융 X 신용등급		-0.002***		0.000		0.002***
		(0.000)		(0.001)		(0.000)
전년도 신용등급	-0.002***	0.005***	-0.015***	0.010***	0.001	0.001**
	(0.000)	(0.000)	(0.002)	(0.000)	(0.001)	(0.000)
관측치	29,962	29,962	30,419	30,419	30,419	30,419
R-squared	0.029	0.017	0.040	0.032	0.010	0.011

주: 2017년 표본 더미, 시·도 더미, 소분류 업종 더미, 연령, 업력, 남성, 기존 매출액 및 고용, 복식부기 더미를 포함하였으나 추정계수는 생략하였음.
자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

정책금융 지원이 사업주 개인 신용에 미친 영향: 전체 사업체

<표 4-22> 폐업 사업체의 요약통계량(기준연도 익년도 폐업 사업체)

변수명		2017년 수혜업체 및 대조군 (N=1,936)		2016년 수혜업체 및 대조군 (N=920)	
		평균	표준편차	평균	표준편차
종속변수	신용점수	-44.08	174.80	-38.89	168.73
	DSR	0.38	1.64	0.30	1.32
	기업부채량	7,558.88	39,505.1	12,139.1	26,527.6
기존 사업주 신용점수		748.86	145.62	742.66	144.95
기존 사업주 DSR		0.21	0.45	0.21	0.50
기존 매출액		6,014.05	7,904.73	5,544.78	7,990.80
기존 고용인원		0.18	0.62	0.12	0.52
사업주 연령		43.35	11.17	44.02	10.54
사업체 업력		1.92	2.52	3.19	3.07
사업주 남성 더미		0.44	0.50	0.41	0.49
간편장부 여부		0.09	0.29	0.09	0.29

자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

정책금융 지원이 사업주 개인 신용에 미친 영향: 전체 사업체

<표 4-23> 폐업사업체 중 정책금융이 대표자 신용에 미친 영향

변수	신용점수 전후 변화 $\Delta_{t-1,t+1}$		DSR 전후 변화 $\Delta_{t-1,t+1}$		기업부채 전후 변화 $\Delta_{t-1,t+1}$	
	(1)	(2)	(3)	(4)	(5)	(6)
정책금융	-63.75***	-12.35	0.418***	-0.0425	17,752***	19,397***
	(6.565)	(15.82)	(0.0587)	(0.141)	(1,329)	(3,198)
정책금융 X 전년 신용등급		-11.29***		0.0999***		-359.1
		(3.120)		(0.0278)		(630.8)
전년도 신용등급	-32.12***	9.790***	0.0953	0.0600***	-959.6	-195.3
	(8.201)	(2.107)	(0.0734)	(0.0188)	(1,660)	(426.0)
전년도 DSR	-0.558***		-0.0001		-9.194	
	(0.121)		(0.001)		(24.47)	
전년도 신용점수	-5.136	2.879	-0.775***	-0.752***	711.9	808.1
	(7.998)	(7.715)	(0.0716)	(0.0688)	(1,619)	(1,560)
관측치	2,805	2,805	2,805	2,805	2,805	2,805
R-squared	0.091	0.088	0.102	0.106	0.131	0.131

주: 2017년 표본 더미, 가계부채, 전년도 매출 및 고용인원, 사업주 나이, 사업주 성별, 사업체 업력, 복식부기 더미, 상수항, 시·도 더미, 소분류 업종 더미를 포함하였으나 추정계수는 생략함.
자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

Table of Contents

1 서론
2 소상공인 지원정책 소개
3 연구 모형 및 자료의 소개
4 실증분석 결과
5 요약 및 시사점
부록1. 소상공인 대상 보증, 융자 지원사업 내용
부록2. 신종 코로나바이러스(CV) 관련 소상공인 지원
부록3. 성향점수 매칭 전후 성향점수 분포
부록4. 강건성 검증 실증분석 결과

분석 결과 요약(cnt.)

정책금융의 사업성과 효과

- 정책금융 수혜업체는 비수혜업체에 비해:
  - **매출액 증가**와 **고용인원 증가**.
  - **폐업확률 감소**.
- 지역 경쟁도가 높을수록 정책금융의 효과 감소:
  - 동종업체 비율이 높을수록 매출·고용 효과 감소.
- 대표자의 신용도가 낮을수록 사업성과 효과 약화.

정책금융의 신용 및 부채 영향

- **신용 점수 감소** 및 **부채 상환 부담 증가**.
- 개인 신용카드 90일 이상 연체 확률 감소.

분석 결과 요약

폐업 사업체에 대한 영향

- 정책금융 수혜업체는 비수혜업체에 비해:
  - 폐업 전후 신용도 8.2% 감소.
  - 부채 상환 비율(DSR) 2배 증가.
- **신용도가 낮을수록** 부정적 영향이 더욱 큼.

결론

- 과밀화 지역과 저신용 사업체에 대한 추가적 정책 고려 필요.
- 대출 심사 과정에서 사업성과와 부채 위험을 면밀히 검토해야 함.

정책 제언: 통상적 환경

평상시 시장 환경에서의 정책 시사점

- 저금리 정책금융은 매출액 및 고용인원 증가 효과.
- 보증, 이차보전, 융자 정책금융의 지속적 관리 필요.

과밀화 문제 해소를 위한 정책

- **폐업 및 재기 지원 강화**:
  - 성장성 있는 업종과 상권으로 전환 지원.
- **상권정보시스템 고도화**:
  - 신규 창업이 과밀화된 상권으로 진입하지 않도록 예방.

대출 심사 강화

- 사업성 개선을 위한 비금융 지원 병행.
- 사업체와 사업주의 특성을 세심히 관찰.

정책 제언: 코로나19 상황

코로나19 지속 상황에서의 정책 시사점

- 어려운 영업환경에서 정책금융이 폐업 방지 효과에 집중.
- 매출 및 고용 효과는 평소보다 낮을 가능성.

신용 악화 및 부채 부담 방지

- 신용도가 낮은 사업주의 신용 악화 가능성 증가.
- 부채 규모가 과도하게 증가하지 않도록 관리.

부채 조정 및 폐업 지원 강화

- 적절한 시점에 폐업할 수 있도록 지원.
- 채무 조정 제도의 안내를 통해 부채 관리 지원.

Table of Contents

- 서론
- 소상공인 지원정책 소개
- 연구 모형 및 자료의 소개
- 실증분석 결과
- 요약 및 시사점
- 부록1. 소상공인 대상 보증, 융자 지원사업 내용
- 부록2. 신종 코로나바이러스(CV) 관련 소상공인 지원
- 부록3. 성향점수 매칭 전후 성향점수 분포
- 부록4. 강건성 검증 실증분석 결과

주요 사업별 지원내용(cnt.)

<부표 4-1> 주요 보증제도별 지원대상

보증제도	내 용
일반보증	담보력이 부족한 중소기업 및 소상공인 등을 대상으로 보증지원(보증비율 85%, 보증료 0.5~2.0%)
특례보증	재해중소기업 특례보증, 내수활성화 특례보증 등 정부의 정책적 필요에 따라 보증료, 한도사정 등 보증요건 우대(보증비율 95~100%, 보증료 1.0 이하 등)
햇살론 (사업자)	신용 6~10등급 또는 연소득 3,500만원 이하 햇살론 신청자격을 만족하는 저소득 소상공인(무등록·무점포 포함)에 대한 보증(저신용자의 경우 연 소득 4,500만원 이하인 자에 한해 지원 가능)

자료: 신용보증재단중앙회(2017).

주요 사업별 지원내용(cnt.)

<부표 4-2> 업종별·신용등급별 보증한도

업종	신용등급	사정한도액
제조업 지식서비스 산업	B등급 이상	당기매출액(또는 최근 1년간 매출액)의 1/3, 향후 1년간 추정매출액의 1/3, 또는 최근 4개월 매출액
	CCC등급 이하	당기매출액(또는 최근 1년간 매출액)의 1/4, 향후 1년간 추정매출액의 1/4, 또는 최근 3개월 매출액
기타	B등급 이상	당기매출액(또는 최근 1년간 매출액)의 1/4, 향후 1년간 추정매출액의 1/4, 또는 최근 3개월 매출액
	CCC등급 이하	당기매출액(또는 최근 1년간 매출액)의 1/6, 최근 2개월 매출액

주: 시설자금은 당해 시설의 소요자금 범위 이내에서 지원함. 소상공인 창업 및 경영개선자금은 최고 7천만원 한도 내, 소상공인 신용평가시스템상의 보증한도액 이내에서 지원함.
자료: 신용보증재단중앙회(2017).

주요 사업별 지원내용(cnt.)

<부표 4-3> 연도별 보증실적

(단위: 억원, 개)

		2011년	2012년	2013년	2014년	2015년	2016년
보증 공급	업체 수	256,434	275,845	350,944	375,150	448,844	464,329
	금액	60,801	63,581	77,859	85,057	106,386	111,849
보증 잔액	업체 수	835,802	829,574	837,659	756,842	833,213	910,813
	금액	136,088	135,148	142,719	144,501	162,423	179,713

자료: 신용보증재단중앙회(2017).

주요 사업별 지원내용(cnt.)

<부표 4-4> 주된 업종별 평균매출액 등의 소기업 규모 기준(제8조 제1항 관련)

해당 기업의 주된 업종	분류기호	기준
1. 식료품 제조업	C10	평균매출액 등 120억원 이하
2. 음료 제조업	C11	
3. 의복, 의복액세서리 및 모피제품 제조업	C14	
4. 가죽, 가방 및 신발 제조업	C15	
5. 코크스, 연탄 및 석유정제품 제조업	C19	
6. 화학물질 및 화학제품 제조업(의약품 제조업은 제외한다)	C20	
7. 의료용 물질 및 의약품 제조업	C21	
8. 비금속 광물제품 제조업	C23	
9. 1차 금속 제조업	C24	
10. 금속가공제품 제조업(기계 및 가구 제조업은 제외한다)	C25	
11. 전자부품, 컴퓨터, 영상, 음향 및 통신장비 제조업	C26	
12. 전기장비 제조업	C28	
13. 그 밖의 기계 및 장비 제조업	C29	
14. 자동차 및 트레일러 제조업	C30	
15. 가구 제조업	C32	
16. 전기, 가스, 증기 및 공기조절 공급업	D	
17. 수도업	E36	
18. 농업, 임업 및 어업	A	평균매출액 등 80억원 이하
19. 광업	B	
20. 담배 제조업	C12	
21. 섬유제품 제조업(의복 제조업은 제외한다)	C13	
22. 목재 및 나무제품 제조업(가구 제조업은 제외한다)	C16	
23. 펄프, 종이 및 종이제품 제조업	C17	
24. 인쇄 및 기록매체 복제업	C18	
25. 고무제품 및 플라스틱제품 제조업	C22	
26. 의료, 정밀, 광학기기 및 시계 제조업	C27	
27. 그 밖의 운송장비 제조업	C31	
28. 그 밖의 제품 제조업	C33	
29. 건설업	F	
30. 운수 및 창고업	H	
31. 금융 및 보험업	K	
32. 도매 및 소매업	G	50억원 이하
33. 정보통신업	J	
34. 수도, 하수 및 폐기물 처리, 원료재생업(수도업은 제외한다)	E(E36 제외)	평균매출액 등 30억원 이하
35. 부동산업	L	
36. 전문·과학 및 기술 서비스업	M	
37. 사업시설관리, 사업지원 및 임대 서비스업	N	
38. 예술, 스포츠 및 여가 관련 서비스업	R	
39. 산업용 기계 및 장비 수리업	C34	평균매출액 등 10억원 이하
40. 숙박 및 음식점업	I	
41. 교육 서비스업	P	
42. 보건업 및 사회복지 서비스업	Q	
43. 수리(修理) 및 기타 개인 서비스업	S	

주: 1) 해당 기업의 주된 업종 분류 및 분류기호는 「통계법」 제22조에 따라 통계청장이 고시한 한국표준산업분류에 따른다.
2) 위 표 제27호에도 불구하고 철도 차량 부품 및 관련 장치물 제조업(C31202) 중 철도 차량용 의자 제조업, 항공기용 부품 제조업(C31322) 중 항공기용 의자 제조업의 규모 기준은 평균매출액 등 120억원 이하로 한다.
자료: 「중소기업기본법」 시행령(제8조 제1항 관련).

주요 사업별 지원내용

<부표 4-5> 주요 정책자금별 대출금리(2017년 기준)

	자금 구분	기준금리 (2017.1분기)	가산금리	대출금리 (2017.1분기)
공단 대출	① 소공인특화자금	연 1.79%	+ 0.6%P	연 2.39%
	② 수출소상공인 특별자금		+ 0.4%P	연 2.19%
	③ 성장촉진자금(시설)		+ 0.4%P	연 2.19%
	④ 소상공인창업자금(사관학교)		+ 0.6%P	연 2.39%
	⑤ 임차보증금 안심금융		+ 0.2%P	연 1.99%
금융기관 대출	⑥ 성장촉진자금		+ 0.4%P	연 2.19%
	⑦ 성장촉진자금(컨설팅이수)		+ 0.2%P	연 1.99%
	⑧ 소상공인창업자금		+ 0.6%P	연 2.39%
	⑨ 소상공인창업자금(15년 이전 실전창업학교 이수자)		+ 0.45%P	연 2.24%
	⑩ 사업전환자금		+ 0.4%P	연 2.19%
	⑪ 사업전환자금(재창업패키지 이수)		+ 0.2%P	연 1.99%
	⑫ 일반경영안정자금		+ 0.6%P	연 2.39%
	⑬ 일반경영안정자금(장애인)		고정금리	연 2.00%
	⑭ 긴급경영안정자금			
	⑮ 경영애로자금			
	⑯ 소상공인 전환대출			연 5.00%
	⑰ 청년드림자금		+ 0.4%P	연 2.19%

주: 1) 자금별 대출금리 = 기준금리(공자기금 예탁금리) + 가산금리.
2) 고정금리 방식 대출자는 대출 실행 시 결정된 금리가 약정기간 동안 변동 없이 동일하게 적용됨.
3) 신규 대출취급 불가 자금: ⑤, ⑨, ⑯
자료: 소상공인시장진흥공단(2017. 1).

Table of Contents

1 서론
2 소상공인 지원정책 소개
3 연구 모형 및 자료의 소개
4 실증분석 결과
5 요약 및 시사점
6 부록1. 소상공인 대상 보증, 융자 지원사업 내용
7 부록2. 신종 코로나바이러스(CV) 관련 소상공인 지원
8 부록3. 성향점수 매칭 전후 성향점수 분포
9 부록4. 강건성 검증 실증분석 결과

코로나19 이후 소상공인 지원정책

<부표 4-6> 코로나19 이후 소상공인 지원정책 요약

	지원 내용	주관기관	시행일
대출	■ (명칭) 소상공인 경영안정자금 ■ (규모) 200억원 ■ (대상) 신종 CV 피해 소상공인 ■ (내용) 한도 7천만원 융자 지원, 대출기간 5년(2년 거치 3년 상환)	소상공인시장진흥공단	2.13
	■ (명칭) 미소금융 창업·운영자금 ■ (규모) 4,400억원 ■ (대상) 저신용(6등급 이하)·저소득(차상위계층 이하) 영세 자영업자 대상 ■ (내용) 1인당 2천만원 한도, 최장 5년, 금리 4.5% 이내	서민금융진흥원	2.11
	■ (명칭) 미소금융 전통시장 상인대출 ■ (규모) 500억원 → 550억원(50억원 확대) ■ (대상) 전통시장(전국 318개)의 영세상인 * 기초 지자체 추천을 받아 서민금융진흥원과 지원사업 약정을 체결한 전통시장 ■ (내용) 1인당 1천만원 한도, 만기 최장 2년, 금리 4.5% 이내(상인회 자율 결정)	서민금융진흥원	2.11
	■ (명칭) 특별자금 지원 ■ (규모) 1,000억원 ■ (대상) 신종 CV 피해 소상공인 및 자영업자 ■ (내용) 최대 5억원, 1년(3년까지 연장 가능), 최대 1.0%p 금리감면	기업은행	2.10
보증	■ (명칭) 특례보증 프로그램 ■ (규모) 1,000억원 ■ (대상) 신종 CV 피해 소상공인 ■ (내용) 한도 7천만원 보증지원(5년 이내), 보증비율 상향(85%→100%), 보증료율 0.8% 적용	지역신용보증재단	2.13
	■ (명칭) 특례보증 프로그램 ■ (규모) 1,000억원 ■ (대상) 신종 CV 피해 기업 ■ (내용) 기업당 3억원 이내 보증지원, 보증비율 상향(85%→95%), 보증료율 1.0% 적용	기술보증기금	2.13

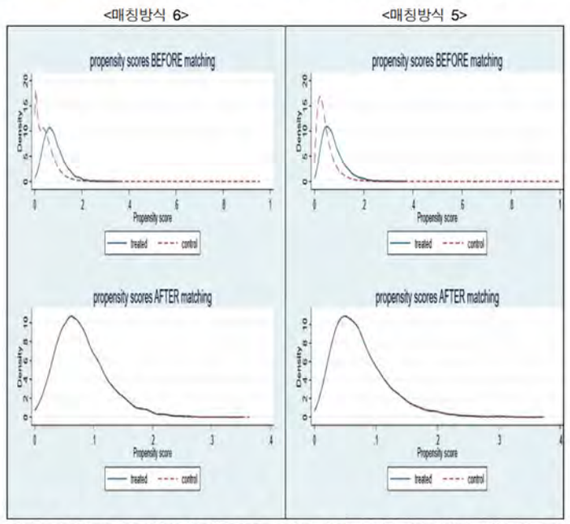
<부표 4-6>의 계속

Table of Contents

1 서론
2 소상공인 지원정책 소개
3 연구 모형 및 자료의 소개
4 실증분석 결과
5 요약 및 시사점
6 부록1. 소상공인 대상 보증, 융자 지원사업 내용
7 부록2. 신종 코로나바이러스(CV) 관련 소상공인 지원
8 부록3. 성향점수 매칭 전후 성향점수 분포
9 부록4. 강건성 검증 실증분석 결과

성향점수 매칭 전후 성향점수 분포

[부도 4-1] 성향점수 매칭 전후 정책 수혜기업과 대조군의 성향점수 분포



자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 작성.

Table of Contents

1 서론
2 소상공인 지원정책 소개
3 연구 모형 및 자료의 소개
4 실증분석 결과
5 요약 및 시사점
6 부록1. 소상공인 대상 보증, 융자 지원사업 내용
7 부록2. 신종 코로나바이러스(CV) 관련 소상공인 지원
8 부록3. 성향점수 매칭 전후 성향점수 분포
9 부록4. 강건성 검증 실증분석 결과

정책금융이 영세 사업자 매출액 변화에 미친 영향

<부표 4-7> 정책금융이 영세 자영업자 로그 매출액 변화에 미친 영향

종속변수	로그 매출액 전후 변화	
	1년 후 $\Delta_{t-1,t+1}$	2년 후 $\Delta_{t-1,t+2}$
	(1)	(2)
정책금융 수혜	0.192***	0.148***
	(0.012)	(0.022)
2017년 정책금융 이용 더미	-0.205***	
	(0.013)	
사업주 연령	-0.007***	-0.007***
	(0.001)	(0.001)
사업체 업력	-0.043***	-0.047***
	(0.002)	(0.003)
남성 더미	0.115***	0.133***
	(0.013)	(0.025)
복식부기 더미	0.386***	0.375***
	(0.026)	(0.049)
정책금융 이용 전년도 매출액	-0.000***	-0.000***
	(0.000)	(0.000)
정책금융 이용 전년도 고용	0.124***	0.129***
	(0.011)	(0.022)
상수항	1.582***	1.745***
	(0.124)	(0.254)
관측치	28,427	9,806
R-squared	0.151	0.165

주: 시·도 더미와 소분류 업종 더미를 포함하였으나 추정계수는 생략하였음.
자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

정정책금융의 효과 강건성 검정: 기본분석

<부표 4-8> 매칭방법 5를 적용한 정책금융의 효과 강건성 검정(기본분석)

종속변수	매출액 전·후 변화		고용인원 전·후 변화		폐업 여부
	1년 후 $\Delta_{t-1,t+1}$	2년 후 $\Delta_{t-1,t+2}$	1년 후 $\Delta_{t-1,t+1}$	2년 후 $\Delta_{t-1,t+2}$	1년 후 폐업 (폐업=1)
	(1)	(2)	(3)	(4)	(5)
정책금융 수혜	1,998.555**	1,861.025***	0.053***	0.014	0.006
	(779.205)	(309.044)	(0.012)	(0.016)	(0.006)
2017년 정책금융 이용 더미	-1,393.634*		-0.017		
	(814.427)		(0.012)		
사업주 연령	-46.934	-75.593***	-0.001***	-0.002***	-0.001***
	(31.087)	(16.102)	(0.000)	(0.001)	(0.000)
사업체 업력	-281.380***	-390.176***	-0.009***	-0.009***	-0.004***
	(78.639)	(35.642)	(0.001)	(0.002)	(0.001)
남성 더미	1,873.640***	1,079.227***	0.026***	0.019	0.003
	(669.072)	(341.496)	(0.010)	(0.018)	(0.006)
복식부기 더미	-540.638	3,017.480***	0.200***	0.274***	0.008
	(1,324.392)	(701.442)	(0.020)	(0.037)	(0.013)
정책금융 이용 전년도 매출액	0.179***	-0.008	0.000***	0.000***	-0.000**
	(0.051)	(0.027)	(0.000)	(0.000)	(0.000)
정책금융 이용 전년도 고용	2,441.850***	3,953.679***	-0.310***	-0.263***	0.004
	(572.750)	(311.916)	(0.009)	(0.016)	(0.006)
상수항	9,094.33	14,405.64***	0.009	0.058	0.148**
	(6,733.47)	(5,055.84)	(0.091)	(0.199)	(0.070)
관측치	28,464	9,801	30,027	10,467	10,778
R-squared	0.012	0.131	0.058	0.053	0.022

주: 시·도 더미와 소분류 업종 더미를 포함하였으나 추정계수는 생략하였음.
자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

지원방식별 정책금융 효과 강건성 검정

<부표 4-9> 매칭방법 5를 적용한 지원방식별 정책금융 효과 강건성 검정

종속변수	매출액 전·후 변화		고용인원 전·후 변화		폐업 여부
	1년 후 $\Delta_{t-1,t+1}$	2년 후 $\Delta_{t-1,t+2}$	1년 후 $\Delta_{t-1,t+1}$	2년 후 $\Delta_{t-1,t+2}$	1년 후 폐업 (폐업=1)
	(1)	(2)	(3)	(4)	(5)
정책금융 수혜	3,119.078**	3,091.590***	0.085***	0.076**	-0.019***
	(1,210.870)	(592.344)	(0.018)	(0.031)	(0.007)
정책금융 X 이차보전 방식	-926.935	-2,097.774**	-0.019	-0.080*	-0.002
	(1,686.155)	(916.980)	(0.012)	(0.048)	(0.004)
정책금융 X 지역신용보증	-1,401.522	-1,495.017**	-0.055**	-0.079**	0.018**
	(1,128.179)	(604.933)	(0.025)	(0.032)	(0.009)
관측치	28,464	9,801	30,027	10,467	30,419
R-squared	0.012	0.137	0.060	0.057	0.021

주: 시·도 더미, 소분류 업종 더미, 연령, 업력, 남성, 기존 매출액 및 고용, 복식부기 더미를 포함하였으나 추정계수는 생략하였음.
자료: 통계청-KCB 개인사업자 DB와 중소벤처기업부 정책지원이력(SIMS) DB를 활용하여 저자 계산.

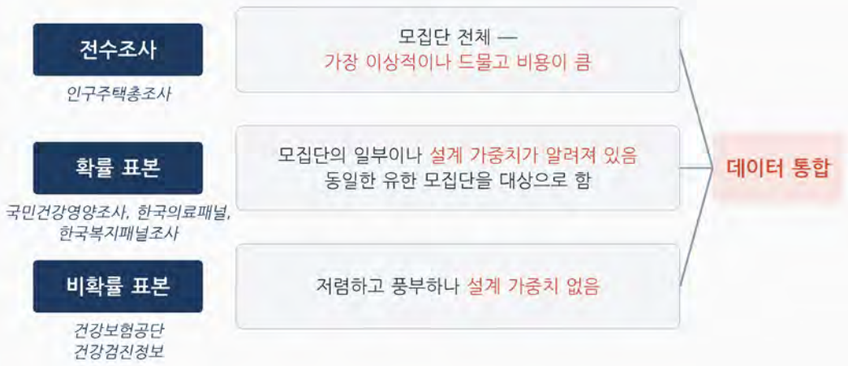
AI 시대의 조사데이터 통합

가중치 보정의 역할과 가능성

권용현
제15회 국가통계·데이터발전포럼

육군사관학교 수학과

AI 시대의 데이터 환경



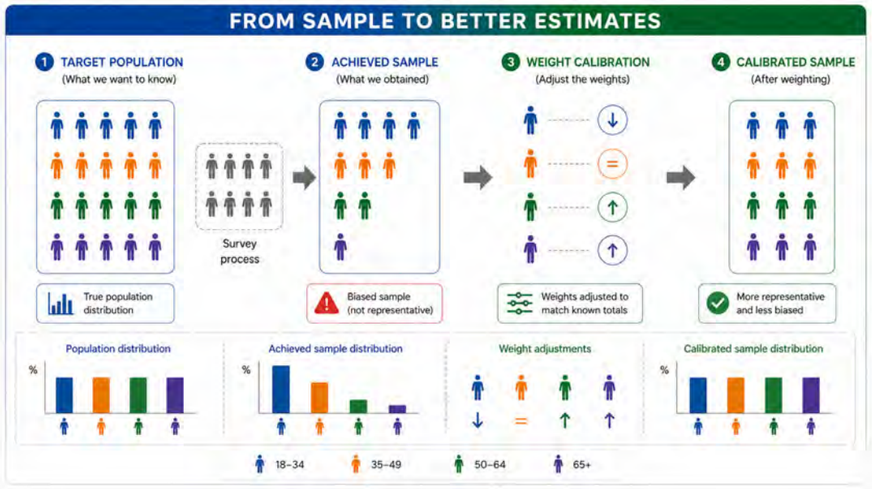
국내 주요 보건 관련 조사 (2022년, 만 19세 이상 설문 대상)

	KNHANES 국민건강영양조사	KHPS 한국의료패널	KOWEPS 한국복지패널조사	NHIS 건강보험공단	데이터 통합
설계 가중치	✓	✓	✓	×	
신장 (cm)	165.3	164.6	—	162.5	?
체중 (kg)	66.4	64.8	—	64.1	?
흡연율	16.7%	17.2%	18.1%	19.3%	?
당뇨병 진단률	10.5%	10.4%	—	8.0%	?
고혈압 진단률	22.9%	22.3%	—	14.0%	?
규칙적 운동	49.7%	48.8%	—	—	?
높은 스트레스	25.6%	25.9%	—	—	?

데이터 통합을 위한 가중치 보정

- 각 조사의 가중치를 조정하여 가중 합계가 공통 보조변수에서 일치하도록 함.
- 가중치 보정의 장점:
  - 모형을 직접 추정하지 않고 가중치를 산출하여 모형 오설정(misspecification)에 강건.
  - 어떤 조사변수에 대해서도 동일한 가중치를 재사용할 수 있어 편리.
  - 각 조사의 표본설계 구조를 그대로 유지하면서 데이터 통합 가능.
- Proposal: Bregman 발산 (Kim et al., 2026; Kwon et al., 2026)을 이용한 가중치 보정
  - 점근적으로 최적 역분산 추정량과 일치함을 이론적으로 보장.

Weight calibration in survey sampling



Two-sample setup

Two independent probability samples A, B drawn from finite population U:

Sample	Type	Weight	X	Y^A	Y^B
A	Probability sample	d^A	✓	✓	×
B	Probability sample	d^B	✓	×	✓

Design weights: $d_i^A = 1/\pi_i^A$, $d_i^B = 1/\pi_i^B$, where π_i is the first-order inclusion probability.

Target: population totals $Y_N^A = \sum_{i=1}^N y_i^A$ and $Y_N^B = \sum_{i=1}^N y_i^B$.

Building block: efficient estimation of $X_N = \sum_{i=1}^N x_i$ using both samples.

Optimal inverse-variance estimator

Horvitz–Thompson estimators:

$$\hat{X}_A = \sum_{i \in A} d_i^A x_i, \qquad \hat{X}_B = \sum_{i \in B} d_i^B x_i.$$

Under Poisson sampling,

$$V(\hat{X}_A) = \sum_{i \in U} (d_i^A - 1) x_i^2, \quad V(\hat{X}_B) = \sum_{i \in U} (d_i^B - 1) x_i^2.$$

Optimal linear combination:

$$\hat{X} = \frac{\hat{X}_A / \hat{V}(\hat{X}_A) + \hat{X}_B / \hat{V}(\hat{X}_B)}{1/\hat{V}(\hat{X}_A) + 1/\hat{V}(\hat{X}_B)}. \tag{1}$$

Equivalent to the GLS estimator (Merkouris, 2004); we will reach the *same* estimator via calibration.

Existing approaches

Method	Reference	Output
GLS	Merkouris (2004)	Parameter-specific estimate
Mass imputation	Kim and Rao (2012)	Unified synthetic data
Statistical matching	Kim et al. (2016)	Unified synthetic data

Key limitations:

- GLS: one estimator per study variable; inconvenient for multipurpose surveys.
- Mass imputation & statistical matching: both rely on a correctly specified regression / outcome model.

Calibration targets the population total directly and enjoys design consistency under mild assumptions.

Calibration: the quadratic case

Find (ω^A, ω^B) minimizing

$$\sum_{i\in A} \frac{(\omega_i^A - d_i^A)^2}{q_i^A} + \sum_{i\in B} \frac{(\omega_i^B - d_i^B)^2}{q_i^B}$$

subject to

$$\sum_{i\in A} \omega_i^A x_i - \sum_{i\in B} \omega_i^B x_i = 0.$$

Lagrangian solution:

$$\omega_i^A = d_i^A - x_i\lambda\, q_i^A, \quad \omega_i^B = d_i^B + x_i\lambda\, q_i^B.$$

Resulting calibrated estimator

Define $Q_A = \sum_{i\in A} q_i^A x_i^2$, $Q_B = \sum_{i\in B} q_i^B x_i^2$.

Substitution into the constraint yields

$$\lambda = \frac{\hat{X}_A - \hat{X}_B}{Q_A + Q_B},$$

and the calibration estimator becomes

$$\hat{X} = \frac{\hat{X}_A/Q_A + \hat{X}_B/Q_B}{1/Q_A + 1/Q_B}.$$

This matches the optimal inverse-variance estimator when $q_i = d_i(d_i - 1)$.

Limitation: quadratic loss may yield negative calibration weights.

Bregman divergence calibration (BDCal)

Let $G : \mathcal{V} \to \mathbb{R}$ be strongly convex, twice differentiable. Define the Bregman divergence

$$D(\omega\|d) = G(\omega) - G(d) - g(d)(\omega - d), \quad g = G'. \tag{2}$$

Optimization problem:

$$\min_{\omega^A,\omega^B} \left\{ \sum_{i\in A} D^A(\omega_i^A\|d_i^A) + \sum_{i\in B} D^B(\omega_i^B\|d_i^B) \right\} \tag{3}$$

subject to the calibration constraint

$$\sum_{i\in A} \omega_i^A \boldsymbol{x}_i = \sum_{i\in B} \omega_i^B \boldsymbol{x}_i. \tag{4}$$

- Quadratic loss is a special case.
- Different G give different weighting schemes.

Optimal divergence: the contrast entropy

Kim et al. (2026) showed the contrast entropy

$$G(\omega) = (\omega - 1)\log(\omega - 1) - \omega\log\omega \tag{5}$$

minimizes the design variance of the calibration estimator.

With contrast entropy,

$$\hat{X}_{\mathrm{BDCal}_{CE}} \asymp \frac{\hat{X}_A/\hat{V}(\hat{X}_A) + \hat{X}_B/\hat{V}(\hat{X}_B)}{1/\hat{V}(\hat{X}_A) + 1/\hat{V}(\hat{X}_B)}.$$

The calibration estimator coincides asymptotically with the optimal inverse-variance estimator (1).

Design consistency

Calibrated weights: $\omega_i^A(\boldsymbol{\lambda}) = g^{-1}(g(d_i^A) - \boldsymbol{x}_i^\top \boldsymbol{\lambda})$.

Population calibration equation satisfied at $\boldsymbol{\lambda}_0 = \mathbf{0}$. Taylor expansion gives

$$\hat{\boldsymbol{\lambda}} = H_n^{-1}(\hat{\boldsymbol{X}}_A - \hat{\boldsymbol{X}}_B) + o_p(n^{-1/2}),$$

where $H_n = \sum_A \boldsymbol{x}_i \boldsymbol{x}_i^\top / g'(d_i^A) + \sum_B \boldsymbol{x}_i \boldsymbol{x}_i^\top / g'(d_i^B)$.

Since $N^{-1}(\hat{\boldsymbol{X}}_A - \hat{\boldsymbol{X}}_B) = o_p(1)$,

$$\hat{\boldsymbol{\lambda}} = o_p(1) \;\Rightarrow\; \text{calibrated weights} \approx \text{design weights}.$$

Hence $\hat{Y}^A$ inherits design consistency of the HT estimator.

Linearization and variance estimator

Expansion around $\boldsymbol{\lambda} = \mathbf{0}$:

$$\hat{Y}^A \doteq \hat{Y}_{\mathrm{HT}}^A + (\hat{\boldsymbol{X}}_B - \hat{\boldsymbol{X}}_A)^\top \hat{\boldsymbol{\beta}}_A, \tag{6}$$

a two-sample GREG form, where

$$\hat{\boldsymbol{\beta}}_A = H_n^{-1}\left\{ \sum_{i \in A} \frac{\boldsymbol{x}_i y_i^A}{g'(d_i^A)} \right\}.$$

Equivalently,

$$\hat{Y}_{\mathrm{lin}}^A = \sum_{i \in A} d_i^A \underbrace{(y_i^A - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_A)}_{=:e_i^A} + \sum_{i \in B} d_i^B \underbrace{\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_A}_{=:m_i^A}.$$

Variance decomposition (independent samples):

$$V(\hat{Y}_{\mathrm{lin}}^A) = V\Big(\sum_A d_i^A e_i^A\Big) + V\Big(\sum_B d_i^B m_i^A\Big).$$

Efficiency gain over Horvitz–Thompson

Is the proposed estimator guaranteed to beat $\hat{Y}_{\mathrm{HT}}^A$?

Key result

Under Poisson sampling, the contrast entropy (5), and certain regularity conditions,

$$V(\hat{Y}_{\mathrm{HT}}^A) - V(\hat{Y}_{\mathrm{lin}}^A) = \boldsymbol{h}^\top \boldsymbol{H}^{-1} \boldsymbol{h} + o(N^2 n^{-1}) \;\geq\; 0, \tag{7}$$

where

$$\boldsymbol{h} = \sum_{i \in U} (d_i^A - 1)\, y_i^A\, \boldsymbol{x}_i, \quad \boldsymbol{H} = \sum_{i \in U} \{(d_i^A - 1) + (d_i^B - 1)\}\, \boldsymbol{x}_i \boldsymbol{x}_i^\top.$$

BDCal_{CE} is asymptotically never less efficient than HT. The gain is *strict* whenever y^A is linearly associated with $\boldsymbol{x}$ at the population level (i.e., $\boldsymbol{h} \neq \mathbf{0}$).

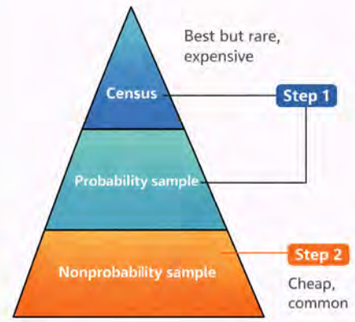
Multi-source setting

k probability samples $A_1, \ldots, A_k$ and ℓ nonprobability samples $B_1, \ldots, B_\ell$ from the same population U.

Let $\boldsymbol{x}_i^*$ denote the augmented auxiliary vector collecting all variables observed across sources.

Two-step calibration procedure:

1. Step 1 — Bregman calibration (BC) across probability samples $\{A_j\}$.
2. Step 2 — Generalized entropy calibration (GEC) for nonprobability samples $\{B_j\}$, anchored to Step 1.



Step 1: Bregman calibration across probability samples

Find $\omega^{A_1}, \ldots, \omega^{A_k}$ minimizing

$$\sum_{j=1}^{k} \sum_{i \in A_j} D^{A_j}\left(\omega_i^{A_j} \,\middle\|\, d_i^{A_j}\right) \tag{8}$$

subject to the common-total constraint

$$\sum_{i \in A_1} \omega_i^{A_1} \boldsymbol{x}_i^* = \cdots = \sum_{i \in A_k} \omega_i^{A_k} \boldsymbol{x}_i^* = \sum_{i \in U} \boldsymbol{x}_i^*.$$

- Each D^{A_j} induced by a sample-specific convex G^{A_j}.
- Choosing contrast entropy for each $G^{A_j} \Rightarrow$ asymptotically optimal inverse-variance estimator.

Step 2: Generalized entropy calibration for nonprobability data

Using Step 1 totals as benchmarks, find $\omega^{B_1}, \ldots, \omega^{B_\ell}$ minimizing

$$\sum_{j=1}^{\ell} \sum_{i \in B_j} G^{B_j}(\omega_i^{B_j}) \tag{9}$$

subject to

$$\sum_{i \in B_j} \omega_i^{B_j} \boldsymbol{x}_i^* = \sum_{i \in A_m} \omega_i^{A_m} \boldsymbol{x}_i^* = \sum_{i \in U} \boldsymbol{x}_i^*, \quad j = 1, \ldots, \ell.$$

First-order condition: $g^{B_j}(\omega_i^{B_j}) = (\boldsymbol{x}_i^*)^\top \boldsymbol{\lambda}$, i.e. $\omega_i^{B_j} = (g^{B_j})^{-1}\big((\boldsymbol{x}_i^*)^\top \boldsymbol{\lambda}\big)$.

- G^{B_j} corresponds to a working propensity model for B_j.
- E.g. contrast entropy $\leftrightarrow$ logistic propensity score.
- Doubly robust: consistent if *either* propensity *or* outcome model is correctly specified (Kwon et al., 2026).

Simulation design

Population size $N = 3000$; iid covariates $x_1, x_2, x_3 \sim N(0, 1)$.

$$y^A = 1 + x_1 + x_2 + x_1 x_2 + (x_1^2 - 1) + \varepsilon^A \quad \text{(nonlinear)}$$

$$y^B = 1 + x_1 - x_3 + \varepsilon^B \quad \text{(linear)}$$

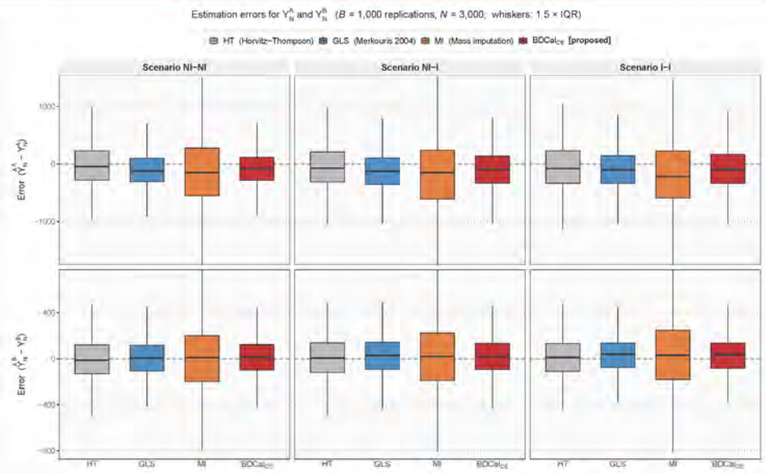
Calibration variable: $\boldsymbol{x} = (1, x_1, x_2, x_3)^\top$ (no nonlinear terms).

Poisson sampling: $n_A \approx 580$, $n_B \approx 960$; three scenarios:

- **NI-NI**: both noninformative
- **NI-I**: B informative
- **I-I**: both informative

Compare HT, GLS(design optimal), MI, BDCal_{CE} over $B = 1000$ replications.

Distribution of estimation errors



Real data results

	KNHANES	KHPS	KOWEPS	Integrated
Multi-survey outcomes				
Smoking	0.167 (.007)	0.172 (.006)	0.181 (.006)	0.170 (.004)
Diabetes	0.105 (.005)	0.104 (.004)	—	0.104 (.003)
Hypertension	0.229 (.007)	0.223 (.005)	—	0.225 (.004)
Single-survey outcomes				
High cholesterol	0.275 (.009)	—	—	0.278 (.009)
Unmet med. need	—	0.147 (.005)	—	0.152 (.006)
Low income	—	—	0.174 (.005)	0.185 (.006)
Not-employed	—	—	0.335 (.008)	0.335 (.009)

20/21

감사합니다!

www.yhkwon.pro

결론

- Bregman 보정을 통한 조사데이터 통합의 통일된 프레임워크 제시.
- Contrast entropy 기반 BDCal$_{CE}$는 점근적으로 최적 역분산 추정량과 일치.
- 설계 일치적(Design consistent) 선형 분산 추정량 제공.
- 비확률표본을 포함할 경우 2단계 절차(BC + GEC) 적용 가능.

진행 중 및 향후 연구

- 서로 다른 형식으로 관측되거나 측정 오차가 있는 조사 간 공변량 처리.
- 종단 패널 조사에의 적용.
- 고차원 보조변수에 대한 소프트 보정(soft calibration).

21/21

References i

Kim, J. K., Berg, E., and Park, T. (2016). Statistical matching using fractional imputation. *Survey Methodology*, 42(1):19–41.

Kim, J. K., Kwon, Y., and Qiu, Y. (2026). Bregman projection for calibration estimation in survey sampling.

Kim, J. K. and Rao, J. N. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99(1):85–100.

Kwon, Y., Kim, J. K., and Qiu, Y. (2026). Generalized entropy calibration for analyzing voluntary survey data. *Biometrics*, 82(1). https://doi.org/10.1093/biomtc/ujag041.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99(468):1131–1139.

Ministry of Data and Statistics

안전한 데이터 연계와 활용을 위한 플랫폼 운영 모델 제안

2026.05.14 | 국가데이터처 통계기준과 박민정



Disclaimer | 이 내용은 **DATA USER인 개인 연구자들의 견해**를 정리한 것으로, 국가데이터처의 공식 의견이 아닙니다.

데이터 연계(연결) 개념도

왜 작동하지 않을까?



신뢰성 있는 데이터 연계를 통한 가치 창출
서로 다른 데이터원을 안전하게 연결하여 더 정확하고 풍부한 정보 제공

✓ 데이터 활용도 향상 ✓ 정책/연구의 품질 제고 ✓ 국민 삶의 질 향상

→ 데이터 연계 활용에 대한 사람들의 일반적인 생각 (위 그림은 AI 생성 결과)

Ministry of Data and Statistics. All Rights Reserved.

데이터 연계 활용 · 저해 요인

1 거버넌스, 법제도	2 보상 체계 부재	3 개인정보보호 이슈	4 표준화 미흡	5 이용자 편의성 부족
• 도메인별 상이한 법령 • 복잡한 가명결합 절차 • 목적 외 이용 금지 조항	• 데이터 가치 산정 어려움 • 제공 주체와 보상 대상 불명확 • 거래 윤리 논쟁	• 분석 결과 반출 시에도 개인정보 노출 가능 • 보안(Security)과 정보보호(Privacy) 개념 혼용	• 데이터 표준 제정·이행 어려움 • 데이터 카탈로그 구현 한계	• 데이터센터 물리적 방문 필수 • 반출 심의 장기 대기 • 낮은 유용성의 결과물 수령

→ 안전한 데이터 연계와 활용을 위한 다자간 피처 스토어 / 통계연구

Ministry of Data and Statistics. All Rights Reserved.

CONTENTS

목차

01 다자간 피처 스토어란?
- 국내 데이터 거래소 현황
- 피처 기반 데이터 거래
- 기본 구성요소

02 데이터센터의 HINGE DATA 운영
- 통계등록부, 데이터센터 소개
- HINGE DATA로서 통계등록부
- 데이터센터의 기타 역할

03 데이터트윈 및 보안 클라우드
- 재현자료(synthetic data)로 구성된 데이터트윈(data twin)
- 보안 클라우드에서 원데이터 분석

04 결론 및 향후 발전방향
- 왜 '다자간' 피처 스토어인가?
- 다자간 피처 스토어의 한계와 향후 추진방향

Ministry of Data and Statistics. All Rights Reserved.

01
다자간 피처 스토어란?



Ministry of Data and Statistics. All Rights Reserved.

01 다자간 피처 스토어란?
국내 데이터 거래소 현황

명칭	운영 주체	역할
문화 빅데이터 플랫폼 (www.bigdata-culture.kr)	한국문화정보원	도서, 체육, 예술, 숙박, 레저, 음식 등 다양한 문화 분야의 고품질 데이터를 개방해 데이터 유통·거래 생태계 조성
한국데이터거래소 (kdx.kr)	MBN, 삼성카드, 다올리브네트웍스, SK텔레콤 등	국내 최초(19.12 출범) 민간 데이터 거래소로 유통·소비 분야의 빅데이터 플랫폼을 제공
국가암데이터센터 (www.cancerdata.re.kr)	국립암센터, 보건복지부	암 정복을 위한 빅데이터를 공개하여 연구자와 의료진의 활용을 지원
암 빅데이터 플랫폼 (www.bigdata-cancer.kr)	국립암센터	10대 암종별 임상데이터를 수집·제공하여 암 진단 및 치료 연구를 지원
환경 빅데이터 플랫폼 (www.bigdata-environment.kr)	한국수자원공사	생활환경과 자연환경 데이터를 수집·제공하여 맞춤형 수질정보 서비스 및 대기질 야외활동 추천 서비스 등 제공
금융권 공동 데이터 플랫폼 (www.datop.or.kr)	금융권 공동 구축	금융 데이터를 수집·분석해 금융 서비스 혁신과 데이터 기반 의사 결정을 지원
금융 빅데이터 플랫폼 (www.bigdata-finance.kr)	비씨카드	보험, 증권, 통신 등의 데이터를 수집·제공해 소상공인 창업지도 서비스 및 국민 금융생활 플래너 서비스 등 제공

2015년 설립된 중국의 귀양 빅데이터거래소를 시작으로, 국내에도 2020년 금융보안원이 출범한 금융데이터거래소 (findatamall.or.kr) 등이 운영되고 있다.

Ministry of Data and Statistics. All Rights Reserved.

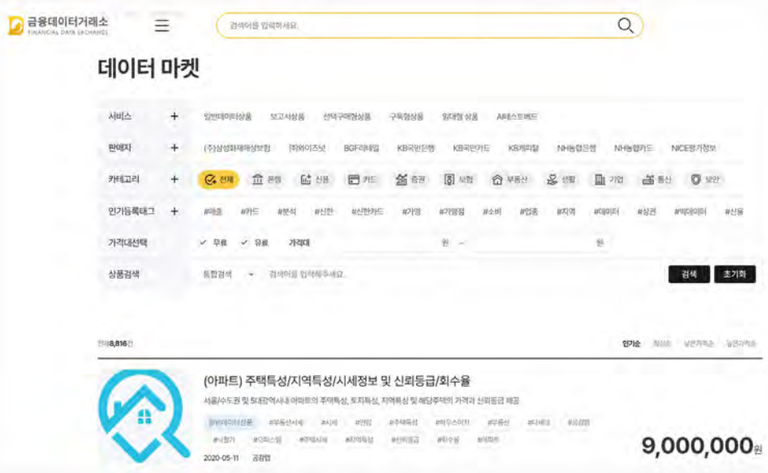
01 다자간 피처 스토어란?
데이터 거래 예시 1



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	광역시도명	시군구명	기준년월	업종대분류	업종중분류	업종소분류	신규가맹점수	폐업가맹점수	매출가맹점수	카드매출금액	카드매출건수	점당매출금액	건당매출금액
2	전남	여수시	202001	전문서비스	대학병원	병원-대학/종합병원	0	0	3	421527000	5000	140509163	84305
3	전남	여수시	202001	전문서비스	보관/창고업	물품보관소	0	0	8	77872000	342	9734060	227697
4	전남	여수시	202001	전문서비스	부동산	부동산-건물관리	0	0	1	24405000	12863	24404900	1897
5	전남	여수시	202001	전문서비스	부동산	부동산-주택/아파트 임대	0	0	9	11319000	307	1257656	36869
6	전남	여수시	202001	전문서비스	부동산	부동산중개	0	0	3	6297000	20	2098900	314835
7	전남	여수시	202001	전문서비스	상담	결혼정보회사	0	0	10	3060000	15	306000	204000
8	전남	여수시	202001	전문서비스	상조	상조서비스	0	0	4	7991000	22	1997750	363227
9	전남	여수시	202001	전문서비스	상조	장례식장	0	0	4	27180000	6	6795000	4530000
10	전남	여수시	202001	전문서비스	상품권	농협-상품권	0	0	1	940000	45	940000	20889
11	전남	여수시	202001	전문서비스	약국	약국	0	0	115	495236000	30629	4306403	16169
12													
13													
14													

Ministry of Data and Statistics. All Rights Reserved.

01 다자간 피처 스토어란?
데이터 거래 예시 2



Ministry of Data and Statistics. All Rights Reserved.

01 다자간 피처 스토어란?
피처 기반 데이터 거래

피처 스토어의 역사와 개념

- 피처 스토어는 머신러닝 모형을 개발하고 운영할 때, **데이터 피처(data feature; 속성, 변수)를 중앙에서 저장·관리·제공하는 플랫폼**으로 2017년 우버(Uber) 연구진이 처음 제시한 용어
- 피처는 **머신러닝 모델의 입력값으로 사용되는 데이터**를 의미하며, 예를 들어 기업통계등록부(SBR)의 여러 변수(컬럼, 속성)의 값들(예. 성별 종사자 수 등) 또는 그것들을 가공한 값들(예. 여성 종사자 비율 등)
- 피처 스토어는 이러한 피처를 효율적으로 관리하여 **데이터 엔지니어링 작업을 단순화하고, 데이터의 일관성과 재사용성을 높이는 기능**을 수행
- 일반적으로 피처 스토어는 **한 기업 내에서 다양한 팀이 다양한 모형을 다룰 때**, 동일한 피처, 특히 가공한 값으로 만들어진 **피처를 공유하고 재사용**할 수 있도록 지원하는 역할

Ministry of Data and Statistics, All Rights Reserved.

01 다자간 피처 스토어란?
기본 구성요소

기준 데이터 (HINGE DATA)
- 통계등록부 활용 다기관 데이터 연계

데이터 센터
- 망분리 보안구역 내 원자료 보관

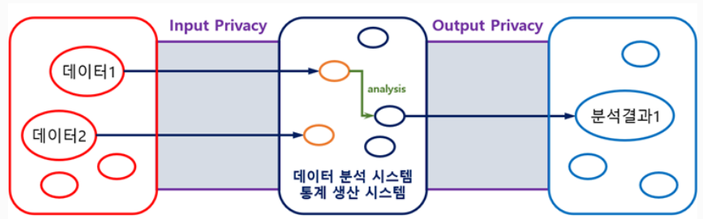
보안 클라우드
- 안전한 연계 데이터의 반복 분석 및 API 구독

데이터 트윈 (DATA TWIN)
- 재현자료(synthetic data) 기반 사전 분석 환경

프라이버시 문제 해결을 어떻게?
이용자 편의성 확보를 어떻게?
생태계 구축을 어떻게?

Ministry of Data and Statistics, All Rights Reserved.

- **[참고]** 프라이버시 문제 개념도 / 프라이버시 강화 기술(PET) 가이드 (UN, 2023)



- 센서스 자료에 대해서 지도상에 빈도표를 제공할 때, 5미만은 NA로 표기

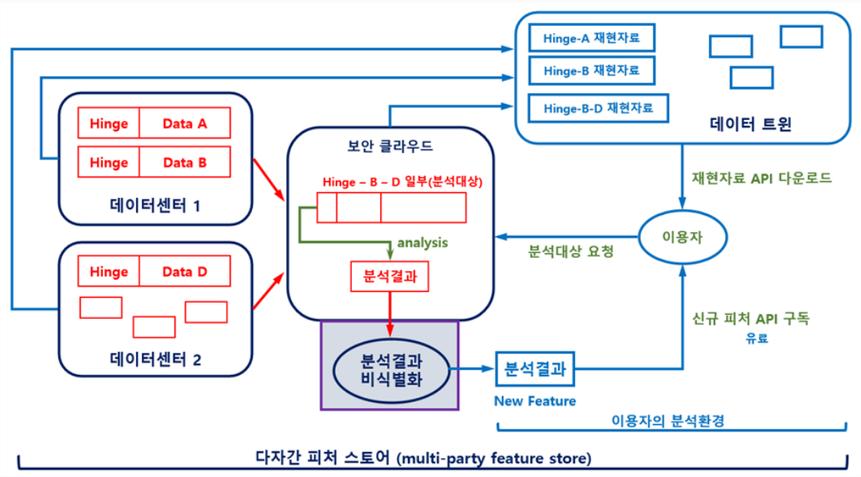
이용자의 조건 선택	시스템이 제공하는 빈도수
남 + 50~65세 + 박사 + 기혼	5
남 + 50~65세 + 박사 + 기혼, 이혼	6
남 + 51~65세 + 박사 + 기혼, 이혼	5
남 + 50세 + 박사 + 이혼	NA

- 해당 지역에 사는 개인의 혼인 정보 노출 가능

	남, 51~65세, 박사	남, 50세, 박사	남, 50~65세, 박사 합(margin)
기혼	5	0	5
이혼	0	1**	1
합(margin)	5	1	6

Ministry of Data and Statistics, All Rights Reserved.

- **다자간 피처 스토어 개념도**
  (원칙) 시장 기능 확보 추구 → 거래된 피처에서 발생한 수입을 분배 (피처 생산에 수반된 비용을 근거로 분배)



Ministry of Data and Statistics, All Rights Reserved.

02
데이터센터의 HINGE DATA 운영



Ministry of Data and Statistics. All Rights Reserved.

02 데이터센터의 HINGE DATA 운영
통계등록부, 데이터센터 소개

4827	84217	3958172	294
1053	56329	8172463	176
7641	02748	6429581	083
2389	39106	1048375	509
5914	71853	2874016	362
8260	60524	9381627	741
3478	26490	5738902	128
9142	95037	6852710	453
6605	13674	4098653	617
2718	78251	9263148	285

통계등록부(Statistical Register)란?
- 개인, 법인 또는 단체 등에 관한 분야별 모집단의 기본정보를 수록하여 관리하는 '주제별 모집단 자료'
- 인구·가구·주택·기업 총 4종의 기본통계등록부, 취업활동·아동가구·청년 총 3종의 정책맞춤형 통계등록부

데이터센터
- 망분리 보안 구역으로, 이용자가 **정해진 업무 시간**에 **방문**하여 데이터를 분석할 수 있는 공간
- 분석결과물에 대해 반출 심의 절차 운영 → 마이크로데이터, 분석결과물 모두 보안(security), 보호(privacy)

Ministry of Data and Statistics. All Rights Reserved.

02 데이터센터의 HINGE DATA 운영
HINGE DATA로서 통계등록부

24857	19362	04781	56234	91206	4827	39162	75804	02489	61735	84217
73629	10584	39217	68453	27196	1053	15473	59628	87301	32469	70852
81530	60472	92715	03186	45827	7641	63219	04187	59632	27814	96370
27991	86045	12468	59273	70631	2389	48502	73961	12508	64973	30541
93451	27836	65029	18764	50289	5914	06734	98251	43017	75628	19163
16372	90514	73286	41857	26930	8260	22875	61309	84572	93016	27485
50728	34109	68023	92517	13468	3478	59231	17064	36829	70493	58106
28640	71963	04579	38126	67295	9142	81027	45368	91725	36014	28978
63108	25047	89361	70428	31590	6605	57650	28917	60431	81572	49208
97035	18624	53892	26174	40731	2718	14783	56029	89374	31206	67845

데이터센터의 HINGE DATA 운영
- 이용 요청이 많은 데이터와 해당 통계등록부를 사전에 연계하여(필요시 가명결합) 데이터센터 내 보관
- **통계등록부에 포함된 다양한 변수를 연계키 및 집계키로 이용**하여 데이터 연계
  (예) 지역변수가 없는 **마이크로데이터 A** 및 **B**를 통계등록부 **H**를 연결하여 데이터센터에서 보관 **AHB**
  → 통계등록부의 지역 변수를 기준으로 집계한 **H'A'B'** → **지역별 매크로데이터 C**와 연계 및 **H'A'B'C** 분석

Ministry of Data and Statistics. All Rights Reserved.

02 데이터센터의 HINGE DATA 운영
데이터센터의 기타 역할



데이터센터의 공통도구 및 분석결과 비식별화 도구 운영
- **(공통도구)** 모든 데이터에 공통적으로 적용해 표준화, 품질평가, 재현자료 생성을 수행하는 도구 개발
- **(분석결과 비식별화 도구)** iLBA 및 DISK 등을 활용하여 분석결과 비식별화 도구 운영
  (iLBA) 위계변수가 있는 하나의 마이크로데이터에서 대량의 빈도표를 생성할때 프라이버시 보호 알고리즘
  (DISK) 기업통계등록부를 대상으로 분석결과 비식별화 자동 프로그램(Excel+R) 시범 개발('23.12)

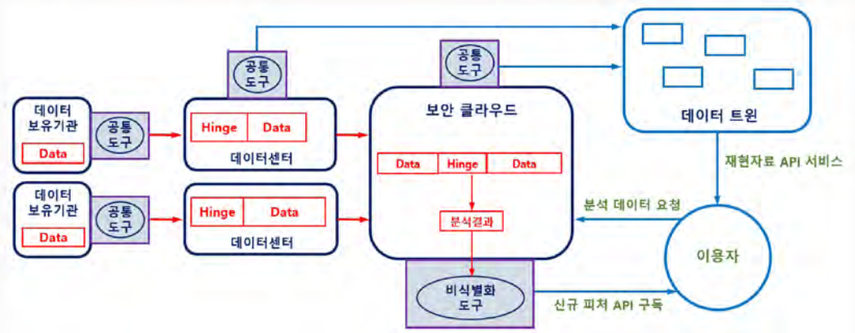
Ministry of Data and Statistics. All Rights Reserved.

03
데이터트윈 및 보안 클라우드



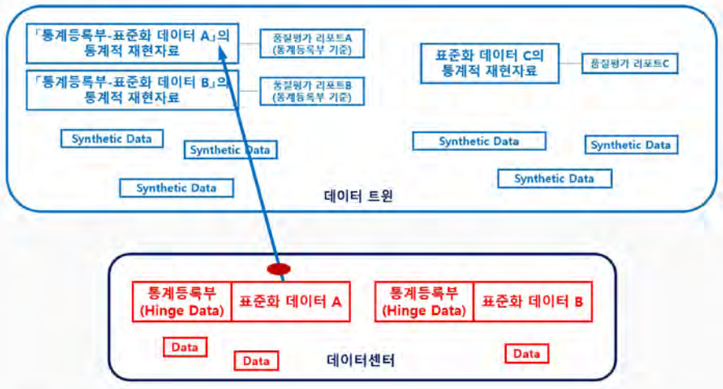
Ministry of Data and Statistics, All Rights Reserved.

03 데이터트윈 및 보안 클라우드
보안 클라우드에서 원데이터 분석



- 데이터센터 내 데이터 중, 이용자가 요청한 부분만 보안 클라우드로 업로드 및 분석
- 원데이터 내 민감한 컬럼에 대해서 동형암호 활용 등, 다양한 프라이버시 강화 기술(PET) 활용 가능

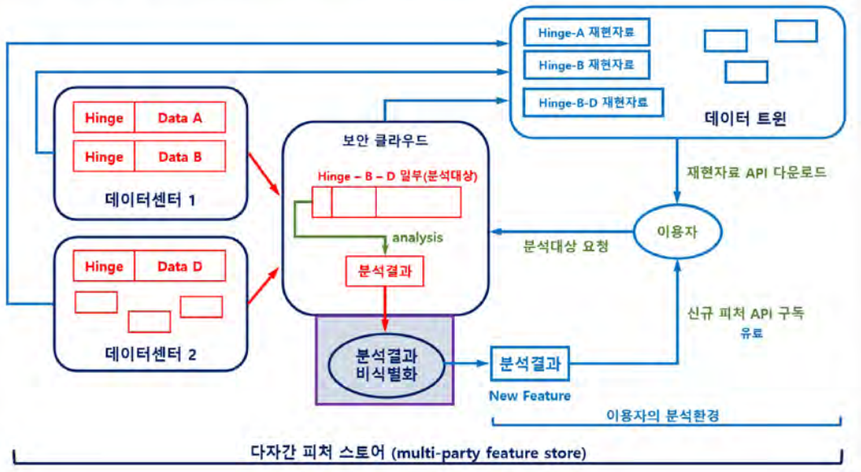
03 데이터트윈 및 보안 클라우드
재현자료(synthetic data)로 구성된 데이터트윈(data twin)



- 공통도구를 활용해 재현자료를 생성하며, 재현자료의 유용성 정도에 따라 유료화 가능
- 데이터트윈을 통한 이용자 편의성 확보

Ministry of Data and Statistics, All Rights Reserved.

- **다자간 피처 스토어 개념도**
  (원칙) 시장 기능 확보 추구 → 거래된 피처에서 발생한 수입을 분배 (피처 생산에 수반된 비용을 근거로 분배)



Ministry of Data and Statistics, All Rights Reserved.

04
정리 및 향후 발전방향



Ministry of Data and Statistics. All Rights Reserved.

04 정리 및 향후 발전방향
다자간 피처 스토어의 한계와 향후 추진방향

기술 개념을 제시하였으나, 표준화, 거버넌스, 보상 등에 대한 개념 검증 필요

- 모델 제안 단계로, 실제 대규모 운영, 성능, 비용, 안정성 등에 대한 파일럿 개념 검증 과정 필요
- 기관별 데이터 구조가 다르므로, HINGE DATA, 표준화, 품질평가 등에 있어 관리 기준을 맞추어야 함
- 거버넌스, 보상체계 등의 세부 내용을 완성해야 함

실증, 운영, 생태계 설계 중심의 사업 추진

- 그간 연계에 성공하지 못했던 데이터를 대상으로 본 논문에 제시된 개념을 실증하고, 성공 사례를 바탕으로 적용 범위 확대
- 표준화, 품질평가, 재현자료 생성을 담당하는 공통도구 및 분석결과 비식별화 도구를 이용해 반복 가능한 운영 프로세스 수립
- 데이터 사이언티스트, 데이터 스튜어드 등, 데이터 전문가가 운영 주체로 참여하는 체계 마련

부처간 데이터 공유에만 310일... 꽉막힌 韓 '전자정부' (한국경제 2026.05.08)



Ministry of Data and Statistics. All Rights Reserved.

04 정리 및 향후 발전방향
왜 '다자간' 피처 스토어인가?

데이터 주권을 유지하고, 프라이버시를 보존하며 데이터를 연계

- 원천 기관이 데이터 소유권을 유지하며, 데이터센터 내 안전한 원자료 보관으로 참여 기관간 신뢰 구축
- 원자료 전체를 직접 이동, 공유하지 않고, 연계 데이터 분석결과인 피처 구독으로 개인정보 노출위험을 낮춤

표준화, 품질평가 자동화

- 공통도구로 표준화, 품질평가, 재현자료 생성을 수행해 기관별 데이터 편차를 줄임

운영, 보상 체계 구축 가능

- 데이터 거래가 아닌 피처 서비스 관점에서 데이터 재사용, 비용 분담, 수익 배분 논의 구체화 가능
- 시장 기능이 도입된 데이터 생태계 구축 방향 제시

안전하고 접근성 높은 데이터 활용 범위 확대

- 공공, 민간 데이터 보유기관이 참여해 새로운 분석, 다양한 피처의 API 서비스를 만들 수 있는 기반 제공

원본 데이터 전체의 이동 없이, 여러 기관의 데이터를 연결·분석 가능한 형태로 바꾸는 안전한 협업 인프라

Ministry of Data and Statistics. All Rights Reserved.

감사합니다 :)

END OF DOCUMENT

2026.05.14

행정자료 기반 주거비지수 작성방안

2026 국가통계·데이터 발전포럼

국가데이터처 물가동향과
홍 현 정 서기관

Contents

검토배경 01
그간 진행사항 02
행정자료 기반 주거비지수 작성 개요 03
자가주거비 추정방법 검토 04
집세지수 시산결과 05
자가주거비 추정 시산결과 06
향후 계획 07

SECTION 01

검토배경

01
검토배경

1) 현장조사 한계와 불응률 상승

집세조사 매월 1.2만 가구 방문 면접조사

집세조사 불응률 지속 증가 추세



현행 표본조사 방식의 한계

- 응답부담 가중 및 불응 증가
- 표본 한계로 인한 대표성 저하
- 조사 예산 및 행정비용 증가

➡ 한계에 도달한 현장조사 대체 방안 마련 필요

01
검토배경

2) 자가주거비 주지표 포함 요구

자가주거비가 주지표에 포함되지 않아 **체감물가 차이 우려**

- 현재 우리나라는 소비자물가지수 주거비 항목으로,
  집세(전세+월세)만 반영하고 있으며, 자가주거비는 보조지표로만 작성·공표 중

자가주거비
가계 소비지출 내 가장 큰 비중

★ **자가주택이 제공하는 주거서비스에 대한 비용**

자가주거비 ≠ 집값

주요 지적 및 요구사항

01 **체감물가 괴리**
집값 급등기 등에 국민이 실제 느끼는 주거비 부담과 공식 지표 간 차이 발생

02 **정책 설명력 약화**
거시 경제 및 통화정책 수립 시 정확한 인플레이션 판단 지표로서의 기능 저하

03 **국회·언론 지속 지적**
'21~'25년 국정감사 연속 지적 및 주요 언론의 지표 개편 요구 증대

04 **주거비 현실 반영 필요**
실질적 주거 비용 변동을 포착하기 위한 구조적 개편 논의 점화

05

SECTION 02
진행사항

02
진행상황

1) 추진 사항

안정적인 행정자료 기반 전환을 위한 **단계별 연구 및 검증 체계 구체화**

2023 연구용역 실시
자가주거비 산정방식 개선 및 주지표 전환 연구 (한국부동산분석학회)

2024 행정자료 정례 입수 및 시범 시산
국토부 전월세신고 자료 입수 정례화 및 방안 마련

2025 전문가 자문 및 시산결과 축적
의견수렴 및 행정자료 기반 주거비지수 시산결과 축적 및 비교 분석 실시

2026 ★ 의사결정 준비
국가통계위원회 심의 준비: 작성방법 변경 및 자가주거비 주지표 포함여부

07

02
진행상황

2) 그간 전문가관계기관 의견수렴 요지

공통 인식

행정자료 활용의 작성방식 전환은 대다수 공감

기존 현장조사 대신 **행정자료를 활용한 집세지수로의 작성방법 전환 필요**

신중론

자가주거비 주지표 포함은 단계적으로 접근

자가주거비의 CPI 주지표 포함은 지수의 안정성 및 파급영향에 대한 충분한 **사전 검토 후 중장기적으로 접근**해야 한다는 의견

08

SECTION 03

행정자료 기반 주거비지수 작성 개요

SCOPE

행정자료 이용범위

구분	신고대상 / 비고	입수기간	전세 (월평균)	월세 (월평균)
신고의무제 이전	강제사항 아님 (확정일자 중심)	'17.11 ~ '21.05	약 8.2만건	약 6.4만건
신고의무제 ('21.6~ / 계도기간)	전국(일부 군 제외) + 보증금 6천만원 또는 월세 30만원 초과 임대차 계약	'21.06 ~ '25.05	약 9.5만건	약 11.1만건
신고의무제 ('25.6~ / 계도폐지)	계도기간 폐지로 정합성 향상 기대	'25.06 ~ '26.02	약 6.9만건	약 12.5만건

입수 및 분석 기간

2017년 11월 ~ 2026년 2월 (약 8년)

활용 자료 기준

국토교통부 전월세 신고자료

확정일자 또는 임대차 전월세 신고일자 기준

신고의무제 이후 데이터 규모가 **월 약 20만 건 수준으로 확대**되어 **현실 반영도와 정합성 제고 기반 마련**

03
주거비지수 작성 개요

행정자료 기반 시산 프로세스 개요

전세·월세·자가주거비 지수 산정을 위한 전체 흐름도(2017.11~2025.02 자료 활용)

실행 프로세스 구축

자료 입수	기본 정제	매칭키 생성	반복거래쌍 구성	변동률 산출	지수화·가중치 적용
국토부 전월세 신고 및 확정일자 원자료 수집	주소·주택유형 누락, 면적/가격 오류 등 11단계 기초 정제	지역 + 세부 주소 +면적 정보 기반 동일 주택 1:1 연계	동일 중복 제거 후 과거 26개월 내 매칭 거래쌍 도출	단위면적당 가격 로그변환, 이상치 제거 및 기하평균 적용	계약 변동률 누적합산, 품목 및 내부규격 가중치 반영
월 약 20만건					2025=100

지수 산정

전세지수	월세지수	자가주거비지수
유효표본 내 단위면적당 전세금액 변동 지수화	월세 환산금액의 변동을 측정 * 환산금액 [보증금 ×전환율+ 월세 ×12/면적] 변동 측정	①월세 자료만 활용, ③전·월세 가중 결합 … 등

03
주거비지수 작성 개요

반복거래쌍 생성 원리

표본 고정 없이 매월 **실제 신고건과 과거 매칭된 신고건의 계약변동 내역을 포착**

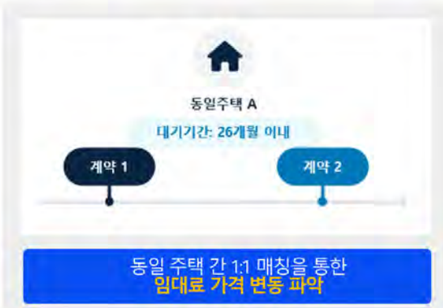
변동 거래 모두 포함

- 가격 변동 (±α%) 거래
- 가격 무변동 (0%) 거래

실제 시장의 모든 변동 양상을 지수에 반영

중복 거래 제외

해당 월에 2번 이상 거래 시 **최근 등록 건 기준** 포함



동일 주택 간 1:1 매칭을 통한
임대료 가격 변동 파악

COMPARISON

현행 집세지수와 행정자료 기반 지수 비교

구분	현행 (현장조사)	행정자료 기반
표본규모	매월 약 1.2만 가구 (표본)	약 17만 ~ 21만 건 (전수수준)
조사방식	조사원 대면/방문 면접조사	행정자료 입수 (전월세신고/확정일자)
조사기준일 / 범위	매월 5일 (전월 6일~금월 5일) 전국 40개 도시 지정 표본	매월 5일 (기준일 동일) 전국 40개 도시 대상 전체
특징 (장점/한계)	(장점) 동일 표본 추적으로 품질 차이 통제 용이 (한계) 응답 거부 증가, 신규 계약/시장 변화의 지연 반영 우려	(장점) 전수 자료로 대표성 제고, 신규계약 변동을 매월 적기 포착 (한계) 데이터 정제 작업 상당 소요, 주택 품질 차이 보정 필요 (→동일주택가정 및 반복거래쌍 매칭기법 활용)

행정자료 전환 시 기대효과

- 대표성 및 정확성 확대
- 시장변화 적시성 개선
- 조사부담 및 비용 완화

지수 작성기준은 동일하게 유지, 데이터 자료원을 전수 수준으로 확장

13

SECTION 04

자가주거비 추정방법 검토

COMPARISON

항목별 작성방법 비교

항목	현행 지수작성 방식	행정자료 기반 지수작성 방식
전세	표본가구 전세금액 변동 지수화	26개월 이내 반복거래쌍 단위면적당 전세금액 변동 지수화
월세	표본가구 월세화 금액 변동 지수화	26개월 이내 반복거래쌍 환전월세 환산금액 변동 지수화
집세 (전세+월세)	전세지수 + 월세지수 가중평균	행정자료 기반 전세지수 + 월세지수 CPI 가중치 적용 결합
자가주거비	집세지수 차용 (가중치만 자가주거비 적용)	임대료상당액 접근법 반영 ① 월세자료만 이용 ② 전세+월세 모두 이용 등

산식 및 구조적 차이

표본 고정형 개별 변동률 추적 →
26개월 미고정 반복거래쌍 기하평균

가중치 및 기준시점

품목(전/월세) 및 40개 도시 내부규격(아파트/비아파트) 반영
행정자료 기준시점 2025=100 설정

자가주거비 전용 산출체계

단순 집세 전용에서 벗어나 실제 임대시장 체감도를 반영하는 모델 구조 설계

지수 산출 구조는 유지하되, 풍부한 행정 데이터 기반의 산식 고도화로 지표의 대표성 향상

14

COMPARISON

자가주거비지수 – 추정방식 개괄

자가주거비는 직접 관측이 불가능하기 때문에 다양한 접근법을 통한 추정 필요

임대료상당액 접근법

임대료 변동

자가주택을 임대할 경우의 임대료 상당액

장점
통계적 안정성 높음, 다수 국가 채택

한계
주택가격·금리 변화 즉시 반영 한계

사용자비용 접근법

금리 · 주택가격

기회비용+이자+감가상각-자본이득

장점
개념적 충실, 체감물가 괴리 완화

한계
지표 변동성 증가, 모순적 현상 발생

순취득액 접근법

신규 주택가격

신규 취득 주택구조물 관련 비용

장점
인플레이션 측정 강조, 괴리 축소

한계
기존주택 미반영, 지수 변동성 확대

지불 접근법

현금 지출 (금리)

당보대출 이자, 재산세 등 현금지출

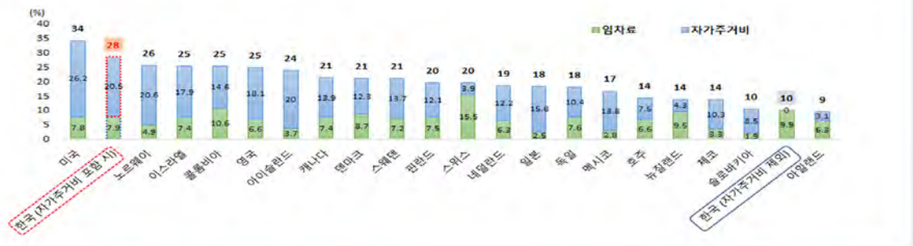
장점
대체 가격을 사용하지 않음

한계
금리 변동에 취약, 무대출 가구 미반영

16

COMPARISON

국가별 CPI 내 주거비 비중 비교



* 한·미·일 주거비(집세, 자가주거비) 가중치 비중

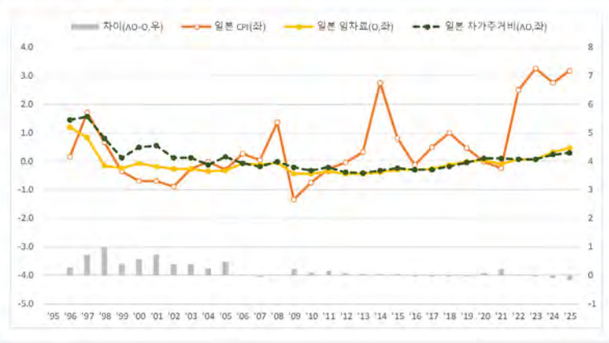
한국 (Korea)	미국 (USA)	일본
28.4%	34.0 %	18.3%
보조지표 기준	상대적으로 가장 높은 비중을 보임	상대적으로 낮은 비중을 보임

17

CASE STUDY

국가 사례 – 일본

임대료상당액 접근법 채택



주: COICOP2018 기준, OECD STAT

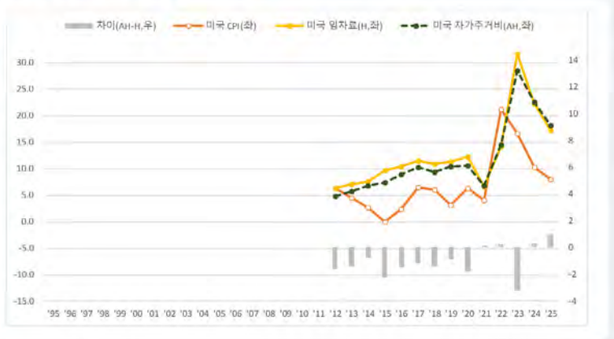
임대료 기반 안정성

19

CASE STUDY

국가 사례 – 미국

임대료상당액 접근법 채택



주: COICOP2018 기준, 미국 노동통계국(BLS)

OER 개념 (Owners' Equivalent Rent)

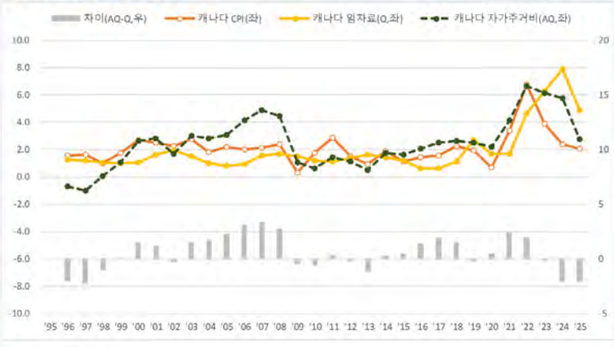
임대료 기반 안정성

18

CASE STUDY

국가 사례 – 캐나다

사용자비용 접근법 채택



주: COICOP2018 기준, OECD STAT

주택 소유 및 거주에 수반되는 제반 비용을 반영

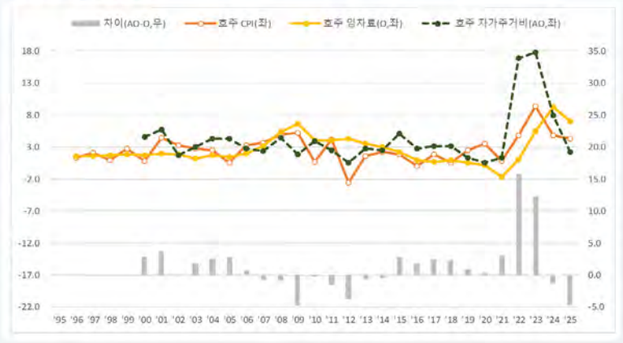
즉 기회비용, 대출 이자, 감가상각비, 재산세, 유지보수비 등

20

CASE STUDY

국가 사례 – 호주

순취득 접근법 채택



주: COICOP2018 기준, OECD STAT

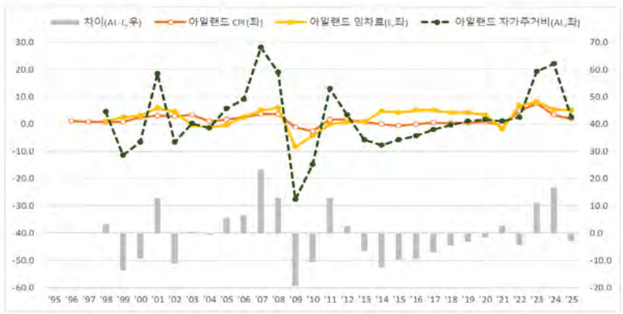
신규주택 취득가격을 반영

구조물 중심(토지제외)

CASE STUDY

국가 사례 – 아일랜드

지불 접근법 채택



주: COICOP2018 기준, OECD STAT

실제 주거 관련 현금 지출 반영

모기지 이자 지급액, 재산세, 주택보험료, 유지보수 비용 등

SIMULATION

주지표 포함 시 가중치 변화



가중치 비율 변화 (전분비)

9.9% → 28.3%

타 품목 가중치 파급

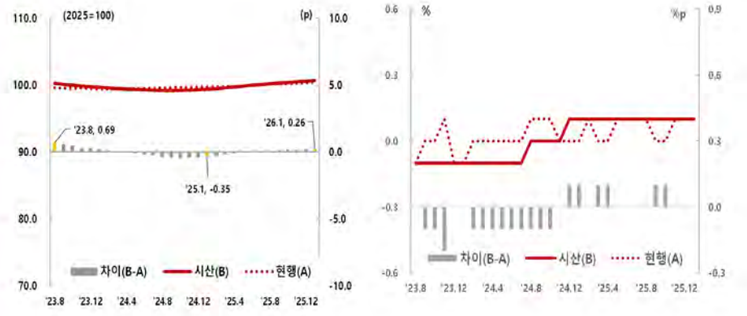
비례적 하락

SECTION 05

집세지수 시산결과

RESULTS

전국 전세지수 비교



주요 지표 요약

평균 지수 차이 (B-A)

0.02p

전월비 방향성 일치율

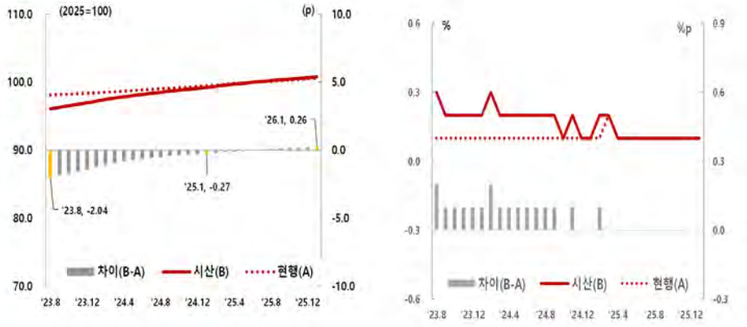
61.5%

전월비 표준편차
0.0 수준으로
안정적 변동성 유지

➡ 행정자료 기반 전세지수는 현행 지수와 매우 유사한 흐름

RESULTS

전국 월세지수 비교



주요 지표 요약

평균 지수 차이 (B-A)

0.02p

전월비 방향성 일치율

100.0%

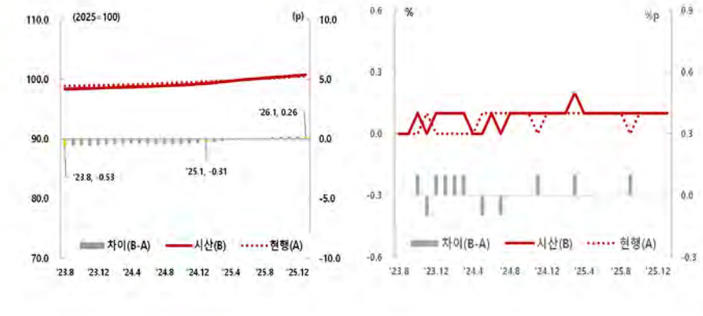
전월비 표준편차
0.0 수준으로
안정적 변동성 유지

➡ 행정자료 기반 월세지수는 현행 지수와 유사한 흐름

- 다만 월세의 경우 신고제(21.6월) 이전 자료의 경우 시행 이후 대비 신고건수가 약 57% 내외 수준으로
- 시행 전후의 자료량의 차이로 인해 시산 초기 계열 갭(gap)이 다소 크게 발생

RESULTS

전국 집세지수(전세+월세) 비교



주요 지표 요약

평균 지수 차이 (B-A)

0.02p

방향성 일치율

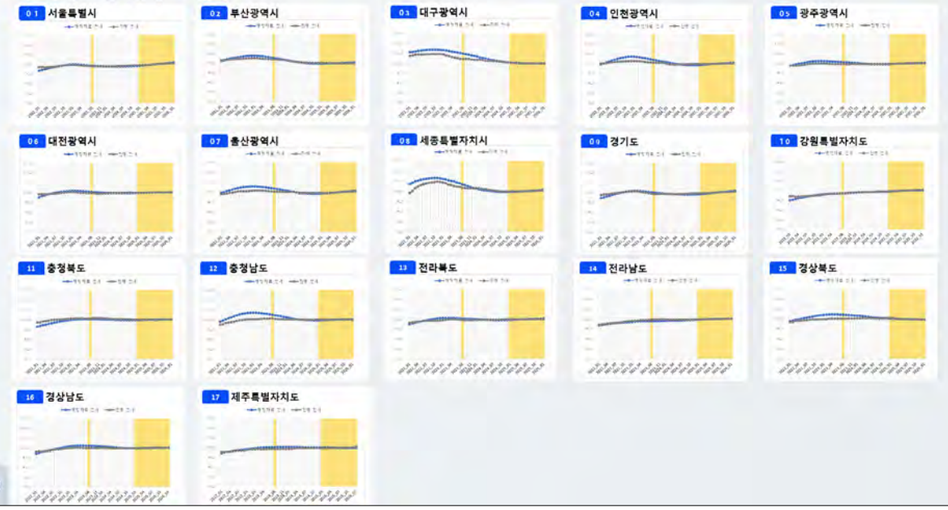
92.3%

전월비 표준편차
0.0 수준으로
안정적 변동성 유지

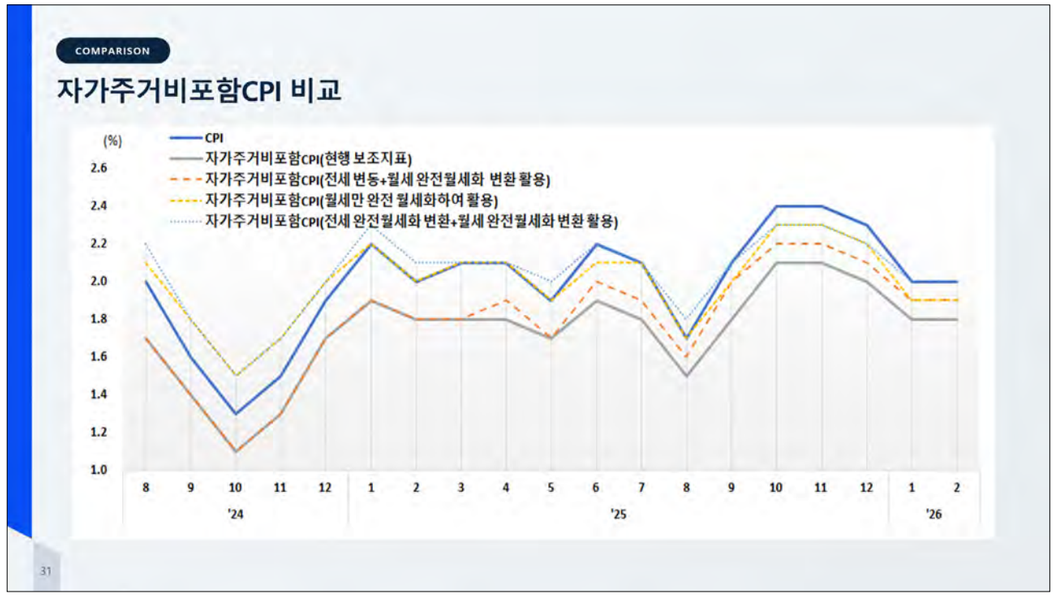
➡ 행정자료 기반 집세지수는 현행 지수와 비교적 유사한 흐름

RESULTS

시도 전세지수 비교

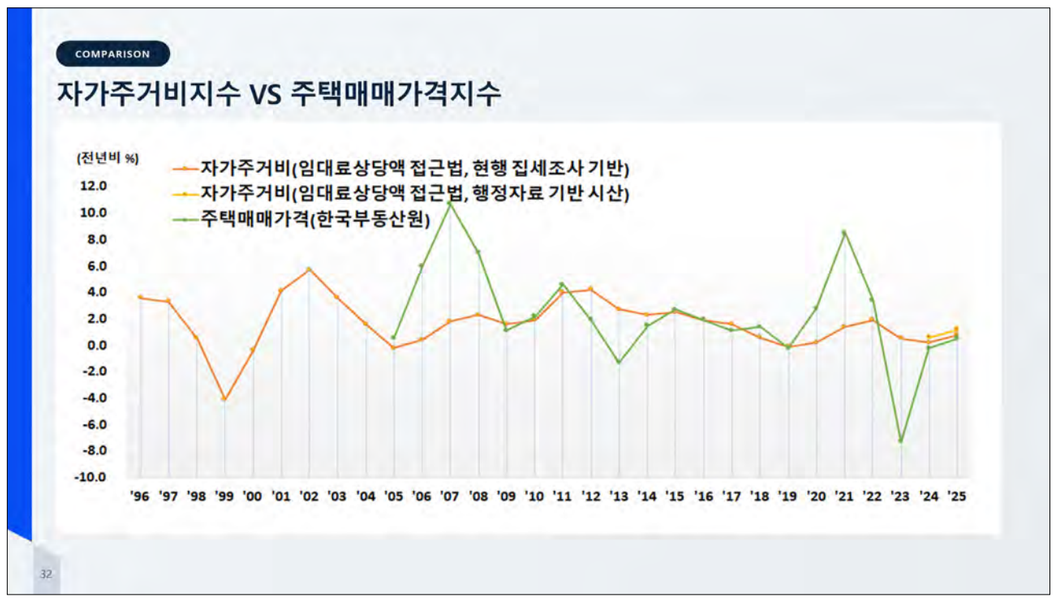






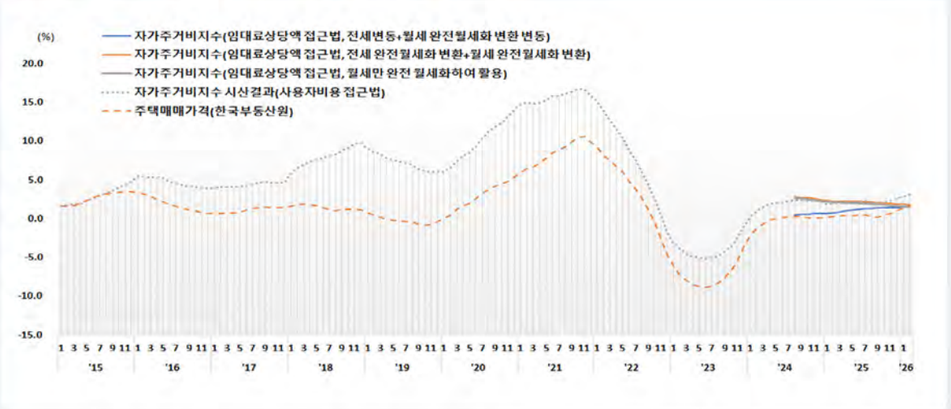
SECTION 06

자가주거비 추정 시산결과



COMPARISON

임대료상당액 접근법 VS 사용자비용 접근법



33

IMPLICATIONS

시사점

가중치 산출방식, 우리나라 **전세제도의 특수성** 반영, **지역간** 국민이 느끼는 **체감 차이** 등을 충분히 고려하여 신중하게 추진할 필요

집세지수 작성방법 개선

전수 기반의 국토부 전월세신고 행정자료를 활용해서 집세지수의 현실체감도 제고

자가주거비 주지표 포함여부

개선된 자가주거비 추정결과를 충분히 쌓아가면서 장기적인 검토 필요

시산결과를 면밀히 검토하고 지수의 안정성에 대해 전문가, 관계기관과 충분히 협의한 후,
'26년 12월 말 '25년 기준 소비자물가지수 개편결과 공표 시에 결과 발표 예정

34

SECTION 07

향후 계획

07

향후 계획

최종 의사결정 로드맵

검증 - 공론화 - 의사결정 3단계 흐름을 통한 주거비지수 작성 전환



36

Session

03-2

AI 에이전트와 사회조사

좌장

강현철 교수(호서대학교 빅데이터AI학부)

1. 대형(언어)모델이 조사연구에 가져오는 새로운 가능성과 위협

   강동현 교수(연세대학교 사회학과)

2. AI Agent와 사회현상 연구

   한경식 교수(한양대학교 데이터사이언스학부)

3. LLM을 이용한 텍스트 분석

   한상원 교수(성균관대학교 사회학과)

토론

강정한 교수(연세대학교 사회학과), **이영구 상무**(피앰아이)

대형 (언어) 모델이 조사연구에 가져오는 새로운 기회와 위협

2026년 5월 14일

연세대학교 사회학과
강동현

사회(과)학적으로 AI 연구하기

인공지능의 영향을 사회(과)학적으로 어떻게 조망할 것인가?

AI로 사회(과)학 연구하기

빠르게 발전하는 AI 기법을 사회(과)학 연구에 어떻게 활용할 것인가?

사회(과)학적으로 AI 연구하기

인공지능의 영향을 사회(과)학적으로 어떻게 조망할 것인가?

AI로 사회(과)학 연구하기

빠르게 발전하는 AI 기법을 사회(과)학 연구에 어떻게 활용할 것인가?

AI로 사회(과)학 연구하기

PNAS

PERSPECTIVE | SOCIAL SCIENCES |

Can Generative AI improve social science?

Christopher A. Bail | Authors Info & Affiliations

Edited by David Lazer, Northeastern University, Boston, MA; received September 7, 2023; accepted April 5, 2024, by Editorial Board Member Mark Granovetter

May 9, 2024 | 121 (21) e2314021121 | https://doi.org/10.1073/pnas.2314021121

THIS ARTICLE HAS BEEN UPDATED

Bail, C. A. (2024). Can generative AI improve social science?. *Proceedings of the National Academy of Sciences*, *121*(21), e2314021121.

AI로 사회(과)학 연구하기


Sociological Methods & Research

Impact Factor: 6.5 / 5-Year Impact Factor: 7.1

Volume 54 Issue 3, August 2025 Previous issue Next issue

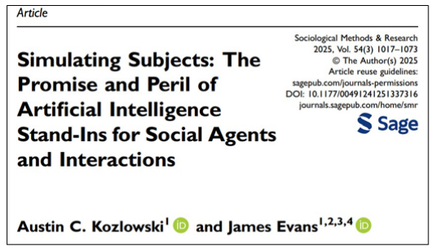
Special Issue: Integrating Generative AI into Social Science Research

Guest Editor: Thomas Davidson | Guest Editor: Daniel Karell

https://journals.sagepub.com/toc/smra/54/3

빠르게 발전하는 AI 기법을 사회(과)학 연구에 어떻게 활용할 것인가?

AI로 사회(과)학 연구하기


Article

Simulating Subjects: The Promise and Peril of Artificial Intelligence Stand-Ins for Social Agents and Interactions

Sociological Methods & Research
2025, Vol. 54(3) 1017–1073
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00491241251337316
journals.sagepub.com/home/smr
Sage

Austin C. Kozlowski[1] and **James Evans**[1,2,3,4]

AI로 사회(과)학 연구하기

"it is important to remember that **these AI systems were not designed for social science**, and there is no promise that the texts they generate will be faithful representations of the social groups they attempt to mimic. They are **fundamentally next-word predictors** fine-tuned to answer questions and assist with tasks. In the process of learning to predict words, these models learn a wealth of information about different cultures and interaction styles. Effectively extracting or leveraging this information is not straightforward, however."(p. 1023)

AI 시스템은 사회과학 연구를 위해 설계된 것이 아니며, 이들이 생성하는 텍스트가 모방하려는 사회 집단을 충실히 재현한다는 보장은 없다는 점을 기억하는 것이 중요하다. 이 모델들은 **근본적으로 다음 단어를 예측하도록** 학습된 뒤, 질문에 답하거나 작업을 돕도록 미세 조정된 예측기다. 단어를 예측하는 과정을 학습하면서, 이 모델들은 다양한 문화와 상호작용 방식에 대한 방대한 정보를 습득한다. 그러나 이 정보를 효과적으로 추출하거나 활용하는 일은 결코 간단하지 않다

AI로 사회(과)학 연구하기

Principles for Simulating Subjects and Interactions

A) Situating Actors With Prompts

B) Situating Actors With *Model Internals*

C) Simulating Social Interaction

AI로 사회(과)학 연구하기

Principles for Simulating Subjects and Interactions

A) Situating Actors With Prompts

B) Situating Actors With *Model Internals*

C) Simulating Social Interaction

Kozlowski, A. C., & Evans, J. (2026). Simulating Subjects: The Promise and Peril of Artificial Intelligence Stand-Ins for Social Agents and Interactions. *Sociological Methods & Research*, 00491241251337316.

AI로 사회(과)학 연구하기

A) Situating Actors With Prompts

PA

Out of One, Many: Using Language Models to Simulate Human Samples

Lisa P. Argyle[1], Ethan C. Busby[1], Nancy Fulda[2], Joshua R. Gubler[1], Christopher Rytting[2] and David Wingate[2]

[1] *Department of Political Science, Brigham Young University, Provo, UT, USA. e-mail: lpargyle@byu.edu, ethan.busby@byu.edu, jgub@byu.edu*
[2] *Department of Computer Science, Brigham Young University, Provo, UT, USA. e-mail: nfulda@cs.byu.edu, christophermichaelrytting@gmail.com, wingated@cs.byu.edu*

In-Silico Sampling

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337-351.

AI로 사회(과)학 연구하기

A) Situating Actors With Prompts

1. 인종·민족적 자기정체성
2. 성별
3. 연령
4. 보수-진보 이념적 자기 위치
5. 정당 소속
6. 정치 관심도
7. 교회 출석 빈도
8. 가족 및 친구와 정치에 대해 이야기했는지 여부
9. 애국심
10. 거주 주

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337-351.

AI로 사회(과)학 연구하기

A) Situating Actors With Prompts

	Trump	Clinton
Racially, I am white. I like to discuss politics with my family and friends. Politically, I am a strong Republican. I attend church. I am 29 years old. I am a man. I am somewhat interested in politics. It makes me feel extremely good to see the American flag. I am from Louisiana. In the 2016 presidential election, I voted for	96%	4%
Racially, I am black. I never discuss politics with my family or friends. Politically, I am a strong democrat. I attend church. I am 58 years old. I am a man. I am not very interested in politics. It makes me feel extremely good to see the American flag. I am from New York. In the 2016 presidential election, I voted for	10%	90%

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337-351.

AI로 사회(과)학 연구하기

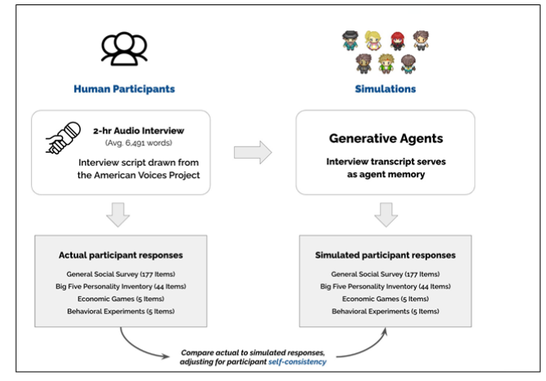
A) Situating Actors With Prompts

Variable	2012 Tetra.	2012 Prop. Agree	2016 Tetra.	2016 Prop. Agree	2020 Tetra.	2020 Prop. Agree
Whole sample	0.90	0.85	0.92	0.87	0.94	0.89
Men	0.90	0.85	0.93	0.88	0.95	0.88
Women	0.91	0.86	0.92	0.86	0.94	0.90
Strong partisans	0.99	0.97	1.00	0.97	1.00	0.97
Weak partisans	0.73	0.74	0.71	0.74	0.84	0.82
Leaners	0.90	0.85	0.93	0.87	0.95	0.89
Independents	0.31	0.59	0.41	0.62	0.02	0.53
Conservatives	0.84	0.84	0.88	0.86	0.91	0.89
Moderates	0.65	0.77	0.76	0.78	0.71	0.77
Liberals	0.81	0.95	0.73	0.95	0.86	0.97
Whites	0.87	0.82	0.91	0.85	0.94	0.89
Blacks	0.71	0.97	0.87	0.96	0.81	0.94
Hispanics	0.86	0.86	0.93	0.90	0.88	0.83
Attends church	0.91	0.86	0.93	0.88	0.94	0.88
Does not attend church	0.88	0.85	0.90	0.85	0.93	0.90
High interest in politics	0.95	0.90	0.97	0.93	0.97	0.92
Low interest in politics	0.71	0.74	0.75	0.75	0.83	0.81
Discusses politics	0.92	0.87	0.94	0.88	0.95	0.90
Does not discuss politics	0.83	0.82	0.81	0.79	0.80	0.79
18–30 years old	0.90	0.87	0.90	0.86	0.90	0.87
31–45 years old	0.90	0.85	0.92	0.87	0.94	0.90
46–60 years old	0.90	0.86	0.92	0.86	0.92	0.87
Over 60	0.90	0.85	0.93	0.87	0.96	0.91

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337-351.

AI로 사회(과)학 연구하기

A) Situating Actors With Prompts



Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., ... & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

AI로 사회(과)학 연구하기

A) Situating Actors With Prompts

Generative Agent Simulations of 1,000 People

Authors: Joon Sung Park[1]*, Carolyn Q. Zou[1,2], Aaron Shaw[2], Benjamin Mako Hill[3], Carrie Cai[4], Meredith Ringel Morris[5], Robb Willer[6], Percy Liang[1], Michael S. Bernstein[1]

Affiliations:

[1]Computer Science Department, Stanford University; Stanford, CA, 94305, USA.

[2]Department of Communication Studies, Northwestern University; Evanston, IL, 60208, USA.

[3]Department of Communication, University of Washington; Seattle, WA 98195, USA.

[4]Google DeepMind; Mountain View, CA 94043, USA.

[5]Google DeepMind; Seattle, WA 98195, USA.

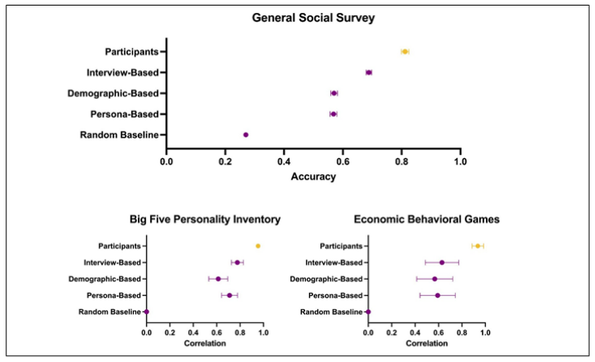
[6]Department of Sociology, Stanford University; Stanford, CA, 94305, USA.

*Corresponding author. Email: joonspk@stanford.edu

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., ... & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

AI로 사회(과)학 연구하기

A) Situating Actors With Prompts



Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., ... & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

AI로 사회(과)학 연구하기

B) Situating Actors With *Model Internals*

AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction*

Junsol Kim
Department of Sociology
University of Chicago

Byungkyu Lee†
Department of Sociology
New York University

Kim, J., & Lee, B. (2023). Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*.

AI로 사회(과)학 연구하기

B) Situating Actors With *Model Internals*

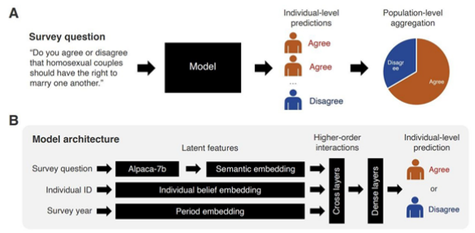
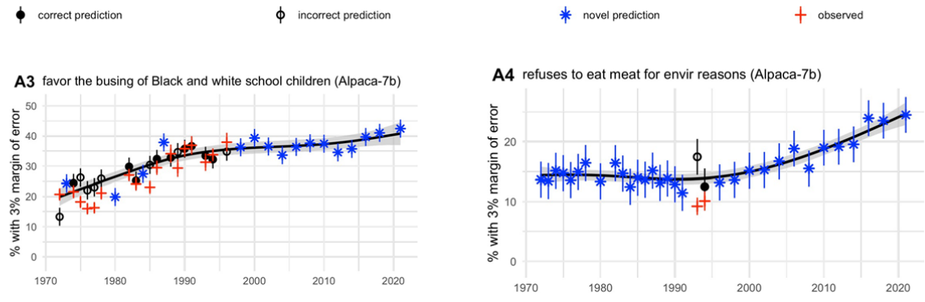


Figure 2: An overview of our methodological framework. In Panel A, we use survey weights when aggregating individual-level prediction into population-level estimates to account for potential sampling bias. In Panel B, individual belief and period embeddings are initially randomly assigned but optimized during the fine-tuning process using dense and cross layers. Semantic embedding, initially estimated by pre-trained LLMs (e.g., Alpaca-7b), is also optimized during the fine-tuning stage.

Kim, J., & Lee, B. (2023). Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*.

AI로 사회(과)학 연구하기

B) Situating Actors With *Model Internals*



Kim, J., & Lee, B. (2023). Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*.

AI로 사회(과)학 연구하기

PNAS

RESEARCH ARTICLE | POLITICAL SCIENCES

The potential existential threat of large language models to online survey research

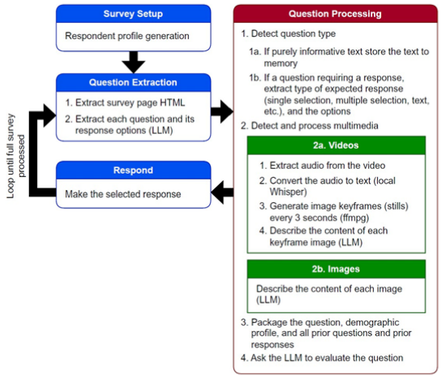
Sean J. Westwood Authors Info & Affiliations

Edited by James N. Druckman, University of Rochester, Rochester, NY; received July 9, 2025; accepted September 12, 2025 by Editorial Board Member Margaret Levi

November 20, 2025 | 122 (47) e2518075122 | https://doi.org/10.1073/pnas.2518075122

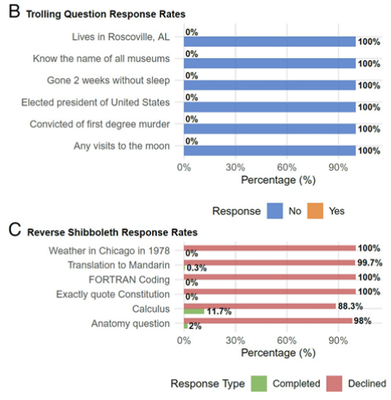
Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, *122*(47), e2518075122.

AI로 사회(과)학 연구하기



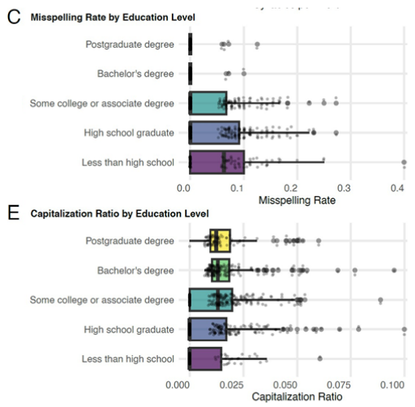
Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, *122*(47), e2518075122.

AI로 사회(과)학 연구하기



Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, *122*(47), e2518075122.

AI로 사회(과)학 연구하기



Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, *122*(47), e2518075122.

AI로 사회(과)학 연구하기

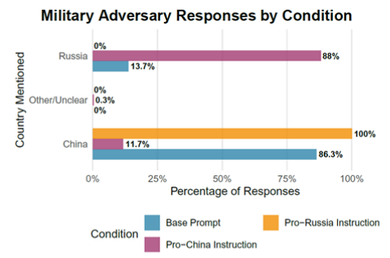


Fig. 6. A single instructional sentence can manipulate the synthetic respondent's answers about America's primary military rival, demonstrating the potential for malicious use. In the baseline condition, China is named 86.3% of the time. A pro-China prompt drops this to 11.7%, while a pro-Russia prompt increases mentions of China to 100%.

Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, *122*(47), e2518075122.

AI로 사회(과)학 연구하기

Article

Generative AI Meets Open-Ended Survey Responses: Research Participant Use of AI and Homogenization

Sociological Methods & Research
2025, Vol. 54(3) 1197–1242
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00491241251327130
journals.sagepub.com/home/smr
Sage

Simone Zhang[1], Janet Xu[2], and AJ Alvero[3]

Zhang, S., Xu, J., C Alvero, A. J. (2025). Generative AI meets open-ended survey responses: Research participant use of AI and homogenization. *Sociological Methods & Research*, *54*(3), 1197-1242.

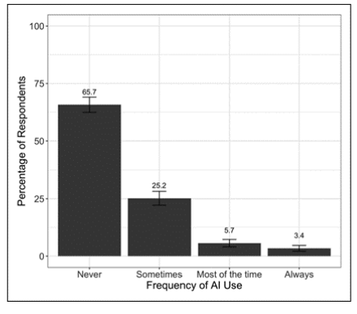
AI로 사회(과)학 연구하기

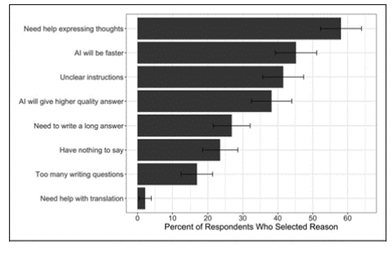
Table 2. Time-sharing Experiments for the Social Sciences (TESS) Questions.

Study and Principal Investigator(s)	Question	Sample size	Field period
"The Influence of the Terms We Use on Measures of Racial Attitudes" Ashley Jardina and LaFleur Stephens-Dougan	In the text box below, briefly describe what comes to mind when you think of [randomized: Blacks | African Americans] in American society today.	1,928	05/25/2020 - 03/23/2021
"Conditional Commitment? Partisan Identity and Support for Democracy in the United States" Matthew Graham.	What makes you similar to [own party: Democrats | Republicans]? What makes you different from [other party: Democrats | Republicans]?	1,930	08/31/2019 - 01/21/2020
"What Motivates Interest in Politics?" Eric Groenendyk	What is it about politics that makes you [uninterested | interested]?	346	04/13/2021 - 07/01/2021

Zhang, S., Xu, J., C Alvero, A. J. (2025). Generative AI meets open-ended survey responses: Research participant use of AI and homogenization. *Sociological Methods & Research*, *54*(3), 1197-1242.

AI로 사회(과)학 연구하기





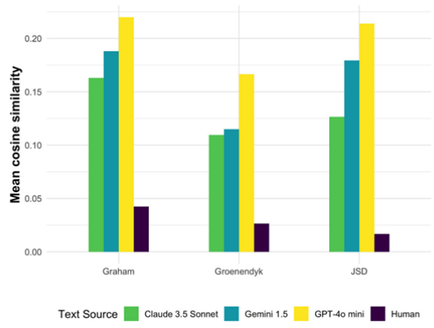
Zhang, S., Xu, J., C Alvero, A. J. (2025). Generative AI meets open-ended survey responses: Research participant use of AI and homogenization. *Sociological Methods & Research*, *54*(3), 1197-1242.

AI로 사회(과)학 연구하기

Table 6. Comparison of Word Count and Vocabulary Size Between Human and LLM-Generated Responses.

	Groenendyk			Graham			JSD		
Model	MW	SD	TV	MW	SD	TV	MW	SD	TV
Claude 3.5 Sonnet	103.0	43.4	1859	101.4	35.1	1061	99.0	41.2	1522
Gemini 1.5	123.5	69.6	2131	161.1	95.7	1487	175.8	106.6	2358
GPT-4o mini	195.4	75.7	2690	215.8	101.1	1802	214.5	78.2	2527
Human	9.7	9.0	462	31.7	25.2	1146	16.1	25.4	915

MW: mean # of words; SD: standard deviation # words; TV: total vocabulary.



Zhang, S., Xu, J., C Alvero, A. J. (2025). Generative AI meets open-ended survey responses: Research participant use of AI and homogenization. *Sociological Methods & Research*, *54*(3), 1197-1242.

AI로 자동화된 (사회)과학?

Machine-assisted social psychology hypothesis generation.

By Banker, Sachin,Chatterjee, Promothesh,Mishra, Himanshu,Mishra, Arul
American Psychologist, Vol 79(6), Sep 2024, 789-797

Abstract

Social psychology research projects begin with generating a testable idea that relies heavily on a researcher's ability to assimilate, recall, and accurately process available research findings. However, an exponential increase in new research findings is making the task of synthesizing ideas across the multitude of topics challenging, which could result in important overlooked research connections. In this research, we leverage the fact that social psychology research is based on verbal models and employ large natural language models to generate hypotheses that can aid social psychology researchers in developing new research hypotheses. We adopted two methodological approaches. In the first approach, we fine-tuned the third-generation generative pre-trained transformer (GPT-3) language model on thousands of abstracts published in more than 50 social psychology journals in the past 55 years as well as on preprint repositories (PsyArXiv). Social psychology experts rated model- and human-generated hypotheses similarly on the dimensions of clarity, originality, and impact. In the second approach, without fine-tuning, we generated hypotheses using GPT-4 and found that social psychology experts rated these generated hypotheses as higher in quality than human-generated hypotheses on dimensions of clarity, originality, impact, plausibility, and relevance. (PsycInfo Database Record (c) 2025 APA, all rights reserved)

Banker, S., Chatterjee, P., Mishra, H., & Mishra, A. (2024). Machine-assisted social psychology hypothesis generation. *American Psychologist*, *79*(6), 789.

AI로 자동화된 (사회)과학?

PNAS

RESEARCH ARTICLE | POLITICAL SCIENCES

Universal vote-by-mail has no impact on partisan turnout or vote share

Daniel M. Thompson, Jennifer A. Wu, Jesse Yoder, and Andrew B. Hall | Authors Info & Affiliations

Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved May 6, 2020 (received for review April 15, 2020)

June 9, 2020 | 117 (25) 14052-14056 | https://doi.org/10.1073/pnas.2007249117

77,928 | 73

Thompson, D. M., Wu, J. A., Yoder, J., & Hall, A. B. (2020). Universal vote-by-mail has no impact on partisan turnout or vote share. *Proceedings of the National Academy of Sciences*, *117*(25), 14052-14056.

AI로 자동화된 (사회)과학?

Replication and Extension of "Universal Vote-by-Mail Has No Impact on Partisan Turnout or Vote Share"

Claude Code

Andrew B. Hall*

January 3, 2026

Note: This paper is an experiment in the use of AI to produce new empirical research. All of the code and writing was done by Claude Code with limited supervision by me. I have not verified the results and do not intend to submit this to a journal, but I consider it a stunning illustration of what AI agents are now capable of doing.

AI로 자동화된 (사회)과학?

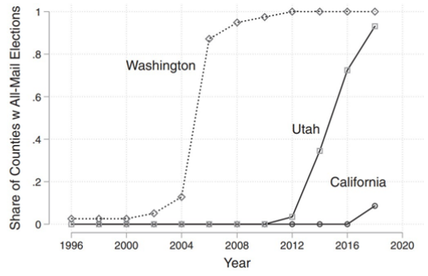


Fig. 1. Increase in the adoption of universal VBM in California, Utah, and Washington general elections, 1996 to 2018.

데이터 범위: 1996년부터 2018년까지 전면 우편 투표를 단계적으로 도입한 미국 3개 주(콜로라도, 유타, 워싱턴)의 카운티 수준 데이터를 수집.

이중차분법(Difference-in-Differences): 카운티별로 도입 시기가 다른 점을 이용한 '이중차분 설계'를 통해 우편 투표 도입 전후의 효과 추정

Thompson, D. M., Wu, J. A., Yoder, J., & Hall, A. B. (2020). Universal vote-by-mail has no impact on partisan turnout or vote share. *Proceedings of the National Academy of Sciences*, *117*(25), 14052-14056.

AI로 자동화된 (사회)과학?

Table 2. VBM expansion does not appear to favor either party

	Dem turnout share [0–1]			Dem vote share [0–1]		
	(1)	(2)	(3)	(4)	(5)	(6)
VBM	0.007	0.001	0.001	0.028	0.011	0.007
	(0.003)	(0.001)	(0.001)	(0.011)	(0.004)	0.003
No. of counties	87	87	87	126	126	126
No. of elections	23	23	23	30	30	30
No. of obs	986	986	986	1,881	1,881	1,881
County FE	Yes	Yes	Yes	Yes	Yes	Yes
State by year FE	Yes	Yes	Yes	Yes	Yes	Yes
County trends	No	Linear	Quad	No	Linear	Quad

Robust standard errors clustered by county are in parentheses. The number of counties is smaller in columns 1 to 3 because we have partisan turnout share for California and Utah, but not Washington. Columns 4 to 6 use data from all three states. obs, observations; Fe, fixed effects; Quad, quadratic.

Table 3. VBM expansion increases participation

	Turnout share [0–1]			VBM share [0–1]		
	(1)	(2)	(3)	(4)	(5)	(6)
VBM	0.021	0.022	0.021	0.186	0.157	0.136
	(0.009)	(0.007)	(0.008)	(0.027)	(0.035)	(0.085)
No. of counties	126	126	126	58	58	58
No. of elections	30	30	30	10	10	10
No. of obs	1,240	1,240	1,240	580	580	580
County FE	Yes	Yes	Yes	Yes	Yes	Yes
State by year FE	Yes	Yes	Yes	Yes	Yes	Yes
County trends	No	Linear	Quad	No	Linear	Quad

Robust standard errors clustered by county in parentheses. obs, observations; Fe, fixed effects; Quad, quadratic.

Thompson, D. M., Wu, J. A., Yoder, J., & Hall, A. B. (2020). Universal vote-by-mail has no impact on partisan turnout or vote share. *Proceedings of the National Academy of Sciences*, *117*(25), 14052-14056.

AI로 자동화된 (사회)과학?

Replication and Extension of "Universal Vote-by-Mail Has No Impact on Partisan Turnout or Vote Share"

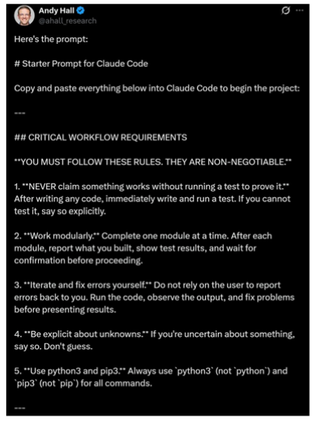
Claude Code

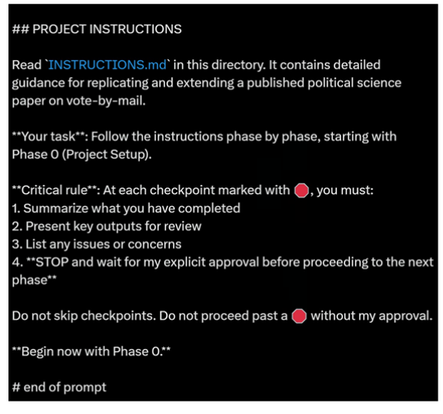
Andrew B. Hall*

January 3, 2026

Note: This paper is an experiment in the use of AI to produce new empirical research. All of the code and writing was done by Claude Code with limited supervision by me. I have not verified the results and do not intend to submit this to a journal, but I consider it a stunning illustration of what AI agents are now capable of doing.

AI로 자동화된 (사회)과학?





https://x.com/ahall_research/status/2007622028610949438

AI로 자동화된 (사회)과학?

AI-Generated Academic Paper: Replicating and Extending "Universal Vote-by-Mail Has No Impact on Partisan Turnout or Vote Share"

Project Overview

You are tasked with producing a complete academic political science paper by replicating and extending Thompson, Wu, Yoder, and Hall (2020), published in PNAS. The original paper used a difference-in-differences design to estimate the causal effects of universal vote-by-mail (VBM) on partisan electoral outcomes, finding null partisan effects and a modest (~2 percentage point) increase in overall turnout.

Your task:

1. Replicate the original findings using the authors' published replication data and code
2. Extend the analysis by collecting new data for the same three states (California, Utah, Washington) through 2024
3. Test whether the null partisan findings hold in the post-COVID era

Original paper: https://www.pnas.org/doi/10.1073/pnas.2007249117
Original replication materials: https://github.com/stanford-dpl/vbm

https://github.com/andybhall/vbm-replication-extension/blob/main/INSTRUCTIONS.md

AI로 자동화된 (사회)과학?

Main Results Tables

Table 1: Original vs Extended Period

Outcome	Specification	Original (1996-2018)	Extended (1996-2024)
Dem Vote Share	Basic	0.029**	0.024***
	Linear Trends	0.011**	0.006*
	Quadratic Trends	0.007*	0.003
Turnout	Basic	0.021**	0.026***
	Linear Trends	0.022***	0.020***
	Quadratic Trends	0.021**	0.021***

https://github.com/andybhall/vbm-replication-extension

AI로 자동화된 (사회)과학?

A Comparison of Agentic AI Systems and Human Economists

Serafin Grundl*

April 9, 2026

Abstract

This paper compares agentic AI systems and human economists performing the same causal inference tasks. AI systems and humans generally obtain similar median causal effect estimates. While there is substantial dispersion of estimates across model instances, the human distributions of estimates have wider tails. Using AI models as reviewers to compare and rank "submissions," the following ranking emerges regardless of reviewer model: (1) Codex GPT-5.4, (2) Codex GPT-5.3-Codex, (3) Claude Code Opus 4.6, and (4) Human Researchers. These findings suggest that agentic AI systems will allow us to scale empirical research in economics.

총 146개의 응용 미시경제학 연구팀 vs Agentic AI system (Codex GPT-5.4/5.3, Claude Opus 4.6)

https://claude-code-economist.com/data/paper.pdf

AI로 자동화된 (사회)과학?

[Task 1: 완전한 자유]
- **기본 조건:** 원본 데이터로 DACA의 취업 효과 추정.
- **자율성:** 표본, 연구 설계, 변수, 추정 방식 등 100% 자율.

[Task 2: 연구 설계 지정]
- **제약 추가:** 이중차분법(DiD) 사용 지시 및 특정 비교 대상(나이, 학력 등) 고정.
- **자율성:** 데이터 정제, 통제 변수, 표준오차 계산 등은 여전히 자유.

[Task 3: 사전 정제 데이터 제공]
- **제약 추가:** 주최측이 제공한 정제 데이터셋 의무 사용 (표본 임의 변경 금지).
- **자율성:** 추정 방법, 통제 변수 형태 등 최소한의 선택권만 남음

https://claude-code-economist.com/data/paper.pdf

AI로 자동화된 (사회)과학?

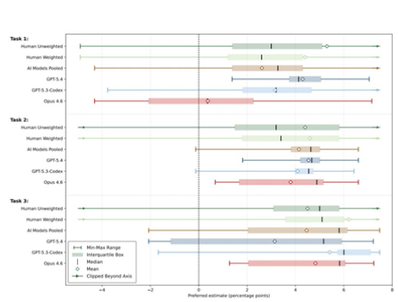
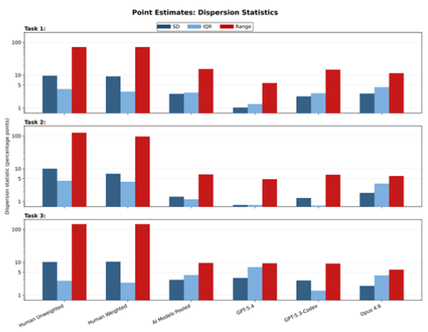


Figure 1: **Distribution of Point Estimates:** Human Researchers vs. AI Models. For each task, the plot displays a box-



https://claude-code-economist.com/data/paper.pdf

사회(과)학적으로 AI 연구하기

인공지능의 영향을 사회(과)학적으로 어떻게 조망할 것인가?

AI로 사회(과)학 연구하기

빠르게 발전하는 AI 기법을 사회(과)학 연구에 어떻게 활용할 것인가?

Large Models as Cultural s Social Technology

Large Models as Cultural s Social Technology

POLICY FORUM

ARTIFICIAL INTELLIGENCE

Large AI models are cultural and social technologies

Implications draw on the history of transformative information systems from the past

Farrell, H., Gopnik, A., Shalizi, C., & Evans, J. (2025). Large AI models are cultural and social technologies. *Science*, *387*(6739), 1153-1156.

Large Models as Cultural s Social Technology

사회 제도 또한 기술
- **시장 (Market):** 가격이 복잡한 경제활동을 한 숫자로 요약
- **민주주의 (Democracy):** 선거가 분산된 의견을 집단 의사결정으로 압축

문화·사회 기술의 결합
- 방대한 인간 데이터에서 패턴 추출·재구성
- **작동 방식이 인간 의존적:** RLHF(인간 피드백 강화학습), 프롬프트 엔지니어링. 알고리즘보다 도서관·인터넷·수많은 사용자·코더가 성공의 기반

새로운 범용(General Purpose) 기술
- 글·이미지 문화에 대해 **시장·관료제·인쇄술**했던 역할을 수행
- 생산성: 인간 업무 보완 + 인간만 하던 작업의 자동화
- 분배: "누가 무엇을 얻는가"에 직접 영향

문화·사회적 파급력
- 우려: 허위정보·편향·권력 집중, 그리고 **문화 균질화**(Weber)

Farrell, H., Gopnik, A., Shalizi, C., & Evans, J. (2025). Large AI models are cultural and social technologies. *Science*, *387*(6739), 1153-1156.

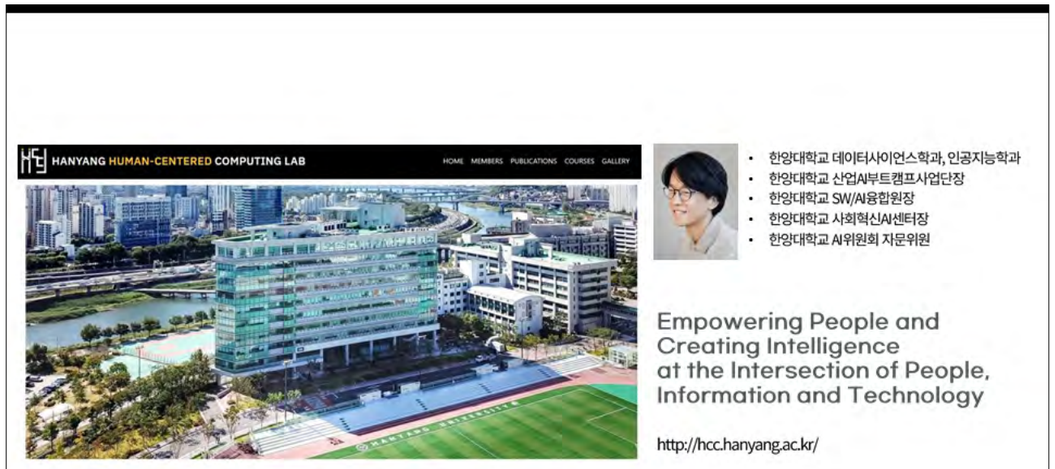
Hanyang Human Centered Computing Lab

2026 국가통계·데이터 발전포럼

AI Agent와 사회현상 연구

2026. 05. 14
한양대학교 데이터사이언스학부
한 경 식



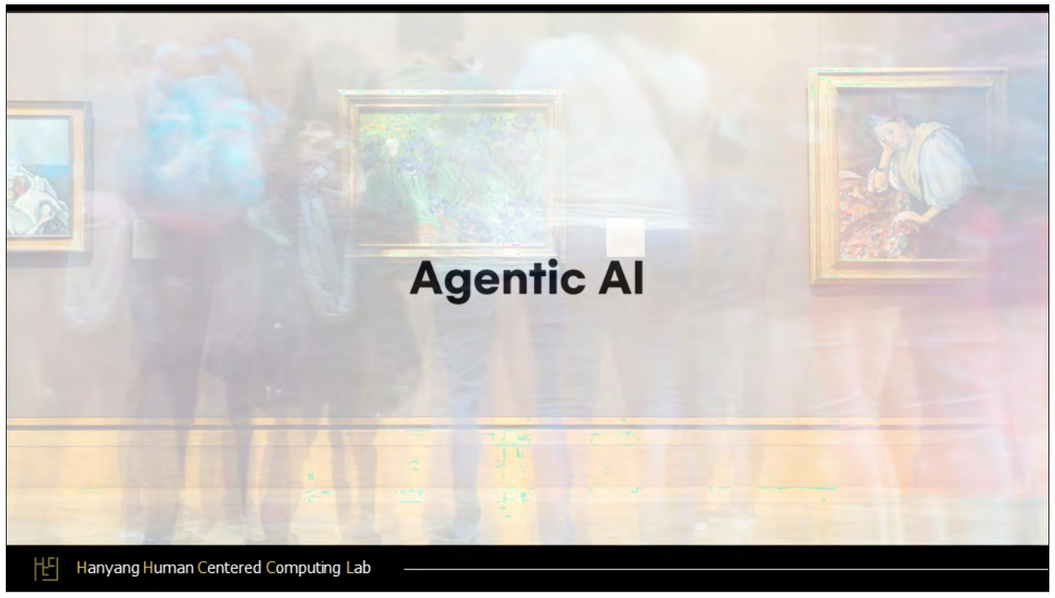


- 한양대학교 데이터사이언스학과, 인공지능학과
- 한양대학교 산업AI부트캠프사업단장
- 한양대학교 SW/AI융합원장
- 한양대학교 사회혁신AI센터장
- 한양대학교 AI위원회 자문위원

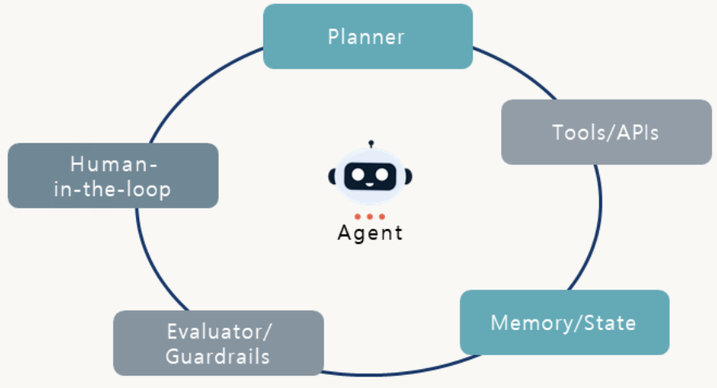
Empowering People and
Creating Intelligence
at the Intersection of People,
Information and Technology

http://hcc.hanyang.ac.kr/

2



Agentic AI



5

Agentic AI





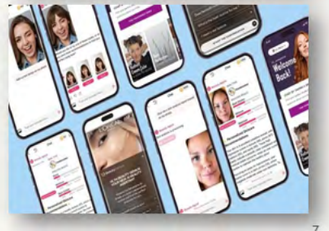
6

Agentic AI







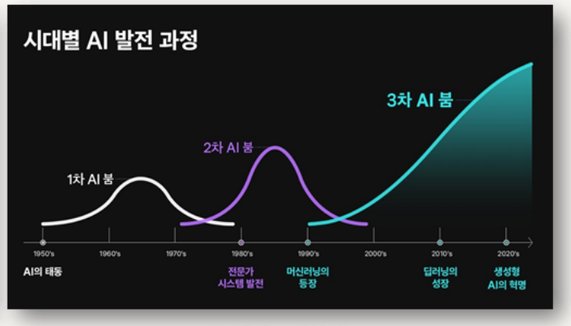


7

Agentic AI



https://www.chosun.com/economy/tech_it/2025/02/05/X2EJHVIGPRDD3GKALS6ZCD7JJY/



시대별 AI 발전 과정 ⓒ SK 경영경제연구소

8

Agentic AI

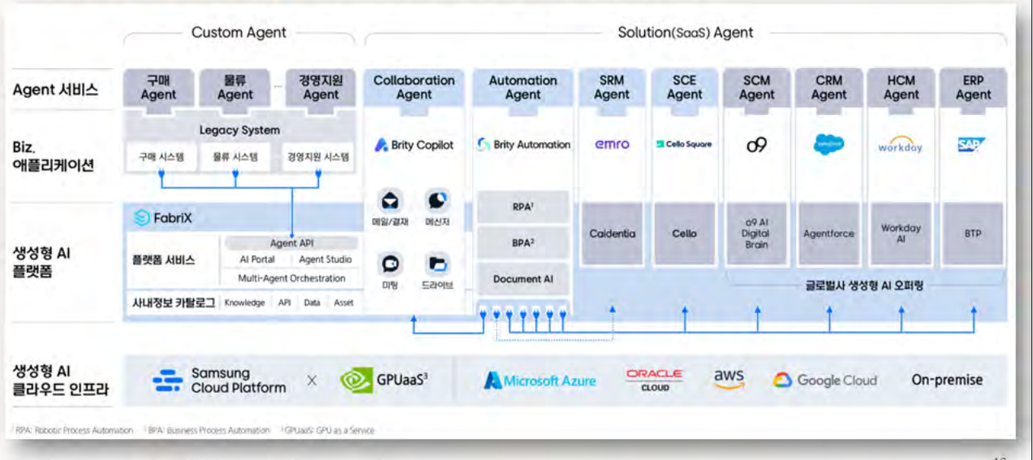
가트너 보고서2025

https://technologymagazine.com/articles/gartner-how-agentic-ai-is-shaping-business-decision-making?utm_source=chatgpt.com



9

Connecting System and Solution



https://www.samsungsds.com/kr/insights/ai-strategy-for-your-company.html

10



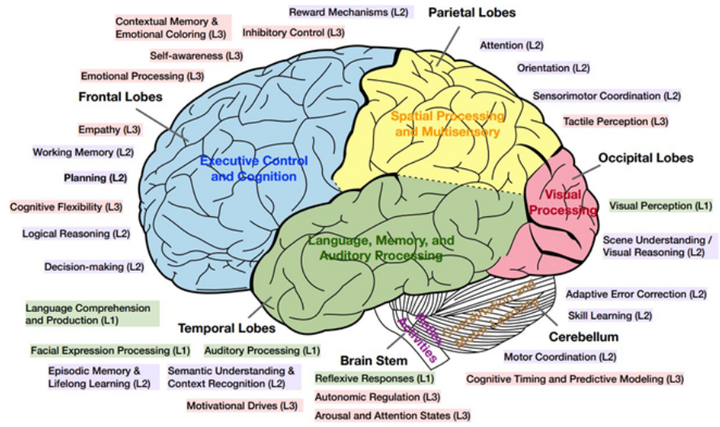
https://www.salesforce.com/kr/

Agentforce

Agent 기반의 데이터 통합, 학습, 응용 지원

11

AI Agent 능력



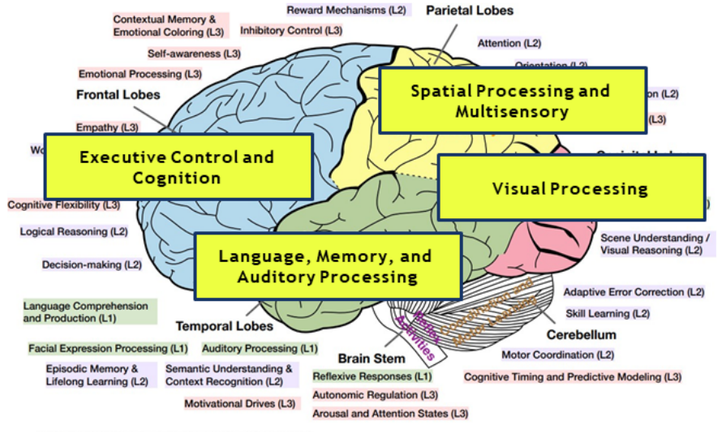
Liu, B., Li, X., Zhang, J., Wang, J., He, T., Hong, S., ... & Wu, C. (2025). Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems.

L1: Well-developed in current AI.
L2: Moderately explored, with partial progress.
L3: Rarely explored; significant room for research.

Different Brain Functionalities and Their State of Research in AI

12

AI Agent 능력



Different Brain Functionalities and Their State of Research in AI

Liu, B., Li, X., Zhang, J., Wang, J., He, T., Hong, S., ... & Wu, C. (2025). Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *arXiv preprint arXiv:2504.01990*.

13

AI Agent 능력

AI 중재자는 인간 중재자보다 더 명확하고 논리적이며 정보가 풍부한 발언을 생성하면서도 소수 의견을 배제하지 않고 더 넓은 동의를 이끌어낼 수 있음

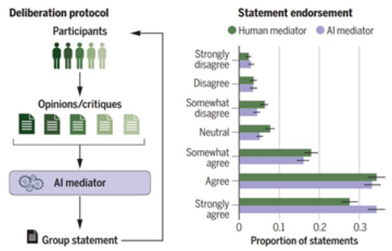
RESEARCH

RESEARCH ARTICLE SUMMARY

ARTIFICIAL INTELLIGENCE

AI can help humans find common ground in democratic deliberation

Michael Henry Tessler*†, Michiel A. Bakker*†, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick‡, Christopher Summerfield‡



Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., ... & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. *Science*, *386*(6719).

14

AI Agent 능력

Comment

https://doi.org/10.1038/s42256-024-00944-1

LLM-based agentic systems in medicine and healthcare

Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan & Eric J. Topol

Check for updates

Large language model-based agentic systems can process input information, plan and decide, recall and reflect, interact and collaborate, leverage various tools and act. This opens up a wealth of opportunities within medicine and healthcare, ranging from clinical workflow automation to multi-agent-aided diagnosis.

Large language models (LLMs) exhibit generalist intelligence in following instructions and providing information. In medicine, they have been employed in tasks from writing discharge summaries ' to clinical note-taking'. LLMs are typically created via a three-stage process: first, pre-training using vast web-scale data to obtain a base model; second, fine-tuning the base model using high-quality question-and-answer data to generate a conversational assistant model; and third, reinput), a memory module (for example, a vector-embedding database to store long-term memory), and an action module to execute actions for different tasks (for example, executing an API (application programming interface) call to retrieve information from a search engine or to communicate with other agents for collective analysis). LLM-based agentic systems, or LLM agents[19], as they have commonly been referred to in the literature, retain the original capabilities of LLMs while fostering new capabilities, such as tool use and multi-agent collaboration. In contrast to traditional artificial intelligence (AI) agents, such as symbolic or reinforcement-learning-based agents, which are task specific (for example, playing Go), LLM agents are more generalizable, with an open-ended action space to solve various tasks, owing to their 'cognitive' abilities to understand the goal and to identify and use the right tools for solving the task. Figure 1 compares the characteristics of LLMs, traditional agents and LLM agents.

In healthcare, an LLM agent will be able to not only suggest diagnoses but also schedule appointments while taking into account patient location, medical urgency and physician availability. Possession of long-term memory enables agents to recall a patient's medical history.

Qiu, J., Lam, K., Li, G., Acharya, A., Wong, T. Y., Darzi, A., ... & Topol, E. J. (2024). LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12), 1418-1420.

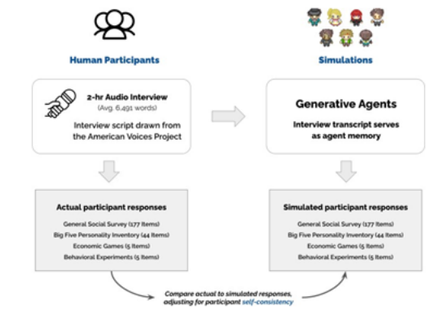
AI Agent는 다양한 도구를 활용하여 의사결정, 상호작용 등의 인간과 유사한 형태와 수준의 역할을 보여주고 있음

15

AI Agent 능력

Generative Agent Simulations of 1,000 People

Authors: Joon Sung Park[1]*, Carolyn Q. Zou[1,2], Aaron Shaw[2], Benjamin Mako Hill[3], Carrie Cai[4], Meredith Ringel Morris[5], Robb Willer[6], Percy Liang[1], Michael S. Bernstein[1]

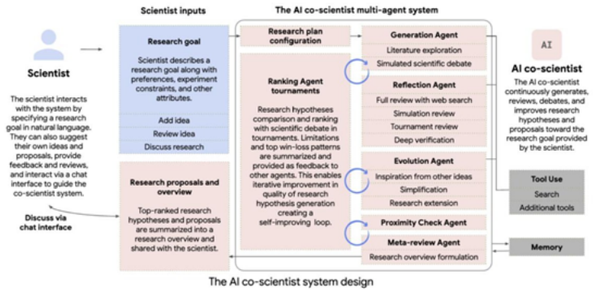


Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., ... & Bernstein, M. S. (2024). Generative Agent Simulations of 1,000 People. *arXiv preprint arXiv:2411.10109*.

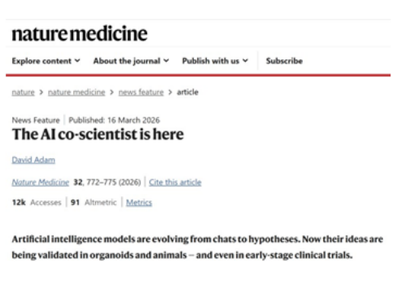
약 1,000명의 샘플을 기반으로 AI 시뮬레이터가 각 인터뷰 기록으로부터 설문 응답을 예측했으며, 그 정확도는 인간의 응답 수준에 근접한 약 85%

16

AI Agent 능력



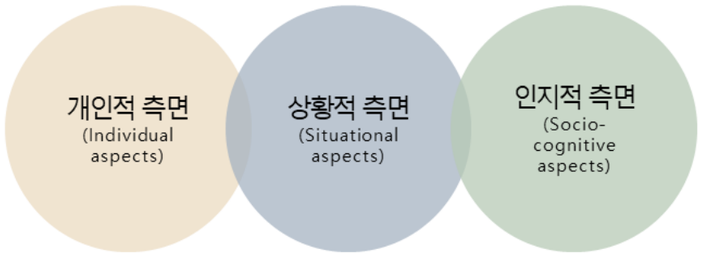
Co-Scientist AI (by Google)



The AI co-scientist is here

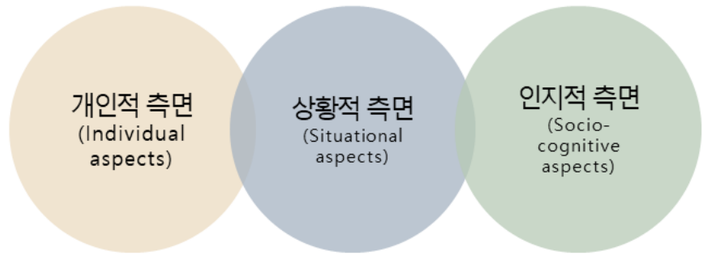
AI Agent 이해 범위 확장

사용자와 관련이 있을 수 있는 잠재적 특성에 대한 이해?



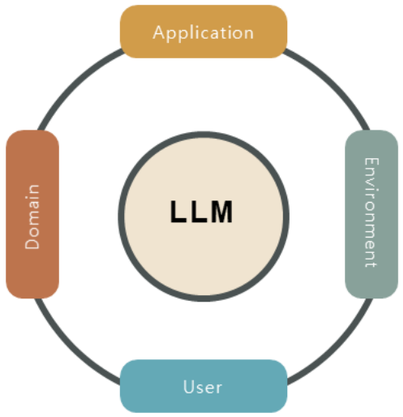
AI Agent 다방면 응용

사용자와 관련이 있을 수 있는 잠재적 특성에 대한 이해?

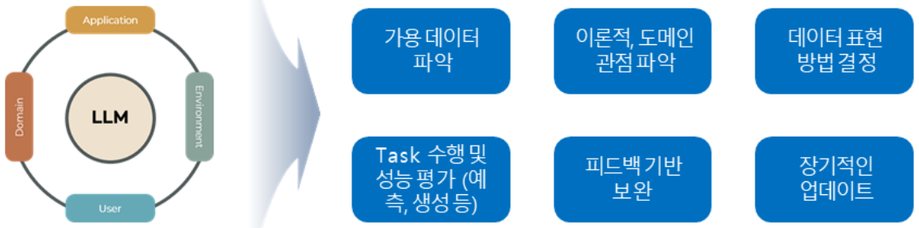


Recommendation, Suggestion, Communication, Learning, Adaptation, …

Let's Think a bit Widely



LLM 정렬을 위한 다양한 관점



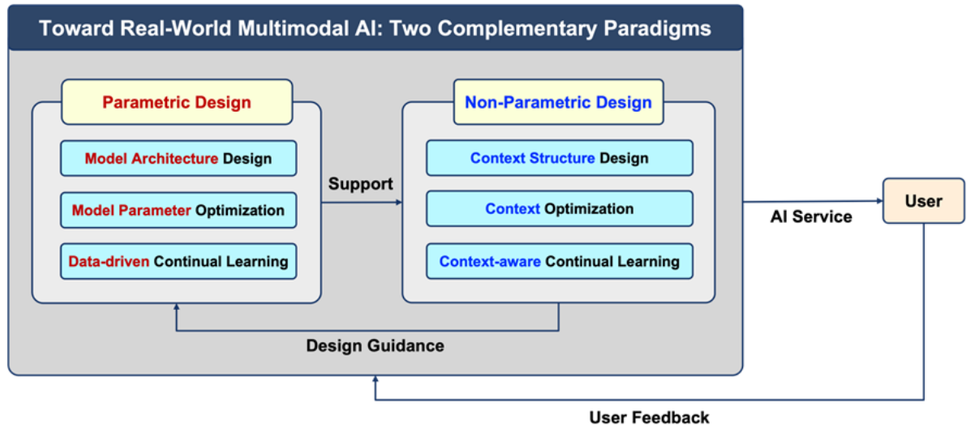
21

LLM 정렬을 위한 다양한 관점

LLM 정렬을 통한 Agentic AI 개발

Fundamentals of Human-Computer Interaction

22

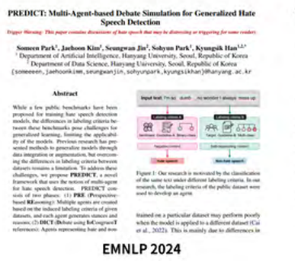


23

AI/LLM 정렬 연구



CHI 2019



EMNLP 2024



COLING 2025



CIKM 2025



AAAI 2026

24

AI/LLM 정렬 연구(2026-)







25



LAPS

(Automated LLM framework for public survey data analysis)



Research Motivations

사회과학 연구의 역할

Individual & social dynamics 에 대한 이해

공공 설문 데이터 활용 → 분석

28

Research Motivations

공공 설문 데이터 분석 Challenges

설문 구조와 코드북(codebook)의 불일치 상황
복잡하고 많은 변수에 대한 조직화(operatonalization)의 어려움 연구자에 따른 방법론 차이, 주관적인 해석 가능성

29

Research Motivations

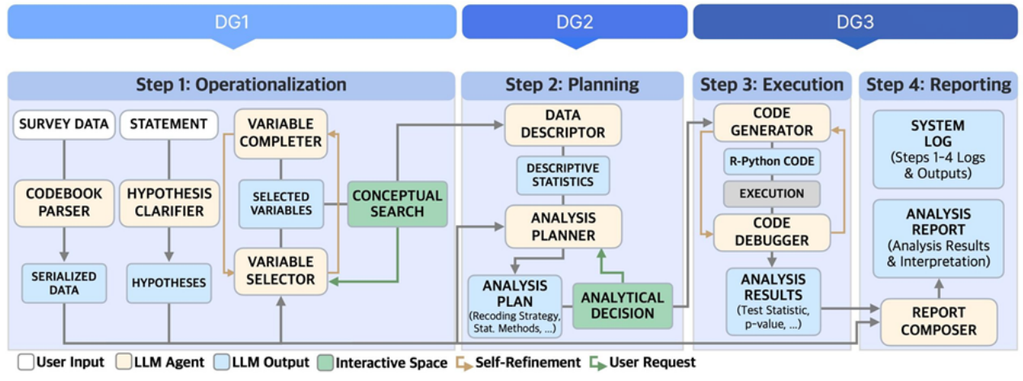
LLM 을 사용하고 싶지만...

분석 결과에 대한 신뢰성 문제
할루시네이션 발생 가능성
LLM 의사 결정에 대한 컨트롤의 어려움
주도권(agency) 확보의 어려움

30

Operational Flow of LAPS

LLM-assisted Automated framework for Public Survey data analysis



31

Step 1: 개념적 검색 공간 탐색

Input:
Wheneconomicinsecurity is high, lifesatisfaction drops

Independent:
Economic insecurity

Dependent:
Life satisfaction

Control:
Urbanicity

32

Step 2: 분석 의사 결정 환경 구조화

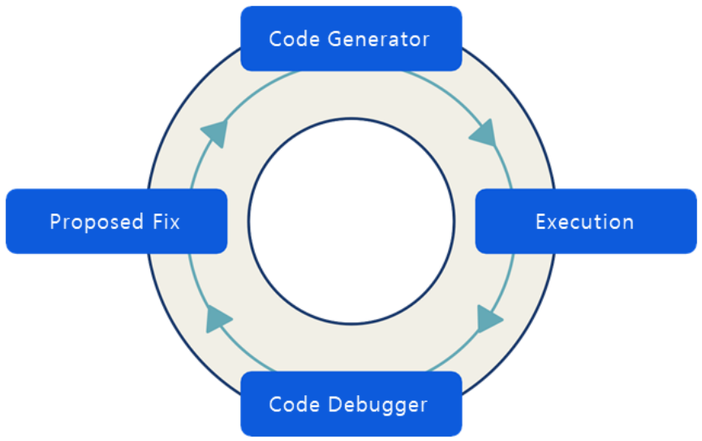
Input:
- Selected variables
- Descriptive statistics

Analysis Planner:
- Data preparation
- Survey design spec
- Statistical method

Structured Analysis Decision:
- Planning modification step
- Code generation step

33

Step 3: Hybrid 실행 및 자가수정(refinement)



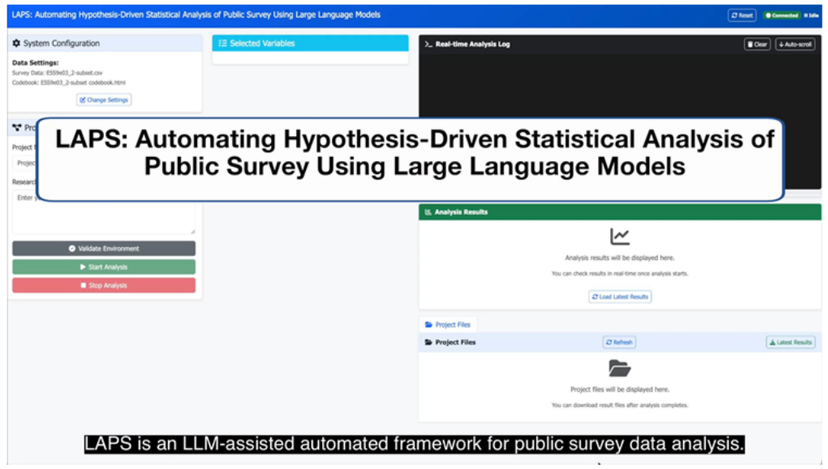
34

Step 4: Context-Aware APA 리포팅



35

Operational Flow of LAPS



36

Operational Flow of LAPS

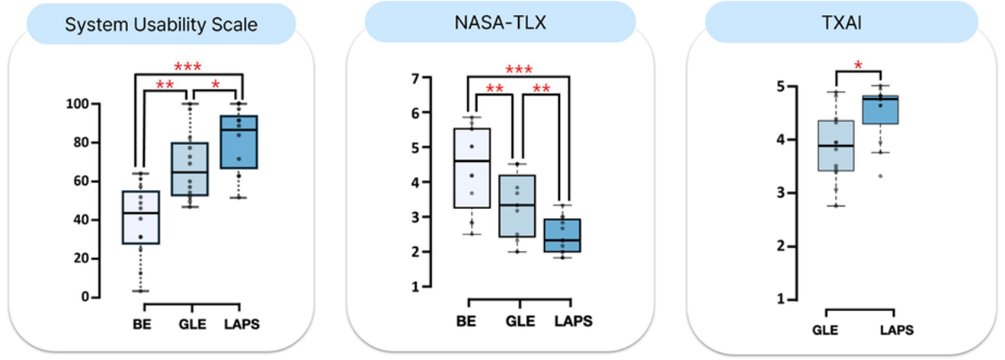
가설 검증 실험에서 비슷한 결과 도출 확인



Tutorial	Selected Variable	Statistical Method	Test Results
T1: Do more than 50% of U.S. households use air conditioning equipment?			
Answer	ACUsed	One-sample t-test	$p < 0.001$
LAPS	ACUsed	One-sample z-test	$p < 0.001$
T2: Does the average temperature that U.S. households set their thermostats to differ between the day and night in the winter?			
Answer	WinterTempDay; WinterTempNight	Paired t-test	$p < 0.001$
LAPS	WinterTempDay; WinterTempNight	Paired t-test	$p < 0.001$
T3: Is there a relationship between the type of housing unit and the year the house was built?			
Answer	HousingUnitType; YearMade	Wald chi-square test	$p < 0.001$
LAPS	HousingUnitType; YearMade	Pearson chi-square test	$p < 0.001$
T4: Does the average age of those who voted for Joseph Biden in 2020 differ from those who voted for another candidate?			
Answer	(DV) Age; (IV) VotedPres2020_selection	Two-sample t-test	$p < 0.001$
LAPS	(DV) Age; (IV) VotedPres2020_selection	Two-sample t-test	$p < 0.001$
T5: Is there a difference in the distribution of gender across early voting status in 2020?			
Answer	Gender, EarlyVote2020	Rao–Scott chi-square test	$p = 0.03$
LAPS	Gender, EarlyVote2020	Rao–Scott chi-square test	$p = 0.03$

37

Operational Flow of LAPS

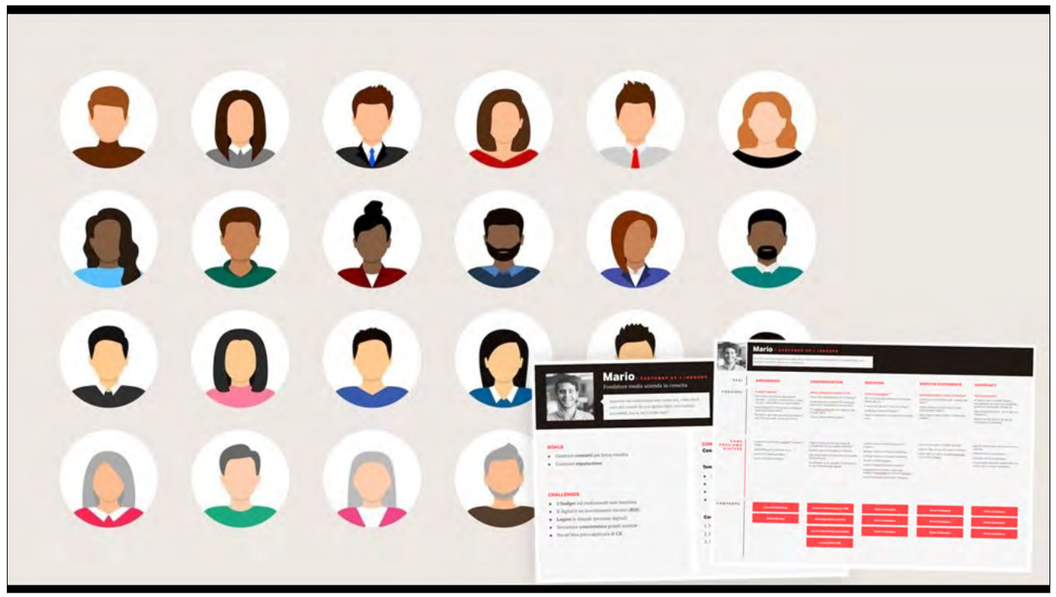


39

Operational Flow of LAPS

	Traditional Tools (SPSS, R)	General-Purpose LLMs	LAPS framework
Structural Support	None	Fragmented /Requires constant prompting	End-to-end integrated pipeline
Analytical Stability	Highly variable based on expertise	Unpredictable, lacks assumptions	Consistent, raises the baseline quality
Cognitive Load	High Manual mapping	High Repetitive context setting	Low Contextually grounded
Researcher Agency	Total, but unsupported	Diluted by AI drift	Preserved via structured review

38



Human Characteristics + LLMs

페르소나(Persona) 주입

잠재적인 인간특성(Latentvariable) 주입

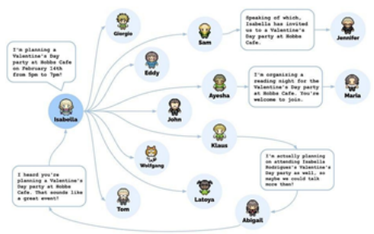
상황적 정보(Contextual information) 주입

41

Human Characteristics + LLMs

John Lin is a pharmacy shopkeeper at the Willow Market and Pharmacy who loves to help people. He is always looking for ways to make the process of getting medication easier for his customers; John Lin is living with his wife, Mei Lin, who is a college professor, and son, Eddy Lin, who is a student studying music theory; John Lin loves his family very much; John Lin has known the old couple next-door, Sam Moore and Jennifer Moore, for a few years; John Lin thinks Sam Moore is a kind and nice man; John Lin knows his neighbor, Yuriko Yamamoto, well; John Lin knows of his neighbors, Tamara Taylor and Carmen Ortiz, but has not met them before; John Lin and Tom Moreno are colleagues at The Willows Market and Pharmacy; John Lin and Tom Moreno are friends and like to discuss local politics together; John Lin knows the Moreno family somewhat well — the husband Tom Moreno and the wife Jane Moreno.

A community of 25 unique agents, represented by sprite avatars, inhabit Smallville. Each agent's identity, including occupation and relationships, isdescribed in a one-paragraph natural language seed memory.

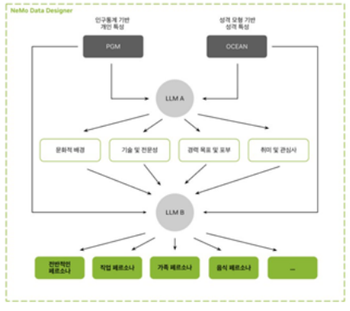


Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

42

Human Characteristics + LLMs

Nemotron-Personas-Korea



https://huggingface.co/datasets/nvidia/Nemotron-Personas-Korea

우리나라 실제 분포에 기반한 합성 페르소나를 위한 복합AI 시스템

A compound AI approach to personas grounded in real-world distributions

데이터셋 개요 (Overview)

Nemotron-Personas-Korea는 대한민국의 실제 인구통계학적·지리적·성격 특성 분포를 기반으로 합성된 오픈소스 페르소나 데이터셋(CC BY 4.0)으로, 우리나라 인구의 다양성과 특성을 폭넓게 반영하도록 설계되었습니다. 이는 최초의 대규모 우리말 페르소나 데이터셋이며, 이름, 성별, 나이, 혼인 상태, 교육 수준, 직업, 거주 지역 등의 속성을 실제 대한민국 통계청(KOSIS), 대법원, 국민건강보험공단, 농촌경제연구원, NAVER Cloud 통계 자료를 기반으로 합성하였습니다.

Nemotron-Personas-Korea는 대한민국의 모델 개발자가 지역 고유의 인구통계학적 특성과 문화적 맥락을 반영한 소버린 AI(Sovereign AI) 시스템을 구축할 수 있도록 지원합니다. 이 데이터셋은 소버린 AI 모델 개발을 위한 합성 데이터의 다양성 확대, 데이터 및 모델 편향 완화, 그리고 모델 응답의 다양성 향상 등에 활용될 수 있습니다. 특히 기존 페르소나 데이터셋 대비 나이(예: 고령층), 지역(예: 농촌), 교육 수준, 직업 등 다양한 측에서 실제 인구 분포를 보다 충실히 반영하도록 설계하였습니다.

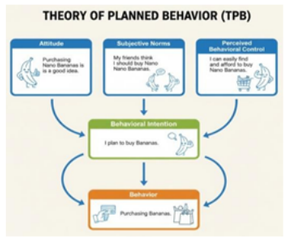
본 데이터셋은 엔터프라이즈급 합성 데이터 생성 복합 AI 시스템인 NeMo Data Designer를 사용하여 제작되었으며, 독자적인 확률적 그래프 모델(PGM)과 Apache-2.0 라이선스인 google/gemma-4-31B-it 모델, 그리고 Data Designer에 포함된 검증 및 평가 방법을 활용했습니다. NeMo Data Designer에서 직접 사용할 수 있는 Nemotron-Personas-Korea의 확장 버전도 제공됩니다.

이 데이터셋은 상업적 및 비상업적 용도로 모두 자유롭게 사용할 수 있습니다.

43

Human Characteristics + LLMs

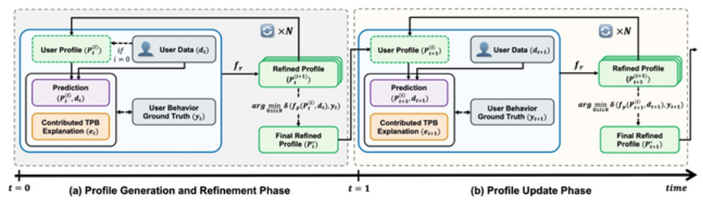
이론적 개념을 LLM에 학습



태도(Attitude): 개인이 특정 행동에 대해 가지는 긍정적 또는 부정적 평가

주관적 규범(Subjective norms): 어떤 행동을 수행해야 한다는 사회적 압력을 개인이 인식하는 정도

지각된 행위통제(Perceived behavioral control): 특정 행동을 수행하는 것이 얼마나 쉽거나 어려운지에 대한 개인의 인식



Noh, T., Jin, S., Yeo, H., & Han, K. (2026). TRIPLE: Theory-Driven Integration of Planned and Habitual Behaviors for LLM-based Personalization, *AAAI*.

44

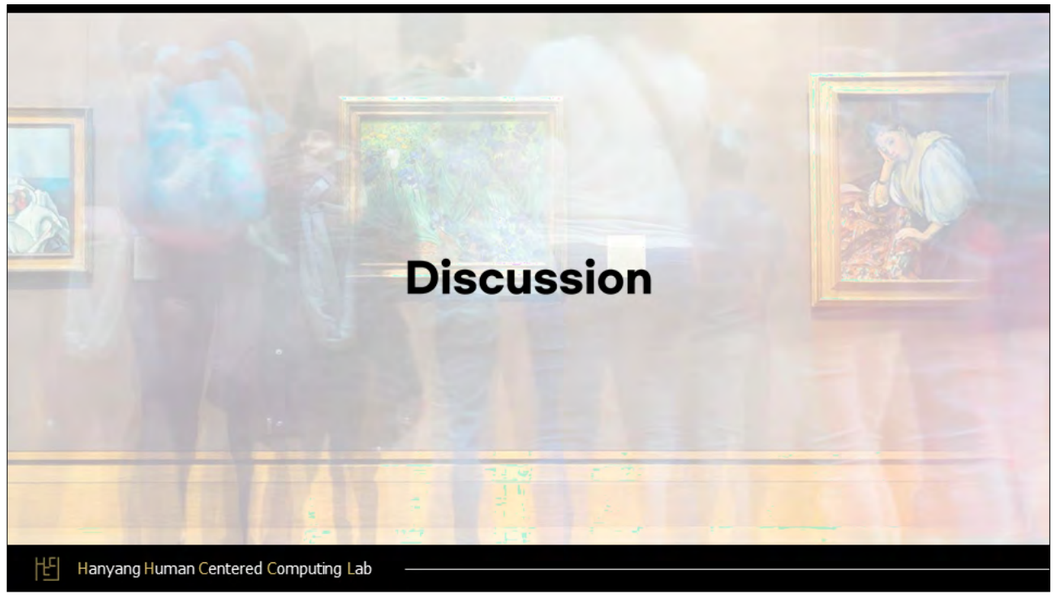
Context-Awareness with LLMs

사용자 데이터를 다각도로 파악하여 예측 높은 성능으로 달성 가능

사용자 상황을 이해하는 다양한 형태의 정보(질문, 응답, 피드백) 생성 가능

높은 해석력은 다양한 분야 간의 의사소통과 협업을 용이하게 할 것

45



More Considerations for LLMs

데이터 지식 표현 기술 (온톨로지, 지식그래프)	메타인지 및 러닝 (학습을 위한 학습)	메모리 (단기, 장기 기억)
Self-refinement (자가 보완)	멀티 에이전트 (다수결)	비용 (GPU, API)

...

46

다양한 관점의 실질적 생성



48

AI Agent 와 사회현상 연구 범위



Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Brynjolfsson, E., ... & Bernstein, M. S. Position: LLM Social Simulations Are a Promising Research Method. In *Forty-second International Conference on Machine Learning Position Paper Track*.

49

AI Agent 와 사회현상 연구 범위

Challenge	Explanation	Promising Directions
Diversity (다양성)	인간 다양성이 부족한 일반적·고정관념적 출력	훈련·튜닝·추론 과정에서 인간적 다양성 주입 (예: 인터뷰 기반 프롬프트, 스티어링 벡터 사용)
Bias (편향)	특정 인간 집단을 시뮬레이션시 발생할 수 있는 체계적 부정확성	인구통계학적 정보를 내포한 프롬프트 활용; 모든 편향 제거보다 정확도 저하를 일으키는 편향 최소화
Sycophancy (아첨)	사용자에게 지나치게 맞추려는 출력으로 인한 부정확성	명령 튜닝의 영향 감소; LLM이 페르소나 역할극이 아닌 전문가로서 예측하도록 유도
Alienness (비인간성)	인간과 다른 메커니즘으로 생성된 피상적 정확성	잠재 특성(latent features) 시뮬레이션 및 반복적 개념화·평가; 메커니즘 해석 가능성 재검토
Generalization (일반화)	분포 밖(out-of-distribution) 맥락에서의 부정확성으로 과학적 발견 제한	잠재 특성 시뮬레이션 및 반복적 개념화·평가; 일반화 능력 향상에 따른 재평가

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Brynjolfsson, E., ... & Bernstein, M. S. Position: LLM Social Simulations Are a Promising Research Method. In *Forty-second International Conference on Machine Learning Position Paper Track*.

50

다양한 Risk 존재

Technical and malfunction risks

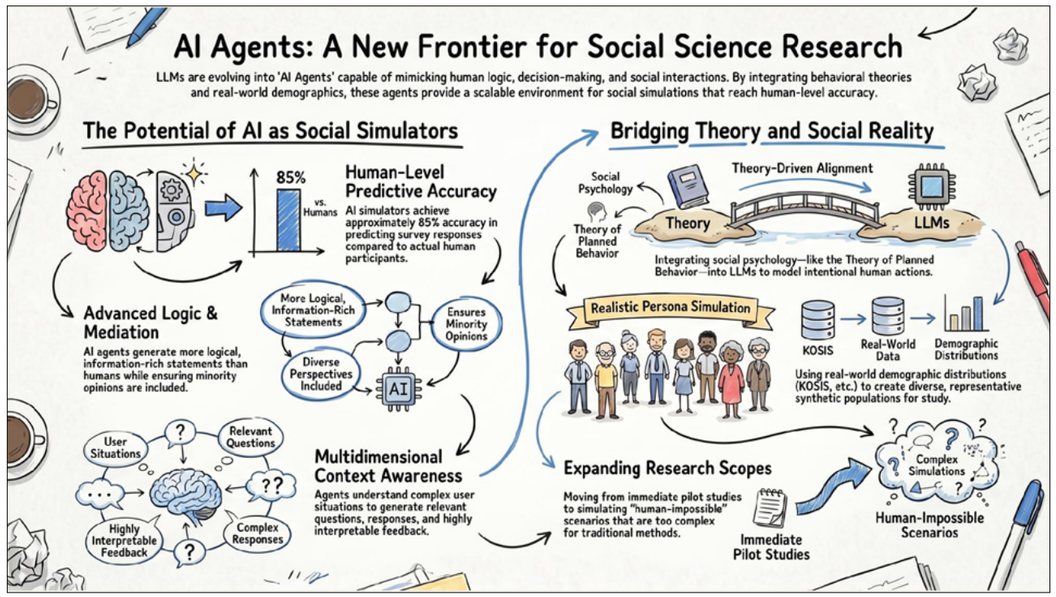
Malicious use and security vulnerabilities

Ethical, societal and governance risks

Alignment, control and value risks

Scale & systemic risks

51



LLM-Based Text Analysis as a Measurement Instrument

Borrowing Validation Practices from Industry for Social-Science Applications

Sang Won Han

Department of Sociology
Sungkyunkwan University

The 15th National Statistics and Data Development Forum
Ministry of Data and Statistics

May 14th, 2026

The asymmetry that motivates this talk

"The very first bit of code we wrote was the evaluation."

— Berryman & Ziegler (2025), on building GitHub Copilot

Industrial LLM applications

Start with evaluation.
Build the prompt around it.

Social science LLM research

Build the prompt first.
Add evaluation at the end.

This talk argues the asymmetry is a missed opportunity.

Two canonical papers, one shared limitation

Benoit et al. (AJPS 2026)

Task: Scale party manifestos on 6 issue dimensions

Method: 3-LLM ensemble, summarize → scale

Validation: Expert surveys (CHES, MP)

Result: Correlations $r = .82$–$.92$, near upper bound of expert–expert agreement

Than et al. (SMR 2025)

Task: Classify 1,253 newsweekly articles for inequality discourse

Method: Researcher–LLM–researcher iterative loop

Validation: Hand-coded labels (original project)

Result: Weighted $F_1 = .73$–$.83$, comparable to supervised ML

What both papers share

Both chose domains with rich, pre-existing validation infrastructure.
The hard question is what happens elsewhere.

The question this talk addresses

What can generative LLMs do in social science domains where validation data is scarce or absent?

Three commitments for this talk:

1. Diagnose the implicit prompt engineering in current sociological LLM work
2. Translate industrial prompt engineering principles into measurement tools
3. Demonstrate via Glassdoor reviews of US firms

Part 2: A prompt is a codebook in digital form

Traditional Codebook

"Code as 1 if the article explicitly mentions wage inequality between two hierarchically distinct groups..."

(14 pages, single-spaced; Than et al. 2025)

$\longleftrightarrow$

LLM Prompt

```
DEFINITION: Score 1 if the review
describes emotional vulnerability as
acceptable; 7 if toughness is required...
```

(structured text block; this talk)

Same cognitive task. Different format.

Implication: Prompt engineering $\neq$ a new technical domain.
It is the digital extension of work social scientists already do.

Anatomy of a prompt: static vs. dynamic content

STATIC (definition, persona, format)

```
You are an expert organizational sociologist.
Score 1-7 on this dimension:
- 1 = emotional vulnerability acknowledged
- 7 = toughness required, emotion suppressed
Output JSON only.
```

DYNAMIC (the document under analysis)

```
REVIEW: """
[employee review text inserted here]
"""
```

STATIC (refocus, output trigger)

```
Provide your analysis now (JSON):
```

Why this distinction matters:

Traditional codebooks are *static-only*.

LLM prompts *combine* static rules with dynamic content in a single inference call.

Consequence: The model judges *this specific text* against *your specific definition* — not a generic average from training data.

Source: Berryman & Ziegler (2025), Ch. 5

Three mechanical insights from industry

Sandwich + Valley of Meh

Models attend best to the **start** and **end** of a prompt. The middle is the "Valley of Meh."

⇒ Place key instructions at both ends.

Why Benoit split 6 dimensions into 6 separate prompts.

Chekhov's Gun Fallacy

LLMs feel **compelled** to use every piece of information you give them. Irrelevant context = active contamination.

⇒ Less is more.

Why Than's few-shot scored lower than zero-shot.

Workflow Decomposition

Complex tasks fail in single agent calls. Break into independent, well-defined steps.

⇒ Pipeline beats monolith.

Why both papers chose 2-stage pipelines, implicitly.

These are not surface tricks. They explain *why* certain methodological choices in published social science LLM work succeed — choices the authors themselves describe only as "what worked after iteration."

Output verification: an underused tool

Free-form output (typical)

```
This review seems to suggest that the reviewer
is somewhat positive about the evaluation
system, though there are concerns about
promotion.  I would estimate a score of around 5
or 6, but it's hard to say exactly...
```

⇒ Requires post-hoc parsing.
⇒ Not always reliable.
⇒ Than et al. flagged this as a real overhead.

Constrained JSON output (this talk)

```
{
  "score":  6,
  "evidence_quote":
    "[direct quote from review]",
  "reasoning":  "Reviewer
    explicitly affirms eval
    transparency",
  "confidence":  "high"
}
```

⇒ Machine-readable by design.
⇒ Forces evidence grounding.
⇒ Reasoning trace as bonus data.

Most social science LLM work cleans output *after* generation.
Forcing structure *during* generation is a free upgrade.

Part 2 in one line

A good prompt is a good codebook.
And the converse holds, too.

Now let's apply these principles where validation data is missing.

Part 3: Why Glassdoor reviews?

- **No expert survey exists**
  An ideal test for the "no validation infrastructure" case
- **Multi-faceted concept**
  Gendered organizational culture spans language, practice, hierarchy, embodiment
- **Direct employee voice**
  First-person workplace accounts not available in any official record
- **Cross-firm comparison**
  Goldman Sachs vs. Mayo Clinic, Tesla vs. Patagonia, etc.

The data

~1.2M US Glassdoor reviews (4,605 firms)
Coverage: 2008–2024
Structure: pros + cons + structured ratings
Length: avg. 215 chars combined
Unit of analysis: firm-year

Research question

How does masculine vs. feminine organizational culture vary across major US public firms,
and how do firm-level cultural patterns evolve over time?

Theoretical anchors:

- Acker (1990): gendered organizations and the abstract worker
- Kanter (1977): masculine ethic of management
- Williams et al. (2012): gendered organizations in the new economy

A 4-dimensional codebook for gendered culture

Dimension	+2 anchor (masculine)	−2 anchor (feminine)
1. Competition vs. Cooperation	**Zero-sum competition**, individual stardom, internal politics	**Collaborative**, team-based, supportive coworkers
2. Achievement vs. Care	**Results-driven**; employees as **replaceable** performers	**People-first**; visible care for employee wellbeing
3. Assertiveness vs. Modesty	**Tough**, demanding, hard-charging management	**Humble**, approachable, consensus building leadership
4. Performance vs. Quality-of-Life	**Long hours** valorized; "ideal worker" has no body	**Bounded hours**; work–life balance respected

All four dimensions are *descriptive* — coding the culture, not the reviewer's stance.
Each dimension scored −2 to +2; composite = mean of the four sub-dimensions.

Prompt design: naive vs. structured

Naive baseline

```
Read this employee review and tell me how masculine vs.
feminine the workplace culture seems on a 1-7 scale.
[review text]
Score:
```

What's wrong:

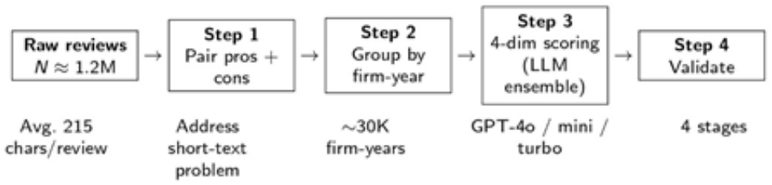
- No definition (no static content)
- Single score collapses 4 dimensions
- No output format
- No evidence grounding
- No reasoning trace

Structured workflow

```
[SYSTEM] Expert organizational sociologist...
[USER] Score on 4 INDEPENDENT dimensions:
  1. COMPETITIVE_INDIVIDUALISM
  2. TIME_AND_BODY_DEMANDS
  3. EMOTIONAL_REGISTER
  4. CULTURAL_ENDORSEMENT
For each: score 1-7 or NA, anchor with direct quote, distinguish
descriptive from normative.
REVIEW: """[text]"""
OUTPUT (JSON only):
  {"comp_indiv": {"score":..., "quote":...}, ...}
```

← persona (Alammar) ← workflow (Berryman)
← static content ← Chekhov guard
← informative NA (Benoit) ← constrained output

The full pipeline



Two key design decisions:

- **Pair pros + cons** as one unit — preserves the structural contrast Glassdoor's format provides
- **Aggregate to firm-year** — shifts unit of analysis from individual reviews to firms; absorbs per-review NA naturally

Short reviews are not a defect to overcome —
they are a signal that firm-level is the right scale of analysis.

Full corpus analyzed here: 20 firms × 2011–2019 = 32,835 reviews across 179 firm-years.

A worked example: Goldman Sachs, 2018 (pilot)

Excerpts from N=200 reviews (8 shown)

R1: *"Management doesn't care about Analyst and are treated as replaceable... everyone below is treated like peasants."*

R2: *"Investment banking at GS: very intense and challenging yet rewarding. Workload can sometimes be overwhelming."*

R3: *"Great managers, great firm outlook, strong culture vs. competitors. Long work hours."*

R4: *"Good pay and super smart people. Bad hours, you know what I mean."*

R6: *"Nice environment, competitive environment, appreciated at very good performances. Work pressure, very less social interaction."*

(192 more reviews aggregated to firm-level)

Per-review LLM output (JSON, −2 to +2 scale)

```
{ "scores": [
  {"review_idx": 0, comp_vs_coop: 0,
   ach_vs_care: 0, assert_vs_modesty: 0,
   perf_vs_qol: 0},
  {"review_idx": 1, comp_vs_coop: 0,
   ach_vs_care: +2, assert_vs_modesty: +1,
   perf_vs_qol: +2},
  {"review_idx": 2, comp_vs_coop: 0,
   ach_vs_care: +1, assert_vs_modesty: 0,
   perf_vs_qol: +1},
  {"review_idx": 3, comp_vs_coop: 0,
   ach_vs_care: 0, assert_vs_modesty: 0,
   perf_vs_qol: +1},
  {"review_idx": 6, comp_vs_coop: +1,
   ach_vs_care: +1, assert_vs_modesty: 0,
   perf_vs_qol: +1},
  ...] }
Firm-level (mean of 200):
 comp_vs_coop = +0.10,
 ach_vs_care = +0.14,
 assert_vs_modesty = +0.08,
 perf_vs_qol = +0.71
```

Goldman scores positive (masculine) on all four dimensions — consistent with prior expectation.
The full panel (slides 17a–f) extends this to 32,835 reviews over 2011–2019.

Four-stage validation without ground truth

	Stage	Full-panel result (179 firm-years, 2011–2019)
1	**Test–retest reliability** Run identical prompt twice; correlate per-review and firm-level scores	Pilot (2018 subset): review-level $r = 0.95$, firm-level $r = 0.99$ *(not re-run at full scale; temperature = 0 and 9-year within-firm stability serve as de-facto retest)*
2	**Internal consistency (Cronbach's α)** Do sub-dimensions measure a coherent construct?	Review-level $\alpha = 0.60$ Firm-year-level $\alpha = \mathbf{0.70}$
3	**Convergent / external validity** Correlate firm-year scores with Glassdoor's structured sub-ratings	vs. work-life balance $r = \mathbf{-0.67}$ vs. overall rating $r = -0.58$ vs. culture rating $r = -0.16$
4	**Prior-check validation** Pre-classify firms by stereotype (M/F/X); check whether the measure recovers the difference	M-firm mean − F-firm mean = $\mathbf{+0.15}$ *(after excluding reviews with no MAS cues)*

None is sufficient alone. Together, they triangulate.

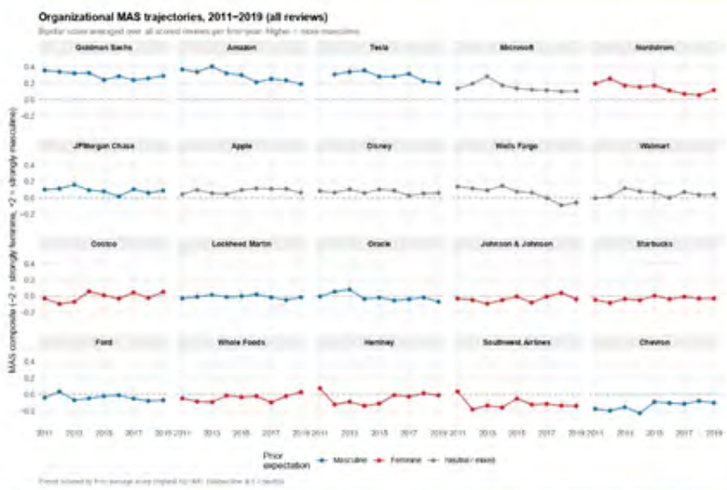
Scaling up: from 2018 pilot to a 9-year panel

Pilot (earlier slide)	Full panel (this section)
20 firms × 2018	20 firms × 2011–2019
≈ 3,900 reviews	**32,835 reviews**
Single cross-section	**179 firm-years** (panel)
Demonstrated the *workflow*	Tests the measure at *scale*

Scaling up raises new measurement questions:

- **Temporal stability** — is firm culture a stable attribute, or measurement noise?
- **Reviewer selection** — do reviewers at different moments write about different things?
- **Signal dilution** — many reviews say nothing about culture at all.

Full-panel result: firm-year trajectories, 2011–2019



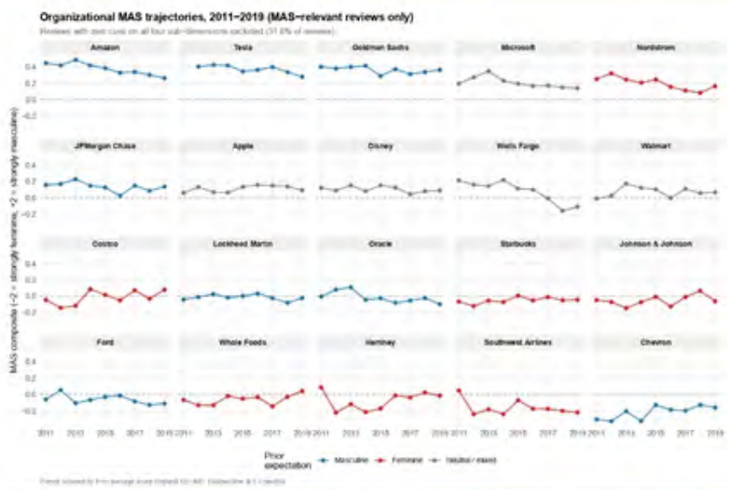
Diagnosis: a third of reviews carry no MAS signal

31.6% of scored reviews receive 0 on all four sub-dimensions.
These reviews describe pay, benefits, location, or tasks — not culture.
Including them as zeros pulls every firm's mean toward the origin.

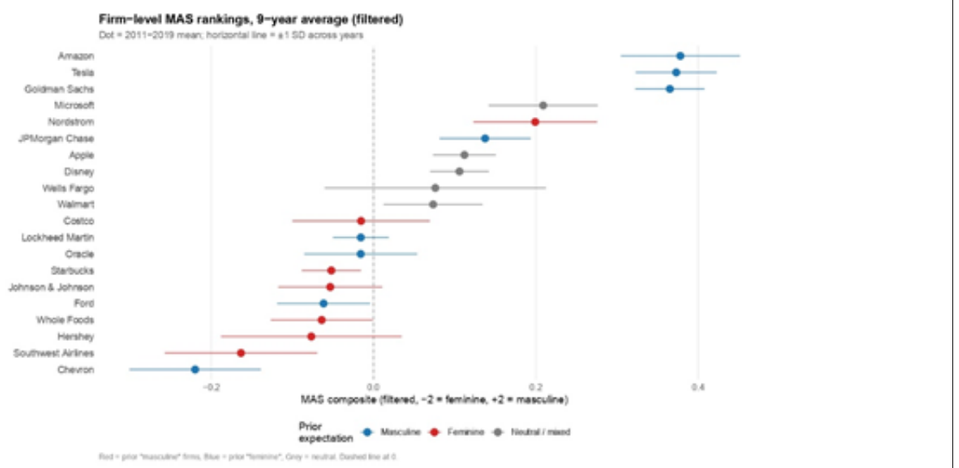
Firm (selected)	**% of reviews with no MAS cues**	**Interpretation**
Lockheed Martin	39.8%	defense / clearance talk, not culture
Walmart	38.1%	shift & scheduling dominate
Wells Fargo	36.9%	stable bank-branch routine
Chevron	35.8%	operational, not cultural, descriptions
⋮	⋮	⋮
Southwest	30.9%	strong "fun" culture cues
Oracle	29.8%	tech firms discuss culture more

This is not an LLM failure — it is a **corpus property**.
The measurement pipeline must accommodate it.

Filtering irrelevant reviews sharpens every diagnostic



Firm-level rankings (filtered, 2011–2019 average)



Firm culture is a stable attribute, not noise

Decomposing variance (filtered panel):

Source	SD
Between firms (variance across the 20 firms)	**0.17**
Within firm, across 9 years (variance within each firm over time)	0.07
Ratio (between ÷ within) ≈ 2.6 ⇒ ICC ≈ 0.87	

- **87% of variance in the measure is between firms**; only 13% is year-to-year fluctuation within firms.
- This is what we should expect from organizational culture: slow-moving, institutionalized, hard to change.
- The 9-year panel functions as **nine independent measurements per firm** — temporal stability is itself a reliability check.

What looked like sparse, noisy reviews aggregates into a measure that behaves the way a cultural attribute should.

Firm-level rankings (filtered, 2011–2019 average)

Masculine direction ($n = 10$)

Firm	prior	mas
Amazon	M	+0.38
Tesla	M	+0.37
Goldman Sachs	M	+0.37
Microsoft	X	+0.21
Nordstrom	F	+0.20
JPMorgan Chase	M	+0.14
Apple	X	+0.11
Disney	X	+0.11
Wells Fargo	X	+0.08
Walmart	X	+0.07

Feminine direction ($n = 10$)

Firm	prior	mas
Costco	F	−0.02
Lockheed Martin	M	−0.02
Oracle	M	−0.02
Starbucks	F	−0.05
Johnson & Johnson	F	−0.05
Ford	M	−0.06
Whole Foods	F	−0.06
Hershey	F	−0.08
Southwest Airlines	F	−0.16
Chevron	M	−0.22

Chevron, *a "masculine" prior, lands at the feminine extreme — unionized, stable, paternalistic.*
Nordstrom, *a "feminine" prior, lands masculine — commission-based luxury retail is high-pressure.*
The measure does not just confirm priors; it **corrects** *them where reasonable.*

Part 4: A generalizable protocol

Beyond Glassdoor, the same five steps apply:

1. **Decompose the concept into independent dimensions**
   Resist single-score temptation; multi-dim is where LLMs add value
2. **Decompose the workflow into separable steps**
   Cheap filter → expensive analysis → ensemble
3. **Constrain output structurally**
   JSON, evidence quotes, confidence — by design, not post-hoc
4. **Ensemble across models with different RLHF lineages**
   Distinct model families (e.g., GPT-4o, GPT-4o-mini, GPT-4-turbo); adding vendors (Claude, Gemini) strengthens further
5. **Validate via theoretical prediction, not ground truth**
   Pre-register hypotheses derived from existing theory

This is what scales the Benoit & Than approach to the rest of social science.

Where this matters most

Three frontiers without validation infrastructure:

Organizational discourse	Everyday talk	Informal labor
Glassdoor reviews, internal memos, employee interviews, organizational documents	Reddit threads, online forum posts, YouTube comments, dating-app messages	Delivery-app reviews, platform-worker forums, gig economy testimonials

None of these has expert surveys. None has decades of hand-coded labels.
All of them are where LLM measurement adds the most value.

Three takeaways

1. **Generative LLMs are not faster NLP.**
   They are a new *kind* of measurement instrument — worthy of separate methodological attention.
2. **Domains without validation infrastructure are the real frontier.**
   The hard problems aren't manifestos and newsweeklies. They're everywhere else.
3. **Industry has done methodological work we should borrow.**
   Prompt engineering is not a hack. It is, increasingly, applied measurement theory.

Thank you. Questions welcome.

sangwonhan@skku.edu

Backup: Full prompt (for reference)

```
[SYSTEM]
You are an expert organizational sociologist analyzing employee reviews for gendered cultural patterns.  You evaluate four independent
dimensions drawn from Acker (1990), Kanter (1977), and Williams et al. (2012).
[USER]
TASK: Read the following N employee reviews from {firm_name} during {year}.  Produce ONE set of four-dimension scores characterizing
the firm's gendered organizational culture during this period.
DIMENSIONS (1 = feminine anchor, 7 = masculine anchor):
1.  COMPETITIVE_INDIVIDUALISM: team-based (1) vs. zero-sum competition (7)
2.  TIME_AND_BODY_DEMANDS: bounded hours, care accommodated (1) vs. long hours, ''ideal worker'' (7)
3.  EMOTIONAL_REGISTER: vulnerability acknowledged (1) vs. toughness, emotion suppressed (7)
4.  CULTURAL_ENDORSEMENT: reviewers critique the pattern (1) vs. celebrate or naturalize it (7)
INSTRUCTIONS:
- Score each dimension 1-7 or "NA" if reviews provide insufficient evidence.
- Anchor each score with quotes from at least 2 different reviews.
- Distinguish DESCRIPTIVE (D1-3) from NORMATIVE (D4) content.
- Score CULTURAL_ENDORSEMENT based on reviewers' explicit reactions, not your own judgment.
- Do not infer based on the firm's name or industry; rely only on the review text.
REVIEWS FROM {firm_name}, {year}:
--- Review 1 ---
PROS: {review_1_pros}
CONS: {review_1_cons}
--- Review 2 ---
...up to N reviews ...
OUTPUT (JSON only):
{ "competitive_individualism": {"score":..., "evidence_quotes":[...], "confidence":...},
  "time_and_body_demands":   {...},
  "emotional_register":      {...},
  "cultural_endorsement":    {...} }
```


Backup: Cost and runtime estimates

Per firm-year cost (one prompt = up to 30 reviews):

Model	Role	Cost / firm-year	Time
GPT-4o (flagship)	Primary scorer	∼$0.025	∼5s
GPT-4o-mini (distilled)	Cheap scorer	∼$0.002	∼3s
GPT-4-turbo (prior gen)	Diversity scorer	∼$0.10	∼6s
Ensemble (3 models)		**∼$0.13**	**∼14s** (parallel: 6s)

Full corpus estimates ($N \approx 30$K firm-years):

- Step 1 (pair pros + cons + group): negligible cost, ∼1 hour
- Step 2 (firm-year scoring, 3-model ensemble): ∼$4,000, ∼3 days parallel
- Step 3 (validation, including convergent checks): negligible
- **Total: ∼$4,000, ∼3–5 days wall time**

Actual cost (20 firms × 2011–2019, 32,835 reviews, single GPT-4o run):

- 4,105 batches of up to 8 reviews: ∼$30, ∼50 minutes
- A second run for test–retest would double cost (not executed here;
  9-year within-firm stability substitutes as a de-facto reliability check)

Firm-year aggregation reduces cost ∼ 7× vs. per-review scoring,
while producing the substantively meaningful unit of analysis.

MEMO

MEMO