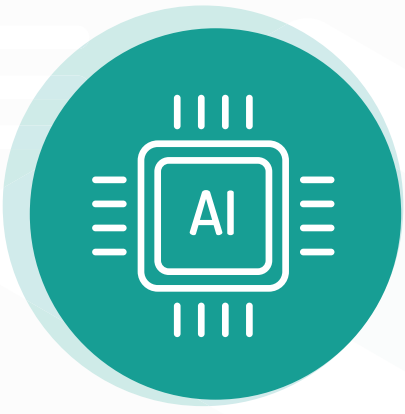
ISSUE REPORT | 2020. 6. 26. IS-098

AI기술의 국가통계 활용 사례 및 국내 도입 촉진 방안

Use Cases of AI Technology in National Statistics & Suggestions for Promoting Domestic Adoption

김정민, 전이슬, 황유림



SPRi 소프트웨어정책연구소
Software Policy & Research Institute

이 보고서는 과학기술정보통신부 정보통신진흥기금 에서 지원받아 제작한 것으로
과학기술정보통신부의 공식의견과 다를 수 있습니다.
이 보고서의 내용은 연구진의 개인 견해이며, 본 보고서와 관련한 의문 사항 또는
수정·보완할 필요가 있는 경우에는 아래 연락처로 연락해 주시기 바랍니다.

소프트웨어정책연구소 데이터·통계연구팀
김정민 선임연구원 jungmink26@spri.kr

요 약 문

국가통계는 사회통합 기능과 정책수립·평가기능을 갖추고 있는 국가의 중요한 공공재로서 통계법을 통해 엄격히 관리되고 있다. 특히 새로운 국가통계를 생산하기 위해서는 반드시 통계청의 통계작성 승인을 득해야 하므로 승인 기준과 품질 척도의 변화는 통계 생산자 모두에게 중요한 사안으로 볼 수 있다.

최근 국내외를 막론하고 AI 및 빅데이터 기술을 국가통계에 활용할 수 있는가에 관한 논의가 본격적으로 진행되는 추세다. 같은 맥락에서 실제 국가통계에 한시적으로 도입해 보거나 테스트하는 사례 또한 증가하고 있다. 이는 결국 AI 및 빅데이터 기술을 통계 생산에 활용할 시 국가통계로 승인 가능한지에 대한 고려로 확대된다. 국내 통계청 또한 올해 상반기 중 조사통계 심사에 맞추어져 있던 승인 심사 문항을 조사 통계 외 생산 방식에도 적용할 수 있도록 일부 개정하는 방안을 추진하고 있으며, 중장기적으로는 국가의 통계 관리 범위를 확대하기 위한 신규 제도 검토를 예고했다. 국가통계 승인 기준의 변화가 가시화되는 시점이다.

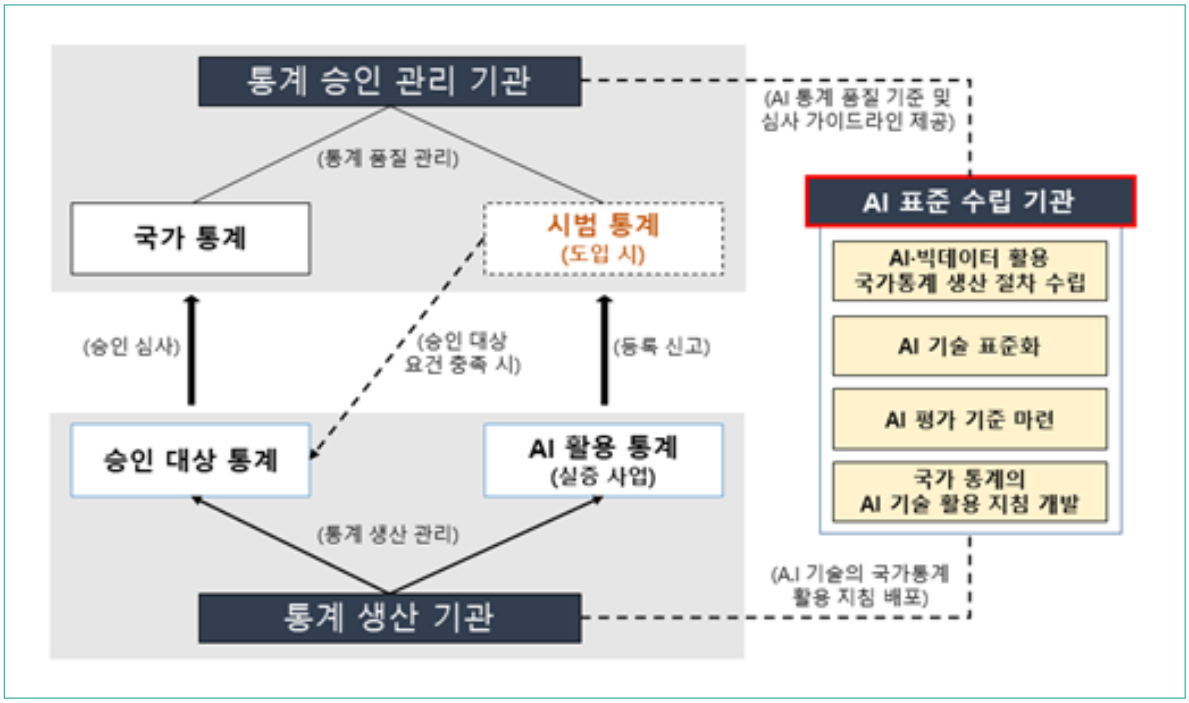
제도적 변화 조짐에도 불구하고 AI 및 빅데이터 기술은 그간 활용했던 통계적 기법과는 방법론의 구조, 평가 기준 등이 상이해 실제 도입에 대한 저항이 존재한다. 그런 연유에서 보고서는 해외의 앞선 논의와 실제 적용 사례를 소개하고 AI기술이 통계 생산에 활용될 시 예상되는 기술적 쟁점이 무엇일지 구체적으로 분석해봄으로써, 도입 촉진에 필요한 요소를 발굴하고 이를 달성하기 위한 방안을 제시하였다.

예상 쟁점 도출을 위해 국가통계에 AI 기술을 도입하는 이슈를 두 가지 시각으로 분리하였다. 조사통계에서 완전히 탈피해 빅데이터에 기반한 통계를 생산하는‘신규 통계 생산’의 시각과 조사통계 생산 과정을 그대로 수용하되 각 생산 과정의 효율성과 성능을 향상시키고자 AI를 도입하는 ‘통계 생산 프로세스의 현대화’가 그것이다. 서로 다른 두 시각에서 바라본 예상 쟁점은 다음과 같이 요약된다.

구 분	예상 쟁점	시사점
신규 통계 생산	AI·빅데이터 기반의 분석결과에 관한 신뢰가 보편적으로 형성되지 않음	신규 통계 생산 시 활용에 적합한 검증된 알고리즘을 공표하고 이에 대한 세부적인 활용 지침 마련이 요구
	신규 통계의 생산방식은 기존의 통계 생산 프레임워크에 맞추어 해석하기에 적합하지 않음	빅데이터 분석 프로세스를 포괄할 수 있는 통계 프로세스 표준의 개정 고려
	재현 불가능한 비결정론적(Non-deterministic) 알고리즘 기반의 결과를 국가통계로 관리 가능한지 여부	인공지능 기술의 특성 분석 및 표준화된 평가 기준 정립이 요구됨
통계 생산 프로세스의 현대화	기존의 방법 대비 우수성을 객관적으로 비교가능한가	전통적 통계 기법-AI 알고리즘의 비교 검증 방안 발굴이 필요
	AI 기술 도입은 통계 생산부터 공표까지의 제한된 생산 기간을 준수할 수 있는 해법인가	다수의 실증 사업을 통한 도입 적합성 진단 필요

분석된 예상 쟁점들과 그에 따른 시사점을 토대로 AI 및 빅데이터 기술의 국가통계 도입이 촉진되기 위해 필요한 두 가지 요소를 도출하였다. 첫 번째로 AI 기술의 대표성 부여를 위한 표준화 된 기술 개발 및 품질 평가 기준 마련을 꼽을 수 있겠으며, 두 번째로는 실증 사업 확대를 통해 국가통계 유형별 도입 적합성을 지속적으로 테스트 해볼 수 있는 장이 마련되어야 한다는 점이다.

끝으로 국가통계에 AI 및 빅데이터 기술 도입을 촉진할 수 있는 실질적 방안을 제시하기 위해, UN 유럽경제위원회「빅데이터 샌드박스」프로젝트의 일환으로 결성된 아일랜드 통계청-기술연구소 간 협력 사례를 참고하였다. 이를 통해 국내 통계청-AI 표준수립기관 간 협력 체계 및 기관별 역할 기능을 제안하였다.



Executive Summary

National statistics are important public goods of the country with functions of social integration and policy establishment·evaluation, and are strictly managed through statistical laws. In particular, in order to produce new national statistics, changes in approval criteria and quality measures can be seen as important for both statistical producers, as new national statistics must be approved by the National Statistical Office.

Recently, discussions on whether AI and Big data technologies can be used for national statistics, both at home and abroad, are underway in earnest. In the same vein, there are also increasing number of cases of temporary adoption or testing of actual national statistics. This eventually extends to the consideration of whether AI and Big data technologies can be approved as national statistics when utilizing them for statistical production. In the first half of this year, the National Statistical Office is also pushing to revise some of the approval screening questions that were tailored to the examination of survey statistics so that they can be applied to production methods other than survey statistics, and in the mid- to long-term, it announced a new system review to expand the scope of the nation's statistical management. This is when changes in the national statistical approval criteria are visible.

Despite signs of institutional change, AI and Big data technologies differ from the statistical techniques used so far, so there is resistance to actual adoption. For such a reason, the report introduced advanced discussions and actual application cases abroad and specifically analyzed what technological issues would be expected when AI technology is used for statistical production, thus exploring the necessary factors for promoting the adoption and presenting measures to achieve them.

The issue of adopting AI technology to national statistics has been divided into two perspectives to draw expected issues. They include the view of 'new statistical production' that produces statistics based on big data, completely deviating from survey statistics, and 'modernization of statistical production processes' that adopts AI to accommodate the process of producing survey statistics but to improve the efficiency and performance of each production process. The expected issues from two different perspectives are summarized as follows.

Categorization	Expected issues	Implications
New statistical production	Trust in analysis results based on AI and Big data is not universally formed	When producing new statistics, it is necessary to publish proven algorithms that are suitable for utilization and to provide detailed utilization guidelines.
	The production method of new statistics is not suitable to be interpreted in accordance with the existing statistical production framework	Consider amending statistical process standards that can encompass the Big data analysis process
	Whether results based on non-deterministic algorithms that cannot be reproducible are manageable as national statistics	Characteristic analysis of artificial intelligence technology and establishment of standardized evaluation criteria are required
Modernization of statistical production processes	Is it possible to objectively compare excellence compared to existing methods?	Discovering methods of comparison verification of Traditional statistical techniques-AI algorithms is required
	Is the adoption of AI technology a solution to comply with the limited production period from statistical production to publication?	Need to diagnose suitability of adoption through a number of demonstration projects

Based on the expected issues analyzed and their implications, two factors were derived to facilitate the adoption of national statistics for AI and Big data technology. First, standardized technology development and quality evaluation criteria for the representativeness of AI technology should be considered, and second, there should be a place to continuously test the suitability of adoption by type of national statistics by expanding demonstration business.

Finally, to present practical measures to promote the adoption of AI and Big data technologies to national statistics, we referred to examples of cooperation between the National Statistical Office and the Institute of Technology in Ireland, which was formed as part of the UN Economic Council's Big Data Sandbox project. Through this, the cooperative system and role functions of each institution were proposed between the Korean National Statistical Office and AI Standard Establishment Agency.

CONTENT

CONTENT

1. 서론

□ **국가통계(National statistics)란,** 국가통계 제도를 통해 배포되는 통계를 의미 [1)]

※ 비공식 통계임을 배포과정에 명시한 경우는 국가통계 범위에서 제외

○ 국가통계는 사회통합 기능과 정책수립·평가기능을 갖추고 있어 [2)] 국가에 매우 중요한 가치를 가지는 공공재임

○ 중요한 자산인 만큼 이에 상응해 요구되는 높은 수준의 품질 척도가 존재하며 정기적인 품질진단을 통해 관리되고 있는 상황

□ **국내 통계법 상 국가통계는,** 통계작성지정기관에 의해 생산되는 통계작성 승인을 획득한 통계를 의미 [3)]

○ 국가통계는 자료수집의 방법에 따라 조사통계(설문조사 기반), 보고통계(행정자료 기반), 가공통계(1차통계 기반)으로 구분

○ 미승인 통계의 공표는 통계법에 저촉되므로, 국가통계로의 승인 여부는 새롭게 작성된 통계의 활용도와 존폐여부를 결정하는 주요한 사안임

□ **최근 국내외적으로 AI 및 빅데이터 기술의 국가통계 활용이 논의됨에 따라 국내 국가통계 승인 관련 사항들에 변화의 기류가 감지**

○ '20년 5월 통계청은 심사기준 개정, 시범통계제도 도입 등을 통해 빅데이터 활용 통계를 국가 관리체계 내 포괄하기 위한 실질적 토대 구축을 예고 [4)]

○ 해당 변화는 AI·빅데이터 등 SW기술과 통계 분야 간 융합을 촉진시키는 계기로 작용될 수 있어, 향후 중요성이 더 강조될 것이라 예상 가능함

그런 이유에서 본 보고서는 AI 및 빅데이터 기술의 국가통계 활용에 관한
해외 국가의 선제적 논의들과 도입 사례를 소개하고,
도입 시 예상되는 쟁점 및 도입 활성화를 위한 역할 체계를 제안하였음

1) OECD statistics 용어사전의 "official statistics" 참조
2) 통계교육원 (2015), "국가통계의 이해"참고
3) 통계법 제3장 제15조(통계작성지정기관의 지정), 제17조(지정통계의 지정 및 지정취소), 제4장 제1절 제18조(통계작성의 승인) 부분 참고
4) 통계청 (2020), 「빅데이터 활용 통계」 등 통계 다양성 확대를 위한 국가통계 승인기준 보완방안(안)

2. 국가통계의 AI 기술 도입 논의와 적용 사례

(1) AI·빅데이터 통계의 도입 논의

> *"Just as haute cuisine must incessantly reinvent itself in order to stay at the forefront of gastronomy, official statistics is also confronted with a rapidly changing context and needs."*
>
> ***"최고급 식당이 미식계의 선두 자리를 계속 유지하기 위해서는 끊임없는 재창조가 필요한 것처럼, 공식 통계 [5] 역시 급속하게 변화하는 상황과 요구사항을 정면으로 마주해야 한다."***
>
> **Walter J. Radermacher, 前 Eurostat 통계청장, 2018**

□ 유럽 국가들을 중심으로 국가통계의 품질 및 조사환경 개선에 대한 공감대가 형성됨에 따라, **이를 해결할 방안으로서 빅데이터 및 인공지능 기술 도입이 추진 및 검토**되고 있음

○ 국가통계 개선이 이슈로 부각된 주요 원인들로 아래와 같은 현상이 거론

① 새로운 유형의 산업이 빠르게 등장함에 따른 산업구조 반영의 어려움
② 기업 활동의 다양화(지역에 기반 하지 않은 산업, 사무실 연락처 부재 등)
③ 조사별 응답자 중복에 의한 회수율 악화
④ 즉시 활용 가능한 민간 중심의 데이터의 급증 등

○ 공통적인 문제 인식과는 별개로 AI 또는 빅데이터 기술 도입에 대한 가치 판단 및 취지는 국가별로 상이하며, 사안에 접근하는 방식 또한 다양

□ 관련 연구들은 UN유럽경제위원회(UNECE)의 국가통계 현대화 논의[6] 가 본격적으로 진행된 2014년 이후 각 국가별로 꾸준히 추진되어 왔음

※ 참고 : 아일랜드는 2014년 UNECE의 국가통계 현대화 일환 프로젝트에서 주도적 역할을 함으로써 타 국가대비 초기 이슈를 선점

○ **(아일랜드:도입 유형 및 기회요인 분석)** 빅데이터 기술을 통계에 활용하는 4가지 유형을 제시*하였으며, 관련 된 기회 및 위협을 분석

5) 일반적으로 공식통계(Official Statistics)와 국가통계(National Statistics)는 동일 의미로 혼용됨
6) UNECE에서는 정기적으로 고위그룹(High level group, HLG) 차원의 국가통계의 현대화 논의가 이루어지며, 그 일환으로 빅데이터 활용 국가통계 프로젝트가 추진되어 왔음

* ① 기존 통계의 완전/부분 대체, ② 추가적인 통계 정보 제공, ③ 통계 추정치의 개선, ④ 완전히 새로운 통계 정보 제공

< 표 1 > 빅데이터를 국가통계에 활용함에 따른 기회 요소(아일랜드)

범주(Category)	기회 요인(Opportunities)
데이터	- 보완, 대체, 개선 및 기존의 데이터에 새로운 요소를 추가 - 서로 다른 데이터 간 연결성 개선 - 전산사회과학, 데이터 과학 및 데이터 산업들 간의 협업 강화
품질/기능	- 시의성 있는 결과 생산 - 품질 및 진실성 향상 - 쉬운 관할(혹은 지역)간 교차 비교 - 새롭고 더 나은 통찰력을 가져다 줄 신규 데이터 분석 - 마이크로 레벨 및 미시 분석으로 확장 및 보완 - 기존 통계의 구성을 재편
생산성/효율성	- 시민과 기업의 설문 응답 피로도 개선 - 업무 최적화 및 생산성 향상 - 비용의 절감 - 고 부가가치 업무에 직원을 재배치 - 국가 통계의 가시성 및 활용 강화

* 참고 : The opportunities, challenges and risks of big data for official statistics(IOS, 2015)

○ **(스위스:도입 필요성을 이론적으로 분석)** 스위스 통계청은 국가통계의 효과적 정책 지원을 위해 조사 통계와 데이터 과학에 기반한 통계를 병행 생산해야 함을 제언

- 의사결정 과정에서 연역적 추론과 귀납적 추론 결과가 함께 고려되어야 효과적임을 근거로 데이터 과학 기반의 통계 필요성을 역설

< 표 2 > 일반 통계와 데이터 과학의 추론 방식에 따른 구분(스위스)

통계 구분	목 적	추론의 방식
전통적 의미의 통계	이론적 개념을 운용해 기존 개념(가설)의 유효성을 설명하고 확인하기 위한 기초 자료	연역적(가설 우선)
데이터 과학	통제 또는 조사자의 감독 없이 새로운 개념(가설 또는 이론)을 개발하기 위해 수집 및 설계	귀납적(데이터 우선)

* 출처 : DGINS(2018)

○ **(독일:프로세스 혁신을 논의)** 독일 연방통계청은 머신러닝 기술을 기존 통계 생산 프로세스에 도입하는데 주안점을 두고 연구 수행

- 2017년 국가차원의 디지털 어젠다(Agenda) 59건 중 선도 프로젝트로 '국가통계의 머신러닝 기술 검증'이 채택되어 관련 연구가 활발히 진행
- 국가통계의 자동화 및 기업 통계 영역에서의 AI 도입이 목표

< 그림 2 > 머신러닝 도입을 통한 기대효과

* 출처 : Machine Learning im Statistischen Bundesamt(독일 연방 통계청, 2019)

○ **(네덜란드:빅데이터의 실질적 활용 방법)** 네덜란드는 빅데이터의 장점을 열거하면서도, 빅데이터를 기존 통계 방법론으로 해석하려는 시각을 견지

- 새로운 방법론의 활용이 사용자에게 투명하게 공개되는 한, 국가통계청은 신규 모델의 사용을 두려워하지 않아야 함을 제언

< 표 3 > 국가통계에서의 빅데이터 활용 방안(네덜란드)

활용 구분	해 석
빅데이터의 불완전성을 그대로 수용	- 빅데이터는 연속성·비교가능성을 유지하기 어렵고 모집단의 범위가 실시간으로 변화할 수 있어 시계열 설명이 불가능하고 비약이 발생될 여지 - 반면 빅데이터는 존재하는 사실 자체로 사회의 흥미를 유발하는 효과
통계적 모델링을 통한 데이터 정형화	- 최근 수리통계학 및 응용통계학자간 빅데이터를 통계적으로 해석하기 위한 다양한 방법이 개발되었음 - 기계학습 기술과 같은 신규 방법론은 전통적 통계 기법과 함께 고려가능

* 출처 : 네덜란드 통계청(2020)

○ **(일본:先실험 後평가)** 일본은 국가통계에 빅데이터를 활용 시 시의성·적시성 등에서 이점이 있는 반면 정확성·편중성 측면에서 우려를 제기

- 2019년 상업통계를 민간 부문 빅데이터를 활용해 대체하는 실험을 추진
  * 해당 사례는 UNECE가 권장하는 민간 POS데이터 활용 통계를 벤치마크

< 표 4 > 민간 부문 빅데이터 활용을 통한 기존 통계 대체가 주는 장단점(일본)

장 점	단 점
- 통계 공표의 가속화 - 집계 빈도를 높일 수 있음 - 품목 및 행위 기반 데이터를 활용하므로 표준 분류체계에 비해 유연한 방식으로 집계가 가능 - 설문 참가자의 부담이 감소 - 통계작업 전반의 효율성 향상	- 정확성과 편중성(Bias)을 제어하기 어려움 - 기업의 합병 및 파산 등의 가능성으로 인해 데이터의 지속적인 가용성을 보장하기 어려움

* 출처 : Can Big Data Change Official Statistics(RIETI, 2019)

□ 이처럼 다양한 국가에서 도입을 위한 논의가 이루어지는 것과 더불어 실증·시범 사업 등 세계 각 국의 발 빠른 실험이 병행되고 있는 상황

○ 이와 관련된 글로벌 단위의 조사로는 독일 연방 통계청 주도로 수행한 **「통계 기관에서의 머신 러닝 사용에 관한 조사」**[7] 가 유일

○ 상기 조사 대상 국가는 아래와 같으며 본고는 조사 결과 중 도입 및 실증 단계에 근접한 유의미한 적용 사례를 선별 후 분석하였음

- 독일 : 독일 내 14개 주 통계청 및 18개 통계 생산 기관
- 해외 : EU 27개 회원국, EFTA 4개국, 비EU 6개국 등

< 표 5 > 통계 기관에서의 머신러닝 사용에 관한 조사의 대상국(독일 제외)

구 분	대상 국가
유럽연합(EU)	오스트리아, 벨기에, 캐나다, 크로아티아, 키프로스, 체코, 덴마크, 에스토니아, 핀란드, 프랑스, 그리스, 헝가리, 아일랜드, 이탈리아, 라트비아, 리투아니아, 룩셈부르크, 몰타, 네덜란드, 노르웨이, 폴란드, 포르투갈, 루마니아, 슬로바키아, 슬로베니아, 스페인, 스웨덴, 영국
유럽자유무역연합(EFTA)	아이슬란드, 노르웨이, 스위스, 리히텐슈타인
비EU 국가	호주, 캐나다, 이스라엘, 일본, 뉴질랜드, 미국
기타	유럽연합통계청(Eurostat), OECD

7) 독일 연방 통계청 (2018), Survey on the use of machine learning in statistical institutions

(2) 해외 각국의 AI·빅데이터 통계 주요 도입 사례

□ (미국) 정형·비정형 설문 응답에 대한 코드 분류를 위해 도입

활용 기관	BLS(노동통계국)	도입 현황	도입
활용 구분	**업무상 상해 및 질병 설문조사의 근로자 상해 응답에 대한 자동 코드 분류**		
세부 내용	- **(업무 개요)** 업무상 상해 및 질병 설문조사를 통해 수집되는 서술형 답변은 매년 수만 건으로, 수천 개의 상해 및 질병 특성 코드 중 상응하는 6개 코드와 매핑이 필요 - **(도입 방법)** 설문 응답을 분석해 각 응답에 가장 a적합한 코드 6개를 추천해줌으로써 자동 분류하는 방식		
활용 기술	정규화 된 로지스틱 회귀(~2018) →딥 러닝(2019~)	도입 목적	생산성, 정확성

활용 기관	Census(통계청)	도입 현황	테스트 단계
활용 구분	**미국 지역사회 설문조사의 산업 및 직업에 대한 코드 분류**		
세부 내용	- **(업무 개요)** 미국 지역사회 설문조사(ACS)는 매년 약 250만명의 개인 산업 및 직업과 관련 된 표준화되지 않은 응답이 수집되며, 이를 산업·직업 분류 코드와 매핑해야 함 - **(도입 방법)** 설문상의 산업 및 직업 관련 응답을 분석해 이에 상응하는 산업·직업 분류 코드로 자동 분류		
활용 기술	로지스틱 회귀	도입 목적	생산성, 정확성

□ (독일) 분류, 데이터 연계, 보조 데이터 확보, 이상치 탐지 등 다방면의 테스트가 진행되고 있으며, 분류와 관련된 업무 일부는 실제 도입

활용 기관	연방 통계청	도입 현황	도입
활용 구분	**국내 법인에 대한 유럽국민계정체계(ESA) 상 기관 분류 할당**		
세부 내용	- **(업무 개요)** 유럽 연합(EU)에 소속된 국가는 유럽국민계정체계가 활용하는 기관 분류에 맞추어 자국 법인을 분류해야 함 - **(도입 방법)** 법인의 회계 활동 정보에 기반해 유형을 파악하고 기관 분류		
활용 기술	서포트 벡터 머신(SVM)	도입 목적	생산성, 연결성

활용 기관	연방 통계청	도입 현황	도입
활용 구분	**공예 분야 행정 통계의 집계 대상 분류**		
세부 내용	- **(업무 개요)** 독일 수공예 분야 통계는 납품 과정에서 발생하는 행정 데이터를 통해 생산되나, 거래 과정에 명시된 모든 기업이 수공예 분야 기업이 아니므로 관계없는 기업을 수기로 찾아 제외시켜야 하는 어려움 존재 - **(도입 방법)** 행정 데이터에 등장하는 모든 법인 정보를 토대로, 제외 후보군을 분류함으로써 과업에 필요한 인력을 축소		
활용 기술	랜덤포레스트, 서포트 벡터 머신	도입 목적	정확성, 비용효율

활용 기관	연방 통계청	도입 현황	테스트 단계
활용 구분	**데이터 연계를 통한 연방 고용청의 패널 데이터 보강**		
세부 내용	- **(업무 개요)** 통계청의 소득 구조 조사(SES)는 시간당 노동자의 임금 정보가 조사되므로 최저 임금에 따른 소득의 영향 정보 도출이 가능한 반면, 연방고용청에서 제공하는 데이터인 통합고용이력(IEB)에는 소득 정보가 누락 - **(도입 방법)** SES 데이터에 기반해 학습한 머신러닝 모델을 IEB에 적용함으로써 데이터를 연계해 최저 임금 변화에 대한 소득 영향 정보를 IEB에 추가		
활용 기술	랜덤 포레스트	도입 목적	연결성

활용 기관	연방 통계청	도입 현황	테스트 단계
활용 구분	**온라인 구인 공고 분류**		
세부 내용	- **(업무 개요)** 기업이 제시하는 임금, 구인 인력의 교육 수준, 채용 담당자의 구인 형태 등 국가 정책 수립에 필요한 정보가 부족한 경우가 많음 - **(도입 방법)** 온라인 구인공고를 웹 스크래핑 후 텍스트를 분석함으로써 구인공고의 특성을 분류하고 관련 정보를 추출		
활용 기술	K 최근접 이웃 클러스터링, 다항분포 나이브 베이즈	도입 목적	부가정보, 인사이트

활용 기관	연방 통계청	도입 현황	테스트 단계
활용 구분	**기업 구조 통계 응답결과의 이상값 식별**		
세부 내용	- **(업무 개요)** 기업 통계 조사의 응답결과에는 타당성이 부족하거나 극단적인 이상치가 포함되어 있어 검토 과정에 많은 시간이 소요 - **(도입 방법)** 검토 대상을 줄이기 위한 방법으로서 머신러닝을 활용해 검토 대상을 추천		
활용 기술	격리 포레스트(Isolation forest)	도입 목적	생산성, 정확성

□ **(캐나다) 분류, 이상 값 검출 등을 위해 도입되었으며, 결측 값 대체를 위한 분류작업에 비지도 학습을 활용한 게 특징**

활용 기관	연방 통계청	**도입 현황**	도입
활용 구분	**월별 소매 거래 조사 및 분기별 소매 상품 조사에 스캔 데이터 활용**		
세부 내용	- **(업무 개요)** 월별 소매 거래 조사는 북미 상품 분류 시스템(NAPCS)의 상품 코드에 기준해 상품 코드별 매출 현황을 통계화 하는 작업 - **(도입 방법)** 민간 기업 스캐너 데이터로 수집되는 상품 설명을 텍스트 분석하여 NAPCS 코드와 매핑		
활용 기술	XGBoost(분산 그라디언트 부스팅), BoW(Bag-of-Words), n-gram 모델	**도입 목적**	기존 통계 대체

활용 기관	연방 통계청	**도입 현황**	도입
활용 구분	**국제 무역 데이터 이상 값 검출**		
세부 내용	- **(업무 개요)** 국제 무역 통계의 데이터 중 단위의 오류, 0과 1로 표기된 이상 값을 처리하는 작업을 위해 수기 검토와 변경 승인 프로세스를 운영 - **(도입 방법)** 머신러닝을 통해 검토 프로세스를 자동화		
활용 기술	XGBoost 트리 모델	**도입 목적**	생산성, 정확성

활용 기관	연방 통계청	**도입 현황**	도입
활용 구분	**인구 조사를 구성하는 이민 허가 변수에 대한 결측 값 대체**		
세부 내용	- **(업무 개요)** 2016년 인구 조사에 이민 허가와 관련된 데이터가 추가 변수로 지정되었으며, 이는 추가 조사가 아닌 별도 데이터와의 연계를 통해 확보되었음. 일부 개인의 경우에는 데이터가 없어 결측 값이 발생하였으며 이를 대체해야 하는 업무가 발생 - **(도입 방법)** 유사 특성을 보이는 조사 데이터를 학습하여 결측 데이터 항목의 대체 값을 추정		
활용 기술	Relief 알고리즘, K 최근접 이웃 클러스터링	**도입 목적**	기존 통계 대체

활용 기관	연방 통계청	**도입 현황**	도입(한시적)
활용 구분	**인구 조사 신규 컨텐츠 수요 분석**		
세부 내용	- **(업무 개요)** 국2016년 조사된 인구 조사의 신규 콘텐츠 설문결과 약 110만 건에 대한 활용처 모색 - **(도입 방법)** 설문결과 텍스트의 맥락 및 주제를 해석하여 요약		
활용 기술	자연어 처리 기술	**도입 목적**	인사이트

□ (네덜란드) 웹 수집 데이터 분석 및 데이터 연계 목적으로 활용

활용 기관	통계청	도입 현황	도입
활용 구분	**인구 조사를 구성하는 이민 허가 변수에 대한 결측 값 대체**		
세부 내용	- **(업무 개요)** 유럽 통계 시스템(ESS) 내 빅데이터 프로젝트의 일환으로서, 기업의 인터넷 정보, 활동, 주소 정보, 소유권 구조 등과 같은 기존 정보를 개선하거나 업데이트하기 위해 추진 [8] - **(도입 방법)** 구글 맞춤형 검색 시스템, 기업 URL 및 연계된 행정데이터를 활용해 대상 기업과 관련한 웹 정보를 수집한 후, 텍스트 마이닝 및 추론 기술을 통해 기업 정보를 정제		
활용 기술	의사결정 트리, 랜덤 포레스트 등	도입 목적	부가정보

활용 기관	통계청	도입 현황	도입
활용 구분	**보건 실태조사를 대상으로 한 혼합형 설문조사의 무응답 대체 기법 보강**		
세부 내용	- **(업무 개요)** 네덜란드는 설문조사의 비용절감 및 회수율 악화를 해결하기 위한 대책을 지속적으로 연구해왔으며, 해결책의 일환으로 혼합형 설문 조사(mixed mode survey) [9]를 추진해왔음 - **(도입 방법)** 실태조사의 실사 이후 발생하는 무응답을 타 조사결과를 통해 대체하는 과정에서 발생 가능한 각종 분류문제(층화, 할당 등)를 머신러닝 기법을 혼용해 해결		
활용 기술	분류 트리	도입 목적	연결성, 정확성

활용 기관	통계청	도입 현황	시범 도입
활용 구분	**인터넷 구매를 통한 자국 소비자의 국경 간 거래 현황 예측**		
세부 내용	- **(업무 개요)** 유럽연합에 소속된 국가들 사이의 인터넷 구매는 설문조사의 샘플링 및 조사 대상의 언어 문제 등으로 인해 거래 총액을 정확하게 측정하거나 추정하기 어려움 - **(도입 방법)** 유럽연합 소속 법인의 세금 정보와 인터넷 데이터를 연계해 국경 간 인터넷 거래를 추정 [10]		
활용 기술	비선형 SVM(RbfSVC), 랜덤 포레스트	도입 목적	부가정보

8) Eurostat (2018), ESSnet Big Data Specific Grant Agreement No 1(SGA-2)

9) 조사 결과 데이터를 확보하는 과정에서 2개 이상의 조사 방법을 채택하는 기법을 의미하며, 상기 사례에서는 1차 실사 결과에서 비롯된 무응답 샘플의 대체에 다른 조사 방식으로 도출된 결과를 활용함으로써 모집단 편향성을 해소하는 접근을 추진

10) Q.A.Meertens, C.G.H.Diks (2018), A Data-Driven Supply-Side Approach for Measuring Cross-Border Internet Purchases

활용 기관	통계청	도입 현황	테스트 단계
활용 구분	**기업 통계 무응답 대체 기법의 개선**		
세부 내용	- **(업무 개요)** 기업 통계를 위한 무응답 대체 기법의 정확성 문제가 지적되었으며, 신뢰도 높은 추정을 위한 개선이 요구되고 있음 - **(도입 방법)** 머신 러닝 기법을 사용해 추정치의 신뢰도 향상 및 자동 무응답 대체를 위한 모델을 구축		
활용 기술	그래디언트 부스팅 머신(GBM)	도입 목적	정확성, 생산성

□ (스위스) 파라데이터(Paradata) [11] 머신러닝 분석 결과를 회수율 제고에 활용

활용 기관	연방 통계청	도입 현황	도입
활용 구분	**조사 무응답 매커니즘의 모델링**		
세부 내용	- **(업무 개요)** 설문조사의 무응답 사례를 최소화하기 위한 노력 - **(도입 방법)** 무응답자의 행위를 모델링하여 유형을 분류하고 실사에 참고		
활용 기술	카이-제곱 자동 상호 작용 탐지(CHAID, 의사결정트리의 종류 중 하나)	도입 목적	생산성, 비용효율

활용 기관	연방 통계청	도입 현황	테스트 단계
활용 구분	**이상치로 의심되는 응답 결과 추천**		
세부 내용	- **(업무 개요)** 수정 규칙이 없는 조사 항목들에 대한 이상치를 검토 - **(도입 방법)** 머신러닝을 통해 잘못된 응답으로 의심되는 케이스를 감지하고 이를 검토자에게 전달하여 검토 시간을 단축		
활용 기술	부스티드 결정 트리, 랜덤 포레스트, 신경망, 나이브 베이즈 등	도입 목적	생산성

활용 기관	연방통계청	도입 현황	테스트 단계
활용 구분	**항공 이미지에 기반한 토지 피복 및 토지 사용 코드 분류 자동화**		
세부 내용	- **(업무 개요)** 지표면의 물리적 상태를 인공위성 촬영 이미지 등을 이용하여 분류 항목별 코드로 구분하는 작업 - **(도입 방법)** 머신러닝을 통해 항공 이미지를 분석하고 각 지형에 상응하는 코드와 자동 매핑		
활용 기술	합성곱 신경망(CNN)	도입 목적	기존 통계 대체

11) 조사과정자료(paradata, process data)는 조사를 진행하는 과정에서 자연적으로 발생하게 되는 파생 자료이며 조사 관리를 위한 보조 자료의 성격을 가짐(임경은 2012)

□ (오스트리아) 특정 통계 조사가 아닌 범용적인 도입에 주안점

활용 기관	통계청	도입 현황	도입
활용 구분	**설문 무응답 대체 및 데이터 통합**		
세부 내용	- **(업무 개요)** 행정 데이터를 포함하는 설문조사 및 프로젝트의 결측값 대체 이슈 및 공통 변수가 존재하는 데이터 셋의 연계 - **(도입 방법)** 군집화 알고리즘을 통해 결측 값을 대체하고 분류 기법을 활용해 데이터 셋 연계		
활용 기술	K 최근접 이웃 클러스터링(결측 값 대체), 랜덤 포레스트(데이터 연계)	도입 목적	정확성, 연결성

활용 기관	통계청	도입 현황	도입(한시적)
활용 구분	**행정 데이터와 SILC 융합을 통한 ICT 조사의 소득 문항 대체**		
세부 내용	- **(업무 개요)** 소득 및 생활 조건 통계(SILC)와 ICT관련 조사에 가계소득을 묻는 문항이 포함되어 있으나 조사 대상이 중복 - **(도입 방법)** 행정자료 및 SILC에 기반해 ICT조사 대상의 가계소득을 추정하는데 머신러닝을 활용하고 ICT조사의 소득 항목 제외		
활용 기술	랜덤 포레스트	도입 목적	연결성

□ (루마니아) 데이터 결합 정확도 향상을 통해 표본의 질을 개선

활용 기관	통계청	도입 현황	도입
활용 구분	**EU-SILC 등록 통계 표본의 질 개선**		
세부 내용	- **(업무 개요)** 유럽연합의 소득 및 생활조건에 관한 연합 통계(EU-SILC)에 등록할 통계는 기타 표본조사 과정에서 수집된 정보와 행정 정보를 결합해 모집단 및 표본 추출이 수행되어야함 - **(도입 방법)** 기타 표본조사 파생 정보와 행정 데이터 간 결합 과정에서 머신러닝을 활용		
활용 기술	랜덤 포레스트	도입 목적	연결성, 정확성

□ (핀란드) 비정형 텍스트 데이터를 토대로 특성을 분류하기 위해 활용

활용 기관	통계청	도입 현황	도입
활용 구분	**경찰청 교통사고 보고 문건에 기반해 사고 내역 자동 분류**		
세부 내용	- **(업무 개요)** 자유 형식으로 작성된 도로 교통사고 보고서에 기반해 상해 유발 사고와 그 외 사고를 분류하는 작업 - **(도입 방법)** 텍스트 마이닝 및 분류 기법을 토대로 비정형 텍스트 기반의 보고서 내용을 분류		
활용 기술	TF-IDF, 로지스틱 회귀모형	도입 목적	생산성

□ (벨기에) 일자리 분석을 위한 코드 분류에 활용

활용 기관	통계청	도입 현황	테스트 단계
활용 구분	**구인공고별 알맞은 경제 활동 통계 분류 코드 예측**		
세부 내용	- **(업무 개요)** 일자리 구인 현황에 대한 분석 수요 - **(도입 방법)** 벨기에 일자리 포털에 등록된 구인공고 상 직무 정보를 수집해 유로연합 경제 활동 통계 분류(NACE) 코드와 매핑하여 통계 생산		
활용 기술	서포트 벡터 머신	도입 목적	부가정보

□ (룩셈부르크) 설문 응답 값 검증 및 분류 부문에 도입

활용 기관	통계청	도입 현황	도입
활용 구분	**민간 기업의 스캐너 데이터를 활용해 상품 분류 코드 추천**		
세부 내용	- **(업무 개요)** 소비자 물가 지수의 소비 기준으로 활용되는 COICOP(목적에 따른 개별 소비 분류) 코드와 실제 소비 현황을 매핑 해야함 - **(도입 방법)** 캐나다의 「월별 소매 거래 조사 및 분기별 소매 상품 조사에 스캔 데이터 활용」 부문과 유사		
활용 기술	다중 선형 회귀(MLR)	도입 목적	기존 통계 대체

활용 기관	통계청	도입 현황	도입
활용 구분	**근로 환경 조사 데이터 검증을 위해 머신러닝 도입**		
세부 내용	- **(업무 개요)** 근로 환경 조사는 개인의 사내 R&D활동 수준을 측정하나, 응답 값에 대한 검증이 필요하다는 지적 - **(도입 방법)** 현재·과거 설문조사 및 행정데이터를 결합해 조사 응답자가 조사 기간 내 R&D활동을 수행했을 확률을 추정하여 조사 결과 검증		
활용 기술	앙상블 모델링(Model Stacking)	도입 목적	정확성, 연결성

3. AI·빅데이터 통계 도입 시 예상 쟁점

□ 국가통계 부문의 AI·빅데이터 기술 도입 목적은 **(1) 신규 통계의 생산, (2) 통계 생산 프로세스의 현대화** 총 두 가지 시각으로 구분 가능

○ 두 시각은 논의의 출발점이 다소 상이해 목적을 달성하기 위한 국가별 정책 및 제도에 차이가 발생할 것임을 추정해볼 수 있음

< 표 6 > 국가통계의 AI·빅데이터 기술 도입의 두 가지 시각

구 분	세부 적용 분야
신규 통계 생산	• **대체(Replace) :** 기존 통계를 완전 또는 부분 대체하는 수단 • **부가(Addition) :** 조사 통계 방식으로 도출이 어려운 부문에 대해 데이터 분석에 기반을 둔 우회적 해결 • **인사이트(Insight) :** 완전히 새로운 분야의 통계 데이터 제시
통계 생산 프로세스의 현대화	• **정확성(Accuracy) :** 통계의 정확도 향상 • **생산성(Productivity) :** 업무 효율화를 위한 자동화 기법 • **연결성(Connectivity) :** 연결 가능한 데이터들의 상호 보완적 활용 • **비용(Price) :** 통계 생산의 비용 절감을 위해 활용

○ 3장에서는 AI·빅데이터 기술이 국가통계 제도권 내에 안착하는데 고려되어야 할 **예상 쟁점 사항**들을 상기 시각으로 구분하여 분석하였음

(1) 신규 통계 생산

□ **(조작적 정의)** 본고에서의 **신규 통계**란 조사통계 기법을 활용하지 않고 빅데이터·AI 기술을 활용해 생산한 통계(국가통계)로 정의

○ **빅데이터(민간, 행정)를 활용한 기초 통계** : 국가 현황 파악에 기초가 되는 1차 통계로서, 통계 생산을 위해 민간 빅데이터 및 행정 통계를 활용

* 완전히 새로운 통계를 발굴하는 것과 더불어 이미 존재하는 조사통계 방식의 기초 통계를 빅데이터·AI기술을 통해 대체하는 경우가 이에 해당

○ **빅데이터(민간, 행정)를 활용한 가공 통계** : 국가 정책 지원, 사회 현상의 진단 등의 목적으로 민간·행정 빅데이터를 가공·분석한 2차 통계

* 소셜 분석, 언론 데이터 분석, 빅데이터·AI 기술을 활용한 지수 개발 등 분석 대상과 기술로서 빅데이터 및 AI를 채택한 경우

○ **기초 통계 데이터를 활용한 가공 통계** : 현존하는 국가 통계 데이터를 빅데이터·AI기술을 통해 2차 분석함으로써 도출하는 신규 통계

* 통계법에 저촉되지 않는 범위 내에서 설문조사 데이터기반 빅데이터·AI 분석을 수행해 별도의 부가적인 지표 및 시사점을 도출하는 경우 이에 해당

□ **(예상 쟁점 #1)** 신규 방식을 수용하는데 있어 방법론의 대표성 및 보편적 신뢰가 형성되지 않은 문제

○ 국가통계는 통계적으로 널리 활용되는 과학적인 작성기법을 사용하여야하나[12], 머신러닝(Machine Learning), 딥 러닝(Deep Learning) 등 AI기술을 통한 분석 결과의 신뢰가 보편적으로 형성되지 않은 상황

○ 본질적인 원인은 수많은 머신러닝 알고리즘 및 파생된 기법들 중 신규 통계 생성에 적합한 기술이 제시되지 않음에서 기인

→ 신규 통계 생산 시 활용에 적합한 검증된 알고리즘을 정의하고 이에 대한 세부적인 활용 지침 마련이 요구됨

□ **(예상 쟁점 #2)** 신규 통계의 생산방식은 기존의 통계 생산 프레임워크에 맞추어 해석하기에 적합하지 않음

○ 국가의 통계업무를 구체화하여 단계별 제시한 국제표준은 GSBPM으로, 해당 모델의 경우 조사통계에 적합한 프로세스를 준용하고 있어 대표적 데이터 마이닝 업계 표준(CRISP-DM[13])과는 구조가 상이

* 한국의 경우 GSBPM을 참고해 국내 환경에 맞추어 재정의한 KSBPM을 준용

○ CRISP-DM은 빅데이터의 제어 및 분석 모델의 유효성 검증과 평가에 주안점을 둔 프로세스로서, 기존 통계모델과 통합하기 어려운 지점을 형성

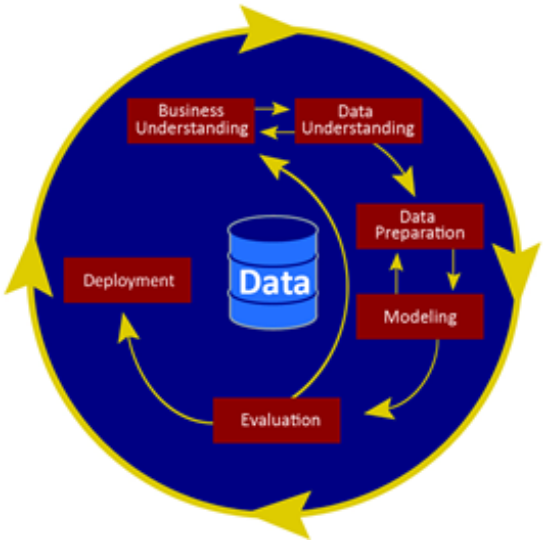
* 스위스 통계청은 GSBPM이 빅데이터 통계(신규 통계)를 포괄하기에 적합하지 않으며, 서로 다른 프로세스의 개발 또는 상호보완적인 연계 방안을 유럽 통계 시스템(ESS)에서 다룰 것을 강력히 권고

→ 빅데이터 분석 프로세스를 포괄할 수 있는 통계 프로세스 표준의 개정 고려

12) 국가통계 기본원칙의 [제 2항 : 신뢰성 제고]의 실천방안 중 발췌

13) Cross-industry standard process for data mining(CRISP-DM) : 데이터 마이닝 분야에서 가장 널리 사용되는 분석 모델로서, 관련 전문가가 사용하는 일반적인 접근 방식을 정의

< 표 7 > CRISP-DM(좌) 및 프로세스 역할 요약(우)



- ▸ 사업 이해(Business Understanding)
  - 사업·기술관점 목표수립, 프로젝트 기획 등
- ▸ 데이터 이해(Data Understanding)
  - 데이터 수집, 특성 확인, 목표데이터 도출 등
- ▸ 데이터 준비(Data Preparation)
  - 분석 데이터 선별·정제·검증·구조화·통합 등
- ▸ 모델링(Modeling)
  - 기법선택, 유효성검증, 모델 생성, 순위 지정 등
- ▸ 평가(Evaluation)
  - 모델링결과 평가, 주요항목 도출, 향후단계 결정 등
- ▸ 전개(Deployment)
  - 전개전략 수립, 지속/유지전략 수립, 보고서 작성 등

* 그림 출처 : IBM SPSS Modeler Essentials(2017)

< 그림 3 > 한국통계업무프로세스모델(V.2.0)

기획	설계	구축	수집	처리	분석	배포	보관	평가
수요 파악	산출물 설계	수집 도구 구현	자료 수집대상 설정	자료통합	산출물 작성	공표자료 점검 및 적재	자료보관규칙 정의	평가 계획 수립
수요 확정	항목 설정	처리시스템 개발, 개선	자료 수집 준비	분류 및 코딩	산출물 검증	공표 산출물 작성	자료 보존 및 관리	수행 및 보고서 작성
산출목표 수립	수집 방법 설계	배포시스템 개발, 개선	자료 수집 진행	자료검토 및 보완	산출물 해석 및 설명 작성	자료 배포 관리		개선과제도출 및 실행계획수립
통계적 개념 정립	모집단 및 표본설계	업무 절차 설정	수집자료 점검 및 완료	결측치 처리	정보보호 및 공개범위 검토	배포 촉진		
자료 가용성 검토	자료처리 방법 설계	시스템 점검		신규변수(항목) 및 단위 도출	산출물 확정	이용자그룹 관리		
통계 작성 계획안 수립	통계 작성 체계 설계	작성 과정 통합 점검		가중치 계산				
		통계 작성 체계 확정		집계				
				자료 집계 및 확정				

* 출처 : 표준 자료처리 모델 GSDEMs의 소개 및 시사점(이의규, 2018)

□ **(예상 쟁점 #3)** 재현 불가능한 비결정론적(Non-deterministic) 알고리즘[14] 기반의 결과를 국가 통계로 관리 가능한지 여부

○ 널리 활용되는 빅데이터·AI 기술 중 결과 도출 과정에서 연구자의 주관적 판단이 반드시 요구되는 알고리즘 존재

* 비 지도학습(Unsupervised Learning)으로 분류되는 클러스터링의 경우 분류의 기준이 되는 군집의 수(k)를 연구자 판단을 통해 결정

○ 빅데이터·AI 알고리즘의 일부는 주관적 판단 없이도 동일한 파라미터에 의해 도출된 결과가 확률적으로 다를 수 있음

- 알고리즘 동작 과정에서 난수(random value)를 계산에 활용하므로, 통계의 재현가능성을 완전히 충족하기 어려운 경우 존재

- 통계적 모델링의 유효성을 검정력(p)으로 판단할 수 있는 것과 달리 AI 알고리즘은 유사한 신뢰 기준이 표준화 되지 않음

○ 한편, 지도학습(Supervised Learning)의 경우 주어진 검증 데이터를 대상으로 학습 모델의 성능 측정이 가능해 상대적으로 활용에 용이할 것으로 예상

→ 인공지능 기술의 특성 분석 및 표준화된 평가 기준 정립이 요구됨

14) 비결정론적 알고리즘이란 예측한 그대로 동작하지 않는 알고리즘을 의미하는데, 확률적으로 최적의 근사 값을 찾는 원리에 기반 하므로 동일 알고리즘 시행 시 동일 결과를 100% 보장할 수 없음

< 표 8 > 신규 통계에 잠재적으로 활용 가능한 머신러닝 알고리즘의 종류와 특성

구 분	설명	대표적 기술군	활용측면의 성능 분석 가능여부
지도학습 (Supervised Learning)	입력에 대한 정답(레이블)을 반복 학습시켜 정답이 제시되지 않은 입력의 정답을 찾는 알고리즘	분류, 회귀, 신경망 [15]	가능 (검증 데이터 기반)
반지도학습 (Semi-Supervised Learning)	정답을 보유한 소수의 입력과 그렇지 않은 다수 입력를 혼용하여 학습함으로써 성능을 높이는 기법	상동	가능 (검증 데이터 기반)
비지도학습 (Unsupervised Learning)	사람의 지도 없이 컴퓨터가 스스로 데이터의 유형을 학습하는 기술	군집화, 분포 추정 [16], 분류 [17]	불가능 (모델에 대한 품질 평가는 가능)

※ 머신러닝 기법으로 현재 상태(State)에서 어떤 행동(Action)이 보상(Reward)을 획득하는데 효과적인지를 학습하는 방식인 강화학습(Reinforcement Learning)이 존재하나, 해당 주제에 적합하지 않다고 판단하여 제외

15) 인공 신경망은 뉴런의 구조에서 착안한 지도학습의 일종으로서, 신경망의 구조가 복잡(다수의 계층으로 구성)하게 설계되어 있는 경우를 심층 신경망(Deep Learning)이라 칭하며, 일반적으로는 분류를 위한 방법론으로 활용됨

16) 데이터의 분포를 설명하기에 최적인 확률 분포에 매핑하여 해석하는 기법

17) 비지도학습을 통한 분류는 최근 GPT-2(비지도학습 기반 문장 예측모델)의 성공으로 조명받기 시작한 분야로서, 향후 지도학습의 대체제로 평가되고 있으나 신뢰성을 갖추기 위한 학습 규모, 학습 시간 등을 고려했을 때 통계 생산 목적의 활용 가치는 아직까지 낮음

(2) 통계 생산 프로세스의 현대화

□ **(조작적 정의) 프로세스 현대화** 는 기존의 조사통계 프로세스를 수용하되 통계 데이터의 품질 향상, 생산 절차의 효율성 등을 개선하기 위해 빅데이터·AI 기술을 도입하는 것을 의미

○ **기능 개선** : 통계 프로세스 상 의사결정 요소들의 성능을 높이는 기법들로서 층화, 대체, 보정 등의 현행 통계 방법론을 보완 또는 대체 [18)]

○ **생산성 증대** : 통계 업무의 비효율을 개선하고 생산성을 높이기 위한 기술 도입으로서 코드화, 이상값(outlier) 검출, 레코드 연계, 데이터 비식별화 지원 등으로 구성

□ **(도입 가능 부문)** 데이터를 제어하고 처리하는 절차 다수에서 AI 기술 도입 가능성 검토가 가능

○ 통계 생산 프로세스 참조 기준으로는 GSBPM(v5.1)을 활용

- GSBPM은 총 8단계로 구성되며 각 단계에는 하위 프로세스가 존재

  * 본 프로세스는 통계 생산 과정에서 유연하게 적용하고 해석할 수 있음 [19)]

○ 해외 문헌을 통해 머신러닝 활용이 가능하다고 판단된 통계 생산 세부 프로세스와 과업은 아래와 같음

< 표 9 > 머신러닝 활용이 가능한 국가통계 생산 과정 및 예상 기술군(GSBPM 기준)

코드 (1레벨)	명 칭	코드 (2레벨)	명 칭	과업 정의	도입 가능 기술군 [20)]
2	설계	2.4	모집단 및 표본 설계	- 모집단을 식별 및 지정 - 적합한 샘플링 기준 및 방법론 결정	- 분류 - 군집화

18) 유엔유럽경제위원회 (2018), The use of machine learning in official statistics 참고

19) statswiki.unece.org, “Understanding the GSBPM”참고

20) 인공신경망의 경우, 설계 목적에 따라 모든 기술군을 포괄가능하므로 별도로 명시하지 않음

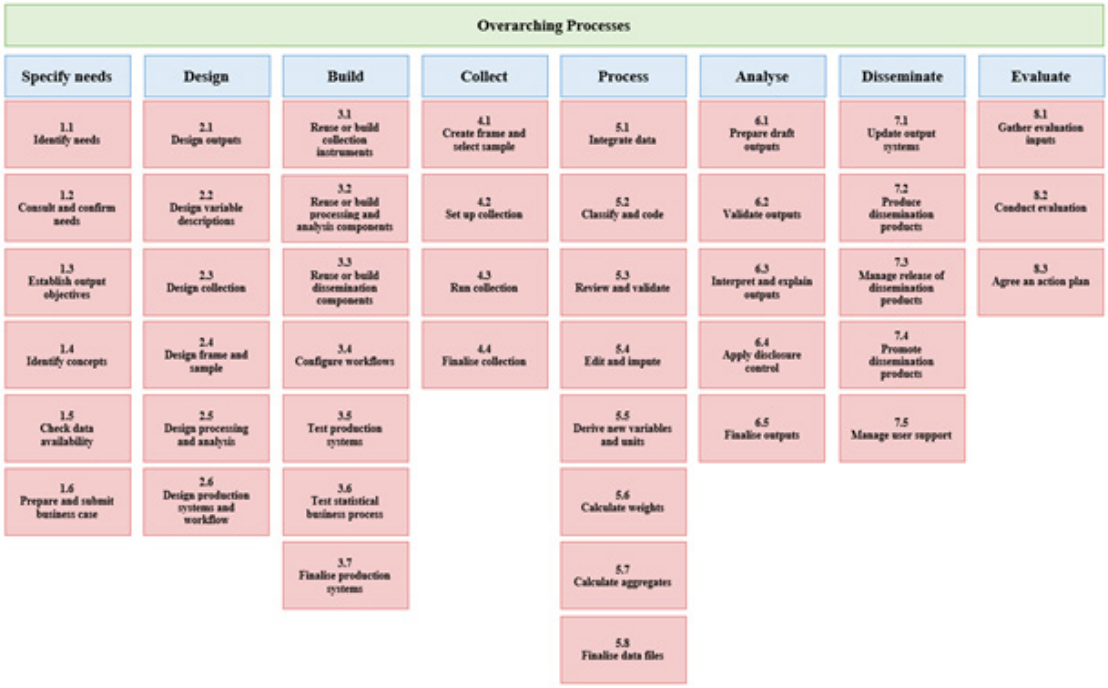
코드 (1레벨)	명 칭	코드 (2레벨)	명 칭	과업 정의	도입 가능 기술군
4	수집	4.1	모집단 생성 및 샘플링	- 2.4에서 지정된 모집단을 생성 - 2.4에서 지정된 샘플링 수행	- 분류
		4.3	수집 진행	- 실사 진행 및 후속조치 점검 - 수기 데이터 입력 및 현장 작업 관리 등	- 분류 - 군집화 - 회귀
5	처리	5.1	자료 통합	- 데이터의 출처가 2개 이상일 시 통합 - 검증된 비 통계 데이터로 설문 결과 일부를 대체하는 과정도 이에 해당	- 군집화
		5.2	분류 및 코딩	- 설문 결과 데이터의 분류 및 코딩	- 분류
		5.3	검토 및 검증	- 데이터를 검사하여 이상치, 항목무응답, 잘못된 코딩 등 잠재적 문제, 오류, 불일치 사항을 식별	- 분류 - 군집화
		5.4	편집 및 대체	- 부정확 데이터 및 누락, 신뢰할 수 없는 데이터를 새로운 값으로 대체 및 제거	- 분류 - 회귀 - 분포 추정
		5.6	가중치 계산	- 설계 단계(2)에서 정의한 가중치를 조사결과에 적용함으로써 모집단에 대한 통계 결과를 도출 - 설문 결과의 무응답 조정 및 변수의 정규화 등	- 분류 - 회귀

코드 (1레벨)	명 칭	코드 (2레벨)	명 칭	과업 정의	도입 가능 기술군
6	분석	6.2	산출물 검증	- 품질 측정 프레임워크에 따라 통계 품질을 검증	- 군집화
		6.3	산출물 해석 및 설명 작성	- 결과가 초기 기대치를 얼마나 잘 반영하는지 평가 - 다양한 관점에서 통계 결과를 해석 - 심층 통계 분석 등	- (미정) [21]
		6.4	정보보호 및 정보공개 검토	- 데이터 배포 과정의 기밀성 훼손 방지 - 필요에 따라 데이터 비식별화 처리 수행	- 회귀 - 분류
7	배포	7.5	이용자 지원 관리	- 통계 서비스에 대한 사용자 질의 및 요청에 대해 합의된 마감일 내에 응답을 제공	- (미정)
8	평가	8.2	평가 실시	- 8.1단계에서 확보한 평가 피드백과 목표한 벤치마킹 대상(만약 존재한다면)과 결과를 비교한 후 평가보고서 작성	- (미정)

※ 자료 참고 : 유엔유럽경제위원회, 독일 통계청

21) 2차 분석의 영역으로서 예상 기술군을 특정하는 것은 부적절하다 판단하였음

< 그림 4 > 일반 통계 비즈니스 프로세스 모델(GSBPM v5.1)



*출처 : unece.org

□ **(예상 쟁점 #1)** 기존의 방법 대비 우수성을 객관적으로 비교가능한가

○ AI 기술 도입이 현실화되기 위해서는 기존 활용 기법 대비 AI 기술을 활용할 시 품질의 상대적 우위를 증명해야하나, **품질 진단 기준이 불명확**

- 1차 통계의 경우 과거 데이터 또한 기존 방법론을 통해 생산된 자료일 가능성이 높아 객관적인 방법론 우위 측정에 한계
  * 통계 주제 및 특성에 따라 기업 정보, 행정 정보 등 기존 조사 외 객관적 정보를 사후 습득할 수 있는 경우에는 해당 이슈에서 자유로울 수 있음

○ 신규 통계의 쟁점(신규 통계 생산-예상 쟁점#3 참조)과 마찬가지로 알고리즘 특성에 의한 신뢰성 문제가 대두될 가능성

- 해외 사례에서 비지도 학습을 국가통계에 활용한 사례가 존재하나, 그럼에도 불구하고 검증 이슈에서 자유로운 것은 아님

→ 전통적 통계 기법-AI 알고리즘의 비교 검증 방안 발굴이 필요

□ **(예상 쟁점 #2)** AI 기술 도입은 통계 생산부터 공표까지의 제한된 생산 기간을 준수할 수 있는 해법인가

○ AI 모델링은 방법론의 종류, 학습 데이터의 규모, 최소 성능 기준 [22] 등에 의해 최종 모델 산출까지의 시간이 유동적

○ 국가 통계는 시의성 [23] 및 정시성 [24]을 갖추기 위해 필연적으로 생산 기한을 준수해야하므로, AI 학습 알고리즘의 학습시간-품질 간 조율 문제가 발생할 가능성

→ 다수의 실증 사업을 통한 통계 유형별 도입 적합성 진단 필요

※ 참고 : 현행 조사기반 국가통계 또한 제한된 생산 기간 준수를 고려하는 사례가 존재하는데, 항목 무응답 대체 기법으로 명시적 형태의 대체 모형(explicit model) [25]이 아닌 내재적 모형(implicit model) [26]을 주로 채택하는 경우가 대표적 [27]

22) AI 성능 지표로서 민감성, 정밀도, 거짓양성률(FPR) 등이 존재(전자신문, 2019.10)
23) 시의성 : 작성 기준시점과 결과공표시점간의 차이를 나타내는 통계의 현실 반영도와 관련된 개념
24) 정시성 : 예정된 공표시기를 정확히 준수하는가에 대한 개념
25) 전체 응답 결과에 상응하는 예상 분포를 추정해 무응답 항목을 보완하는 방법으로서, 모델링 결과가 존재해 수치 검증이 가능한 장점을 가지나, 추정 난이도 및 소요 기간으로 인해 선호되지 않음
26) 무응답 항목을 추정할 수 있는 별도 데이터를 참조하여 결측 값을 대체하는 방법으로서, 추정 난이도가 낮고 소요기간이 짧아 선호되는 기법임. 대표적으로 핫 덱, 콜드덱 기법이 존재
27) 명시적 형태의 대체 모형을 추정해 항목 무응답을 처리하는 기법은 생성 모델의 검증이 가능한 장점이 있는 반면, 모델링 과정의 기간 소요, 추가 예산 등의 현실적 어려움이 존재

4. 국가통계 AI 도입 촉진 방안

(1) 국내 통계 제도의 개편 방향

□ 국내 통계청은 **단기적으로 현행법 내 승인기준의 일부 보완, 중장기적으로는 신고제도 및 시범통계 제도 도입**을 검토할 것임을 예고

○ 단기적 관점의 승인기준 보완은 기존 통계작성승인 심사표상의 **(1) 조사통계에 국한된 심사문항을 범용적인 문구로 변경하고 (2) 기존 통계 대체를 위한 빅데이터 도입의 경우 중복을 허용**할 수 있도록 하는 조치

* 예1) (현행) 조사표 설계나 조사 항목이 동일 또는 유사한가?
→ (수정) 수집 항목 및 결과표 항목이 동일 또는 유사한가?
* 예2) (추가) 기존 통계를 보완하는 사항이 있는가?

○ 중장기적 관점에서 고려될 신규 제도 도입은 국가가 관리하는 통계의 범위를 확대하는 방향을 의미하나, 아직 구체화 되지 않음(아래 참조)

□ **(시범통계제도)** 통계청에서 언급한 시범통계제도란 **영국의 실험통계(experiment statistics) 제도에서 착안**한 국가통계 관리체계 개편을 의미

○ 영국의 실험통계는 새로 개발되거나 혁신적인 공식통계로 정의[28] 되어 AI·빅데이터 활용 통계를 국가에서 관리하는데 용이한 제도

※ 영국은 공식통계 범위 내 국가통계와 실험통계를 별도로 두어 관리하나, 국내는 국가에서 관리하는 통계가 국가통계로 일원화

※ 실험통계는 추후 국가통계로서의 품질 조건을 충족할 경우 국가통계로 승격이 가능하며 통계 생산의 지속 및 폐지가 비교적 자유로움

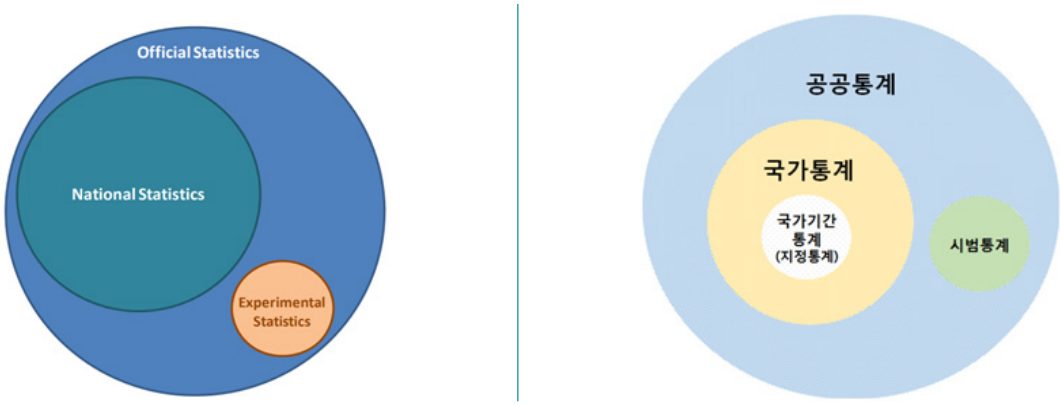
○ 해당 제도는 국가통계에 新기술을 활용할 수 있는 시기를 앞당기는데 의의가 있어 일종의 샌드박스(Sandbox) 제도[29] 범주에 해당

○ 국내 통계청 또한 국가통계 승인 대비 완화된 기준에 입각한 시범통계 제도 운영을 구상하고 있어 향후 귀추가 주목

28) GSS Guidance on Experimental Statistics(2019.7) 인용

29) 국민의 생명·안전에 위해가 되지 않는 한 마음껏 도전하고 새로운 시도를 해볼 수 있도록 기회를 부여하기 위한 제도를 의미

< 그림 5 > (좌)영국의 공식통계 개념도와 (우)한국의 시범통계 도입 구상도



*출처 : (좌)Government Statistical Service(2019), (우)정보통신정책연구원(2019)

□ **이처럼 AI 활용 통계의 국가통계 승인 기회를 개방하는 것은 긍정적**이나, 실제 **AI 도입을 유도하고 활성화 하는 측면에서는 해결해야할 쟁점이 산적**

○ 3장에서 제시한 쟁점들은 국내 국가통계 제도의 개편이 완료되더라도 여전히 해소되기 어려운 AI 자체의 본질적 사안으로서 별도의 해법이 요구됨

< 표 10 > AI·빅데이터 활용 통계 도입의 예상쟁점 및 시사점(3장 참조)

구 분	예상 쟁점	시사점
신규 통계 생산	AI·빅데이터 기반의 분석결과에 관한 신뢰가 보편적으로 형성되지 않음	신규 통계 생산 시 활용에 적합한 검증된 알고리즘을 공표하고 이에 대한 세부적인 활용 지침 마련이 요구
	신규 통계의 생산방식은 기존의 통계 생산 프레임워크에 맞추어 해석하기에 적합하지 않음	빅데이터 분석 프로세스를 포괄할 수 있는 통계 프로세스 표준의 개정 고려
	재현 불가능한 비결정론적 (Non-deterministic) 알고리즘 기반의 결과를 국가 통계로 관리 가능한지 여부	인공지능 기술의 특성 분석 및 표준화된 평가 기준 정립이 요구됨
통계 생산 프로세스의 현대화	기존의 방법 대비 우수성을 객관적으로 비교가능한가	전통적 통계 기법-AI 알고리즘의 비교 검증 방안 발굴이 필요
	AI 기술 도입은 통계 생산부터 공표까지의 제한된 생산 기간을 준수할 수 있는 해법인가	다수의 실증 사업을 통한 도입 적합성 진단 필요

○ 통계 혁신을 위한 제도 개편과 함께, **활용 촉진을 위한 AI·빅데이터 기술의 대표성 확보가 필요한 시점**

(2) 국가통계 AI 도입 촉진 방안

□ **(제안)** 통계청의 정책 변화 방향에 발맞추면서도 실질적인 AI 도입 성과를 높이는데 필요한 기능과 역할을 수행하는 **AI 표준 수립 기관 지정**

□ **(제안 취지)** 향후 국가통계 관리 체계 변화에 대응해 **AI 기술의 대표성을 부여하고 실증 사업 확대를 유도**하기 위함

< 표 11 > 국가통계의 AI기술 도입 촉진을 위해 필요한 요소

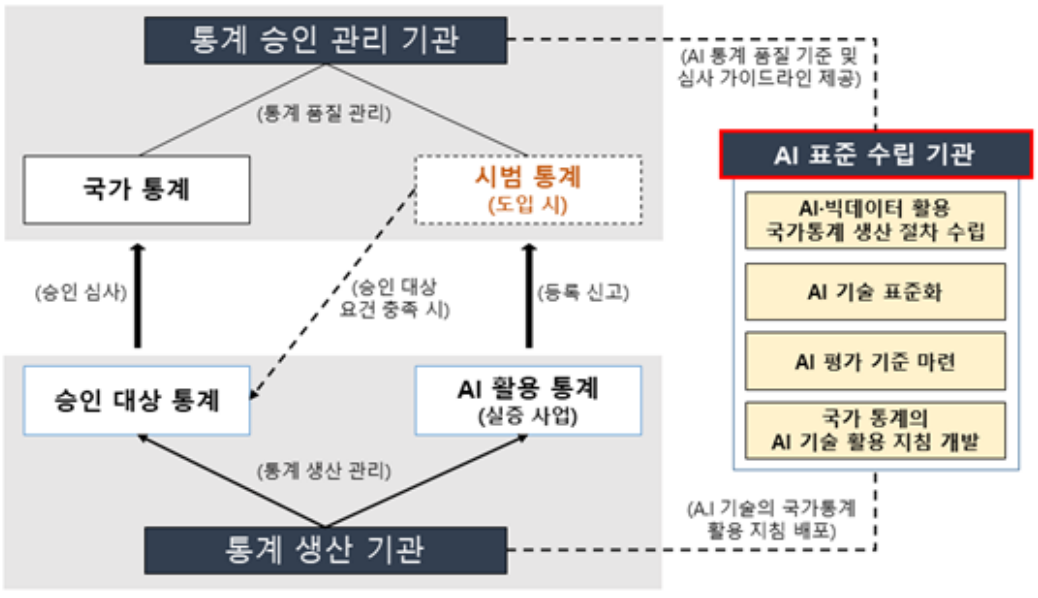
구 분	시사점
신규 통계 생산	검증된 AI 알고리즘 공표 및 세부적 활용 지침
	AI·빅데이터 활용 국가통계 생산 절차 수립
	기술 특성 분석 및 평가 기준 표준화
통계 생산 프로세스의 현대화	기존 통계기법-AI 알고리즘의 비교 검증 방안 발굴
	다수의 실증 사업을 통한 도입 적합성 진단 필요

▶ AI 기술의 대표성 부여

▶ 실증 사업의 확대

□ **(실행 방안)** 통계승인관리기관—연구기관 간, **AI 중심의 국가통계 혁신을 위한 협력 체계** [30]

< 그림 6 > AI 중심의 국가통계 혁신을 위한 협력 체계(안)



*출처 : 소프트웨어정책연구소

30) 2014년 UNESE 주도의 빅데이터 샌드박스 프로젝트 추진을 위해 아일랜드 통계청(CSO)과 슈퍼컴퓨팅 연구소(ICHEC) 간 협력 체계가 구성 된 사례 참고

5. 요약 및 시사점

□ 유럽 국가들을 중심으로 국가통계의 품질 및 조사환경 개선에 대한 공감대가 형성됨에 따라, **이를 해결할 방안으로서 빅데이터 및 인공지능 기술 도입이 추진 및 검토**되고 있음

○ 국가통계의 AI·빅데이터 도입 필요성이 수면위로 떠오른 주요 원인들로 아래와 같은 사회 현상이 거론

① 새로운 유형의 산업이 빠르게 등장함에 따른 산업구조 반영의 어려움
② 기업 활동의 다양화(지역에 기반 하지 않은 산업, 사무실 연락처 부재 등)
③ 조사별 응답자 중복에 의한 회수율 악화
④ 즉시 활용 가능한 민간 중심의 데이터의 급증

○ 이를 해결하기 위하여 독일, 네덜란드, 스위스 등 세계 각국의 국가통계 AI·빅데이터 기술 도입 사례가 증가하고 있는 상황임

□ **국가통계의 AI·빅데이터 기술 도입은 아래 두 가지 시각으로 구분**

○ **(신규 통계 생산)** 조사통계 기법을 활용하지 않고 빅데이터·AI 기술을 활용해 통계 생산

○ **(통계 생산 프로세스의 현대화)** 기존의 조사통계 프로세스를 수용하되 통계 데이터의 품질 향상, 생산 절차의 효율성 등을 개선하기 위해 빅데이터·AI 기술을 도입

구 분	세부 적용 분야
신규 통계 생산	• **대체(Replace) :** 기존 통계를 완전 또는 부분 대체하는 수단 • **부가(Addition) :** 조사 통계 방식으로 도출이 어려운 부문에 대해 데이터 분석에 기반을 둔 우회적 해결 • **인사이트(Insight) :** 완전히 새로운 분야의 통계 데이터 제시
통계 생산 프로세스의 현대화	• **정확성(Accuracy) :** 통계의 정확도 향상 • **생산성(Productivity) :** 업무 효율화를 위한 자동화 기법 • **연결성(Connectivity) :** 연결 가능한 데이터들의 상호 보완적 활용 • **비용(Price) :** 통계 생산의 비용 절감을 위해 활용

□ **국가통계에 AI 기술 도입 시 예상되는 쟁점들은 아래와 같으며 AI 기술의 대표성 부여와 실증 사업의 확대 필요성을 시사**

구 분	예상 쟁점
신규 통계 생산	AI·빅데이터 기반의 분석결과에 관한 신뢰가 보편적으로 형성되지 않음
	신규 통계의 생산방식은 기존의 통계 생산 프레임워크에 맞추어 해석하기에 적합하지 않음
	재현 불가능한 비결정론적(Non-deterministic) 알고리즘 기반의 결과를 국가 통계로 관리 가능한지 여부
통계 생산 프로세스의 현대화	기존의 방법 대비 우수성을 객관적으로 비교 가능한가
	AI 기술 도입은 통계 생산부터 공표까지의 제한된 생산 기간을 준수할 수 있는 해법인가

AI 기술의 대표성 부여

실증 사업의 확대

○ **(AI 기술의 대표성 부여)** 국가통계 활용 목적의 AI 기술 개발 및 표준화, 성능 평가기준 마련 등

○ **(실증사업의 확대)** 다양한 유형의 AI 활용 통계 생산 실험을 통해 AI 기술들의 국가통계 도입 적합성 및 실효성 진단

□ **국가통계 AI 도입 촉진을 위해 기술 검증·평가·활용 지침 마련 등을 수행하는 협력 기관 지정을 제안**

○ 통계승인관리기관—연구기관 간 AI 중심의 국가통계 혁신을 위한 협력 체계를 구축함으로써, 실질적인 제도 도입의 성과를 상승시키고 국가통계 혁신의 시기를 앞당기는데 기여할 것으로 기대

참고문헌

1. 국내문헌

통계교육원 (2015), 국가통계 이해.

통계청 (2020), 「빅데이터 활용통계」등 통계 다양성 확대를 위한 국가통계 승인기준 보완방향(안).

통계청 (2015), 통계조정업무 매뉴얼.

임경은 (2012), 조사과정자료 수집 및 활용을 위한 가이드라인 수립 방안.

이의규 (2018), 표준 자료처리 모델 GSDEMs의 소개 및 시사점.

전자신문 (2019), [기고] AI 모델과 AI 기반 시스템, 도대체 어떻게 평가해야 하나,

통계청 (2019), 빅데이터 활용통계의 국가통계 승인관리방안 연구

2. 국외문헌

Radermacher, W. J. (2018), Official statistics in the era of big data opportunities and threats.

UNECE (2018), The use of machine learning in official statistics

Rob Kitchin (2015), The opportunities, challenges and risks of big data for official statistics.

Loison, B., & Kuonen, D. (2018), Are Current Frameworks in the Official Statistical Production Appropriate for the Usage of Big Data and Trusted Smart Statistics?

Martin Beck, Florian Dumpert, Joerg Feuerhake (2018), Machine Learning in Official Statistics

Federal Statistical Office Germany (2019), Introduction to Big Data in Official Statistics

Statistics Netherland (2020), Discusstion paper : Big data in official statistics

Eurostat (2017), ESSnet Big Data Specific Grant Agreement No 1 (SGA-1)

Edith, desiree de leeuw. (2008), Mixed mode surveys: When and Why

Meertens, Q. A., Diks, C. G. H., van den Herik, H. J., & Takes, F. W. (2018). A data-driven supply-side approach for measuring cross-border internet purchases.

Wirth, R., Hipp, J. (2000), CRISP-DM: Towards a standard process model for data mining.

Government Statistical Service (2018), GSS Guidance on Experimental Statistics

INFOSTAT Slovakia (2013), Introducing New Tool for Official Statistics – Genetic Programming

3. 기 타

The OECD Glossary of Statistical Terms, https://stats.oecd.org/glossary/

통계법, law.go.kr/법령/통계법

Big Data Sandbox, https://joinup.ec.europa.eu/solution/big-data-sandbox

RIETI, Can Big Data Change Official Statistics,https://www.rieti.go.jp/en/events/bbl/19031401.html

국가통계 기본원칙 전문, http://kostat.go.kr/portal/korea/kor_ko/2/index.action

GPT-2: 1.5B Release –OpenAI, https://openai.com/blog/gpt-2-1-5b-release/

Understanding the GSBPM, https://statswiki.unece.org/

Enabling Big Data approaches to gather official statistics, https://www.ichec.ie/partnerships/public-sector/cso

Big Data Sandbox, https://joinup.ec.europa.eu/solution/big-data-sandbox/

주 의

이 보고서는 소프트웨어정책연구소에서 수행한 연구보고서입니다.
이 보고서의 내용을 발표할 때에는 반드시
소프트웨어정책연구소에서 수행한 연구결과임을 밝혀야 합니다.

AI기술의 국가통계 활용 사례 및 국내 도입 촉진 방안

Use Cases of AI Technology in National Statistics &
Suggestions for Promoting Domestic Adoption

경기도 성남시 분당구 대왕판교로 712번길 22 글로벌 R&D센터 연구동(A)
Global R&D Ceneter 4F, 22, Daewangpangyo-ro 712beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do

www.spri.kr