Special Feature

빅데이터 시대의 노동통계 : 아프리카를 비롯한 저소득 국가 사례를 중심으로

이가현 (Clinical Linguistics Information Processing(CLIP) Laboratory)

■ 서 론

사람은 누구나 경제 활동을 영위하면서 살아가지만, 경제 상황을 눈으로 직접 살펴볼 수 있는 것은 아니다. 각국 정부에 있어서 이렇게 눈에 보이지 않는 경제 상황을 지속적으로 파악하는 것은 매우 중요한데, 첫 번째 이유는 경제가 우리 생활에 미치는 영향이 크기 때문이며, 두 번째 이유는 경제에 문제가 있을 경우 더 큰 문제가 발생하지 않도록 각종 정책으로 시장에 개입해야 하기 때문이다.

지속적으로 경제 상황을 파악하기 위한 방법으로는 여러 가지가 있겠지만, 그중 가장 기본이 되는 것은 각국 통계청을 통해 정기적으로 수집되는 공식 경제통계를 꼽을 수 있다. 흔히 정확하고 신속한 공식 통계 수집은 경제 문제를 해결하는 첫걸음으로 여겨지고 있고(Lohr, 2009), 현상에 대한 정확한 파악 없이는 문제에 대한 적절한 해결 방안 도출이 불가능하기 때문에, 각 국가에서는 이러한 공식 경제통계를 수집하는 데 많은 시간과 비용을 투입하고 있다.

한편, 최근 들어 가용할 수 있는 데이터의 양이 기하급수적으로 늘어나고 이를 처리할 수 있는 기술이 발전함에 따라 사회 전반에 걸쳐 빅데이터를 활용한 분석이 널리 활용되고 있다. 빅데이터 분석은 기존에는 물리적 혹은 기술적 한계로 인해 접근할 수 없었던 방대한 양의 데이터를 활용하여 전통적 분석 기법으로는 파악할 수 없던 것들을 데이터 마이닝이나 기계학습과 같은 컴퓨터 공학적 방법을 통해 알아내는 것을 의미한다. 이러한 빅데이터 및 빅데이

터 분석 기법의 등장은 경제통계 수집의 방법도 변화시켰는데, 이 글에서는 아프리카를 비롯한 저소득 국가에서 전통적인 통계 수집 기법이 가진 한계를 극복하기 위해 빅데이터가 활용된 사례를 분석해보고자 한다.

이 글에서는 먼저 전통적인 통계 수집 방법에 대해 알아보고 그러한 전통적인 통계 수집 방법이 가진 한계점을 살펴본다. 그 후 빅데이터의 개념과 특성을 알아보고, 이를 활용한 통계 수집 기법이 기존의 한계를 어떻게 극복했는지 살펴본다. 사례 분석에서는 빅데이터를 통한 경제통계 수집이 아프리카 지역을 비롯한 저소득 국가에서 어떻게 활용되었는지 살펴보고, 마지막 단락에서는 빅데이터를 활용한 통계 수집 기법의 한계와 앞으로의 활용 가능성에 대해 평가해보고자 한다.

■ 전통적인 통계 수집

전통적인 통계 수집 방법

전통적인 경제통계 수집은 하향식으로 이루어져 왔다. 예를 들어 정책 결정을 위해 필요한 정보가 있으면, 그 정보를 도출하기 위한 분석기법을 미리 정하고 분석에 필요한 데이터를 직·간접적인 방법을 통해 수집해 오는 식이다. 즉 전통적인 통계 수집은 특정 정보에 대한 필요에 의해 데이터 수집이 시작되기 때문에 데이터가 사전에 정의되며, 그에 맞는 데이터 수집 계획이 수립된다. 데이터 수집을 위해 주로 사용되는 방식은 표본 추출을 통한 계획된 설문조사이다. 애초부터 데이터 수집의 의도가 정해져 있기 때문에 수집된 데이터를 분석하는 데 시간이 오래 걸리지 않지만, 수집된 데이터의 활용 가능성은 제한적이다. 전체 과정 중에서 데이터 수집 계획의 수립과 데이터 수집에 가장 많은 시간이 할애된다.

전통적인 통계 수집 방식의 한계

전통적인 방식을 통한 경제통계 데이터의 수집은 적시성, 정확성 및 신뢰성이 떨어진다는

문제점이 있으며, 이러한 문제점은 저소득 국가의 통계 데이터에서 더 심각하게 나타나는 경향이 있다.

- 적시성 : 면적이 넓고 IT 인프라가 발달하지 못한 사하라 남부 아프리카에서 설문을 기반으로 통계 데이터를 수집하게 되면, 수집된 데이터는 평균적으로 5년 이상 지체되어 있는 경우가 많다(Glugale, 2013). 이 경우 적시성이 없는 통계 데이터를 기반으로 의사결정을 내리게 되면, 시의적절한 상황 대응이 불가능하게 되어 정책의 효과가 미미해진다. 더욱이 해당 정책의 효과를 확인하는 데 다시 5년 이상이 걸리기 때문에, 이미 수행한 정책을 수정하는 것도 힘들어진다.
- 정확성 : 적시성과 더불어 아프리카를 비롯한 저소득 국가의 통계 데이터들은 정확성 또한 담보되기 어렵다. 특히 아프리카 통계 수집 시스템은 지난 수십 년간 업데이트되지 않은 경우가 많아서 정확성이 현격히 떨어진다. 일례로, 가나의 경우 2010년 새로운 통계 수집 기법이 도입되면서 기존의 통계 데이터를 업데이트하였는데, 이전 방식에 의해 수집·파악되었던 것에 비해 경제 규모가 약 60%가 증가하여, 단숨에 중위소득 국가로 편입된 사례가 있다(Glugale, 2013).
- 신뢰성 : 통계 데이터의 신뢰성 또한 문제가 되기 쉽다. 경제통계는 경제와 관련된 문제일 뿐만 아니라 정치적인 문제이기도 하며 사하라 이남 아프리카와 같이 정치적으로 안정되지 못한 국가에서는 공식 통계에 대한 신뢰성 문제가 더 크게 나타난다(Krätke and Byiers, 2014). 특히 임금 수준이나 실업률 같이 정치적으로 민감한 사안들은 있는 그대로 발표되었다고 신뢰할 수 없고, 정치적인 의도를 가지고 조작되는 경우가 많다. 이런 경우에는 비정부기관의 데이터를 통해 노동시장 상황을 조망하는 것이 더 객관적일 수 있다.

■ 빅데이터 기반 통계 수집

배 경

빅데이터의 개념

네이버 데이터랩 및 구글 트렌드에서 제공하는 자료에 따르면 빅데이터라는 용어는 2011년경 최초로 언론사 보도를 통해 국내에 소개되었으며 학술지에도 비슷한 시기에 등장하였다. 2012년경부터 빅데이터라는 용어가 본격적으로 등장하기 시작하였는데, 이 당시는 데이터의 발생 속도가 데이터의 저장 및 처리 속도를 급격히 추월하여 관련 기술 및 대응 방안에 대한 논의가 활발히 이루어지던 시기였다. 그 후 검색 빈도를 통해 파악한 빅데이터에 대한 국내 관심도는 꾸준히 증가하는 추세에 있다. 빅데이터를 문자 그대로 해석하면 방대한 용량의 자료를 의미한다. 하지만 빅데이터는 단순히 데이터의 용량만을 따지는 개념은 아니다. 빅데이터란 기존의 데이터 수집, 저장, 분석 기법으로 다루기 힘든 정도의 용량, 속도, 다양성을 가진 데이터를 의미하며 문맥에 따라 빅데이터 관련 기술 및 산업을 총칭하기도 한다. 빅데이터는 '기존의 기법으로 다루기 힘든' 데이터를 의미하므로, 데이터 분석, 관리, 저장 등 관련 기술과의 비교를 통해서만 파악될 수 있는 상대적인 개념이며 관련 기술이 발전함에 따라 정의가 달라질 수 있는 유동적인 개념이다.

빅데이터의 특성

빅데이터의 특성을 정의할 때 가장 보편적으로 활용되는 것은 가트너(Gartner)사의 애널리스트인 더그 레이니(Doug Laney)가 정의한 3V(Volume, Velocity, Variety)이며, 최근 2V(Veracity, Value)가 추가로 정의되어 3V와 함께 빅데이터의 특성으로 사용되고 있다.

- 고용량(high-volume) : 데이터의 물리적 측면에 대한 특성으로 데이터의 크기가 크다는 것을 의미한다. 데이터의 크기가 클수록 더 많은 저장 공간이 필요하며 데이터를 처리하는 데 걸리는 시간이 늘어난다. 빅데이터의 이러한 특성은 데이터의 효율적인 저장을 위한 기술의 발전을 촉진한다.

- 고속도(high-velocity) : 데이터의 교환이 빈번하게 일어나는 특성을 의미한다. 실시간으로 발생하고 기기 간에 데이터가 교환되는 속도가 빨라짐에 따라 단일 데이터의 크기는 크지 않지만 전통적인 기술로는 분석이 불가능한 데이터가 생겨나게 되었다.
- 다양성(high-variety) : 데이터의 종류가 다양하고 다른 종류로의 변환이 잦은 특성을 말한다. 구조화된 데이터의 분석이 주를 이루었던 과거와는 달리 최근에는 텍스트, 이미지, 음성, 비디오 등 수집되는 데이터의 종류가 다양화되었으며, 수집되는 양 또한 증가하였다. 빅데이터 패러다임하에서는 이러한 비구조화된 자료도 분석의 대상에 포함된다.
- 신뢰성(high-veracity) : 수집된 데이터의 질이 높음을 의미한다. 전통적인 방식으로 생성된 데이터와 달리 빅데이터는 자료의 소스가 다양해지고 생성하는 사람이 많아짐에 따라 데이터의 질을 논할 때 여러 가지 요소가 고려될 수 있는데, 데이터가 얼마나 일관적인가, 수집과정에서의 에러는 없는가, 데이터가 현상을 얼마나 잘 반영하고 있는가가 주요 관심 대상이 된다. 빅데이터 분석을 논할 때마다 자주 언급되는 GIGO(Garbage in, garbage out)란 문구에서 알 수 있듯 부정확한 데이터를 가지고 하는 분석은 아무런 정보를 창출하지 못한다.
- 고가치(high-value) : 데이터를 활용했을 때의 가치가 높음을 의미한다. 앞서 언급한 바와 같이 데이터는 그 자체로서 가치를 가진다기보다는 가공 및 분석을 통해서 유용하고 가치 있는 '정보'로 변환된다. 빅데이터는 분석을 통해 동일한 데이터에서도 여러 가지 정보가 추출될 수 있으며, 그 정보가 갖는 가치가 큰 데이터이다.

빅데이터에 기반한 통계 수집의 특징

빅데이터를 통한 통계 수집의 특징은 전통적인 통계 수집 방식과 비교했을 때 더 뚜렷해진다. 전통적인 통계 수집 방식이 잘 설계된 표본 추출에 의한 하향식 방법이라면 빅데이터를 통한 통계 수집은 상향식 방법이라고 할 수 있다. 필요한 정보를 미리 정의하고 분석 기법 및 데이터를 확정하는 것이 아니라 데이터에 따라 분석 기법 및 도출 가능 정보가 정해진다. 빅데이터 기반 통계 수집에 활용되는 해당 통계 수집만을 목적으로 하는 새로운 데이터가 아니라, 다른 목적을 위해 존재하는 기존 데이터를 활용하는 경우가 많다. 또한 전통적인 방식으로 수집

된 데이터가 통계적 기법을 통해 분석되는 것과는 달리 빅데이터 기반 통계 분석에서는 데이터 마이닝 또는 다양한 기계학습 알고리즘을 통한 추론적 분석이 사용된다. 데이터라는 현상이 발생한 원인이나 과정에 대한 분석보다는 어떤 현상이 발생하고 있는지를 분석하는 것이 주요 주제이다. 이러한 기법들은 기존 통계로 알 수 없었던 것들을 파악하는 데 도움을 준다. 예를 들면, Marinescu and Wolthoff(2015)는 잡포털 빅데이터에 포함된 지리 정보를 활용하여 노동자들의 구직활동 패턴이 지리적인 요소에 의해 차이가 있음을 밝혀냈다. 빅데이터는 데이터 자체가 통계를 목적으로 작성된 것이 아니며, 방대한 양의 데이터는 의도를 가지고 조작되기 힘들기 때문에 신뢰성이 담보된다고 여겨진다. 전통적인 통계 수집이 데이터 수집 계획의 수립 및 데이터의 수집에 중점을 둔다면, 빅데이터 기반 분석에서는 분석하는 과정에 모든 초점이 맞춰지게 된다.

초기 빅데이터를 기존 통계 수집의 보완적 또는 대체적 용도로 사용하려는 시도는 2009년 무렵 시작되었으며(Lohr, 2009), 2019년 현재 28개 EU 회원국이 빅데이터에 기반한 실시간 노동시장 모니터링 시스템을 가동 중이다(UK Commission for Employment and Skills, 2019). 이 중 특히 구인광고와 관련된 설문 및 수집된 웹데이터 등이 포함된 잡포털 빅데이터(Cedefop, 2014)는 공식 통계로의 편입을 추진 중이다.

빅데이터의 종류

통계 수집에 활용할 수 있는 빅데이터는 다음과 같이 분류될 수 있다.

Human-generated

사람에 의해 생성된 데이터이다. 입력에 오류가 많고 정형화되어 있지 않다는 단점이 있지만, 상황 파악을 위한 적시성을 가지고 있으며 자료에 접근이 용이하다. 사회관계망서비스(SNS) 데이터 등이 이에 속한다.

Process-mediated

기존에 존재하는 상업 시스템에서 나오는 데이터이다. 자료가 잘 정제되어 있으며 비교적

정확하여 데이터의 질이 높지만, 접근이 어려운 경우가 대부분이며 다른 시스템에서 생성된 데이터 간에는 일관성이 떨어져 분석이 어려운 경우가 많다. 매출 데이터, 의료 데이터 등이 이 분류에 속한다.

Machine-generated

사물인터넷(IoT) 네트워크상의 각종 센서 및 클라우드와 같은 컴퓨터 시스템상에서 기계에 의해서 자동으로 생성되는 데이터이다.

■ 빅데이터 기반 통계 수집 사례

케냐의 노동 수요-공급 불균형 분석

필요성

사하라 이남 아프리카 지역의 저소득 국가는 경제 상황에 대한 전반적인 정보가 부족하다. 특히 노동 수요 및 공급에 대한 적시성 있는 정보가 부재하여, 일자리 개수, 새로 생성되는 일자리의 종류, 분야별 구인 현황 등 기본적인 정보도 얻기 힘든 상황이다. 이러한 정보의 비대칭성은 심각한 시장 불균형을 초래하였으며, 오랫동안 사회 전반에 걸쳐 자원의 효율적 공급 및 배분을 저해하였다. 미래 인적 자원 개발이 저해되는 것은 이들 국가를 빈곤에서 벗어날 수 없게 만드는 심각한 문제이다(Samans and Zahidi, 2017). 이러한 상황에서는 빅데이터를 통한 통계 수집과 같은 혁신적인 정보 수집 기법이 필요하다. Ketamo & Passi-Rauste(2019)는 노동통계의 적시성 결여로 인해 시장에서 원하는 노동 수요와 대학이 수행하는 노동 공급 사이의 불균형을 빅데이터 분석 기법을 통해 적시성 있게 분석하는 연구를 수행하였다.

분석 방법

노동시장의 수요 및 공급 분석을 위해 가장 먼저 해야 할 것은 필요한 데이터를 어디서 어떻게 수집할지를 정하는 것이다. 노동시장 수요 분석을 위해 해당 연구에서는 케냐에서 가장

활발히 활용되는 3개의 공공 구직사이트를 선정하였다. AI 기반 알고리즘이 선정된 사이트를 방문하여 자동으로 구인광고를 수집·분석하였고, 이를 통해 노동시장에서 현재 요구되는 기본능력을 먼저 도출하였다. 한편, 노동시장 공급도 같은 방식으로 분석되었는데, 노동시장에 공급되는 인력의 기본능력은 각 대학에서 제공하는 커리큘럼을 수집하여 자연어 처리 기법을 적용하여 분석하였다.

분석 결과

노동시장에서 요구되는 능력과 공급되는 인력이 가진 능력을 비교하면 시장에서 요구되고 있으나 대학에서 아직 가르치지 못하는 기술, 또는 대학에서 가르치고는 있으나 시장에서 더는 사용되지 않고 있는 기술을 알 수 있다. Ketamo & Passi-Rauste(2019)의 분석 결과 노동시장에서 요구되는 기술과 실제로 대학에서 가르치는 기술 사이에는 상당한 불균형이 발생하고 있다. 대학에서 가르치는 윈도우, 리눅스, 객체 지향 프로그래밍, 정보보안 등은 이미 시장에서 표준화되어 차별화되지 못하는 기술들이며, 아마존 AWS, 클라우드 컴퓨팅 등은 시장에서 활발히 요구되나 대학에서 가르치고 있지 않는 기술이므로 추가적으로 커리큘럼에 포함시킬 필요가 있다.

휴대폰 사용이 남아프리카 지방 노동시장에 미치는 영향

필요성

앞서 언급된 바와 같이 남아프리카와 같은 저소득 국가, 그중에서도 더욱 낙후된 지방의 노동시장은 전통적인 통계 수집 기법으로는 파악이 어렵다. 따라서 Klonner and Nolen(2010)은 휴대폰과 같은 기술의 보급이 지방 노동시장에 미치는 영향을 분석하고 휴대폰 네트워크를 바탕으로 노동시장을 파악하려는 시도를 하였다.

분석 방법

빅데이터 분석을 위해 남아프리카 전체 시장 점유율이 56%에 달하는 최대 통신사 보다콤(Vodacom)으로부터 연도별 통신망 서비스 범위 데이터를 수집하였다.

분석 결과

휴대폰의 보급은 노동시장에 극적인 변화를 가져온 것으로 밝혀졌다. 일단 휴대폰 사용이 불가능했던 지역이 서비스 가능 지역으로 편입되면, 고용이 약 15% 증가하는 효과가 있다. 한편, 이러한 고용 증대 효과는 남성이 아닌 여성에게 집중되었다. 네트워크 서비스 지역 확대는 고용의 증대뿐 아니라 소득의 증가도 수반하였다. 서비스 가능 지역으로 편입된 경우 특히 농업에서의 고용이 감소했는데, 고용 감소는 주로 남성에게 나타났다.

인도 잡포털 빅데이터의 정책적 활용

필요성

교육은 노동생산성 향상, 빈곤 퇴치 및 경제 성장에 도움을 주지만 시장 수요에 맞지 않는 교육(특히, 기술에 대한 교육)은 경제 성장에 큰 저해 요소로 작용한다. 그동안 인도의 노동시장 통계는 국가표본조사(National Sample Survey)를 통해 이루어져 왔다. 하지만 링크드인(LinkedIn), 인디드(Indeed), 몬스터(Monster), 커리어빌더(Career Builder)와 같은 글로벌 잡포털 및 로컬 잡포털에서 생성되는 빅데이터를 적절한 분석 기법을 활용해 분석하면 노동시장 상황을 적시성 있게 파악하는 데 도움이 된다. 해당 연구에서는 빅데이터를 활용한 분석 기법이 기존 노동통계를 보완하여 5가지 분야, ① 노동시장 모니터링 및 분석, ② 시장에서 요구되는 기본능력 분석, ③ 구직행동 분석, ④ 시장에서 요구되는 기본능력 예측, ⑤ 실험연구에서 정책적으로 유용하게 활용될 수 있음을 보여 주었다.

분석 방법

이 분석은 2007년부터 2015년 사이에 인도 잡포털 사이트인 바바잡(Babajob)을 통해 게시된 85만 8천 개의 구인광고를 수집하여 이루어졌다.

분석 결과

분석 결과 인도 노동시장에는 지역 및 직종에 따라 성별 임금불균형이 다르게 나타나는 것으로 파악되었다. 집사, 유모와 같은 일부 가사도우미 직군에서는 여성의 임금이 훨씬 높게

나타났으나, 금융, 교육, 요리를 비롯한 대부분의 직군에서는 여성의 임금이 남성의 임금에 미치지 못하는 것으로 나타났다. 또한 보모, 사무직, 기술직 등의 제한된 직군에서는 남녀 임금이 평등한 것으로 분석되었다. 지역에 따라 남녀 임금 격차 불균형에서도 차이가 큰 편으로 구르가온(Gurgaon) 지역은 남녀 임금이 거의 균등하였으나, 델리(Delhi) 같은 지역은 여성의 평균 임금이 남성 임금의 80% 정도밖에 미치지 못하는 것으로 분석되었다.

또한 구인광고에 포함된 자격 요건을 분석하여 전문성에 따라 요구되는 자격 요건의 차이를 분석하였는데 전문직, 비전문직 모두 경험과 기술, 언어와 의사소통 능력을 중시하였으나, 특히 전문인력 구인에서 경험 및 기술의 비중이 상대적으로 높았다.

구인광고가 게시된 시점으로부터 경과한 시간과 서류 전형을 통과하는 비율은 반비례 관계에 있는 것으로 나타났다. 즉, 게시되자마자 지원한 지원자일수록 서류 전형을 통과할 가능성이 큰 것으로 나타났다.

구인광고를 통해 파악한 평균임금 수준은 해가 지날수록 꾸준히 상승하고 있으나 뚜렷한 계절성을 띠는 것으로 나타났다. 여름에 임금 수준이 가장 높고 겨울에는 임금 수준이 떨어지는 것으로 분석되었다.

■ 빅데이터 활용의 한계

모든 방법론이 저마다의 문제점을 가지고 있듯이, 빅데이터에 기반한 통계 수집에도 명확한 한계가 있다. 이번 단락에서는 빅데이터 분석의 일반적 한계점에 대해서 알아보고 저소득 국가에서의 빅데이터 활용에는 어떠한 제약이 있는지 파악하고자 한다.

일반적인 한계점

데이터 품질

컴퓨터 공학 분야에서 흔히 얘기되는 GIGO(Garbage In, Garbage Out)란 좋은 분석 결과를 위해서는 품질이 높은 데이터가 필요하다는 의미이다. 빅데이터는 데이터의 용량이 방대

한 만큼 많은 오류와 노이즈 데이터를 포함하고 있으며, 데이터를 정제하는 데도 많은 비용이 든다. 빅데이터 수집 및 분석에 드는 총비용 중 50~80%가 데이터 정제에 소요된다는 연구 결과도 있다(Lohr, 2014). 아무리 훌륭한 분석 방법도 데이터의 품질이 담보되지 않으면 무용지물이 될 수밖에 없다. 그러나 저소득 국가의 경우 자금 부족으로 인해 양질의 데이터를 생산하는 것이 어려운 실정이다.

데이터 및 분석 방법 선택의 오류

연산 기법의 발달로 인해 컴퓨터화된 분석 방법이 대세를 이루고 있고, 그중에서도 기계학습을 필두로 한 인공지능 기법이 활발히 활용되고 있다. 하지만 충분히 크지 않은 데이터에 기계학습과 같은 방법론을 적용했을 때도 문제가 생길 수 있다. 모든 기계학습 방법론은 과적합 위험을 내포한다. 과적합이란 기계학습 분석 수행 시 학습이 너무 많이 수행되어 학습 대상 데이터(트레이닝 데이터)에 대한 모델의 일반성이 떨어지는 현상을 말한다. Loukides(2014)가 지적한 바와 같이 많은 분석에서 원래의 분석 의도에 맞지 않는 데이터가 잘못 사용되고 있으며 인과관계를 설명해주지 못하는 데이터를 가지고 인과관계를 설명하려고 하는 오류가 나타난다. 일반적으로 저소득 국가에서는 사용 가능한 데이터의 크기가 크지 않은 경우가 많으며, 숙련된 연구자가 부족하기 때문에 상대적으로 크기가 작은 빅데이터를 분석하기 위한 적절한 방법을 택하는 데도 어려움을 겪는 것이 현실이다.

데이터 환경의 변화

데이터가 항상 같은 환경에서 생성되지 않는다는 것도 빅데이터의 활용을 저해하는 요소이다. 현재 활발히 사용되고 있는 페이스북, 인스타그램, 트위터 등의 SNS 플랫폼의 사용층은 언제든지 달라질 수 있다. 현재는 인스타그램이 20~30대를 대변하는 SNS 플랫폼이라는 인식이 강하다. 하지만 지금으로부터 5~10년의 시간이 흘러도 여전히 주요 사용자층이 20~30대일지는 분명하지 않다. 환경의 변화를 고려하지 않고 인스타그램이 여전히 20~30대를 대변한다고 가정하는 오류 등이 데이터 환경 변화를 고려하지 않은 잘못된 데이터 사용의 예가 될 수 있다.

저소득 국가에서의 활용 한계

대표성 문제

빅데이터는 주로 그 데이터의 생성자 및 환경을 대변하기 때문에 데이터를 생성시킬만한 여력을 갖지 못한 사람이나 지역 또는 산업이 존재할 경우 해당 부분은 빅데이터에 의해 표현되지 못한다는 한계가 있다. 일례로 빅데이터 분석에 활발히 활용되고 있는 SNS 데이터는 주로 휴대폰과 같은 모바일 플랫폼을 통해 생성되므로 선진국의 경우에는 지역이나 산업을 막론하고 노동자들의 스마트 기기의 사용을 전제하고 빅데이터를 분석할 수 있다. 하지만 인터넷 및 이동통신의 보급 상황을 미루어 볼 때, 아프리카를 비롯한 저소득 국가에서는 스마트 기기의 보편적 사용을 전제할 수 없는 상황이다. SNS 데이터로 노동시장 전반에 대한 분석을 수행할 경우 휴대폰이 보급되지 않은 지역 또는 산업에 분포한 노동시장의 분석은 제한될 것이므로, 이러한 데이터는 대표성이 결여된다.

정확성 및 신뢰성 문제

사람에 의해서 생성되는 빅데이터의 정확성 및 신뢰성은 데이터를 생성하는 사람들의 인식 수준에 의해서도 좌우된다. 따라서 데이터를 생성하는 사람들의 교육 수준이 낮거나 편향된 인식을 갖고 있는 경우 생성된 데이터의 정확성 및 신뢰성에 문제가 생기는 경우가 많다. 일례로 아프리카 지역의 질병 발현을 SNS로 예측하는 모델은 지금까지 활발히 사용되어 왔지만 사스(SARS)와 같은 호흡기 질환이 유행했을 때 이 질병의 유행을 파악하는 데는 실패하였는데, 그 이유는 사람들이 사스와 독감을 제대로 구분하지 못하여 사스를 독감이라고 언급하는 경우가 많았기 때문이다.

■ 결 론

지금까지 빅데이터를 통해 전통적인 노동통계 수집의 한계를 극복하는 사례들을 알아보았다. 빅데이터를 통한 통계 수집의 유용성 및 정확도에 따라 노동통계를 담당하는 공공기관의

역할 또한 달라질 수 있다. 이를테면, 전통적인 방식의 통계 수집 속도가 늦어 적시성이나 정확성이 떨어지는 일부 통계 지표들의 경우, 빅데이터로부터 추출한 통계를 보완하거나 정확성에 대한 검증을 거쳐 공식 통계로 활용하거나 혹은 자료가 수집되기 전 미리 예상치를 공표하는 방법 등이 있을 수 있다. 또는 적시성을 요하는 통계 수집에는 빅데이터를 활용하거나 전통적인 방식에 의해 집계된 공식 통계는 추후 빅데이터를 분석하는 방법론 검증의 용도로만 활용하는 방안도 고려해볼 만하다. 아프리카 및 저소득 국가들은 자료의 적시성, 정확성 및 신뢰성이 떨어지기 때문에 빅데이터를 활용한 통계지표 도입을 적극적으로 검토하여야 한다. 어찌 되었건, 빅데이터가 전통적인 노동통계 수집 방식을 보완하는 것은 분명하며, 이를 적극적으로 활용하여 정책에 반영하는 것은 정책 담당자들의 몫으로 남겨져 있다. KLI

참고문헌

- Cedefop(2014), *Real-time Labour Market Information on Skill Requirements : Feasibility Study and Working Prototype*, Cedefop Reference number AO/RPA/VKVET-NSOFRO/Real-time LMI/010/14, Contract notice 2014/S 141-252026 of 15/07/2014, 2014.
- Glugale, M.(2013), "Fix Africa's Statistics", *Huffpost*, 18 December, available at : https://www.huffpost.com/entry/fix-africas-statistics_b_2324936
- Ketamo, H. and A. Passi-Rauste(2019), *Labor Market Analysis and Curriculum Gap Assessment Using Big Data in Kenya,* Final report for the World Bank contract 7192067.
- Klonner, S. and P. J. Nolen(2010), "Cell Phones and Rural Labor Markets : Evidence from South Africa", *Proceedings of the German Development Economics Conference, Hannover 2010 56*, Verein für Socialpolitik, Research Committee Development Economics.
- Krätke, F. and B. Byiers(2014), *The Political Economy of Official Statistics*, Partnership in Statistics for Development in the 21st Century.
- Lohr, S.(2009), "For Today's Graduate, Just One Word : Statistics", *New York Times*, 5 August, available at : www.nytimes.com/2009/08/06/technology/06stats.html

· Lohr, S.(2014), "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights", *New York Times*, 17 August, available at : www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-toinsights-is-janitor-work.html

· Loukides, M.(2014), "The Backlash Against Big Data, Continued", *O'Reilly Radar*, 11 April, available at : http://radar.oreilly.com/2014/04/the-backlash-against-big-data-continued-2.html

· Marinescu, I. E. and R. P. Wolthoff(2015), "Opening the Black Box of the Matching Function : The Power of Words", *IZA Discussion Papers*, No. 9071.

· Nomura, S., S. Imaizumi, A. C. Areias, and F. Yamauchi(2017), *Toward Labor Market Policy 2.0 : The Potential for Using Online Job-Portal Big Data to Inform Labor Market Policies in India*, Policy Research Working Paper; No. 7966. Washington, D.C. : World Bank. https://doi.org/10.1596/1813-9450-7966

· Samans, R. and S. Zahidi(2017), *The Future of Jobs and Skills in Africa*, World Economic Forum.

· UK Commission for Employment and Skills(2019), "LMI for All", 2015. Last accessed March 2019 at : www.lmiforall.org.uk/

· Zardetto, D.(2016), *The Implication of Big Data for Official Statistics*, European Commission, available at : https://circabc.europa.eu/sd/a/c35f00b3-3103-4873-a5a8-68386822af03/DAY%201_ITEM%202_The%20implication%20of%20big%20data.pdf

· Vale, S(2020), *Classification of Types of Big Data*, available at : https://statswiki.unece.org/display/bigdata/Classification+of+Types+of+Big+Data