기초연구 2018-03

문화·체육·관광 데이터 연계를 통한 빅데이터 생산 및 활용방안 연구

박근화

한국문화관광연구원
Korea Culture & Tourism Institute

연구책임 박 근 화(한국문화관광연구원 수석전문위원)
공동연구 장　　훈(한국문화관광연구원 부연구위원)
한 정 임(한국문화관광연구원 차석전문원)
김 정 림(한국문화관광연구원 차석전문원)
외부연구 김 희 경(동국대학교 연구초빙교수)
김 광 섭(동국대학교)

서 문

4차 산업혁명의 시대에 들어선 지금, 사회는 인공지능, 사물인터넷, 빅데이터, 로봇산업 등에 주목하고 있으며, 이러한 첨단산업의 발전으로 사회는 물론 우리의 삶도 빠르게 변화하고 있습니다. 그리고 사회의 발전과 더불어 개인의 삶과 가정의 중요성이 확대되고 있어, '여가생활보장'과 '일과 가정의 양립'이 가능하도록 법률 제정은 물론 다양한 정책적 지원이 제공되고 있습니다.

이러한 정책적 지원을 위해서는 다양한 분야에서 많은 정책연구와 함께 정책과제를 분석하고 연구하기 위한 통계자료가 필요하지만, 데이터는 여전히 부족합니다. 정보의 시대에서 21세기의 중요 자원인 데이터를 생산 및 활용하는 방안마련이 무엇보다도 중요한 시기라 할 수 있습니다.

본 연구는 빅데이터의 요소 중에서 가치(Value)의 입장에서 데이터를 활용을 증대할 수 있는 방안으로 서로 다른 데이터를 연계하여 하나의 파일로 만들어 분석함으로써, 새로운 가치있는 정보를 산출할 수 있는 방안을 제시하였습니다. 데이터 연계에 대한 연구는 최근에 일부 진행되었으나, 국내에서는 아직 미개척 분야라 할 수 있습니다. 이 연구를 활용하여 데이터 연계를 통한 다양한 정보를 산출하고 의미 있는 연구를 진행하는데 도움이 되기를 바랍니다.

본 연구 수행에 도움을 전문가 여러분과 연구원 등의 관계자 분들에게 깊은 감사의 말씀을 드립니다.

2018년 12월

한국문화관광연구원

원 장 **김 정 만**

연구개요

1. 서론

가. 연구 배경 및 목적

1) 연구 배경

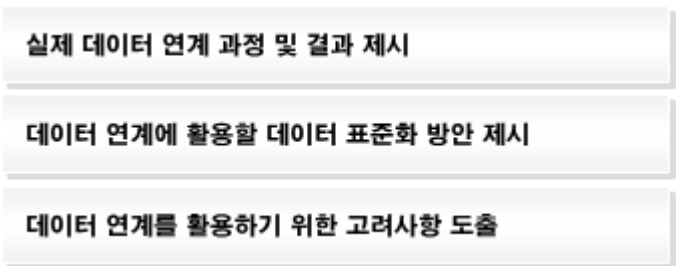
□ 현재 생산되고 있는 문화·체육·관광 분야의 통계를 이용하여 다양한 정보를 활용하고자 하지만, 실제로 활용할 수 있는 데이터가 충분하지 않을 뿐만 아니라 생산된 통계의 활용성도 부족함

○ 기존에 생산된 데이터를 서로 연계하여 활용이 가능하다면 새로운 정보의 생산 및 분석에 투입되는 예산은 물론 시간, 인력을 절감할 수 있음

○ 문화·체육·관광 분야에서도 데이터를 연계하여 데이터 활용도를 높일 수 있는 방안 마련이 필요하며, 문화·체육·관광 분야들 간의 데이터만이 아닌 다른 분야의 데이터도 연계할 수 있다면 매우 좋은 정보를 생산하여 활용할 수 있을 것임

2) 연구 목적

□ 본 연구의 목적은 다양한 데이터를 연계하여 새로운 가치 있는 정보를 제공한다는 개념에서의 (빅)데이터 생산 방법을 제시하고, 실제 데이터 연계를 통해 연계하기 전에는 분석할 수 없었던 새로운 정보를 제공하는 것임



[그림 1] 연구 목적

나. 연구 범위 및 방법

1) 연구 범위

□ 연구의 대상적 범위는 연계의 기준 자료인 문화 · 체육 · 관광관련 통계자료와 연계할 대상이 되는 연계자료 즉, 문화 · 체육 · 관광 관련 분야 및 다른 분야의 행정 자료 또는 조사 자료임

□ 연구의 내용적 범위는 데이터 연계 방안 및 절차와 연계된 데이터의 활용 방안 도출임

○ 데이터 연계 방안 및 절차에서는 데이터 연계에 대한 구분, 데이터 연계에서 사용하는 유사성 측도, 데이터 연계 방법, 데이터 연계에 대한 타당성 검증에 대한 내용을 다룸

○ 연계된 데이터의 활용방안에서는 실제 두 개 이상의 다른 데이터 파일에 데이터 연계 방법을 적용하여 분석하고, 데이터 연계를 활용하기 위해 고려해야 할 사항으로 개인정보문제와 공통 변수를 구성하는 방안을 제시함

2) 연구 체계

□ 연구는 총 7개 장으로 구성됨

□ 국내 · 외 데이터 연계 사례분석, 문화 · 체육 · 관광 관련 데이터 현황 분석, 전문가 자문회의 등을 통해 문화 · 체육 · 관광 관련 데이터 연계의 활용성과 전문성을 확보함

○ 국내·외 데이터 연계 사례분석을 통한 문화·체육·관광 데이터 연계 방안 도출

○ 문화·체육·관광 관련 데이터 현황 분석을 통해 데이터 연계에 활용 가능한 데이터를 파악하고 데이터 연계를 위한 데이터 표준화 방안

과 데이터 연계를 위한 공통문항 개선방안 도출

○ 실제 데이터 연계를 수행하고 있거나, 본 연구에서 활용한 데이터 생산 기관의 담당자와 데이터 연계 방법 및 연계 결과에 대한 검증을 위한 통계학자 등으로 구성된 전문가 자문으로 문화·체육·관광 데이터 연계에 대한 전문성 확보

2. 데이터 연계 개념

가. 데이터 연계란

□ 데이터 연계(data linkage)는 서로 개별적으로 생산되어 별개의 파일(또는 데이터베이스(DB))로 존재하는 데이터를 하나로 연결하여 통합 파일을 만듦으로써 새로운 정보를 활용할 수 있도록 하는 방법임

○ 데이터 연계는 데이터 연결 기준 따라 크게 5가지로 구분할 수 있으며, 정확 연계, 판단 연계, 확률적 연계, 통계적 연계, 데이터 연결임

○ 본 연구에서는 이러한 5가지 연계 방법 중에서 연구의 목적, 연계 방법의 유사성 등을 고려하여 정확 연계와 통계적 연계만을 고려함

〈표 1〉 연결 기준에 따른 데이터 연계 구분

구분	정의	연결 기준
정확 연계	두 데이터에 있는 동일한 고유식별정보를 이용하여 연결	고유식별정보
판단 연계	연구자의 데이터에 대한 이해를 바탕으로 연결	연구자의 지식
확률적 연계	정확 연계에서 일치하지 않는 케이스들이 발생할 때, 각 변수들의 연결 가능성을 계산하여 연결	공통 변수로 계산한 가능성
통계적 연계	통계적 방법으로 가장 유사한 케이스를 찾아 연결	유사성 측도
데이터 연결	둘 이상의 파일에서 변수들간의 연관성이 있을 경우 하나가 변화할 때 같이 변화가 가능하도록 연결	변수들 간의 연계성

출처 : 이영섭 외 4인(2009), 통계조사 자료와 행정 자료간의 통계적 연계 기법에 관한 연구

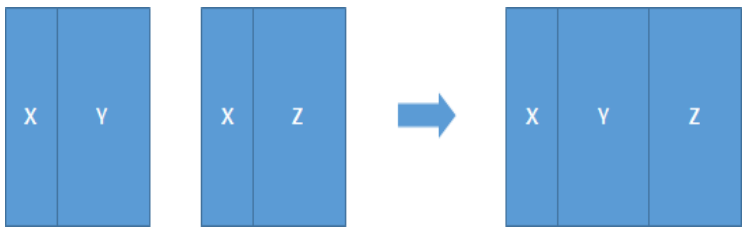
나. 정확 연계 방법

□ 정확 연계는 서로 다른 데이터에서 동일한 대상을 찾아 연결하여 통합 파일을 만드는 방법임

○ 정확하면서도 매우 효율이 좋은 방법으로 동일한 대상을 찾아 연결하기 때문에 각 개체를 정확하게 파악할 수 있는 고유식별이 가능한 변수가 있어야 함

○ 고유식별이 가능한 변수를 이용하기 때문에 개인정보보호 문제로 일반적으로 활용하기 어려움

○ 정확 연계에 대한 평가는 기준 파일의 데이터가 얼마나 연계 파일과 연계되었는가의 연계율로 판단함



[그림 2] 정확 연계의 형태

다. 통계적 연계 방법

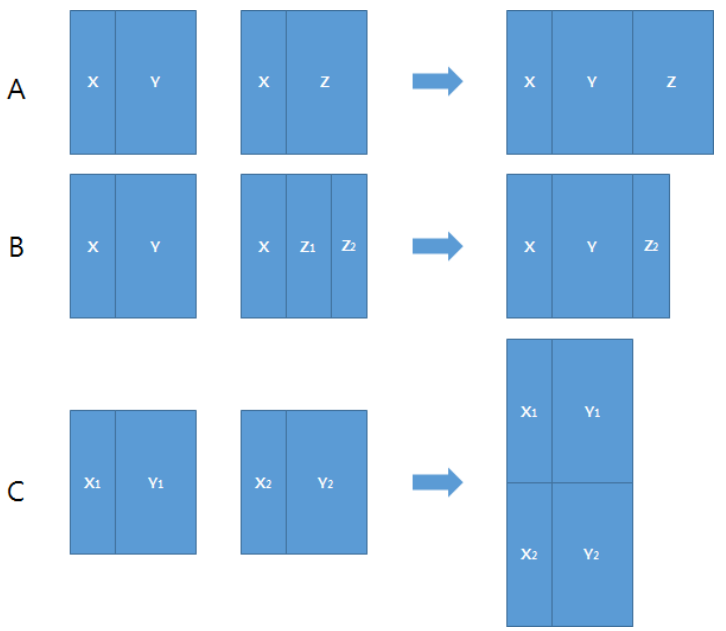
□ 통계적 연계는 서로 다른 데이터 파일에서 유사한 성향을 가진 데이터를 연결하는 방법임

○ 우리가 일반적으로 접하는 개인식별정보가 없는 데이터, 조사 자료 등에서 추가적으로 알고 싶은 정보가 있을 때 적용하여 해당 정보를 얻을 수 있음

○ 통계적 연계는 유사한 데이터를 연계하는 것이기 때문에 정확 연계

보다 정확성이 부족하고 사용하기 위한 조건이 까다로우며, 통계적 연계를 적용위한 기본 가정을 충족해야 함

○ 통계적 연계는 정확 연계처럼 동일한 대상을 연계하는 것이 아니라 유사한 성향을 지닌 대상들끼리 연계하는 것이기 때문에 데이터 연계 후의 생성된 공통 파일에 대한 평가가 중요하며, 통계적 연계에서의 평가는 정확성(accuracy), 예측성(predictability)과 대표성(representation)의 문제로 생각할 수 있음



[그림 3] 통계적 연계의 형태

3. 사례분석

가. 사례분석 개요

○ 문화·체육·관광분야 데이터를 연계하기에 앞서 타 분야에서 데이터 연계에 활용되는 데이터와 연계 방법은 무엇인지 그리고 다른 나라

에서는 어떠한 방향으로 데이터 연계를 진행하고 있는지 살펴보고 각 사례별 주요사항을 도출하고자 함

〈표 2〉 사례분석 방법

국내사례	해외사례	
데이터 연계 연구, 활용	미국, 호주	영국, 미국, 캐나다, 뉴질랜드
해당 연구 개요 검토	해당 연구 개요 검토	연계 수행 기관 및 환경 개요 검토
데이터 연계 방법 검토	데이터 연계 방법 검토	데이터 연계 방법 검토
기준 데이터, 연계 데이터의 특성 검토	기준 데이터, 연계 데이터의 특성 검토	데이터 접근 방법 검토

나. 시사점

○ 통계적 연계 방법은 정확한 연계는 아니지만 데이터 연계를 활용하는데 효율적임. 단, 통계적 연계 방법은 기본 가정이 만족하는지를 반드시 평가하고 만족할 경우에만 사용해야 함

○ 정확 연계는 매우 좋은 방법이나, 연계율이 낮으면 효율성이 낮음. 그리고 정확 연계에서 사용하는 고유식별정보가 있는 경우 데이터에 대한 보안의 문제가 발생하기 때문에 이에 대한 다양한 정책을 마련해야 함

○ 해외에서는 데이터의 정보에 따라 다양한 방법으로 데이터를 연계하고 있는데, 이러한 방법에 대한 연구가 필요함

○ 데이터 활용을 높이기 위해서는 데이터에 접근하는 다양한 방법을 마련하고 보안이나 데이터의 활용 정도 등에 따라 제공하는 데이터의 범위를 달리할 필요가 있음

4. 데이터 연계에서 데이터 이해

가. 연계에 사용할 문화·체육·관광 데이터의 이해

□ 본 연구에서 통계적 연계에 사용할 기준 데이터는 데이터의 정확성, 신뢰성을 고려하여 문화 · 체육 · 관광 관련 국가승인통계 총 25종 중에서 조사통계 17종으로 함

○ 통계적 연계는 공통 변수가 있어야 가능하기 때문에 조사통계를 활용함

○ 보고통계는 특정한 목적에 맞는 정보를 입력시스템을 통해 축적된 데이터이기 때문에 공통 변수가 충분하지 않거나 보편적으로 이용하기에 적절하지 않은 통계가 대부분이며, 가공통계는 모두 고유식별정보가 있는 통계를 가공한 것으로 엄밀하게 정확 연계에 해당하는 통계임

〈표 3〉 문화·체육·관광 관련 국가승인통계 현황

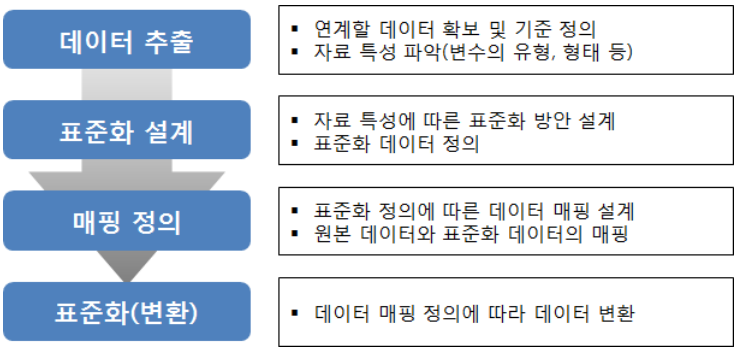
기관	조사통계	보고통계	가공통계
문화체육관광부	16	4	0
그 외 관련 기관	1	2	3
합계	17	6	3

나. 데이터 표준화

□ 데이터 표준화는 연계에 활용할 데이터의 공통 변수를 동일한 기준으로 맞추는 것으로 통계적 연계에서는 공통 변수가 많을수록 다양한 조합을 통해 유사한 성향을 가진 케이스들을 연계할 수 있기 때문에 실제 데이터 연계만큼이나 중요한 과정임

○ 문화·체육·관광 관련 국가승인통계 중에서 국민 대상의 조사통계 5종을 살펴본 결과, 연계에 활용한 공통 변수들이 동일한 기준으로 작성되어 있지 않음

○ 국민을 대상으로 하는 문화·체육·관광 관련 국가승인 조사통계를 이용하여 데이터 연계를 위해서는 반드시 표준화 과정이 필요하며, 향후 데이터 생산 시에도 데이터 활용도를 감안하여 국민 대상 조사통계의 공통 변수인 인구통계학적 변수에 대한 표준화를 고려해야 함



[그림 4] 데이터 표준화 절차

5. 데이터 연계 활용

□ 문화 · 체육 · 관광 관련 데이터를 기준 데이터로 하여 다른 분야의 데이터와 실제 연계를 진행한 연계 과정과 함께 분석 결과를 제시함

○ 정확 연계에서는 카드 데이터와 기상청 데이터 연계를 통한 기후에 따른 이동지출 분석과 문화체육관광 분야 사업체 표본틀과 통계청 행정 데이터 연계를 통한 문화체육관광산업 경영활동 현황 분석을 진행함

○ 통계적 연계에서는 국민여가활동조사와 문화향수실태조사의 데이터 연계를 통한 문화적 경험과 여가활동의 상관관계 분석과 국민여가활동조사와 한국의료패널조사 데이터 연계를 통한 여가유형별 경험에 따른 의료비 지출 분석을 진행함

〈표 4〉 데이터 연계 활용 개요

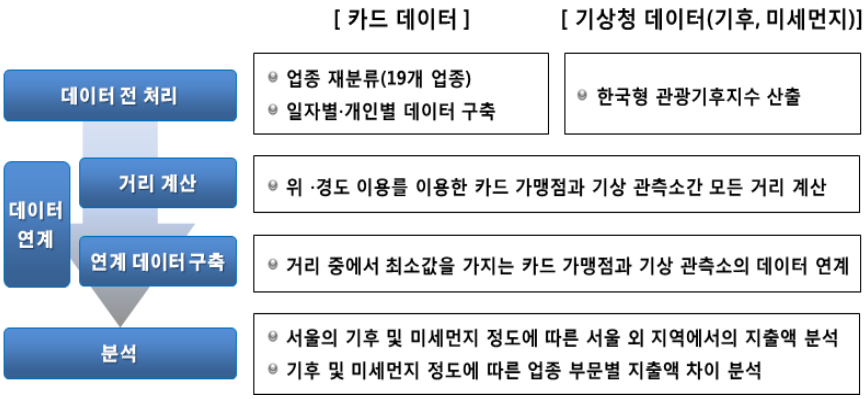
연계 방법	연계 변수	데이터 특성		분석
		기준 데이터	연계 데이터	
정확 연계	위 · 경도	카드 데이터	기상청 데이터	기후에 따른 이동지출 분석
정확 연계	사업체고유번호	문화체육관광 분야 사업체 표본틀	통계청 행정 데이터	문화체육관광산업 경영활동 현황 분석
통계적 연계	연령, 지역, 학력, 동거가구원수, 혼인상태(배우자유무), 동거자녀수, 월평균본인소득	국민여가활동조사	문화향수실태조사	문화적 자본과 여가활동의 관계 분석
통계적 연계	연령, 지역, 학력, 동거가구원수, 혼인상태(배우자유무), 월평균가구소득	국민여가활동조사	한국의료패널조사	여가활동의 건강효과 분석

가. 실제자료를 이용한 정확 연계

□ 카드 데이터와 기상청 데이터 연계를 통한 기후에 따른 이동지출 분석

○ 카드 데이터와 기상청 데이터 연계를 통한 기후에 따른 이동지출 분석에서는 정확 연계 방법으로 카드 데이터와 기상청 데이터를 연계하여 통합 데이터를 구축한 후, 기상 상태(기후, 미세먼지)를 몇 단계의 등급으로 구분하고 구분된 기상 등급에 따른 여가 관련 신용카드 이동 지출의 차이를 비교·분석함

○ 데이터 연계 방법을 살펴보면, 카드 데이터의 가맹점 위·경도와 기후 데이터의 관측소 위·경도의 거리를 계산하고, 카드 가맹점별로 94개의 관측소 중에서 가장 짧은 거리의 관측소를 선정하여 해당 관측소의 기후 정보를 연계함

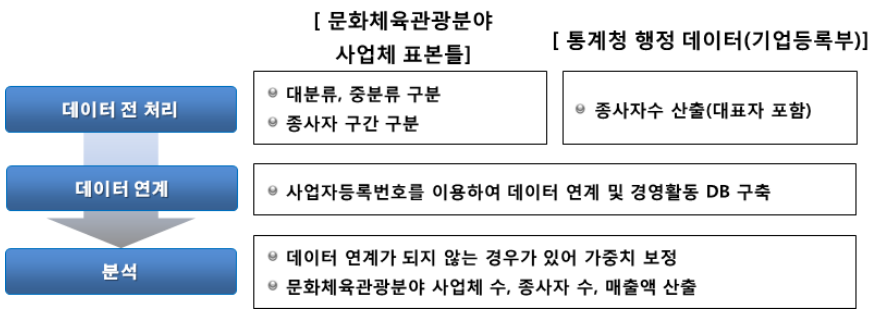


[그림 5] 카드 데이터와 기상청 데이터 연계

□ 문화체육관광 분야 사업체 표본틀과 통계청 행정 데이터 연계를 통한 문화체육관광산업 경영활동 현황 분석

○ 문화체육관광 분야 사업체 표본틀과 통계청 행정 데이터 연계를 통한 문화체육관광산업 경영활동 현황 분석에서는 실제로 문화체육관광산업통계(국가승인통계) 작성을 위해 수행한 정확 연계 방법을 살펴봄

○ 일반적으로 사업체 자료의 정확 연계에는 사업자등록번호를 이용하고 있으나, 문화체육관광 분야 사업체 표본틀에는 사업자등록번호 대신 사업체의 고유한 식별 변수로 사업체고유번호가 있어 이를 연계 변수로 사용하여 데이터를 연계함



[그림 6] 문화체육관광 분야 사업체 표본틀과 통계청 행정 데이터 연계

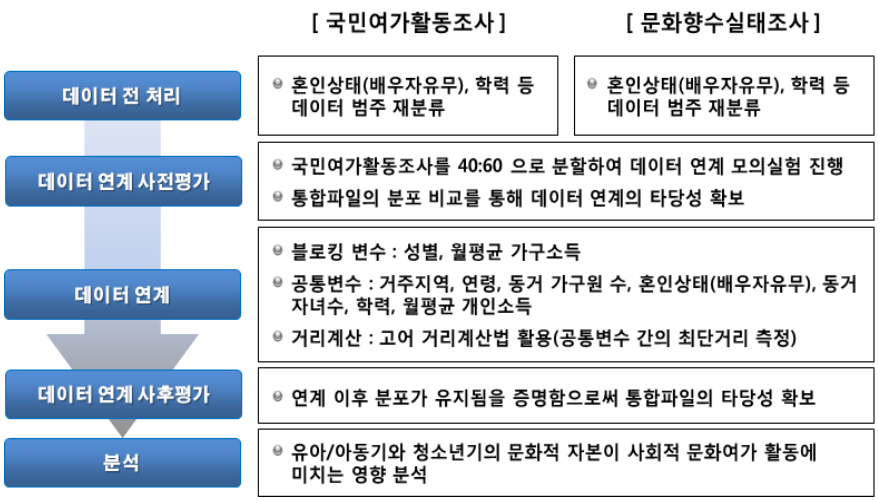
나. 실제자료를 이용한 통계적 연계

□ 국민여가활동조사와 문화향수실태조사를 이용한 문화적 자본과 여가활동의 관계 분석

○ 국민여가활동조사와 문화향수실태조사의 연계를 통해 유년기와 청소년기에 형성된 문화적 자본이 향후 여가활동에 미치는 영향을 파악하고자 함

○ 기준자료(국민여가활동조사)를 분할하여 데이터 연계 사전 평가를 통해 공통 변수의 분포가 유지됨을 보여줌으로써 데이터 연계 활용의 타당성을 확보함

○ 성별과 월평균 가구소득을 블로킹 변수로 설정하여 동일한 성별과 가구소득 구간 내에서 동거가구원수, 성별, 연령 등 공통 변수에 대한 최단거리 데이터를 연결하는 고어 거리계산법을 활용하여 데이터를 연계함



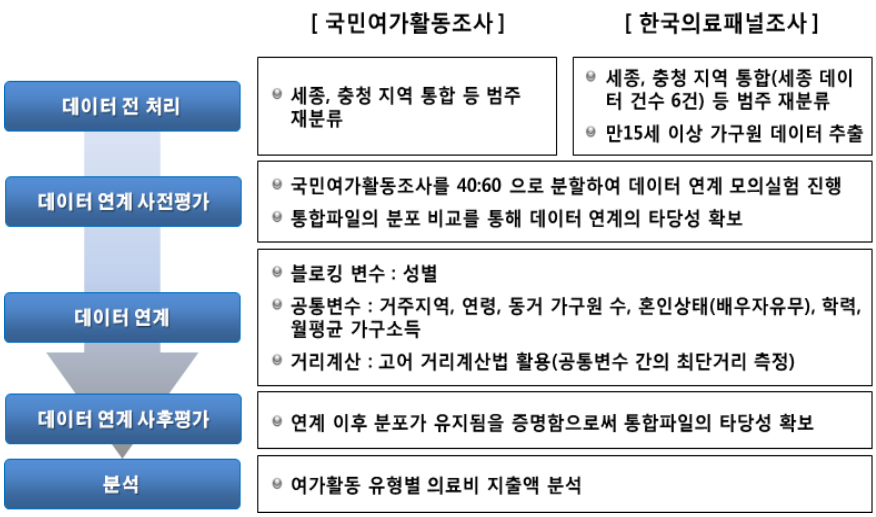
[그림 7] 국민여가활동조사와 문화향수실태조사 데이터 연계

□ 국민여가활동조사와 의료패널 연계를 통한 여가활동의 건강효과 분석

○ 국민의 여가활동과 의료비 지출의 관계를 파악하기 위하여 국민여가활동조사와 한국의료패널조사 데이터를 연계하여 여가활동 유형별 경험 여부에 따른 의료비 지출액이 차이에 대해 분석함

○ 한국의료패널조사 데이터에 개인식별 변수가 존재하나 국민여가활동조사에는 개인식별 변수가 존재하지 않기 때문에 정확 연계가 아닌 통계적 연계 기법을 활용함

○ 데이터 연계는 성별을 블로킹 변수로 설정하여 동일 성별 내에서 거주지역, 연령, 동거가구원수 등의 공통 변수 간의 최단거리 데이터를 연결하는 고어 거리계산법을 활용함



[그림 8] 국민여가활동조사와 한국의료패널조사 데이터 연계

6. 데이터 연계 활용을 위한 고려사항

□ 데이터 연계 활용을 위한 고려사항은 다음과 같음

[그림 9] 데이터 연계 활용을 위한 고려사항

가. 데이터 연계에서 개인정보보호

□ 데이터를 연계 또는 활용할 때에는 사전에 개인의 특성을 파악할 수 있는 정보를 제거하거나 개인정보보호기법 등을 활용하여 개인정보가 유출되지 않도록 조치를 취한 후 활용해야 함

○ 개인정보의 노출위험과 자료의 활용가치 사이의 적절한 합의점이 필요하며, 개인정보를 보호하면서 데이터를 연계할 수 있는 방법을 적용할 필요가 있음

○ 데이터 연계에서 활용할 수 있는 개인정보보호 기법으로는 기존의 식별 정보를 다르게 코딩하여 데이터 값을 변경하는 비식별화 기법 등을 적용할 수 있음

나. 문화·체육·관광 데이터 분류

□ 타 기관에서 생산하는 통계로부터 문화 · 체육 · 관광 관련 분야의 데이터를 추출하여 활용하기 위해 데이터를 분류할 수 있는 분류체계가 필요함

○ 문화·체육·관광 관련 분야 여부를 결정하는 것은 소비 또는 공급, 행위 등이며, 일반적으로 활동과 업종(산업)에 의해 구분됨. 특히 업종(산업) 관련 분류가 가장 많이 활용되고 있음

○ 문화체육관광부에서는 통계청의 표준산업분류를 기준으로 하여 문화·체육·관광 분야의 산업분류를 마련하여 활용하고 있으므로, 이를 기준으로 다양한 분류를 연계하여 활용함으로써 다른 통계의 데이터를 활용할 수 있는 방안을 마련할 수 있음

다. 데이터 연계를 위한 공통 변수

□ 통계적 연계는 공통 변수에서 최적의 조합을 찾고 유사성 척도의 연계 방법으로 통합파일을 생성하는 것이 중요하며, 이를 위해 표준화 과정이 필요함

○ 국민 대상의 조사통계 데이터를 연계하기 위해 공통 변수로 활용할 수 있는 응답자 현황에 대한 질문의 표준화 방안을 마련하고 제시함

○ 연계목적이 명확한 경우 목적변수를 선정하여 모형에 대한 예측값으로 유사성을 측정하는 방법도 있으며, 이 경우 목적 변수를 질문 항목에 추가함으로써 데이터 연계의 효율성을 높일 수 있음

7. 결론 및 제언

○ 문화·체육·관광 분야의 데이터를 하나로 모을 수 있는 제도적 장치와 데이터를 관리할 수 있는 물리적 공간(데이터베이스)이 마련하고, 문화체육관광부의 통계담당부서에서 통합적으로 관리할 수 있

도록 함으로써 데이터의 활용도를 높일 수 있음

○ 문화·체육·관광 관련 데이터를 구분하고 분류할 수 있는 체계를 마련하여 다양한 기관에서 생산되는 데이터로부터 문화·체육·관광 분야의 데이터를 추출할 수 있음. 이는 새로운 데이터를 생산하는 것과 동일한 효과를 가짐

○ 문화·체육·관광 분야의 데이터를 보유하고 있는 기관과 데이터 활용을 위한 MOU 체결을 통해 데이터의 정책적 활용도를 높일 수 있으며, 정확 연계 시의 개인정보보호 관련 문제를 해결하기 위한 데이터 보안 센터를 마련하는 등의 정책이 필요함

○ 통계적 연계 시 공통변수로 활용하기 위한 응답자 기본정보 관련 문항의 구성에 대한 표준방안을 마련함으로써 데이터의 활용도를 높일 수 있음

표 목차

그림 목차

제1장

서론

제1절

연구 배경 및 목적

1. 연구 배경

2016년 세계 경제 포럼(WEF: World Economic Forum)에서 '제4차 산업혁명'이란 용어를 사용하면서, 지금의 시대를 4차 산업혁명시대로 지칭하고 있다. 4차 산업혁명은 인공지능, 사물인터넷, 빅데이터, 클라우드 등 첨단 정보통신기술(IT: Information Technology)을 기반으로 다양한 분야들이 융합되어 혁신적인 변화를 일으키는 것을 의미한다. 그리고 이러한 변화는 모바일 등의 첨단 IT 환경의 급속한 발전과 더불어 빠르게 성장·진행되고 있다.

4차 산업혁명의 근간이 되는 기술 분야 중, 빅데이터(Big data)는 최근 몇 년 동안 사회에서 큰 주목을 받고 있는 분야이다. 빅데이터가 주목을 받은 이유는 사회 변화와 IT 발전으로 데이터가 기하급수적으로 증가했지만, 활용하지 못하고 쌓여만 가던 데이터를 분석하여 새로운 가치 있는 정보를 제공하기 때문이다. 즉, 빅데이터 분석을 통해 소량의 정형데이터(structured data)에서는 발견하지 못한 중요한 정보를 파악하고 새로운 가치 창출이 가능하다.

빅데이터 초창기에는 빅데이터를 구성하는 요소는 BI/DW 리서치 기관인 TDWI가 제시한 크기(Volume), 속도(Velocity), 다양성(Variety)으로 규정하였으며, 최근에는 빅데이터를 이용함으로써 얻게 되는 가치(Value)를 새로운 빅데이터의 추가 요소로 생각하게 되었다. 이렇게 빅데이터에서 가치의 중요성이 대두되면서 빅데이터에 대한 개념도 다양한 방면으로 인식하게 되었다. 전통적인 관점에서 빅데이터는 매우 용량이 큰 데이터를 의미하지만, 실제로 대용량의 관점에서 빅데이터를 가지고 있는 기관은 일부에 지나지 않는다.

대용량의 관점에서의 빅데이터는 대부분의 사업체나 이용자(연구자, 분석가 등)들은 쉽게 활용할 수 없다. 그러나 가치 창출의 개념에서의 빅데이터는 많은 사람들이 데이터의 크기에 상관없이 빅데이터를 활용할 수 있는데, 이를 가능하게 하는 방안이 서로 다른 데이터를 연계하여 분석하는 것이다. 실제로 빅데이터는 단지 큰(Big) 데이터를 의미하는 것이 아니라 다양한 형태의 다양한 종류의 데이터가 쌓여있는 대용량의 데이터를 의미한다. 즉, 단일 데이터로 큰 데이터를 의미하는 것이 아닌 다양한 종류의 데이터 집합체를 의미하는 것이다.

서로 다른 데이터를 결합 또는 연결하여 분석하는 것을 데이터 연계(data linkage)라 하며, 데이터 연계를 통해 데이터의 활용도를 넓힐 뿐만 아니라, 각각의 데이터에서 파악할 수 없었던 새로운 정보를 얻을 수 있게 된다. 최근에는 다양한 정보들을 모아 가치 있는 새로운 정보를 발견하기 위한 데이터 연계와 관련한 연구가 조금씩 진행되고 있다. 빅데이터에서는 다양한 종류의 데이터를 다양한 형태로 결합하여 분석하고 있으며, 데이터를 직접적으로 연결하는 데이터 연계는 빅데이터의 다양한 결합 방법 중에서 효율성이 매우 뛰어난 방법이라고 할 수 있다.

사회가 빠르게 변화하고 기술이 발전하면서 여가, 문화, 예술, 관광, 스포츠 등의 문화·체육·관광 분야에 대한 관심은 지속적으로 증가하고 있다. 문화·체육·관광 분야의 관심이 증가함에 따라 현황 파악 또는 연구를 위한 통계 자료의 요구도 함께 증가하고 있다. 따라서 문화·체육·관광 분야에 대한 통계를 산출하고 분석하여 정책이나 마케팅 등에 활용할 수 있는 정보를 제공하는 것은 매우 중요한 일이다. 그러나 문화·체육·관광 분야에 대한 통계나 데이터가 충분하지 않기 때문에 데이터 연계를 통해 가치 있는 정보를 찾아 제공하는 것이 필요하다.

현재 생산되고 있는 문화·체육·관광 분야의 통계를 이용하여 다양한 정보를 활용하고자 하지만, 실제로 활용할 수 있는 데이터가 충분하지 않을 뿐만 아니라 생산된 통계의 활용하는 정도도 부족하다. 즉, 생산되고

있는 통계들은 각각의 목적에 의해 생산되고 있지만, 통계를 생산한 목적 이외 관점에서의 활용도는 많이 부족하다. 또한 현재의 통계생산 방식으로는 기존에 생산된 통계와 관련된 새로운 정보가 필요한 경우, 이를 알 수 있는 방법은 새로운 통계를 다시 생산하는 방법밖에 없다. 즉, 추가적인 정보를 얻기 위해서는 새로운 통계를 생산해야 하는데 이를 위해서는 많은 예산과 시간, 인력이 투입 되어야 한다. 적은 정보라도 새로운 정보를 얻기 위해서는 많은 비용과 노력이 들어가기 때문에 새로운 통계를 생산하는 것은 쉬운 일이 아니다. 이러한 경우 기존에 생산된 데이터를 서로 연계하여 활용이 가능하다면 새로운 정보의 생산 및 분석에 투입되는 예산은 물론 시간, 인력을 절감할 수 있다.

다양한 형태의 데이터를 연계하기 위해서는 연계에 활용할 데이터에 대한 정보를 파악하고 있어야 한다. 우리나라와 같이 분산형 통계제도를 채택하고 있는 국가의 경우에는 통계의 생산뿐만 아니라 흩어져 있는 통계의 효율적 활용을 위한 데이터 연계는 더욱 중요하다. 최근에 UN, EU 등 국제기구에서도 응답부담 감소를 위해 행정 자료 활용을 적극 권장하고 있는데, 이는 행정 자료들 간에는 데이터의 연계가 가능하기 때문이다. 우리나라에서도 행정 자료 활용 방안을 마련하여 이를 제공함으로써 활용률이 증가하고 있다. 이러한 정책의 하나로 행정안전부에서 공공데이터 포털(https://www.data.go.kr/)을 통해 공공데이터를 수집 및 제공하고 있으며, 통계청 역시 통계청이 수집한 자료를 다양한 분야의 기관에서 활용 가능하도록 하고 있다. 따라서 문화·체육·관광 분야에서도 데이터를 연계하여 데이터 활용도를 높일 수 있는 방안 마련이 필요하며, 문화·체육·관광 분야들 간의 데이터만이 아닌 다른 분야의 데이터도 연계할 수 있다면 매우 좋은 정보를 생산하여 활용할 수 있을 것이다.

2. 연구 목적

본 연구의 최종 목적은 다양한 데이터를 연계하여 새로운 가치 있는 정보를 제공한다는 개념에서의 (빅)데이터[1] 생산 방법을 제시하고, 실제 데이터 연계를 통해 연계하기 전에는 분석할 수 없었던 새로운 정보를 제공하는 것이다. 이를 위한 세부 목적은 크게 다음의 세 가지로 제시할 수 있다.

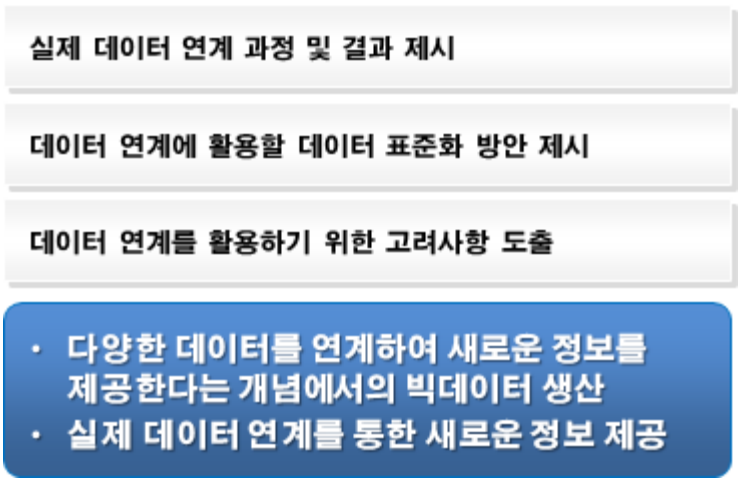
첫째, 데이터 연계에 대한 정의와 개념 그리고 기본적인 이론과 함께 실제 데이터를 이용하여 적용·활용·분석하는 과정, 결과를 제시함으로써 데이터 연계에 대한 이해를 돕는 것이다. 실제 데이터 이용에는 정확 연계와 통계적 연계 방법[2]을 적용하고, 이들의 구체적인 연계 과정과 각 과정에서의 고려사항 등을 살펴보도록 한다.

둘째, 데이터 연계에서 활용할 데이터에 대한 설명과 데이터 연계에서 공통 변수에 대한 데이터 표준화 방안 제시이다. 데이터 표준화는 서로 다른 두 데이터의 대상, 기준시점, 동일한 의미를 갖는 변수들의 기준을 맞추는 작업이다. 이는 기준을 맞추는 데이터의 대상과 시점 등을 조화시키는 것과 같은 의미를 가지는 변수들을 동일한 기준으로 조화시키는 것으로 구분할 수 있다.

셋째, 개인정보문제 해결 방안, 문화·체육·관광 관련 분류체계 마련, 문화체육관광부 국가승인통계 중 조사통계의 응답자 특성 문항들의 표준화 방안 등의 데이터 연계를 활용하기 위한 고려사항 도출이다. 개인정보 문제는 데이터 활용 시에 중요하게 고려되는 부분으로 특히, 데이터 연계에서는 더욱 민감한 사항이기 때문에 개인정보에 대한 정의와 데이터 연계에서 개인정보문제 해결 방안을 살펴본다. 그리고 데이터 연계를 위해

1) 가치의 기준에서는 빅데이터(Big data)라고 할 수 있으며, 다양한 연계를 통해 데이터가 커진다는 의미로서도 빅데이터라고 정의할 수 있다. 그러나 빅데이터라는 용어를 사용하지 않고 데이터를 생산하는 방안이라고 해도 문제되지 않는다.
2) 정확 연계와 통계적 연계는 제2장에서 설명하도록 한다.

서는 다양한 데이터를 활용하게 되는데, 문화·체육·관광 관련 내용만을 활용하기 위해서는 명확한 분류 기준이 필요하다. 따라서 한국표준산업분류를 기준으로 문화·체육·관광 관련 분류와 이와 관련된 분류를 연결함으로써 연계 기준 방안을 제시하고자 한다. 또한 데이터 연계에서 활용하는 데이터가 조사통계인 경우, 응답자 특성 항목 표준화 방안을 도출해 향후 데이터 연계 또는 분석에서의 활용성을 높이고자 한다.



[그림 1-1] 연구 목적

제2절

연구 범위 및 체계

1. 연구 범위

본 연구의 대상적 범위는 데이터 연계의 기준 자료인 문화·체육·관광 관련 통계자료와 연계할 대상이 되는 연계 자료 즉, 문화·체육·관광 관련 분야 및 다른 분야의 행정 자료[3] 또는 조사 자료이다. 이들 자료는 개인을 대상으로 하는 자료와 사업체를 대상으로 하는 자료 그리고 그 밖의 특수한 대상에 대한 자료 등으로 구분할 수 있다.

연구의 내용적 범위는 크게 두 가지로 구분 할 수 있으며, 첫 번째는 데이터 연계 방안 및 절차에 대한 것이고 다른 하나는 연계된 데이터의 활용 방안 도출이다. 데이터 연계 방안 및 절차에서는 데이터 연계에 대한 구분, 데이터 연계에서 사용하는 유사성 측도, 데이터 연계 방법, 데이터 연계에 대한 타당성 검증에 대한 내용을 구체적으로 다루도록 한다. 두 번째에 해당하는 데이터 연계의 활용 방안 도출은 세 가지로 나눌 수 있다. 하나는 실제 서로 다른 두 개 이상의 데이터 파일에 데이터 연계 방법을 적용하여 통합 파일을 만드는 것이고, 다른 하나는 통합 파일을 이용하여 알고 싶은 정보를 찾거나, 추가 분석을 통해 새로운 가치 있는 정보를 생성하는 것이다. 그리고 마지막으로 데이터 연계를 활용하기 위해 고려해야 할 사항으로 개인정보문제와 공통 변수를 구성하는 방안 등을 제시하는 것이다.

3) 행정 자료는 공공의 행정 자료와 민간의 행정 자료로 구분할 수 있는데, 대부분의 빅데이터는 민간의 행정 자료가 포함된다.

2. 연구체계

가. 연구 구성

서론에 이어 제2장에서는 데이터 연계에서 사용하는 연계 방법에 대해 살펴보도록 한다. 두 개의 데이터 파일을 하나로 합치는 데이터 연계 방법과 데이터 연계를 구분하는 방법을 설명하며, 데이터 연계를 위해 사용하는 유사성 측도와 데이터 연계를 적용하기 위한 기본 가정을 설명한 후 실제로 연계하는 과정을 제시한다.

제3장에서는 사례분석을 통해 데이터 연계에 활용되는 데이터와 연계 방법은 무엇인지 그리고 다른 나라에서는 어떠한 방향으로 데이터 연계를 진행하고 있는지를 살펴본다. 사례분석은 크게 국내사례와 해외사례로 구분하며, 국내사례는 데이터 연계와 관련한 연구를 중심으로, 해외사례는 데이터 연계가 활발하게 이루어지고 있는 국가를 중심으로 데이터 연계와 데이터 관련 정책 등의 사례를 살펴보고 분석한다.

제4장에서는 문화·체육·관광 관련 데이터 중에서 실제로 데이터 연계에 활용할 데이터 설명과 데이터 연계 전에 데이터 표준화 방안을 도출하고 실제로 연구에서 사용할 데이터의 연계 방법을 구체적으로 제시한다.

제5장에서는 제4장에서 살펴본 데이터를 연계하고 새로운 통합 파일을 만들어 새로운 정보 생성 과정과 그 결과를 제시한다. 여기서는 정확 연계와 통계적 연계 방법을 제시하며, 정확 연계는 외부데이터와 문화체육관광산업 표본틀, 통계적 연계는 문화체육관광부의 국가승인통계 중에서 국민을 대상으로 하는 조사통계 데이터를 기반으로 한다.

제6장에서는 데이터 연계를 활용하기 위해 고려해야 할 사항을 살펴보도록 한다. 이는 개인정보문제 해결 방안, 문화·체육·관광 관련 분류체계 마련 그리고 문화체육관광부 국가승인통계 중 조사통계의 응답자 특성 문항들의 표준화 방안이다.

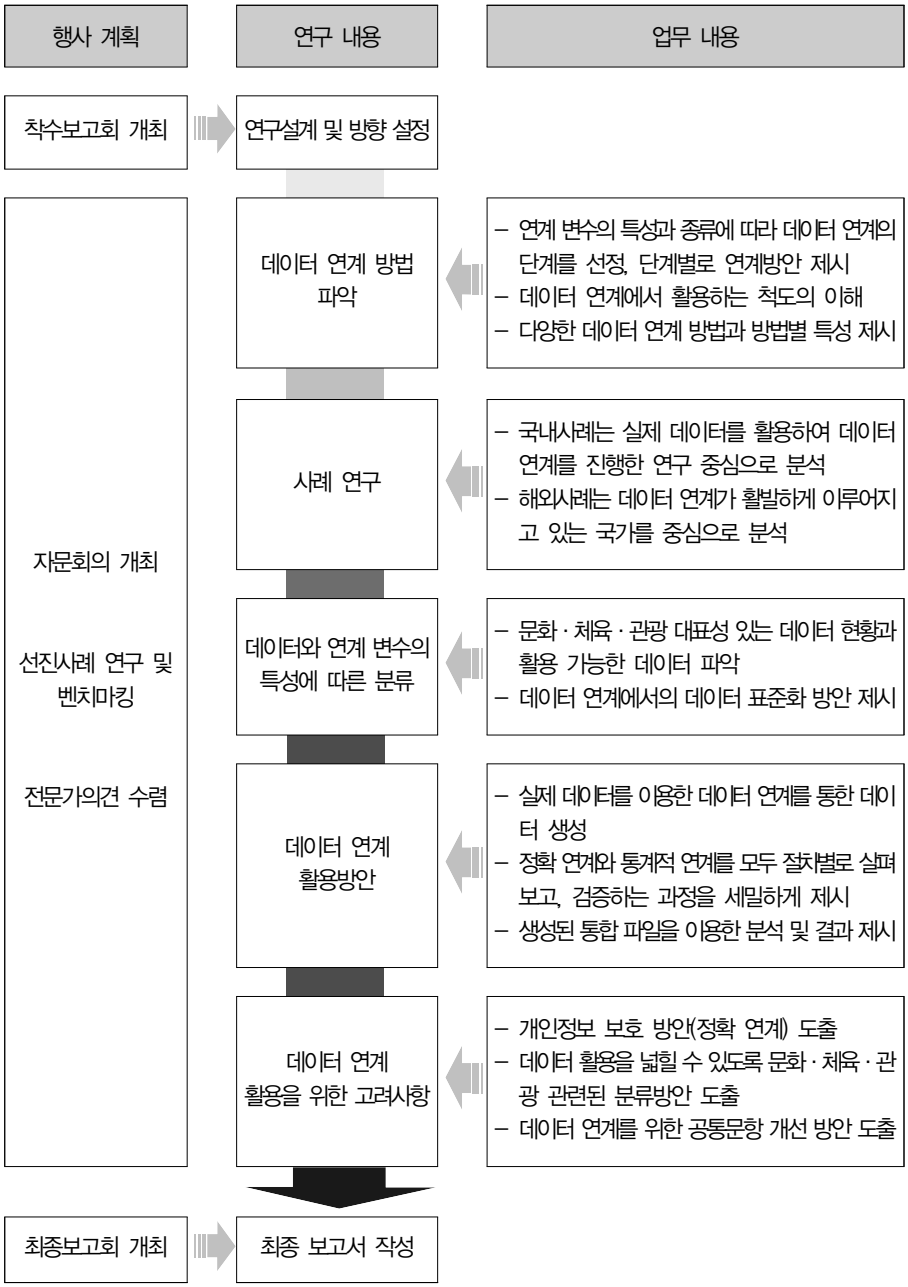
개인정보문제는 많은 연구에서 다루고 있으며, 본 연구에서는 데이터 연계와 관련하여 필요한 사항만을 살펴본다. 문화·체육·관광 관련 분류체계는 많은 통계자료에서 문화체육관광 관련된 데이터만을 식별 또는 분리할 수 있는 기준으로 문화체육관광 관련 분류체계 정립 방안을 제시하며, 이와 함께 문화체육관광부의 조사통계에서 가각 다른 형태로 구성되어 있는 응답자 특성 문항의 표준화 방안도 제시한다.

제7장에서는 문화·체육·관광 분야에서의 데이터 연계를 통한 빅데이터 생산과 활용에 대한 연구를 정리하고 데이터 활용 방안과 데이터 연계를 통한 (빅)데이터 생산 및 활용을 위한 체계의 발전 방안을 함께 논의하도록 한다.

나. 연구 방법

연구 방법은 [그림 1-2]에 제시한 연구 수행 체계에 맞춰 진행한다. 먼저 국내·외 데이터 연계 사례분석을 통해 문화·체육·관광 데이터 연계 방안을 도출한다. 국내 사례분석은 데이터 연계 방법 자체에 대한 연구보다는 실제 데이터를 활용하여 데이터 연계를 진행한 연구 중심으로 검토하며, 해외 사례분석은 국내 사례분석과 유사하게 실제 데이터를 활용하여 연계를 진행한 국가를 중심으로 데이터 연계 방법과 함께 개인정보보호와 연관이 있는 데이터 접근에 대한 내용까지 검토하도록 한다. 또한 문화·체육·관광 관련 데이터 현황 분석을 통해 데이터 연계에 활용 가능한 데이터를 파악하고 데이터 연계를 위한 데이터 표준화 방안과 공통문항 개선방안을 도출한다.

그리고 실제 데이터 연계를 수행하고 있거나, 본 연구에서 활용한 데이터 생산 기관의 담당자와 데이터 연계 방법 및 연계 결과에 대한 검증을 위한 통계학자 등으로 구성된 전문가 자문을 통해 문화·체육·관광 데이터 연계에 대한 전문성을 확보한다.


행사 계획
연구 내용
업무 내용
착수보고회 개최
연구설계 및 방향 설정
데이터 연계 방법 파악
– 연계 변수의 특성과 종류에 따라 데이터 연계의 단계를 선정, 단계별로 연계방안 제시
– 데이터 연계에서 활용하는 척도의 이해
– 다양한 데이터 연계 방법과 방법별 특성 제시
사례 연구
– 국내사례는 실제 데이터를 활용하여 데이터 연계를 진행한 연구 중심으로 분석
– 해외사례는 데이터 연계가 활발하게 이루어지고 있는 국가를 중심으로 분석
자문회의 개최
선진사례 연구 및 벤치마킹
전문가의견 수렴
데이터와 연계 변수의 특성에 따른 분류
– 문화 · 체육 · 관광 대표성 있는 데이터 현황과 활용 가능한 데이터 파악
– 데이터 연계에서의 데이터 표준화 방안 제시
데이터 연계 활용방안
– 실제 데이터를 이용한 데이터 연계를 통한 데이터 생성
– 정확 연계와 통계적 연계를 모두 절차별로 살펴보고, 검증하는 과정을 세밀하게 제시
– 생성된 통합 파일을 이용한 분석 및 결과 제시
데이터 연계 활용을 위한 고려사항
– 개인정보 보호 방안(정확 연계) 도출
– 데이터 활용을 넓힐 수 있도록 문화 · 체육 · 관광 관련된 분류방안 도출
– 데이터 연계를 위한 공통문항 개선 방안 도출
최종보고회 개최
최종 보고서 작성

[그림 1-2] 연구 수행 체계

제2장

데이터 연계 개념

제1절

데이터 연계의 개념

1. 가치 측면에서의 빅데이터

가. 빅데이터의 요소

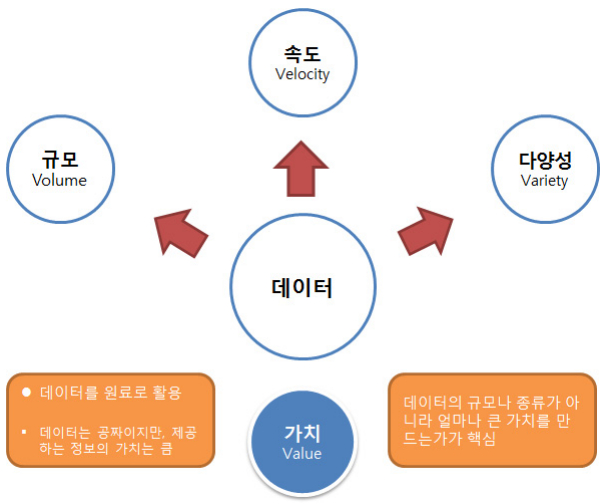
4차 산업혁명의 요소 기술[4] 중에서 중요한 위치에 있는 빅데이터는 최근 몇 년 동안 우리사회 전 분야에서 많은 주목을 받을 뿐만 아니라, 다양한 분야에서 빅데이터를 구축하고 분석하려는 노력이 있어 왔다. 이러한 빅데이터를 구성하는 요소는 [그림 2-1]에서 제시한 것처럼 데이터의 양(Volume), 데이터의 빠른 증가와 처리 속도(Velocity), 자료의 다양성(Variety) 그리고 새로운 가치(Value)를 들 수 있다. 초창기에는 데이터의 양(Volume), 속도(Velocity), 다양성(Variety)의 3요소를 빅데이터 구성 요소로 정의하였지만, 이후 가치(Value)를 추가로 4V[5]를 구성 요소로 보고 있다. 이는 데이터의 규모나 종류가 어떻게 되고, 얼마나 빠르게 처리하는가 보다는 얼마나 큰 가치를 만드는가가 핵심으로 변했기 때문이다.

현재 빅데이터는 정보통신(IT: Information Technology) 뿐만 아니라 다양한 분야에서 중요한 역할을 하고 있으며 향후 더 많은 가치를 창출할 것으로 기대되고 있다. 따라서 빅데이터에 대한 관심이 증가하고 있으며, 빅데이터를 이용하기 위한 노력으로 빅데이터 관련 기술도 발전하고 있다.

빅데이터란 기존의 방식으로 저장·관리·분석하기 어려울 정도로 큰 규모의 자료를 의미한다(오기환 외 2인, 2015). 즉, 빅데이터는 큰(Big)데

4) 4차 산업 혁명의 요소 기술로는 빅데이터, 인공지능(Artificial Intelligence, AI), 로봇공학, 양자암호, 사물인터넷(IoT), 무인 운송 수단, 3D 프린팅 등이 있다(https://ko.wikipedia.org/).
5) 각 요소의 첫 이니셜을 따서 3V, 또는 4V로 빅데이터를 정의한다.

이터를 의미하며, 또한 큰 데이터의 저장·관리와 분석능력이 결합함으로써 나타난 기술이라고도 할 수 있다. 그러므로 빅데이터의 가장 기본은 큰 데이터이여야 하지만 방대한 크기의 데이터는 우리 주변에서 찾아보기가 쉽지 않으며, 규모로서의 빅데이터를 보유하고 있는 곳은 포털업체, 통신업체, 금융관련 서비스를 제공하는 일부의 업체 또는 기관에 불과하다. 따라서 이러한 대용량의 데이터를 생성하고 이를 저장하고 있는 일부 기관에 소속된 연구자 또는 IT 담당자들이나 빅데이터를 접할 수 있다.



[그림 2-1] 빅데이터의 요소

출처 : HR 애널리스트(http://hrd04.tistory.com/)에서 각색

그러나 다양한 기관, 많은 사람들이 여전히 빅데이터를 중요하게 인식하고, 활용하려는 노력을 하고 있다. 이러한 이유는 데이터의 용량이 아닌, 데이터를 이용하여 새로운 정보를 생성하고 가치를 실현할 수 있기 때문이다. 빅데이터 이전에 데이터마이닝(Data Mining)의 정의는 대용량의 데이터로부터 지식을 추출하는 것으로 대용량의 가공하지 않은 데이터로부터 소량의 귀중한 정보를 찾는 과정을 의미한다(박우창 외 3인, 2004). 가치 없는 처치 곤란한 대용량의 데이터였을 때는 중요하게 인식하지 않다가, 분석을 통해 가치 있는 정보를 발견할 수 있었기 때문에

중요성을 인식하게 된 것이다. 빅데이터 역시, 데이터의 양이 엄청나게 크기 때문이 관심 있는 것이 아니고 큰 데이터에서 새로운 정보를 생산하여 가치를 창출하기 때문에 중요하게 관심을 갖는 것이다.

나. 빅데이터의 다양성과 가치

빅데이터는 단일 데이터로 구성되어 있는 것만이 아닌, 다양한 종류의 데이터의 결합에 의해 구성되어 있는 경우가 대부분이다. 최근 우리 주변의 빅데이터 분석이라고 제시하고 있는 결과물들을 살펴보면 대부분이 빅데이터의 기술을 이용하거나, 다양한 데이터를 연결하여 새로운 정보를 얻었을 뿐 용량이 큰 빅(Big) 데이터를 이용한 경우는 많지 않다.

최근의 빅데이터는 용량의 개념이 아닌 새로운 정보 또는 가치가 크다는 측면의 개념으로 인식하는 경우가 많은데, 적은 용량의 데이터를 가지고도 빅데이터 분석을 적용하여 새로운 정보를 찾는 노력을 하는 이유가 바로 다양한 데이터 연계(data linkage) 때문이다. 현재 우리는 빅데이터 분석 결과를 신문 또는 뉴스, 인터넷 등에서 접하고 있는데 대부분이 다양한 형태, 다양한 출처의 데이터를 수집하여 연관성이나 패턴 등을 분석하여 우리가 몰랐거나, 단일 데이터에서는 알 수 없었던 정보를 제공하고 있다. 이렇듯, 최근에는 다양한 데이터를 연계하여, 새로운 정보를 파악하고 이를 활용함으로써 다양한 이익을 얻을 수 있는 노력들을 하고 있다.

빅데이터 3요소 중에서 다양성은 빅데이터 4번째 요소인 가치를 창조하는데 있어 가장 중요한 요소이다. 특히, 다양한 종류의 자료를 결합하여 새로운 가치를 창출하고 미래를 예측하는 작업이 빅데이터 분석의 핵심요소이다. 빅데이터 시대가 빠르게 우리에게 다가올 수 있었던 가장 큰 원동력은 자료의 다양성이며, 자료의 다양성을 증가시키고 새로운 가치를 창출하는데 있어 가장 중요한 것은 데이터를 연계 및 활용하는 방법론의 발달이다(오미애, 2015).

과거에는 숫자 형태로 되어 있는 정형화된 데이터만을 분석하여 통계

로서 활용하였지만, IT기술의 발달로 숫자 데이터는 물론 텍스트(Text), 이미지(Image), 동영상 등의 다양한 비정형 데이터를 분석하는 방법이 개발되고 있다. 다양한 형태, 다양한 소스의 정보들을 결합함으로써 더 가치 있는 정보를 제공할 수 있다. 이렇듯이 데이터의 다양성은 가치 있는 정보를 제공하는 기반이 되며, 다양한 데이터의 결합을 통한 분석은 새로운 가치 있는 정보를 산출하게 된다.

2. 데이터 연계 필요성

IT의 발달과 함께 다양한 종류의 데이터가 생성되고 있으며, 다양한 데이터는 서로 결합하여 분석됨으로써 가치 있는 정보를 제공하고 있다. 그러나 이러한 빅데이터 분석들은 환영받으면서도 회의적인 시각들이 존재하고 있다. 데이터의 기준이 각각 상이함으로써 이를 연계하여 분석한 결과가 대표성과 정확성을 담보하지 못하기 때문이다. 이러한 이유로 빅데이터 분석의 기준을 마련하여 대표성을 확보하려는 노력이 지속적으로 이뤄지고 있다. 이러한 노력의 하나가 동일한 기준으로 데이터를 표준화하여 결합·분석하는 것이다.

연구자나 분석가들은 데이터를 많이 활용하게 된다. 따라서 다양한 데이터를 기관의 DB나 시스템에 축적하고 있는 경우가 많다. 그러나 다양하면서 많은 데이터를 가지고 있어도 새로운 분석 연구(또는 프로젝트)를 할 때, 이에 필요한 데이터가 없어서 일하기 어렵다는 얘기들을 많이 한다. 이는 현재 생산되어 연구에 활용되고 있는 데이터의 대부분이 특정한 목적 또는 업무를 위해 조사된 자료, 시스템을 통해 취합한 행정 자료 또는 실험을 통해 얻은 실험 자료 등인데, 이러한 데이터는 생산되었을 때는 목적 또는 업무에 맞게 활용되겠지만 다른 연구를 할 때는 일부의 정보가 없는 이유로 데이터 활용에 한계가 있는 경우가 많다. 이러한 경우 일부의 필요한 정보가 추가적으로 있어야 새로운 연구에 활용 될 수 있다.

새로운 추가 정보를 얻을 수 있는 방법에는 2가지가 있는데, 하나는 새롭게 데이터를 생산하는 것이고, 다른 하나는 다른 데이터에 있는 정보를 결합하여 분석 데이터를 생성하는 것이다. 데이터 결합은, 각각 다른 원천 데이터에 일부씩 존재하고 있는 데이터로부터 필요한 정보를 추출하여 기존의 데이터와 결합하여 분석에 필요한 데이터를 만드는 것이다. 데이터를 결합할 때, 각각의 데이터를 연결할 수 있는 고유의 연계 변수가 있다면, 연계 변수를 이용하여 기존의 데이터에 다른 데이터에 있는 새로운 변수를 추가한다. 이렇게 확장된 통합 파일을 분석하여 연관성, 패턴 등을 파악하거나 모델링을 하여 새로운 정보를 얻을 수 있다. 데이터를 연결할 수 있는 연계 변수가 있다면 직접적인 연결이 가능하기 때문에, 데이터 결합(data merge)에 대한 이해만 있다면 어려움 없이 연결할 수 있다. 그러나 대부분 결합하여 활용하고 싶은 데이터가 있다 하더라도 이들 각각의 데이터를 직접적으로 연결할 수 있는 연계 변수가 없는 경우가 대부분이다. 연계 변수가 없는 경우, 대부분 데이터를 연결하려는 시도를 하지 않고 데이터를 다시 생산하는 방안을 생각하지만 새로운 데이터를 생산하기 위한 시간과 예산 그리고 인력 등 많은 추가비용이 발생하기 때문에 현실적으로 쉬운 일은 아니다.

데이터 연계는 필요한 정보가 여러 데이터에 나누어져 있을 경우에 이를 하나의 데이터로 통합 또는 연결하여 새로운 정보와 해석을 위해 필요한 방법으로, 데이터 간의 연계를 통해 생성된 통합 데이터는 풍부한 정보를 제공할 수 있다. 데이터 연계는 다른 형태의 다양한 정보를 융합하여 새로운 가치 있는 정보를 찾는 하나의 방법으로 가치 측면의 빅데이터를 생산하는 다양한 방법 중의 하나이다. 이러한 데이터 연계를 많은 나라에서 데이터개방을 통해 공공데이터는 물론 민간 영역의 데이터까지 활용할 수 있는 방안을 마련하는 이유이다. 우리나라 행정안전부에서 공공데이터포털(https://www.data.go.kr/)은 대표적인 예라 할 수 있으며, 현재에도 데이터 이용을 통한 가치 창출이 가능하도록 데이터 공유 확산을

위한 노력이 이뤄지고 있다.

또한 서로 다른 데이터를 연결하여 새로운 통합 데이터를 생성함으로써 데이터의 가치를 높이려는 시도는 연구뿐만이 아니라 실제 많은 국가에서 다양한 분야에서 적용하고 있다. 이는 데이터를 수집, 저장, 분석하는 IT기술과 데이터 연계 방법론에 대한 연구가 지속적으로 되어 왔기 때문에 가능한 일이다. 데이터 연계로 다양한 정보를 가진 데이터를 활용할 수 있게 된다면, 새로운 정보, 인사 또는 마케팅 등의 다양한 분야에서 의사결정에 도움을 줄 수 있을 것이다.

3. 데이터 연계란?

데이터 연계(data linkage)는 서로 개별적으로 생산되어 별개의 파일(또는 데이터베이스(DB))로 존재하는 데이터를 하나로 연결하여 통합 파일을 만듦으로써 새로운 정보를 활용할 수 있도록 하는 방법이며, 이를 데이터 매칭(data matching) 또는 데이터 통합(data integration or data fusion)으로도 표현한다. 즉, 데이터 연계는 서로 다른 복수의 데이터 파일을 결합하여 하나의 완전한 통합 데이터를 만드는 방법으로 정의되며, 이때의 통합된 파일은 결합을 통해 풍부한 정보를 제공하는 것을 의미한다(오미애, 2015). 데이터 연계에 대한 명칭과 의미는 조금씩 차이가 있는데, 최근에는 대체적으로 구분하지 않고 사용하고 있다. 변종석 외 4인(2013)의 연구에서는 데이터 매칭과 데이터 통합을 구분하여, 데이터 매칭은 자료관점에서 서로 다른 자료를 결합하는 과정을 의미하고, 데이터 통합은 서로 다른 자료를 결합하여 새로운 데이터셋을 제공하기 위한 데이터 연계 과정의 모든 활동을 포괄적으로 포함하는 개념으로 사용되기에 엄밀한 의미에서 차이가 있다고 하였다. 그러나 일부 연구자들은 데이터 매칭과 데이터 통합은 동일한 개념으로 사용되기도 한다(National Research Council 1992, Kamakura and Wedel 1997).

관점에 따라 다양하게 사용되고 있지만, 본 연구에서는 혼동을 피하기 위해 데이터 연계로 통일하여 사용하기로 한다. 즉, 데이터 연계는 데이터 매칭, 데이터 퓨전이나 데이터 통합과 동일한 의미로 사용함으로써 용어에 대한 혼란을 피하도록 한다. 실제로는 데이터 연계와 데이터 매칭은 다소 다른 의미로 사용되는데, 데이터 연계는 데이터를 연계하는 전체적인 프로세스 과정을, 데이터 매칭은 연계 방법을 사용하여 데이터를 결합하는 의미로 구분하여 사용되고 있다. 그러나 이는 명확하게 구분하여 설명하기가 어려울 뿐만 아니라, 독자들이 혼동할 우려가 있기 때문에 본 연구에서는 구분하지 않고 데이터 연계로만 사용하도록 한다.

데이터 연계는 서로 다른 데이터를 하나의 파일로 연결시키는 방법으로 연결시키고자 하는 목적에 부합하도록 연결 규칙을 마련하고 그 규칙에 따라 케이스[6] 하나하나를 연결시키는 것을 의미한다. 이 때 데이터를 연결시키는 규칙을 정하기 위한 중요한 조건이 있는데, 바로 데이터를 연결하기 위한 공통 변수이다. 공통 변수가 고유식별이 가능한 값을 가지는 경우에는 연계 변수로 활용하여 직접적으로 연결을 하고, 고유한 값을 가지고 있지 않다면 공통 변수들을 이용한 연결 규칙을 생성하여 연결시키도록 한다.

4. 데이터 연계의 구분

가. 연결 기준에 따른 구분

서로 다르게 생성된 2개 이상의 데이터 파일을 결합하는 방법인 데이터 연계는 데이터를 연결하는 기준 따라 구분할 수 있으며, 현재 가장 일반적으로 사용하는 데이터 연계에 대한 구분은 영국의 “National Statistics code of Practice on Data Matching(2003)”에서 제시한 5가지이다. 5가

6) 하나의 케이스(case)는 조사대상이 되는 한 사람의 데이터를 의미한다. 즉, 한 줄(line)의 데이터를 의미한다.

지는 정확 연계(Exact Matching), 판단 연계(Judgemental Matching), 확률적 연계(Probability Matching), 통계적 연계(Statistical Matching) 그리고 데이터 연결(Data Linking)이다(이영섭 외 4인, 2009). 이 때 연결하는 기준은 데이터가 가지는 공통 변수의 특성과 연구자가 판단에 따라 달라지며, 연계하는 규칙은 선택된 데이터 연계 방법론 중에서 최적 방법으로 선택하게 된다.

〈표 2-1〉 연결 기준에 따른 데이터 연계 구분

구분	정의	연결 기준
정확 연계	두 데이터에 동일한 고유식별정보가 있어 이를 이용하여 연결	고유식별정보
판단 연계	두 데이터가 정확히 일치하는 것은 없지만, 연구자의 데이터에 대한 이해를 바탕으로 연결	연구자의 지식
확률적 연계	정확 연계에서 일치하지 않는 케이스들이 발생할 때, 각 변수들의 연결가능성을 계산하여 연결	공통 변수로 계산한 가능성
통계적 연계	고유식별정보가 없는 경우 통계적 방법으로 가장 유사한 케이스를 찾아 연결	유사성 측도
데이터 연결	둘 이상의 파일에서 변수들간의 연관성이 있을 경우 하나가 변화할 때 같이 변화가 가능하도록 연결	변수들 간의 연계성

출처 : 이영섭 외 4인(2009), 통계조사 자료와 행정 자료간의 통계적 연계 기법에 관한 연구

〈표 2-1〉에서 구분한 5가지 데이터 연계를 살펴보면, 정확 연계는 서로 다른 데이터 파일에 각각 주민등록번호, 사업자등록번호 등과 같은 고유식별 변수가 공통으로 있는 경우에 적용할 수 있다. 정확 연계는 연계하고자 하는 각각의 파일에 동일한 고유식별 변수가 있을 때, 연결하는 각각의 파일에서 고유식별 값이 완전히 일치하는 케이스끼리 연결하는 방법이다. 정확 연계는 각각의 데이터에 있는 동일한 대상(예를 들어 동일인, 동일 사업체, 동일 가구 등)의 정보가 있을 때, 고유식별 값을 이용하여 동일한 대상을 결합시키기 때문에 가장 정확한 데이터 연결 방법이다. 또한 고유식별이 가능한 변수가 없을지라도 주소, 전화번호, 성(surname), 성별 등과 같이 다양한 변수들을 연결하였을 때 고유식별이 가능하다면

정확 연계를 할 수 있다. 정확 연계는 고유식별이 가능해야 활용 할 수 있기 때문에 개인정보보호와 같은 민감한 문제가 발생한다. 따라서 이러한 데이터 이용에는 제약이 있을 수밖에 없다. 즉, 정확 연계는 개인정보와 같은 민감한 정보를 이용하기 때문에 데이터에 대한 보안이 보장되는 환경에서 데이터 연계의 권한을 부여받은 사람만이 활용 할 수 있다.

판단 연계는 공통인 변수들 사이에 정확히 일치하지는 않지만, 연구자가 자료에 대해 잘 알고 있는 경우 연구자의 주관적 판단에 의한 자료를 연계하는 방법이다. 주관적 판단이 어려울 경우, 일부분에 대한 별도의 조사 또는 분석을 통해 적절하다고 판단되는 케이스들을 연결함으로써 데이터를 결합하는 방법이다. 즉, 정확히 일치하는 대상도 아니고 특별한 규칙도 없지만 연구자가 데이터에 대한 이해를 통해 주관적으로 연결하는 방법이다. 판단 연계는 대부분 정확 연계를 한 후 일부 케이스들이 연계되지 않았을 때, 연계되지 않은 케이스들을 검토하여 이름 또는 주소에 오타나 글자의 일부가 잘렸을 때, 연구자의 경험 또는 지식을 바탕으로 직관적으로 데이터를 연계한다. 이러한 경우는 서양에서 사람의 이름과 주소를 이용한 연계에서 많이 적용된다. 서양 사람의 이름은 길고 명확하기 때문에 이름과 주소만 가지고도 정확 연계와 같은 효과를 낼 수 있는데, 이때 데이터의 한두 개의 케이스가 철자가 다를 경우에 연구자가 사전 정보 또는 새로운 정보를 이용하여 연계를 결정하게 된다. 물론 정확 연계와 상관없이 사용할 수도 있다.

확률적 연계는 정확 연계의 공통 변수 일부에 오류가 있는 경우, 정확한 정도에 따라 가중치를 주고 확률적으로 데이터를 결합하는 방법이다. 즉, 같은 대상에게 다른 조사를 진행한 경우, 고유식별자로 사용할 공통변수는 있지만, 일부 케이스들의 고유식별자 값에 오류가 있어 일치하지 않는 케이스들의 정보를 확률로 계산해서 가장 일치할 가능성이 높은 케이스들을 연결하는 방법이다. 이는 통계적 연계와 유사하며, 확률적 연계 방법은 대부분 정확 연계를 한 후 연계가 안 된 케이스들을 연결할 때

주로 사용한다는 특징이 있다.

통계적 연계는 연결하고자 하는 각각의 데이터에 있는 공통 변수에 고유식별이 가능한 변수가 없는 경우에 사용하는 방법으로 정확하게 일치하는 대상은 아니지만 유사한 케이스를 찾아 연결하는 방법이다. 이 방법은 정확 연계처럼 정확한 대상을 찾아 연결하는 것이 아니라 변수들의 기본 성질을 유지하는 형태로 연결시킨다. 따라서 데이터가 가지고 있는 정보를 100% 유지하기는 어렵지만 새로운 데이터를 생산하는 비용(예산, 인력 시간 등)에 비해 효율적이다. 하지만 고유식별 변수가 없어 유사한 성향을 가진 케이스들을 연결하기 때문에 통계적 연계 방법으로 연계하기 위해서는 다양한 연계 방법과 절차들 중에서 가장 적합한 방법을 선택해야 하며, 가장 적합한 방법을 선택하기 위한 방법과 절차가 어렵고 까다롭다는 문제가 있다.

데이터 연결은 둘 이상의 파일에서 변수들 간의 연관성을 만들어 바로 데이터 갱신(update)이 가능하도록 하는 데이터 결합 방법이다. 즉, 데이터를 직접적으로 결합하여 하나의 데이터로 만드는 것이 아니고, 데이터 간의 연관성을 만듦으로써 하나의 데이터가 변화하면 연계된 데이터에도 변화를 반영할 수 있도록 하는 방법이다. 이는 데이터의 관계를 이용한 관계형 데이터베이스(relational database)를 구현할 때 주로 사용하는 방법으로 변수들 간의 관계가 명확할 때 사용하게 된다.

이러한 5가지 연계 중에서 본 연구에서는 두 개 이상의 데이터를 연결하여 하나의 데이터로 만드는 방법으로 정확 연계와 통계적 연계만을 고려하여 살펴보기로 한다. 그 이유는 데이터 연결은 데이터를 결합하는 것이 아닌 데이터를 데이터베이스 관점에서 연결하는 것이기 때문에 하나의 파일로 결합하는 것과는 거리가 있다. 그리고 판단 연계와 확률적 연계는 단독으로 사용할 때는 통계적 연계와 비슷하지만 실제로는 정확 연계를 한 후에 연계가 되지 않은 케이스들은 연결할 때 주로 사용하기 때문에 본 연구의 데이터 연계 방법에서는 제외하였다.

나. 데이터 특성에 따른 구분

데이터 연계는 결합할 데이터의 특성에 따라 구분할 수 있으며, 데이터 특성에 따른 구분은 행정 데이터(administrative data)와 조사 데이터(survey data)로 구분할 수 있다. 사전적 의미로의 행정 데이터는 공공기관이 직무상 작성·취득하여 관리하고 있는 데이터베이스를 의미하며, 조사 데이터는 다양한 조사방법을 활용하여 개별 개체의 기록(record)을 포함하는 데이터를 말한다(오미애 외 4인, 2014). 그러나 최근에는 행정자료를 공공자료에 국한하여 사용하지 않고, 민간에서 생산되는 행정 자료와 공공에서 생산되는 행정 자료를 포함하는 의미로 사용한다. 따라서 행정 자료는 시스템상에서 생성되는 데이터 축적의 개념에서 빅데이터를 포괄하는 개념이라 할 수 있다. 즉, 행정 자료는 정형의 빅데이터로 표현할 수 있다.

데이터 연계는 연결하려는 데이터의 특성에 따라 4가지로 구분되는데, 행정 데이터와 행정 데이터, 행정 데이터와 조사 데이터, 조사 데이터와 행정 데이터 그리고 조사 데이터와 조사 데이터이다. 데이터 특성을 살펴보면, 행정 데이터는 대부분 시스템상에서 수집이 되기 때문에 데이터베이스로 구축되어 있으며, 특히 공공 데이터의 경우는 고유식별이 가능한 변수를 포함할 가능성이 높다. 조사 데이터의 경우는 설문조사 방식으로 데이터를 수집하며, 응답자의 특성을 묻는 질문이 얼마나 자세한지에 따라 연계의 효율이 달라진다. 그러나 응답자들의 부담 등의 이유로 개인 신상에 대한 질문을 많이 묻는 것은 어렵다. 특히 고유식별이 가능한 정보를 설문 문항에 넣을 수도 없으며, 이를 수집한다 하더라도 이를 이용해서는 안 된다. 그러나 통계청에서 진행하는 센서스[7]와 같은 조사는 국가에서 시행하는 경우이기 때문에 예외라 할 수 있다.

7) 국민을 대상으로 하는 '인구센서스'와 사업체를 대상으로 하는 '전국사업체조사'가 이에 해당한다.

〈표 2-2〉 데이터 특성에 따른 데이터 연계 구분

연계 방법	데이터 특성	
	기준 파일	연계 파일
정확 연계	행정 데이터	행정 데이터
	행정 데이터	조사 데이터
	조사 데이터	행정 데이터
	조사 데이터	조사 데이터
통계적 연계	행정 데이터	행정 데이터
	행정 데이터	조사 데이터
	조사 데이터	행정 데이터
	조사 데이터	조사 데이터

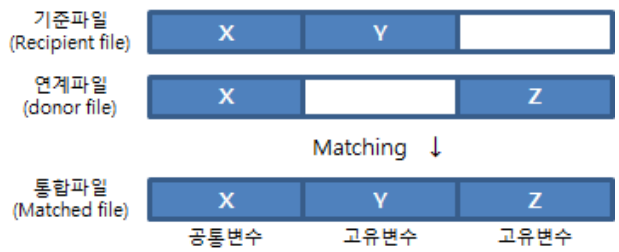
출처 : 오미애 외 4인(2014), 보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안

제2절

데이터 연계의 체계

1. 데이터 연계의 용어[8)]

데이터 연계는 서로 다른 데이터 파일을 하나의 파일로 결합함으로써 새로운 데이터를 생산하는 방법인데, 더 자세히 설명하면 [그림 2-2]에서와 같이 주로 사용하고자 하는 데이터 파일에, 필요한 변수가 있는 데이터를 찾아 결합하는 것이다. 즉, 주로 사용하는 파일에 새로운 데이터에서 필요한 변수를 결합하는 의미로 이해할 수 있다. 서로 다른 두개의 데이터가 있다는 가정 하에 데이터를 연계할 때, 기준이 되는 파일을 기준 파일(주체파일(host file) 또는 수용파일(recipient file)이라고도 함)로, 추가적인 정보를 얻기 위해 연결하려는 대상을 연계 파일(제공파일(donor file)이라고도 함)로 정의한다.



[그림 2-2] 데이터 연계

출처 : 이영섭 외 4인, 통계연구(2009), 통계조사 자료와 행정 자료 간의 통계적 매칭기법에 관한 연구

[그림 2-2]에서 처럼 기준 파일은 변수 X, Y로 구성되어 있으며 연계파일은 X, Z로 구성되어 있다고 할 때, 두 파일에 모두 존재하는 변수

8) 데이터 연계에서 사용하는 용어는 영어의 단어로는 차이가 없지만, 한국어로 해석하는 과정에서 조금씩 차이가 있게 사용하고 있는데, 본 연구에서는 오미애 외 4인(2014)이 사용했던 용어를 주로 이용하였다.

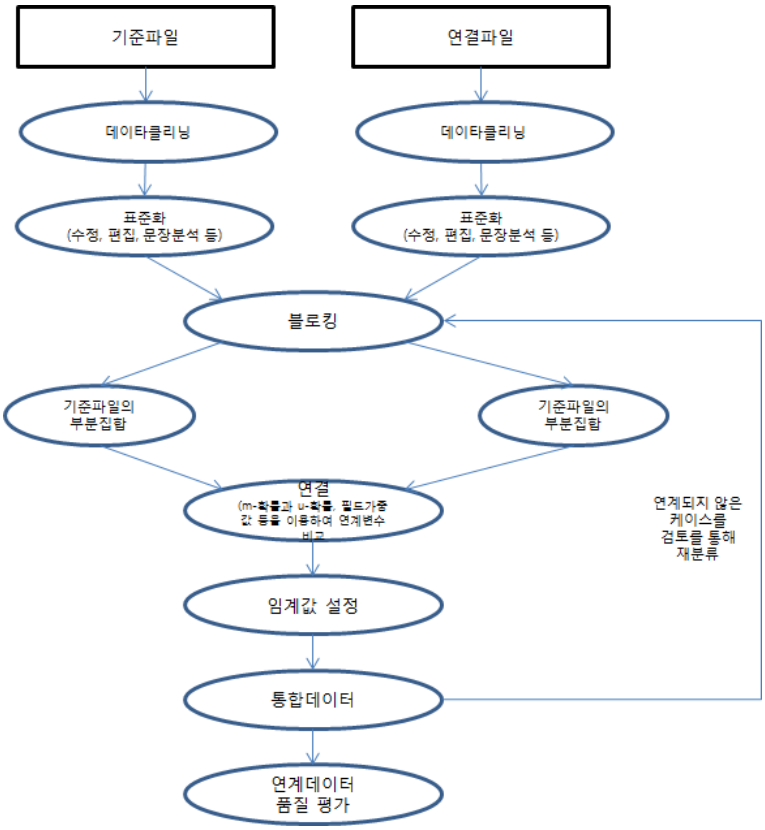
X는 공통 변수(common variable)라고 하며 기준 파일에만 존재하는 변수 ***Y***와 연계 파일에만 존재하는 변수 ***Z***를 고유 변수(unique variable)라고 한다. 데이터 연계는 공통 변수 ***X***를 이용하여 연계 파일에 있는 ***Z***를 기준 파일에 추가하는 것이며, 공통 변수 중에서 연계에 사용된 변수를 '연계 변수(matching variable)라고 정의한다. 또한 연계 변수를 통해 연계된 하나의 파일을 통합 파일(또는 결합파일(matched file))이라 한다.

여기서는 데이터 연계를 위한 기본 용어만을 설명하며, 데이터 연계하는 과정에서 사용하는 통계 기법 용어는 해당 방법을 설명할 때 제시하도록 한다.

2. 데이터 연계 프로세스

데이터 연계의 기본 과정은 정확 연계와 통계적 연계 모두 동일하다. 세부적으로는 정확 연계는 고유식별 변수가 있어 이를 이용하기 때문에 거의 하나의 방법이라면, 통계적 연계는 매우 다양한 방법이 있어 데이터의 특징에 따라 사용하는 방법도 달라지기 때문에 차이가 있다. 데이터 정확 연계, 통계적 연계 모두 연계의 전체적인 체계는 데이터 연계하는 과정을 도식화한 [그림 2-3]과 같으며, 데이터 연계 과정에 대한 기본적인 이해를 위해 프로세스의 처음부터 순차적으로 차례대로 세부 내용을 살펴보도록 한다.

설명을 위해 기준 파일과 연계 파일이 각각 있다고 가정한다. 실제로는 세 개 이상의 데이터 파일의 경우에도 데이터 연계를 할 수 있는데, 이 경우에도 두 개의 데이터 파일을 먼저 연계한 후 다른 데이터 파일을 연계하기 때문에 두 개의 데이터파일을 연결하는 과정을 설명하는 것으로 충분하다.



[그림 2-3] 데이터 연계 프로세스

출처 : 오미애 외 4인(2014)을 바탕으로 재작성함

데이터 연계에서 가장 먼저 해야 하는 것은 데이터 클리닝이다. 데이터 클리닝은 말 그대로 데이터를 깨끗하게 하는 과정으로 데이터의 오류가 없도록 수정하여 바로잡는 과정이다. 데이터에 오류가 있다면 데이터 연계 기법을 아무리 정교하게 적용한다고 하더라도 잘못된 데이터에 적용하기 때문에 당연히 잘못된 결과를 생성할 수밖에 없다. 따라서 데이터클리닝 과정은 가장 기본적이면서도 가장 중요한 과정이라 할 수 있다.

데이터 클리닝이 끝나면 표준화 작업을 진행하게 된다. 표준화 작업은 쉽게 설명하면 기준 파일과 연계 파일의 동일한 의미를 갖는 변수를 같은 기준으로 만드는 과정이라 할 수 있다. 예를 들어 〈표 2-3〉과 같이 기준

파일의 성별이 남자는 1이고 여자는 2라고 입력되고 연계 파일은 남자는 M으로 여자는 F로 입력되었다면, 통계프로그램에서는 다르게 인식하게 된다. 따라서 같은 기준이 되도록 연계 파일의 성별을 남자는 1로, 여자는 2로 수정하여 두 파일의 성별이 동일한 값이 되도록 표준화 한다. 표준화 과정은 변환, 생성, 추출 등을 포함하는데 이러한 과정 모두 동일한 의미를 가지는 변수의 값이 서로 다른 두 개의 파일에 다르게 입력되었다면 이를 동일한 기준으로 맞춰주는 것을 의미한다.

〈표 2-3〉 표준화 과정의 성별 예시

구분	기준 파일	연계 파일
표준화 전	남자: 1, 여자: 2	남자: M, 여자: F
표준화 후	남자: 1, 여자: 2	남자: 1, 여자: 2

다음으로 블로킹 기법을 적용하는데, 블로킹은 명목화된 변수를 선정함으로써 필터링을 하여 데이터 연계를 위한 값들의 비교 횟수를 줄이면서도 연계의 효율을 높일 수 있는 방법이다. 이를 자세히 설명하면, 10,000개의 케이스가 있는 두 개의 데이터를 연계한다면 자신의 값과 가장 가까운 데이터를 찾기 위해서 100,000,000(10,000×10,000)개의 비교가 이뤄지게 되며, 연계에 활용되는 케이스는 10,000개이므로 99,000,000개의 쌍들은 '비연결'로 판단하게 된다.

최근에는 데이터의 크기가 점차 커지고 있는데, 데이터의 수가 증가할수록 비교의 쌍은 제곱의 형태로 증가하게 된다. 따라서 비교 횟수를 줄이기 위해 특정한 값으로 필터(filtered)한 특정 케이스들만 비교하도록 하는 방법을 고려하는데 이렇게 특정한 조건으로 필터링(filtering) 하는 것을 블로킹(blocking)이라 하며, 필터링한 변수를 블로킹 변수라고 한다. 비교 횟수는 블로킹 변수로 선택된 범주만큼 비례적으로 줄어든다. 예를 들어 10개의 범주와 하나의 범주마다 1,000개 케이스가 있다고 가정할 때, 하나의 범주에서 비교되는 쌍은 1,000,000개이므로 전체 비교 횟수의 10분의 1로 줄어들며, 이는 연계하는 시간과 컴퓨터 메모리의 부

하를 줄여 적은 메모리에서도 연계 작업을 무난하게 할 수 있게 된다.

블로킹 변수를 결정하는 것은 매우 중요한 일이며, 두 데이터를 연계할 때 동일한 변수의 값은 반드시 같은 데이터로 연계하고자 하는 변수를 블로킹 변수로 선정한다. 예를 들어 사람에 대한 데이터를 연계할 경우, 성별 변수를 블로킹으로 사용한다면 기준 파일의 남자 데이터와 연계 파일의 남자 데이터 간에 연계가 되도록 하며 기준 파일의 여자 데이터와 연계 파일의 여자 데이터를 간에 연계가 되도록 한다. 블로킹 방법은 연계의 효율을 높이기 때문에 많이 이용하며, 한 변수가 아닌 여러 변수를 사용하여도 된다. 즉, 성별과 시도 단위의 지역을 블로킹 처리를 한다면, 2×17=34개의 부분 집합으로 데이터를 분할하여 연계를 하게 된다. 따라서 언뜻 생각하면, 블로킹 변수를 많이 사용하면 품질 좋은 연계가 일어날 것 같지만, 연계하는 시간과 시스템의 작동에는 효율이 좋을지 모르나 통합 파일의 효율에도 도움이 될지는 알 수가 없다. 각각의 파일 안에 있는 변수들의 분포가 다르기 때문에 블로킹 변수를 많이 사용할 경우 일부의 변수에 있어 왜곡이 발생할 수 있다. 따라서 블로킹변수를 어떤 것을 사용할지는 데이터 연계에서 매우 중요한 과제 중 하나이다.

데이터 연계 규칙은 크게 정확 연계와 통계적 연계로 구분되며, 이는 다음 절에서 각각 자세히 설명하도록 한다. 연계 규칙을 적용하여 통합 데이터를 생성하고, 이때 연계 되지 않는 케이스가 있다면 다시 새로운 연계 규칙을 적용하는데, 모든 데이터를 반드시 연계할지에 대한 결정은 상황에 따라 다르며, 최종 결정은 연구자가 판단한다.

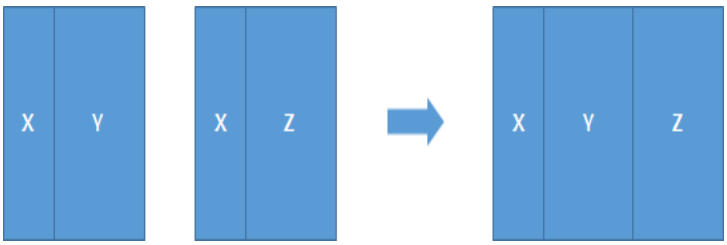
데이터 연계가 끝나면 연계의 기본 가정을 만족하였는지, 데이터의 관계가 변하지 않았는지 등의 데이터 연계 품질을 평가하게 된다. 연계한 통합 파일의 품질이 좋지 않을 경우, 이를 이용한 분석은 왜곡된 결과를 산출할 수밖에 없기 때문에 통합 파일과 연계 파일을 비교하여 통합 파일의 품질이 좋지 않다면 다른 방법으로 데이터 연계를 시도해야 한다.

제3절

정확 연계 방법

1. 정확 연계의 형태

정확 연계는 서로 다른 데이터에서 동일한 대상을 찾아 연결하여 통합 파일을 만드는 방법으로, 정확하면서도 다양한 정보를 얻기 때문에 매우 효율이 좋은 방법이다. 정확 연계는 동일한 대상을 찾아 연결하기 때문에 각 개체를 정확하게 파악할 수 있는 고유식별이 가능한 변수가 있어야 한다. 정확 연계의 형태는 [그림 2-4]와 같다.



[그림 2-4] 정확 연계의 형태

[그림 2-4]를 살펴보면, 공통 변수 $\boldsymbol{X}$가 하나의 변수일수도 여러 개의 변수들로 구성된 $\boldsymbol{X}=\{X_1, X_2, \cdots, X_p\}$ 형태의 벡터(vector)일수도 있다. 한 개의 변수일 경우에는 주민등록번호와 같은 고유한 식별이 가능한 변수이고, 벡터의 경우는 고유식별 변수는 없지만 여러 개의 변수를 조합할 경우 고유한 식별이 가능하게 되는 경우로 이름, 생년월일, 성별, 소속, 주소 등과 같은 정보를 통해 개인을 식별하는 경우가 해당된다.

2. 정확 연계의 절차

정확 연계도 [그림 2-3]의 프로세스를 기본적으로 따르며, 연계 전에

이루어지는 데이터 클리닝과 표준화 작업의 세부 절차는 다음과 같다.

① 블로킹 변수를 선정하여, 데이터셋을 블로킹의 범주 수만큼의 부분 집합으로 분할한다.

② 분할된 부분집합별로 고유식별정보를 이용하여 정렬한다.

③ 위에서부터 동일한 고유식별정보를 가진 케이스들끼리 연결하여 공통 파일을 만든다.

④ 연계되지 않은 케이스가 있다면 고유식별 값이 오타나 누락 등이 있는지 살펴보고, 오류가 있다면 수정하여 다시 연계하며 오류가 없다면 연계되지 않은 케이스를 제외하고 통합 파일을 구축한다.

정확 연계에서 블로킹을 하는 이유는 연계의 효율을 높이기 위한 것이 아니라 많은 데이터를 연계하기 위해, 시간과 시스템의 부하를 줄일 수 있도록 부분집합으로 데이터를 분할하여 케이스들의 고유한 값이 동일한지를 비교하는 경우의 수를 줄이기 위함이다.

정확 연계는 동일한 대상을 연계하여 공통 파일을 만들기 때문에 매우 정확하고 좋은 방법이다. 그러나 정확 연계를 할 때도 고려해야 할 사항이 두 가지 있다. 첫 번째는 개인정보보호 문제이다. 정확 연계는 고유식별이 가능하기 때문에 항상 개인정보문제가 주요한 이슈가 된다. 두 번째는 기준 파일의 모든 케이스(또는 대부분의 케이스)가 연계 파일과 연계가 되지 않는다면, 연계된 공통 파일의 분포가 기준 파일과 같지 않기 때문에 정보의 변질이 발생한다. 즉, 정확 연계는 연계 파일이 기준 파일의 대상을 모두 포함하고 있다면 문제가 없겠지만, 기준 파일의 대상을 포함하지 못해 기준 파일의 많은 케이스들이 연계되지 않은 상태에서 공통 파일을 활용한다면 기준 파일의 대상과 전혀 다른 대상으로 분석하는 것과 같다. 따라서 정확 연계에서 기준 파일이 모두 연계되지 않았다면 연계된 케이스들만으로 새로운 의미를 부여한 통합 파일로 분석을 하든, 연계되지 않은 케이스들을 판단 연계 또는 확률적 연계(또는 통계적 연계)를 적용하

여 남은 케이스들은 모두 연계되도록 할지를 판단해야 한다.

3. 정확 연계에 대한 평가

정확 연계는 동일한 대상을 찾아 연계하는 방법이기 때문에 연계가 잘 되었는지에 대한 평가는 별도로 할 필요가 없다. 연계가 되었다면 동일한 대상끼리 연결되었을 것이고, 연계되지 않았다면 동일한 대상이 없다는 것이다. 그러나 정확 연계에서는 연계가 얼마나 되었는지에 대한 평가는 필요하다. 기준 파일의 데이터가 얼마나 연계 파일과 연계되었는지에 대한 평가는 연계율로 판단한다. 기준 파일의 일부만 연계되었다면, 아무리 동일한 대상을 연계하였다고 하더라도 통합 파일을 활용하는 것은 어렵다. 일부분의 데이터만 연계된 통합 파일은 기준 파일의 분포를 유지하지 못하기 때문이다.

정확 연계는 연계된 통합 파일의 분포가 기준 파일의 분포와 동일성을 유지해야 연계된 결과를 활용할 수 있다. 정확 연계는 이용할 수 있다면 가장 좋은 연계 방법이지만, 모두가 연계되는 경우가 아닐 때에는 통합 파일의 대표성을 담보할 수 없다. 이러한 경우 확률적 연계 또는 통계적 연계 방법을 적용하여 연계되지 않은 기준 파일의 나머지 데이터를 연계하여 기준 파일의 분포가 유지될 수 있도록 해야 한다.

제4절

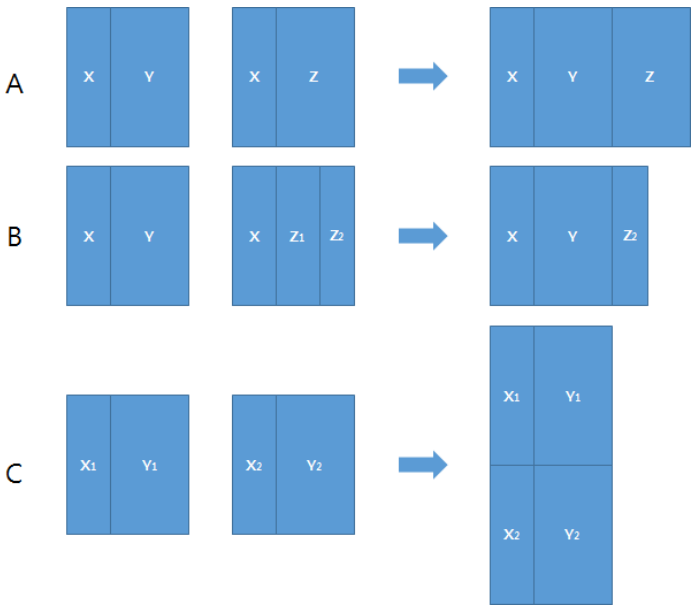
통계적 연계 방법

정확 연계는 동일한 대상을 찾아 연계하는 방법으로 고유식별이 가능한 정보를 바탕으로 연계하기 때문에 개인정보문제로 인해 데이터에 접근하여 이용하는 것은 어려울 수 있지만 연계 방법은 간단하다. 그러나 통계적 연계는 동일한 대상을 연결하는 것이 아닌 유사한 대상을 연결하는 것으로 연계하는 과정에 대한 절차를 어떻게 하는지에 따라 다른 결과가 나올 수 있기 때문에 연계 방법도 매우 다양하며 절차도 복잡하다.

정확 연계는 고유식별정보를 이용하기 때문에 개인정보보호문제로 데이터를 이용하는 것만으로도 문제가 될 소지가 있으며, 통계청과 같이 데이터 보안센터가 있는 기관에서만 활용할 수 있다. 따라서 정확 연계는 대부분의 연구자들은 실제로 거의 활용하기 어렵다. 우리가 접하는 대부분의 데이터는 개인식별정보가 없는 데이터이며 그 중의 대부분은 조사자료이다. 그런데 조사 자료들은 예산과 시간, 그리고 응답자의 부담 등을 이유로 조사하는 문항을 많이 만들어 조사할 수 없다. 즉, 조사 이후에 부족한 정보 또는 추가적으로 알고 싶은 정보가 있을 때 이를 충족시키는 방안의 하나가 통계적 연계이다. 통계적 연계는 정확 연계보다 데이터의 정확성이 부족하며 일부의 정보 손실이 있지만 서로 다른 데이터를 연결하여 활용함으로써 추가 정보를 얻을 수 있다는 장점이 있다.

1. 통계적 연계의 형태

통계적 연계는 서로 다른 데이터 파일에서 유사한 성향을 가진 데이터를 연결하는 방법인데, 데이터를 연계하려는 목적에 따라 연계 방법은 물론 연계의 형태도 구분되며, [그림 2-5]에 제시하였다.



[그림 2-5] 통계적 연계의 형태

[그림 2-5]에서 제시한 것처럼 통계적 연계의 형태는 3가지 경우를 고려할 수 있으며 각 경우마다 연계 방법을 적용할 때 고려사항에 차이가 있다. A의 경우는 정확 연계와 같이 기준 파일과 연계 파일 모두를 한 개의 통합 파일로 연결하는 것으로 데이터 연계의 기본 구조이다. B의 경우는 연계 파일의 고유 변수 모두가 아닌 일부 또는 한두 개의 변수에 관심이 있는 경우로, 모든 파일을 가져오지 않고 관심 있는 변수만을 연계하기 때문에 특정 변수를 목적 변수로 생각하여 연계하면 된다. 다음으로 C[9]의 경우는 데이터 연계의 특별한 형태로 두 파일의 고유 변수가 같은 경우이다. 이때는 ***Y***도 공통 변수로 생각할 수 있지만 ***X***만을 공통 변수로 사용하여 연계하는 특별한 경우이다. 데이터를 연계하고 나면 데이터 병합(data merge)이 아닌 데이터를 추가하는 형태가 된다.

A와 B는 연계 파일의 모든 변수를 연계할지, 아니면 일부를 연계할지

9) C방법은 현재 많이 사용하고 있지만 본 연구와는 다소 차이가 있기 때문에 여기서 간단히 설명하고 이후에는 다루지 않도록 한다.

에 대한 차이를 제외하고는 연계하는 방법과 연계 후 분석 방법은 동일하다. 이는 연계되는 변수의 개수가 차이가 날 뿐 변수가 추가되는 것은 동일하기 때문이다. 그러나 C의 경우는 연계할 때의 방법은 비슷한 방법을 사용할 수도 있지만 분석에 필요한 모든 변수가 동일하기 때문에 변수를 추가하는 것이 아닌 케이스를 추가하는 형태이다. C의 경우는 1집단과 2집단의 고유 변수 Y를 비교하고자 하는 명확한 목적이 있으며, C를 '성향점수(propensity score)를 이용한 연계'라 하며 많은 연구에서 사용되고 있는 방법이다.

2. 통계적 연계에서의 기본 가정

통계적 연계를 사용할 때에는 반드시 지켜야 할 가정과 연계하기 전에 연구자가 선택하는 가정이 있다. 통계적 연계를 사용할 때 꼭 지켜야 하는 기본 가정을 지키지 않는다면 통계적 연계를 통해 생성된 통합 파일을 신뢰할 수 없게 된다. 즉, 통합 파일을 통해 추가적인 정보를 도출하기 위해서는 반드시 지켜져야 하는 가정이기 때문에 통합 파일이 기본적인 가정을 충족하는지 반드시 평가해야 한다.

연구자가 선택하는 가정은 반드시 지켜야 하는 것은 아니지만, 최적의 연계를 위해서 연계 파일의 케이스를 중복하여 연계할지 또는 중복없이 연계할지, 기준 파일의 모든 케이스를 연계할지 등을 판단을 하는 것이다.

가. 통계적 연계의 기본 가정

통계적 연계는 정확한 대상을 연계하는 것이 아니고 유사한 데이터를 연계하는 것이기 때문에 정확 연계보다 사용하기 위한 조건이 까다롭다. 통계적 연계를 적용하기 위해서는 다음과 같은 가정이 충족되어야 한다. 첫 번째는 기준 파일과 연계 파일은 동일한 모집단에서 추출된 데이터이

어야 한다는 것이고, 두 번째는 공통 변수 $\boldsymbol{X}$가 주어졌을 때 고유 변수인 $\boldsymbol{Y}$와 $\boldsymbol{Z}$는 조건부 독립이어야 한다는 것이다. 이는 다음과 같이 표현된다.

$$P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{X}) = P(\boldsymbol{Y}|\boldsymbol{X}) \cdot P(\boldsymbol{Z}|\boldsymbol{X})$$

첫 번째 기본 가정인 기준 파일과 연계 파일의 모집단이 동일해야 한다는 것은 기준 파일과 연계 파일은 같은 분포를 가져야 한다는 것이다. 과거에는 연계할 데이터의 표본은 동일한 모집단에서 조사되었다는 강한 가정을 하였고, Arbeitsgemeinschaft Media Analyse(1996)에서 제시한 연계율은 연계할 두 데이터의 표본은 공통 변수들의 평균에 유의한 차이가 나서는 안 된다고 하였다(오미애 외 4인, 2014). 그러나 van der Puttern et al.(2002)은 두 연계 데이터가 반드시 같은 모집단에서 추출될 필요는 없지만 동일한 분포를 가져야 한다고 제시하고 있다(이영섭 외 4인, 2009).

두 번째인 조건부 독립성 가정(conditional independent assumption)을 하는 이유는 기준 파일과 연계 파일이 같이 조사된 것이 아니기 때문에 두 데이터 파일에 있는 $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$의 결합확률분포함수(joint probability distribution function)를 추정할 수 없는데 공통 변수 $\boldsymbol{X}$가 주어졌을 때 고유 변수인 $\boldsymbol{Y}$와 $\boldsymbol{Z}$가 조건부 독립이 성립한다고 가정하면 다음과 같이 결합확률분포함수를 표현할 수 있게 된다.

$$\begin{aligned}\hat{f}_{XYZ}(x,y,z) &= f_{XY}(x,y) f_{Z|X}(z|x) \\ &= f_{Y|X}(y|x) f_X(x) f_{Z|X}(z|x) \\ &= f_{YZ|X}(y,z|x) f_X(x)\end{aligned}$$

결론적으로 조건부 독립성의 가정이 충족된다면 통계적 연계를 한 후 각각의 데이터로부터 추정할 수 없었던 $\boldsymbol{Y}$와 $\boldsymbol{Z}$의 관계를 파악할 수 있게 된다. 실제로 $\boldsymbol{Y}$와 $\boldsymbol{Z}$의 직접적인 관계를 알 수 없지만 공통 변수 $\boldsymbol{X}$가

주어진다면 Y와 Z의 관계를 파악할 수 있게 된다.

나. 통계적 연계에서 선택적 가정

통계적 연계를 적용할 때 연구자가 선택하는 것으로 연계할 때 연계 파일의 중복을 제한할지에 대한 부분과 기준 파일의 모든 케이스를 연계할지에 대한 부분으로 구분할 수 있다. 이중에서 첫 번째인 중복여부의 결정은 통계적 연계 방법을 구분하는 기준으로도 사용되는데 연계 파일의 케이스들을 한 번씩만 연계하도록 사용을 제한할지에 대한 여부에 따라 제한적 연계 방법(constrained matching)과 비제한적 연계 방법(unconstrained matching)으로 구분된다.

데이터 연계에서 연계 파일의 케이스들이 중복되지 않도록 제한을 두기 때문에 기준 파일의 모든 케이스와 연계되는 연계 파일의 케이스들은 한 번씩만 연계된다. 따라서 연계 파일의 고유 변수 Z의 분포가 연계 후 공통 파일에서도 잘 유지되는 장점이 있다. 반면 데이터가 연계되는 횟수의 제한이 있기 때문에 기준 파일의 대상과 연결되는 연계 파일의 대상 간에 유사성이 크지 않은데도 연계가 되는 단점이 있다(김희경, 2010).

데이터 연계에서 연계 파일의 케이스들을 중복이 허용되도록 제한을 두지 않는 것은 연계 파일의 케이스들이 기준 파일의 여러 케이스와 연계하는 것을 허용하는 방법이다. 즉, 기준 파일의 케이스는 한 번씩만 연계되지만 연계 파일의 케이스는 여러 번 사용될 수 있도록 한다. 이는 횟수가 중요한 것이 아니라 기준 파일 케이스와 공통 변수가 가장 유사한 대상을 연계 파일의 케이스에서 찾아 연계시켜 유사한 대상들을 연계하는 방법이기 때문이다. 연계 파일의 한 케이스가 여러 기준 파일의 케이스들과 유사하다면 여러 번 연계될 수 있다. 따라서 기준 파일의 각 케이스와 가장 가까운 관측치가 결합되는 장점이 있다. 그러나 연계 파일의 케이스가 동일한 횟수로 연계되지 않기 때문에 고유 변수의 분포가 연계 후 통합

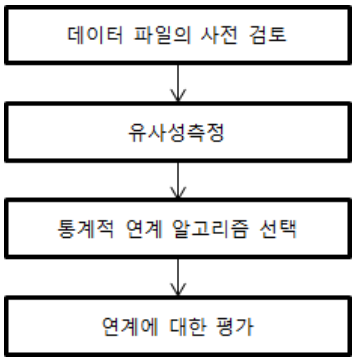
파일에서 달라질 수 있다는 단점이 있다.

일반적인 통계적 연계는 기준 파일의 입장에서 연계 파일을 잘 연계시키는 것이기 때문에 기준 파일의 각 케이스들과 가장 유사한 성향을 가진 연계 파일의 케이스를 연결시킨다. 이때 기준 파일의 각 케이스들은 한 번씩 사용하지만 연계 파일의 케이스가 기준 파일의 여러 케이스와 유사성이 크다면 여러 번 연계에 사용되는 것을 허용한다. 이는 유사성이 높은 데이터끼리 연계되어야 통합 파일을 이용한 분석의 신뢰도가 높아지기 때문이다. 이때 연계 파일의 공통 변수와 고유 변수와의 관계가 통합 파일에서도 유지될 수 있도록 해야 통합 파일에서 고유 변수 Y와 Z의 관계를 이용한 분석을 수행할 수 있다.

다음으로 기준 파일에 있는 케이스 모두를 연계할지에 대한 판단을 해야 하는데 모든 파일을 연계하면 기준 파일의 개체들의 특성 분포가 유지되기 때문에 기준 파일 모두를 연계하는 것이 가장 좋다. 그러나 모든 케이스를 연계하다보면 일부 케이스들은 유사하지도 않은데 연계될 수 있는 문제가 있다. 따라서 기준 파일의 모든 케이스를 연계하는 것을 원칙으로 하고 통합 파일을 평가할 때, 유사성이 많이 다른 케이스들이 연계된 경우는 절삭(trimmed)시킬지에 대한 판단을 해야 한다. 기준 파일의 일부 데이터를 절삭하여 연계시키면 기준 파일과 통합 파일의 대상들의 분포가 달라질 수 있기 때문에 가중값 조정 등의 다양한 방법을 고려하여 분포를 유지할 수 있는 방법을 적용할 필요가 있다.

3. 통계적 연계의 수행 과정

통계적 연계의 수행 절차는 [그림 2-6]과 같다. 먼저 자료를 검토하며 통계적 연계를 적용하는 유사성(또는 상이성)의 정도를 측정하는 측도를 선정하고, 통계적 연계 알고리즘에 따라 연계한 후 평가를 통해 최적의 통계적 연계를 하도록 한다.



[그림 2-6] 통계적 연계의 절차

가. 데이터 파일에 대한 사전 검토

데이터를 이용하여 새로운 분석이나 통계모델을 만들 때 데이터에 대한 검토를 먼저 진행한다. 통계적 연계를 할 경우에도 데이터에 대한 검토 작업을 먼저 해야 한다. 사전 검토 작업은 연계 프로세스에서 제시한 데이터 클리닝과 표준화 작업의 개념을 포함하고 있다. 통계적 연계는 동일한 대상을 연계하는 것이 아니기 때문에 가장 유사한 성향을 찾기 위해서는 두 데이터파일에 대한 사전검토 작업이 세밀하게 이뤄져야 한다.

D'Orazio et al.(2006)은 '데이터 파일에 대한 사전 검토는 연계하고자 하는 다른 두 개의 데이터 파일이 서로 같은 기준으로 맞춰 주는 것과 공통 변수를 선택하는 것이다'라고 하였다. 두 개의 데이터 파일이 서로 같은 기준으로 맞춰 주는 과정을 조화과정(harmonization step)이라고 하는데, 통계적 연계를 수행할 때 가장 먼저 진행해야 할 부분이다. 조화과정에서는 단위의 조화과정(unit harmonization step)과 변수의 조화과정(variable harmonization step)이 있다(김희경, 2010). 따라서 사전 검토 작업을 다음과 같은 3개의 단계로 요약할 수 있다.

① 단위의 조화과정

② 변수의 조화과정

③ 공통 변수 중에서 연계에 활용할 변수선택

단위의 조화과정은 조사단위가 다른 경우 이를 동일한 단위로 맞춰주는 작업이다. 예를 들어 개인단위 데이터와 가구단위 데이터가 있는 경우, 데이터의 기준이 다르기 때문에 이를 개인단위나 가구단위로 맞추어야 한다. 단위가 다르게 되면 대상이 다르기 때문에 연결도 어렵겠지만 연계시키더라도 왜곡된 결과를 도출할 수밖에 없다.

변수의 조화과정은 두 개의 데이터파일은 서로 다른 방법 또는 다르게 생성되었기 때문에 동일한 변수일지라도 표현은 다를 수 있다. 예를 들면 연령을 구간으로 표현할 때 기준 파일은 10살 간격으로 구분하고, 연계파일은 5살 간격으로 구분할 때, 기준이 다르기 때문에 동일한 규칙을 적용하기 어렵다. 따라서 이러한 경우에 다른 정보가 없다면 10살 간격을 5살 간격으로 바꾸는 것은 되지 않으므로 10살 간격의 동일한 구간으로 설정해야 한다. 이러한 과정이 변수의 조화과정이다.

다음은 공통 변수들 중에서 어떤 변수를 연계 변수로 이용할 지를 선택하는 과정이다. 선택하는 방법은 연계를 하는 방법과 목적에 따라 다소 다를 수 있지만 통계적 연계에서 기본 가정을 만족시키는 것을 원칙으로 할 수 있다. 즉, 기본 가정인 '공통 변수가 주어졌을 때 기준 파일과 연계파일의 고유 변수들이 조건부독립을 만족한다'는 조건이 성립하도록 연계변수는 두 데이터를 동일하게 포괄하는 변수가 되도록 선정한다. 다음으로 데이터 연계를 하는 이유가 기준 파일에 없는 필요한 변수를 연계 파일에서 가져오는 것이기 때문에 공통 파일에서도 분포가 유지되는 변수를 선정한다면 품질이 좋은 통합 파일을 만들 수 있다. 결론적으로 두 파일을 모두 잘 설명하는 공통 변수들 중에서 연계 파일의 고유 변수 중에서 관심의 대상이 되는 변수를 잘 설명할 수 있는 변수를 연계 변수로 선정한다.

나. 유사성 척도

통계적 연계 시 유사한 케이스들을 연결 시켜주는 기준이 되는 척도가 유사성(similarity)이다. 유사성 척도는 값이 적을수록 유사하며, 클수록 다르다는 것을 의미한다. 즉, 유사성은 연계를 위한 공통 변수들의 값으로 측정하며, 유사성 거리 척도가 0에 가까운 가장 유사한 케이스들끼리 연계하게 된다.

〈표 2-4〉 유사성 거리 척도

유사성 척도	측정식
유클리드 거리	$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T} = \sqrt{\sum_{i=1}^{p}(x_{1i} - x_{2i})^2}$
제곱 유클리드 거리	$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T = \sum_{i=1}^{p}(x_{1i} - x_{2i})^2$
맨하탄 거리	$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{i=1}^{p} \mid x_{1i} - x_{2i} \mid$
민코우스키 거리	$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left[\sum_{i=1}^{p} \mid x_{1i} - x_{2i} \mid^m\right]^{\frac{1}{m}}$
마할라노비스 거리	$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)\Sigma^{-1}(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T}$ 여기서 Σ는 $\boldsymbol{X}_1$과 $\boldsymbol{X}_2$의 분산-공분산 행렬이다
체비쉐프 거리	$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \max(\mid x_{1i} - x_{2i} \mid)$
코사인 거리	$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{i=1}^{p}(x_{1i} - x_{2i}) \Big/ \sqrt{\sum_{i=1}^{p} x_{1i}^2 \sum_{i=1}^{p} x_{2j}^2}$
고어(Gower) 거리	$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{i=1}^{p}\delta_{ijk} d_{ijk} \Big/ \sum_{i=1}^{p}\delta_{ijk}$ 여기서 범주형자료일 때 $d_{ijk} = \begin{cases} 0 & x_{ik} = x_{jk} \\ 1 & x_{ik} \neq x_{jk} \end{cases}$ 수치형 자료일 때 $d_{ijk} = \dfrac{\mid x_{ik} - x_{jk} \mid}{R_k}$ 여기서 R_k는 k번째 변수의 범위 그리고 $\delta_{ijk} = \begin{cases} 0 & x_{ik} = NA \ , \ x_{jk} = NA \ , \ x_{ik} = x_{jk} \\ 1 & otherwise \end{cases}$

유사성 척도	측정식
정확 연계	$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{i=1}^{p} D_{(12)k}$, $D_{(12)k} = \begin{cases} 1 & x_{1k} = x_{2k} \\ 0 & x_{1k} \neq x_{2k} \end{cases}$

연계에 사용되는 두 데이터의 공통 변수 중 연계 변수로 선택된 변수가 p개 있다고 가정하면 두 데이터 파일 각각의 연계 변수 벡터를 다음과 같이 나타낼 수 있다.

$$\boldsymbol{x}_1 = \{x_{11}, x_{12}, \cdots, x_{1p}\} \ , \ \boldsymbol{x}_2 = \{x_{21}, x_{22}, \cdots, x_{2p}\}$$

두 벡터들 간의 거리를 측정하여 유사성의 정도를 파악하게 되는데, 〈표 2-4〉에는 유사성을 측정하는 주요한 측정식을 제시하였다. 많은 유사성 척도가 〈표 2-4〉에 제시되어 있지만, 이들 중에서 유클리드거리, 맨하튼 거리, 정확 연계 등이 많이 사용된다. 최근에는 데이터 연계에서 고어 거리 척도도 많이 사용되고 있다.

이론적으로는 유사성 척도 중에서 공통 변수의 특성을 고려하여 가장 타당한 척도를 사용하고 가장 유사한 값을 가지는 대상들을 연계한다. 그러나 실제로 어느 척도가 가장 좋은지에 대한 판단은 어려우므로 사전에 테스트를 통해 품질이 좋은 통합 파일을 생성하도록 연계하는 척도를 선정한다.

4. 통계적 연계 방법

통계적 연계 방법은 어떠한 방법을 사용하는지에 따라서 데이터 연계의 효율에 차이가 있을 수 있으며, 이러한 통계적 연계 방법은 데이터 연계의 목적, [그림 2-5]에 제시한 통계적 연계 형태에 따라서 결정된다.

[그림 2-5]의 A와 같은 형태는 연계 변수들의 유사성 척도를 위주로 연계하고 주요한 고유 변수의 유사성 척도를 보조적으로 연계에 이용한다면, B의 형태는 결합시키고자 하는 고유 변수를 특성을 잘 반영할 수 있는 방법을 사용하여 유사성 척도를 계산하고 연계한다.

통계적 연계 방법은 목적 변수 여부에 따라 비모수적(nonparametric) 방법과 모수적(parametric)방법으로 구분된다. 모수적 방법은 목적 변수를 종속 변수로 두는 회귀(regression) 모형과 같은 통계적 모형을 추정하여 예측값을 추정하고 추정된 예측값으로 유사성을 측정하는 방법이고, 비모수적방법은 모형을 설정하지 않고 연계 변수들 간의 유사성 척도만을 이용하는 방법이다. 모수적방법과 비모수적방법 모두 매우 다양한 기법들이 많이 있지만, 여기서는 많이 사용하고 있는 기법들 위주로 설명하도록 한다. 비모수적방법의 대표적인 방법인 핫덱(hot deck)방법과 모수적 방법인 회귀식을 이용한 방법 위주로 각각 살펴보도록 한다.

가. 비모수적 연계 방법(nonparametric matching method)

1) 핫덱(hot deck) 방법

핫덱방법은 랜덤 핫덱(random hot deck), 순위 핫덱(rank hot deck), 거리 핫덱(distance hot deck) 방법이 있는데, 각 방법을 구분하는 것은 연계 변수의 척도로 〈표 2-5〉와 같다.

〈표 2-5〉 핫덱 연계 방법

구분	연계 변수의 척도
랜덤 핫덱	명목형
순위 핫덱	순서형
거리 핫덱	연속형

이중에서 연계 변수가 명목형 자료일 때 사용하는 랜덤 핫덱 방법은 기준 파일의 각 케이스에 대해 연계 파일의 케이스를 랜덤하게 선택하여 연계하는 방법이다. 랜덤하게 기준 파일의 케이스와 연계 파일의 케이스를 연결시키기 때문에 조건이 없는 방법이라 할 수 있다. 이러한 경우 모든 케이스가 연계대상이 되기 때문에 연계 조합이 매우 다양하며 연계한 결과에 대한 신뢰도 떨어진다. 그러나 랜덤 핫덱이 많이 사용되는 이유

는 하나 이상의 변수를 블로킹 변수로 선정하여 그 범주이내에서 랜덤 핫덱을 구현하기 때문이다. 블로킹변수를 잘 사용할 경우, 랜덤성 때문에 분포를 유지하는 데이터 연계 결과를 제공한다. 따라서 블로킹 변수를 어떤 것을 사용하는지가 매우 중요하다. 예를 들어 설명하면, 블로킹 변수로 성별을 사용한 경우 기준 파일의 남자 케이스는 연계 파일의 남자케이스에서 랜덤하게 선택하여 연계한다. 만약 성별과 연령대를 블로킹으로 사용하면, 기준 파일의 케이스와 동일한 성별과 연령대에 있는 연계 파일 케이스에서 랜덤하게 선택하여 연계시킨다. 너무 많은 블로킹 변수를 사용하면, 기준 파일의 케이스와 연계할 가능성이 있는 연계 파일의 케이스가 급격하게 줄어들어 랜덤성이 없어지므로 적절하게 블로킹 변수를 선정하는 것이 중요하다.

2) 거리 핫덱(hot deck) 방법

거리 핫덱 연계 방법은 연계 변수가 연속형인 경우 주로 사용하는 방법으로 최근접이웃(nearest neighbour) 방법이라고도 한다. 기준 파일의 각 케이스는 연계 변수를 이용하여 연계 파일의 각각의 케이스들과 거리를 모두 측정한 후 가장 가까운 케이스를 찾아 연계하는 방법이다. 즉, 기준변수의 첫 번째 케이스와 연계 파일의 모든 케이스와의 거리를 계산한 후 가장 거리가 짧은 케이스를 찾아서 짝(pair)으로 연결시키는 방법이다. 이는 다음과 같은 절차를 따른다.

단계 1] 기준 파일의 케이스를 랜덤하게 순서를 정한 후, 가장 첫 번째에 있는 케이스를 선정한다.

단계 2] 선정된 케이스와 연계 파일의 모든 케이스와의 유사성 거리를 측정한 후 가장 값이 적은 대상을 찾아 통합 파일로 보낸다.

단계 3] 기준 파일의 두 번째 케이스를 선정하여 연계 파일의 모든 케이스와의 유사성 거리를 측정한 후 가장 값이 적은 대상을 찾아 통합 파일로 보낸다.

단계 4] 기준 파일의 모든 케이스를 연계할 때까지 수행한다.

이때 만약 2개 이상의 케이스가 유사성 척도가 동점이 나오면, 동점의 케이스들 중에서 랜덤하게 선택한다. 공통 변수 중에서 적절한 블로킹 변수를 설정할 경우, 거리 핫덱 방법은 더 좋은 연계 결과를 제공한다.

나. 모수적 연계 방법

통계적 모형을 구축하여 연계하는 방법인 모수적 연계 방법은 공통 변수뿐만 아니라 고유 변수의 정보까지 연계 변수로 사용한다. 모수적 연계 방법은 기준 파일에 연계 파일에서 관심 있는 변수를 목적변수(종속변수)로 하여 예측모형을 추정한 후, 모형의 예측값을 추정한다. 그리고 기준 파일과 연계 파일의 모든 케이스들의 예측값을 구한 후, 유사성 값을 계산하여 가장 가까운 케이스를 연결시킨다.

모수적 연계 방법은 연계 파일의 고유 변수들 중에 가장 관심이 있는 변수를 목적 변수(종속 변수)로 하고, 공통 변수들 중에서 연계 변수를 독립변수로 이용한다. 모형은 기본적으로 회귀모형을 적용하고, 종속변수가 연속형이면 회귀분석모형, 범주형이면 로지스틱 회귀분석을 적용한다.

1) 회귀분석 연계 방법

회귀분석 연계 방법은 원칙적으로 연계 변수와 연계하고자 하는 고유 변수가 모두 연속형 자료일 때 사용하는 방법이지만, 고유 변수가 범주형 자료[10]까지 확장할 수 있다.

다음은 회귀분석 연계 방법을 적용하는 단계를 제시하였다. 이때 연계 파일의 고유 변수들 중에서 기준 파일과 연계할 주된 변수를 연계하고자 하는 것이기 때문에 연계 파일에서 회귀모형을 추정한 후 이 회귀모형을

10) 종속변수가 이분형자료이면 로지스틱 회귀분석이고, 다항자료이면 다항로지스틱 회귀분석을 적용한다.

통해 예측값을 산출하며 절차는 다음과 같다.

단계 1] 연계 파일에서 주요한 공통 변수가 $X_1, X_2, \cdots, X_p$로 p개가 존재할 때, 이들을 독립변수(또는 설명변수)로 하고, 데이터 연계에서 고유 변수들 중에서 제일 중요하면서 활용성이 높다고 판단되는 변수 Z를 종속변수로 선정하여 회귀모형을 추정한다. 추정된 회귀식은 다음과 같이 표현될 수 있다

$$\hat{Z} = a + b_1 X_1 + \cdots + b_p X_p$$

그러면 연계 파일의 j번째 케이스 n_{2_j}[11]에 대한 예측값(또는 추정값은 $\hat{z}_{2j}$로 표현된다. 그리고 계산은 다음과 같이 된다.

$$\hat{z}_{2j} = a + b_1 x_{1j} + \cdots + b_p x_{pj}$$

단계 2] 추정된 회귀식에 기준 파일의 공통 변수의 값들을 적용하여 기준 파일의 모든 케이스에 대하여 예측값을 추정한다. 이때 기준 파일의 i번째 케이스 n_{1i}에 대한 예측값 $\hat{z}_{1i}$로 나타낸다. 계산은 다음과 같이 된다.

$$\hat{z}_{1i} = a + b_1 x_{1i} + \cdots + b_p x_{pi}$$

단계 3] 기준 파일의 i번째 케이스의 예측값 $\hat{z}_{1i}$와 가장 차이가 적은 연계 파일의 예측값을 갖는 케이스를 찾는다[12]. 가장 가까운 연계된 값이 $\hat{z}_{2j}$라면 연계 파일의 j번째 케이스가, 기준 파일의 i번째 케이스와 가장 유사한 것으로 판단하여 연계를 시킨다.

단계 4] 기준 파일에서 i번째 케이스를 제외하고, 나머지 $n_1 - 1$개의 케이스의 값을 가지고, [단계 1]에서 [단계 3]까지를 반복하여, 모든 기준 파일이 연계되도록 한다.

---

11) 기준 파일을 1로 표현하고, 연계 파일을 2로 표현하였을 때, 기준 파일의 데이터의 케이스 수를 n_1개, 연계 파일의 케이스 수를 n_2개로 나타낼 때, 연계 파일의 j번째 케이스는 n_{2_j}로 표현된다.

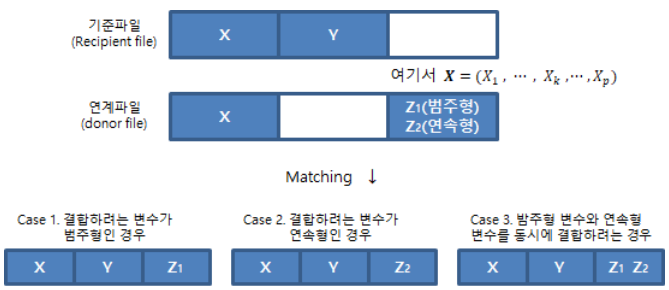
12) 이때 사용한 방법은 거리핫덱 방법으로 최근접 이웃 연계 방법(nearest neighbour matching method) 이다.

변형된 방법으로 [단계 3]에서 연계할 때 연계 파일의 값은 예측값이 아닌 원래의 값을 적용하여 거리를 계산하여 가장 차이가 적은 케이스를 연계하는 변형된 방법도 있다.

2) 혼합 연계 방법

혼합 연계 방법은 회귀분석방법을 적용할 때, 동점이 발생하면 2단계에 걸쳐 연계하여 가장 유사한 케이스를 연결하는 방법을 적용하는데 이를 혼합 연계(mixed matching)이라 한다. 혼합 연계(mixed matching)는 두 단계의 과정으로 연계하기 때문에 단계적 연계 방법이라고도 한다.

단계적 연계 방법을 설명하기 위해 이영섭 외 4인(2009)에서 제시한 데이터구조를 이용하여 설명하면, 데이터구조는 [그림 2-7]과 같다고 가정한다. [그림 2-7]에서 제시하였듯 연계하고자 하는 연계 파일의 고유변수가 범주형인 경우와, 연속형인 경우 그리고 범주형과 연속형의 혼합인 경우를 생각할 수 있다. 각각의 방법이 비슷하기 때문에 범주형인 경우만을 설명하도록 한다.



[그림 2-7] 혼합 연계에서 가정된 데이터 구조

출처 : 이영섭 외 4인(2009), 통계조사 자료와 행정 자료간의 통계적 연계 기법에 관한 연구

위의 [그림 2-7]의 3가지 경우에서 case1만을 고려하면, 설명의 편의상 고유 변수가 범주형 중에서 0과 1의 값을 갖는 이항형이라고 가정한다.

단계 1] 종속변수의 데이터 형태가 0과 1인 이항형이기 때문에, 로지스틱회귀분석을 적용하여 데이터의 유사성을 측정한다. 즉, 연

계 파일에서 결합하고자 하는 고유 변수들 중에서 주요변수인 Z_1을 종속변수로 하고, 공통 변수 $X_1, X_2, \cdots, X_p$를 독립변수로 하여 회귀모형을 추정한다. 추정된 회귀모형에 기준 파일과 연계 파일 각각의 케이스들의 예측값을 구하고, 이를 비교하여 유사성을 측정한다. 추정된 회귀식을 이용하여 기준 파일의 i번째 케이스와 연계 파일 j번째 케이스의 유사성은 다음과 같이 계산한다.

$$d^F(1i, 2j) = \mid \hat{z}_{1i} - \hat{z}_{2j} \mid$$

여기서 $\hat{z}_{1i}$는 연계 파일에서 추정된 회귀식에 기준 파일의 i번째 케이스의 공통 변수를 적합 시켜 예측한 값이고, $\hat{z}_{2j}$는 연계 파일에서 추정된 회귀식에 기준 파일의 j번째 케이스의 공통 변수를 적합시켜 예측한 값이다. 이렇게 계산된 모든 $d^F(1i, 2j)$들 중에서 가장 작은 값을 갖는 기준 파일의 i번째 케이스와 연계 파일의 j번째 케이스를 연결하여 결합한다.

만약, 기준 파일 i번째 케이스가 단계 1 에서 측정한 $d^F(1i, 2j)$를 통해 연계 파일의 j번째 케이스가 1개로 결정되면 연결시키지만, 유사성이 가장 작은 $d^F(1i, 2j)$값이 동점인 경우가 여러 케이스로 나타난다면, 여러 개의 연계 파일의 케이스가 선택된다. 이러한 경우에 단계 2 또는 단계 3으로 가서 단계 1에서 사용하지 않은 공통 변수를 이용하여 가장 유사한 값을 찾아 연계하도록 한다.

단계 1을 한 후에 사용하지 않은 공통 변수가 범주형이 있다면 단계 2로 가고, 범주형이 없고 연속형만 있다면 단계 3으로 바로 넘어간다.

단계 2] 단계 1의 로지스틱회귀분석 결과 추정된 회귀식에 포함하지 않은 공통 변수 중에서 범주형 변수들을 이용하여 다음과 같은 방법으로 유사성을 측정한다.

$$d^S(1i, 2j) = \sum_k I(x_k^{1i}, x_k^{2j})$$

여기서 x_k^{1i}는 기준 파일의 i번째 케이스의 k번째 공통 변수 값을 의미하며, x_k^{2j}는 연계 파일 j번째 케이스의 k번째 공통 변수 값을 의미한다. 또한 지시함수(indicator function) $I(\cdot)$는 다음과 같다.

$$I(a, b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b \end{cases}$$

$d^S(1i, 2j)$가 가장 작은 기준 파일 i번째 케이스와 연계 파일 j번째 케이스를 연계한다.

위에서 언급하였듯이, 공통 변수 중에서 범주형이 없다면 단계 2를 거치지 않고 단계 3으로 넘어간다. 또한 단계 2에서도 유사성 $d^S(1i, 2j)$가 같은 값이 나온다면 단계 3으로 넘어가 공통 변수 중에서 단계 1에서 추정한 로지스틱 회귀분석에 사용하지 않은 연속변수를 이용하여 유사성을 측정한다.

연속형 변수는 단위도 다르고, 값의 척도(scale)도 다르기 때문에, 값의 범위가 크고, 값의 편차가 큰 값들이 영향을 많이 미치게 된다. 따라서 단위와 무관하면서, 범위와 척도 모두 동일한 기준으로 바꾸는 표준화작업을 우선 진행하도록 한다. 표준화작업은 각 케이스의 공통 변수 X_k값에 평균 $\overline{X}_k$를 빼고 표준편차 S_k로 나누어 구하는데, 표준화를 하고 나면 평균이 0이고 표준편차가 1로 바뀐다. 표준화작업은 다음과 같다.

$$\frac{X_k^i - \overline{X}}{S_k} \sim (0, 1)$$

여기서 X_k^i는 i번째 케이스의 k번째 공통 변수를 나타낸다. 표준화작업이 끝났다면 단계 3을 진행한다.

단계 3] 단계 1에서 사용하지 않은 변수 중에서 연속형 변수를 표준화

한 후에, 이 값들을 이용하여 다음과 같은 유사성을 측정한다.

$$d^T(1i, 2j) = \sum \mid z\left(x_k^{1i}\right) - z\left(x_k^{2j}\right) \mid$$

여기서 $z\left(x_k^{1i}\right)$는 기준 파일의 i번째 케이스의 k번째 공통 변수를 표준화한 값을 의미하고, $z\left(x_k^{2j}\right)$는 연계 파일의 j번째 케이스의 k번째 공통 변수를 표준화한 값을 의미한다.

여기서, $d^T(1i, 2j)$가장 작은 기준 파일 i번째 케이스와 연계 파일 j번째 케이스를 연계한다.

5. 통계적 연계에 대한 평가

통계적 연계는 정확 연계처럼 동일한 대상을 연계하는 것이 아니라 유사한 성향을 지닌 대상들끼리 연계하는 것이기 때문에 데이터 연계 후의 생성된 공통 파일에 대한 평가가 중요하다.

통계적 연계 결과에 대한 평가에 대한 많은 학자들에 의해 다양한 연구를 통해 제안되어 왔다. 이들 제안된 평가방법들을 정확성(accuracy), 예측성(predictability)과 대표성(representation)의 문제로 압축해서 생각할 수 있다(van Pelt. 2001). 그리고 통계적 연계를 위한 기본 가정이 충족되었는지 살펴보아야 한다. 데이터 연계에서 기본 가정은 모집단이 동일하며, 공통 변수가 조건부로 주어졌을 때 기준 파일과 연계 파일의 고유 변수는 독립이라는 것이다.

판단하는 기준 중에서 정확성 또는 예측성은 연계결과가 적절한지 확인하고 평가하는 기준으로, 기대되는(또는 알고 있는) 목표와 연계 결과의 차이를 측정하여 예측력(정확성)을 판단한다. 판단하는 기준은 2가지로 생각할 수 있다. 사전에 두 데이터가 제대로 연계가 되었다면 특정한 변수의 값 또는 변수들의 관계가 어떤지에 대한 정보가 있다면 이를 기준으로 평가할 수 있다. 다음으로 여러 연계 방법들을 사용하였을 때 가장

작은 오차를 가지는 방법을 선택하는 것인데, 연속형 변수인 경우는 평균제곱오차(MSE)로 범주형 변수인 경우 오분류율(error rate)을 척도로 사용할 수 있다. 그리고 데이터 연계에 대한 대표성에 대한 평가는 공통파일이 기준 파일과 비교할 때 공통 변수들의 분포가 변화되었는가의 문제를 말한다. 데이터 연계에 대한 결과는 기준 파일의 평균과 표준편차가 유지되었는지, 통합 파일과 기준 파일의 고유 변수, 공통 변수의 분포가 유지되었는지 등을 살펴보고 판단한다.

그러나 정확성, 예측성 그리고 대표성 모두 평가하는 것은 쉽지 않다. 특히 정확성과 예측성은 실제 자료를 이용하여 평가하는 것은 불가능하다. 그래서 모의실험과 같은 방법을 적용하는데, 대부분 연계하기 전에 기준 파일을 이용하여 평가한다. 그리고 연계 후에 데이터 연계의 기본 가정을 만족하는지에 대한 평가를 나눠서 진행할 수 있다. 통계적 연계가 어려운 것이 연계 파일과 기준 파일의 직접적인 관계를 알 수 없기 때문인데, 이를 판단하기 위해 다양한 절차를 이용하여 평가하도록 한다.

가. 데이터 연계 전의 평가

데이터를 연계하기 전의 평가에서도 두 가지 정도로 구분하여 평가할 수 있다. 시기적으로는 데이터 연계를 계획할 때와 연계하기 전에 모의실험(simulation test)으로 구분된다. 데이터 연계를 계획할 때에는 기준 파일과 연계 파일이 동일한 모집단에서 추출되었는지, 아니면 동일한 분포를 가졌다고 판단할 수 있는지 살펴본다.

데이터 연계 전에 기준 파일을 가지고 모의실험을 하여 어떠한 연계 방법이 적절한지, 연계 후에 분포가 유지될 수 있는지, 최적의 연계 방법이 어떠한 방법인지를 사전에 파악하도록 한다. 이를 자세히 설명하면 다음과 같다.

단계 1] 기준 파일의 데이터를 3대 7(또는 4대 6)으로 구분하여, 적은 용량의 파일은 기준 파일로, 큰 용량의 파일은 연계 파일로

구분한다.

단계 2] 다양한 연계 방법을 적용하여 오차를 계산한다.

단계 3] 단계 1과 단계 2를 반복적으로 계산하여 제곱근 평균오차(RMSE : root mean square error)를 계산하여, 가장 작은 오차를 계산한다.

단계 4] 통합 파일의 공통 변수의 분포와 연계 파일에서 연계된 케이스들의 공통 변수들의 분포가 동일한지를 검정한다. 이는 선택편의(selection bias)가 없는지 평가한다.

단계 5] 기준 파일에서 고유 변수들 간의 관계가 통합 파일에서도 유지되는지 평가한다.

단계 6] 단계 4와 단계 5가 만족되지 않는다면 RMSE가 두 번째로 작은 방법을 적용하는 등 다양한 방법을 적용하여 만족시키도록 한다.

위의 단계를 거쳐 가장 최적의 방법과 데이터 연계에 따른 통합 파일이 원래의 기준 파일의 성질이 유지된다면 데이터 연계는 충분한 효과를 발휘할 것이라고 생각할 수 있다. 사전 평가 후에 데이터 연계를 실제로 진행하고 그에 대한 평가를 진행하게 된다.

나. 데이터 연계 후의 평가

데이터 연계를 수행한 후 평가는 매우 어렵다. 그 이유는 기준 파일과 연계 파일의 변수가 통합 파일로 있었던 적이 없기 때문에 결과를 알 수가 없기 때문이다. 따라서 데이터를 연계한 후에도 기존의 관계가 유지되는지를 살펴보는 정도로 만족해야 한다. 이러한 이유로 데이터 연계 전의 평가를 진행함으로써 평가 방법과 평가 전과 후의 간접적인 관계에 대한 신뢰를 할 수 있도록 한다.

연계에서 사용하는 데이터는 이미 존재하는 데이터이다. 따라서 기준 파일과 연계 파일의 대상들은 구성에서 차이가 대체적으로 있다. 구성하는 대상의 분포가 다르면 대상들 간의 차이가 나기 때문에 같은 성향의 대상이라고 보는 것은 무의미하다. 따라서 데이터 연계를 함으로써 선택편의(selection bias)를 없애는 것이 가장 우선적으로 해야 할 일이다.

다음으로 공통 파일이 주어졌다는 가정하에 고유 변수 Y와 Z가 독립이라는 가정을 만족해야 하는데 이는 직접적인 평가는 어렵다. 이러한 가정이 필요한 이유는 고유 변수 Y와 Z의 직접적인 연관성을 판단할 수 없기 때문에 간접적으로 판단하기 위해 공통 변수 X를 이용하여 고유 변수 Y와 Z의 간접적인 연관성을 판단을 할 수 있기 때문에 제시하고 있는 가정이다. 따라서 다음의 단계적으로 평가를 진행해야 한다.

단계 1] 데이터가 연계된 후에 통합 파일의 공통 변수와 연계 파일에서 연결된 케이스들의 공통 변수의 분포가 동일한지를 평가한다.

단계 2] 단계 1이 만족된다면 연계 파일의 공통 변수와 고유 변수의 관계가 통합 파일에서도 유지가 되는지 평가한다.

단계 1은 데이터를 연계한 후, 연계된 공통 파일연계 파일의 공통 변수의 분포와 기준 파일의 공통 변수의 분포가 동일하다고 볼 수 있는지를 검정한다. 명목자료의 경우 동질성검정을 수행하여 χ^2의 값이 유의하게 나타나지 않아야 한다.

단계 2는 공통 변수가 동질적이라고 검정이 된다면 조건부 독립의 기본 가정을 이용하여 데이터 연계전의 분포와 연계 후의 분포가 유지되는지를 검토한다. 즉, 연계 파일의 공통 변수와 고유 변수의 관계가 연계 후에 공통 변수와 고유 변수의 관계가 유지된다면 통합 파일은 원래의 특성을 유지하기 때문에 좋은 통계적 연계가 이뤄졌다고 판단할 수 있다. 조건부독립이 성립한다면 공통 변수가 주어졌을 때 고유 변수 Y와 Z의 연관성을 볼 수 있는데, 이는 X와 Y가 상관이 있고 X와 Z의 상관이

있다면 X가 주어졌을 때 Y와 Z도 상관이 있다고 판단한다.[13] 즉, 고유변수 Y와 Z의 상관여부는 직접적으로 판단할 수 없지만, 조건부독립의 가정 하에서는 Y와 Z의 관계를 공통 변수 이용하여 간접적으로 판단할 수 있다.

13) 직접적인 관계가 아니기 때문에 설명력은 다소 떨어질 수밖에 없다.

제3장

데이터 연계 사례분석

제1절

사례분석 방법

제2장에서 데이터 연계를 위한 다양한 방법의 데이터 연계 방법을 살펴보았다. 이러한 방법을 적용하여 문화·체육·관광 관련 데이터를 기준 파일로 하고, 관심 있는 정보를 포함하고 있는 파일을 연계 파일로 설정하여 통합 파일을 만들 수 있다. 따라서 데이터 연계를 통해 통합 파일을 만들어 활용하기 위해서는 실제 데이터 연계 관련된 사례들을 살펴볼 필요가 있다.

따라서 3장에서는 문화·체육·관광분야 데이터를 연계하기에 앞서 타 분야에서 데이터 연계에 활용되는 데이터와 연계 방법을 살펴보고, 다른 나라에서는 어떠한 방향으로 데이터 연계를 진행하고 있는지 사례를 분석하고, 각 사례별 주요사항을 도출하고자 한다.

사례분석은 크게 국내사례와 해외사례로 구분하며, 국내사례는 데이터 연계와 관련한 연구를 중심으로, 해외사례는 데이터 연계가 활발하게 이루어지고 있는 국가를 중심으로 데이터 연계와 데이터 관련 정책 등의 사례를 살펴보고 분석하였다.

먼저, 국내사례에서는 데이터 특성(기준 파일, 연계 파일)과 연계 방법을 중심으로 검토하였다. 여기서 데이터 특성은 행정 데이터와 조사 데이터로 구분하였는데, 여기서의 행정 데이터(Administrative data)와 조사 데이터(Survey data, Micro data)는 각각 공공기관 또는 민간이 직무상 작성·취득하여 관리하고 있는 데이터베이스와 다양한 조사방법을 활용하여 개별 개체(단위, unit)의 기록(record)을 포함하는 데이터로 정의한다. 일반적으로 행정 데이터는 공공기관의 데이터를 의미하나, 최근 통계청 등에서 민간이 관리하고 있는 공공 성격의 데이터를 행정 데이터에 포함하여 사용하고 있기 때문에, 본 연구에서도 행정데이터의 정의를 민

간영역까지 포함하는 의미로 사용하였다. 그리고 연계 방법은 정확 연계와 통계적 연계로 구분하였으며, 기준 파일과 연계 파일에 고유식별정보가 있다면 정확 연계 방법, 기준 파일과 연계 파일 중 하나라도 고유식별정보가 존재하지 않는다면 유사한 성향을 가진 개체를 연계하는 통계적 연계 방법의 활용이 가능하다.

해외사례는 데이터 연계 방법과 데이터 접근 방법에 대해서 살펴보았다. 데이터 연계 방법은 실제 데이터를 활용하여 연계를 진행한 사례를 중심으로 어떠한 특성의 데이터(조사 데이터, 행정 데이터)를 기준 파일과 연계 파일로 나누어서 데이터 연계를 진행하였는지 정리하였고, 데이터 접근 방법의 경우에는 데이터 연계 및 활용을 위한 개인정보 보호 측면에서 데이터 접근 방법에 대한 내용까지 함께 검토하였다.

〈표 3-1〉 사례분석 방법

국내사례	해외사례	
데이터 연계 연구, 활용	미국, 호주	영국, 미국, 캐나다, 뉴질랜드
해당 연구 개요 검토	해당 연구 개요 검토	연계 수행 기관 및 환경 개요 검토
데이터 연계 방법 검토	데이터 연계 방법 검토	데이터 연계 방법 검토
기준 데이터, 연계 데이터의 특성 검토	기준 데이터, 연계 데이터의 특성 검토	데이터 접근 방법 검토

제2절

국내사례

데이터 연계에 대한 국내사례는 연계 방법을 정확 연계와 통계적 연계로 구분하고, 어떠한 특성의 데이터(조사 데이터, 행정 데이터)를 기준 파일과 연계 파일로 나누어 데이터 연계 연구를 진행하였는지 구체적으로 살펴보고자 한다.

국내 사례분석은 통계적 연계 방법 자체에 대한 연구보다는 실제 데이터를 활용하여 통계적 연계를 진행한 연구 중심으로 검토하였으며, 분야의 경우에는 문화·체육·관광분야에서는 실제 데이터를 활용하여 연계를 진행한 연구가 전무한 실정으로 보건, 사회 등의 전 분야로 확대하여 개요, 데이터, 연계 방법, 결과로 나누어서 정리하였다.

〈표 3-2〉 국내사례 요약

연구자/작성기관	연구명	연계 방법	데이터 특성	
			기준 데이터	연계 데이터
최현수, 오미애(2015), 한국자료분석학회	데이터 연계방법론을 활용한 납세 및 복지수급 여부에 따른 복지인식 비교 분석	통계적 연계	한국복지패널 (조사 데이터)	재정패널 (조사 데이터)
정미옥, 최필근(2014), 통계개발원	사회조사 자료연계 방법 연구	정확 연계	2012년 사회조사 (조사 데이터)	2013년 사회조사 (조사 데이터)
		통계적 연계		
변종석 외 4인(2013), 통계개발원	다양한 출처 자료 처리 및 통계 생산방안 연구 (세부 과제1 : 자료 연계 및 통합 기법 연구)	통계적 연계	생활시간조사 (조사 데이터)	경제활동인구조사 (조사 데이터)
이영섭 외 4인(2009), 통계연구	통계조사 자료와 행정자료 간의 통계적 매칭 기법에 관한 연구	정확 연계	사업체기초조사 (조사 데이터)	국민연금자료 (행정 데이터)
		통계적 연계		

1. 데이터 연계방법론을 활용한 납세 및 복지수급 여부에 따른 복지인식 비교 분석

가. 개요

'데이터 연계방법론을 활용한 납세 및 복지수급 여부에 따른 복지인식 비교 분석'은 납세 및 복지수급 여부에 따른 집단별 복지인식에 대한 차이를 분석하고자 복지수급, 복지인식에 대한 정보를 포함하고 있는 한국복지패널조사 데이터와 납세에 대한 정보를 포함하고 있는 재정패널조사 데이터를 통계적 연계 방법을 활용하여 통합하였다.

데이터 연계로 만든 통합 파일에서 두 데이터 각각의 핵심 변수들을 활용하여 개인의 납세 및 복지수급여부에 따라 4개의 집단을 구분하고, 집단별로 근로 능력이 있는 사람에 대한 최저 생활보장 필요성, 중간 계층의 세금 수준에 대한 인식, 사회복지 확대를 위한 증세 필요성 등 다양한 복지인식에 대한 차이를 비교·분석함으로써 향후 조세 및 복지정책 마련을 위한 근거자료 제공을 주요한 목적으로 하고 있다.

나. 데이터

데이터 연계에 활용하고 있는 데이터를 살펴보면, 한국복지패널조사는 한국보건사회연구원과 서울대학교 사회복지연구소가 2006년부터 수행하고 있는 조사로 전국을 모집단으로 추출한 400개 조사구(제주도 포함)중 7,000가구를 표본으로 추출하여 패널을 구축하였다. 한국복지패널조사는 국민의 생애주기별 삶의 역동성과 욕구의 변화 등을 동태적으로 파악하여 정책의 대응성과 반응성을 높이는데 그 목적이 있으며, 조사내용은 건강 및 의료, 경제활동 상태, 사회보험 및 개인연금 수급, 생활실태 및 만족, 생활습관 등을 포함하고 있다.

재정패널조사는 조세 및 복지정책 수립을 위한 기초자료 제공을 목적

으로 2008년부터 5,000가구의 패널을 구축하여 매년 조사하고 있으며, 경제활동 상태, 소득 및 연금, 보험 관련 지출 현황, 소득세 유형 및 소득 공제 현황 등의 내용으로 구성되어 있다.

두 데이터는 각각의 연구 목적에 맞게 특정 집단을 과대표집[14]하였기 때문에 두 데이터가 완전히 동일한 분포를 가지고 있지는 않지만, 모집단은 모두 인구주택총조사로 동일하다.

다. 연계 방법

1) 공통 변수 표준화

한국복지패널조사 데이터(기준 데이터)와 재정패널조사 데이터(연계 데이터)를 연계하기 위해 먼저 공통 변수들을 표준화하였다. 연령은 10세 단위, 교육 수준은 5레벨, 경제활동 상태는 7레벨, 가구소득은 가구 균등화 소득 개념을 사용하여 공통 변수를 표준화화였다.

〈표 3-3〉 한국복지패널조사와 재정패널조사의 공통 변수 표준화

구분	한국복지패널조사 (기준 데이터)	재정패널조사 (연계 데이터)	공통 변수 표준화
성별	1. 남 2. 여	1. 남 2. 여	– 남 – 여
출생연도	_____년	_____년	10세 단위 나이
교육 수준	1. 미취학(만 7세 미만) 2. 무학(만 7세 이상) 3. 초등학교 4. 중학교 5. 고등학교 6. 전문대학 7. 대학교 8. 대학원(석사) 9. 대학원(박사)	1. 안받았음 2. 초등학교 3. 중학교 4. 고등학교 5. 전문대학교 6. 대학교 7. 대학원(석사) 8. 대학원(박사)	– 안받았음 – 초등학교 – 중학교 – 고등학교 – 전문대 이상

14) 한국복지패널조사는 표본추출 시 중위소득 60% 미만 저소득층에 전체 표본의 50%를 할당하였으며, 재정패널조사는 소득수준 상위 10% 이상 고소득층과 차상위계층 이하 저소득층에 각각 300가구를 추가로 과대표집하였다.

구분	한국복지패널조사 (기준 데이터)	재정패널조사 (연계 데이터)	공통 변수 표준화
경제활동 상태	1. 상용직임금근로자 2. 임시직임금근로자 3. 일용직임금근로자 4. 자활근로, 공공근로, 노인일자리 5. 고용주 6. 자영업자 7. 무급가족종사자 8. 실업자(지난 4주간 적극적으로 구직활동) 9. 비경제활동인구	1. 상용직 2. 임시직 3. 일용직 4. 고용원이 없는 자영업자 5. 고용원을 둔 사업주 6. 무급가족종사자	– 상용직 – 임시직 – 일용직 – 자영업자 – 고용주 – 무급가족종사자 – 비경제활동인구
개인소득	개인소득	개인소득	개인소득
가구소득	경상소득	경상소득	경상소득/ $\sqrt{(\text{가구원 수})}$

자료 : 오미애(2014), 보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안, 한국보건사회연구원

2) 공통 변수 유의성 검토

통계적 연계는 조건부 독립에 대한 강한 가정을 전제하기 때문에 연계 변수의 연계 데이터 고유 변수에 대한 유의성을 검토(Kim, Park, 2014) 해야 하므로 공통 변수와 재정패널조사 데이터의 납세여부 변수와의 관계를 설명하기 위해 로지스틱 회귀분석을 수행하였다.

한국복지패널조사 데이터와 재정패널조사 데이터의 공통 변수 중에서 연계하는데 사용한 공통 변수는 성별, 연령대, 경제활동 상태, 개인소득, 가구 균등화 경상소득이다. 이들 공통 변수를 설명변수로 하고 재정패널조사 데이터의 납세여부를 종속변수로 로지스틱 회귀분석을 실시한 결과, 설명변수는 모두 유의한 것으로 나타났다. 즉, 한국복지패널조사 데이터와 재정패널조사 데이터를 연계하는데 성별, 연령대, 경제활동 상태, 개인소득, 가구 균등화 경상소득은 납세여부 관련 변수들과 밀접한 관계를 가지고 있으며, 연계된 통합 데이터를 활용한 분석들도 유의미한 결과를 도출할 수 있을 것으로 판단하였다.

3) 데이터 연계

두 데이터를 연계할 때, 성별, 연령대, 경제활동 상태는 각 변수의 범주들로 블록화(block)하여 정확히 일치하도록 연계하였다. 그리고 두 데이터의 개인 소득과 가구 균등화 경상소득인 경우에는 가장 유사할 경우를 연계하기 위해서 마할라노비스 거리함수를 이용한 랜덤 핫덱 방법을 사용하였다.

기준 파일은 2012년 한국복지패널조사의 응답자 중 복지인식 응답자 4,185명으로 설정하였으며, 연계 파일인 2012년 재정패널조사의 응답자는 7,550명으로 이 중에서 중복을 포함하여 4,185명의 정보가 한국복지패널조사 데이터와 연계되었으며, 오류값을 제외한 통합 데이터의 최종 분석 대상은 4,182명이다.

라. 결과

연계 후 공통 변수의 분포를 살펴보면, 성별, 연령대, 경제활동 상태는 블로킹 변수로 사용하였기 때문에 한국복지패널조사 데이터와 재정패널조사 데이터 비율은 정확히 일치하였다. 그리고 개인소득과 가구 균등화 경상소득의 분포는 큰 차이가 없었으며, 교육수준, 혼인상태는 연계 전 분포보다 격차가 줄어들었다.

또한 한국복지패널조사 데이터와 재정패널조사 데이터는 모두 기초생활보장, 기초노령연금, 자녀양육수당 수급 여부 자료를 모두 포함하고 있어 통합 데이터에서 이를 비교한 결과, 데이터의 일치율은 각각 89%, 74%, 82%로 연계 결과에 대한 신뢰성은 어느 정도 확보된 것으로 판단되었다.

2. 사회조사 자료연계 방법 연구

가. 개요

'사회조사 자료연계 방법 연구'는 5개 부문씩 2년에 걸쳐서 조사되고 있는 사회조사 데이터를 연계하여 총 10개 부문 교차분석이 가능한 통합 데이터를 구축함으로써 사회조사 데이터의 분석 및 활용가치를 제고하고자 하였다. 현재는 보건과 노동 부문이 다른 해에 조사되기 때문에 개인의 건강상태에 따른 고용 상황이나 노동 만족도 등을 연관하여 분석할 수 없었으나 통합 데이터를 통해 이러한 다양한 분석이 가능하도록 하고 있다.

사회조사의 경우 표본설계의 주기가 5년으로 표본 대상 가구가 5년 동안 유지될 수 있기 때문에 데이터 연계는 정확 연계와 통계적 연계의 두 가지 방법을 적용하여 정확 연계 결과를 기준으로 통계적 연계 결과의 정합성을 평가하였다.

나. 데이터

사회조사는 국민의 삶의 질과 관련된 사회적 관심사와 주관적 의식에 관한 내용을 조사하여 삶의 수준과 사회적 변동을 파악하고 이를 정책의 기초자료로 활용하기 위해 1979년부터 통계청에서 매년 조사하고 있다. 기본항목(성별, 연령, 교육정도 등)과 10개 부문을 5개 부문(짝수년 : 보건, 교육, 안전, 가족, 환경, 홀수년 : 복지, 사회참여, 문화와 여가, 소득과 소비, 노등)씩 격년으로 조사하고 있으며, 표본 가구내 만 13세 이상 가구원을 조사 대상으로 한다.

다. 연계 방법 1 : 정확 연계

사회조사는 표본틀로 사용하고 있는 인구주택총조사의 조사 주기에 맞춰서 5년마다 표본설계를 하고 있다. 따라서 연계에 활용한 2012년과

2013년 데이터의 표본가구 모두가 동일하지는 않겠지만 상당 부분 중복될 것이므로 정확 연계를 진행하였다. 정확 연계의 결과는 통계적 연계를 평가하기 위한 기준으로 활용할 예정으로 식별 정보만을 연계 변수로 사용하였다.

사회조사의 표집 단위는 가구이고, 조사 단위는 가구원으로 서로 단위가 다르기 때문에 정확 연계는 가구 연계와 가구원 연계의 2단계로 나누어 수행하였다.

1) 가구 연계

가구는 조사구 번호, 거처 번호, 가구 번호를 조합하여 생성한 가구 id를 기준으로 연계하였다. 그리고 가구 id에 의해 연계된 가구들이 실제로 동일한 가구인지 판단하기 위해서 주소를 추가로 활용하였으며, 2012년 17,424가구와 2013년 17,664가구 중 최종적으로 12,441가구가 연계되었다.

2) 가구원 연계

가구원 연계는 가구 단위에서 연계된 12,441가구 내의 동일한 가구원을 찾는 연계로 연계 변수로 가구원 번호 대신 가구원 이름을 사용하였다. 이는 가구원 번호는 부여하는 규칙이 있긴 하지만, 매년 가구원 변동이 있을 경우 변경 되었을 여지가 있기 때문이다. 가구원 이름으로 연계 후, 성별, 나이, 혼인상태, 교육 정보, 직업 등의 가구원 정보와 가구주와의 관계, 가구 내 다른 가구원 등 가구 구성과 관련한 정보를 비교하여 최종적으로 10,584가구 내의 21,251명이 가구원이 연계되었다.

〈표 3-4〉 2012년, 2013년 사회조사 데이터 정확 연계 결과

		원 데이터	가구 연계 데이터	가구원 연계 데이터
가구 수	2012년	17,424	12,441	10,584
	2013년	17,664		
가구원 수	2012년	36,888	26,841	21,251
	2013년	37,648	27,041	

자료 : 정미옥, 최필근(2014), 사회조사 자료연계 방법 연구, 통계개발원

라. 연계 방법 2 : 통계적 연계

정확 연계된 2012년 사회조사(기준 데이터) 가구원 21,251명을 대상으로 식별 정보 없이 특성 정보만을 이용하여 2013년 사회조사(연계 데이터) 전체 가구원 37,648명과 통계적 연계를 수행하였다.

통계적 연계는 호주(ABS, 2013)와 뉴질랜드(SNZ, 2006 & 2013)에서 주로 사용하는 방법을 적용하였으며, 이 방법은 범주형 자료에서 주로 이용되는 핫덱 방법과 유사하다. 하지만 각 항목의 중요도를 고려하여 가장 유사한 데이터와 연계한다는 점에서 일반 핫덱 방법에 비해서 더 세밀한 방법이라고 할 수 있다.

1) 변수 가중값 산출

변수 가중값은 두 데이터 짝으로부터 특성 변수에 대한 값을 비교하여 구한 각 변수에 대한 중요도를 의미하며, m-확률과 u-확률을 이용하여 산출할 수 있다. m-확률은 두 데이터가 동일인일 경우, 각 변수의 값이 일치할 확률, u-확률은 두 데이터가 동일인이 아닐 경우, 각 변수의 값이 일치할 확률을 뜻한다. 예를 들어, 성별의 경우 m-확률은 1, u-확률은 0.5 근처의 값을 가지며, 주소의 경우 동일인이 아닌데도 불구하고 주소가 동일한 경우는 거의 없기 때문에 u-확률은 0에 가까운 값을 가지게 된다.

○ 일치 변수 가중값 = $\log_2(\frac{m}{u})$

○ 불일치 변수 가중값 = $\log_2(\frac{1-m}{1-u})$

* m-확률 = P(field agree|record belong to the same entity)
  : 두 데이터가 동일인일 경우, 각 변수의 값이 일치할 확률
* u-확률 = P(field agree|record belong to the different entity)
  : 두 데이터가 동일인이 아닐 경우, 각 변수의 값이 일치할 확률

'사회조사 자료연계 방법 연구'에서 m-확률은 2012년 사회조사와 2013년 사회조사를 정확 연계하여 구축한 통합 데이터에서 변수별 일치 비율, u-확률은 1을 항목의 범주 수로 나눈 값으로 산출하며, 그 값은 〈표 3-5〉와 같다.

〈표 3-5〉 2012년, 2013년 사회조사 데이터 연계 변수의 m-확률, u-확률

항목(범주 수)		역할	m-확률	u-확률	일치 변수 가중값	불일치 변수 가중값
성별	2	블로킹 변수				
연령	14	블로킹 변수				
지역	25	블로킹 변수				
혼인상태	4	연계 변수	0.972	0.250	1.96	-4.74
교육정도	5	연계 변수	0.799	0.200	2.00	-1.99
경제활동 상태	2	연계 변수	0.828	0.500	0.73	-1.54
직업	6	연계 변수	0.766	0.167	2.20	-1.83
종사상 지위	4	연계 변수	0.849	0.250	1.76	-2.31
점유형태	5	연계 변수	0.856	0.200	2.10	-2.47

자료 : 정미옥, 최필근(2014), 사회조사 자료연계 방법 연구, 통계개발원

2) 연계 가중값 산출

연계 가중값은 앞서 산출한 변수 가중값을 근거로 작성한다. 연계 가중값은 모든 변수의 일치 여부를 판단하여 일치하는 변수는 일치 변수 가중

값, 불일치하는 변수는 불일치 변수 가중값을 적용한 후, 모든 변수 가중값을 합산하여 구한다. 즉, 연계 가중값이 클수록 유사한 사람이 연계될 가능성이 높아진다.

3) 데이터 연계

성별, 연령, 지역은 블록킹 변수로 사용하였으며, 혼인상태, 교육정도, 경제활동 상태, 직업, 종사상 지위, 점유형태의 특성 변수를 이용하여 통계적 연계를 진행하였다. 데이터 연계는 가장 연계 가중값이 높은 사람을 연계하되, 연계 가중값의 최고값 동점자가 있을 경우에는 랜덤하게 한명을 선택하여 연계하였다.

마. 결과

통계적 연계 결과를 평가하기 위해, 정확 연계를 통해 연계한 21,251명의 데이터와 이 중에서 40%(8,501명)는 통계적 연계, 나머지 60%(12,750명)는 정확 연계로 구축한 데이터를 비교하였다. 각 연계 데이터에서 2012년과 2013년 항목을 교차분석하기 위해 부문별 자료들 간에 상호 연관성이 있는 건강 평가별 주말이나 휴일의 여가 활용, 전공과 직업 일치도별 고용의 안정성, 전반적인 가족 관계 만족도별 사회적 관계망 등의 주요 문항들의 교차분석을 실시하였다.

6개 주요 문항에 대한 교차분석 결과를 살펴보면, 정확 연계 데이터 중 40%를 통계적 연계 데이터로 대체한 경우에 결과가 정확하게 일치하는 것은 아니지만, 전체적인 경향은 유사하게 유지되어 동일한 결과 해석이 가능하다는 것을 확인하였다.

3. 다양한 출처 자료 처리 및 통계 생산방안 연구(세부 과제1 : 자료 연계 및 통합 기법 연구)

가. 개요

'다양한 출처 자료 처리 및 통계 생산방안 연구'의 세부 과제인 자료 연계 및 통합 기법 연구에서는 통계청에서 수행하고 있는 센서스 및 통계조사 데이터의 연계를 통해 2차 자료를 생산할 수 있도록 데이터 연계 방법을 검토하고, 시범적으로 통계청에서 수행하고 있는 경제활동인구조사와 생활시간조사를 연계하여 살펴본다.

나. 데이터

데이터 연계에 사용된 데이터는 2009년 경제활동인구조사와 생활시간조사로 연계 방안들의 비교가 목적이기 때문에 분석의 편의를 위해서 전체 자료가 아닌 서울 지역의 자료만을 분석에 사용하였다.

경제활동인구조사는 국민의 경제활동(취업, 실업, 노동력 등) 특성을 파악하기 위한 조사로 만 15세 이상의 성인을 대상으로 하고 있으며, 일에 관한 사항, 구직에 관한 사항, 이전 직장(일)에 관한 사항, 일에 관한 사항을 포함하고 있다.

생활시간조사는 국민들이 하루 24시간을 어떤 형태로 보내고 있는지를 파악하여 국민의 생활방식과 삶의 질을 측정할 수 있는 기초 자료 제공을 목적으로 만 10세 이상의 가구원을 대상으로 5년 주기로 조사하고 있다. 조사 내용으로는 가구 관련 사항(주택 종류 및 점유 형태, 가구소득 등), 개인 관련 사항(경제활동 상황, 교육 정도, 개인소득 등)과 함께 시간일지가 포함되어 있으며, 시간일지는 하루 24시간의 활동형태를 10분 간격으로 주 행동과 동시 행동, 함께한 사람, 행위 장소, 이동 수단 등을 작성하도록 하고 있다.

다. 연계 방법

1) 공통 변수 검토

경제활동인구조사는 만 15세 이상, 생활시간조사는 만 10세 이상을 대상으로 하므로 데이터 연계를 위해서 생활시간조사의 만 15세 미만인 자료는 제외하였다. 그리고 공통 변수 중에서 성별, 혼인상태 등의 값이 없는 경우를 제외하고 최종 데이터는 경제활동인구조사 109,232명, 생활시간조사 4,712명이 포함되었다. 그리고 공통 변수를 표준화하여 각 조사별로 공통 변수의 분포를 살펴보면, 부업시간을 제외한 변수가 두 데이터 간에 차이가 거의 없는 것으로 파악되었다.

〈표 3-6〉 경제활동인구조사, 생활시간조사 공통 변수의 분포

공통 변수		경제활동인구조사 (기준 데이터)	생활시간조사 (연계 데이터)
성	남	50,916(46.61%)	2,212(46.94%)
	여	58,316(53.39%)	2,500(53.06%)
혼인 상태	미혼	33,282(30.47%)	1,328(28.18%)
	배우자 있음, 사별, 이혼	75,950(69.53%)	3,384(71.82%)
교육 정도	무학, 초등학교, 중학교	20,488(18.76%)	834(17.7%)
	고등학교	38,191(34.96%)	1,666(35.36%)
	대학(4년제 미만, 4년제 이상)	44,775(40.99%)	1,948(41.34%)
	대학원 석사과정, 박사과정	5,778(5.29%)	264(5.6%)
연령	평균	42.99	43.06
	표준편차	16.55	16.85
	최소값	15.00	15.00
	1사분위수	30.00	30.00
	중위수	42.00	42.00
	3사분위수	54.00	54.00
	최대값	99.00	93.00
주업시간	평균	26.54	26.83
	표준편차	25.71	26.83
	최소값	0.00	0.00
	1사분위수	0.00	0.00

공통 변수		경제활동인구조사 (기준 데이터)	생활시간조사 (연계 데이터)
	중위수	33.00	28.00
	3사분위수	48.00	50.00
	최대값	109.00	105.00
부업시간	평균	0.04	0.23
	표준편차	0.81	2.16
	최소값	0.00	0.00
	1사분위수	0.00	0.00
	중위수	0.00	0.00
	3사분위수	0.00	0.00
	최대값	47.00	48.00

자료 : 변종석 외 4인, 다양한 출처 자료 처리 및 통계 생산방안 연구, 통계개발원

2) 공통 변수 유의성 검토

공통 변수를 활용한 통합 데이터의 구축은 경제활동인구조사에서 얻어진 경제활동 관련 변수를 생활시간조사에 활용하기 위해서이며, 이를 위해서는 공통 변수의 경제활동인구조사의 경제활동 관련 변수에 대한 설명력을 살펴보아야 한다. 즉, 경제활동인구조사의 경제활동 관련 변수가 공통 변수에 의해 충분히 설명되어야만 공통 변수를 이용한 통합 데이터의 분석이 타당성을 가지며, 이와 함께 조건부 독립성 조건이 근사적으로 만족된다.

경제활동 관련 변수 중에서 실업률과 공통 변수와의 관계를 살펴보기 위해서 경제활동인구조사의 실업률을 종속변수로, 성, 혼인 상태, 교육정도, 연령, 주업시간, 부업시간을 설명변수로 하는 로지스틱 회귀분석을 수행하였다. 로지스틱 회귀분석 결과, 공통 변수가 모두 높은 설명력을 갖는 것으로 나타났다. 즉, 통합 파일의 공통 변수는 경제활동인구조사의 경제활동 여부와 밀접한 관계를 나타내고 있으며, 공통 변수를 이용한 통합 파일을 활용한 분석들은 적절할 것으로 판단하였다.

3) 데이터 연계

경제활동인구조사와 생활시간조사의 데이터 연계는 성별, 혼인 상태, 교육 정도는 각 변수의 범주들로 블록화(block)하여 정확히 일치시켰으며, 연령, 주업시간, 부업시간을 공통 변수로 활용하였다.

데이터 연계는 거리 핫덱과 랜덤 핫덱을 고려하였으며, 두 비모수적인 연계 방법만을 비교하기 위해 9가지 세부 방법으로 다시 구분하였다.

〈표 3-7〉 고려된 9가지 연계 방법

구분	연계 방법
1	거리 핫덱, 맨하튼 거리 함수
2	거리 핫덱, 마할라노비스 거리 함수
3	거리 핫덱, 비유사성 지수
4	랜덤 핫덱, 맨하튼 거리 함수, Exact 옵션
5	랜덤 핫덱, 마할라노비스 거리 함수, Exact 옵션
6	랜덤 핫덱, 비유사성 지수, Exact 옵션
7	랜덤 핫덱, 맨하튼 거리 함수, Span 옵션
8	랜덤 핫덱, 마할라노비스 거리 함수, Span 옵션
9	랜덤 핫덱, 비유사성 지수, Span 옵션

자료 : 변종석 외 4인, 다양한 출처 자료 처리 및 통계 생산방안 연구, 통계개발원

라. 결과

9가지 연계 방안에 대한 평가는 두 가지 관점에서 진행하였다. 첫 번째는 통합 파일에서의 경제활동인구조사의 분포와 연계 전의 경제활동인구조사의 분포를 비교하는 것이며, 두 번째는 경제활동인구조사와 생활시간조사에 모두 포함되어 있는 공통 변수이지만 연계에서 활용하지 않은 변수들의 값을 비교하는 것이다.

먼저, 경제활동인구조사 주요 문항의 연계 전, 후의 분포를 살펴보면, 취업기간의 경우에는 9가지 연계 방법이 비슷한 결과로 연계 전, 후의 분포가 유사하게 나타났다. 그리고 실업자 수는 거리 핫덱, 실업률은 랜덤 핫덱이 연계 전, 후의 분포가 유사한 것을 확인할 수 있었다. 주요 문항들

의 분석 결과, 경제활동인구조사와 생활시간조사의 데이터 연계에서는 분석 결과의 편향이 커지는 것을 최소화하기 위해서는 랜덤 핫덱, 분석 결과의 변동을 줄이기 위해서는 거리 핫덱이 적절할 것으로 판단하였다.

다음으로 경제활동인구조사와 생활시간조사에 모두 포함되어 있지만 연계 변수로 활용하지 않은 '지난 1주일간 1시간 이상 노동 여부'와 '종사상 지위' 변수를 실제 값과 연계된 데이터에서의 값을 비교한 결과, 지난 1주일간 1시간 이상 노동 여부는 약 99% 내외, 종사상 지위는 70% 내외로 9가지 연계 방법들의 차이는 거의 존재하지 않았다.

〈표 3-8〉 연계 결과 비교 : 지난 1주일간 1시간 이상 노동 여부

구분	연계 방법	일치 여부	
		일치	불일치
1	거리 핫덱, 맨하튼 거리 함수	99.75	0.25
2	거리 핫덱, 마할라노비스 거리 함수	99.83	0.17
3	거리 핫덱, 비유사성 지수	99.83	0.17
4	랜덤 핫덱, 맨하튼 거리 함수, Exact 옵션	99.75	0.25
5	랜덤 핫덱, 마할라노비스 거리 함수, Exact 옵션	99.77	0.23
6	랜덤 핫덱, 비유사성 지수, Exact 옵션	99.81	0.19
7	랜덤 핫덱, 맨하튼 거리 함수, Span 옵션	99.75	0.25
8	랜덤 핫덱, 마할라노비스 거리 함수, Span 옵션	99.75	0.25
9	랜덤 핫덱, 비유사성 지수, Span 옵션	99.85	0.15

자료 : 변종석 외 4인, 다양한 출처 자료 처리 및 통계 생산방안 연구, 통계개발원

4. 통계조사 자료와 행정 자료 간의 통계적 매칭기법에 관한 연구

가. 개요

'통계조사 자료와 행정 자료 간의 통계적 연계기법에 관한 연구'에서는 조사 데이터와 행정 데이터를 연계하여 양질의 통합 데이터를 생성하기 위한 효율적인 통계적 연계 방법 및 평가 방법에 대한 연구를 수행하였다.

그리고 사업체기초통계조사와 국민연금자료의 정확 연계와 통계적 연계를 수행하고 그 결과를 평가함으로써 사업체기초통계조사의 종사자 수를 국민연금자료의 가입자 수로 대체하여 사용할 수 있는지를 검토하였다.

나. 데이터

데이터 연계에는 사업체기초통계조사와 국민연금자료가 활용되었으며, 사업체기초통계조사, 국민연금자료 모두 지역을 서울로 한정하여 사용하였다.

사업체기초통계조사는 전국의 지역별 사업체의 규모 및 분포를 파악하여 관련 정책 수립 및 평가 등의 기초 자료로 활용되며, 사업체 대상의 통계조사의 모집단 명부로 사용되는 조사이다. 사업체기초통계조사에는 사업체의 소재지, 사업장 형태, 업종, 종사자 수, 연간 매출액 등의 내용을 포함하고 있다. 국민연금자료는 가입자, 징수, 급여, 기금 현황 등을 파악하기 위한 행정 자료로 소재지, 사업장 형태, 업종, 가입자 수, 징수액 등의 내용으로 구성되어 있다.

다. 연계 방법 1 : 정확 연계

사업체기초통계조사 데이터(741,229개)와 국민연금자료(223,186개)에 모두 포함되어 있는 사업자등록번호를 연계 변수로 정확 연계를 진행하였다. 하지만 사업자등록번호가 동일하더라도 대표자 성명이나 사업체명이 다른 경우가 존재해 추가적으로 대표자 성명을 연계 변수로 추가하였다.

그 결과, 124,826개가 정확 연계 되었으며, 사업체기초통계조사 데이터의 종사자 수와 국민연금자료의 가입자 수가 일치하는 비율을 살펴보면, 종사자 수가 증가함에 따라 일치하는 비율이 크게 감소하는 경향으로 나타나 연계의 정확성 측면에서 검토가 필요하다고 판단하였다.

라. 연계 방법 2 : 통계적 연계

통계적 연계는 사업체기초조사를 기준 데이터(741,229개)로 하고 국민연금자료(223,186개)를 연계 데이터로 하며, 사업체기초조사의 종사자수를 기준 데이터의 고유 변수, 국민연금자료의 가입자 수를 연계 데이터의 고유 변수로 하여 연계하였다. 하지만 두 데이터 간의 연계를 위한 공통 변수가 부족하고 기준 데이터가 연계 데이터보다 크다는 문제점을 가지고 있어 정확 연계로 구축된 통합 데이터 124,826개를 기준 데이터와 연계 데이터로 나누어 통계적 연계를 수행하였다. 이는 통계적 연계 이후에 연계에 대한 평가가 용이하다는 장점을 가지고 있다.

소재지, 업종, 사업형태의 3가지 범주형 변수를 연계 변수로 통계적 연계를 수행하였다. 그리고 이 경우에는 연계 변수로 활용할 수 있는 변수가 적고, 연계 변수가 모두 범주형으로 데이터 연계 시 많은 동점자가 발생할 가능성이 크므로, 데이터 연계 방법 중 활용 가치가 큰 랜덤 핫덱 방법을 활용하였다.

마. 결과

통계적 연계 결과는 대표성과 정확성 측면에서 평가해 볼 수 있다. 연계 데이터인 국민연금자료에서의 가입자 수의 분포와 통계적 연계로 구축된 통합 데이터에서의 가입자 수의 분포를 평균, 중위수, 표준편차, 평균 절대값 오차(Mean absolute error : MAE) 등의 기준으로 비교해 본 결과 두 분포가 거의 유사하게 나타나 대표성 측면에서 데이터 연계 결과가 타당한 것으로 보았다.

또한 정확 연계된 데이터를 기준으로 통계적 연계 데이터의 정확성을 비교해 본 결과, 실제 값과 연계 데이터의 값이 정확하게 일치하는 비율은 31.94%에 불과하나 그 차이가 5 이하인 경우가 전체의 74.97%로 비교적 높은 것으로 나타났다. 따라서 통계적 연계의 경우 적은 수의 연계 변수를

사용하여 랜덤 핫덱 방법을 적용하였으나 대표성과 정확성 측면에서 신뢰할 만한 결과를 얻은 것으로 판단하였다.

〈표 3-9〉 랜덤 핫덱 방법 적용 결과(예시)

[사업체기초조사(기준 데이터)]

연계 변수1 (소재지)	연계 변수2 (업종)	연계 변수3 (사업형태)	유일변수 (종사자 수)
신정3	교육 서비스업	법인	79
신정3	교육 서비스업	법인	85
신정3	교육 서비스업	법인	85
신정3	교육 서비스업	법인	7
신정3	교육 서비스업	법인	70

+

[국민연금자료(연계 데이터)]

연계 변수1 (소재지)	연계 변수2 (업종)	연계 변수3 (사업형태)	유일변수 (가입자 수)
신정3	교육 서비스업	법인	18
신정3	교육 서비스업	법인	6
신정3	교육 서비스업	법인	5
신정3	교육 서비스업	법인	16
신정3	교육 서비스업	법인	16

[통합 데이터]

연계 변수1 (소재지)	연계 변수2 (업종)	연계 변수3 (사업형태)	유일변수 (종사자 수)	연계된 변수 (가입자 수, A)	정확 연계 데이터 (가입자 수, B)	차이 | A - B |
신정3	교육서비스업	법인	79	16	18	2
신정3	교육서비스업	법인	85	5	6	1
신정3	교육서비스업	법인	85	6	5	1
신정3	교육서비스업	법인	7	16	16	0
신정3	교육서비스업	법인	70	16	16	0

자료 : 이영섭 외 4인(2009), 통계조사 자료와 행정 자료간의 통계적 연계 기법에 관한 연구, 통계개발원

제3절

해외사례

해외사례에서는 데이터 연계 방법과 데이터 접근 방법에 대한 사례로 구분하였다. 데이터 연계 방법에 대한 사례에서는 국내사례분석처럼 실제 데이터를 활용하여 연계를 진행한 사례들 중에서 몇 나라의 국가사례를 중심으로 어떠한 특성의 데이터(조사 데이터, 행정 데이터)를 기준 파일과 연계 파일로 나누어 데이터 연계를 진행하였는지 정리하였다.

데이터 연계가 활성화되고 있는 국가에서는 정부, 공공기관이 보유한 데이터간의 연계뿐만 아니라 민간에서 보유한 데이터와도 상호 연계할 수 있는 시스템 또는 환경을 제공해 주고 있다. 이러한 데이터 연계는 각 국가별 통계청에서 수행할 수도 있고, 별도의 전담부서를 두고 진행할 수도 있다. 데이터 접근 방법 사례에서는 데이터 연계 수행 기관을 중심으로 데이터 연계 방법과 함께 개인정보 보호와 연관이 있는 데이터 접근에 대한 내용까지 같이 살펴보았다.

〈표 3-10〉 해외사례 요약

	국가	연구자/수행기관	연계 방법	데이터 특성	
				기준 데이터	연계 데이터
데이터 연계	미국	Muennig(2011)	통계적 연계	사회조사 (조사 데이터)	국민사망자료 (행정 데이터)
	호주	호주 통계청 (Australian Bureau of Statistics)	통계적 연계	2006 센서스 (조사 데이터)	2011년 센서스 (조사 데이터)
	국가	수행기관	데이터 접근 방법		
데이터 접근 방법	영국	행정 데이터연구센터 (ADRC)	접근이 허용된 시설 내에서 가능 (영국 국가 통계청, 북아일랜드 통계연구소 등)		
	미국	미국 인구조사국 (U.S. Census Bureau)	접근이 허용된 시설 내에서 가능 (연방통계연구데이터센터)		
	캐나다	캐나다 통계청 (Statistics Canada)	공공기관 및 캐나다 통계청이 협약하는 고등교육기관 내에서 가능		

	국가	수행기관	데이터 접근 방법
			(공개 사용이 가능한 마이크로데이터) 접근이 허용된 시설 내에서 가능 (연구데이터센터)
	뉴질랜드	뉴질랜드 통계청 (Statistics New Zealand)	접근이 허용된 시설 내에서 가능 (보안 데이터랩))

1. 데이터 연계 방법 사례

가. 미국[15)]

1) 개요

어떠한 사회적 환경이 개인 건강에 어떠한 영향을 주는지를 파악하여 국민의 건강 증진에 도움이 되는 자료를 생산하고자 2011년 미국에서는 사회조사(General Social Survey) 데이터와 국민사망자료(National Death Index)를 연계하는 작업을 진행하였으며, 두 자료의 데이터 연계를 통해 사망 원인과 사회적 지위, 신념, 사회적 인식 간의 관계를 파악할 수 있게 되었다.

미국의 사회조사는 1972년 시카고대학교 여론조사센터에서 처음 진행되어 1993년까지는 매년, 1994년부터는 2년 주기로 조사가 진행되고 있다. 조사 내용은 매년 동일한 내용의 핵심 내용과 사회적 관심사에 따른 특별히 선정된 관심 내용으로 구성되어 있으며, 18세 이상 인구를 조사 대상으로 약 1,500여명을 조사하고 있다(정미옥, 최필근, 2014).

2) 데이터 연계

데이터는 1978년에서 2002년까지의 사회조사 데이터(연계 데이터)를 1979년에서 2008년까지의 국민사망자료(기준 데이터)에 연계하였다.

15) Muennig(2011) 내용을 참조하였다.

① 정확 연계

사회조사 데이터와 국민사망자료는 먼저 사회보장번호를 기준으로 정확 연계를 진행하였다. 사회보장번호는 개인을 식별하는 항목으로 정확 연계에 활용할 수 있는 중요한 항목이나, 사회조사에서는 사회보장번호 작성이 필수가 아니어서 응답 자료 중에서 21% 만이 사회보장번호를 포함하고 있다. 정확 연계로 구축된 데이터는 통계적 연계의 결과를 평가하고 정확도를 향상을 위한 작업에 도움을 줄 수 있다.

② 통계적 연계

다음으로 각 데이터 주요 변수의 일치 여부를 점수화하여 데이터를 연계하는 통계적 연계를 진행하였다. 연계 변수들 간의 일치 여부를 비교하여 변수별 확률점수를 산정하였으며, 연계 변수로 사용된 변수는 사회보장번호, 성과 이름, 생년월일, 성별, 인종, 출생지역 등이다.

변수별 확률점수는 각 자료에서의 특성별 발생 비율의 역수에 $\log_2$를 취해 산정하며, 예를 들어, 남성의 비율이 46.3%이며 남성에 대한 성별의 확률점수는 $\log_2(\frac{1}{0.463})$이 된다. 그리고 연계 변수들 간에 값이 일치하면 양(+)의 부호, 일치하지 않으면 음(−)의 부호를 적용하여 연계 변수들 간의 점수를 모두 합산하여 개인별 확률점수를 산출한다.

$$\begin{aligned}\text{개인별 확률점수} = \; & W_{\text{사회보장번호}} + W_{\text{이름}\times\text{성별}\times\text{출생연도}} \\ & + W_{\text{중간이름}\times\text{성별}}{}' + W_{\text{인종}} + W_{\text{성}} + W_{\text{성별}} \\ & + W_{\text{결혼상태}\times\text{성별}\times\text{나이}}{}' + W_{\text{출생일}} + W_{\text{출생월}} \\ & + W_{\text{출생연도}} + W_{\text{출생지역}} + W_{\text{거주지역}}\end{aligned}$$

개인별 확률점수에 따라 5단계의 등급으로 구분하여 연계 유무를 결정한다. 1등급은 사회보장번호(9자리 숫자), 이름, 중간이름, 성, 생년월일, 출생지역 등이 모두 일치하며, 2등급은 사회보장번호의 9자리 중에서 7자리 이

상이 일치하고, 1등급의 나머지 변수 중에서 1개 이상이 일치하지 않는다. 3등급은 사회보장번호는 알 수 없지만, 이름, 중간이름, 성, 생년월일, 성별, 인종 등에서 8개 이상의 변수가 일치하며, 4등급은 8개 미만의 변수가 일치한다. 그리고 마지막으로 5등급은 사회조사에 사회보장번호가 기입되어 있으나, 국민사망자료에 일치하는 사회보장번호가 없는 자료를 의미한다.

1등급에 속한 자료는 두 데이터가 정확 연계되어 사망으로 분류되며, 5등급에 속한 자료는 생존으로 분류되어 연계에서 제외하였다. 그리고 2~3등급에 속한 자료는 정확하게 연계할 확률은 최대화하면서 잘못 연계할 확률은 최소화하는 지점이 결정되도록 하는 컷오프(cut-off) 점수를 산정하여 최종 연계 여부를 결정하였다.

연계 결과를 살펴보면, 사회조사 데이터 32,830명 중 국민사망자료와 연계된 데이터는 9,271명이로 이 중에서 6,504명(약 70%)은 정확 연계, 2,767명(약 30%)은 통계적 연계 방법으로 연계되었다.

〈표 3-11〉 미국 사회조사, 국민사망자료 연계 데이터의 연계 비율

연도	표본 수	사망자 수	사망(연계) 비율(%)
1978년	1,509	689	45.7
1980년	1,274	583	45.8
1982년	1,715	746	43.5
1983년	1,349	550	40.7
1984년	1,411	552	39.1
1985년	1,439	632	43.9
1986년	1,363	558	40.9
1987년	1,725	626	36.9
1988년	1,451	513	35.4
1989년	1,486	497	33.5
1990년	1,346	427	31.7
1991년	1,486	454	30.5
1993년	1,547	338	21.9
1994년	2,949	587	19.9
1996년	2,835	478	16.9

연도	표본 수	사망자 수	사망(연계) 비율(%)
1998년	2,712	404	14.9
2000년	2,650	357	13.5
2002년	2,583	280	10.8
전체	32,830	9,271	28.2

자료 : Muennig(2011)

나. 호주[16)]

1) 개요

호주 통계청은 센서스 자료의 활용성을 제고하기 위해 2006년에 센서스 종단 데이터(Australian Census Longitudinal Dataset: ACLD)를 구축하였다. 센서스는 많은 비용과 시간이 투입되는 조사로 자료의 활용성을 제고하기 위해 센서스간의 연계를 통해 자료의 활용성을 강화하였다. 먼저 2005년 센서스 시범 예행조사(dress rehearsal)와 2006년 센서스를 연계하여 연계 방법에 대한 실현 가능성을 확인하고(Solon, 2009), 이 방법을 2006년과 2011년 센서스 연계에 적용하였다.

2) 데이터 연계

먼저, 두 자료간의 통계적 연계를 위해 각 자료에 포함되어 있는 주소, 이름, 나이 등의 공통 변수를 표준화하는 작업을 진행하였다. 그리고 데이터 연계는 두 자료에 개인 식별 자료가 포함되어 있지 않기 때문에 통계적 연계 방법을 적용하였으며, 연계 변수로는 이름, 주소, 생년월일, 성별, 출생 국가, 연령, 결혼 여부 등이 활용되었다.

데이터 연계에는 m-확률(두 자료가 동일인일 경우, 각 변수의 값이 일치할 확률)과 u-확률(두 자료가 동일인이 아닐 경우, 각 변수의 값이 일치할 확률)을 이용하였다. 두 확률을 통해 각 자료 모든 변수의 일치

16) Australian Bureau of Statistics(2013) 내용을 참조하였다.

여부를 판단하여 변수 가중값을 산출하고, 이들 변수 가중값을 모두 합산하여 연계 가중값을 산출한 뒤, 연계 가중값이 높은 사람을 연계하였다.

2005년 센서스 시범 예행조사(dress rehearsal)와 2006년 센서스의 통계적 연계의 결과 평가에는 3가지 자료가 활용되었다. 먼저 Gold 방법은 이름, 주소, 특성 변수(성, 나이 등), Silver 방법은 이름과 주소를 사용하지 않고 해쉬 값(hash values)[17], 특성 변수, Bronze 방법은 특성 변수만을 사용하여 자료를 연계하였다. Gold, Silver 및 Bronze 방법의 상대적 적합성을 조사한 결과, 이름과 주소 정보를 사용하여 연결하면 연계의 품질이 좋았지만, Bronze 방법은 종단 분석에 충분한 품질의 데이터를 구축할 수 있다는 결론을 지었으며, 2006년과 2011년 센서스 연계에도 Bronze 방법을 적용하였다.

연계 결과 2006년 센서스에서 추출한 979,661명 중에서 800,759명(81.7%)이 연계되었으며, 연계 결과는 〈표 3-12〉와 같다.

〈표 3-12〉 2006년과 2011년 센서스 연계의 연계율

구분		2006년 센서스 자료	센서스 종단 데이터	연계율
성	남성	480,285	390,487	81.3
	여성	499,372	410,274	82.2
연령	15세 미만	194,017	170,834	88.1
	15세 ~ 19세	66,247	51,220	77.3
	20세 ~ 24세	66,512	49,327	74.2
	25세 ~ 29세	62,249	48,642	78.1
	30세 ~ 39세	140,271	117,655	83.9
	40세 ~ 49세	142,911	123,946	86.7
	50세 ~ 59세	126,285	108,962	86.3
	60세 ~ 69세	86,385	71,906	83.2
	70세 ~ 74세	31,004	23,678	76.4
	75세 이상	63,784	34,586	54.2
토착 상태	비원주민	942,253	775,419	82.3

17) 이름과 성을 조합하여 숫자로 변환한 값을 의미한다.

구분		2006년 센서스 자료	센서스 종단 데이터	연계율
	원주민	19,694	13,340	67.7
	토래스 해협 섬 주민	1,449	923	63.7
	원주민 & 토래스 해협 섬 주민	839	543	64.7
	명시되어 있지 않음	15,416	10,530	68.3
거주 지역	New South Wales	323,136	263,369	81.5
	Victoria	244,095	203,668	83.4
	Queensland	192,606	154,013	80.0
	South Australia	75,481	62,239	82.5
	Western Australia	95,795	77,921	81.3
	Tasmania	23,787	19,583	82.3
	Northern Territory	8,469	6,226	73.5
	tralian Capital Territory	16,186	13,680	84.5
전체		979,661	800,759	81.7

자료 : Australian Bureau of Statistics(2013)

2. 데이터 접근 방법 사례

가. 영국[18)]

1) 개요

영국의 행정 데이터연구네트워크(Administrative Data Research Network : ADRN)는 사회, 경제 연구자들에게 안전한 환경에서 연계된 비식별 행정 데이터에 대한 접근을 제공하기 위한 네트워크로 이를 통해 사회에 대한 지식과 이해를 증진시키고, 정책 결정자에게 도움을 주고자 한다.

ADRN은 행정 데이터를 직접 보유하고 있는 것이 아니며, ADRN은 연구자들을 대신하여 요청할 데이터의 범위를 검토하고, 데이터 보유기

18) https://www.adrn.ac.uk/ 내용을 참조하였다.

관과 협의를 진행하며, 신뢰할 수 있는 제3자(TTP)를 통해 데이터를 연계하고, 연계된 비식별 데이터에 접근할 수 있는 보안 환경을 제공한다.

2) 데이터 연계 및 접근 방법

개인정보 보호를 위해 서로 다른 데이터의 연계는 다음 절차를 따라 이루어진다.

① 1 단계 : 연구 제안서의 승인이 이루어지고 연구자가 훈련을 받은 후, ADRN은 프로젝트와 관련된 데이터의 제공을 위해 데이터 보유기관과 협상을 진행한다.

② 2 단계 : 데이터 보유기관(데이터를 보유한 정부 부처)은 각 레코드에 고유한 참조번호(reference number)를 부여하며, 이름, 생년월일 등 사람들을 직접 식별할 수 있는 식별자를 분리한다.

③ 3-1 단계 : 데이터 보유기관들은 식별정보를 고유 참조번호로 대체한 데이터를 행정 데이터연구센터(Administrative Data Research Centres : ADRC)중 하나로 보낸다.

④ 3-2 단계 : 동시에, 직접 개인을 식별할 수 있는 정보는 각 레코드의 고유 참조 번호와 함께 신뢰할 수 있는 제3자(TTP)에게 보낸다. 이때, 연구 데이터는 포함되지 않는다.

⑤ 4 단계 : TTP는 고유 참조번호와 식별정보를 사용하여 이 정보들을 연계한다. 그리고 개인 식별정보를 삭제한 후 연계된 고유 참조번호만을 남긴다.

⑥ 5 단계 : 색인키(index key)는 서로 다른 데이터 집합에서 어떤 참조번호가 같은 사람과 관련되는지를 보여주며, TTP는 색인키를 ADRC에 보낸다.

⑦ 6 단계 : ADRC는 색인키를 사용하여 서로 다른 기관들이 보내온 데이터 집합을 연계한다. 그리고 색인키와 참조번호를 지운 후에 연구자에게 연계된 데이터에 대한 접근을 제공한다.

이 시스템은 개인 식별정보와 연구 데이터의 분리를 유지한다. 즉, TTP는 단지 식별정보와 참조번호만을 볼 수 있으며, 연구 데이터를 볼 수 없다. ADRN 직원은 단지 연구 데이터와 색인키만을 볼 수 있을 뿐, 개인 식별정보는 볼 수 없다. 연구자는 보안시설에서 자신이 요청한 데이터만을 볼 수 있으며, 색인키와 개인 식별정보는 볼 수 없다. 여기서 신뢰할 수 있는 제3자(TTP)는 데이터 연계를 위한 보안시설을 가지고 있는 조직을 의미하는데, 국가통계청(Office for National Statistics : ONS) 이나 북아일랜드 통계연구소(Northern Ireland Statistics and Research Agency : NISRA) 등이 이에 해당한다. 만일 어떤 기관이 TTP로서의 역할과 ADRC의 파트너 역할을 동시에 수행하고 있다면, 두 업무는 철저하게 분리되어야 한다. 해당 조직의 서로 다른 부서에서 완전히 다른 직원들이 각각의 역할을 수행해야 한다. 이러한 역할 분리를 통해 데이터 기밀성이 유지될 수 있도록 기술적, 운영적 통제를 해야 한다.

나. 미국[19)]

1) 개요

미국 인구조사국(U.S. Census Bureau)은 행정 데이터를 광범위하게 활용하여 미국 인구와 경제에 관한 통계를 작성하고 있으며, 1940년대부터 외부 데이터를 활용해 왔다.

인구조사국은 평가자와 정책 분석가의 행정 데이터 접근을 개선하기 위해 데이터 연계 기반(Data Linkage Infrastructure)을 확대하고 있다. 인구조사국은 통계 목적으로 정부 기록에 접근하고 데이터를 수집할 수 있는 권한을 가지고 있으며, 데이터 연계 기반은 정책 분석과 연구를 위한 데이터를 확보하고 안전하게 분석적인 접근을 할 수 있도록 한다. 데이터

19) https://www.census.gov/about/adrm/linkage.html의 Data Linkage Infrastructure 내용을 참조하였다.

연계 기반은 정부 기록을 확보하기 위한 규약, 파일에 대한 메타데이터와 문서화, 연계 방법 및 결과, 데이터 저장 및 제공, 기존 연계 결과의 공유 등을 포함한다.

2) 데이터 연계 및 접근 방법

데이터 연계를 위해 인구조사국의 행정기록연구응용센터(Center for Administrative Records Research and Applications : CARRA)는 개인식별 확인 시스템(Person identification Validation System : PVS)을 통해 서로 다른 개인 식별을 위한 식별 보호키(Protected Identification Key : PIK)을 부여하는데, 이 PIK가 데이터 연계에 이용된다.

PVS는 '입수한 파일(incoming file)'을 '참조 파일(reference file)'과 연계하기 위해, '입수한 파일'에 포함된 이름, 주소, 생년월일, 사회보장번호와 같은 개인 식별번호에 기반을 둔 확률적 연계 방법을 활용한다. PVS 시스템을 통한 연계는 '참조파일'이 필요한데, 참조파일은 사회보장국(Social Security Administration : SSA)의 숫자식별파일(Numerical Identification file : SSA Numident)로부터 만들어진다. 이 숫자식별파일은 하나의 사회보장번호(SSN)에 대응한 모든 기록을 가지고 있는데, 이로부터 인구조사국의 숫자식별파일(Census Numident), 즉 참조파일이 만들어진다. 참조파일은 각 SSN에 해당하는 하나의 레코드를 보유하고 있으며, 생년월일과 이름 등 모든 변종을 별도의 파일로 보유한다. 이에 상응하는 개인 레코드에는 고유한 식별 보호키(PIK)가 부여되며, 이것이 PVS 시스템을 이용하는 모든 파일의 개인 연계 변수로 활용되는 것이다. '입수한 파일'과 '참조파일'의 연계가 이루어지면, '입수한 파일'에 PIK가 덧붙여진다.(Deborah Wagner, Mary Layne, 2014)

또한, CARRA는 주소 연계도 할 수 있으며, 인구조사국의 경제연구센터(Center for Economic Studies : CES)는 서로 다른 데이터셋의 기업들을 연계하는데 조세 ID, 고용인식별번호(Employer Identification

Number : EIN)을 이용한다.

제안서의 승인과 훈련이 완료되면, 연방조사국은 보안 컴퓨팅 환경 내에서 접근을 제공하며, 대부분의 경우 연방통계연구데이터센터(Federal Statistical Research Data Centers : FSRDCs)에서 하게 되는데, 원격접근이 허용되는 경우도 있다. 연구자들은 승인된 데이터 파일의 읽기전용의(read-only) 비식별화된 버전에 접근하며, 그들은 승인된 연구자들이 공유하는 프로젝트 단위의 작업 공간에서 모든 작업을 수행하게 된다. 마이크로데이터에 대한 모든 분석은 이 컴퓨팅 환경에서 이루어져야 한다. FSRDC는 인구조사국에 의해 관리되는 시설로서, 연방통계청(federal statistical agencies)과 주요 연구기관들의 협력 관계로 운영되는 사용이 제한된 마이크로데이터(restricted-use microdata)를 통계 목적으로만 접근을 허가하는 보안시설이다. 현재 25개의 연구데이터센터(RDC)가 있으며, 50여 개의 대학, 비영리연구소, 정부기관과 협력 관계를 맺고 있다.

다. 캐나다[20)]

1) 개요

캐나다 통계청(Statistics Canada)은 2012년 행정 자료, 조사 자료 등의 자료를 연계하고 연구지원 체계(The Canadian Centre for Data Development and Economic Research : CDER)를 마련하는 등 데이터 연계를 활발히 진행하고 있다.

캐나다 통계청의 SDLE(Social Data Linkage Environment)는 데이터 연계를 위한 환경으로 데이터 연계를 통해 중요한 문제를 해결하고 사회경제 정책에 정보를 제공하기 위해 기존의 행정 및 조사 데이터의 활용을 촉진시키며, 추가적인 자료를 수집하지 않고 연계된 분석 데이터의 생성

20) https://www.statcan.gc.ca/eng/sdle/index의 Social Data Linkage Environment 내용을 참조하였다.

을 통해 보건, 교육, 소득과 같은 다양한 영역에 걸친 분석을 가능하게 한다.

SDLE의 목적은 캐나다 사회경제 통계 연구를 촉진하는 것이며, 세부적으로 다음의 역할을 수행하고 있다. 새로운 데이터를 수집하지 않고 기존 조사와의 관련성 및 행정 데이터의 활용을 크게 증가시킨다. 그리고 추가적인 데이터 수집 없이 새로운 정보를 생성하며, 높은 수준의 개인 및 정보 보호를 유지한다. 마지막으로 데이터 연계 프로세스 및 방법의 표준화된 접근 방식을 증진한다.

2) 데이터 연계 및 접근 방법

캐나다 SDLE의 핵심은 DRD(Derived Record Depository)라 할 수 있으며, DRD는 캐나다 통계청이 선정한 자료를 연계하여 기본적인 개인 식별 정보만을 포함하는 국가적인 차원에서의 개인 데이터베이스이다. 이때, DRD 구축을 위해서 세금 기록, 출생·사망 등의 등록 기록, 이민 기록 등이 사용되며, 이러한 자료에 대한 업데이트는 지속적으로 DRD와 연결된다.

DRD에는 단지 기본적인 개인 식별 정보만이 저장되며, 저장되는 개인 식별 정보에는 성, 이름, 생년월일, 성별, 사회보장번호, 부모 이름, 결혼 여부, 주소(우편번호를 포함한), 전화번호, 이민날짜, 사망일 등이 있다. 저장된 개인 식별 정보를 통해 각 개인은 익명화된 SDLE 식별자(SDLE identifier)를 부여받게 되는데, 이는 SDLE 외에서는 어떠한 의미도 가지지 않는 정보이다.

데이터 연계는 고유 식별자가 있는 경우에는 정확 연계를 진행하며, 이름, 성별, 생년월일, 우편번호 등 고유하지 않은 식별자가 있는 경우에는 G-Link라는 소프트웨어를 이용하여 확률적 연계를 진행한다.

캐나다 통계청은 연계 데이터를 요청하는 성격에 따라 다양한 데이터 접근 방법을 제공하고 있다(이은우, 2017).

- Data Liberation Initiative(DLI) : 고등 교육기관과 캐나다 통계청의 협약에 의해 진행되고 있는 사업으로, 고등교육기관에 광범한 데이터와 메타데이터를 제공한다. 교수와 학생들은 공개사용이 가능한 마이크로데이터 파일(public use microdata files), 데이터베이스, 지리파일에 무제한 접근 가능하다.
- 연구데이터센터(Research Data Centres : RDC) 프로그램 : RDC 프로그램은 대학의 안전한 환경 내에서 행정 마이크로데이터 파일과 인구 및 가구 설문조사 데이터에 대한 직접 접근을 제공한다. 이 센터는 통계청 직원이 관리하며, 승인된 프로젝트의 연구자에게만 제공된다. RDC는 캐나다 전국에 위치해 있다.
- 실시간 원격접근(Real Time Remote Access : RTRA) 시스템 : RTRA 시스템은 온라인을 통해 이용자들이 안전한 공간에 위치해있는 마이크로데이터 파일에 대해 실시간으로 SAS 프로그램을 실행할 수 있도록 한다. 연구자들은 마이크로데이터에 직접 접근하여 콘텐츠를 볼 수는 없으며, 대신 SAS 프로그램을 통해 결과물만을 추출할 수 있다. 따라서 통계법상 '직원 간주'의 지위를 획득할 필요는 없으며, 연구 제안서를 제출할 필요도 없다.

라. 뉴질랜드[21)]

1) 개요

뉴질랜드 통계청(Statistics New Zealand)은 다양한 부문 간 관계에 대한 연구, 비용과 시간의 절약 등을 위해 데이터 연계를 적극적으로 활용하고 있으며, 10여 년 이상 데이터 통합 업무를 수행해 왔다.

데이터 연계와 관련해서는 홈페이지를 통해 데이터통합매뉴얼[22)] 등

21) https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/의 The Integrated Data Infrastructure 내용을 참조하였다.
22) http://archive.stats.govt.nz/methods/data-integration/data-integration-manual.aspx

데이터 연계 및 통합과 관련된 법적, 기술적 정보를 제공하고 있다. 또한, 뉴질랜드 통계청은 자신이 보유한 데이터를 통한 연구를 활성화하기 위해 통합 데이터기반(The Integrated Data Infrastructure : IDI)을 운영하고 있는데, 이는 사람 및 가정에 대한 마이크로데이터를 포함하고 있는 연구 데이터베이스이다.

2) 데이터 연계 및 접근 방법

각 데이터셋에서 활용 가능한 변수에 따라 서로 다른 데이터 연계 방법이 적용된다. IDI에서는 정확 연계 및 확률적 연계 방법이 모두 사용된다. 국가건강색인(National Health Index : NHI) 번호와 같은 공통된 고유 개인 식별자가 있을 경우 정확 연계 방법이 이용되며, 그렇지 않을 경우, 이름, 생년월일, 성별 등 인구 정보 변수를 이용한 확률연계 방법이 이용된다. 오류를 줄이기 위해 업데이트마다 연계 품질 검사가 수행되고 연계를 책임지는 직원은 변수의 일부(subset)에만 접근할 수 있으며, 고유식별자는 연계 전에 암호화된다.

이름, 주소 등 개인 식별정보는 삭제되며, IRD 번호(조세번호), NHI 번호 등 고유 식별자는 암호화(다른 번호로 대체)된다. 연구를 위해 다음과 같은 세 개의 SQL 데이터베이스가 제공된다.

첫째, 온전한 데이터베이스(clean database) : 통상 데이터 제공기관의 이름순으로 정렬된 모든 데이터 테이블을 포함하고 있다. 또한, 서로 다른 소스의 정보를 연계한 데이터도 제공된다. 연구자들은 이 모든 데이터셋에 접근할 수는 없으며, 연구에 필요한 데이터셋에만 접근이 허용된다.

둘째, 메타데이터베이스 : 특정 데이터 모음에 대한 검색 코드와 분류를 포함하고 있다.

셋째, 작업장 데이터베이스 (sandpit database) : 연구자들이 프로젝트팀원들과 테이블, 데이터셋, 프로그래밍 코드 등을 공유할 수 있는 공간이다.

공익 목적의 연구 프로젝트에만 IDI에 대한 접근이 허용된다. 연구자들은 자신의 연구 프로젝트에 필요한 데이터에만, 그리고 보안 데이터랩(secure Data Lab)을 통해서만 데이터에 접근할 수 있다.

연구자들은 보안 데이터랩 환경에서만 마이크로데이터에 접근할 수 있다. 이 공간에서 연구자들은 프로그램 실행, 데이터셋의 생성, 자료 저장, 다른 동료 연구자와의 공유, 추가적인 메타데이터에 대한 접근 등을 할 수 있다. 그러나 데이터랩의 컴퓨터는 인터넷 및 프린터 연결이 되지 않는다.

데이터랩은 뉴질랜드 통계청에 위치해 있지만, 연구자들은 자신의 작업공간에서 데이터랩 서버로의 보안 연결을 신청할 수 있다. 승인된다면, 통계청 사무실에서와 마찬가지의 소프트웨어 및 데이터가 제공된다. 이러한 원격 시설은 6개월 이상 소요되는 프로젝트, 그리고 과거에 기밀성과 보안 관련하여 좋은 기록을 갖고 있는 경우에 고려될 수 있다. 실현 가능성, 보안, 모니터링과 관련한 엄격한 조건을 충족해야 하는데, 예를 들어 데이터랩은 공공공간이나 사적 주거공간에는 설치될 수 없다.

제4절

소결

제3장에서는 국내에서 진행된 데이터 연계와 관련한 연구와 데이터 연계가 활성화되어 있는 국가에서 진행하고 있는 데이터 연계 및 데이터 정책을 중심으로 데이터 연계 방법, 기준·연계 데이터의 특성, 그리고 데이터 접근 방법을 살펴보았으며. 이를 간략하게 정리하면 다음과 같다.

국내에서는 대부분 정확 연계를 바탕으로 연구가 이루어지고 있으며, 본 연구에서는 국내사례 중에서 통계적 연계를 사용한 연구 위주로 정리하였다. 최현수 외 1인(2015)과 변종석 외 4인(2013)의 연구에서는 통계적 연계 방법을 설명하고 통계적 연계 방법을 실제로 적용 및 평가하였다. 두 연구 모두 연계의 기본 가정을 만족하는지를 평가하여 통계적 연계 방법의 적용이 타당하다고 결론지었다. 그리고 두 데이터에 공통으로 있는 변수들 중에 연계에 사용하지 않은 변수들의 정확한 연계 정도를 비교함으로써 통계적 연계에 대한 효율성을 평가하였다. 정확 연계가 아니기 때문에 이러한 평가는 한계가 있지만, 정확 연계를 할 수 없는 상황에서 통계적 연계도 효율성이 있다는 결과를 제시하였다.

정미옥 외 1인(2014)과 이영섭 외 4인(2009)의 연구에서는 정확 연계를 우선적으로 수행하였으며, 연계율을 살펴보면 정미옥 외 1인(2014)의 연구는 가구 연계를 기준으로 71.4%의 연계가 이뤄졌으며, 이영섭 외 4인(2009)의 연구는 전국사업체기초조사 자료를 기준으로 할 경우 16.8%의 연계율을 보였다. 정미옥 외 1인(2014)의 연구는 연계율이 70%를 넘어 보정 방안을 고려할 수 있겠지만, 이영섭 외 4인(2009)의 자료는 활용하는데 한계가 있을 수밖에 없다. 그리고 두 연구 모두 통계적 연계의 정확한 절차를 따르지 않아 통계적 연계에 대한 사례를 보는 것에는 한계가 있지만, 정확 연계를 하고 정확 연계한 자료를 이용하여 통계적 연계의

효율성을 검정하고 통계적 연계가 효율성이 있다고 판단하였다.

정확 연계가 동일한 대상을 연결시키는 정확한 연계 방법이지만 연계율이 매우 떨어질 경우에는 대표성을 담보하기 어렵기 때문에 통계적 연계 방법을 고려하는 것이 더 효율적일 수 있다. 그리고 국내 연구의 대부분은 랜덤 핫덱 또는 거리 핫덱 방법을 응용한 방법 등의 비모수적 방법을 주로 사용하고 있다. 특히 랜덤 핫덱을 사용한 방법은 블로킹 변수들을 조절하여 좋은 결과를 도출하였다.

해외사례에서 데이터 연계 방법을 살펴본 국가는 미국과 호주로 미국은 시카고대학의 여론조사센터, 호주는 통계청에서 수행한 사례이다. 두 사례 모두 정확 연계와 통계적 연계를 함께 적용한 방법으로 미국의 사례는 고유식별정보가 있는 대상과 없는 대상, 그리고 고유식별정보가 없을 경우에 정보의 양에 따라 연계 방법을 구분하여 연계를 수행하였다. 또한 호주 통계청의 경우에는 고유식별정보는 없지만 고유식별과 같은 정보가 있는 대상과 그렇지 않은 대상을 구분하여 데이터를 연계하였다. 해외사례는 우리나라와 차이가 있는 것은 이름과 대략적인 주소만 있으면 거의 고유식별정보와 같은 효과가 있다는 것이다. 따라서 정보의 양에 따라 데이터를 연계하는 방법이 발전되어 왔으며, 통계적 연계 방법을 사용하여도 충분한 품질의 데이터를 구축할 수 있다고 제시하고 있다. 그리고 이 경우에도 핫덱 방법을 응용한 방법을 사용하고 있다.

해외사례에서 개인정보 보호와 연관이 있는 데이터 접근 방법을 살펴본 국가들은 데이터 연계뿐만 아니라 데이터에 대한 접근에 대한 정책을 마련하고 있다. 영국 같은 경우에는 데이터 관리를 맡고 있는 행정 데이터 연구네트워크(ADRN)가 데이터에 대한 의뢰를 받으면, 데이터의 범위를 협의하고 데이터 연계를 전문 집단에게 맡기고 연구데이터와 행정 데이터를 분리하여 제공하고 있다. 미국의 인구조사국은 개인식별정보를 다루고 있기 때문에, 데이터의 개인보호문제가 발생하지 않도록 서로 다른 식별 보호키를 이용하여 데이터 연계에 이용하고 있으며, 이를 활용하기

위해서는 보안센터에서 승인을 받은 대상이 보안 컴퓨팅 환경에 접속하여 활용해야 한다. 캐나다와 뉴질랜드의 경우에는 개인정보를 다루는 경우 보안을 전제로 데이터를 활용하도록 하고 있으며, 연계 방법은 정확 연계와 확률적(통계적) 연계 방법을 동시에 사용하여 제공하고 있다. 두 나라 모두 데이터에 대한 보안을 우선시하면서도 중요도와 활용도 등을 고려하여 데이터를 제공하는 방법과 제공범위가 차이가 있다는 것을 알 수 있다.

국내·외 사례분석의 통해 다음과 같이 정리 할 수 있다.

① 통계적 연계 방법은 정확한 연계는 아니지만 데이터 연계를 활용하는데 효율적이다. 단, 통계적 연계 방법은 기본 가정이 만족하는지를 반드시 평가하고 만족할 경우에만 사용해야 한다.

② 통계적 연계 방법은 모수적 방법이 아닌 비모수적 방법인 핫덱 방법을 사용하여도 효율성이 좋다.

③ 정확 연계는 매우 좋은 방법이나, 연계율이 낮으면 효율성이 낮다.

④ 정확 연계에서 사용하는 고유식별정보가 있는 경우 데이터에 대한 보안의 문제가 발생하기 때문에, 이에 대한 다양한 정책을 마련해야 한다.

⑤ 해외에서는 데이터의 정보에 따라 다양한 방법으로 데이터를 연계하고 있는데, 이러한 방법에 대한 연구가 필요하다.

⑥ 데이터 활용을 높이기 위해서는 데이터에 접근하는 다양한 방법을 마련하고 보안이나 데이터의 활용 정도 등에 따라 제공하는 데이터의 범위를 달리할 필요가 있다.

제4장

데이터 연계에서 데이터 이해

제1절

문화·체육·관광 영역 데이터 연계 방향

1. 문화·체육·관광 데이터 연계 방안

문화·체육·관광 관련 분야는 사회가 발전할수록 많은 관심과 중요성은 확대되고 있다. 따라서 정책적이나 경제적 관점에서 많은 정보와 함께 이를 뒷받침할 수 있는 많은 연구와 분석이 필요하다. 예를 들어 최근에 일과 가정 모두 잘 할 수 있도록 '남녀고용평등과 일·가정 양립 지원에 관한 법률'과 '국민여가활성화기본법'을 제정 또는 개정함에 따라 여가 및 문화, 관광, 체육활동 시간은 증대될 것이라고 기대할 수 있는데 실제 그러한지 분석 또는 연구가 필요할 것이다. 또한 2004년부터 시행되고 있는 '주5일 근무제'와 2018년 7월부터 300인 이상 대기업부터 우선적으로 시행되고 향후 전체 사업장으로 확대될 예정인 '주 52시간 근무제'에 따른 효과 분석도 필요하다.

이와 같이 문화, 예술, 관광 그리고 스포츠 관련 정책이 증대됨에 따라 많은 연구와 분석이 필요하며 이를 위해서는 관련 데이터가 필요하지만 데이터는 부족한 실정이다. 이러한 이유로 새로운 정책에 대한 연구를 수행할 때마다 필요한 데이터를 일시적으로 생산하고 있다. 그러나 데이터를 생산하기 위해서는 많은 예산과 인력, 시간이 투입되어야 하나 1회성 연구를 위한 데이터를 생산하기 위해 실제 많은 예산과 시간, 인력을 투입하는 것은 쉬운 일이 아니다. 또한 일시적으로 생산되는 데이터의 대부분은 통계청에서 제시하고 있는 품질 기준을 따르지 않고 생산되기 때문에, 객관적인 자료로 신뢰하기 어려운 부분도 있다.

자료의 대표성을 담보하기 위해서는 통계청의 국가승인통계 또는 행정 자료를 이용할 수 있다면 가장 좋다. 그러나 국가승인통계의 자료는 필요한 연구나 분석하기에 충분하지 않거나 개인정보 등의 문제로 데이터

를 활용하지 못하는 경우가 많다. 그렇다고 적은 예산과 짧은 시간에 데이터를 생산할 경우, 결과에 대한 신뢰성 문제가 발생한다.

문화·체육·관광 분야의 데이터 연계를 통해 대표성과 신뢰성 문제는 해결할 수 있으며, 기준 데이터를 대표성 있는 자료를 이용하고 연계 데이터의 중요한 변수를 활용하면 된다. 즉, 기준 데이터가 통계청의 품질진단을 충족하거나 행정 자료가 대표성을 만족한다면, 데이터 연계를 통한 통합 파일 역시 대표성을 만족하게 된다. 단, 이 경우는 통합 파일이 데이터 연계의 조건을 모두 만족하는 경우를 전제로 한다.

따라서 본 연구에서는 데이터 연계는 기준 파일이 대표성을 충족하고 신뢰할 수 있는 데이터를 이용하도록 한다. 그리고 데이터 연계를 위한 조건이 충족되는지에 대한 검정을 충분히 하는 것을 전제로 한다.

정확 연계에서는 기준 파일의 대부분이 연계된다면 데이터 활용에 문제가 없으며, 동일한 대상이기 때문에 당연히 대표성도 만족한다. 그러나 통계적 연계는 동일한 대상이 아닌 유사한 대상을 연계하기 때문에 정보의 누락이 있을 수 있지만, 기준 파일이 설계(design)가 되어 생산되는 데이터이기 때문에 신뢰할 수 있다.

2. 문화·체육·관광 데이터 현황 이해

문화·체육·관광 국가승인통계는 〈표 4-1〉, 〈표 4-2〉 그리고 〈표 4-3〉에 제시한 것처럼 문화체육관광부의 국가승인통계 20종과 문화재청 1종, 산하기관 중 통계작성기관 4곳에서 4종, 관련 협회에서 1종을 생산하고 있다. 즉, 문화·체육·관광 국가승인통계는 25종이며, 조사통계가 17종이며, 보고통계가 6종, 가공통계가 2종이다.

〈표 4-1〉 문화체육관광부 국가승인통계 현황

순번	분야	종류	주기	대상	통계명
1	문화예술	조사	2년	국민	문화향수실태조사
2		조사	3년	예술인	문화예술인실태조사
3		조사	2년	국민	국민여가활동조사
4		조사	2년	사업체	공연예술실태조사
5	문화산업	조사	1년	국민	국민독서실태조사
6		조사	1년	사업체	콘텐츠산업통계조사
7		조사	1년	사업체	광고산업통계조사
8		조사	1년	사업체 근로자	근로자휴가실태조사
9		보고	1년	시설	전국도서관통계조사
10		보고	월	간행물	정기간행물등록현황
11		조사	3년	청각장애인	한국수어 사용실태조사
12	관광	조사	1년	국민	국민여행실태조사
13		조사	1년	외래객	외래관광객실태조사
14		조사	1년	사업체	관광사업체기초통계조사
15		보고	월	관광지	주요 관광지점 입장객 통계
16		보고	1년	시설	호텔업운영현황
17	체육	조사	2년	국민	국민생활체육참여실태조사
18		조사	2년	국민	국민체력실태조사
19		조사	1년	장애인	장애인생활체육실태조사
20		조사	1년	사업체	스포츠산업실태조사

〈표 4-2〉 문화재청 국가승인통계 현황

순번	분야	종류	주기	대상	통계명
1	문화	보고	1년	지자체	문화재관리현황

〈표 4-3〉 문화체육관광부 관련기관 및 협회 국가승인통계 현황

분야	종류	주기	대상	통계명	기관
문화체육관광	가공	년	사업체	문화체육관광산업 통계	한국문화관광연구원
문화	조사	1년	사업체	신문잡지산업실태조사	한국언론진흥재단
관광	가공	월	출입국자	한국관광통계	한국관광공사
관광	보고	월	사업체	여행사 국제관광객 유치·송출 통계	(사)한국여행업협회

그리고 통계청, 노동부 등에서 생산하고 있는 통계에서도 문화·체육·관광 관련된 내용의 데이터가 있을 수 있다. 통계청이나 노동부 등은 모든 국민이나 사업체 단위의 통계를 생산하고 있기 때문에 문화·체육·관광 영역의 내용이나 사업체에 해당하는 부분을 포함하는 통계를 생산하고 있다. 이러한 통계는 두 가지로 구분할 수 있는데, 하나는 대상은 동일하고 문화·체육·관광 관련 문항이 별도로 있는 경우이고, 다음은 문항은 동일한데 대상이 문화·체육·관광 관련 영역의 데이터가 포함된 경우이다. 전자는 국민을 대상으로 하는 통계인 경우가 해당되고, 후자는 사업체를 대상으로 하는 산업통계가 해당된다.

국민을 대상으로 하는 통계에서는 전국에 사는 국민을 대상으로 표본설계를 하여 생산하기 때문에 질문 중에서 문화·체육·관광 관련된 질문이 있다면 해당하는 질문을 나타내는 변수와 인구통계학적 변수만으로 새로운 데이터를 만들어 기준 파일로 활용할 수 있다.

사업체 대상의 통계자료는 문화체육관관 분야에 영위하는 사업체만을 추출하여 통계에 활용할 수 있는데, 이 경우에는 두 가지 주요 고려사항이 있다. 하나는 사업체 관련 경영활동이 있는 경우에는 개인정보보호 등의 문제로 데이터 제공을 하지 않는 경우의 문제이고, 다른 하나는 문화·체육·관광 관련된 데이터를 추출하는 문제로 이를 위해서는 명확한 분류기준이 마련되어야 한다.

제2절

연계에 사용할 문화·체육·관광 데이터의 이해

1. 문화·체육·관광 분야의 기준 파일로 사용할 데이터

데이터 연계를 하는 목적은 가지고 있는 데이터(기준 파일)에 다른 데이터(연계 파일)를 결합하는 것이다. 이때 관심있는 변수는 연계 파일 변수 중에서 연구나 분석에 필요한 변수가 된다. 따라서 연계 파일의 변수가 데이터 연계의 목적이 되는 것은 맞지만, 기준 파일의 데이터가 통합 파일의 기준이 되기 때문에 기준 파일이 대표성을 충족하는 것이 우선적이며 중요하다.

일반적으로 기준 파일은 모든 케이스들을 1회씩만 연계시키는 것을 원칙으로 하며 연계 파일은 기준 파일의 케이스와 가장 유사한 성향을 가진 케이스를 연계시키기 때문에 2회 이상 연계되는 것을 대부분 허용한다. 따라서 연계 파일의 전체 케이스가 기준 파일의 케이스보다 커야 데이터 연계가 효율적이다.

따라서 기준 파일이 우선적으로 대표성 등을 확보해야 하고, 연계 파일을 잘 연계시키기 위해서는 공통 변수가 적절해야 한다. 정확 연계에서는 고유식별이 가능하다면 공통 변수가 1개만 있어도 된다. 반면에 통계적 연계에서는 많은 공통 변수가 필요한데 공통 변수가 많을수록 다양한 조합을 통해 유사한 성향을 가진 케이스들을 연계하여 좋은 품질의 통합 파일을 생성할 수 있다.

문화·체육·관광 관련 데이터를 이용한 정확 연계에서 기준 파일로 활용이 가능한 데이터는 통계청이나 노동부 또는 민간 금융시관의 데이터 정도밖에 없다. 따라서 대부분의 데이터는 개인정보 문제로 실제 사용하는 것은 거의 불가능하다. 그리고 사전에 명확하게 내가 필요한 정보를

얻기 위해서는 명확하게 문화·체육·관광 관련된 자료만을 분리할 수 있도록 분류에 대한 정의가 있어야 할 것이다.

통계적 연계는 국가승인통계를 이용할 수밖에 없다. 25종의 국가승인통계 는 조사통계가 17종이며, 보고통계가 6종, 가공통계가 2종이며, 이 중에서 보고통계의 경우 특정한 목적에 맞는 정보를 입력시스템을 통해 모은 데이터이기 때문에 공통 변수가 충분하지 않거나 보편적으로 이용하기에 적절하지 않은 통계가 대부분이다.

그리고 가공통계는 모두 고유한 식별정보가 있는 통계를 가공한 것으로 엄밀히 말하면 정확 연계에 해당하는 통계이다. 한국관광통계는 출입국통계를 가공한 통계인데, 실제 원자료(raw data)를 이용하는 것이 아닌 통계값이 주어지면 이를 가공하는 통계이다. 다음으로 문화체육관광산업통계는 통계청의 '전국사업체기초자료'에서 문화체육관광산업 분류체계를 적용하여 문화체육관광산업 표본틀을 만든 후, 이를 통계청의 보안센터에 통계청의 기업등록부(BR : Business Register)자료와 연계하여 산출하고 있다. 따라서 가공통계는 원자료가 없는 데이터이기 때문에 실제 연구나 분석에서 활용하는 것은 불가능하다.

따라서 통계적 연계에서는 조사통계자료만을 이용할 수밖에 없으며, 문화·체육·관광 데이터 연계에서는 기준 파일로 활용되며 이들 조사통계들끼리도 연계가 가능하다. 데이터 연계는 같은 단위의 대상끼리 연계가 되기 때문에 국민, 사업체, 그리고 특정대상 통계로 구분하여 살펴본다.

본 연구에서는 문화체육관광부의 국가승인통계 중에서 국민을 대상으로 하는 통계데이터만을 세부적으로 살펴보기로 한다. 공통 변수가 연계된 공통 파일의 데이터 품질에 많은 영향을 주기 때문에 공통 변수에 대하여 세세히 검토하여 살펴보도록 한다.

2. 국민대상의 국가승인통계 데이터

국민대상의 문화체육관광부 국가승인통계 데이터는 국민여가활동조사, 문화향수실태조사, 국민독서실태조사, 국민여행실태조사, 국민생활체육참여실태조사, 국민체력실태조사 6종이 있다. 이중에서 국민체력실태조사는 집합시키는 조사이며 공통항목에 대한 조사가 거의 이뤄지고 있지 않아 이를 제외한 5종만을 검토하도록 한다.

가. 국민여가활동조사

국민여가활동조사는 전국의 만 15세 이상인 국민을 대상으로 국민의 여가활동 참여 실태를 조사하는 것으로 격년으로 짝수년도에 조사를 하고 있다. 조사의 대상기간은 조사시점으로부터 지난 1년을 대상으로 하고 있다. 〈표 4-4〉에 국민여가활동조사에 개요를 제시하였다.

〈표 4-4〉 국민여가활동조사 개요

구분	설명
통계종류	일반통계, 조사통계
법적근거	통계청 승인통계(승인번호 제113014호, 2007.05.01)
조사목적	다양하고 변화되는 국내 여가환경변화에 따라 국민의 여가수요에 미치는 활동 실태를 분석하여 생활양식의 변화 및 삶의 질 수준 파악
조사주기	2년
조사대상	전국 17개 시·도에 거주하는 만 15세 이상 남녀(제주도, 세종시 포함)
조사방법	– 확률표본 – 면접조사
조사기간	조사실시년도 9월~10월
작성체계	조사업체 → 한국문화관광연구원 → 문화체육관광부
자료검색	– 문화체육관광부 자료공간(www.mcst.go.kr) – 문화예술지식정보시스템(policydb.kcti.re.kr)

다음의 〈표 4-5〉는 국민여가활동조사 응답자 특성을 제시한 표이다. 응답자 특성을 나타내는 변수들이 공통 변수로 주로[23] 사용할 변수들이

기 때문에 응답자 특성 변수를 잘 이해하고 있을 필요가 있다. 특히, 조사통계에서는 응답자 특성의 일부 변수들을 가지고 표본설계를 하기 때문에 모집단의 대표성을 확보하기 위해서는 표본설계에 사용된 변수와 가중값을 계산할 때 사용된 변수를 잘 파악해야 한다.

〈표 4-5〉 국민여가활동조사의 응답자 특성

구분	내용
지역	11. 서울 23. 인천 26. 울산 32. 강원 35. 전북 38. 경남 21. 부산 24. 광주 29. 세종 33. 충북 36. 전남 39. 제주 22. 대구 25. 대전 31. 경기 34. 충남 37. 경북
성별	1. 남자 2. 여자
연령	만 세
최종학력	1. 무학 3. 중학교 5. 대학(4년제 미만) 7. 대학원 석사 과정 2. 초등학교 4. 고등학교 6. 대학교(4년제 이상) 8. 대학원 박사 과정
졸업 여부	1. 졸업 3. 수료 5. 중퇴 2. 재학 4. 휴학
혼인상태	1. 미혼 3. 사별 5. 기타 2. 배우자 있음 4. 이혼
경제활동여부	1. 예 2. 아니오
경제활동 없는 경우의 직업	1. 전업주부 2. 학생 3. 기타
경제활동 있는 경우의 직업(주업)	1. 관리자 5. 판매 종사자 8. 장치, 기계 조작 및 2. 전문가 및 관련 종사자 6. 농림어업 숙련 종사자 조립 종사자 3. 사무 종사자 7. 기능원 및 관련 9. 단순노무종사자 4. 서비스 종사자 기능 종사자 10. 군인
종사상의 지위	1. 임금, 봉급 근로자 3. 고용원이 없는 자영업자 2. 고용원을 둔 사업주 4. 무급가족 종사자
동거가구원수	총 명
월평균 본인 소득 월평균 가구 소득	1. 소득없음 8. 월평균 600만원 ~ 700만원 미만 2. 월평균 100만원 미만 9. 월평균 700만원 ~ 800만원 미만 3. 월평균 100만원 ~ 200만원 미만 10. 월평균 800만원 ~ 900만원 미만 4. 월평균 200만원 ~ 300만원 미만 11. 월평균 900만원 ~1,000만원 미만 5. 월평균 300만원 ~ 400만원 미만 12. 월평균 1,000만원 이상 6. 월평균 400만원 ~ 500만원 미만 99. 무응답 7. 월평균 500만원 ~ 600만원 미만

23) 주로라는 표현을 사용한 이유는 연계 파일에 없는 변수라면 공통 변수가 될 수 없기 때문이다.

다음의 〈표 4-6〉은 국민여가활동조사의 목적에 맞는 질문한 내용으로 데이터 연계에서는 주로[24] 고유 변수가 사용되며, 관심 변수는 모수적 방법(회귀분석 등)을 통해 예측값을 이용하여 데이터 연계에 활용된다.

〈표 4-6〉 국민여가활동의 세부 내용

구분	항목명	내용
여가활동 참여실태	한번이상 참여한 여가활동 유형	문화예술관람/참여, 스포츠관람/참여, 관광, 취미 · 오락, 휴식, 사회 및 기타
	최다참여 여가활동(1~5순위)	유형, 동반자, 빈도, 소요시간, 비용, 만족도
	여가활동 목적	여가활동의 주된 목적
	최고만족도 여가활동	가장 만족스런 활동 1~3순위
	여가경력	반복참여여가활동, 참여기간, 노력도
	여가비용	월평균금액, 여가비용충분도, 적정 여가비용
평일 및 휴일 (휴가, 연휴) 여가활동	평일 유형별 여가활동	(평일)여가활동 유형, 만족도, 희망 여가활동
	휴일 유형별 여가활동	(휴일)여가활동 유형, 만족도, 희망 여가활동
	여가시간	(평일/휴일) 하루평균 여가시간, 여가시간 충분도, 여가시간 자유도, 희망 여가시간
	휴가 활용	휴가사용여부, 사용휴가일수, 분기별 휴가일수, 휴가중 여가활동
	연휴 및 대체휴일 활용	연휴기간 중 여가활동 1~3순위
여가공간	이용 또는 희망하는 여가공간	최다이용/희망 여가공간(1~3순위)
	여가공간만족도(공공시설)	여가시설 존재인식도, 이용충분도, 프로그램 인지도 및 충분도, 여가시설 이용여부, 서비스 만족도
	여가공간만족도(민간시설)	민간 여가관련 공간/용품/서비스에 대한 평가 및 만족도
다양한 여가활동	동호회 활동	동호회 참여여부, 주된 활동, 부수적 활동
	사회성 여가활동	자원봉사활동 경험 및 분야
	스마트기기 이용시간, 활용 여가활동	평일 및 휴일 스마트기기 활용시간, 주요 여가활동
여가인식 및 만족도	여가 정책	정부정책 중요도, 중요순위, 만족도
	일(학업)과 여가의 균형	일과 여가의 균형 정도
	향후 여가시간 활용	여가시간 활용계획
	여가생활 만족도	여가생활 만족도, 불만족 이유
	여가인식	여가활동 필수요건정도, 긍정적 영향정도
	행복 수준	행복수준(10점 척도)

24) 여기서도 주로라는 표현을 한 이유는 일부 변수는 중복되는 경우가 있을 수 있는데 이러한 경우에는 공통 변수가 된다.

나. 문화향수실태조사

문화향수실태조사는 전국의 만 15세 이상인 국민을 대상으로 문화예술 관람 및 교육활동, 문화관련 활동에 대한 조사를 통해 문화향유 실태를 파악하는 조사이다. 조사는 국민여가활동조사와 같이 격년으로 짝수년도에 조사를 하고 있다. 조사의 대상기간은 조사시점으로부터 지난 1년을 대상으로 하고 있다. 〈표 4-7〉에 문화향수실태조사 개요를 제시하였다.

〈표 4-7〉 문화향수실태조사 개요

구분	설명
통계종류	일반통계, 조사통계
법적근거	통계청 승인통계(승인번호 제113001호, 1991.05.07)
조사목적	우리나라 국민의 문화 활동 향유의 필요성 및 인식이 높아짐에 따라 실태 파악을 위한 문화향유 경로와 방식에 대하여 통계적으로 분석하여 궁극적으로 국민 문화향유 진흥 도모
조사주기	2년
조사대상	전국 17개 시·도에 거주하는 만 15세 이상 전국민(제주도,세종시 포함)
조사방법	– 확률표본 – 면접조사
조사기간	조사실시년도 9월~10월
작성체계	조사업체 → 한국문화관광연구원 → 문화체육관광부
자료검색	– 문화체육관광부 자료공간(www.mcst.go.kr) – 문화예술지식정보시스템(policydb.kcti.re.kr)

〈표 4-8〉은 문화향수실태조사 응답자 특성을 제시한 표이다. 공통변수로 사용할 응답자 특성은 국민여가활동조사의 응답자 특성문항과 비교할 때 장애등록여부 변수를 제외하고는 동일하다.

〈표 4-8〉 국민여가활동조사의 응답자 특성

구분	내용
지역	11. 서울 23. 인천 26. 울산 32. 강원 35. 전북 38. 경남 21. 부산 24. 광주 29. 세종 33. 충북 36. 전남 39. 제주 22. 대구 25. 대전 31. 경기 34. 충남 37. 경북
성별	1. 남자 2. 여자
연령	만 세
최종학력	1. 무학 3. 중학교 5. 대학(4년제 미만) 7. 대학원 석사 과정 2. 초등학교 4. 고등학교 6. 대학교(4년제 이상) 8. 대학원 박사 과정
졸업 여부	1. 졸업 3. 수료 5. 중퇴 2. 재학 4. 휴학
혼인상태	1. 미혼 3. 사별 5. 기타 2. 배우자 있음 4. 이혼
경제활동여부	1. 예 2. 아니오
경제활동 없는 경우의 직업	1. 전업주부 2. 학생 3. 기타
경제활동 있는 경우의 직업(주업)	1. 관리자 5. 판매 종사자 8. 장치, 기계 조작 및 2. 전문가 및 관련 종사자 6. 농림어업 숙련 종사자 조립 종사자 3. 사무 종사자 7. 기능원 및 관련 9. 단순노무종사자 4. 서비스 종사자 기능 종사자 10. 군인
종사상의 지위	1. 임금, 봉급 근로자 3. 고용원이 없는 자영업자 2. 고용원을 둔 사업주 4. 무급가족 종사자
장애등록여부	1. 해당사항 없다 2. 장애등록을 하였다 3. 등록하지 않았다
동거가구원수	총 명
월평균 본인 소득 월평균 가구 소득	1. 소득없음 8. 월평균 600만원 ~ 700만원 미만 2. 월평균 100만원 미만 9. 월평균 700만원 ~ 800만원 미만 3. 월평균 100만원 ~ 200만원 미만 10. 월평균 800만원 ~ 900만원 미만 4. 월평균 200만원 ~ 300만원 미만 11. 월평균 900만원 ~1,000만원 미만 5. 월평균 300만원 ~ 400만원 미만 12. 월평균 1,000만원 이상 6. 월평균 400만원 ~ 500만원 미만 99. 무응답 7. 월평균 500만원 ~ 600만원 미만

〈표 4-9〉는 문화향수실태조사의 목적에 맞도록 설문 설계하여 조사하는 문항을 정리한 것이다. 사용데이터 연계에서는 주로 고유 변수에 해당하는 내용이며, 연계할 데이터에 문화활동 관련된 문항이 있을 경우에는 문화활동 관련된 척도를 개발하여 공통 변수로 사용할 수 있다.

〈표 4-9〉 문화향수실태조사의 세부 내용

구분	항목명	내용
문화예술관람 및 참여	문화예술행사 관람실태	관람횟수, 만족도, 의향, 참여경험,
	문화예술행사 직접관람 실태	관람지역, 관람경유, 동반자, 시간대, 정보습득매체, 보완점, 관람기준, 관람방해요소
	매체이용 문화예술행사 관람	경험, 이용매체, 만족도
	문화예술행사 참여실태	경험, 만족도, 의향, 동반자, 시간대
	문화예술관련 지출	지출 항목(구입 및 대여, 관람, 기타)
문화예술교육	문화예술교육 경험	과거, 지난 1년간의 경험, 만족도, 의향
	문화예술교육 실태	교육기관, 보완점, 교육방식, 교육 방해 요소
문화예술활동 공간이용실태	문화예술공간 이용실태	이용횟수, 위치구분, 만족도, 참석횟수
	문화예술공간 방문의향	문화공간 참여방해요소, 참여의향
문화관련 활동	자원봉사활동 및 기부활동	경험, 참여횟수, 금전기부경험
	문화관련 동호회참여	참여경험, 동호회성격, 활동공간, 참여빈도
역사문화유적지 및 축제	역사문화유적지 방문실태	경험, 만족도, 유적지구분, 의향
	축제관람실태	경험, 축제 주제, 만족도

다. 국민독서실태조사

국민독서실태조사는 전국의 만 19세 이상의 성인과 전국 초·중·고 학생을 대상으로 독서량, 독서생활, 독서환경 등에 대한 조사를 실시한다. 조사주기는 2년이며 지난 1년간을 대상으로 한다. 〈표 4-10〉은 국민독서실태조사의 조사개요이다.

〈표 4-10〉 국민독서실태조사 개요

구분	설명
통계종류	일반통계, 조사통계
법적근거	통계청 승인통계(승인번호 제113018호, 2008.09.02)
조사목적	국민의 독서실태와 변화 추이를 파악하는 국가승인통계로서 표준적인 독서지표를 작성하여 국민 독서 진흥을 위해 정부,언론계,교육계,도서관계,출판산업계, 독서계 등 사회 각계에서 기본 통계로 활용하고자 함
조사주기	2년
조사대상	– 성인 : 전국의 19세 이상 성인남녀 – 학생 : 전국의 초등학생(4~6학년), 중학생, 고등학생

구분	설명
조사방법	– 확률표본 – 성인 : 조사원에 의한 1:1 개별면접조사(타계식 조사) – 학생 : 조사원(또는 교사)의 통제에 의한 자기기입식(자계식 조사)
조사기간	조사대상년도 11월~12월
작성체계	조사업체 → 한국문화관광연구원 → 문화체육관광부
자료검색	– 문화체육관광부 자료공간(www.mcst.go.kr) – 문화예술지식정보시스템(policydb.kcti.re.kr)

〈표 4-11〉은 국민독서실태조사에서 성인[25]에 대한 응답자 특성을 제시한 표이다. 주로 공통 변수로 사용할 응답자 특성을 살펴보면 연령이 만 19세 이상이며, 항목별로 살펴보면 가족 구성에 대한 정보가 부족한 부분이 있다.

〈표 4-11〉 국민독서실태조사의 응답자 특성

구분	내용
지역	1. 서울 4. 인천 7. 울산 10. 강원 13. 전북 16. 경남 2. 부산 5. 광주 8. 세종 11. 충북 14. 전남 17.제주 3. 대구 6. 대전 9. 경기 12. 충남 15. 경북
성별	1. 남자 2. 여자
연령	만 세
최종학력	1. 교육을 안 받았음 4. 고등학교 7. 대학원 석사 과정 2. 초등학교 5. 대학(4년제 미만) 8. 대학원 박사 과정 3. 중학교 6. 대학교(4년제 이상)
졸업 여부	1. 졸업 3. 수료 5. 중퇴 2. 재학 4. 휴학
직업분류1	1. 관리자 9. 단순노무 종사자 2. 전문가 및 관련 종사자 10. 군인 3. 사무 종사자 11. 자영업 종사자 4. 서비스 종사자 12. 학생 5. 판매 종사자 13. 전업주부 6. 농림어업 숙련 종사자 14. 은퇴, 무직 7. 기능원 및 관련 기능 종사자 15. 기타 8. 장치/기계 조작 및 조립 종사자
직업분류2	1. 농업/수산업/축산업 9. 가정주부 2. 자영업 10. 중학생

구분	내용	
	3. 판매직/서비스직	11. 고등학생
	4. 기능공/숙련공	12. 대학생
	5. 일반작업직	13. 대학원생
	6. 사무직/기술직	14. 은퇴/무직
	7. 경영관리직	99. 모름/무응답
	8. 전문직	
월평균 본인 소득	1. 100만원 미만	5. 400만원 ~ 500만원 미만
	2. 100만원 ~ 200만원 미만	6. 500만원 ~ 600만원 미만
	3. 200만원 ~ 300만원 미만	7. 600만원 ~ 700만원 미만
	4. 300만원 ~ 400만원 미만	8. 700만원 이상

〈표 4-12〉는 국민독서실태조사의 목적에 맞도록 설문 설계한 내용을 작성한 것이다. 데이터 연계에서 주로 고유 변수에 해당하는 부분이다.

〈표 4-12〉 국민독서실태조사의 세부 내용

구분	항목명
독서생활	하루 평균 여가시간, 독서 시간, 독서 장애 요인, 본인 유용성 평가
독서량과 책의선택	독서량, 독서량에 대한 평가, 독서 목적, 독서의 계기 및 기회, 독서장소, 도서 선택 시 이용 정보, 도서 입수 경로, 도서 선호 분야, 도서 선호 분야, 도서 구입처, 도서 구입비
독서환경	독서 대화, 독서 권장을 받은 경험, 직장의 도서 환경, 공공도서관 이용 경험(주로 이용하는 공공도서관의 종류, 공공도서관 이용 빈도, 공공도서관 이용 목적), 도서관 비이용 이유,
독서 프로그램 및 독서 동아리(독서 모임)참여	독서 프로그램 참여 경험(참여해본 독서 프로그램, 독서 프로그램 참여 계기), 독서 프로그램 비참여 이유, 참여 희망 독서 프로그램, 독서 동아리(독서 모임) 참여 경험(참여 경험이 있는 독서 동아리, 독서 동아리 참여 계기), 독서 동아리 비참여 이유, 참여 희망 독서 동아리
독서 활성화 방안	독서 활성화 방안(독서환경, 독서문화, 독서운동, 독서복지), 도서구입비 세제 혜택 인지도, 도서 구입비 세제 혜택의 도움 정도

라. 국민여행실태조사

국민여행실태조사는 가구 내에 만 15세 이상 동거 가구원을 대상으로 국내와 국외 여행관련 조사를 실시한다. 공표주기는 1년이면 조사주기는 반년(6개월)로, 상반기 여행과 하반기 여행에 대한 조사를 각각 구분하여

25) 학생들 대상은 특정대상에 대한 통계로, 일반 국민들 대상인 성인에 대해서만 제시하였음

실시한다. 조사방법은 응답자가 국내외 여행을 다녀온 직후에 여행기록부(종이, 온라인)에 자기기입하는 방식이다. 〈표 4-13〉에 국민여행실태조사에 대한 설명을 정리하여 제시하였다.

〈표 4-13〉 국민여행실태조사 개요

구분	설명
통계종류	일반통계, 조사통계
법적근거	통계청 승인통계(승인번호 제314001호, 1976.09.18)
조사목적	우리나라 국민의 여행실태를 종합적으로 파악하여 국가관광에 관한 정책수립과 연구, 분석 등을 위한 기초 자료를 제공함으로써 관광을 통한 지역발전을 도모하고, 궁극적으로 국민복지 및 국민의 삶의 질 제고
조사주기	반기(6개월)
조사대상	전국 만 15세 이상의 가구원
조사방법	– 확률표본 – 면접조사(응답자가 국내외 여행을 다녀온 직후 종이 여행기록부 또는 온라인 여행기록부 작성하는 자기기입법)
조사기간	– 상반기 : 조사대상년도 7월~8월 – 하반기 : 조사대상년도 익년 1월~2월
작성체계	조사업체 → 한국문화관광연구원 → 문화체육관광부
자료검색	– 문화체육관광부 자료공간(www.mcst.go.kr) – 관광지식정보시스템(www.tour.go.kr)

〈표 4-14〉는 국민여행실태조사 응답자 특성을 제시한 표이다. 응답자 특성의 특이사항은 세종시가 없다는 것이다. 이는 최초설계가 2010년도에 이뤄져서 같은 대상에 대하여 매년 조사하여 온 조사이기 때문에[26] 세종시가 조사대사에서 빠진 것이다. 그러나 여행방문지 응답에는 세종시가 반영되어 있다. 또한 직업분류를 2가지로 구분하여 종사분야까지 조사하였다.

26) 2018년도부터는 월별조사로 바뀌었고, 매월 새롭게 표본설계하여 표본을 추출하여 조사하고 있다.

〈표 4-14〉 국민여행실태조사의 응답자 특성

구분	내용
지역	1. 서울 4. 인천 7. 울산 10. 충북 13. 전남 38. 경남 2. 부산 5. 광주 8. 경기 11. 충남 14. 경북 3. 대구 6. 대전 9. 강원 12. 전북 15. 경남
성별	1. 남자 2. 여자
연령	만 세
최종학력	1. 무학 4. 고등학교 7. 대학원 석사 과정 2. 초등학교 5. 대학(4년제 미만) 8. 대학원 박사 과정 3. 중학교 6. 대학교(4년제 이상) 9. 모름/무응답
졸업 여부	1. 졸업 3. 수료 5. 중퇴 2. 재학 4. 휴학
혼인상태	1. 미혼 3. 사별 2. 배우자 있음 4. 이혼
직업분류1	1. 관리자 10. 군인 2. 전문가 및 관련 종사자 11. 가정주부 3. 사무 종사자 12. 중학생 4. 서비스 종사자 13. 고등학생 5. 판매 종사자 14. 대학생 6. 농림어업 숙련 종사자 15. 대학원생 7. 기능원 및 관련 기능 종사자 16. 은퇴/무직 8. 장치/기계 조작 및 조립 종사자 99. 모름/무응답 9. 단순노무자
직업분류2	1. 농업/수산업/축산업 9. 가정주부 2. 자영업 10. 중학생 3. 판매직/서비스직 11. 고등학생 4. 기능공/숙련공 12. 대학생 5. 일반작업직 13. 대학원생 6. 사무직/기술직 14. 은퇴/무직 7. 경영관리직 99. 모름/무응답 8. 전문직
동거가구원수	총 명
월평균 가구 소득	만원

〈표 4-15〉는 국민여행실태조사의 목적에 맞도록 설문 설계한 내용을 작성한 것이다. 데이터 연계에서는 주로 고유 변수에 해당하는 내용이며, 여행 관련된 내용이 있는 데이터 결합할 경우 효율이 높은 연계를 할 수 있다.

〈표 4-15〉 국민여행실태조사의 세부 내용

구분	항목명	내용
가구대표 국내여행	일반적 사항	당일/숙박 여부, 여행기간, 여행시기, 여행목적
	관광여행용 문항	정보습득매체, 동행자 정보, 여행지역, 여행지 선택이유, 교통수단, 숙박시설, 숙박일수, 활동, 만족도, 재방문의향, 추천의향, 지출비용, 사전예약 서비스, 단체경비, 여행상품 구매/이용 여부, 여행경비 상세, 세부항목별 만족도 등
	비관광여행용 문항	동행자 정보, 여행지역, 교통수단, 숙박시설, 숙박일수, 여행총비용
개인 국내여행	일반적 사항	당일/숙박 여부, 여행기간, 여행시기, 여행목적
	관광여행용 문항	정보습득매체, 동행자 정보, 여행지역, 여행지 선택이유, 교통수단, 숙박시설, 숙박일수, 활동, 만족도, 재방문의향, 추천의향, 지출비용, 사전예약 서비스, 단체경비, 여행상품 구매/이용 여부, 여행경비 상세, 세부항목별 만족도 등
	비관광여행용 문항	동행자 정보, 여행지역, 교통수단, 숙박시설, 숙박일수, 여행총비용
가구대표 해외여행	일반적 사항	해외여행구분(당일/숙박/북한), 여행기간, 여행시기, 여행목적
	여행 목적(지)	정보습득매체, 동행자 정보, 여행지역, 여행지 선택이유, 교통수단, 숙박시설, 숙박일수, 활동, 만족도, 재방문의향, 추천의향
	여행 지출 경비	지출비용, 사전예약 서비스, 단체경비, 여행상품 구매/이용 여부, 여행경비 상세
	여행소감	세부항목별 만족도 등
개인 해외여행	일반적 사항	해외여행구분(당일/숙박/북한), 여행기간, 여행시기, 여행목적
	여행 목적(지)	정보습득매체, 동행자 정보, 여행지역, 여행지 선택이유, 교통수단, 숙박시설, 숙박일수, 활동, 만족도, 재방문의향, 추천의향
	여행 지출 경비	지출비용, 사전예약 서비스, 단체경비, 여행상품 구매/이용 여부, 여행경비 상세
	여행소감	세부항목별 만족도 등

마. 국민생활체육참여실태조사

국민생활체육참여실태조사는 대한민국 내 만 10세 이상의 국민을 대상으로 매년 실시하는 조사이다. 조사시점 당시 지난 1년을 대상 기간으로 조사가 진행되며, 국민의 생활체육 참여 현황과 여건, 건강상태 등을 파악할 수 있다. 최신 보고서는 2017년 발간되었다. 〈표 4-16〉에 국민생활체육참여실태조사 개요를 정리하였다.

〈표 4-16〉 국민생활체육참여실태조사 개요

구분	설명
통계종류	일반통계, 조사통계
법적근거	통계청 승인통계(승인번호 제113003호, 1991.10.29)
조사목적	국민의 생활체육 참여실태를 파악하여 효과적이고 체계적인 체육정책의 수립을 위한 기초자료 제공
조사주기	1년
조사대상	전국 만 10세 이상의 가구원
조사방법	– 확률표본 – 면접조사
조사기간	조사대상년도 8월~10월
작성체계	조사업체 → 한국문화관광연구원 → 문화체육관광부
자료검색	– 문화체육관광부 자료공간(www.mcst.go.kr) – 한국스포츠정책과학원(www.sports.re.kr)

〈표 4-17〉은 국민생활체육참여실태조사 응답자 특성을 제시한 표이다. 국민여가활동조사와 비슷한 구조인데, 월평균가구소득이 50만원 단위의 간격으로 구성된다. 응답자 특성 문항들은 데이터 연계에서 주로 공통문항으로 사용되게 된다.

〈표 4-17〉 국민생활체육참여실태조사의 응답자 특성

구분	내용
지역	11. 서울 23. 인천 26. 울산 32. 강원 35. 전북 38. 경남 21. 부산 24. 광주 29. 세종 33. 충북 36. 전남 39. 제주 22. 대구 25. 대전 31. 경기 34. 충남 37. 경북
성별	1. 남자 2. 여자
연령	만 세
최종학력	1. 무학 3. 중학교 5. 대학(4년제 미만) 7. 대학원 석사 과정 2. 초등학교 4. 고등학교 6. 대학교(4년제 이상) 8. 대학원 박사 과정
졸업 여부	1. 졸업 3. 수료 5. 중퇴 2. 재학 4. 휴학
혼인상태	1. 미혼 3. 사별 5. 기타 2. 배우자 있음 4. 이혼
경제활동여부	1. 예 2. 아니오

구분	내용
경제활동 없는 경우의 직업	1. 전업주부 2. 학생 3. 무직 4. 기타
경제활동 있는 경우의 직업(주업)	1. 관리자 2. 전문가 및 관련 종사자 3. 사무 종사자 4. 서비스 종사자 5. 판매 종사자 6. 농림어업 숙련 종사자 7. 기능원 및 관련 기능 종사자 8. 장치, 기계 조작 및 조립 종사자 9. 단순노무종사자 10. 군인
가구원수 (본인포함)	1. 1명 2. 2명 3. 3명 4. 4명 5. 5명 6 기타(명)
자녀현황	1. 없음 2. 1명 3. 2명 4. 3명 5. 5명 이상
월평균 가구 소득	1. 100만원 미만 2. 100만원 ~ 150만원 미만 3. 150만원 ~ 200만원 미만 4. 200만원 ~ 250만원 미만 5. 250만원 ~ 300만원 미만 6. 300만원 ~ 350만원 미만 7. 350만원 ~ 400만원 미만 8. 400만원 ~ 450만원 미만 9. 450만원 ~ 500만원 미만 10. 500만원 ~ 550만원 미만 11. 550만원 ~ 600만원 미만 12. 600만원 이상

〈표 4-18〉은 국민생활체육참여실태조사의 목적에 맞도록 설문 설계한 내용을 작성한 것이다. 데이터 연계에서는 주로 고유 변수에 해당하는 내용이이다.

〈표 4-18〉 국민생활체육실태조사의 세부 내용

구분	항목명
건강 및 체력상태	건강상태 척도(5점), 건강유지 중요점, 체력상태 인식, 체력유지요인 수행도(5점), 활동 지장 여부
체육활동 및 여건	체육시설 위치 인식 여부, 이용 시설 유형 및 이유, 희망 이용 시설(1~3순위), 운동용품 구입 상세, 체육교육 수강 상세, 체육정보 습득 매체(1~3순위), 동호회 상세, 운동처방 또는 상담 서비스
체육활동 참여현황	체육활동별 참여여부, 규칙적 체육활동 빈도 및 상세, (학생대상) 체육활동 경험 상세, 체육활동 참가 이유 및 요인, 동반자, 안전사고 예방 활동, 병원치료 상세, 한달 평균 경비 등

제3절

데이터 표준화

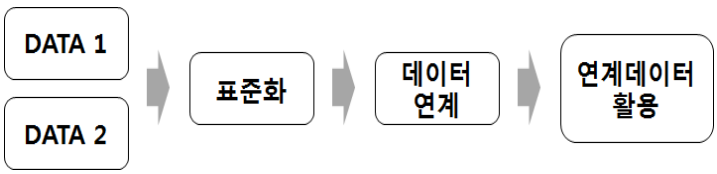
데이터 연계에서 두 데이터를 결합하는 연계 방법이 가장 관심을 가지게 된다. 그러나 실제 데이터 연계 방법만큼 중요한 과정이 데이터 연계 전의 두 데이터를 동일한 기준으로 맞추는 조화과정과 연계 후의 통합 파일을 이용해서 분석해도 될지에 대한 판단을 하는 검정 과정이다.

데이터 표준화 과정은 데이터 연계에서 데이터 연계 전의 처리과정의 일부이다. 이러한 데이터 전처리 과정은 데이터 클리닝 과정과 데이터 표준화 과정으로 구분되는데, 데이터 클리닝 과정은 데이터를 검증하여 오류가 없도록 수정하는 과정으로 데이터 연계가 아닌 경우에도 반드시 해야 하는 과정이다. 그러므로 이번 절에서는 데이터 표준화하는 과정을 살펴보도록 한다.

일반적으로 데이터베이스(DB : Data Base)를 구축할 때, 새로운 데이터를 적재(loading)하기 전에, 데이터 형식을 어떻게 적재할 지에 대한 기준을 정하고 이 기준으로 데이터를 맞추기 위해 변환(transformation) 등의 과정을 거쳐, 표준화된(Standardized) 데이터를 DB에 적재시킨다. 서로 다른 두 개의 데이터를 연계하기 전에 표준화하는 과정은, 데이터베이스에 있는 두 개의 데이터를 적재할 때 고려하는 과정과 동일하다. 같은 의미를 지니는 변수는 동일하게 표현 되도록 하며, 변수명칭도 사전에 명명규칙을 만드는 등의 과정을 거쳐 표준화시킨다. 이렇게 표준화된 두 데이터는 연계 방법에 따라 통합 파일을 만들게 된다.

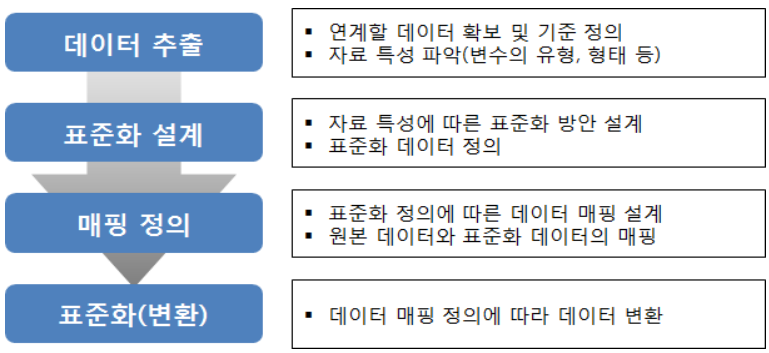
따라서 서로 다른 데이터를 연계하여 통합된 하나의 데이터를 만든다는 것은 통합되기 전 각각의 데이터를 동일한 기준으로 표현하였다는 것을 의미한다. 즉, 데이터를 연계할 때 동일 변수명의 같은 표현의 값은 반드시 일치하는 값을 의미해야 한다. 그러나 표준화시키지 않은 경우,

각각의 데이터는 개별적으로 생산되고, 부호화(coding) 시키기 때문에 동일 항목이 동일 데이터 값을 갖는 경우는 매우 드물다. 따라서 데이터를 연계하기에 앞서 진행해야 하는 것이 바로 데이터 표준화이다. 즉, 데이터 연계는 이러한 표준화 작업 이후에 진행해야 한다.



[그림 4-1] 데이터 연계의 도식화

데이터 표준화를 위해서는, 연계하는 데이터의 특성을 파악한 후, 데이터 표준화 방안을 마련하고 표준화 규칙을 정립할 필요가 있다. 설계된 표준화 규칙을 데이터에 반영하기 위해서는 데이터 매핑(mapping)을 정의하고, 이를 데이터에 적용시키는 데이터 변환 작업을 통해 표준화 파일을 생성하면, 데이터 연계를 위한 표준화 작업이 완료된다. 표준화 작업의 과정은 [그림 4-2]에 제시하였다.



[그림 4-2] 데이터 표준화 절차

1. 데이터 추출

데이터 연계를 위해서는 먼저 수집된 데이터의 대상과 시점에 대한 기준이 동일해야 한다. 연계한 데이터는 하나의 통합된 데이터가 되는 것이므로, 동일한 대상과 시점을 기준으로 수집된 것이어야 한다. 데이터를 구성하는 대상을, 각 케이스를 살펴보아 파악하도록 한다. 연계할 데이터가 각각 다른 대상 또는 특정대상으로 한정되어 생산되었다면, 대상이 동일한 기준으로 데이터를 맞춰 생산할 수 있는지 살펴보는 것이 전제되어야 한다. 데이터를 동일한 대상의 기준으로 맞출 수 없다면 데이터를 연계하는 것은 불가능하다. 따라서 일부 정보의 손실이 있다 하더라도, 동일한 대상으로 두 데이터를 표현할 수 있는 것이 중요하다. 예를 들어 예술인 대상의 통계와 미술가들의 통계가 있다면 예술인들 중에서 미술가들의 데이터만을 추출하여 미술가 데이터와 연계할 수 있다. 다른 예로 가구조사와 가구원 조사가 있다면, 두 데이터를 가구조사 또는 가구원 조사로 조사단위를 수정한 후 데이터 연계를 진행할 필요가 있다.

다음으로 데이터 기준시점은 데이터가 수집된 목적과 시점에 따라 다를 수 있다. 예를 들면 2016년 실시된 국민여가활동조사와 문화향수실태조사의 기준시점은 2015년 8월부터 2016년 7월이다. 하지만, 국민여행실태조사 기준시점은 데이터를 생성하기 때문에 해당 연도의 1월부터 12월까지이다. 이러한 경우 문화향수실태조사와 국민여행실태조사 데이터를 연계한다면 두 데이터를 동일한 시점으로 맞춰야할지, 아니면 그대로 사용하여도 문제없는지를 판단해야 한다. 만약 시점을 맞춰야만 데이터 연계의 의미가 있다면, 매년 조사하는 국민여행실태조사의 데이터를 2015년의 8월 이후 데이터와 2016년의 7월 이전 데이터로 새로 병합하여 새로운 데이터를 생성하는 과정이 필요하다.

2. 표준화 설계

연계할 데이터의 대상과 시점에 대해 통일하였다면, 표준화 방안에 대한 설계가 필요하다. 데이터는 생산하는 목적과 과정 그리고 생산기관(또는 책임자)의 특성을 반영하고 있다. 따라서 데이터를 표준화하기 위해서는 데이터가 가진 고유의 특성을 파악하는 것이 선행되어야 한다. 데이터 고유의 특성값은 수집된 데이터의 항목명(Column), 항목별 데이터값(Code), 데이터의 유형(Type), 길이(Length) 등이 해당된다. 파악된 데이터의 특성을 토대로 각자의 데이터를 총괄할 수 있는 데이터 표준화 방안을 마련한다.

연계할 데이터의 변수명이 같다면 동일한 데이터라고 생각하는 것은 기본적인 개념이다. 따라서 같은 의미의 변수에 대해서는 변수 명칭을 통일하여 표준화를 시켜야 한다. 하지만 실제 데이터의 항목명은 가장 많이 사용하는 성별, 연령, 지역(시도) 등에 대해서도 항목명이 상이한 경우가 대부분이다. 따라서 항목명을 표준화 할 때 항목명 생성 규칙(Naming Rule)을 정립하고, 표준용어를 사용하면, 데이터 연계 이후 분석 활용에도 효율적이다. 항목명 생성 규칙은 단위 단어의 요약명칭을 비롯하여 명칭의 정의까지 포함한다.

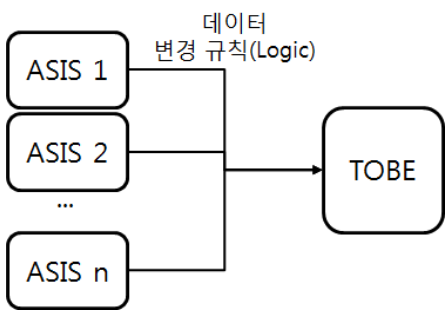
〈표 4-19〉 항목명 생성 규칙 예시

항목명	정식명칭	설명
CD	CODE	– 데이터 값이 코드형 데이터인 경우, 항목명에 CD를 추가 – 예시) 시도코드 : SIDO_CD
DT	DATE	– 날짜 데이터로, 세분화 하여 연월일의 형태로 이루어진 데이터에만 DT를 붙이고, 연도 항목이나 월 항목 등에 대해서는 다른 항목명으로 생성 – 예시) 여행일자 : TVL_DT
ADDR	ADDRESS	– 주소를 나타내는 항목에 붙이며, 상세주소를 별도로 관리하는 경우는 상세를 뜻하는 명칭에 대한 항목명을 추가하여 사용
…	…	…

항목명 표준화 외에도 항목별 데이터 값의 표준화 역시 필요하다. 데이터 값의 유형은 여러 가지로 구분할 수 있지만, 크게 두 가지로 보자면 연속형과 범주형이다. 데이터의 유형이 연속형일 때 데이터의 규모가 같아야 하기 때문에 데이터 간의 단위가 동일한지 확인해야 한다. 단위가 동일하다면 같은 숫자는 같은 크기를 나타내게 된다. 항목이 범주형인 경우에는 각 범주가 갖는 값의 의미를 파악하여, 같은 의미를 갖는 데이터 값에 대해서는 동일 범주를 갖도록 재범주화를 진행해야 한다.

3. 매핑(mapping)[27] 정의

데이터에 대한 표준화를 설계하였다면, 이후에는 실제 데이터를 표준화 데이터로 변환하는 작업이 필요하다. 각 데이터의 값들이 어떤 표준값으로 변환되는지를 표현하는 것이 데이터 매핑이다. 즉, 데이터 매핑이란 기존(AS-IS)의 데이터를 특정한 규칙(Logic)에 따라 새로운(TO-BE) 데이터로 대응시키는 일종의 함수라고 할 수 있다.



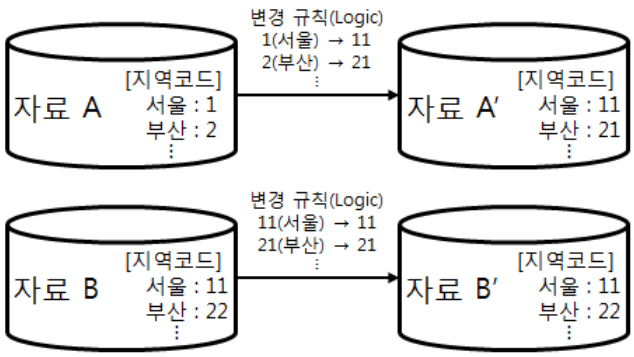
[그림 4-3] 데이터 매핑의 개념

각각의 데이터는 자료를 수집하는 목적(연구 목적) 또는 연구자의 성향에 따라, 변수명을 정하고 각 범주를 나누는 기준이 달라진다. 예를 들어 대다수의 조사에서 지역 정보를 수집하지만, 이러한 지역 정보에

27) 매핑(mapping)이란 단어는 지도를 만든다는 뜻과 사상(寫像), 함수라는 뜻을 함께 가지고 있다

대한 실제 데이터값, 즉 코드값은 개별 연구자(혹은 기관)가 정하는 기준에 따라 다르게 분류되어 저장된다.

[그림 4-4]의 데이터 매핑 예시를 보면, 자료 A는 지역코드는 서울이 '1', 부산이 '2' 등으로 할당되어 있으나, 자료 B는 서울이 '11', 부산이 '21' 등으로 할당되어 있다. 따라서 이에 대한 표준화는 서울 '11', 부산 '21'로 진행하였으며, 자료 A에 대한 변경규칙은 서울 '1'은 '11'로, 부산 '2'는 '21'로 세운 것을 알 수 있다. 이렇듯 동일한 데이터이지만, 값이 다르게 들어가 있는 것에 대해 앞서 설계한 표준화에 따라 변환 규칙을 정립하는 것을 데이터 매핑이라 할 수 있다.



[그림 4-4] 데이터 매핑의 예시

가. 매핑정의서[28)]

데이터의 표준화 과정에서 각 자료(조사, 행정 등)별 데이터(AS-IS)를 표준화한 데이터(TO-BE)로 변환하기 위한 규칙을 정의하는 것을 데이터 매핑이라 칭하였으며, 이를 기록하여 문서화한 것이 매핑정의서이다. 즉, 개별 데이터가 어떠한 매핑에 의하여 표준화 데이터로 변환되었는지 기록한 것이 매핑정의서라 할 수 있다.

28) 매핑정의서는 일반적으로 IT 업계에서 테이블을 생성하기 위한 규칙을 정의하기 위하여 주로 사용된다. 본 연구에서는 설명을 위해서만 사용하며, 실제사례에서는 매핑정의서를 작성하지는 않도록 한다.

일반적으로 매핑정의서에는 원자료에 대한 자료명, 항목명, 변환규칙, 변환된 자료명, 변환항목명 등이 포함되며, 이를 통해 원자료가 어떠한 규칙에 의해 변환되었는지 확인 가능하다. 또한 매핑정의서의 기록은 데이터 변환의 이력을 관리 할 수 있는 장점도 있다.

〈표 4-20〉 매핑정의서 예시

원시 데이터(AS-IS) 자료명 : A				표준화 데이터(TO-BE) 자료명 : STD_A				변환규칙	비고
항목명	값	유형	길이	항목명	값	유형	길이		
SQ1	1	숫자	2	SIDO_CD	11	숫자	2	1 -〉 11	'서울'
SQ1	2	숫자	2	SIDO_CD	21	숫자	2	2 -〉 21	'부산'
SQ1	3	숫자	2	SIDO_CD	22	숫자	2	3 -〉 22	'대구'
...				...				...	...

나. 매핑테이블(Mapping Table)

매핑테이블이란, 경우에 따라 다르겠으나 본 연구에서는 각 코드별로 표준화된 값을 DB(DataBase)화 시켜놓은 것을 의미한다. 따라서 일종의 참조표라 할 수 있다. 이러한 매핑테이블은 다양한 코드값들을 표준화된 값과 미리 매핑을 시켜놓음으로써, 동일한 코드값을 갖는 자료들의 변환을 보다 손쉽게 작업할 수 있다는 장점이 있다.

[그림 4-4]에서 예시로 들었던 지역코드의 경우, 〈표 4-21〉과 같이 매핑테이블이 만들어질 수 있다. 이러한 매핑테이블은 변수별로 매핑이 보기 쉽게 정리되어 있으며, 매핑의 관리가 용이한 장점이 있다. 또한 데이터 매핑 관련 프로그램 만들 때 매핑의 정의를 매번 직접 입력하는 작업을 줄임으로써, 사람에 의한 실수를 줄일 수 있고, 표준화 데이터의 변경이 일어날 경우에도 프로그램의 변경 없이 매핑테이블 데이터만 변경하면 되므로, 작업을 단순화할 수 있다.

〈표 4-21〉 매핑테이블 예시

구분	데이터	원시 데이터(AS-IS)		표준화 데이터(TO-BE)	
		코드	코드명	코드	코드명
지역코드	A	1	서울	11	서울
		2	부산	21	부산
		3	대구	22	대구
		...	...	...	...
	B	11	서울	11	서울
		21	부산	21	부산
		22	대구	22	대구
		...	...	...	...

4. 변환

연계할 데이터를 동일한 시점과 대상을 기준으로 확보하였고, 이에 대한 표준화 방안을 마련하여 매핑까지 정의했다면, 정의된 매핑에 의하여 데이터를 변환해야 한다. 이 단계에서 연계에 활용할 표준화된 데이터가 생성된다. 데이터의 변환은 앞서 정의한 매핑에 의하여 이루어지며, 원자료를 매핑규칙에 적용하여 신규 데이터를 생성한다.

표준화된 데이터는 변환에서 끝나는 것이 아니라, 변환된 데이터가 매핑 정의에 맞도록 변환된 것인지에 대한 검증이 필요하다. 이러한 검증은 일반적으로 범주형 항목의 경우 범주별 빈도가 변환 전과 후에 일치하는지를 파악함으로써 이루어진다. 변환 전과 후의 범주값에 대한 비교는 교차표 형태로 검증하는 것이 가장 정확하다. 연속형 변수의 경우는 합계가 변환 전과 후에 변화가 있는지를 비교함으로써 검증하는 것이 일반적이나, 데이터 연계 시에는 연속형 변수에 대한 변환은 실질적으로 이루어지기가 힘들다. 단, 연속형 변수를 범주화하는 경우는 변환 이후에 범주별 최소값과 최대값을 파악함으로써 검증할 수 있다.

제4절

소결

문화·체육·관광 관련 정책 또는 연구에 활용할 데이터에 대한 필요성은 중요하게 인식하고 있지만, 실제로 생산되고 있는 데이터는 매우 부족한 실정으로 새로운 정책이나 연구를 할 경우 기존의 데이터를 활용하려는 노력이 많이 이뤄지고 있다. 그러나 새로운 정책을 위한 연구를 할 때 필요한 정보 모두를 가지고 있는 데이터는 많지 않다. 이러한 경우 데이터 연계를 이용하여 통합 파일을 만들어 사용할 수 있다.

정책적 활용을 위한 데이터는 대표성을 만족해야 활용하는데 신뢰할 수 있다. 따라서 본 연구에서는 통계청의 국가승인통계만을 데이터 연계 대상으로 하며, 문화·체육·관광 관련 국가승인통계는 전체 25종으로 조사통계가 17종, 보고통계가 6종, 가공통계가 2종이다. 이중에서 데이터 연계에 활용할 데이터는 조사통계 데이터이며, 그 이유는 통계적 연계에서는 공통 변수가 있어야 가능하기 때문이다. 그리고 가공통계인 '문화체육관광산업통계'는 사업체고유번호를 이용하여 정확 연계한 통계이다.

문화·체육·관광 데이터를 연계하기 위해서 조사통계 중 국민을 대상으로 하는 통계들을 자세하게 살펴보았는데, 공통 변수로 사용할 수 있는 변수들이 대부분 범주형 자료라는 특성이 있다. 이는 조사 통계이기 때문에 응답자들의 피로를 고려한 설문 설계에 기인한다. 범주형 자료가 많을 경우 공통 변수들을 이용하여 유사성을 측정할 때 동점이 많이 발생하기 때문에 1대 1 연계가 되기 쉽지 않은 문제가 있다.

국민대상의 조사통계에서 공통 변수로 이용할 수 있는 변수들을 살펴보면 동일한 기준으로 되어있지 않다는 것을 파악할 수 있다. 따라서 데이터 표준화를 통해 매핑하려는 항목 내의 범주가 동일한 의미를 가지는 항목인지 살펴볼 필요가 있다. 예를 들어, 직업군의 경우 데이터를 수집하

는 목적에 따라 직업군을 다른 분류 기준을 적용시킬 수 있으므로 이를 표준화하고 매핑을 정의할 때 항목별 범주별 정의, 분류 기준을 명확히 해야 한다.

표준화한다는 것은 두 데이터가 동일한 기준이 아니라는 뜻이기 때문에 동일한 기준으로 데이터를 맞추는 과정이다. 동일한 기준으로 맞출 경우 세밀한 기준보다는 대략적인 기준이 맞추기가 더 쉽다. 예를 들어 나이 구간을 '15~19세, 20~24세, … , 65~69세, 70세 이상'으로 구분한 데이터와 '10대, 20대, 30대, 40대, 50대, 60대, 70대, 80대 이상'으로 된 데이터를 연계할 때, 연령대라는 변수의 구간은 '10대, 20대, 30대, 40대, 50대, 60대, 70대 이상'으로 표준화 시키게 된다. 즉, 가지고 있는 세밀한 정보보다 부족한 정보로 표준화가 이루어지게 되는 것이다.

데이터를 생산할 때 표준화 규칙이 정해져 있다면 정보를 그대로 활용할 수 있지만, 데이터를 각각 생산한 이후 표준화 시킨다면 가지고 있는 정보를 그대로 활용하는 것이 아니기 때문에 일부의 정보 손실이 발생할 수밖에 없다. 따라서 데이터를 생산하기 전에 데이터 활용도를 감안하여 인구통계학적 변수에 대한 표준화를 고려할 필요가 있다.

제5장

데이터 연계 활용

제1절

실제자료를 이용한 데이터 연계

제5장에서는 앞서 살펴봤던 문화·체육·관광 데이터 중에서 몇 개의 데이터를 기준 데이터로 활용하여 다른(분야) 데이터를 연계시키고, 연계하는 과정을 세밀하게 살펴보도록 한다. 실제 데이터 연계 역시 정확 연계와 통계적 연계로 구분하며, 실제 데이터 연계 과정과 함께 분석 결과까지 검토하도록 한다. 물론 개인정보보호 등의 문제로 데이터를 활용하는데 한계가 있고, 분석 결과를 일반화하는데 제약이 있을 수 있지만, 연계 방안을 구체적으로 살펴봄으로써, 향후 데이터 연계를 이용하여 분석하는 방향과 활용가능성 측면에 초점 맞춰 살펴보도록 한다.

정확 연계에서는 카드 데이터와 기상청 데이터 연계를 통한 기후에 따른 이동지출 분석과 문화체육관광 분야 사업체 표본틀과 통계청 행정 데이터 연계를 통한 문화체육관광산업 경영활동 현황 분석을 진행하였으며, 이들 각각의 연계를 위한 연계 변수에는 카드 가맹점/관측소별 위·경도와 사업체고유번호가 사용되었다.

통계적 연계에서는 국민여가활동조사와 문화향수실태조사의 데이터 연계를 통한 문화적 자본과 여가활동의 관계 분석과 국민여가활동조사와 한국의료패널조사 데이터 연계를 통한 여가활동의 건강효과 분석을 진행하였다. 블로킹 변수는 성별 변수가 두 연계에 공통으로 사용되었으며, 문화향수실태조사 데이터 연계 시에는 월평균가구소득 변수를 추가적으로 사용하였다. 공통 변수로는 지역, 연령, 혼인상태(배우자 유무), 동거가구원수, 학력 등의 변수가 사용되었다. 이에 대한 내용은 〈표 5-1〉에 제시하였다.

〈표 5-1〉 실제 데이터를 활용한 데이터 연계 개요

연계 방법	연계 변수	데이터 특성		분석
		기준 데이터	연계 데이터	
정확 연계	위 · 경도	카드 데이터	기상청 데이터	기후에 따른 이동지출 분석
	사업체고유번호	문화체육관광 분야 사업체 표본틀	통계청 행정 데이터	문화체육관광산업 경영활동 현황 분석
통계적 연계	연령, 지역, 학력, 동거가구원수, 혼인상태(배우자유무), 동거자녀수, 월평균본인소득	국민여가활동조사	문화향수실태조사	문화적 자본과 여가활동의 관계 분석
	연령, 지역, 학력, 동거가구원수, 혼인상태(배우자 유무), 월평균가구소득	국민여가활동조사	한국의료패널조사	여가활동의 건강효과 분석

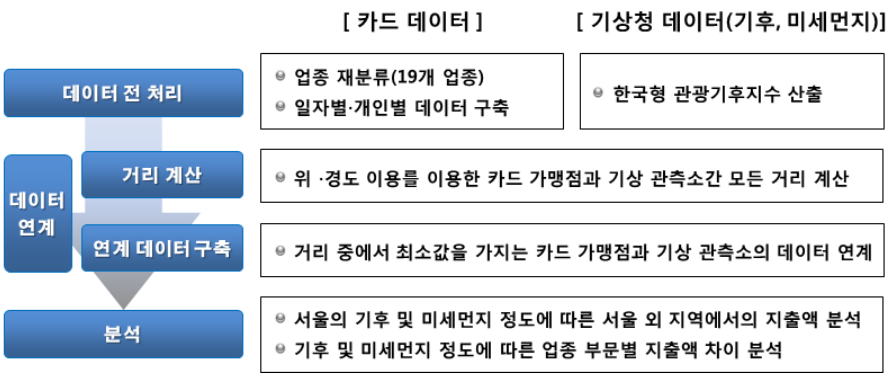
제2절

실제자료를 이용한 정확 연계

1. 카드 데이터와 기상청 데이터 연계를 통한 기후에 따른 이동지출 분석

카드 데이터와 기상청 데이터 연계를 통한 기후에 따른 이동지출 분석에서는 정확 연계 방법을 적용하여 카드 데이터와 기상청 데이터를 연계하여 통합 데이터를 구축한 후, 기상 상태(기후, 미세먼지)를 몇 단계의 등급으로 구분변수를 만들고, 구분된 기상 등급에 따른 여가 관련 신용카드 이동 지출의 차이를 비교·분석하고자 한다.

카드 데이터와 기상 데이터 연계는 크게 1) 연계를 위한 데이터 전처리, 2) 위·경도를 이용한 카드 가맹점과 기상 관측소 간 거리 계산과 거리 중에서 최소값을 가지는 데이터 연계, 3) 연계 데이터를 활용한 분석으로 나누어지며, [그림 5-1]과 같다.



[그림 5-1] 카드 데이터와 기상청 데이터 연계

가. 데이터 설명

1) 카드 데이터

카드 데이터는 내국인이 신한카드를 이용한 카드 사용 내역에 대한 데이터를 사용하였다. 신한카드 카드 지출액 데이터에는 성별, 연령대, 거주 지역(시·도, 시·군·구), 소득, 직업군, 혼인 여부, 자녀 여부 등의 인적 정보와 카드 지출 가맹점 지역(시·도, 시·군·구), 카드 지출 가맹점 위·경도, 업종, 지출 금액의 카드 사용 정보를 포함하고 있다.

본 연구에서 적용하려고 하는 이동 지출은 거주 지역 외의 지역에서 사용한 카드 지출 금액이다. 데이터 활용은 개인정보보호 등의 문제로 지역은 서울 거주자 5,500명, 기간은 2016년 3월 ~ 5월까지로 한정하였다. 이때, 분석 대상이 되는 5,500명은 2016년 3월 ~ 5월에 카드 사용 실적이 있는 서울 거주자로 서울 지역 25개 시군구별로 220명 씩 랜덤으로 추출한 후, 개인 식별정보를 제거한 상태의 자료이다.

2) 기상청 데이터

기상청 데이터는 기상자료 개방 포털(https://data.kma.go.kr/)에서 제공하고 있는 다양한 기상 데이터 중에서 기후 데이터와 미세먼지 데이터를 활용하였다. 기후 데이터에는 시간별, 일별, 월별 등으로 주요 관측소 지점에 대한 정보(위도, 경도)와 관측소별 평균 기온, 최저 기온, 평균 상대 습도, 최소 상대 습도, 일 강수량, 평균 풍속 등에 대한 정보가 포함되어 있다. 미세먼지 데이터에는 주요 관측소 지점에 대한 정보(위도, 경도)와 관측소별 미세먼지농도($\mu g/m^3$)에 대한 정보가 포함되어 있으며, 미세먼지 측정 관측소와 기후 데이터 측정 관측소에는 차이가 있다.

기후 데이터에는 측정 관측소가 94개, 미세먼지 데이터에는 측정 관측소가 28개가 있으며, 2016년 3월 ~ 5월까지로 기간을 한정[29]하였기 때문에 기후

29) 기상청 데이터와 연계하는 카드 데이터의 개인정보보호 문제로 기간을 한정한다.

데이터는 총 8,648개, 미세먼지 데이터는 2,576개의 데이터를 포함한다.

〈표 5-2〉 기후 데이터

일시	지점	위도	경도	평균 기온 (°C)	최고 기온 (°C)	일 강수량 (mm)	평균 풍속 (m/s)	최소 상대 습도 (%)	평균 상대 습도 (%)	합계 일조 시간 (hr)
2016-03-01	90	38.2509	128.5647	-0.3	5	0	2.1	21	37	10.2
2016-03-02	90	38.2509	128.5647	5.8	9.6	0	2.4	39	48	8.8
2016-03-03	90	38.2509	128.5647	7.2	11.5	0	2.2	52	79.1	9.9
2016-03-04	90	38.2509	128.5647	4.6	6	0	2.2	80	92.8	0
...										
2016-03-01	95	38.1479	127.3042	-4.7	1.1	0	1.9	21	51.6	9.7
...										

자료 : 기상자료 개방 포털(https://data.kma.go.kr/)

나. 데이터 전 처리

1) 카드 데이터

카드 데이터에는 카드사의 기준에 따라 업종을 분류하고 있기 때문에, 통계청의 공식분류체계와는 차이가 난다. 따라서 카드 데이터를 활용하기 위해서는, 분석의 목적에 따라 업종을 재분류할 필요가 있다. 본 연구에서 업종은 여가와 관련된 업종으로 한정하며, 한국문화관광연구원에서 발간하고 있는 '국민여가 관련 신용카드 지출액 현황 분석' 보고서 기준으로 종합 쇼핑, 외식, 유흥, 공연 관람, 미술·공연 참가, 운동경기 관람, 골프, 스키, 숙박 등의 19개 업종으로 재분류하였다.

그리고 카드 데이터의 경우에는 개인별이 아니라 결재 이용 건별로 데이터가 구축되어 있기 때문에 기상청 데이터와의 연계를 위해 일자별·개인별 데이터 구축이 필요하다. 결재 이용 건별 카드 데이터를 일자별, 고객 개인별로 동일한 결제 지역(서울, 광역시와 도의 경우에는 시단위), 동일한 업종(19개 업종)에서 발생한 매출액은 합산하여 일자별·개인별

50,600개의 카드 데이터를 구축하였다. 그리고 위·경도의 기준이 되는 가맹점의 경우에는 매출액이 합산되는 동일한 결제 지역, 동일한 업종에서 결제 금액이 가장 많이 발생한 가맹점의 위·경도를 포함하였다.

2) 기상청 데이터

다양한 기상 자료들을 포함하고 있는 기상청 데이터를 데이터 연계에 이용하기 위해서는 기상 상태의 좋고 나쁨을 판단할 수 있는 근거 및 자료가 필요하며, 이를 위해 한국형 관광기후지수를 별도로 산출하였다.

기상청에서는 다양한 기상정보 및 관광정보를 활용하여 관광하기 적합한 날씨를 지수화한 한국형 관광기후지수를 개발하여, 관광객에게 일정 및 관광지 선택의 의사결정에 도움을 주고 있다. 한국형 관광기후지수는 해외 관광기후지수 모형을 기반으로 날씨 영향도가 높은 국내 관광지를 선정하고 기상 요소별 가중치를 재산정하여 산출된다. 한국형 관광기후지수는 기상 자료와 관광지의 월별 관광객 수를 활용하여 개발한 지수로 기상 자료 역시 월별 평균 기상 자료를 활용하여 일별로 정확한 관광에 적합한 날씨를 판단하기에는 어려움이 있는 것으로 판단되나, 현재 다양한 기상 자료들을 활용하여 기상 상태의 좋고 나쁨을 나타낼 수 있는 별도의 지표·지수가 없기 때문에 본 연구에서는 시범적으로 활용해보고자 한다.

〈표 5-3〉 한국형 관광기후지수 산출식

해외 관광기후지수 = 2 × (4(HN) + (HD) + (W) + 2(R) + 2(SI)
한국형 관광기후지수 = 2 × (2.46(HN) + 2.37(HD) + 1.63(W) + 1.79(R) + 1.74(SI)
* HN : 한낮 열쾌적성　　* HD : 평균 열쾌적성
* W : 바람 지수　　* R : 강수 지수
* SI : 일사 지수

기상청의 기후 데이터를 이용하여 2016년 3월 ~ 5월까지의 일자별, 관측소별 한국형 관광기후지수를 산출하였으며, 세부 요소별 산출 방법

은 〈표 5-3〉과 같다.

열쾌적성은 기온과 습도를 이용한 열지수(Heat Index: HI) 수식[30]을 적용하여 한낮열지수(최고기온, 최소상대습도)와 일평균열지수(평균기온, 평균상대습도)를 산출하고, 이를 각 구간 값에 따라 점수화하여 한낮 열쾌적성과 일평균 열쾌적성을 산출한다. 바람지수는 기 개발 관광기후지수의 바람지수 점수 구간 값을 준용하되, ㎞/h단위를 ㎧단위로 변환 및 시간 해상도를 일치하여 지수화 하였다. 강수 지수는 강수량 범주를 적용하였으며, 강수 유무를 기준으로 강수량 발생 시 (-) 점수를 부여하는 등 강수지수의 구간 값을 조정하여 산출한다. 일사 지수는 운량별 평균 일조시간으로 06시~18시 동안의 운량에 따른 일별 하늘상태 평균을 적용하여 일평균 일사지수를 산출한다(기상청, 2015).

〈표 5-4〉 한국형 관광기후지수 4단계 등급

범위		등급	내용
90 ~ 100	이상적임	매우 좋음	야외 관광 활동을 하기에 매우 적합한 날씨
80 ~ 89	훌륭함		
70 ~ 79	매우 좋음	좋음	야외 관광 활동을 하기에 적합한 날씨
60 ~ 69	좋음		
50 ~ 59	괜찮음	보통	야외 관광 활동을 하기에 지장이 없는 날씨
40 ~ 49	좋지도 나쁘지도 않음		
30 ~ 39	나쁨	나쁨	야외 관광 활동을 하기에 적합하지 않은 날씨
20 ~ 29	매우 나쁨		
10 ~ 19	극히 나쁨		
〈 10	불가능함		

한국형 관광기후지수는 관광하기에 적합한 기상 기준에 따라 〈표 5-4〉에 제시한 것처럼 4단계 등급으로 구분하였다.

그리고 미세먼지는 2016년 미세먼지 예보 등급을 활용하여 1단계(좋

30) 열지수(Heat Index: HI) = $-42.379 + (2.04901523 \times T) + (10.14333127 \times R) - (0.22475541 \times T \times R) - (0.00683783 \times T^2) - (0.05481717 \times R^2) + (0.00122874 \times T^2 \times R) + (0.00085282 \times T \times R^2) - (0.00000199 \times T^2 \times R^2)$
(T : 기온(℉), R : 상대습도(%))

음, 15㎍/㎥ 이하), 2단계(보통, 15㎍/㎥ 초과 ~ 50㎍/㎥ 이하), 3단계(나쁨, 50㎍/㎥ 초과 ~ 100㎍/㎥ 이하), 4단계(매우 나쁨, 100㎍/㎥ 초과)로 미세먼지 단계를 구분하였다[31]. 이는 〈표 5-5〉에 제시하였다.

〈표 5-5〉 미세먼지 4단계 등급

범위	등급
100㎍/㎥ 초과	매우 나쁨
50㎍/㎥ 초과 100 ㎍/㎥ 이하	나쁨
15㎍/㎥ 초과 50㎍/㎥ 이하	보통
15㎍/㎥ 이하	좋음

다. 데이터 연계

카드 데이터와 기상청 데이터의 연계에는 기상청 데이터의 관측소별 위·경도와 카드 데이터의 가맹점 위·경도를 이용하였다. 카드 데이터의 가맹점 위·경도와 기상청 데이터의 관측소별 위·경도가 정확하게 일치하는 경우에 연계하는 것이 정확 연계이지만, 실제로 두 데이터의 위·경도가 일치하는 경우는 존재하지 않지만 같은 구역에 있는 대상들을 연결할 수 있다[32]. 동일한 위·경도 정보를 포함하고 있지 않은 지역별 날씨 예보의 경우에도 관측소별 위·경도를 이용하여 정보를 제공하고 있다는 측면에서 카드 데이터와 기상청 데이터의 연계는 정확 연계에 포함된다.

카드 데이터의 가맹점 위·경도와 기후 데이터의 관측소 위·경도의 거리를 계산하고, 카드 가맹점별로 94개의 관측소 중에서 가장 짧은 거리의 관측소를 선정하고 해당 관측소의 기후 정보를 연계하였다. 그리고 기후 데이터 관측소와 미세먼지 데이터의 관측소에는 차이가 있기 때문에 동일한 방법으로 카드 데이터의 가맹점 위·경도와 미세먼지 데이터의 관측소 위·경도의 거리를 계산하고, 카드 가맹점별로 28개의 관측소 중에서 가장 짧은 거리의

31) 미세먼지 예보 등급은 2018년 3월에 기준이 강화되어 좋음(15㎍/㎥ 이하), 보통(15㎍/㎥ 초과 ~ 35㎍/㎥ 이하), 나쁨(35㎍/㎥ 초과 ~ 75㎍/㎥ 이하), 매우 나쁨(75㎍/㎥ 초과)로 개정되었다.
32) 기상관측소와 가맹점은 정확히 일치할 수 없다. 그러나 하나의 기상관측소의 계측범위에 드는 가맹점은 기상 관점에서는 동일한 날씨 정보를 갖는다고 가정할 수 있다.

관측소를 선정하고 해당 관측소의 미세먼지 정보를 연계하였다.

〈표 5-6〉 카드 데이터와 기상청 데이터 연계 방법

카드 데이터			기상청 데이터		
카드 가맹점	위도	경도	기상 관측소	위도	경도
A_1	a_{11}	a_{12}	B_1	b_{11}	b_{12}
A_2	a_{21}	a_{22}	B_2	b_{21}	b_{22}
A_3	a_{31}	a_{32}	B_3	b_{31}	b_{32}
...			...		

▼

- 1단계 : 카드 가맹점과 기상 관측소간 모든 거리 계산($\sqrt{(a_{i1}-b_{i1})^2+(a_{i2}-b_{i2})^2}$)
- 2단계 : 거리 중에서 최소값을 가지는 카드 가맹점과 기상 관측소의 데이터 연계

라. 분석결과

기후 및 미세먼지에 따른 이동 지출 구조를 파악하기 위하여 두 가지의 분석을 하였다. 첫 번째는 서울의 기후 및 미세먼지 정도에 따른 서울 외 지역에서의 지출액의 차이가 있는지를 분석하였고, 두 번째는 서울 외 지역 에서 지출액이 있는 경우, 관광기후지수와 미세먼지 등급에 따른 업종 부문(취미·오락, 문화생활, 스포츠, 여행)별 지출액에 차이가 있는지를 알아보았다. 기후 및 미세먼지 정도에 따른 지출액 차이를 분산분석을 통해 살펴보았으며, 유의미한 차이가 있는 경우에는 회귀분석을 실시하여 어떠한 인과 관계가 있는지를 검토하였다.

본 절에서는 카드 데이터 이용에 있어서 개인정보보호 문제로 서울 거주자의 2016년 3월 ~ 5월의 여가 관련 카드 지출액만을 분석에 활용하였기 때문에 전체 결과 해석보다는 연계 방안을 중심으로 향후 분석 가능성 측면에 초점을 맞춰서 살펴보아야 할 것이다.

1) 서울의 기후 및 미세먼지 정도에 따른 서울 외 지역에서의 지출액 분석

서울의 기후 및 미세먼지 정도에 따른 서울 외 지역에서의 지출액에 차이가 있는지 분산분석을 통하여 검정한 결과를 〈표 5-7〉에 제시하였다. 서울의 경우 2016년 3월 ~5월에는 관광기후지수 등급과 미세먼지 등급 모두 4단계인 경우는 없었으며, 서울의 관광기후지수 등급과 미세먼지 등급에 따라 서울 외 지역에서의 카드 지출액에는 차이가 있는 것으로 나타났다.

〈표 5-7〉 서울의 기후 및 미세먼지 정도 따른 서울 외 지역에서의 지출액 분산분석

구분	등급	사례수	카드 지출액(원)		F	유의확률
			평균	표준편차		
관광기후지수 등급	1	242,000	1,291.71	26,389.86	20.834	〈 0.001
	2	258,500	997.21	14,870.62		
	3	5,500	2,354.91	25,264.71		
미세먼지 등급	1	27,500	1,640.91	23,180.90	8.255	〈 0.001
	2	269,500	1,154.66	24,361.45		
	3	209,000	1,086.22	16,148.23		

주 : 관광기후지수 등급이 높을수록 관광하기 적합한 날씨를 의미하며, 미세먼지 등급이 높을수록 미세먼지 농도가 낮음

서울의 관광기후지수 등급과 미세먼지 등급이 아닌 관광기후지수와 미세먼지 농도를 독립변수, 서울 외 지역에서의 카드 지출액을 종속변수로 하는 회귀분석을 실행한 결과는 〈표 5-8〉과 같다. 서울의 관광기후지수가 낮을수록, 미세먼지 농도가 높을수록 서울 외 지역에서의 지출액이 많아지는 것으로 나타났다. 지역, 기간 등이 한정되어 있어 전체로 판단하기에는 한계가 있어, 향후 지역, 기간 등을 확대하여 분석하여 일반화시킬 필요성이 있다.

〈표 5-8〉 서울의 기후 및 미세먼지 정도 따른 서울 외 지역에서의 지출액 회귀분석

	계수	표준오차	t	유의확률
상수	928.45	79.24	11.717	〈 0.001
관광기후지수	-3.58	1.31	-2.745	0.006
미세먼지 농도	5.91	1.29	4.596	〈 0.001

2) 기후 및 미세먼지 정도에 따른 업종 부문별 지출액 차이 분석

서울 외 지역에서 지출액이 있는 경우, 해당 지역의 기후 및 미세먼지 정도에 따른 업종별 지출액에 차이가 있는지 분산분석을 통하여 검정하였다. 검정한 결과는 〈표 5-9〉와 〈표 5-10〉에 제시하였으며, 여행 부문(숙박, 교통, 여행사 등)의 경우에 카드 지출 지역의 관광기후지수 등급에 따라 카드 지출액에 차이가 있는 것으로 나타났다.

〈표 5-9〉 관광기후지수 등급에 따른 업종 부문별 지출액 분산분석

업종 부문	관광기후 지수 등급	사례수	카드 지출액(원)		F	유의확률
			평균	표준편차		
취미 · 오락	1	6,854	35,203.24	79,711.32	0.830	0.477
	2	5,702	32,968.43	130,234.60		
	3	680	32,927.94	67,437.46		
	4	25	13,760.00	12,774.58		
문화생활	1	6,854	202.80	3,922.27	0.899	0.441
	2	5,702	115.40	1,642.81		
	3	680	211.76	2,336.09		
	4	25	200.00	707.11		
스포츠	1	6,854	5,187.04	58,601.82	0.146	0.932
	2	5,702	4,656.09	40,176.02		
	3	680	5,375.00	39,063.62		
	4	25	2,800.00	14,000.00		
여행	1	6,854	5,671.29	36,121.10	3.244	0.021
	2	5,702	3,767.27	35,342.39		
	3	680	4,327.94	18,739.43		
	4	25	160.00	800.00		

주 : 관광기후지수 등급이 높을수록 관광하기 적합한 날씨를 의미함

〈표 5-10〉 미세먼지 등급에 따른 업종 부문별 지출액 분산분석

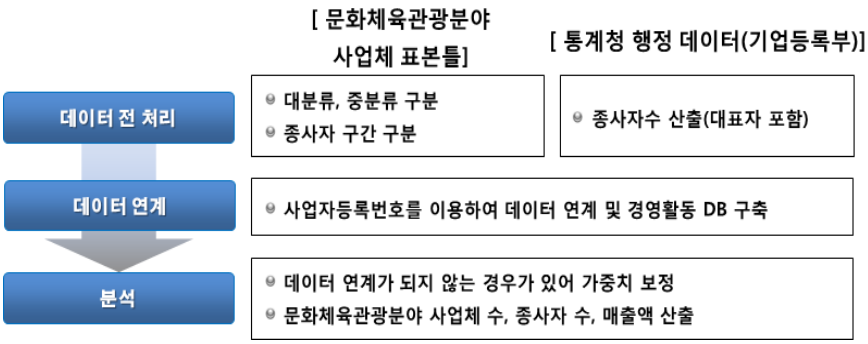
업종 부문	메시먼지 등급	사례수	카드 지출액(원)		F	유의확률
			평균	표준편차		
취미 · 오락	1	199	28,301.51	45,443.49	1.166	0.321
	2	6,985	32,917.54	68,328.66		
	3	5,358	36,020.90	142,326.29		
	4	719	32,605.01	49,899.67		
문화생활	1	199	145.73	1,551.76	0.140	0.936
	2	6,985	169.94	3,155.51		
	3	5,358	152.48	3,024.46		
	4	719	228.09	2,775.89		
스포츠	1	199	2,703.52	16,150.41	1.253	0.289
	2	6,985	4,985.54	55,800.64		
	3	5,358	4,578.76	37,212.45		
	4	719	8,248.96	79,630.26		
여행	1	199	5,281.41	22,828.11	1.667	0.172
	2	6,985	4,648.68	27,025.14		
	3	5,358	4,539.19	36,367.51		
	4	719	7,588.32	74,802.34		

주 : 관광기후지수 등급이 높을수록 관광하기 적합한 날씨를 의미함

2. 문화체육관광 분야 사업체 표본틀과 통계청 행정 데이터 연계를 통한 문화체육관광산업 경영활동 현황 분석

문화체육관광 분야 사업체 표본틀과 통계청 행정 데이터 연계를 통한 문화체육관광산업 경영활동 현황 분석에서는 실제로 국가승인통계인 문화체육관광산업통계 작성을 위해 수행한 정확 연계 방법을 설명한다.

문화체육관광 분야 사업체 표본틀과 통계청 행정 데이터 연계는 크게 1) 연계를 위한 데이터 전 처리, 2) 사업자등록번호를 이용하여 데이터 연계 및 경영활동 DB 구축, 3) 데이터 연계가 되지 않은 일부 데이터에 해당하는 부분에 대한 가중치 보정을 함으로써 문화체육관광 분야 사업체 수, 종사자 수, 매출액 산출로 나누어지며, [그림 5-2]와 같다.



[그림 5-2] 문화체육관광 분야 사업체 표본틀과 통계청 행정 데이터 연계

가. 데이터 설명

1) 문화체육관광 분야 사업체 표본틀

문화체육관광 분야 사업체 표본틀은 통계청의 전국사업체조사 데이터에서 문화체육관광 분야 사업체를 선별하여 구축한 데이터이다. 전국사업체조사 데이터의 '무엇을 가지고(원재료, 영업장소 등)', '어떤 방법으로(주요 영업, 생산 활동)', '생산, 제공하였는가(최종 재화, 용역)' 변수를 활용하여 문화체육관광산업을 영위하는 사업체를 선별하였으며, 이때 선별하는 기준으로 〈표 5-11〉에 제시한 문화·체육·관광 분야 산업분류를 활용하였다. 문화체육관광 분야 사업체 표본틀에는 문화·체육·관광 분야 산업분류와 함께 지역, 종사자수 등의 정보를 포함하고 있다.

〈표 5-11〉 문화·체육·관광 분야 산업분류

활용 산업분류	설명
저작권산업 특수분류	저작권산업특수분류는 세계지적재산권기구(World Intellectual Property Organization, WIPO)의 「저작권기반산업의 경제적기여도 조사 가이드」를 근거로 저작권산업의 정책수립에 필요한 분류체계를 수립하기 위하여 2011년에 제정되었다. 저작권산업특수분류의 포괄 범위는 세계지적재산권기구(WIPO)의 저작권산업 포괄 범위를 참고하여 핵심 저작권산업을 중심으로 저작권산업에 대한 기여도와 역할에 따라 상호의존, 부분적용, 지원산업으로 구분하고 있다.

활용 산업분류	설명
콘텐츠산업 특수분류	콘텐츠산업특수분류는 콘텐츠미디어산업 분류, 2009 UNESCO Framework for Cultural Statistics 등을 토대로 2010년 제정되었으며, 1차 개정까지 이루어졌다. 콘텐츠산업특수분류에서는 콘텐츠산업을 크게 12개 산업으로 구분하고 있으며, 이를 기준으로 국가승인통계 콘텐츠산업통계조사가 생산되고 있다.
스포츠산업 특수분류	스포츠산업특수분류는 국민체육진흥법, 체육시설의 설치·이용에 관한 법률 등의 스포츠산업 정의, 중앙생산물분류(CPC 2.0)의 스포츠관련 산업 및 생산물을 바탕으로 2000년에 제정되었으며, 3차 개정까지 이루어졌다. 스포츠산업특수분류는 '국민체육진흥법에 따른 체육활동을 지원하는 제조업, 건설업, 관련 서비스업과 스포츠서비스를 제공하기 위해서 재화와 서비스를 생산 유통하는 산업을 포함하고 있으며, 이를 기준으로 국가승인통계 스포츠산업실태조사가 생산되고 있다.
관광산업 특수분류	관광산업특수분류는 세계관광기구(UNWTO)와 유엔통계위원회가 공동으로 작성한 국제관광표준분류(Standard International Classification of Tourism Activities)를 기초로 2000년에 제정되었으며, 3차 개정까지 이루어졌다. 관광산업특수분류의 포괄범위는 저작권산업특수분류와 유사하게 핵심 관광산업을 중심으로 관광산업에 대한 기여도와 역할에 따라 상호의존, 부분적용, 지원산업으로 구분하고 있다.
문화예술 산업분류	문화예술산업의 경우에는 한국표준산업분류에서 공연예술 위주로 창작, 예술 및 여가관련 서비스업(90)으로만 분류되어 있고 별도의 산업특수분류가 제정되어 있지 않아 자체적으로 구축한 산업분류를 활용하고 있다.

자료 : 한국문화관광연구원(2018), 2016년 기준 문화체육관광산업통계

〈표 5-12〉 문화체육관광 분야 사업체 표본틀

전국사업체조사 데이터						문화체육관광산업 해당 여부
사업체 고유번호	지역	한국 표준산업분류	무엇을 가지고	어떠한 방법으로	생산, 제공하였는가	
111111	서울	만화출판업	출판사	일반인 대상으로	만화책출판	O
222222	대전	종합스포츠 시설 운영업	휘트니스	체력단련시설을 갖추고	체력단련서비스	O
333333	서울	비알콜 음료점업	카페	접객시설	커피, 쥬스	X
444444	서울	계약배달 판매업	신문	스포츠한국 무료 배포	일간지배포	O

전국사업체조사 데이터						문화체육 관광산업 해당 여부
사업체 고유번호	지역	한국 표준산업분류	무엇을 가지고	어떠한 방법으로	생산, 제공하였는가	
555555	부산	피부미용업	피부관리실	피부마사지	피부관리서비스	X
…						

▼

문화체육관광 분야 사업체 표본틀						
사업체 고유번호	지역	한국 표준산업분류	무엇을 가지고	어떠한 방법으로	생산, 제공하였는가	문화체육 관광분야 분류
111111	서울	만화출판업	출판사	일반인 대상으로	만화책출판	문화산업
222222	대전	종합 스포츠시설 운영업	휘트니스	체력단련시설 을갖추고	체력단련서비스	스포츠 산업
444444	서울	계약배달 판매업	신문	스포츠한국 무료 배포	일간지배포	문화산업
…						

2) 통계청 행정 데이터

통계청에서는 2017년 기준으로 행정안전부, 국토교통부, 국세청 등 73개 기관에서 주민등록자료, 건축물대장, 4대보험, 과세자료 등 행정 자료 208종을 입수하여 행정 자료통합관리시스템[33]에 110종 DB로 구축(통계청, 2017)하였으며, 구축한 DB 중에서 기업등록부(Business Register: BR)를 활용하였다. 기업등록부는 통계청이 입수한 행정 자료(사업자등록자료)와 통계청의 조사 자료(전국사업체조사)를 연계하여 구축한 데이터로 사업체고유번호[34], 주 업종 코드, 종사자수, 매출액 등의 정보를 포함하고 있다.

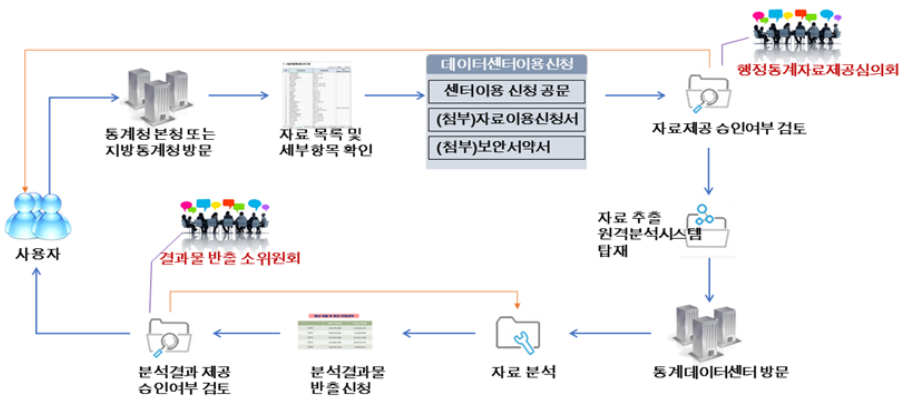
통계청의 행정 데이터는 개인정보보호 문제로 데이터 보안이 가능한

33) 통계청에서 입수한 행정 자료를 DB로 구축하여 체계적으로 관리하고 통계작성에 효율적으로 이용할 수 있도록 지원하는 시스템이다.

34) 통계청에서는 사업체마다 고유한 번호를 부여하여 이를 관리하고 있으며, 동일한 사업체는 매년 동일한 사업체고유번호를 부여받는다.

통계청의 통계빅데이터센터에서만 이용이 가능하며, 통계빅데이터센터를 이용하기 위해서는 사전에 이용 신청 및 승인을 받아야 한다. 빅데이터센터 이용 절차는 [그림 5-3]과 같다(통계청, 2016).

- (1단계) 센터 이용 절차 및 이용자료 상담
- (2단계) 자료 이용 신청
- (3단계) 자료 요청 항목에 관한 자료제공 여부 결정
- (4단계) 좌석 배정, 자료 추출 원격시스템 탑재
- (5단계) 센터 방문, 자료 분석 및 분석결과 반출 요청
- (6단계) 반출 승인 후 반출



[그림 5-3] 통계빅데이터센터 이용 절차

자료 : 통계청(2016), 통계데이터센터 이용 가이드

나. 데이터 전 처리

1) 문화체육관광 분야 사업체 표본틀

문화체육관광 분야의 산업별 경영활동 현황을 파악하기 위해서 문화체육관광 분야 사업체 표본틀의 분류를 대분류(문화산업, 예술산업, 스포츠산업, 관광산업), 중분류(19개)로 구분하였다.

그리고 종사자 규모별 사업체 수, 매출액을 산출하기 위해 표본틀의

종사자수를 7개의 구간(1인 ~ 4인, 5인 ~ 9인, 10인 ~ 19인, 20인 ~ 49인, 50인 ~ 99인, 100인 ~ 299인, 300인 이상)으로 나누었다. 종사자 규모는 통계청의 기업등록부 자료에도 포함되어 있으나, 표본틀과는 다소 차이가 있으며, 본 연구에서는 표본틀 기준으로 구간을 구분하였다.

2) 통계청 행정 데이터

문화체육관광 분야 사업체 표본틀과 통계청 기업등록부를 연계하여 경영활동 현황을 파악하기 위해서는 기업등록부의 사업체 조직 형태, 사업체 수, 매출액 등의 데이터를 활용해야 하며, 기업등록부를 이를 산출 가능한 데이터로 변경하였다.

본 연구에서의 종사자 수는 대표자를 포함한 임금근로자, 무급가족 종사자, 기타 종사자를 포함한 산업에 종사하는 전체 종사자 수를 의미한다. 하지만 기업등록부에서 종사자 수는 개인 사업체의 경우, 대표자가 제외되어 있으며, 법인의 경우에는 대표자 개념이 없어 전체 종사자 수를 의미한다. 따라서 기업등록부에서 조직 형태가 개인 사업체의 경우에는 기업등록부의 전체 종사자수에서 1을 더해서 전체 종사자 수를 변경하였다. 그리고 전체 종사자 수와 함께 남·여 종사자 수를 산출하기 위해서, 상용근로자(남·여), 임시 및 일용 근로자(남·여)로 구분되어 있는 변수를 합쳐서 남·여 종사자 수를 각각 생성하였다. 이 때에도 개인 사업체의 경우에는 종사자 수에 1을 더해야 하며, 1을 더하는 기준은 대표자의 성별 변수를 활용하여 대표자가 남성일 경우에는 남성 종사자수에 1을 더하고 여성일 경우에는 여성 종사자수에 1을 더했다.

다. 데이터 연계

문화체육관광 분야 사업체 표본틀과 통계청 기업등록부 연계에는 2015년 기준 데이터를 활용하였다. 일반적으로 사업체 자료의 정확 연계에는 사업자등록번호를 이용하고 있으나, 문화체육관광 분야 사업체 표

본틀에는 사업자등록번호 대신 사업체의 고유한 식별 변수로 사업체고유번호가 있어 이를 연계 변수로 사용하였다.

문화체육관광 분야 사업체 표본틀의 사업체고유번호와 기업등록부의 사업체고유번호를 연계하여 문화체육관광산업에 해당하는 사업체의 경영활동 DB를 구축하였다. 데이터 연계를 통해 구축된 문화체육관광 분야 사업체의 경영활동 DB에는 산업 대분류, 산업 중분류별 종사자수, 매출액 정보가 포함된다.

〈표 5-13〉 문화체육관광 분야 사업체 표본틀과 통계청의 행정 데이터 연계 방법

문화체육관광 분야 사업체 표본틀			연계	기업등록부		
사업체 고유번호	산업 대분류	산업 중분류		사업체 고유번호	종사자수	매출액
111111	문화산업	출판산업	←	111111	10명	10천만원
222222	스포츠산업	스포츠시설업	←	222222	20명	20천만원
444444	문화산업	출판산업	↙	555555	30명	30천만원
555555	예술산업	공연		666666	40명	40천만원
…				…		

▼

문화체육관광 분야 사업체의 경영활동 DB				
사업체고유번호	산업 대분류	산업 중분류	종사자수	매출액
111111	문화산업	출판산업	10명	10천만원
222222	스포츠산업	스포츠시설업	20명	20천만원
444444	문화산업	출판산업	–	–
555555	예술산업	공연	30명	30천만원
…				

라. 분석결과

데이터 정확 연계를 통해 문화체육관광 분야 사업체의 경영활동 DB를 구축하였으나, 실제로 문화체육관광 분야 사업체 표본틀과 기업등록부는 일부 연계가 되지 않는 경우가 있어 이에 대한 부분은 가중값으로 처리하여 추정하여 결과를 산출하였다. 가중값을 처리할 때 기준은 중분류(19개 구간)와 종사자 수 구간(7개 구간) 기준으로 다음과 같이 산정하였다.

가중값 = (해당구간의 연계된 사업체 수) / (해당 구간의 전국사업체 자료 수)

문화체육관광 분야 사업체 표본틀과 통계청 기업등록부 연계를 통한 2015년 기준 경영활동 현황은 〈표 5-14〉와 같으며, 2015년 기준 문화체육관광 관련 산업의 사업체 수는 약 48만개, 종사자 수는 약 173만명, 매출액은 약 302조원 가량으로 나타났다. 이 때, 하나의 사업체가 다수의 문화체육관광 관련 산업을 영위할 수 있기 때문에, 문화체육관광산업 전체의 합계는 산업별 결과의 합이 아닌 전체 결과를 사용해야 한다. 즉, 〈표 5-14〉에서의 문화체육관광산업의 값이 문화체육관광 관련 산업 전체의 합이다.

〈표 5-14〉 2015년 기준 경영활동 현황 총괄

(단위: 개, 명, 백만원)

		사업체 수	종사자 수	매출액
문화체육관광산업		478,586	1,732,935	302,508,091
문화산업	출판산업	29,223	133,778	20,078,105
	음악산업	33,766	42,549	3,770,623
	영화산업 및 방송산업	4,872	75,122	26,148,636
	광고산업	18,383	90,081	15,063,948
	게임산업	15,074	59,939	11,411,895
	시각그래픽 및 캐릭터	29,370	92,347	12,846,368
예술산업	문화유산 및 문화시설	3,757	43,221	3,711,565
	문학 및 출판	36,653	163,389	23,329,462
	공연	45,360	79,910	10,361,093
	시각예술	43,811	157,408	17,584,786
	공예	101,946	285,957	55,586,507
스포츠산업	스포츠시설업	40,165	172,959	25,980,209
	스포츠용품업	57,443	149,693	40,622,150
	스포츠서비스업	36,903	112,324	26,850,124
관광산업	관광숙박업 및 식당업	52,671	249,330	23,281,748
	여행사 및 관광운수업	11,258	115,945	25,028,309
	문화오락 및 레저산업	6,638	87,488	15,655,915
	관광쇼핑업	1,148	7,365	5,926,785
	국제회의 및 전시업	420	4,406	747,645

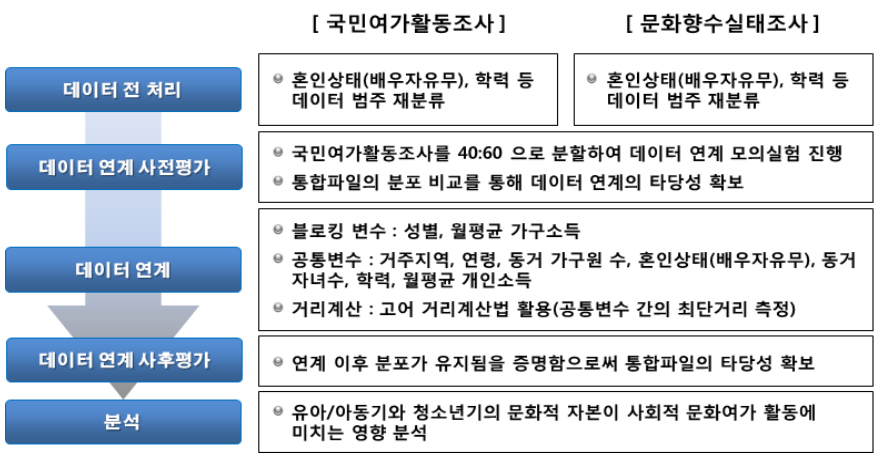
자료 : 한국문화관광연구원(2018), 2016년 기준 문화체육관광산업통계

제3절

실제자료를 이용한 통계적 연계

1. 국민여가활동조사와 문화향수실태조사를 이용한 문화적 자본과 여가활동의 관계 분석

국민여가활동조사와 문화향수실태조사는 데이터의 구성과 조사항목 등에서 유사한 부분이 있지만, 국민여가활동조사의 조사항목이 여러 여가활동에 대한 내용을 포괄적으로 다루고 있다면, 문화향수실태조사는 여가활동 중의 문화향유에 초점을 맞춰 세부적으로 질문하고 있다. 이처럼 유사하지만 세부적으로 차이가 있는 두 데이터의 연계를 통해 유년기와 청소년기에 형성된 문화적 자본이 향후 여가활동에 미치는 영향을 파악해 보고자 한다.



[그림 5-4] 국민여가활동조사와 문화향수실태조사 데이터 연계

가. 데이터 설명

1) 국민여가활동조사

국민여가활동조사는 국민의 여가활동 참여 실태와 만족도 등 여가행태를 파악할 수 있는 내용으로 설문이 구성되어 있어, 참여한 여가 유형 및 여가시간과 여가비용 등을 파악할 수 있다. 또한 국민들의 문화여가생활을 통한 행복수준을 측정하기 위한 문화여가행복지수[35]는 국민여가활동조사의 일부 항목을 활용하여 산출할 수 있다(문화여가행복지수, 2016). 본 절에서는 2016년도에 조사된 국민여가활동조사 데이터를 기준 파일로 하여 데이터를 연계하고자 한다.

2) 문화향수실태조사

문화향수실태조사는 국민여가활동조사에서는 다루지 않는 문화 향유의 경험이나 빈도, 만족도, 제반 상세사항 등에 대해 조사를 시행한다. 또한 조사기준 시점에 대한 문화예술교육의 경험뿐만 아니라, 유아/아동기 및 청소년기의 문화예술교육 여부에 대해서도 질문하고 있다. 이러한 유아/아동기 및 청소년기의 문화예술교육 경험이 문화적 참여, 여가비용과 여가시간 등의 여가활동에 영향을 미치는지 알아보고자 한다.

문화향수실태조사 데이터를 연계 파일로 하여 국민여가활동조사와 연계하기 때문에, 동일한 기준시점으로 조사된 2016년 문화향수실태조사 데이터를 대상으로 한다.

35) 문화여가행복지수는 시간 및 비용에 대한 지표에 대한 개인여건지수, 시설, 프로그램, 산업에 대한 자원지수, 자주하는 여가활동의 빈도 및 개수를 지표로 하는 참여지수, 여가활동에 대한 인식과 영향에 대한 태도지수, 여가생활 만족도를 나타내는 만족도지수 이렇게 5가지 지수의 평균값을 표준화하여 산출한다.

나. 데이터 전처리

국민여가활동조사와 문화향수실태조사의 데이터는 동일한 기준으로 조사설계를 하였기 때문에, 조사기준은 물론 인구통계학적 배경변수가 동일하다[36]. 따라서 데이터 연계할 때 공통 변수들 간 데이터의 값들을 조정하는 전처리 과정이 필요하지 않다는 장점이 있다. 하지만 두 데이터의 변수들 간의 형태에 대한 데이터 전처리 외에도 데이터 연계를 위한 전처리가 필요하다. 이를 위해 동거 자녀 수, 혼인상태, 종사상의 지위 등의 항목에 대해 데이터 전처리를 진행하였다.

두 조사에서는 동거 자녀수에 대해 미취학아동과 취학 자녀, 그리고 성인 자녀를 구분하여 조사하였기 때문에 별개의 변수로 구분하였다. 개별적으로 데이터 연계를 하는 것보다 전체 자녀수로 연계에 활용하는 것이 더 타당하다고 판단되어, 동거 자녀수로 합한 통합변수를 생성하여 활용하였다. 그리고 혼인상태에 대해서는 미혼, 배우자 있음, 사별, 이혼, 기타의 5가지 유형으로 분류하여 조사를 진행하는데, 연계에서는 미혼과 사별, 이혼, 기타를 '배우자 없음'으로 통합하여 배우자 유무라는 새로운 변수를 생성하여 활용하였다.

학력은 최종 학력과 이수 여부를 조합하여 '초졸 이하', '중졸', '고졸', '대졸 이상' 으로 구분하였다. 즉, 최종 학력의 이수 여부가 재학, 수료, 휴학, 중퇴에 해당하는 경우는 이전 학력에, 졸업인 경우에만 해당 학력에 포함하였다. '초졸 이하' 에는 무학과 초등학교 재학 등이 포함되어 있으며, '대졸 이상'은 대학교와 석사, 박사과정의 이수여부가 졸업인 경우만을 포함하였다.

월평균 개인/가구 소득에 대해서는 고소득일수록 전체적인 비율이 낮아지는 경향이 존재하여 고소득 일부 구간을 통합하였다. 그리고 월평균 가구소득은 소득이 없는 경우와 100만원 미만인 경우의 성향에 차이가

36) 문화향수실태조사에 '장애인 여부'의 항목이 하나 더 있는 것 외에 공통 변수는 모두 일치한다.

존재하지 않을 거라 판단되어 두 구간을 통합하였다.

국민여가활동조사와 문화향수실태조사의 데이터 연계를 위한 데이터 전처리 방법은 〈표 5-15〉에 구체적으로 제시하였다.

〈표 5-15〉 공통 변수 표준화 내역

구분	국민여가활동조사 (기준 데이터)	문화향수실태조사 (연계 데이터)	공통 변수 표준화
혼인상태	1. 미혼 2. 배우자 있음 3. 사별 4. 이혼 5. 기타	1. 미혼 2. 배우자 있음 3. 사별 4. 이혼 5. 기타	– 배우자 없음 – 배우자 있음
동거 자녀 수	1. 자녀없음 2. 미취학아동 자녀 __명 3. 취학 자녀 __명 4. 성인 자녀 __명	1. 자녀없음 2. 미취학아동 자녀 __명 3. 취학 자녀 __명 4. 성인 자녀 __명	동거 자녀 __명
최종 학력	1. 무학 2. 초등학교 3. 중학교 4. 고등학교 5. 대학교(4년제 미만) 6. 대학교(4년제 이상) 7. 대학원 석사 과정 8. 대학원 박사 과정	1. 무학 2. 초등학교 3. 중학교 4. 고등학교 5. 대학교(4년제 미만) 6. 대학교(4년제 이상) 7. 대학원 석사 과정 8. 대학원 박사 과정	– 초등학교 졸업 이하 – 중학교 졸업 이하 – 고등학교 졸업 이하 – 대학 졸업 이상
이수 여부	1. 졸업 2. 재학 3. 수료 4. 휴학 5. 중퇴	1. 졸업 2. 재학 3. 수료 4. 휴학 5. 중퇴	
월평균 본인 소득	1. 소득없음 2. 100만원 미만 3. 100만원 ~ 200만원 미만 4. 200만원 ~ 300만원 미만 5. 300만원 ~ 400만원 미만 6. 400만원 ~ 500만원 미만 7. 500만원 ~ 600만원 미만 8. 600만원 ~ 700만원 미만 9. 700만원 ~ 800만원 미만	1. 소득없음 2. 100만원 미만 3. 100만원 ~ 200만원 미만 4. 200만원 ~ 300만원 미만 5. 300만원 ~ 400만원 미만 6. 400만원 ~ 500만원 미만 7. 500만원 ~ 600만원 미만 8. 600만원 ~ 700만원 미만 9. 700만원 ~ 800만원 미만	– 소득없음 – 100만원 미만 – 100만원 ~ 200만원 미만 – 200만원 ~ 300만원 미만 – 300만원 ~ 400만원 미만 – 400만원 ~ 500만원 미만 – 500만원 이상

구분	국민여가활동조사 (기준 데이터)	문화향수실태조사 (연계 데이터)	공통 변수 표준화
	10. 800만원 ~ 900만원 미만 11. 900만원 ~1,000만원 미만 12. 1,000만원 이상	10. 800만원 ~ 900만원 미만 11. 900만원 ~1,000만원 미만 12. 1,000만원 이상	
월평균 가구 소득	1. 소득없음 2. 100만원 미만 3. 100만원 ~ 200만원 미만 4. 200만원 ~ 300만원 미만 5. 300만원 ~ 400만원 미만 6. 400만원 ~ 500만원 미만 7. 500만원 ~ 600만원 미만 8. 600만원 ~ 700만원 미만 9. 700만원 ~ 800만원 미만 10. 800만원 ~ 900만원 미만 11. 900만원 ~1,000만원 미만 12. 1,000만원 이상	1. 소득없음 2. 100만원 미만 3. 100만원 ~ 200만원 미만 4. 200만원 ~ 300만원 미만 5. 300만원 ~ 400만원 미만 6. 400만원 ~ 500만원 미만 7. 500만원 ~ 600만원 미만 8. 600만원 ~ 700만원 미만 9. 700만원 ~ 800만원 미만 10. 800만원 ~ 900만원 미만 11. 900만원 ~1,000만원 미만 12. 1,000만원 이상	- 100만원 미만 - 100만원 ~ 200만원 미만 - 200만원 ~ 300만원 미만 - 300만원 ~ 400만원 미만 - 400만원 ~ 500만원 미만 - 500만원 ~ 600만원 미만 - 600만원 ~ 700만원 미만 - 700만원 이상

다. 데이터 연계 사전 평가

일반적으로 조사 자료는 자료의 특성상 데이터 연계에 사용되는 공통 변수들이 대체적으로 인구통계학적 변수들이며, 이는 대부분 범주형 변수로 구성되어 있다. 따라서 데이터 연계에 사용되는 여러 가지 통계적 기법들 중 범주형 데이터를 이용한 데이터 연계 방법을 주로 활용하게 된다. 데이터 연계에 활용할 수 있는 방법이 다양하기 때문에 어떠한 통계적 연계 방법을 적용할지에 대한 평가가 필요하다. 통계적 연계 방법별 평가는 연계에 사용된 연계자료의 데이터 건수와 연계 전/후의 분포 비교를 통해 진행할 수 있다. 데이터 연계의 사전평가는 최적의 연계 방법을 선정하는 것도 있지만, 기준자료 관점에서 데이터를 연계할 때 활용이 가능한지에 대한 사전 평가의 의미를 가진다고 할 수 있다.

기준자료와 연계자료를 연계할 때, 연계자료의 케이스를 중복이 허용할지를 선택해야 하는데, 통계적 연계의 효율을 높이기 위해서는 중복을

허용하는 것이 대체적으로 더 좋은 결과물을 제공하기 때문에, 이 연구에서의 기준자료는 하나의 케이스가 한 번씩만 사용되지만, 연계자료는 하나의 케이스가 기준 자료의 여러 케이스와 유사하다면 중복으로 사용된다. 그리고 기준자료의 모든 케이스는 연계자료와 연계될 수 있지만, 유사성의 정도가 떨어지는 것을 억지로 연계할 경우에 연계의 효율이 떨어질 수 있어, 모든 케이스를 다 연계할지에 대한 부분은 데이터 연계를 진행하면서 지속적인 평가를 통해 결정한다.

데이터를 연계하는 다양한 유사성 척도들 중에 하나를 선정하고, 이를 다양한 연계 방법들 중 하나의 방법을 선정하여 가장 작은 값, 즉, 두 데이터 사이의 거리가 가장 짧은 데이터를 결합하는 것이다. 따라서 실제 연계를 진행하기에 앞서, 거리를 계산하기 위한 다양한 방법 중 어떤 척도와 방법을 적용할지에 대한 비교를 진행한다. 여기서 비교한 거리계산 방법은 고어 거리계산법(Gower's Distance)[37], 맨해튼 거리계산법(Manhattan Distance)[38], 유클리디안 거리계산법(Euclidean Distance)이다.

모의실험은 국민여가활동조사 데이터를 40:60으로 분할하여 활용하였다. 40에 해당하는 데이터는 기준자료, 60에 해당하는 데이터는 연계자료의 역할을 하도록 하였다. 이는 일반적으로 연계자료가 기준자료보다 더 커야 효율이 있기 때문에 연계자료를 더 크게 분할하였다. 공통변수로는 동거 가구원수, 성별, 연령, 혼인상태, 동거 자녀수, 종사상 지위(비경제활동자와 통합한 변수), 월평균 본인/가구 소득 변수들을 활용하였다. 연계자료에만 존재하는 유일변수는 여가시간과 여가비용 변수를 활용하였다.

고어 거리계산법과 맨해튼 거리계산법은 블로킹 변수를 설정한 경우와 설정하지 않은 두 가지 경우로 구분하여 연계를 진행하였다. 블로킹

37) 고어 거리계산법(Gower's Distance) : 범주형과 연속형 변수가 모두 존재하는 경우 사용한다.
38) 맨해튼 거리계산법(Manhattan Distance) : 범주형 변수에 대해 더미변수(dummy variable)를 생성하여 거리를 측정

변수는 성별을 사용하였다. 기준자료와 연계자료의 연계 이후 분포를 비교해보면, 성별의 블로킹 여부에 따라 분포가 달라지는 것으로 보이진 않았다. 하지만, 이는 데이터에 따라 달라질 수 있으며, 유사성 척도 계산할 때 시스템 부하가 커질 경우가 발생할 수도 있기 때문에 동일한 효율이라면 블로킹 변수를 활용하는 것이 효과적이다. 각 데이터 연계 방법별로 데이터 연계 전/후 분포의 동질성 검정 결과를 보면, 고어 거리계산법과 유클리디안 거리계산법은 거리계산에 사용한 7개 변수 중 1개의 변수에 대해서만 분포가 동일하지 않지만, 맨해튼 거리계산법의 경우 4개의 변수에 대한 분포가 동일하지 않게 나타났다.

연계된 통합 파일의 데이터 건수를 살펴보면, 분포에 대한 결과와 마찬가지로 성별에 대한 블로킹은 데이터 사용건수에 크게 영향을 미치지 않는 것으로 보인다. 하지만, 사용되는 데이터 건수를 보면, 고어 거리계산법보다는 맨해튼 거리계산법이 더 많은 데이터를 활용하는 것으로 나타났으며, 맨해튼 거리계산법보다도 성향점수에 의한 유클리디안 거리계산법이 더 많은 데이터를 사용하는 것으로 나타났다. 그러나 크게 차이가 나는 것은 아니기 때문에 데이터 연계 방법 선택을 하는 결정에 영향은 미비한 것으로 판단된다.

〈표 5-16〉 사전평가를 위한 5가지 연계 방법 및 결과

구분	연계 방법	연계 전후 분포 비교 (총 8개 변수)		데이터 연계	
		분포 동일한 변수 개수	동일 비율	사용한 데이터 개수	연계 사용률(%)
1	고어 거리 함수	7	87.5	2,732	43.0
2	랜덤 핫덱, 고어 거리 함수	7	87.5	2,725	42.8
3	맨하튼 거리 함수	4	50.0	2,877	42.2
4	랜덤 핫덱, 맨하튼 거리 함수	3	37.5	2,851	44.8
5	성향점수, 유클리디안 거리 함수	7	87.5	2,858	44.9

연속형 변수의 분포를 보면, 대부분 변수들의 평균이 기준 데이터와

큰 차이가 없으나, 월평균 여가비용 변수를 보면, 맨해튼 거리계산법에 의한 차이가 가장 큰 것으로 나타났다.

〈표 5-17〉 국민여가활동조사 연속형 변수의 분포 및 RMSE 비교

공통 변수		국민여가 활동조사 (기준 데이터)	연계 데이터				
			고어	고어 (블로킹 성별)	맨해튼	맨해튼 (블로킹 성별)	유클리디안 (블로킹 성별)
연령	평균	46	46	46	46	46	46
	표준편차	18	17	17	18	18	18
	최소값	15	15	15	15	15	15
	1사분위수	32	33	33	33	33	33
	중위수	46	47	47	47	47	46
	3사분위수	59	59	59	60	60	60
	최대값	96	96	96	96	96	90
동거 자녀수	평균	1	1	1	1	1	1
	표준편차	1	1	1	1	1	1
	최소값	0	0	0	0	0	0
	1사분위수	0	0	0	0	0	0
	중위수	0	0	0	0	0	0
	3사분위수	2	2	2	2	2	2
	최대값	7	5	5	5	5	7
월평균 여가비용	평균	139,330	135,264	133,750	130,453	132,202	138,420
	표준편차	140,104	127,941	127,400	130,927	127,486	136,941
	최소값	0	0	0	0	0	0
	1사분위수	50,000	50,000	50,000	50,000	50,000	50,000
	중위수	100,000	100,000	100,000	100,000	100,000	100,000
	3사분위수	200,000	200,000	180,000	150,000	200,000	200,000
	최대값	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000
하루평균 여가시간 (평일)	평균	3	3	3	3	3	3
	표준편차	2	2	2	2	2	2
	최소값	0	0	0	0	0	0
	1사분위수	2	2	2	2	2	2
	중위수	3	3	3	3	3	3
	3사분위수	4	4	4	4	4	4
	최대값	16	16	16	15	12	12
하루평균 여가시간 (휴일)	평균	5	5	5	5	5	5
	표준편차	2	2	2	2	2	2
	최소값	0	0	0	0	0	0
	1사분위수	3	3	3	3	3	3

공통 변수		국민여가활동조사(기준 데이터)	연계 데이터				
			고어	고어(블로킹 성별)	맨해튼	맨해튼(블로킹 성별)	유클리디안(블로킹 성별)
	중위수	5	5	5	5	5	5
	3사분위수	6	6	6	6	6	6
	최대값	16	16	16	16	16	16
RMSE	여가비용(금액)	–	3,882.1	4,276.7	5,364.7	5,814.4	3,073.8
	여가시간(평일)	–	0.047	0.045	0.035	0.033	0.026
	여가시간(휴일)	–	0.044	0.036	0.041	0.046	0.036

범주형 변수와 연속형 변수의 분포, RMSE(평균제곱근오차), 그리고 사용되는 데이터 개수를 종합적으로 비교해 봤을 때, 데이터 연계를 통한 분포의 변화가 거의 없는 것으로 나타나 데이터 연계는 효율적이라 판단할 수 있다. 또한 사전 평가를 진행한 데이터 연계 방법 중에서 고어 거리 계산법 또는 성향점수를 이용한 방법을 이용한 방법이 적절한 것으로 나타났으므로 실제 데이터 연계에 활용하도록 한다.

라. 데이터 연계

기준자료인 국민여가활동조사는 10,602건의 데이터를 가지고 있으나, 데이터 전처리 과정에서 공통 변수 중 종사상 지위(비경제활동자와 통합한 변수) 항목에 무응답인 건이 한 건 존재하였다. 따라서 이 건을 제외한 10,601건을 연계에 활용하였다. 연계자료인 문화향수실태조사는 10,716건의 데이터 전체에서 공통 변수에 대한 무응답인 건이 존재하지 않아 모든 데이터를 사용하였다.

데이터 연계는 성향점수[39]를 산출하고, 블로킹 후에 유클리디안 거리

39) 데이터 연계는 기준자료의 데이터를 기준으로 하지만, 연계를 통해 추가적으로 얻고자 하는 연계자료의 고유 변수에 관심이 있기 때문에 성향점수 산출 모형(Model)은 연계자료의 데이터를 토대로 구축한다. 따라서 연계자료(문화향수실태조사)의 주요변수인 유아/아동기와 청소년기의 문화예술교육 경험여부를 종속변수로 하는 로지스틱 회귀분석 모형을 성향점수 산출 모형으로 구축하였다.

를 계산하여 최단거리 데이터를 연계하는 방식과 공통 변수들 간의 거리를 고어 거리계산법으로 최단거리 데이터를 연계하는 방식 두 가지로 진행하였다. 블로킹 변수로는 성별과 월평균 가구 소득을 사용하여, 성별과 월평균 가구소득이 동일한 대상자들만 데이터 연계가 되도록 하였다. 그 결과, 성향점수의 유클리디안 거리 계산에 의한 연계보다는 고어 거리계산에 의한 연계가 공통 변수들 간의 분포 유지에 더 적당하였기 때문에 고어 거리계산에 의한 연계 데이터를 최종적으로 활용하였다.

고어 거리계산과 성향점수 산출에 사용된 공통 변수로는 거주지역, 연령, 동거가구원 수, 혼인상태(배우자 유무), 동거 자녀수, 학력, 월평균 개인 소득 변수를 이용하였다.

마. 데이터 연계 사후 평가

정확 연계는 동일한 사람이나 사업체 등 같은 대상에 대한 정보를 가진 데이터를 연결하여 분석하기 때문에, 데이터 작성 시점 또는 환경에서 고려할 사항에 대한 점만 파악하여 조정했다면, 데이터 연계 후 통합 파일에 대한 검증은 필요가 없다. 그러나 통계적 연계를 이용한 데이터 분석에서는 데이터 연계 과정에서 연계자료의 데이터 분포에 변동이 발생하기 때문에 연계 후에 데이터의 변질이 발생하게 된다. 따라서 통계적 연계 이후의 분포에 대한 평가가 필요하다.

연계 데이터 통합 파일에 대한 사후 평가는 기준자료와 연계자료의 공통 변수에 대한 분포비교와 연계 자료의 연계 전/후 고유 변수의 분포 비교를 통해 진행한다. 공통 변수의 분포를 비교함으로써 두 자료 간의 연계가 적절하게 진행되었는지 파악하고, 고유 변수의 분포를 비교함으로써 연계 전/후의 공통 변수와 고유 변수 간의 관계가 유지되는지 파악하기 위함이다.

먼저 공통 변수에 대한 기준자료와 연계자료의 분포를 비교해보면, 범주형 변수의 경우, 지역과 혼인상태, 학력, 그리고 월평균 본인 소득에

대한 분포가 유지되는 것을 알 수 있다.

〈표 5-18〉 기준자료와 연계자료의 범주형 공통 변수 분포 비교

변수	범주	기준자료 (국민여가 활동조사)	연계자료 (문화향수 실태조사)	χ^2 통계량 (유의확률)
지역	서울	1,257	1,384	25.278 (0.065)
	부산	751	746	
	대구	632	648	
	인천	675	698	
	대전	492	497	
	광주	501	493	
	울산	426	395	
	세종	175	122	
	경기	1,357	1,446	
	강원	462	421	
	충북	512	511	
	충남	590	566	
	전북	551	539	
	전남	532	495	
	경북	669	654	
	경남	719	709	
	제주	300	277	
혼인상태	배우자없음	3,770	3,741	0.173 (0.688)
	배우자있음	6,831	6,860	
학력	초졸 이하	1,248	1,203	1.463 (0.691)
	중졸 이하	1,471	1,442	
	고졸 이하	4,182	4,218	
	대졸 이상	3,700	3,738	
월평균 본인 소득	소득없음	3,597	3,670	5.241 0.513
	100만원 미만	1,022	941	
	100만원 ~ 200만원 미만	1,989	2,020	
	200만원 ~ 300만원 미만	1,742	1,763	
	300만원 ~ 400만원 미만	1,460	1,449	
	400만원 ~ 500만원 미만	419	397	
	500만원 이상	372	361	

연계 통합 자료의 연속형 변수의 분포는 평균과 표준편차, 그리고 중위수와 사분위수 등을 토대로 비교하였다. 연령과 동거가구원수, 동거자녀수에 대해 비교한 결과 연속형 변수의 분포에서도 큰 차이가 없는

것으로 나타났다.

〈표 5-19〉 기준자료와 연계자료의 연속형 공통 변수 분포 비교

변수		연계 통합 자료(10,601건)	
		기준자료(국민여가활동조사)	연계자료(문화향수실태조사)
연령	평균	46.23	45.90
	표준편차	17.649	17.403
	최소값	15	15
	1사분위수	32.00	32.00
	중위수	47.00	46.00
	3사분위수	59.00	59.00
	최대값	96	89
동거 가구원수	평균	3.06	3.05
	표준편차	1.178	1.121
	최소값	1	1
	1사분위수	2.00	2.00
	중위수	3.00	3.00
	3사분위수	4.00	4.00
	최대값	10	8
동거 자녀수	평균	0.79	0.78
	표준편차	0.965	0.925
	최소값	0	0
	1사분위수	0.00	0.00
	중위수	0.00	0.00
	3사분위수	2.00	2.00
	최대값	7	5

바. 분석결과

일반적으로 유년기 혹은 청소년기의 문화적 경험이 문화적 자본이 되어 사회적 문화여가 활동에 영향을 미친다고 알려져 있다. 본 분석을 통해 유년기의 문화예술교육 경험 여부에 따라 문화예술 분야의 여가활동에 영향을 미치는지 파악해보았다.

아래의 〈표 5-20〉은 유아/아동기의 문화예술교육 경험 여부에 따른 여가활동 참여에 차이가 있는지 파악하기 위한 카이제곱 검정 결과이다. 대부분의 여가활동 경험 여부는 유아/아동기의 문화예술교육 경험 여부

에 따라 동일한 분포를 가진다고 말할 수 없다. 즉, 이는 유아/아동기의 교육 경험 여부에 따라 여가활동 경험 여부에 차이가 있다고 말할 수 있다. 따라서 성인이 되기 전에 형성된 문화자본이 문화적 참여에 영향을 미치는 것을 알 수 있다. 단, 문화예술 관람 및 참여에 대한 다양한 여가활동 중 전통예술 활동은 유아/아동기의 문화예술교육 경험 여부에 따라 활동 여부에 차이가 있다는 통계학적 근거가 부족했으며, 악기연주/노래 또는 춤/무용에 참여하는 여가활동에 대해서도 마찬가지로 차이에 대한 정밀한 검증이 필요한 것으로 나타났다. 따라서 이 세 가지 여가활동에 대해서는 유아/아동기의 문화예술교육 경험 여부에 따른 차이가 있다고 말할 수 있는 근거가 부족하였다.

〈표 5-20〉 유아/아동기의 문화예술교육경험과 여가활동의 분포

여가활동			전체	유아/아동기 교육 경험		통계량 (Chi)	유의확률
				경험있음	경험없음		
문화예술 관람활동	전시회 관람	비활동	9,678	16.8	83.2	89.509	0.000
		활동	923	29.3	70.7		
	박물관 관람	비활동	9,600	17.0	83.0	50.883	0.000
		활동	1,001	26.1	73.9		
	음악연주회 관람	비활동	10,245	17.6	82.4	13.842	0.000
		활동	356	25.3	74.7		
	전통예술 공연 관람	비활동	10,220	18.1	81.9	12.580	0.000
		활동	381	11.0	89.0		
	연극공연 관람	비활동	9,756	17.0	83.0	63.504	0.000
		활동	845	27.9	72.1		
	무용공연 관람	비활동	10,444	17.7	82.3	9.871	0.002
		활동	157	27.4	72.6		
	영화 관람	비활동	4,394	8.6	91.4	438.152	0.000
		활동	6,207	24.4	75.6		
	연예공연 관람	비활동	9,791	17.1	82.9	51.753	0.000
		활동	810	27.2	72.8		
문화예술 참여활동	문학행사 참여	비활동	10,320	17.6	82.4	11.868	0.001
		활동	281	25.6	74.4		

여가활동			전체	유아/아동기 교육 경험		통계량 (Chi)	유의확률
				경험있음	경험없음		
	글짓기/독서토론	비활동	10,351	17.5	82.5	47.771	0.000
		활동	250	34.4	65.6		
	미술활동	비활동	10,245	17.6	82.4	10.878	0.001
		활동	356	24.4	75.6		
	악기연주/노래	비활동	9,932	17.7	82.3	3.738	0.059
		활동	669	20.6	79.4		
	전통예술	비활동	10,469	17.9	82.1	0.129	0.810
		활동	132	16.7	83.3		
	사진	비활동	9,105	16.4	83.6	90.863	0.000
		활동	1,496	26.6	73.4		
	춤/무용	비활동	10,517	17.8	82.2	4.009	0.057
		활동	84	26.2	73.8		

〈표 5-21〉은 청소년기의 문화예술교육 경험 여부에 따른 여가활동 참여에 차이가 있는지 파악하기 위한 카이제곱 검정 결과이다. 그 결과는 유아/아동기의 문화예술교육 경험 여부에 따른 여가활동 여부와 유사하게 나타났다. 대부분의 문화예술 관람/참여 여가활동에서 청소년기의 문화예술교육 경험이 영향을 미치는 것으로 나타났으나, 유아/아동기와 마찬가지로 악기연주/노래와 전통예술, 춤/무용에 참여하는 여가활동에 대해서는 분포의 차이가 나타나지 않았다. 또한 그 외에도 전통예술공연 관람에 대한 여가활동 역시 청소년기 문화예술교육경험에 따른 분포의 차이가 나타나지 않아 청소년기 문화예술교육이 해당 여가활동에 영향을 미친다고 말할 수 있는 근거가 부족하였다.

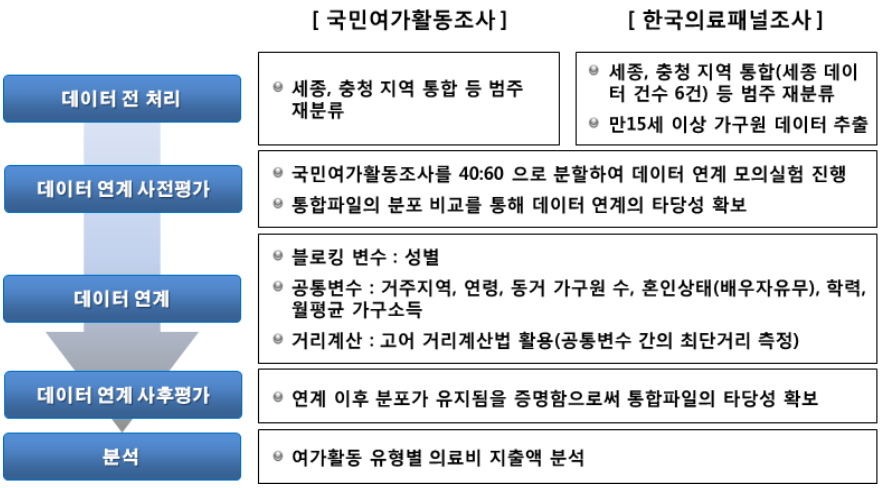
〈표 5-21〉 청소년의 문화예술교육경험과 여가활동의 분포

여가활동			전체	유아/아동기 교육 경험		통계량 (Chi)	유의확률
				경험있음	경험없음		
문화예술 관람활동	전시회 관람	비활동	9,678	17.2	82.8	79.460	0.000
		활동	923	29.0	71.0		
	박물관 관람	비활동	9,600	17.5	82.5	31.930	0.000

여가활동			전체	유아/아동기 교육 경험		통계량 (Chi)	유의확률
				경험있음	경험없음		
		활동	1,001	24.8	75.2		
	음악연주회 관람	비활동	10,245	17.9	82.1	16.583	0.000
		활동	356	26.4	73.6		
	전통예술 공연 관람	비활동	10,220	18.3	81.7	2.375	0.139
		활동	381	15.2	84.8		
	연극공연 관람	비활동	9,756	17.4	82.6	53.983	0.000
		활동	845	27.6	72.4		
	무용공연 관람	비활동	10,444	18.0	82.0	13.142	0.001
		활동	157	29.3	70.7		
	영화 관람	비활동	4,394	9.3	90.7	397.621	0.000
		활동	6,207	24.5	75.5		
	연예공연 관람	비활동	9,791	17.5	82.5	40.829	0.000
		활동	810	26.5	73.5		
문화예술 참여활동	문학행사 참여	비활동	10,320	18.0	82.0	16.353	0.000
		활동	281	27.4	72.6		
	글짓기/ 독서토론	비활동	10,351	17.9	82.1	25.517	0.000
		활동	250	30.4	69.6		
	미술활동	비활동	10,245	18.0	82.0	13.346	0.000
		활동	356	25.6	74.4		
	악기연주/ 노래	비활동	9,932	18.1	81.9	2.141	0.146
		활동	669	20.3	79.7		
	전통예술	비활동	10,469	18.2	81.8	0.477	0.565
		활동	132	15.9	84.1		
	사진	비활동	9,105	17.0	83.0	67.300	0.000
		활동	1,496	25.8	74.2		
	춤/무용	비활동	10,517	18.2	81.8	3.615	0.060
		활동	84	26.2	73.8		

2. 국민여가활동조사와 의료패널 연계를 통한 여가활동의 건강효과 분석

우리나라 국민의 여가활동과 의료비 지출의 관계를 파악하는 분석을 위해 국민여가활동조사와 한국의료패널조사의 데이터를 연계하였다. 데이터 연계를 통해 통합 데이터를 구축하고, 여가활동유형별 경험여부에 따른 의료비 지출액의 차이에 대해 분석하였다.



[그림 5-5] 국민여가활동조사와 한국의료패널조사 데이터 연계

가. 데이터 설명

1) 국민여가활동조사

국민여가활동조사의 내용은 앞서 설명하였기에 생략하도록 하며, 이번 통계적 데이터 연계 역시 국민여가활동조사의 데이터를 기준 파일로 한다.

2) 한국의료패널조사

한국의료패널조사는 한국보건사회연구원과 국민건강보험공단이 공동으로 주관하여 전국적으로 진행되는 국가승인통계 조사로, 보건의료

이용과 비용지출의 수준 및 배분을 추정하고, 의료전달시스템 및 보험체계의 동태적 변화에 대한 데이터베이스를 구축하여, 의료이용 및 의료비에 대한 실증자료를 바탕으로 한 보건의료정책의 수립·시행을 목적으로 하는 조사이다. 한국의료패널조사는 의료이용형태와 의료비 지출 규모에 관한 정보뿐 아니라 의료이용 및 의료비 지출에 영향을 미치는 요인들을 포괄적으로 분석할 수 있는 패널조사이다.

한국의료패널조사에는 인구통계학적 특성, 사회보험 및 연금, 민간보험, 건강 의식 및 행태, 보건 의료 접근성 및 만족도, 의료 이용 및 지출 등의 문항을 포함하고 있다. 본 연구에서는 한국의료패널 2008년~2015년 연간데이터(Version 1.4) 중에서 2015년 데이터를 활용하였으며, 사용한 주요 변수는 인구통계학적 특성 문항(성, 연령, 혼인상태, 경제활동상태 등), 개인/가구 의료비 등이다.

한국의료패널조사 연간데이터는 가구정보와 가구원정보, 응급서비스 이용정보, 입원서비스 이용정보, 외래서비스 이용정보, 가구원 경제활동정보 등이 각각의 데이터 파일로 구성되어 있다. 따라서 필요한 정보들로 구성된 하나의 데이터 파일로 구성하는 작업이 필요하다. 각각의 데이터 파일은 데이터의 성격에 따라 차수별 가구식별번호[40]와 가구원고유번호[41]를 고유키(Primary Key) 변수로 가지고 있으므로, 이 두 변수를 기준으로 데이터 파일을 통합하였다.

나. 데이터의 표준화

국민여가활동조사와 한국의료패널조사는 조사의 목적부터 다르기 때문에 조사기준과 인구통계학적 배경변수 등이 대부분 상이하였다. 따라

40) 차수별 가구식별번호 : 한국의료패널조사 1차년도(2008년)에 가구별로 부여된 가구고유번호(5자리)에 가구생성차수와 가구분리일련번호(원가구로부터 분가한 분가가구에 대한 일련번호)를 결합한 변수로, 분가정보를 포함하고 있는 가구식별번호이다. (2008~2015 한국의료패널 연간데이터 사용안내서)

41) 가구원고유번호 : 개인(가구원) 단위로 병합할 때 사용하는 변수로, 개인이 분가를 할 때에도 가구원고유번호는 그대로 유지된다.

서 기본적인 데이터 전처리 후에 연계를 위한 데이터 전처리를 다시 함으로써 2단계에 걸친 데이터 전처리 과정을 거쳤다.

항목별로 표준화 작업 내역을 살펴보면, 성별 변수는 두 데이터가 동일한 코드로 범주화되어 있으며, 동거 가구원 수도 동일한 의미의 수치형 변수이기 때문에 별도의 표준화 작업 없이 항목명(Column)만 일치시켰다. 연령의 경우, 국민여가활동조사는 연령을 직접 기입하도록 되어 있으나, 한국의료패널조사는 출생년도를 기재하게 되어 있어, 출생년도를 만 나이로 변환하여 사용하였다.

지역 변수의 경우, 지역코드가 서로 상이하여 국민여가활동조사 데이터를 기준으로 한국의료패널조사 코드값을 변경하여 표준화하였다. 이후 데이터의 빈도를 파악해봤을 때, 한국의료패널조사의 데이터에 세종시 거주자의 빈도가 6명에 불과하였다. 국민여가활동조사의 세종시 거주자는 175명으로 지역항목을 유일한 공통 변수로 활용할 경우, 175명의 데이터에 6명이 중복으로 연계되는 상황이 발생하게 된다. 이는 연계의 효율을 떨어뜨리는 요인이 될 수 있기 때문에, 충북, 충남, 세종을 충청도라는 하나의 코드값으로 통합하였다.

혼인상태 항목은 사실혼을 포함한 '혼인' 범주와 '별거' 범주를 '배우자 있음' 범주에 포함시켰으며, 한국의료패널조사 데이터의 코드값을 국민여가활동조사 범주에 맞춰 변환하였다. 이를 데이터 연계에 활용할 때에는 미혼과 사별, 이혼, 기타를 '배우자 없음'으로 통합하여 배우자 유무에 대한 코드값으로 변형하여 활용하였다.

한국의료패널조사 데이터의 학력 범주는 국민여가활동조사 데이터에 비해 세분화되어 있다. 따라서 한국의료패널조사의 학력을 국민여가활동조사의 학력 범주에 맞춰 동일하게 구분하였다. 하지만 한국의료패널조사의 학력 범주에서는 대학(4년제 미만)과 대학교(4년제 이상)의 구분이 불가하여, 국민여가활동조사의 대학(4년제 미만)과 대학교(4년제 이상) 범주를 통합하였다. 또한 연계에 활용하기 위해 학력과 이수 여부를 조합

하여 '초졸 이하', '중졸', '고졸', '대졸 이상' 으로 구분하였다. 그리고 최종 학력의 이수 여부가 재학, 수료, 휴학, 중퇴에 해당하는 경우는 이전 학력에, 졸업인 경우에만 해당 학력에 포함 시켰다. '초졸 이하' 에는 무학과 초등학교 재학 등이 포함되어 있으며, '대졸 이상'은 대학교와 석사, 박사과정을 포함시키고, 대학교는 이수여부가 졸업인 경우만을 포함하였다.

한국의료패널조사는 개인 및 가구의 연간근로소득에 대해 만원 단위의 실수치를 조사하였으나, 국민여가활동조사는 월평균 개인/가구 소득에 대해 100만원 단위별로 범주화 되어있다. 따라서 한국의료패널조사의 연간근로소득을 월평균 금액으로 환산한 후에 국민여가활동조사의 범주에 맞춰 범주화하였다. 또한 월평균 개인/가구 소득에 대한 고소득 범주의 경우 전체적인 비중이 낮아 고소득 일부 구간을 통합하였으며, 월평균 가구소득은 소득이 없는 구간과 100만원 미만인 구간을 통합하였다.

〈표 5-22〉 공통 변수 표준화 내역

구분	국민여가활동조사 (기준 데이터)	한국의료패널조사 (연계 데이터)	공통 변수 표준화
연령	만_____세	출생년도	10세 단위 나이
지역	11. 서울 21. 부산 22. 대구 23. 인천 24. 광주 25. 대전 26. 울산 29. 세종 31. 경기 32. 강원 33. 충북 34. 충남 35. 전북 36. 전남 37. 경북 38. 경남 39. 제주	11. 서울 26. 부산 27. 대구 28. 인천 29. 광주 30. 대전 31. 울산 36. 세종 41. 경기 42. 강원 43. 충북 44. 충남 45. 전북 46. 전남 47. 경북 48. 경남 50. 제주	 – 서울 – 부산 – 대구 – 인천 – 광주 – 대전 – 울산 – 경기 – 강원 – 전북 – 전남 – 경북 – 경남 – 제주 – 충청

구분	국민여가활동조사 (기준 데이터)	한국의료패널조사 (연계 데이터)	공통 변수 표준화
혼인상태	1. 미혼 2. 배우자 있음 3. 사별 4. 이혼 5. 기타	1. 혼인 중(사실혼 포함) 2. 별거(이혼 전제) 3. 사별/실종 4. 이혼 5. 미혼	– 배우자 없음 – 배우자 있음
학력	1. 무학 2. 초등학교 3. 중학교 4. 고등학교 5. 대학교(4년제 미만) 6. 대학교(4년제 이상) 7. 대학원 석사 과정 8. 대학원 박사 과정	01. 미취학 아동 02. 무학(해독 불가) 03. 무학(해독 가능) 11~16. 초등 1학년~ 6학년 21~23. 중등 1학년~3학년 31~33. 고등 1학년~3학년 41~46. 대학(교) 1학년~6학년 51. 대학원 석사 52. 대학원 박사	– 초등학교 졸업 이하 – 중학교 졸업 이하 – 고등학교 졸업 이하 – 대학 졸업 이상
이수 여부	1. 졸업 2. 재학 3. 수료 4. 휴학 5. 중퇴	1. 졸업 2. 재학 3. 휴학 4. 수료 5. 중퇴	
월평균 본인 소득	1. 소득없음 2. 100만원 미만 3. 100만원 ~ 200만원 미만 4. 200만원 ~ 300만원 미만 5. 300만원 ~ 400만원 미만 6. 400만원 ~ 500만원 미만 7. 500만원 ~ 600만원 미만 8. 600만원 ~ 700만원 미만 9. 700만원 ~ 800만원 미만 10. 800만원 ~ 900만원 미만 11. 900만원 ~ 1,000만원 미만 12. 1,000만원 이상	_____만원	– 소득없음 – 100만원 미만 – 100만원 ~ 200만원 미만 – 200만원 ~ 300만원 미만 – 300만원 ~ 400만원 미만 – 400만원 ~ 500만원 미만 – 500만원 이상
월평균 가구 소득	1. 소득없음 2. 100만원 미만 3. 100만원 ~ 200만원 미만 4. 200만원 ~ 300만원 미만 5. 300만원 ~ 400만원 미만 6. 400만원 ~ 500만원 미만 7. 500만원 ~ 600만원 미만 8. 600만원 ~ 700만원 미만	_____만원	– 100만원 미만 – 100만원 ~ 200만원 미만 – 200만원 ~ 300만원 미만 – 300만원 ~ 400만원 미만 – 400만원 ~ 500만원 미만 – 500만원 ~ 600만원 미만 – 600만원 ~ 700만원 미만 – 700만원 이상

구분	국민여가활동조사 (기준 데이터)	한국의료패널조사 (연계 데이터)	공통 변수 표준화
	9. 700만원 ~ 800만원 미만 10. 800만원 ~ 900만원 미만 11. 900만원 ~ 1,000만원 미만 12. 1,000만원 이상		

다. 데이터 연계 사전 평가

데이터 연계에 대한 사전 평가는 국민여가활동조사 데이터를 기준으로 진행하며, 이는 앞서 진행한 국민여가활동조사와 문화향수실태조사 데이터 연계 시의 사전 평가와 동일하다. 따라서 본 절에서는 따로 다루지 않기로 한다.

라. 데이터 연계

기준자료인 국민여가활동조사는 만 15세 이상 가구원 대상 응답자 10,602명 중에서 경제활동 관련 변수가 무응답인 1명을 제외한 10,601명을 대상자로 하였다. 한국의료패널조사는 조사 가구원 전체를 대상으로 하고 있어 전체 응답자 18,130명 중 만 15세 이상 가구원인 15,731명을 대상으로 데이터 연계를 진행하였으나, 개인지출 의료비[42]가 무응답인 건이 1건 존재하여 이를 제외한 15,730건을 최종 연계자료로 활용하였다.

한국의료패널조사 데이터는 가구식별번호와 가구원고유번호라는 고유키(PK, Primary Key)로 개인 식별이 가능하다. 하지만 국민여가활동조사는 개별 데이터를 식별할 수 있는 고유키 변수가 존재하지 않는다. 또한 고유키 변수가 존재한다 하더라도 주민등록번호나 사업자등록번호와 같이 두 데이터에서 공통으로 사용할 수 있는 개인 식별 고유키 변수가

42) 한국의료패널조사에는 개인지출 의료비 변수가 2개로, 하나는 응급 · 입원 · 외래의료비와 처방약값을 합한 것이고, 나머지 하나는 응급 · 입원 · 외래의료비와 처방약값에 관련 교통비와 입원 간병비까지 포함한 금액이다.

아니기 때문에 정확 연계의 기법을 활용한 데이터 연계는 불가능하다. 따라서 통계적 연계 기법을 활용하여 데이터를 연계하였다.

기준자료 국민여가활동조사와 연계자료 한국의료패널조사의 데이터 연계는 블로킹과 고어 거리계산법에 의한 공통 변수의 거리계산의 두 단계로 진행되었다. 블로킹 단계에서는 성별을 블로킹 변수로 하였다. 따라서 동일한 성별의 범주 내에서 기준자료와 연계자료에서의 공통 변수 간의 고어 거리계산법에 의한 거리가 최소인 데이터끼리 연계하였다. 고어 거리계산에 사용된 공통 변수는 거주지역, 연령, 동거가구원수, 혼인상태(배우자 유무), 학력, 월평균 가구 소득 변수이다.

블로킹과 거리계산의 두 단계에 걸친 데이터 연계를 통해 국민여가활동조사 데이터에 한국의료패널데이터를 결합한 하나의 통합 파일을 생성하였다. 생성된 통합 파일을 활용하여 참여한 여가활동 유형에 따른 의료비 지출에 차이가 있는지 분석하도록 한다.

마. 데이터 연계 사후 평가

연계 이후의 평가에 대해서는 공통 변수의 분포에 대한 평가 및 공통 변수와 연계자료 고유 변수의 분포에 대한 평가 두 가지 단계로 진행하였다. 공통 변수에 대한 기준자료와 연계자료의 분포를 비교해보면, 범주형 변수의 경우, 지역과 혼인상태, 학력, 그리고 월평균 가구 소득에 대한 분포가 유지되는 것을 알 수 있다.

〈표 5-23〉 기준자료와 연계자료의 범주형 공통 변수 분포 비교

변수	범주	기준자료 (국민여가 활동조사)	연계자료 (한국의료 패널조사)	χ^2통계량 (유의확률)
지역	서울	1,257	1,260	1.772 (1.000)
	부산	751	753	
	대구	632	636	
	인천	675	674	
	대전	492	478	
	광주	501	508	

변수	범주	기준자료 (국민여가 활동조사)	연계자료 (한국의료 패널조사)	χ^2통계량 (유의확률)
	울산	426	395	
	경기	1,357	1,379	
	강원	462	464	
	전북	551	547	
	전남	532	540	
	경북	669	670	
	경남	719	723	
	제주	300	294	
	세종,충남,충북	1,277	1,280	
혼인상태	배우자없음	3,770	3,764	0.007
	배우자있음	6,831	6,837	(0.931)
학력	초졸 이하	1,248	1,246	
	중졸 이하	1,471	1,469	0.013
	고졸 이하	4,182	4,190	(1.000)
	대졸 이상	3,700	3,696	
월평균 가구 소득	100만원 미만	861	854	
	100만원 ~ 200만원 미만	1,075	1,081	
	200만원 ~ 300만원 미만	1,593	1,599	
	300만원 ~ 400만원 미만	2,437	2,431	0.136
	400만원 ~ 500만원 미만	2,231	2,237	(1.000)
	500만원 ~ 600만원 미만	1,356	1,352	
	600만원 ~ 700만원 미만	583	577	
	700만원 이상	465	470	

연계 통합 자료의 연속형 변수의 분포는 평균과 표준편차, 그리고 중위수와 사분위수 등을 토대로 비교하였다. 연령과 동거가구원수에 대해 비교한 결과 연속형 변수의 분포 역시 큰 차이가 없는 것으로 나타났다.

〈표 5-24〉 기준자료와 연계자료의 연속형 공통 변수 분포 비교

변수		연계 통합 자료(10,601건)	
		기준자료(국민여가활동조사)	연계자료(한국의료패널조사)
연령	평균	46.23	46.10
	표준편차	17.649	17.531
	최소값	15	15
	1사분위수	32.00	33.00
	중위수	47.00	46.00

변수		연계 통합 자료(10,601건)	
		기준자료(국민여가활동조사)	연계자료(한국의료패널조사)
	3사분위수	59.00	59.00
	최대값	96	92
동거 가구원수	평균	3.06	3.14
	표준편차	1.178	1.112
	최소값	1	1
	1사분위수	2.00	2.00
	중위수	3.00	3.00
	3사분위수	4.00	4.00
	최대값	10	7

바. 분석결과

여가활동이 국민의 신체적, 정신적으로 긍정적인 영향을 미침으로써 긍정적인 사회경제적 영향을 미칠 것이라는 것이 일반적인 통론이다. 본 분석에서는 여가활동이 우리나라 국민의 의료비 지출에 미치는 영향을 파악하기 위하여 여가유형별 여가활동 경험 여부에 따른 의료비 지출액을 분석하였다.

국민여가활동조사의 여가활동 유형별로 한 번 이상 참여한 여가활동이 존재하면 해당 유형에 참여한 것으로 인지하였다. 분석은 여가활동 유형별 참여여부를 독립변수로 놓고 개인/가구 의료비 지출액을 종속변수로 하는 독립표본 t-검정을 통해 분석하였다.

〈표 5-25〉 여가활동유형별 활동 내역

여가활동유형	상세 여가활동 내역
문화예술관람활동	- 전시회 관람 (미술, 사진, 건축, 디자인 등) - 박물관, 음악연주회(클래식, 오페라) 관람 - 전통예술공연 관람 (국악, 민속놀이 등) - 연극공연 관람(뮤지컬 포함) - 무용공연, 영화, 연예공연(쇼, 콘서트, 마술 쇼 등) 관람
문화예술참여활동	- 문학행사참여, 글짓기/독서토론 - 미술활동 (그림, 서예, 조각, 디자인, 도예, 만화 등) - 악기연주/노래교실, 전통예술 배우기 (사물놀이, 줄타기 등) - 사진촬영 (디지털카메라 포함), 춤/무용 (발레, 한국무용, 현대무용 등

여가활동유형	상세 여가활동 내역
스포츠관람활동	– 스포츠 경기 직접관람– 경기장 방문관람 (축구, 야구, 농구, 배구 등) – 스포츠 경기 간접관람– TV, DMB를 통한관람 (축구, 야구, 농구, 배구 등) – 격투기 경기관람 – 온라인게임 경기 현장관람 (e–스포츠 경기 포함)
스포츠참여활동	– 농구, 배구, 야구, 축구, 족구, 테니스, 스쿼시, 당구 · 포켓볼, 볼링, 탁구, 골프 – 수영, 윈드서핑, 수상스키, 스킨스쿠버 다이빙, 래프팅, 요트, 스노보드, 스키 등 – 아이스스케이트, 아이스하키 등 – 헬스(보디빌딩)/에어로빅, 요가/필라테스/태보 – 배드민턴/줄넘기/맨손 · 스트레칭 체조/훌라후프, 육상/조깅/속보 – 격투기운동 (태권도, 유도, 합기도, 검도, 권투 등) – 댄스스포츠 (탱고, 왈츠, 자이브, 맘보, 폴카, 차차차 등) – 사이클링/산악자전거, 인라인스케이트, 승마, 암벽등반, 철인삼종경기, 서바이벌
관광활동	– 문화유적방문 (고궁, 절, 유적지 등) – 자연명승 및 풍경 관람, 삼림욕 – 국내캠핑, 해외여행, 소풍/야유회, 온천/해수욕, 유람선 타기 – 테마파크가기/놀이공원/동물원/식물원 가기, 지역축제 참가, 자동차 드라이브
취미 · 오락활동	– 수집활동(스크랩 포함), 생활공예(십자수, 비즈공예, D.I.Y, 꽃꽂이 등) – 요리하기/다도, 애완동물 돌보기, 노래방 가기, 인테리어(집, 자동차 등) – 등산, 낚시, 미니홈피/블로그 관리, 인터넷 검색/채팅/UCC 제작/SNS – 게임(인터넷, 닌텐도, PSP, PS3 등), 보드게임/퍼즐/큐브 맞추기 – 바둑/장기/체스, 겜블(경마, 경륜, 카지노, 카드놀이, 고스톱, 마작 등)/복권구입 – 쇼핑/외식, 음주, 독서/만화책(웹툰) 보기 – 미용(피부관리, 헤어관리, 네일 아트, 마사지, 성형 등) – 어학 · 기술 · 자격증 취득 공부 · 학원 등 이용
휴식활동	– 산책 및 걷기, 목욕/사우나/찜질방, 낮잠, TV시청(DMB/IPTV 포함) – 비디오(DVD) 시청, 라디오 청취, 음악 감상, 신문/잡지보기 – 아무것도 안 하기
사회 및 기타 활동	– 사회봉사활동, 종교활동, 클럽/나이트/디스코/카바레 가기 – 가족 및 친지방문, 잡담/통화하기/문자보내기, 계모임/동창회/사교(파티)모임 – 이성교제(데이트)/미팅/소개팅, 친구만남/동호회 모임 – 위에서 분류되지 않은 기타 여가 활동

문화예술관람활동의 참여여부별 의료비 지출액에 대한 독립표본 t–검정 결과를 보면, 참여여부에 따른 의료비 지출액에는 차이가 있다고 말할 수 있는 통계학적 근거가 부족하였다.

〈표 5-26〉 문화예술관람활동 여부에 따른 의료비 평균 비교

구분	전체	문화예술관람활동		t-검정 통계량	유의 확률
		비활동	활동		
개인의료비[43]	560,995	510,963	562,228	-0.652	0.515
개인의료비(교통비 포함)[44]	573,450	519,181	574,788	-0.691	0.489
가구의료비	1,545,975	1,554,557	1,545,764	0.063	0.950
가구의료비(교통비 포함)	1,578,908	1,577,238	1,578,949	-0.012	0.990

문화예술참여활동의 참여여부에 따른 의료비 지출액에는 개인 의료비뿐만 아니라 가구 의료비에도 차이가 존재하였다. 또한 평균을 비교해보면 문화예술참여활동에 참여한 경우 의료비 지출액이 더 적은 것을 알 수 있다. 따라서 〈표 5-27〉의 결과와 비교해보면, 문화예술관람활동보다는 문화예술참여활동이 사회경제적으로 긍정적인 영향을 미침을 알 수 있다.

〈표 5-27〉 문화예술참여활동 여부에 따른 의료비 평균 비교

구분	전체	문화예술참여활동		t-검정 통계량	유의 확률
		비활동	활동		
개인의료비	560,995	596,351	452,192	6.328	0.000
개인의료비(교통비 포함)	573,450	609,904	461,272	6.378	0.000
가구의료비	1,545,975	1,580,395	1,440,054	3.155	0.002
가구의료비(교통비 포함)	1,578,908	1,615,168	1,467,325	3.263	0.001

스포츠관람활동 여부에 따른 의료비 지출액 역시 개인 의료비와 가구 의료비의 차이가 존재하였다. 마찬가지로 평균을 비교해보면 스포츠관람활동에 참여했을 때의 의료비 지출액이 참여하지 않았을 때보다 더 적은 것을 알 수 있다.

43) 개인/가구 의료비 : 응급, 입원, 외래 의료비 및 처방약값을 포함한 금액
44) 개인/가구 의료비(교통비 포함) : 개인/가구 의료비에 응급(앰뷸런스), 입원, 외래 교통비와 입원간병비를 합산한 금액

〈표 5-28〉 스포츠관람활동 여부에 따른 의료비 평균 비교

구분	전체	스포츠관람활동		t-검정 통계량	유의 확률
		비활동	활동		
개인의료비	560,995	636,452	498,535	5.692	0.000
개인의료비(교통비 포함)	573,450	652,131	508,322	5.805	0.000
가구의료비	1,545,975	1,610,472	1,492,588	2.717	0.007
가구의료비(교통비 포함)	1,578,908	1,647,145	1,522,424	2.813	0.005

반면, 스포츠참여활동 여부에 따른 의료비 지출액은 개인의료비에는 차이가 존재하였으나, 가구의료비에는 차이가 있다고 말할 수 없어 가구의료비에는 영향이 없다고 할 수 있다.

〈표 5-29〉 스포츠참여활동 여부에 따른 의료비 평균 비교

구분	전체	스포츠참여활동		t-검정 통계량	유의 확률
		비활동	활동		
개인의료비	560,995	664,818	483,047	7.366	0.000
개인의료비(교통비 포함)	573,450	683,016	491,190	7.594	0.000
가구의료비	1,545,975	1,576,307	1,523,203	1.231	0.218
가구의료비(교통비 포함)	1,578,908	1,613,787	1,552,721	1.386	0.166

관광활동은 스포츠참여활동과 마찬가지로 활동 여부에 따라 개인의료비에는 차이가 존재하였으나, 가구의료비에는 차이가 없는 것으로 나타났다. 개인의료비는 사람이 경험하지 않은 사람보다 지출액이 더 적은 것으로 나타났다.

〈표 5-30〉 관광활동 여부에 따른 의료비 평균 비교

구분	전체	관광활동		t-검정 통계량	유의 확률
		비활동	활동		
개인의료비	560,995	639,268	548,849	2.706	0.007
개인의료비(교통비 포함)	573,450	658,382	560,272	2.844	0.004
가구의료비	1,545,975	1,505,837	1,552,204	-0.741	0.459
가구의료비(교통비 포함)	1,578,908	1,544,570	1,584,236	-0.620	0.535

취미·오락활동 역시 스포츠참여활동, 관광활동과 마찬가지로 활동 여

부에 따라 개인의료비의 차이가 존재하였으나, 가구의료비의 차이에 대해서는 통계학적 근거가 부족하다. 개인의료비 지출액 역시 취미·오락활동을 경험했을 때 지출액이 더 적은 것으로 나타났다.

〈표 5-31〉 취미·오락활동 여부에 따른 의료비 평균 비교

구분	전체	취미 · 오락활동		t-검정 통계량	유의 확률
		비활동	활동		
개인의료비	560,995	828,085	554,200	3.550	0.000
개인의료비(교통비 포함)	573,450	847,503	566,479	3.608	0.000
가구의료비	1,545,975	1,533,535	1,546,292	-0.093	0.926
가구의료비(교통비 포함)	1,578,908	1,576,268	1,578,975	-0.019	0.985

휴식활동은 활동 여부에 따른 의료비 지출액에 뚜렷한 차이가 나타나지 않았다.

〈표 5-32〉 휴식활동 여부에 따른 의료비 평균 비교

구분	전체	휴식활동		t-검정 통계량	유의 확률
		비활동	활동		
개인의료비	560,995	376,473	561,676	-0.930	0.352
개인의료비(교통비 포함)	573,450	381,327	574,160	-0.947	0.343
가구의료비	1,545,975	1,335,698	1,546,752	-0.599	0.549
가구의료비(교통비 포함)	1,578,908	1,410,356	1,579,530	-0.470	0.638

사회 및 기타 활동 역시 휴식활동과 마찬가지로 활동 참여 여부에 따른 가 부족한 것으로 나타났다.

〈표 5-33〉 사회 및 기타활동 여부에 따른 의료비 평균 비교

구분	전체	사회 및 기타 활동		t-검정 통계량	유의 확률
		비활동	활동		
개인의료비	560,995	510,963	562,228	-0.652	0.515
개인의료비(교통비 포함)	573,450	519,181	574,788	-0.691	0.489
가구의료비	1,545,975	1,554,557	1,545,764	0.063	0.950
가구의료비(교통비 포함)	1,578,908	1,577,238	1,578,949	-0.012	0.990

제4절

소결

제5장에서는 앞서 살펴보았던 데이터 연계 방법을 실제 데이터에 적용해서 살펴보았다. 정확 연계는 카드 데이터와 기상청 데이터 연계와 문화체육관광 분야 사업체 표본틀과 통계청 행정 데이터 연계하는 두 가지 경우를 살펴보았다. 카드 데이터와 기상청 데이터 연계는 날짜, 위도와 경도를 이용하여 정확 연계를 하였으며, 서울에 거주하는 사람이 서울이 아닌 지역에서 카드사용을 한 대상들을 분석하였다. 이때 서울의 기후가 나쁠수록 다른 지역에서 지출한 금액이 많은 것으로 나타났으며, 이동한 지역의 관광기후지수 등급에 따라 여행 관련 분야의 차이는 나타난 것을 제외하고는 이동한 지역의 날씨는 카드지출에 영향을 주지 않는 것으로 나타났다. 이러한 결과는 서울의 날씨가 나쁘면 다른 지역으로 이동해서 돈을 사용하는 것에 차이가 있으며, 이미 다른 지역으로 이동했을 경우 그 지역의 날씨는 지출에 영향을 주지 않는 것으로 해석할 수 있다.

문화체육관광 분야 사업체 표본틀과 통계청 행정 데이터 연계는 사업자등록번호를 식별정보로 이용하여 동일한 사업체를 연계함으로써 국가승인통계를 생산한다. 이때 사업체 식별정보인 사업체고유번호를 이용하기 때문에 데이터 이용 절차가 까다롭고, 이용하는 장소도 보안센터에서만 가능하다. 또한 통계 결과에 대한 반출도 심의를 거쳐 제공받게 된다.

정확 연계는 동일한 대상을 연계하기 때문에 매우 정확한 정보를 알 수 있지만, 민감한 정보들이 있기 때문에 이용하는데 제약 조건이 많다. 카드 데이터도 랜덤하게 일부만 추출하여 개인 식별정보를 제거한 후 제공받았으며, 카드 데이터 자체가 민감한 정보를 내포하고 있기 때문에 많은 검토와 개인정보에 대한 부분에 대한 심의를 받는데 오랜 시간이 걸렸다. 통계청의 행정 데이터와의 연계는 통계청의 행정 데이터가 민감

한 정보를 포함하고 있기 때문에 이용을 승인하는 과정과 결과를 반출하는 과정의 심의를 받고, 분석도 데이터보안센터에서 하게 된다. 이와 같이 정확 연계는 개인정보문제 때문에 활용의 제약이 크다. 그렇지만 활용만 할 수 있다면 그 어떤 정보보다도 자세한 정보를 이용할 수 있게 된다.

통계적 연계는 국민여가활동조사와 문화향수실태조사를 이용한 문화적 자본과 여가활동의 관계 분석과 국민여가활동조사와 의료패널 연계를 통한 여가활동의 건강효과 분석을 하는 과정으로 진행하였다. 결과를 살펴보면, 문화예술관람활동은 유아/아동기의 예술교육에 따라 모두 유의한 차이가 있는 것으로 나타났으며, 문화예술참여활동은 악기연주/노래와 전통예술은 유아/아동기의 예술교육에 따라 모두 유의한 차이가 없는 것으로 나타났으며, 그 외 다른 활동은 차이가 있는 것으로 나타났다. 여기서, 특이한 결과는 유아/아동기의 예술교육이 있는 경우가 모두 긍정적으로 활동하는 것으로 나타났으나, 전통예술 관람의 경우는 음의 영향이 있는 것으로 나타났다.

국민여가활동조사와 의료패널 연계를 통한 여가활동의 건강효과 분석 결과는 문화예술관람활동 여부에 따른 의료비 지출액 차이는 없는 것으로 나타났으며, 문화예술참여활동 여부에 따른 의료비 차이는 있는 것으로 나타나 문화예술활동을 하는 사람이 활동하지 않는 사람보다 의료비를 덜 쓰는 것으로 나타났다. 스포츠관람활동 여부에 따른 의료비 지출액 역시 활동하는 사람이 활동하지 않는 사람보다 의료비를 덜 쓰는 것으로 나타났다. 스포츠참여활동 여부에 따른 의료비 지출액은 개인의료비에는 차이가 존재하였으나, 가구의료비에는 차이가 없는 것으로 나타났다. 관광활동 또한 개인의료비에는 차이가 존재하였으나, 가구의료비에는 차이가 없는 것으로 나타났다. 취미·오락활동 역시 개인의료비에는 차이가 존재하였으나, 가구의료비에는 차이가 없는 것으로 나타났다. 그러나 휴식활동과 사회 및 기타 활동은 개인과 가구 의료비 모두 유의한 차이가 없는 것으로 나타났다.

통계적 연계 결과를 살펴보면, 통계적 연계를 통해 생성된 통합파일의 분석 결과는 의미 있게 나타난 것을 알 수 있다. 이는 통계적 연계는 정확 연계처럼 정확한 정보를 주지는 않더라도 유사한 분포를 유지하는 대상들을 연계하기 때문에 결과는 신뢰할 수 있다. 단지, 통계적 연계는 데이터 표준화 과정과 연계전의 평가와 연계 후의 평가 과정들을 거쳐야하기 때문에 복잡하다.

통계적 연계를 위해서는 공통 변수가 충분해야 하며, 두 데이터의 공통 변수가 같은 기준으로 구성되도록 표준화시켜야 한다. 즉, 표준화 과정이 중요한데 표준화가 잘못되면 많은 정보가 손실된다. 따라서 사전에 유사하게 공통 변수가 구성되어 있다면 데이터 연계는 쉽게 할 수 있을 것이다. 그리고 사전평가에서는 기준 파일이 데이터 연계로 활용이 가능한지와 어떤 방법이 통계적 연계에 적합한지를 평가한다. 사후평가는 통합 파일이 유사한 대상들이 결합되었는지, 즉 결합한 대상들의 분포가 유사하여 선택 편의가 발생하지 않는지를 평가한다.

제6장

데이터 연계 활용을 위한 고려사항

제1절

데이터 연계에서 주요 고려사항

데이터 연계 중 정확 연계는 쉬운 방법이지만, 개인정보의 문제로 활용이 어렵다. 제5장의 신한카드 데이터와 기상 데이터 연계는 공통 변수로 위도와 경도 그리고 날짜를 이용하기 때문에 개인정보를 활용하지는 않았으나, 신한카드 데이터 자체에 개인정보를 내포하고 있기 때문에 데이터를 활용하는데 한계가 있었다. '문화체육관광산업 통계'는 사업체고유번호를 이용하여 기업의 모든 정보를 연계할 수 있기 때문에 보안을 위하여 통계청의 빅데이터센터 내에서만 작업할 수 있으며, 통계 결과값만을 반출하여 활용할 수 있다는 한계가 있었다.

고유식별정보의 부재 시 유사 데이터를 연계하는 통계적 연계는 동일한 대상을 연결하는 것이 아니라 데이터의 분포가 유사한 대상들을 연결하기 때문에 절차와 방법이 모두 어렵고 복잡하다. 그러나 데이터 연계를 통해 효율적인 통합파일을 생성할 수 있다면 가치 있는 정보를 포함한 데이터를 생산하는 것과 동일하다.

제5장에서 문화·체육·관광 관련 자료를 이용한 데이터 연계를 적용하며 다양한 문제를 인식할 수 있다. 정확 연계는 개인정보 문제가 가장 먼저 제기되는 것을 알 수 있다. 다음으로 문화·체육·관광 분야의 데이터를 분리하는 것이다. 신한카드 데이터에서는 문화·체육·관광 관련 업종을 분리하고 그 업종에서의 지출을 찾아내 새로운 데이터셋을 만들었으며, '전국사업체기초자료'에서도 문화·체육·관광 분야의 데이터를 찾아야 하는 것이다. 이는 문화·체육·관광에 대한 명확한 분류 기준이 있어야 가능하다.

통계적 연계는 효과적인 연계를 위해 연계에 사용되는 공통 변수들이 다양하고 좋은 정보를 가지고 있어야 한다. 하지만 문화·체육·관광

분야의 대표성 있는 데이터 중에서 데이터 연계가 가능한 데이터는 조사통계 데이터밖에 없다. 조사 데이터의 공통 변수는 대부분 명목데이터이며, 항목의 명칭과 범주구분 기준이 제각각이기 때문에 정보의 손실이 발생하게 된다. 따라서 공통 변수에 대한 명확한 기준이 정립되어 있다면 충분한 정보를 활용할 수 있으므로 데이터 연계에 활용하기 좋을 것이다.

이밖에도 데이터 연계 시에 고려해야할 사항이 많겠지만, 데이터 연계 방법을 살펴보고 문화·체육·관광 분야의 데이터를 이용한 데이터 연계를 진행하여 분석한 결과를 볼 때, 주요한 고려사항으로 세 가지를 언급할 수 있다. 첫 번째는 개인정보 문제이고, 두 번째는 문화·체육·관광 분야의 데이터를 추출할 수 있는 분류 기준, 마지막으로 통계적 연계에서 활용할 조사통계의 인구통계학적 항목에 대한 범주를 사전에 표준화시키는 방안이다.

제2절

데이터 연계에서 개인정보보호

데이터를 활용할 때에는 항상 개인정보 문제를 생각해야 한다. 이는 개인의 사생활을 침범하게 될 수 있으며, 법적인 문제가 야기되기 때문이다. 특히 본 연구에서 다루고 있는 정확 연계는 고유식별 정보를 이용하기 때문에 개인정보 문제가 반드시 고려되어야 한다. 따라서 개인정보의 의미를 살펴보고, 데이터 연계에서 개인정보 문제를 해결할 수 있는 방안을 살펴보도록 한다.

1. 데이터 연계에서 개인정보보호의 개념

정보통신기술의 발전에 따라, 온라인 시스템 상에서 금융, 공공, 게임 등의 서비스를 활용하기 위해 개인정보를 제공하고 회원 가입하여 사용하고 있다. 하지만 빠르고 편리한 서비스를 이용하기 위해 제공한 개인정보는 정보보안의 부실, 개인에 의한 유출 또는 정보 활용 시에 일부 유출되기도 한다. 정보가 유출되고 악용되는 사례가 빈번하게 발생함에 따라 개인정보의 중요성은 점차 증대되고 있다.

개인정보에 대한 정의는 두 개의 법률상에 명확하게 정의되어 있는데, 개인정보보호법과 정보통신망 이용촉진 및 정보보호 등에 관한 법률이다. 이에 대한 정의는 〈표 6-1〉에 제시하였다.

〈표 6-1〉 법률상에 제시된 개인정보의 정의

법	정의
개인정보보호법	– "개인정보"란 살아 있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보(해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것을 포함한다)를 말한다.
정보통신망 이용촉진 및	– 생존하는 개인에 관한 정보로서 성명 · 주민등록번호 등에 의하여 특정한

법	정의
정보보호 등에 관한 법률	개인을 알아볼 수 있는 부호 · 문자 · 음성 · 음향 및 영상 등의 정보(해당 정보만으로는 특정 개인을 알아볼 수 없어도 다른 정보와 쉽게 결합하여 알아볼 수 있는 경우에는 그 정보를 포함한다)

즉, 개인정보는 생존하는 자연인에 관한 정보[45]이며, 개인과 관련된 사실적 정보(예: 이름, 주소, 주민등록번호, 직업 등)만이 아니라 개인에 대한 타인의 의견·평가·견해 등과 같은 주관적인 정보(예: 신용평가정보, 사회적 지위)도 관련성이 있으면 개인정보에 포함된다.

개인정보는 "특정 개인을 식별하거나 식별가능"해야 하므로, 이미 통계적으로 변환되어 개인을 식별할 수 있는 인자가 제거된 상태라면 개인정보가 아니다. 그러나 다른 정보와 결합해서 개인 식별이 가능한 경우에는 개인정보로 규정하고 있다(행정안전부, 2009). 예를 들어 주민등록번호는 개개인마다 고유하기 때문에 이를 활용하면 쉽게 개인을 식별할 수 있으므로 명확한 개인정보이다. 회사명과 거주지역은 개인정보가 아니지만 그 회사에 해당 지역에 살고 있는 사람이 한 명 밖에 없다면 개인의 식별이 가능하기 때문에 개인정보가 된다.

따라서 개인정보는 세 가지 형태로 구분할 수 있다. 첫 번째는 주민등록번호와 같은 고유식별 정보가 있는 경우이고, 두 번째는 특정기관에서 부여한 번호체계로 고유한 정보이지만 일반적인 사람들이나 시스템에서 식별이 불가능한 경우이며, 마지막으로 서로 다른 정보를 연결하여 개인 식별이 가능한 경우이다.

2. 데이터 연계를 위한 개인정보보호 방안

데이터를 이용할 때 항상 염두에 두어야 하는 것이 개인정보보호 문제이다. 특히 데이터 연계에서 개인정보에 대한 문제는 항상 제기될 수밖에

45) 법인의 상호, 영업소재지, 대표이사의 성명, 이사 · 감사 등 임원정보, 자산, 영업실적 등의 정보는 보호받는 개인정보의 범위에 해당된다고 볼 수 없다.

없다. 그 이유는 다수의 자료가 연계됨으로써 개인의 신상이 밝혀질 수 있으며, 사생활침해 또는 개인정보를 이용한 다양한 방법으로 손실이 발생할 수 있기 때문이다. 특히 정부나 공공기관 또는 금융기관에서 보유하고 있는 데이터를 활용할 때에는 신중을 기해야 한다.

데이터를 연계 또는 활용할 때에는 사전에 개인의 특성을 파악할 수 있는 정보를 제거하거나 개인정보보호기법 등을 적용하여 개인정보가 유출되지 않도록 조치를 취한 후 활용해야 한다. 개인정보보호는 반드시 지켜져야 하지만, 개인정보문제에 대한 너무 많이 고려는 데이터 활용에 제약이 많아지게 된다. 이러한 경우의 가장 큰 문제는 데이터를 활용하지 못하게 되는 것이다. 또 하나의 제약은 데이터에 비밀보호 조치를 취함으로써 일부 변수만 활용 가능하도록 하여, 실제 필요한 정보에는 접근할 수 없는 경우이다. 그리고 데이터를 이용하여 분석하였지만 그 결과를 활용하지 못하는 경우 등 다양한 제약이 있을 수 있다. 이와 같이 개인정보보호로 인해 데이터 활용에 제약을 두게 되면, 개인정보문제는 발생하지 않을 수 있지만, 데이터를 이용하여 중요한 정보는 파악할 수 없게 된다.

따라서 개인정보 노출위험과 자료의 활용가치의 적절한 합의점이 필요하다[46]. 데이터의 가치가 높다는 이유로 개인정보의 노출을 감수하고 데이터를 제공하는 것에 동의하는 것은 어려운 일이지만, 개인정보보호 때문에 데이터 연계를 할 수 없다면, 데이터 연계에 대한 시도조차 하지 못함으로 중요한 정보를 파악조차 못하게 되어, 매우 큰 가치의 손실을 볼 수 있다. 따라서 개인정보를 보호하면서도 데이터 연계를 활용할 수 있는 방법을 적용하는 방안이 필요하다.

데이터 연계에서 활용할 수 있는 개인정보보호 기법으로는 노출위험을 없애는 비식별화 기법이 있다. 비식별화 기법은 개인식별 정보를 이용하여 정확 연계를 할 때 개인식별 정보에 별도의 규칙을 정하여 기존의

46) 현재의 우리나라 정책에서는 아직 개인정보보호가 우선시 되고 있다.

식별 정보를 다르게 코딩함으로써 개인정보를 식별하지 못하게 수정하는 것이다. 이는 암호화와 같은 방법을 적용하여 가명으로 처리하는 것이 대표적인 방법으로 두 가지로 구분된다. 하나는 일정한 규칙을 정해 값을 수정하여 활용하고 필요할 때 원래의 값을 찾을 수 있도록 하는 방법이고, 다른 하나는 값을 다시 환원할 필요가 없을 경우에 규칙을 정하지 않고 임의로 값을 변경하는 방법으로 구분된다.

또한 여러 개의 값을 합하여 그룹으로 제시하는 방법, 개인정보를 가지고 있는 변수의 값을 일부 삭제하는 방법, 각각의 값들을 범주화로 처리하는 방법, 일부데이터를 마스킹하는 방법 등이 있으며, 이는 데이터 연계 이후 데이터를 제공할 때 활용하는 방법이다. 그룹으로 값을 제시하는 방법은 개별 데이터의 값을 원값으로 제공하는 것이 아닌 5명 또는 목록에 따라 합산한 값을 제공하는 방법이다. 이 방법은 정보의 전체적인 관점에서 사용하는 것은 문제없지만, 세부분석을 하는 데는 활용하기 어려운 단점이 있다.

데이터를 삭제하는 방법은 대부분 개인의 식별과 직접적인 관련이 있는 경우에 사용한다. 예를 들어 주민등록번호가 95XXXX-1XXXXXX일 때, 95년생 남자로 처리할 수 있다. 데이터 범주화는 데이터를 범주 변수로 만드는 방법인데, 실제로 데이터를 처리할 때 많이 사용하는 방법이다. 데이터 마스킹은 값을 변형하는 대신에 값의 일부를 파악할 수 없는 기호로 처리하는 방법으로, '임꺽정'의 경우 '임OO'으로 처리할 수 있다.

제3절

문화·체육·관광 데이터 분류

문화·체육·관광 분야의 데이터는 문화체육관광부와 소속기관 또는 산하기관 등 관련기관에서 생산하는 것만 존재하는 것은 아니다. 우리나라 전체를 대상으로 생활이나 근로, 매출 등의 내용을 생산 또는 기록하고 있는 통계청, 노동부 그리고 국세청 등의 기관에서도 관련 내용을 포함하고 있다. 이러한 타 국가기관 등에서 생산되는 데이터는 전반적인 분야의 데이터를 다루고 있기 때문에 데이터를 활용하기 위해서는 문화·체육·관광 관련 분야의 데이터를 추출해야 한다. 통계청의 '생활시간조사'나 '가계동향조사'의 경우 문화·체육·관광 관련 문항의 내용을 별도로 추출하면 문화·체육·관광 관련 데이터가 되며, '사업체기초통계조사'의 경우 전체 산업 중에서 문화·체육·관광 분야의 산업에 해당하는 사업체만을 분리하면 문화·체육·관광 산업에 종사하는 사업체 자료가 된다.

타 기관의 데이터를 활용할 경우 생기는 이점 중의 하나가 정확 연계 또는 정확한 대상만을 연계하여 분석할 수 있다는 점이다[47]. 통계청에서 실시하는 조사는 인구총조사 데이터를 기반으로 조사구를 선정하기 때문에 인구총조사 데이터와 연계[48]할 수 있다. 데이터 연계는 일차적으로 정확 연계를 하고 정확 연계가 이루어지지 않는 데이터에 대해서는 주소나 전화번호 등의 정보를 이용한 확률적 연계를 하여 대부분 동일한 대상이 연계되기 때문에 정확한 정보를 추가적으로 확보할 수 있다. 특히, '사업체기초통계조사'의 경우는 사업자등록번호라는 고유식별 변수가 존재하기 때문에 통계청, 국세청 등의 데이터와 연계함으로써 사업체에 대한 모든 정보를 활용할 수 있다. 이 경우, 보안 관련 문제는 더욱 강하게

47) 이는 개인정보문제가 발생하기 때문에 실제 활용하기 위해서는 통계청 등의 관련 기관의 승인을 득해야 하며, 이용할 때에도 보안에 대한 방안이 마련되어야 가능하다.
48) 이사를 가거나 조사를 거부할 경우 등의 일부 데이터는 연계되지 않는다.

제기되며, 방안도 반드시 마련해야 한다. 보안상의 문제로 인해 정확 연계를 하지 못하고 통계적 연계를 적용하는 경우에도 확대된 정보를 활용하는 관점에서는 이점이 있다고 할 수 있다.

데이터를 대상별로 구분하면, 크게 국민 대상의 개인 데이터와 기업 또는 사업체 데이터로 구분할 수 있다[49]. 또한 공급과 소비측면에서 구분하자면, 공급측면의 경우 산업의 관점이고, 소비측면의 경우 활동 또는 업종의 관점에서 접근할 수 있다. 문화·체육·관광 관련분야 여부를 결정하는 것은 개인을 구분하는 것이 아닌, 소비 또는 공급, 행위 등에 의해 구분하며 일반적으로 활동과 업종(산업)에 의해 구분된다. 특히 업종(산업) 관련 분류가 가장 많이 활용된다.

산업분류는 통계청의 표준산업분류(Korean Standard Industrial Classification, KSIC)가 기준이 되는데, 표준산업분류는 「통계법」에 의거해 통계 자료의 정확성 및 국가 간의 비교성을 확보하기 위하여, 유엔에서 분류기준을 정하여 권고하고 있는 국제표준산업분류(International Standard Industrial Classification, ISIC)를 기초로 작성한 통계분류이다(박근화 외 4인, 2017). 그러나 문화·체육·관광 분야는 일반적인 산업 영역과 차별화된 특성을 지니고 있기 때문에, 표준산업분류를 활용하기 어렵다. 따라서 문화·체육·관광 산업의 분류를 재구성하는, 특수산업분류를 이용해야 한다.

문화체육관광산업은 'R 예술, 스포츠 및 여가관련 서비스업'을 비롯하여 한국표준산업분류상 여러 업종에 산재되어 있어, 한국표준산업분류만으로는 문화체육관광산업 현황을 파악하기 어렵다. 예를 들어 표준산업분류에서는 크게 'R 예술, 스포츠 및 여가관련 서비스업'에는 창작, 예술 및 여가관련 서비스업과 스포츠 및 오락관련 서비스업이 포함된다. 그런데 실제 한국표준산업분류를 살펴보면, 문화체육관광산업에 포함되는

49) 예술인, 장애인, 외국인 등의 특정 대상 입장에서의 통계도 있지만, 이러한 통계는 각각의 대상에서는 일부이기 때문에, 여기서는 개인과 사업체만을 고려하도록 한다.

‘신문 배달업’은 한국표준산업분류상 ‘계약 배달 판매업’, ‘게임기 도매업’은 ‘장난감·취미용품 도매업’으로 분류되어 있다(박근화 외 2인, 2017).

이에 문화체육관광부는 별도의 문화체육관광 분야 산업에 대한 분류체계를 생성하고, 한국표준산업분류와 연계표를 만들어 활용하고 있다. 〈표 6-2〉에 문화체육관광 분야 산업분류 구성을 제시하였다.

〈표 6-2〉 문화·체육·관광 분야 산업분류 구성

(단위 : 개)

문화 · 체육 · 관광 분야 산업분류		대분류	중분류	소분류	세분류
산업 특수분류	저작권산업특수분류	4	12	56	308
	콘텐츠산업특수분류	12	51	139	–
	스포츠산업특수분류	3	8	20	65
	관광산업특수분류	4	22	57	106
자체 산업분류	광고산업분류	4	10	21	54
	문화예술산업분류	4	13	37	103

출처 :박근화 외 2인(2017). 2017년도 문화체육관광 분야 사업체 표본틀 구축 보고서

이렇게 문화체육관광 분야의 산업분류를 한 이후에는 다른 분류도 연결하여 분류체계를 마련할 수 있다. 관세청의 HS 코드(국제통일상품분류체계 ; Harmonized Commodity Description and Coding System)에서 문화체육관광 분야의 상품코드를 이용하여 별도의 통계를 산출할 수 있으며, 이는 ‘문화체육관광산업 통계’에서도 활용하고 있다.

다양한 연구에서 분류체계를 필요로 한다. 관광위성계정이나 문화체육관광 분야의 GDP 기여도 등에 대한 연구 시에 산업연관표를 이용하는데, 표준산업분류체계와 산업연관표의 기본부문을 연계하여 산출한다. 〈표 6-3〉은 산업연관표상에서 관광산업 분류를 일부 발췌한 것을 제시한 것이다.

〈표 6-3〉 산업연관표상에서의 관광산업 분류

구분	관광산업분류명	산업연관표(기본부문 384)		
		코드	분류	부가가치율
관광쇼핑업	면세점	303	소매서비스	53.571
	외국인전용 관광기념품 판매업			
	관광 인증 쇼핑업			
국제회의 및 전시업	국제회의 기획업	359	기타 사업지원서비스	58.565
	국제회의 시설업			

출처 : 박근화 외 4인(2017), 문화체육관광산업의 규모 추정 방안 연구에서 일부 발췌

데이터를 연계하는 이유는 개별적으로 파악할 수 없던 정보를 데이터 연계를 통해 다양한 관계 등을 파악하여 새로운 정보를 제시하기 위함이다. 따라서 다양하게 데이터가 많이 있으면, 데이터 연계는 활용의 폭이 더욱 넓어지게 된다. 현재 문화·체육·관광 관련 대표성 있는 통계는 25종에 불과하고, 데이터 연계에 활용할 수 있는 통계는 조사통계에 국한되어 있기 때문에, 다른 분야의 데이터를 분류하여 활용할 수 있는 방안을 마련할 필요가 있다. 문화·체육·관광 분야의 산업분류와 정확 연계를 이용하여 생산하고 있는 통계가 제5장에서 제시한 '문화체육관광산업통계'이다.

향후에는 문화체육관광산업의 분류체계를 다양한 분류체계와 연계하고, 이를 이용하여 다른 분야의 통계에서 문화·체육·관광 관련 자료를 추출하여 활용하는 방안도 고려할 필요가 있을 것이다.

제4절

데이터 연계를 위한 공통 변수

본 연구에서는 데이터 연계에 대하여 문화·체육·관광 관련 데이터를 기준으로 다른 데이터를 연계하는 과정을 살펴보고, 연계한 통합파일을 분석하여 기존에 알 수 없었던 새로운 정보를 파악하는 방안을 제시하였다. 데이터 연계에서는 무엇보다도 연계가 잘 되는 것이 중요하다. 정확 연계에서는 동일한 대상을 정확하게 연계할 수 있는 고유식별 정보가 있으면 되지만, 통계적 연계에서는 다양한 공통 변수들 중에서 연계가 잘 이루어질 수 있는 조합을 찾고 유사성 척도의 연계 방법을 통해 최적의 통합파일을 찾는 것이 중요하다.

통계적 연계에서 공통 변수를 이용한 연계에 대한 고민은 반드시 필요하기 때문에 본 제4절에서는 통계적 연계로 한정하여 살펴보도록 한다. 데이터 연계를 위해서는 표준화가 전제 되어야 하며, 동일 모집단에서 생성된 대상(개인, 가구 또는 사업체 등)에 대한 데이터라면 공통 변수가 데이터 연계에 적합해야 한다.

문화·체육·관광 분야에서 데이터 연계에 활용할 수 있는 데이터는 대부분 조사통계 데이터이기 때문에 동일한 모집단에서 대상의 기준이 동일한 데이터인지 확인해야 한다. 또한 공통문항(보통 인구통계학적 변수)의 정보가 충분한지, 범주의 범위가 일치하지 않아 활용을 위해서 변환 등의 표준화가 필요한지 살펴보아야 한다.

통계적 연계에서는 기본적으로 표준화 과정이 필요하다. 각각의 조사에서는 개별적으로 필요한 통계를 생산하기 때문에 질문과 세부 항목이 서로 다를 수밖에 없다. 그러나 대상이 동일한 조사의 경우 질문과 세부 항목을 동일하게 구성한다면 표준화 과정에서 손실되는 정보의 양이 적을 것이다.

본 절에서는 문화체육관광부의 국가승인통계 중 국민을 대상으로 하는 조사에 대해 살펴보고 공통항목으로 사용될 인구통계학적 변수들에 대한 개선방안을 제시하고자 한다. 국가승인통계의 국민을 대상으로 하는 통계는 대부분 인구총조사 자료의 조사구를 이용하여 표본설계를 하고 추출된 가구를 찾아가서 응답을 받게 된다. 즉, 인구총조사의 조사구를 표본틀로 이용한 통계이다. 이렇게 문화체육관광부에서 생산하는 국민대상의 국가승인통계는 6종으로, '문화향수실태조사', '국민여가활동조사', '국민독서실태조사', '국민여행실태조사', '국민생활체육실태조사', '국민체력실태조사'가 있다. 이 중 '국민체력실태조사'는 조사방식과 조사대상 선정에 차이가 있어 이를 제외한 5종을 대상으로 공통문항의 표준화 방안을 제시하도록 한다.

1. 기준 파일의 공통 변수

문화·체육·관광 관련 데이터를 기준 파일로 하는 데이터 연계를 할 경우, 공통 변수는 설문지의 응답자 특성을 묻는 질문이 된다. 데이터를 연계할 때 응답자 특성변수를 사용하여 유사성 거리를 계산하는 주된 이유는 응답자 특성이 유사한 대상을 연계함으로써 유사한 성향의 대상을 연계할 수 있기 때문이기도 하지만, 통계학적인 관점에서는 이러한 특성이 조사하기 전에 이미 결정되어져 있어 조사 또는 실험의 영향을 받지 않는 변수이기 때문에[50] 이를 사용한다.

본 연구에서 다루는 기준 파일은 표본설계를 통해 표본을 랜덤하게 추출하기 때문에 기준 파일의 조사대상은 대표성이 있다. 또한 연계 파일의 데이터가 연계를 통해 유사한 대상의 데이터가 붙는다면 선택편의가 없는 데이터가 될 수 있다. 따라서 응답자 특성별 비교 또는 응답자 특성을 고려한 분석결과는 신뢰할 수 있다. 선택편의가 없는 데이터 연계를

50) 실험 또는 연구자에 의해 변형될 수 없는 정보를 가진 변수를 공변수(covariable)라고 하며, 이러한 변수들의 분포를 맞추면, 선택편의(selection bias)가 없어지는 효과가 생긴다.

위해서는 공통 변수가 데이터 연계에 활용하기에 적합한 정도의 정보를 포함하고 있어야 한다. 따라서 각 통계의 생산기준과 응답자 특성변수를 검토하고 차이를 살펴보도록 한다.

가. 조사의 기준

국민 대상의 조사 5종에 대한 작성주기와 대상 기간은 〈표 6-4〉에 제시한 것처럼 조금씩 차이가 있다. 작성 주기는 각각 2년과 1년으로 되어 있어 다르며, 대상기간 역시 통계별로 차이가 있기 때문에 데이터 연계 등의 데이터 활용 시에 주의해야 한다.

가. 조사의 기준

국민 대상의 조사 5종에 대한 작성주기와 대상 기간은 〈표 6-4〉에 제시한 것처럼 조금씩 차이가 있다. 작성 주기는 각각 2년과 1년으로 되어 있어 다르며, 대상기간 역시 통계별로 차이가 있기 때문에 데이터 연계 등의 데이터 활용 시에 주의해야 한다.

〈표 6-4〉 국민대상 통계의 작성주기, 대상기간

통계명	작성 주기	대상 기간
문화향수실태조사	2년	전년도 8. 1 ~ 해당년도 7. 31
국민여가활동조사	2년	전년도 8. 1 ~ 해당년도 7. 31
국민독서실태조사	1년	전년도 10. 1 ~ 해당년도 9. 30
국민여행실태조사	1년	해당년도 (1. 1 ~ 12. 31)
국민생활체육실태조사	1년	전년도 8. 24 ~ 해당년도 8. 23

문화체육관광 분야의 중요성이 증대됨에 따라, 향후 통계 작성 주기는 1년으로 작성해야 할 것이며, 대상 기간도 응답자들이 대상기간이 전년과 조사년도가 중복됨에 따른 혼돈을 피하기 위해서는 1월 1일부터 12월 31일까지로 하는 것이 적절할 것이다. 그러나 현재는 〈표 6-4〉와 같으므로 같은 기준년도의 데이터끼리 연계를 하고 대상 기간 역시 완전하게

동일하진 않더라도 가급적 비슷한 대상 기간으로 데이터를 변환해서 사용할 수밖에 없다. 물론 통계의 활용에 따라 대상 기간이 동일하지 않아도 문제되지 않는다면, 원래의 데이터를 그대로 연계하면 된다. 단, 기준년도가 다를 경우, 최소한의 정보[51]는 조정하여 사용해야 한다.

나. 응답자 특성의 차이

문화·체육·관광 관련 데이터를 연계할 때, 다른 분야 데이터와의 연계도 필요하지만 우선적으로 문화·체육·관광 분야의 데이터끼리 연계하여 사용하는 방안을 마련할 필요가 있다. 이는 문화·체육·관광 데이터의 정보 부족 문제에 대해 데이터 연계를 통해 극복하고자 하는 것이기 때문에, 통계적 연계에서 사용할 응답자 특성 문항을 분석할 필요가 있다.

국민을 대상으로 하는 5종의 통계조사의 응답자 특성 문항을 각각 어떻게 조사되고 있는지 변수별로 정리를 하는데, 먼저 성별을 살펴보면 5종 모두 응답자를 남자와 여자로 분류하고 있다. 다음으로 연령을 살펴보면 연령은 모두 주관식으로 응답을 받고 있으나, 분석 및 공표 시에는 대부분 10세 간격을 유지하고 있다. 다음의 〈표 6-5〉는 연령에 대해 각 통계에서 구분한 내용을 제시한 것이다. 국민독서실태조사는 학생을 별도로 조사하기 때문에 성인대상의 통계는 만18세 이상을 대상으로 하고 있으며, 국민생활체육참여실태조사는 만10세 이상을 대상으로 한다. 그 외 나머지 통계에서는 만15세 이상을 대상으로 하였다.

〈표 6-5〉 국민대상 통계의 연령 분류

통계명	연령 분류
문화향수실태조사	만15~19세, 20대, 30대, 40대, 50대, 60대, 70대 이상
국민여가활동조사	만15~19세, 20대, 30대, 40대, 50대, 60대, 70대 이상
국민독서실태조사	만18~29세, 30대, 40대, 50대, 60대 이상
국민여행실태조사	만15~19세, 20대, 30대, 40대, 50대, 60대 이상
국민생활체육실태조사	10대, 20대, 30대, 40대, 50대, 60대, 70대 이상

51) 예를 들어 연령을 기준년도에 맞춰 조정하는 등이 해당한다.

학력의 분류현황은 〈표 6-6〉에 제시되어 있다. 4종의 통계에서 최종학력과 이수현황을 별도로 질문하는 형태로 구성되어 있으나, 국민여행실태조사만 재학과 졸업으로 구분하고 무학에 대한 부분을 초졸이하에 포함시켰으며 대학원 학력에 대해서 석사와 박사를 구분하지 않았다.

〈표 6-6〉 국민대상 통계의 학력 분류

통계명	최종학력 분류	이수여부
문화향수실태조사	무학, 초등학교, 중학교, 고등학교, 대학교(4년제 미만), 대학교(4년제 이상), 대학원 석사과정, 대학원 박사과정	졸업, 재학, 수료, 휴학, 중퇴
국민여가활동조사	무학, 초등학교, 중학교, 고등학교, 대학교(4년제 미만), 대학교(4년제 이상), 대학원 석사과정, 대학원 박사과정	
국민독서실태조사	무학, 초등학교, 중학교, 고등학교, 대학교(4년제 미만), 대학교(4년제 이상), 대학원 석사과정, 대학원 박사과정	
국민여행실태조사	초졸 이하, 중학교(재/졸), 고등학교(재/졸), 전문대(재/졸), 대학교(재/졸), 대학원(재/졸)	
국민생활체육실태조사	무학, 초등학교, 중학교, 고등학교, 대학교(4년제 미만), 대학교(4년제 이상), 대학원 석사과정, 대학원 박사과정	졸업, 재학, 수료, 휴학, 중퇴

혼인상태 분류현황은 〈표 6-7〉에 제시하였다. 혼인상태에 대해 국민독서실태조사와 국민여행실태조사는 문항이 존재하지 않으며, 나머지 3종에서는 동일하게 질문하고 있다

〈표 6-7〉 국민대상 통계의 혼인상태 분류

통계명	혼인상태 분류
문화향수실태조사	미혼, 배우자 있음, 사별, 이혼, 기타( )
국민여가활동조사	미혼, 배우자 있음, 사별, 이혼, 기타( )
국민독서실태조사	–
국민여행실태조사	–
국민생활체육실태조사	기혼, 미혼, 사별, 이혼, 기타

직업분류는 국민여행실태조사를 외에는 대부분 한국표준직업분류를 기준으로 분류하고 있으며, 군인과 자영업 종사자에 대한 문항만 일부 차이가 존재하였다.

〈표 6-8〉 국민대상 통계의 직업 분류

통계명	직업 분류
문화향수실태조사	전업주부, 학생, 기타, 관리자, 전문가 및 관련 종사자, 사무 종사자, 서비스 종사자, 판매 종사자, 농림어업 숙련 종사자, 기능원 및 관련 기능 종사자, 장치 · 기계조작 및 조립 종사자, 단순 노무 종사자, 군인
국민여가활동조사	전업주부, 학생, 기타, 관리자, 전문가 및 관련 종사자, 사무 종사자, 서비스 종사자, 판매 종사자, 농림어업 숙련 종사자, 기능원 및 관련 기능 종사자, 장치 · 기계조작 및 조립 종사자, 단순 노무 종사자, 군인
국민독서실태조사	관리자, 전문가 및 관련 종사자, 사무 종사자, 서비스 종사자, 판매 종사자, 농림어업 숙련 종사자, 기능원 및 관련 기능 종사자, 장치·기계조작·조립직 종사자, 단순 노무 종사자, 군인, 자영업 종사자, 학생, 전업 주부, 은퇴, 무직, 기타( )
국민여행실태조사	전문 · 관리, 사무, 서비스 · 판매, 농어업, 기능 · 노무
국민생활체육실태조사	관리자, 전문가 및 관련 종사자, 사무 종사자, 서비스 종사자, 판매 종사자, 농림어업 숙련 종사자, 기능원 및 관련 기능 종사자, 장치 · 기계조작 및 조립 종사자, 단순 노무 종사자, 전업주부, 학생, 무직

소득분류는 개인소득과 가구소득으로 구분되며, 월 평균 소득을 묻고 있다. 개방형 질문으로 정확한 소득에 대한 응답을 받는다면 좋겠지만, 응답자들이 소득에 대한 응답하는 것에 대한 부담을 가지기 때문에 대부분의 조사에서 구간으로 분류하여 응답을 받고 있다. 이에 대해 〈표 6-9〉에 제시하였다. 구간을 살펴보면 문화향수실태조사와 국민여행실태조사는 소득없음을 포함하고 있으며 100만원 단위로 1,000만원까지 구간을 구분하고 있다. 국민독서실태조사는 100만원 단위로 구분하며, 700만원까지 구분하고 있다. 국민여행실태조사는 100만원 미만에 대해서는 50만원 단위로 구분하였으며, 그 이상에 대해서는 600만원까지 100만원 단위로 구분하였다. 국민생활체육실태조사는 100만원 미만에 대해서는 한 구간으로 구분하되, 100만원 이상에 대해서는 600만원까지 50만원 단위로 구분하였다.

〈표 6-9〉 국민대상 통계의 소득 분류

통계명	소득 분류
문화향수실태조사	소득없음, 100만원 미만, 100~200만원 미만, 200~300만원 미만, 300~400만원 미만, 400~500만원 미만, 500~600만원 미만, 600~700만원 미만, 700~800만원 미만, 800~900만원 미만, 900~1,000만원 미만, 1,000만원 이상
국민여가활동조사	소득없음, 100만원 미만, 100~200만원 미만, 200~300만원 미만, 300~400만원 미만, 400~500만원 미만, 500~600만원 미만, 600~700만원 미만, 700~800만원 미만, 800~900만원 미만, 900~1,000만원 미만, 1,000만원 이상
국민독서실태조사	100만원 미만, 100~200만원 미만, 200~300만원 미만, 300~400만원 미만, 400~500만원 미만, 500~600만원 미만, 600~700만원 미만, 700만원 이상
국민여행실태조사	50만원 미만, 50~100만원 미만, 100~200만원 미만, 200~300만원 미만, 300~400만원 미만, 400~500만원 미만, 500~600만원 미만, 600만원 이상
국민생활체육실태조사	100만원 미만, 100~150만원 미만, 150~200만원 미만, 200~250만원 미만, 250~300만원 미만, 300~350만원 미만, 350~400만원 미만, 400~450만원 미만, 450~500만원 미만, 500~550만원 미만, 550~600만원 미만, 600만원 이상

지역에 대한 분류는 5종의 통계 모두에서 17개 시도로 모두 분류하고 있으며, 그 외에도 국민여가활동조사와 문화향수실태조사는 직장을 다니는 사람들에게 종사상의지위와 직장에서 주당 근로시간을 추가로 질문하고 있다. 또한 동거가구원수에 대해 질문하고 있다.

다. 인구총조사의 응답자 특성

국민을 대상으로 하는 통계로는 인구총조사 자료가 가장 큰 자료이다. 인구총조사의 조사표를 살펴보면, 각 가구원에 대해 인적사항을 자세히 질문을 하고 있는데, 주요응답자 관련된 특성항목을 살펴보면 〈표 6-10〉과 같다. 인구총조사의 조사 목적이 우리나라 인구의 전반적인 현황을 살펴보는 것이기 때문에 응답자 특성문항이 별도로 두지 않고 있으며 대부분의 질문에 응답자 특성이 포함되어 있다. 표에 제시한 것 외에도 가구원 수나 종교와 같은 질문이 있다.

〈표 6-10〉 인구총조사의 응답자 특성 조사표 내용

구분	조사표		조사 목적
성별	남성 여성		남녀 간의 인구 · 경제 · 사회적 특성을 비교하기 위한 항목으로 이를 다른 항목과 연계하고 성별로 세분화하여 더욱 폭 넓은 분석결과 제공
연령	주관식 응답		나이는 인구의 구조를 파악하는 중요한 항목으로 학령인구, 병역인구, 생산연령 인구, 가임여성 인구, 고령인구 등 인구를 특성집단별로 구분하는 수단으로 활용
최종학력	학력	무학(미취학 포함) 초등학교 중학교 고등학교 대학(4년제 미만) 대학교(4년제 이상) 대학원 석사과정 대학원 박사과정	국민의 교육수준을 성별, 연령별로 파악함으로써 우리나라 인구의 질적 수준 및 특성 분석
	이수여부	졸업 재학 수료 휴학 중퇴	
혼인상태	미혼 배우자 있음 사별 이혼		인구의 혼인상태 및 혼인시기는 인구규모의 변동은 물론 가구형성과 해체를 파악하고 예측하며, 저 출산의 원인과 대책 마련에 활용
직업	주관식 응답		우리나라의 직업 구조를 파악하기 위해 사업체의 종류별로 취업자가 하고 있는 일의 종류 조사
종사상 지위	임금 근로자 고용원이 없는 자영업자 고용원이 있는 자영업자 무급가족 종사자		국민의 경제활동 참가 수준은 물론 근로자, 자영자 등 일하는 형태를 파악하여 경제활동구조에 대한 종합적인 정보 제공
지역	주관식 응답		주민등록상의 거주지가 아닌 실거주지를 조사하는 항목으로 지역통계 수요에 부응

2. 데이터 연계를 위한 설문 설계에서 공통 변수 구성 방안

조사통계를 이용한 데이터 연계는 활용성이 매우 높을 것으로 예상되는데, 이는 조사통계가 설계(design)를 통해 생산되는 통계이기 때문이

다. 표본설계를 통해 조사하여 생산된 데이터이기 때문에 대표성과 정확성을 담보하고 있으며, 기준데이터로 활용하기에 적합하고 데이터 분석은 물론 추정도 가능한 장점이 있다. 특히 최근에 비정형 데이터인 빅데이터에 대한 신뢰성 문제와 활용의 장점 등에 대한 이견들이 논쟁이 되는 경우가 종종 있는데, 조사 데이터와 연계를 통해 신뢰성과 활용성을 모두 확보할 수 있을 것이다.

조사 데이터를 이용한 데이터 연계를 위해서는 충분하고 명확한 공통변수가 필요하다. 즉, 다양한 문항을 통한 정보가 있어야 한다. 공통 변수로 활용할 수 있는 문항이 부족하다면 데이터를 연계하기 어렵고, 문항이 충분히 많다 하더라도, 항목의 범주 분류 기준이 제각각이라면 데이터 연계 시 활용에 제약이 있을 수밖에 없다.

가. 연계분석을 위한 통계작성 기준

모든 통계는 작성할 때 작성시점, 조사대상 기간, 작성주기, 조사대상 등의 기준을 정하게 된다. 특히 데이터 연계를 고려한다면, 통계작성 기준을 가장 우선적으로 살펴봐야 할 필요가 있다. 데이터 연계를 위해서는 통계작성 대상은 동일해야 한다. 다음으로 작성시점도 가능한 동일해야 연계가 가능하다. '가능한'이라는 표현을 사용한 이유는 시점의 차이가 전혀 문제가 되지 않거나 시차를 이용한 분석을 할 경우가 있어서이며, 시차를 이용한 분석의 경우 시점의 차이에 대한 명확한 기준을 마련할 필요가 있다.

국민을 대상으로 하는 조사통계에서 국민여행실태조사만이 가구와 개인을 구분하여 조사하고 있으며 그 외 통계는 모두 개인을 대상으로 한다. 따라서 국민여행실태조사는 다른 통계 데이터와 연계할 경우 개인을 대상으로 하는 통계만을 사용해야 한다. 또한 작성시점의 차이로 인하여 시점이 완벽하게 동일해야 하는 경우는 연계가 어렵다. 따라서 가능하다면 조사시점의 조정을 통해 통계작성 시점을 통일할 필요가 있다. 박근화

외 1인(2013)의 데이터 연계방안 연구에서는 작성시점의 마지막 날은 12월 31일로 하는 것이 가장 타당하다고 제시하였다. 개인을 대상으로 하는 조사는 최소한의 가구원 정보를 확보한다면 가구대상 통계 데이터와 연계할 수 있는 방안도 마련할 수 있을 것이다. 따라서 가구원수, 가구 구성 등에 대한 최소한의 정보를 추가하여 응답자 부담을 최소화하는 응답항목을 마련할 필요가 있다.

나. 국민대상 통계의 응답자 특성항목 구성 방안

앞서 살펴본 통계별 응답자 특성문항은 성별, 연령 그리고 학력, 혼인상태, 직업, 가구소득, 개인소득 등으로 구성되어 있지만, 조사통계의 목적 등에 따라 응답자 특성문항과 분류에 차이가 있다.

통계를 산출할 때, 응답자 특성문항을 기준으로 구분하여 작성하기 때문에, 문항이 다르면 산출되는 통계와 산출되지 않는 통계가 나뉘게 되어 이를 연계하여 해석할 수 없다. 〈표 6-7〉에 제시한 혼인상태를 예로 들면, 혼인상태에 대한 질문이 있는 통계와 없는 통계는 당연히 혼인 관련된 통계를 연계하여 분석할 수 없다.

따라서 각자 생산된 통계 데이터를 연계하여 하나의 데이터로 만드는 데이터 연계는 기본적으로 공통 변수로 사용하는 응답자 특성 항목이 동일해야 좋은 연계가 가능하다. 그렇다면 설문지에 포함되는 응답자 특성 항목에 대한 고려가 필요하다. 본 연구에서는 문화체육관광부의 국가승인통계의 특성항목과 국민대상통계의 기본적인 표본틀로 활용되고 있는 인구총조사의 질문을 바탕으로 제시하고자 한다.

통계청의 인구총조사 자료는 조사목적 자체가 국민들의 가구구성과 주거행태, 그리고 개인별 특성 등을 파악하고자 생산하는 통계이기 때문에 응답자 특성항목으로 별도로 구분하지는 않지만, 인구특성에 대해 가장 많은 정보를 가지고 있는 기본적인 통계이다. 따라서 인구총조사 자료와 연계하는 방안을 우선적으로 고려할 필요가 있다. 인구총조사 자료는

고유식별정보를 가지고 있으며, 통계적 연계도 가능할 정도의 개인정보를 충분히 가지고 있어 데이터 연계에 대한 최적의 조건을 가지고 있다. 따라서 〈표 6-10〉에서 제시한 인구총조사 자료의 응답자 특성 내용을 가능한 포함하여 국민대상 통계조사 설문지 응답자특성 항목을 구성하는 것이 타당할 것이다.

설문지의 기본 구성항목은 인구총조사 통계의 응답자 특성항목과 동일하게 구성하기 위하여 성별, 연령, 최종학력, 혼인상태, 직업, 지역을 포함하며, 인구총조사에 존재하는 항목은 아니지만 데이터 연계 시에 활용도가 높은 가구소득과 개인소득 등을 포함할 수 있다. 이에 대해서는 〈표 6-10〉의 인구총조사 항목과 문화체육관광부의 국가승인통계에서 많이 사용되는 항목을 기준으로 제시하였다. 소득구분은 소득의 다양성과 데이터 연계에서 중요하게 영향력을 고려하여 분류하였다. 데이터 연계에서는 정보가 세밀할수록 데이터 연계의 정도(precision)가 높아지기 때문에 개방형(주관식) 질문이 좋지만, 응답자들의 응답 기피에 대한 가능성도 고려해야 한다. 따라서 연령 항목만을 개방형으로 질문 하도록 구성하였으며, 응답자들이 응답자특성 항목에 대한 거부감으로 인해 조사를 기피하는 등의 특별한 사유가 없다면 〈표 6-11〉에 제시한 질문은 기본적으로 포함할 필요가 있다.

〈표 6-11〉 응답자 특성 조사표 내용 구성 방안

항목	분류	
성별	남성, 여성	
연령	주관식 응답	
최종학력	학력	무학(미취학 포함), 초등학교, 중학교, 고등학교, 대학(4년제 미만), 대학교(4년제 이상), 대학원 석사과정, 대학원 박사과정
	이수여부	졸업, 재학, 수료, 휴학, 중퇴
혼인상태	미혼, 배우자 있음, 사별, 이혼	
직업	전업주부, 학생, 기타, 관리자, 전문가 및 관련 종사자, 사무 종사자, 서비스 종사자, 판매 종사자, 농림어업 숙련 종사자, 기능원 및 관련 기능 종사자, 장치 · 기계조작	

항목	분류
	및 조립 종사자, 단순 노무 종사자, 군인
종사상 지위	임금 근로자, 고용원이 없는 자영업자, 고용원이 있는 자영업자, 무급가족 종사자
지역	서울, 부산, 대구, 인천, 광주, 대전, 울산, 세종, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주
가구소득	소득없음, 100만원 미만, 100~200만원 미만, 200~300만원 미만, 300~400만원 미만, 400~500만원 미만, 500~600만원 미만, 600~700만원 미만, 700~800만원 미만, 800~900만원 미만, 900~1,000만원 미만, 1,000만원 이상
개인소득	

조사의 특성에 따라 부가적으로 필요한 질문을 추가하는 것은 문제가 없다. 예를 들어 '문화향수실태조사'의 경우는 '장애인 여부'에 대한 질문이 추가되어 있다. 이와 같이 필요에 의해, 또는 응답자의 응답부담에 영향을 미치지 않는다면 항목을 추가하여 활용하는 것이 좋다.

다. 연계목적을 고려한 문항

데이터를 연계할 때 공통 변수만 활용하기도 하지만, 목적변수를 같이 이용하기도 한다. 즉, 인구특성변수만을 이용한 공통 변수의 유사성을 이용한 방법과 연계목적이 명확할 경우 연계목적 변수를 선정하여 목적변수를 잘 설명할 수 있는 모형을 구축하고 모형에 적용하여 산출된 예측값을 계산하여 기준 파일의 케이스별 예측값과 가장 가까운 예측값을 갖는 연계 파일의 케이스를 연계하는 방법이 있다. 따라서 데이터 연계를 위한 유사성 측정 시, 공통 변수만을 사용하는 것이 아니라 목적변수를 종속변수로 하고 공통변수를 독립변수로 활용할 수도 있다.

목적변수를 활용한 데이터 연계를 고려하면, 설문을 설계할 때 응답자 특성문항 뿐만 아니라 주요목적과 관련된 문항의 구성 방안에 대한 고민이 필요하다. 다른 통계와의 연계를 염두에 두고 있을 경우 동일한 질문항목을 추가하는 것도 하나의 방법이다. 목적변수를 선정하여 구축된 모형에 대한 예측값을 산출할 때 연계 파일에 있는 변수들로 모형을 선정하

고 기준 파일의 데이터를 모형에 적용하여 산출된 예측값으로 거리를 측정하여 케이스를 연계하는 것이 일반적이지만, 목적변수가 기준파일과 연계파일 모두에 존재한다면 각각의 파일에서 모형을 구축하고 예측값을 산출하여 유사성 측정을 통해 연계하는 것이 더 좋은 결과를 얻게 된다.

통계청 등의 기관에서는 예산 낭비 등의 이유로 유사한 질문을 갖는 통계를 산출하지 않도록 권고하고 있다. 그러나 데이터 연계에서는 유사한 것이 아닌 동일한 질문이 있다면 더 좋은 연계가 가능하다. 따라서 전체 문항 관점에서의 유사 질문은 배제해야겠지만, 얻고자 하는 관계 또는 정보가 있다면 이와 관련된 통계와의 연계를 염두에 두고 동일한 질문을 일부 추가하는 것도 데이터 연계 관점에서는 좋은 방법이라 할 수 있다.

제4절

소결

빅데이터의 중요성이 증대됨에 따라 데이터를 보다 효율적으로 이용할 수 있는 방안에 대한 다양한 연구가 진행되고 있다. 기관(민간, 공공)은 MOU 등을 통해 서로의 데이터를 공유하기도 하고, 데이터를 보유하고 있는 기관과 분석능력이 있는 기관이 서로 협력하여 새로운 정보를 창출하는 등 데이터의 다양한 활용 방안을 마련하고 있다. 많은 기관들이 데이터의 활용도를 높이기 위해서 강구하는 것이 다양한 데이터 분석을 통해 새로운 정보를 얻는 것이며, 그 방안의 하나가 데이터 연계이다.

데이터를 활용할 때 항상 대두되는 문제 중의 하나가 개인정보보호에 대한 것이다. 개인정보보호문제는 개인에 대한 정보가 식별 가능할 때 발생하기 때문에 주로 정확 연계에서 발생한다. 고유 식별정보는 반드시 노출되지 않도록 해야 하며, 데이터 연계도 보안 시스템이 마련된 장소에서 작업되어야 한다. 데이터 연계 후의 통합파일을 사용하기 위해서는 가명처리방법을 이용하여 제공해야 하며, 식별정보가 아닌 항목도 개인을 식별할 수 있는 정보가 포함되어 있는지 검토하여 데이터를 제시하는 방법에 대한 판단이 필요하다. 데이터를 제시하는 방법에는 그룹을 나눠 값을 제시하거나 범주화하는 방법과 데이터를 삭제 또는 마스킹 처리 하는 방법 등이 있다.

통계적 연계는 고유 식별정보를 사용하지 않기 때문에 개인정보보호 문제와는 거리가 있지만, 데이터 연계를 통해 특정대상으로 오인할 소지가 있거나 여러 값들의 결합으로 인해 개인 정보가 노출될 가능성이 있으므로 개인정보보호에 대해 고민할 필요가 있다.

문화·체육·관광 관련 국가승인통계는 25종에 불과하며, 그 중 데이터 연계에 활용할 수 있는 데이터는 조사통계 17종이다. 이는 많은 정보를 파악하는 것에 한계가 있으므로, 다른 통계로부터 문화·체육·관광 관련

분야의 데이터를 추출하여 활용할 수 있다면 매우 효율적일 것이다. 특히 통계청이나 고용노동부, 국세청 또는 민간 금융 업체 또는 통신 데이터 등의 데이터를 활용할 수 있다면 상당히 좋은 정보를 활용할 수 있을 것이다. 이를 위해서는 문화·체육·관광 분야의 데이터를 분류할 수 있는 분류체계가 필요하다. 문화체육관광부에서는 문화·체육·관광 분야의 산업분류를 마련하여 활용하고 있다. 이를 기준으로 다양한 분류를 연계하여 활용한다면 다른 통계의 데이터를 활용할 수 있는 방안을 마련할 수 있을 것이다. 향후 데이터의 활용도를 높이기 위해 분류체계의 연계방안을 마련할 필요가 있다.

조사통계자료를 이용한 통계적 연계는 공통 변수들이 다양하고 많을수록 유사한 케이스끼리 연계되는 연계의 효과가 높아질 것이다. 그러나 조사통계에서는 응답자의 응답부담 등의 이유로 질문의 수를 확대하는 것은 조심스럽다.

국민대상의 조사통계를 살펴보면 응답자 현황 대부분이 공통 변수로 활용된다. 응답자 현황은 범주형 자료가 대다수이며, 이는 활용할 때 변별력이 연속형 자료에 비해 떨어진다. 따라서 제4절에서는 데이터 연계를 위해서 공통 변수로 활용할 수 있는 응답자 현황에 대한 질문의 표준화 방안을 마련하고 제시하였다. 또한 특정 목적을 위해 데이터 연계를 할 경우 설문 설계에서 데이터를 연계할 수 있도록 다른 통계와 동일한 질문을 넣는 방안도 제시하였다.

데이터 연계에서 공통 변수는 다음의 조건을 만족한다면 좋을 것이다.

① 연계시킬 데이터 간에는 공통 변수가 충분히 있어야 한다.

② 연계시킬 데이터의 공통 변수의 형태와 의미는 같아야 한다.

③ 공통 변수가 가지는 정보는 가능한 많을수록 좋다.

④ 공통 변수는 응답자특성변수와 같은 배경변수 뿐만 아니라 조사의 목적에 해당하는 주요변수가 있다면 데이터 연계의 효율이 높아질

수 있다.

즉, 데이터 연계에서 공통 변수가 표준화 되어 있다면 현재의 정보를 그대로 활용할 수 있을 것이며, 범주형 자료보다는 연속형 자료가 더욱 정교한 데이터 연계를 가능하게 한다. 그리고 응답자 특성을 고려한 문항만이 아닌 데이터 연계의 목적과 관련된 변수를 활용할 수 있다면, 이를 이용한 데이터 연계는 더욱 정교하고 목적에 부합하는 통합파일을 생산할 가능성이 높아질 것이다.

문화·체육·관광 관련 국가승인통계의 조사통계 응답자 특성문항은 대상(개인, 사업체, 기타[52])별로 다르게 구성하지만 대상이 동일하다면 동일한 기준으로 질문문항을 구성할 필요가 있다. 또한 응답자부담이 없는 범위에서 가능한 다양하고 세밀한 질문으로 구성하는 것이 좋다.

52) 여기서 기타는 예술인, 장애인, 외국인 등을 포함한다.

제7장

결론 및 제언

제1절

결론

OECD에서 발표한 2017년도 자료에 따르면 우리나라 근로자들의 노동시간은 2,024시간으로 멕시코(2,257시간)와 코스타리카(2,179시간) 다음으로 긴 것으로 나타났다. 2004년부터 주 5일제를 도입하고 있지만, 실제로는 야근 등이 잦아 잘 지켜지지 않는 것이다. 이에 정부에서는 최근 '일 가정 양립'을 위한 지원 정책을 다양하게 제공하고 있으며, '주 52시간 근무제'가 시행될 수 있도록 근로기준법을 개정하였다. 이러한 노동시간의 단축은 개인 여가시간의 증가로 연결되며, 이는 다시 문화와 관광, 스포츠 등 다양한 분야에 대한 관심으로 이어진다. 이에 따라 국민의 여가활동을 지원하기 위한 문화와 관광, 스포츠 분야의 정책적 연구가 이루어질 것이며, 다양한 연구를 위해서는 국민의 여가활동에 대한 수요나 제약사항 등을 파악할 수 있는 자료들이 필요하다.

문화·체육·관광 분야에 대한 수요가 증대됨에 따라 정책적 지원을 위한 다양한 연구가 필요하며, 이를 위해서는 정책적 활용을 위한 데이터가 뒷받침 되어야 한다. 문화·체육·관광 분야의 통계적 자료들이 많이 생산되고는 있지만, 현실적으로는 여전히 부족한 상황이다. 특히 정부나 공공기관에서 활용할 수 있는 신뢰도 높은 데이터는 턱없이 부족한 실정이다. 하지만, 지속적으로 활용하지 않는 데이터를 매번 새롭게 생산하는 것은 많은 비용(예산, 인력, 시간 등)이 소요되어 경제적 효율성이 떨어진다. 특히 정책은 시의성이 중요한데, 새로운 데이터를 빠르게 생산하는 것은 매우 어려운 일이다.

정책적 시의성을 확보하기 위해 데이터를 빠르게 확보하고자 하는 노력은 지속적으로 있어왔다. RDD(random digit dialing) 방식의 전화조사, 조사업체들이 보유하고 있는 온라인 패널(online panel) 대상의 온라

인조사(online survey), SNS (social network services/sites) 데이터를 빅데이터 기법으로 분석하는 SNS 분석 등의 조사(survey)는 일시적인 정보로 사용할 수는 있지만, 우리나라 국민 대상의 정책적 지원을 위한 정보로는 자료의 대표성에 대한 보장이 어려운 한계가 있다.

최근 빅데이터의 중요성이 커짐에 따라 활용도는 확대되고 있으나, 분석 결과에 대한 신뢰성 문제는 항상 야기되어 왔다. 그 이유는 (대부분의) 빅데이터는 표준화되지 않은 데이터이기 때문이다. 또한 대체적으로 빅데이터와 동일시되는 SNS 데이터는 모든 사람을 대상으로 하는 것이 아니라 특정 건에 대해 관심을 가진 사람을 대상으로 하기 때문에 데이터의 대표성이 부족하다. 따라서 마케팅 측면에서 활용하기에는 효율적일 수 있지만 국민에 대한 대표성이 필요한 정책적 측면에서는 활용하기에 부적절하다.

본 연구에서 다루고 있는 데이터 연계는 시의성 있는 정보가 필요할 때 활용할 수 있는 대안이 될 수 있다. 각각 존재하는 데이터를 연결하여 하나의 데이터로 만드는 데이터 연계는 통합 데이터를 생성함으로써 서로의 연관관계 등을 파악할 수 있으며, 이를 통해 개별 데이터로는 알 수 없던 새로운 정보를 얻을 수 있다. 또한 대표성이 확보된 국가승인통계를 기준 파일로 활용하고 대표성이 확보되지 않은 데이터를 연계 파일로 하는 데이터 연계는 통계결과에 대한 신뢰성을 제시할 수 있다. 데이터 연계는 다양한 절차와 방법을 적용할 수 있기 때문에 데이터 연계를 통한 통합 파일을 제시하는 것이 쉽지 않지만, 새로운 통계를 생산하는 비용에 비해 적은 비용으로 데이터 생산이 가능하다는 장점이 있다.

사회에서 화두가 되고 있는 빅데이터는 양적인 측면에서 큰(big) 데이터를 의미하기도 하지만 단일 데이터로써의 큰 데이터가 아닌, 다양한 형태의 많은 데이터를 빠르게 분석하여 가치 있는 정보를 찾아 활용하는 것을 의미하기도 한다. 즉, 다양한 형태의 데이터가 쌓여 있는 것을 분석하여 새로운 정보를 찾고 그 결과를 제공하는 과정을 모두 포함하는 개념

이다. 이러한 빅데이터 분석은 여러 데이터를 연결하여 분석하는 것을 포함하는데 데이터를 연결하는 방법 중의 하나가 바로 데이터 연계이다.

데이터를 연결하는 방법에는 개별 데이터의 패턴을 분석하는 방법과 데이터를 직접 연계하는 방법 등이 있다. 전자의 경우, 데이터를 분석하여 그 안의 패턴을 찾아내거나 데이터 간의 패턴 또는 관계를 분석하는 것은 대용량의 데이터가 있어야 가능하다. 하지만 후자의 경우, 데이터를 직접 연계하여 분석하는 절차와 방법은 복잡하지만 적은 데이터로도 가능하기 때문에 매우 효율적이다. 본 연구에서는 데이터를 직접 연계하는 후자의 방법을 다루고 있다. 이는 가치의 개념에서 빅데이터라 할 수 있으며, 또한 빅데이터 분석 방법의 하나라고도 할 수 있다.

본 연구에서 다루는 데이터 연계는 정확 연계와 통계적 연계로 나눌 수 있다. 정확 연계는 통계청과 같은 국가통계를 다루고 있는 기관에서 주로 많이 사용하고 있는 방법으로 주민등록번호나 사업자등록번호와 같은 고유식별 정보를 이용하여 동일한 대상을 연계하는 방법이다. 고유식별 정보는 대상자를 각각 식별할 수 있는 정보로 반드시 ID와 같은 하나의 변수로 구성되는 것은 아니며, 여러 변수를 결합하여 생성할 수도 있다.

정확 연계는 정확한 정보를 하나의 데이터로 통합하여 분석하기 때문에 세밀한 정보를 파악할 수 있지만, 개인정보와 같은 민감한 사항에 대한 문제가 항상 대두될 수밖에 없다. 따라서 정확 연계 방법을 적용하기 위한 데이터는 보안센터 내에서 보안망(security network)을 통해서만 다룰 수 있다. 이처럼 정확 연계는 활용성이 매우 뛰어나지만 실제로 활용하기 위해서는 많은 제약사항이 존재한다.

본 연구에서 다룬 정확 연계 중, '문화체육관광산업통계'는 문화체육관광사업체 표본틀과 통계청의 기업등록부(BR : business register) 자료를 연계하여 생성하는 국가승인통계로, 데이터 연계 시 정확 연계의 방법을 이용하며, 일부 연계가 되지 않은 데이터는 가중값 조정 등을 통해 가공한, 정확한 통계라 할 수 있다. 문화체육관광산업통계는 통계청의

보안센터인 빅데이터센터에서 데이터 연계 및 통계 생산에 대한 작업을 하고 결과의 반출 승인을 요청하여 심의를 통과한 이후 결과파일을 제공받아 활용하고 있다. 신한카드 데이터와 기상청 데이터의 정확 연계는 는 위도와 경도, 날짜 정보를 이용하여 정확 연계를 하였기 때문에 연계 시의 개인정보에 대한 문제의 소지는 없지만, 카드데이터 자체가 민감한 정보이다. 따라서 데이터를 비식별 조치하고 일부 데이터를 랜덤하게 추출하여, 승인을 받은 이후에 제공받았다. 신한카드와 연구원의 MOU에도 불구하고, 개인정보에 대한 민감성으로 인해 데이터를 활용하기까지 많은 시간과 검토가 이뤄질 수밖에 없었다. 이렇듯 정확 연계는 쉽고 간단하게 적용 가능하며 많은 정보를 얻을 수 있다는 큰 장점이 있지만, 연계를 통해 민감한 정보까지 얻게 될 수도 있어 심각한 개인정보문제가 발생할 수 있으며, 이에 따라 활용하는데 제약을 많을 수밖에 없다.

통계적 연계는 정확 연계와 같이 고유식별 정보를 활용하는 것이 아니기 때문에 활용하는데 제약은 없지만 다양한 검증 절차가 필요하다. 또한 동일한 대상을 연결하는 것이 아니라 유사한 대상을 연계하여 정보를 제공하기 때문에 데이터 연계 과정에서 일부 정보가 손실될 수 있음을 반드시 밝혀야 한다.

통계적 연계의 적용은 문화체육관광부의 국가승인통계인 '국민여가활동조사' 데이터를 기준 파일로 활용하였으며, '문화향수실태조사' 데이터와 '한국의료패널조사' 데이터를 각각의 연계 파일로 활용하였다. 본 연구에서는 데이터 연계 방법의 중요성을 높이 판단하여 자세하게 설명하였으며, 실제 데이터를 이용한 연계 시에는 데이터 검증방법을 특히 세밀하게 다루었다. 이는 동일한 대상을 연계한 것이 아니라는 이유로 통계적 연계를 통한 통합파일의 데이터에 대한 신뢰가 낮은 경우가 많기 때문이다.

다양한 연구와 영국, 호주, 캐나다 등의 많은 나라에서 통계적 연계의 효율성을 인식하여 활용하고 있다. 정확 연계가 아니기 때문에 통계적 연계를 국가의 공식 통계로 활용하지는 않지만, 많은 연구에서 활용하고

있으며 그 결과에 대한 신뢰가 높아지고 있다. 우리나라에서는 통계적 연계에 대한 연구가 일부 진행되고 있지만, 활용은 대부분 정확 연계 후에 통계적 연계를 적용하는 방식으로, 엄밀히 말하면 통계적 연계가 아닌 확률적 연계 방법이라 할 수 있다.

본 연구는 국내에서 아직 활발하게 활용되지 못하고 있는 데이터 연계의 두 가지 방법(정확 연계와 통계적 연계) 모두를 자세히 제시하는데 의의를 두고 있다. 통계적 연계는 데이터 활용성 등을 고려하면 꼭 필요하며 사용될 가능성이 높다. 본 연구의 제5장에서는 데이터의 연계 과정을 자세하게 제시하고 있으며, 오픈소스(open source) 통계프로그램인 R을 사용하였다. 빅데이터의 영향으로 오픈소스인 R은 다양한 데이터 연계 방법에 대한 함수를 제공하기 때문에 쉽게 적용할 수 있을 것이다.

제2절

제언

데이터 연계는 4차 산업혁명과 더불어 21세기의 중요한 자원인 데이터의 활용을 극대화할 수 있는 방안이다. 4차 산업혁명으로 인한 사회의 발전은 여가의 수요 증대에 영향을 미치며, 이는 문화와 관광, 스포츠 등 여가 향유에 대한 중요성의 확대로 연결된다. 향후 여가와 복지 등의 분야에 대한 정책의 필요성이 커짐에 따라 정책 마련을 위한 데이터의 필요성 역시 커질 것이다. 하지만 매번 데이터를 생산할 수는 없으므로, 데이터 연계는 문화·체육·관광 분야의 데이터를 효율적으로 활용하기 위해 매우 중요한 수단이다.

우리나라에서의 데이터 연계는 그에 대한 인식이 부족하고, 데이터의 활용도가 매우 낮은 실정이다. 또한 데이터 연계 시에 많은 제약 사항이 존재하여, 이번 절을 통해 데이터 연계를 확대하기 위한 방안들을 제언하고자 한다.

데이터 활용의 효율성을 높이기 위해서는 신뢰할 수 있는 많은 데이터가 있어야 한다. 문화·체육·관광 관련 데이터는 많이 생산되지만, 대표성을 확보할 수 있을 정도로 신뢰할 만한 데이터는 부족하다. 국가승인통계와 정기적으로 생산되는 비승인통계 데이터를 제외하면 대부분의 데이터가 1회성 연구 또는 정책적 활용을 위해 생산되는데 그치고 있어 경제적 측면(예산, 인력, 시간 등)에서 비효율적이며, 데이터의 관리 또한 되지 않고 있다. 따라서 문화·체육·관광 분야의 데이터를 하나로 모을 수 있는 제도적 장치와 데이터를 관리할 수 있는 물리적 공간(데이터베이스)이 마련되어야 한다.

우리나라 통계는 분산형 통계로 부처에서 필요한 통계를 각각 생산하고 있으며, 그 중 국가승인통계만이 통계청의 관리 하에 국가통계포털

(KOSIS)에서 통합 서비스되고 있기 때문에 국가승인통계 외에는 통합적으로 관리할 수 있는 제도적 장치가 없다. 이러한 이유로 국가 예산으로 생산되는 문화·체육·관광 관련 데이터는 현재 통합 관리되고 있지 않고 있다. 문화체육관광 관련된 데이터의 수요는 물론 빅데이터 관점에서도 통합관리가 필요한 시점이다. 따라서 문화체육관광부의 통계담당부서에서는 문화체육관광 관련 데이터를 통합적으로 관리하는 방안을 마련해야 할 것이다.

빅데이터의 중요성과 필요성은 점차 확대되고 있는 반면, 문화·체육·관광 분야의 빅데이터 분석은 SNS분석과 통신데이터 분석이 대부분이다. 즉, 고차원적인 빅데이터는 거의 활용되고 있지 않다. 이는 빅(Big) 데이터가 없기 때문이기도 하며, 다양한 목적으로 생산되고 있는 데이터가 체계적으로 관리되고 있지 않아 문화체육관광 관련된 분야의 빅데이터 분석은 민간에 의존할 수밖에 없기 때문이다. 따라서 문화·체육·관광 관련 데이터에 대한 체계적인 통합 관리방안이 필요하며, 이는 빅데이터의 활용으로 연결될 것이다.

데이터 연계를 위해서는 문화·체육·관광 관련 데이터에 대해 구분하고 분류할 수 있는 체계가 필요하다. 다양한 기관에서 데이터가 생산되고 있으며, 타 기관에서 생산하는 데이터에 문화·체육·관광 분야의 데이터가 포함되어 있는 경우가 많이 있다. 따라서 이를 구분하거나 분류하는 것은 새로운 문화·체육·관광 데이터를 생산하는 것과 동일한 효과를 가진다.

문화체육관광부에서는 문화·체육·관광 분야의 산업분류를 별도로 관리하여 표준산업분류와 연계할 수 있도록 하고 있다. 이처럼 다양한 분류체계 마련은 많은 분야에서 활용이 가능할 것이다. 즉, 문화·체육·관광 분야의 통합 분류체계를 마련함으로써, 다양한 분야에서 생산되고 있는 데이터로부터 문화·체육·관광 분야의 데이터를 분류하고 추출하여 활용할 수 있을 것이다.

또한 문화·체육·관광 분야의 데이터를 소유하고 있는 기관과 데이터

활용을 위한 MOU 체결을 통해 정확 연계의 발판을 마련할 필요가 있다. 통계청을 포함하여 카드데이터, 이동통신, 문화예술 및 연예 등의 분야에 대한 데이터를 서비스하는 기관과 MOU를 체결하여 생산된 데이터를 활용할 수 있다면 데이터의 활용이 훨씬 높아질 것이다. 데이터 경제라는 말과 같이, 타 기관의 중요한 데이터를 사용가능해진다면 데이터의 정책적 활용도는 더욱 높아질 것이다.

데이터 연계에서는 정확 연계가 할 수 있다면 정확 연계를 하는 것이 더 좋다. 그러나 개인정보보호 등 관련 문제가 발생할 수 있어 데이터의 활용에 제약이 있다. 따라서 데이터 활용을 위한 MOU 체결을 위해서는 다양한 데이터 정책이 마련되어야 한다. 그에 대한 일례로, 데이터 작업을 위한 보안센터를 마련하거나, 데이터의 단계를 구분하여 데이터 단계별로 접근 및 활용 범위를 정의하여 제공하는 것을 들 수 있다.

통계적 연계 방법은 어렵지만 개인정보문제에서는 다소 자유롭기 때문에 활용하는 데 제약은 없다. 또한 대표성 있는 통계를 기준 파일로 활용한다면 신뢰성이 부족한 데이터는 데이터 연계를 통해 대표성을 충족할 수 있는 장점이 있다. 따라서 비정형 데이터인 빅데이터와의 연계방안이 마련된다면 통계적 연계는 매우 훌륭한 데이터 연계 방법이 될 것이다. 그러나 통계적 연계는 연계하는 과정에서 충족시켜야 할 가정이 제약조건이 된다. 즉 통계적 연계의 가정이 제약조건이다. 제약조건 중에서 공통변수가 각각의 케이스들을 잘 연계할 수 있는 수준의 정보를 가지고 있어야 한다는 조건은 절대적이다. 따라서 문화·체육·관광 분야의 데이터를 생산할 때, 특히 조사통계의 경우 공통 변수로 활용이 가능한 응답자 기본정보 관련 문항의 구성에 대한 표준방안을 마련할 필요가 있다. 연계하고자 하는 분야의 데이터가 충분한 공통 변수를 가지고 있지 않다면 데이터 연계가 어렵겠지만, 최소한 문화·체육·관광 관련 데이터는 연계가 가능하도록 함으로써 데이터가 충분히 활용될 수 있도록 해야 한다.

이번 절에서 제안한 것은 대부분 데이터 정책과 관련된 부분이다. 문화·

체육·관광 분야의 데이터를 생산 및 활용하기 위한, 데이터 관련 정책은 반드시 마련되어야 한다. 영국의 통계청은 데이터정책을 마련하여 데이터의 등급에 따라 활용을 할 수 있는 방안을 제시하여, 데이터의 활용도를 높이고 있다. 데이터의 중요성이 점차 증대되며, 데이터를 자원으로 생각하고 있는 시점에서 데이터 관련 정책은 반드시 마련되어야 할 것이다.

참고문헌

1. 국내외 문헌

김대건, 박민규(2014), 마이크로 매칭 방안 비교 연구–경제활동인구조사와 생활시간조사 자료에 적용, 한국자료분석학회, Vol.16, No.5.

김희경(2010), 가중 k–최근접이웃방법을 이용한 통계적 매칭 기법에 관한 연구, 동국대학교.

박근화, 이강욱, 이용관, 한정임, 송정련(2017), 문화체육관광 규모 추정방안 연구, 한국문화관광연구원.

박근화, 이관제(2003), 비용효과에 대한 관측자료 연구에서의 로버스트 방법을 이용한 추론, 한국자료분석학회, Vol.5, No.3.

박근화, 한정임(2013), 종합문화통계 생산을 위한 통계 연계방안 연구, 한국문화정보센터.

박근화, 한정임, 송정련(2017), 2017년도 문화체육관광분야 사업체 표본틀 구축 보고서.

박우창, 승현우, 용환승, 최기헌(2004), 데이터 마이닝 개념 및 기법, 자유아카데미.

배현주(2015), 국민건강보험 빅데이터 연계 기후변화 건강영향평가, 환경부.

변종석, 박민규, 박인호, 임경은. 최재혁(2013), 다양한 출처 자료 처리 및 통계생산방안 연구, 통계개발원.

오기환, 전수연, 최명호(2015), Big Data의 이해와 활용, 디지에코 보고서.

오미애(2015), 보건복지분야 데이터 연계 필요성 및 활용방안, 보건복지포럼.

오미애, 최현수, 김용대, 이용희, 진재현(2014), 보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안, 한국보건사회연구원.

이영섭, 김선웅, 안홍엽, 임경은, 김희경(2009), 「통계조사 자료와 행정 자료간의 통계적 연계 기법에 관한 연구」, 통계연구, 제14권 제1호, 통계개발원.

이은우(2017), 데이터 연계, 겨합합 지원제도 도입방안 연구, 개인정보보호위원회.

정미옥, 백지선, 김현식, 최은영(2016), 사회조사와 인총 표본조사 연계자료 분석

및 활용, 통계개발원 2016년 하반기 연구보고서 통계개발원.

정미옥, 최필근(2014), 사회조사 자료연계 방법 연구, 통계개발원 2014년 하반기 연구보고서, 통계개발원.

최현수, 오미애(2015), 데이터 연계방법론을 활용한 납세 및 복지수급 여부에 따른 복지인식 비교 분석, 한국자료분석학회, Vol.17, No.4.

통계청(2016), 통계데이터센터 이용 가이드.

통계청(2018), 행정 자료관리 주요동향.

한국문화관광연구원(2018), 2016년 기준 문화체육관광산업통계.

함영진(2014), 맞춤형 사회보장 서비스 제공을 위한 복지정보 및 고용정보 활용 방안, 한국보건복지정보개발원.

행정안전부(2009), 김대리, 개인정보보호 달인되기.

Arbeitsgemeinschaft Media-Analyse. (1996). Die Media-Analyse der Arbeitsgemeinschaft Media-Analyse eV.

Australian Bureau of Statistics(2013), Australian Census Longitudinal Dataset : Methodology and Quality Assessment 2006-2011, ABS cat. no. 2080.5.

BSA.ORG(2015), 데이터는 왜 중요한가?.

D'Orazio, Marcello. Di Zio. Marco and Scanu. Mauro(2006). Statistical tching Theory and Practice, Wiley.

Economist Intelligence Unit. The Deciding Factor: Big Data & Decision Making. Cap Gemini, 2012. https://www.capgemini.com/resources/the-decidingfactor-big-data-decision-making.

Kamakura, W.A. and Wedel, M.(1997), Statistical Data Fusion for Cross-Tabulation. Journal of Marketing Research, 34, 485-498.

Muennig, P., Johnson, G., Kim, J., Smith, T. W. and Rosen, Z.(2011), "The general social surveynational death index: an innovative new datasest for the social sciences", MBC Research Notes 4:385.

National Research Council(1992), Combining Information : Statistical

Issues and Oppoyunities for Research. Washington. D.C. : National Academy Press.

National Statistics(2003). National Statistics code of Practice Protocal on Data Matching.

Solon, R. and G. Bishop(2009), "A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset", ABS cat. no. 1351.0.55.025.

Statistics New Zealand(2006), Data Integration Manual.

Statistics New Zealand(2013), Developing a historical longitudinal census dataset in New Zealand.

Van Der Putten, Peter , Kok, Joost N., and Gupta, Amar (2002), Why the information explosion can be bad for data mining, and how data fusion provides a way out, Second SIAM International Conference on Data Mining, Arlington, April, 11-13.

Van Pelt, X.(2001), The Fussion Factory: A Constrained Data Fussion Approach. Master of Science. Thesis, Leiden Institute of Advanced Computer Science, The Netherlands.

2. 웹사이트

뉴질랜드 통계청 https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure

미국 인구조사국 https://www.census.gov/about/adrm/linkage.html

영국 행정 데이터연구센터 https://www.adrn.ac.uk

위키백과 포털 https://ko.wikipedia.org/

캐나다 통계청 https://www.statcan.gc.ca/eng/sdle/index

행정안전부 공공데이터 포털 https://www.data.go.kr/

3. 법률 및 규정

개인정보보호법

정보통신망 이용촉진 및 정보보호 등에 관한 법률

ABSTRACT

A Study on the Production and Utilization of Big Data through the Data Linkage in Culture, Sports, and Tourism Field

The purpose of this study is to propose ways to link data that has already been produced and to produce and utilize Big Data with these data. Accordingly, we looked at data on culture, sports and tourism field and actually conducted data linkage with these data.

The data linkage is to create an integrated file by linking separate data produced individually, which can be largely divided into precise linkage and statistical linkage. The precise linkage is a method of linking through a unique key variable when there is a unique key variable that can identify an individual. And the statistical linkage is a method of statistically linking data when there is no personally identifiable inherent key variable.

In this study, we have studied the stepwise method of data linkage in detail. Prior to the data linkage, we identified the data related to culture, sports, and tourism field and data linkage was carried out using some data among them.

Using precise linkage, we analyzed the movement-expenditure on climate by linking card data and meteorological data and analyzed the current status of management activities of the culture, sports and tourism industry through linking sample frames of businesses in this field with statistical office's administrative data. The link between card data and meteorological data is to connect the shortest distance data by calculating the latitude and longitude of the card data merchant's point and the distance between the latitude and longitude of the observing station of

the climate data. The link between business sample frame and the statistical office' administrative data was connected using a unique key variable called the business type.

Using statistical linkages, we analyzed relationships between cultural capital and leisure activities using Survey of national Leisure Activity and Survey of Cultural Perception and analyzed health effects of leisure activities through linking the Survey of National Leisure Activities with medical panels. The data link between the national leisure activity survey and the culture improvement status survey was connected using the Gore Distance calculation method, which sets gender and monthly average household income as blocking variables and links data on common variables such as residential area and age. The data link between the National Leisure Activity Survey and the Korea Medical Panel Survey was established as a blocking variable and then linked using the Gore Distance calculation method, which calculates the shortest distance of common variables such as region and age.

One of the considerations in linking data is to address personal information problems. When linking data, information that can identify characteristics of an individual should be removed in advance, or personal information should be taken to prevent leakage by means such as the Personal Protection Act before using the data.

참여 연구진

연구책임자

박 근 화(한국문화관광연구원 수석전문위원)

공동연구자

장　　훈(한국문화관광연구원 부연구위원)

한 정 임(한국문화관광연구원 차석전문원)

김 정 림(한국문화관광연구원 차석전문원)

외부연구자

김 희 경(동국대학교 연구초빙교수)

김 광 섭(동국대학교)

문화·체육·관광 데이터 연계를 통한
빅데이터 생산 및 활용방안 연구

발 행 인　김 정 만

발 행 처　한국문화관광연구원
서울시 강서구 금낭화로 154
전화 02-2669-9800　팩스 02-2669-9880
http://www.kcti.re.kr

인 쇄 일　2018년 12월 28일

발 행 일　2018년 11월 30일

인 쇄 인　더크리홍보 주식회사

ISBN : 978-89-6035-748-8　93300