Research Report

RE-099

국가통계 AI도입 활성화 방안 연구

A study on how to activate the use of AI technology in national statistics

김정민

2021. 07.

SPRi 소프트웨어정책연구소

이 보고서는 2020년도 과학기술정보통신부 정보통신진흥기금을 지원받아 수행한 연구결과로 보고서 내용은 연구자의 견해이며, 과학기술정보통신부의 공식입장과 다를 수 있습니다.

연 구 기 관 : 소프트웨어정책연구소

과제책임자 : 김정민 선임연구원

목 차

<표 목차>

<제목 차례>

<요약문>

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.

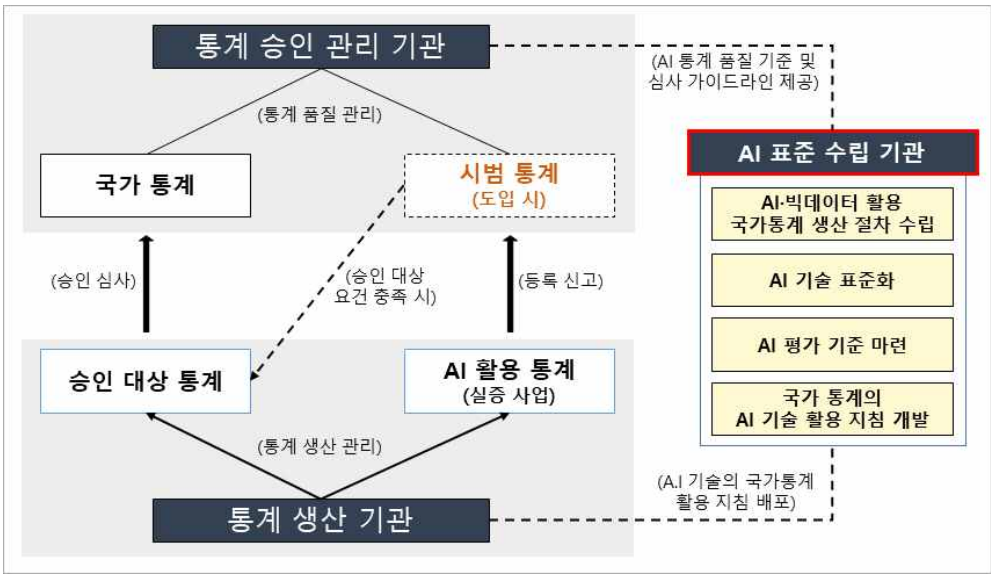
최근 국내외를 막론하고 AI 및 빅데이터 기술을 국가통계에 활용할 수 있는가에 관한 논의가 본격적으로 진행되는 추세다. 같은 맥락에서 실제 국가통계에 한시적으로 도입해 보거나 테스트하는 사례 또한 증가하고 있다. 이는 결국 AI 및 빅데이터 기술을 통계 생산에 활용할 시 국가통계로 승인 가능한지에 대한 고려로 확대된다. 국내 통계청 또한 올해 상반기 중 조사통계 심사에 맞추어져 있던 승인 심사 문항을 조사 통계 외 생산 방식에도 적용할 수 있도록 일부 개정하는 방안을 추진하고 있으며, 중장기적으로는 국가의 통계 관리 범위를 확대하기 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. 결과적으로 아래와 같은 AI 표준 수립 기관의 지정될 시 해당 이슈 해결에 효과적임을 제안하였다.



이와 더불어 4-5장에서는 금융감독원의 전자공시 외감기업 전수에 해당하는 빅데이터 확보를 전제로 생산 가능한 두 가지 AI·빅데이터 활용 통계 생산 방안을 제안하였다. 첫 번째는 디지털 산업을 분류하기 위해 OECD의 디지털 산업 정의와 국내 전자공시 빅데이터를 활용하는 방안이며, 두 번째는 디지털 전환지수를 개발하는 관점에서 전자공시 빅데이터를 활용하는 연구 모델이다. 국내 차원의 빅데이터 활용 통계에 대한 도입 가능성이 높아짐에 따라 향후 실제 국가 승인을 목표로 제작될 수 있을 것으로 예상된다.

제1장 연구 배경

국가는 일련의 정책을 개발하거나 행정적인 조치를 취하고자 할 때 그것이 온당한 방향성을 가지는지 검토하는 과정을 거친다. 대다수의 정책과 행정 개선이 그러하며 여기에는 관련 학술적 이론의 검증과 더불어 국내의 객관적인 상황을 파악하는 절차가 필수적으로 포함되어 있다. 뿐만 아니라 실행된 정책의 성패 여부를 판단하는데 있어서도 현재 정책 수혜처의 일반 현황과 평가에 활용할만한 정량적 근거를 찾는다. 우리는 이러한 목적을 달성하기 위하여 주로 일련의 동향을 한눈에 파악할 수 있는 수치 데이터인 통계를 인용하는 게 일반적이다. 통계자료는 국가 운영을 위해 없어서는 안 될 객관적 지표로서 국가가 새로운 정책을 개발 및 관리하고자 할 때 필수적으로 활용된다. 때문에 정책과 국정 운영 방향에 맞추어 끊임없이 새로운 통계가 개발되고 생산되어 왔다. 이처럼 국가 운영에 활용하기 위해 개발 및 관리되는 통계를 우리는 국가통계(National Statistics)라 부른다. 명시적 개념으로는 국가통계 제도를 통해 배포되는 통계를 의미하는데, 사회통합 기능과 정책의 수립 및 평가 기능을 갖추고 있기 때문에 국가의 중요한 공공재로 분류된다. 통계법에는 이와 같은 국가통계 생산을 담당하는 주체에 대한 기준, 국가통계의 목적으로 생산된 통계를 승인하는 절차 및 체계, 국가통계로 승인된 통계에 관한 품질의 관리 방법 등이 포함되어 있다.

통계법 상의 국가통계는 통계작성지정기관에 의해 생산되는 통계작성 승인을 획득한 통계를 의미한다. 즉 통계작성기관이 통계를 생산하더라도 승인을 획득하지 못할 경우 원칙적으로 수치를 공표 및 배포하는 것이 금지되어 있다. 이처럼 국가통계의 승인 획득 여부가 새롭게 작성된 통계의 존폐를 결정하는 주요한 사안이다 보니 통계를 생산하고자 하는 복수의 정부·공공기관들은 자체적으로 생산한 통계의 통계작성 승인 획득을 목표로 한다. 때문에 국가통계의 승인 심사기준은 그들에게 중요한 달성 목표로 작용하고 있다.

최근 국내외를 막론하고 인공지능 및 빅데이터 기술을 국가통계에 활용할 수 있는지에 관한 논의가 본격적으로 진행되고 있다. 같은 맥락에서 실제 국가통계 개발에 해당 기술들을 한시적으로 도입해보거나 테스트하는 사례 또한 증가하고 있는데, 이

는 결국 인공지능 및 빅데이터 기술을 통계 생산에 활용할 시 국가통계로서 승인 가능한지에 관한 고려로 확대되는 추세다. 국내 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. 현행 국가통계는 설문조사를 통해 수집한 데이터 기반의 조사통계와 행정자료에 기반한 행정통계가 주류를 이루고 있으며 이러한 전통적 통계 생산방식은 오랜기간동안 준용되어왔으므로 관리자, 통계 생산자 모두에게 익숙한 방식이다. 가령 이와 다른 방식을 통해 좀 더 개선된 통계 산출물을 도출할 수 있다손 치더라도 그것을 보편화하고 많은 사람들에게 전파하는 일은 쉬운 일이 아닐 것이다. 게다가 그러한 방법론이 경험 기반의 데이터 품질 향상의 접근을 택한다면 더욱 문제가 될 소지가 높다. 빅데이터, AI등 신기술의 품질 고도화는 컴퓨팅 분야에서 비롯되다보니 모델 설계를 통한 품질 제고보다는 특정 방법론의 파라미터 조정이나 알고리즘 수정을 통해 해결하려는 시각이 강한데, 이는 기존 통계 생산 방식과 차이가 있다. 즉 사회조사방법론과는 문제를 해결하기 위한 접근방식에서 차이를 보이며 품질을 검증하는 기준 정립에도 세부적인 관점이 상이하다. 그렇기 때문에 최근의 국가통계의 SW기술 도입 현안에 대하여 단순히 제도적 기회를 개방해주는 측면만으로는 AI나 빅데이터 기술의 통계 분야 도입이 활성화되는데는 한계가 존재한다. 제도적 기회뿐만 아니라 어떤 요건이 부합되어야만 실제 통계 생산자들이 혼란을 겪지 않고 AI기술에 기반한 통계 개발이 가능한지에 관해 논의하는 과정은 그런 이유에서 필수적이다.

본 보고서는 상기 문제에 착안하여 수행되었다. 먼저 2장을 통해 국가통계에 AI·빅데이터 기술을 도입하는 것과 관련한 해외의 앞선 논의를 되짚어보고 실제 주요 도입 사례를 소개한다. 이어서 3장에서는 AI·빅데이터를 활용하는 방향성을 두 가지 시각으로 구분하여 각각의 차이를 정의하고, 도입 시 예상 쟁점은 어떤 것들이 있는지를 분석하였다. 또한 2,3장을 통해 도출 된 이슈들을 종합해보고 현재 국내 통계 체계 개편 방향과의 접점을 모색해보았다. 4장의 경우 별도의 내용으로 구성하였는데, AI·빅데이터 기술을 활용하여 SW분야의 새로운 국가통계를 개발하기 위한 몇 가지 실현가능한 아이디어를 제안하였다.

제2장 국가통계의 AI·빅데이터 기술 도입 논의와 적용사례

국가통계의 빅데이터 활용은 기본적으로 활용할 수 있는 빅데이터가 존재한다는 가정하에 현실화될 수 있다. 여기에서 활용 가능한 빅데이터란 통계 생산의 기반 자료로의 가치를 가진 데이터를 의미하는 것으로, 자세하게는 목적성에 부합해야하고 대표성을 가지며 지속적인 데이터 획득 등의 문제에서 자유로워야 하는 제약을 가진다. 실제 이런 조건을 모두 만족하는 데이터를 구하기란 쉽지 않다. 그렇다보니 대안적인 접근은 국가에서 관리하는 행정자료 및 기존 통계 생산을 위해 수집한 설문 데이터를 2차적으로 분석하여 새로운 가치를 창출하는 방식이었다. 물론 이 또한 데이터 공개 이슈에서 자유로울 수 없으므로 시작 단계에서부터 쉽게 달성되기 어려운 난제임에 틀림없다.

전 세계의 국가들 중, 유로연합(EU)에 소속된 국가의 경우 이와 같은 문제에서 비교적 자유로운 편이다. 그들은 오픈 데이터(Open Data)를 지향함으로써 타 권역 국가 대비 기존 조사 통계를 산출하는데 있어 파생되는 다양한 데이터를 쉽게 확보 가능하다. 또한 국가 행정자료를 큰 어려움 없이 연구 목적 등으로 제공 받을 수 있는 체계가 구축되어 있다. 국가 행정자료는 전 국민 대상, 또는 목적성에 부합하는 데이터 전수에 해당하는 매우 방대한 빅데이터이고 국가의 신변에 위협이 발생하지 않는 이상 지속성을 확보하고 있으므로, 해당 국가들의 빅데이터 통계 도입의 속도를 높이는데 영향을 끼쳤다. 그러므로 2장에서 언급하는 국가 다수는 일반적으로 선진국으로 인식되는 특정 국가보다는 유로연합에 소속되거나 관련성을 가지는 국가를 중심으로 소개하였다.

제1절 각 국의 문제인식과 관련 논의

유로연합의 장을 역임한 Walter J. Radermacher는 그의 2018년 저서를 통해 최고급 식당이 미식계의 선두 자리를 계속 유지하기 위해서 끊임없는 재창조 노력을 하는 것처럼 공식 통계 역시 급속하게 변화하는 혁신 요구를 정면으로 마주해야 함을 주장하

였다. 이와 같은 발언은 4차산업혁명으로 급격한 사회변화가 이루어지고 있는 상황에서 통계 분야는 기존 방식을 혁신하려는 움직임이 소극적인 작금의 상황을 경고하는 의미로 해석가능하다. 그의 발언은 단순히 국가 통계를 관장하는 수장으로써의 덕담으로 받아들일 수도 있겠으나, 현실은 대다수 유럽 국가를 중심으로 관련 문제가 대두된 결과라 볼 수 있다.

유럽 국가들은 일찍이 국가통계의 품질 및 조사환경 개선에 대한 공감대를 형성해 왔다. 국가통계 개선이 이슈로 부각된 주요 원인은 4가지 정도로 정리해 볼 수 있는데, 각각의 원인은 비단 유럽 뿐만 아니라 통계를 생산하고 관리하는 모든 국가의 부처에서 우려하는 부분이다.

첫 번째는 산업구조의 반영이다. 2000년대 들어 다양한 신기술의 등장과 더불어 제조 중심의 사회에서 서비스 중심의 사회로 변화함에 따라 매년 새로운 유형의 산업이 등장하기 시작하였다. 불행하게도 국가통계는 유동성을 가진 팩터에 관한 반영에 보수적으로 접근할 수밖에 없는 한계가 존재한다. 통계의 시계열성을 보존하기 위해 조사통계의 근간이 되는 분류체계의 빈번한 변경을 최대한 지양하다보니 실제 사회를 진단하는데 통계 결과 활용이 적합하지 않은 문제가 발생하였다.

두 번째는 기업 활동의 형태 다양화이다. 확정된 오프라인 공간을 두고 사업을 영위하는 전통적인 기업의 업태와는 달리 최근에는 온라인 기반의 사업이 활성화 되어 있어 통계 조사를 위한 소재지 파악 측면에서 점점 어려워지는 형국이다. 또한 지역에 기반하지 않는 산업이 증가하고 있고 사무실의 내선번호가 별도 존재하지 않는 기업이 상당수로 점점 더 모집단을 구축하기 어려워지는 상황이다.

세 번째는 조사별 응답자 중복에 의한 회수율 악화이다. 설문 회수율은 통계의 품질을 결정하는데 상당히 중요한 지표이다. 회수율이란 모집단을 설계하고 이에 기반해 추출한 표본집단에 관한 실사 과정에서 실제 응답이 이루어진 비율을 의미한다. 국가통계의 수가 증가함에 따라 한 응답자가 다수 통계조사에 응해야 하는 환경이 조성되기도 하는데, 이것은 응답자의 설문 응답 부담을 늘리는 문제를 야기한다. 다수의 통계 조사에서 50% 회수율을 넘기는 사례는 찾아보기 힘들기 때문에 조사 통계의 품질 악화 이슈도 끊임없이 제기되고 있는 실정이다.

네 번째는 즉시 활용 가능한 민간 중심의 데이터 증가이다. 국민 관점에서는 국가통계의 활용을 선호하지 않는다. 왜냐하면 상기 언급했던 바처럼 국가통계는 통계 조사의

분류체계가 유연성을 띄기 어려우며 품질 제고 등으로 인한 공표 주기가 비교적 길기 때문에 시의성을 중요하게 생각하는 사안들에 대해 대체적으로 좋은 참고 지표가 되기 어렵다. 오히려 민간 기업을 중심으로 한 단발성 통계조사나 특정 서비스 영역의 소비자를 독점하고 있는 기업의 현황 통계를 참고하는 사례가 늘면서 국가통계 자체의 가치에 위협이 되는 상황이다.

이와 같은 이유로 유럽 각국은 AI 또는 빅데이터 기술을 통계에 도입해 일종의 대안통계를 제공하는 것에 관심을 기울였고, 관련하여 다양한 논의들이 이루어지고 있다. 물론 이를 어떠한 접근 방식과 관점에서 추진하느냐에 관한 시각은 국가별 상이하며 국가마다의 데이터 공개 규정 및 일반적인 인식 등에 따라 논의의 수준도 다르다. 관련 연구들의 시작은 UN유럽경제위원회(UNECE)를 중심으로 이루어졌는데, 논의의 주제는 국가통계의 현대화로 명명된다. 빅데이터가 세간의 주목을 받기 시작했던 2014년 이후 관련 위원회에 소속된 국가들은 이를 발전시키기 위한 논의를 시작했다.

- **아일랜드**

아일랜드는 2014년 UNECE의 국가통계 현대화 일환의 시범 프로젝트에서 자국의 슈퍼컴퓨팅 연구소(ICHEC)의 컴퓨팅 자원을 지원함으로써 타 국가대비 이와 관련한 초기 이슈를 선점한 국가다. 그렇다보니 프로젝트 수행과정에서 파생된 연구 논문들을 중심으로 관련 논의의 물꼬를 텄다는 점에서 주목해볼 필요가 있다. 특징적인 점으로는 빅데이터 기술을 통계에 활용하는 유형으로 4가지를 제안했다는 점으로 각각의 유형별 기회요인과 위협을 분석 및 제시하였다.

① 기존 통계의 완전/부분 대체 : 현재 공표되는 통계 조사 중 빅데이터를 활용해 유사한 지표를 도출할 수 있는 경우가 이에 해당하며 대체를 위해 기존 통계 생산 방식과 빅데이터 생산 방식 간 비용 효율 및 대표성, 정확성 등을 비교해볼 필요

② 추가적인 통계 정보 제공 : 현재 공표되는 통계 조사에서 도출되기 힘든 수치들(오차범위 밖에 존재하는 미시통계, 함께 조사 시 효과적인 부가지표 등)을 빅데이터 활용을 통해 제공함으로써 기존 통계 자료를 뒷받침할 수 있는 통계 개발

③ 통계 추정치의 개선 : 기존 통계 조사 과정에서의 잘못된 값을 추정하는 과정 및 설문 무응답에 관한 데이터 보정 등 통계적 추정이 필요한 절차에서 추정치의 품질을 높이기 위해 빅데이터를 활용

④ 완전히 새로운 통계 정보 제공 : 행정통계, 조사통계 방식으로는 도출되기 어려우나 정보의 수요가 발생하는 다양한 사안에 관해 빅데이터 기반의 통계 생산을 통한 해결책을 탐색하는 것

아일랜드 국립대학의 Rob Kitchin 교수는 2015년 학술저널에서 빅데이터를 국가통계에 활용함에 따른 기회 및 위협 요소를 도출하였는데, 이는 총 3가지의 범주: 데이터, 품질/기능, 생산성/효율성 측면으로 요약가능하다.

<표2-1> 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

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

해당 문헌은 통계의 빅데이터 활용을 일련의 파괴적인 혁신으로 규정한다. 이를 뒷받침하듯 Letouze & Jutting의 2014년 저서에서는 빅데이터를 활용하는 것이 단순히 기술적 고려사항이 아니라 정책적인 의무로 받아들여야 함을 시사하였다. 또한 '빅데이터 샌드박스(Big data Sandbox) 환경 구축을 강조하였는데 이는 아일랜드가 UNECE의 국가통계 현대화 프로젝트에서 기여한 중요한 역할 중 하나다.

빅데이터 샌드박스란 유로연합에 소속된 국가들의 빅데이터를 아일랜드의 슈퍼컴퓨팅 연구소의 데이터베이스로 함께 공유하여 데이터베이스 접근 권한을 가진 조직 내

에서 다양한 통계 실험을 추진하고 그 성과를 공유하는 장이다. 여기에서는 몇 가지 정형화 된 공통 목표를 공유한다.

① 원격 접근 및 처리의 타당성 테스트 : 전 세계의 통계 조직은 중앙 서버에 있는 빅데이터에 접근하고 이를 분석할 수 있다. 이 접근 방식은 현실성을 가지는가? 또는 이를 위해 해결해야할 현안은 무엇인가

② 기존 통계 표준/모델/방법론 : 빅데이터를 활용하는 통계 생산에서 기존 통계 표준 및 모델, 방법론을 동일하게 적용할 수 있는가

③ 소프트웨어 : 통계 조직에서 유용하게 활용할 수 있는 빅데이터 활용 소프트웨어 도구는 무엇인가

④ 빅데이터 활용 기법의 실효성 : 빅데이터의 잠재적인 활용을 가정하였을 때 장단점은 무엇인지, 기계학습 등 AI기법의 활용가능성 등

⑤ 국제 협력 기구 : 빅데이터 활용의 기술적 측면에 대한 아이디어와 경험 공유를 위한 국제 협력 기구를 구축

이와 같은 논의 결과 아일랜드에서는 현실화 측면에서 해결해야할 정책적 난제가 존재함을 인정하고 있다. 특히 빅데이터의 관리, 기술 표준화, 방법론 등의 접근방식과 예외처리에 관한 규정의 필요성, 새로운 규제 환경 내에서 빅데이터의 품질을 인증하는 주체는 누가 되어야 하는지, 빅데이터의 라이선스 문제, 단기적인 실험에서 그치는 빅데이터 활용 통계를 장기적 관점에서 유시하기 위한 모범사례 발굴 등을 해결해야할 과제로 꼽고 있다.

· 스위스

스위스는 국가통계의 효과적인 정책 지원을 위해 무엇이 보완되어야 하는지에 관해 논한다. 그들은 빅데이터를 새로운 통계의 패러다임으로 인식하기보다는 통계 산출의 논리적 구조를 다양화하기 위한 도구로서 바라보는 것이 특징이다. 정책적 의사결정에 도움이 되는 지표 산출을 위해서는 연역적 추론과 귀납적 추론 결과가 함께 고려되어야 효과적임을 강조하고 있으며 이런 측면에서 데이터 과학 이론에 주목하고 있다.

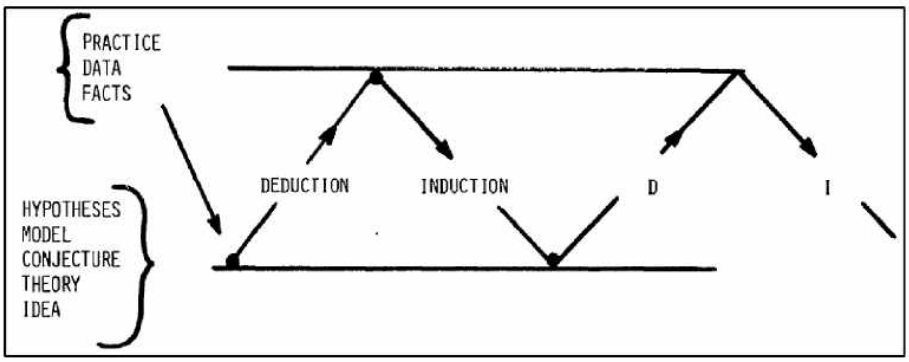
<표2-2> 일반 통계와 데이터 과학의 추론 방식에 따른 구분(스위스)

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)

스위스의 입장은 유로연합 통계국(EuroStat)의 국가 통계 기관장 그룹(DGINS)의 문헌을 통해 여실히 드러난다. 2018년 DGINS 문서에 의하면 그들은 전통적으로 통계가 타당성을 설명하고 확인하기 위해 수집된 1차 데이터를 분석하는 것과 관련되어 있다고 해석한다. 세부적으로는 특정한 가설을 이론적 개념 하에 세운 후 평가하거나 테스트하는 것을 목표로 하므로 이러한 분석 패러다임을 아이디어(이론)에 우선한 하향식(Top-down) 추론, 즉 연역적 추론으로 정의하고 있다. 반면 데이터 과학은 다른 이유로 수집 된 데이터 또는 관찰·탐색을 통해 발견한 2차 데이터를 분석하는 것으로 정의한다. 이는 특정 가설 또는 이론을 평가하거나 검증하기 위해 통제되는 데이터가 아니며, 반대로 해당 데이터들을 통해 새로운 가설 또는 이론을 창조하는 상향식(Bottom-up) 분석법으로 해석하고 있다. 이러한 분석 패러다임은 이론보다는 데이터에 우선하는 귀납척 추론으로 정의된다. 그들은 연역적 추론과 귀납적 추론은 상호보완적 역할을 할 수 있기 때문에 데이터 기반의 의사결정을 가능하게 하는데 두 가지 통계 생산방식을 병행해야 함을 주장하고 있다. 정보에 입각한 정책 결정의 과정은 아래 그림과 같이 해석하고 있다.

<그림2-1> 훈련데이터-가설 모델 간 지속적 보완 사이클



* 출처 : DGINS(2018)

• 독일

독일 연방의 경우 이론적 논의 보다는 국가 전략 일환의 실증 연구가 중심이 되고 있어 타 국가 대비 전향적인 도입이 이루어지는 상황이다. 그들은 2017년 말 국가차원의 디지털 어젠다 59건을 선정하였는데, 그 중 우선시되는 프로젝트로 국가통계의 머신러닝 기술 도입에 관한 검증이 채택되었다. 표면화 된 프로젝트이고 2019년까지 문헌연구, 실증연구 등이 함께 수행되었기 때문에 좀 더 구체적인 실험 등을 엿볼 수 있다. 특징적인 부분으로는 빅데이터 활용 관점을 건너뛰고 인공지능 기술 도입에 주안점을 두고 있다는 점인데, 다른 국가들과는 다른 파격적 접근이다. 인공지능 기술을 통계에 적용한다는 의미는 기술의 특성 상 통계 데이터 산출 방법의 변경에 주안점을 둔 전략으로 볼 수 있다.

<그림2-2> 머신러닝 도입을 통한 기대효과

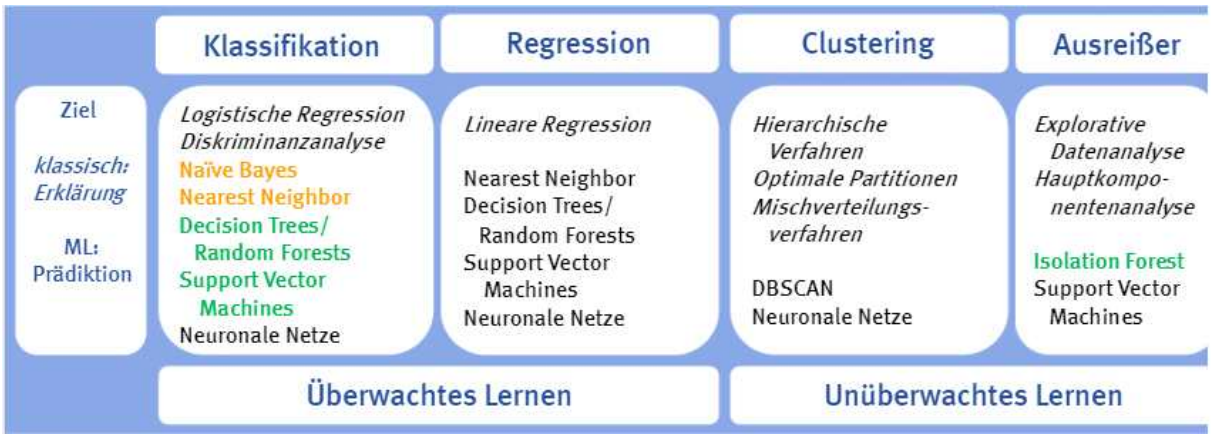
*출처: 독일연방통계청(2019)

그들은 머신러닝 기술 도입을 통한 통계분야의 긍정적 효과로서 3가지에 주목한다. 첫 번째는 통계 품질의 향상이다. 이는 기존 통계 생산 절차에서 주관적인 판단이 요구되는 특정 절차들을 머신러닝으로 대체해 더욱 객관적이고 정확성을 가지는 결과를 산

출하거나, 기존의 통계를 신뢰할 수 있는 빅데이터에 기반한 통계로 대체함으로써 품질을 높이는 것을 의미한다. 두 번째로는 보다 효율적인 프로세스를 구축하는데 있다. 전통적 통계를 생산하는 전반적 과정은 매 생산주기마다 동일한 업무라도 같은 자원 소요를 수반하는데 반해, 머신러닝의 도입은 잠재적으로 이러한 반복 작업을 자동화 할 수 있음에 주목한다. 세 번째는 분석 옵션의 다변화이다. 머신러닝 분석의 기반이 되는 빅데이터는 설문조사와는 달리 특정한 목적에 맞추어 선별 제작된 데이터가 아니다. 그렇기 때문에 이를 기반해 생성된 통계는 자연스래 연계성을 가지는 타 데이터와의 분석 여지를 가지게 된다. 또한 머신러닝 기법은 분석 목적에 따라 다양한 알고리즘 저변을 보유하고 있다. 새로운 방법론의 적용은 결과적으로 이러한 알고리즘을 적재적소에 활용할 수 있다는 장점을 가진다.

세부적으로 살펴보면, 그들은 머신러닝 알고리즘 중 활용을 고려해 볼법한 대표적 기술 군을 정의하고 있다. 아래는 독일 연방 이 정의한 4가지 머신러닝 기술군에 해당하는 대표적 알고리즘이다.

<그림2-3> 머신러닝의 4가지 기술 분류별 대표적 알고리즘(독일 연방 , 2019)



*출처: 독일연방통계청(2019)

· 네덜란드

네덜란드는 여타 국가들과 마찬가지로 빅데이터 활용가능성에 초점을 맞추고 있다. 그들은 2020년 보고서를 통해 빅데이터가 만연한 사회 환경의 변화를 시사하며, 빅데이터에 통계의 미래가 있음을 인정한다. 그들은 빅데이터 활용 통계를 소프트웨어적 관점이 아닌 기존의 통계 방법론상으로 해석하려는 시각을 견지하고 있는데, 빅데이터 활용을 활성화하는데 동의하면서도 통제할 수 없음을 우려한다. 결과적으로 그들은 빅데이

터 기반의 통계를 두 가지 관점에서 구분해 바라보고 있음을 알 수 있다.

<표2-3> 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)

빅데이터의 불완전성은 쉽게 품질을 증명하기 어려워 데이터의 신뢰성을 확보하는데 한계가 있음에서 기인한다. 가령 특정 통계를 지속적으로 생산하기 위해 통계의 기반이 되는 목표 모집단부터 표본 설계까지의 절차를 검증하기 위한 다양한 방법론이 활용된다. 그에 반해 빅데이터는 원 데이터 자체의 품질을 진단하는 반면 분석 과정에 관한 검증이 없다보니 연속성을 가지는 통계 생산의 관점에서 과거 산출한 빅데이터 기반 통계와 신규 산출 결과를 동일 기준에 놓고 해석하는데 논쟁의 여지가 발생한다. 또한 빅데이터는 데이터의 방대한 규모를 통해 객관성을 확보하는 특성을 가지는데, 여기에서의 데이터 규모라는 것이 유동적이라는 것에 눈여겨볼 필요가 있다. 규모 변화에도 불구하고 균일한 분포임이 검증되어야만 이를 통해 산출되는 결과의 신뢰성이 보장 가능할 것이나 이에 대한 고려가 부족하다. 상기 이유들로 인해 다양한 비약이 발생될 여지가 있다. 그럼에도 불구하고 빅데이터는 존재하는 사실 자체를 여과 없이 드러내는데 그 가치가 있다는 의견도 존재하므로 빅데이터의 활용 과정에서 통계적 해석이 다소 불편할지라도 결과 자체를 폄훼하기 곤란함을 해당 문헌은 고려하고 있다.

네덜란드 통계청은 통계적 모델링을 통해 빅데이터를 통제하는 방법 또한 제시한다. 이와 같은 논의의 배경은 최근 들어 수리통계, 응용통계학 분야에서 빅데이터를 통계적으로 해석하기 위한 방법론들이 다수 개발되었다는데서 비롯된다. 전통적인 통계 기법은 신뢰성 확보를 위한 확률적인 검증 방법이 구조화 되어 있으므로 이를 준용하는 것이 안전하다고 판단한다. 기계학습 기술과 같은 신규 방법론 또한 기존의 통계 기법과 함께 활용 가능하도록 그들을 수정하는 것이 가능하다고 보는 시각이다.

• **일본**

일본은 일찍이 유로연합과 긴밀한 협업을 통해 국가통계 현대화를 위한 다양한 프로젝트에 관한 실질적인 정책 반영을 꾀한 국가이다. 유로연합은 소비자 물가 통계를 민간 소매점의 스캐너 데이터를 활용해 대체하는 것을 권장해왔다. 같은 맥락에서 유로연합 국가들은 스캐너 데이터를 활용해 소비자 물가 통계를 대체하려는 시도를 추진하였는데 일본 또한 이를 시도하였다. 단지 실험 차원에서의 도입이 아니었던 것이 괄목할 만한 부분인데, 정부차원에서 해당 통계를 국가통계로서 인정하였다. 이와 같은 지원에 힘입어 그들은 국가통계에 빅데이터를 활용 시 어떤 장점과 단점이 발생하는지를 실증적으로 분석할 수 있었다. 간단히 요약해 그들은 빅데이터 통계가 시의성과 적시성을 가진다는 점에서 장점을 가지나 정확성과 편중성 측면에서 아직까지는 해결과제가 존재함을 인정하였다.

<표2-4> 민간 부문 빅데이터 활용을 통한 기존 통계 대체가 주는 장단점(일본)

장점	단점
- 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)

빅데이터 활용 통계의 장단점을 언급한 부분은 여타 유럽 국가들에서도 그 사례를 쉽게 찾아볼 수 있으나 일본이 지목한 단점들은 타 문헌에서 직접적으로 언급되지 않았던 것들이라 눈여겨볼 필요가 있다. 첫 번째로 편중성의 문제로 편중성은 현재의 빅데이터 및 인공지능 솔루션에서 대표적으로 해결되어야 할 현안으로 주목받고 있다. 빅데이터 분석 기법 및 인공지능 알고리즘의 한계는 데이터를 있는 그대로 학습하여 정확도를 높이는 것을 목표로 한다는 점에 있다. 이는 알고리즘이 성능적으로 최적화되더라도 분석 기반이 되는 데이터가 특정 편향을 가질 시 이러한 편향을 결과에 그대로 반영한다는 문제로 확대된다. 객관성을 보장해야하는 통계 분야에서는 특히 중요한 부분

이라 볼 수 있으며, 충분히 관련 주제로 논의해볼 가치가 있다. 두 번째로는 데이터의 안정성 문제이다. 실제 소비자 물가 통계에 스캐너 데이터를 활용해 더욱 효율적인 통계 데이터 생산이 가능하다면 원론적으로 문제 여지는 없다. 반면 민간 기업의 데이터에 의존한다는 점은 국가가 기업의 지속성을 담보해야한다는 문제에 직면한다. 가령 국가의 대다수 소매 기록을 담당하는 특정 기업이 존재한다고 하였을 때, 해당 기업의 존폐여부에 의해 소비자 물가 데이터의 시계열 안정성이 흔들릴 수 있다. 이와 같은 부분은 분명 실질적인 빅데이터 활용 통계 개발에서 염두에 두어야 할 부분으로 판단된다.

· 한국

한국 은 통계 작성 기본 원칙 및 실무 가이드라인을 통해 꾸준히 빅데이터 활용 통계의 승인 가능성 및 관련 시도 활성화에 대한 검토를 진행해왔다. 그럼에도 불구하고, 2019년까지의 공식적인 입장은 빅데이터 활용 통계 자체의 신뢰성과 각종 해결 안건으로 인해 국가통계로의 승인에 한계가 있음을 시사했다. 그러나 올해 통계청은 통계 관리체계의 개정을 통해 빅데이터 활용 통계를 국가통계의 관리 체계 내 포함하기 위한 제도 개편을 예고하였다.

<표2-5> 빅데이터 통계 관리를 위한 유형별 시나리오 평가

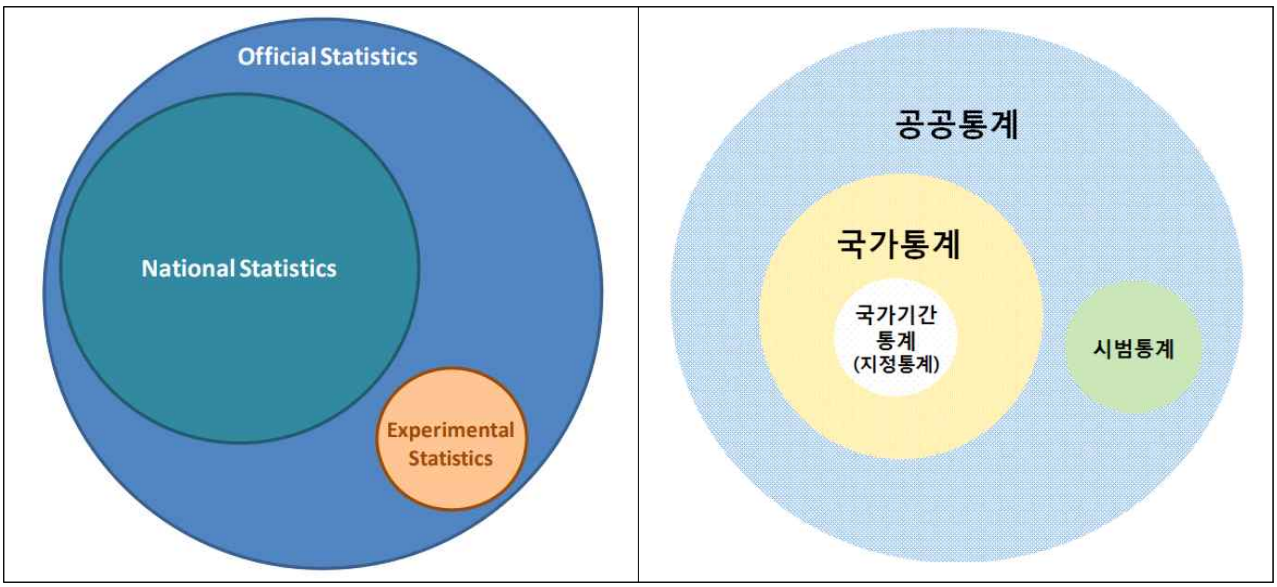
구분	법령 미개정 (지침, 매뉴얼 위주)	법령 개정 (통계법, 시행령, 시행규칙 등 개정)
현 관리체계 유지	- 법 개정 없이 빅데이터 활용 통계 관리방안 제시(△)	- 법 개정을 통한 빅데이터 활용통계 승인관리방안 제시(○)
관리체계 개편	- 법 개정 없이 국가통계관리 체계 개편(안) 적용 가능성 검토(X)	- 법령 개정을 전제로 공공통계, 시범통계 관리방안 제시(○)

*출처: 정보통신정책연구원(2020)

논의의 핵심은 통계 관리체계이다. 빅데이터 활용 통계는 현재 국가통계로 승인되는 경우가 극소수이고 승인 되는 사례 또한 한시적 승인에 가까워 전무하다고 보는 게 타당하다. 문제는 통계 관리체계는 통계법으로 명시되어 있는 사안이고 해당 법령을 개정하지 않는 이상 빅데이터 또는 인공지능 활용 통계를 위한 별도의 제도를 적용하기 어렵다는 점이다. 현재 관리체계를 유지하는 선에서 법령 개정 없이 빅데이터 활용 통계

의 관리방안을 제공할 시, 본질적으로 국가통계로의 승인 여부가 모호해질 것이므로 실효성이 없는 가이드라인이 될 가능성이 농후하다. 그러므로 현 관리체계의 유지여부와 무관하게 법령 개정이 필수적이다.

<그림2-4> (좌)영국의 공식통계 개념도와 (우)한국의 시범통계 도입 구상도



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

올해 5월 에서 발표한 내용에 따르면 조사통계 및 행정통계가 아닌 그 외 통계의 경우에도 국가통계로의 승인심사를 가능하도록 심사기준 문구를 변경하는 선에서 제도 개편이 이루어질 것으로 보이며, 여기에는 법령 자체의 개정은 예정되어있지 않다. 그러나 향후 영국의 공공통계 제도를 밴치마크하는 방안이 논의될 것으로 보이는데 이는 통계법 개정과 함께 추진되어야 하는 부분이다. 여기에서 신설되는 시범통계란 빅데이터 활용 통계를 국가 관리범위 내 포함시키기 위한 제도이다. 실질적인 논의는 2021년 중 활발히 진행될 것이라 예상된다.

제2절 UN의 국가통계의 머신러닝 활용 논의

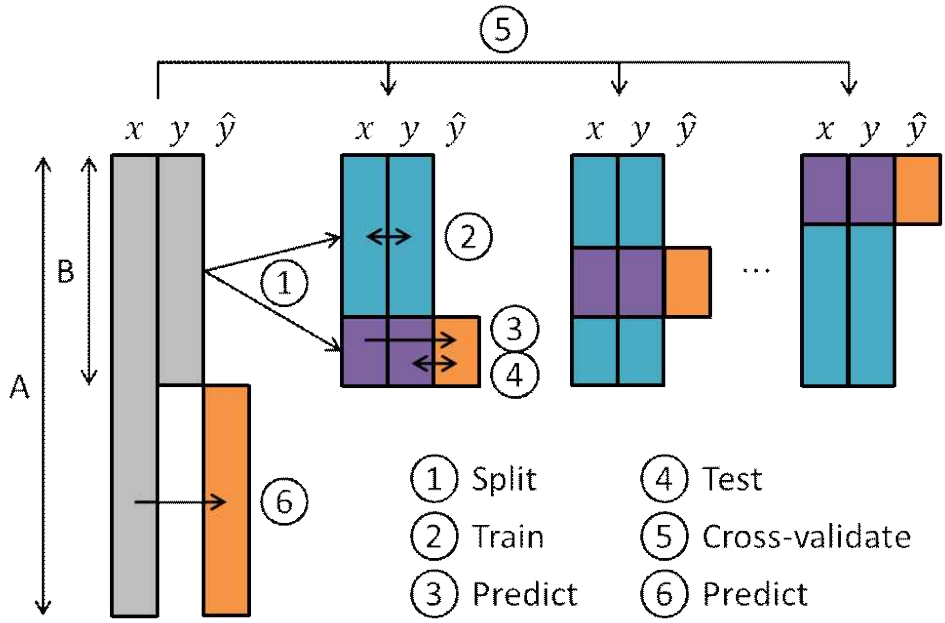
UN유럽경제위원회의 머신러닝 분과는 2018년 말 국가의 관리자 및 정책 입안자에게 국가통계의 생산에 머신러닝을 사용할 수 있는 기회에 대해 알리고, 머신러닝에 익숙하지 않은 국가 통계 전문가에게 이해를 돕기 위한 문서를 공개하였다. 해당 문서는 유럽 국가 전체를 대상으로 통계 생산에 머신러닝 활용을 직접적으로 추천하였다는 점에서 의의가 있다. 현재 국내외를 비롯한 국가들 다수가 빅데이터의 통계 활용을 검토하는 수준에서 머무르는 게 현실이나, 여기에서 한 발 더 나아간 논의가 이루어지고 있다는 점에서 주목해볼 필요가 있다. 해당 문서는 국가통계에 머신 러닝을 사용할 수 있는 두 가지 유형을 소개하고 있다. 또한 관련 유형의 통계 정확도에 대한 직접적인 고찰이 포함되어 있다. 본 절은 이와 관련된 문서의 내용을 요약 및 소개한다.

• 머신러닝의 정의와 활용 방향

머신 러닝(ML)은 컴퓨터가 명시적으로 프로그래밍 된 규칙(rules)을 준용하는 대신, 학습하고 습득한 지식을 새로운 환경에 일반화하도록 하는 학문이다. 보조 정보와 관심 변수(레이블이 지정된 입력)가 둘 다 존재한다면 성능 검증이 가능하므로 지도 학습을 수행할 수 있다. 보조 정보만 있거나 레이블이 지정되지 않은 입력만 있는 경우에는 피드백이 제공되지 않는 비지도 학습을 해야 한다. 지도학습에서는 관심 변수가 정성적/범주인 경우 컴퓨터가 분류 문제를 해결하는 방법을 학습해야하며, 관심 변수가 정량적/숫자인 경우 컴퓨터는 회귀 문제를 해결하는 방법을 학습한다. 비지도학습의 경우 군집화 문제를 해결하는 과정으로 정의된다.

아래 그림은 지도학습의 도식적 개요를 나타내며, 이는 국가 통계와 관련된 통계 생산의 맥락으로도 해석할 수 있다. 국가 통계에서는 표적 모집단에서 관심 변수를 정확히 추정하는 것을 목표로 한다. 여기에 해당하는 예로 청년층의 실업률 또는 항공 산업의 이산화탄소 배출량 추정 등이 있을 수 있다. 보조 정보 x는 표적 모집단 A의 모든 요소에 대해 확인되지만 관심 변수 y는 표본 B의 요소에 대해서만 관측된다. 머신 러닝을 사용하여 표본 B를 통해 모집단 A에 대한 추론을 수행할 수 있다.

<그림2-5> 지도학습의 유형에 관한 도식



*출처 : UNECE(2018)

머신 러닝은 그 외에도 다양하게 사용할 수 있다. 예를 들어 A를 확률 표본으로, B를 응답자 집합으로 간주하고 머신 러닝을 사용하여 단위 비응답을 수정할 수 있다. 다른 시나리오의 경우 A는 응답자 집합, B는 항목 x와 y에 대해 응답한 집합이며, 항목 y에 대해 항목 비응답을 대체하는 데 사용가능하다. 네 번째 옵션은 A에 대리 변수 x의 관측 값이 있고 B에는 관심 변수 y의 관측 값이 있는 경우로, 머신 러닝을 사용하여 측정 오차를 모델링할 수 있다. 마지막으로, A가 시계열이고 B가 과거 데이터인 경우 실시간 예측(nowcasting)에 사용된다.

구체적으로 지도학습은 6개 단계로 구분된다.

1. 레이블이 지정된 입력(B)은 임의로 훈련 세트(파란색)와 테스트 세트(보라색)으로 분할
2. 모델 또는 알고리즘은 훈련 세트에서 x와 y 간의 관계를 학습
3. 모델 또는 알고리즘은 훈련 세트의 x에서 y, $\hat{y}$(주황색)를 예측하는 데 사용.
4. 예측 값 $\hat{y}$는 훈련 세트에서 관측값 y와 비교하여 평가된다. 예측 오차가 최소화될 때까지 여러 (초)매개변수를 사용하여 2 ~ 4단계를 반복한다.
5. 과적합을 방지하기 위해 각 분할에 대해 1 ~ 4단계를 반복한다. 내부 루프에서 초

매개변수(hyper parameter)를 조정하고 외부 루프에서 일반화 오류를 추정하기 위해 중첩 교차 검증을 수행할 수 있다

6. 비 관측 값은 모델 또는 알고리즘을 사용하여 표본 외부의 x를 통해 예측하며, 대체로 분할에서 가장 작은 예측 오차를 제공

중요한 단계는 훈련 세트에서 x와 y 간의 관계를 학습하는 것이다(2단계). 이를 위한 다양한 알고리즘이 존재하며, 예를 들어 랜덤 포레스트, 신경망, 서포트 벡터 머신 등이 고려될 수 있다. 이러한 알고리즘은 인과 설명이 아니라 예측 오차를 최소화하는 데 초점을 맞춘다는 점에서 기존 회귀 기법과는 다르며, 고차원 공간에서 비선형 관계를 모델링하는 데 더욱 적합할 수 있다. 또한 x는 디지털 형식의 데이터라는 조건만 충족하면 기존의 숫자 정보부터 자연어 텍스트, 음성 녹음, 이미지 또는 동영상 모두 가능하다. 다양한 성능 지표를 사용하여 예측 오차를 수량화할 수 있으며, 대표적으로 회귀에서의 평균 제곱 오차와 분류에서의 Matthews 상관 계수를 들 수 있다.

지도학습의 성공적인 사용은 x의 예측 능력과 훈련 데이터의 크기에 따라 결정된다. 표본 설문조사 데이터는 훈련 데이터 크기가 제한적이지만(n) 일반적으로 보조 변수의 수가 많다(P). 한편, 빅데이터는 대량의 데이터(N)에서 이름이 유래했지만, 대부분의 경우 변수의 수는 제한적이다(p). 따라서 N과 P를 사용하는 행정 통계에서 가장 유용할 것으로 기대할 수 있다.

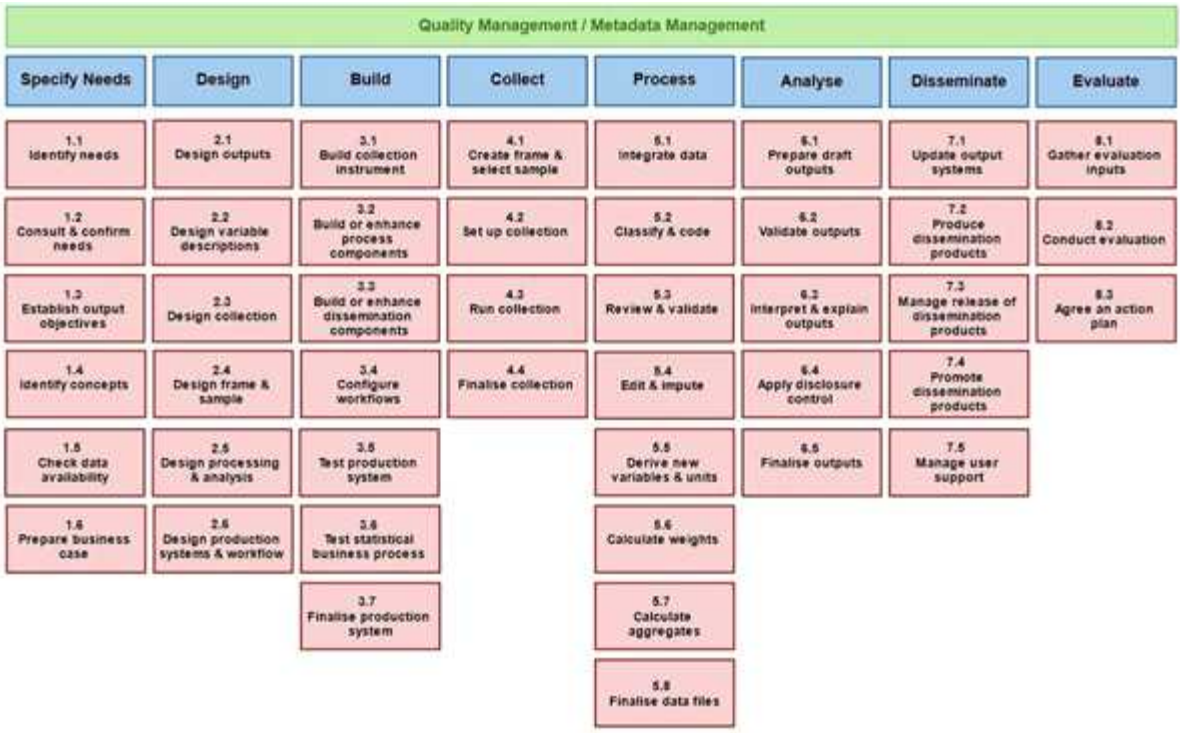
· 1차 데이터(조사통계) 생산에 머신러닝 활용

일반 통계 생산 과정을 모델화한 가장 대표적인 자료는 GSBPM이다. 이는 UN유럽경제위원회의 후원 하에 개발된 표준 통계개발 방법론이며, 국내외를 불문 해당 표준을 준용해 각국에 실정에 맞는 방법론 및 모델을 지정하고 있다. GSBPM은 다양한 통계 생산에서 참조해볼 수 있는 표준일 수 있으나 통상적으로는 조사통계 생산 및 관리에 적합한 표준으로 알려져 있다. 해당 문헌은 GPBPM의 생산절차에 맞추어 1차 데이터(조사통계 기반의 통계 생산) 처리 기준에서의 머신 러닝 활용에 대해 다루고 있다.

그림과 같이 GSBPM은 8단계로 구성되어 있으며 각 단계에는 하위 프로세스가 있다. 첫 번째 단계(요구사항 지정)는 데이터 사용자를 참여시켜서 데이터 요구사항을 확인하게 하는 것과 관련된 활동을 다루기 때문에, 이 단계에서는 머신 러닝을 적용할 기회가

그다지 많지 않다.

<그림2-6> GSBPM 프로세스 모델



두 번째 단계(설계)에는 통계 결과, 개념, 방법론, 수집 도구 및 작업 프로세스가 정의되어 있다. 하위 프로세스 2.4(프레임 및 표본 설계)는 머신 러닝 적용 가능성이 있는 영역이다. 표본으로부터의 데이터 수집과 관련된 통계 프로세스에만 적용되는 이 하위 프로세스는 표본 추출 프레임 식별을 포함한다. 표본 추출 프레임의 소스에는 레지스터, 관리 데이터, 인구 조사 및 기타 설문조사가 포함됩니다. 이러한 소스는 군집화 알고리즘을 사용할 수 있는 레코드 연계 프로세스를 통해 결합될 수 있다. 표본 추출 프레임의 준비 시 무엇보다도 중요한 것은 설계 정보의 품질(산업, 지역, 활동, 직업 등)이다. 이러한 '코드화 활동'은 분류 알고리즘을 적용하는 것을 최우선적으로 고려할 수 있다. 코드화 유형 활동뿐만 아니라 프레임의 설계 정보 품질을 검증하는 경우에도 머신 러닝을 사용할 수 있다. 예를 들어 군집화 방법을 사용하여 설계 정보의 이상 값을 찾을 수 있다.

네 번째 단계(수집)에는 프레임 개발, 표본 추출, 수집 실행의 하위 프로세스가 포함되어 있으며, 그 중 일부에 머신 러닝을 적용할 수 있다. 여러 설문조사에서는 더욱 정확하게 결과를 추정하기 위해 프레임을 계층화한다. 하위 단계 4.1에서 분류 기법을 사

용하여 모집단을 계층화할 수 있다.

다음에 머신 러닝을 적용할 수 있는 단계는 4.3이다(수집 실행). 실제 수집 작업에 사용하기는 어려울 수 있지만, 수집 활동의 관리에 사용할 수는 있다. 많은 국가 통계 기관은 수집 프로세스의 효율을 높이기 위해 적극적인 수집 관리 전략을 채택하고 있다. 이러한 전략 중 다수가 추정 또는 예측을 요하는 응답 확률을 사용한다. 전체 표본에 사용 가능한 정보를 통해 개별 단위의 응답 확률을 예측하는 경우 기존 방법(예: 로지스틱 회귀)와 머신 러닝 방법(예: 회귀 알고리즘)을 둘 다 사용할 수 있다. 이렇게 예측한 응답 확률을 사용하면 수집 활동을 가장 효율적으로 관리할 수 있다. 응답 확률을 추정하는 데 사용되는 응답 성향 모델의 적합도를 개선하기 위해, 일반적으로 비슷한 행동을 보이는 단위 그룹에 대해 해당 모델을 추정한다. 분류 방법은 각 그룹의 성향 모델을 추정하기 전에 이러한 그룹을 '최적으로' 정의하는 데 사용할 수 있다.

적응적 수집 설계 시 설문조사 방법, 인센티브 사용 등의 기능을 하위 모집단에 맞게 조정할 수 있다. 보조 정보 또는 파라데이터를 군집화, 분류 또는 회귀 알고리즘과 함께 사용하여 응답률 향상을 위해 단위를 여러 개의 하위 모집단으로 분류할 수 있다.

또한 데이터 수집 중에(특히 전자적 수집 사용 시) 데이터를 보고 시점에 검증할 수 있다. 수집 중에 군집화 방법을 사용하여 이상하거나 오차가 있는 데이터 포인트를 찾아서 응답자가 수정하도록 하거나, 감독형 머신 러닝을 사용하여 올바른 값을 예측할 수 있다. 마지막으로, 많은 국가 통계 기관이 응답자가 의견 또는 문의사항을 기록할 수 있도록 한다. 자연어 처리 도구를 사용하여 이러한 항목을 처리하고 즉시 응답해야 하는 항목을 찾을 수 있다.

다섯 번째 단계(프로세스)에서는 머신 러닝 방법을 사용할 기회가 많다. 하위 프로세스 5.1(데이터 통합)에서는 여러 소스의 데이터를 통합한다. 따라서 모든 데이터 소스에서 공통된 고유 ID를 사용할 수 없는 경우 확률론적 레코드 연계 기법을 사용할 수 있다. 통합을 완료한 후에는 머신 러닝 방법을 사용하여 데이터를 정리할 수 있다(이상값, 오차, 불일치하는 레코드 등을 식별). 하위 프로세스 5.2(분류 및 코드화)에서는 데이터를 표준 분류(산업, 지역, 상품 등)로 분류 및 코드화하며, 머신 러닝을 활용할 수 있는 전형적인 부분이다.

하위 프로세스 5.4(편집 및 대체)에서는 잘못된 데이터, 결측 데이터 또는 신뢰할 수 없는 데이터를 대체한다. 감독형 머신 러닝을 사용하여 결측 데이터 또는 잘못된 데이

터를 대체할 수 있다. 동질적인 단위를 대체 클래스로 분류하면 대체 결과를 개선할 수 있기 때문에, 층화에 사용되는 것과 유사한 머신 러닝 방법을 사용하여 이러한 클래스를 정의할 수 있다.

하위 프로세스 5.6(가중치 계산)에서도 대체 클래스와 비슷한 방식으로 클래스가 생성되지만 목적은 조금 다르다. 단위 비응답이 존재하는 경우 재가중 클래스는 응답 확률이 유사한 단위로 구성된다. 일반적으로 로지스틱 회귀 모델을 사용하여 이러한 응답 확률을 계산한 다음 해당 응답 성향을 기준으로 동질 그룹이 구성된다. 이러한 그룹의 정의 작업에도 머신 러닝 방법을 사용할 수 있다. 2차 데이터가 있는 경우, 여러 설문조사에서 위에 언급한 머신 러닝 방법을 통해 정의할 수 있는 여러 보정 그룹에 보정 측정 법칙을 사용한다. 대체클래스 및 보정 그룹을 정의하는 경우 머신 러닝 기법을 통해 최상의 변수를 선택하여 클래스 또는 그룹 정의(하위 집합 선택)에 사용할 수 있다.

여섯 번째 단계(분석)은 통계 결과를 준비하는 단계로, 결과의 유효성 검증과 해석, 노출 제어를 포함한다. 유효성 검증 및 해석 작업은 일반적으로 분석가가 수행하지만, 이상 추정 값을 찾기 위해 머신 러닝 방법을 활용할 수도 있다. 또한 개인화된 데이터에 적용된 머신 러닝 알고리즘의 분류 오차를 통해 노출 제어 시 프라이버시와 효용 간의 균형을 조정할 수 있다.

· 2차 데이터(가공통계) 생산에 머신러닝 활용

2차 데이터는 통계 목적으로 수집되는 것은 아니지만 국가 통계 기관에 유용한 통계 정보를 포함하고 있을 수 있다. 이러한 데이터는 빅데이터, 레지스터 데이터 및 기타 소스에서 수집된 이러한 데이터의 조합 등이며, 1차 데이터도 포함될 수 있다. 그런 연유에서 해당 부분을 통해 빅데이터, 레지스터 데이터 및 멀티 소스 데이터에 머신 러닝 기법을 사용하는 것에 대해 논의한다.

① 빅데이터

빅데이터의 ‘볼륨’과 ‘다양성’ 차원은 공식 통계의 기존 계산 수단을 무력화할 수 있다. 여기서는 ‘볼륨’과 ‘다양성’으로 인해 야기되는 문제를 살펴보고, 머신 러닝으로 이러한 문제를 해결할 수 있는 가능성에 대해 논의한다.

데이터의 볼륨

설계 기반 및 모델 보조 조사 표본 추출 이론, 회귀 추정 등 국가 통계의 기존 방법은 확률 표본을 처리하기 위해 고안된 것이다. 따라서 이러한 방법을 구현하는 메인스트림 알고리즘은 데이터가 소량이고 고품질인 경우에 적합하다. 그러나 이러한 알고리즘은 일반적으로 계산 복잡도가 높기 때문에 대량의 데이터에서는 성능이 제한된다.

예를 들어 선형 회귀의 경우 n개의 관측값과 p개의 변수가 있는 데이터 집합에서는 최소 제곱 추정법의 시간 복잡도가 O(p3+ n p2)이다. 이는 Newton-Raphson 알고리즘과 같은 방법론을 통해 보정 작업을 처리하는 경우에 더욱 불리하게 작용할 수 있음을 의미한다. 이러한 알고리즘은 계산 강도가 지나치게 높아서 빅데이터에는 실용적이지 못한 문제도 있지만(특히 p가 크거나 n과 p가 모두 큰 경우) 더 큰 문제는 분할 정복(병렬화) 방식을 사용하여 최적화하기가 어렵다는 점이다. 따라서 계산 통계의 여러 메인스트림 알고리즘에서는 MapReduce, Spark/Hadoop 등의 우수한 빅 데이터 처리 모델과 빅데이터 소프트웨어 프레임워크를 쉽게 활용할 수 없다.

계산 강도가 높고 병렬화하기 어렵다는 문제 외에도, 국가 통계의 기존 방법 중 다수는 이상값과 오차 값에 매우 민감하기 때문에 조사 기간 중 조사 데이터를 편집하고 대체하기 위해 엄청난 양의 작업을 수행해야 한다. 대부분 손상과 왜곡이 심하고 제대로 선별되지 않아서 수많은 이상 값과 오차 값이 존재하는 빅데이터의 경우 이러한 문제는 더욱 악화된다. 또한 빅데이터는 그 크기만으로도 완전하고 철저한 데이터 점검과 정리를 수행하기가 불가능하다.

최신 머신 러닝 접근 방식은 기존 통계 방법에 비해 빅데이터 소스의 확장성을 지원하기가 더 용이하다. 이는 처음에는 모순적인 것으로 생각될 수 있다. 랜덤 포레스트나 딥 러닝 등의 우수한 머신 러닝 방법은 계산 강도가 높고 일부 상황에서는 기존 방법보다 더 많은 작업이 필요하기 때문이다. 그러나 머신 러닝 방법은 쉽게 병렬화할 수 있어 빅데이터의 처리 시 가장 먼저 고려할 수 있다. 예를 들어 랜덤 포레스트 방법은 포레스트를 구성하는 트리가 서로 다른 표본에서 독립적으로 작성되고 훈련되었기 때문에 쉽게 병렬화할 수 있다. 따라서 완전 분산된 규격품 형태의 랜덤 포레스트 구현이 다수 존재하기 때문에 빅데이터를 매우 효율적으로 처리할 수 있다.

마찬가지로 딥 러닝 모델도 계산 강도가 높기는 하지만 신경망의 각 층에서 뉴런이 독립적으로 처리되기 때문에 본질적으로 병렬화가 가능하다. 그러므로 딥 러닝 애플리

케이션은 대규모 병렬 컴퓨팅에 최적화된 특수 하드웨어 아키텍처를 활용할 수 있다. 이렇게 하면 계산 시간을 크게 줄일 수 있어 빅데이터에서 매우 복잡한 딥 러닝 모델을 효율적으로 훈련할 수 있다.

또한 다수의 머신 러닝 방법은 일반적으로 대부분의 기존 통계 방법에 비해 이상값과 오차 데이터에 덜 민감하다. 그 이유는 이러한 방법의 첨단 구현에 하위 표본 추출 접근법이 사용되기 때문이다. 앞에서 언급한 대로 랜덤 포레스트 내의 각 트리는 원래 데이터의 하위 표본에 적합한 크기이기 때문에 원래 설명 변수의 무작위 하위 집합만 포함한다. 적합화 후에는 모든 트리의 예측을 집계하거나 해당 평균을 구한다. 하위 표본 추출(훈련 시) 및 평균 계산(예측에서)은 입력 데이터를 암묵적으로 평활화하여 이상 값과 오차의 영향을 크게 완화한다. 실제로 상대적으로 드문 값 조합인 이상 값과 오차로 인해 오염 및 편향되는 하위 표본/트리는 소수에 불과하기 때문에, 전체 모델은 거의 영향을 받지 않다. 딥 러닝 애플리케이션도 매우 유사한 하위 표본 추출 메커니즘을 사용하므로 이상 데이터에 거의 영향을 받지 않는다.

머신 러닝으로의 전환이 타당하다는 것을 뒷받침하는 마지막이자 결정적인 이유는, 첨단 빅데이터 분석 프레임워크에서 선형 회귀와 같은 단순한 기존 모델을 포괄하는 머신 러닝 기법을 채택하기 때문이다.

데이터의 다양성

국가 통계의 조사 데이터와 관리 데이터는 고도로 정형화되어 있다. 기존 처리 파이프라인에서 일반적으로 사용되는 데이터 모델은 “사례 × 변수” 행렬이다. 여기서 사례는 레코드로 표시되고 변수는 레코드의 필드로 표시되며, 레코드에는 고정된 수의 정부호 유형 필드가 있다. 빅데이터의 ‘다양성’ 차원에는 느슨하게 정형화된 데이터와 비정형 데이터까지도 포함된다. 일반화된 선형 모델과 같은 기존 방법은 “사례 × 변수” 데이터에 적합하며, 따라서 관측 변수가 사례별로 다를 수 있는 데이터에는 쉽게 사용할 수 없다. 자연어 텍스트와 같이 완전한 비정형 데이터인 경우 “변수”의 자연 개념이 더 이상 존재하지 않기 때문에 기존 방법을 사용하면 더 큰 문제에 부딪힌다. 대신 원시 데이터에서 의미 있는 특성을 어떻게든 추출해야 한다. 인간 분석가가 수행하는 경우 이 데이터 준비 단계를 특성 공학이라고 한다. 반대로, 일부 머신 러닝 기법(대표적으로 딥 러닝)은 원시 데이터에서 특성을 자동으로 추출할 수 있기 때문에 긴급한 작업을 할 때 유용하다.

② 행정통계

행정 데이터는 흔히 빅데이터로 통칭해서 부른다. 모집단을 나열하면 크기가 큰 데이터 집합이 생성되기 때문이다. 이에 대한 구분 기준은 크기가 아니라 구조이다. 행정는 모집단 내에서 식별 가능한 개체의 전체 목록이다(Wallgren과 Wallgren, 2007). 이러한 정의를 적용하면 센서 데이터는 식별 가능한 모집단 단위에 연계할 수 있는지 여부에 따라 (반복 측정) 행정 데이터가 될 수도 있고, 그렇지 않을 수도 있다. 관리 행정는 통계 목적이 아닌 관리 목적으로 다른 조직에서 관리되기 때문에 통계청의 관점에서는 2차 데이터이다. 국가 법 체계를 통해 행정 데이터에 대한 접근권을 통계청에 부여할 수 있다. 통계 행정는 통계 분석을 위한 매우 훌륭한 소스이다. 특히 소규모 영역, 빈도가 매우 낮은 이벤트, 경시적 프로세스를 연구하는 경우에 유용하다(Connelly 등, 2016). 사례 수가 많고(큰 n) 보조 변수 집합이 크면(큰 p) 통계 행정에도 머신 러닝을 다양하게 활용할 수 있다(Thompson 2018). 장점은 그 외에도 많다.

사회과학자들은 사회적 현상을 이해하는 데 관심이 있지만 실험을 수행할 수 없는 경우가 많다. 예를 들어 새로운 장소로 옮겨가는 사람과 한 장소에 계속 머무는 사람의 차이점은 무엇인지? 일자리를 찾거나 잃는 사람들과 관련된 보조 변수는 무엇인지? 어떤 종류의 회사가 해외 무역을 하는지? 파산하는 회사는 어떤 특징이 있는지? 사회적 참여와 신뢰는 여러 하위 모집단별로 어떻게 다른지(CBS 2015) 등이다. 행정자료의 관측 데이터를 통해 이러한 질문에 대한 답을 얻을 수 있다. 행정자료는 데이터의 비실험적인 특성으로 인해 인과 관계보다는 연관성에 초점을 맞춘다 (Hand, 2018). 얼핏 보기에는 청년과 노인, 저학력자와 고학력자, 부유층과 빈곤층, 원주민과 이주민, 종교가 있는 사람과 종교가 없는 사람, 비정규직과 정규직, 도시 주민과 농어촌 주민이 양분된 것으로 보인다. 그러나 이러한 잠재적 결정 요인은 매우 밀접하게 상관되어 있고 난해하다. 특정 요인의 영향을 분리하려면 다른 모든 요인에 대해 수정해야 하지만 이는 실제적으로 불가능하다.

이러한 상황에서는 단계적 회귀를 기계적으로 적용하여 적합도와 절약 간의 균형을 최적화하는 모델을 선택한다. 차원 축소 기법과 혼합 영향 모델을 사용하여 추정할 매개변수의 수를 압축할 수 있다. Lasso와 Ridge 회귀는 과적합을 방지하기 위해 회귀 계수의 강도에 추가로 벌점을 부과한다. 모델 평균화는 최상의 모델에 대한 지원이 강력하지만 명백하지 않은 경우에 적용 가능하다(Symonds와 Moussalli, 2011).

매개변수 모델 기반의 방법은 강력하고 풍부한 인사이트를 제공하지만 두 가지의 큰 문제가 있다. 첫 번째는 잠재적 예측 변수가 많으면 가능한 모델의 수가 급속하게 늘어나서 통제하기가 어려워지는 점이다. 두 번째는 비선형 관계와 고차원 상호작용을 명시적으로 정의해야 하는 점이다. 공식 통계에서 비선형 관계와 복잡한 상호작용을 모델링하기 위한 대체 방법으로서 신경망과 같은 머신 러닝 방법을 활용할 여지를 모색해야 한다.

비감독형 머신 러닝은 다양한 기능이 있는 고차원 공간에서 개체를 군집화할 수 있다. 감독형 머신 러닝은 고차원 공간에서의 군집화를 사용하여 결측 관측값을 대체하거나(de Waal 등, 2011) 비관측 하위 모집단에 대한 관계를 외삽할 수 있다.

③ 멀티 소스(Multi-source) 통계

다원적 통계는 하나 이상의 설문조사, 관리 행정 또는 빅 데이터 집합의 조합과 같이 복수의 데이터 소스를 기반으로 한다. 다원적 통계에 적용 가능한 머신 러닝 접근법을 결정하는 중요 요인은 데이터 소스의 통합 가능 정도이다. 여기에서는 연계를 세 가지 수준으로 구분하고 각 수준에서 머신 러닝을 사용할 수 있는 가능성에 대해 논의한다. 특히 국가 통계의 더욱 전통적인 설문조사 및 관리 소스와 관련하여 빅데이터 집합을 중점적으로 다룹니다.

미시적 수준 연계 - 미시적 통합

미시적 수준의 연계는 여러 데이터 집합의 개별 단위를 서로 연관시킬 수 있는 경우에 수행하며, 대부분 고유 ID가 필요하다. 빅데이터 집합에서 관측된 단위를 하나 또는 여러 관리 행정의 단위에 연계할 수 있는 경우, 관리 데이터를 통해 빅데이터에서만 관측되는 변수를 예측하기 위해 추정 및 예측 모델에 사용할 수 있는 보조 변수를 제공하여 빅데이터 레코드를 보강할 수 있다. 또한 관리 데이터를 표준 프레임으로 사용할 수 있다. 이렇게 하면 빅데이터 소스의 데이터 생성 메커니즘을 설명할 수 있게 되어 빅데이터 기반의 모집단 추정에 존재할 수 있는 편향을 없애거나 줄일 수 있다. 관련된 예로는 차량의 연간 주행 거리 예측을 위해 여러 머신 러닝 방법을 비교한 Buelens 등(2018)의 연구가 있다.

비무작위 선택으로 인해 발생하는 빅데이터의 선택 편향을 없애는 데 사용할 수 있는 다른 방법으로는 빅데이터와 설문조사 데이터의 연계가 있다. 이러한 환경에서는 빅데이터 소스에서 설문조사 대상 단위의 측정값을 가져올 수 있다. 정확한 연계가 불가능한 경우 표본 매칭을 적용하여 반드시 동일하지는 않은, 유사하고 대표적인 단위를 찾을 수 있다(Baker 등, 2013). 이 방법도 빅 데이터 집합을 단독으로 사용하는 경우보다 대표성을 향상시키는 것이 목표이다.

거시적 수준 연계 - 거시적 통합

거시적 수준 연계는 여러 소스의 단위를 개별적으로 연계하거나 매칭할 수는 없지만 특정 집계 수준에서 연관시킬 수 있는 데이터 연계이다. 예를 들어 동일한 지방자치체의 주민, 또는 동종 업계나 크기 클래스에 속한 기업이다. Tennekes와 Offermans(2014)는 지리적 위치를 정확하게 찾을 수는 있지만 개인에게 연계할 수는 없는 휴대폰 메타데이터를 사용한다. 이들은 휴대폰 빅데이터를 지자체 수준의 관리 행정에 연계하는 집계 수준에서의 편차 보정을 제안한다.

집계 수준의 빅 데이터 소스를 공변량으로 사용하여 설문조사 기반의 추정을 명시적으로 모델링하는 것은 Marchetti 등 (2015)에 의해 제안되고 Pappalardo 등 (2015) 등에 의해 적용되었다. 여기서 살펴본 사항은 소지역 추정 프레임워크 내에서 지역 수준의 모델을 적용하는 것으로(Rao와 Molina, 2015) 단위 수준의 연계는 필요하지 않지만 소스 간에 존재할 수 있는 상관을 이용하는 것이다. 이러한 접근법을 확장하여 기존 회귀모델 유형을 머신 러닝 예측 모델로 대체할 수 있다. 해당 접근법을 다룬 문헌은 없는 것으로 보인다.

Van den Brakel 등 (2017)은 빅 데이터 시계열을 독립 공변량 계열로 사용하는 구조적 시계열 모델링 접근법을 통해 설문조사 기반 추정의 정확도를 높였다. 빅데이터 시계열은 소셜 미디어 메시지에서 파생되며, 메시지 텍스트에 포함된 감정을 반영한다. Twitter 메시지와 같은 메시지는 개인에게 연계할 수 없고 충분히 세부적인 수준에서 집계할 수 없지만, 이 접근법은 시간 상관성을 사용할 수 있다.

제3절 해외 AI·빅데이터 활용 통계 도입 사례

각국의 국가통계 혁신 논의는 논의로만 그치는 것이 아니라 실제 도입을 위한 시범연구를 추진하거나 실제 통계로 활용되는 경우도 적지 않다. 물론 해외의 AI·빅데이터 분석 기반의 통계 데이터의 신뢰 기준이 국내에도 그대로 적용될 수 있다고 판단하기에는 논란의 여지가 있으나, 실제 도입에 의한 현장의 리스크 등을 파악하는 과정 또한 전면적인 도입 이전에 검토해볼 필요가 있기에 유의미한 시도로 볼 수 있다.

한편 각국의 새로운 생산 방식에 의한 국가통계 개발은 아직까지 표면적으로 공개되어있지 않은 사항들이 다수로서 그 사례를 정리하는 일이 쉽지 않다. 다행히도 독일은 자국의 국가통계의 머신러닝 기술 도입과 관련해 참고가 될 수 있는 다양한 근거, 즉 다수 국가들을 대상으로 한 관련 동향을 파악하였다. 이 자료는 통계 기관에서의 머신러닝 사용에 관한 조사로 명명되어, 독일내 14개 주 에 소속된 18개 통계 생산 기관 뿐만 아니라 유로연합의 27개 회원국, EFTA의 4개국, 아시아 등 비EU 국가 6개국의 도입 사례를 소개하고 있다.

<표2-6> 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	유럽연합(Eurostat), OECD

본 절은 해당 조사 결과에 따른 각국의 국가통계에 대한 AI 도입 시도들에 관해 소개할 것이며, 각 사례에 대한 국가별 추이를 파악해본다.

• **미국**

미국은 주로 경제, 노동, 그리고 농업 통계에 특화 된 국가통계 혁신 검토가 활발히 진행 중인 국가이다. 농업과 같은 특정 산업에 관련한 머신러닝의 통계 활용이 검토되는 주된 이유로는 미국의 산업 구조적 특성과 맞물려 농업과 관련 된 연간 통계의 조사 응답자의 응답 오류 및 무응답 사례가 비교적 문제로 지적되고 있으며 수확량 예측 등의 실질적인 예측 통계의 수요, 타 조사 통계와의 연계 수요가 많기 때문으로 분석된다. 미국은 머신러닝 기법을 전통적인 통계의 품질을 높이기 위한 보완적인 해법으로 접근하고 있으며 이와 관련 된 사례는 도입에 가깝거나 이미 도입된 경우도 존재한다. 반면 빅데이터에 기반해 미래를 예측하는 형태의 분석은 테스트 단계로 아직 실제 도입까지는 시간이 소요될 것으로 예상된다.

주요 사례는 아래처럼 설문조사의 응답결과 중 비정형 데이터의 특성을 분류함으로써 통계산출에 기초가 되는 통계 분류를 보완하는 형태이다. 해당 접근 중 업무상 상해 및 질병 설문조사에 적용된 방법은 이미 국가통계에 실제 도입된 상황이다.

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 적합한 코드 6개를 추천해줌으로써 자동 분류하는 방식		
활용 기술	정규화 된 로지스틱 회귀(~'18)→딥 러닝('19~)	**도입 목적**	생산성, 정확성

활용 기관	Census()	**도입 현황**	테스트 단계
활용 구분	**지역사회 설문조사의 산업 및 직업에 대한 코드 분류**		
세부 내용	- **(업무 개요)** 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

코드 분류에 활용된 주요 방법론은 로지스틱 회귀이나, 노동통계국의 사례에서는 2019년 이후 딥 러닝 기법으로 대체된 것을 알 수 있다.

<표2-7> 국가통계 머신러닝 도입 사례(미국)

기관	프로젝트 이름	적용 작업	방법
상무부 경제분석국	- 사전 GDP 추정을 위한 소스 데이터 최신 상황 예측	- 회귀	- 앙상블 방법
노동부 노동통계국	- 업무상 상해 및 질병 설문조사에서의 근로자 상해 서술에 대한 자동 코드 분류	- 다항 텍스트 분류	- 정규화된 로지스틱 회귀, 심층 신경망
	- 직업 요건 설문조사를 위한 직무 서술의 자동 코드 분류 및 검토	- 다항 텍스트 분류	- 정규화된 로지스틱 회귀
	- 급부 및 보장 범위 개요 문서에서 급부 정보 자동 추출	- 텍스트 분류, 정보 추출	- 랜덤 포레스트
	- 치명적 상해 사례와 직업 안전 건강 관리청(OSHA) 레코드의 자동 연계	- 레코드 연계	- 랜덤 포레스트
	- 치명적 상해 사례 정보와 웹 페이지 문서의 자동 연계	- 레코드 연계	- 랜덤 포레스트
	- 직업별 고용 통계(OES) 설문조사에 대한 무응답 분석	- 무응답 성향 모델링	- 회귀 트리
	- 직업별 고용 통계(OES) 직업 자동 코드 분류	- 직급 및 기타 가용 정보를 여러 직업 코드 중 하나에 지정	- 확률적 경사 하강법(SGD)을 사용한 로지스틱 회귀
	- 직업별 고용 통계(OES) 설문조사에 대한 무응답 분석	- 모델 보조 추정	- 회귀 트리
	- 경시적 직업별 고용 통계(OES) 설문조사에 대한 무응답 분석	- 무응답 성향 모델링	- 선형 모델을 사용한 회귀 트리
	- 정보 표본 추출에서의 자율 학습을 사용한 이상값 검출	- 이상값 검출	- K 평균
	- 면접자 기록의 텍스트 분석	- 면접자 기록의 군집화	- 모델 기반 군집화, K 평균 군집화, 베이지안 계층적 군집화

기관	프로젝트 이름	적용 작업	방법
통계국	- 자동 코더를 사용하여 지역사회 설문조사의 산업 및 직업에 대한 코드 분류	- 다중 분류	- 로지스틱 회귀
농무부 국립농업(NASS)	- 농업 총 조사의 표본 설계에 정보 제공	- 표본 추출 설계에 정보를 제공하기 위해 응답 확률 점수 개발	- 랜덤 포레스트와 부스팅
	- 농업 총 조사의 대체에 정보 제공	- 응답 확률 점수의 개발, 대체에 정보 제공	- 랜덤 포레스트와 부스팅
	- 농경지 데이터 계층, 행정 데이터, 설문조사 데이터를 기반으로 한 수확량 예측	- 수확량을 최대한 정확하게 예측하기 위해 다양한 데이터 혼합	- TBD(베이지안도 가능)
	- 농업 총 조사를 위한 머신 러닝	- 응답 확률 점수의 개발, 대체에 정보 제공	- 랜덤 포레스트와 부스팅
	- 질의서 통합	- 질의서 통합	- 계층적 군집화
	- 무응답자 예측	- 분류	- 의사결정 트리
	- 보고 오류 분석	- 분류	- 의사결정 트리

• **독일**

독일은 연방을 중심으로 산하 연구기관, 대학 등을 통해 관련 연구가 한창이다. 연구기관 및 대학 등을 통해 개발된 결과는 연방 을 경유하여 실질적으로 국가에서 활용할 수 있는지 여부를 판단하며, 이에 기반하여 효용성 있는 결과는 테스트 과정을 거쳐 실제 도입을 가늠하는 형태로 체계를 갖추고 있다. 그들은 다양한 분야에 걸쳐 국가통계의 머신러닝 도입 실험을 추진하고 있는데, 그 중 연방의 고유 프로젝트는 행정통계 자료의 부정확성을 보완하거나 새로운 정보를 추가하는 것, 기존 통계를 대체할 목적으로 추진되고 있다. 이와 달리 연구기관 및 대학의 프로젝트를 통해 수행되는 주제는 신뢰성 검증에 논의가 필요한 비정형 데이터에 관한 분류 및 예측 분야를 주로 수행한다.

주요 사례로는 각종 조사결과 및 행정자료의 분류와 관련된 것이 주류로 꼽힌다. 그 중 서로 상이한 두 가지 통계를 연계해 고용청의 패널 데이터를 보강하는 사례와 온라인 구인 공고를 분류하여 직업 정보를 추출하는 것은 타 국가의 사례에 비해 더 심층적인 접근이 이루어지는 것으로 판단된다.

활용 기관	연방	**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	**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	도입 현황	테스트 단계
활용 구분	**데이터 연계를 통한 연방 고용청의 패널 데이터 보강**		
세부 내용	- **(업무 개요)** 의 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	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	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

사례가 다양한 만큼 활용에 고려된 기술 또한 다양하다. 그 중 대표적인 알고리즘은 서포트 벡터 머신인데, 통계학의 회귀분석과 유사한 구조로 통계 분야에 적용하기에 진입장벽이 낮은 편이다. 흥미로운 부분은 K 최근접 이웃 클러스터링과 같은 비지도 학습 기법 또한 고려된다는 점이다. 머신러닝 기법 중 비지도 학습의 경우 훈련 데이터를 학습하는 방식이 아닌 데이터 자체의 특성에 기반한 분류를 수행하는 원리로, 결과의 재현성 및 결과 해석 측면에서 신뢰성을 검증하기에 비교적 어렵다는 한계가 존재한다. 그럼에도 불구하고 확률적 통계 모델인 “잠재 디리클레 할당법”이 아닌 클러스터링을 채택했다는 점은 향후 실제 도입으로 갈 수 있을지 여부에 관심을 기울여야 할 대목이다.

<표2-8> 국가통계 머신러닝 도입 사례(독일)

기관	프로젝트 이름	적용 작업	방법
독일 연방	비즈니스 행정에 포함된 기업을 기관 부문에 지정	이진 분류	SVM
	수공예 통계에서 관련이 없는 기업 인식	이진 분류	SVM, 랜덤 포레스트
	자녀 양육으로 인한 고용 중단 추정	이진 분류	SVM, 랜덤 포레스트, 로지스틱 회귀
	소득 구조 조사에서 근로자의 시민 여부(독일인/외국인) 추정	이진 분류	SVM(주요 방법)
	통합 고용 이력	이진 분류	랜덤 포레스트
	머신 러닝 방법론	다수	K 최근접 이웃, 나이브 베이즈, 랜덤 포레스트, SVM, ANN 등
	온라인 구인 광고 분류	텍스트 분류	K 최근접 이웃, 다항 나이브 베이즈
	소비자 물가 통계에서 스캐너 데이터 사용	분류	이 도구에는 세 가지 절차가 있으며 SVM, 랜덤 포레스트, 로지스틱 회귀에서 해당 절차를 테스트할 예정이다.
	이상값 식별: 격리 포레스트	이상값 식별	격리 포레스트
	타당성 검사 자동화 기술 검증(소득 통계)	오차 검출 및 대체	HoloClean(기본 확률 모델이 계수 그래프, 깁스 표본 추출을 통한 추론)
	소득 구조 조사(VSE)에서 BA(연방노동청)의 SV 키에 포함된 시간제 고용 코드 분류를 검사	이진 분류	랜덤 포레스트
RKI / ZBS 6	분광 조사 데이터를 기반으로 한 생물학적 표본(예: 세포, 조직, 미생물)의 식별, 구별 및 분류	구별, 식별, 분류	ANN SVM
RKI / MF1	생물정보학	분류(이진 및 다중 문제), 회귀, 군집화	랜덤 포레스트, 딥 러닝
RKI / FG-31	전염병 발생 자동 감지	회귀, 분류 및 정보 검색	HMM, GLM, NLP,
RKI / P4	RKI 건강 모니터링(KiGGS)에 대한 머신 러닝 방법 적용	군집화, 이진 및 다중 분류	군집 방법, 의사결정 트리, 신경망
유럽경제연구센터(ZEW)	TOBI - 새로운 혁신 지표 기준으로서의 텍스트 데이터 기반 산출 지표	텍스트 분석 및 분류	기타

기관	프로젝트 이름	적용 작업	방법
	Science4KMU	확률론적 점수화 모델	신경망
헤센 주	기업 웹사이트의 웹 스크래핑	데이터 추출, 일관성 검사	웹 스크래핑, 그래프 기반 신경망
독일연방은행	유가증권 투자에 대한 통계: 머신 러닝	분류	랜덤 포레스트
	FDSZ의 레코드 연계	분류	랜덤 포레스트
	거시 경제 시계열에 계절별 패턴이 존재하는지 확인하기 위한 통계 테스트 조합	분류	랜덤 포레스트, 조건부 랜덤 포레스트
	개인사업자의 산업 분류	분류	서포트 벡터 머신과 랜덤 포레스트의 오차 학습 시퀀스를 사용하는 역전파 인공 신경망 (이전 이름: 로짓
IAB FB B2	노동 시장 지역 유형화(비교 유형 SGB III / SGB II)	군집화	Ward / K 평균 알고리즘 기반 군집화
IAB, 만하임 대학교	새로운 직무 코드 분류 방법	다중 분류	여러 알고리즘의 비교, 쌓기, 부스팅
고용연구소(IAB)	행정 데이터의 교육 정보 수정	다중 분류	의사결정 트리, 랜덤 포레스트, SVM, 부스팅, 앙상블, 쌓기
	실업 기간 예측	회귀	의사결정 트리, 랜덤 포레스트, SVM, 신경망, 앙상블, 쌓기
	머신 러닝 방법을 통해 현장 실험을 기반으로 실업 기간을 예측하고 일자리 소개 조치를 평가	회귀 및 분류	의사결정 트리, 랜덤 포레스트, SVM, 부스팅, 앙상블, 쌓기
	CART 절차를 사용한 경제 활동 추정	다중 분류	의사결정 트리
IAB FB C1, 브리스톨 대학교	실업 기간 예측	분류 및 예측	의사결정 트리, 로짓, Lasso, SVM
고용연구소(IAB)	근무 시간에 관한 결측 정보 대체	분류	분류 트리, 교차 검증
연방 이민난민청	프로필 분석	텍스트 문단의 집계 및 분류	시맨틱 텍스트 분석
GESIS - 라이프니츠 사회과학연구소	소셜 미디어상의 선거 운동: 정치인, 지지자, Facebook과 Twitter를 통한 정치 커뮤니케이션의 중개	토픽 모델링	준지도 분석
	포퓰리스트가 인기를 얻게 될 때: 우익 단체 페기다(Pegida)와 독일 정당들의	토픽 모델링	LDA

기관	프로젝트 이름	적용 작업	방법
	Facebook 사용 비교		
	네트워크 추론에서의 표본 추출 편향 수량화 목표	관계 분류	베이즈(Bayes) + 완화
	사용자가 웹에서 온톨로지를 탐색하는 방법: NCBO의 BioPortal 사용 기록에 대한 연구	군집화, 차원 축소	K 평균 + 주성분 분석(PCA)
	HDP를 위한 실제적 붕괴 확률론적 변이 추론	베이지안 비매개변수 혼합 멤버십 군집화	PCSVB0
	다중 언어로 분류된 토픽 모델	토픽 모델	PLL-TM
	Predicting structured metadata from un-structured metadata(비정형 메타데이터에서 정형 메타데이터 예측)	메타데이터 예측	LDA, SVM
	문서 색인화를 위해 백과사전적 배경 지식으로 온톨로지 보강	문서 분류	LLDA, SVM
	크라우드 워커의 동기 측정: 다차원 크라우드 워커 동기 척도	척도 개발	구조방정식 모델
	유의어 사전과 분류 시스템의 확률론적 연결을 위한 시스템	개념 연결	PLL-TM
	Why We Read Wikipedia(우리가 위키피디아를 읽는 이유)	분류	그래디언트 부스팅
	음악 검색에 대해 문화 및 사회 경제적 요인에서 장르 선호도 예측	회귀	그래디언트 부스팅, 랜덤 포레스트
	도시 공간에서의 이동 패턴 감지 및 특징 확인: 맨해튼 택시 데이터에 대한 연구	군집화	텐서 분해
	위키피디아의 랜덤 워크에서 시맨틱 추출: 학습 방법과 계수 방법 간 비교	의미적 관련성	딥 러닝
	티켓 발행 시스템을 사용하여 기업의 애플리케이션 품질을 확인하기 위한 텍스트 범주화	텍스트 분류	SVM, 의사결정 트리 등
	RDF	순위 지정 학습	RankLib 라이브러리의 다양한 순위 지정 학습 알고리즘

· **캐나다**

캐나다는 연방에서 모든 국가통계 AI·빅데이터 도입을 관장한다. 그들 또한 상기 소개한 사례국가와 마찬가지로 데이터 이상치 검출, 통계 조사의 분류 등을 위한 머신러닝 도입을 추진하고 있다. 특징적인 점으로는 추진하는 프로젝트의 수에 비해 도입 유형은 획일화 되어 있다는 점이다. 그들은 이상치 검출과 분류체계 코드 분류 분야에 머신러닝 도입을 집중하고 있는데, 분류 목적에 편중되어 있다보니 서포트 백터 머신, XGBoost 등의 전형적인 딥러닝 기법을 주로 활용하고 있다. 이와 달리 몇몇 예는 민간 데이터를 활용하기 위한 사전 작업을 수행하고 있는 것으로 나타났다. 대표적으로 소매통계를 빅데이터 활용 통계로 대체하기 위하여 관련 사이트의 웹 스크래핑 및 구글 맵을 활용해 지역별 데이터를 구축하는 등 기반 빅데이터 축적에도 힘을 주고 있다. 또한 경제 분석을 위해 기업간 거래 네트워크를 구축하는 등의 작업도 진행 중이다. 그러나 추진되는 프로젝트에 비해 실제 적용되는 경우는 소수이며, 미국과 인접한 국가로서 농산물 수확량을 추정하는 시도 등의 경우가 실제 도입된 사례 중 하나이다.

주요 사례를 살펴보면 북미 상품 분류 시스템(NAPCS)의 상품 코드에 기준하여 민간 기업의 스캐너 데이터로 수집되는 비정형 데이터에 기반한 코드 분류가 있으며, 국제무역 시스템에 들어오는 관련 데이터 중 오기 또는 잘못된 데이터들을 검출하는 데에도 활용되고 있다. 캐나다와 같은 국가는 이민자들에 관한 현황 통계 또한 중요한 집계 이슈로서 이를 정확히 통계로 집계하기란 사실상 쉽지 않다. 그들은 이러한 통계조사의 결측치에 관한 추정을 딥러닝 기법을 통해 해결하려는 시도를 수행해 이를 실제 도입하고 있는 등 다양한 분야에서 딥 러닝을 적용 중이다.

<table>
<tr><td>활용 기관</td><td>연방</td><td>도입 현황</td><td>도입</td></tr>
<tr><td>활용 구분</td><td colspan="3">월별 소매 거래 조사 및 분기별 소매 상품 조사에 스캔 데이터 활용</td></tr>
<tr><td>세부 내용</td><td colspan="3">- (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</td></tr>
<tr><td>활용 기술</td><td>XGBoost(분산 그라디언트 부스팅), BoW(Bag-of-Words), n-gram 모델</td><td>도입 목적</td><td>기존 통계 대체</td></tr>
</table>

활용 기관	연방	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	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	연결성, 정확성

활용 기관	연방	도입 현황	도입(한시적)
활용 구분	**인구 조사 신규 컨텐츠 수요 분석**		
세부 내용	- **(업무 개요)** 2016년 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

활용 알고리즘은 딥 러닝 분류 기법 다수가 다양하게 분포되어 있다. 또한 추가적으로 설문조사 서술형 문답 및 다양한 텍스트 데이터, 즉 비정형 데이터에 관한 인사이트 추출에도 관심이 많아 형태소 분석을 포함한 자연어 처리 기술 다수가 활용되고 있다. 몇 가지 예의 경우 유전 알고리즘과 같은 규칙 기반의 머신러닝 알고리즘을 사용하는 예도 존재한다.

<표2-9> 국가통계 머신러닝 도입 사례(캐나다)

기관	프로젝트 이름	적용 작업	방법
캐나다 연방	소비자 물가 스캐너 데이터 사용	분류	SVM
	월별 소매 거래 조사 및 분기별 소매 상품 조사에 스캐너 데이터 사용	분류	XGBoost 선형, BoW(Bag-of-Words) 문자 n-gram 모델
	국제 무역 데이터 이상값 검출	이상값 검출	XGBoost 트리 모델
	BAEO(경영 활동, 경비 및 결과) 설문조사 의견 텍스트 마이닝	분류	선형 SVM, BoW(Bag-of-Words) 모델 사용
	결제 데이터 타당성 프로젝트	분류, 대체	SVM
	운송 통계: 트럭 운송 상품 원산지 및 목적지 설문조사(TCOD)	텍스트 분류, 이미지 분류, 보조 데이터 기반의 결측 정보에 대한 예측 모델링, 레코드 연계	최근 시작됨
	기업 통계	데이터 추출, 자연어 처리, 레코드 연계	최근 시작됨
	R&D 설문조사 프레임 개선을 위한 웹 스크래핑	데이터 추출, 자연어 처리, 레코드 연계	최근 시작됨
	소매 통계	데이터 추출, 자연어 처리, 레코드 연계	웹 스크래핑 라이브러리
	제조업과 생산자 물가 상품	데이터 추출, NAPCS에 대한 분류	TBD
	농산물 수확량 추정	예측 모델링	1단계: 라소(LASSO) 및 기타 2단계: TBD
	캐나다 주택 통계 프로그램, 국적 및 출생국 관련 프로젝트	분류, 대체	TBD
	관광객 지출 지표를 개발하기 위한 머신 러닝 탐색	회귀	TBD
	경제 분석: 고성장 기업과 폐업 기업 확인	분류	랜덤 포레스트
	경제 분석: 기업 네트워크 확인	분류/네트워크 분석	TBD
	의사소통과 전파: 웹에서의 사용자 경험 개선	챗봇	TBD
	의사소통과 전파: 웹에서의 사용자 경험 개선	TBD	TBD
	사망률 예측	예측 모델링	TBD

기관	프로젝트 이름	적용 작업	방법
	서술적 설명에서 사망 원인에 대한 주요 데이터 확보	분류	다수(SVM, 신경망, Adabost 나이브 베이즈)
	희생 조사(서술적 설명에서 주요 데이터 확보)	서술적 설명의 분류	텍스트 인식(가능한 경우), 자연어 처리, TBD
	일반화된 시스템의 R&D	표본 할당/선택	유전 알고리즘
	인구 조사	대체	특성 선택을 위한 ReliefF 알고리즘 및 최근접 이웃 대체를 위한 가중치 적용
	2016년도 인구 조사 콘텐츠 질의를 통해 제출된 의견 조사	주요 용어 식별 및 이진 분류	자연어 처리
	인구 조사 및 기타	분류	자율 머신 러닝을 통한 의견 군집화
	대체 방법의 R&D	대체	다수
	산업 및 직업의 텍스트 설명에 코드를 지정하기 위한 머신 러닝 탐색	통계 코드 분류	다수
	인구 조사 프로그램 혁신 프로젝트	정의 예정	TBD
	미시적 시뮬레이션을 위한 합성 데이터 생성	분류, 매칭	K 최근접 이웃
	합성 데이터 파일	합성 데이터(공시 통제)	CART 모델 및 랜덤 포레스트
	수집을 위한 우선순위 지정 점수		베이지안 계층적
	행정 데이터 시뮬레이션	모델링	TBD
	레코드 연계를 위한 특성 추출 자동화	분류, 특성 선택	지도, 자율, 차원 축소 기법(예: PCA)
	레코드 연계 소프트웨어인 G- Link 개발	Fellegi-Sunter 방법론을 사용한 확률론적 레코드 연계	지도: K 평균 감독: 2단계 방법 K 평균 + 프로빗
	레코드 연계를 위한 머신 러닝	이진 분류	SVM, 신경망, 지도 로지스틱
	NAICS/NOC 자동 코더	NAICS/NOC로 분류	BoW(Bag-of-words) 및 통계 분류 기준
	물질 유형별 음주운전 식별을 위한 머신 러닝 탐색	분류	TBD

· **네덜란드**

앞서 국가통계의 혁신과 관련한 국가별 논의에서 소개했던 네덜란드는 전통적인 통계 기법과 관리체계를 준수하는 선에서 빅데이터·AI 기술을 포괄할 수 있도록 다양한 방법을 모색하고 있음을 알 수 있었다. 이들 또한 캐나다와 마찬가지로 관련 프로젝트를 에서 전담하고 있는데, 그들은 설문조사 방식의 통계 생산에서 가장 큰 이슈로 주목받고 있는 회수율 문제를 해결하기 위한 독창적인 접근법이나 기업 특성을 수집하는 등 타국보다 좀 더 창의적인 방식들을 시도하고 있는게 특징이다. 특히 유럽처럼 국경을 두고 수출입이 이루어지는 대륙국가의 특성에 맞춘 통계 현안들을 빅데이터를 활용해 해결하려는 시도들이 이루어지고 있다. 이와 같은 시도들은 다른 국가에서 찾아보기 힘든 것들이라 국내 실정과는 부합하지 않으나 참고할만한 내용으로 판단된다.

주요 사례로는 위에도 언급했던 것처럼 국경 간 인터넷 구매에서 발생하는 거래내역을 추정하기 위하여 유로연합 소속 법인의 세금 정보와 인터넷 데이터를 연계하는 방식을 택한 것이 대표적이다. 또한 기업 통계의 무응답 대체에 있어 제기되는 신뢰성 문제라던지 설문조사 회수율 제고를 위해 혼합형 설문조사(Mixed mode survey)를 추진하는 경우에 각 조사의 통합 과정의 통계 유의성 확보를 위한 분류 문제를 머신러닝으로 해결하는 예를 꼽을 수 있다.

활용 기관		**도입 현황**	도입
활용 구분	**기업 특성 수집을 위한 웹 스크래핑**		
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[1)] - **(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

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

활용 기관		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)[2]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		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 조사대상의 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[3]		
활용 기술	비선형 SVM(RbfSVC), 랜덤 포레스트	도입 목적	부가정보

활용 기관		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

네덜란드 또한 활용 알고리즘은 전형적인 지도학습을 활용한다. 그 중 다소 잘 알려진 알고리즘을 준용하는 것으로 사료되며 그에 상응하는 것으로 SVM, 분류 트리, GBM 등이다.

2) 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

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

<표2-10> 국가통계 머신러닝 도입 사례(네덜란드)

기관	프로젝트 이름	적용 작업	방법
네덜란드	비확률 표본을 통한 추론	회귀	최근접 이웃, 신경망, 회귀 트리, 서포트 벡터 머신
	이주 행동 조사	이진 분류(다항 분류로 확장 가능)	랜덤 포레스트
	신뢰도 높은 추정과 머신 러닝 방법을 통해 대체 개선	비즈니스 통계의 대체	GBM(그래디언트 부스팅 머신)
	URL 검색 기업	이진 분류	의사결정 트리 / 랜덤 포레스트
	기업 분류 기준	이진 및 다중 분류 기준	의사결정 트리 / 랜덤 포레스트
	소비자 물가 지수의 상품 범주로 상품 분류	다중 분류	랜덤 포레스트
네덜란드 (TenForce와 협업)	소비자 물가 지수(CPI)를 위한 슈퍼마켓 상품 분류	다중 분류	SVM / 랜덤 포레스트
네덜란드	도메인 이름 정보 통신 회사		
	네덜란드 , 건강 설문조사에 적응 데이터 수집 적용	분류	분류 트리
	웹사이트 텍스트를 통한 산업 예측	다수 사례	다양한 머신 러닝 기법(SVM, 랜덤 포레스트 등)
	사이버 보안	이진 분류	SVM
	머신 러닝을 사용한 지속가능 회사 분류 자동화	이진 분류	알려진 바 없음
	딥 러닝 - 결측 값 추정	예측	딥 러닝(신경망)
	유럽 연합(EU) 웹 매장에서의 역외 인터넷 구매	이진 분류	반자동 머신 러닝

• **스위스**

스위스의 사례는 기존 조사통계 방식을 최대한 보존하는 선에서 머신러닝 기법을 보조의 수단으로 활용한다. 그들의 활용 사례들을 분석해보면 이러한 경향이 확연히 드러나는데, 가령 파라데이터의 활용을 그 예로 들 수 있다. 파라데이터는 설문조사 과정에서 축적되는 응답자의 설문 응답 반응이나 행동 패턴등을 총칭하는 것으로써 이는 설문조사 자체의 생산 절차에서는 활용되지 않는 변방에 위치하는 데이터로 볼 수 있다. 스위스는 이러한 파라데이터를 활용하여 설문조사에 응답하지 않는 표본이 응답에 성실히 수행하는 표본과 어떤 차이점을 가지는지를 정량적으로 분석하는데 머신러닝을 활용한다. 또한, 타 국가에서도 흔히 볼 수 있는 이상치 탐지 부분에서도 이런 경향이 드러나는데 이상치 탐지를 통해 관련 데이터를 제거하거나 보정하는 것이 아니라 이를 검토를 담당하는 담당자에게 우선 검토 목록으로 추천하는데 활용한다. 이처럼 그들은 머신러닝 도입을 하나의 효율적 도구로 접근함을 알 수 있다.

활용 기관	연방	**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	**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	연방	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

활용 기술군은 다양하나 특이점으로는 목적에 맞는 실험적인 알고리즘 활용을 추진한다는데 있다. 가령 의사결정트리, 랜덤 포레스트 등 잘 알려진 머신러닝 기법에는 기초가 되는 저변 알고리즘이 존재하나 여기에서 한발 더 나아가 저변 알고리즘을 특정 목적에 맞게 응용한 알고리즘을 활용한다는 점에서 주목해볼 필요가 있다.

<표2-11> 국가통계 머신러닝 도입 사례(스위스)

기관	프로젝트 이름	적용 작업	방법
스위스 연방	무응답 메커니즘의 모델링	응답 동질성 그룹의 이진 분류	CHAID
	의심스러운 응답 감지	이진 분류	일반화된 부스티드 모델, 랜덤 포레스트, 신경망, 나이브 베이즈, 트리 알고리즘 등
	그룹화된 답변에서 기업 거래액 분석	분류	랜덤 포레스트
	NOGA(스위스 NACE) 코드 분류의 자동화	분류	다수
	항공 이미지에 기반한 토지 피복 및 토지 사용 코드의 자동화	분류	합성곱 신경망(CNN)
	사회 보장 제도에서의 '경력' 군집화	군집화	다수

· 기타 도입 사례

앞서 언급한 국가 외 다양한 국가 및 국가연합에서 통계의 머신러닝 도입을 위한 시도들이 추진되고 있다. 아래는 추진 국가 또는 기관별 프로젝트 및 연구 현황을 요약하였다. 유로연합과 OECD의 경우 실제 특정 분야에 적용하기 보다는 표준화 된 기술 도입 연구에 가까운 사례로 판단할 수 있다.

<표2-12> 국가통계 머신러닝 도입 사례(유로연합, OECD, 기타국가)

국가(혹은 연합)	프로젝트 이름	적용 작업	방법
호주	복합 분류를 위한 자동 코드 분류	다층적 다중 분류	SVM, 부트스트랩 집계, 텍스트 모델
호주	연결 패턴에 의한 통계 엔티티 연결 및 그룹화	그래프 구조 분류, 모티프 탐지	SVM, 그래프 커널 모델
호주	토지 이용 및 수확량 예측	다중 분류	CNN
호주	행정 데이터 세트 편집	다중 예측, 이상 항목 감지	확률론적 그래픽 모델
호주	인구 조사 거주 예측	이진 분류	의사결정 트리
호주	다중 데이터 세트 연결	연결 스파인의 엔티티 분석 생성	경험적 베이즈(Empirical Bayes)
라트비아 중앙	인구 수 및 주요 인구 통계 지표-	이진 분류	로지스틱 회귀(GLM)를 실제로 사용 중이다. 구현 단계 중에 확률 그래디언트 부스팅(GBM), 서포트 벡터 머신(SVM), 규칙적 판별 분석(RDA), 다층 퍼셉트론(MLP), 방사 기저 함수 네트워크(RBF)를 테스트
아일랜드 중앙	오픈 소스 인덱싱 유틸리티를 통한 자동 코드 분류	다중 분류	

국가(혹은 연합)	프로젝트 이름	적용 작업	방법
유럽연합	대규모 설문조사 데이터 세트를 사용한 유로 지역 GDP 예측 - 랜덤 포레스트 접근법	회귀	랜덤 포레스트
유럽연합	신경망과 베이지안 네트워크를 통한 범주 데이터 대체	대체	로지스틱 회귀, 신경망, 베이지안 네트워크
헝가리 중앙	조세 포탈자 감지	분류	k-최근접 이웃
프랑스 경제통계연구원(INSEE)	임금/유급 감지 고용주 급여 계산 신고 통계 데이터베이스의 시간 이상 항목	비지도 알고리즘	퍼지 연관, 격리 포레스트, 국소 이상값 요인
프랑스 경제통계연구원(INSEE)	매체 정보를 사용한 단기 비즈니스 지표 실시간 예측	텍스트 마이닝, 머신 러닝, 의미 분석	딕셔너리 기반 접근법, 딥 러닝(Word2vec), 벌점 회귀
			(ElasticNet)
프랑스 경제통계연구원(INSEE) 및 내무부 통계국	고소에 포함된 가족 내 폭력 사건 식별	텍스트 마이닝, 분류	LDA, 랜덤 포레스트
프랑스 경제통계연구원(INSEE) 및 파리경제학교(IPP)	미시적 시뮬레이션 모델의 경력 및 임금 예측을 위한 머신 러닝	다항 분류	분류 트리, 랜덤 포레스트, 부스팅
프랑스 경제통계연구원(INSEE)	가중치 재조정을 통해 무응답 문제에 대한 머신 러닝 알고리즘 사용	지도 알고리즘	랜덤 포레스트, 배깅, 부스팅, CART, 그래디언트 부스팅 트리, 그룹 라소
이탈리아	인터넷 스크래핑을 통한 설문조사 대체		나이브 베이즈 및 기타
루마니아	기업 통계에서의 관리 데이터 사용	데이터 대체	트리 기반 방법: 회귀 트리, 랜덤 포레스트, 부스팅
루마니아	웹 스크래핑된 데이터를 사용한 물가 지수 추정	문자열 매칭	레벤슈타인 거리
루마니아	EU-SILC 개선을 위한 실행계획	빈곤 지표	랜덤 포레스트
루마니아	노동 시장에서의 잠재적 인적 자본 모델링	분류	로지스틱 회귀

국가(혹은 연합)	프로젝트 이름	적용 작업	방법
일본	가계 소득 및 지출 조사의 자동 코드 분류를 위한 감독 다중 분류 기준	다중 분류	나이브 베이즈
OECD	알고리즘과 담합	분류	다수 방법
영국	군집 토픽 식별을 사용한 비감독 문서 군집화	군집화	Doc2vec
룩셈부르크	스캐너 데이터	분류	MLR
	BERD(기업 R&D)	분류	모델 쌓기
오스트리아	대체	대체	K 최근접 이웃
	통계 매칭	통계 매칭	랜덤 포레스트, K 최근접 이웃
	오스트리아 고소득층의 빈곤 및 사회적 배제 위험 추정	추정/분류	부스팅 트리 알고리즘
	행정 데이터를 기반으로 SILC(소득 및 생활 조건 통계)와 유사한 소득 추정	추정/회귀	랜덤 포레스트
벨기에	일자리의 NACE 코드 예측	이진 분류	SVM,
덴마크	이민자 교육 수준 대체	대체	랜덤 포레스트
핀란드	사고 보고서 기계 독해	이진 분류	tf-idf +로지스틱 회귀
	산업과 직업의 자동 코드 분류	다중 분류	랜덤 포레스트
아이슬란드	제품 설명에 따라 세분화된 제품 ID 지정	분류	랜덤 포레스트
	데이터 처리 자동화 증가	다수	해당 없음
노르웨이	행정 기반 급여 통계의 편집에 사용하기 위한 랜덤 포레스트 탐색	분류	랜덤 포레스트
폴란드	농작물 탐지	세분화	K 최근접 이웃, SVM
	삶의 만족도	분류	자연어
포르투갈	빅 데이터 ESSNet	다중 분류	SVM(선형 커널)
	빅 데이터 ESSNet	다중 분류	퍼셉트론(선형)
	빅 데이터 ESSNet	다중 분류	언어 식별을 위한 신경망 모델
	분류 트리를 사용하여 오류가 포함된 레코드 식별	오류 감지	의사결정 트리

국가(혹은 연합)	프로젝트 이름	적용 작업	방법
스페인	UFAES(신규 방법론)	알고리즘 보조 설문조사 표본 추출	랜덤 포레스트
	정량적 변수의 선택적 편집	예측	랜덤 포레스트, SVM, 스플라인 회귀, K 최근접 이웃 회귀
	정성적 변수의 선택적 편집	이진 분류	랜덤 포레스트, SVM, 로지스틱 회귀, K 최근접 이웃 회귀
스웨덴	Essnet 빅 데이터 WP(작업 패키지)2 - 기업 특징 웹 스크래핑 - 구인 광고 사용 사례	이진 분류	SVC,
	Essnet 빅 데이터 WP(작업 패키지)2 - 기업 특징 웹 스크래핑 NACE의 사용 사례	다중 분류	케라스 순차 신경망
	머신 러닝 방법을 사용하여 직급 코드 분류 자동화	이진 분류	K 최근접 이웃
뉴질랜드	개인에 대한 지리적 지역 지정	분류	의사결정 트리
	자연어 처리를 사용한 건축 승낙 분류	다중 분류	지도 학습: 일반화된 선형 모델
	건축 승낙 분류	분류	벡터화된 n-gram을 사용하는 일반화된 선형 모델
	산업 분류의 코드 지정	분류	
	서포트 벡터 머신을 사용한 인구 조사 변수의 자동 코드 분류	다중 분류	SVM
	분류 및 회귀 트리를 통한 대체	분류 / 대체	의사결정 트리
	대체 매칭 변수 결정	대체/변수 선택	랜덤 포레스트
	동질적인 대체 클래스 생성	대체	의사결정 트리
	편집 규칙 파생	편집	연관성 분석

제4절 국내 AI·빅데이터 활용 통계 도입 사례

· 임금근로일자리행정통계(2012~)

해당 통계는 일자리 정책수립과 취업준비자에게 일자리선택에 필요한 기초자료를 제공할 목적으로 개발 된 2개 이상의 행정통계를 연계 및 분석해 도출되는 국가통계이다. 물론 통계청에서 관리하는 국가 행정통계를 활용해 해당 데이터의 가공을 최소화하고 제공된 데이터 그 자체의 가치를 최대한 보존하였다는 점에서 빅데이터 활용 통계로 바라볼 수 있는지에 관해서는 논쟁의 여지가 있다. 그러나 해당 통계는 2010년 추진되어 2012년에 국가통계 중 행정통계로서 정식 승인 받았다는 점에서 의의가 있으며, 현재에 이르러서도 일자리행정통계로 명칭이 변경되어 약 8년째 지속 발행되고 있어 참고 가치가 충분하다.

일자리행정통계의 구조는 공공기관이 보유한 행정자료들을 성·연령별, 기업체 규모별, 기업체형태별, 산업분류별로 구분하고 이를 통합해 별도 논의된 임금근로일자리의 구성 체계에 담아낸 행정자료 기반 2차 분석 데이터로 볼 수 있다. 여기에 활용된 통계자료는 국민연금, 고용보험, 산재보험, 근로소득지급명세, 사업자등록자료, 부가가치세, 법인세 정보 등으로 관리 주체가 매우 다양하다. 그러므로 해당 통계의 생산 및 관리는 통계청 자체적으로 수행할 수밖에 없는 한계가 있다.

<표2-13> 일자리행정통계 생산 방법

행정자료DB 구축		⇒	⇒		
①종사자DB	②사업자DB	③통계작성용 원시자료생성 (raw data)		④에디팅 (Editing)	⑤마이크로데이터 (Microdata)
종사자정보 통합 (i)국민연금-고용보험 (ii)국민·고용-국세 <연계키> 개인식별번호 사업자등록번호	사업자정보 통합 (4대보험-국세) <연계키> 사업자등록번호	종사자-사업자 통합 <연계키> 사업장관리번호 사업자등록번호	행정자료를 통계자료로 전환 <주요내용> 중복검사(editing) 비임금근로자 제거	무등록자료 대체 (Imputation) ⇓ Microdata생성	

*출처 : 통계청(2012)

일자리 행정통계를 생산하는데 있어 필요한 각각의 통계자료가 모두 확보되더라도 그

들을 연계할 수 있는 방법이 필수적으로 마련되어야만 한다. 그런 의미에서 그들은 개인식별번호와 사업자등록번호를 데이터베이스의 기본 키(PK)로 지정해 이를 기반하여 각종 데이터를 통합하는 작업을 거친다. 해당 작업이 성공적으로 이루어졌을 때 통계작성용 원시자료생성이 완료되며, 그 후 일자리행정통계에서 제공할 지표들에 관한 모델을 적용하고 산출된 데이터를 통계자료로 변환하는 작업을 수행한다. 이 과정에서 연계가 녹록치 않거나 등록되지 않은 결측치에 관한 처리가 필요하며 이와 같은 부분을 해결해 마이크로데이터 생성까지를 완료하는 절차로 통계화를 추진한다.

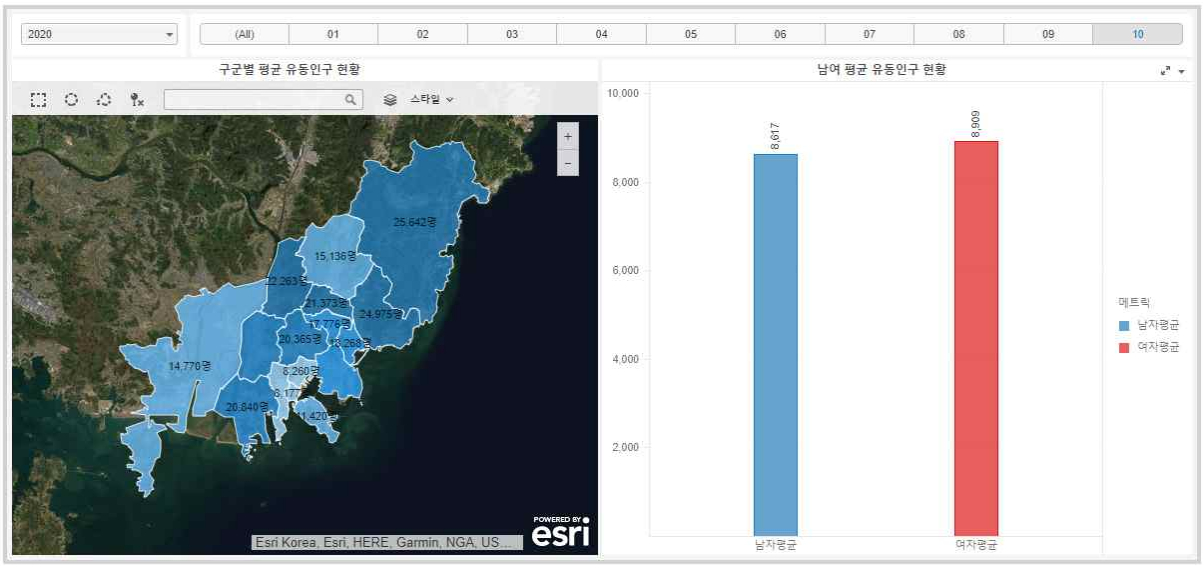
해당 자료는 행정자료를 종합 및 가공했다는 측면에서 몇 가지 용어에 관한 조작적 정의가 필요수적이다. 대표적으로 4가지 개념을 정의하였는데 첫 번째로 임금근로일자리(Wage & Salary employment position)이 있다. 이는 기업체에서 현물 또는 현금을 대가로 상품생산이나 서비스 활동을 하는 임금근로자가 '점유한 고용위치'를 의미한다. 두 번째는 지속일자리(Sustainability employment position)이다. 이는 동일한 기업체에서 당해연도와 전년도에 걸쳐 지속적으로 동일 근로자에 의해 일자리가 점유된 경우를 뜻한다. 세 번째로는 신규·대체일자리(New & Employee Substitution employment position)이다. 이는 통계기준시점 현재 당해연도와 전년도 사이에 조직생성이나 조직내 신규 및 대체로 인해 변화된 일자리를 의미한다. 여기에서 조직생성 부분은 새로운 법인 설립 등 새로운 조직 생성에 따른 신규일자리로 정의해 구분하며 조직내 신규 및 대체의 경우는 동일 조직에서 조직 확장이나 근로자의 입퇴직 등으로 인한 일자리 변화로 구분해 집계하고 있다. 마지막으로는 일자리증감이다. 이는 전년 대비 일자리 수의 증가 혹은 감소한 정도를 집계하는데 계산식은 매우 간단하다. 전년도 말일의 일자리 수를 당년도 현재 일자리 수에서 차감하는 형태로 시계열을 구성한다.

· 부산 서비스인구통계(2014)

해당 사례는 현재 부산광역시의 빅데이터 포털을 통해 지속적인 개선 및 응용 데이터를 제공하고 있는 대표적인 빅데이터 활용 통계이다. 통계는 인구 통계를 보다 미시적으로 보여줄 수 있는 다양한 지표를 제공하는데, 대표적으로는 지역에 따른 주중/주말 평균 유동인구, 구군별 일평균 유동인구, 지역에 따른 계절별 유동인구, 지역에 따른 시간대별 평균 유동인구 등이다. 이와 같은 데이터는 기존 설문조사로 집계하기 어려우며, 실시간 지표를 제공하는 것은 구조상 불가능 하다. 이를 달성하기 위해 서비스인구

통계는 과기부의 통신사별 점유율 통계 정보와 SKT의 모바일 위치 정보를 연계하여 SKT 통신사를 활용하는 인구의 모바일 위치 정보 변화를 파악하고 통신사 점유율을 통해 부산시 전체 인구의 변화를 추정하는 형태의 통계 자료이다. 이 사례의 특이점은 민간 빅데이터를 활용한 최초의 승인통계라는 점인데, 모집단 누락 등의 신뢰성 문제로 인해 1년간 승인 후 영구 승인을 받는 데는 실패하였다는 점이다. 민간 데이터를 활용하였을 시 문제로 발생할 수 있는 것이 바로 모집단에 관한 품질 평가 기준을 충족하기 어렵다는 점으로, 승인 연장 과정에서 이와 같은 부분이 부각된 사례로 평가 가능하다. 그럼에도 불구하고 지자체를 통해 제공되는 통계이므로 현재까지도 시각화된 형태로 간략한 정보 제공이 이루어지고 있다.

<그림2-7> 부산 서비스인구통계내 지역에 따른 성별 평균 유동인구 정보



*출처 : 부산시 빅데이터 포털(2020)

- **교통접근성 지표(2017)**

교통접근성 지표는 국가통계로 원안이 승인된 현재까지의 유일한 통계이다. 한국교통연구원을 통해 추진된 해당 지표는 일반통계 중 가공통계로서 승인되었다. 작성 목적은 교통부문 여객과 화물의 원활한 이동성 및 접근성 확보와 사회경제활동의 지원에 필요한 최적 교통시설 확보 등을 위한 기초자료를 제공하기 위함이다. 이는 국가 및 지방정부 차원에서 우리나라의 교통접근체계의 수준을 평가하는데 활용할 수 있어 지역별 교통 환경 개선을 위해 유의미한 참고자료로 활용이 가능하다. 해당 지표는 교통과 관련 된 다

양한 공공데이터를 활용해 개발되어 앞서 소개한 통계청 일자리행정통계 대비 필요 데이터의 확보차원에서 용이한 장점을 가진다. 그들은 집계구, 서비스시설, 도로교통망, 대중교통망 등과 관련 된 총 18개 공공데이터를 취합하여 매년 1회 공표하는 것이 특징이다. 빅데이터의 출처가 산재해있고 그들의 공표 시기가 서로 다르며 실시간성을 띄지 않는 특성을 고려하다보니 실시간성을 확보하지 못한 점이 한계이나, 명확한 지표 개발을 위한 산식을 모두 제공하고 있어 국내에서 가장 체계적인 분석 사례로 볼 수 있다.

그들은 통계지표의 산출 방법론으로 총 3가지 산식을 공개하였다. 첫 번째는 평균접근시간에 관한 개념인데 가장 인접한 서비스시설까지 도달하기 위한 평균 소요시간을 의미한다.

<그림2-8> 평균 접근시간 계산식

$$\text{평균접근시간}_j = \frac{\sum_{j_i \in \Lambda_i} \left(Pop_{j_i} \times Min\left(T_{j_i \to W} \right) \right)}{\sum_{j_i \in \Lambda_i} Pop_{j_i}}$$

j : 각 행정구역(시군구, 읍면동 등)
$\Lambda_l = \{j_1, j_2, \ldots, j_k\}$: l번째 행정구역 내 전체 집계구 집합
Pop_{j_i} : j_i 집계구의 인구
$W = \{w_1, w_2, \ldots, w_n\}$: 대상시설 집합
$T_{j_i \to W}$: j_i 집계구 중심에서 대상시설 집합 $W = \{w_1, w_2, \ldots, w_n\}$으로의 통행시간 값들 $\{T_{j_i \to w_1}, T_{j_i \to w_2}, \ldots, T_{j_i \to w_k}\}$

*출처 : 한국교통연구원(2018)

두 번째 지표는 접근 가능 인구 비율이다. 이는 특정시간(15,30,45,60분) 내 각 서비스시설로 도달할 수 있는 이용자의 비율을 의미한다.

<그림2-9> 접근 가능 인구 비율 계산식

$$\text{접근가능인구비율}_j = \frac{\sum_{j_i \in \Lambda_i} \left(Pop_{j_i} \times I\left(Min\left(T_{j_i \to W} \right) < T_{\max} \right) \right)}{\sum_{j_i \in \Lambda_i} Pop_{j_i}}$$

I : Index 함수(조건을 만족할 시 '1', 만족하지 못할 시 '0')
$T_{\max}$: 대상시설로의 한계통행시간(15, 30, 45, 60분)

*출처 : 한국교통연구원(2018)

세 번째 지표는 접근 가능 시설 수이다. 이는 특정시간(접근 가능 인구 비율과 동일 기준) 내 도달할 수 있는 서비스시설 수의 평균값을 의미한다.

<그림2-10> 접근 가능 시설 비율 계산식

$$\text{접근가능시설수}_j = \frac{\sum_{j_i \in \Lambda_i} \left(Pop_{j_i} \times \sum_{w_k \in W} I\left(T_{j_i \to w_k} < T_{\max} \right) \right)}{\sum_{j_i \in \Lambda_i} Pop_{j_i}}$$

*출처 : 한국교통연구원(2018)

· 기타 미승인 통계

현재 통계청을 통해 기 개발 되었으나 국가통계로의 승인에 난항을 겪는 빅데이터 활용 통계는 총 3건으로, 외부 공표가 되어 있지 않아 관련 정보를 찾기 쉽지 않다. 3건 모두 통계청이 직접 개발하는 사례로 일자리행정통계와 마찬가지로 다 부처에서 관장하는 행정자료를 연계하는 형태로 추진되었다. 미승인 된 사유를 추정해 보았을 때, 일자리 행정통계 대비 연계 자료가 미약해 기존에 보유한 행정통계와 별개의 통계로 승인하기에는 규모차원의 한계가 있을 것이라 추측되며, 신용정보 및 개인정보활용 측면의 이슈 발생여지가 있다는 점에서 승인이 쉽지 않을 것임을 예측해볼 수 있다.

<표2-14> 빅데이터 활용 개발 통계 중 미승인 통계 예

통계 사례	활용 통계	수행기관
학자금대출자의 경제활동 특성 분석	- 인구 센서스 정보(통계청) - 학자금 대출 잔액 정보(신용정보원)	통계청
개인사업자 부채 분석	- 통계기업등록부 정보(통계청) - 등록센서스 정보(통계청) - 대출정보(KCB)	통계청
생활안전사고 분석	- 인구센서스 정보(통계청) - 구조활동 빅데이터 정보(소방청)	통계청

*출처 : 정보통신정책연구원(2020)

제3장 국가통계의 AI·빅데이터 기술 도입 쟁점과 시사점

앞선 국가별 논의사항 및 관련 도입 사례들을 종합해 보았을 때 다양한 적용분야와 방법론이 동시다발적으로 등장하는 것을 알 수 있다. 그중에서는 실제 국가통계로의 승인을 받아 생산되는 통계도 있는가하면 아직까지 실험 단계에 머무른 경우도 다수 존재한다. 이처럼 다양한 국가통계 부문의 AI·빅데이터 기술의 도입은 유형이 다양하므로 도입 유형 및 목표에 따라 이에 상응하는 필요 정책 및 고려사항도 상이해질 수 밖에 없다. 본 보고서는 이와 같은 도입 유형을 두 가지로 요약하였다. 첫 번째는 신규 통계의 생산이고 두 번째는 통계 생산 프로세스의 현대화이다.

<표3-1> 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

신규 통계의 생산 관점의 도입은 기존의 조사통계 및 행정통계를 새로운 데이터와 방법론으로 완전 대체하거나 추가 정보를 제공하기 위해 부가적인 분석을 하는 것, 또는 이와 별개로 실태조사 등이 수행되기에는 저변이 마련되지 않은 신규 이슈에 관해 인사이트를 가져다줄 수 있는 통계 데이터를 생산하는 것 등으로 요약된다. 이는 기존의 통계 방법론과 궤를 달리하며 빅데이터 및 AI 기술을 통계 생산의 주요 방법론으로 채택하는 경우를 의미한다.

반면 통계 생산 프로세스의 현대화 관점은 앞선 2장에서 소개된 독일의 사례와 유사하다. 기존의 통계 생산 과정의 각 프로세스 중 머신러닝 기술이 도입될 시 효율성을 높일 수 있는 부분을 탐색하고, 기존 추진되었던 방법론을 머신러닝 기술로 일부 대체하는 접근이다. 이는 결국 기존 통계 생산에 활용하는 방법론과 품질의 비교우위를 증명해야만 활용 가능성을 점쳐볼 수 있어 위의 신규 통계 생산 관점과는 사뭇 고려사항이 다를 것임을 예상해 볼 수 있다.

본 장에서는 AI·빅데이터 기술이 국가통계 제도권 내에 안착되어 실제 활용이 촉진되기위해 고려되어야 할 예상 쟁점을 기술 도입의 두 가지 유형으로 구분해 짚어보았다. 또한 각 유형별 고려되어야할 예상 쟁점을 종합해 전반적으로 국내에 필요한 기능이 무엇인지 분석해 본 후, 이를 현실화하기 위해서 거버넌스 측면의 요구사항 등을 논의해본다.

제1절 AI·빅데이터 통계 도입 시 예상 쟁점

1. 신규 통계 생산의 관점

본 보고서는 신규 통계에 대해 조사통계 기법을 활용하지 않고 빅데이터·AI 기술을 활용해 생산한 통계(국가통계)로 정의하였다. 이와 같은 유형의 통계는 세부적으로 세 가지로 구분된다.

첫 번째는 빅데이터를 활용한 기초통계이다. 이는 국가 현황 파악에 기초가 되는 1차 통계로서 통계 생산을 위해 민간 빅데이터 및 행정 통계를 활용하는 것으로 볼 수 있다. 구체적으로는 완전히 새로운 통계를 발굴하는 것과 더불어 이미 존재하는 조사통계 방식의 기초통계를 빅데이터 및 AI 기술을 통해 대체하는 경우가 해당된다.

두 번째는 빅데이터를 활용한 가공통계이다. 여기에는 국가 정책 지원, 사회 현상의 진단 등의 목적으로 민간 및 행정 빅데이터를 가공 또는 분석한 사례가 해당한다. 구체적으로는 소셜 분석, 언론 데이터 분석, 빅데이터 및 AI기술을 활용한 지수 개발 등이 해당되며 이와 더불어 분석 대상과 기술로서 빅데이터 및 AI를 채택한 경우이다.

세 번째는 기초 통계 데이터를 활용한 가공통계이다. 이는 현존하는 국가 통계 데이터를 빅데이터·AI 기술을 통해 2차 분석함으로써 도출되는 신규 통계를 의미한다. 구

체적으로는 통계법에 저촉되지 않는 범위 내에서 설문조사 데이터에 기반한 빅데이터 분석을 수행해 별도의 부가적인 지표 및 시사점을 도출하는 경우로 볼 수 있다.

해당 통계 생산 관점에서 AI 및 빅데이터 기술이 도입될 시 예상쟁점은 총 3가지를 꼽을 수 있다.

① 대표성 및 보편적 수용

통계청에서 고시하는 국가통계 기본원칙의 제 2항에는 신뢰성 제고와 관련한 실천방안이 기술되어 있다. 여기에는 국가통계는 통계적으로 널리 활용되는 과학적인 작성기법을 사용해야함이 명시되어 있는데, 이와 관련 된 이슈가 바로 AI활용에서 나타날 수 있는 대표성과 수용성 문제이다. 머신러닝, 딥 러닝 등으로 대표되는 AI기술과 텍스트 마이닝, 자연어 처리 기술 등으로 대표되는 빅데이터 분석 방법론은 사회조사 방법론의 검증 체계를 따르지 않으므로 기존 방법론과는 검증의 관점이 상이하다. 가령, 빅데이터 분석에 활용되는 알고리즘과 딥 러닝 알고리즘 등은 결과의 정확성을 높이기위해 특정 파라미터의 유동적 변환을 전제로 하는데, 이는 검증의 최우선 기준이 정확도에 집중되어 있음에 기인하기 때문이다. 그러나 일반적으로 활용되는 통계생산에 활용되는 방법론의 경우 결과의 정확도 이전에 모델 형성 과정의 대표성과 당위성이 중요하다. 검증 된 모델을 활용하게 되면 모델의 선정 타당성을 더 평가의 높은 기준으로 보기 때문에 정확도를 별도 검증하는 것에서 비교적 자유롭다.

여기에서 문제가 되는 것은 머신러닝 알고리즘 및 파생된 기법들의 경우 국가통계와 같은 정형화 된 틀 내에서 활용된 것이 아니다보니 통계에 활용 가능한 대표적인 기술군이 정의되지 않았다는 것에서 비롯된 대표성 부재이다. 또한 각 알고리즘은 작금의 상황에서도 정확도를 높이기 위한 다양한 파생 기법으로 발전되고 있는 상황이나, 각 기술 중 통계 생성에 어떤 기술이 적합한지에 관하여 별도 제시된 연구가 없다. 2장을 통해 소개한 UN유럽경제위원회의 머신러닝 활용 문헌에 약간의 대표적 알고리즘에 관한 단서를 찾을 수는 있으나 이 또한 개념적인 설명일 뿐 확실한 기준을 제시하고 있지는 않다는 점이 한계이다.

결과적으로, 신규 통계 생산 시 활용에 적합한 검증된 알고리즘을 정의하고 이에 대한 세부적인 활용 지침 마련이 요구된다고 볼 수 있다.

② 통계 생산 프레임워크

신규 통계의 생산은 기존의 통계 생산 프레임워크에 맞추어 해석하기에 적합하지 않다. 가령 국가의 통계업무를 구체화하여 단계별로 제시한 국제 표준 GSBPM은 유연성을 일부 보유하고 있으나 조사통계에 적합한 프로세스를 준용하고 있다. 국내의 국가통계 생산 절차의 기준으로 활용하는 KSBPM의 경우 또한 마찬가지인데, 이는 GSBPM의 상세 내용과 프레임워크 구조를 상당부분 준용한 표준이므로 거의 유사하다고 판단 가능하다. 데이터마이닝, 또는 AI관련 알고리즘을 적용해 통계를 생산하는 절차를 이에 대응가능한지에 관하여 분석해보면 기본 절차, 분석 방법 등이 상이한 것을 가늠해볼 수 있다.

<그림3-1> 한국통계업무프로세스모델(V.2.0)

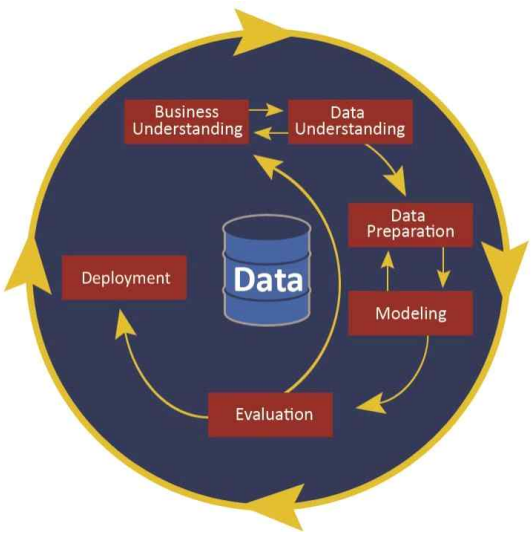
기획	설계	구축	수집	처리	분석	배포	보관	평가
수요 파악	산출물 설계	수집 도구 구현	자료 수집대상 설정	자료통합	산출물 작성	공표자료 점검 및 적재	자료보관규칙 정의	평가 계획 수립
수요 확정	항목 설정	처리시스템 개발, 개선	자료 수집 준비	분류 및 코딩	산출물 검증	공표 산출물 작성	자료 보존 및 관리	수행 및 보고서 작성
산출목표 수립	수집 방법 설계	배포시스템 개발, 개선	자료 수집 진행	자료검토 및 보완	산출물 해석 및 설명 작성	자료 배포 관리		개선과제도출 및 실행계획수립
통계적 개념 정립	모집단 및 표본설계	업무 절차 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)

KSBPM의 모델은 순차적 단계를 따르고 있는 것이 특징이다. 물론 여기에서 각 단계는 수요에 따라 유연하게 운영할 수 있음이 명시되어 있으나, 각 프로세스가 분절된 업무 및 관리 역할을 기준해 나뉘어 있으므로 정상적인 통계 생산 루틴 하에서는 이를 위반하거나 수정할 이유가 없다. 그러나 AI관련 알고리즘을 적용해 통계를 생산하는 경우에는 생산 과정에서 반드시 내부의 순환 고리가 요구된다. 가령 대표적인 데이터마이닝 업계 표준 중 하나인 CRISP-DM을 보면 확연히 다름을 알 수 있는데, 빅데이터의 제어 및 분석 모델의 유효성을 검증하는 과정이 순환고리로써 반복 수행되어야 함을

강조하고 있다. 이는 현 통계업무프로세스모델로는 도식화 하기 어려운 부분을 형성하고 있다. 이와 관련해 스위스 통계청의 경우 GSBPM이 빅데이터 통계를 포괄하기에 적합하지 않으며, 서로 다른 프로세스의 개발 또는 상호보완적인 연계 방안을 유럽 통계시스템(ESS)에서 다뤄야 한다 언급하였다. 국내 또한 마찬가지 쟁점이 예상된다.

<표3-2> CRISP-DM(좌) 및 프로세스 역할 요약(좌)



▶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)

③ 비결정론적(Non-deterministic) 알고리즘

널리 활용되는 빅데이터 및 AI기술들은 확률론에 기초하는 사례가 많다. 결과 도출 과정에서 연구자의 주관적 판단이 반드시 요구되는 알고리즘이 많은데, 이는 알고리즘이 최적화를 위한 최근접 근사 값을 얻기 위한 전략을 취하기 때문이다. 머신러닝 알고리즘은 크게 지도학습과 비지도학습으로 구분되는데, 특히 훈련 데이터를 활용하지 않고 데이터의 특성을 파악해 분류하는 비지도학습의 경우 분류의 기준이 되는 군집의 수(K)를 연구자의 판단을 통해 결정하는 경우가 많다. 예외적인 경우 또한 군집의 수를 결정하는 방법이 학술적 관행에 의해 결정된 타 기준을 만족하는 특정 K를 탐색하는 방법을 택한다. 이는 잠재적으로 국가통계에 도입하는데 있어 불확실한 요소로 작용할 수 있으므로 기법 활용의 신뢰성 문제가 제기될 가능성이 있다.

또한, 빅데이터 및 AI알고리즘의 일부가 주관적 판단 없이도 동일 파라미터에 의해 도출된 결과가 확률적으로 다를 수 있다. 구체적으로 알고리즘 동작 과정에서 지역 최

적화 문제(Local Optima)를 피하기 위한 난수(random variable)를 활용하는 경우가 존재한다. 이는 즉 같은 알고리즘과 같은 파라미터에 기반한 시행임에도 불구 동일한 결과를 재현할 수 없는 비결정론적 문제에 직면한다. 이는 통계의 신뢰성을 판단하는 기준 중 재현가능여부에 정면으로 위배되는 사항이다.

그럼에도 불구하고 이는 해결이 원천적으로 불가능한 이슈는 아니다. 기존의 통계 모델링 과정 또한 유사한 계산을 수행하는 경우가 존재하기 때문이다. 그러나 기존 통계적 모델링의 유효성은 검정력(P)으로 판단할 수 있도록 일종의 가이드를 제시하고 있다. 이는 통계학에서 매우 오래된 통계적 유의성과 관련한 이슈와 무관하지 않다. AI알고리즘은 현재 이와 유사한 신뢰 기준이 표준화되지 않음에서 관련 쟁점이 발생할 것이라 예상해볼 수 있다.

<표3-3> 신규 통계에 잠재적으로 활용 가능한 머신러닝 알고리즘의 종류와 특성

구 분	설 명	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[4]	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[5], 분류[6]	불가능 (모델에 대한 품질 평가는 가능)

주) 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

4) 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

5) 데이터의 분포를 설명하기에 최적인 확률 분포에 매핑하여 해석하는 기법

6) 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

한편 지도학습의 경우 주어진 검증 데이터를 대상으로 학습 모델의 성능 측정이 가능하므로 비지도학습의 사례와 달리 상대적으로 통계에 활용하기 용이할 것으로 추정된다. 반지도학습은 정확도 재고를 위하여 비지도학습과 지도학습을 혼용하는 방식으로서, 지도학습의 특성 일부를 보유하고 있기 때문에 학습 모델의 성능 측정이 가능하다. 이는 AI알고리즘에 관한 검증 방안이 표준화될 시 지도학습과 마찬가지로 도입에 유리한 요건을 갖추고 있음을 알 수 있다.

2. 통계 생산 프로세스의 현대화 관점

본 보고서는 통계 생산 프로세스의 현대화에 관해 기존의 조사통계 프로세스를 수용하되 통계 데이터의 품질 향상, 생산 절차의 효율성 등을 개선하기 위해 빅데이터·AI 기술을 도입하는 것으로 정의하였다. 이와 같은 유형은 두 가지로 구분된다.

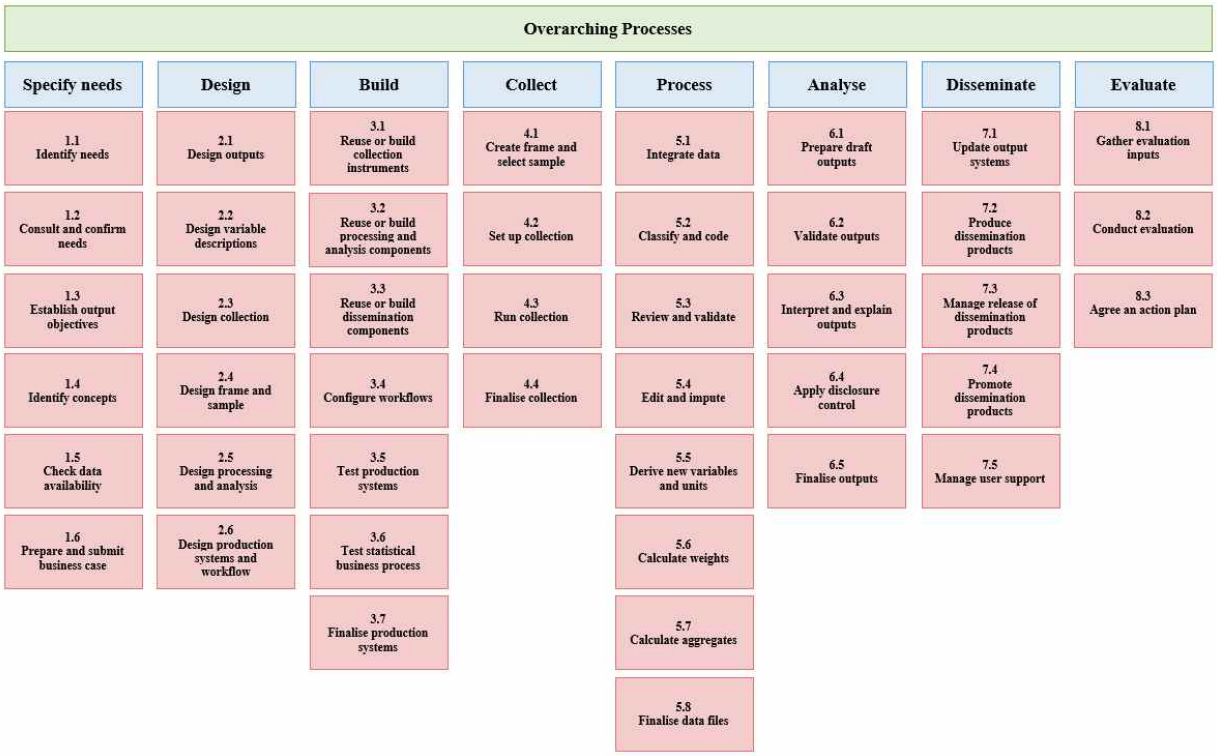
첫 번째는 기능 개선 목적이다. 통계 프로세스 상 의사결정 요소들의 성능을 높이는 기법들을 의미하는데, 일반적으로 표본 층화, 이상 값 대체 및 보정 등의 현행 통계 방법론에 적용될 수 있다.

두 번째는 생산성 증대 목적이다. 통계 업무는 정형화 된 절차를 가지고 있으나 매 생산주기마다 동일한 업무량을 수반할 수밖에 없는 비효율성을 지니고 있다. 통계 업무의 비효율을 개선하고 생산성을 높이기 위한 기술 도입이 바로 해당 목적이며, 구체적으로 자동 코드화, 이상값(outlier) 검출, 레코드 연계, 데이터 비식별화 지원 등이 여기에 해당한다.

신규 통계의 생산 관점과는 달리 통계 생산 프로세스는 보다 정형화 된 절차가 존재하므로 각 절차 중 어떤 분야에서 빅데이터 및 AI 알고리즘이 적용 가능한지 검토해볼 필요가 있다. 신규 통계의 생산 관점에서 언급했던 통계업무표준인 GSBPM을 통해 이를 검토해볼 수 있다. 현재 GSBPM은 5.1 개정이 완료된 상황으로 총 8단계로 구성되며 각 단계에는 하위 프로세스가 존재한다. 각 통계 생산을 수행하는 실무자 입장에서 각 업무 단위로서 참고할 수 있는 부분이 하위프로세스이며 이는 세부적인 기준과 생산하는 통계의 특성에 따라 실제 추진 방법이 유연하게 변경될 수 있다.

본 절에서는 해외 문헌을 검토함으로써 머신러닝 활용이 가능하다고 판단 된 통계 생산의 세부 프로세스와 과업에 관하여 정리하였다.

<그림3-2> 일반 통계 비즈니스 프로세스 모델(GSBPM v5.1)



*출처 : unece.org

<표3-4> 머신러닝 활용이 가능한 국가통계 생산 과정 및 예상 기술군 (GSBPM 기준)

코드 (1레벨)	명 칭	코드 (2레벨)	명 칭	과업 정의	도입 가능 기술군[7]
2	설계	2.4	모집단 및 표본 설계	- 모집단을 식별 및 지정 - 적합한 샘플링 기준 및 방법론 결정	- 분류 - 군집화
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

코드 (1레벨)	명 칭	코드 (2레벨)	명 칭	과업 정의	도입 가능 기술군[7]
		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
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)[8]
		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

7) 인공신경망의 경우, 설계 목적에 따라 모든 기술군을 포괄가능하므로 별도로 명시하지 않음
8) 2차 분석의 영역으로서 예상 기술군을 특정하는 것은 부적절하다 판단하였음

통계 생산 프로세스의 현대화 관점에서 제기될 것으로 예상되는 쟁점은 총 두 가지로 요약해 볼 수 있다.

① 검증 가능성

통계 생산 프로세스를 개선하기 위한 새로운 기술의 도입은 기존 방식 대비 우수성이 반드시 확인되어야만 유효성을 증명할 수 있다. 문제는 AI기술의 품질 진단 기준이 불명확하다는 점이다. 1차 통계, 즉 조사 통계의 경우 과거 데이터 또한 기존 방법론을 통해 생산된 자료일 가능성이 높다. 그러므로 새로운 기술을 도입해 데이터의 품질을 높이고자 할 때 참조할 수 있는 데이터 또한 기존 방법론에 의해 생산된 데이터를 기준으로 평가해야만 한다. 이는 결과적으로 객관적인 방법론 우위 측정에는 한계가 발생함을 의미한다. 물론 예외의 경우도 존재한다. 통계의 주제 및 특성에 따라서는 기업 정보, 행정 정보 등 기존 조사결과 외 객관적 정보로 비교 대상을 대체할 수 있는 유형도 존재한다. 이런 경우에는 해당 이슈에서 자유로울 수 있다.

이와 더불어 신규 통계 생산의 예상 쟁점 중 하나와 마찬가지로 알고리즘 특성에 의한 신뢰성 문제가 대두될 가능성이 존재한다. 흥미로운 점은 비지도 학습을 국가통계에 활용한 사례가 해외에는 존재한다는 것이다. 그럼에도 불구하고 해당 기법을 활용한 국가통계 사례는 현재 해당 국가의 국가통계 규정상 정식 승인은 되지 않았으며 국내 도입상의 검증 이슈에서 자유롭지 않을 것으로 예상된다.

통계 생산 프로세스 개선에 빅데이터 및 AI기술을 활용하는데 있어 전통적 통계 기법과 AI알고리즘 간 비교 검증 방안이 발굴 되어야 함을 시사한다.

② 생산 주기

두 번째 예상 쟁점은 AI기술 도입이 통계 생산부터 공표까지의 제한된 생산 기간을 준수할 수 있는 해법인지에 관한 것이다. AI 모델링은 방법론의 종류, 학습 데이터의 규모, 최소 성능 기준 등에 의해 최종 모델 산출까지의 시간이 유동적인 특징이 존재한다. 그럼에도 불구하고 국가 통계는 시의성 및 정시성을 갖춰야만 하는 특성이 있다. 실제 국내 국가 통계 생산과정을 살펴보면 제한된 생산 기간 준수를 위해 더 정확한 방법을 채택하지 않는 사례가 존재한다. 가령 항목 무응답 대체의 정확성은 매우 중요

한 통계 산출 과정 중 하나이나, 무응답 대체 기법으로서 명시적 형태의 대체 모형(explicit model)이 아닌 내재적 모형(implicit model)을 주로 선택한다. 여기에서 명시적 형태의 대체 모형은 무응답 외 데이터 응답의 분포를 통해 결측치를 분포에 맞추어 할당하는 방법이다. 이는 통계의 정확성을 높일 수 있는 기법임에도 불구하고 분석의 난이도가 높고 예산 및 제한된 시간을 준수하기 어려워 준용한 경우가 흔치 않다.

AI기술의 통계 생산 과정의 일부로의 도입은 결국 상기 이유와 마찬가지로 생산 기간을 준수할 수 있는지 여부를 따져봐야 한다. 즉 AI 학습 알고리즘의 학습시간-품질 간 조율 문제가 수면위로 떠오를 가능성이 높다.

제2절 국내 통계 제도의 개편 방향과 정책적 시사점

국내 통계청은 단기적으로 현행법 내 승인기준의 일부 보완과 더불어 중장기적인 신고제도 및 시범통계제도 도입을 검토할 것임을 예고하였다. 결국 두 가지의 보완 방향을 가지고 해당 이슈를 발전시키고자 함을 확인할 수 있는데, 첫 번째는 통계작성 심사표상의 문구를 일부 수정함으로써 조사통계에 국한된 심사 문항을 범용적인 문구로 변경하고 기존 통계 대체를 위한 빅데이터 도입의 경우 중복성 검사에서 자유롭도록 제도적 예외를 부여하는 부분이라 볼 수 있다. 2020년 5월 발표된 관련 개정안은 현재 개정진행 중인 것으로 보이며 현 통계 관리체계를 유지하는 선에서의 개정이므로 잠정적으로 적용될 것으로 짐작해볼 수 있다.

<표3-5> 기존 통계 대체관점의 빅데이터·AI 도입 시 예외 승인 기준(예)

구 분	예외 기준
범위	- 주기나 공표 범위가 기존의 유사 중복에 비하여 상세 하거나 보완하는 내용이 있는가? → 예 : 1년 주기를 분기 또는 월 단위 통계로 단축 가능여부 → 예 : 기존의 통계에 비해 소지역 추정이 가능한지 여부 등
목적	- 기존의 통계와 작성 목적이 동일한가? → 예 : 자동차 검사 기준 시군구별 자동차 주행거리 측정을 내비게이션, 센서 자료 기반의 도로 기반 자동차 주행거리 측정으로 변경할 수 있는지 여부

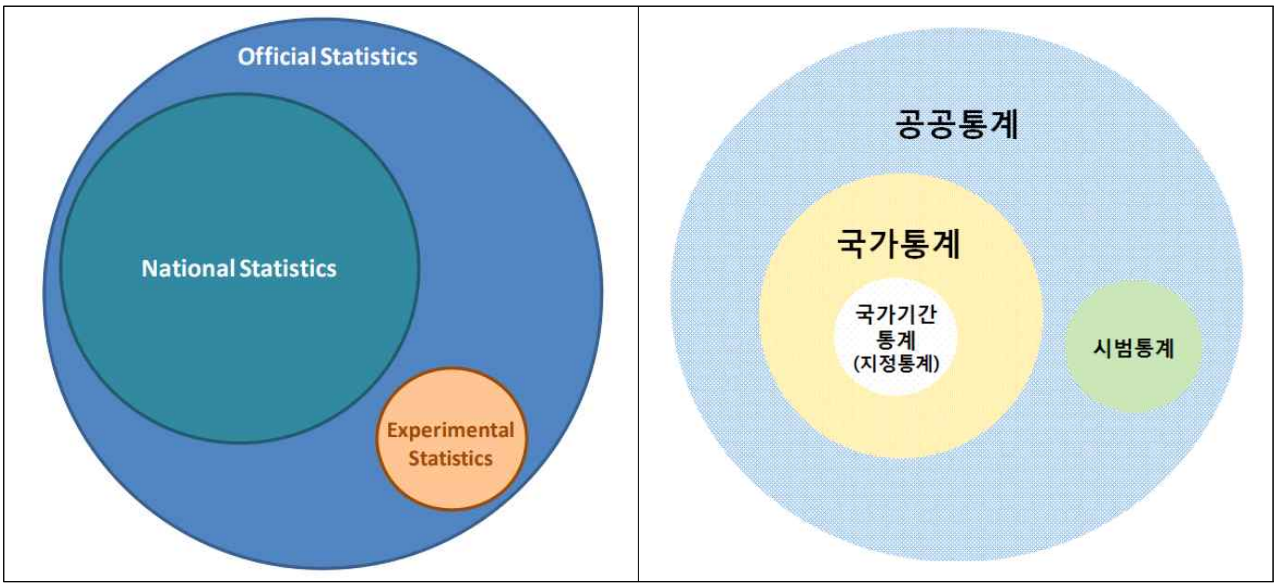
*출처 : 정보통신정책연구원(2019)

두 번째 방향은 시범통계 제도의 신설과 이에 따른 국가통계 관리 체계의 개편이다. 이는 현행 국가통계로 일원화 되어 관리하는 통계 체계를 더 넓은 의미의 공공통계 내 속하는 하나의 개념으로 보고 공공통계에 시범통계라는 새로운 의미의 국가 관리형 통계 체계 운영을 꾀하는 것이다. 여기에서 시범통계란 영국의 실험통계(experiment statistic)를 벤치마크한 국가통계 관리체계를 의미한다.

영국의 실험통계는 새로 개발되거나 혁신적인 통계 개발 시도를 국가가 관리할 수 있도록 하는 제도로서 반드시 국가통계로의 승인을 득하지 않더라도 발전 가능성이 보이거나 실험적인 통계 생산이 필요하다고 판단될 시, 국가통계 승인 심사에 비해 간소화된 심사 기준 하에 선발 및 관리할 수 있다. 즉 AI 및 빅데이터를 활용하는 통계를 국가 차원에서 생산 지원하는데 용이한 제도라고 할 수 있다. 실험통계는 향후 체계내에서 관리되는 통계 데이터가 국가통계로서의 품질 조건을 만족할만큼 발전되었다고 판단될 시 국가통계로 승격 가능한 구조이다. 게다가 통계 생산의 지속 및 폐지가 비교적 자유롭기 때문에 통계 생산자의 능동적 참여를 기대해볼 수 있는 장점을 지닌다.

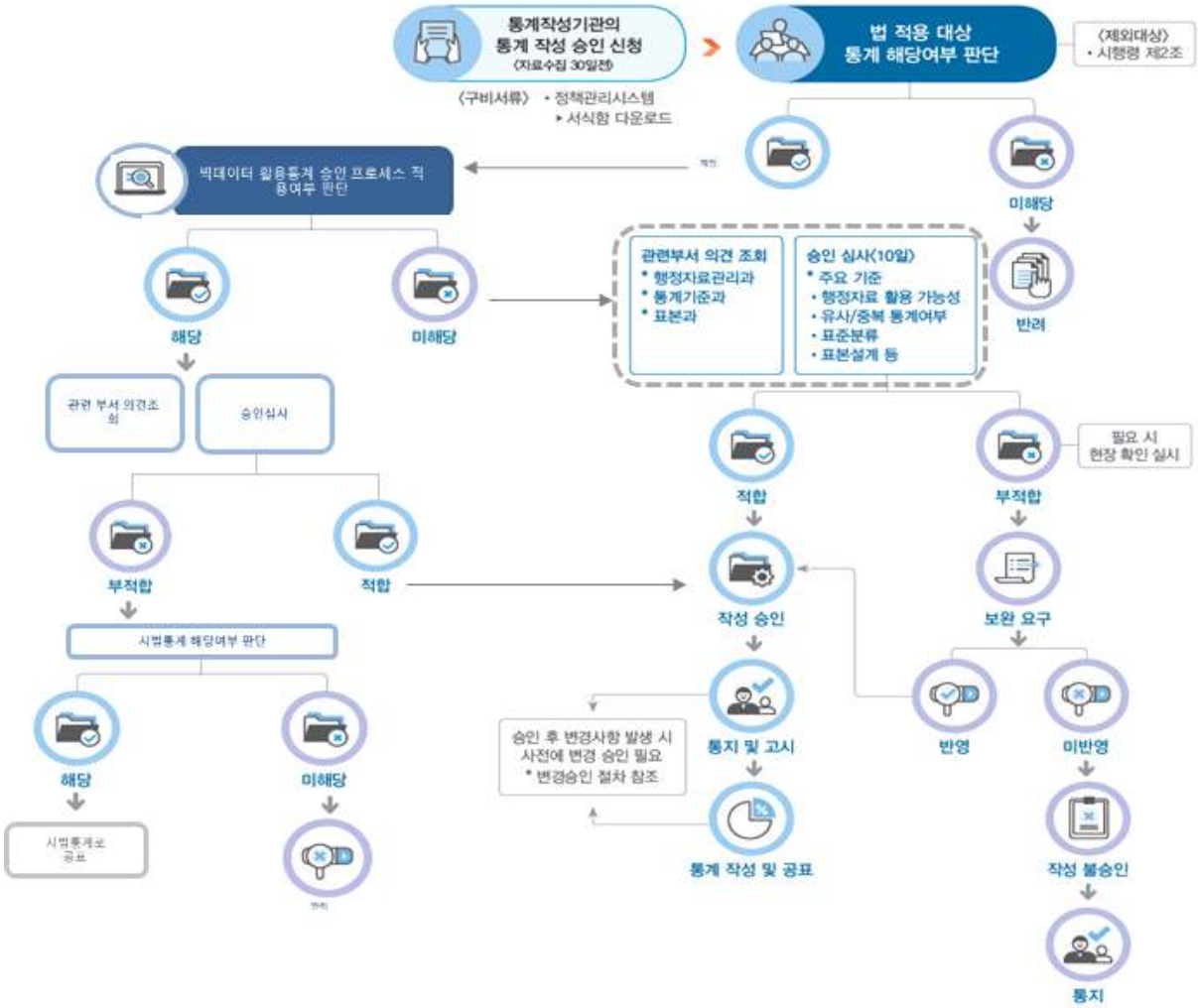
해당 제도는 국가통계에 신기술을 활용해 볼 수 있는 시기를 앞당기는데 의의가 있다. 그러므로 일종의 샌드박스 제도의 범주에 해당한다. 국내 통계청은 국가통계 승인 대비 완화된 기준에 입각한 시범통계 제도 운영을 구상하고 있는 상황으로 당장의 논의는 불가능할 것으로 전망되나, 장기적으로 해당 체계가 구축될 것으로 예상되는 상황이다.

<그림3-3> (좌)영국의 공식통계 개념도와 (우)한국의 시범통계 도입 구상도



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

<그림3-4> 시범통계 승인 절차(예시)



*출처 : 정보통신정책연구원(2019)

이처럼 AI활용 통계의 국가통계 승인 기회를 개방하려는 국내 통계청의 노력은 긍정적으로 평가된다. 이는 시스템적인 기회 부여에 해당하며 불가피한 이유로 빅데이터 및 AI기반의 통계 생산이 필요할 시 국가통계 승인을 고려할 수 있는 여지를 제공하는데 의의가 있다. 그렇다면 국내 통계청의 제도적 개편은 1절에서 언급한 통계 생산의 관점 중 어디에 해당하는가. 현재로서는 신규 통계 생산 관점에서 관련 통계를 정의하고 있음을 알 수 있다. 그러나 신규 통계 생산의 관점은 두 가지 접근법 중 본격적인 도입이 논의되는 과정에서 오히려 더 진입장벽이 높은 방법이다. 결과적으로 실제 AI 및 빅데이터 기반의 활용 통계를 생산하도록 실무자를 유도하고 이를 통해 관련 통계 생산을

활성화 하는 측면에서는 해결해야할 쟁점이 산적해 있음을 알 수 있다. 본 보고서의 1절에서 제시한 쟁점들은 대체적으로 국내 국가통계 제도의 개편이 완료되더라도 여전히 해소되기 어려운 AI 자체의 본질적 사안을 담고 있다. 그러므로 우리는 아래와 같은 시사점을 도출해 볼 수 있다.

<표3-6> 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

① **검증된 알고리즘 공표와 활용 지침 마련**: AI 및 빅데이터 활용 통계의 생산을 유도하기 위해서는 정확히 어떠한 알고리즘이 활용되었을 시 신뢰성을 담보할 수 있으며 어떻게 활용해야만 국가통계로서의 승인 가능성을 판단할 수 있을지에 관한 표준이 필요하다. 여기에는 지침 형태의 가이드라인 뿐만 아니라 관련 표준화 문서 및 활용 가능 알고리즘 별 세부 개발 요구사항이 포함된다.

② **AI·빅데이터 활용 국가통계 생산 절차 수립**: 통계 생산 뿐만아니라 모든 공적 업무에는 업무를 수행하는데 필요한 프레임워크가 존재한다. 현존하는 KSBPM과 이에 영향을 받는 세부 통계 개발 지침들은 조사통계에 기반하고 있고 이러한 체계가 순차적 프로세스를 지향해 AI·빅데이터 활용 통계 생산을 포괄하는데 한계가 존재하므로 이를 해결할 수 있는 신규 생산 프레임워크를 제시하거나 관련 문제를 해소할 수 있는 합리적인 대안이 필요하다.

③ **기술 특성 분석 및 평가 기준의 표준화**: AI 및 빅데이터 분석 방법론을 활용하기

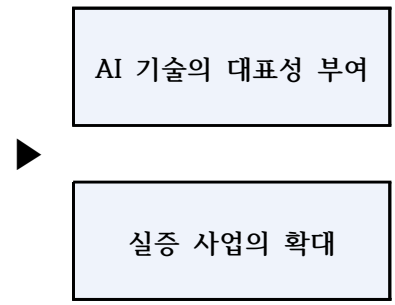
이전에 알고리즘 자체의 구조와 동작형태, 재현가능성 충족을 위한 기준마련 또는 이를 신뢰할 수 있도록 기술적인 도구의 제공 등이 이루어져야 한다. 여기에는 AI 및 빅데이터 방법론을 공적 이슈에 활용하는데 요구되는 다양한 방면의 소프트웨어 연구개발이 병행 추진 되어야 할 것이다.

④ **기존 통계기법-AI 알고리즘의 비교 검증 방안 발굴**: 기존 통계의 완전 대체 및 프로세스 현대화 과정의 일부 통계 생산 방법론 변경 등의 이슈에서 기존 방법론과의 비교 검증이 가능한 일련의 방법 및 절차가 개발되어야 한다. 여기에는 거버넌스적인 절차 뿐만아니라, 방법론 간 상이함에 따른 객관적 품질 측정 방안 또한 포함된다.

⑤ **다수의 실증 사업을 통한 도입 적합성 진단**: 국가통계는 시의성과 정시성을 모두 만족시킬 수 있어야만 효과적으로 관리될 수 있다. AI 및 빅데이터 알고리즘을 활용할 시, 각 통계 생산의 목적에 맞추어 특화된 파라미터 설정과 모델 학습 기간을 요하므로 이러한 시간적 소요와 결과 품질간 균형점을 찾기위한 실험적인 도입 연구가 지속되어야만 관련 기준 정립이 가능하다.

<표3-7> 국가통계의 AI기술 도입 촉진을 위해 필요한 요소

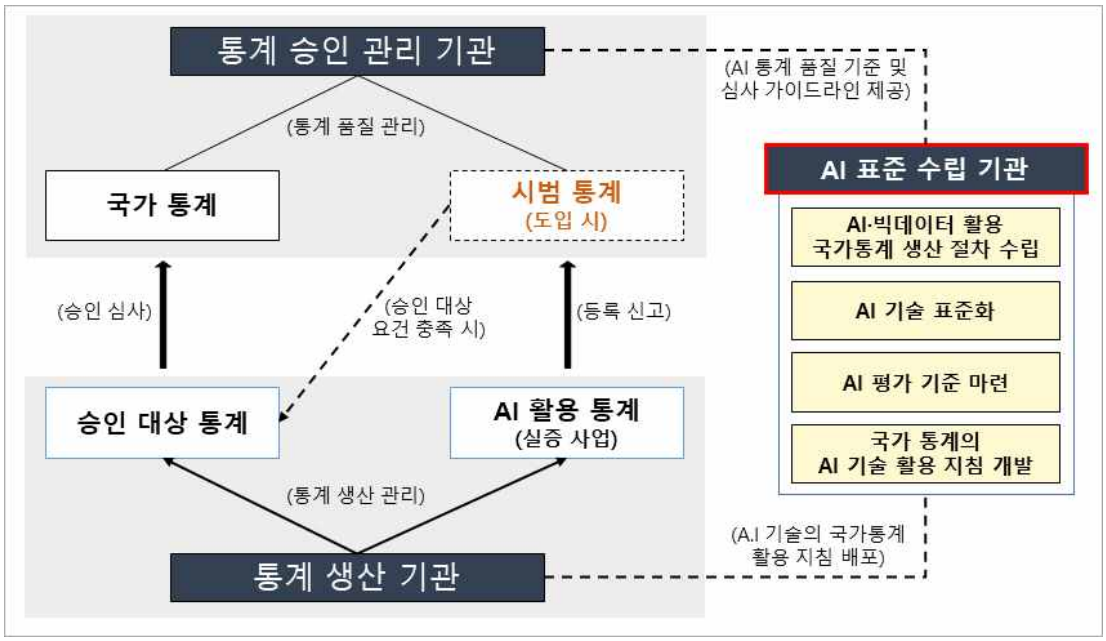
구 분	시사점
신규 통계 생산	검증된 AI 알고리즘 공표 및 세부적 활용 지침
	AI·빅데이터 활용 국가통계 생산 절차 수립
	기술 특성 분석 및 평가 기준 표준화
통계 생산 프로세스의 현대화	기존 통계기법-AI 알고리즘의 비교 검증 방안 발굴
	다수의 실증 사업을 통한 도입 적합성 진단 필요



이와 같은 연구 소요는 총 두 가지의 정책적 방향성으로 정리해볼 수 있다. 첫 번째는 AI기술의 대표성을 부여하는 것이며, 두 번째는 실증사업을 확대하는 것인데, 이를 해결하기 위해서는 현행 통계청의 제도적 정비와 더불어 실질적인 연구개발 및 사업수행 체계가 조성되어야 할 것이다. 이에 부합하는 해외의 정책으로 UN유럽경제위원회가 주도한 빅데이터 샌드박스 프로젝트가 존재한다. 해당 프로젝트는 유럽연합 가입국가들이 동일한 플랫폼 내에서 빅데이터 활용 통계에 관해 논의하고 실험적 통계를 발굴하는 것을 목표로 하고 있는데, 우리가 주목해볼 부분은 이와 같은 플랫폼이 구성되는데

영향을 준 협력 체계이다. 빅데이터 샌드박스 프로젝트의 플랫폼과 관련 논의는 아일랜드 통계청(CSO)이 주도하였는데, 관련 플랫폼의 제공 및 프로젝트 수행에 아일랜드의 슈퍼컴퓨팅 연구소인 ICHEC가 협력 관계를 맺음으로써 유기적인 빅데이터 활용 통계 개발을 꾀할 수 있었다. 이에 착안하여 국내 또한 관련 이슈의 활성화를 위하여 제도적 장치를 검토하는 것뿐만 아니라 실제 연구개발의 기능을 가진 AI·빅데이터 표준 수립 기관의 협력이 필요할 것으로 판단하였다. 아래는 국내 환경에서 시범 통계의 개념이 도입되었음을 가정하였을 때 국가통계의 승인 체계의 변화와 관련 쟁점들을 해결할 수 있는 표준 기관과의 협력 체계를 도식화 하였다.

<그림3-5> AI 중심의 국가통계 혁신을 위한 협력 체계(안)



결론적으로 통계승인관리기관-연구기관 간 AI 중심의 국가통계 혁신을 위한 협력체계 구축은 실질적인 제도 도입의 성과를 상승시키고 국가통계의 혁신 시기를 앞당기는데 기여할 것이라 예측해볼 수 있다. 이와 같은 AI 표준 수립 기관은 AI·빅데이터 활용 국가통계 생산 절차 수립과 기술 표준화, 평가 기준 마련과 함께 국가통계의 AI 기술 활용 지침 개발 등의 역할을 수행할 수 있고 이는 승인관리기관에게는 품질 기준과 심사 전문성에 입각한 심사 가이드라인 제공해주는 측면에서, 생산기관에게는 AI 기술의 국가통계 활용 방법을 제공해 적극적인 참여 유인책으로 작용할 수 있어 모두에게 이로운 효과를 파생할 것으로 예상된다.

제4장 [부록 1] AI·빅데이터를 활용한 디지털산업 분류 방안

본 장은 해당 정책연구의 일환으로 검토 된 AI・빅데이터 활용 유스케이스 발굴 목적으로 제안된 연구 아이디어로서, 금융감독원 전자공시시스템(DART)의 전자공시 빅데이터를 활용한 신규 통계 지표 산출 방안을 소개하였다.

제1절 연구 배경과 목적

1. 연구 배경

전 세계적으로 경제의 디지털화(digitalization)가 빠른 속도로 진행되면서 경제 전반에서 디지털경제(digital economy)가 차지하는 비중도 커지고 있다. 이에 따라 경제, 산업, 사회 전반의 문제와 한계를 분석하고 정책을 입안하는 데 있어 디지털경제를 정확하게 측정하고자 하는 수요도 증가하고 있다. 특히 각 국가의 정부는 디지털경제를 새로운 경제성장 동력으로 인식하고, 디지털산업 육성을 통한 국가경제 성장률과 경쟁력 제고를 도모하고 있으며, 경제정책의 목표와 성과 관리를 위해 디지털경제를 정확하게 측정하고 싶어 한다.

디지털경제 측정은 두 가지 방법으로 이루어진다. 첫 번째는 디지털경제의 발달 정도를 간접적으로 추정하기 위해 디지털전환지수(digital tranformation indicators)를 산출하는 것이다. 두 번째는 디지털경제의 구성과 규모를 직접적으로 파악하기 위해 국내총생산(gross domestic product, GDP)을 산출하는 것처럼 디지털경제 규모를 측정하는 것이다. 그러나 2014년부터 디지털경제 측정의 필요성이 본격적으로 논의되어 왔음에도 불구하고 디지털경제를 측정하는 수준은 아직까지 시범적인 단계에 머물러 있다(안성희 · 유철종, 2019).

특히 최근 들어 디지털경제에 대한 정의가 합의되는 과정에 있으나 디지털경제 규모를 통일된 기준으로 측정하고 비교하는 단계까지는 이르지 못한 상태다. 또한 디지털

경제 규모를 측정하는 방법도 일부 국가에 의해 실험적으로 제시되거나 각 국가마다 디지털경제를 측정할 수 있는 통계 기반과 역량이 달라 국제적으로 통일된 기준을 마련하는 데 시간이 걸리는 것으로 보인다. 예를 들면, OECD(2020)는 그간의 오랜 논의와 작업을 종합하여 디지털경제 측정을 위한 공통 체계(a common framework for measuring digital economy)를 제안하고 있지만, 디지털경제 규모를 정확하게 측정하기 위해서는 풀어야할 과제가 여전히 많다는 점을 인식하고 우선적으로 G20 회원국 간 비교 가능한 디지털경제 통계를 산출하기 위한 로드맵(roadmap)을 제시한 상태다. UNCTAD(2019)도 기존의 정보경제보고서(Information Economy Report)를 디지털경제보고서로 개명하고, 각 국가의 디지털경제 지표와 규모를 산출하여 보고하고 있으나, 디지털경제 규모를 측정하는 데 여러 가지 한계가 존재함을 적시하고 있다.

G20 회원국 중에서 디지털경제 규모를 측정하는 데 있어 가장 앞서 나가는 국가는 미국, 캐나다, 호주, 중국으로 파악된다(ABS, 2019; BEA, 2020; CAICT, 2020; OECD, 2020; Statistics Canada, 2019). 우리나라의 경우 경제통계를 관할하고 한국은행이 OECD의 디지털경제 측정 전문가 회의(Working Party on Measurement and Analysis of the Digital Economy, WPMADE)에 참가하여 국제적 논의에 참여하는 정도이고, 디지털경제 규모 측정과 관련하여 실험적 연구나 구체적 계획은 아직까지 발표되지 않았다(안성희 · 유철종, 2019). 또한 우리나라의 경우 디지털경제 규모 측정과 관련하여 학계나 민간의 연구도 해외의 연구가 활발한 것과 대조적으로 거의 전무한 것으로 파악된다.[9)]

2. 연구 목적

본 장에서는 상기와 같은 논의와 배경을 바탕으로 디지털경제 규모를 측정하기 위한 사전작업으로 우리나라에서 「주식회사등의외부감사에관한법률」에 따라 외부감사 대상에 해당되는 기업의 경영공시 데이터를 AI · 빅데이터 기술로 분석해 디지털산업을 분류하는 방안을 제시하고자 한다. 이는 디지털경제에 포함시켜야 할 산업의 범위가 명확하게 결정되지 않으면 디지털경제 규모를 정확하게 측정할 수 없고, 디지털 재화와

9) 국내 문헌 대부분은 대개 디지털경제 측정과 관련된 국제적 논의를 정리하는 수준에 그치고 있고, 구체적인 측정 방법이나 결과를 제시한 문헌은 없는 것으로 파악된다(김건우, 2016; 안성희 · 유철종, 2019; 이금노, 2020; 한국은행, 2017a; 한국은행, 2017b; 황석원, 2020)

서비스를 생산하거나 제공하는 산업의 범위에 따라 디지털경제 규모도 다르게 측정되기 때문이다.

제2절 기존의 디지털산업 분류 방법과 한계

1. 디지털경제와 디지털산업 정의

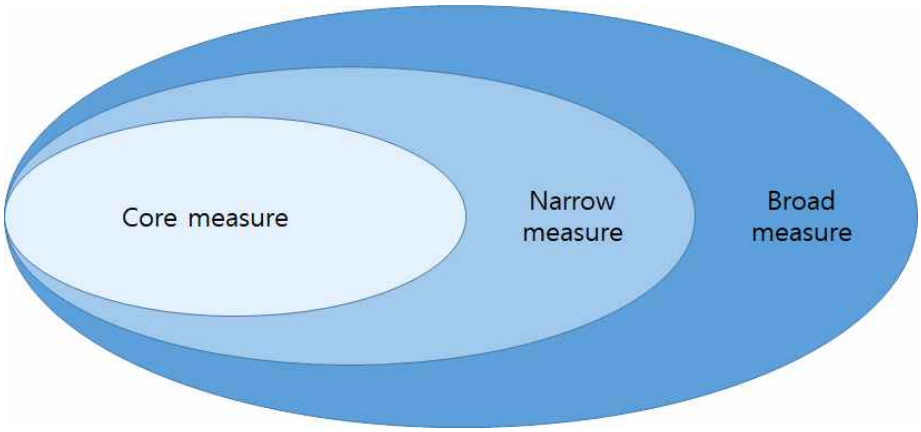
(1) 디지털경제 정의

2014년부터 디지털경제 측정에 대한 필요성이 국제적으로 제기되고 논의되기 시작한 이후 2020년 말 현재 디지털경제에 대한 정의는 어느 정도 합의되어 가는 과정에 있다. 그럼에도 불구하고 OECD(2020)가 지적한 것처럼 디지털경제 정의는 여전히 모호할 수 있으므로 디지털경제 규모를 정확하게 측정하려면 그 정의를 구체화하는 작업이 필요하다. 그러기 위해서는 먼저 그간 진전된 디지털경제에 대한 정의를 자세히 살펴보고 이해할 수 있어야 한다.

한국은행(2017b)은 디지털경제를 정보통신기술(Information and Communication Technology, ICT)에 기반한 경제라고 간단하게 정의한 바 있으며, 이후 디지털경제 정의를 별도로 제시하지 않은 것으로 파악된다. 이와 달리 디지털경제 정의에 대한 국제적 논의는 2017년부터 상당히 구체적으로 진전되었다. 특히 OECD(2020)는 그간의 국제적 논의를 종합한 결과를 토대로 디지털 재화와 서비스의 생산, 디지털경제를 구성하는 산업, 디지털 재화와 서비스의 거래 등 다양한 측면을 포괄할 수 있는 정의를 계층화(tiered way)하여 제시하였다.

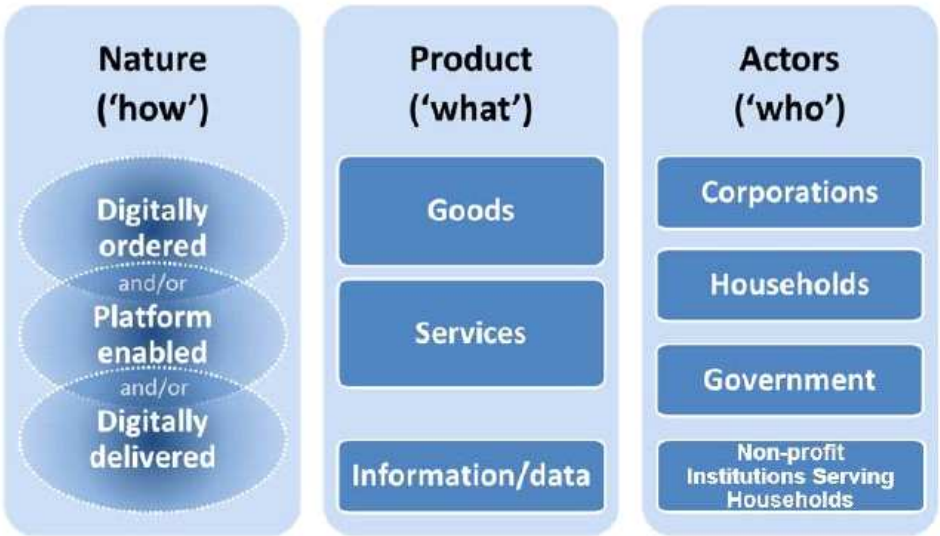
OECD(2020)가 제시한 생산(production) 활동 측면에서의 디지털경제 정의는 다음과 같다. 디지털경제는 디지털 기술, 디지털 인프라, 디지털 서비스 및 데이터 등 디지털 투입물의 사용에 의존하거나 그 사용에 의해 상당히 강화되는 모든 경제 활동을 뜻한다. 생산 활동 측면에서의 디지털경제 정의는 <그림2-1>에서 볼 수 있듯이 핵심 정의(core measure), 협의(narrow measure), 광의(broad measure)로 구분될 수 있다. 디지털경제의 핵심 정의에는 ICT 재화와 디지털 서비스의 생산자로부터의 경제 활동만을 포함한다. 협의에는 디지털 투입물에 의존하는 기업으로부터의 경제활동까지 포함한다. 광의에는 디지털 투입물의 사용에 의해 상당히 강화되는 경제 활동까지 포함한다.

<그림 4-1> 디지털경제 정의



또한 OECD(2020)는 생산 활동이 아닌 거래(transaction) 활동 관점에서 디지털경제를 '전자적으로 주문되거나 전자적으로 배송되는 모든 경제 활동(all economic activities that is digitally ordered and/or digitally delivered)' 으로 정의할 수 있다고 보았다. 이는 <그림2-2>에서 살펴볼 수 있듯이 IMF(2018)가 제시한 디지털경제의 대안적 정의에서 '플랫폼에서 이용되는(platform-enabled)'가 제외된 것처럼 보이지만 그렇지 않다. 플랫폼에서 이용되는 거래는 전자적으로 주문되고 배송되는 거래와 동일하다.

<그림 4-2> 디지털 거래의 세 차원



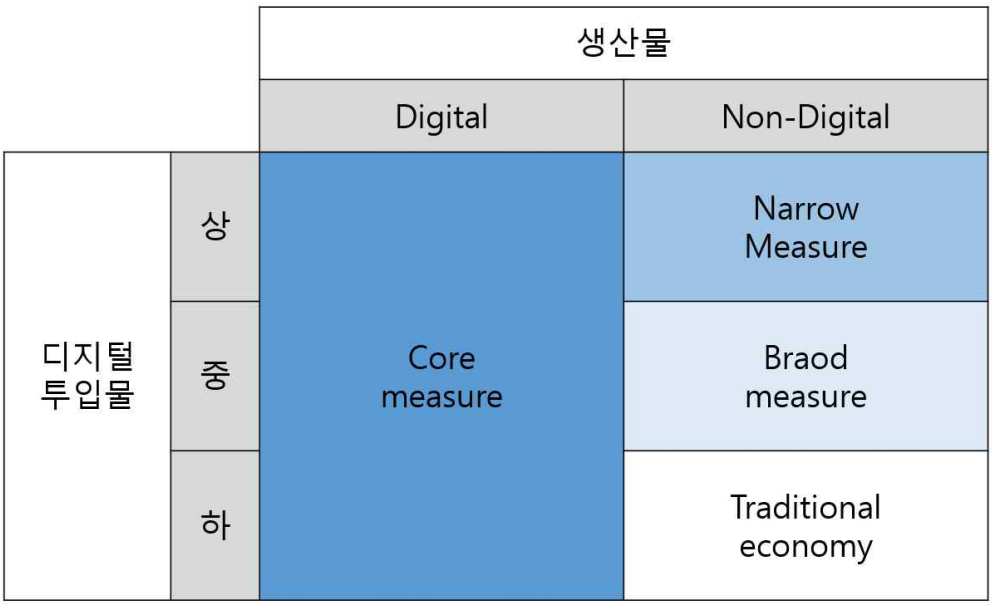
자료: IMF(2018), Frontier and Matei(2017)

(2) 디지털산업 정의

디지털경제 정의에 따라 디지털산업 범위도 달라진다. OECD(2020)는 디지털경제를 생산 활동과 거래 활동 관점에서 각각 정의하였다. 이에 따라 디지털산업을 정의하는 방식도 달라진다. 예를 들면, 디지털경제를 생산 활동 관점에서 정의할 경우 기존의 산업 분류를 활용할 수 있다는 장점이 있으나, 디지털 투입물의 투입 정도에 따라 디지털산업 여부가 결정되기도 있다. 디지털경제를 거래 활동 관점에서 정의할 경우 디지털 거래 여부를 판단해야 디지털산업을 정의할 수 있다.

생산 활동 관점에서 디지털경제를 정의할 경우 <그림2-3>에 나타난 바와 같이 최종 생산물이 디지털 재화 또는 서비스에 해당되는지 여부와 디지털 투입물이 생산요소로서 투입된 정도에 따라 디지털산업 여부가 결정된다. 예를 들면, 최종 생산물이 디지털 재화나 서비스이면 그 산업은 디지털산업에 속한다. 그러나 최종 생산물이 디지털 재화나 서비스가 아니면 OECD(2020)는 디지털 투입물의 투입 정도가 상(high)이면 협의에 따라, 중(medium)이면 광의에 따라 디지털산업에 속한다고 보았다. 디지털투입물이의 투입 정도가 하(low)이면 디지털산업에 속하지 않는다.

<그림 4-3> 생산 활동 관점에서의 디지털산업 정의



자료: OECD(2020)

또한 거래 활동 관점에서 디지털경제를 정의할 경우 전자적으로 주문되거나 전자적으로 배송되는 재화나 서비스를 생산하는 산업은 디지털산업에 속하나, 전자적으로 주문되지도 않고 전자적으로 배송되지 않는 산업은 디지털산업에 속하지 않는다. 예를 들면, 디지털 콘텐츠는 대표적으로 전자적으로 주문되고 배송되는 재화이다. 디지털 금융서비스도 마찬가지다. 전자상거래를 통해 거래되는 재화나 서비스는 전자적으로 주문되는 대표적인 사례다. 숙박, 비행기, 택시 예약, 배달 서비스도 이에 해당될 수 있다. 반면에 점포에서 개설된 이동통신 서비스를 제외하고는 전자적으로 주문되지 않으나 전자적으로 배송되는 재화나 서비스를 찾기 어렵다.

<그림 4-4> 거래 활동 관점에서의 디지털산업 정의

		배송	
		Digital	Non-Digital
주문	Digital		
	Non-Digital		

자료: OECD(2020)

2. 현행 디지털산업 분류 방법

OECD(2020)가 제시한 디지털경제 정의에 따라 디지털산업을 개념적으로 정의할 수는 있으나, 디지털산업을 식별하고 분류하는 작업은 완전히 별개의 문제다. 생산 활동 측면에서 디지털 투입물의 투입 정도를 측정할 수 있어야 하고, 그 정도를 판단할 수 있는 기준이 있어야 한다. 또는 거래 활동 측면에서 각 기업의 재화나 서비스가 전자적으로 주문되거나 전자적으로 배송되는지를 파악할 수 있어야 하고, 그 정도를 판단할 수 있는 기준이 있어야 한다. 이 점을 감안하고 디지털경제 규모를 측정하기 위해 디지털산업을 분류한 미국, 캐나다, 호주, 중국, 영국 사례를 토대로 현행 디지털 분류 방법을 고찰한다.

(1) 미국

미국 상무부(Department of Commerce, DOC) 산하의 경제분석국(Bureau of Economic Analysis, BEA)은 2018년 3월에 처음으로 디지털경제 정의와 측정에 관한 보고서를 발표하였고, 2019년 4월에 디지털경제 규모 측정치를 업데이트하였으며, 2020년 8월에 다시 새로운 디지털경제 규모 측정치를 발표하였다(BEA, 2018; BEA, 2019, BEA, 2020). 경제분석국은 2018년 3월 보고서에서 주로 디지털(primarily digital) 재화와 서비스만을 생산하거나 공급하는 산업만을 고려하였으나, 2020년 8월 보고서에서는 부분적으로 디지털(partially digital) 재화와 서비스를 생산하거나 공급하는 경제 활동까지 포함하고자 하였다(BEA, 2020).

경제분석국이 디지털경제 규모를 측정하기 위해 채택한 디지털산업 분류 방법은 다음과 같다. 먼저 디지털경제를 정의하였고, 다음으로 공급사용표(Supply-Use Tables, SUTs)를 활용해 디지털경제에 속하는 재화와 서비스를 식별하였으며, 이렇게 식별된 재화와 서비스를 생산하거나 공급하는 산업을 미국의 산업분류코드(NAICS)를 활용해 디지털산업으로 분류하였다. 또한 부분적으로 디지털 재화나 서비스를 생산하거나 공급하는 경제 활동을 포착하기 위해 디지털 재화와 서비스의 비중을 각종 경제 데이터를 활용하여 추정하였다.

우선 BEA(2018)와 BEA(2020)는 디지털경제를 ‘주로 인터넷과 이와 관련된 정보통신기술(ICT)에 기반한 경제 활동’으로 정의하였다. 그러나 이러한 정의만으로는 디지털경제의 경계를 명확하게 구분 짓기 쉽지 않다. 이에 따라 BEA(2018)는 디지털산업을 구체적으로 정의하는 방식으로 디지털경제 정의의 모호성을 보완하고자 하였다. 예를 들면, 컴퓨터 네트워크가 존재하고 운영되기 위해 위한 디지털 지원 인프라(digital-enabling infrastructure), 그 시스템을 통해 일어나는 디지털 거래(e-commerce), 디지털경제 사용자가 제작하고 접근하는 콘텐츠(digital media)와 관련된 산업이 디지털산업으로 분류될 수 있다고 보았다.

그러나 산업분류코드에 기반하여 분류되는 재화와 서비스 중에는 디지털(digital)과 비디지털(non-digital)을 구분하기 어려운 것들도 많다. 더구나 이를 구분할 수 있는 데이터가 항상 존재하는 것도 아니다. 이를 감안하여 BEA(2018)에서는 디지털 재화나 서비스를 생산하거나 공급하는 산업만을 디지털산업으로 간주하였으며, 개념적으로는 디지털산업에 포함되나 데이터의 한계로 경제 활동의 측정할 수 없는 경우를 제외하였

다.

BEA(2020)에서 제시한 디지털산업의 범위는 BEA(2018)보다 더 포괄적이다.[10] 우선 디지털산업의 범주를 <표 5-1>에 나타난 바와 같이 디지털 지원 인프라(digital-enabling infrastructure), 전자상거래(e-commerce), 가격이 있는 디지털 서비스(priced digital services)로 재정의하였다. 그러나 BEA(2018)에서와 같이 디지털 지원 인프라 중 구조물을 제외하였다.[11] 전자상거래에는 BEA(2018)에서와 달리 판매자와 구매자 간의 직접 상거래를 중개하는 플랫폼 사업자를 포함하였다. 가격이 있는 디지털 서비스 중 디지털 중개 서비스는 포함하지 않았다.

<표 4-1> 미국의 디지털산업 분류 현황(2020년 8월 기준)

범주	하위 구성	포함 정도
디지털 지원 인프라	하드웨어(Hardware)	거의 포함
	소프트웨어(Software)	거의 포함
	구조물(Structures)	미포함
전자상거래	B2B(business-to-business)	거의 포함
	B2C(business-to-consumers)	거의 포함
가격 있는 디지털 서비스	클라우드 서비스(cloud services)	거의 포함
	디지털 중개 서비스(digital intermediary services)	미포함
	그 외 디지털 서비스(all other digital services)	거의 포함

자료: BEA(2020)

또한 BEA(2020)에서는 전자상거래와 클라우드 서비스와 관련된 경제 활동을 정확하게 포착하기 위해, B2C 전자상거래에 대해서는 연간 소매거래 조사(Annual Retail Trade Survey, ARTS)를, B2B 전자상거래에 대해서는 연간 도매거래 조사(Annual Wholesale Trade Survey, AWTS)를, 클라우드 서비스에 대해서는 Economic Census 데이터를 활용해 각각의 디지털경제 규모의 비중을 추정하였다. 그 결과, BEA(2018)에서 3,156억 달러였던 2017년 전자상거래 경제 규모가 7.838억 달러(B2C가 2,315억 달러,

10) 디지털산업의 구체적인 산업분류코드(NAICS)는 BEA(2018)과 BEA(2020)의 부록을 참고한다.
11) 디지털 지원 구조물로는 디지털 재화나 서비스를 생산하거나 공급하는 건물 또는 이를 지원하는 건물이 해당된다.

B2B가 5,523억 달러)로 두 배 이상 증가하였고, 클라우드 서비스의 경제 규모는 2018년에 1,110억 달러로 추정되었다.

(2) 캐나다

캐나다 통계청(Statistics Canada)는 2019년 5월에 처음으로 캐나다의 디지털경제 규모를 측정할 목적으로 디지털경제 활동 측정에 관한 보고서를 발표하였다(Statistics Canada, 2019). Statistics Canada(2019)는 디지털경제를 BEA(2018)과 동일하게 정의하였다. 또한 디지털산업을 분류하는 방법도 기본적으로 BEA(2018)의 방법을 따랐다. 그러나 Statistics Canada(2019)가 제시한 디지털산업은 BEA(2018)보다 더 포괄적이다. 예를 들면, Statistics Canada(2019)는 BEA(2018)에서 다루지 않은 교육서비스와 디지털 금융서비스도 디지털경제 활동 중 하나로 인식하였다. 또한 부분적 디지털 재화나 서비스도 광범위하게 포괄하고자 하였다.

통계청은 BEA(2018)와 BEA(2020)과 다르게 산업분류코드에 기반하여 디지털산업을 분류하지 않고 공급사용표(SUTs)에 기반하여 디지털 재화와 서비스(이후, 디지털 재화로 통일)를 분류하는 방법을 택하였다.[12] 그렇다하더라도 디지털산업의 범주는 BEA(2018)와 유사하게 디지털 지원 인프라(digital-enabling infrastructure), 전자적으로 주문되는 거래(digitally-ordered transactions), 전자적으로 배송되는 재화(digitally-delivered products)로 정의하였다. 다음으로 캐나다의 공급사용표(SUTs)를 활용해 각 범주에 속한 디지털 재화가 완전한(full) 또는 부분적(partial) 디지털 재화에 해당하는지를 구분하여 인식하였다.[13]

우선 통계청은 <표 5-2>에서 살펴볼 수 있듯이 완전한 디지털 재화와 부분적 디지털 재화를 구분하고자 하였다. 예를 들면, 컴퓨터와 그 주변기기나 부품들은 완전한 디지털 재화에 속한다. 소프트웨어나 인터넷 접속 서비스도 완전한 디지털 재화에 속한다. 이와 달리 디지털 구성요소(digital components)와 비디지털 구성요소(non-digital components)를 모두 지닌 도서나 신문 같은 디지털 콘텐츠는 부분적 디지털 재화에 속한다.

12) 통계청은 산업분류코드 대신에 공급사용표에서 사용하는 재화코드(Supply and Use Product Code)에 기반하여 디지털 재화와 서비스를 인식하였다.
13) 디지털 재화의 구체적인 목록은 Statistics Canada(2019)의 부록을 참고한다.

통계청은 공급사용표를 활용해 각 범주에 속한 디지털 재화를 인식하는 과정에서 다음과 같은 사항을 고려하였다. 먼저 디지털 지원 인프라 범주에 완전한 디지털 재화로서 모든 컴퓨터 하드웨어와 소프트웨어, 통신 재화와 서비스, 이를 지원하는 서비스를 포함하였다. 다만 앞서 설명한 바와 같이 지원 서비스 중 교육 서비스는 부분적 디지털 재화로 인식하였다. 디지털경제를 지원하는 정부부처와 규제당국의 재화와 서비스는 포함되지 않았다. 미국과 같이 디지털 구성요소와 비디지털 구성요소를 정확하게 구분하기 어려워 구조물(structures)과 사물인터넷(internet of things, IOT)도 포함하지 않았다.

<표 4-2> 캐나다의 디지털 재화 분류 현황(2019년 5월 기준)

범주	하위 구성	포함 정도
디지털 지원 인프라	하드웨어(Hardware)	full
	소프트웨어(Software)	full
	통신(Telecommunications)	full
	지원 서비스(Support services) - 데이터 관련, 컴퓨터 장치 임대 및 리스, 컴퓨터 시스템 설계 및 관련 서비스 - 교육 서비스	full partial
전자적 주문 거래	도매(Wholesale trade)	partial
	소매(Retail trade)	partial
전자적 배송 재화	신문, 잡지, 도서 출판물	partial
	영화, 음악, 방송 콘텐츠	full
	은행 및 예금신용 서비스	partial
	여행 관련 서비스	full

주석: 하위 구성은 세부 목록을 토대로 임의로 구성
자료: Statistics Canada(2019)

전자적으로 주문되는 거래 범주에는 기본적으로 전자상거래로 판매하고 소비할 수 있는 재화가 모두 포함되어 있다. 그러나 전자상거래에 의한 디지털경제 활동을 인식하는 Statistics Canada(2019)의 접근방식은 BEA(2018)과 다르다. 우선 BEA(2018)처럼 전자상거래가 속한 산업분류코드를 사용하지 않았다. 그 대신에 각각의 재화의 도매

또는 소매 거래 중 전자적으로 주문되는 거래 비중을 연간 도매거래 조사(Annual Wholesale Trade Survey), 연간 소매거래 조사이(Annual Retail Trade Survey)와 연간 비매장 소매거래 조사(Annual Retail Non-store Survey)를 활용해 추정하는 방식을 택하였다.

전자적으로 배송되는 거래 범주에는 영화, 음악, 방송 등 미디어 콘텐츠와 이와 관련된 콘텐츠 저작권, 공급 및 광고가 완전한 디지털 재화로 포함되었다. 또한 무료 미디어 콘텐츠 관련 광고 이윤과 개인 소득 데이터를 활용해 무료 미디어 콘텐츠를 통한 디지털경제 활동도 포함하고자 하였다. 한편 비디지털 구성요소를 지닌 신문, 잡지, 도서 등 출판물은 부분적 디지털 재화로 인식하였고, 디지털 버전의 출판물 판매 비중을 활용해 각각의 디지털경제 활동의 비중을 추정하였다. 마지막으로 전자적으로 배송되는 거래 범주에 부분적 디지털 재화로 은행과 신용조합의 디지털 금융서비스를 포함하고자 하였다. 다만 금융서비스의 디지털 구성요소와 비디지털 구성요소를 구분하는 것이 쉽지 않은 점을 고려해 명확하게 구분되는 금융서비스만을 고려하였다.

(3) 호주

호주 통계국(Australian Bureau of Statistics, ABS)은 2019년에 처음으로 호주의 디지털경제 활동 측정에 관한 보고서를 발표하였다(ABS, 2019). ABS(2019)는 기본적으로 미국의 BEA(2018)의 방법론을 따랐다. 디지털산업의 범주를 디지털 지원 인프라(digital enabling infrastructure), 디지털 미디어(digital media), 전자상거래(e-commerce)로 구분하였고, 호주의 공급사용표 데이터를 활용하여 세 범주에 속하는 디지털 재화를 식별하고, 각 디지털 재화를 생산하거나 공급하는 디지털산업을 호주의 산업분류표에 따라 분류하였다.[14]

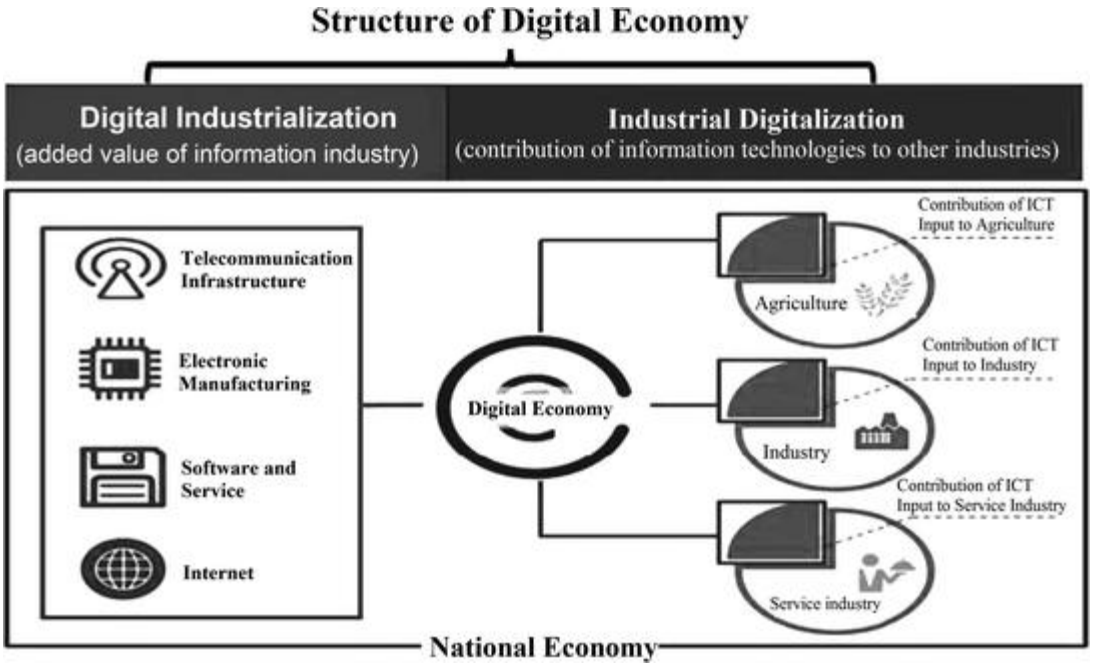
(4) 중국

중국정보통신기술연구소(中国信息通信研究院, CAICT)는 2017년 7월에 처음으로 중국의 디지털경제 규모를 추정한 보고서(中国数字经济发展白皮书)를 발표하였으며, 2020년 7월에 중국의 디지털경제 발전 현황을 업데이트한 보고서(Digital Economy

14) 디지털 재화와 디지털산업의 구체적 목록은 ABS(2019)를 참고한다.

Development in China)를 발표하였다(CAICT, 2017a; CAICT, 2020). 우선 중국은 여타 국가와 다르게 디지털경제를 <그림 5-5>에서 볼 수 있듯이 매우 광범위하게 정의하고 있다. 디지털경제의 구조를 디지털 산업화(digital industrialization)와 산업 디지털화(industry digitization)로 구분하고, 정보통신기술(ICT) 등 기초적인 디지털산업뿐만 아니라 전통적인 산업의 디지털경제 활동 비중을 계량분석을 통해 추정하는 방식으로 사실상 거의 모든 산업을 디지털산업에 포함시켰다.

<그림 4-5> 중국의 디지털경제 구조와 범위



자료: CAICT(2017a, Chinese Academy of Cyberspace Studies(2017)

예를 들면, CAICT(2017a는 전통적인 제조업, 서비스업, 농업에서의 디지털경제 비중을 각각 17.0%, 29.6%, 6.2%로 산출하였다. 또한 <표 5-3>에서 살펴볼 수 있듯이 제조업의 10개 부문, 서비스업의 15개 부문, 농업의 4개 부문의 디지털경제 비중을 각각 산출하여 보고하였다. 이 점에서 중국의 디지털산업 분류는 앞서 살펴본 국가보다 매우 광범위하다.

<표 4-3> 중국의 전통적인 산업의 디지털경제 비중

구분	산업	비중(%)
제조업 (工业)	문화 및 사무용 기계(文化、办公用机械)	58.8
	기기 및 계량기(仪器仪表)	47.3
	기타 전기기계 및 기자재(其他电气机械和器材)	25.6
	기타 전용 설비(其他专用设备)	24.0
	송전 배전 및 제어 설비(输配电及控制设备)	23.1
	기타 통용 설비(其他通用设备)	22.7
	가구(家用器具)	20.9
	금속가공 기계(金属加工机械)	20.3
	전기기계(电机)	18.7
	선박 및 관련 장치(船舶及相关装置)	18.4
서비스업 (服务业)	보험(保险)	46.2
	방송, TV, 영화 및 영상 제작(广播、电视、电影和影视录音制作)	45.4
	전문 기술 서비스(专业技术服务)	40.5
	화폐금융 및 기타 금융서비스(货币金融和其他金融服务)	40.3
	자본시장 서비스(资本市场服务)	40.2
	공공관리 및 사회조직(公共管理和社会组织)	38.0
	우편(邮政)	35.4
	기타 서비스(其他服务)	34.1
	교육(教育)	33.2
	사회보장(社会保障)	32.3
	임대(租赁)	31.6
	수상운송(水上运输)	29.3
	철도운송(铁路运输)	28.7
	문화예술(文化艺术)	28.3
	과학기술 진흥 및 응용 서비스(科技推广和应用服务)	26.6
농업 (农业)	임산물(林产品)	10.6
	수산물(渔产品)	8.2
	농산물(农产品)	6.4
	축산물(畜牧产品)	3.9

자료: CAICT(2017a

(5) 영국

영국의 경우 2013년 7월에 민간기구인 Growth Intelligence(GI)와 National Institute of Economic and Social Research(NIESR)에서 빅데이터를 활용한 디지털경제 측정에 관한 보고서를, 2014년 8월에 국가통계청(Office for National Statistics, ONS)에서 디지털경제 측정에 관한 자문 보고서를 발표한 바 있다(GI and NIESR, 2013; ONS, 2014). 또한 디지털, 문화, 미디어 및 스포츠부(Department of Digital, Culture, Media and Sport, DCMS)에서 2010년부터 디지털 섹터에 국한하여 디지털경제 활동과 관련된 통계를 발표하고 있다.

우선 GI and NIESR(2013)는 기업의 섹터 상황(sector context), 재화 유형(product type), 판매 과정(sales process), 고객 유형(client)을 빅데이터로 분석해 디지털산업에 속할 확률을 산출하고, 일정 확률 이상의 기업을 디지털산업으로 분류하는 방법을 채택하였다. 그 결과, 2013년 6월에 영국 정부가 표준산업코드(Standard Industrial Classification, SIC)를 활용해 식별한 12만 개 기업보다 두 배 이상 많은 약 27만 개 기업이 디지털경제에서 활동하고 있는 것으로 파악되었다.

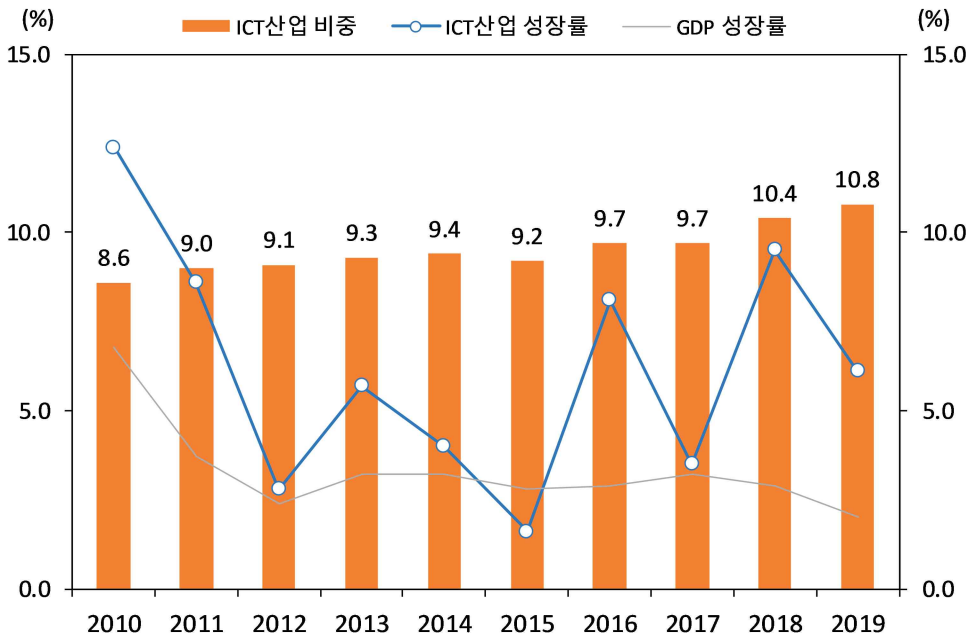
한편 ONS(2014)와 DCMS(2020)에서는 디지털경제를 정보통신기술(ICT)과 디지털 컨텐츠와 관련된 산업을 디지털산업으로 분류하는 방식으로 정의하고 있다. 그 외 디지털경제 활동에 참여하는 기업이나 산업을 고려하지 않고 있다. 예를 들면, DCMS(2020)는 디지털 섹터를 전자제품 및 컴퓨터 제조업(manufacturing of electronics and computers), 컴퓨터 및 전자제품 도매업(wholesale of computers and electronics), 출판업(publishing)(번역 및 통역 활동 제외(excluding translation and interpretation activities)), 소프트웨어(software publishing), 영화, TV, 비디오, 라디오 및 음악(film, TV, video, radio and music), 통신(telecoms), 컴퓨터 프로그래밍, 컨설팅 및 관련 활동(computer programming, consultancy and related activities), 정보서비스 활동(information service activities), 컴퓨터 및 통신장비 수리(repair of computers and communication equipment)로 정의하고 있다.

3. 현행 방법의 한계와 개선 필요성

현재 디지털경제를 정확하게 측정하고 디지털경제 활동을 실체적으로 포착할 목적으로 디지털산업을 구체적으로 분류하는 국가는 많지 않다. 정보통신기술(ICT)과 직접적

으로 관련 있는 산업을 대략적으로 디지털산업 또는 디지털섹터로 구분하고 있는 국가가 더 많다. 우리나라도 과학기술정보통신부(이하, 과기부)가 주관하여 ICT산업 통계를 발표하고 있는 정도다. 예를 들면, 과기부 산하 정보통신정책연구원(KISDI)에서 <그림 5-6>에서 살펴볼 수 있듯이 ICT산업 관련 통계가 발표되고 있다.

<그림 4-6> 한국의 ICT산업 비중과 성장률(실질 기준)



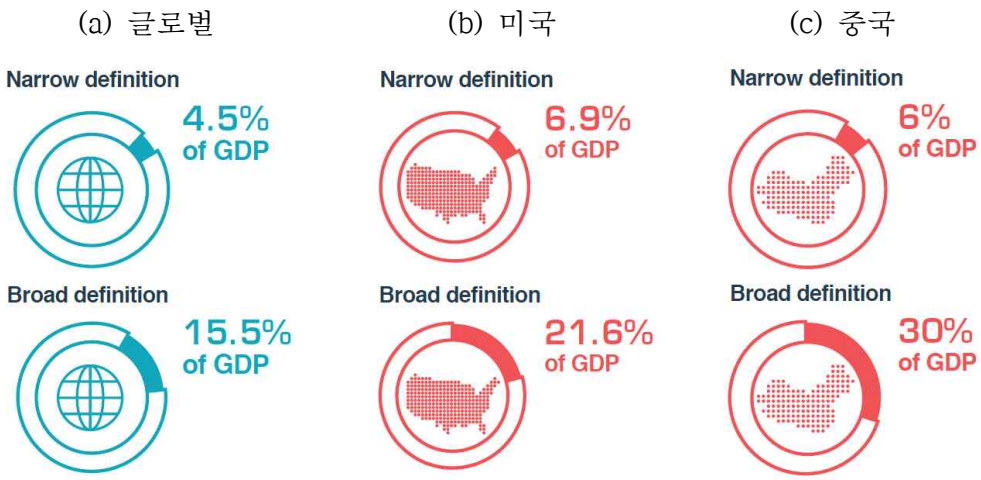
자료: KISDI ITSTAT

이와 달리 ICT산업뿐만 아니라 전통적인 산업의 디지털화도 빠르게 진행되고 있는 추세를 반영하여 미국, 캐나다, 호주, 중국, 영국 등 일부 국가에서는 디지털산업을 좀 더 폭넓게 인식하려는 작업을 진행하고 있다. 그럼에도 불구하고 경제 전반의 디지털화가 이전과 다른 양상으로 전개되면서 디지털과 비디지털 영역이 모호해지고, 디지털산업을 폭넓게 인식하고자 할수록 디지털산업을 명확하게 획정하는 작업은 더 어려워지고 있다. 이는 경제협력개발기구(Organization for Economic Cooperation and Development, OECD) 주도 하에 국제적으로 논의되는 디지털경제 측정 방법론을 따르는 미국, 캐나다, 호주에서조차도 디지털산업이 달리 정의되고 분류되고 있다는 사실로 입증된다.

미국, 캐나다와 호주는 공통되게 사전적으로 디지털산업 범위를 정의하였다(BEA, 2020; Statistics Canada, 2019; ABS, 2019). 이를 OECE(2020)가 제시한 디지털경제 정의에 따라 일반화하면 디지털 지원 인프라, 전자적으로 주문되는 재화와 서비스, 전자적으로 배송되는 재화와 서비스로 구분될 수 있다. 그러나 이와 같은 디지털산업 분류는 전통적인 산업의 디지털화를 제대로 포착하지 못하는 한계를 지닌다. 이는 미국의 경우 BEA(2018)과 BEA(2002)에서 디지털산업 분류와 디지털경제 측정 결과가 다르고, 중국의 경우 디지털산업 범위가 다른 국가보다 넓게 정의되고 디지털경제가 국내총생산(GDP)에서 차지하는 비중도 확연히 높게 측정되는 이유이기도 하다.

예를 들어, UNCTAD(2019)에 따르면 <그림 5-7>에서 살펴볼 수 있듯이 2017년 기준 전 세계의 GDP에서 디지털경제가 차지하는 비중은 디지털산업 범위를 협소하게 정의할 경우 4.5%이나 폭넓게 정의할 경우 15.5%로 추정된다. 미국은 6.9%와 21.6%로, 중국은 6.0%와 30.0%로 추정된다.

<그림 4-7> 디지털경제 정의에 따른 디지털경제 규모 차이



자료: UNCTAD(2019)

또한 전통적인 산업의 디지털화가 빠르게 진행되고 있는 반면, 각 산업과 이에 속한 기업의 디지털과 비디지털 구성요소를 구분할 수 있는 데이터가 축적되어 있지 않았기 때문에 각 국가는 디지털경제를 정확하게 측정할 목적으로 디지털산업을 소극적으로 정의할 유인을 갖는다. 결과적으로 디지털경제가 과소하게 측정될 수 있고 디지털 정책 수립과 관리에 있어 잘못된 의사결정을 유도할 수 있다.

디지털산업을 분류하는 방법도 디지털경제 측정에 상당한 영향을 미칠 수 있다. 중국이 CAICT(2017a)의 방법론을 따라 G20 회원국의 디지털경제 규모를 측정한 결과를 살펴보면 OECD의 방법론을 따른 미국, 캐나다, 호주가 측정한 디지털경제 규모와 확연한 차이를 보인다(CAICT(2017b). 예를 들면, 미국은 BEA(2020)에서 2016년 국내총생산(GDP) 대비 디지털경제의 비중을 8.9%로 보고하였으나 CAICT(2017b)는 58.3%로, 캐나다는 Statistics Canada(2019)에서 2017년 비중을 5.5%로 보고하였으나 CAICT(2017b)는 약 21%로, 호주는 ABS(2019)에서 2016년 비중을 약 5.8%로 보고했으나, CAICT(2017b)는 약 12%로 보고하였다.

디지털경제 규모를 측정하는 이유는 국내적으로는 디지털정책 수립과 관리를 등을 위해, 국제적으로는 디지털경제 규모의 비교에 활용하기 위해서다. 그렇다면 디지털산업을 분류하고 디지털경제를 측정하는 방법은 OECD 등과 같이 국제적으로 논의되는 방법론을 따를 필요가 있다. 그러나 디지털경제의 국제적 비교 활용성만을 고려하여 디지털산업을 소극적으로 분류할 이유는 없다. 국내의 정책적 수요에 맞게 디지털경제 규모를 정확하게 측정하려면 전통적인 산업의 디지털화까지 적극적으로 디지털산업 범위를 결정할 필요가 있다.

제3절 AI・빅데이터를 활용한 디지털산업 분류 방안

1. 디지털산업 분류 방법론

본 연구는 기본적으로 OECD(2020)가 제시하고, 미국, 캐나다, 호주가 채택한 방법론을 따르기로 한다. 다만 우리나라의 경우 전통적인 산업의 디지털화가 빠르게 일어나고 있는 점을 고려하여 BEA(2020), Statistics Canada(2019) 및 ABS(2019)에서와 같이 사전적으로 디지털산업 범위를 정의하지 않고, GI and NIESR(2013)에서와 같이 AI・빅데이터 기술을 활용해 각 산업의 디지털전환척도를 추정하고 일정 수준 이상을 초과하는 경우 디지털산업으로 식별하는 방식으로 디지털산업 범위를 결정하기로 한다.

이를 위해 먼저 생산과 거래 활동 관점에서 OECD(2020)가 제시한 디지털경제 정의를 모두 사용하고자 한다. 이를 통해 디지털산업을 식별하는 데 용이성과 정확성을 제고할 수 있다. 생산 활동 관점에서 제시된 OECD(2020)의 핵심 정의(core measure)에 따라 디지털 재화나 서비스를 생산하거나 공급하는 산업을 디지털산업으로 쉽게 식별할 수 있기 때문이다. 또한 디지털 투입물의 투입 정도를 고려하지 않아도 거래 활동 관점에서 전자적으로 주문되거나 거래되는 재화와 서비스를 생산하거나 공급하는 산업까지 디지털산업으로 식별할 수 있기 때문이다.

다음으로 미국, 캐나다, 호주와 같이 공급사용표(SUTs)를 활용해 디지털경제 정의를 충족하는 재화와 서비스를 식별한다.[15] 여기까지는 OECD가 제시하고, 미국, 캐나다, 호주가 채택한 방법론과 크게 다르지 않다. 그러나 본 연구에서는 전통적인 산업의 디지털화가 빠르게 진행되고 있는 점을 감안하여 다음과 같은 방법으로 디지털경제 정의를 충족하는 재화와 서비스를 식별하고자 한다.

먼저 「주식회사등의외부감사에관한법률」에 따라 외부감사 대상에 해당되는 기업의 경영공시 데이터를 AI・빅데이터 기술로 분석해 각 기업의 디지털전환척도를 추정해 생산 활동 관점에서 디지털 투입물의 투입 정도를 간접적으로 측정하고, 그 값이 일정 수준 이상을 초과하는 기업이 생산하거나 공급하는 재화나 서비스를 식별한다. 다음으로 사전적인 제한 없이 거래 활동 관점에서 각종 전자상거래 품목 데이터, 전자금융거

15) 한국은행 경제통계시스템에서 제공하는 산업연관분석 통계에 따르면 2015년 기준 공급사용표의 재화(상품)분류는 총 381개로 구성되어 있고, 산업분류표의 산업분류는 총 278개로 구성되어 있다.

래 데이터 등을 활용하여 전자적으로 주문되거나 거래되는 재화와 서비스를 식별한다. 마지막으로 두 방법으로 식별된 재화와 서비스를 디지털경제 정의를 충족하는 재화와 서비스로 식별한다.

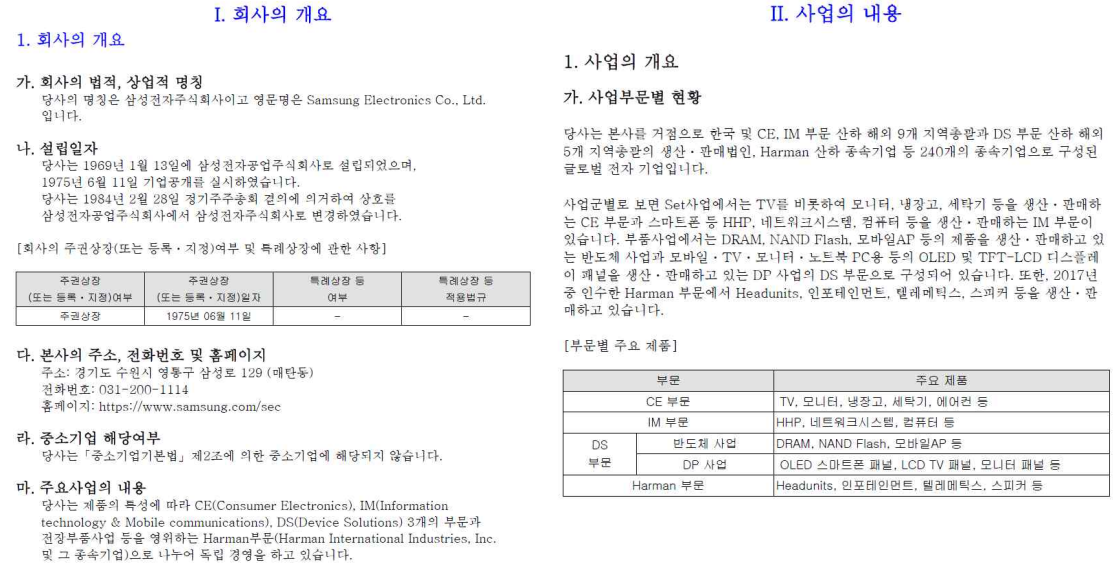
본 연구가 채택한 방법론은 미국, 캐나다, 호주의 방법론과 비교할 때 두 가지 장점이 있다. 첫째, 전자적으로 주문되지 않거나 거래되지 않더라도 생산 활동 관점에서 디지털 투입물의 투입 정도를 고려함으로써 디지털경제 정의를 충족하는 재화와 서비스로 식별할 수 있다. 예를 들면, 스마트 팩토리(smart factory)와 스마트 파밍(smart farming) 기술을 활용해 생산되는 재화를 식별할 수 있다. 둘째, 생산 활동 과정에서 디지털 투입물이 거의 없더라도 거래 활동 관점에서 전자적 주문 또는 배송 여부를 고려함으로써 디지털경제 정의를 충족하는 재화와 서비스로 폭넓게 식별할 수 있다. 예를 들면, 중고차와 같이 플랫폼을 통해 거래되는 재화를 식별할 수 있다.

마지막으로 이렇게 식별된 재화와 서비스를 기준으로 디지털산업을 식별하고 디지털산업 범위를 결정한다. 디지털산업에 따라서는 완전한 디지털 재화 또는 서비스를 제공하는 경우도 있으나, 디지털산업에 속하는 전통적인 산업의 경우 디지털화 정도에 따라 부분적 디지털 재화 또는 서비스를 제공할 가능성이 크다. 그러나 본 연구는 디지털산업 범위를 결정하는 것을 목표로 하고 있기 때문에 디지털산업을 식별하는 데 있어 정확한 디지털전환비율을 산출하거나 디지털경제 규모를 측정할 수 있는 가능성을 고려하지 않기로 한다.

2. AI・빅데이터 활용 데이터

본 연구에서 사용하는 데이터는 크게 세 가지로 구성된다. 우선 AI・빅데이터 기술을 활용해 전통적인 산업에 속하는 기업의 디지털전환척도를 추정하기 위해 금융감독원에서 운영하고 있는 전자공시시스템(DART)에 게시된 외부감사대상 기업의 경영공시 보고서를 사용할 예정이다. 경영공시 보고서에는 기업의 기술적 요소를 분석할 수 있는 비정형 데이터(unstructured data)가 포함되어 있다. 특히 사업보고서에는 기업이 생산하거나 공급하는 재화나 서비스를 구체적으로 파악할 수 있는 회사의 개요와 사업의 내용이 자세히 기술되어 있다. 예를 들어, <그림 5-8>에서 살펴볼 수 있듯이 삼성전자의 주요 제품과 기술 등을 파악할 수 있는 내용이 포함되어 있다.

<그림 4-8> 삼성전자의 2019년도 사업보고서 일부



I. 회사의 개요

1. 회사의 개요

가. 회사의 법적, 상업적 명칭
당사의 명칭은 삼성전자주식회사이고 영문명은 Samsung Electronics Co., Ltd. 입니다.

나. 설립일자
당사는 1969년 1월 13일에 삼성전자공업주식회사로 설립되었으며, 1975년 6월 11일 기업공개를 실시하였습니다.
당사는 1984년 2월 28일 정기주주총회 결의에 의거하여 상호를 삼성전자공업주식회사에서 삼성전자주식회사로 변경하였습니다.

[회사의 주권상장(또는 등록·지정)여부 및 특례상장에 관한 사항]

주권상장 (또는 등록·지정)여부	주권상장 (또는 등록·지정)일자	특례상장 등 여부	특례상장 등 적용법규
주권상장	1975년 06월 11일	-	-

다. 본사의 주소, 전화번호 및 홈페이지
주소: 경기도 수원시 영통구 삼성로 129 (매탄동)
전화번호: 031-200-1114
홈페이지: https://www.samsung.com/sec

라. 중소기업 해당여부
당사는「중소기업기본법」제2조에 의한 중소기업에 해당되지 않습니다.

마. 주요사업의 내용
당사는 제품의 특성에 따라 CE(Consumer Electronics), IM(Information technology & Mobile communications), DS(Device Solutions) 3개의 부문과 전장부품사업 등을 영위하는 Harman부문(Harman International Industries, Inc. 및 그 종속기업)으로 나누어 독립 경영을 하고 있습니다.

II. 사업의 내용

1. 사업의 개요

가. 사업부문별 현황

당사는 본사를 거점으로 한국 및 CE, IM 부문 산하 해외 9개 지역총괄과 DS 부문 산하 해외 5개 지역총괄의 생산·판매법인, Harman 산하 종속기업 등 240개의 종속기업으로 구성된 글로벌 전자 기업입니다.

사업군별로 보면 Set사업에서는 TV를 비롯하여 모니터, 냉장고, 세탁기 등을 생산·판매하는 CE 부문과 스마트폰 등 HHP, 네트워크시스템, 컴퓨터 등을 생산·판매하는 IM 부문이 있습니다. 부품사업에서는 DRAM, NAND Flash, 모바일AP 등의 제품을 생산·판매하고 있는 반도체 사업과 모바일·TV·모니터·노트북 PC용 등의 OLED 및 TFT-LCD 디스플레이 패널을 생산·판매하고 있는 DP 사업의 DS 부문으로 구성되어 있습니다. 또한, 2017년 중 인수한 Harman 부문에서 Headunits, 인포테인먼트, 텔레메틱스, 스피커 등을 생산·판매하고 있습니다.

[부문별 주요 제품]

부문		주요 제품
CE 부문		TV, 모니터, 냉장고, 세탁기, 에어컨 등
IM 부문		HHP, 네트워크시스템, 컴퓨터 등
DS 부문	반도체 사업	DRAM, NAND Flash, 모바일AP 등
	DP 사업	OLED 스마트폰 패널, LCD TV 패널, 모니터 패널 등
Harman 부문		Headunits, 인포테인먼트, 텔레메틱스, 스피커 등

자료: 금융감독원 전자공시스템

두 번째로, 전자상거래 품목, 전자금융거래 등 전자적으로 주문되거나 거래되는 재화나 서비스를 파악할 수 있는 데이터를 사용할 수 있다. 예를 들면, 통계청이 매월 발행하는 온라인쇼핑 동향 보고서를 활용할 수 있다(통계청, 2020. 12. 3). 이 보고서에는 <표 5-4>에서 살펴볼 수 있듯이 숙박, 비행기, 택시 예약, 배달 서비스 등 전자적으로 주문되는 서비스를 포함해 23개 상품군에 대한 전자상거래 현황이 담겨져 있다. 또한 한국은행이 매년 발행하는 금융정보화 추진 현황 보고서를 활용할 수 있다(한국은행, 2020). 이 보고서에는 전 금융권역의 디지털 투입물의 투입 정보와 디지털 금융서비스 현황을 파악할 수 있는 정보가 포함되어 있다. 그 외에도 플랫폼을 통해 전자적으로 주문되거나 배송되는 재화와 서비스를 식별하기 위해 다양한 데이터를 수집하여 활용할 계획이다.

마지막으로 개인사업자 비중이 높은 국내 산업구조의 특성을 고려하여 중소기업벤처부가 매년 발행하는 소상공인 실태조사 등과 같이 소상공인 관련 데이터를 사용할 수 있다(중소기업벤처부 · 통계청, 2019). 특히 중소기업벤처부는 2020년 9월에 소상공인 성장 · 혁신을 위한 디지털 전통시장, 스마트 상점, 스마트 공방 등 디지털 전환 지원 방안을 발표한 바 있다(중소기업벤처부, 2020. 9. 1). 이 점을 감안할 때 전자공시스템에 게시된 주식회사 중심의 기업 경영공시 데이터로 포착되지 않은 소상공인의 디지

털화도 고려되어야 한다.

<표 4-4> 중국의 전통적인 산업의 디지털경제 비중

상품 분류	조사 범위
컴퓨터 및 주변기기	PC, 노트북, 프린터, 스피커, CD형태 등 유형의 소프트웨어
가전·전자·통신기기	TV, 냉장고, 세탁기, 디지털카메라, 휴대폰 등
서적	각종 도서 (e-Book은 콘텐츠에 해당하여 조사에서 제외)
사무·문구	사무용품, 문구류, 다이어리/앨범, 종이류/복사지, 필기구 등
의복	의복류 (남성복, 여성복, 스포츠웨어 등)
신발	신발 (구두, 운동화, 샌들, 실내화 등)
가방	가방 (핸드백, 가방, 여행용 등)
패션용품 및 액세서리	모자, 장갑, 스카프, 시계, 금반지, 각종 액세서리 등
스포츠·레저용품	운동용품, 레저용품, 등산화, 등산배낭 등
화장품	화장품, 향수, 화장관련 소품 등
아동·유아용품	기저귀, 유모차, 그네, 아기침대, 보행기, 카시트, 인형, 완구 등
음·식료품	공산품류(커피, 차, 음료, 생수, 설탕, 식용유 등), 김치, 장류 및 장아찌류 등
농축수산물	곡물, 육류, 어류, 채소, 과실, 신선식품류 등
생활용품	주방용품, 침구, 비누, 샴푸, 세제, 화장지, 꽃, 화분 등
자동차 및 자동차용품	자동차, 오토바이, 튜닝/선팅용품, 내비게이션, 블랙박스, 엔진오일, 워셔액 등 자동차 관련용품
가구	가구 (장농, 화장대, 신발장, 책상, 의자 등)
애완용품	애완용품 (사료, 장난감, 장신구 등)
여행 및 교통서비스	항공권, 교통티켓(버스, 기차), 렌터카, 숙박시설 등
문화 및 레저서비스	영화, 공연 등의 예약서비스
e쿠폰서비스	해당금액에 상응하는 서비스나 상품을 제공하는 바코드형식의 상품권
음식서비스	온라인 주문 후 조리되어 배달되는 음식 (피자, 치킨 등 배달서비스)
기타서비스	인화 등 주문제작, 이사, 청소 등 용역서비스, 각종 렌탈서비스
기타	문화상품권, 의료기구(안마의자제외), 골동품, 종교용품, 성인용품, 음반·비디오·악기 등

주석: 콘텐츠(음원, 이모티콘, 전자책 등)에 해당하는 무형의 상품은 제외

자료: 통계청(2020. 12. 3)

제5장 [부록 2] 전자공시 데이터 기반 디지털전환 지수 산출 방안

본 장은 해당 정책연구의 일환으로 검토 된 AI·빅데이터 활용 유스케이스 발굴 목적으로 제안된 연구 아이디어로서, 금융감독원 전자공시시스템(DART)의 전자공시 빅데이터를 활용한 신규 통계 지표 산출 방안을 소개하였다.

제1절 배경 및 필요성

1. ICT 신기술의 부상과 ICT시장·산업의 변동성과 불확실성 확대

최근 ICT신기술에 의한 혁신이 4차 산업혁명을 주도하면서 국내 ICT 산업의 지속적인 성장을 위해서는 인공지능기술, 빅데이터, 클라우드, 사물인터넷, 소프트웨어 및 콘텐츠 등 지능정보기술에 기반 한 산업혁신이 중요해지고 있다. 무엇보다 2020년 신종 코로나바이러스 확산으로 인한 전세계적 혼란이 ICT산업의 패러다임 변화를 가속화 시키고 있는 현시점에서 시의성 있는 ICT정책 수립이 중요하다. 이를 위해서는 시시각각 변화하는 다양한 데이터로부터 ICT산업·시장현황을 종합적이고, 객관적으로 진단하고 예측하는 과학적 의사결정과정이 필요하다. 하지만 ICT시장의 불확실성을 실시간으로 포착하여 정책수립을 지원하는 빅데이터기반 정책지원체계는 상당히 미흡한 수준이다.

2. 우리나라 ICT 산업의 특수성

국내 ICT산업에 한정하여 생각해보면, 우리나라 ICT산업은 국내 시장포화, 중국의 추격, 주력품목 시장의 높은 변동성으로 인해 최근 ICT 산업의 성장세, 국민경제성장 기여 모두 하락 추세를 보이고 있다. 향후 4차 산업혁명 확산을 위해서는 신성장산업의 출현, 서비스업에서의 ICT 활용 증가 등이 주력 제조업종의 설비투자 둔화를 상쇄할 수 있어야 한다. 이는 최근 활발히 논의되고 있는 전통제조업의 디지털전환 이슈와 맞

닿아 있는 부분이다. 특히 우리나라 ICT 산업 구조는 HW 중심의 특정품목(반도체, 디스플레이, 휴대폰)중심으로 중국, 미국 등 일부 국가에 수출이 편중된 구조적 문제가 존재하고 있어서 중장기적인 산업 성장 동력이 우려되는 상황이다. 이러한 상황에서 ICT 산업 현황 및 시의적절한 전망요인 분석은 단기·중기 ICT 산업정책 수립 및 관련 기업들의 전략수립에 매우 중요한 요인이라 할 수 있다.

3. 빅데이터를 활용한 미래예견적 정책수립의 중요성 증대

이처럼 국내 ICT산업을 둘러싼 불확실성이 급증하는 상황에서 전통적인 정형자료 기반 계량경제(Econometric)모형에 기초한 산업분석 및 전망은 자료 구축에 많은 시간이 소요되며 이용 가능한 자료의 형태가 제한적인 관계로 불확실성이 높은 오늘날의 기업환경에서 실효성 있는 정책을 제안하는 것을 어렵게 만들고 있다. 더욱이 기존의 전망 및 예측모형은 방대한 분량의 실시간 비정형자료를 적시에 활용하지 못함으로써 시장 불확실성의 실시간 포착을 통한 위기 대응과 기회요인 포착에 한계가 있다는 것을 의미한다. 이와 같은 상황에서 최신 데이터 사이언스 방법론을 활용하여 공공기관·국책연구기관 등에 旣축적된 정형 데이터와 논문, 보고서, 뉴스 등 비정형 데이터를 연계·분석하여 미래를 예측하고 상황별 시뮬레이션에 기초한 정책수립이 무엇보다 중요하다고 할 수 있다.

제2절 선행연구 및 DART 자료의 특징

1. 경제사회 빅데이터 분석·예측, 선행연구 현황

경제예측·분석모형에 활용된 빅데이터의 특징은 크게 두 가지인데 정형자료 대비 1) 주기가 짧고(high-frequency), 2) 분화된(disaggregated) 데이터라는 특징을 갖는다. 빅데이터 중 비정형데이터(혹은 텍스트데이터)를 활용한 연구들은 주로 온라인 데이터를 활용하여 주요 토픽을 추출하고 추출된 토픽의 의미를 분석하거나 실물경제 변수에 선행하는 토픽을 활용하여 선행지수를 구축한 연구들이 있다. 한편, 정형데이터를 활용한 연구들은 차원의 저주(curse of dimensionality)를 극복하기 위해 동적요인모형(Dynamic Factor Model)을 통해 주요인을 추출하여 현재를 추정(Nowcasting)함으로써 시의성을 확보하는 연구이다. 주기가 다른 자료들을 통합하여 예측모형을 보정하는

MIDAS(mixed data sampling)기반 연구들도 빅데이터 시대 도래와 함께 주목받는 연구 모형이다. 최근 들어 각국 중앙은행 중심으로 빅데이터를 활용한 다양한 거시경제 예측모형 연구가 시도되고 있고 동시에 비정형 데이터로부터 유의미한 정보를 추출하여 이를 계량모형에 반영하려는 연구 또한 계속 되고 있다

<표5-1> 빅데이터를 활용한 경제사회 예측 선행연구 분석

구 분		선행연구 분석 기준		
		연구목적	연구방법	주요연구내용
주요선행연구	1	○**과제명 : 빅데이터를 활용한 환경분야 정책수요 분석** - 연구자(년도): 이미숙 외 (2014) - 소셜 빅데이터를 활용한 환경 정책 수요분석 및 탐색	- 환경분야 전반과 12개 세부 환경분야에 대한 소셜 빅데이터 분석(빈도분석, 키워드 분석, 감성분석)을 수행 - 2012년 11월부터 2014년 4월 까지 1년 6개월 동안 뉴스, 블로그, 트위터 채널에서 한국어로 발현된 7억 5538만건을 대상으로 키워드 선정과 문서 추출 과정을 거쳐 약 547만 건을 분석	- 환경분야 전반에 걸쳐 트위터 채널의 빈도가 높게 나타나 수요자 채널로서 활용가능성이 높음을 발견 -감성분석 결과 환경분야 전반에 대해 부정 감성이 긍정 감성보다 우세하게 나타나 국민들의 환경에 대한 인식이 부정적임을 확인
	2	○**연구명 : 소셜 데이터 기반 실시간 식자재 물가 예측 모형** - 연구자(년도): 김재우 외 (2017) - 소셜 빅데이터를 활용한 실시간(Nowcasting) 물가 예측 모형 개발 및 인도네시아 적용	- 15개월간 전체 78,518개의 트윗을 제시한 28,800개의 계정을 분석 - 트윗 내용의 반복적 유사도(linguistic similarity)와 게시 빈도를 통해 광고 봇 계정을 제거 - 동질 마르코프 가정(Homogeneous Markov process)을 적용하여 모델 상정하고 성능 개선	- 트위터와 같은 온라인 빅데이터를 수집 및 분석하여 주요 소비재 시장물가를 실시간으로 단기 예측하는 알고리즘을 개발하고 실제 15개월간 인도네시아를 대상으로 주요 식자재의 일별 물가 추이를 예측
	3	○**과제명 : ICT정책에서 빅데이터 분석의 활용방안 연구** - 연구자(년도): 김경훈 외 (2017) - 사례분석을 통해 정책연구에서 빅데이터 도입방안을 연구하고 부처 보도자료의 텍스트를 분석하여 텍스트마이닝을 통한 정책이슈 제언 가능성을 점검	- 2013년 2월부터 2017년 4월까지 미래창조과학부와 방송통신위원회의 보도자료를 취합하여 토픽 모델링 수행 - 2013년 1월부터 2016년 5월까지 소셜빅데이터를 총 257,515건의 텍스트 문서를 크롤링 하고 이중 ICT와 연관있는 63,102건을 분석 대상으로 삼아 주제분석, 감성분석, 요인분석 수행	- ICT 보도자료 텍스트 마이닝 및 토픽 모델링 - ICT 순기능/역기능 소셜 빅데이터 분석과 영향요인 식별 - ICT 경제정책불확실성지수 개발 방법론 연구
	4	○**연구명: Macroeconomic Nowcasting and Forecasting with Big Data** - 연구자(년도): Bok et al. (2017) - 가용한 월간 자료를 바탕으로 분기별로 업데이트 되는 데이터를 실시간 추정(Nowcasting)	- 24개의 매월 발표되는 공인데이터 기반, 동적요인모형(Dynamic Factor Model)를 활용하여 4개 분야 주요요인을 추출하고 이를 이용하여 GDP 성장률을 추정	- 빅데이터(24개 월별데이터)로부터 동적요인모형을 통해 글로벌, 소프트(경제주체감성지수), 실물, 노동부문 요인을 추출 - 뉴욕 연방준비은행은 4개의 주성분을 활용하여 일(Day) 단위로 GDP성장율 전망치를 수정하고 매주 발표
	5	○**과제명 : 빅데이터 기반 GDP 성장률 예측 모형 개발**	- 거시경제 시계열 자료(82종)를 가중평균하여 설명력이 강한 3개의	- 82종 거시 시계열 데이터로부터 주성분분석을 통해 3개의 주요

구 분		선행연구 분석 기준		
		연구목적	연구방법	주요연구내용
		- 연구자(년도): 금융감독원(2018) - 발표주기가 상대적으로 짧은 데이터를 활용하여 GDP성장률 추정	요인을 추출 - 칼만필터링을 통해 주기가 긴 데이터 추정하여 균형(balanced) 데이터 구축	인을 추출 - VAR 모형을 통해 GDP와 세 개 주요인의 동적관계를 추정 - 업데이트된 시계열자료와 VAR 추정결과를 종합하여 GDP성장률 전망
	6	○**연구명: Using large data sets to forecast sectoral employment** - 연구자(년도): Gupta et al. (2014) - 규모가 큰 벡터자기회귀모형에서 파라미터의 Bayesian shrinkage를 활용한 전망 방법론	- 8개의 산업별 실업률과 143개의 월별 데이터를 활용하여 주요인을 추출 - 산업별 실업률 전망을 위해 14개의 전망모형을 구축하고 성능을 비교	- AR, VAR, VEC, 그리고 각 모형의 베이지안 버전과 요인확장형(factor-augmented) 버전의 모형을 활용하여 8개 산업별 실업률 추정 - 각 모형의 In-sample fitting과 out-of-sample forecasting 성능을 비교한 결과 모든 모형의 평균이 하나의 최적모형 보다 성능이 좋은 것으로 나타남
	7	○**연구명: 전력산업 미래전망을 위한 빅데이터 활용 방안 연구(에너지경제연구원 기초과제)** - 연구자(년도): 박찬국 외. (2017) - 전력산업 전망을 위한 텍스트 데이터를 활용한 미래신호탐색 모형 제안	- 텍스트 데이터를 활용하여 사회적 관심도, 관심도의 변화, 전력과의 연관성을 중심으로 한 미래신호 탐색 모형 - 기계학습의 일종인 나이브베이즈를 활용하여 특정 기간별 약신호에서 강신호로 전환될 확률을 예측	- 2010년부터 2016년까지 영문 언론기사에서 4개의 키워드('electric industry', 'electricity industry', 'power industry', 'energy industry')를 중심으로 수집 - 과거 2년 이상의 데이터를 활용하여 미래 2년을 전망한 결과, 강신호 전환 예측 정확도가 80%를 넘어섬 - 그러나, 단기 예측정확도는 상대적으로 많이 떨어짐

2. DART 자료의 특징

기업전자공시스템(DART)에서는 코스피시장이나 코스닥시장에 상장된 주식회사부터 비상장주식에 이르기까지 국내 거의 모든 기업들이 자기회사의 경영상태에 대한 공시를 수행한다. 때문에 DART에서는 한국에 있는 외감이상의 거의 모든 기업들의 재무정보를 확인할 수 있고 동시에 정기공시를 통해 분기·반기·연도 등 다양한 주기의 사업보고서 또한 확인할 수 있다.

빅데이터 관점에서 살펴보면 기업단위 분석 나아가 산업단위 분석에 필수적인 재무정보를 확인할 수 있는 동시에 기업의 주요사업과 관련된 사업부문별 현황정보, 영업의 개황 등을 비정형 텍스트정보로 확인할 수 있다. 즉 DART는 단일 자료원에서 기업의 주요한 의사결정과 관련된 정형자료와 비정형자료가 동시에 제시되기 때문에 이

들을 수집하여 효과적으로 활용할 경우 정형·비정형 통합 모형을 구축할 수 있는 장점이 있다.

제3절 문제인식과 디지털전환 지수 모형제안

1. 표본조사 기반 ICT기업경기조사의 한계

기업경기실사지수(BSI)는 월별/분기별로 보고되는 특성으로 인하여 무역 분쟁과 같은 경제 이벤트나 이슈를 시의적절하게 반영하지 못하는 문제 등이 있다. 또한 ICT기업경기조사(BSI)는 표본조사를 통하여 작성되고 있기 때문에 조사항목이 실제 경기 인식을 반영하지 못하는 문제가 있다.

<그림5-1> 정보통신제조업 기업경기실사지수

자료: 통계청 산업별 생산지수(KOSIS)

이러한 상황에서 빅데이터 분석으로부터 현재의 표본조사 기반ICT경기실사지수를 대체하는 정보(지수)를 포착할 수 있다면 현재 ICT산업의 경제 예측 시스템의 한계점을 효과적으로 개선할 수 있으리라 본다. 특히나 기업환경이 시시각각 변화하는 상황에서 단주기(high frequency)의 비정형자료로 부터 소위 우리나라의 “디지털전환정도”를 포착할 수 있는 지수를 구축하고 이를 ICT수출과 같은 경제예측모형에 설명변수로 도입하여 활용할 수 있다면, 그 자체로서 의미있는 변수를 획득할 수 있고 동시에 ICT예측모형의 설명력을 제고할 수 있다.

물론 기존의 정형화된 거시·금융 자료는 여전히 ICT 경제예측의 주요 설명변수로서 역할이 있다. 다만, 기존 데이터에 빅데이터 기법 등을 통해 수집 가공된 비정형적인 데이터를 활용하여, 기계학습을 바탕으로 모형을 수립하고 표본조사 기반의 ICT산업의 경기 실사지수를 대체하는 지표들을 생성할 경우 ICT경기를 나타내는 지수로서 나아가 다양한 실물경제를 적기에 예측하는 설명변수로서 더 나은 적격성을 띌 수 있다. 물론 기존 정형 데이터와 비정형 데이터를 통해 생성된 지수를 교차검증(cross validation) 등의 방법을 이용하여 모형의 예측력을 평가하고 최적의 모형을 찾는 과정이 동반되어야 한다. 이에, DART의 비정형텍스트 자료를 활용하여 우리나라의 디지털전환정도를 포착할 수 있는 지수구축 방안에 대해 소개하고자 한다.

2. 디지털전환 지수 구축 제안 모형

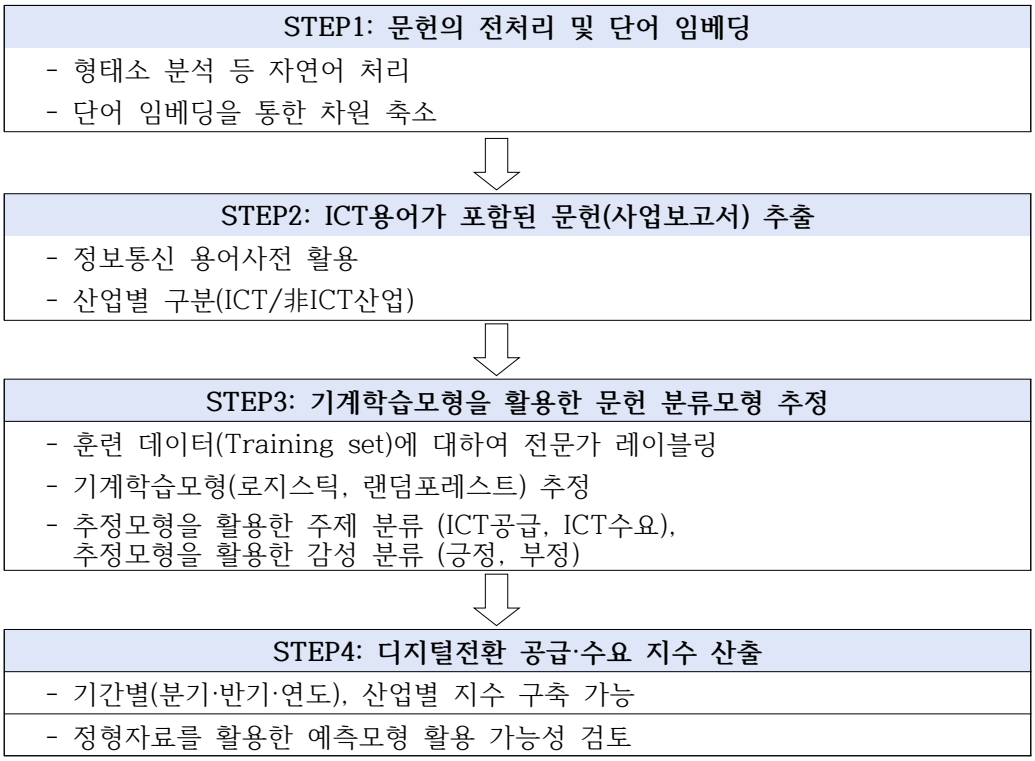
본 모형은 주요사업의 확장 또는 사업의 축소에 관계 된 기업 주요 의사결정을 다루고 있는 DART의 사업보고서(반기,분기보고서 포함) 내 영업의 개황 등 현재의 사업내용과 관련된 텍스트 자료를 활용하여 디지털전환 지수를 구축하는 것을 목적으로 한다.

기본적으로 산업의 구분을 두지 않고 전체 사업보고서를 취합하거나 산업별 사업보고서를 한데 모아 ICT 키워드의 증감 정도를 지수화 하여 디지털 전환을 가늠해 볼 수도 있겠으나, 향후 기업성과 및 거시경제 예측 등 지수의 활용도를 고려하여 디지털 전환을 공급과 수요측면을 구분하여 구축하는 것을 제안하고자 한다. 디지털전환 공급지수(DT-SI: Digital Transformation Supply Index)는 디지털 전환을 기술적으로 가능케 하는 클라우드, AI, 5G 등의 ICT신기술과 관련한 ICT기업들의 역량을 측정하는 지수이고, 디지털전환 수요지수((DT-DI: Digital Transformation Demand Index)는 기업 내 디지털 전환을 시도하는 非ICT 기업들의 준비정도를 측정하는 지수이다.

지수구축 과정을 간단히 요약하면, 우선 분석에 활용되는 자료가 텍스트자료인 만큼 기초적인 자연어 처리과정(Natural Lnaguage Processing)을 통해 문서를 정제하고 형태소분석을 진행 한 뒤, word2vec 이나 doc2vec을 활용하여 문헌단위로 문서를 임베딩하는 과정이 필요하다. 이후 확보한 문헌 중 훈련데이터(Training set) 목적으로 활용하고자 하는 일부 문헌에 대해 전문가 자문을 통해 ICT 수요·공급 여부, 긍정·부정 여부를 확인할 수 있는 라벨링 과정을 진행한다. 지도학습에 필요한 테스트데이터를 만

드는 과정이다. 라벨링 된 자료를 학습데이터로 하여 로지스틱 모형의 파라미터를 추정하고, 추정모형을 바탕으로 나머지 테스트데이터를 분류 한다. 끝으로 공급(긍정/부정)·수요(긍정/부정)으로 분류된 문헌을 토대로 기간별 산업별 지수를 생성한다.

<표5-2> DART 자료를 활용한 디지털전환 지수 구축 과정



제4절 모형 분석 절차

1. 모형의 분석 단계

① 자료의 수집 및 정제

자료는 DART 사이트를 크롤링 함으로써 확보할 수 있다. 다만, 기업 내 사업보고서에서 공시기간동안 ICT 관련 키워드(정보통신용어사전)를 한번도 포함하고 있지 않은 기업은 향후 분석에서 제외한다. 이후 자연어 처리 과정이 필요하다. 자연어 처리(NLP: Natural Language Process)에서는 우선적으로 데이터 전처리 과정이 선행되어야 한다. 텍스트 데이터 전처리는 용도에 맞게 텍스트를 사전에 처리하는 작업으로, 텍스트 마이닝 결과에 큰 영향을 주는 중요한 요인이다. 본 연구에서는 알파벳, 숫자, 특수문자, 구두점 등을 제거하는 정제 (cleaning) 작업을 한 뒤, 형태소 분석을 하였다. 형태소 분석이란 형태소를 비롯하여 어근, 접두사, 접미사, 품사 (POS, part-of-speech) 등 다양한 언어적 속성의 구조를 파악하는 것이다. 단어의 표기는 같지만, 품사에 따라 단어의 의미가 달라지기 때문에 형태소 분석 과정에서 단어가 어떤 품사로 쓰였는지 구분해 놓는 작업을 하며, 이를 태깅 (part-of-speech tagging)이라고 한다. 한글의 품사 태킹을 위한 토큰화 과정에서는 Python 패키지인 KoNLPy의 Okt 한글 형태소 분석기를 이용할 수 있다.

② 워드 임베딩 (실제로는 doc2vec을 활용한 문헌의 임베딩)

단어 임베딩은 텍스트를 거리 (metric) 공간으로 수치화하는 자연어 처리의 한 방법이다. 대표적인 방식으로는 원-핫 인코딩 (one-hot encoding)이 있다. 원-핫 인코딩은 단어 집합의 크기를 벡터의 차원으로 하며, 표현하고 싶은 단어의 인덱스에 1의 값을 부여하고, 다른 인덱스에는 0을 부여하는 단어의 벡터표현 방식이다. 이렇게 될 경우 분석에서 취급하는 단어 수 만큼의 벡터 차원이 필요하다는 단점이 있다. 최근에는 이와 같은 한계점을 보완한 분산표현이 폭 넓게 활용되고 있다. 분산 표현은 ‘비슷한 위치에 있는 단어들은 비슷한 의미를 갖는다’ 라는 분포 가정하에 만들어진 표현 방법이다. 예를 들어, 강아지라는 단어는 주로 ‘귀엽다’, ‘사랑스럽다’ 와 같은 단어

와 등장하는데, 강아지라는 단어가 분산 표현으로 벡터화되면 이 주변 단어들은 강아지와 유사한 단어로 인식된다. 분산 표현 방식 중 대표적으로 사용되는 것이 word2vec (Mikolov et al., 2013)과 word2vec 방식을 문서로 확장한 doc2vec (Le & Mikolov, 2014)이다. 본 분석에서는 문헌단위로 분석을 해야하기 때문에 문서를 수치화한 문서임베딩 방식인 doc2vec을 활용하는 것을 제안한다.

<표5-3> doc2vec 학습결과 (100차원) 벡터 예시

반도체	-1.3670751	-1.850202	…	-1.1217545
상승	-0.9209298	-2.502876	…	-2.0222008
시장	-0.95774364	-0.03745903	…	0.18656346
분기	-1.8828163	-0.93941045	…	-1.0199012
경쟁력	-0.2441145	-0.82625777	…	-0.4863924
수출	-1.7562301	-1.5138226	…	-2.2640836
한국	-0.25101385	-2.0243618	…	-0.65352184
미국	-0.6530417	-1.3561974	…	0.7473832
업체	0.3464992	0.50071573	…	-1.2836359
제조	0.08519085	1.2477363	…	-0.14513953
부진	0.56372315	0.41104081	…	-2.3206246
이익	-1.6077964	-2.8104615	…	-3.6957588
산업	-1.1442248	-1.9410778	…	-1.3190506
중국	-0.62526476	-2.5239367	…	-0.00837229
일본	-0.6424548	-0.3905851	…	0.9587376

③ 기계학습(로지스틱)을 활용한 분류모형 구축

DART 사업보고서에서 다루고 있는 사업의 내용이 ICT수요 관련인지 ICT 공급관련인지, 나아가 개별 주제(공급·수요)를 긍정적으로 평가하고 있는지 부정적으로 평가하고 있는지 감성여부를 판단하는 분류모형을 구축해야 한다. 분류모형 중에는 일반적으로 가장 많이 사용되는 로지스틱 회귀(logistic regression) 모형을 활용할 수 있고, 이 외에도 랜덤포레스트(Random forest)와 같은 이진분류 모형도 활용할 수 있다.

예를 들어 온라인에서 수집한 반도체 관련 뉴스기사에 대하여 개별 뉴스(문헌)를 하나의 관측치로 간주하고, 해당 뉴스의 주요 내용이 반도체 수요 관련인지, 반도체 공급 관련인지 나아가 수요측·공급측 각각에 대하여 긍정적 의미인지 부정적 의미인지 판단할 수 있다. 온라인 뉴스 문헌이 수집되었다고 할 경우 앞서 설명한 규칙을 따라 다음 표와 같이 분류할 수 있다.

<표5-4> 반도체 관련 뉴스기사 공급·수요 긍정·부정 분류 예시

문서 번호	수요	공급	링크	기사제목
1	1	0	http://www.segye.com/newsView/...	삼성 다시 반도체 절대 강자로… “D램 점유율 50% 육박”
2	1	0	http://www.etnews.com/...	클라우드 요충지는 한국...글로벌 데이터센터 몰려온다
3	0	0	https://news.mt.co.kr/...	“日, 반도체 핵심소재 '액체 불화수소' 韓 수출 지연”
4	0	1	https://www.mk.co.kr/news/business/view/...	삼성, 최고성능 SSD 출시…스토리지 새 역사
5	0	1	https://www.mk.co.kr/news/business/view/...	삼성, 파운드리 `GAA` 기술개발 로드맵 발표
6	0	1	https://www.mk.co.kr/news/realestate/view/...	SK하이닉스 122조 투자…용인에 세계 최대 반도체 생산라인
7	2	0	https://www.yna.co.kr/view/...	中도 '비틀'…통상전쟁에 TV·휴대폰·반도체 모두 '역성장'
8	2	1	http://www.digitaltoday.co.kr/news/...	가트너, “올해 전세계 반도체 매출 9.6%↓"
9	0	2	http://www.mirae-biz.com/news/articleView...	글로벌 메모리 반도체 업체, 장기 불황 대비…감산 러시
10	1	0	http://www.etnews.com/...	삼성, '업계 최초' 0.7㎛ 픽셀 이미지 센서 개발
11	2	0	http://biz.chosun.com/site/data/...	세계 D램 시장 장악한 삼성·SK…가격 하락으로 매출은 계속 줄어...

*** 수요 공급 코딩에서, 1은 긍정적, 2는 부정적 기사임을 의미**

현재 다루고 있는 DART사업보고서 분석의 경우로 생각해보면, 10,000개의 사업보고서가 수집되어 분석에 활용하고자 할 경우, 400개의 문헌을 학습데이터로, 나머지 9,600개의 문헌을 테스트데이터로 활용할 수 있다. 전문가들이 판단하여 공급문헌 200개 수요문헌 200개에 대해 레이블링을 진행한다. 레이블링 된 400개의 문헌을 활용하여 로지스틱 모형의 모수를 추정하고, 나머지 9,600개의 테스트데이터에 대해서 분류를 진행한다. 감성분석에 대해서도 주제분류와 마찬가지로 위의 예시와 같은 분류모형을 구축·추정하고 문헌을 분류할 수 있다.

로지스틱모형 외에도 앞서 언급한 랜덤포레스트 모형, 또는 기계학습에 활용되는 다양한 이진선택 모형을 적용하여 모형을 추정하고 분류기로 활용할 수 있다. 다만 통계모형별로 다음과 같은 정오표를 통해 분류의 정확도를 확인하고 가장 좋은 성능의 분류기를 채택하여 활용하여야 할 것이다.

<표5-5> 기계학습모형 별 주제분류 결과 비교 예시

실제 / 예측	로지스틱 회귀		랜덤 포레스트	
	수요	공급	수요	공급
수요	56	7	62	1
공급	10	11	12	8

이미 레이블링 된 학습데이터를 활용하여, 주제별(공급/수요), 감성별(긍정/부정)분류 기계학습 모형을 추정하고 나면, 이후 임베딩 된 모든 문헌에 대해서는 분류기가 자동으로 판단 할 수 있다.

④ 4단계: 디지털전환 공급·수요 지수 산출

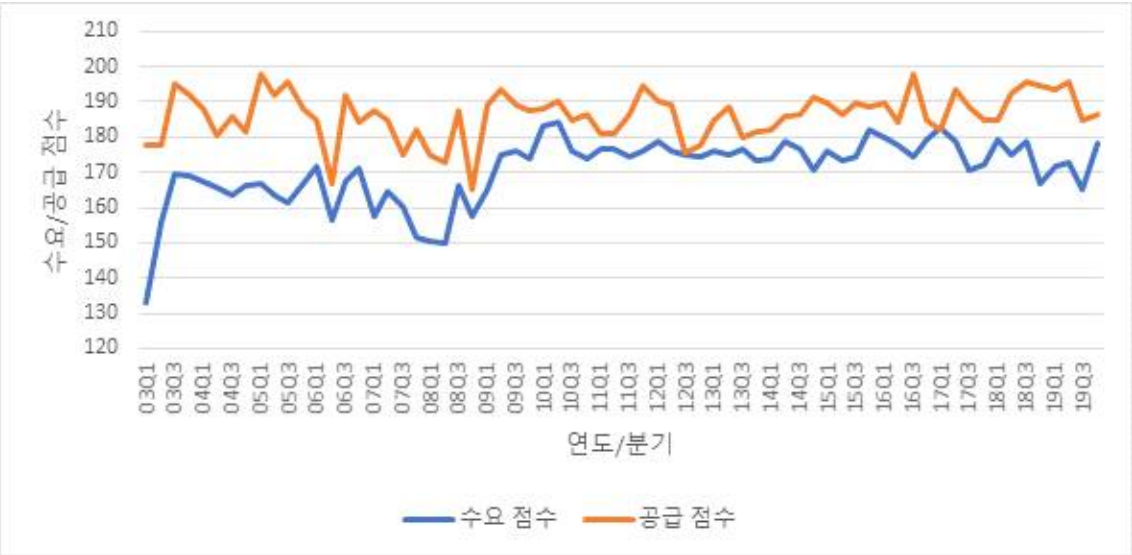
수요, 공급 각각에 대해서 감성분석 후 긍정으로 분류된 문헌과 부정으로 분류된 문헌을 활용하여 각 주제(공급/수요)별로 지수를 구축 할 수 있다. 지수를 산출하는 방법에는 다양한 방식이 있을 수 있으나, 가장 직관적인 방식은 다음과 같다.

$$\frac{\text{긍정문헌 개수} - \text{부정문헌 개수}}{\text{긍정문헌 개수} + \text{긍정문헌 개수}} \times 100 + 100$$

'지수'는 연구자가 설정한 기간에 따라, 정해진 기간별 (분기별/반기별/연도별)로 계산할 수 있다. 지수 점수 공식에 근거하면 긍정인 문헌만 존재한다면 200점이 되고, 부정인 문헌만 존재한다면 0점이 된다. 따라서 위 지수를 통해 해당 분기의 각 주제(공급/수요)의 지수가 0점에서 100점이면 부정적, 100점에서 200점이면 긍정적이라고 판단할 수 있다.

특정한 기간에 대해서 주제별 지수를 산출하면 다음과 같이 추이를 살펴 볼 수 있다.

<그림5-2> 수요·공급 주제별 지수 추이 예시



제5절 디지털전환 지수의 활용

1. 디지털전환 지수의 활용

지수는 그 자체로서 우리나라의 공급 관점, 수요 관점에서 평가한 디지털전환 준비 정도 또는 역량을 가늠할 수 있는 척도의 의미를 지닌다. 하지만 이를 활용하여 실물 경제의 예측모형의 성능을 개선하는데도 활용할 수 있다.

통계학에서 회귀 분석 (Regression Analysis)라 함은 종속변수 y의 변동을 설명변수 x의 변동으로 설명하고자 하는 분석이다. 만약 설명변수의 변동이 적고 종속 변수의 변동이 큰 데이터로 회귀 분석을 수행하고자 한다면, 독립변수의 변동이 상대적으로 큰 데이터보다 적절한 결과를 도출하기 힘들다. 이는 단순 선형 회귀 (Simple Linear Regression)에서 추정량의 분산이 커진다는 점과 일맥상통한다.

종속변수와 독립변수는 그 형태가 여러 가지 있다. 예컨대, $(-\infty,+\infty)$의 값을 가지는 연속형 변수 (continuous variable), 그 중에서도 $(0,1)$의 값을 가지는 변수, (예, 아니오) 값을 가질 수 있는 범주형 변수 등이 있다. 비 회귀 (Ratio Regression)은 이 중에서도 종속변수와 독립변수가 비율의 형태를 취함으로써 $(0,+\infty)$의 값을 가지는 모델로, 이를 Model Equation Form으로 적으면 아래와 같다.

$$E(y_t|x_t) = \beta x_t$$

2. 비추정(ratio estimation) 모형

다음의 비추정 모형을 통해 산업별 재무성과지수 또는 ICT실물지수(예, 정보통신 서비스업 고용자 수)에 대한 기본 예측모형을 제안할 수 있다.

$$\frac{Y_t}{Y_{t-1}} \approx \beta \times \left(\frac{X_{t-1}}{X_{t-2}} \right) \qquad (1)$$

식 (1)에서 Y_t와 X_t는 각각 t시점에서의 정보통신서비스업 고용자 수와 주제(수요/공급) 점수를 의미한다. 그런데 회귀모형에서 $y_t = \frac{Y_t}{Y_{t-1}}$와 $x_t = \frac{X_{t-1}}{X_{t-2}}$으로 종속변수와

독립변수를 각각 정의하기 때문에 x_t는 실제로 t-1시점과 t-2 시점의 정보만을 필요로 한다.

x_t로 표현되는 지수는 앞서 설명하였다시피 기업의 사업보고서를 기업단위 또는 산업단위로 합계하여 텍스트 분석 방법으로 전환한 것으로 0점-200점 사이의 값을 가지도록 설계된 것이다. Y_t를 직접 예측하는 대신에 비(ratio)를 예측하는 것으로 제안한 이유는 일반적인 예측모형 (forecasting model)과 차이가 있기 때문이다. 비(ratio)를 추정하는 형태로 모형을 설계한 이유는 (i) 변수 스케일에 따른 효과를 최대한 제어하면서도, (ii) 시계열적 연속성을 확보할 수 있기 때문이다.

나아가 모형의 성능은 Y_t/Y_{t-1}의 변화를 잘 캡쳐할 수 있는 X_{t-1}/X_{t-2}의 함수 형태를 찾는 것이 중요하다고 할 수 있다 식 (2)와 같이 목적함수를 정의하고 나면 예측력이 더 좋은 함수 g를 찾을 수 있을 것이다. 이때, 적절한 함수 g는 다항함수 (polynomial functions)를 이용하거나 Reproducing Kernel Hilbert Space (RKHS)를 통하여 많이 찾을 수 있다.

$$\frac{Y_t}{Y_{t-1}} \approx \beta g\left(\frac{X_{t-1}}{X_{t-2}}\right) \qquad (2)$$

3. 성과변수(타깃변수)

앞서 구축한 디지털전환 지수를 활용하여 비추정 모형을 제안 할 경우 성과척도는 다양한 모습을 띌 수 있으리라 본다. DART에서 가용 가능한 재무수준의 성과를 예측하는데 활용할 수도 있고, 인력, 생산 수출 등 ICT거시경제지표를 예측하는데 활용할 수도 있다.

<표5-6> 성과변수(타깃변수)로 활용할 수 있는 재무성과 평가척도와 변수

평가척도	변수명	산식
안정성	부채비율(%)	부채총계/자본총계×100
	자기자본비율(%)	자기자본/총자본×100
	차입금자기자본비율(%)	차입금/자기자본×100
수익성	자기자본순이익률(%)	당기순이익/자본총계×100
	매출액영업이익률(%)	영업이익/매출액×100

	매출액순이익률(%)	당기순이익/매출액×100
성장성	매출액증가율(%)	(당기매출액-전기매출액)/전기매출액×100
	영업이익증가율(%)	(당기영업이익-전기영업이익)/전기영업이익×100
	총자산증가율(%)	(당기총자산-전기총자산)/전기총자산×100
유동성	유동비율(%)	유동자산/유동부채×100
	당좌비율(%)	당좌자산/유동부채×100
	순운전자본비율(%)	(유동자산-유동부채)/자산총계×100
활동성	총자산회전율(배)	매출액/총자산
	재고자산회전율(배)	매출액/재고자산
	매출채권회전율(배)	매출액/매출채권
생산성	1인당기계장비율	기계장치/종업원수
	1인당매출액	매출액/종업원수
	1인당부가가치	(세전이익+금융비용+인건비+복리후생비+조세공과+임차료+감가상각비+리스료-이자수익) /종업원수
혁신성	매출액대비연구개발비(%)	연구개발비/매출액×100

<표5-7> 성과변수(타깃변수)로 활용할 수 있는 ICT관련 거시경제 지표

구분	지표	내용	출처	공표 주기
경제 지표	경제성장률	전년대비 연간 실질(2010년 기준) GDP 성장률	한은 국민계정	분기
	소비자물가 상승률	기준년도(2015년)의 소비자 물가 대비 해당연도의 상대적 물가 수준	통계청, 소비자물가조사 IITP ICT소비자물가지수	연
	수출입액	관세청 통관자료 및 무역통계(KITA)를 기초로 작성	산업부 수출입실적 IITP ICT수출입통계	월
	설비투자 증가율	설비투자: 기업이 장단기 경영계획하에 재생산을 목적으로 기계장치, 운반차량 및 건물 등의 설비를 도입하는 것을 의미하며 지출측면 GDP 중 총고정투자의 일부분을 차지	한은 국민계정	분기
	ICT생산(매출)	생산액 : 국내소재 사업체에서 발생한 생산액으로 부가세를 포함한 공장도가격의 총액	과기정통부 ICT주요품목 동향조사	월
일자리 상황	취업자 수	취업자 : 수입을 목적으로 주당 1시간 이상 일한 자, 18시간 이상 일한 무급가족종사자, 일시휴직자 등	통계청 경제활동인구조사	월
	종사자 수	종사자는 자영업주 및 무급가족종사자, 상용종사자, 임시 및 일용종사자 등을 포함하며, 조사기준연도 12월 31일 그 사업체에 고용되어 있는 남녀종사자를 기준으로 함	과기정통부 ICT인력동향조사	연
	고용률	고용률=(취업자수/생산가능인구)×100, 생산가능인구 : 만15세~64세에 해당하는 인구 * 생산가능인구 = 경제활동인구(취업자, 실업자) + 비경제활동인구(주부, 학생, 구직단념자)	통계청 경제활동인구조사	월
	실업률	실업률=(실업자/경제활동인구)×100		
일자리 창출	벤처기업 수	벤처기업: 벤처투자기업의 기준 요건은 벤처투자기관으로부터 투자받은 금액이 자본금의 10% 이상, 투자금액이 5천만원 이상	벤처인	월
	취업유발계수	10억원의 재화를 생산할 때 유발되는 취업자 수	한은 산업연관표	연
	종사자 수 증가	전년동기대비 증가한 종사자 수	과기정통부 ICT주요품목 동향조사	연
	신규채용 (상용근로자)	조사기준월(월력상) 초일부터 마지막 영업일 사이에 입직자 중 채용한 근로자 수	고용부 사업체노동력조사	월
	신설법인 수	신설법인 : 상법상의 영리법인(주식회사, 유한회사, 합자회사, 합명회사)으로 법원(상업등기소)에 설립등기를 마친 법인(개인기업 제외)	중기청 신설법인동향	월
	고용보험 신규취득	신규업체에서 고용보험 피보험자격을 취득한 자의 수, 생애 최초로 취업을 하여 고용보험에 가입한 피보험자 외 재취업이나 이직 등에 따라 발생하는 피보험자를 포함	고용정보원 고용보험통계	월

[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

LG Business Insight(2016), 디지털 경제, 과소평가되고 있다.

한국은행(2019), 디지털경제 측정에 대한 국제적 논의 현황.

한국소비자원(2020), 디지털 소비자의 편익 · 비용 이슈와 시사점.

중소기업벤처부 · 통계청(2019), 2018년 기준 소상공인 실태조사.

중소벤처기업부(2020), 중기부, 소상공인 디지털 전환 지원 본격 추진:「소상공인 성장・혁신 방안 2.0」 발표, 보도자료.

통계청(2020), 2020년 10월 온라인쇼핑 동향

한국은행(2017), GDP통계의 디지털 및 공유 경제 반영 현황 및 향후 개선 계획

한국은행(2019),「산업연관표」 통계정보고서.

한국은행(2020), 2019년도 금융정보화 추진 현황, 금융정보화추진협의회.

과학기술정책연구원(2020), 디지털 경제와 소비자 후생의 측정: GDP-B

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

Australian Bureau of Statistics (ABS), 2019, Measuring Digital Activities in the Australian Economy, written by Zhao, P.

Bureau of Economic Analysis (BEA), 2018, Defining and Measuring the Digital Economy, Working Paper, written by Barefoot, K., Curtis, D., Jolliff, W.,

Nicholson, J. R. and Omohundro, R..

Bureau of Economic Analysis (BEA), 2019, Measuring the Digital Economy: An Update Incorporating Data from the 2018 Comprehensive Update of the Industry Economic Accounts.

Bureau of Economic Analysis (BEA), 2020, New Digital Economy Estimates, written by Nicholson J. R..

China Academy of Information and Communications Technology (CAICT), 2017, G20国家数字经济发展研究报告(2017年).

China Academy of Information and Communications Technology (CAICT), 2017, 中国数字经济发展白皮书(2017年).

China Academy of Information and Communications Technology (CAICT), 2020, Digital Economy Development in China.

Chinese Academy of Cyberspace Studies, 2019, World Internet Development Report 2017, Springer.

Department for Digital, Culture, Media and Sport (DCMS), 2020, DCMS Sectors Economic Estimates 2018 (provisional): Gross Value Added.

Growth Intelligence and National Institute of Economic and Social Research (NIESR), 2013, Measuring the UK's Digital Economy with Big data, written by Nathan, M. and Rosso, A..

IMF, 2018, Measuring the Digital Economy, Staff Report.

Millar, J. and Grant, H., 2017, Valuing the Digital Economy of New Zealand.

OECD, 2019, Measuring the Digital Transformation: A Roadmap for the Future.

OECD, 2020, A roadmap toward a common framework for measuring the digital economy, Report for the G20 Digital Economy Task Force.

Office for National Statistics (ONS), 2014, Consultation on Measuring the Digital Economy.

Statistics Canada, 2019, Measuring Digital Economic Activities in Canada: Initial

Estimates.

UNCATD, 2019, Digital Economy Report 2019.

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

1. 이 보고서는 소프트웨어정책연구소에서 수행한 연구보고서이다.
2. 이 보고서의 내용을 발표할 때에는 반드시 소프트웨어정책연구소에서 수행한 연구결과임을 밝혀야 한다.

비매품/무료
95060
9 791197 090448
ISBN 979-11-970904-4-8 (PDF)