

OpenAlex 기반 지식 생산 구조의 실증 분석: LLM 분류를 통한 비정형 데이터의 변수화 및 통계적 검증

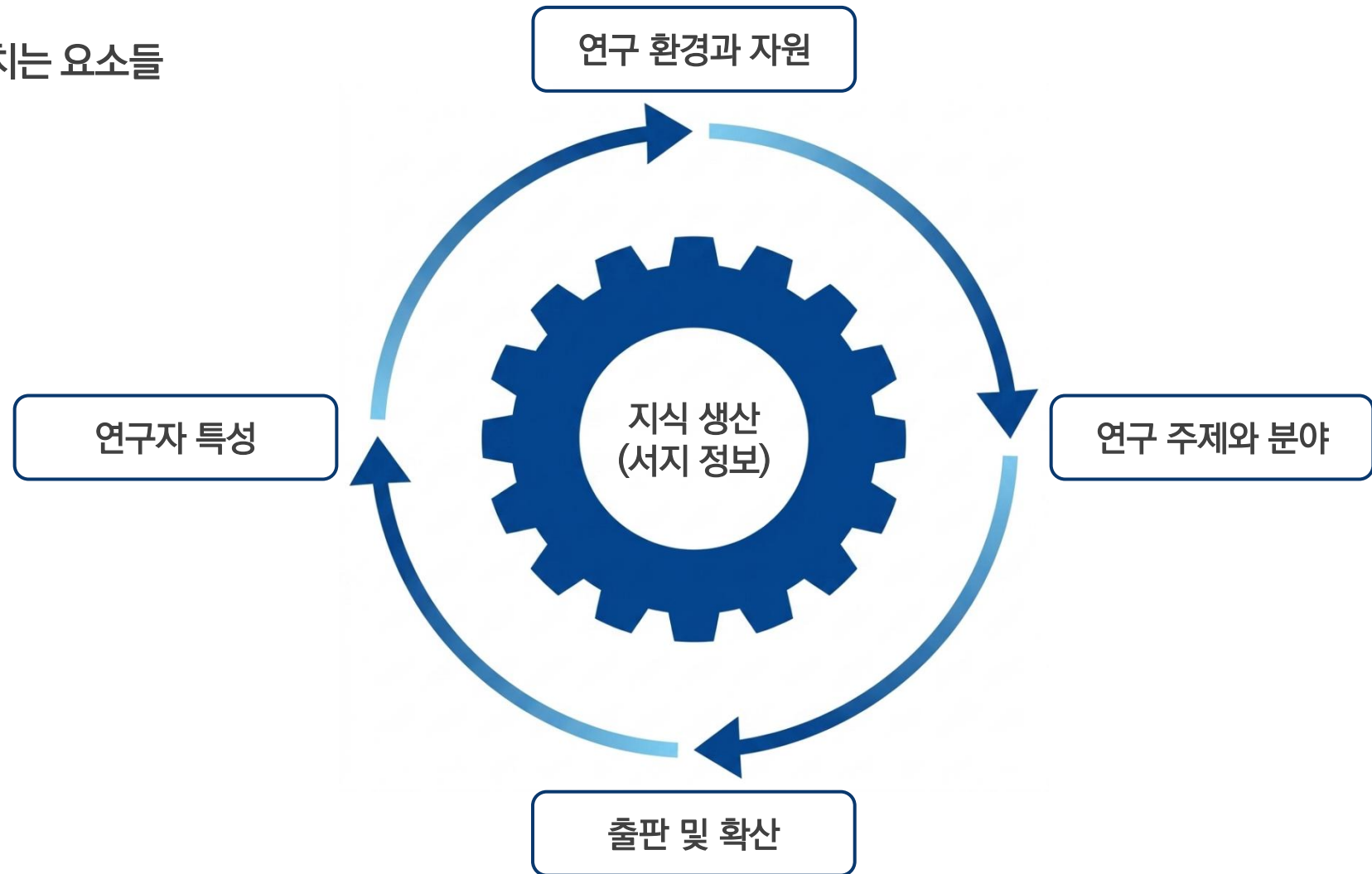
권은낭

한국과학기술원 디지털인문사회과학센터 연수연구원

eunrang_kwon@kaist.ac.kr

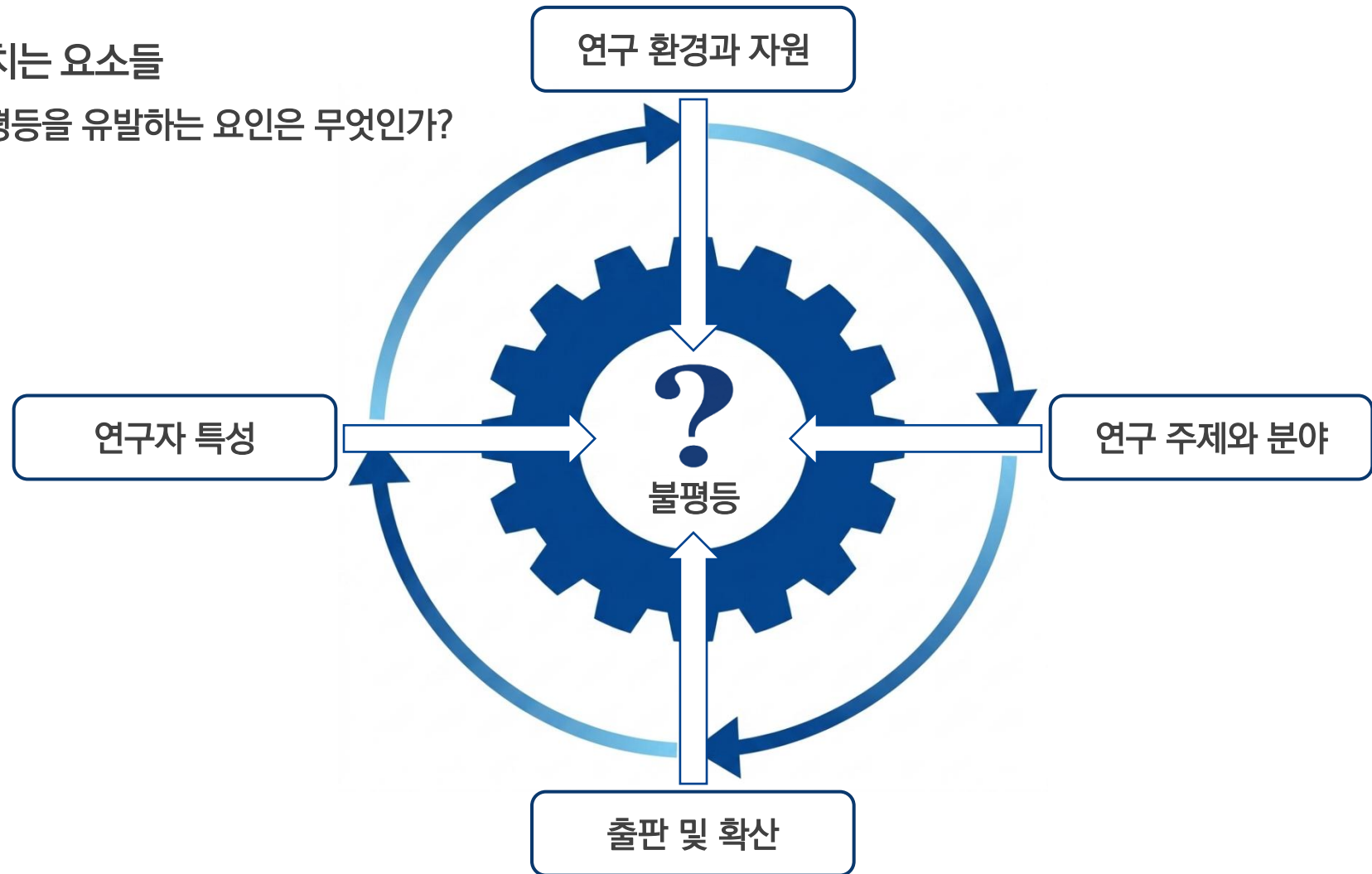
연구질문: 지식 생산의 구조적 불평등

지식 생산에 영향을 미치는 요소들



연구질문: 지식 생산의 구조적 불평등

지식 생산에 영향을 미치는 요소들
그렇다면 지식 생산의 불평등을 유발하는 요인은 무엇인가?



연구사례: 지식 생산의 구조적 불평등

✓ 사례1

- 여성 연구자의 지식 생산
- 연구자의 환경 및 특성을 고려했을 때, 코로나 이전과 비교해 여성 연구자의 논문 생산성이 감소했는가?
- 외부 데이터셋을 활용한 변수 생성

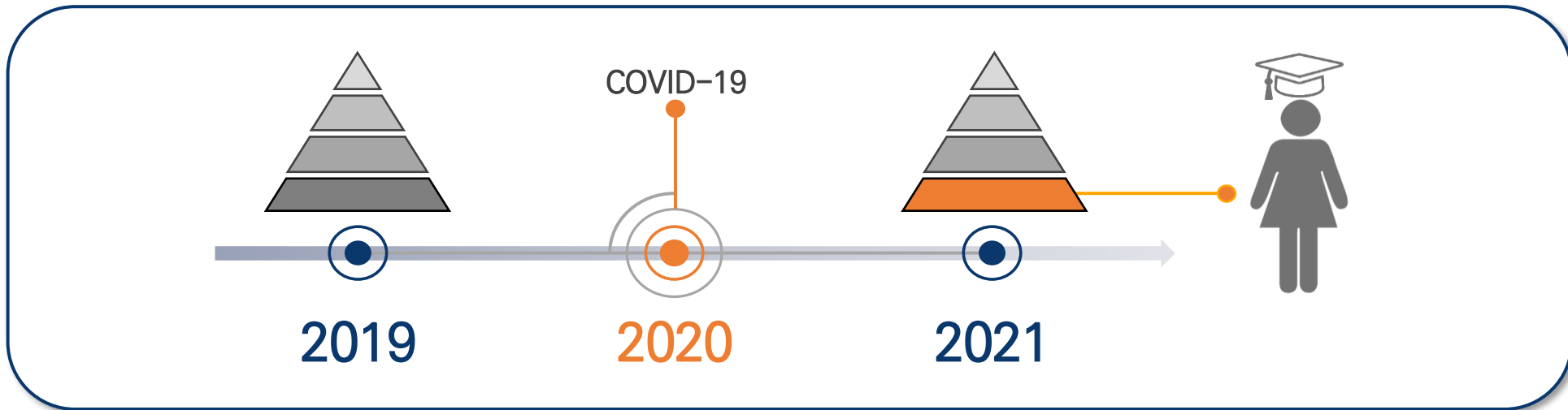
✓ 사례2

- 중심국과 비중심국의 지식 생산
- 중심국-비중심국 위계 구조가 사회학 지식 생산에 어떠한 영향을 미치는가?
- OpenAI를 활용한 변수 생성

사례1: 코로나 시기 여성 연구자의 논문 생산성

사례1: 연구자의 환경 및 특성을 고려했을 때, 코로나 이전과 비교해 여성 연구자의 논문 생산성이 감소했는가?*

외부 데이터셋을 활용한 변수 생성



*Kwon, E., Yun, J., & Kang, J. H. (2023). The effect of the COVID-19 pandemic on gendered research productivity and its correlates. *Journal of Informetrics*, 17(1), 101380.

사례1: 코로나 시기 여성 연구자의 논문 생산성

❖ 데이터 설명

데이터셋 구축

- 연구자의 환경 및 특성을 고려한 데이터셋을 구축하기 위해 서지 정보와 외부 데이터를 결합.

서지 정보 (Microsoft Academic Graph, MAG))

- 2016년부터 2020년 사이의 Microsoft Academic Graph (MAG) 데이터를 활용.
- 학술 연령(academic age): 저자의 첫 번째 저널 논문과 현재 사이에 경과된 시간.
- h-index: 연구자가 발표한 논문 중 최소 h번 이상 인용된 논문이 h편 이상이고, 나머지 논문들은 h번 미만으로 인용되었을 때의 h값.

MAG와 외부 데이터의 결합*

- 저자의 성별: Genderize.io (웹 애플리케이션)를 사용하여 저자의 이름(first name) 기반 성별 추론.
- 국가별 성평등 수준: Linkedin의 Gender Gap index (GGI, 성격차 지수) 활용.
- 국가별 백만 명당 총 확진자 수/사망자 수: The Johns Hopkins Coronavirus Resource Center의 코로나19 공표 자료 활용.
- 코로나19 기간 동안 국가별 이동성: Google의 직장 및 주거 이동성 (Workplace and Residential mobility) 자료 활용.

*각 논문의 국가는 첫번째 저자의 소속기관이 위치한 국가

사례1: 코로나 시기 여성 연구자의 논문 생산성

❖ 데이터 설명

Genderize.io 를 활용한 저자 이름 기반한 성별 분류 (<https://genderize.io/documentation>)

Articles

The Politics of International Aid to the Gaza Strip

Tamer Qarmout & Daniel B-land

Pages 32-47 | Published online: 21 Dec 2020

🗨 Cite this article 🔗 <https://doi.org/10.1525/jps.2012.XLI.4.32>

🌐 Tamer Qarmout 🔍

Tamer Qarmout is male with 99% certainty

cURL

Python

JavaScript

PHP

Ruby

Go

```
curl "https://api.genderize.io?name=peter&apikey=YOUR_API_KEY"
```

```
{
  "name": "peter",
  "gender": "male",
  "probability": 1.0,
  "count": 1346866
}
```

사례1: 코로나 시기 여성 연구자의 논문 생산성

❖ 분석 방법

CID (Change in Difference)

- CID의 정의: 두 집단 간 이중차분법 (difference-in-difference, DID) 의 변형으로 정의
- 이는 최근 몇 년 동안 여성 저자의 논문 생산이 지속적으로 증가하거나 감소한 추세를 넘어, 2020년에 나타난 순변화를 포착할 수 있음

〈CID에 대한 계산식〉

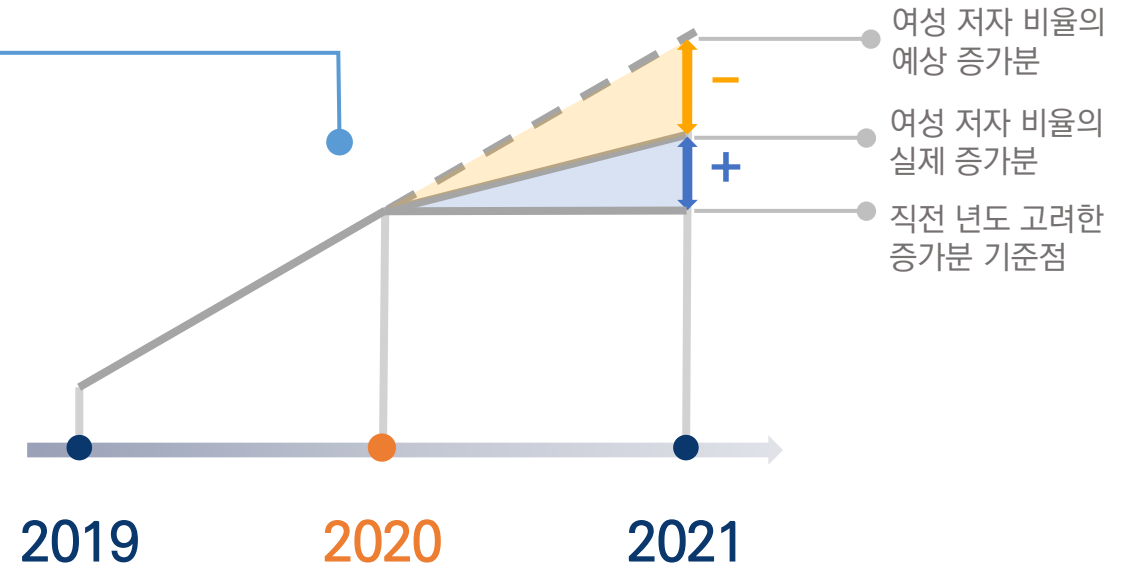
$A(X)$: X년도의 총 논문 수

$B(X)$: X년도 여성 저자의 논문 수

$F(X) = B(X)/A(X)$: 여성 저자의 비율

$$CID(X) = [F(X) - F(X-1)] - [F(X-1) - F(X-2)]$$

$$\text{ex) } CID(2020) = [F(2020) - F(2019)] - [F(2019) - F(2018)]$$

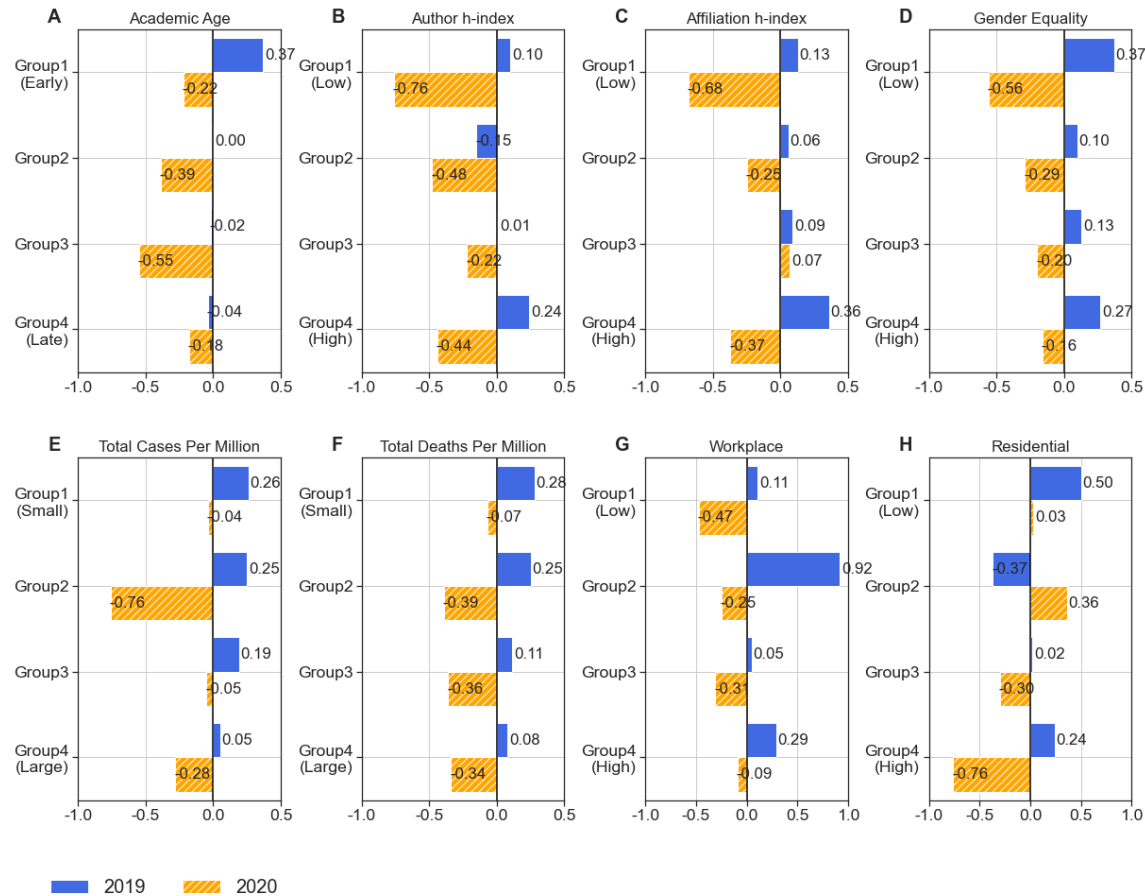


XGboost를 이용한 요인 중요도 추정

연도 및 저자 역할별로 여성 저자의 존재 여부를 예측하기 위해 변수들의 요인 중요도(feature importance)를 분석함

사례1: 코로나 시기 여성 연구자의 논문 생산성

연구자 특성별 여성 연구자의 논문 생산성



연구자 특성별 성별화된 연구 생산성

생산성 감소가 두드러진 여성 제 1 저자 기준으로 검증,
각 변수를 수준(level)에 따라 4개의 그룹으로 나눔.

Fig. 3. CID for % female first authors by first author's characteristics.

사례1: 코로나 시기 여성 연구자의 논문 생산성

연구자 특성별 여성 연구자의 논문 생산성

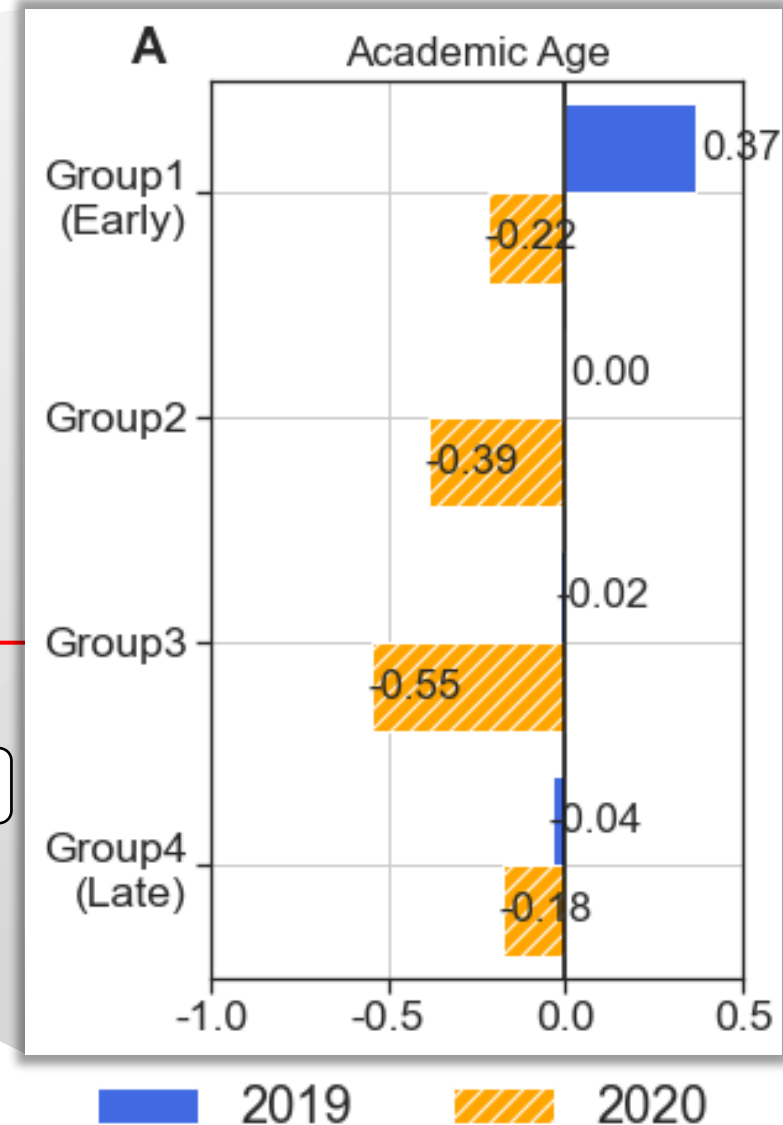


Fig. 3. CID for % female first authors by first author's characteristics.

사례1: 코로나 시기 여성 연구자의 논문 생산성

연구자 특성별 여성 연구자의 논문 생산성

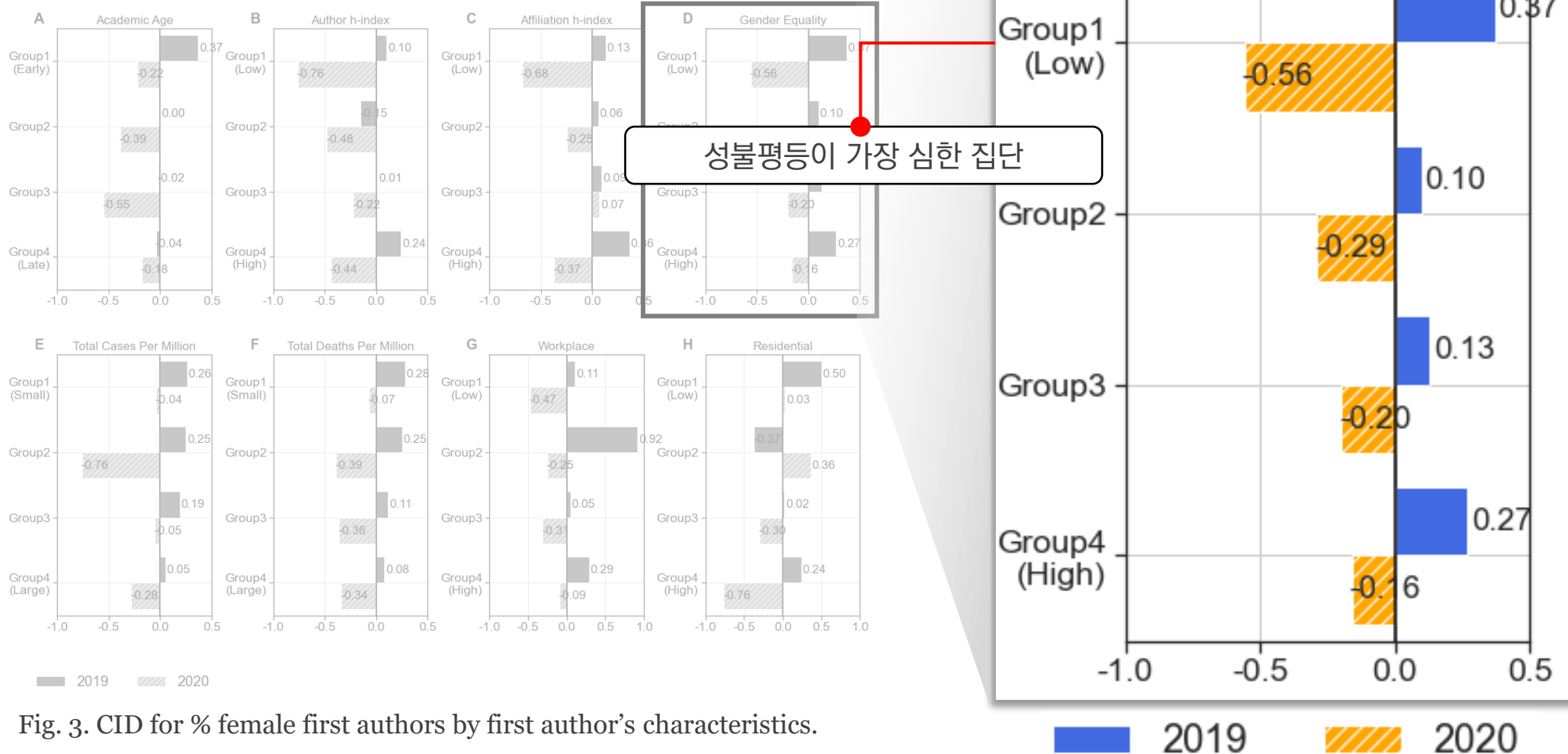


Fig. 3. CID for % female first authors by first author's characteristics.

사례1: 코로나 시기 여성 연구자의 논문 생산성

연구자 특성별 여성 연구자의 논문 생산성

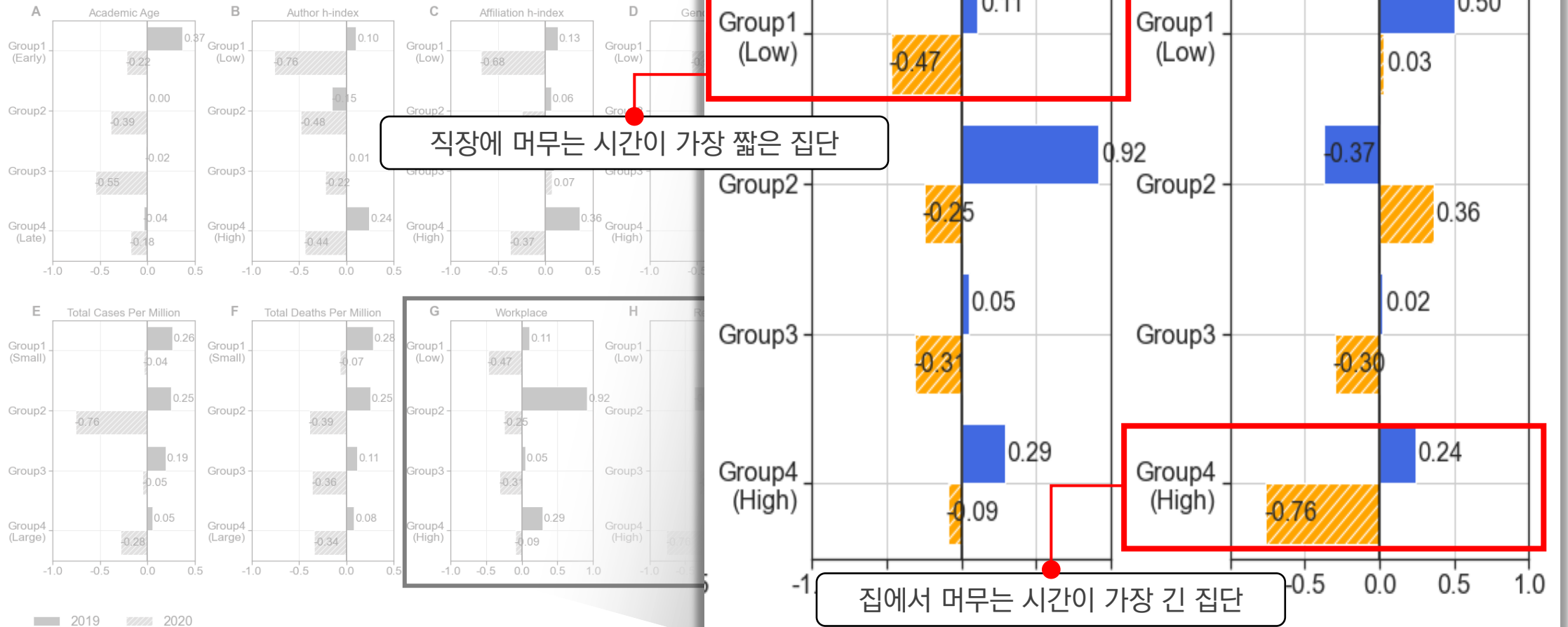
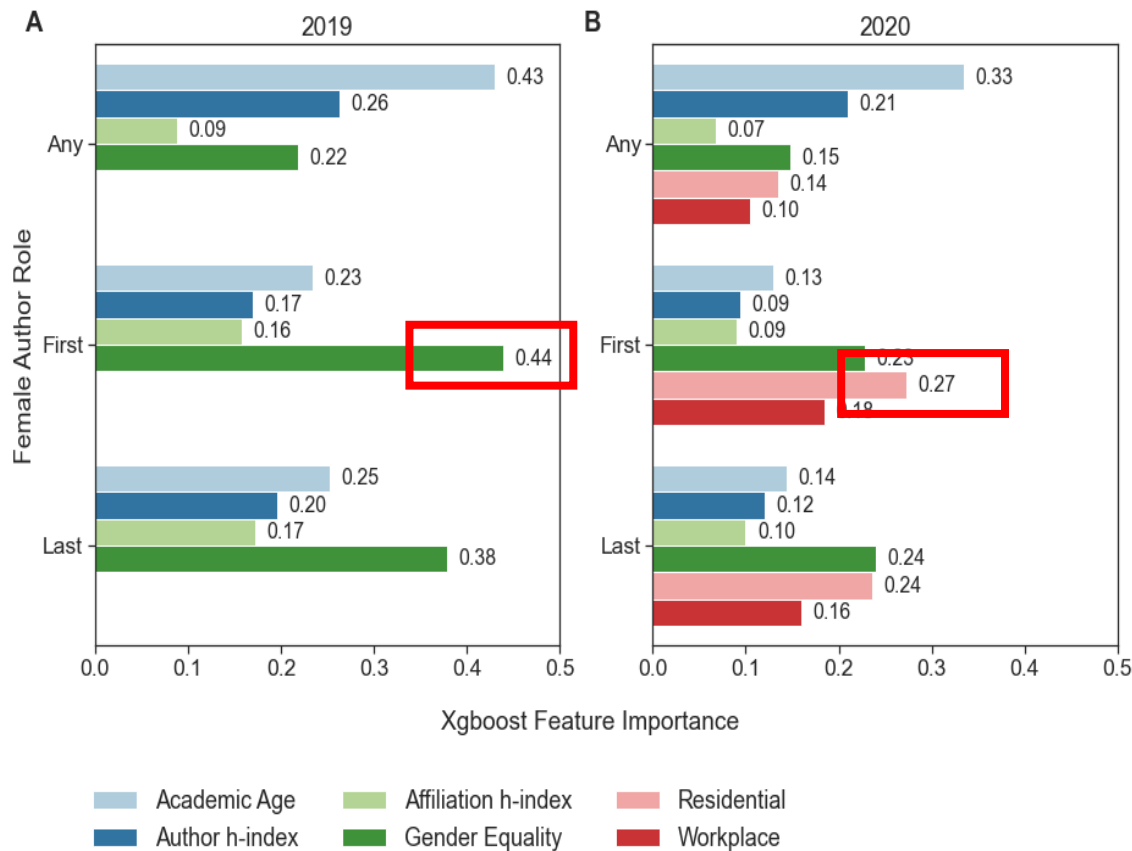


Fig. 3. CID for % female first authors by first author's characteristics.

사례1: 코로나 시기 여성 연구자의 논문 생산성

연구자 특성들이 생산성에 미치는 상대적 중요도



연구자 특성들은 성별화된 연구 생산성을 얼마나 잘 설명하는가?

저자 역할별 논문 내 여성 저자의 존재 여부에 대한 요인 중요도 예측
코로나19 이전인 2019년의 분석은 이동성 변수 제외함

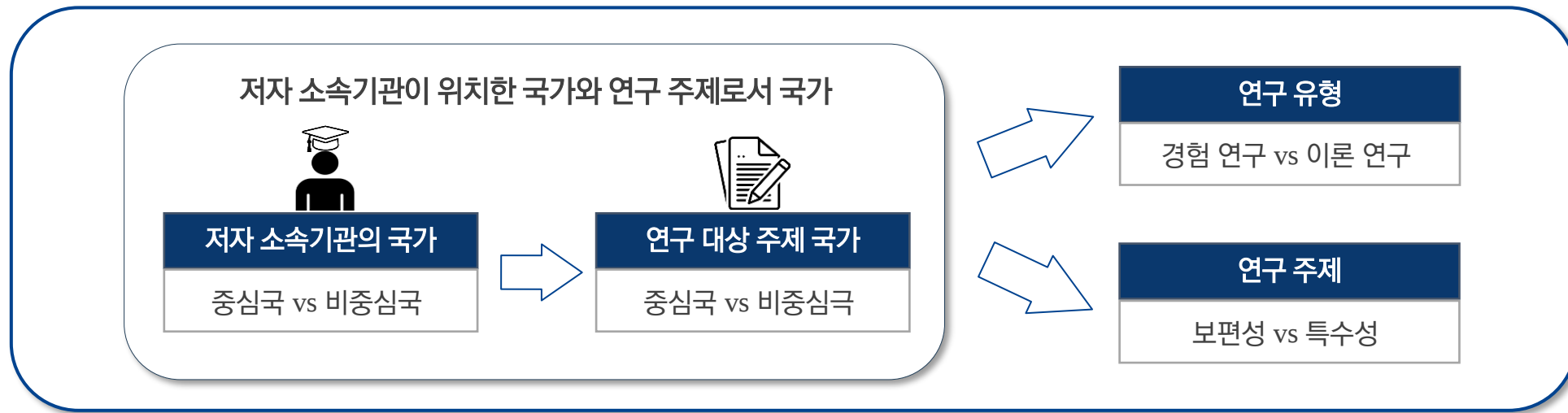
- 높은 설명력을 보인 요인의 경우,
2019년 성평등 수준에서 2020년 주거 이동성으로 요인 변화.
- 이를 통해, 코로나19 기간 동안 국가 수준의 이동 제한 조치가
논문에서 주도적인 역할을 맡은 여성 연구자의 연구 생산성에
영향을 미쳤음을 보여줌.

Fig. 5. Feature importance by female author's characteristics measured with XGBoost

사례2: 중심국과 비중심국에 따른 사회학 지식 생산

사례2: 중심국-비중심국 위계 구조가 사회학 지식 생산에 어떠한 영향을 미치는가?*

OpenAI를 활용한 변수 생성



*Kwon, E., Kim, L., & Jeon, J. 2025 Organized Unknowns in Global Sociology: Systematic Blind Spots of Sociology in Core and Peripheral Countries, 2025 후기 한국사회학회.

사례2: 중심국과 비중심국에 따른 사회학 지식 생산

❖ 데이터 설명

데이터셋 구축

- 데이터 출처: [OpenAlex](#) (서지정보 데이터베이스)
- 분석 대상 기간: 20년 (2004년 – 2024년)
- 대상: 75개 SSCI 저널 (사회학 분야 / Subject Area 1)
- 포함 기준: 초록(Abstract)이 포함된 논문
- 최종 표본: 전처리 후 총 38,716개의 초록

변수 구성

- **저자 소속 국가**: OpenAlex에서 제공하는 제1저자의 소속 기관 메타데이터로부터 추출함.
- **주제 국가(연구 대상 국가) 및 연구 유형**: OpenAI를 활용한 자동화된 초록 분석을 통해 분류함.
- **연구 주제**: 구조적 토픽 모델링(STM)을 통한 잠재적 주제 분석 및 토픽 모델링.

사례2: 중심국과 비중심국에 따른 사회학 지식 생산

변수 구성의 예시

Articles

The Politics of International Aid to the Gaza Strip

Tamer Qarmout & Daniel B-land

Pages 32-47 | Published online: 21 Dec 2020

Abstract

International aid to the Palestinian Authority is conditioned in part on democratization and good governance. However, since Hamas's victory in the 2006 Palestinian Legislative Council elections and its takeover of the Gaza Strip, aid agencies have supported the international boycott of the Hamas government. This article argues that aid agencies, by operating in Gaza while boycotting its government, subvert their mandates and serve the political interests of donors and the PA rather than the humanitarian and development needs of Gazans. As a consequence, assistance has, inadvertently and unintentionally, increased Gazans' dependence on humanitarian aid, impeded economic development, and enabled Israel to maintain its occupation and the blockade of Gaza.

Country of Author Affiliation

OpenAlex

- Doha Institute for Graduate Studies, Qatar → Periphery

Country of Study

OpenAI

- State of Palestine
- Israel → Periphery

Research Type

OpenAI

- Empirical

Research Topic

STM

- Public Policy, Political Science

사례2: 중심국과 비중심국에 따른 사회학 지식 생산

OpenAI 프롬프트 엔지니어링

- 주제 국가 추출 (사용 모델: GPT-5nano)
 - 초록에 언급된 지역, 국적 또는 특정 국가를 추출하여 표준화된 국가명으로 매핑함.
- 연구 유형 분류 (사용 모델: GPT-5)
 - 상호 배타적인 세 가지 범주인 실증(Empirical), 이론(Theory), 방법론(Method) 중 하나로 단일 레이블 분류함.
 - 의미론적 추론: 정확한 분류를 보장하기 위해 각 연구 유형에 대한 명확한 정의와 키워드 예시를 프롬프트에 포함함:
 - Empirical: Analyzes data/observations to report findings.
 - Theory: Develops/refines conceptual frameworks or propositions.
 - Method: Proposes/evaluates a tool, procedure, dataset, metric, or methodology.

변수의 정확도

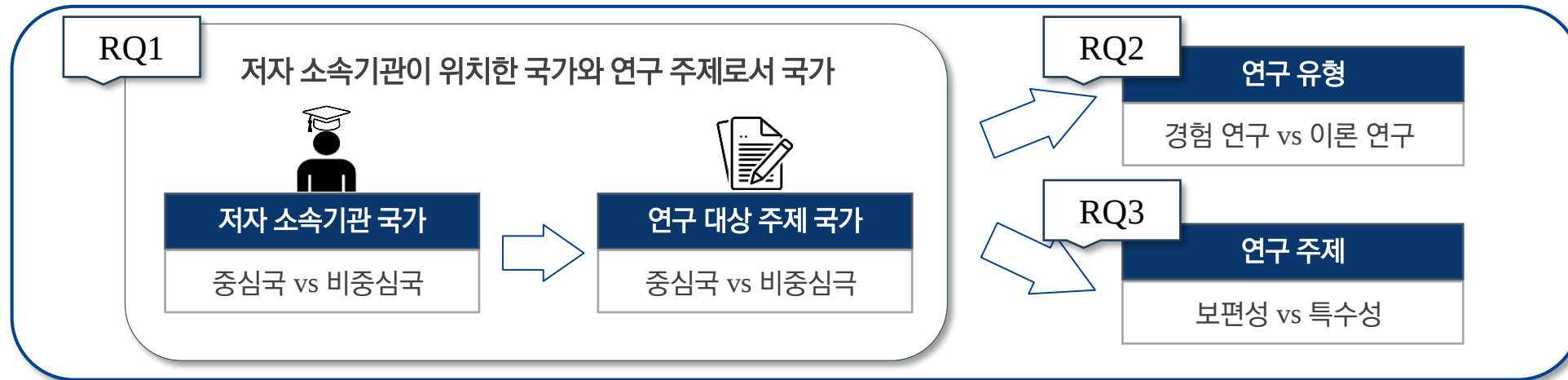
- 분류 성능을 평가하기 위해 무작위로 선정된 100편의 논문으로 골드 스탠더드(gold-standard) 데이터셋을 수동으로 구축함.
- 분류 정확도(Accuracy) 결과: 주제 국가: 0.98, 연구 유형: 0.88의 성능을 보임.

연구 유형 분류: Model Performance Evaluation			
Model	Zero-shot	Few-shot (1-sample)	Few-shot (3-sample)
GPT-5	0.884	0.876	0.872
GPT-5-mini	0.857	0.841	0.846
GPT-5-nano	0.828	0.820	0.780

사례2: 중심국과 비중심국에 따른 사회학 지식 생산

연구질문

- 연구질문 1) 저자 국가와 동일한 국가를 주제로 더 많이 연구하는가?
- 연구질문 2) 비중심국 연구자의 비중심국 연구는 경험적 연구일 확률이 더 높은가?
- 연구질문 3) 저자 국가와 주제 국가별 주로 연구하는 토픽은 무엇인가?



분석방법

- 지리적 변수 및 연구 유형 변수의 연계: 통계 분석
 - 카이제곱 검정(Chi-square Test) 및 로지스틱 회귀분석(Logistic Regression)
- 토픽 분포 분석: 구조적 토픽 모델링 (STM, Structural Topic Modeling)
 - 문서 수준의 메타데이터를 추정 과정에 반영하면서 연구 초록으로부터 잠재된 주제를 추출하도록 설계된 텍스트 분석 방법

사례2: 중심국과 비중심국에 따른 사회학 지식 생산

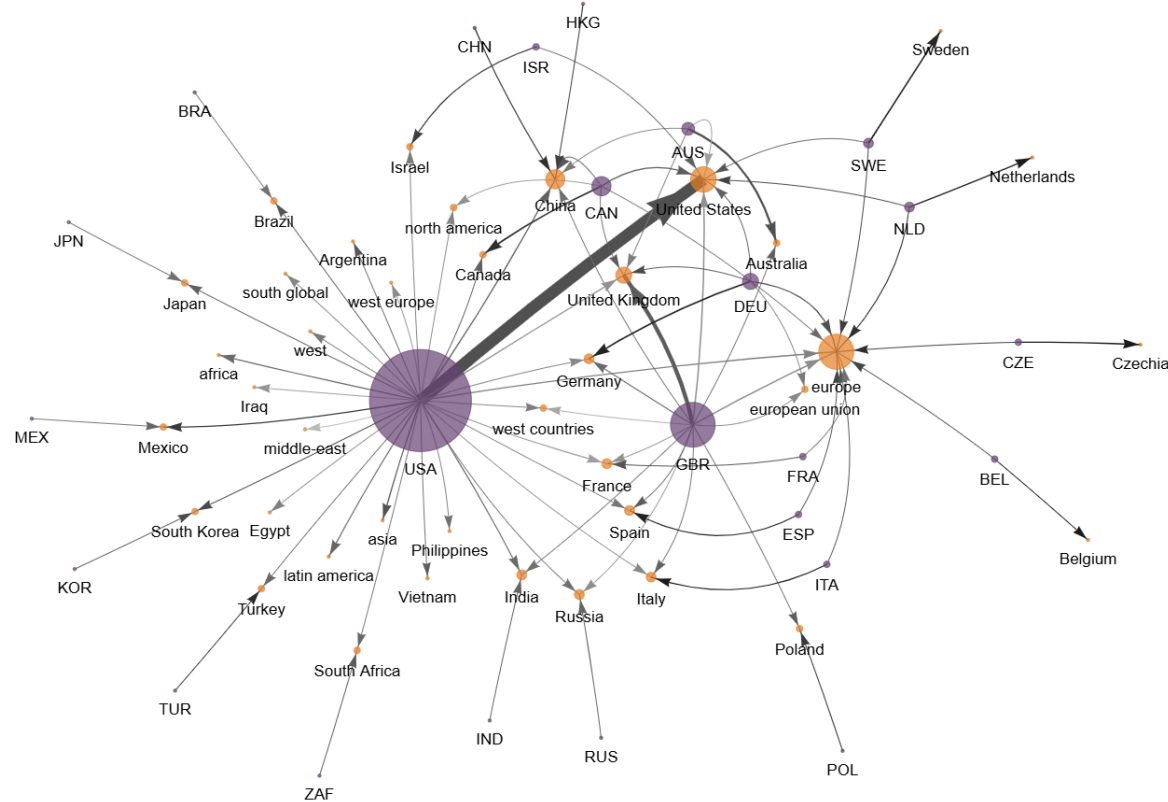


Figure 2. Mapping the Linkages between Country of Author and Country of Study



연구 동향 파악을 위한 네트워크 분석

- 저자 소속 국가와 주제 국가 간의 네트워크를 분석함.

주요 저자 소속 국가:

- 미국, 캐나다, 유럽 및 기타 영어권과 유럽 국가들.

주요 주제 국가:

- 미국, 영국, 유럽, 중국 등.
- 중국은 비중심국이지만 주요 연구 대상 국가로 나타났으며, 미국, 캐나다와 같은 중심국 연구자들에 의해 주로 연구됨.

그룹별 연구 경향:

- 중심국(미국, 유럽 등) 출신 연구자들은 주로 중심국 연구, 비중심국 저자와 비중심국 연구 대상 국가의 비율은 상대적으로 낮게 나타남.

사례2: 중심국과 비중심국에 따른 사회학 지식 생산

Table 2. Chi-square Test Results: Country of Author and Country of Study

Variable		Group				χ^2	DF	P-value	
Country of Study		Country of Author				7,379.4	2	0.000***	
		Value	Core	Periphery	Total				
		Core	Observed	15,735	737				
			Expected	14,531.9	1,940.1				16,472
			Residual	9.98***	-27.32***				
		Periphery	Observed	2,278	2,257				
			Expected	4,000.9	534.1				4,535
			Residual	-27.2***	74.55***				
		Not mentioned	Observed	10,285	784				
			Expected	9,765.3	1,303.7				11,069
	Residual		5.3***	-14.39***					
Total	Observed	28,298	3,778	32,076					

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

Table 3. Logistic Regression Analysis Predicting Periphery-Study

Dependent Variable: Periphery-focused studies (Core or Not-mentioned=0)		Model
		β (SE)
Country of Author (Periphery, Core=0)	OR=15.14	2.717(0.089)***
Journal Impact Factor		-0.203(0.089)*
Rescaled Publication Year		0.015(0.008)
Constant		-2.034(0.289)***
Observation		32,076
Log-likelihood		-10275.214
Pseudo R ²		0.214

Note: * $p < .05$, ** $p < .01$, *** $p < .001$. Standard errors are reported in parentheses.

〈연구 질문 1〉

저자 국가 → 주제 국가, 비중심국 연구자들은 비중심국을 연구하는가?

카이제곱 검정

- 중심국 연구는 중심국 저자, 비중심국 연구는 비중심국 저자 우세
- 주제 국가가 나타나지 않는 경우는 중심국 저자가 더 많이 연구함

로지스틱 회귀분석

- 비중심국 연구자는 중심국/알수없음에 비해
비중심국 연구 가능성 높음

따라서, 저자 국가가 주제 국가에 미치는 영향을 살펴보면
비중심국 출신이 비중심국을 주제로 더 많이 연구함

사례2: 중심국과 비중심국에 따른 사회학 지식 생산

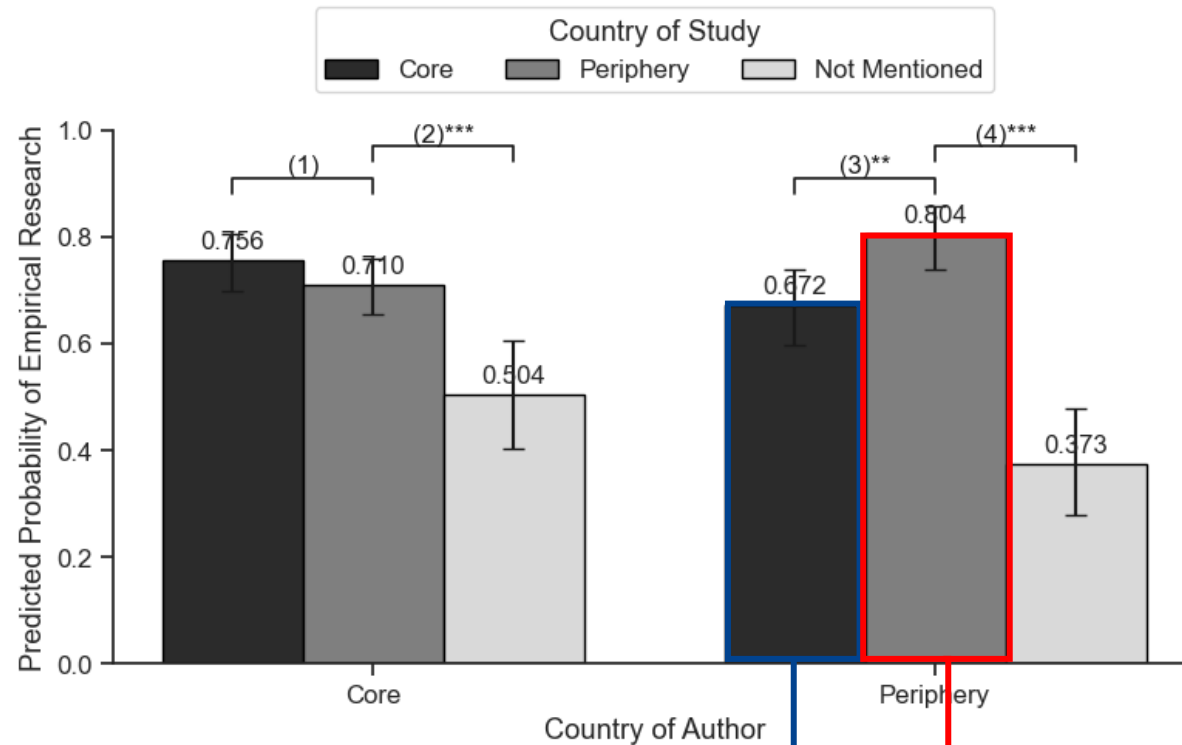


Figure 3. Predicted Probability of Empirical Research by Geographic Variables

〈연구 질문 2〉

저자 국가 X 주제 국가 → 연구 유형

비중심국 저자는 비중심국을 연구할 때 경험 연구를 더 많이 하는가?

상호작용항의 시각화:

저자 국가와 주제 국가별 예측된 경험 연구 가능성

(1) 중심국 저자: 주제 국가에 따라 경험 연구 가능성 차이 없음

(3) 비중심국 저자: 중심국보다 비중심국 연구 시 경험연구 가능성 높음

비중심국 연구자 X 비중심국 주제 연구

비중심국 연구자 X 중심국 주제 연구

사례2: 중심국과 비중심국에 따른 사회학 지식 생산

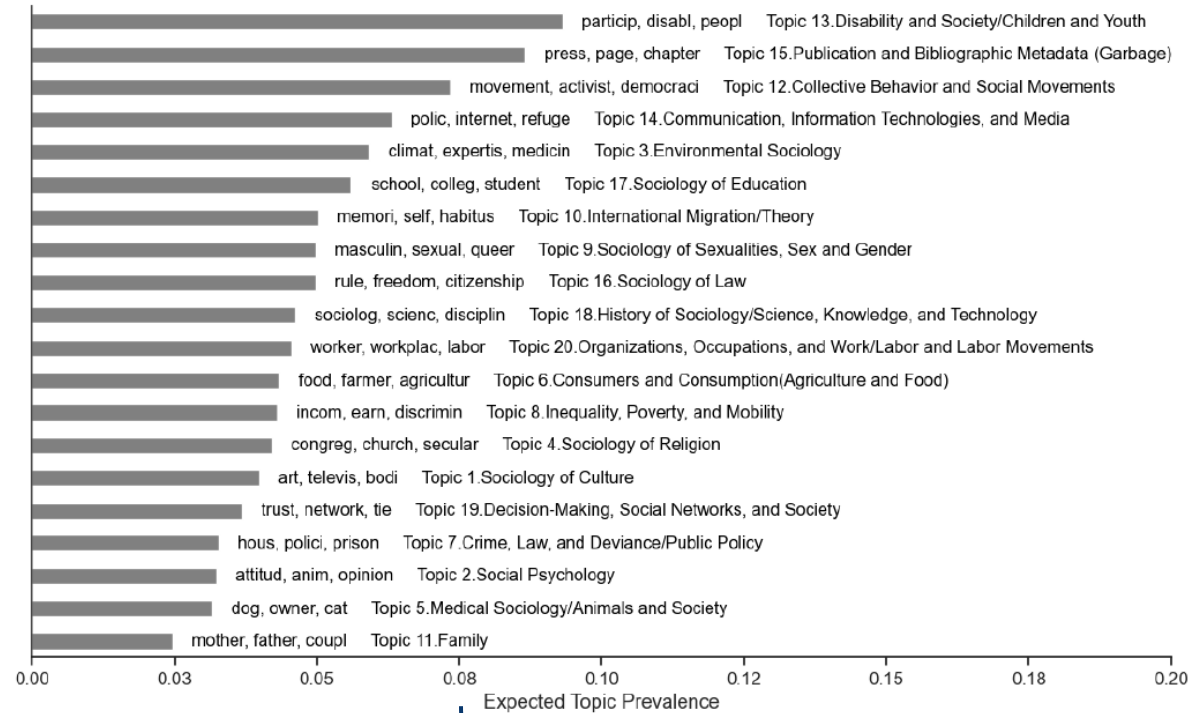
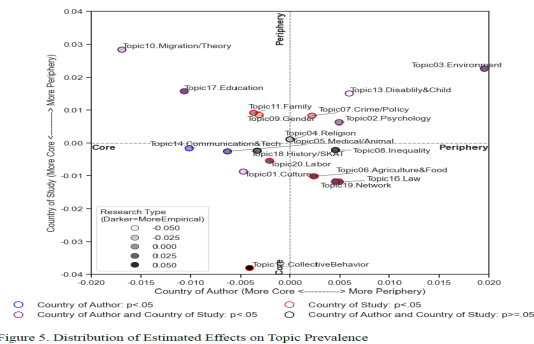


Figure 4. Expected Topic Prevalence of 20 Topics

저자 국가 X 주제 국가



〈연구 질문 3〉

저자 국가 X 주제 국가 → 연구 주제

저자 국가와 주제 국가별로 주로 연구하는 토픽은 무엇인가?

STM을 통해 논문 초록 내 주제를 추출

저자 국가, 주제 국가, 연구 유형을 고려한 분석 결과

- 저자 국가와 주제 국가별로 우세한 토픽을 살펴봄

- (1사분면) 비중심국 저자 X 비중심국 연구:

장애/아동, 심리학, 환경, 범죄/정책

→ 사회적 갈등과 같은 실질적 현안에 주목

- (2사분면) 중심국 저자 X 주변국 연구:

이민/이론, 교육, 가족, 젠더

→ 중심국 저자는 주변국 사회를 노동력 공급원이나 개발 대상, 비서구적 가족 및 젠더 역할 탐구의 장소로 활용

사례2: 중심국과 비중심국에 따른 사회학 지식 생산

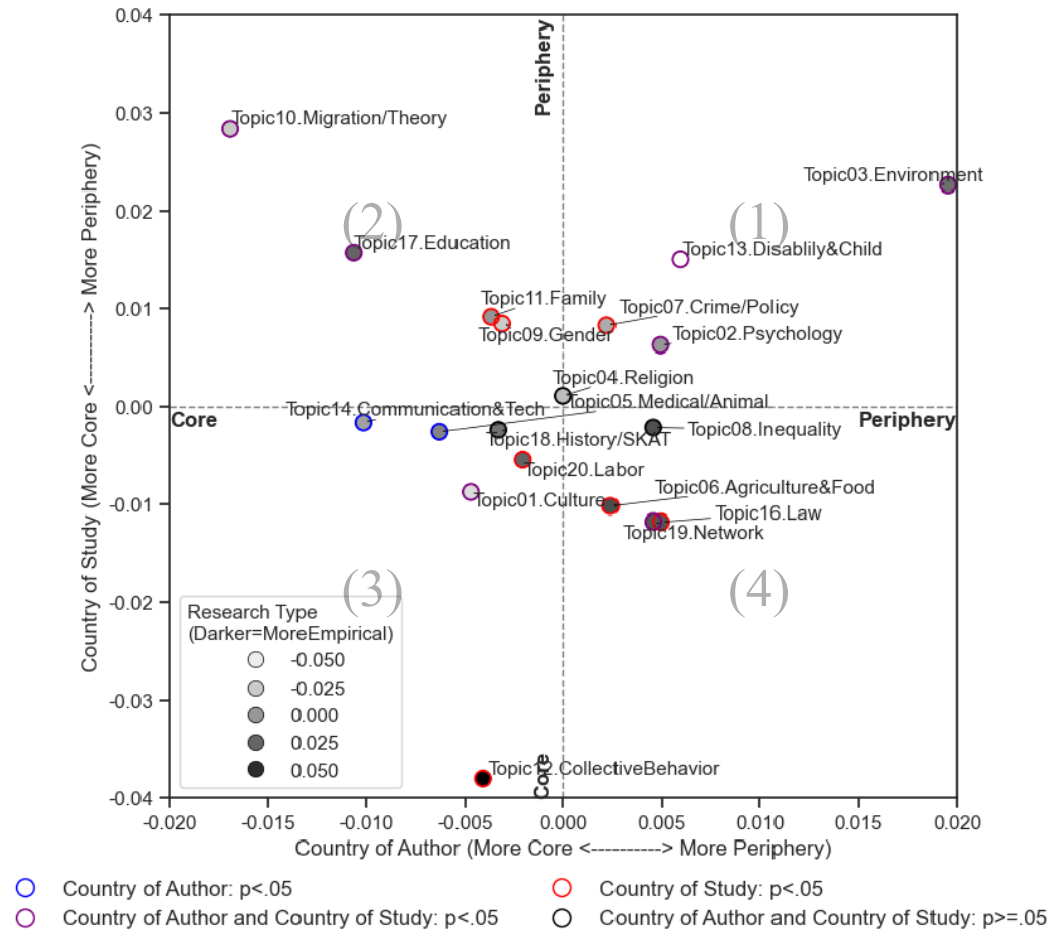


Figure 5. Distribution of Estimated Effects on Topic Prevalence

〈연구 질문 3〉

저자 국가 X 주제 국가 → 연구 주제

저자 국가와 주제 국가별로 주로 연구하는 토픽은 무엇인가?

STM을 통해 논문 초록 내 주제를 추출

저자 국가, 주제 국가, 연구 유형을 고려한 분석 결과

- 저자 국가와 주제 국가별로 우세한 토픽을 살펴봄
- (1사분면) 비중심국 저자 X 비중심국 연구:
장애/아동, 심리학, 환경, 범죄/정책
→ 사회적 갈등과 같은 실질적 현안에 주목
- (2사분면) 중심국 저자 X 주변국 연구:
이민/이론, 교육, 가족, 젠더
→ 중심국 저자는 주변국 사회를 노동력 공급원이나 개발 대상, 비서구적 가족 및 젠더 역할 탐구의 장소로 활용

경청해주셔서 감사합니다.