국제사회보장리뷰 2020 가을호 Vol. 14, pp. 5~15

미국의 사회과학 및 정책 연구를 위한 행정 빅데이터 활용*

Using Administrative Big Data to Solve Problems in Social Science and Policy Research

시 숭(펜실베이니아대학교 교수), 토머스 콜먼(시카고대학교 교수)
Xi Song(University of Pennsylvania), Thomas S. Coleman(University of Chicago)

이 글은 개인 수준의 공공 행정 데이터의 유용성이 미국을 포함해 전 세계적으로 급증한 상황을 설명한다. 이러한 데이터 세트들은 독립된 소스로 활용될 수도 있고 여러 다른 소스에 걸쳐 연결될 수도 있다. 이 새로운 유형의 소스들로 인해 사회, 인구 통계 및 경제 변화에 대한 변혁적 연구와 정책 평가 및 기타 실험적 분석이 용이해질 것이다. 이 글에서는 미국 행정 빅데이터의 현주소와 사회과학 및 정책 연구를 발전시킬 수 있는 빅데이터의 잠재성, 그리고 이 데이터를 실제로 사용할 때의 장점과 해결해야 할 과제에 대해 논의한다. 그뿐만 아니라, 현재 미국에서 진행 중인 몇 가지 대규모 행정 데이터 프로젝트를 보여 주고, 다른 국가들이 앞으로 이와 같은 노력을 할 수 있는 계기를 제공하고자 한다.

1. 들어가며

빅데이터는 데이터의 양, 다양성, 속도 측면에서의 급격한 성장을 특징으로 하며 사회, 행동 및 정책 연구 분야에 대변혁을 일으키고 있다. 소셜 미디어, 온라인 등록, 거래, 정보 보관, 위성 및 GPS 추적 장치, 자연 언어, 정보 네트워크 등에서 막대한 양의 데이터가 발생한다. 자동화, 기계 학습, 정보통신기술에 따라 디지털 아카이브가 기하급수적으로 증가했다. 문자, 그림, 동영상, 위치 정보 등과 같은 새로운 유형의 데이터로 인해 사회과학 연구 분야에서 데이터의 개념이 확장되었다. 크라우드소싱 플랫폼이 많아지면서 절감된 비용으로 더

* 영문으로 작성된 원고로 원문은 한국보건사회연구원 영문홈페이지 참고(https://www.kihasa.re.kr/english/main.do)

쉽고 빠르게 데이터를 수집할 수 있게 되었다. 실시간 데이터가 자연적으로 발생하면서 데이터 생성 과정에서 연구자의 역할이 축소되었다. 이 글은 빅데이터 유형 중 하나인 행정 마이크로데이터를 집중적으로 다룬다. 사회과학 연구자는 행정상 개인정보(인구조사 정보, 납세 신고 파일, 인구동태 통계, 교구 정보, 선거인 등록, 의료비 청구, 가계도, 인구 명부 등)로부터, 그리고 이 같은 마이크로데이터가 서로 이질적인 행정 및 기타 데이터 소스 전반에 걸쳐 통합됨으로써 지금까지 연구상 큰 도움을 받아 왔고 앞으로도 그러할 것이다.

사회과학 분야에서의 실증적 연구는 전통적으로 연구 또는 정책 목적을 위해 의도적으로 설계되고 수집된 서베이 데이터를 기반으로 이루어졌다. 이와 대조적으로 오늘날 세계에는 행정 및 민간 영역 모두의 데이터 소스로부터 생성된 '빅데이터'가 넘쳐나고 있다. 이 글을 통해 사회과학 연구에서의 현 공공 행정 데이터 활용 정책과 관행을 미국의 사례에 비추어 대략적으로 살펴보려 한다. 우선, 서베이 데이터와 행정 데이터를 대조해 보고, 학문적 연구와 정책 분석을 목적으로 행정 데이터를 활용하고 서로 연결함으로써 얻을 수 있는 장점과 주의 사항들을 살펴 미국에서 진행 중인 데이터 연계 프로젝트 사례의 배경과 성과, 개인정보 보호에 관한 도전 과제 및 시행 이슈 등을 논의하고자 한다. 마지막으로, 다른 유형의 빅데이터와 대조해 본 다음 미국에서의 향후 발전 모습과 다른 국가에서 이루어지고 있는 유사한 노력들에 미칠 수 있는 영향에 대해 생각해 보겠다.

2. 서베이 데이터

서베이 데이터라는 용어에 대해 널리 통용되는 정의는 없다. 서베이 방법론의 고전인 Groves 외(2009)의 저서에서는 인구조사를 서베이의 초창기 유형으로 여기는 반면, 민간 부문에서는 일반적으로 서베이와 여론 조사라는 용어가 혼용된다. 이 글에서는 서베이를 전체 인구에 대한 정보 제공의 목적으로 부분 집단을 대상으로 수집한 데이터로 정의한다(Slemrod, 2016). 일반적으로 서베이는 연구자 또는 기관이 관심을 두고 있는 특정한 기본적 구조 또는 관계에 대한 정보를 도출하기 위해 설계된다. 이때 서베이 데이터를 관련된 이론적 구조와 연결 짓기 위한 숙고와 노력이 이루어진다. 서베이는 (1) 포함오차, (2) 무응답오차, (3) 표본오차, (4) 서베이 측정오차(응답 부정확성)라는 네 가지 주요 오차 요인에 빠지기 쉽다. 서베이 데이터는 수치 또는 문자 형태일 수도 있고, 서베이 연구 초기(1930년대)에 실

시된 대면 또는 우편 설문조사부터 오늘날의 완전한 전자 설문조사에 이르기까지 다양한 형식으로 수집될 수도 있다(Groves, 2004). 공식적인 정부 기관에서 독점적으로 취합하고 보유한 행정 데이터와 비교할 때, 서베이 데이터는 등록, 연구, 마케팅 또는 정치적 예측을 목적으로 공공 또는 민간 기관에 의해 수집될 수 있다.

3. 행정 데이터

서베이 데이터와는 대조적으로 행정 데이터는 일반적으로 인구동태 신고 사항, 인구조사 정보, 세금 징수[미국 국세청(IRS: Internal Revenue Service)], 주 실업보험, 퇴직자 연금[미국 사회보장국(Social Security Administration)의 임금 및 연봉 기록], 의료 혜택 범위[메디케어·메디케이드 서비스 센터(CMS: Centers for Medicare and Medicaid)], 교육 등과 같은 행정 프로그램을 관리·운영하는 업무의 일환으로, 공식적인 정부 기관이 취합 및 보유한다. 행정 데이터는 서베이 데이터와 관련하여 다음과 같이 세 가지 중요한 이점이 있다(Card et al., 2010).

1. 대규모의 데이터: 일반적으로 표본 또는 부분 집단보다는 전체 인구를 포함시키므로 표본오차를 줄이고 희귀한 현상에 대한 연구를 가능하게 한다.
2. 고품질의 정보: 일반적으로 응답에 따라 응답자는 긍정적 또는 부정적인 영향을 받게 된다. 즉, 혜택(예를 들어, 실업보험 혜택)이 제공되는 식의 긍정적인 결과나, 무응답 또는 잘못된 응답을 할 경우 불이익(거짓으로 소득세 신고를 할 경우 벌금 및 징역)이 돌아가는 식의 부정적인 결과가 따른다.
3. 종단 데이터: 주로 동일한 개체를 오랜 시간에 걸쳐 다루다 보니 자연적으로 패널 데이터 구조를 얻는다.

4. 행정 데이터의 연구상 이점

행정 데이터는 사회과학 연구자에게 다양한 이점을 제공한다. 대규모 표본을 통해 소규모 하위 그룹, 희귀한 현상, 그리고 기존 서베이에서는 놓칠 수도 있는 그룹들을 연구할 수 있는 폭넓은 기회가 만들어진다. 무응답률이 더 낮고 데이터의 품질도 뛰어나기 때문에 이미 수용된 연구 결과를 재검토할 수 있고 새로운 관점을 발전시킬 수도 있다. 다수의 코호트를 포함시킴으로써 시간 변화에 따른 사회 변화를 연구할 수 있게 된다. 또, 행정 데이터는 1차 데이

터 수집 시 저장해 둘 수 있다.

미국의 과세 데이터를 활용한 최상위 소득 점유율과 불평등에 대한 연구는 Piketty와 Saez(2003)가 도입한 것으로, 대규모 표본과 (가상) 코호트의 장점을 잘 보여 주는 대표적인 예이다. 최상위 1% 또는 0.1%의 소득자는 보통 두 가지 이유로 일반적인 사회조사에서 잘 드러나지 않는다. 첫째, 서베이 표본은 주로 일반 대중을 대상으로 하기 때문에 극단에 해당하는 인구 집단은 표본 규모가 너무 작아 이들의 소득에 대해 신뢰할 만한 추론을 도출하기 어렵다. 둘째, 상대적으로 규모가 작은 서베이에서는 응답자의 신원을 보호하기 위해 주로 일정한 값 이상은 부호 처리하여 소득을 표시한다(예를 들어, 20만 달러가 넘는 모든 소득은 실제 액수가 아닌 "200k 달러 초과"로 표시한다). 미국 국세청(IRS)의 행정 데이터와 납세 신고서 모집단 전체를 열람할 수 있게 되자 소득 분포 최상위 1%와 0.1%에 대한 연구가 가능해졌다. 이러한 데이터는 약 100년 전의 기록으로까지 확장되어 미국 불평등의 역사와 현주소에 대해 재고할 필요가 있다는 과제를 남겼다.

Meyer와 Mittag(2019)에서도 빈곤 측정 기준을 재검토하고 개선하기 위해 행정 데이터를 활용하였다. 이들의 연구 결과에 따르면, 개량된 측정 기준을 적용하면 정부의 빈곤 퇴치 프로그램의 외형적 유효성이 상당히 달라진다. 미국의 현재 인구 서베이(CPS: Current Population Survey), 구체적으로 말해 3월 보충 자료[연간 사회경제 보충 자료(ASEC: Annual Social and Economic Supplement)]는 개인 및 가구 소득과 관련된 서베이의 일반적인 출처이다. 구체적으로 Meyer와 Mittag(2019)는 현재 인구 서베이(CPS), 뉴욕의 사회복지기관, 미국 주택도시개발부(HUD: Federal Department of Housing and Urban Development)의 데이터를 모두 통합해 정부의 현금 지원 프로그램에 대해 서베이로 조사된 지급액과 행정 지급액을 비교했다. 영양 보충 지원 프로그램인 SNAP(Supplemental Nutrition Assistance Program, 푸드 스탬프)와 빈곤 가족 지원 프로그램인 TANF(Temporary Assistance for Needy Families), 일반부조(General Assistance), 그리고 연방정부의 주거 지원에 대해 현재 인구 서베이(CPS)로 조사된 액수는 실제로 지급된 액수보다 상당히 낮았다. 이와 같은 불일치는 소득 분포의 최하위층에 가장 크게 나타나 빈곤율에 상당한 영향을 미쳤다. 이는 빈곤 측정 도구에 대한 정책 논의를 진전시키는데 도움이 되었다(OMB, 2020).

5. 행정 데이터의 연구상 과제

행정 데이터가 위에 언급한 서베이 데이터의 오차들을 축소하거나 제거할 수는 있겠지만, 서베이 데이터의 모든 문제점 또는 연구상의 모든 문제들을 해결하지는 못한다. 행정 데이터에는 측정오차, 보고오차, 정보 매칭 및 관련 통계 단위 구성의 문제, 또는 분석가가 요구한 측정 기준과 특정한 행정상의 측정 기준이 달라서 생겨나는 문제들이 있을 수 있다(Groen, 2012). 그러므로 데이터의 정의와 생성 과정을 이해하고 문서화하는 것이 중요하다. 왜냐하면 행정 데이터는 주로 특정한 행정적 목적을 위해 설계되는 것이지, 연구자가 연구하고자 하는 구조에 반드시 일치하는 것은 아니기 때문이다(Connelly et al., 2016). 일부 선행연구에서는 이와 관련하여 행정 데이터 활용에 필요한 데이터 매칭과 정제 기법에 대해 간략한 설명을 제공한다(Goerge & Lee, 2002).

최상위 소득 점유율에 관한 연구에 사용된 미국 국세청(IRS) 과세 자료는 데이터 정의와 관련해 어떤 점에 주의를 기울여야 하는지를 보여 주는 좋은 예가 된다. 혹자는 납세 신고서에 신고한 소득이 간단명료한 소득 수치를 제공하는 것이라 추정할 수도 있다. 그러나 과세 자료는 전혀 완벽하지 않다. Slemrod(2016)는 여러 쟁점을 제시하고 있는데, 이 중 두 가지는 특히 장기간의 변화 비교와 관련이 있다(Auten & Splinter, 2019; Guvenen & Kaplan, 2017). 첫째, 과세 규정의 변화는 실제 소득 변화가 없어도 개별 주체의 소득 신고 방식을 변화시킬 수 있다. 1986년 세제개혁법(Tax Reform Act)으로 인하여 개인 소득으로 신고되는 소득에 대한 법정 세율이 바뀌게 되어, 근본적인 소득 변화가 없음에도 개별 신고서에 개인 소득으로 신고할 경우 적용되는 인센티브가 달라지게 되었다. 소득 분포 최상위와 최하위에 적용되는 인센티브의 차등적 변화는 장기간의 변화 비교를 복잡하게 만든다. 둘째, 과제 자료를 통해 측정 단위와 관련한 쟁점을 살펴볼 수 있다. 납세 신고서에 의한 소득 측정이 자연스러운 해결책처럼 보이고 학계에서도 이런 방식을 자주 채택하지만, 정책의 관심은 납세 신고서가 아닌 개별 주체 또는 가구에 집중된다. 특히 결혼율과 같은 인구 통계의 변화는 측정 단위를 변화시킬 수 있는데, 이 경우도 마찬가지로 소득 분포 최상위와 최하위에서 차등적으로 변화한다. 이러한 변화는 집계되는 최상위 소득 점유율에 유의미한 영향을 줄 수 있다. 이러한 문제들은 모두 해결 가능한 것이지만, 연구자들은 이러한 이슈를 인식하고 다룰 수 있도록 열린 자세를 갖추어야 한다.

행정 데이터 사용 시의 도전 과제 중 하나는 납세자(tax filer)를 위한 인구 통계학적 데이터와 같은 보조 데이터가 부족한 경우가 많다는 점이다. 데이터 세트 전반에 걸쳐 행정 데이터를 이어 주면 보조 데이터가 부족한 경우와 같은 문제를 처리하는 데 도움이 될 수 있다. 많은 북유럽 국가들(특히 노르웨이, 핀란드, 스웨덴, 덴마크)은 이러한 링크드(linked) 데이터 세트를 구축해 놓았으며(United Nations, 2007), 포괄적 링크드 데이터 세트 개발에서 미국보다 앞서 있다. 데이터 연결 이슈에 관해서는 다음 장에서 논한다.

6. 행정 및 서베이 데이터 연결

1996년 사이언스지에 게재된 필립 스미스와 바버라 보일 토리의 선견지명이 있는 논문 「행동 및 사회과학의 미래(The Future of the Behavioral and Social Sciences)」에서 이들이 제시한 첫 번째 도전 과제는 "현재의 데이터 세트 통합"이다.

행정 데이터를 수집 및 통합하고 이용을 제공하는 중앙기관으로서 많은 국가의 합리적인 모델로 꼽히는 곳이 덴마크 통계청(Statistics Denmark)이다. 그러나 여러 가지 이유로 이 모델은 미국에 매력적인 모델이 아니다(Card et al., 2010). 첫째, 미국 정부는 다양한 기관이 연방정부, 주정부, 지방정부의 3단계로 분권화되어 있다. 둘째, 공유 및 개인정보와 관련하여 각기 다른 기관이 상이한 법에 의해 보장받고 있으며, 중앙집권화하는 데는 상당히 어려운 입법 조치가 요구된다. 셋째, 미국은 수년간 오랜 전통으로 중앙집권 정부, 특히 단일 기관에 권력이 집중되는 것에 대한 불신이 있으며, 행정 데이터 공유를 위한 개발은 당연히 이런 전통을 존중해야 한다. 분권화된 미국이 취한 접근법에 따른 결과의 예로 개인정보 문제는 중앙집권식의 커뮤니케이션 전략이 아닌 임시 기반으로 처리된다.

미국에서는 70개 이상의 연방정부 기관과 훨씬 더 많은 수의 민간 기관에서 서베이 및 행정 데이터를 수집했다. 그러나 각 서베이는 자체 목적과 특정 집단을 위해 수집되는 경우가 많다. 이런 데이터 소스들은 대부분 직접 비교나 연결이 불가능하다. 학계의 관점에서 연구자들은 정부 및 제3기관에서 기록 연계 및 통계적 매칭을 통해 취합한 거대한 양의 마이크로 데이터를 중요한 연구 문제를 해결하는 데 재사용하고 싶어 한다. 정부의 관점에서 공공 정책 입안자 및 정부 관료들은 프로그램 평가와 정책 설계 및 결정을 개선해야 한다는 부담을 받는다.

미국의 민간 연구 및 연방 기관은 상이한 행정 데이터 자료와 타(他) 사회 서베이를 연계하는 동일한 형식의 공공 데이터베이스를 구축하기 시작했다. 이러한 연구 노력 중 미국 국립연구회의(NRC: National Research Council)와 인구조사국의 위탁을 받은 일단의 사회학자 및 경제학자들이 진두지휘한 미국의 '기회연구'(AOS: American Opportunity Study) 프로젝트를 통해 1960~2010년 사이의 인구조사와 미국 '지역사회 서베이'(ACS: American Community Survey)를 연결했다(Grusky et al., 2019). 미시간대학교의 사회과학자 그룹이 구축한 '종단적, 세대 간 가족의 전자 마이크로-데이터베이스 프로젝트(LIFE-M: Longitudinal, Intergenerational Family Electronic Micro-Database Project)'는 출생, 사망, 결혼 기록과 1940년 인구조사를 합친 것이다(Bailey et al., 2019). 미네소타 인구센터의 공공 이용 통합 마이크로-데이터 시리즈(IPUMS: Minnesota Population Center Integrated Public Use Miccro-data Series)는 세계 각국으로부터 방대한 인구 데이터 집합을 수집하고 배포했다(Ruggles et al., 2015). 이러한 데이터 세트의 연계로 변수의 수와 표본의 시간 범위가 상당히 확대되어 통계 분석을 위한 신뢰할 만한 데이터가 더욱 많이 생성되었다.

연방 차원에서 이러한 수요를 해결하기 위해 미국 인구조사국은 인구조사 및 서베이 데이터 수집 작업 평가, 데이터 품질 및 안전 개선, 여러 데이터 소스의 정보 결합을 통해 단일 데이터 자료로 얻을 수 없는 연구 및 정책 활용의 목적으로 새로운 결과를 생성하는 행정 기록 연구 및 응용 센터(CARRA: Center for Administrative Records Research and Application)를 설립하였다. 행정 기록 연구 및 응용 센터(CARRA)는 많은 유럽 국가의 정부에서 오랫동안 적용한 데이터 공유 방식에 따라 데이터 세트에 있는 개인의 기록을 연결하는 매칭 소프트웨어를 개발했다. 인구조사 종단 인프라 프로젝트(CLIP: Census Longitudinal Infrastructure Project)를 사용한 연계절차와 데이터 프라이버시 이슈는 다음과 같다.

인구조사 종단 인프라 프로젝트(CLIP)는 미국 인구조사국[1]이 수집한 10년간의 인구조사, 서베이, 행정 기록들을 가지고 링크드 데이터 파일 세트를 만들었다. 연결 시스템은 보호식별키(PIK: Protected Identification Keys)를 이용한다. 이것은 인구조사국 기록의 특성을 사회보장국(SSA: Social Security Administration)의 NUMIDENT(숫자 식별 시스템) 파일 및 기타 행정 데이터로 구성된 참고 파일 내 기록의 특성과 비교하는 확률 매칭 알고리즘인 개인 신원 확인 시스템(PVS: Person

1) https://www.census.gov/about/adrm/linkage/projects/clip.html

Identification Validation System)에 의해 부여되었다. 이러한 변수들은 사회보장번호(SSN: Social Security Numbers), 이름, 출생일, 주소, 부모 성명 등을 포함하며, 데이터 안의 이용 가능한 정보에 따라 달라진다. 각 보호식별키(PIK)는 특정 개인을 고유하게 식별한다. 따라서 연구자들은 보호식별키(PIK)가 부여된 다양한 데이터 소스 전반에 걸쳐 개인을 찾아낼 수 있다. Ferrie, Massey와 Rothbaum(2016), Massey 외(2018)는 개인 신원 확인 시스템(PVS)에 기반한 보호식별키(PIK)를 이용한 데이터 연결, 연결된 표본 대표성, 편중 가능성에 대해 상세한 내용을 체계적으로 기록했다.

서베이 데이터는 여전히 사회과학 연구의 중추이기는 하지만 이전보다 더 잘 통합되고 있다. 최근에는 데이터 세트 전반의 개인 기록들을 연결하거나 서베이를 기타 공공 행정 데이터 소스로 연결하는 알고리즘 및 매칭 소프트웨어의 개발이 이루어졌다. 예를 들어, 사회보장국은 인구조사국과 협력하여 사회보장 혜택 및 수익 데이터를 연결하고, 국세청(IRS) 소득 파일을 두 개의 주요 서베이인 현재 인구 서베이(CPS) 및 소득 및 프로그램 참여 서베이(SIPP: Survey of Income and Program Participation)와 연결했다(McNabb et al., 2009). 미국에서 가장 오래된 가구 서베이인 미국패널소득조사(PSID: Panel Study of Income Dynamics)는 자체 서베이 데이터를 국가보건통계청(National Center for Health Statistics)과 메디케어·메디케이드 서비스 센터가 보유한 사망 및 보건 데이터, 미국 주택도시개발부의 주택 보조금 데이터, 국립교육통계센터(National Center for Education Statistics)의 학업 성취도 데이터 등 외부 데이터와 연결했다(McGonagle et al., 2012). 이 프로젝트는 미국 인구조사를 여러 다른 주요 연령 및 생애 과정 서베이에 연결하기 위한 광범위한 노력의 일환으로, 후자에 포함되는 서베이로는 건강과 은퇴 연구(HRS: Health and Retirement Study), 위스콘신 종단 연구(WLS: Wisconsin Longitudinal Study), 국가 사회생활·건강·노화 프로젝트(NSHAP: National Social Life, Health, and Aging Project), 국가 건강 및 노화 트렌드 연구(NHATS: National Health and Aging Trends Study) 등이 있다(Warren et al., 2020). 이러한 데이터 연결 인프라는 첨단 사회과학 연구 발전의 잠재력을 가지고 있을 뿐 아니라 정책 설계 및 평가를 용이하게 한다.

7. 기타 빅데이터

앞서 언급한 바와 같이, 행정 데이터는 사회과학 및 정책 연구에 대변혁을 일으킨 빅데이터의 한 종류이다. 인터넷과 컴퓨터 작업의 증가를 통해 상업적·사회적 거래의 부산물로 만

들어진 다양한 종류의 빅데이터가 - 학술 연구에 활용할 수 있도록 - 데이터 자료를 풍부하게 만든다. 잘 알려진 유기적 데이터의 예로는 구글 검색 결과, 소셜미디어 메시지, 교통 카메라 기록, 휴대전화 위치 추적, 인터넷 사이트 방문 이력 등이 있으며 이 외에도 문자, 사진, 오디오, 비디오 형태의 다양한 디지털 데이터 등이 있다. 유기적 빅데이터가 이 글의 주된 관심사는 아니지만 여러 이유로 언급할 가치가 있다. 유기적 빅데이터는 비즈니스를 변화시켜 오고 있으며 앞으로도 그러할 것이고, 가치 있는 연구 통찰력을 제공할 것이다.

행정 데이터가 가진 많은 장점은 유기적 빅데이터에도 동일하게 해당된다. 고품질 정보(적절히 사용되는 경우)를 포함하고 있는 대형 표본이 장점 중 하나이다. 이 외 부가적 혜택과 도전과제들도 있다. 데이터는 생성 빈도가 높고 실시간으로 생성되는 경우가 많아 기존의 사회과학 연구에서는 상상조차 어려울 정도로 표본 규모가 크고, 새로운 전산 및 통계 기술을 필요로 한다. 데이터는 기존의 횡단면 또는 패널 데이터보다 덜 구조적(또는 더 복잡한 구조)인 경우가 많아 데이터 저장과 검색에 새로운 접근법이 요구된다.

이러한 유기적 데이터가 개인적인 수준에서 행정 데이터 및 서베이 데이터와 연계되는 것이 어려울 수는 있지만, 인근 지역, 특정 영역, 지역적으로 접근이 어려운 곳에 대한 종합적인 정보를 제공할 수 있다. 예를 들어 Alexander, Polimis와 Zagheni(2020)의 연구에서는 소셜 미디어와 서베이 데이터를 연계하여 미국의 이주 유입 현황을 보여 준다. Basellini 외(2020)의 연구에서는 국가통계에서의 코로나바이러스감염증-19 사망 자료와 구글의 이동 데이터 연계에 대해 설명한다. 유기적 빅데이터와 행정 데이터 연계에 대한 장점과 문제들이 많지만, 유기적 데이터의 창의적인 활용은 오래된 궁금증으로부터 새로운 장을 열고 과학적 연구에 대한 지평을 넓힐 것이다.

8. 나가며

행정 데이터의 새로운 소스들은 장래성이 있지만, 기존 서베이들을 대체하기보다는 보완하는 역할을 한다. 데이터 혁명은 사회 및 정책 연구에 새로운 도전 과제를 던져 준다. 미래의 연구는 공동 프로토콜과 방법론을 개발하고, 지식 축적을 장려하고, 수학·통계학·컴퓨터 과학의 도구들을 하나로 모으고, 이론·알고리즘·메커니즘·실습 등을 활용하고, 사회적·법적·윤리적 고려 사항에 대처함으로써 서로 다른 데이터 소스들을 더욱 효율적으로 통합하는

데 집중해야 할 것이다. 행정 마이크로데이터는 경제 발전, 결혼 및 가족의 변화 추이, 도시화, 국내 및 국제 이주, 노화 및 국민 건강 등의 기본적인 경향을 이해하는 데 필수적이다. 여러 나라에서도 공공데이터 연계 프로젝트들이 시도되고 있으며, 이는 사회과학 연구의 범위를 확장하고, 시급한 사회문제들을 해결하는 정책 개발에 기회를 제공하고 있다.

참고문헌

Alexander, M., Polimis, K., & Zagheni, E. (2020). Combining social media and survey data to nowcast migrant stocks in the United States. arXiv preprint arXiv:2003.02895.

Auten, G., & Splinter, D. (2018). Income inequality in the United States: Using tax data to measure long-term trends. Washington, DC: Joint Committee on Taxation.

Bailey, M., Cole, C., Henderson, M., & Massey, C. (2017). How Well Do Automated Linking Methods Perform? Lessons from US Historical Data (No. w24019). National Bureau of Economic Research.

Basellini, U., Alburez-Gutierrez, D., Del Fava, E., Perrotta, D., Bonetti, M., Camarda, C. G., & Zagheni, E. (2020). Linking excess mortality to Google mobility data during the COVID-19 pandemic in England and Wales.

Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). Expanding access to administrative data for research in the United States. American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas.

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. Social science research, 59, 1-12.

Ferrie, J., Massey, C., & Rothbaum, J. (2016). Do grandparents and great-grandparents matter? Multigenerational mobility in the US, 1910-2013 (No. w22635). National Bureau of Economic Research.

Goerge, R. M., & Lee, B. J. (2002). Matching and cleaning administrative data. New Zealand Economic Papers, 36(1), 63-64.

Groen, J. A. (2012). Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. Journal of Official Statistics (JOS), 28(2).

Groves, R. M. (2004). Survey errors and survey costs (Vol. 536). John Wiley & Sons.

Groves, R. M. (2011). Designed Data and Organic Data. Technical Report, U.S. Census Bureau. Library Catalog.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). Survey methodology (Vol. 561). John Wiley & Sons.

Grusky, D. B., Hout, M., Smeeding, T. M., & Snipp, C. M. (2019). The American Opportunity Study: A New Infrastructure for Monitoring Outcomes, Evaluating Policy, and Advancing Basic Science. RSF: The Russell Sage Foundation Journal of the Social Sciences, 5(2), 20-39.

Guvenen, F., & Kaplan, G. (2017). Top income inequality in the 21st century: Some cautionary notes (No. w23321). National Bureau of Economic Research.

Massey, C. G., Genadek, K. R., Alexander, J. T., Gardner, T. K., & O'Hara, A. (2018). Linking the 1940 US Census with modern data. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 51(4), 246-257.

McGonagle, K. A., Schoeni, R. F., Sastry, N., & Freedman, V. A. (2012). The Panel Study of Income Dynamics: Overview, recent innovations, and potential for life course research. Longitudinal and life course studies, 3(2).

McNabb, J., Timmons, D., Song, J., & Puckett, C. (2009). Uses of administrative data at the Social Security Administration. Soc. Sec. Bull., 69, 75.

Meyer, B. D., & Mittag, N. (2019). Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness, and holes in the safety net. American Economic Journal: Applied Economics, 11(2), 176-204.

OMB. 2020. "Request for Comment on Considerations for Additional Measures of Poverty." Regulations.Gov. April 14, 2020. https://www.regulations.gov/docket?D=OMB-2019-0007.

Piketty, T., & Saez, E. (2003). Income inequality in the United States, 1913?1998. The Quarterly journal of economics, 118(1), 1-41.

Ruggles, S. (2014). Big microdata for population research. Demography, 51(1), 287-297.

Ruggles, S., McCaa, R., Sobek, M., & Cleveland, L. (2015). The IPUMS collaboration: Integrating and disseminating the world's population microdata. Journal of demographic economics, 81(2), 203-216.

Slemrod, J. (2016). Caveats to the research use of tax-return administrative data. National Tax Journal, 69(4), 1003.

United Nations. Economic Commission for Europe. (2007). Register-based statistics in the Nordic countries: review of best practices with focus on population and social statistics. United Nations Publications.

Warren, J. R., Pfeffer, F. T., Helgertz, J., & Xu, D. (2020). Linking 1940 US Census Data to the Panel Study of Income Dynamics: Technical Documentation.