문화·관광 분야 가명처리 데이터 활용방안 연구

2021-05

기초연구

A Study on the Utilization of Pseudonymized Data in Culture and Tourism

권태일
송정연
김성준

한국문화관광연구원
Korea Culture & Tourism Institute

기초연구 2021-05

문화·관광 분야 가명처리 데이터 활용방안 연구

A Study on the Utilization of Pseudonymized Data in Culture and Tourism

권태일·송정연·김성준

한국문화관광연구원
Korea Culture & Tourism Institute

연구책임

권태일 정책정보센터 통계관리팀/연구위원
송정연 정책정보센터 통계관리팀/차석전문원

공동연구

김성준 문화산업연구센터/연구원

연구개요

1. 서론

가. 연구 배경 및 목적

1) 연구배경

- 2020년 6월 데이터를 기반으로 한 과학적 행정 구현을 국정과제로 설정하고 복지 분야에서 맞춤행정을 위한 데이터 센터를 운영하는 것을 시작으로, 「데이터 기반 활성화에 관한 법률」[1]이 재정됨
  - 개인이나 조직의 경험이나 직관에 따라 정책을 수립하는 방식에서 나아가 '데이터를 정책 수립 및 의사결정에 활용함으로써 객관적이고 과학적 행정을 수행'할 수 있는 '데이터 기반 행정'의 제도적 기반이 마련되었음

- 2020년 8월 데이터 3법(데이터 규제 완화 3법)[2]의 개정안 통과로, 가명정보를 통한 활용 가능한 데이터의 범주가 확대됨에 따라 통계작성, 사회과학 및 정책 연구라는 제한적 목적 하에서 정보주체의 동의 없이 가명정보를 처리할 수 있는 환경이 조성됨

- 이렇듯 개인 정보를 가명 처리한 데이터(이하. 가명처리 데이터) 활용의 기반이 마련됨에 따라 경제, 사회 및 복지, 의료, 문화·관광 등 다양한 분야에서 활용 가능한 공공 및 민간 데이터의 수가 증가할 뿐만 아니라, 기존에 활용되고 있는 데이터의 활용성이 개선되어 그 가치가 높아질 것으로 예상됨
  - 가명정보를 활용하게 될 경우, 기존 빅데이터에서 파악 가능한 유형화된 데이터를 넘어 보다 다양한 인구통계학적 변수를 활용할 수 있으므로 기존 데이터로 파악하기 어려웠던 새로운 집단의 정의 및 분석이 가능하게 될 것으로 판단됨

- 이에 따라 현재 '가명처리 데이터'를 이용해 정책적으로 유의미한 통계를 생산하기 위한 제도적 기반(예. 개인정보처리 가이드라인 마련, 가명정보 결합 전문기관 지정 등)마련이 활발하게 이루고 지고 있음
  - 2020년 9월 개인정보위원회에서 '가명정보 처리 가이드라인'을 발간하여 가명

---

1) 데이터기반행정 활성화에 관한 법률(약칭: 데이터기반행정법) [시행:2020. 12. 10] [법률: 제17370호. 2020년 6월 9., 재정]
2) 데이터 이용을 활성화하는 「개인정보 보호법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률(약칭: 정보통신망법)」, 「신용정보의 이용 및 보호에 관한 법률(약칭: 신용정보법)」등 3가지 법률을 통칭

처리 및 가명정보 결합과 반출에 관한 절차를 구체적으로 안내하고 있음

- 또한 2020년 10월 보건복지부에서 보건의료분야 '가명정보결합 전문기관'[3]결합전문기관으로 국민건강보험공단, 건강보험심사평가원, 한국보건산업진흥원을 지정한 것을 시작으로 2020년 11월 개인정보위원회에서 통계청과 삼성SDS를 '21년 1월 과학기술정보통신부에서 한국지능정보사회진흥원(NIA), SK주식회사, 더존비즈온을 가명정보 결합전문기관으로 지정함

■ 이에 문화·관광분야 또한 이러한 환경 변화에 맞춰 가명처리 데이터 활용의 필요성을 점검[4]하고, 기존 데이터 분석으로 발견할 수 없었던 새로운 정책적 활용 방안을 제시하기 위한 연구가 필요한 시점임

2) 연구목적

■ 본 연구의 목적은 가명데이터의 활용이 원활해짐에 따라 선도적으로 문화관광분야에서 가명처리데이터 활용에 필요한 일련의 프로세스를 직접 검토하고 분석하는 절차를 제공함으로써 향후 관련분야 데이터 활용 및 후속연구를 위한 기초적 내용을 제공하는 것임

- 연구목적을 달성하기 위해 설계(design), 분석(analysis), 진단(diagnosis)으로 구분한 3가지 연구진행 프로세스를 설정하였으며 이를 토대로 세부적인 연구내용을 제시함

〈표 1〉 연구 목적에 따른 세부 사항

연구 목적	세부사항
가명처리데이터의 구조파악 및 현황검토를 통한 **설계(design)**	- i) 가명데이터의 명확한 개념 및 활용을 위한 구조(structure) 제시 - ii) 가명데이터 특징 검토 및 분서을 위한 고려사항 제시 - iii) 사례분석을 통한 가명데이터 활용 및 결합방안 제시
가명처리데이터 설계에 기반한 **분석(analysis)**	- i) 가명데이터의 문화관광분야 활용컬럼 도출 및 트렌드 분석 - ii) 이종 간 데이터 결합을 통한 활용방안 도출

3) 서로 다른 개인정보 처리자간의 가명정보 결합을 수행하기 위해 개인정보 보호위원회(이하 "보호위원회"라 한다) 또는 관계 중앙행정기관의 장이 지정하는 전문기관을 의미함

4) 문화·관광 분야 또한 데이터 3법 시행 후 신한카드-SK텔레콤 간 MOU를 체결('20.8)하고 통신과 카드 데이터에 대한 결합 1호 신청('20.9)하였으며, 연구원에서도 데이터 결합 및 활용에 선도적으로 대응하기 위해 신한카드, SK텔레콤과 3자간 MOU를 체결 ('20.9)함.

연구 목적	세부사항
설계, 분석단계의 한계를 기반으로 지속가능한 데이터 활용을 위한 **진단(diagnosis)**	- i) 가명데이터 효율적 활용을 위한 가이드라인 제시 - ii) 가명데이터 원활한 활용을 위한 제도적 개선방안 제시

나. 연구 범위 및 방법

1) 연구 범위

- **(데이터 활용 범위)** 본 연구는 가명데이터 구조분석을 통해 활용방안을 도출하는 것이 주목적이므로 데이터의 활용 범위를 명확화가 필요
  - 가명정보 활용을 위해 현재 제공 중인 문화·관광분야 공공 및 민간 데이터를 종합검토하고 데이터 결합 시 가장 활용도가 높은 데이터 가 무엇인지 살펴보고 데이터의 활용 범위를 설정하고자 함
- **(공간적 범위)** 활용 가능한 자료 및 분석목적에 따라 설정
  - 예를 들어 전국분석 및 특정지역 분석 목적에 따라 통신데이터의 경우 집계구, 조사구, 시군구, 관광지점 등으로 구분되며 신용카드의 경우 시군구, 특정지점, 가맹점 등의 단위로 설정
- **(내용적 범위)** 설계(design), 분석(analysis), 진단(diagnosis)으로 구분
  - **설계단계:** 가명정보의 개념, 가명처리 및 가명데이터 결합, 가명데이터 및 결합 데이터 활용 시 고려사항, 가명데이터 활용을 위한 공공·민간데이터 현황 및 활용 사례
  - **분석단계:** 설계단계에서 도출된 최종 활용 데이터를 기반으로 가명데이터 만을 대상으로 한 분석과 가명데이터의 이종 간 데이터 결합을 통한 이원화된 분석을 통한 문화관광분야 활용방안을 도출
  - **진단단계:** 설계 및 분석단계를 통해 도출된 가명데이터 활용의 한계점 및 향후 보다 합리적이고 효율적인 데이터 활용을 위한 정책적 방안을 제시

2) 연구 수행 방법

- 문헌연구를 통해 개인정보 가명처리에 대한 개념, 절차, 데이터 활용범위 점검 등

의 이론적 개념 검토, 활용가능한 가명처리 데이터 현황 및 국내외 활용 사례 점검

- 데이터 분석에 앞서서 활용 가능한 가명정보(key)변수를 선정하고, 문화관광분야에서 활용 가능한 변수를 분류하고, 각 변수의 로직과 이상치를 점검하는 등 통계분석을 위한 데이터 전처리를 우선적으로 진행
- 처리된 데이터를 바탕으로 문화관광분야 핵심 현황파악을 위한 분석주제를 도출하고 ⅰ) 가명데이터만을 이용한 분석과 ⅱ) 가명데이터에 기반한 이종 간 결합데이터를 이용한 이원화된 통계분석 실시하여 시사점을 도출
- 통신데이터와 카드데이터를 직접 운영하는 관련분야 전문가들의 의견 및 협조를 기반으로 연구를 수행
- 문화관광분야에서의 활용을 위한 관련분야 전문가 자문 및 통계적 모델링과 더불어 도출된 결과에 대한 활용 등의 업무를 수행 경험이 있는 데이터 분석 전문가를 대상으로 자문회의를 실시

2. 가명정보와 가명처리의 이론적 개념

가. 가명정보의 개념

1) 가명정보의 정의

- **(정의)** 개정된 개인정보보호법상 가명정보란 개인정보를 가명처리함으로써 원래의 상태로 복원하기 위한 추가정보의 사용·결합 없이는 특정 개인을 알아볼 수 없는 정보를 의미함
  - 이때, 가명처리란 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가정보 없이는 특정 개인을 알아볼 수 없도록 하는 것임
- **(활용범위)** 가명정보는 통계작성(상업적 목적을 포함), 연구(산업적 연구를 포함), 공익적 기록보존 등을 위하여 가명정보를 제공하는 경우에는 개인인 신용정보주체의 동의 없이 가명정보를 활용할 수 있음(개인정보보호법 제32조제6항제9호의2)

나. 가명처리 및 데이터 결합절차

■ 문화관광분야의 가명처리 데이터를 활용한 연구 수행에 있어서 개인정보를 침해하지 않는 선에서 개인 정보를 처리하는 과정과, 더 나아가 이종의 가명처리 데이터를 결합하는 절차를 살펴봄으로써, 향후 문화관광분야의 가명처리 데이터를 활용한 연구 및 통계 생산의 가이드라인을 제시하고자 함

1) 개인정보 가명처리

■ 개정된 개인정보 보호법에 따르면 개인정보처리자는 개인정보의 유형, 성격 등을 고려하여 법령을 준수하는 범위 내에서 가명처리 절차와 방법을 자율적으로 판단하여 처리할 수 있지만 시행 초기임을 고려하여 산업현장에서 실제 가명처리를 수행할 때 법적 불확실성을 해소하기 위해 가이드라인을 따를 것을 권고하고 있음

- 가명처리의 절차는 1) 사전준비 2) 가명처리 3) 적정성 검토 및 추가 가명처리 4) 활용 및 사후관리의 단계로 구성

2) 가명데이터 결합

■ 개인정보보호법 제28조의3에 따라 통계작성, 과학적 연구, 공익적 기록보존 등을 위한 서로 다른 개인정보처리자 간의 가명정보의 결합은 보호위원회 또는 관계 중앙행정기관의 장이 지정하는 결합전문기관에서 수행

■ 결합전문기관은 서로 다른 개인정보처리자 간의 가명정보 결합을 수행하기 위해 개인정보 보호위원회 또는 관계 중앙행정기관의 장이 지정하는 전문기관으로, 개인정보보호법에서는 '결합전문기관', 신용정보법에서는 '데이터전문기관'에서 가명정보의 결합을 수행함

- 단, 개인정보보호법에서는 신용정보법과 상이하게 '결합키관리기관'[5]이 별도로 지정(시행령 제29조의3제2항)

■ 타 기관의 정보를 결합하여 활용하고자 하는 경우 반드시 개인정보호법과 신용정

5) 결합키 결합정보를 생성하여 결합전문기관에 제공하는 등 가명정보의 안전한 결합을 지원하는 업무를 하는 한국인터넷진흥원(또는 보호 위원회가 지정하여 고시하는 기관)을 의미함

보법에 근거하여 결합전문기관 또는 데이터 전문기관을 통해 가명정보 결합을 수행해야함

- 절차 대부분은 유사하지만 결합키 관리 및 반출 단계 등에서 일부 차이가 존재하므로 개인정보의 결합 또는 개인신용정보를 포함한 개인정보의 결합 시에는 이를 주의해야함

■ 실제로 데이터 3법 개정이후 가명정보 결합 현황을 살펴보면 일반적으로 결합절차 진행 시 상당한 시간이 소요되므로 이에 대한 사전적 고려가 필요함

나. 가명처리 및 데이터 결합 시 특징

1) 개인단위의 원시자료 활용 범위 확대

■ 개인정보를 가명처리하는 경우 기존의 익명의 집계데이터로 생산했던 데이터도 개인단위의 원시자료 형태로 활용할 수 있게 된다. 이 경우 총량으로만 생산되던 통계에 개인의 특성을 고려한 미시적 접근이 가능함

2) 시의성 높은 대규모 패널 데이터 구축

■ 개인정보보호법 개정 전, 익명화된 집계 데이터로만 활용되던 민간데이터를 개인단위 데이터로 활용할 수 있게 되는 경우 많은 표본 수 확보와 더불어 시의성 높은 대규모 개인 패널 데이터를 구축할 수 있게 됨

3) 분석 가능한 유효표본 수 확대

■ 현재 개인단위로 활용할 수 있는 데이터 대부분이 조사 통계 자료인데, 사실상 조사를 통해 많은 표본을 확보하기 위해서는 많은 예산과 시간이 소요됨

■ 하지만 개인정보보호법의 개정으로 데이터를 가명처리하고 이를 연구목적에 활용할 수 있게 됨에 따라 기존의 다양한 민간 데이터를 익명화된 집계 데이터가 아닌 개인단위의 원시데이터를 대상으로 분석할 수 있어 대규모 표본 데이터를 확보할 수 있게 됨

- 이에 개인의 세부특성별 분석 뿐 아니라 소지역 단위의 개인별 평균 통계를 산출할 수 있어 통계 활용성이 높아질 것으로 기대됨

4) 이종 데이터 결합을 통한 활용 정보 확대

- 가명정보 및 결합제도의 도입으로 이종 산업분야에서 관리되고 있는 다양한 형태의 데이터를 서로 결합하여 분석 할 수 있게 됨

- 결합합분석은 가명정보 도입으로 가장 기대되는 이점 중 하나로 데이터의 가치가 더욱 높아지고 결합정보를 활용해 기존에는 할 수 없었던 연구를 수행하여 새로운 지식과 가치를 발견하여 새로운 정책 수립에 기여할 수 있을 것으로 예상됨

〈표 4〉 가명정보 도입 전후 비교 및 기대효과

구분	가명정보 도입 전	가명정보 도임 후	기대효과
개인 단위 원시자료 활용	· 기존 익명화된 집계 데이터를 활용하여 총량 위주의 추세 분석 · 각 특성별 규모를 파악할 수 없어 세부 특성별 비교 분석이 어려움	· 개인단위 원시자료로 활용할 수 있게 됨에 따라 1인당 평균 치 통계를 산출 · 이를 바탕으로 세부 특성별 통계치를 비교 분석할 수 있게 됨	**개인단위 통계 생산을 통해 보다 세부적인 변화와 추세를 파악할 수 있어 개인 특성을 고려한 정책 수립 관련 시사점 제공**
시의성 높은 대규모 패널 데이터 구축	· 조사기반 패널 데이터는 자료 수집 및 공표에 시간 소요 多 · 표본의 규모가 크지 않아 패널 이탈 시 종단 분석 시 어려움 有	· 민간데이터 내 고객의 시계열 정보를 활용, 대규모 패널 데이터 구축 가능 · 민간데이터의 경우 시의성이 높아 시의성 있는 패널자료 확보 가능	**패널데이터 분석을 활용 사회변화에 대응한 적시성 높은 정책적 대응 가능**
분석 가능한 유효표본 수 확대	· 조사 데이터는 세부특성별 혹은 소지역 단위 통계 생산을 위한 유효표본 확보가 어려움 · 민간데이터의 경우, 개인 단위 원시자료 활용 불가능	· 개인 단위 원시자료 형태인 대규모 표본의 민간데이터 확보 가능 · 개인의 다양한 특성을 고려한 분석이 가능 · 시군구 또는 읍면도 단위 통계생산 가능	**세부적인 특성 및 소지역 통계 산출을 통해 맞춤형 정책 수립 시 근거 자료로 활용**
이종 데이터 결합을 통한 활용정보 확대	· 개인정보보호를 문제로 개인단위 정보 결합 제한 · 서로 다른 자료의 유사한 개체를 결합하는 통계적 연계만 가능	· 데이터 결합해 다양한 정보를 동시에 활용함으로써 다양한 시사점 도출 가능 · 가명처리를 통해 정확연계가 가능하여 정확한 데이터 확보	**이종 데이터 결합을 통해 기존 데이터로 볼 수 없는 다양한 정책 시사점 제시**

다. 가명처리 및 데이터 결합 시 고려사항

- 가명정보를 안전하게 활용하기 위해서는 가명정보처리 시 요구하는 법적 준수사항을 충분히 숙지하여 동의 없는 가명정보의 처리(개인정보보호법 제28조의2) 과정에서 개인정보 오·남용을 방지하고 관련 법령을 준수할 수 있도록 주의해야 함

- 또한 데이터 결합 시 발생될 수 있는 다양한 문제에 대해서도 심도있는 검토가 필요하다. 마지막으로 가명 데이터 활용 및 데이터 결합 후 분석 과정에서 발생되는 문제 또한 고려해야함

〈표 5〉 가명처리 데이터 활용 시 고려사항

고려사항	세부내용
분석과정	· 데이터 활용범위 축소 및 고급 통계 분석의 한계 - 가명처리 과정에서 데이터 정밀도 하락 - 이에 통계모형 활용 시, 결과의 신뢰도 측면에서 신중한 접근 필요 · 데이터 분석 환경 구성의 어려움 - 가명처리 및 분석 시스템은 기존 시스템과 물리적 분리가 필요 · 분석결과 활용 - 데이터 용량상의 문제와, 개인식별 가능성을 줄이기 위해 통계, 집계 형태로 활용 · 표본의 대표성 문제 - 민간 데이터 가명처리 시, 전 국민 대상의 표본 확보 여전히 불가능
데이터 결합	· 결합률 - 결합률을 높이기 위해서 최신 정보로 사전 업데이트 필요 · 결합기관간 의사소통 - 데이터 결합 시 발생할 수 있는 문제점 미리 인지하는 것이 필요 · 결합신청자들 간 다른 데이터 활용정책 - 결합신청자들은 각자의 업종 상황이나 정책을 이해 요구 · 복잡한 데이터 결합절차 - 데이터 결합 시, 절차가 복잡하고 처리 시간이 소요됨 - 이에 결합신청자간 적극적 의사소통 및 합의가 필요 · 데이터 신뢰성 문제 - 결합 시, 동일한 정보가 일치하지 않는 경우 발생 - 데이터의 신뢰도와 오차율을 바탕으로 신뢰도가 높은 데이터 선택 필요

3. 가명처리 데이터 활용 현황분석 및 결합사례 검토

가. 가명정보 활용 데이터 현황

1) 공공데이터 현황

■ 2021년 5월 기준, 문화관광분야의 총 23,313개의 공공데이터가 공개되어 있어 파일 다운로드 또는 오픈API(Open API)를 통해 이용할 수 있음

- 현재 해당 포털에서 개인정보를 가명처리하여 데이터를 제공하고 있지는 않다. 하지만 개인 정보를 활용한 익명의 집계 데이터를 제공하고 있어, 향후 연구 목적에 따라서 집계 데이터 수집 전 개인 단위의 원시자료 활용도 가능할 것으로 판단됨
- 다만, 해당 개인정보를 가명처리한 데이터의 형태로 활용하기 위해서는 데이터 제공기관과의 업무협의를 기반으로 이용기관으로서 가명정보 활용과 관련된 제도적, 법률적 근거를 갖추어야 할 것임

〈표 6〉 문화체육관광 분야 공공데이터 포털 개방 현황

(단위: 개)

분류명	합계
문화예술	5,099
문화재	10,584
체육	1,750
관광	5,880
문화체육관광 일반	3,666
합계(문화체육관광)	23,313

출처: 송철재(2021)

■ 공공데이터 개방과 활용 지원 노력의 일환으로 문화·관광 분야에서 운영되고 있는 주요 데이터 개방플랫폼은 문화센터, 관광지식정보시스템, 문화예술지식정보시스템, 문화 빅데이터 플랫폼, 문화데이터 광장, 한국관광데이터랩 등이 있음

- 해당 플랫폼에서 제공하고 있는 데이터는 개방플랫폼에서 자유롭게 다운로드할 수 있는 익명정보가 대다수를 차지하고 있으며, 개별 협의를 통한 가명정보의 이용이 가능한 것으로 보이나 아직은 활용사례가 거의 없는 것으로 보임

- 향후 가명정보의 유통이 지금보다 활성화 된다면 위와 같은 플랫폼의 역할이 점점 더 중요해지고 플랫폼을 통한 가명정보의 활용이 빠르게 확산될 것으로 예상됨

〈표 7〉 문화·관광 분야 공공데이터 플랫폼

운영기관	플랫폼명
문화체육관광부	통합문화이용권정보시스템, 빅데이터 기반 언어 말뭉치시스템, 도서관 빅데이터플랫폼, 사서의사결정시스템, 예술인경력증명시스템, 국제문화교류정보시스템, 지역문화 통합정보시스템, 디지털국악아카이브시스템, 현대사디지털아카이브시스템, 국립중앙극장, 공연예술자료관리시스템, 불법온라인도박관리시스템, 정책여론수렴시스템, 콘텐츠수출마케팅플랫폼
한국문화관광연구원	관광지식정보시스템, 문화셈터, 문화예술지식정보시스템
한국관광공사	TourAPI 3.0, 한국관광데이터랩
한국문화정보원	문화 공공데이터 통합활용 시스템, 문화 빅데이터 플랫폼
한국문화예술위원회	문화예술 데이터관리시스템

■ 문화체육관광부는 행정업무 기반의 다양한 수집정보를 정보시스템을 통해서 관리하고 있으며, 현재 문화체육관광부의 정보시스템을 통해 수집되는 공공데이터는 총 331개로 파악됨

- 〈표 8〉는 문화체육관광부의 정보시스템을 통해 수집되는 공공데이터는 총 331개중 주요 문화·관광 관련 개인정보 데이터 현황을 간략히 제시함

〈표 8〉 문화·관광 관련 개인정보 데이터 현황

구분	제공 내용	제공기관
문화누리카드 DB	o (주요내용) 기초생활수급자, 차상위 계층을 위한 문화누리카드 발급 및 이용 실적 관리 o (개인정보) 이름, 집주소, 이메일, 집 연락처, 핸드폰, 주민번호, 외국인 등록번호, (※ (필수) 문화누리카드번호, 문화누리카드비밀번호, CI/DI)	한국문화예술위원회
예술의 전당 무료 및 유료 회원	o (주요내용) 예술의전당 공연 구매 이력, 그 외 서비스 제공을 위한 기타 개인 정보 o (개인정보) 이름, 집주소, 이메일, 핸드폰, 생년월일	예술의전당
카지노 이용 고객 정보	o (주요내용) 카지노 이용고객의 등록, 출입 이력관리, 영업활동에 이용 o (개인정보) 이름, 집주소, 핸드폰, 생년월일, 여권번호, 외국인등록번호, 기타(국적, 직업)	그랜드코리아레저(주)
골프장 홀인원 시스템 예약 정보	o (주요내용) 골프장 이용 고객의 등록, 출입, 이력 관리 o (개인정보) 이름:필수, 핸드폰(연락처):필수	한국관광공사

구분	제공 내용	제공기관
박물관 교육프로그램 운영 및 참여	o (주요내용) 교육 프로그램 참여 인원 통계 분석 자료 활용 o (개인정보) 이름:필수, 집주소:필수, E-Mail:필수, 핸드폰(연락처):필수	국립 중앙박물관
공공도서 책이음 시스템	o (주요내용) 책이음서비스 이용자 서비스 제공 o (개인정보) 이름:필수, 집주소:필수, E-Mail:필수, 핸드폰	국립중앙도서관
예술인경력정보시스템	o (주요내용) 예술인의 경력 정보 시스템, o (개인정보) 이름:필수, 집주소:필수, 직장주소:필수, E-Mail:필수, 집연락처:필수, 핸드폰(연락처):필수, 주민번호:필수	한국예술인복지재단

■ 그 외 기타 행정 업무를 통해 생성 혹은 취득한 데이터도 존재함

- 문화체육관광부를 제외한 소속 및 산하기관 48개 중 21개 기관을 대상으로 데이터 전수조사를 실시한 결과 총 452개의 데이터를 보유 6)

〈표 9〉 문화·관광 분야 공공데이터 플랫폼

(단위: 개, %)

운영기관	데이터 수
문화예술	345(100.0)
체육	40(100.0)
관광	45(100.0)
문화관광 일반	22(100.0)

출처: 송철재(2021)

2) 민간데이터 현황

■ 민간 분야에서 문화·관광 관련 연구를 위한 접근 가능한 데이터는 주로 개인의 소비 패턴 및 행동 특성을 연구하기에 유용한 데이터를 보유한 이동통신사·신용카드·유통사의 데이터로 구분할 수 있음

〈표 10〉 문화·체육·관광 관련 민간 분야 데이터 현황

구분	제공 내용	제공기관
이동통신 데이터	실거주지, 지점별 체류시간, 빈도 등	SK텔레콤, KT
신용카드 데이터	업종별, 개인특성별, 지역별 카드 지출액 제공	신한, BC, 국민
유통사 데이터	연령·지역별 구매내역 등	GS 리테일 등

6) 문화체육관광부 소속 및 산하기관을 대상으로 수행된 데이터 보유 전수조사 결과(송철재, 2021)

3) 조사데이터 현황

■ 조사 데이터 문화체육관광부에서는 문화, 체육, 관광 관련 정책에 대한 연구를 위해 주제별로 다양한 조사 목적의 조사데이터를 관리하고 있음

〈표 11〉 문화·관광 관련 개인단위 조사 데이터 현황

구분	조사 목적	주요 지표
국민문화 예술활동조사	o 우리나라 국민의 문화 활동 향유의 필요성 및 인식이 높아짐에 따라 실태 파악을 위한 문화향유 경로와 방식에 대하여 통계적으로 분석하여 궁극적으로 국민 문화향유 진흥 도모	- 문화예술행사 관람 및 관람 의향 - 문화예술행사 매체 이용 실태 및 참여 활동 - 문화예술교육 경험 및 의향 - 문화예술 공간이용 실태 및 방문의향 - 문화관련활동
국민여행조사	o 우리나라 국민의 여행실태를 종합적으로 파악, 국가 관광에 관한 정책수립과 연구·분석 등을 위한 기초 자료를 제공	- 여행목적(지) - 여행지출액 - 여행소감 - 기타 여행 문항
국민여가 활동조사	o 국민들이 여가를 어떻게 인식하고, 여가생활을 하고 있는지를 조사, 생활양식의 변화 및 삶의 질적 수준을 파악하여 정부의 여가정책 수립을 위한 기초자료 제공	- 여가활동 참여실태 - 사회성 여가활동 - 동호회 활동 - 휴가, 연휴 활용 - 여가공간 - 여가자원 활용실태 - 여가생활 만족도
주요관광지점 입장객통계	o 주요 관광지점 이용객 통계의 생산·배포를 통한 관광객 수요 추정 및 관광시설 공급판단의 기초자료로 활용	- 국내·외 입장객 및 시설이용객 현황(유료관광지) - 방문객 현황(무료관광지)
외래관광객조사	o 방한 외래관광객의 여행성향 및 실태의 변화추이를 정기적으로 조사·비교·분석함으로써 외래관광객 유치 증대 및 수용태세 개선을 위한 관광정책 수립의 기초자료로 활용	- 한국여행실태(방문 횟수, 시기, 방문 목적 등) - 한국여행 소비 실태(지출 경비, 쇼핑 품목, 쇼핑 장소) - 한국여행 평가(한국 여행에 대한 만족도 등)
근로자휴가조사	o 국내에서 산업 활동을 영속중인 사업체와 근로자를 대상으로 휴가실태조사를 실시하여 관련분야 정책수립시 합리적 의사결정 지표로 활용될 신뢰할 수 있는 통계자료 제공	- 연차휴가일수(연차휴가 소진율) - 휴가사용 환경 - 휴가 만족도 등

나. 가명처리 데이터 활용사례

1) 해외사례

■ 주요국의 가명정보는 주로 데이터를 가명처리 한 후 결합하기 위한 사례들 위주로 진행

〈표 12〉 해외 가명정보 활용 사례 요약

국 가	활용분야	결합 대상 데이터	결합데이터 이용기관
미 국	o 사고 유형별 안전 장치 영향도 분석	(보험사) 차량 사고처리 정보 (제조사) 차량별 안전장치정보	o 보험사: 보험료 할인상품 개발 o 제조사: 안전장치 기능 개선
영 국	o 소셜데이터와 주가지수 상관관계 분석	(포털사) 소셜데이터(비정형) (증권사) 주가정보	o 투자은행: 주가예측 로보어드바이저 개발
	o 의료분야 분석 연구	전자의무기록과 보건의료 행정데이터(1차의료, 2차의료, 질병레지스트리)를 연계한 'CALIBER'이라는 데이터 플랫폼을 구축	o Farr Institute
캐나다	o 재정, 사회, 경제 관련 연구	(의료) 캐나다 암등록 자료 전국 인구 건강조사 (경제) 노동 및 소득 동태조사 (교육) 중고등학생 정보 시스템 등 행정 및 조사 데이터를 연계하여 데이터연계 환경(SDLE) 프로세스 구축	o 캐나다 통계청 (데이터연계 환경(SDLE) 프로세스)
	o 건강, 웰빙 관련 연구(교육, 데이터 연계 등 지원)	(의료) 의약보험, 의료서비스계획 지불 정보 등 (인구) 출생, 사망, 결혼 통계 등 인구 통태 통계, 영주권자, 소득구관 등 행활과정 통계 (경제) 직업 정보 그 외 교육, 환경 관련 데이터 이용한 연계 데이터 제공	o 브리티시컬럼미아주 인구 데이터(PopData BC)
캐나다	o 지역 기반의 건강 서비스 분석, 보건정책, 데이터 과학 등 연구	(보건) 입원환자 병원 기록, 전자의무 기록, 수술 기록 등 (시설) 보건서비스 기관정보 (금융) 의료서비스 비용 (인구) 난민 및 이민 상태, 인구조사 프로필, 인구 추정 및 추계 그 외 기타 비보건 부문 데이터를 이용하여 연계	o 온타리오주 임상평가과학연구소

자료: 1. 캐나다 통계청(Statistics Canada) 홈페이지의 사회적 데이터연계 환경(SDLE) 내용 요약 (https://www.statcan.gc.ca/eng/sdle/overview/)
2. 자료: Ark et al. (2019)
3. 자료: Schull et al. (2019)

2) 국내사례

■ 가명정보 활용의 기반이 마련됨에 따라 다양한 분야의 공공데이터 활용 가명정보 결합사례가 진행되고 있음

〈표 13〉 공공데이터를 활용한 가명정보 시범사례

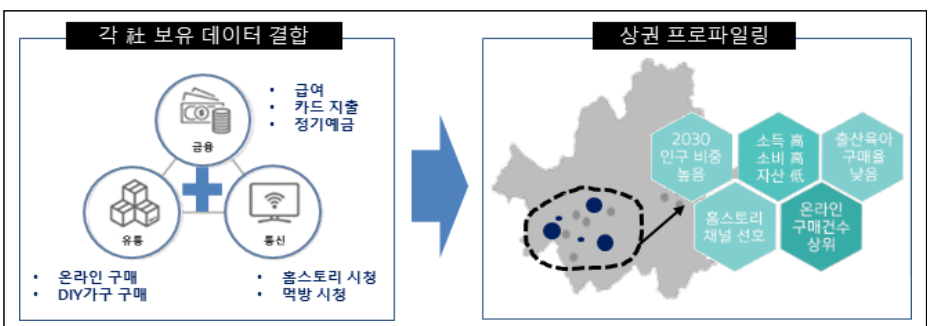
분야	시범사례	이용기관	결합전문기관
의료+인구	① 암 질병 치료효과 분석	국립암센터	통계청
	② 암 환자의 합병증 및 만성질환 예측 연구	국립암센터	건보공단
금융+보훈	③ 국가보훈대상자 신용실태 연구	국가보훈처	신용정보원
소득+복지	④ 노후소득보장 종합연구	사회보장위원회	국세청
통신+유통	⑤ 불법스팸 실태연구	인터넷진흥원	삼성SDS
레저+건강	⑥ 맞춤형 산림치유 프로그램 분석	임업진흥원	건보공단

■ 민간분야에서는 데이터 활용 수요를 일찍이 인식하고 데이터3법이 시행됨에 따라 상호간의 협력을 통해 데이터 결합 사례를 추진해왔음

① 카드 이용정보 ✚ 기지국 접속 정보 ➡ 여행/관광 정보 분석(신한카드, SK텔레콤)

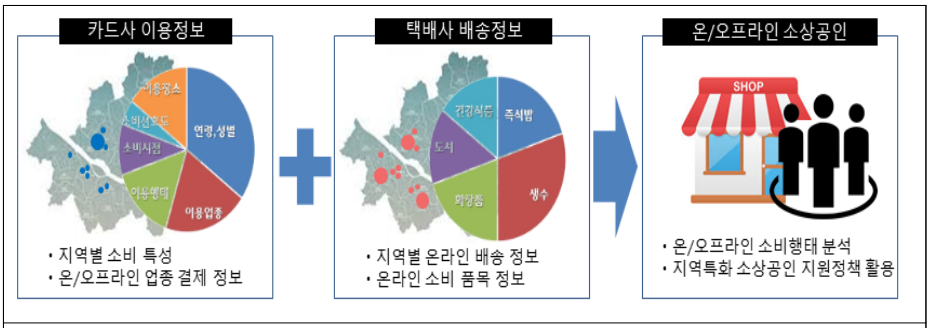
◈ 카드사 보유 카드 이용정보와 통신사의 고객 기지국 접속 정보를 결합하여 여행/관광 정보 분석 ➡ (공공기관) 여행/관광 정책수립, (기업 등) 고객 특성별 선호 여행지 정보 제공

② 소득/소비/자산 정보 ✚ 온라인 채널 택배정보 ✚ IPTV 시청정보 ➡ 상권별 소비행태 분석



◈ 은행의 소득/소비/자산 정보와 택배사의 택배정보, 통신사의 IPTV 시청정보를 결합하여 상권별 거주자 소비행태를 분석 ➡ (소상공인) 주거지 인근 상권 마케팅 전략 수립, (공공기관) 정책 수립 및 행정 서비스 제공

③ 카드 이용정보 ✚ 택배 정보 ➡ 소비행태 분석(KB카드, CJ올리브네트웍스)



◈ 카드 이용 정보와 택배사의 온라인 채널 택배정보를 결합하여 온·오프라인 소비행태 분석 ➡ (공공기관 및 지자체) 온·오프라인 소상공인 지원정책 수립 활용

3) 조사데이터 활용

■ 온라인 시장조사업체를 중심으로 자체 조사데이터 간 익명데이터 결합 활용을 시도하고 있음 온라인 시장조사업체 마이크로밀 엠브레인은 인구통계학적 기준에 맞춰 추출한 샘플에 대한 App 이용정보, 오프라인 방문정보, 카드 결제정보를 각각 수집하여 다음과 같이 제공하고 있음

〈표 14〉 온라인 시장조사업체 결합 데이터 정보

구분	칼럼
APP 이용정보	사용일자, App 카테고리, App 이름, 총 사용시간
오프라인 방문정보	매장방문일자, 매장카테고리, 매장 층수
카드결제 정보	문자수신날짜, 문자수신시간, 결제타입, 결제형태, 사용처 카테고리, 총 결제금액, 실 결제금액 등

자료: 금융데이터거래소

다. 소결

■ 본 연구는 가명처리 데이터를 활용하여 실증 분석을 실시하여 (ⅰ)기존 데이터의 활용범위 확대됨을 보이고 (ⅱ) 가명정보를 이용한 데이터 결합을 실시하여, 데이터의 문화관광 부문의 새로운 정책적 가치를 도출하는 것이 주요 목적

■ 이러한 연구 목적을 달성하기 위하여 각 데이터의 현황을 검토하고, 데이터 선정 기준을 설정하여 최종적으로 ㈜SK텔레콤, ㈜신한카드 데이터를 선정함

〈표 15〉 가명처리 데이터 활용방안 도출을 위한 데이터 선정

공공데이터	민간데이터	조사데이터
(+) 정부의 정책을 평가할 수 있는 유의미한 변수 多 (+) 활용 가능한 정보 확인을 위한 접근성 용이 (-) 가명정보 반출 및 제공이 활발하지 않아 이용의 제약 (-) 데이터 결합 시 충분한 유효표본 확보가 어려움	(+) 관련 분석 및 서비스 제공 활발, 상대적으로 적은 시간 소요 (+) 충분한 유효표본 확보와 적시성 있는 통계 생산 가능 (+) 대규모의 패널 데이터 구성이 가능 (-) 보안 규정으로 인하여 활용 가능한 컬럼 확보 어려움 (-) 추가적인 데이터 및 변수 가공 필요	(+) 정부의 정책을 평가할 수 있는 유의미한 변수 多 (+) 활용 가능한 정보 확인을 위한 접근성 용이 (-) 적시성 있는 통계 생산 어려움 (-) 데이터의 결합 시 추가적인 개인정보 수집 및 활용 재동의필요

고려사항 1: 기 생산된 통계 및 연구와 차별화된 정책적 가치 도출 가능성 검토

⇩

고려사항 2: 연구기간, 연구 예산, 접근가능성 등을 고려한 실제 활용성 점검

⇩

㈜SK텔레콤 이동통신 데이터, ㈜신한카드 카드결제액 데이터 선정

4. [가명처리 데이터] 문화·관광 부문 활용방안

가. 데이터 구성 절차도

- 가명처리 데이터를 활용하기 위해서는 기존과 달리 다양한 법률적, 행정적, 실무적인 절차들이 요구됨
- 이에 본 연구에서는 가명처리 데이터를 구성하기 위 한 업무를 크게 분석 환경구성, 데이터 구성, 반출 및 분석단계로 구분하여 진행하였으며, 이는 다음 〈표 16〉과 같이 도식화 할 수 있음
  - 여기서 가명처리 데이터 분석을 위해서 SK텔레콤과 신한카드사는 결합 대상 데이터를 제공하는 '제공기관', 연구원은 결합된 데이터를 이용하는 '이용기관'임

〈표 16〉 가명처리 데이터 활용을 위한 업무 절차도

구분	주체
1단계: 데이터 활용 관련 업무 협의 데이터의 활용을 위한 데이터 제공기관과 이용기관 간, 업무 협의	제공기관, 이용기관
⇩	
2단계: 가명처리 데이터 분석환경 조성 가명정보 활용 관련 내부관리계획 수립 및 점검 가명정보 활용 분석 환경 점검	이용기관
⇩	
3단계: 가명처리 데이터 구성 (*상향식 접근(bottom-up)방식) 가명처리 데이터 활용주제 선정(이용기관) 활용 가능한 컬럼 보유 여부 확인(이용기관, 제공기관) 선정된 컬럼의 활용 관련 범무적 검토(제공기관) 최종컬럼 확정 후 가명 및 익명처리(제공기관)	제공기관, 이용기관
⇩	
4단계: 가명처리 활용 및 반출 적정성 평가 반출 신청서 작성(이용기관) 가명처리 데이터 반출 및 활용 적정성 평가위원회 개최(제공기관) 비밀보호협약서 작성(이용기관, 제공기관)	제공기관, 이용기관
⇩	
5단계: 데이터 반출 및 문석 개인정보 보호를 위해 암호화 전송(제공기관) 지정된 분석 환경 내 업무 진행(이용기관)	이용기관

나. 최종 가명처리 데이터 구성

1) 컬럼 선정

- 가명처리 데이터의 관광분야의 활용방안을 도출할 수 있는 적합한 컬럼을 선정하고자 다음 사항을 고려하였음
- **(고려사항 1)** 관광 및 문화콘텐츠 분야의 시사점 도출에 적합한 컬럼 선정
  - 전체 보유 컬럼을 반출할 경우 분석 상의 어려움 뿐만 아니라 개인 식별의 가능성이 커지므로 관광 및 문화콘텐츠 분야의 유의미한 시사점을 보여줄 수 있다고 판단되는 컬럼 위주로 구성함
- **(고려사항 2)** 충분한 시계열 자료 확보가 가능한 컬럼 선정

- 관광 및 문화콘텐츠 분야의 변화를 살펴보고 향후를 진단하기 위하여 충분한 시계열 자료 확보가 가능한지 여부를 고려함

■ **(고려사항 3)** 개인정보 침해하지 않는 수준의 컬럼 선정

- 관광 및 문화콘텐츠 분야 시사점 도출에 있어 유용한 정보로 판단된다 하더라도 개인정보를 침해한다고 판단되는 경우 이를 활용하지 않는 것을 원칙으로 함

① 통신데이터 컬럼

■ 통신 데이터는 다음과 같이 구성함

〈표 17〉 [가명처리 데이터] 통신 데이터 최종 컬럼 구성

컬럼	기간	세부내용
ⅰ. 개인 특성		
성별	'19.1~'21.6	시군구 코드
연령대	'19.1~'21.6	1인 가구, 2인 가구, 3인 가구, 4인 가구, 5인 가구 이상
세대구분 1	'19.1~'21.6	M세대(1981~96년생), Z세대(1997~2010년생), 그 외 세대
세대구분 2	'19.1~'21.6	청년(18~34세), 중년(35~49세), 장년(50~64세), 노년(65세 이상)
가구원수	'20.1~'21.1	1인 가구, 2인 가구, 3인 가구, 4인 가구, 5인 가구 이상
거주지역	'19.1~'21.6	시군구 코드
연간소득 ** 8개 소득 구간*	'19.1~'21.6	1억이상, 7천-1억미만, 5천-7천미만, 4천-5천미만 3천-4천미만, 2천-3천미만, 1천-2천미만, 1천 미만
ⅱ. 관광 부문: 이동횟수		
평일 이동 횟수	'19.8~'21.6	평일 중에서 법정공휴일을 제외한 날 한 달 간 평일 총 이동 횟수
휴일 이동 횟수	'19.8~'21.6	2개 요일(토일) 또는 법정공휴일 한 달간 휴일 총 이동횟수
ⅲ. 문화 부문: 온라인 콘텐츠 이용		
평일 온라인콘텐츠이용량	'19.8~'21.6	전체, 동영상, 음악, 게임, 도서, 웹툰, 커뮤니티 등
휴일 온라인콘텐츠이용량	'19.8~'21.6	전체, 동영상, 음악, 게임, 도서, 웹툰, 커뮤니티 등
ⅲ. 문화 관련 온라인 콘텐츠 관련:		
구독서비스 정보	'19.8~'21.6	구독서비스 명
구독기간 이용 기간	'19.8~'21.6	개월

② 카드데이터 컬럼

■ 관광분야 실증분석을 위한 주요 컬럼으로 관광 관련 카드 지출액으로 선정함

- 이때 관광은 다음 〈표 18〉의 14개 업종의 지출액을 관광 관련 카드지출액으로 정의함[7)]

〈표 18〉 카드데이터 오프라인 콘텐츠 분야 업종 분류 정의

부 문	세부 분류
관광 부문	01. 여행업 02.관광호텔(4,5성급) 03.관광호텔(3성급) 04.콘도미니엄 05.일반숙박업 06.카지노 07.유원시설업 08.면세점 09.항공사 10.대중교통 11.렌터카 12.관광기념품판매업 13. 음식점업 14.레저스포츠 체험업

■ 문화콘텐츠 분야 실증분석을 위한 주요 컬럼으로는 온오프라인 콘텐츠 부문 소비지출액을 선정함

- 다음 〈표 19〉의 업종 지출액을 오프라인 콘텐츠 부문 소비지출액으로 정의함

〈표 19〉 카드데이터 오프라인 콘텐츠 분야 업종 분류 정의

부 문	세부 분류
오프라인 콘텐츠	01. 인쇄,출판 02. 공연장,극장 03. 음반테이프 04. 서적 05. 전자오락실 06. PC게임방 07. 노래방 08. 수련원,체험장 09. 실내골프장 10. 비디오방/전화방

- 다음 〈표 20〉에 해당하는 사업자로부터 발생한 지출액을 온라인 콘텐츠 부문 소비지출액으로 정의하였음

〈표 20〉 카드데이터 온라인 콘텐츠 분야 해당 사업자 분류

부 문	세부 분류
01. 영상	해당 분야 지출액
02. 음악	해당 분야 지출액
03. 게임	해당 분야 지출액
04. 출판	해당 분야 지출액

■ 그 외 관광 및 온라인 콘텐츠 활동에 영향을 줄 수 있는 다양한 개인적 특성을 포함하여 다음 〈표 21〉과 같이 통신 데이터 최종 컬럼을 구성함

7) 해당 분류는 현재 연구원에서 신한카드 데이터를 활용하여 매월 생산하고 있는 「COVID19 문화관광콘텐츠 영향」과 「관광레저소비지출경제동향」의 관광업종 분류를 적용하였다.

〈표 21〉 [가명처리 데이터] 카드 데이터 최종 컬럼 구성

컬럼	기간	세부내용
ⅰ. 개인 특성		
성별	'19.1~'21.6	시군구 코드
연령대	'19.1~'21.6	1인 가구, 2인 가구, 3인 가구, 4인 가구, 5인 가구 이상
세대구분 1	'19.1~'21.6	M세대(1981~96년생), Z세대(1997~2010년생), 그 외 세대
세대구분 2	'19.1~'21.6	청년(18~34세), 중년(35~49세), 장년(50~64세), 노년(65세 이상)
가구원수	'20.1~'21.1	1인 가구, 2인 가구, 3인 가구, 4인 가구, 5인 가구 이상
거주지역	'19.1~'21.6	시군구 코드
연간소득 ** 8개 소득 구간*	'19.1~'21.6	1억이상, 7천-1억미만, 5천-7천미만, 4천-5천미만 3천-4천미만, 2천-3천미만, 1천-2천미만, 1천 미만
ⅱ. 오프라인 카드 지출액		
전체 월별	'19.1~'21.6	전체 발생한 신용카드 지출의 월별 총 금액(단위: 원)
관광 월별	'19.1~'21.6	관광 분야의 신용카드 지출의 월별 총 금액(단위: 원)
콘텐츠 월별	'19.1~'21.6	오프라인콘텐츠 분야의 신용카드 지출의 월별 총 금액(단위: 원)
ⅲ. 온라인 콘텐츠 지출액		
영상 월별 카드 지출	'19.1~'21.6	영상관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)
음악 월별 카드 지출	'19.1~'21.6	음악관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)
게임 월별 카드 지출	'19.1~'21.6	게임관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)
출판 월별 카드 지출	'19.1~'21.6	출판관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)

2) 데이터 구성

① 통신데이터 구성

- 통신데이터의 총 표본수는 분석대상 기간 동안 SK텔레콤 가입자 24,240,343명이며 데이터 수집 기간은 2019년 8월 ~ 2021년 6월까지임

〈표 22〉 SK텔레콤 가명처리 데이터 정보

부 문	세부 분류
총 표본수	24,240,343명 *분석대상 기간 동안 SK텔레콤 가입자
활용 컬럼 수	616개, *개인 정보, 이동횟수, 구독서비스 기간 등을 월별로 구성
분석대상 기간	2019년 8월 ~ 2021년 6월 *활용 컬럼 중 가장 오래된 정보와 최신의 정보를 반출 여부를 설정

- 수집된 데이터는 가명처리 후 다음 〈표 21〉과 같은 패널 데이터 구조로 변환함

〈표 23〉 SK텔레콤 통신 데이터 최종 구성 - 가명처리 및 패널 데이터 구조변환

(가명처리 전) 개인정보를 포함한 SK텔레콤 시계열 데이터(예시)

(대상: SK텔레콤 가입자)

ID	20년 6월 연령	20년 7월 연령	…	21년 5월 연령	…	20년 6월 이동량	20년 7월 이동	…	21년 5월 이동량
김관광	29세	29세	…	30세	…	10회	24회	…	36회
이문화	41세	42세	…	42세	…	140회	150회	…	170회

(가명처리 후) 개인정보를 가명처리한 후 SK텔레콤 패널데이터 구조화(예시)

(대상: SK텔레콤 가입자)

ID	Time	성별	연령대	이동량	구독서비스 이용우뮤
A	19년 7월	여성	20대 이하	10회	이용함
A	19년 8월	여성	20대 이하	24회	이용함
……	……	……	……	……	……
A	21년 5월	여성	30대	36회	이용함
B	19년 7월	남성	40대	24회	이용함
B	19년 8월	남성	40대	19회	이용안함
……	……	……	……	……	……
B	21년 5월	남성	40대	15회	이용안함

② 카드데이터 구성

- 카드데이터의 총 표본수는 분석대상 기간 동안 신한카드 가입자 8,453,672명 이며 데이터 수집 기간은 2019년 8월 ~ 2021년 6월까지임

〈표 24〉 신한카드 가명처리 데이터 정보

부 문	세부 분류
총 표본수	8,453,672명 *분석대상 기간 동안 신한카드 가입자
활용 컬럼 수	633개 *개인 정보, 소비지출액 월별 컬럼 생성
분석대상 기간	2019년 8월 ~ 2021년 6월 *활용 컬럼 중 가장 오래된 정보와 최신의 정보를 반출 여부를 설정

- 카드 데이터 또한 수집된 데이터는 가명처리 후 다음 〈표 25〉과 같은 패널 데이터 구조로 변환함
  - 가명처리 과정에서 지출액이 큰 경우 분석에서 제외하여 가명처리함

〈표 25〉 신한카드 통신 데이터 최종 구성 - 가명처리 및 구조변환

(가명처리 전) 개인정보를 포함한 신한카드 시계열 데이터(예시)

(대상: SK텔레콤 가입자)

ID	20년 6월 연령	20년 7월 연령	…	21년 5월 연령	…	20년 6월 관광지출	20년 7월 관광지출	…	21년 5월 이동량
김관광	29세	29세	…	30세	..	161,000	144,000	..	156,000
이문화	41세	42세	…	42세	..	13,456,000	13,586,000	..	
박정책	75세	75세	…	76세	..	56,000	78,000	..	34,000

(가명처리 후) 개인정보를 가명처리한 후 신한카드 패널데이터 구조화(예시)

(대상: SK텔레콤 가입자)

ID	Time	성별	연령대	관광 소비지출	온라인 콘텐츠
A	19년 7월	여성	20대 이하	161,000	22,000
A	19년 8월	여성	20대 이하	144,000	22,000
……	……	……	……	……	……
A	21년 5월	여성	30대	156,000	33,000
C	19년 8월	여성	70대 이상	56,000	0
……	……	……	……	……	……
C	21년 5월	여성	70대 이상	78,000	6,000
C	21년 6월	여성	70대 이상	34,000	6,000

다. 실증분석 주제 선정

■ 가명처리 데이터가 가지는 강점을 보여줄 수 있는 관광과 문화콘텐츠 분야의 분석 가능 한 주제를 검토하고, 다음 표와 같이 선정함

〈표 26〉 가명처리 데이터 문화관광 부문 활용방안 도출을 위한 실증분석 주제 선정

SK텔레콤과 신한카드 지출액 '가명처리 데이터'를 활용

개인단위 원시자료 활용	패널 데이터 구축	유효표본 수 확대
(관광부문) - 1인 평균 관광 이동행태 분석 - 1인 평균 관광 소비지출 분석 **(문화콘텐츠 부문)** - 1인 평균온라인 콘텐츠 이용율 분석 - 1인 평균 콘텐츠 소비 지출 분석	**(관광부문)** - 코로나 19 전후 관광 이동 변화 분석 **(문화콘텐츠 부문)** - 코로나 19 전후 콘텐츠 및 구독 서비스 이용 변화 분석	**(관광부문)** - 세부특성별 관광행태 분석 **(문화콘텐츠 부문)** - 세부 특성별 온라인 콘텐츠 및 구독스비스 이용 행태 분석

라. 관광부문 활용방안

1) 개인단위 원시자료 활용을 통한 현황 분석

- 기존의 익명 데이터의 경우 유효표본수나 각 세부 특성별 표본을 파악할 수 없기 때문에 1인당 평균 지출액을 산출하기 어려웠음
- 이에 가명처리 데이터를 통해서 개인단위의 원시데이터를 확보함에 따라 1인당 소비지출액을 분석할 수 있는 기반이 마련되어 1인당 평균적인 관광소비지출액을 파악할 수 있게 됨

〈표 27〉 개인단위 원시 자료 활용을 통한 관광 이동량 분석 결과 비교

(익명집계데이터 활용 시) 한달 간 이동총량

(대상: SK텔레콤 가입자, 단위: 전체/백만 회)

구분		1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
전체	'19	-	-	-	-	-	-	-	263.3	263.3	351.0	272.6	333.4
	'20	344.6	238.3	150.4	282.3	384.5	216.6	252.2	268.6	114.2	346.8	243.5	206.6
	'21	203.0	255.4	236.0	218.8	419.6	275.9	-	-	-	-	-	-

자료: SK텔레콤 익명집계데이터

(가명처리데이터 활용 시) 한달 간 이동총량의 1인당 평균

(대상: SK텔레콤 가입자, 단위: 1인/회)

구분		1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
전체	'19	-	-	-	-	-	-	-	10.9	10.9	14.5	11.2	13.8
	'20	14.2	9.8	6.2	11.6	15.9	8.9	10.4	11.1	4.7	14.3	10.0	8.5
	'21	8.4	10.5	9.7	9.0	17.3	11.4	-	-	-	-	-	-
남성	'19	-	-	-	-	-	-	-	9.9	9.9	13.2	10.2	12.6
	'20	12.9	8.2	4.9	9.8	13.8	7.7	9.1	9.6	3.8	12.6	8.7	7
	'21	7.0	9.2	8.5	7.9	15.5	10.1	-	-	-	-	-	-
여성	'19	-	-	-	-	-	-	-	11.8	11.8	15.7	12.2	14.8
	'20	15.4	11.3	7.4	13.3	17.8	10.1	11.6	12.5	5.6	15.9	11.2	9.9
	'21	9.6	11.8	10.9	10.1	19.0	12.5	-	-	-	-	-	-

자료: SK텔레콤 가명처리데이터

2) 패널데이터를 활용한 현황 변화 분석

- 가명처리 데이터가 가지는 핵심적 강점은 시의성 높은 대규모 패널데이터를 구축할 수 있다는 점임
- 다음 〈표 28〉은 코로나 19 전후 이동량 관련 분석을 통해 가명처리 데이터를 활용하는 경우 단순 추세 분석에서 확장하여 종단적 분석이 가능함을 보여주는 예임

〈표 28〉 패널화된 가명처리데이터를 활용한 관광 카드지출액 분석 결과 비교

(익명집계데이터 활용 시) 코로나 전후 한달 간 이동총량 변화 비교 분석

(대상: SK텔레콤 가입자, 단위: 백만 회, %)

구분	한달 간 이동 총량		증감률
	2019년 11월 ~ 2020년 1월	2020년 2월 ~ 2021년 4월	
전체	316.9	223.7	-29.4
남성	138.3	89.1	-35.6
여성	178.6	134.6	-24.7

자료: SK텔레콤 익명집계데이터

(가명처리 후) 코로나 전 후 집단별 이동량 감소율 비교 분석

(대상: SK텔레콤 가입자, 단위: %)

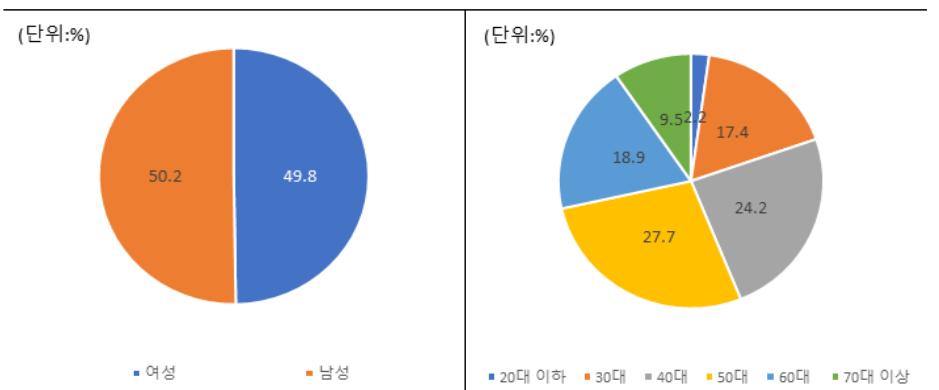
구분	(비교1) 2019년 11월 ~ 2020년 1월 vs 2020년 2월 ~ 2020년 4월 까지		(비교2) 2019년 11월 ~ 2020년 1월 vs 2020년 11월 ~ 2021년 1월 까지	
	이동량 감소	이동량 유지 / 증가	이동량 감소	이동량 유지 / 증가
전체	68.8	31.2	70.1	29.9
남성	71.3	28.7	71.3	28.7
여성	66.5	33.5	68.9	31.1

자료: SK텔레콤 가명처리데이터

- 뿐만 아니라 가명처리 데이터를 활용하게 되면 각 개인별로 코로나 19 이후 이동량이 감소했는지 유지 혹은 증가했는지 파악할 수 있음

[그림 1] 패널화된 가명처리데이터를 활용한 코로나19 이후 이동량 감소 집단의 특성 분포

(대상: SK텔레콤 가입자, 단위 %)



성별 | 연령대별

자료: SK텔레콤 가명처리데이터

3) 유효표본 수 확대에 따른 세부특성별 현황 분석

■ 개인정보를 가명처리한 데이터를 활용하는 경우 유효표본수가 확대되어 개인의 다양한 특성을 고려한 분석이 '세부특성별 분석'이 가능하게 됨

- 〈표 29〉 국민여행조사와 SK텔레콤 가명처리 데이터의 표본수를 비교해 보면 가명처리 데이터의 표본 수가 매우 크기 때문에 개인의 다양한 특성을 함께 고려할 수 있어 맞춤형 정책 수립과 관련된 자료로 활용성이 높아질 것으로 예상할 수 있음

〈표 29〉 가명처리 후 유효표본 수 확대에 따른 확보 가능한 개인 특성별 유효표본 수 비교

(단위: 명)

2020년 기준 국민여행조사 표본수		→	2020년 12월 SK텔레콤 가입자 기준	
전체 표본 수	50,710		전체 표본 수	24,240,343
남성	25,251		남성	**********
여성	25,459		여성	**********
청년층	12,782		청년층	**********
중년층	14,246		중년층	**********
장년층	13,363		장년층	**********
노년층	8,641		노년층	**********
1인 가구	7,346		1인 가구	**********
2인 가구	12,531		2인 가구	**********
3인 이상 가구	30,833		3인 이상 가구	**********

■ 이에 다음과 같은 1인 가구의 성별 X 세대별 X소득별 평균 연간 이동량와 같은 보다 세부적인 여행행태 분석이 가능해 짐

〈표 29〉 가명처리 후 유효표본 수 확대에 따른 1인 가구 특성별 여행 행태 분석

(기존 조사데이터 활용 시) 1인 가구의 성별 X 세대별 월평균 총 여행일 수

(대상 전체*, 단위: 1인/일, %)

구분		18년	20년	증감률('18년 대비)
남성	청년층	1.2	1.0	-15.2
	중년층	1.2	0.9	-29.0
	장년층	1.1	0.6	-47.8
	노년층	0.6	0.3	-43.9
여성	청년층	1.1	1.1	3.9
	중년층	1.1	0.9	-14.6
	장년층	1.0	0.5	-50.6
	노년층	0.5	0.2	-56.0

참고: 표본설계를 통해 만15세 이상 전 국민을 대상으로 함
자료: 2018-2020 국민여행조사

(가명처리데이터 활용 시) 1인 가구의 성별 X 세대별 X소득별 월평균 총 이동횟수

(대상: SK텔레콤 가입자, 단위: 1인/회)

구분		2020년 상반기			2021년 상반기			증감률('20년 상반기 대비)		
		3천미만	3-5천	5천 이상	3천미만	3-5천	5천 이상	3천미만	3-5천	5천 이상
남성	청년층	10.4	12.4	13.9	9.4	11.4	12.8	-9.8	-8.5	-7.4
	중년층	14.2	14.9	15.8	13.0	13.6	14.6	-8.3	-8.4	-7.5
	장년층	13.4	14.1	14.4	12.2	12.9	13.5	-9.1	-8.3	-6.1
	노년층	10.8	14.1	15.0	10.1	13.2	14.1	-6.0	-6.9	-5.6
여성	청년층	8.0	9.3	10.1	7.1	8.5	9.5	-10.3	-8.1	-5.6
	중년층	11.3	11.8	12.7	10.3	10.9	11.7	-9.0	-8.1	-7.7
	장년층	10.5	10.7	11.1	9.8	10.2	10.7	-6.9	-5.0	-3.3
	노년층	10.2	11.7	12.0	10.1	11.6	11.9	-1.4	-1.4	-0.6

자료: SK텔레콤 가명처리데이터

마. 문화콘텐츠 부문 활용방안

1) 개인단위 원시자료 활용을 통한 현황 분석

■ 문화콘텐츠 분야 또한 가명처리 데이터를 통해서 개인단위의 원시데이터를 확보함에 따라 1인당 평균적인 관광소비지출액을 파악할 수 있게 됨

〈표 30〉 개인단위 원시 자료 활용을 통한 온라인 콘텐츠 지출액 분석 결과 비교

(익명집계데이터 활용 시) 온라인 콘텐츠 월별 지출액 총량 추이

(대상: 신한카드 가입자, 단위: 백만원)

구분		1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
전체	'19	·	·	·	·	·	·	·	6,547	7,346	6,425	6,502	7,021
	'20	9,194	7,933	8,655	7,733	8,082	8,317	8,682	8,843	9,450	9,511	8,156	9,791
	'21	4,553	4,097	5,085	4,577	4,903	4,656	·	·	·	·	·	·

자료: 신한카드 익명데이터,

(가명처리데이터 활용 시) 온라인 콘텐츠 1인당 평균 월별 지출액

(대상: 신한카드 가입자, 단위: 1인/원)

구분		1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
전체	'19	·	·	·	·	·	·	·	1,898	2,130	1,863	1,885	2,035
	'20	2,665	2,300	2,509	2,242	2,343	2,411	2,517	2,563	2,739	2,757	2,364	2,838
	'21	1,320	1,188	1,474	1,327	1,421	1,350	·	·	·	·	·	·
남성	'19	·	·	·	·	·	·	·	2,697	3,038	2,585	2,624	2,787
	'20	3,802	3,198	3,459	3,006	3,107	3,257	3,375	3,448	3,775	3,711	3,123	3,808
	'21	1,323	1,198	1,522	1,373	1,465	1,391	·	·	·	·	·	·
여성	'19	·	·	·	·	·	·	·	1,032	1,145	1,080	1,084	1,220
	'20	1,433	1,326	1,479	1,413	1,515	1,495	1,588	1,604	1,617	1,724	1,542	1,787
	'21	1,316	1,176	1,423	1,277	1,374	1,305	·	·	·	·	·	·

자료: 신한카드 가명처리데이터,

2) 패널데이터를 활용한 현황 변화 분석

■ 문화콘텐츠 부문 또한 기존 데이터로 분석하지 못했던 개인의 변화 특성을 분석할 수 있게 됨

〈표 31〉 **(가명처리 데이터 활용 시)** 구독서비스 이용 변화 집단의 특성 분포

(대상: SK텔레콤 가입자, 단위: %)

구 분	성별		연령		가구원수		
	여성	남성	MZ	그외	1인	2인	3인
1. 코로나 19 이후 이용자 특성 : 2019년 부터 2020년 1월까지 사용 안하다가 2020년 2월부터 1회 이상 사용한 집단	24.6%	24.7%	24.4%	24.8%	22.8%	23.5%	23.9%
2. 코로나 19 이후 이용자 특성 : 2019년 부터 2020년 1월까지 사용 안하다가 2020년 2월 부터 지속적으로 사용한 사람	0.4%	0.5%	0.6%	0.4%	0.5%	0.5%	0.5%

자료: SK텔레콤 가명처리데이터

3) 유효표본 수 확대에 따른 세부특성별 현황 분석

■ 개인정보를 가명처리한 데이터를 활용하는 경우 유효표본수가 확대되어 개인의 다양한 특성을 고려한 분석이 '세부특성별 분석'이 가능하게 됨

- 〈표 32〉 미디어패널조사와 SK텔레콤 가명처리 데이터의 표본수를 비교해 보면 가명처리 데이터의 표본 수가 매우 크기 때문에 개인 세부특성 분석이 가능해 맞춤형 정책 수립의 유의미한 근거자료로 활용할 수 있을 것임

〈표 32〉 가명처리 후 유효표본 수 확대에 따른 확보 가능한 개인 특성별 유효표본 수 비교

(단위: 명)

2020년 한국미디어패널조사 표본수

전체 표본 수		10,302
성별	남성	5,146
	여성	5,156
연령	20대 이하	2,682
	30대	1,509
	40대	1,745
	50대	1,827
	60대	1,377
	70대 이상	1,162
소득	3천6백만원 미만	8,555
	3천6백만원 이상 4천8백만원 미만	1,155
	4천8백만원 이상 6천만원 미만	327
	6천만원 이상	267

→

2020년 12월 SK텔레콤 가입자 기준

전체 표본 수		24,240,343
성별	남성	**********
	여성	**********
연령	20대 이하	**********
	30대	**********
	40대	**********
	50대	**********
	60대	**********
	70대 이상	**********
소득	4천만원 미만	**********
	4천만원 이상 5천만원 미만	**********
	5천만원 이상 7천만원 미만	**********
	7천만원 이상	**********

■ 이에 다음과 같은 싱글 X 세대별 X소득별 온라인 콘텐츠 이용률 추이와 같은 보다 세부적인 여행행태 분석이 가능해 짐

〈표 33〉 **(가명처리 데이터 활용 시)** 인구특성별 온라인 콘텐츠 이용율

(대상: SK텔레콤 가입자, 단위: 명, %)

구 분 (결혼여부/세대/소득구간)	연평균 관측치수	2019.8	2020.2	2020.6	2021.6	기간 증감
싱글*MZ세대*5000-7000미만	511,951	80.0%	89.7%	93.2%	98.6%	18.7%p
싱글*MZ세대*4000-5000미만	897,909	80.7%	89.4%	93.0%	98.7%	18.0%p
싱글*MZ세대*3000-4000미만	1,547,329	79.9%	88.6%	92.5%	98.5%	18.6%p
싱글*MZ세대*2000-3000미만	2,751,256	78.3%	85.3%	90.3%	97.9%	19.6%p
싱글*MZ세대*1000-2000미만	487,876	77.8%	83.2%	88.4%	97.0%	19.2%p

자료: SK텔레콤 가명처리데이터

5. [데이터 결합] 문화·관광 부문 활용방안

가. 데이터 결합 절차도

■ 데이터 결합 절차는 개인정보위원회 「가명정보 처리 가이드라인(2020)」에서 제시한 가명정보 결합절차를 기반으로 진행하되, 제공기관과 결합기관의 요구 사항에 따라 추가적인 행정절차를 진행하였으며 다음 〈표 34〉과 같이 도식화 할 수 있음

〈표 34〉 데이터 결합 업무 절차도

구분	주체
1단계: 데이터 결합 관련 업무 협의 데이터의 안전한 활용을 위해 제공기관과 이용기관 간, 협약체결	제공기관-이용기관
⇩	
2단계: 가명정보 활용 내부관리 계획 및 분석 환경 점검 가명정보 활용 관련 내부관리계획 수립 및 점검 가명정보 활용 분석 환경 점검	이용기관
⇩	
3단계: 데이터 결합 신청 가명정보 관리 환경 및 이행 확약서 작성(이용기관) 결합 대상 데이터 작성(제공기관) 데이터 결합 적정성 평가(결합기관)	제공기관, 이용기관, 결합기관
⇩	
4단계: 데이터 결합 및 적정성 평가 적적성 평가위원회 개최, 데이터 결합 및 이용에 대한 적정성 평가(결합기관)	제공기관, 이용기관, 결합기관
⇩	
5단계: 데이터 반출 및 문석	이용기관

나. 최종 결합 데이터 구성

■ 결합 데이터는 기존의 SK텔레콤과 신한카드 가명처리데이터를 결합하였기 때문에 앞서 선택한 컬럼과 동일한 구조를 가짐

〈표 35〉 [가명처리 데이터] 통신 데이터 최종 컬럼 구성

컬럼	기간	세부내용
ⅰ. 개인 특성		
성별	'19.1~'21.6	시군구 코드
연령대	'19.1~'21.6	1인 가구, 2인 가구, 3인 가구, 4인 가구, 5인 가구 이상
세대구분 1	'19.1~'21.6	M세대(1981~96년생), Z세대(1997~2010년생), 그 외 세대
세대구분 2	'19.1~'21.6	청년(18~34세), 중년(35~49세), 장년(50~64세), 노년(65세 이상)
가구원수	'20.1~'21.1	1인 가구, 2인 가구, 3인 가구, 4인 가구, 5인 가구 이상
거주지역	'19.1~'21.6	시군구 코드
연간소득 ** 8개 소득 구간*	'19.1~'21.6	1억이상, 7천-1억미만, 5천-7천미만, 4천-5천미만 3천-4천미만, 2천-3천미만, 1천-2천미만, 1천 미만
ⅱ. 관광 부문: 이동횟수		
평일 이동 횟수	'19.8~'21.6	평일 중에서 법정공휴일을 제외한 날 한 달 간 평일 총 이동 횟수
휴일 이동 횟수	'19.8~'21.6	2개 요일(토일) 또는 법정공휴일 한 달간 휴일 총 이동횟수
ⅲ. 문화 부문: 온라인 콘텐츠 이용		
평일 온라인콘텐츠이용량	'19.8~'21.6	전체, 동영상, 음악, 게임, 도서, 웹툰, 커뮤니티 등
휴일 온라인콘텐츠이용량	'19.8~'21.6	전체, 동영상, 음악, 게임, 도서, 웹툰, 커뮤니티 등
ⅲ. 문화 관련 온라인 콘텐츠 관련:		
구독서비스 정보	'19.8~'21.6	구독서비스 명
구독기간 이용 기간	'19.8~'21.6	개월
ⅱ. 오프라인 카드 지출액		
전체 월별	'19.1~'21.6	전체 발생한 신용카드 지출의 월별 총 금액(단위: 원)
관광 월별	'19.1~'21.6	관광 분야의 신용카드 지출의 월별 총 금액(단위: 원)
콘텐츠 월별	'19.1~'21.6	오프라인콘텐츠 분야의 신용카드 지출의 월별 총 금액(단위: 원)
ⅲ. 온라인 콘텐츠 지출액		
영상 월별 카드 지출	'19.1~'21.6	영상관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)
음악 월별 카드 지출	'19.1~'21.6	음악관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)
게임 월별 카드 지출	'19.1~'21.6	게임관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)
출판 월별 카드 지출	'19.1~'21.6	출판관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)

다. 실증분석 주제 선정

- 결합 데이터가 가지는 강점을 보여줄 수 있는 관광과 문화콘텐츠 분야의 분석 가능한 주제를 검토하고, 다음 표와 같이 선정함

〈표 36〉 결합 데이터 활용방안 도출을 위한 실증분석 주제 선정

이종 데이터 결합을 통한 활용 가능한 정보의 확대	
활용 가능 정보 확대	**유효표본 수 확대**
(관광부문) - 1인 평균 관광 이동량에 따른 관광소비지출 현황 - 1인 평균 관광 이동량에 따른 관광 여행사, 면세점, 항공사, 관광숙박업 지출 현황 **(문화콘텐츠 부문)** - 1인 평균 이동량에 따른 콘텐츠(온라인, 오프라인) 소비 추이 분석 - 지역별 이동량에 따른 온라인 콘텐츠 소비 추이 분석	**(관광부문)** - MZ세대의 이동량별, 성별 관광 이동량에 따른 관광소비지출 현황 - 시군구별 관광 이동량에 따른 관광소비지출 현황 **(문화콘텐츠 부문)** - MZ세대의 이동량별, 성별, 소득별 온라인 콘텐츠 소비액 분석 - 지역별 MZ세대의 이동량별, 성별, 소득별 온라인 콘텐츠 소비액 분석

라. 관광부문 활용방안

1) 이종 간 결합 정보를 활용한 현황 분석

- 이종데이터를 결합하게 되면 활용가능한 정보가 확대되어 새로운 결과를 도출할 수 있는 가능성이 높아짐
- 이에 본 연구에서는 SK텔레콤 통신데이터와 신한카드 지출액 데이터를 결합하여 휴일 이동량에 따른 소비지출 현황과 상관관계를 분석함

〈표 37〉 이종 간 결합 정보를 활용한 이동량과 관광소비지출 분석 결과 비교

(단일 통신 데이터) 1인당 분기 총 휴일 이동량

대상: SKT가입자, 단위: 1인/회)

이동량	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
휴일 이동량	39	30	36	26	33	29	38

자료: SK텔레콤 가명처리 데이터

(단일 카드 데이터) 1인당 분기 총 관광부문 카드지출

(대상: 신한카드 가입자, 단위: 1인/원)

구분	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
관광 지출액	474,310	361,698	394,944	385,762	353,729	314,907	398,888

자료: 신한카드 가명처리 데이터

(통신 + 카드 결합 데이터) 1인당 이동량에 따른 분기 총 관광부문 카드지출

(대상 SK텔레콤, 신한 동시가입자 단위: 1인/원)

이동량	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
1분위	391,136	285,700	296,840	280,475	262,971	233,391	300,266
2분위	502,885	396,054	411,345	399,322	376,480	336,582	414,852
3분위	622,452	509,620	540,874	540,602	510,905	459,255	550,020

자료: SK텔레콤, 신한 결합데이터

〈표 38〉 **(통신 + 카드 결합 데이터)** 1인당 이동량 별 관광소비지출 상관관계 분석

(대상: SK텔레콤, 신한 동시가입자, 피어슨상관계수)

구분	이동량	전체지출액
이동량	1.0000	
전체지출액	0.0599*	1.0000

전체

구분	이동량	관광지출액
이동량	1.0000	
관광지출액	0.1793*	1.0000

관광부문 지출액

자료: SK텔레콤, 신한 결합데이터

2) 유효표본 수 확대에 따른 현황 분석

- 결합데이터의 대규모 유효표본을 활용할 경우 또 하나의 강점은 시군구 단위의 통계를 생산할 수 있다는 점임
- 이에 국민여행조사의 17개 시도의 표본수와 SK텔레콤과 신한카드 결합데이터의 시군구 단위의 표본 수를 비교하면 〈표 38〉와 같음

〈표 38〉 데이터 결합 후 유효표본 수 확대에 따른 확보 가능한 개인 특성별 유효표본 수 비교

(단위: 명)

2020년 기준 국민여행조사 표본수		→	2020년 12월 통신,카드 동시가입자	
전체 표본 수	50,710		전체 표본 수	3,449,478
남성	25,251		남성	1,655,252
여성	25,459		여성	1,794,226
MZ세대	19,055		MZ세대	1,497,015
그외세대	31,655		그외세대	1,952,463
1인 가구	7,346		1인 가구	1,113,186
2인 가구	12,531		2인 가구	739,941
3인 이상 가구	30,833		3인 이상 가구	1,048,631

2020년 기준 국민여행조사 표본수		→	2020년 12월 통신,카드 동시가입자	
전체 표본 수	50,710		전체 표본 수	3,449,478
남성X1인가구XMZ세대	1,738		남성X1인가구XMZ세대	323,955
남성X1인가구X그외세대	1,547		남성X1인가구X그외세대	297,502
여성X1인가구XMZ세대	948		여성X1인가구XMZ세대	268,520
여성X1인가구X그외세대	3,113		여성X1인가구X그외세대	223,209

■ 본수가 적은 경상북도 울릉도 군 거주자 대상으로 이동 수준에 따라 구분하면 다음 〈표 39〉과 같이 유효표본을 확보할 수 있음

〈표 39〉 경상북도 울릉군 거주자의 이동량 별 유효표본 수

(대상 SK텔레콤, 신한 동시가입자 단위: 명)

구분	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
전체	313	321	327	330	331	331	339
1분위	87	74	88	89	82	73	78
2분위	94	87	104	103	118	128	118
3분위	132	160	136	138	132	130	142

■ 이를 바탕으로 다음과 같이 울릉도 거주자의 이동량 별 관광소비 지출액을 산출할 수 있게 됨

〈표 40〉 경경상북도 울릉군 거주자의 이동량 별 관광소비 지출액

(대상 SK텔레콤, 신한 동시가입자 단위: 1인/만원)

구분	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
1분위	56	43	41	39	39	36	40
2분위	36	42	41	39	35	38	40
3분위	32	30	26	29	31	21	26

자료: SK텔레콤, 신한 결합데이터

마. 문화콘텐츠 부문 활용방안

1) 이종 간 결합 정보를 활용한 현황 분석

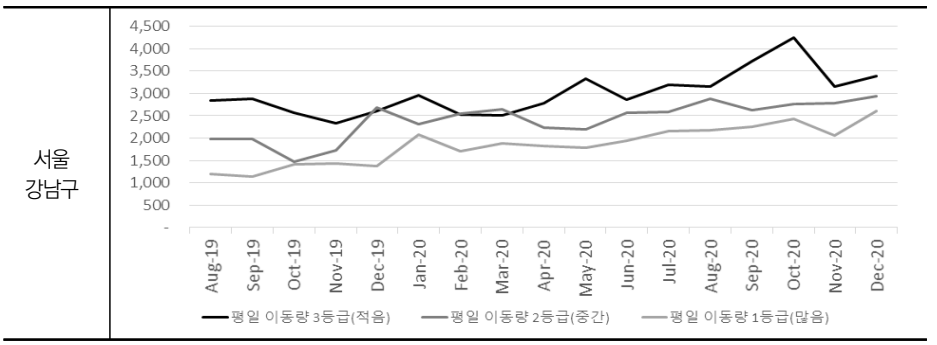
■ 문화콘텐츠 분야 또한 이종데이터를 결합하게 되면 활용가능한 정보가 확대되어 새로운 결과를 도출할 수 있는 가능성이 높아짐

■ 다음은 SK텔레콤 통신데이터와 신한카드 지출액 데이터를 결합하여 평일 이동량에 따른 온라인 콘텐츠 업종 소비액 추정 결과임

- 기존 각각의 데이터로 볼 수 없던 휴일 이동량과 콘텐츠 업종 소비액의 상관성을 유추해 볼 수 있음

[그림 2] 지역별 평일 이동량 등급에 따른 온라인 콘텐츠업종 소비액 추이(2019.8-2020.12.)

(대상 SK텔레콤, 신한 동시가입자 단위: 원)



자료: SK텔레콤, 신한 결합데이터

2) 유효표본 수 확대에 따른 현황 분석

■ 이종의 데이터를 결합하더라도 여전히 큰 유효표본을 확보할 수 있어 다음과 같은 MZ세대의 이동량별, 성별, 소득별 온라인 콘텐츠 소비액 분석과 같은 보다 세부적인 분석이 가능함

〈표 41〉 연소득 3천만원 이상 5천만원 미만 MZ세대'의 평일 이동량 등급에 따른 온라인 콘텐츠 소비액 추이 및 증감

(대상 SK텔레콤, 신한 동시가입자 단위: 원,%)

구 분		2019.08.	2019.12.	2020.02.	2020.06.	2020.12.	기간증감 ('19.8 - '20.12.)
남성	1등급	3,807	3,658	4,666	4,530	5,676	49.1
	2등급	4,616	5,283	6,050	6,096	7,015	52.0
	3등급	8,017	7,718	10,083	10,810	10,004	24.8
여성	1등급	1,116	1,337	1,389	1,627	2,009	80.0
	2등급	1,139	1,406	1,710	1,896	2,354	106.6
	3등급	1,818	2,301	2,594	2,726	2,855	57.1

자료: SK텔레콤, 신한 결합데이터

6. 결론 및 시사점

가. 결론

■ 본 연구는 데이터 기반 행정의 제도적 기반 마련 및 데이터 3법(데이터 규제 완화 3법)[8]의 개정안 통과를 통해 가명정보를 통한 활용 가능한 데이터의 범주가 확대됨에 따라 선제적으로 데이터 활용 가능성을 검토하고 활용방안에 대한 구체적인 가이드라인 제공을 위해 실시함

■ 또한 기존의 연구 목적과는 차별적으로 설계(design), 분석(analysis), 진단(diagnosis)이라는 틀 안에서 본 연구를 통해 도출하고자 하는 세부사항을 검토함

- **(설계)** 첫 번째는 단일 가명데이터만을 활용한 방안이며 두 번째는 이종 간 가명데이터를 매칭키를 통해 결합한 후 활용하는 방안을 검토함
- **(분석)** 앞서 설계한 통신사와 카드사의 가명처리 데이터와 가명처리된 데이터를 매칭키로 결합한 결합데이터를 기반으로 문화관광분야 활용컬럼 도출 및 트렌드 분석을 실시
- **(진단)** 본 연구의 설계단계, 분석단계 등을 종합적으로 검토하여 향후 가명데이터 기반의 연구 수행시 필요한 주요 내용들을 정리하고 추후 검토 및 보완해야 할 부분들에 대해 제시

■ 본 연구의 목적에 따른 자료 검토와 실증분석을 통해 도출한 최종 결론은 다음과 같음

(ⅰ) 가명처리 데이터를 통한 분석의 다양화

- 데이터 활용을 위한 여러 가지 제도개선 등이 수반되면서 연구 및 통계 생산 목적으로 이용할 수 있게 되었음
- 이로 인해 기존의 빅데이터 분석에서 볼 수 없었던 보다 다양한 분석 및 결과 도출이 가능해짐

(ⅱ) 시의성 있는 데이터 분석 가능

- 실시간 데이터를 확보하고 분석할 수있으며, 심지어 일 단위를 넘어 시간단위까

8) 데이터 이용을 활성화하는 「개인정보 보호법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률(약칭: 정보통신망법)」, 「신용정보의 이용 및 보호에 관한 법률(약칭: 신용정보법)」등 3가지 법률을 통칭

지도 분석이 가능한 데이터로 활용이 가능함

- 본 연구에서는 데이터 용량에 따른 분석 시간 등을 고려하여 월 단위 개인 집계 데이터로 분석을 시도했지만 향후 보다 디테일한 분석을 위해서는 충분히 시의성 있는 데이터 확보를 통한 분석이 가능해질 것임

(ⅲ) 데이터 결합을 통한 확장성

- 물리적 결합이 어려웠던 이종 간의 데이터를 결합하여 사용할 경우 분석을 위한 확장성이 엄청나게 커짐
- 또한 한 개의 데이터로 설명하지 못했던 데이터 기반의 분석결과를 데이터 결합을 통해 보완함으로써 결과의 활용 및 확장성은 계속해서 커질 것임

나. 진단

1) 가이드라인

■ 데이터 활용에 대한 전반적인 내용을 정리하여 공유하는 것을 통해 향후 관련 데이터 활용에 있어서 동일한 오류를 범하지 않고 보다 효율적으로 연구할 수 있는 기반마련이 가능해질 것임

■ 이에 본 장에서는 가명처리 데이터의 활용과 가명처리 데이터를 통한 데이터 연계까지 연구를 통해 정리하고 검토한 내용을 다음과 같이 도식화하여 제시함

2) 개선방안 및 향후 과제

① 개선방안

■ 활용 컬럼의 명확한 정의 검토

- 현재 활용되는 대부분의 빅데이터(통신 또는 카드)의 경우 데이터를 제공하는 기관에서 설정한 기준을 그대로 준용하여 활용하고 있으며 이에 대한 검증도 이루어지지 않고 있음
- 이에 보다 효율적인 데이터 활용을 위해서는 제공되는 데이터에 대한 생산 방식에 대한 이해 및 구조화된 컬럼에 대한 명확한 기준마련이 필요함

[그림 3] 가이드라인

단계	주요내용
Ⅰ. 환경 조성	가명정보 관리 내부관리계획 수립 가명정보 관리 내부환경 조성
Ⅱ. 가명처리 데이터 검토	가명처리 데이터 제공기관 및 결합기관 선정
	가명처리 데이터 활용 주제 선정
	최종컬럼 선정 – [활용 가능한 컬럼 보유 여부 확인] 최종컬럼 선정 – [활용 적정성 평가] 개인정보 유출 관련 법률적 검토
	가명 및 익명처리
Ⅳ. [가명처리 데이터] 반출 및 분석	적정성 평가
	반출신청서 및 폐기확약서 작성
	비밀보호 협약서 작성
	데이터 전송 및 분석
Ⅳ. 가명처리 데이터 연계	데이터 연계 협약 체결
	데이터 연계 신청
	적정성 평가
	데이터 결합
Ⅴ. [결합 데이터]반출 및 분석	가명정보 내부관리 계획 기반 데이터 반출 및 분석

■ 표본설계를 통한 빅데이터 신뢰성 확보

- 기존의 빅데이터 연구에서 조사결과의 추정을 위해 통신 및 카드데이터 가입자 정보 및 비율을 통해 보정함으로써 모수추정 값에 대한 신뢰가 항상 문제점으로 대두됨
- 이러한 문제를 해결하고 가명처리된 빅데이터를 보다 효율적으로 활용하기 위해서는 기존의 단순 비례 보정방식이 아닌 모집단을 대상으로 한 정교한 표본설계를 통한 모수추정 방안 마련이 필요함

■ 가명데이터 활용을 위한 표준화 방안 마련

- 본 연구는 민간이터 가명처리 및 결합을 중심으로 진행되었음. 이에 향후 다양한 데이터를 기반으로 한 사례연구를 통해 가명데이터 활용을 위한 표준화 방안이 마련되어야 하며 데이터 종류에 따라 산발적으로 데이터 결합을 하는 것 보다는 명확한 주제를 기반으로 한 관광분야, 문화분야, 인구분야, 복지분야 등 분야별 데이터의 표준을 정립하는 것도 필요할 것

② 향후 과제

■ 설문조사데이터의 대체방안 검토

- 가명처리 데이터의 활용도가 높아졌다고 해도 아직은 현실적으로 빅데이터가 조사데이터를 완벽하게 대체하기는 어려운 상황임
- 다만 통계청[9] 등에서도 빅데이터를 활용한 통계생산 방안 등을 검토하고 있는 만큼 연구원 차원에서도 이에 대한 선제적 검토를 통해 데이터 확보의 효율화 방안을 모색해야 할 것

■ 공공데이터의 활용방안 마련

- 본 연구에서는 민간데이터 기반의 연구를 수행하였지만 공공에서 제공하고 있는 많은 데이터 검토를 통해 공공데이터 활용기반 마련을 위한 노력이 필요할 것임
- 데이터 종류, 데이터 구조, 데이터 컬럼 등 공공데이터가 가지고 있는 종합적인

9) 통계청에서는 SK텔레콤의 빅데이터 등 새로운 통계의 활성화를 위해 「통신 모바일 인구이동량 통계」를 실험적 통계 1호로 서비스('21.9.16)하고 코로나19 발생전후의 통계수요 도출 검토

정보를 정리하고 향후 데이터 활용을 위한 기반을 마련한다면 보다 폭넓은 분석 결과 도출이 가능해질 것으로 판단됨

- 빅데이터 패널 구축을 통한 종단 연구 및 예측분석 모델 마련
  - 가명데이터 기반의 패널 유지가 용이하게 된다면 인구 특성별 여행의 라이프사이클 등에 대한 시계열예측 가능할 것이며 이를 AI가 예측할 수 있는 학습데이터로 축적하여 빅데이터 분석기법(클러스터링 및 유동인구예측)을 이용한 여행객 규모예측 등이 가능해질 것
  - 유동인구데이터를 딥러닝 모델(Mask R-CNN모델) 또는 t-SNE/ UMAP 등으로 예측하여 연령별・성별・시간대별 유동인구데이터의 평균값뿐만 아니라, 유동인구의 갑작스런 증감행태 등이 파악 가능한 표준편차, 중간값, 1/2/3/4분위 이동량 등을 분석하여 정책수립의 기초자료로 활용할 수 있을 것으로 판단됨

- 빅데이터 활용 변화에 대한 지속적 검토
  - 현재 문화관광분야에서 활용하고 있는 빅데이터(가명데이터 포함)의 경우 데이터 도출을 위한 기술적 환경변화에 따라 데이터 도출값이 달라질 수 있는 우려사항이 있음
  - 최근 데이터 집계방식으로 삼각측량, 서브기지국 활용, 5G 변화에 따른 집계수준 변화, WIFI 데이터 활용 등 다양한 변화를 모색하고 있기 때문에 이러한 기술이 적용될 경우 데이터를 어떤 방식으로 활용할 것인지에 대한 검토도 필요할 것임

- N종간 데이터 연계방안 검토
  - 해당 연구에서는 문화관광분야에 기본적으로 가장 많이 활용되는 데이터를 기반으로 연구를 수행하였지만 향후에는 보다 확대된 차원의 분석 및 시사점 도출을 위해 N종간 데이터 검토도 필요할 것
  - 이를 위해서는 가명처리된 개인화된 결합 가능 데이터들이 어떠한 칼럼들을 지니고 있으며 어떻게 상호 결합될 수 있는지에 대해서 다수의 연구자들과 산업관계자들이 함께 검토하고 필요한 데이터에 대한 확보를 요구할 수 있는 시스템이 구축되어야 할 것임

■ 가명정보 활용을 위한 통합기반 마련

- (가명정보 처리 전담부서 설치)가명처리 관련 업무 총괄·관리 및 의사결정을 위한 전담부서를 지정하여 가명처리 목적 적합성 검토, 가명처리, 가명처리 적정성 검토, 가명정보취급자에 대한 관리·감독 등을 수행하여야 함
- (데이터 통합 관리 기반 마련)데이터의 효율적 관리 및 활용을 위해서는 현재 기관에서 보유중인 데이터 및 향후 확보가 예상되는 데이터에 대한 관리가 체계적으로 이루어져야 함
- (보호조치가 되어있는 가명처리·분석 환경 구축) 명처리 및 분석 환경을 구축하고자 할 경우, 「개인정보 보호법」에서 요구하는 가명정보 보호조치 수준이 개인정보 보호조치 수준과 유사하므로 개인정보처리시스템 보호조치 항목에 준하여 환경을 구축하여야 함

목차

표 목차

그림 목차

제1장

서론

제1절 연구 배경 및 목적

1. 연구 배경

2020년 6월 데이터를 기반으로 한 과학적 행정 구현을 국정과제로 설정하고 복지분야에서 맞춤행정을 위한 데이터 센터를 운영하는 것을 시작으로, 「데이터 기반 활성화에 관한 법률」[1]이 재정되었다. 따라서 개인이나 조직의 경험이나 직관에 따라 정책을 수립하는 방식에서 나아가 '데이터를 정책 수립 및 의사결정에 활용함으로써 객관적이고 과학적 행정을 수행'할 수 있는 '데이터 기반 행정'의 제도적 기반이 마련되었다. 문화체육관광부에서도 데이터 기반 활성화의 일환으로 2020년 6월 개별적으로 관리했던 데이터를 일원화하여 관리하고 통계에 기초한 정책 체계를 구축하기 위해 정책분석팀을 신설하였다.

또한 2020년 8월 데이터 3법(데이터 규제 완화 3법)[2]의 개정안 통과로, 가명정보를 통한 활용 가능한 데이터의 범주가 확대됨에 따라 통계작성, 사회과학 및 정책 연구라는 제한적 목적 하에서 정보주체의 동의 없이 가명정보를 처리할 수 있는 환경이 조성되었다. 여기서, 가명정보란, 개인정보를 가명처리함으로써 원래 상태로 복원하기 위한 추가 정보의 사용 또는 결합 없이는 특정 개인을 알아볼 수 없는 정보를 의미하는데, 기존의 개인정보보호법 하에서는 특정 서비스를 제공하기 위해 개인정보처리를 받는 경우, 해당 개인 정보는 반드시 특정 서비스를 제공하기 위해서만 허용되어, 개인정보 처리의 목적이 변경되는 경우 다시 개인의 동의가 필요하기 때문에 비용과 시간 소요가 비교적 컸다. 하지만 데이터 3법이 통과됨에 따라 가명화된 개인정보를 보다 자유롭게 사용할 수 있게 되어, 추가적인 절차 없이 가명정보를 활용한 데이터를 활용할 수 있는 법률적 제도가 마련되었다.

---

1) 데이터기반행정 활성화에 관한 법률(약칭: 데이터기반행정법) [시행:2020. 12. 10] [법률: 제17370호. 2020년 6월 9., 재정]

2) 데이터 이용을 활성화하는 「개인정보 보호법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률(약칭: 정보통신망법)」, 「신용정보의 이용 및 보호에 관한 법률(약칭: 신용정보법)」등 3가지 법률을 통칭

이렇듯 개인 정보를 가명 처리한 데이터(이하. 가명처리 데이터) 활용의 기반이 마련됨에 따라 경제, 사회 및 복지, 의료, 문화·관광 등 다양한 분야에서 활용 가능한 공공 및 민간 데이터의 수가 증가할 뿐만 아니라, 기존에 활용되고 있는 데이터의 활용성이 개선되어 그 가치가 높아질 것으로 예상된다. 가명정보는 식별가능성이 완전히 제거된 '익명정보'와는 구별되는 개념으로, 정보의 활용 가치 측면에서는 익명정보보다 유용하게 활용되며 정보 업데이트나 정보 간 결합이 불가능하여 활용도가 낮아 연구의 제약요인으로 대두되었던 문제점을 개선할 수 있을 것으로 판단된다. 또한, 가명정보를 활용하게 될 경우, 기존 빅데이터에서 파악 가능한 유형화된 데이터를 넘어 보다 다양한 인구통계학적 변수를 활용할 수 있으므로 기존 데이터로 파악하기 어려웠던 새로운 집단의 정의 및 분석이 가능하게 될 것으로 예상된다. 특히, 개인 혹은 가구, 사업체 단위의 데이터를 이용할 수 있게 되어 단순 빈도분석 외 통계적 모델링(statical modeling)을 이용해 정책 효과에 대한 인과관계 및 예측 분석이 가능하다는 장점이 있다.

또한 가명정보를 주요변수(key variable)로 하여 각기 다른 형태의 데이터를 분석 가능한 하나의 데이터 형태로 구성하는 '데이터 연계'[3]가 가능하게 되어, 이는 기존 분석 결과와는 차별화된 새로운 정책적 가치를 창출하고 향후 미래를 예측하는데 활용 할 수 있을 것으로 판단된다. 사실상 기존의 데이터 연계는 가명정보와 같은 주요 변수(key variable)가 없기 때문에 서로 다른 자료에서 유사한 개체를 결합하는 통계적 매칭(statistical matching)을 이용하여 연계[4]하였다. 하지만 가명정보를 데이터 연계의 주요변수로 활용하게 되는 경우, 서로 다른 자료에서 동일 개체를 결합하는 정확 매칭(exact matching)이 가능하여 보다 정확한 데이터 정보를 구성할 수 있다는 장점이 있다. 이러한 데이터 연계는 기존의 단일 데이터 정보에 추가정보를 추가하여 보다 많은 정보를 확보할 수 있을 뿐만 아니라 서로 다른 형태로 구성된 데이터를 매칭해 시계열 데이터를 생성할 수 있게 되므로 정책적 이슈를 설명하는데 보다 유의미한 근거자료로 활용될 수 있을 것이다.

이러한 데이터 활용의 장점이 부각됨에 따라 '가명처리 데이터'를 이용해 정책적으로 유의미한 통계를 생산하기 위한 제도적 기반(예. 개인정보처리 가이드라인 마련, 가명정

3) 데이터 연계(data linkage) 또는 데이터 매칭 (data matching)이란, 서로 다른 복수의 데이터 파 일을 결합하여 보다 풍부한 정보를 제공해 줄 수 있는 하나의 완전한 통합데이터를 만드는 방법을 의미함

4) 기존에는 가명처리된 정보가 없었기 때문에, 유사 개체를 결합하는 통계적 매칭과 관련 된 연구들이 진행된 바 있음(참고: 박근화(2019) 문화체육관광 데이터 연계를 통한 빅데이터 생산 및 활용방안 연구)

보 결합 전문기관 지정 등)마련이 활발하게 이루고 지고 있다. 2020년 9월 개인정보위원회에서 '가명정보 처리 가이드라인'을 발간하여 가명처리 및 가명정보 결합과 반출에 관한 절차를 구체적으로 안내하고 있다. 또한 2020년 10월 보건복지부에서 보건의료분야 '가명정보결합 전문기관'[5]결합전문기관으로 국민건강보험공단, 건강보험심사평가원, 한국보건산업진흥원을 지정한 것을 시작으로 2020년 11월 개인정보위원회에서 통계청과 삼성 SDS를 '21년 1월 과학기술정보통신부에서 한국지능정보사회진흥원(NIA), SK 주식회사, 더존비즈온을 가명정보 결합전문기관으로 지정되었다.

이에 문화·관광분야 또한 이러한 환경 변화에 맞춰 가명처리 데이터 활용의 필요성을 점검[6]하고, 기존 데이터 분석으로 발견할 수 없었던 새로운 정책적 활용 방안을 제시하기 위한 연구가 필요한 시점이다. 데이터 분석을 위하여 개인 정보를 가명처리하고 더 나아가 각 데이터를 연계하는 것은 시간과 비용이 요구되는 작업으로 기존 데이터 분석 결과와 특징적인 차별점이 없다면 무의미한 연구라고 할 수 있다. 따라서 가명처리 데이터를 검토하고 분석 및 연계하는 것이 어떠한 장점이 있으며 얼마나 새로운 결과를 도출할 수 있을지에 대한 검토연구의 필요성이 커지고 있다.

2. 연구 목적

본 연구의 목적은 가명데이터의 활용이 원활해짐에 따라 선도적으로 문화관광분야에서 가명처리데이터 활용에 필요한 일련의 프로세스를 직접 검토하고 분석하는 절차를 제공함으로써 향후 관련분야 데이터 활용 및 후속연구를 위한 기초적 내용을 제공하는 것이다. 연구목적을 달성하기 위해 설계(design), 분석(analysis), 진단(diagnosis)으로 구분한 3가지 연구진행 프로세스를 설정하였으며 이를 토대로 세부적인 연구내용을 제시하였다.

설계(design): 본 연구에서 다루어질 가명처리데이터에 대한 전반적인 사항은 기존에 다루어진 적이 없는 주제인 만큼 가명처리데이터 활용을 위한 기초적인 설계단계가 매

5) 서로 다른 개인정보 처리자간의 가명정보 결합을 수행하기 위해 개인정보 보호위원회(이하 "보호위원회"라 한다) 또는 관계 중앙행정기관의 장이 지정하는 전문기관을 의미함

6) 문화·관광 분야 또한 데이터 3법 시행 후 신한카드-SK텔레콤 간 MOU를 체결('20.8)하고 통신과 카드 데이터에 대한 결합 1호 신청('20.9)하였으며, 연구원에서도 데이터 결합 및 활용에 선도적으로 대응하기 위해 신한카드, SK텔레콤과 3자간 MOU를 체결 ('20.9)함.

우 중요하다. 이를 통해 향후 관련 분야 연구의 기반이 마련될 수 있기 때문이다. 설계 단계의 하위목적은 다음과 같다. 첫째, 가명정보, 가명데이터, 가명데이터를 통한 결합데이터에 대한 명확한 개념 및 구조를 제시함으로써 가명데이터 전반에 대한 이론적 개념을 정립한다. 둘째, 가명데이터의 특징(장점)을 검토하고 분석을 위한 전반적인 고려사항을 제시함으로써 효율적 활용을 위한 지침을 정비한다. 셋째, 공공·민간·조사데이터의 종합검토를 통해 가명데이터의 활용 및 결합방안을 제시한다.

분석(analysis): 분석 단계에서는 설계 단계에서 검토된 가명데이터 및 결합데이터를 통해 데이터 활용 가능성을 위한 실증분석을 실시한다. 분석단계의 하위목적은 다음과 같다. 첫째, 분석에 활용될 가명데이터를 선정하고 문화관광 차원에서 활용 가능한 의미있는 컬럼을 도출한다. 이를 통해 가명데이터를 통해 분석 가능한 차별화된 분석결과를 제시한다. 둘째, 가명정보를 통해 이종 간 데이터의 결합을 검토하고 빅데이터의 정량적 결과 도출을 넘어 정성적 결과도출에 대한 구체적 제시 내용을 제시한다.

진단(diagnosis): 설계 단계를 거쳐 분석을 통한 결과를 도출하는 과정에서 가명데이터 및 결합데이터의 문화관광분야 활용에 대한 한계점을 분명히 발생될 것이다. 이러한 한계점을 극복하기 위한 진단단계의 하위목적은 다음과 같다. 첫째, 가명처리 데이터 활용 시 발생할 수 있는 개인정보, 분석 및 활용 관련 주요 이슈를 정리하고 향후 가명정보를 통한 문화·관광 데이터의 지속적인 연구를 위한 활용 가이드라인을 제공한다. 둘째, 아직까지 직접 가명처리된 데이터를 연구원에서 분석할 수 없는 인적·물리적·행정적 한계를 제시하고 향후 해당 분야 연구의 지속가능성을 위한 제도적 개선방안 및 정책적 활용방안을 제시함으로써 향후 연구를 위한 확장성을 도모한다.

〈표 1-1〉 연구진행 과정 및 내용

연구진행 과정(프로세스)	연구내용
가명처리데이터의 구조파악 및 현황검토를 통한 **설계(design)**	- i) 가명데이터의 명확한 개념 및 활용을 위한 구조(structure) 제시 - ii) 가명데이터 특징 검토 및 분서을 위한 고려사항 제시 - iii) 사례분석을 통한 가명데이터 활용 및 결합방안 제시
가명처리데이터 설계에 기반한 **분석(analysis)**	- i) 가명데이터의 문화관광분야 활용컬럼 도출 및 트렌드 분석 - ii) 이종 간 데이터 결합을 통한 활용방안 도출
설계, 분석단계의 한계를 기반으로 지속가능한 데이터 활용을 위한 **진단(diagnosis)**	- i) 가명데이터 효율적 활용을 위한 가이드라인 제시 - ii) 가명데이터 원활한 활용을 위한 제도적 개선방안 제시

제2절 연구 범위 및 방법

1. 연구 범위

가. 데이터 활용범위

본 연구는 가명데이터 구조분석을 통해 활용방안을 도출하는 것이 주목적이므로 데이터의 활용 범위를 명확하게 할 필요가 있다. 본 연구에서는 가명정보 활용을 위해 현재 제공 중인 문화관광분야 공공 및 민간 데이터를 종합검토하고 데이터 결합 시 가장 활용도가 높은 데이터가 무엇인지 살펴보고 데이터의 활용 범위를 설정하였다. 데이터의 제공가능성과 분석주제 도출적정성을 기준으로 최종적으로 민간에서 제공하는 가명처리 기반의 통신데이터와 카드데이터를 활용을 위한 분석데이터로 설정하였다. 또한 확보된 데이터를 기반으로 코로나 전·후의 문화관광의 행태변화를 파악하기 위해 코로나 발생 전인 2019년 6월부터 최근시점인 2021년 6월까지 2년간의 데이터를 분석데이터로 검토하였다.

나. 공간적 범위

가명정보의 활용을 위한 공간적 범위는 활용 가능한 자료 및 분석목적에 따라 설정된다. 예를 들어 전국분석 및 특정지역 분석 목적에 따라 통신데이터의 경우 집계구, 조사구, 시군구, 관광지점 등으로 구분되며 신용카드의 경우 시군구, 특정지점, 가맹점 등의 단위로 설정된다. 본 연구에는 활용된 데이터의 특성(많은 유효표본수 확보)상 공간적 범위는 시군구의 소지역까지 추정 가능한 강점을 내포하고 있다. 이에 실증분석을 위한 공간적 범위를 소지역 추정단위까지로 검토하였다.

다. 내용적 범위

본 연구의 내용적 범위는 설계(design), 분석(analysis), 진단(diagnosis)으로 구분할 수 있다. 설계단계의 경우 본 연구의 2장과 3장으로 구성되며 가명정보의 개념, 가명처리 및 가명데이터 결합, 가명데이터 및 결합데이터 활용 시 고려사항, 가명데이터 활용을 위한 공공·민간데이터 현황 및 활용 사례로 구성한다. 분석단계의 경우 본 연구의 4장과 5장으로 구성되며 설계단계에서 도출된 최종 활용 데이터를 기반으로 가명데이터만을 대상으로 한 분석과 가명데이터의 이종 간 데이터 결합을 통한 이원화된 분석을 통한 문화관광분야 활용방안을 도출한다. 진단단계는 본 연구의 6장 결론 및 한계를 정리한 내용으로 설계 및 분석단계를 통해 도출된 가명데이터 활용의 한계점 및 향후 보다 합리적이고 효율적인 데이터 활용을 위한 정책적 방안을 제시한다.

2. 연구 수행 방법

가. 문헌연구

본 연구는 문헌 연구를 통해 개인정보 가명처리개념, 가명처리 방법 및 절차, 가명처리 데이터 활용 범위에 대한 이론적 내용을 검토하고 가명정보 활용이 가능한 문화·관광분야 공공 및 민간 데이터 현황과 해당 데이터의 구조를 파악하였다. 또한 가명데이터 활용과 관련된 시사점을 도출하기 위하여 가명처리 데이터 관련 국내외 사례분석을 실시하였다.

나. 데이터 전처리 및 통계 분석

데이터 분석에 앞서서 데이터 제공기관으로부터 확보한 전체 컬럼을 기반으로 문화관광분야에서 활용 가능한 변수를 분류하고, 각 변수의 로직과 이상치를 점검하는 등 통계분석을 위한 데이터 전처리를 우선적으로 진행하였다, 전 처리된 데이터를 바탕으로 문화관광분야 핵심 현황파악을 위한 분석주제를 도출하고 ⅰ) 가명데이터만을 이용한 분석과 ⅱ) 가명데이터에 기반한 이종 간 결합데이터를 이용한 이원화된 통계분석 실시하여 시사점을 도출하였다.

다. 전문가 자문 실시

본 연구에 활용된 가명처리 데이터의 현황 및 직접적인 활용을 위해서는 데이터 제공기관 담당자의 직접적인 협의가 필요하다. 이를 위해 통신데이터와 카드데이터를 직접 운영하는 관련분야 전문가들의 의견 및 협조를 기반으로 연구를 수행하였다. 또한 문화관광분야에서의 활용을 위한 관련분야 전문가 자문 및 통계적 모델링과 더불어 도출된 결과에 대한 활용 등의 업무를 수행 경험이 있는 데이터 분석 전문가를 대상으로 자문회의를 실시하였다. 가이드라인 및 행정·제도적 기반 마련을 위한 전문가 섭외를 통해 추가적인 자문회의를 수행하였다.

〈표 1-2〉 전문가 자문회의 추진 개요

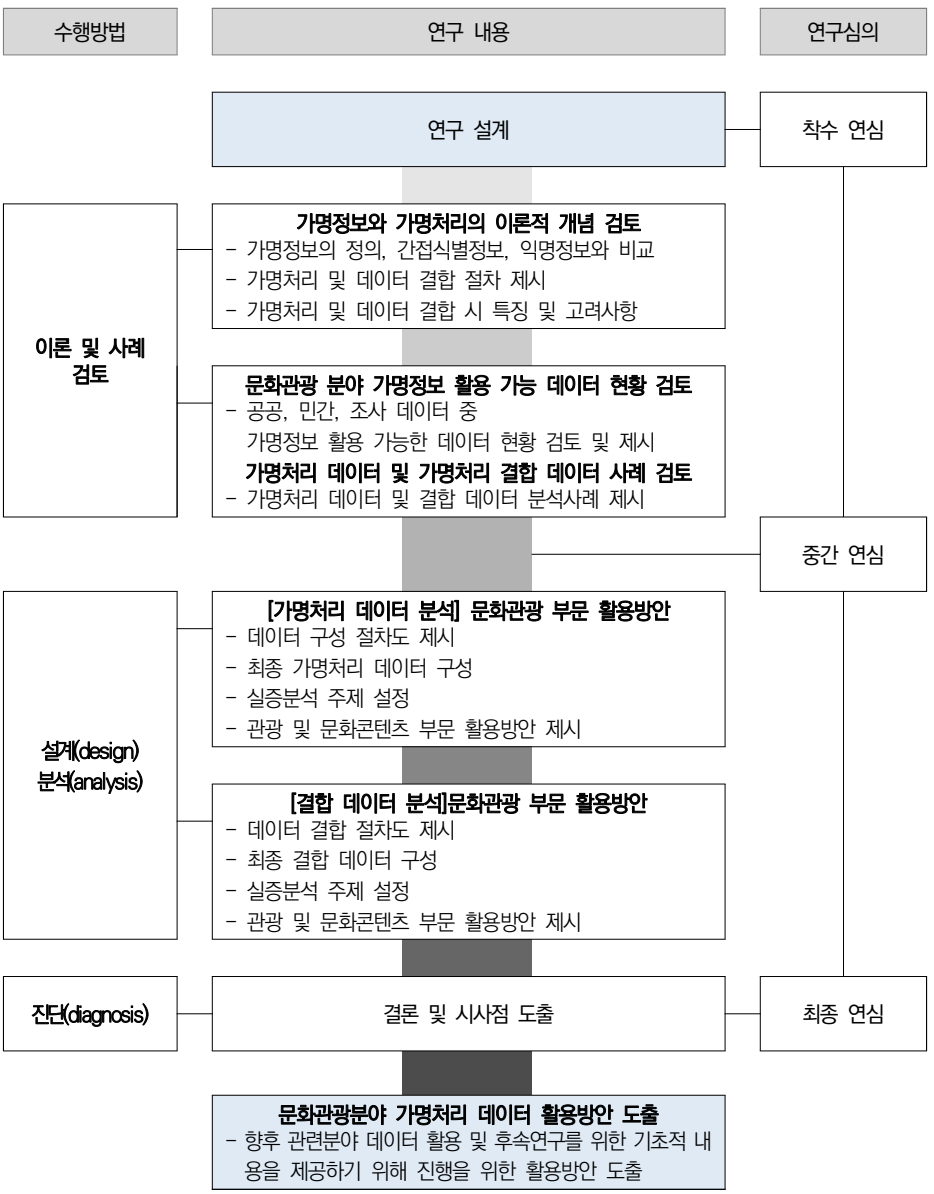
날짜	참석자	주요내용
3월 17일	SK 박○○ SK 김○○ 신한 이○○	[SK, 신한 가명처리 데이터 활용 협의] - 개인정보 가명처리 데이터 활용 및 데이터 연계 가능 여부 확인
4월 7일	SK 박○○ SK 김○○	[데이터 결합 관련 자문] - 가명처리 데이터 접근 관련 절차 및 분석 방안, 소요 일정 및 예산
4월 20일	SK 박○○ SK 김○○	[통신데이터 원자료 활용 1차 협의] - 관광 및 문화콘텐츠 분야 관련 컬럼 검토
5월 7일	SK 박○○ SK 류○○ SK 김○○	[통신데이터 원자료 활용 2차 협의] - 관광 및 문화콘텐츠 분야 관련 컬럼 검토
5월 13일	Smart Media Rep 윤○○	[1차 - 분석 주제 및 모형 도출: 문화콘텐츠 부문] - 가명 데이터를 활용한 분석 주제 도출 - 결합 데이터를 활용한 분석 주제 도출
5월 27일	가천대 이○○ 세종대 이○○ Smart Media Rep 윤○○ SK 류○○ SK 김○○	[2차 - 분석 주제 및 모형 도출: 관광 및 문화콘텐츠 부문] - 가명 데이터를 활용한 분석 주제 도출 - 결합 데이터를 활용한 분석 주제 도출
5월 27일	가천대 이○○ 세종대 이○○	[3차 - 분석 주제 및 모형 도출: 관광 부문] - 가명 데이터를 활용한 분석 주제 도출 - 결합 데이터를 활용한 분석 주제 도출
6월 8일	SK 김○○○	[SK 가명데이터 활용 보안 관련 협의] - 최종 데이터 구성 및 컬럼 - 분석 가능 여부 최종 확인

날짜	참석자	주요내용
6월 10일	SK 박○○ 신한 이○○	[가명처리 데이터 제공 절차 관련 업무 협의] - 가명결합' 관련 데이터 제공 절차 - '공급계약' 관련 절차 - 그 외 기타 논의사항
8월 25일	신한 안○○ 신한 김○○ 신한 이○○ SK 이○○	[sk-신한 데이터 자문 및 연계 업무 협의] 1. 일정 2. 활용 컬럼 확정 3. 활용 방안 논의 4. 연계 관련 자문 (분석방법, 소요 시간, 분석 진행 시 유의점 등)
9월 7일	SK 김○○ 신한 이○○	[문화·관광 분야 가명처리 데이터 활용방안 연구 데이터 활용 협약]
10월 7일	컬쳐미디어랩 김○○ 세종대 김○○ 가천대 이○○	[가명처리 및 결합 데이터 관광 및 콘텐츠 활용 주제자문회의] - 현재까지 주요 내용 자문 - 실증분석 주제 관련 의견 수렴
10월 7일	서울보건환경연소 김○○ 아태인구연구원 김○○ 빅데이터와미래연구소 강○○ (주)매지스 강○○	[가명처리 및 결합 데이터 통계 활용 관련 주제자문회의] - 현재 생산된 가명처리 데이터 활용 관련 문제점 - 향후 가명처리 데이터 활용 관련 개선사항 및 시사점
10월 22일	SK 김○○ SK 이○○ SK 이○○ 신한 이○○	[결합 데이터 분석 관련 자문회의] - 1. 현재 연계 데이터에 설명 - 2. 비식별 처리 관련 논의 - 3. 데이터 분석 실행
10월 26일	신한 이○○	[카드 데이터 가명처리 분석 관련 자문회의] - 가명처리 데이터 분석 실행 - 이상치 점검
10월 29일	SK 김○○○	[통신 데이터 가명처리 분석 관련 자문회의] - 가명처리 데이터 분석 실행 - 이상치 점검

3. 연구 체계

본 연구는 연구 설계, 이론적 개념 및 사례 검토, 실증 분석, 결론 및 시사점 도출 순으로 진행하였으며 다음과 같이 도식화 할 수 있다,

[그림 1-1] 연구의 수행방법 및 과정



제2장

가명정보와 가명처리의 이론적 개념

제1절 가명정보의 개념

데이터가 새로운 가치를 창출하고 경제성장을 이끄는 동력으로 부상하면서 4차 산업혁명 시대의 핵심자원으로 강조되고 있다. 이에 빅데이터, AI 등 다양한 융·복합 산업에서의 데이터 이용수요가 증가하고 있으며, 다양한 산업분야에서 데이터를 활용한 전략을 적극적으로 모색하고 있다. 하지만 데이터 이용에 핵심 정보는 개인에 관한 정보를 포함하는 경우가 많기 때문에 해당 데이터의 활용 확대를 강조하는 경우, 개인정보보호 침해 위험이 높아질 수 있다는 우려가 꾸준히 제기 되어 왔다.

이에 따라 정부는 개인정보를 안전하게 활용하기 위해 가명정보라는 개념을 도입하여 법에 근거하여 개인정보를 연구 목적으로 활용할 수 있는 길을 열었다. 정보주체의 동의 없이도 데이터를 안전하게 활용할 수 있도록 가명처리 및 데이터 결합제도 등을 새롭게 도입한 데이터 3법[7]이 시행됨('20년 8월) 에 따라 가명정보 활용에 대한 법적근거가 마련되어 체계적인 데이터 활용기반이 마련되었다. 이후 순차적으로 가명정보의 안전한 활용을 위한 제도적 기반으로 가명정보 활용 및 결합에 필요한 절차, 결합전문기관 지정 등에 관한 관련 기관별 고시[8] 제·개정이 이루어졌고 현장에서 공공기관 및 민간기관이 실제 가명처리·가명정보 결합 등 업무에 참고할 수 있는 분야별 가이드라인이 배포되었다.

이에 본 절에서는 가명정보에 대한 정의와 개인정보 가명처리 절차, 가명처리 된 이종의 데이터를 결합하는 방법을 살펴봄으로 써 가명정보의 이론적 개념을 소개하고자 한다. 또한 가명처리 데이터를 활용할 경우 기존 데이터와 어떠한 차별성을 가지는지 비교함으로써, 향후 관련 분야의 연구 및 통계 생산과 관련한 가명처리 데이터의 활용가치가 무엇인지 제시하고자 한다.

---

7) 「개인정보 보호법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」, 「신용정보의 이용 및 보호에 관한 법률」

8) 가명정보의 결합 및 반출 등에 관한 고시(개인정보위, '20.9.1.), 공공기관의 가명정보 결합 및 반출 등에 관한 고시(개인정보위, '20.12.2.), 신용정보업 감독규정 고시 개정(금융위, '20.8.5.)

1. 가명정보의 정의

개정된 개인정보보호법상 “가명정보”란 ‘개인정보를 가명처리 함으로써 원래의 상태로 복원하기 위한 추가정보의 사용 · 결합 없이는 특정 개인을 알아볼 수 없는 정보’를 의미한다.[9] 가명정보의 정의에 따라 가명정보는 추가정보의 사용 · 결합을 통해 개인정보로 복원할 수 있음을 의미하게 된다. 따라서 기존의 가명정보가 되기 위해서는 개인정보와 복원에 활용되는 추가정보가 필요하게 된다. 만약 수집 단계에서 개인 정보가 없다면 이는 가명정보가 아닌 간접식별정보 혹은 익명 정보에 해당하게 된다. 즉, 가명정보가 되기 위해서는 최조 수집된 개인정보와 복원 가능한 추가정보가 존재해야 한다.

2. 가명정보와 간접식별정보, 익명정보의 비교

“간접식별정보”란 “다른 정보”와 쉽게 결합하여 특정 개인을 알아볼 수 있는 정보를 의미한다. “다른 정보”란 해당 정보를 제외한 그밖의 모든 정보를 의미하므로 개인정보처리자가 현재 보관하고 있는 정보는 물론 합리적으로 입수 가능한 정보까지 포함한다. 반면, “추가 정보”는 넓게 보아도 개인정보처리자가 현재 보관하고 있는 정보로 한정되어야 하며(연계정보 및 원본정보), 좁게 보면 개인정보처리자가 별도 보관 중인 “연계정보”만을 의미한다고 보아야 한다. 따라서 어떤 데이터셋이 가명정보로 인정받기 위해서는 적어도 개인정보처리자가 별도로 보관하고 있는 “추가 정보” 이외의 다른 정보와 결합해서 특정 개인을 식별할 수 없게 조치하여야 한다. [10]

익명정보와 가명정보는 둘 다 특정 개인을 알아 볼 수 없는 정보라는 점에서는 공통적이다. 그러나 “가명정보”는 ‘추가 정보의 사용 · 결합 없이는 특정 개인을 알아볼 수 없는 정보’임에 비해서(제2조제1호다 목), “익명정보”는 ‘시간 · 비용 · 기술 등을 합리적으로 고려할 때 다른 정보를 사용하여도 더 이상 개인을 알아볼 수 없는 정보’이다(제58조의2)[11]

9) 개인정보보호법 제2조 1의2. "가명처리"란 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가 정보가 없이는 특정 개인을 알아볼 수 없도록 처리하는 것을 말한다.

10) 인터넷진흥원(2020)

11) 개인정보 보호법은 “익명정보”의 개념을 정의하고 있지도 않고 “익명정보”라는 용어를 사용하고 있지도 않지만 제58조의2에 해당하는 정보를 일반적으로 “익명정보”로 부르고 있다.

가명정보와 익명정보 모두 가명정보의 개념을 보다 명확하게 이해하기 위해서 〈표 2-1〉에 제시된 개인정보, 가명정보, 간접식별정보, 익명정보의 예시를 살펴보고자 한다. 첫 번째 김관광이라는 사람의 이름, 생년월일, 핸드폰번호 등 식별정보를 가진 개인정보 원본에서 민감한 정보를 삭제 또는 대체하여 나타낸 두 번째 예시의 가명처리 한 정보는 암호화된 핸드폰 번호로 구별되는 추가정보를 가지고 있지만, 해당 정보만으로는 개인 식별이 불가능하므로 가명정보에 해당한다. 즉, 사실상 가명화된 정보만으로는 개인 식별이 불가능하며 추가정보를 활용하게 되면 개인정보로 복원할 수 있으므로 가명정보라 할 수 있다.

다음으로 간접식별정보의 예시를 보면 이름이 이니셜로 대체되고 핸드폰 번호가 삭제되고 주소가 범주화되었으나 생년월일 등 나머지 정보들을 활용하여 충분히 개인 식별이 가능할 것으로 예상되므로 이는 간접식별정보에 해당하게 된다.

마지막으로 추가 컬럼의 범주화 처리 등으로 서울시에 거주하는 가족이 3명이고 카드지출액 4~5백만원인 개인은 한명으로 특정되지 않으므로 다른 정보와 결합하더라도 개인을 식별할 수 없어 익명정보로 정의된다.

〈표 2-1〉 개인정보(원본정보), 가명정보, 익명정보 예시

① 개인정보 예시(원본정보)

이름	생년월일	핸드폰	주소	직업	가족	8월 카드 지출액
김관광	860802	010-999-3333	서울시 강남구 천서로123	국회의원	배우자, 아들1,	4,567,900원

② 가명정보 예시 (사실상 해당 정보만으로는 개인식별 불가능. 다만, 핸드폰 번호 암호화 알고리즘 과 같은 추가정보를 을 알면 특정 개인을 알아볼 수 있으므로 가명정보에 해당)

이름	생년월일	핸드폰	주소	직업	가족	8월 카드 지출액
*삭제	*삭제	*암호화	서울시 강남구 *범주화	*삭제	배우자, 아들1,	456만원

③ 개인정보 예시(원본정보)

이름	생년월일	핸드폰	주소	직업	가족	8월 카드 지출액
KKK	860802	*삭제	서울시 강남구 *범주화	고위공무원	배우자, 아들1,	4,567,900원

④ 익명정보 예시 (그 자체로, 또는 다른 정보와 결합해서도 개인을 식별할 수 없음)

이름	생년월일	핸드폰	주소	직업	가족	8월 카드 지출액
*삭제	*삭제	*삭제	서울시 *범주화	*삭제	가족3명	400만원-500만원

3. 가명정보의 활용범위

가명정보는 통계작성, 연구, 공익적 기록보존 등을 위하여 가명정보를 제공하는 경우에는 개인인 신용정보주체의 동의 없이 가명정보를 활용할 수 있다(개인정보보호법 제32조 제6항, 제9호의 2). 이 경우 통계작성에는 시장조사 등 상업적 목적으로 수행하는 통계작성을 포함하며, 연구에는 대학, 연구소 등 연구기관 뿐 아니라 기업 등이 수행하는 산업적 연구를 포함한다. 다만, 특정 개인을 식별할 수 있는 형태의 통계작성, 연구, 공익적 기록 보존 등의 행위는 모두 허용되지 않는다.

가명정보의 활용 목적

□ **통계 작성:** 통계란 특정 집단이나 대상 등에 관하여 작성한 수량적인 정보를 의미
· 시장조사와 같은 상업적 목적의 통계 처리도 포함
* 직접(1:1) 마케팅 등을 위해 특정 개인을 식별할 수 있는 형태의 통계는 해당하지 않음

※ 통계 작성의 예시
지자체가 연령에 따른 편의시설 확대를 위해 편의시설(문화센터, 도서관, 체육시설 등)의 이용 통계(위치, 방문자수, 체류시간, 연령, 성별 등)를 생성·분석하여 적합한 지역에 신규 편의시설을 선정하고자 하는 경우

□ **과학적 연구**: 과학적 연구는 기술의 개발과 실증, 기초연구, 응용연구 및 민간 투자 연구 등 과학적 방법을 적용하는 연구를 의미
- 과학적 연구는 과학적 방법을 적용하는 연구를 말하며 자연과학, 사회과학, 의료 등 다양한 분야에서 가능
- 여기서 과학적 방법은 체계적이고 객관적인 방법으로 검증 가능한 질문에 대해 연구하는 것을 의미
- 과학적 연구는 기술의 개발과 실증·기초 연구·응용 연구뿐만 아니라 새로운 기술·제품·서비스 개발등 산업적 목적을 위해서도 수행이 가능하며, 민간 투자 연구·기업 등이 수행하는 연구도 가능

※ 과학적 연구의 예시
코로나19 위험 경고를 위해 생활패턴과 코로나19 감염률의 상관성에 대한 가설을 세우고, 건강관리용 모바일앱을 통해 수집한 생활습관, 위치정보, 감염증상, 성별, 나이, 감염원 등을 가명처리하고 감염자의 데이터와 비교·분석하여 가설을 검증하는 경우

□ **공익적 기록보존**: 공공의 이익을 위하여 지속적으로 열람할 가치가 있는 정보를 기록하여 보존하는 것을 의미
- 공공기관이 처리하는 경우에만 공익적 목적이 인정되는 것은 아니며 민간기업, 단체 등이 일반적인 공익을 위하여 기록을 보존하는 경우도 공익적 기록보존 목적이 인정 됨

※ 공익적 기록 보존의 예시
연구소가 현대사 연구 과정에서 수집한 정보 중에서 사료가치가 있는 생존 인물에 관한 정보를 기록·보존하고자 하는 경우

출처: 개인정보위원회a(2020).

제2절 가명처리 및 데이터 결합 개념 및 절차

앞선 1절에서는 가명정보의 이론적 개념과 가명정보의 도입으로 기존의 데이터를 통계 생산 및 연구에 활용할 수 있음을 확인하였다. 하지만 정보주체의 동의 없이 처리할 수 있는 환경이 조성되었다고 하더라도 제약 없이 자유롭게 활용 가능한 것은 아니다. 가명정보를 안전하게 활용하기 위해서는 가명정보처리 시 요구하는 법적 준수사항을 충분히 숙지하여 동의 없는 가명정보의 처리(개인정보보호법 제28조의2) 과정에서 개인정보 오·남용을 방지하고 관련법령[12]을 준수할 수 있도록 주의해야 한다. 이에 본 절에서는 가명처리에 대한 개념과 문화·관광분야의 가명처리 데이터를 활용한 연구 수행에 있어서 개인정보를 침해하지 않는 수준에서 개인 정보를 처리하는 과정, 더 나아가 이종의 가명처리 데이터를 결합하는 절차를 살펴봄으로써, 향후 문화·관광분야의 가명처리 데이터 및 결합 데이터 활용 가이드라인 도출을 위한 검토를 실시하였다.

1. 개인정보 가명처리

가. 가명처리 개념

기술적인 의미에서 '가명처리'란 개인정보 중 주요 식별 요소를 다른 값으로 대체하여 개인 식별을 어렵게 하는 방법을 의미한다. 하지만 개인정보보호법에서 정의하는 '가명처리'란 개인정보를 대체하는 것 외에 일부를 삭제하는 등의 방법으로 추가정보 없이는 특정 개인을 알아볼 수 없도록 하는 것으로 정의하여 개인정보를 가명화하기 위한 모든 비식별 조치를 포괄하는 보다 넓은 개념으로 해석된다. 이에 본 연구에서 가명처란 개인정보의 가명화하기 위해 가능한 모든 비식별 조치로 정의하고자 한다.

12) 개인정보보호법 및 신용정보의 이용 및 보호에 관한 법률

나. 가명처리 대상

가명처리 대상을 정의하기에 앞서 가명정보의 비식별 조치를 어느 대상까지 정의할 것인지를 판단하기 위해서는 개인정보의 유형을 먼저 살펴볼 필요가 있다. 개인정보는 식별의 용의성에 따라 고유식별정보와 준식별정보로 구분할 수 있다.

"고유식별정보"란 개인을 고유하게 구별하기 위하여 부여된 식별정보로서 주민등록번호, 여권번호, 운전면허번호, 외국인등록번호에 해당하는 정보를 의미한다,[13]"준식별정보"는 연령, 성별, 거주 지역, 국적 등과 같이 해당 데이터만으로는 직접적으로 특정 개인을 식별할 수는 없지만, 다른 정보와 결합하여 개인을 식별할 수 있는 정보를 의미한다.[14]

가명처리 대상 및 범위에 대해서는 고유식별정보를 가명화 하면 된다는 의견[15]과 고유식별정보 외 모든 식별정보를 가명처리 대상으로 보는 의견이 존재한다.[16]이러한 차이에도 불구하고 개인정보 보호법에서 가명정보는 '추가정보가 없이는 특정 개인을 알아볼 수 없도록' 명시하고 있으므로 해당 정의를 충족하는 수준에서 가명처리를 진행하면 된다고 하겠다. 즉, 추가정보 없이 식별이 불가능한 수준으로 개인정보를 가명처리 해야 하며 외부의 정보와 결합해서도 식별이 불가능할 수준으로 가명처리가 되어야 함을 의미한다.

다. 가명처리 방법

개인정보를 가명처리하는 방식은 사실상 개인정보 비식별 방식과 기술적으로 비슷한 방법이 적용된다. 이는 가명처리(pseudonymisation), 총계처리(aggregation), 데이터 값 삭제(data reduction), 범주화(data suppression), 데이터 마스킹(data masking)으로 구분되며 주요 내용 및 세부기술, 각 기법의 장단점은 다음과 같이 요약된다.[17]

---

13) 「개인정보 보호법」 제24조제1항 및 「개인정보 보호법 시행령」 제19조 .
14) 개인정보 비식별화에 대한 적정성 자율평가 안내서, 한국정보화진흥원(2019)
15) WP29, Opinion 05/2014 onAnonymisation Techniques, 2014.4, pp.20-21
16) ENISA, Pseudonymisation techniques and best practices, 2019.11, p.21
17) 개인정보 비식별 조치 가이드라인.한국인터넷진흥원(2018)

〈표 2-2〉 가명처리 방법

처리 기법	주요 내용 및 세부 기술	세부기술
가명처리 (Pseudonymisation)	**- 개인정보 중 주요 식별 요소를 다른 값으로 대체하여 개인 식별을 어렵게 하는 방법** (+) 완전 비식별화 가능, 데이터 변형 및 변질이 적음 (-) 대체값을 기준으로 하는 분석에 한계가 있음	휴리스틱 가명화 k-가명화 암호화 교환방법
총계처리 (Aggregation)	**- 데이터 총합 또는 평균값으로 대치하여 데이터 값이 보이지 않도록 하는 방법** (+) 민감한 정보에 대한 비식별화가 가능, 이에 다양한 통계 분석 데이터 마련이 가능함 (-) 집계화 되어 정교한 분석에 한계, 또한 집계 수량이 적을 경우 데이터 결합 시 개인정보 예측 위험이 높음	총계처리기본방식 부분집계 라운딩 데이터재배열
데이터 값 삭제 (Data Reduction)	**- 데이터 셋에서 구성된 값 중에서 필요없는 값 또는 개인 식별에 중요한 값을 삭제하는 방법** (+) 민감한 개인식별 정보에 대하여 완전히 삭제되어 개인정보 예측의 위험이 낮음 (-) 데이터가 삭제되어 데이터 분석이 제한적이며, 신뢰성을 저하시킬 수 있음	속성값 삭제 속성값 부분 삭제 데이터 행 삭제 단순익명화
범주화 (Data Suppression)	**- 데이터의 값을 범주에 값으로 변환하여 명확한 값을 감추는 방법** (+) 범주나 범위는 통계형 데이터 형식으므로 다양한 분석 및 제공이 가능함 (-) 범주, 범위로 표현됨에 따라 정확한 수치 값에 따른 분석, 특정한 분석 결과 도출이 어려우며, 데이터 범위 구간이 좁혀질 경우 추척, 예측이 가능함	범주화 기본방식 랜덤 올림 방법 범위 방법 세분정도 제한방법 제어올림 방법
데이터 마스킹 (Data Masking)	**- 공개된 정보 등과 결합하여 개인을 식별하는데 기여할 확률이 높은 주요 식별자를 보이지 않도록 처리하는 방법** (+) 완전 비식별화가 가능하고, 원시데이터 구조에 대한 변형 최소화 (-) 과도하게 데이터를 마스킹할 경우 필요한 정보를 활용하는 것이 제한적이며, 반대로 낮을 경우 특정한 값의 추적 예측이 가능해짐	임의잡음 추가방법 공백과 대체 방법

출처: 한국인터넷진흥원(2018) 재가공.

1) 가명처리(Pseudonymisation)

가명처리는 개인 식별이 가능한 정보를 직접적으로 식별할 수 없도록 다른 값으로 대체하는 방법으로 주로 이름이나 학교, 근무지 등이 대체의 대상이 된다. 예를 들어, 홍관광, 한국문화관광연구원 재직 중인 사람을 홍길동, 회사원으로 개인정보의 값을 대체하게 되면 개인 식별이 어렵게 됨을 알 수 있다. 이는 완전 비식별화가 가능하고 데이터의 변형이 적다는 장점이 있는 반면, 대체된 값으로 가명 처리 함으로써 고유 특징에 따른 분석을 진행하는데 한계가 있다.

세부 기술로는 식별자에 해당하는 값을 몇가지 규칙을 통해 개인정보를 가명화 하는 휴리스틱 가명화(Heuristic Pseudonymization), 동일한 속성 값을 가지는 데이터를 k 개 이상 유지하여 데이터를 공개하는 K-가명화(K-anonymity), 정보 가공 시 일정한 규칙의 알고리즘을 적용하여 암호화함으로써 개인정보를 대체하는 암호화(Encryption), 기존 데이터베이스의 레코드를 사전에 정해진 외부 변수(항목값)과 연계하는 교환 방법(Swapping) 있다.

2) 총계처리 (Aggregation)

총계처리란, 개인정보에 대하여 통계 값(전체 또는 부분)을 적용하여 특정 개인을 판단할 수 없도록 하는 비식별 방법이다. 이는 개인과 직접 관련된 날짜정보, 수입 및 지출 등의 개인 민감 정보를 비식별화 하는데 주로 적용된다. 예를 들어 이는 개인 민감 정보의 비식별화가 가능하며 연령대별 소득대별 등의 통계 분석 시 유용하게 사용될 수 있다. 다만 데이터 값이 집계화되면서 정교한 분석이 어렵고 집계 수량이 적을 경우 데이터 결합 과정에서 개인 정보 식별 가능성이 높아지게 된다는 단점이 있다.

세부 기술로는 수집된 정보에 민감한 개인정보가 있을 경우 데이터 집단 또는 부분으로 집계(총합, 평균 등)처리하는 총계처리(Aggregation) 방식이 가장 일반적이다. 그 외 다른 데이터 값에 비해 오차 범위가 큰 항목을 통계 값(평균 등)으로 변환하는 부분총계(Micro Aggregation)방식, 집계 처리된 값에 대하여 라운딩(Rounding) 하여 최종 집계 처리하는 방식, 기존 정보값은 유지하면서 개인이 식별되지 않도록 데이터 재배열(Rearrangement)하는 방식이 있다.

3) 데이터 값 삭제(Data Reduction)

데이터 값 삭제는 개인정보 식별이 가능한 특정 데이터 값을 삭제 하여 개인정보를 비식별화 하는 조치를 의미한다. 보통 개인을 식별하기 쉬운 주민등록번호 및 운전면허 번호와 같은 고유 식별정보가 그 대상이 된다. 해당 방법을 통해 민감한 개인 식별 정보와 완전히 삭제되기 때문에 개인에 대한 예측, 추론이 어렵게 된다. 다만 데이터 삭제로 인하여 분석의 다양성, 유효성, 신뢰성을 저하시킨다는 한계가 있다.

세부 기술로는 원시 데이터에서 민감한 속성값 등 개인 식별 항목을 단순히 제거하는

속성값 삭제(Reducing Variables)방법, 민감한 속성값에 대하여 전체를 삭제하는 방식이 아닌 해당 속성의 일부값을 삭제함으로써 대표성을 가진 값으로 보이도록 하는 속성값 부분 삭제(Reducing Partial Variables) 방법, 타 정보와 비교하여 값이나 속성의 구별이 뚜렷하게 식별되는 정보 전체를 삭제하는 데이터 행 삭제(Reducing Records), 식별자뿐만 아니라 잠재적으로 개인을 식별할 수 있는 준식별자를 모두 제거함으로써 프라이버시 침해 위험을 줄이는 준식별자 제거를 통한 단순 익명화(Trivial Anonymization) 방법이 있다.

4) 범주화(Data Suppression)

범주화는 단일 식별 정보를 해당 그룹의 대푯값으로 변환(범주화) 하거나 구간값으로 변환(범위화)하여 고유정보 추적 및 식별을 방지하는 방법을 말한다. 이는 주소나 생년월일등과 같은 식별정보나, 주민등록번호 및 운전면허번호, 기관 및 단체 등의 이용자 계정을 비식별화 하는데 주로 적용된다. 해당 기법은 데이터 값을 범주화 하는 것으로 통계분석이 가능하다는 점이 강점이나 정확한 수치를 이용한 분석이 불가능하기 때문에 분석의 정교성이 떨어지며, 데이터 범위 구간이 좁혀지는 경우 예측이 가능하여 개인 식별의 위험이 존재한다.

세부 기법으로는 데이터의 평균 또는 범주의 값으로 변환하는 범주화(Data Suppression) 기본방식, 개인식별 정보에 대한 수치데이터를 임의의 수 기준으로 올림(round up) 또는 절사(round down)하는 기법으로서 민감성이 높은 정보에 대하여 대표값(범주화)으로 처리하는 랜덤 올림 방법(Random Rounding), 개인식별 정보에 대한 수치데이터를 임의의 수 기준의 범위(range)로 설정하는 기법으로서 해당 값의 분포(범위, 구간)로 표현하는 범위 방법(Data Range), 랜덤 올림 방법에서 어떠한 특정 속성값을 변경시킬 때 행과 열의 합이 일치하지 않는 단점을 해결하기 위해 행과 열이 맞지 않는 것을 제어하여 일치시키는 제어 올림 방법 (Controlled Rounding)이 있다.

5) 데이터 마스킹(Data Masking)

데이터 마스킹은 공개된 정보 등과 결합하여 개인을 식별하는데 기여할 확률이 높은 주요 개인 식별자를 보이지 않도록 처리하여 개인 식별을 하지 못하는 방법을 의미한다.

이는 쉽게 개인을 식별할 수 있는 정보, 고유식별정보 등에 주로 적용되며 완전 비식별화가 가능하기 때문에 원시데이터의 구조에 대한 변형이 적다. 하지만 과도한 마스킹 적용 시 필요한 정보로 활용하기 어려우며, 반대로 마스킹 수준을 낮추게 되면 특정한 값을 추정할 가능성이 높아진다.

세부 기법으로는 개인 식별이 가능한 정보에 임의의 숫자 등 잡음을 추가(더하기 또는 곱하기)하는 임의 잡음 추가 방법(Adding Random Noise), 특정 항목의 일부 또는 정부를 공백 또는 대체문자(‘ * ’, ‘ _ ’ 등이나 전각 기호)로 바꾸는 공백(blank)과 대체(impute) 방법이 있다.

라. 가명처리 절차

개정된 개인정보 보호법에 따르면 ‘개인정보처리자’[18]는 개인정보의 유형, 성격 등을 고려하여 법령을 준수하는 범위 내에서 가명처리 절차와 방법을 자율적으로 판단하여 처리할 수 있다. 하지만 가명처리가 시행 초기임을 고려하여 현장에서 실제 가명처리를 수행할 경우 법적 불확실성을 해소할 수 있도록 가명처리 가이드라인을 따를 것을 권고하고 있다. 본 연구에서는 개인정보위원회에서 발간한 가명정보 처리 가이드라인에서 안내하고 있는 일반적인 가명처리 절차 및 방법을 기반으로 실증 분석을 진행하고자 하였다. 만약 활용하고자 하는 가명정보가 금융·보건·교육 분야인 경우 아래에 소개되는 일반적인 절차 외에도 해당 분야의 법령 및 가이드라인을 우선적으로 준수해야하기 때문에 반드시 해당분야의 가이드라인의 내용을 반드시 확인해야 한다.

개인정보위원회의 가명정보 처리 가이드라인에서 안내하는 가명처리의 절차는 ① 사전준비 ② 가명처리 ③ 적정성 검토 및 추가 가명처리 ④ 활용 및 사후관리의 단계로 구성된다. 개인정보처리자가 개인정보를 가명처리하기 위해 필요한 각 단계별 처리사항은 다음 〈표 2-3〉과 같이 요약된다. 개별 연구자는 가명정보의 특성 · 목적 및 분야별 가이드라인 등을 고려하여 추가절차를 포함하거나 일부 절차를 간략화 할 수 있다.[19]

---

18) 가명정보처리자는 개인정보를 생산하는 주체와는 다른 개념으로, 가명정보 처리자는 업무를 목적으로 개인 정보파일을 운용하기 위하여 스스로 또는 다른 사람들 통하여 개인 정보를 처리하는 공공기관, 법인, 단체 및 개인을 의미한다.

19) 통계법 등 관련법령에 따라 개인정보를 수집·이용하는 경우에는 당해 법령에 따라 처리

〈표 2-3〉 가명처리 단계별 절차

구분	세부절차	주요 내용
사전준비		가명처리 대상 항목 및 처리수준을 정의하고 처리 목적이 적합한지 여부 확인 및 필요한 서류 작성 ※ 가명정보를 제3자에게 제공하는 경우 이용목적 및 방법, 재식별 위험관리 등의 내용을 포함한 계약을 체결 할 수 있음
가명처리	대상선정	개인정보파일에서 가명처리에 필요한 항목 추출 ※ 목적달성에 필요한 최소 항목을 처리하여야 함
	위험도 측정	가명정보처리자의 개인정보 보호수준 및 다른 정보 보유여부 등을 검토하고, 항목별 위험도 분석을 통해 위험도를 측정
	가명처리 수준 정의	가명처리 검토 결과보고서'를 기반으로 가명정보의 활용목적 달성에 필요한 수준을 고려하여 가명처리 수준 정의
	가명처리	'가명처리 수준 정의표'를 기반으로 가명처리 수행
적정성 검토 및 추가 가명처리		'가명처리 수준 정의표'의 기준에 따라 적절한 수준으로 가명처리가 이루어졌는지, 재식별 가능성은 없는지 등에 대한 판단 ※ 가명정보의 적정성 검토는 외부전문가로 구성된 적정성 평가단을 구성하여 검토할 수 있음
활용 및 사후관리		적정하다고 판단되면 가명정보를 본래 활용목적을 위해서 처리할 수 있으며, 법령에 따라 기술적·관리적·물리적 안전조치 이행

출처: 개인정보위원회a(2020)

1) 사전준비

가명처리를 위한 사전준비 단계에서는 가명정보 활용 목적을 명확히 하고 가명처리를 수행할 것인지를 결정하여야 한다. 가명처리의 목적은 법률이 허용하는 범위 내에서 처리하는 목적을 최대한 구체적이고 명확하게 작성하여야 하며 가명정보 활용 위험성 및 법·제도 부합여부, 기술적 가능여부 등을 고려하여 가명처리 여부를 결정한다. 사전 준비단계에서 가명처리를 위탁하거나 가명처리한 정보를 제 3자 제공하는 경우에는 필요에 따라 재식별 금지에 관한 사항, 기타 가명처리 시 유의사항을 포함하여 계약서를 작성할 수 있다. 또한 가명처리에 관한 내부관리계획을 사전에 수립하여야 한다. (시행령 제29조의 5)

2) 가명처리

가명처리 단계는 세부적으로 ①대상선정, ②위험도 측정, ③가명처리 수준 정의, ④가명처리와 같은 4단계로 이루어진다.

① 대상선정

사전 준비 단계에서 수립한 목적을 달성하기 위해 개인정보 파일에서 필요한 항목을 추출하며 이 때 목적달성에 필요한 최소 항목을 처리해야 한다.

〈표 2-4〉 가명정보 처리 대상 정보 추출 예시

개인정보파일	가명처리 목적
이름, 휴대전화번호, 성별, 이메일, 주소, 구매상품, 구매액, 장바구니목록	성별과 지역에 따른 구매액 상관관계를 시계열 분석하고자 함 • 가명정보 처리 대상 항목: 이름, 휴대전화번호, 성별, 주소(시군구), 구매액 * 분석목적과는 상관없는 정보는 대상선정에서 제외

출처: 개인정보위원회a(2020)

② 위험도 측정

가명정보처리자의 개인정보 보호수준 및 다른 정보 보유 여부 등을 검토하고, 항목별 위험도 분석을 통해 위험도를 측정한다.

〈표 2-5〉 위험도 측정

검토사항	내 용
처리제공 환경검토	처리 목적에 따라 처리(제공)환경과 제공받는 자의 개인정보 보호수준 및 다른 정보 보유여부 등을 검토 - 제공받는 부서가 다른(개인)정보를 보유한 경우, 제3자로부터 다른 정보를 받아 함께 활용하는 경우 등을 고려
항목별 위험도분석	추출한 결과 정보의 항목별 위험도를 분석 - 개인식별 가능성이 높은 항목을 분류하여 가명처리 방법 및 수준을 결정하는데 참고할 수 있도록 항목별 위험도를 분석
결과보고서 작성	가명정보처리자는 처리(제공)환경과 항목별 위험도 분석을 참고하여 가명처리에 대한 위험도 평가 결과를 도출

출처: 개인정보위원회(2020)

가명정보는 처리(제공) 환경에 따라 가명정보처리자가 내부에서 활용(자체활용 또는 내부 제공·결합)하는 경우와 제3자에게 제공하는 경우로 구분할 수 있으며, 이에 따라

위험도 측정 결과가 달라질 수 있다. 내부 활용하는 경우는 예를 들어 가명정보처리자가 보유한 개인정보를 가명처리 또는 내부 결합하여 직접 활용 또는 다른 부서에 제공하는 경우를 들 수 잇는데, 이 때는 가명정보취급자의 소속부서에서 이미 보유하고 있는 정보 및 처리시점을 기준으로 제공 받는 다른 정보를 고려하여 위험도를 검토해야 한다. 제3자에게 제공하는 경우는 개인정보처리자가 보유한 개인정보를 가명처리하여 특정 제3자에게 제공하는 것을 의미하며 일반적으로 내부 활용보다 위험도가 높게 측정된다. 제3자의 개인정보보호수준을 고려하여 가명정보 제공으로 인하여 발생할 수 있는 재식별 위험을 최소화하기 위하여 노력하여야 하며 이를 반영하여 위험도를 측정한다. 또한 사전준비 단계에서 작성된 계약서 또는 서약서에 재식별에 대해 명시한 사항이 있다면 이를 고려할 수 있다.

추출한 결과정보의 항목별 위험도를 분석할 때에는 개인식별 가능성이 높은 항목을 분류하여 가명처리 방법 및 수준을 결정하는데 참고할 수 있도록 항목별 위험도를 분석한다. 개인 식별 가능성이 높은 정보는 고유식별정보(여권번호, 외국인등록번호, 운전면허번호), 성명, 전화번호, 전자우편 주소와 같은 식별 정보, 성별, 연령, 국적, 혈액형, 신장, 직업, 위치정보 등 가명정보처리자의 입장에서 개인을 알아볼 수 있는 식별 가능 정보, 국내 최고령, 최장신, 고액 체납금액 고액급여수급자, 희귀 성씨, 희귀 직업 등 특이정보를 예로 들 수 있다. 마지막으로 가명정보처리자는 처리(제공)환경과 항목별 위험도 분석을 참고하여 가명처리에 대한 위험도 평가 결과를 도출하여 결과보고서를 작성하여 관리할 수 있다.

③ 가명처리 수준 정의

가명정보처리자는 '가명처리 검토 결과보고서'를 기반으로 가명정보의 활용목적 달성에 필요한 수준을 고려하여 가명처리 수준을 정의한다.

④ 가명처리

'가명처리 수준 정의표'를 기반으로 가명처리를 수행하며 가명처리 단계에서 생성되는 추가정보는 가명정보와 분리하여 별도로 저장하여야 한다. 추가정보의 분리보관은 가명정보의 기술적 보호조치 관련 법령을 준수하여야 한다.

〈표 2-6〉 가명처리 예시

이름	연락처	주택 구분	법정동 코드	시도	시군구	읍면동	지번	건물명	전세 (천원)	보증금 (천원)	월세 (천원)	전용면적	공급면적
김철수	090-1234-5678	아파트	2636010700	서울특별시	동작구	사당동	1388-4	대우마리나	-	25,000	750	104.00	84.00
이영희	090-2489-3579	오피스텔	3611011000	대전광역시	서구	둔산동	656	푸르지오시티	81,250	-	-	56.45	24.32
박민호	090-9976-5432	아파트	4311410100	부산광역시	해운대구	우동	111-13	평화	125,000	-	-	100.00	84.00

⇓ 추출

(대상선정)

- 목적: 부동산 임대소득 계산 및 인근지역 시세자료 파악을 위한 연구

이름	연락처	주택 구분	시도	시군구	읍면동	지번	전세(천원)	보증금 (천원)	월세(천원)	전용면적	공급면적
김철수	090-1234-5678	아파트	서울특별시	동작구	사당동	1388-4	-	25,000	750	104.00	84.00
이영희	090-2489-3579	오피스텔	대전광역시	서구	둔산동	656	81,250	-	-	56.45	24.32
박민호	090-9976-5432	아파트	부산광역시	해운대구	우동	111-13	125,000	-	-	100.00	84.00

⇓

(위험도 측정)

- 처리환경 검토와 개인정보 항목별 위험도 분류에 따라 가명처리 수준 정의
  1) 처리환경: A사의 부동산 시세정보를 B기관에 제공(계약)
  2) 항목별 위험도 분류: 소유자명, 연락처는 개인정보로 분류하고 가명처리(암호화)

이름	연락처	주택 구분	시도	시군구	읍면동	지번	전세(천원)	보증금 (천원)	월세(천원)	전용면적	공급면적
김철수	090-1234-5678	아파트	서울특별시	동작구	사당동	1388-4	-	25,000	750	104.00	84.00
이영희	090-2489-3579	오피스텔	대전광역시	서구	둔산동	656	81,250	-	-	56.45	24.32
박민호	090-9976-5432	아파트	부산광역시	해운대구	우동	111-13	125,000	-	-	100.00	84.00
(소유자 명, 연락처) + Salt 암호화 식별정보						삭제	라운딩				
						식별 가능정보					

⇓

가명처리

ID	주택 구분	시도	시군구	읍면동	전세(천원)	보증금 (천원)	월세(천원)	전용면적	공급면적
wd4e8502C1qe89rwqe	아파트	서울특별시	동작구	사당동	-	25,000	800	104.00	84.00
r5wle2Sxizi4wd64qwz	오피스텔	대전광역시	서구	둔산동	81,300	-	-	56.45	24.32
ghe6W1525ax40e24jx	아파트	부산광역시	해운대구	우동	125,000	-	-	100.00	84.00

3) 적정성 검토 및 추가 가명처리

이 단계에서는 가명정보처리자가 정의한 가명처리 수준에 따라 적절히 가명처리가 되었는지 확인하고 이 때 가명정보의 적정성 검토는 개인정보처리자의 판단에 따라 외부 전문가로 구성된 적정성 평가단을 구성하여 검토할 수 있다. 가명처리한 가명정보가 가명정보 활용 목적을 달성할 수 있는지 여부 검토하고 특이정보 등의 확인 등 개인 식별 가능성이 있다고 판단되면 가명처리 단계를 재 수행 하여 추가로 가명처리를 해야 한다.

〈표 2-7〉 적정성 검토 사항

구분	주요내용
가명처리의 적정성	**가명정보처리자가 정의한 가명처리 수준에 따라 적절히 가명처리가 되었는지 확인** ※ 가명처리 대상 항목별로 처리결과 전체를 검토(대용량 정보의 경우 중간에 처리 안 된 부분이 있는지 확인 등을 포함함) ※ 개인식별 가능성이 높은 정보 포함 여부 및 가명처리 수준 적정성 등을 검토
목적달성 가능성	**가명정보가 활용목적을 달성할 수 있는지 여부 확인** • 생성된 가명정보로 활용목적 달성이 어렵거나 가명처리 수준이 부족하다고 판단되는 경우에는 가명처리 단계를 반복하거나 부분적으로 추가 가명처리를 수행 • 데이터의 분포, 내용 등을 고려하여 특이정보(식별가능성)가 있다고 판단한 경우 해당 데이터에 대한 적절한 조치를 취하여야 함 ※ 개인정보를 가명처리하여 개인을 알아볼 수 없게 처리했더라도 '특이정보'를 통해 개인식별이 가능한 경우에는 재차 가명처리 필요

4) 활용 및 사후관리

가명정보처리자는 누구든지 특정 개인을 알아보기 위한 목적으로 가명정보를 처리할 수 없으며(법 제28조의5제1항) 가명정보 처리 과정에서 개인식별 가능성이 증가하는지 여부 등을 지속적으로 모니터링 하여 안전하게 처리하여야 한다(법 제28조의5제2항). 또한 특정 개인이 식별되는 경우 즉시 처리중지, 회수, 파기 등 위와 같은 위험을 제거하기 위해 적절한 조치를 수행하여야 하고 관련 법령에서 요구하는 추가정보의 분리 보관, 접근권한의 분리, 기록 작성·보관 및 공개의 의무를 준수하여야 한다.

2. 가명데이터 결합

가. 데이터 결합 개념

데이터 결합은 개별적으로 생산되어 별개의 파일(또는 데이터베이스(DB))로 존재하는 데이터를 하나로 연결하여 통합 파일을 생성하여 새로운 정보를 활용할 수 있도록 하는 방법을 말한다. 다만, 가명데이터의 결합은 개별적으로 생산되어 별개의 파일 데이터 결합은 동일한 '개인정보처리자'[20]내 서로 다른 개인정보처리자에 의해서 각각 관리되고 있는 데이터를 추가 정보 없이 개인을 알아볼 수 없도록 가명처리한 후, 이종 이상의 데이터를 결합하여 새로운 데이터 셋을 생성하는 것을 의미하기 때문에 정확연계의 개념으로 이해할 수 있다.[21] 이 또한 데이터 3법이 개정됨에 따라 정보주체의 동의가 없더라도 과학적 연구 또는 통계작성 등의 목적으로 가명처리하는 것이 허용되었기 때문에 가명처리 데이터 셋을 결합하는 것 또한 가능하게 되었다.

나. 데이터 결합 방법

가명처리 데이터 내 가명처리된 개인의 고유식별자를 기준으로 이종 이상의 데이터를 결합하기 때문에 정확연계로 진행된다. 특정한 고유식별키를 이용하여 정확연계 후 데이터의 정보를 어떻게 조합하는지에 따라서 결합방법은 크게 내부결합(Inner Join)과 외부결합(Outer Join)으로 나뉜다.

1) 내부결합(Inner Join)

내부결합은 각 데이터에 고유식별키(key)가 동일한 경우 값을 가져오는 방법으로 각 데이터의 교집합으로 볼 수 있다. 만약 한 쪽 데이터에만 정보가 있을 경우, 내부결합하게 되면 최종 결합물에 해당 정보는 제외된다. (참고: [그림 2-1])

---

20) 가명정보처리자는 개인정보를 생산하는 주체와는 다른 개념으로, 가명정보 처리자는 업무를 목적으로 개인 정보파일을 운용하기 위하여 스스로 또는 다른 사람들 통하여 개인 정보를 처리하는 공공기관, 법인, 단체 및 개인을 의미한다.

21) 데이터 결합에는 데이터 연결 기준 따라 정확 연계, 판단 연계, 확률적 연계, 통계적 연계, 데이터 5가지로 구분할 수 있으며, 가명처리 데이터는 가명처리 된 개인의 고유식별자를 기준으로 결합되므로 정확연계의 일종으로 볼 수 있다.

2) 외부결합(Outer Join)

외부결합의 경우 여러 데이터 중 한 데이터에만 정보가 있고 다른 데이터에는 정보가 없는 경우 이를 모두 가져오는 방법으로 각 데이터의 합집합으로 볼 수 있으며 다음과 같이 도식화 할 수 있다(참고: [그림 2-2]).

[그림 2-1] 이종 데이터의 내부 결합 예시

(가명처리 전 이종 데이터)

〈기업1(통신회사)〉

성명	전화번호	이동량	콘텐츠사용량
김관광	02-200-6000	43회	623
아뮨화	05-000-0009	87회	305
박정책	02-457-9000	27회	743
정통계	031-123-456	14회	867

〈기업2(카드회사)〉

성명	전화번호	관광지출	문화지출
김문화	02-200-6000	32,014	8,357
강관리	054-000-000	56,042	78,453
박정책	062-000-000	16,125	1,441
정통계	031-123-456	54,863	23,765

(가명처리 후 데이터 내부결합)

결합 → 기업1·기업2

키값	이동량	콘텐츠사용량	관광지출	문화지출
B0001(김문화)	40회	600	32,000	8,000
B0002(박정책)	25회	700	22,000	16,000
B0003(정통계)	10회	800	54,000	23,000

[그림 2-2] 이종 데이터의 외부 결합 예시

(가명처리 전 이종 데이터)

〈기업1(통신회사)〉

성명	전화번호	이동량	콘텐츠사용량
김관광	02-200-6000	43회	623
아뮨화	05-000-0009	87회	305
박정책	02-457-9000	27회	743
정통계	031-123-456	14회	867

〈기업2(카드회사)〉

성명	전화번호	관광지출	문화지출
김문화	02-200-6000	32,014	8,357
강관리	054-000-000	56,042	78,453
박정책	062-000-000	16,125	1,441
정통계	031-123-456	54,863	23,765

(가명처리 후 데이터 외부결합)

결합 → 기업1·기업2

키값	이동량	콘텐츠사용량	관광지출	문화지출
B0001(김문화)	40회	600	32,000	8,000
B0002(박정책)	25회	700	22,000	16,000
B0003(정통계)	10회	800	54,000	23,000
B0004(이문화)	87회	305	-	-
B0006(강관리)	-	-	56,000	79,000

현재까지 진행된 결합 사례들은 대부분 내부결합을 진행한 것으로 확인된다. 향후 외부 결합을 활용할 경우, 정보가 결합되지 않은 부분은 가명처리 데이터 자체만으로도 다양한 분석을 진행할 수 있기 때문에 활용성 확장이 가능할 것으로 예상된다.

다. 데이터 결합 절차

가명처리 데이터는 개인정보보호법 제28조의3에 따라 통계작성, 과학적 연구, 공익적 기록보존 등을 위한 서로 다른 개인정보처리자 간의 가명정보의 결합은 보호위원회 또는 관계 중앙행정기관의 장이 지정하는 전문기관에서 데이터를 결합하여 활용할 수 있다.[22)]

여기서 데이터 결합전문기관은 서로 다른 개인정보처리자 간의 가명정보 결합을 수행하기 위해 개인정보 보호위원회 또는 관계 중앙행정기관의 장이 지정하는 전문기관으로 '21.4월 기준 13개의 전문기관이 지정되었다. (참고: 〈표 2-8〉) 이때 결합하고자 하는 데이터가 개인정보보호법에 의해 관리되는 경우는 '결합전문기관', 신용정보법에 의해 관리되는 경우는 '데이터전문기관'에서 데이터 결합을 진행하게 된다. 만일 결합하고자 하는 데이터 중 하나라도 신용정보법에 의해 관리되는 경우 '데이터전문기관'에서 데이터 결합을 진행하게 된다.

단, 개인정보보호법에서는 신용정보법과 상이하게 '결합키관리기관'이 별도로 지정되어 있다.(시행령 제29조의3제2항) 여기서 '결합키관리기관'이란 결합키 결합정보를 생성하여 결합전문기관에 제공하는 등 가명정보의 안전한 결합을 지원하는 업무를 하는 한국인터넷진흥원(또는 보호 위원회가 지정하여 고시하는 기관)을 말한다.

〈표 2-8〉 결합전문기관

결합전문기관	데이터전문기관 (신용정보법에 따른 금융분야 데이터 전문기관)
(개인정보보호위원회) 통계청 삼성SDS('20.11.27) **(보건복지부)** 건강보험심사평가원, 한국보건산업진흥원, 국민건강보험공단('20.10.29) **(국토교통부)** 한국도로공사 **(과기부)** 한국지능정보사회진흥원, SK주식회사, 더존비즈온('21.1.8)	**(금융위원회)** 금융보안원, 한국신용정보원('20.8.6) 국세청('20.12.22) 금융결제원('21.3.8)

22) 개인정보보호법 제28조의3(가명정보의 결합 제한), 개인정보보호법 시행령 제29조의2(개인정보처리자 간 가명정보의 결합 등), 가명정보의 결합 및 반출 등에 관한 고시

가명정보의 결합 절차 개인정보보호법과 신용정보법 절차의 대부분은 유사하지만 결합키 관리 및 반출 단계 등에서 일부 차이가 존재한다. (참고: 〈표 2-9〉) 따라서 개인정보의 결합 또는 개인신용정보를 포함한 개인정보의 결합 시에는 이를 주의해야 한다.

〈표 2-9〉 개인정보보호법 vs 신용정보법 가명정보 결합절차 비교

개인정보보호법 절차	신용정보법 절차
[사전준비] • 목적설정 • 결합대상 결정 • 가명처리 적합성 검토 • 관리계획 수립	
[가명처리] • 위험도 평가 • 가명처리수준정의 • 적정성검토(심의위원회 선택), (필요시) 추가 가명처리	
[결합신청] • 신청자간 사전협의, 신청자료 준비 • **신청번호 등록** • **결합키, 결합정보 분리 생성(2종)** • **정보전달 (결합키관리기관 + 결합전문기관)**	[결합신청] • 신청자간 사전협의, 신청자료 준비 • **결합키, 결합정보 통합 생성(1종)** • **정보전달 (데이터전문기관)**
[결합] • **결합정보 생성·제공(결합키관리기관)** • 결합 (결합전문기관)	[결합] • 결합 (데이터전문기관)
[반출] • **데이터분석, 추가 가명처리(신청자 수행)** • 반출심사위원회 승인 • 반출	[반출] • **데이터분석, 추가 가명처리(결합기관 수행)** • 적정성 평가위원회 심사 • 반출
[사후관리] • 안전조치 이행 등	
※ 제출서류(가이드라인) • 결합신청서 • 결합신청자 관련 서류 • 결합대상 가명정보 관련 서류 • 결합목적 증명서류 • 안전조치계획 및 이를 증빙할 수 있는 서류	※ 제출서류 ▷ (지침) • 결합신청서 • 결합대상 정보집합물의 기초 자료 - 결합대상 정보집합물 정보 ▷ (결합기관별 운영 실무) • 결합대상 정보집합물 자료 • 이용기관 정보 • 가명처리수행 확약서, 사후관리이행 확약서

출처: 개인정보위원회a(2020).

제3절 가명처리 및 데이터 결합 시 특징

기존에는 개인정보를 가명처리한 정보를 활용할 수 있는 근거가 명확하지 않았기 때문에 데이터를 활용하고 연구하는데 제약이 있었고, 데이터를 보유한 공공기관 또는 민간 기업은 데이터 공개에 보수적이었다. 특히 국내에서는 개인정보유출 위험에 대한 우려로 개인정보보호의 중요성이 매우 강조되었기 때문에 연구목적으로 데이터를 제공하고 협력관계를 구축하여 공동연구를 수행하는 것에도 매우 소극적이었다. 하지만 가명정보 활용에 대한 법적 기반이 마련되고 개인정보를 가명처리하여 통계생산 및 연구에 활용할 수 있게 됨에 따라서 네 가지 측면에서 데이터의 활용 범위가 확대될 것으로 예상해 볼 수 있다.

1. 개인단위의 원시자료 활용 범위 확대

개인정보를 가명처리하는 경우 기존의 익명의 집계데이터로 생산했던 데이터도 개인단위의 원시자료 형태로 활용할 수 있게 된다. 이 경우 총량으로만 생산되던 통계에 개인의 특성을 고려한 미시적 접근이 가능해진다. 예를 들어 기존 신용카드 지출액 자료의 경우, 익명화된 집계 데이터를 활용하여 지출 총액의 추세 분석할 수 밖에 없었다. 이에 업종별 1인당 카드 지출액을 산출할 수 없었으며, 대신 카드사용카드 건수를 보조자료로 이용하여 1인당 카드 지출액을 대략적으로 파악하는 수준에 머물러 있었다. 하지만 이를 개인화 하면 1인당 카드지출액을 추산할 수 있어 집단별 분석결과를 비교할 수 있게 된다. 뿐만 아니라 보다 인과관계 분석(casual effect analysis)이나 예측모형 분석(prediction analysis)과 같은 통계적 모델링을 사용할 경우 보다 정교한 분석이 가능해질 것으로 예상된다.

〈표 2-10〉 가명처리 데이터 활용 가능 전 vs 후 확보 가능한 데이터 구조 변화 예시

(가명처리 데이터 활용 전) 익명화된 집계 데이터 → (가명처리 데이터 활용 후) 개인단위 데이터

날짜	업종구분	지출액
8월	관광	11466
9월	관광	3466
10월	관광	1433
9월	문화	512
10월	문화	964
11월	문화	1346

성명	전체지출	관광지출	문화지출
김관광	520	400	650
이문화	600	450	450
최예술	650	500	520
박정보	100	150	115
문정책	140	145	100
하통계	200	300	450

참고: 본 내용은 데이터 일부를 발췌한 것이 아닌 저자가 작성한 가상의 예시임

2. 시의성 높은 대규모 패널 데이터 구축

민간사업체는 고객으로 등록되어있는 기간 내 정보를 시계열로 수집·보관하고 있다. 따라서 개인정보보호법 개정 전, 익명화된 집계 데이터로만 활용되던 민간데이터를 개인단위 데이터로 활용할 수 있게 되는 경우 많은 표본 수 확보와 더불어 시의성 높은 대규모 개인 패널 데이터를 구축할 수 있게 된다. (참고: 〈표 2-11〉)

〈표 2-11〉 가명처리 데이터 활용 가능 전 vs 후 활용 민간 데이터 활용 변화 예시

(가명처리 데이터 활용 전) 익명화된 집계 데이터 → (가명처리 데이터 활용 후) 개인단위 패널 데이터

날짜	업종구분	지출액
8월	관광	11466
9월	관광	3466
10월	관광	1433
9월	문화	512
10월	문화	964
11월	문화	1346

성명	Time	관광지출	문화지출
김관광	**8월**	400	650
김관광	**9월**	450	450
김관광	**10월**	500	520
이문화	**8월**	150	115
이문화	**9월**	145	100
이문화	**10월**	300	450

참고: 본 내용은 데이터 일부를 발췌한 것이 아닌 저자가 작성한 가상의 예시임

뿐만 아니라 개인의 가족정보를 활용한다면 가구 단위의 패널 데이터를 구축하는 방향으로 확장이 가능할 것으로 보인다. 만약 가구 구성원 모두 동일한 민간 사업체의 고객인 경우 각 개인의 가구정보를 이용·연결하여 하나의 가구 데이터를 구축할 수 있다.

다만 가구 구성원 모두 민간사업체 고객으로 등록되어 있어야 하므로 현실적으로 가구 단위의 패널을 구축하는데 비교적 많은 시간과 노력이 요구될 것으로 보인다.

이어서 현재 조사기반으로 생산되고 있는 패널데이터는 데이터 수집 및 공표까지 걸 약 1~2년이 소요되기 때문에 비교적 적시성 높은 통계를 생산하기 어렵다. 하지만 민간 데이터의 경우 데이터를 가공하고 반출까지 비교적 짧은 기간이 소요되기 때문에, 이를 활용할 경우 매우 시의성 높은 패널 데이터를 구축할 수 있게 된다.[23] 이를 통해 개인정보를 가명처리하여 데이터화 하게 된다면 보다 시의성 높은 패널 데이터를 생산할 수 있어 환경 변화에 빠르게 대처할 수 있는 시의적절한 정책마련에 유의미한 근거자료로 활용 될 수 있을 것이다.

다만 민간 데이터의 경우 내부 고객이 탈회한 경우, 기존에 축적되어있는 개인의 내부 정보를 모두 삭제되기 때문에 균형패널(Balance panel)이 구축된다. 다만 민간 데이터의 유효표본수 규모가 비교적 많기 때문에 데이터 제공기간의 고객의 이탈이 발생하더라도 패널 데이터를 활용한 종단분석에 큰 영향을 주지 않을 것으로 예상된다.

3. 분석 가능한 유효표본 수 확대

이어서 가명처리 데이터를 활용할 경우 기존 보다 비교적 많은 표본을 확보할 수 있어 보다 다양한 분석이 가능해진다, 현재 개인단위로 활용할 수 있는 데이터 대부분이 조사통계 자료인데, 사실상 조사를 통해 많은 표본을 확보하기 위해서는 많은 예산과 시간이 소요된다. 이에 대부분의 조사통계는 가용예산과 조사 소요시간 등 조사에 필요한 제반 여건과 작성되는 통계의 가장 핵심적인 지표의 과거 추정량에 대한 상대표준오차(Relative standard error: RSE)를 이용하여 표본 크기를 결정하는 경우가 많다. 즉, 개인의 세부특성 별 분석에 초점을 맞춰서 표본 설계를 하기 보다는 주요 지표 분석에 초점을 맞춰 가장 효율적으로 수준으로 표본의 크기를 결정하게 된다. 하지만 이 경우, 제한적인 표본으로 각 통계별 주요지표의 대표성을 확보할 수 있으나, 분석 대상을 개인의 특성별로 세분화하는 경우 대표성 있는 결과를 기대하기 어렵다.

---

23) 민간데이터마다 시점은 다르나 신용카드 지출액의 경우, 통상적으로 내국인 지출액 데이터는 3일, 외국인 데이터는 4일 이후 분석 가능한 형태의 데이터로 구성할 수 있다.

한편 기존의 문화관광분야에서 주로 활용되는 이동통신 및 카드지출액과 같은 민간 데이터의 경우 내부 고객정보를 파악하고 있기 때문에 매우 많은 표본 수 확보가 가능하지만 개인정보보호법으로 인해 개인단위의 데이터에 대한 접근이 불가능하였다. 따라서 이를 원시 데이터(Raw Data)가 아닌 개인정보를 익명화하여 가공된 집계 데이터를 기반으로 총량 위주의 통계를 생산하여 추세(Trend)를 파악하는 데 제한적으로 활용되었다.

하지만 개인정보보호법의 개정으로 데이터를 가명처리하고 이를 연구목적에 활용할 수 있게 됨에 따라 기존의 다양한 민간 데이터를 익명화된 집계 데이터가 아닌 개인단위의 원시데이터를 대상으로 분석할 수 있어 대규모 표본 데이터를 확보할 수 있게 되었다. 예를 들어, SK텔레콤과 신한카드사의 고객의 개인정보를 가명화 하여 개인 단위의 데이터를 활용할 수 있게 되면 가입자 각각 2,500만명, 850만명의 개안단위 데이터를 확보할 수 있으며, 각 데이터를 결합 하더라도 SK텔레콤과 신한카드 동시가입자 약 350만 표본을 확보할 수 있다.

다음 〈표 2-12〉 은 통계청의 인구총조사 자료와 문화체육관광 승인통계 중 가장 규모가 큰 국민여행조사의 표본수와 SK텔레콤과 신한카드 동시 가입자 수를 비교한 결과이다. 이를 통해 민간데이터를 개인 단위의 원시데이터 형태로 활용한다면 개인의 특성을 세분화하더라도 충분한 표본 수를 확보할 수 있음을 유추해 볼 수 있다.

〈표 2-12〉 가명처리 데이터 확보 후 유효표본 수 확대에 따른 확보 가능한 표본 수

(단위: 천명)

2020 인구총조사		sk, 신한 결합데이터		2020 국민여행조사	
전체 표본 수	50,710	전체 표본 수	3,449	전체 표본 수	51
남성	25,251	남성	1,655	남성	25
여성	25,459	여성	1,794	여성	25
MZ세대	18,855	MZ세대	1,497	MZ세대	19
그외세대	28,988	그외세대	1,952	그외세대	32
		1인 가구	1,113	1인 가구	7
		2인 가구	740	2인 가구	13
		3인 이상 가구	1,049	3인 이상 가구	31

유효표본 수 확대에 따라 기대되는 또 다른 강점 중 하나는 시군구 단위 통계 생산이 가능해 진다는 점이다. 기존의 조사통계 데이터는 제한된 표본 수 때문에 17개 시도 단

위의 통계를 생산하는데 머물렀다. 또한 기존의 민간데이터의 경우 시군구 단위의 통계가 산출되었지만 표본 수를 알 수 없기 때문에 총량위주의 통계 생산에 머물러 있었다. 다음 〈표 2-13〉는 인구총조사와 SK텔레콤과 신한카드 결합 데이터 지역별 표본 수를 비교한 것이다. SK텔레콤과 신한카드 결합 데이터를 이용하면 시군구 까지도 충분한 표본확보가 가능하여 소지역 통계 생산에 유의미한 자료로 활용될 수 있음을 알 수 있다.

〈표 2-13〉 데이터 결합 후 확보 가능한 지역 단위 유효표본 현황

2020 인구총조사(단위: 천명)

순위	행정지역	표본수
1	경기도 화성시	881
2	경기도 부천시	833
3	경기도 남양주시	696
4	제주특별자치시	671
5	서울특별시 송파구	643
…	….	….
246	강원도 양구군	21
247	전라북도 장수군	21
248	인천 옹진군	19
249	경상북도 영양군	16
250	경상북도 울릉군	8

sk, 신한 결합데이터(단위: 명)

순위	행정지역	표본수
1	경기도 화성시	64,487
2	서울특별시 송파구	62,392
3	서울특별시 강남구	59,300
4	서울특별시 강서구	56,797
5	경기도 부천시	54,718
…	….	….
246	전라북도 장수군	427
247	전라북도 진안군	408
248	경상북도 영양군	397
249	경상북도 군위군	394
250	경상북도 울릉군	339

다만, 개인화된 민간 데이터 활용이 가능하다고 해서 기존 민간데이터가 가지는 표본의 대표성 문제가 해결되지 않는다.[24] 이러한 점을 보완하기 위해서는 현재 민간데이터가 가지는 표본 대표성 문제해결이 선행적으로 이루어져야할 것이다.

24) 통신데이터를 예를들면, 특정 업체 가입자가 전국에 거주하는 국민을 대표할 수 있다고 보기는 어렵다. 통신 3사(SK텔레콤, KT, LGU) 가입이 개인이 임의로 배정받는 것이 아니라 개인의 선택에 의해 결정되므로 선택에 의한 표본의 편의가 발생하게 된다.

4. 이종 데이터 결합을 통한 활용 정보 확대

가명정보 및 결합제도의 도입으로 데이터 결합에 활용할 수 있는 가능성이 열리게 됨에 따라 은행·카드·보험·금융투자 등 금융업권의 데이터와 통신정보·위치정보·보건의료정보 등 이종 산업분야에서 관리되고 있는 다양한 형태의 데이터를 서로 결합하여 분석할 수 있게 되었다. 결합합분석은 가명정보 도입으로 가장 기대되는 이점 중 하나로 결합전문기관을 통해 결합한 가명정보는 데이터의 가치가 더욱 높아지고 결합정보를 활용해 기존에는 할 수 없었던 연구를 수행하여 새로운 지식과 가치를 발견하고 새로운 비지니스 기회 발굴 등 미래 혁신 성장에 기여할 수 있다.

데이터 결합는 개별적으로 생산되어 별개의 파일(또는 데이터베이스(DB))로 존재하는 데이터를 하나로 연결하여 통합 파일을 생성하여 새로운 정보를 활용할 수 있도록 하는 방법을 말하는데 데이터 결합에는 데이터 연결 기준 따라 정확 연계, 판단 연계, 확률적 연계, 통계적 연계, 데이터 5가지로 구분할 수 있다.(참고: 〈표 2-14〉)

정확연계는 서로 다른 데이터 파일에서 동일한 대상을 찾아 연결하는 방법으로 각 개체를 정확하게 파악할 수 있는 고유식별정보가 있어야 하며 정확하면서 매우 효율이 좋은 방법이다. 통계적 연계는 서로 다른 데이터 파일에서 유사한 성향을 가진 데이터를 연결하는 방법으로 정확 연계보다 정확성이 부족하고 사용하기 위한 조건이 까다로우며, 통계적 연계를 적용하기 위한 기본 가정을 충족해야 한다. 일반적으로 개인식별정보가 없는 데이터, 조사 자료 등에서 추가적으로 알고 싶은 정보가 있을 때 적용하는 방법이다.

〈표 2-14〉 데이터 연계유형

유형	내용	연결 기준
정확 연계	서로 다른 데이터셋이 같은 고유식별자(예: 주민번호)를 사용할 때 해당 식별자를 매개(연계키)로 두 레코드를 연결	고유식별정보
판단 연계	연구자의 데이터에 대한 이해를 바탕으로 연결	연구자의 지식
확률 연계	정확 연계에서 일치하지 않는 케이스들이 발생할 때, 각 변수들의 연결 가능성을 계산하여 연결	공통변수로 계산한 가능성
통계적 연계	통계적 방법으로 가장 유사한 케이스를 찾아 두 레코드를 연결	유사성 측도
데이터 연결	둘 이상의 파일에서 변수들간의 연간성이 있을 경우 하나가 변화할 때 같이 변화가 가능하도록 연결	변수들과의 연계성

출처: 박근화(2018).

기존에는 개인정보보호 문제로 개인단위의 정보를 결합하는 건이 제한되어 결합키(고유식별정보)를 이용하지 않고 서로 다른 자료에서 유사한 개체를 결합하는 통계적 연계(statistical matching)를 이용하여 연계할 수밖에 없었다. 하지만 가명정보의 결합은 식별할 수 있는 연계키를 활용하여 각기 다른 두 가지 형태의 데이터를 정확 연계하는 것으로 〈표 2-15〉 에서와 같이 성명 및 전화번호를 연계키로 활용하여 개인의 이동과 온라인콘텐츠 사용이력과 카드 지출 정보를 정확하게 연계하여 이동량 특성에 따른 소비 지출현황을 보다 정교하게 분석할 수 있다. 이처럼 기존 연구에 비해 연계의 효율성과 정확도를 높여 연구의 질을 높일 수 있게 되었다. 가명정보 도입으로 서로 다른 개인정보 처리자가 처리한 가명정보를 결합전문기관을 통해 정확연계(exact matching)의 방법으로 결합이 가능하게 되면서 더 정확한 정보를 분석에 활용할 수 있게 되는 장점이 있다.

〈표 2-15〉 가명처리 데이터 활용 가능 전 vs 후 이종데이터 결합을 통한 정보 확대 예시

(가명처리 데이터 활용 전)

〈기업1(통신회사)〉

성명	전화번호	이동량	콘텐츠사용량
권율	02-200-6000	43회	623
이순신	02-457-9000	27회	743
홍길동	031-789-0123	14회	867

〈기업2(카드회사)〉

성명	전화번호	관광지출	문화지출
권율	02-200-6000	32,014	8,357
이순신	02-457-9000	22,123	16,456
홍길동	031-789-0123	54,863	23,765

↓

(가명처리 데이터 활용 후)

결합 → 기업1·기업2

키값	이동량	콘텐츠사용량	관광지출	문화지출
B0001(권율)	40회	600	32,000	8,000
B0002(이순신)	25회	700	22,000	16,000
B0003(홍길동)	10회	800	54,000	23,000

참고: 본 내용은 데이터 일부를 발췌한 것이 아닌 저자가 작성한 가상의 예시임

제4절 가명처리 및 데이터 결합 시 고려사항

가명정보를 안전하게 활용하기 위해서는 가명정보처리 시 요구하는 법적 준수사항을 충분히 숙지하여 동의 없는 가명정보의 처리(개인정보보호법 제28조의2) 과정에서 개인정보 오·남용을 방지하고 관련 법령을 준수할 수 있도록 주의해야 한다. 이에 개인정보를 보호하기 위한 개인정보 처리 시 개인정보 처리를 위한 안전조치의무, 관련된 금지의무 등을 검토해야 한다. 또한 데이터 결합 시 발생될 수 있는 다양한 문제에 대해서도 심도 있는 검토가 필요하다. 마지막으로 가명 데이터 활용 및 데이터 결합 후 분석 과정에서 발생되는 문제 또한 고려해야한다.

1. 개인정보 가명처리 시 고려사항

가. 개인정보 처리를 위한 안전조치 의무

가명정보 역시 개인정보이기 때문에 개인정보 보호법에 따라 개인정보의 안전조치 의무를 다해야한다. 개인정보 보호법에 따르면 개인정보처리자[25]는 특정 목적으로 가명정보를 활용하되 3자에게는 개인을 알아볼 수 없는 통계 데이터 형태로 제공해야 한다고 명시하고 있다. (개인정보 보호법 「제28조의제1항」, 「제28조의제2항」) 개인정보처리자는 가명데이터를 이용할 때 아래와 〈표 2-16〉 같은 의무사항을 지켜야 한다.(개인정보 보호법 「제28조의4제1항」 및 「제29조의5제1항」) 개인정보처리자는 가명정보를 다루는데 개인정보에 준하는 조치들을 취해야 한다.

25) "개인정보처리자"란 업무를 목적으로 개인정보파일을 운용하기 위하여 스스로 또는 다른 사람을 통하여 개인정보를 처리하는 공공기관, 법인, 단체 및 개인 등을 말한다(개인정보보호법 제2조)

〈표 2-16〉 개인정보처리자의 가명정보 조치방안

조치	설명
안전성 확보	가명정보를 처리하기 위한 시스템 안전성 확보
가명정보와 추가정보의 분리 보관	가명정보와 추가정보는 분리되어 보관되어야 하며 추가정보가 불필요한 경우 추가정보를 파기해야 함
가명정보와 추가정보의 접근 권한 분리	가명정보와 추가정보의 접근 권한은 분리되어야 함. 다만 소상공인법 제2조에 따라 소상공인으로 가명정보를 취급할 자를 추가로 둘 여력이 없고 접근 권한의 분리가 어려운 정당한 사유가 있을 경우 업무 수행에 따른 최소한의 접근 권한만 부여하고 접근 권한의 보유 현황을 기록으로 관리하는 등의 형식으로 접근 권한을 관리/통제 해야 함

나. 가명처리 시 금지의무

다음은 가명데이터를 처리하는 과정에서 하지 말아야 할 사항이다. 우선 개인정보 보호법에 의하면 누구든지 특정 개인을 알아보기 위한 목적으로 가명정보를 처리해서는 안된다 규정하고 있다.(개인정보 보호법 「제28조의 5제 1항」) 즉, 반드시 특정 개인을 식별할 수 있어서는 안되며 이를 위반하는 경우 5년 이하의 징역 또는 5천만원 이하의 벌금에 처해진다.(개인정보 보호법 「제71조 제4호의 3」) 또한 이를 위반하여 개인을 알아볼 수 있는 정보가 생성되었음에도 이용을 중지하지 않거나 이를 회수/파기하지 않은 자에게는 3천만원 이하의 과태료가 부과된다.(개인정보 보호법 「제75조제2항 제7호」). 따라서 반드시 가명정보를 운영하거나 다루는 주체는 특정 개인을 식별하려는 시도나 노력을 해서는 안된다.

2. 분석 과정 시 고려사항

가. 데이터 활용의 범위 축소

가명처리는 '개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 등의 방법으로 추가정보 없이는 특정 개인을 알아볼 수 없도록 처리하는 것'을 의미한다. 가명정보만으로 특정 개인을 알아보거나 추가정보를 결합하여 개인을 식별할 수 있다면 가명처리가 잘못된 경우이다. 이 때문에 가명정보를 결합하기 전에 데이터를 식별할 수 없도록 처리하는데 가명처리 기술과 가명처리로 인하여 데이터 활용의 어려움이 발생할 수 있

다. 여기서 가명처리 방식에 대한 공식이나 규제는 따로 없기 때문에 처리해야 하는 데이터 크기나 분석 방향에 따라 알맞은 기술을 사용한다. 가명처리를 위한 다양한 기술이 있지만 많이 사용하는 기술을 사용하여 처리하면 다음과 같다.

첫째, 개인정보를 범주화(비슷한 부류로 묶는 것) 하는 것이 대표적인 방법 중 하나이다. 예를 들어 나이의 경우 숫자 데이터이기 때문에 올림, 내림, 반올림 등을 이용하여 라운딩(Rounding) 처리하여 집계화 한다. 두 번째로, 문자 데이터를 범주화 하는 방법이다. 문자데이터를 범주화 하는 경우는 '고양이', '개'를 '동물'로 표현하는 것처럼 문자 정보의 상위 개념으로 범주화 한다. 세 번째는 핸드폰 번호의 가명처리의 방법이다. 핸드폰 번호는 그 자체로 식별이 되기 때문에 부분삭제 기술을 사용한다. 핸드폰 번호 뒤 4자리만 남기고 나머지를 삭제하면 개인을 식별할 수 없게 된다. 또한 다른 데이터에 비해 확연하게 구분되는 데이터가 있다면 그 행 항목을 전체 삭제하거나 특이치 데이터만 삭제할 수 있다. 〈표 2-17〉에서 제시된바와 같이 가명처리 후 데이터의 경우 여가비가 다른 사람들에 비해 유난히 높은 박명수님의 여가비를 삭제함으로써 특정인의 식별을 최소화하는 방안을 검토해야 한다.

〈표 2-17〉은 최종 가명처리된 데이터이다. 이를 비교해 보면 가명처리가 되면서 데이터가 많이 달라진 것을 알 수 있다. 대표적으로 주소가 다른 개인이 같은 분류로 구분되면서 분석의 정밀도가 떨어질 것을 쉽게 예상할 수 있다. 가명데이터 목적이 개인 마케팅이 아닌 통계작성, 과학적 연구, 공익적 기록보존이 목적이기 때문에 감수해야 하지만 분석하는 입장에서 데이터 정밀도 문제에 대한 신중한 고려가 필요하다고 하겠다.

나. 고급 통계 분석의 한계

본 연구에서는 결합정보를 가지고 머신러닝이나 딥러닝 기술을 이용하여 예측이나 분류 작업을 진행하는 것이 가능해졌다고 언급하였지만 현실적으로 어려운 부분 또한 존재한다. 앞에서 설명한 바와 같이 가명, 익명 처리 과정에서 데이터 정밀도가 떨어지고 이로 인하여 학습 데이터들이 패턴이 무뎌지게 된다. 또 가명데이터 분석을 할 수 있는 시스템이 머신러닝이나 딥러닝을 돌리기에 충분한 환경이 아닐 확률이 높다. 가명처리를 위한 시스템은 기존 시스템과 분리되어야 하므로 때문에 오직 가명처리를 위한 시스템을 따로 구축해야 하는데 데이터 이용기관에 따라 다르겠지만 이 목적 하나만을 위해

시스템에 큰 투자를 하기가 쉽지 않다. 따라서 가명 결합 데이터를 이용하여 인과관계 분석 및 예측모형을 활용하는 것은 결과의 신뢰도 측면에서 신중하게 접근해야 한다. 그럼에도 불구하고 가명처리된 개인화된 데이터의 시계열적 축적 및 예측분석에 활용할 수 있는 방안마련은 꼭 필요할 것이다.

〈표 2-17〉 개인정보 비식별을 위한 가명처리 예시

[가명처리 전 데이터]

이름	연령	주소	핸드폰번호	여가비
김창수	27	서울시 용산구 서빙고로 17	010-1111-2222	35,000
이영자	39	서울시 용산구 신계동 6-1	010-1111-3333	27,000
박명수	24	성남시 분당구 구미동 7-2	010-5555-2222	590,000
최영희	31	서울시 용산구 녹평대로 11길 54	011-7777-4444	63,000

[가명처리 후 데이터]

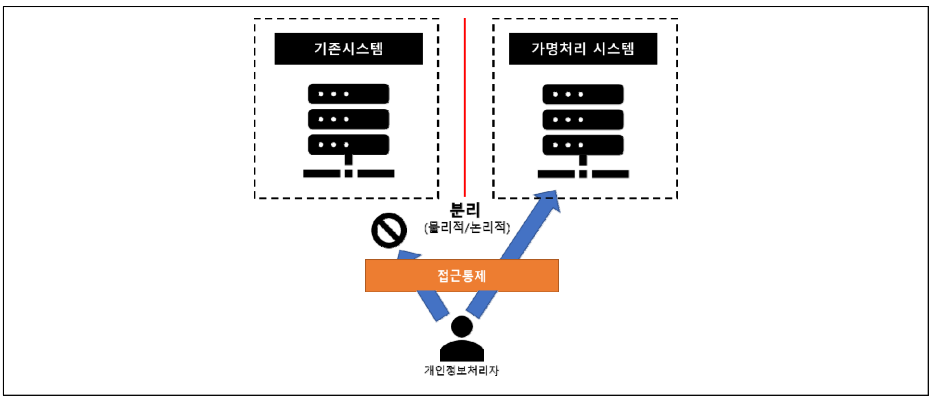
이름	연령대	주소	핸드폰번호	여가비
김창수	**20**	서울시 용산구	2222	35,000
이영자	**30**	서울시 용산구	3333	27,000
박명수	**20**	성남시 분당구	2222	-
최영희	**30**	서울시 용산구	4444	63,000

참고: 본 내용은 데이터 일부를 발췌한 것이 아닌 저자가 작성한 가상의 예시임

다. 데이터 분석 환경 구성의 어려움

개인정보처리자는 개인정보보호법과 개인정보보호법 시행령에 따라 가명정보를 안전하게 관리하기 위한 내부 관리계획을 수립하고 실행해야 한다. 아래 [그림 2-3]은 가이드에 따른 분석환경을 도식화 한 것으로 가명처리 시스템은 기존 시스템과 물리적으로 분리되어야 한다. 물리적 분리가 어려운 경우는 엄격한 통제 정책을 마련하고 논리적으로 분리하여야 한다. 다음으로 가명 결합에 접근 권한을 부여해야 한다. 가명정보취급자 혹은 개인정보처리자는 기존 시스템에 접근할 수 없어야 하고 가명처리 시스템에 접근할 수 있는 인원은 최소한으로 구성해야 한다. 이 모든 것은 엄격한 접근 통제 하에서 진행되어야 한다.

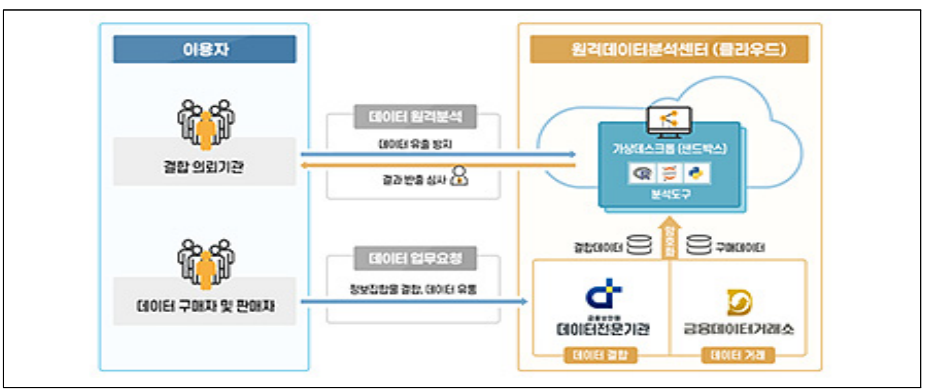
[그림 2-3] 가명정보 가이드라인에 따른 분석환경



출처: 금융보안원 가명정보 분석환경 가이드라인

하지만 이러한 통제 시스템을 구축할 수 없다면 데이터분석센터를 활용하는 방법이 있다. 가명정보를 이용하려는 사람은 클라우드 비용을 지불하고 결합데이터를 분석할 수 있다. 아래 [그림 2-4] 는 금융보안원에서 제공하는 클라우드 기반 원격 데이터분석 시스템이다. 이 시스템의 장점은 복잡한 가명처리 시스템을 구성하지 않고 결합신청 과정에서 관여했던 결합데이터에 접근할 수 있다는 것이다. 그리고 R 스튜디오(R Studio)나 쥬피터 노트북(Jupyter Notebook)을 제공하기 때문에 클라우드에서 분석하고 분석 결과를 반출할 수 있다. 데이터 반출 과정에서 정보보호법에 위반되는 것을 막기 위해서 반출심사 후 분석 결과만 반출 가능하고 가명정보에 대한 기술적 보호조치를 준수할 수 있기 때문에 편하게 데이터를 활용할 수 있다.

[그림 2-4] 금융보안원 원격데이터분석센터를 통한 분석 환경



출처: 금융보안원 가명정보 분석환경 가이드라인

라. 분석결과 활용

가명처리 데이터의 분석 결과는 통계, 집계화된 형태로 활용할 필요가 있다. 첫째 가명결합의 대상은 우리나라 국민으로 우리나라 인구를 생각해 보면 못해도 십만, 많으면 천만대 데이터를 처리해야하기 때문에 통계, 집계화 하지 않고 사용하는 것은 현실적으로 어렵다. 둘째로 개인 식별의 위험성을 방지할 수 있다. 구조적으로 결합하는 과정에서 개인 식별의 위험이 줄어들지만 혹시나 통계, 집계가 아닌 개별 데이터 형태로 활용하는 과정에서 개인정보보호법을 위반할 가능성이 존재한다. 따라서 통계, 집계화를 통해 가명처리의 목적에 맞게 분석결과를 활용할 수 있도록 유의해야한다.

가명처리데이터의 분석이 확대되면서 최근 참여연대·민주사회를 위한 변호사모임(민변) ·서울YMCA·진보네트워크센터가 SK텔레콤을 상대로 개인정보 가명처리 중지를 요구하는 소송을 제기하였다. 소송 단체들은SK텔레콤에 보유 개인정보를 가명처리했는지 여부와 가명처리를 했다면 그 대상이 된 사람이 개인정보 일체를 열람할 수 있는지에 대한 의견을 물어봄과 동시에 가명처리 중단도 함께 요구했다. 향후 가명처리데이터의 활용도가 더욱 커짐에 따라 이러한 개인정보관련 이슈도 같이 발생될 것이다. 기존에 발생된 문제점을 깊이 있게 파악하고 대응할 수 있는 방안 마련도 필요할 것이다.

마. 표본의 대표성 문제

표본의 대표성 문제는 기존의 민간데이터 활용 시 발생하게 되는 문제 중 하나이다. 이는 가명처리를 하게 되더라도 해결되지 않는다. 예를 들어 SK텔레콤 가입자를 대상으로 한 가명처리 데이터를 분석하여 1인당 평균의 이동량을 산출했을 때, 이는 '전국민 1인 당' 평균 이동량이 아닌 'SK텔레콤 가입자 1인당' 평균 이동량으로 해석된다. 통신사 가입은 임의할당(Random)되는 것이 아닌 개인의 선택에 의해 결정되므로 표본 자체에 선택편의(Selection Bias)가 존재한다고 볼 수 있다. 따라서 표본 대표성 문제를 해결하기 위한 추가적인 연구가 필요하다고 하겠다.

가명처리된 빅데이터의 표본 대표성 문제를 해결하고 활용도를 높이기 위해 한국문화관광연구원에서는 2017년 '관광분야 빅데이터 활용체계 및 실증분석 연구(권태일·이충희)' 등을 통해 해결방안 등을 제시하였다. 향후 가명처리된 개인데이터의 경우 표본의 거주지 정보파악을 토대로 통계청의 조사구·집계구 기준의 표본설계를 실시하고 결과값

을 추정함으로써 기존에 제기된 임의할당으로 인한 대표성의 문제를 해결하고 보다 객관적인 데이터로써의 활용방안을 모색해야 할 것이다.

3. 데이터 결합 시 고려사항

개인정보 외에도 데이터 분석 시, 데이터를 결합하는 과정에서 문제가 발생할 수 있으며 구체적인 내용은 다음과 같다.

가. 결합률

결합신청자의 개인정보로 결합키를 만들었을 경우 결합률이 떨어지는 경우가 많다. 대부분 한쪽, 혹은 양쪽의 결합신청자의 데이터가 최신 정보로 갱신되지 않기 때문에 발생하는 문제들이다. 다음 〈표 2-18〉 와 같은 경우 결합률이 떨어지게 된다. 결합키는 결합신청자들이 가지고 있는 데이터 중에서 중복되지 않는 특별한 데이터를 사용하는데 아래 예시에서는 김관광님의 전화번호 정보를 한쪽에서 업데이트 하지 않았기 때문에 이 경우에는 결합이 되지 않는다. 따라서 데이터 결합률을 높이기 위해서는 가장 최신 정보가 반영될 수 있도록 사전적 노력이 요구된다.

〈표 2-18〉 결합 되지 않는 경우(예시)

구분	이름	주소	전화번호
결합신청자	김관광	서울시 용산구	010-xxxx-1111
	이문화	성남시 분당구	010-xxxx-7777
결합신청자	김관광	서울시 용산구	010-xxxx-2222
	이문화	성남시 분당구	010-xxxx-7777

참고: 본 내용은 데이터 일부를 발췌한 것이 아닌 저자가 작성한 가상의 예시임

나. 결합신청자 간 의사소통

업종이 비슷하거나 규모가 비슷한 결합신청자들은 결합 과정에서 상대적으로 효율적으로 의사소통할 수 있는 장점이 있다. 사용하는 데이터에 대한 이해도 빠르고 기업 문화나 업황에 대해 이미 알고 있기 때문이다. 하지만 이미 진행해 본 연구를 하게 되거나 일반적인 결과를 도출할 확률이 높은 단점도 있다. 본 연구에서는 데이터 결합을 위해 통신데이터와 카드데이터의 결합을 시도하였으나 정해진 컬럼 내에서는 분석의 제한 및 한계가 발생될 수 있기 때문에 보다 다양한 컬럼의 확보 및 분석의 다양화를 모색할 필요가 있다.

반면에 업종이 다르거나 기업문화가 다른 결합신청자들의 경우 의사소통 문제가 발생할 확률이 높다. 데이터에 대한 이해와 업종 성격을 이해하는 시간이 필요하기 때문이다. 하지만 이질적인 데이터를 결합할수록 기존과 차별적인 분석이 가능해진다. 데이터 결합 과정에서는 발생할 수 있는 문제점들을 미리 인지하는 것이 필요하다고 하겠다.

다. 결합신청자마다 다른 데이터 활용 정책

금융 업종, 의료 업종 그리고 IT 업종이 데이터를 결합하려고 하는 경우를 생각해 보자. IT 업종은 상대적으로 데이터 수집이 용이하고 민감 정보가 적다. 민감 정보가 있더라도 데이터 활용을 위해서 통계화 하거나 민감도를 낮추는 작업이 진행되어 있다. 반면 의료나 금융 업종의 데이터는 민감한 데이터가 많고 업종 자체의 보수성으로 인하여 데이터 활용에 제약이 심하다. 또 오래된 업종일수록 디지털화가 떨어지고 과거의 데이터 신뢰성이 낮을 수 있지만 쉽게 구할 수 없는 가치 높은 데이터를 많이 보유하고 있다. 따라서 결합신청자들은 각자의 업종 상황이나 정책을 이해하고 최대한 의사소통에 문제가 없도록 노력해야 한다. 또한 활용하려는 데이터에 대하여도 많은 의견 교환이 필요하다.

라. 복잡한 데이터 결합 절차

데이터 결합 시, 여러 이해 당사자가 얽혀 있고 민감한 개인정보를 다루기 때문에 절차가 복잡하고 처리 시간이 소요된다. 이를 고려하여 데이터 결합을 계획하는 단계에서부터 적극적으로 결합신청자들 사이에서 합의를 도출해야 한다. 만약 완벽한 계획 없이

결합이 완료된 시점에서 분석 주제가 바뀌거나 필요한 데이터가 추가로 필요할 경우 다시 가명처리에 대한 절차를 거쳐야하기 때문에 업무의 비효율을 야기한다. 본 연구에서는 데이터 결합절차 수행을 위해 데이터 제공기관과의 수차례 논의를 통해 최종 합의점을 도출하였으며, 최종적으로는 데이터 제공기관과 데이터 이용기관이 모두 만족할 수 있는 방안제시를 통해 결합에 대한 합의점을 찾아야 한다. 특히 공공기관에서 데이터를 이용할 경우 데이터 구입비용 등과 향후 개인정보 활용에 따른 문제 발생 등에 대해 사전에 충분히 검토하고 협의해야 할 것이다. [26]

마. 데이터 신뢰성 문제

데이터 분석에서 원천 데이터의 신뢰성은 중요한 문제이다. 결합신청자 A와 결합신청자 B가 데이터를 결합하고 나서 아래 〈표 2-19〉 와 같은 가명 데이터를 얻었다고 가정해 보자. 결합신청자 A와 결합신청자 B는 각각 개별적으로 수집한 성별과 개별적으로 추정한 소득추정 값을 가지고 있다. 둘이 결합되었을 때 성별과 소득추정 값이 다른 경우가 발생할 수 있다. 이런 경우 데이터를 결합하기 전에 개별 결합신청자의 데이터에 대한 신뢰도와 오차율을 바탕으로 신뢰도가 높은 데이터를 선택하고 같은 성격의 데이터는 결합 시 활용하지 않는 방안이 있다.

〈표 2-19〉 다른 결합신청자들로부터 비슷한 성격의 데이터가 결합된 경우

결합키	결합신청자 A 성별	결합신청자 B 성별	결합신청자 A 소득	결합신청자 B 소득
kdfd35	남자	여자	상위 10%	상위 15%
4z3dfa	남자	남자	상위 20%	상위 60%

참고: 본 내용은 데이터 일부를 발췌한 것이 아닌 저자가 작성한 가상의 예시임

26) 개인정보 규제가 데이터 활용 및 결합에 미치는 영향에 대한 연구로 김상광·김선경. (2020), 김상광. (2020)가 있음

제5절 소결

제2장에서는 데이터3법에서 새롭게 마련된 제도 기반의 핵심 개념인 가명정보의 개념을 이해하고, 가명정보 활용한 데이터를 연구에 사용하기 위해 필요한 절차를 살펴봄으로써 향후 가명정보 데이터 기반의 연구를 진행하는데 참고자료가 될 수 있도록 가명처리의 기술적인 내용들을 포함하여 정리하였다.

또한 가명정보를 활용할 수 있게 됨에 따라 기존에는 접근할 수 없었던 개인단위의 데이터를 확보할 뿐만 아니라 더 나아가 데이터를 결합하여 연구에 활용하는 방안을 살펴봄으로써, 가명처리 데이터를 통해 발견할 수 있는 새로운 가치가 무엇인지 살펴보았으며 이는 다음 〈표 2-20〉과 같이 요약된다.

〈표 2-20〉 가명정보 도입 전후 비교 및 기대효과

구분	가명정보 도입 전	가명정보 도임 후	기대효과
개인 단위 원시자료 활용	• 기존 익명화된 집계 데이터를 활용하여 총량 위주의 추세 분석 • 각 특성별 규모를 파악할 수 없어 세부 특성별 비교 분석이 어려움	• 개인단위 원시자료로 활용할 수 있게 됨에 따라 1인당 평균 치 통계를 산출 • 이를 바탕으로 세부 특성별 통계치를 비교 분석할 수 있게 됨	**개인단위 통계 생산을 통해 보다 세부적인 변화와 추세를 파악할 수 있어 개인 특성을 고려한 정책 수립 관련 시사점 제공**
시의성 높은 대규모 패널 데이터 구축	• 조사기반 패널 데이터는자료 수집 및 공표에 시간 소요 多 • 표본의 규모가 크지 않아 패널 이탈 시 종단 분석시 어려움 有	• 민간데이터 내 고객의 시계열 정보를 활용, 대규모 패널 데이터 구축 가능 • 민간데이터의 경우 시의성이 높아 시의성 있는 패널자료 확보 가능	**패널데이터 분석을 활용 사회변화에 대응한 적시성 높은 정책적 대응 가능**

구분	가명정보 도입 전		가명정보 도입 후	기대효과
분석 가능한 유효표본 수 확대	• 조사 데이터는 세부특성별 혹은 소지역 단위 통계 생산을 위한 유효표본 확보가 어려움 • 민간데이터의 경우, 개인 단위 원시자료 활용 불가능	➡	• 개인 단위 원시자료 형태인 대규모 표본의 민간데이터 확보 가능 • 개인의 다양한 특성을 고려한 분석이 가능 • 시군구 또는 읍면도 단위 통계생산 가능	**세부적인 특성 및 소지역 통계 산출을 통해 맞춤형 정책 수립 시 근거 자료로 활용**
이종 데이터 결합을 통한 활용정보 확대	• 개인정보보호를 문제로 개인단위 정보 결합 제한 • 서로 다른 자료의 유사한 개체를 결합하는 통계적 연계만 가능	➡	• 데이터 결합해 다양한 정보를 동시에 활용함으로써 다양한 시사점 도출 가능 • 가명처리를 통해 정확연계가 가능하여 정확한 데이터 확보	**이종 데이터 결합을 통해 기존 데이터로 볼 수 없는 다양한 정책 시사점 제시**

하지만 이러한 강점에도 불구하고 가명처리 데이터 활용 시, 개인정보 보호를 위한 다양한 조치가 필요하며 데이터 분석과 결합물을 활용하는데 있어서 다음과 같은 한계점이 존재한다.

〈표 2-21〉 가명처리 및 데이터 결합 시 고려사항

고려사항	세부내용
분석과정	• 데이터 활용범위 축소 및 고급 통계 분석의 한계 - 가명처리 과정에서 데이터 정밀도 하락 - 이에 통계모형 활용 시, 결과의 신뢰도 측면에서 신중한 접근 필요 • 데이터 분석 환경 구성의 어려움 - 가명처리 및 분석 시스템은 기존 시스템과 물리적 분리가 필요 • 분석결과 활용 - 데이터 용량상의 문제와, 개인식별 가능성을 줄이기 위해 통계, 집계 형태로 활용 • 표본의 대표성 문제 - 민간 데이터 가명처리 시, 전 국민 대상의 표본 확보 여전히 불가능
데이터 결합	• 결합률 - 결합률을 높이기 위해서 최신 정보로 사전 업데이트 필요 • 결합기관간 의사소통 - 데이터 결합 시 발생할 수 있는 문제점 미리 인지하는 것이 필요 • 결합신청자들 간 다른 데이터 활용정책 - 결합신청자들은 각자의 업종 상황이나 정책을 이해 요구 • 복잡한 데이터 결합절차 - 데이터 결합 시, 절차가 복잡하고 처리 시간이 소요됨 - 이에 결합신청자간 적극적 의사소통 및 합의가 필요 • 데이터 신뢰성 문제 - 결합 시, 동일한 정보가 일치하지 않는 경우 발생 - 데이터의 신뢰도와 오차율을 바탕으로 신뢰도가 높은 데이터 선택 필요

그럼에도 불구하고 가명정보 및 결합제도가 도입됨에 따라 연구에 활용할 수 있는 데이터의 범위가 확대됨에 따라 데이터의 가치가 강조되고 정부의 제도적, 정책적 지원에 힘입어 빅데이터 활용은 점차 활성화 될 것으로 예상된다. 이에 가명처리 데이터의 중요성 못지않게 개인정보보호를 기반으로 한 활용을 위하여 관련 법령의 내용을 충분히 숙지하고 준수할 필요가 있다. 또한 가명정보 활용 및 결합을 추진할 경우 개인정보 보호를 위해 가명처리 절차 및 결합전문기관을 통한 세부적인 절차에 의해 진행된다는 점에서, 향후 연구계획 수립 시 결합 절차에 소요되는 시간 및 비용에 대한 사전적 고려가 필요하다고 하겠다.

다음 장에서는 가명정보 활용이 가능한 문화·관광분야의 공공 및 민간분야의 데이터를 살펴봄으로써, 추후 문화·관광분야의 가명처리 데이터 활용방안 도출에 적합하다고 판단되는 데이터를 검토하고 활용방안을 제시하고자 한다.

제3장

문화관광 분야 가명정보 활용 데이터 현황 및 활용사례 검토

제1절 가명정보 활용 데이터 현황

정부는 디지털 뉴딜 정책의 데이터 댐 구축 프로젝트를 추진 중이며 데이터 댐은 공공기관이나 민간 기업이 데이터를 수집하고 이를 가공하여 유용한 정보로 재구성한 집합시스템을 의미한다. 정부의 데이터를 최대한 수집하여 산업 현장에서 잘 활용할 수 있도록 하는 노력과 함께 공공기관에서는 자체 보유하고 있는 공공데이터 개방 노력을 확대하고 있다. 문화·관광분야에서도 행정업무 기반의 다양한 수집정보를 처리하고 있으며, 개방 가능한 데이터를 모아서 제공하고 데이터 분석을 지원하고자 하는 노력을 기울이고 있다. 이러한 데이터들은 개인 정보를 가명처리 할 수 있게 됨에 따라 그 활용성이 더욱 증가할 것으로 보인다. 본 장에서는 가명정보를 활용할 수 있는 문화체육관광 분야의 데이터 현황을 공공, 민간, 조사데이터로 구분하여 살펴보고 향후 활용 가명처리 데이터 및 데이터 결합을 통해 활용 가능한 데이터의 현황정보를 제공하고자 한다.

1. 공공 데이터 현황

가. 현황

1) 공공데이터 포털

공공데이터포털(Data Portal)[27]은 행정안전부에서 운영하는 공공데이터 통합제공 시스템이다. 대한민국 정부가 보유한 다양한 공공데이터를 개방하여 누구나 편리하고 손쉽게 활용할 수 있게 하는 것이 목적이다. 행정안전부 공공데이터정책과에서 관련 정책을 추진하고 있다. 공공데이터포털은 「공공데이터의 제공 및 이용에 관한 법률」(2013. 10월 시행) 제 21조에 의해 대한민국 행정안전부 장관에 의해 구축 운영되고 있다.[28]

27) 통합데이터 포털: https://www.data.go.kr/
28) 「공공데이터의 제공 및 이용 활성화에 관한 법률」 제21조

2021년 5월 기준, 문화관광분야의 총 23,313개의 공공데이터가 공개되어 있어 파일 다운로드 또는 오픈API(Open API)를 통해 이용할 수 있다. [29]

〈표 3-1〉 문화체육관광 분야 공공데이터 포털 개방 현황

(단위: 개)

분류명	합계
문화예술	5,099
문화재	10,584
체육	1,750
관광	5,880
문화체육관광 일반	3,666
합계(문화체육관광)	23,313

출처: 송철재(2021).

현재 해당 포털에서 개인정보를 가명처리하여 데이터를 제공하고 있지는 않다. 하지만 개인 정보를 활용한 익명의 집계 데이터를 제공하고 있어, 향후 연구 목적에 따라서 집계 데이터 수집 전 개인 단위의 원시자료 활용도 가능할 것으로 판단된다. 다만, 해당 개인정보를 가명처리한 데이터의 형태로 활용하기 위해서는 데이터 제공기관과의 업무협의를 기반으로 이용기관으로서 가명정보 활용과 관련된 제도적, 법률적 근거를 갖추야 할 것이다.

2) 문체부 소속·산하기관 운영 개방 플랫폼

공공데이터 개방과 활용 지원 노력의 일환으로 문화·관광 분야에서 운영되고 있는 주요 데이터 개방플랫폼은 문화셈터, 관광지식정보시스템, 문화예술지식정보시스템, 문화빅데이터 플랫폼, 문화데이터 광장, 한국관광데이터랩 등이 있다.

한국문화관광연구원에서 운영하는 문화셈터는 문화체육관광부 국가승인통계(20종)에 대한 통계표와 보고서, 원시자료를 제공하고 있다. 또한 승인통계 뿐만 아니라 문화·체육·관광 관련 월간 경제지표 보고서, 신용카드 지출액 분석 보고서 등의 다양한 통계 보고서 제공하고 있다.

29) 출처: 송철재(2021)

관광지식정보시스템은 지식기반 사회로의 이양 및 정보기술 발전 추세에 따라 관광부문의 정보화 사업추진 전략을 제시한 국가관광정보화 추진 전략계획(문화관광부, 2002년)에 근거하여 연차사업으로 한국문화관광연구원에서 운영하고 있다. 다양한 관광과 관련된 통계지표와 보고서뿐만 아니라 관련된 데이터를 함께 제공하고 있다.

문화예술지식정보시스템은 국내외 문화예술, 문화콘텐츠 관련 지식 정보와 통계 정보를 제공하기 위해 문화체육관광부 산하 정책연구기관인 한국문화관광연구원이 구축한 DB시스템이다. 이는 급변하는 정책 환경 변화에 선제적으로 대응하고, 중장기적인 문화예술 발전과 경쟁력 강화를 위해 국내외 정부, 지방자치단체, 연구소, 학계, 산업계 등 각 영역에서 생산된 관련 정보를 수집하고 분석하여 공유·확산하는 것을 목표로 2012년부터 진행되었다.

한국문화정보원에서 운영하는 문화 빅데이터 플랫폼은 데이터 산업 육성을 위한 플랫폼으로 데이터 유통 및 개방 등 유통지원과 데이터 활용을 위한 데이터 마케팅 사업을 지원하고 있다. 한국문화정보원이 보유한 데이터 셋 뿐만 아니라 한국청소년활동진흥원, 부산정보산업진흥원, 한국문화예술위원회, 국민체육진흥공단, 코리아크레딧뷰로, 야놀자 등 공공 및 민간분야의 다양한 데이터를 결합하여 제공하고 있다.

문화데이터광장은 문화체육관광부 소속 및 산하 66개 기관, 타 부처 및 지자체 74개 기관(총 140개)의 결합을 통해 약 8,500만건 메타데이터를 생산하여 유용한 정보를 선별하여 대국민 서비스를 제공하는 플랫폼으로 문화데이터의 접근과 입수, 활용이 쉽도록 정보를 개방하고 문화데이터의 민간 활용을 확대하기 위해 노력하고 있다.

한국관광데이터랩은 이동통신, 신용카드, 내비게이션, 관광통계, 조사연구 등 다양한 관광 빅데이터 및 융합분석 서비스를 제공하는 관광특화 빅데이터 플랫폼으로 관광기업, 지자체, 업계, 학계 등 관광산업 이해관계자들이 데이터에 기반한 과학적인 관광정책 수립 및 관광 비즈니스를 수행하도록 지원하고자 '21년 2월에 서비스를 개시하였다.

〈표 3-2〉 문화·관광 분야 공공데이터 플랫폼

운영기관	플랫폼명
문화체육관광부	통합문화이용권정보시스템, 빅데이터 기반 언어 말뭉치시스템, 도서관 빅데이터플랫폼, 사서의사결정시스템, 예술인경력증명시스템, 국제문화교류정보시스템, 지역문화 통합정보시스템, 디지털국악아카이브시스템, 현대사디지털아카이브시스템, 국립중앙극장, 공연예술자료관리시스템, 불법온라인도박관리시스템, 정책여론수렴시스템, 콘텐츠수출마케팅플랫폼
한국문화관광연구원	관광지식정보시스템, 문화셈터, 문화예술지식정보시스템
한국관광공사	TourAPI 3.0, 한국관광데이터랩
한국문화정보원	문화 공공데이터 통합활용 시스템, 문화 빅데이터 플랫폼
한국문화예술위원회	문화예술 데이터관리시스템

위의 플랫폼에서 제공하고 있는 데이터는 개방플랫폼에서 자유롭게 다운로드할 수 있는 익명정보가 대다수를 차지하고 있으며, 개별 협의를 통한 가명정보의 이용이 가능한 것으로 보이나 아직은 활용사례가 거의 없는 것으로 보인다. 향후 가명정보의 유통이 지금보다 활성화 된다면 위와 같은 플랫폼의 역할이 점점 더 중요해지고 플랫폼을 통한 가명정보의 활용이 빠르게 확산될 것으로 예상된다.

3) 그 외 기타 공공 데이터

앞서 언급된 공개된 행정 데이터 외, 문체부 소속 산하기관이 고유의 행정업무 수행을 통해 생성하거나 취득한 데이터 또한 가명정보 활용이 가능하다. 첫째, 문화체육관광부는 행정업무 기반의 다양한 수집정보를 정보시스템을 통해서 관리하고 있으며, 현재 문화체육관광부의 정보시스템을 통해 수집되는 공공데이터는 총 331개로 파악되고 있다. 주로 정부 서비스 운영실적(회원관리 포함), 정부 정책 프로그램 참여 및 성과, 정부 정책 수혜자 정보 등으로 이루어져 있다.

해당 정보는 정보주체의 동의 또는 법령에 근거한 개인정보수집 및 처리 목적에 따라 운영되며, 목적 달성이 완료될 때까지의 보유기간을 설정하고 보유기간이 도래한 경우 지체 없이 삭제 및 파기하여 관리된다.[30] 데이터 3법 개정 전에는 문화·관광 분야 정책 프로그램의 참여한 개인정보는 수집·이용 목적에 해당하는 경우에만 이용할 수 있어 다양한 정책 효과 분석 연구에 활용하는 데 제약이 있었던 반면 이제는 해당 개인정보를

30) 파기된 데이터는 연구에 활용될 수 없으므로 연구에 필요한 데이터가 있다면 연구대상 기간을 고려하여 미리 가명처리를 수행하는 것도 고려해 볼 수 있다.

안전하게 가명처리하여 활용 할 수 있어 연구 기회가 확대를 기대할 수 있게 되었다.

〈표 3-3〉은 문화체육관광부의 정보시스템을 통해 수집되는 공공데이터는 총 331개 중 주요 문화·관광 관련 개인정보 데이터 현황을 보여준다. 주요 내용을 살펴보면 문화누리카드 DB는 기초생활 수급자 및 차상위 계층을 위한 문화 누리카드 발급 현황정보를 보유하고 있다. 이어서 예술의 전당 회원정보, 카지노 고객정보 등 문화·관광 분야에서 운영하고 있는 서비스 회원 및 참여 정보를 관리하고 있으며 이를 통해 예술의 전당 공연 구매 이력, 카지노, 골프장, 박물관, 도서서비스 이용 현황, 예술인 등록정보 등을 확인할 수 있다. 이 밖에도 문화·관광 관련 해당 개인정보시스템 내에는 다양한 문화·관광 프로그램 이용정보, 정부 정책 프로그램 지원정보 등을 관리하고 있으며 이러한 내부 시스템 내 정보를 결합하여 박물관, 도서서비스를 이용하는 사람이 예술의 전당에서 어떤 공연을 즐기는 지 등의 문화활동 예측 분석을 수행하거나 기초생활 수급자가 경험하는 문화·관광 활동을 분석함으로서 정책 효과를 파악하고 정책방향을 수립하는데 지원할 수도 있을 것으로 예상된다.

〈표 3-3〉 문화·관광 관련 개인정보 데이터 현황

구분	제공 내용	제공기관
문화누리카드 DB	o (주요내용) 기초생활수급자, 차상위 계층을 위한 문화누리카드 발급 및 이용 실적 관리 o (개인정보) 이름, 집주소, 이메일, 집 연락처, 핸드폰, 주민번호, 외국인 등록번호, (※ (필수) 문화누리카드번호, 문화누리카드비밀번호, CI/DI)	한국문화예술위원회
예술의 전당 무료 및 유료 회원	o (주요내용) 예술의전당 공연 구매 이력, 그 외 서비스 제공을 위한 기타 개인 정보 o (개인정보) 이름, 집주소, 이메일, 핸드폰, 생년월일	예술의전당
카지노 이용 고객 정보	o (주요내용) 카지노 이용고객의 등록, 출입 이력관리, 영업활동에 이용 o (개인정보) 이름, 집주소, 핸드폰, 생년월일, 여권번호, 외국인등록번호, 기타(국적, 직업)	그랜드코리아레저(주)
골프장 홀인원 시스템 예약 정보	o (주요내용) 골프장 이용 고객의 등록, 출입, 이력 관리 등 영업활동에 이용 o (개인정보) 이름:필수, 핸드폰(연락처):필수	한국관광공사
박물관 교육프로그램 운영 및 참여	o (주요내용) 교육 프로그램 참여 인원 통계 분석 자료 활용 o (개인정보) 이름:필수, 집주소:필수, E-Mail:필수, 핸드폰(연락처):필수	국립 중앙박물관
공공도서 책이음 시스템	o (주요내용) 책이음서비스 이용자 서비스 제공 o (개인정보) 이름:필수, 집주소:필수, E-Mail:필수, 핸드폰	국립중앙도서관

구분	제공 내용	제공기관
예술인경력정보시스템	o (주요내용) 예술인의 경력 정보 시스템, o (개인정보) 이름:필수, 집주소:필수, 직장주소:필수, E-Mail:필수, 집연락처:필수, 핸드폰(연락처):필수, 주민번호:필수	한국예술인복지재단

둘째로 시스템으로 외 기타 행정 업무를 통해 생성 혹은 취득한 데이터도 존재한다. 문화체육관광부를 제외한 소속 및 산하기관 48개 중 21개 기관을 대상으로 데이터 전수조사를 실시한 결과 총 452개의 데이터를 보유한 것으로 나타났다. 31)해당 데이터 또한 개인정보를 포함하고 있다면 가명처리를 통해 원시자료 형태의 데이터로 활용이 가능할 것이다. 다만, 데이터 보유기관과의 업무 협의 및 가명처리 수준 등의 행정적, 법률적 절차들이 요구된다.

〈표 3-4〉 문화·관광 분야 공공데이터 플랫폼

(단위: 개, %)

운영기관	데이터 수
문화예술	345(100.0)
체육	40(100.0)
관광	45(100.0)
문화관광 일반	22(100.0)

출처: 송철재(2021).

나. 특성

공공기관이 보유한 정보는 정책 수혜 정보를 포함하고 양적으로도 풍부한 데이터 이므로 가명활용에 대한 연구 수요는 높으나, 연구 목적으로 활용 시 유효표본 확보가 어려울 것으로 예상된다. 〈표 3-3〉의 문화관광 관련 개인정보를 포함한 데이터 현황을 살펴보면, 개인의 정보가 전 국민을 대상으로 수집된 것이 아니라 기초생활 수급자, 차상위 계층, 또는 특성 도서관 및 박물관 이용자로 한정되어 있다. 따라서 연구 목적으로 활용할 경우 가명정보를 활용한 이종 데이터와 결합 등의 작업에서 충분한 유효표본 확보가 어려울 것으로 보인다.

31) 문화체육관광부 소속 및 산하기관을 대상으로 수행된 데이터 보유 전수조사 결과(송철재, 2021)

그 외, 공공기관이 적극적으로 가명정보를 제공할 유인은 부족한 편이다. 정부정책효과에 대한 분석은 많은 연구자들이 관심이 많은 연구 주제이고 의미 있는 연구임에도 불구하고 공공기관은 보유하고 있는 공공정보를 외부에 공개한 사례가 거의 전무한 실정이다. 정부정책의 일환으로 공공데이터 개방을 적극적으로 추진해 오고 있음에도 불구하고 가명정보 제공 실적은 미미하며 여전히 가명정보 제공에 대한 환경은 크게 변하지 않은 실정이다. 또한 가명정보 제공에 대한 사회적 공감을 가지고 있다 하더라도 공공기관은 가명정보 재식별 위험부담을 크게 가지고 있어 내부 승인 절차를 거친 반출이 이루어지기 어렵다. 또한 전문 인력도 부족하여 가명·익명 처리 제도에 대한 이해도가 낮은 수준으로 가명처리 수행에 적극적이지 않은 점도 활용에 어려운 요인으로 꼽힌다. 해당 자료의 경우 공공기관의 가명정보 제공 및 활용을 확대하여 민간수요에 부응하고, 민간 정보와의 결합을 통한 시너지 효과를 제고하는 노력이 필요할 것으로 보인다.

2. 민간 데이터 현황

민간분야에서는 마케팅 목적 등에 빅데이터 활용이 강조되면서 공공분야 보다 데이터를 활용한 가치창출에 좀 더 적극적인 움직임을 보이고 있다. 고객이 원하는 가치를 파악하는 것이 곧 기업의 경쟁력과 직접 연결되므로 매 순간 수집 및 축적되는 고객 데이터를 바탕으로 분석하여 고객 맞춤형 서비스를 제공하려는 시도를 하고 있다. 또한 기업 내 데이터 전략 수립 및 분석을 위한 전담부서를 만들고 내부 고객 데이터 표준화 관리, 다른 기업 데이터와 결합하여 분석 프로젝트를 진행하는 사례도 늘어나고 있다.

〈표 3-5〉은 민간 분야에서 문화·관광 관련 연구를 위한 접근 가능한 데이터 현황을 정리한 표이다. 주로 개인의 소비 패턴 및 행동 특성을 연구하기에 유용한 데이터를 보유한 이동통신사·신용카드·유통사의 데이터로 구분할 수 있다. 이동통신사 데이터는 통신사와 결합한 영화 관람·커머스 소비내역·미디어 이용현황 등의 내용을 보유하고 있어 개인별 공연 및 미디어 분야의 소비행태를 확인할 수 있다. 또한 신용카드사 데이터는 신용카드 결제내역을 통해 업종별 개인의 소비패턴을 확인할 수 있어 문화 관광과 관련되어 있는 업종의 소비특성을 분석할 수 있다. 마지막으로 유통사 데이터는 유통채널을 통한 구매 정보에 접근하여 문화 관광분야 관련 선호하는 구매 유형 등을 살펴볼 수 있다. 특히 이러한 데이터들을 결합한 데이터를 활용한 분석은 더 많은 정책적 인사이트와 가치를 찾을 수 있을 것이다.

〈표 3-5〉 문화·체육·관광 관련 민간 분야 데이터 현황

구분	제공 내용	제공기관
이동통신 데이터	실거주지, 지점별 체류시간, 빈도 등	SK텔레콤, KT
신용카드 데이터	업종별, 개인특성별, 지역별 카드 지출액 제공	신한, BC, 국민
유통사 데이터	연령·지역별 구매내역 등	GS 리테일 등

가. 현황

1) 통신 및 인구 데이터

스마트폰, 태블릿 등 모바일 없는 생활은 이제 상상할 수 없기 때문에 이동통신사는 대표적인 데이터중심 회사이다. 통신사의 고객정보에는 이동 동선, 모바일 서비스 이용 현황, 온라인 콘텐츠트 이용량, 구독 서비스정보 뿐만 아니라 다양한 정보를 활용하여 추정한 연령, 직업군(자영업자 등), 여행 지수 등을 분석에 함께 활용할 수 있다.

〈표 3-6〉은 통신 및 인구와 관련된 데이터 정보를 제공하는 분석 가능한 민간 데이터 항목을 나타낸다. 크게 이동성 지표, 온라인 컨텐츠 사용량, 인구특성, 위치, 가족, 그 외 구독서비스 이용형태 등으로 구분할 수 있다. 이동성 지표에는 이동 동선을 기반으로 추정한 재택 및 외출 지수, 이동거리 및 반경, 여행 이벤트. 여행객 지표, 여행 지수, 여행지 유사도 항목 정보로 얼마나 자주 여행을 하고 다양한 여행지를 즐기는지 비슷한 여행지를 중복해서 방문하는지 등을 나타내는 정보를 포함한다. 온라인 컨텐츠 사용량은 동영상, 음악, 게임, 도서, 웹툰, 커뮤니티 등 종류별로 사용횟수랑 사용량을 제공하여 개인별 온라인 소비 현황을 파악할 수 있다. 인구특성은 직업 및 연령정보 위치특성은 출퇴근 이동 등 이동거리 및 범위, 활동성 정보, 그 외 가족 서비스 및 구독서비스 이용 현황정보도 이용할 수 있다. 해당데이터를 활용하여 활동성이 높은 사람의 여행 타입 및 온라인 컨텐츠 사용량과 활동성의 관계 등을 분석이 가능할 것으로 판단된다.

〈표 3-6〉 이동통신 데이터 현황(SK텔레콤)

구분	컬럼명
이동성지표	재택 및 외출 지수
	이동거리
	이동반경
	인구접촉 지표

구분	컬럼명
	여행 이벤트
	여행객지표
	여행지수
	여행지유사도
위치	출퇴근특성-거주지와직장과의거리
	출퇴근 특성 - 이동시간
	출퇴근 특성 - 이동횟수
	평일/휴일 이동횟수
	평일/휴일 이동거리
	교통수단 이용성(1개월)
	활동성 지표
	지하철 이용성
구독서비스 이용 관련	구독서비스 이용 유무
	구독서비스 이용 기간
	구독서비스 이용 유형
위치	출퇴근특성-거주지와직장과의거리
	출퇴근 특성 - 이동시간
	출퇴근 특성 - 이동횟수
	평일/휴일 이동횟수
	평일/휴일 이동거리
	교통수단 이용성(1개월)
	활동성 지표
	지하철 이용성
인구 특성	실사용자 연령
	직장인 유무
	자영업자 유무
	가구원 수 등 가구특성

2) 신용카드데이터

카드는 소비의 주요 수단으로서 카드회사는 고객의 소비현황 및 패턴을 분석하기에 적합한 정보를 가지고 있다. 〈표 3-7〉는 신한카드회사가 제공하는 신용카드 관련 데이터 항목이다. 고객특성을 나타내는 성별, 연령, 주소, 소득(추정), 직업정보, 월 소비규모 데이터에 접근할 수 있고, 카드이용정보를 변수로 활용하여 추정한 취미, 앱구매등급,

라이프 스테이지 정보도 이용할 수 있다. 분석에 활용할 수 있는 주요 변수로는 문화 관광분야 업종분류에 따라 결제장소와 결제시점 및 결제금액 정보를 제공한다. 카드회사의 데이터를 활용하여 공연장 극장을 자주 찾는 고객의 특성을 카테고리화 하고 이들의 관련 소비 문화를 예측하는 등 문화 관관 관련 업종에 개인별 특성에 따른 소비 패턴을 분석하고 예측할 수 있다.

〈표 3-7〉 신용카드 데이터 항목(신한카드)

구분	컬럼명
고객특성	성별
	연령
	거주지 주소
	직장 주소
	추정소득
	직업군
	월 소비 규모
	라이프 스테이지
	취미 1순위
	취미 2순위
	앱구매 등급
	디지털 음악이용
	게임 이용등급
	자동차 이용여부
결제시점(When)	기준연도
	기준연월
	기준일자
	요일
결제장소(Where)	광역시도
	시군구
업종(What)	신한카드 세부 업종분류
지출금액	신용카드지출금액

3) 유통데이터 (GS칼텍스)

GS칼텍스가 제공하는 유통데이터는 국적, 점포, 구매시간대, 품목 등의 분류에 따라 구매건수, 구매금액, 구매수량 등의 집계정보를 제공한다. 이용자가 어떤 지역 편의점에 많이 방문하고 어떤 상품을 얼만큼 구매하는지를 파악할 수 있는 정보이다. 이 정보를 활용하여서는 이용자 국적별, 지역별, 시간대별 주력 이용 상품 등의 분석이 가능하다.

〈표 3-8〉 유통 데이터 항목(GS칼텍스)

구분	컬럼명
기준분류 (Grouping)	기준년도
	해당국적
	점포지역_광역시도
	점포지역_시군구
	점포지역_읍면동
	구매시간대
	품목_중분류_코드
	품목_중분류_명
	품목_소분류_코드
	품목_소분류_명
집계 (Sum)	구매건수
	구매금액
	구매수량

나. 특성

민간데이터는 방대한 고객의 데이터를 보유하고 있으며 데이터 활용의지도 적극적인 편으로 가명정보의 결합 분석 수요가 높다. 때문에 분석을 위해 제공하는 가명정보에 접근하고 활용할 수 있는 기회가 타 분야에 비해 높은 편이다. 또한 민간데이터는 신뢰할 수 있는 데이터로 질이 높고 분석에 활용할 수 있는 데이터 항목이 풍부하다는 장점이 있다. 또한, 공공 및 조사 데이터 보다 비교적 유효 표본이 크게 때문에 성별·연령별 또는 1인 가구·세대별과 같은 개인의 세부 특성을 고려한 분석이 비교적 용이하다. 실시간 데이터가 축적되기 때문에 가명처리절차를 신속하게 진행한다면 시의성 있게 빠른 분석을 할 수 있다는 것도 큰 이점이다. 신속한 분석은 트렌드나 유행이 민감하게 급변

하는 시대에 대응에 필수적이라 할 수 있다.

하지만 민간기업 역시 고객데이터를 활용한다는 점에서 개인정보보호 이슈를 고려해야 하고, 활용 가능한 데이터 항목이 많다는 것은 식별 위험이 높아진다는 것을 의미하므로 데이터 항목별로 다소 높은 가명처리 수준의 데이터로 제공될 가능성이 높다. 예를 들어 분석에 주요변수로 활용되는 소득, 이용금액 등은 범주화 처리를 해서 제공된다면 분석 시 정보 손실량이 발생하여 분석결과에 민감도가 떨어질 수 있다는 점을 연구 시 고려해야 한다.

3. 조사 데이터 현황

가. 현황

조사를 기반으로 생산된 데이터 또한 개인 혹은 가구 정보를 가명처리하여 제공하고 있기 때문에, 일종의 가명처리 데이터로 구분된다고 할 수 있다. 또한 조사대상자의 개인정보수집활용과 관련된 동의가 있다면 개인정보를 활용하여 이종의 데이터와 결합도 가능하다.

조사 데이터 문화체육관광부에서는 문화, 체육, 관광 관련 정책에 대한 연구를 위해 주제별로 다양한 조사 목적의 조사데이터를 관리하고 있다, 주기적으로 조사된 데이터는 외부에 승인 통계로 공표되고 있으며, 다양한 목적으로 연구에 활용되고 있다.

〈표 3-9〉 에서는 문화체육관광부에서 운영하는 조사 중 문화관광 분야에 해당하는 승인통계 6종으로 매년 주기적으로 조사하여 외부에 공표하고 있다. 조사목적에 따라 주요 조사항목이 구분되어 있으며, 조사 목적이 분명하기 때문에 해당 목적에 따라 조사항목이 상세하게 구성되어 있음을 확인할 수 있다.

〈표 3-9〉 문화·관광 관련 개인단위 조사 데이터 현황

구분	조사 목적	주요 지표
국민문화예술활동 조사	o 우리나라 국민의 문화 활동 향유의 필요성 및 인식이 높아짐에 따라 실태 파악을 위한 문화향유 경로와 방식에 대하여 통계적으로 분석하여 궁극적으로 국민 문화향유 진흥 도모	- 문화예술행사 관람 및 관람 의향 - 문화예술행사 매체 이용 실태 및 참여 활동 - 문화예술교육 경험 및 의향 - 문화예술 공간이용 실태 및 방문의향 - 문화관련활동

구분	조사 목적	주요 지표
국민여행조사	o 우리나라 국민의 여행실태를 종합적으로 파악, 국가 관광에 관한 정책수립과 연구·분석 등을 위한 기초 자료를 제공	- 여행목적(지) - 여행지출액 - 여행소감 - 기타 여행 문항
국민여가활동조사	o 국민들이 여가를 어떻게 인식하고, 여가생활을 하고 있는지를 조사, 생활양식의 변화 및 삶의 질적 수준을 파악하여 정부의 여가정책 수립을 위한 기초자료 제공	- 여가활동 참여실태 - 사회성 여가활동 - 동호회 활동 - 휴가, 연휴 활용 - 여가공간 - 여가자원 활용실태 - 여가생활 만족도
주요관광지점 입장객통계	o 주요 관광지점 이용객 통계의 생산·배포를 통한 관광객 수요 추정 및 관광시설 공급판단의 기초자료로 활용	- 국내·외 입장객 및 시설이용객 현황(유료관광지) - 방문객 현황(무료관광지)
외래관광객조사	o 방한 외래관광객의 여행성향 및 실태의 변화추이를 정기적으로 조사·비교·분석함으로써 외래관광객 유치 증대 및 수용태세 개선을 위한 관광정책 수립의 기초자료로 활용	- 한국여행실태(방문 횟수, 시기, 방문 목적 등) - 한국여행 소비 실태(지출 경비, 쇼핑 품목, 쇼핑 장소) - 한국여행 평가(한국 여행에 대한 만족도 등)
근로자휴가조사	o 국내에서 산업 활동을 영속중인 사업체와 근로자를 대상으로 휴가실태조사를 실시하여 관련분야 정책수립시 합리적 의사결정 지표로 활용될 신뢰할 수 있는 통계자료 제공	- 연차휴가일수(연차휴가 소진율) - 휴가사용 환경 - 휴가 만족도 등

나. 특성

조사데이터는 조사목적에 해당하는 상세항목을 연구에 활용할 수 있고, 설문항목에 따라 개인의 의향 및 선호도를 직접 파악할 수 있다는 장점이 있다. 하지만 설문에 응한 응답에 의존하는 데이터로 설문지 구성 및 조사 방법에 따라 다소 신뢰도가 떨어질 수 있다는 단점이 있다. 그리고 이러한 조사 데이터는 처리 및 제공에 상당히 많은 시간이 소요되고 가명처리하여 가명정보 결합 분석 추진을 검토할 경우 결합키로 활용할 개인식별정보를 처리하기 위해 개인정보수집활용에 대한 재동의가 필요하다. 개인정보 수집활용 재동의에는 많은 비용과 시간이 투입되어야 함으로 향후 지속적인 데이터 결합 가능성에 대한 검토를 통해 향후 활용방안에 대한 모색이 필요하다.

제2절 가명정보 활용 사례

앞서 가명정보의 개념이 도입됨에 따라서 데이터 활용가치가 어떻게 확장되고 개선될 수 있는지 확인하였다. 또한 실제 가명처리 데이터 활용 시 발생할 수 있는 문제점을 바탕으로 고려해야할 사항을 살펴보았다. 예상되는 문제점에도 불구하고 가명처리 데이터의 결합 활용을 통해 얻을 수 있는 정책적 가치에 대한 기대에 따라 이를 활용하는 사례가 국내외로 진행되고 있다.

1. 해외 사례

주요국의 가명정보는 주로 데이터를 가명처리 한 후 결합하기 위한 사례들 위주로 진행되고 있다. 이러한 데이터 결합은 금융업, 제조업, 의료업을 등 민간분야 뿐만 아니라 공공 분야에서도 진행되고 있는 것으로 파악되며 주요 사례는 다음과 같이 요약된다. 미국사례를 살펴보면 민간의 가명처리 데이터를 결합하여 보험사 차량사고처리 정보와 제조사의 차량별 안전장치 정보를 활용하여 사고 유형별 안전 장치 영향도를 분석하였다. 영국의 경우 포털사와 증권사의 데이터를 결합하여 소셜 데이터와 주가지수의 상관관계를 분석한 바 있으며 의료분야에서는 의료 행정 데이터들을 연계하여 'CALIBER'이라는 데이터 플랫폼을 구축하였다. 캐나다의 경우 의료, 경제, 교육 등 다양한 분야에서 가명정보를 활용 한 결합 사례들이 진행되고 있다. 의료 및 경제 분야의 조사 데이터와 교육 행정데이터를 연계하여 재정, 사회, 경제와 관련한 통합 연구를 진행한바 있으며, 그 외 의료와 금융, 인구 자료를 연계하여 지역 기반의 건강 서비스 분석 및 보건정책 연구, 데이터 과학 등 다양한 분야의 연구 자료로서 활용하고 있다.

〈표 3-10〉 해외 가명정보 활용 사례 요약

국 가	활용분야	결합 대상 데이터	결합데이터 이용기관
미 국	o 사고 유형별 안전장치 영향도 분석	(보험사) 차량 사고처리 정보 (제조사) 차량별 안전장치정보	o 보험사: 보험료 할인상품 개발 o 제조사: 안전장치 기능 개선
영 국	o 소셜데이터와 주가지수 상관관계 분석	(포털사) 소셜데이터(비정형) (증권사) 주가정보	o 투자은행: 주가예측 로보어드바이저 개발
	o 의료분야 분석 연구	전자의무기록과 보건의료 행정데이터(1차의료, 2차의료, 질병레지스트리)를 연계한 'CALIBER'이라는 데이터 플랫폼을 구축	o Farr Institute
캐나다	o 재정, 사회, 경제 관련 연구	(의료) 캐나다 암등록 자료 전국 인구 건강조사 (경제) 노동 및 소득 동태조사 (교육) 중고등학생 정보 시스템 등 행정 및 조사 데이터를 연계하여 데이터연계 환경(SDLE) 프로세스 구축	o 캐나다 통계청 (데이터연계 환경(SDLE) 프로세스)
	o 건강, 웰빙 관련 연구(교육, 데이터 연계 등 지원)	(의료) 의약보험, 의료서비스계획 지불 정보 등 (인구) 출생, 사망, 결혼 통계 등 인구 통태 통계, 영주권자, 소득구관 등 행활과정 통계 (경제) 직업 정보 그 외 교육, 환경 관련 데이터 이용한 연계 데이터 제공	o 브리티시컬럼미아주 인구데이터(PopData BC)
	o 지역 기반의 건강서비스 분석, 보건정책, 데이터 과학 등 연구	(보건) 입원환자 병원 기록, 전자의무기록, 수술 기록 등 (시설) 보건서비스 기관정보 (금융) 의료서비스 비용 (인구) 난민 및 이민 상태, 인구조사 프로필, 인구 추정 및 추계 그 외 기타 비보건 부문 데이터를 이용하여 연계	o 온타리오주 임상평가과학연구소

자료: 1. 캐나다 통계청(Statistics Canada) 홈페이지의 사회적 데이터연계 환경(SDLE) 내용 요약 (https://www.statcan.gc.ca/eng/sdle/overview/)
2. 자료: Ark et al. (2019)
3. 자료: Schull et al. (2019)

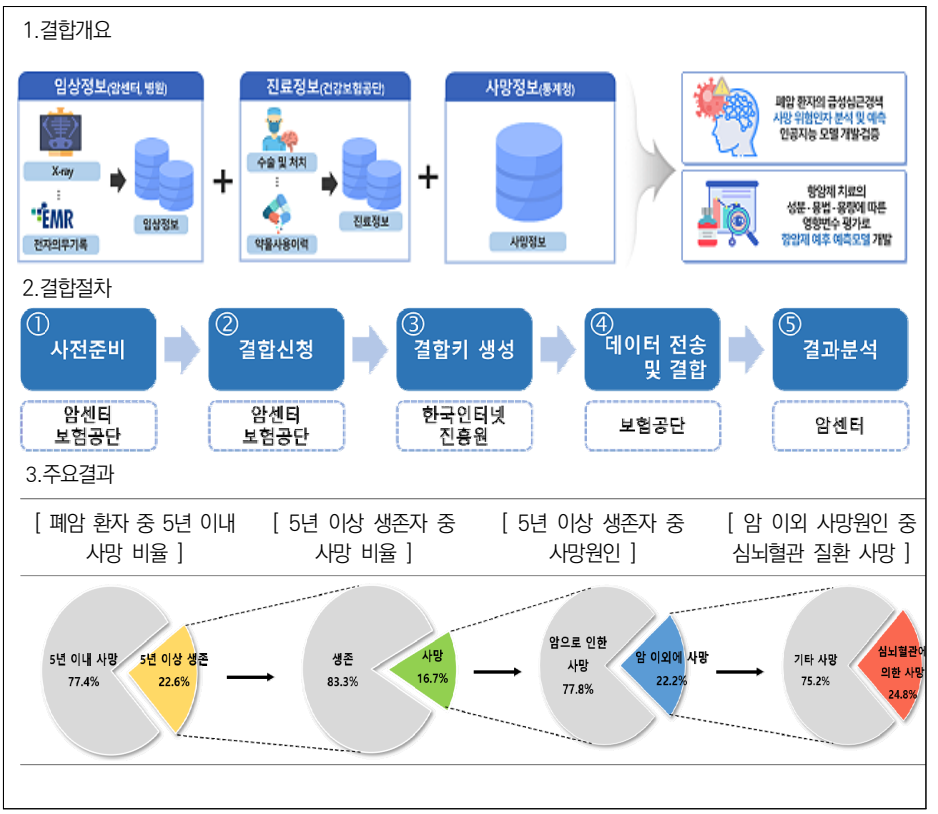
2. 국내 사례

가. 공공분야 가명정보 활용 사례

정부는 가명정보 결합에 대한 수용도를 높이기 위하여 공익적 목적의 실질적 사례를 우선적으로 발굴·추진하였다. 시범사례로 발굴된 분야는 의료가명처리데이터와 인구가명처리데이터의 결합, 금융가명처리데이터와 보훈가명처리데이터의 결합, 소득가명처리데이터와 복지가명처리데이터의 결합, 통신가명처리데이터와 유통가명처리데이터의 결합, 마지막으로 레저가명처리데이터와 건강가명처리데이터의 결합 총 6가지이다. 이들 시범사례 중 2021년 10월 현재 성과물을 제시한 분야는 의료와 인구, 통신 분야의 가명정보를 결합한 2가지 사례이다.

먼저 의료가명처리데이터와 인구가명처리데이터의 결합사례를 살펴본다. 의료가명처리데이터와 인구가명처리데이터를 결합·활용하는 시범기관으로 국립암센터가 지정되었다. 국립 암센터는 보건복지부에서 발표한 「보건의료 데이터 활용 가이드라인」에 따라 암센터가 보유한 2만 명의 폐암 환자 가명처리 정보와 건강보험관리공단이 보유한 이들의 암 진료 정보를 결합해 '폐암 치료효과 분석 및 폐암 환자에서의 합병증·만성질환 발생 및 사망 예측모델' 개발을 추진하였다. 암센터는 이번 가명처리정보 결합을 통해 폐암환자에 대한 연구 수행 시 단일 병원 데이터만을 활용하여 연구할 경우 환자가 여러 병원을 이용할 가능성을 고려할 수 없기 때문에 정확한 분석이 어려우며, 건강보험공단의 진료정보만을 이용할 경우에는 다양한 진료행위 간 영향력을 고려할 수 없는 점을 보완할 수 있었다고 분석하였다. 가명정보의 결합을 통해 도출한 주요 분석결과는 다음과 같다. 국립암센터에 내원한 폐암 환자 중 1년 이내 사망은 38.2%, 3년이내 사망은 누적 67.3%, 5년 이내 사망은 누적 77.4%, 10년 이내 사망은 누적 87.5% 였으며, 폐암 진단을 받고 5년 이상 생존 후 연구대상기간 내 사망한 환자의 22.2%가 암 이외의 원인으로 사망하였고, 이 중 심뇌혈관질환으로 인한 사망이 24.8%를 차지하였다(국립암센터, 2021.05).

[그림 3-1] 국립 암센터 가명정보 결합 시범사례 결합방법 및 주요결과

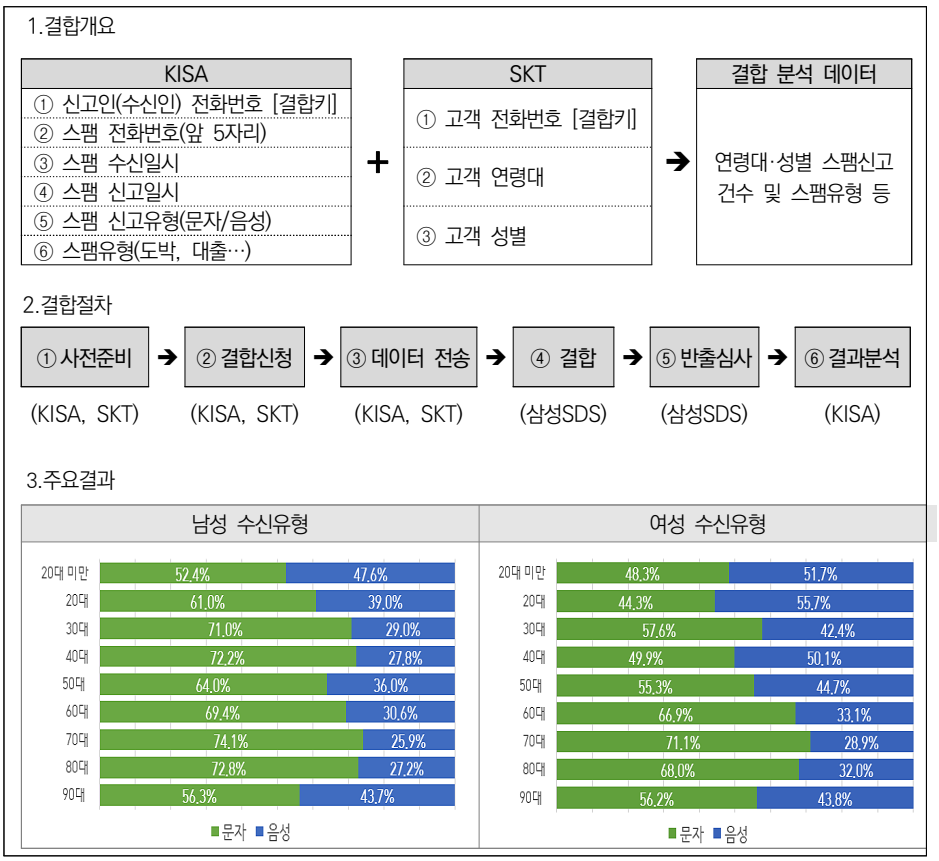


자료: 국립암센터(2021.05.28.), '국립암센터, 가명정보 결합 시범사례 첫 성과발표' 및 '두 번째 성과발표' 재구성

다음으로 통신가명처리데이터와 스팸신고가명처리데이터의 결합사례를 살펴본다. 통신가명처리데이터와 스팸신고가명처리데이터를 결합·활용하는 시범기관으로 한국인터넷진흥원(KISA)가 지정되었다. 한국인터넷진흥원은 진흥원이 가진 2020년도 스팸신고정보 중 데이터 결합을 하는 통신사 가입자가 신고한 1,377만 건의 스팸정보와 해당 이동통신사가 보유한 고객의 성별, 연령정보 등을 가명처리 후 결합하였다. 가명정보 결합을 통해 도출한 주요 분석결과를 살펴보면 다음과 같다. 성별 스팸신고 비율은 남성 64.4%, 여성 35.6%이었으며, 연령대별로는 50대 28.6%, 60대 22.8%, 40대 22.7% 순서를 나타냈다. 이와 함께 남성은 문자 스팸 수신율이 높았고, 여성은 음성 스팸 수신율이 상대적으로 높았다는 내용도 도출하였다. 한국인터넷진흥원의 가명정보 결합은 특정 통신사의 이용자만을 대상으로 분석했기 때문에 대표성을 가지긴 어려운 측면이 있으나, 표본

이 충분히 크다는 점(1,377만 건), 공공분야와 민간분야의 가명처리 데이터가 결합한 첫 사례라는 점에서 의의가 있다.

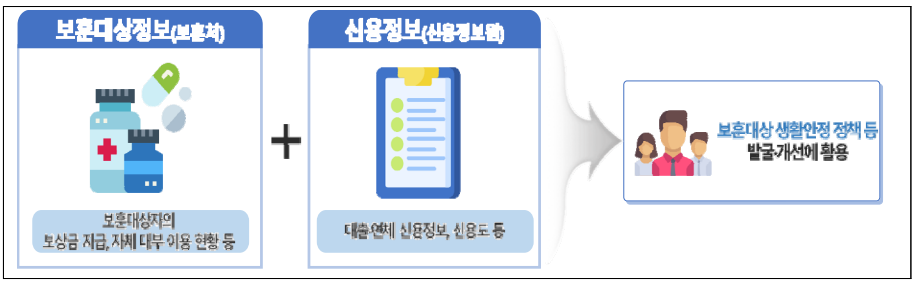
[그림 3-2] 한국인터넷진흥원 가명정보 결합 시범사례 결합방법 및 주요결과



자료: 개인정보보호위원회(2021.06.25.), '불법스팸 실태 분석을 위한 가명정보 결합 시범사례 결과 발표' 등 재구성

그 밖에 4개의 시범사업은 현재 결과도출을 위한 사업을 진행하고 있는 상황이다. 국가보훈처는 보훈처가 보유하고 있는 보훈대상정보(보상금지급, 자체대부이용현황 등)와 신용정보원이 보유하고 있는 개인신용정보(대출·연체 신용정보, 신용도 등)를 결합하여 보훈대상자의 경제적 형편과 관련된 대출 및 연체 이력 등 개인 신용정보를 파악하여 국가보훈처 정책의 생활안정효과를 분석하고 있다.

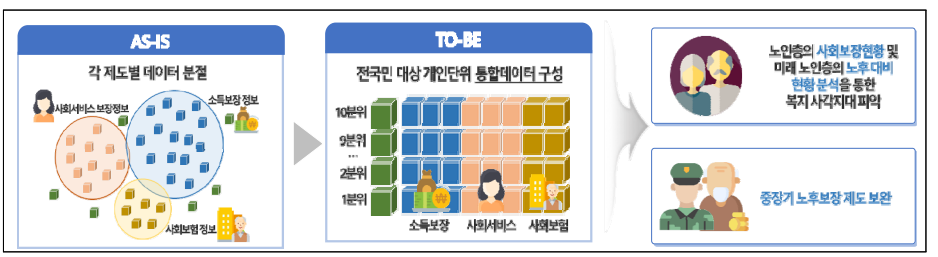
[그림 3-3] 국가보훈처 가명정보 결합 시범사례 결합 추진방안



자료: 국립암센터(2021.05.28.), '국립암센터, 가명정보 결합 시범사례

사회보장위원회는 인구통계학적 정보(소득·연령 등)와 노후보장 데이터(공적, 사적)를 결합하여 소득보장체계(기초보장-국민연금-퇴직연금-개인연금) 분석하기 위해 국세청의 자료를 가명 처리하여 결합·활용하는 분석을 진행하고 있다. 결합대상정보는 과세정보(소득), 주민등록정보, 지방세정보(재산), 사회보장정보(국민연금 등), 금융정보(퇴직연금, 개인연금 유무 등) 등이다. 사회보장위원회는 가명처리 정보의 결합을 통해 현재 노후보장제도의 평가와 개선과제를 발굴하고, 현재 노령층의 사각지대 및 미래 노령층의 노후대비 현황을 파악하는데 활용하기 위한 방안을 마련하기 위한 방안을 마련하기 위한 연구를 진행 중이다.

[그림 3-4] 사회보장위원회 가명정보 결합 시범사례 결합 추진방안



자료: 개인정보보호위원회(2021.11.), '가명정보 결합 추진현황 및 시범사례'

한국임업진흥원은 산림치유 참가자의 정보와 질병 등 건강정보를 결합하여 건상상태와 산림치유 프로그램 참여 간 상관관계를 분석하는 연구를 추진하고 있다. 결합대상정보는 임업진흥원이 보유하고 있는 산림치유 프로그램 참가자의 정보(연령, 성별, 직군, 거주지역 등)와 건강보험관리공단이 보유하고 있는 해당 참가자의 건강정보(질병력, 상해, 유전, 증후군 이력 등)이다. 진흥원은 가명정보 결합을 통해 특성별로 대상자 군을

구분하고, 향후 치유목적의 산림자원-의료연계 산업이 육성될 수 있도록 만드는 정책의 근거자료로 활용할 계획이다.

[그림 3-5] 한국임업진흥원 가명정보 결합 시범사례 결합 추진방안

자료: 개인정보보호위원회(2021.11.), '가명정보 결합 추진현황 및 시범사례'

〈표 3-11〉 공공분야 가명정보 결합 시범사업 현황

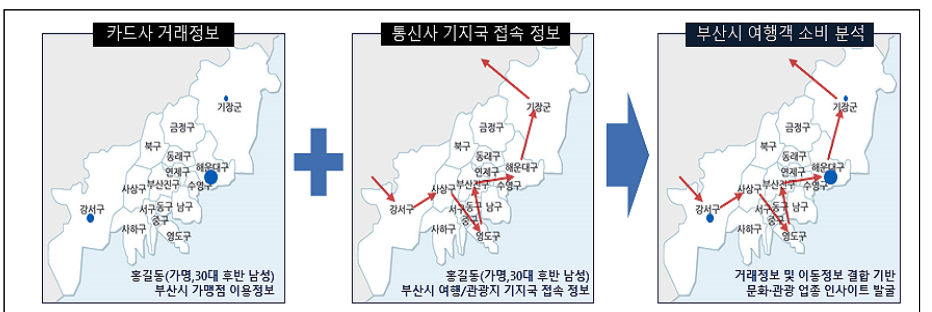
분야	시범사례	정보보유기관 및 대상정보	기관
1.의료+인구	① 암 질병 치료효과 분석	(암센터, 병원) 암 치료 임상정보 (건보공단) 진료정보 (통계청) 사망원인정보	국립암센터
	② 암 환자의 합병증 및 만성 질환 예측 연구	(암센터, 병원) 암 치료 임상정보 (건보공단) 진료정보	
2.금융+보훈	③ 국가보훈대상자 신용실태 연구	(보훈처) 생활안정지원 (신정원) 개인신용정보	국가보훈처
3.소득+복지	④ 노후소득보장 종합연구	(사보원, 연금공단 등) 사회보장정보 (행안부) 재산정보 (국세청) 소득정보 등	사회보장위원회
4.통신+유통	⑤ 불법스팸 실태연구	(인터넷진흥원) 스팸정보 (통신사) 스팸의심정보,가입정보	한국인터넷진흥원
	⑥ 지역별, 상권별, 상품별 소비패턴 분석	(통신사) 위치정보 (유통사) 연령·지역별 구매내역 등	통신사
5.레저+건강	⑦ 맞춤형 산림치유 프로그램 분석	(임업진흥원) 산림치유자정보 (건보공단) 의료정보	한국임업진흥원

자료: 개인정보보호위원회(2021.11.), '가명정보 결합 추진현황 및 시범사례' p.3 표 재구성.

나. 민간분야 가명정보 활용 사례

민간분야[32]에서는 데이터3법이 시행됨에 따라 기업 상호 간의 협력을 통해 가명정보 결합을 추진하고 있다. 문화관광 분야 첫 번째 사례로 심원섭외(2018) 에서 데이터 결합의 필요성을 제시한 적은 있으나 실제 결합사례는 2020년 9월 한국문화관광연구원에서 처음으로 진행하였다, 이는 신용카드사업자가 보유한 가입자별 카드이용정보와 이동통신사 이용자의 기지국 접속 정보를 결합하여 여행 및 관광 정보를 분석하였다.[33] 이동통신사업자의 가입자별 기지국 접속정보를 활용해 주요 관광지에 위치한 사람들이 거주민인지 관광객인지 여부를 구별하였고, 관광객으로 구별된 가입자의 이동 정보를 파악하여 관광객의 주요 관광지 방문 순서를 파악하였다. 신용카드사업자의 가입자별 카드이용정보를 활용해 가입자가 방문한 관광지와 주요 상권에서 얼마나 지출하였는지, 어떤 업종을 이용하였는지, 얼마나 오래 이용하였는지 등의 정보를 파악하였다. 이렇게 두 민간사업자가 보유한 가명처리 정보를 활용하면 관광분야 정책추진 기관에게 여행 및 관광 정책수립 시 활용할 수 있는 실증분석 자료를 제공할 수 있을 뿐만 아니라 관련 기업은 고객 특성별로 맞춤형 여행정보를 제공하거나 관련된 사업을 추진해 수익성을 높일 수 있을 것으로 기대된다.

[그림 3-6] 민간분야(통신·금융) 가명처리정보 결합활용 사례



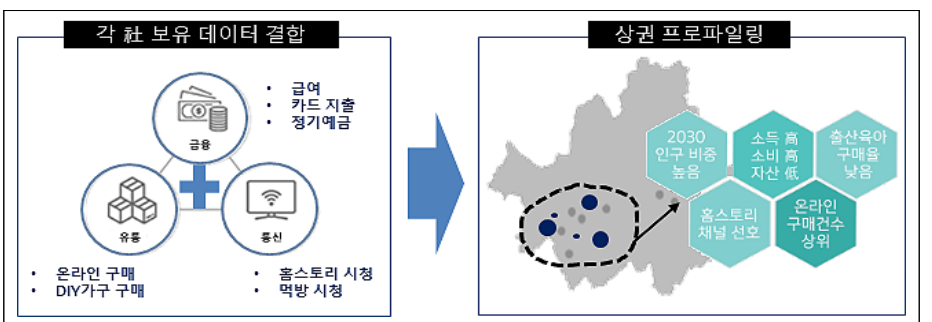
1 카드 이용정보 ✚ 기지국 접속 정보 ➡ 여행/관광 정보 분석(신한카드, SK텔레콤)
카드사 보유 카드 이용정보와 통신사의 고객 기지국 접속 정보를 결합하여 여행/관광 정보 분석 ➡ (공공기관) 여행/관광 정책수립, (기업 등) 고객 특성별 선호 여행지 정보 제공

32) 민간분야의 사례는 공공분야 사례와 다르게 과정과 성과를 별도로 공개하지 않아 추진사례 중심으로 서술하였다.

33) 한국문화관광연구원과 SK텔레콤, 신한카드는 문화·체육·관광 분야의 가명처리 데이터 및 가명처리 데이터 간 결합 활성화를 위해 2020년 9월 업무협약을 체결하여 가명처리 데이터 결합을 진행하였다. 3개 기관의 업무협약 체결은 가명데이터 간 결합 활용 1호 사례로 문화·체육·관광 분야 중 관광분야에 대한 분석을 위한 목적으로 데이터 간 결합을 시도이다.

다음은 통신·금융·유통분야의 가명처리정보 결합추진 사례를 살펴본다. 신한은행(금융), CJ올리브네트웍스(유통), LG유플러스(통신)는 금융정보와 비금융정보 간 결합을 통해 ① 빅데이터 네트워크 및 얼라이언스 구축으로 데이터 공동수집, 활용체계 마련, 고객행동 공동연구 진행, ② 소비자 중심의 빅데이터 기반 머신러닝, 딥러닝 등 AI를 통한 맞춤형 데이터 플랫폼 개발, ③ 데이터 신사업 발굴 및 추진을 위한 데이터 유통상품 개발, 디지털 마케팅 협력 등 3대 과제를 추진하기 위한 협약을 2020년 11월 체결하였다. 결합되는 가명정보는 2,500만 명의 금융정보(소득, 소비, 자산 등), 2,700만 명에 대한 유통서비스/재화 이용/구입정보, 1,600만 명의 통신서비스 이용자 정보이다. 이들 기업들은 먼저 다음과 같은 가명정보 결합을 통해 시범 분석을 추진하고 있다. 통신사업자의 IPTV 시청 정보, 급여·카드·예금 등의 금융정보, 지역별 거주자의 물품구매 정보 등을 결합하여 상권별 거주자 소비행태 분석을 시도 중이다. 이들 기업들은 향후 분석결과를 통해 상권별 공동 마케팅 전략을 추진한다는 계획을 수립하였다.

[그림 3-7] 민간분야(통신·금융·유통) 가명처리정보 결합활용 사례



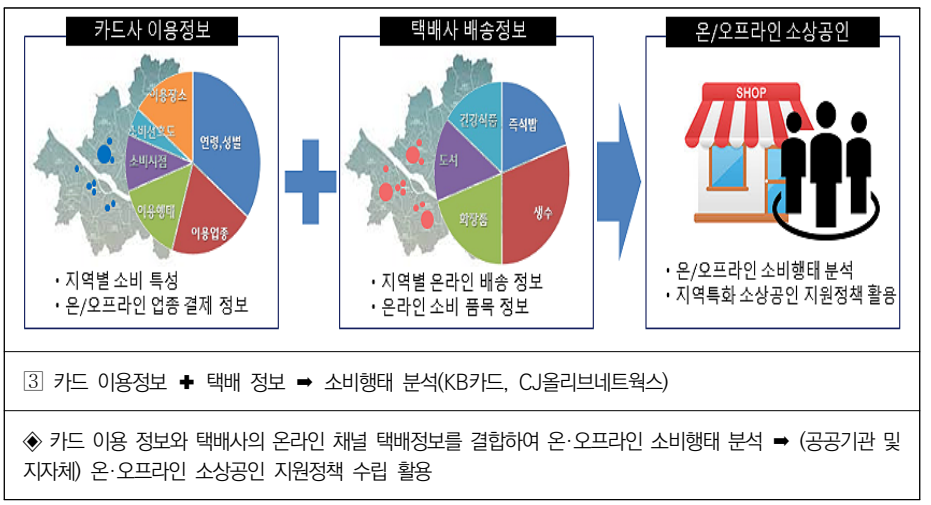
2 소득/소비/자산 정보 ✚ 온라인 채널 택배정보 ✚ IPTV 시청정보 ➡ 상권별 소비행태 분석

은행의 소득/소비/자산 정보와 택배사의 택배정보, 통신사의 IPTV 시청정보를 결합하여 상권별 거주자 소비행태를 분석 ➡ (소상공인) 주거지 인근 상권 마케팅 전략 수립, (공공기관) 정책 수립 및 행정 서비스 제공

다음으로 금융정보와 유통정보의 결합 사례를 살펴본다. KB국민카드(금융)와 CJ올리브네트웍스(유통)은 공동마케팅을 위해 금융정보와 비금융정보의 결합을 추진하였다. 이들의 결합에서 활용된 정보는 금융정보 중 지역별 소비의 특성, 온·오프라인 업종 결제정보이고, 비금융정보 중 지역별 온라인 배송 정보 및 온라인 소비 품목 정보이다. 신용카드 결제정보와 배송정보를 결합하면 특정 사업자의 물품이 어느 지역으로 많이 유통

되고 있는지를 파악할 수 있으며, 시간대별 이용정보와 지역정보를 활용하면 특정지역에 특정 업종에 대한 타겟 마케팅에도 활용할 수 있어 기업의 수익극대화에 도움이 될 수 있을 것으로 보인다. 이와 함께 해당 정보를 공공분야에서 이용할 경우 사업자 업종별, 지역별 맞춤형 마케팅 컨설팅을 소상공인에게 제공할 수 있을 것으로 기대된다.

[그림 3-8] 민간분야(금융·유통) 가명처리정보 결합활용 사례



현재 추진 중이라고 밝힌 통신정보와 관광정보의 결합 사례도 존재한다. 통신사업자(KT)와 관광플랫폼(야놀자)는 통신사업자가 보유한 가명 처리된 이동데이터와 야놀자가 보유한 가명처리된 관광서비스 이용자의 이용정보를 결합하여 신규 사업을 추진하기 위한 계획을 지난 2021년 9월 수립하였다. 해당 관광플랫폼은 국내 숙박도매시장에서 시장점유율이 높은 사업자이고, 해당 통신사업자는 시장점유율이 유선통신 서비스시장 1위, 무선통신 서비스시장의 2위 사업자라는 점에서 각 사의 가명정보를 결합해 활용하더라도 대표성을 가질 수 있을 것으로 보인다. 이들 기업들은 가명정보 결합을 통해 개인의 취향을 파악하여 매칭하는 방식을 넘어서 위치와 시간에 따라 추가적인 제안을 할 수 있는 비즈니스 모델을 구축하기 위한 협업을 추진하고 있다.

이렇듯 주요 사례를 통해 다양한 형태의 정보가 결합 가능해짐에 따라, 소비 및 구매기록, 이동동선 정보, 식당 방문기록, 건강보험, 진료기록 등 방대한 자료를 바탕으로 고부가가치 데이터를 생성 활용할 수 있는 기반이 마련되었음을 확인할 수 있다.

민간기업은 다양한 기관과의 협업을 통한 가명 정보 결합, 분석이 가능하게 되어 빅데이터에 근거한 고객 라이프 스타일 특성 분석, 소비 패턴의 추정, 안전하고 정확한 진단·검사·치료법 개발, 자산관리 포트폴리오 추천 등으로 고객의 삶의 질 향상에 기여할 수 있을 것이다. 산업계에서는 결합정보를 바탕으로 제품·서비스 수요 발굴 및 모형(모델) 검증을 통해 신규 서비스 개발이 가속화될 것으로 보인다. 정부 및 공공기관에서는 가명정보의 결합·활용을 통해 데이터에 기반을 둔 공공정책 수립과 정밀한 정책 효과성 평가 등에 활용할 수 있을 것으로 기대하고 있으나 데이터 접근성, 분석환경 등의 제약으로 민간기업의 적극적인 협조가 필요하다. 또한 민간기업이 가명처리정보 제공과 결합에 동의하더라도 민간기업의 방대한 자료를 분석할 수 있는 안전한 환경의 구축도 필요하다.

〈표 3-12〉 민간분야 가명정보 결합추진 사례

분야	사례내용	기대효과	결과
통신·금융	주요 관광지 이동, 소비지출분석	맞춤형 관광정보 제공	추진중
통신·금융·유통	상권별 소비행태분석	기업 간 공동마케팅 체계구축	추진중
금융·유통	지역별·업종별 소비행태분석	온라인 소상공인 컨설팅 활용	추진중
통신·금융·관광	위치·시간별 추천알고리즘 설계	고객 세분화로 수익극대화	추진중

다. 조사분야 가명정보 활용 사례

통계청은 사업자등록번호[34]를 연계키로 활용해 기업단위 국가승인통계 및 행정정보를 활용해 신규 통계를 제공하고 있다. 다양한 사례 중 하나로 통계청이 제공하고 있는 영리법인통계는 다음 표와 같이 여러 기관이 생산·보유하고 있는 조사·행정정보를 법인등록번호, 사업자등록번호를 활용해 결합하여 작성하고 있다. 이렇게 작성된 통계를 통해 영리법인 기업체의 모집단을 구축하고, 해당 기업에서 종사하고 있는 종사자DB를 구축할 수 있으며, 기업의 규모와 재무와 관련된 사항도 한 번에 파악할 수 있게 하였다.

34) 최근 대통령 직속기구인 '4차 산업혁명위원회'가 사업자등록번호의 일반 대중에 대한 공개를 허용(2021. 05.13.)하면서 기업단위 데이터의 결합도 활성화되고 있다. 위원회의 결정 이전에는 사업자등록번호를 주민등록번호에 준하는 번호로 관리하였고, 법인의 비밀사항 중 하나로 보았다. 하지만 이번 결정을 통해 사업자등록번호는 민간에 개방되었고, 이를 활용한 기업단위 데이터 연계가 활성화될 것으로 보인다.

〈표 3-13〉 통계청 영리법인통계작성을 위한 수집자료 목록

제공기관	자료명	입수일정
국세청	사업자등록자료, 부가가치세자료	매월
국세청	법인세자료	5월
국세청	근로소득지급명세서	7월
국민건강보험공단	건강보험자료	5월
국민연금공단	국민연금자료	매월
한국고용정보원	고용보험자료	매월
근로복지공단	산재보험자료	매월
공정거래위원회	상호출자제한기업집단 소속기업	5월, 9월, 12월

자료: 통계청(2019), 영리법인통계 통계정보보고서

민간분야에서도 조사정보 간 결합을 추진하고 있다. 매달 여러 기업과 기관에서 요청하고 있는 조사를 위탁 수행하는 일부 온라인 시장조사업체는 자제적으로 조사한 조사정보 간 결합 활용을 시도하고 있다. 한 온라인 시장조사업체(마이크로밀 엠브레인)는 다양한 기관에서 요청하고 있는 조사정보 중 App 이용정보, 오프라인 방문정보, 카드결제정보 등이 개별적으로 수집하고 있다는 점을 인식하고, 통계수집 목적으로 보유하고 있는 조사대상의 개인정보를 연계키로 활용해 개별적으로 수집되고 있는 조사정보를 결합하였다. 이 과정을 통해 전체 조사대상 샘플 중 각 조사별로 중복되어 있는 조사대상을 구분하여 이들의 응답정보를 결합해 1,000명에 대한 월 단위 패널 데이터를 구축하였다. 시장조사업체의 자체조사 데이터 결합 및 활용은 향후 관련 법령이 완화되어 가명데이터로 결합이 가능할 경우, 많은 장점을 가지게 될 것으로 보인다. 현재는 가명데이터 결합 시 개인 식별가능성이 높아지기 때문에 개인식별이 불가능하도록 최대한 데이터를 범주화하고 있다. 현재 규제체계에서 조사데이터를 가명처리 후 결합한다면 모든 응답지가 범주화되면서 결합이전보다 나은 결과를 도출하기 어려울 것으로 보인다. 향후 조사데이터를 조사된 결과 자체로 결합할 수 있도록 제도가 개선된다면 기존 조사된 결과의 활용가능성이 크게 확대될 뿐만 아니라 각 분야별 설문조사 결과 간 결합 등이 가능해져 분야 간 연계분석도 활성화될 것으로 보인다.

〈표 3-14〉 온라인 시장조사업체 결합 데이터 정보

구분	칼럼
APP 이용정보	사용일자, App 카테고리, App 이름, 총 사용시간
오프라인 방문정보	매장방문일자, 매장카테고리, 매장 층수
카드결제 정보	문자수신날짜, 문자수신시간, 결제타입, 결제형태, 사용처 카테고리, 총 결제금액, 실 결제금액 등

자료: 금융데이터거래소

제3절 소결

3장에서는 접근가능한 문화·관광분야 데이터를 기반으로 가명정보를 활용할 수 있는 공공·민간분야의 데이터를 검토함으로써, 향후 문화관광분야의 새로운 통계 생산과 연구에 활용할 수 있는 데이터가 무엇인지 살펴보았다. 소결에서는 이를 바탕으로 앞으로 제시될 4장과 5장에 활용될 가명처리 데이터 기반의 실증분석 데이터를 선정하고자 한다.

본 연구는 가명처리 데이터를 활용한 실증 분석을 통해 (ⅰ)기존 데이터 활용범위의 확장성을 검토하고 (ⅱ) 가명정보를 이용한 데이터 결합을 실시하여, 가명처리 데이터 기반의 문화관광 부문의 새로운 정책적 가치를 도출하는 것이 주요 목적이다. 이러한 연구 목적에 따라 연구에 활용고자 하는 데이터 선정을 위해 크게 두 가지 사항을 고려하였다.

고려사항 1: 기 생산된 통계 및 연구와 차별화된 정책적 가지 도출 가능성

우선, 기 생산된 통계 및 연구와 차별화된 정책적 가치를 도출할 가능성이 있는지 여부를 고려하고자 하였다. 개인 정보가 가명처리된 데이터를 사용하는 경우, 데이터 전처리 과정에서 많은 시간과 예산이 소요되고, 추가적인 법률적 검토도 필요하다. 이에 가명처리된 데이터를 사용하는 것이 기존 데이터 분석과 정책적으로 차별성을 가질 수 있어야 하므로 실증분석을 위한 데이터 선정 시 이러한 가능성을 염두에 두는 것이 필수적이라고 판단된다.

고려사항 2: 연구기간, 연구예산 등 현실적 여건을 고려한 실제 활용가능성

활용하고자 하는 가명처리 데이터가 정책적으로 유의미한 분석이 가능하다고 판단된다 하더라도 현실적으로 실제 이용이 가능한지 유무가 중요하다고 할 수 있다. 따라서 현실적인 연구 기간과 예산, 데이터의 접근 가능성을 종합적으로 고려하여

최종 활용 데이터를 선정할 필요가 있다.

이에 각 데이터의 특성과 주요 고려사항을 반영하여 문화관광분야야 가명처리 활용방안 도출을 위한 데이터는 SK텔레콤 이동통신 데이터와 신한카드 카드결제액 데이터를 선정하였다. (참고: 〈표 3-15〉) 공공분야의 데이터는 정부 정책평가를 위한 유의미한 분석 변수를 가지고 있으나 아직 가명정보 반출 및 제공이 활발하게 이뤄지지는 않고 있어 이용하기에는 제약이 많다. 또한 이종의 데이터와 결합 시, 유효표본 확보의 어려움이 있을 것으로 판단된다. 예를 들어 문화누리카드 DB를 이용하여 차상위 계층의 문화관광분야 지출액을 분석하고자 신용카드 지출액 데이터와 결합하는 경우, 충분한 유효표본을 확보하는데 한계가 있다. 그 외, 조사데이터의 경우는 연구 및 통계생산을 목적으로 설계되어 있기 때문에 연구에 활용될 수 있는 유의미한 변수가 많다고 하지만 다른 데이터와 결합하여 활용 하려면 개인정보 수집·활용 재 동의를 받아야 하므로 시간 및 비용이 발생하게 된다.

한편 민간 데이터의 경우, 연구 및 통계분석을 목적으로 생산하고 있는 데이터가 아니기 때문에 이용 시 연구나 분석의 목적에 맞춰 데이터와 변수를 재가공 해야 하여 별도의 시간과 노력이 요구된다. 또한 기업 및 내부 고객에 대한 정보이기 때문에 개인정보 보안을 문제로 활용할 수 있는 정보를 확보하는데 어려움이 있다. 하지만 현재 공공 및 조사데이터 보다 가명정보를 활용한 분석 및 서비스 제공이 활발하게 이루어져 관련된 연구 진행에 비교적 적은 시간이 소요된다. 또한 매일 새로운 데이터가 적재되기 때문에 가장 시의성 높은 통계를 생산할 수 있다는 이점이 있다. 이에 본 연구에서는 문화관광분야의 가명처리 데이터 활용방안을 도출을 위한 실증 분석시, 민간 데이터를 이용하여 분석하고자 한다. 특히 본 연구에서는 민간 데이터 중 SK텔레콤의 통신 데이터와 신한카드의 소비 지출액 데이터 2종을 활용방안 도출을 위한 가명처리 데이터로 선정하였다. SK 텔레콤과 신한카드의 경우 연구원과 MOU를 통해 업무 협약을 맺은 바 있을 뿐만 지속적으로 연구원과 업무를 진행하였기 때문에 연구를 위한 추가적인 데이터 및 변수의 가공에 걸리는 시간이 타 사의 데이터를 이용했을 때 보다 더 적을 것으로 예상된다. 뿐만 아니라 해당 두 기업의 데이터로 가명정보를 활용하여 결합한 가명처리 데이터를 생산한 경험이 있기 때문에 연구의 현실적인 기간과 예산을 고려하여 최종적으로 SK텔레콤과 신한카드 데이터를 선정하는 것이 적합하다고 판단하였다. 마지막으로 통신데이터와 카드데이터가 가지고 있는 주요 정보가 문화관광분야에서 활용할 수 있는 다양한 분석변수를 포함하고

있기 때문에 향후에도 관련분야의 분석 및 연구에서 양사의 데이터는 주요데이터로 활용될 것이라 판단된다.

〈표 3-15〉 가명처리 데이터 활용방안 도출을 위한 데이터 선정

공공데이터	민간데이터	조사데이터
(+) 정부의 정책을 평가할 수 있는 유의미한 변수 多 (+) 활용 가능한 정보 확인을 위한 접근성 용이 (-) 가명정보 반출 및 제공이 활발하지 않아 이용의 제약 (-) 데이터 결합 시 충분한 유효표본 확보가 어려움	(+) 관련 분석 및 서비스 제공 활발, 상대적으로 적은 시간 소요 (+) **충분한 유효표본 확보와 적시성 있는 통계 생산 가능** (+) **대규모의 패널 데이터 구성이 가능** (-) 보안 규정으로 인하여 활용 가능한 컬럼 확보 어려움 (-) 추가적인 데이터 및 변수 가공 필요	(+) 정부의 정책을 평가할 수 있는 유의미한 변수 多 (+) 활용 가능한 정보 확인을 위한 접근성 용이 (-) 적시성 있는 통계 생산 어려움 (-) 데이터의 결합 시 추가적인 개인정보 수집 및 활용 재동의필요

고려사항 1: 기 생산된 통계 및 연구와 차별화된 정책적 가치 도출 가능성 검토

⇩

고려사항 2: 연구기간, 연구 예산, 접근가능성 등을 고려한 실제 활용성 점검

⇩

㈜SK텔레콤 이동통신 데이터, ㈜신한카드 카드결제액 데이터 선정

최종적으로 분석데이터로 선정된 통신데이터와 카드데이터를 기반으로 제4장과 5장에서 가명처리 데이터의 활용방안을 도출을 위해 이동통신데이터와 카드 지출액데이터를 활용·결합 하여 분석할 수 있는 주제를 살펴보고 가명처리 데이터를 이용한 실증분석을 통해 문화관광 부문 활용 방안을 제시하도록 하겠다.

제4장

[가명처리 데이터] 문화·관광 부문 활용방안

제1절 데이터 구성 절차도

지금까지 가명처리 데이터의 이론적 개념과 기존통계와 다른 차별적 특징, 정책적 기대효과를 이론적으로 살펴보았다면, 4장부터는 실증분석을 통해 '관광 문화콘텐츠 분야' 연구 및 통계에 있어 가명처리 데이터의 실제 활용가능성과 정책적 가치를 살펴보고자 한다.

가명처리 데이터를 활용하기 위해서는 기존과 달리 다양한 법률적, 행정적, 실무적인 절차들이 요구된다. 본 연구에서는 가명처리 데이터를 구성하기 위 한 업무를 크게 분석 환경구성, 데이터 구성, 반출 및 분석단계로 구분하여 진행하였으며, 이는 다음 〈표 4-1〉과 같이 도식화 할 수 있다.

〈표 4-1〉 가명처리 데이터 활용을 위한 업무 절차도

구분	주체
1단계: 데이터 활용 관련 업무 협의 데이터의 활용을 위한 데이터 제공기관과 이용기관 간, 업무 협의	제공기관, 이용기관
⇩	
2단계: 가명처리 데이터 분석환경 조성 가명정보 활용 관련 내부관리계획 수립 및 점검 가명정보 활용 분석 환경 점검	이용기관
⇩	
3단계: 가명처리 데이터 구성 (*상향식 접근(bottom-up)방식) 가명처리 데이터 활용주제 선정(이용기관) 활용 가능한 컬럼 보유 여부 확인(이용기관, 제공기관) 선정된 컬럼의 활용 관련 범무적 검토(제공기관) 최종컬럼 확정 후 가명 및 익명처리(제공기관)	제공기관, 이용기관
⇩	
4단계: 가명처리 활용 및 반출 적정성 평가 반출 신청서 작성(이용기관) 가명처리 데이터 반출 및 활용 적정성 평가위원회 개최(제공기관) 비밀보호협약서 작성(이용기관, 제공기관)	제공기관, 이용기관
⇩	
5단계: 데이터 반출 및 문석 개인정보 보호를 위해 암호화 전송(제공기관) 지정된 분석 환경 내 업무 진행(이용기관)	이용기관

1. 데이터 활용 관련 업무 협의

가명처리 데이터 활용을 위하여 우선적으로 데이터 제공 기관과 이용 기간과의 업무 협의 및 역할에 대한 정의가 필요하다. 이에 본 연구에서 가명처리 데이터 분석을 위해서 SK텔레콤과 신한카드사는 결합 대상 데이터를 제공하는 '제공기관', 연구원은 결합된 데이터를 이용하는 '이용기관'으로 다음 〈표 4-2〉 과 같이 업무를 구분하였다.

〈표 4-2〉 데이터 결합을 위한 결합 데이터 제공기관, 이용기관 주요 업무 내용

데이터 제공기관 -SK텔레콤, 신한카드-	데이터 이용기관 -한국문화관광연구원-
- 데이터 제공 - 데이터 익명 및 가명처리 - 데이터 반출 적정성 심사	- 가명처리 데이터 분석환경 조성 - 가명처리 데이터 활용 주제 선정 - 데이터 활용 및 분석

2. 가명처리 데이터 분석 환경 조성

가명처리 데이터를 분석하기 위해서는 데이터 제공기관과 이용기관 모두 가명정보 및 추가정보를 안전하게 관리하기 위한 내부 관리계획을 수립·시행하여야 한다.[35] 만약 가명처리 내부관리계획이 마련되어 있지 않다면 데이터를 제공하거나 이용할 수 없다. 이에 본 연구원에서는 2021년 7월 개인정보 보호 내부관리 계획 내 다음 내용을 포함한 가명정보 보호 내부관리 계획을 수립·시행하여 가명처리 데이터 이용 및 제공기관으로써 자격을 갖춘 바 있다.

가명정보 처리 내부관리 계획 포함 주요 내용

가. 가명정보 또는 추가정보의 관리책임자 지정에 관한 사항
나. 추가정보 별도 분리 보관
다. 가명정보 또는 추가정보의 안전성확보조치에 관한 사항
라. 가명정보처리자의 교육에 관한 사항
마. 가명정보 처리 기록 작성 및 보관에 관한 사항
바. 개인정보 처리방침 공개에 관한 사항
사. 가명정보의 재식별 금지에 관한 사항
※ 상기 내용에 포함되지 않은 항목은'개인정보

35) 개인정보보호법「제29조의5 1항 1호」

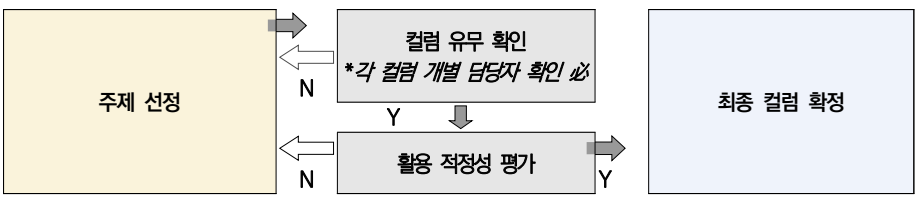
또한 가명정보 관리 내부계획에 따라서 실제로 가명처리 데이터를 직접 분석 및 이용할 수 있는 환경을 조성하였다. 가명정보는 개인정보와 물리적으로 분리하여 저장할 수 있는 공간을 확보하였으며, 별도의 외부 인터넷 접속이 차단된 PC에 저장하여 분리·보관하였다. 가명정보가 저장된 PC는 물리적 접근이 제한된 공간에서 관리하고 접근은 지정된 가명정보 취급자(분석가) 총 3명만 가능하도록 암호화 하였다.

3. 가명처리 데이터 구성

본 분석에서는 가명처리 데이터의 제공기관은 SK 텔레콤과 신한카드사, 이용기관은 연구원으로 역할을 구분하였는데, 여기서 가명처리 데이터를 구성하기 위한 활용 가능한 컬럼 정보에 대한 접근은 데이터 제공기관의 내부 규정를 기반으로 진행된다.

SK텔레콤과 신한카드 모두 각 사가 보유한 활용 가능한 정보를 공개하지 않는 것을 원칙으로 하고 있다. 따라서 이용기관인 연구원에서 분석주제를 우선적으로 설정하고, 제공기관 내 관련한 컬럼정보가 있는지 확인하는 상향식 접근(bottom-up)방식으로 업무를 진행하였다. 만약 활용 가능한 컬럼 정보가 없다면 다시 주제를 설정하게 되고, 정보가 있다면 활용이 가능한지에 대한 내부 법률적 검토를 거쳐 최종적 활용 컬럼을 확정하였다.

〈표 4-3〉 데이터 결합을 위한 결합 데이터 제공기관, 이용기관 주요 업무 내용



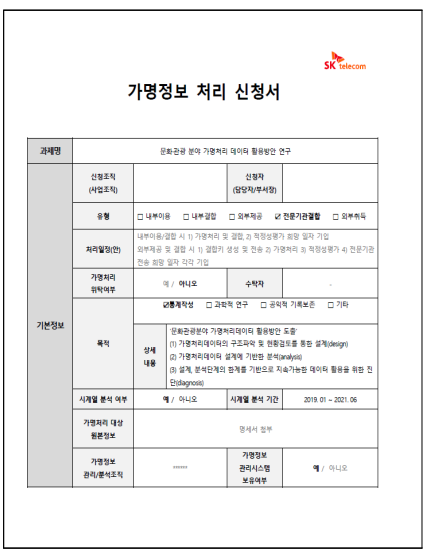
또한 각 사는 개인정보보호 및 비밀 유지를 위하여 내부적으로 각 컬럼정보를 각기 다른 서버에서 관리하고 있으며, 접근 권한 또한 각 서버에 따라 다르게 지정하고 있다. 따라서 최종적으로 하나의 데이터 셋으로 구성하기 위해서는 각 서버에 접근할 수 있는 권한을 가진 관리자에게 활용 유무를 확인한 후, 필요한 각 컬럼정보를 추출해 하나의 데이터 셋으로 구성하는 별도의 내부적인 결합 절차가 필요해 많은 시간이 소요된다.

이렇듯 데이터 제공기관의 우수한 개인정보 관리체계로 개인정보에 대한 유출과 침해 가능성이 최소화되는 반면, 가명처리 데이터를 구성하는 과정은 비효율적으로 진행되는 구조로 담당자의 업무 부담이 가중될 수 밖에 없다.

4. 가명처리 데이터 활용 및 반출 적정성 평가

해당 과정을 통해 선정된 컬럼을 바탕으로 개인정보를 가명처리한 후 데이터를 구성하게 되면, 데이터 제공기관 내 평가위원회를 개최하여 반출 및 활용의 적정성을 최종적으로 확인하게 된다. 이에 연구원은 가명정보 활용에 대한 목적과 분석방법을 포함한 신청서를 데이터 제공기관에 제출하고, 데이터 활용에 관한 비밀보호계약을 체결하였다. 단, 가명처리 데이터의 경우 전수 반출이 아닌 결합데이터 용으로 일부를 반출하였다.36)

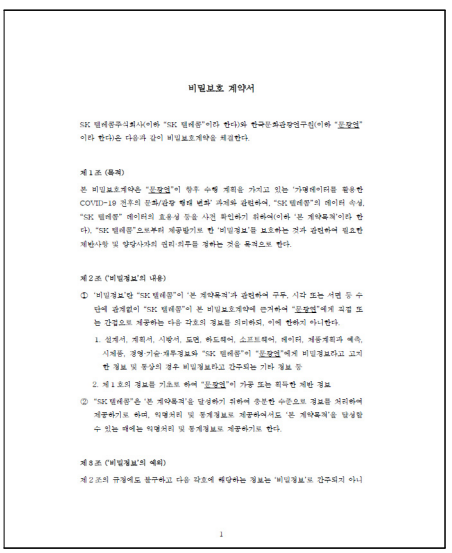
[그림 4-1] ㈜SK텔레콤-한국문화관광연구원 가명처리 신청서 및 비밀보호 계약



SK telecom

가명정보 처리 신청서

과제명	문화관광 분야 가명처리 데이터 활용방안 연구				
기본정보	신청조직 (사업조직)		신청자 (담당자/부서장)		
	유형	□ 내부이용 □ 내부결합 □ 외부제공 ✓ 전문기관결합 □ 외부취득			
	처리절차(안)	내부이용/결합 시 1) 가명처리 및 결합, 2) 적정성평가 희망 일자 기입 외부제공 및 결합 시 1) 결합키 생성 및 전송 2) 가명처리 3) 적정성평가 4) 전문기관 전송 희망 일자 각각 기입			
	가명처리 위탁여부	예 / 아니오	수탁자	-	
	목적	상세 내용	필통계작성 □ 과학적 연구 □ 공익적 기록보존 □ 기타 문화관광분야 가명처리데이터 활용방안 도출 (1) 가명처리데이터의 구조파악 및 현황검토를 통한 설계(design) (2) 가명처리데이터 설계에 기반한 분석(analysis) (3) 설계, 분석단계의 한계를 기반으로 지속가능한 데이터 활용을 위한 진단(diagnosis)		
	시계열 분석 여부	예 / 아니오	시계열 분석 기간	2019. 01 – 2021. 06	
	가명처리 대상 원본정보	명세서 참부			
	가명정보 관리/분석조직	-------	가명정보 관리시스템 보유여부	예 / 아니오	



비밀보호 계약서

SK 텔레콤주식회사(이하 "SK 텔레콤"이라 한다)와 한국문화관광연구원(이하 "문광연"이라 한다)은 다음과 같이 비밀보호계약을 체결한다.

제 1 조 (목적)

본 비밀보호계약은 "문광연"이 향후 수행 계획을 가지고 있는 '가명데이터를 활용한 COVID-19 전후의 문화/관광 행태 변화' 과제와 관련하여, "SK 텔레콤"의 데이터 속성, "SK 텔레콤" 데이터의 효용성 등을 사전 확인하기 위하여(이하 '본 계약목적'이라 한다), "SK 텔레콤"으로부터 제공받으 한 '비밀정보'를 보호하는 것과 관련하여 필요한 제반사항 및 당당사자의 권리·의무를 정하는 것을 목적으로 한다.

제 2 조 ('비밀정보'의 내용)

① '비밀정보'란 "SK 텔레콤"이 '본 계약목적'과 관련하여 구두, 시각 또는 서면 등 수단에 관계없이 "SK 텔레콤"이 본 비밀보호계약에 관계하여 "문광연"에게 직접 또는 간접으로 제공하는 다음 각호의 정보를 의미하며, 이에 한하지 아니한다.

1. 설계서, 계획서, 사양서, 도면, 하드웨어, 소프트웨어, 데이터, 제품계획과 예측, 시제품, 경영·기술·재무정보와 "SK 텔레콤"이 "문광연"에게 비밀정보라고 고지한 정보 및 통상의 경우 비밀정보라고 간주되는 기타 정보 등

2. 제 1 호의 정보를 기초로 하여 "문광연"이 가공 또는 획득한 제반 정보

② "SK 텔레콤"은 '본 계약목적'을 달성하기 위하여 충분한 수준으로 정보를 처리하여 제공하기로 하며, 익명처리 및 통계정보로 제공하여서도 '본 계약목적'을 달성할 수 있는 때에는 익명처리 및 통계정보로 제공하기로 한다.

제 3 조 ('비밀정보'의 예외)

제 2 조의 규정에도 불구하고 다음 각호에 해당하는 정보는 '비밀정보'로 간주되지 아니

1

36) 카드사와 통신사 동시 가입자만 추출하여 사용하였으며 이를 대상으로 추후 데이터 결합을 진행

5. 데이터 반출 및 분석

최종 데이터 반출 승인 후, 데이터 제공기관은 데이터 전문 결합기관인 금융보안원의 데이터분석(https://data.fsec.or.kr/)에 본 데이터를 업로드하고 연구원은 지정된 분석 단말기를 이용하여 방화벽 설정을 통해 데이터 분석센터만 접근 가능하도록 설정하여 데이터를 다운 받은 후 네트워크를 차단하여 반출하였다.

제2절 최종 가명처리 데이터 구성

1. 컬럼 선정

가명처리 데이터의 관광분야의 활용방안을 도출할 수 있는 적합한 컬럼을 선정하고자 다음 사항을 고려하였다.

고려사항 1: 관광 및 문화콘텐츠 분야의 시사점 도출에 적합한 컬럼 선정

실제로 통신 데이터에는 다양한 개인 및 고객정보가 존재하며 이는 연구 목적 보다는 내부의 마케팅 용도 또는 고객 관리 등 다양한 목적으로 활용되고 있다. 이에 본 연구에는 관광분야의 시사점 도출이 가능하다고 판단되는 컬럼 일부를 선택적으로 사용하였다. 물론 데이터의 모든 정보를 활용하는 방법도 있지만, 컬럼의 정보가 많아질수록 개인을 유추할 수 있는 정보가 많아지는 것이기 때문에 개인정보를 침해할 가능성이 높아진다. 또한 전체 컬럼의 반출을 승인받게 되더라도 데이터의 크기가 매우 커지기 때문에 현재의 분석 시스템 환경에서 이를 활용하는 것이 현실적으로 불가능하다. 따라서 관광 및 문화콘텐츠 분야의 유의미한 시사점을 보여줄 수 있다고 판단되는 컬럼 위주로 구성하도록 하였다

고려사항 2: 충분한 시계열 자료 확보가 가능한 컬럼 선정

두 번째로 관광 및 문화콘텐츠 분야의 변화를 살펴보고 향후를 진단하기 위하여 충분한 시계열 자료 확보가 가능한지 여부를 고려하였다. 해당 분야의 유의미한 시사점을 도출할 수 있다 하더라도 충분한 기간 동안의 정보를 사용할 수 없다면 그 변화를 파악하기 어렵다. 따라서 최소 2년 간 시계열 자료 확보가 가능한지 여부를 판단하여 선정하였다.

고려사항 3: 개인정보 침해하지 않는 수준의 컬럼 선정

가명처리 데이터를 활용하는 경우, 연구자는 반드시 개인정보보호법을 준수해야한다. 따라서 관광 및 문화콘텐츠 분야 시사점 도출에 있어 유용한 정보로 판단된다 하더라도 개인정보를 침해한다고 판단되는 경우 이를 활용하지 않는 것을 원칙으로 하였다. 또한 단일 컬럼은 개인정보를 침해하지 않지만 여러 가지 컬럼 정보를 동시에 활용하는 경우 개인을 식별할 수 있는 가능성이 높아지므로, 최종 컬럼 구성 후 개인정보를 침해하지 않는지 여부를 종합적으로 검토하였다.

가. 통신데이터 컬럼

관광분야 실증분석을 위한 주요 컬럼으로 휴일 이동횟수를 선정하였다. 관광이 개인 또는 가족단위의 이동을 기반으로 이루어진다는 점에서 이동횟수를 통해 관광활동 현황을 유추해 볼 수 있다. 또한 출퇴근, 등교와 같이 정기적인 이동은 관광으로 보기 어렵다. 따라서 비교적 휴일에 여가 및 여행이 활발하다는 점을 고려하여 휴일 이동횟수를 관광활동을 간접적으로 파악할 수 있는 지표로 선정하였으며 구체적 정의는 다음 〈표 4-4〉 와 같다. 통신사의 고객정보에는 이동관련 정보를 포함하고 있기 때문에 이에 대한 파악이 가능하며 이는 관광의 현황 및 변화를 파악하는데 유의미하게 활용될 수 있을 것으로 판단된다. 또한 2019년 8월 부터 가장 최근의 데이터가 누적되어 있어 충분한 시계열 자료의 확보도 가능하다, 단, 세부적인 이동 동선 정보는 개인정보를 침해할 가능성이 매우 높기 때문에 최종적으로 데이터 구성 시 월별 총 이동횟수로 컬럼을 가공하였다.

〈표 4-4〉 통신 데이터 이동횟수 정의

항목	세부내용
이동횟수	**한달 휴일의 총 이동 횟수(단위: 횟수)** 1) 유의미한 체류 이력: 30분 이상 체류한 위치 파악 2) 유의미한 체류 이력의 의 개수 - 1 = **이동횟수** : (예) 하루동안, [집] ->[회사] ->[저녁약소장소] ->[집] = 이동회수 3

문화콘텐츠 분야 실증분석을 위한 주요 컬럼으로는 평일/휴일 온라인콘텐츠 이용량을 선정하였으며 정의는 다음 〈표 4-5〉와 같다. 해당 정보는 개인의 휴대폰에 설치된 온라인 콘텐츠 관련 앱 또는 웹의 사용 정보를 기반으로 해당 정보를 파악할 수 있다. 또한

고객의 구독서비스 관련 상품 결제 이력이나 관련 앱/웹 이용현황을 바탕으로 추정된 구독서비스 이용 관련 컬럼도 추가하였다. 선정된 컬럼들은 2019년 8월 부터 정보가 누적되어 있기 때문에 시계열 분석이 가능하다. 이 또한 최종 데이터를 구성할 때 월별 이용량으로 가공하여 개인정보를 침해할 가능성을 줄일 수 있도록 하였다.

〈표 4-5〉 통신 데이터 온라인 콘텐츠 이용량 정의

항목	세부내용
온라인 콘텐츠 이용량	**한달 간 평일 또는 휴일의 총 이용량(단위: Packet)** 영상, 음악, 게임, 도서 및 잡지, 웹툰, 커뮤니티, SNS,블로그 관련 사용량

그 외 관광 및 온라인 콘텐츠 활동에 영향을 줄 수 있는 다양한 개인적 특성을 포함하여 다음 〈표 4-6〉와 같이 통신 데이터 최종 컬럼을 구성하였다.[37]

〈표 4-6〉 [가명처리 데이터] 통신 데이터 최종 컬럼 구성

컬럼	기간	세부내용
ⅰ. 개인 특성		
성별	'19.1~'21.6	시군구 코드
연령대	'19.1~'21.6	1인 가구, 2인 가구, 3인 가구, 4인 가구, 5인 가구 이상
세대구분 1	'19.1~'21.6	M세대(1981~96년생), Z세대(1997~2010년생), 그 외 세대
세대구분 2	'19.1~'21.6	청년(18~34세), 중년(35~49세), 장년(50~64세), 노년(65세 이상)
가구원수	'20.1~'21.1	1인 가구, 2인 가구, 3인 가구, 4인 가구, 5인 가구 이상
거주지역	'19.1~'21.6	시군구 코드
연간소득 ** 8개 소득 구간*	'19.1~'21.6	1억이상, 7천-1억미만, 5천-7천미만, 4천-5천미만 3천-4천미만, 2천-3천미만, 1천-2천미만, 1천 미만
ⅱ. 관광 부문: 이동횟수		
평일 이동 횟수	'19.8~'21.6	평일 중에서 법정공휴일을 제외한 날 한 달 간 평일 총 이동 횟수
휴일 이동 횟수	'19.8~'21.6	2개 요일(토일) 또는 법정공휴일 한 달간 휴일 총 이동횟수
ⅲ. 문화 부문: 온라인 콘텐츠 이용		
평일 온라인콘텐츠이용량	'19.8~'21.6	전체, 동영상, 음악, 게임, 도서, 웹툰, 커뮤니티 등
휴일 온라인콘텐츠이용량	'19.8~'21.6	전체, 동영상, 음악, 게임, 도서, 웹툰, 커뮤니티 등
ⅲ. 문화 관련 온라인 콘텐츠 관련:		
구독서비스 정보	'19.8~'21.6	구독서비스 정보
구독기간 이용 기간	'19.8~'21.6	개월

37) 세부적인 각 컬럼 설명 및 가명처리 내용은 [부록 2] 가명처리 데이터 설명서 205p 참고

나. 카드데이터 컬럼

관광분야 실증분석을 위한 주요 컬럼으로 관광 관련 카드 지출액으로 선정하였으며 관광분야에 대한 정의는 다음 〈표 4-7〉 과 같다.[38]

〈표 4-7〉 카드데이터 관광 분야 업종 분류 정의

<table>
<tr><th colspan="2">부 문</th><th>세부 분류(신한분류)</th><th>부 문</th><th>세부 분류(신한분류)</th></tr>
<tr><td colspan="2">01. 여행업</td><td>01. 관광여행사</td><td rowspan="3">13. 음식점업</td><td rowspan="3">17. 한식, 18. 양식, 19. 일식
20. 중식, 21. 뷔페
22. 패스트푸드,
23. 제과점,
24. 커피전문점</td></tr>
<tr><td rowspan="2">관광숙박</td><td>02.관광호텔(4,5성급)</td><td>02. 특급호텔</td></tr>
<tr><td>03.관광호텔(3성급)</td><td>03. 1급호텔, 04. 2급호텔</td></tr>
<tr><td colspan="2">04.콘도미니엄</td><td>05. 콘도미니엄</td><td rowspan="9">14. 레저
스포츠 체험업</td><td rowspan="9">25. 운동경기,레져용품
26. 레저스포츠
27. 스포츠센타/레포츠클럽
28. 운동경기관람
29. 실외골프장, 30. 실내골프장
31. 테니스장, 32. 볼링장
33. 스키장, 34. 수영장,
35. 당구장36. 자전거(성인용),
37. 보트판매
38. 수중장비.
39. 공연장,극장
40. 이용,미용,
41. 피부미용실
42. 온천장,
43. 수련원,체험장</td></tr>
<tr><td colspan="2">05.일반숙박업</td><td>06. 모텔, 여관, 기타숙박</td></tr>
<tr><td colspan="2">06.카지노</td><td>44. 카지노</td></tr>
<tr><td colspan="2">07.유원시설업</td><td>07. 종합레저타운/놀이동산
08. 수족관 09. 동물농장</td></tr>
<tr><td colspan="2">08.면세점</td><td>10. 면세점</td></tr>
<tr><td colspan="2">09.항공사</td><td>11. 항공사</td></tr>
<tr><td colspan="2">10.대중교통</td><td>12. 철도, 13. 고속버스,
14. 여객선</td></tr>
<tr><td colspan="2">11.렌터카</td><td>15. 렌터카</td></tr>
<tr><td colspan="2">12.관광기념품판매업</td><td>16. 관광민예, 설문용품</td></tr>
</table>

콘텐츠 분야 지출액 또한 관광 분야와 마찬가지로 연구원에서 신한카드 자료를 이용하여 매월 생산중인 「콘텐츠소비지출 동향 분석 자료」를 참고하여 정의하였으며 다음 〈표 4-8〉과 같이 분류된다.

〈표 4-8〉 카드데이터 오프라인 콘텐츠 분야 업종 분류 정의

부 문	세부 분류
01. 핵심	01. 인쇄,출판 02. 공연장,극장 03. 음반테이프 04. 서적 05. 전자오락실 06. PC게임방 07. 노래방 08. 수련원,체험장 09. 실내골프장 10. 비디오방/전화방
02. 간접	11. 종합레저타운/놀이동산 12. 레저스포츠 13. 문구용품 14. 인형및완구아동용자전거 15. 정보통신기기,컴퓨터 16. 컴퓨터소프트웨어

38) 해당 분류는 현재 연구원에서 신한카드 데이터를 활용하여 매월 생산하고 있는 「COVID19 문화관광콘텐츠영향」과 「관광레저소비지출경제동향」의 관광업종 분류를 적용하였다.

온라인 콘텐츠지출액은 영상콘텐츠, 음악,게임, 출판 4개 영역으로 구분하여 각각의 지출액 정보를 활용하였다. 이때 각 분야별 사업자를 다음 〈표 4-9〉과 같이 사전에 정의하여 해당 사업자에 발생한 지출액을 온라인 콘텐츠 지출액으로 정의하였다.

〈표 4-9〉 카드데이터 온라인 콘텐츠 분야 해당 사업자 분류

부 문	세부 분류
01. 영상	영상 관련 온라인 콘텐츠 분야 지출액
02. 음악	음악 관련 온라인 콘텐츠 분야 지출액
03. 게임	게임 관련 온라인 콘텐츠 분야 지출액
04. 출판	출판 관련 온라인 콘텐츠 분야 지출액

이에 최종적으로 구성된 신한카드 데이터는 다음과 〈표 4-10〉과 같다. 개인특성은 통신 데이터와 동일하게 구성하였다.[39]

〈표 4-10〉 [가명처리 데이터] 카드 데이터 최종 컬럼 구성

컬럼	기간	세부내용
ⅰ. 개인 특성		
성별	'19.1~'21.6	시군구 코드
연령대	'19.1~'21.6	1인 가구, 2인 가구, 3인 가구, 4인 가구, 5인 가구 이상
세대구분 1	'19.1~'21.6	M세대(1981~96년생), Z세대(1997~2010년생), 그 외 세대
세대구분 2	'19.1~'21.6	청년(18~34세), 중년(35~49세), 장년(50~64세), 노년(65세 이상)
가구원수	'20.1~'21.1	1인 가구, 2인 가구, 3인 가구, 4인 가구, 5인 가구 이상
거주지역	'19.1~'21.6	시군구 코드
연간소득 ** 8개 소득 구간*	'19.1~'21.6	1억이상, 7천-1억미만, 5천-7천미만, 4천-5천미만 3천-4천미만, 2천-3천미만, 1천-2천미만, 1천 미만
ⅱ. 오프라인 카드 지출액		
전체 월별	'19.1~'21.6	전체 발생한 신용카드 지출의 월별 총 금액(단위: 원)
관광 월별	'19.1~'21.6	관광 분야의 신용카드 지출의 월별 총 금액(단위: 원)
콘텐츠 월별	'19.1~'21.6	오프라인콘텐츠 분야의 신용카드 지출의 월별 총 금액(단위: 원)
ⅲ. 온라인 콘텐츠 지출액		
영상 월별 카드 지출	'19.1~'21.6	영상관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)
음악 월별 카드 지출	'19.1~'21.6	음악관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)
게임 월별 카드 지출	'19.1~'21.6	게임관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)
출판 월별 카드 지출	'19.1~'21.6	출판관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)

39) 세부적인 각 컬럼 설명 및 가명처리 내용은 [부록 2] 가명처리 데이터 설명서 참고

2. 데이터 구성

가. 통신데이터 구성

SK텔레콤으로 부터 2019년 8월 ~ 2021년 6월까지 SK텔레콤 가입자를 약 2천 400만명을 대상으로 하며, 개인 정보와 이동횟수, 구독서비스 및 온라인콘텐츠 이용과 관련된 컬럼을 월별로 정보를 제공받아 총 616개 컬럼을 구성하였다. 분석대상 기간은 2019년 8월 ~ 2021년 6월로 설정하였다. 분석대상 기간은 활용하고자 하는 컬럼을 수집할 수 있는 최대한의 기간으로 설정하였다. SK텔레콤의 경우 개인정보의 내부 관리 계획에 의하여 일정 기간이 지나면 개인의 정보를 삭제하고 있다. 또한 최근에 개인별 특성 변수와 관련하여 성별 연령별 외 다양한 특성을 생성하고 있어 아직까지 누적된 정보가 부족한 실정이다. 따라서 이를 고려하여 현실적인 분석 기간을 설정하였다.

〈표 4-11〉 SK텔레콤 가명처리 데이터 정보

부 문	세부 분류
총 표본수	24,240,343명 *분석대상 기간 동안 SK텔레콤 가입자
활용 컬럼 수	616개, *개인 정보, 이동횟수, 구독서비스 기간 등을 월별로 구성
분석대상 기간	2019년 8월 ~ 2021년 6월 *활용 컬럼 중 가장 오래된 정보와 최신의 정보를 반출 여부를 설정

SK텔레콤 데이터에서 개인 정보를 가명처리하기 위해 연령이나 가구원수, 소득과 같은 연속된 데이터 값을 가지는 컬럼의 경우 범주화하였다. 여기서 이동량의 경우 관광활동을 판단하는 주요 지표로 판단하여 범주화하여 정보를 축소하지 않고 상단과 하단의 이상치를 재가공하여 식별의 가능성을 낮추도록 하였다.[40] 또한 온라인콘텐츠 사용량은 개인별 편차가 매우 커서 식별의 가능성이 높다고 판단하여, 평균치를 중심으로 표준화하였다. 즉, 평균치 사용자는 0의 값을, 평균보다 적게 사용한 경우 음(-)의 값, 많이 사용한 경우 양(+)의 값을 가지도록 지수화[41] 하여 식별의 가능성을 최소화 하였다,

40) 상위 1% 이하 5단위 구간화, 상위 1% 초과 상단 코딩 $f(x) = int(x/5) \times 5$

41) 지수화(z-score)를 적용하여, 소수점 이하 2자리 반올림

$$f(x) = round((x - avg(x))/stddev(x), 2)$$

다음 〈표 4-12〉는 SK텔레콤 통신 데이터의 가명처리 전후를 비교한 것이다. 기존의 개인정보가 식별가능한 시계열 데이터를 가명처리하여 최종적으로 패널 데이터로 구조를 변환하여 분석에 활용하였다. 단, 분석은 총 2천 4백만명 전수를 사용하지 않고 향후 가명 결합 건을 고려하여 통신 및 카드 동시가입자 만을 대상으로 하였다.

〈표 4-12〉 SK텔레콤 통신 데이터 가명처리 예

(가명처리 전) 개인정보를 포함한 SK텔레콤 시계열 데이터(예시)

(대상: SK텔레콤 가입자)

ID	20년 6월 연령	20년 7월 연령	…	21년 5월 연령	…	20년 6월 이동량	20년 7월 이동	…	21년 5월 이동량
김관광	29세	29세	…	30세	…	10회	24회	…	36회
이문화	41세	42세	…	42세	…	140회	150회	…	170회
박정책	75세	75세	…	76세	…	16회	10회	…	19회

(가명처리 후) 개인정보를 가명처리한 후 SK텔레콤 패널데이터 구조화(예시)

(대상: SK텔레콤 가입자)

ID	Time	성별	연령대	이동량	구독서비스 이용우뮤
A	19년 7월	여성	20대 이하	10회	이용함
A	19년 8월	여성	20대 이하	24회	이용함
……	……	……	……	……	……
A	21년 5월	여성	30대	36회	이용함
B	19년 7월	남성	40대	24회	이용함
B	19년 8월	남성	40대	19회	이용안함
……	……	……	……	……	……
B	21년 5월	남성	40대	15회	이용안함
C	19년 8월	여성	70대 이상	16회	이용안함
……	……	……	……	……	……
C	21년 5월	여성	70대 이상	10회	이용함
C	21년 6월	여성	70대 이상	19회	이용안함

나. 카드데이터 구성

신한카드로 부터 2019년 8월 ~ 2021년 6월까지 신한카드 가입자를 약 850만명을 대상, 개인 정보와 관광 및 온라인콘텐츠 소비 정보를 포함한 컬럼을 월별로 정보를 제공받아 총 633개 컬럼을 구성하였다. 분석대상 기간은 2019년 8월 ~ 2021년 6월로 설정하였다. 통신 데이터와 마찬가지로 분석대상 기간은 활용하고자 하는 컬럼을 수집할 수 있는 최대한의 기간으로 설정하였다.

〈표 4-13〉 신한카드 가명처리 데이터 정보

부 문	세부 분류
총 표본수	8,453,672명 *분석대상 기간 동안 신한카드 가입자 중. 단, 지출액 이상치는 식별가능성이 높아 제외
활용 컬럼 수	633개 *개인 정보, 전체 및 관광, 온라인콘텐츠 등의 소비지출액 월별 컬럼 생성
분석대상 기간	2019년 8월 ~ 2021년 6월 *활용 컬럼 중 가장 오래된 정보와 최신의 정보를 반출 여부를 설정

신한카드 지출액 데이터 또한 연속된 데이터 값을 가지는 컬럼의 경우 범주화하여 가명처리하였다. 여기서 지출액의 경우 카드데이터 활용 시 주요 변수이기 때문에 범주화하여 정보를 축소하지 않고 천원 단위에서 반올림하는 것으로 이상치를 가진 개인은 제거하여 식별의 가능성을 낮추도록 하였다.

다음 〈표 4-14〉는 신한카드 지출액 데이터의 가명처리 전후를 비교한 것으로 개인정보를 가명처리한 후 패널 데이터 구조로 변환한 예시이다. 여기서 이문화의 관광지출액은 약 1천만원이 넘기 때문에 이상치로써 개인 식별의 가능성이 매우 높아진다. 따라서 본 가명처리 데이터에는 이러한 개인은 제거하는 것으로 가명처리하였다.

카드 가명처리 데이터 분석 또한 총 8백 50만명 전수를 사용하지 않고 향후 가명 결합 건을 고려하여 통신 및 카드 동시가입자 만을 대상으로 하였다.

〈표 4-14〉 신한카드 가명처리 데이터(예시)

(가명처리 전) 개인정보를 포함한 신한카드 시계열 데이터(예시)

(대상: 신한카드 가입자)

ID	20년 6월 연령	20년 7월 연령	…	21년 5월 연령	…	20년 6월 관광지출	20년 7월 관광지출	…	21년 5월 이동량
김관광	29세	29세	…	30세	..	161,000	144,000	..	156,000
이문화	41세	42세	…	42세	..	13,456,000	13,586,000	..	
박정책	75세	75세	…	76세	..	56,000	78,000	..	34,000

(가명처리 후) 개인정보를 가명처리한 후 신한카드 패널데이터 구조화(예시)

(대상: 신한카드 가입자)

ID	Time	성별	연령대	관광 소비지출	온라인 콘텐츠
A	19년 7월	여성	20대 이하	161,000	22,000
A	19년 8월	여성	20대 이하	144,000	22,000
……	……	……	……	……	……
A	21년 5월	여성	30대	156,000	33,000
C	19년 8월	여성	70대 이상	56,000	0
……	……	……	……	……	……
C	21년 5월	여성	70대 이상	78,000	6,000
C	21년 6월	여성	70대 이상	34,000	6,000

제3절 실증분석 주제 선정

다음은 SK텔레콤과 신한카드 지출액 '가명처리 데이터'를 활용하여, 가명처리 데이터가 가지는 강점을 보여줄 수 있는 관광과 문화콘텐츠 분야의 분석 가능 한 주제를 검토하고, 이 중 몇 가지 선정하여 실제 실증분석을 진행하고자 한다. 이를 통해 이론적으로 살펴보았던 가명처리 데이터가 가지는 강점을 실증적으로 확인해 봄으로써 관광 및 문화콘텐츠 분야의 실제 활용 가치를 진단하고자 한다.

강점 1: 개인단위 원시자료를 활용한 활용 범위 확대

개인정보를 가명처리하는 경우 기존의 익명의 집계데이터로를 개인단위의 원시자료 형태로 활용할 수 있게 됨을 확인하였다. 이 경우 총량으로만 생산되던 통계에 개인의 특성을 반영한 추가적인 통계를 생산 할 수 있다는 점에서 기존 보다 데이터 활용 가치가 높다는 것을 알 수 있다. 이에 문화관광 분야의 다음과 같은 주제의 실증분석이 가능할 것으로 판단된다.

개인 단위 원시자료를 활용한 분석 가능 주제(안)

- 1인 평균 관광 이동행태 분석
- 1인 평균 관광 관련 지출 추세 분석
- 1인 평균 온라인 콘텐츠 이용률 분석
- 1인 평균 온라인 콘텐츠 이용량 분석
- 1인 평균 오프라인 콘텐츠 소비 지출 분석
- 1인 평균 온라인 콘텐츠 소비 지출 분석
- 1인 평균 구독서비스 이용률 분석
- 1인 평균 구독서비스 이용횟수
- 1인 평균 구독서비스 이용기간 분석

장점 2: 유효표본 수 확대에 따른 분석

개인정보를 가명처리한 데이터를 활용하는 경우 얻을 수 있는 또 하나의 장점은 민간데이터를 개인단위 원시데이터로 확보할 수 있게 됨에 따라 유효표본수가 확대되어 개인의 다양한 특성을 고려한 분석이 '세부특성별 분석'과 '소규모 지역 통계생산' 가능해진다는 점이다. 이에 다음과 같은 주제를 분석할 수 있을 것이다.

유효표본 확대에 따른 분석 가능 주제(안)

- 1인 가구의 세부특성별 관광 이동량 분석
- 1인 가구의 세부특성별 관광 소비지출 분석
- 1인 가구의 세부특성별 오프라인 콘텐츠 이용률 분석
- 1인 가구의 세부특성별 온라인 콘텐츠 이용률 분석
- MZ세대의 세부특성별 관광 이동량 분석
- MZ세대의 세부특성별 관광 소비지출 분석
- MZ세대의 세부특성별 오프라인 콘텐츠 이용률 분석
- MZ세대의 세부특성별 온라인 콘텐츠 이용률 분석
- 시군구 단위의 세부특성별 온라인 콘텐츠 이용률 분석
- 시군구 단위의 세부특성별 관광 이동량 분석
- 시군구 단위의 세부특성별 관광 소비지출 분석
- 시군구 단위의 세부특성별 오프라인 콘텐츠 이용행태 분석
- 시군구 단위의 세부특성별 온라인 콘텐츠 이용행태 분석
- 온라인 콘텐츠 핵심 사용자의 세부특성별 온라인 콘텐츠 이용행태 분석
- 온라인 콘텐츠 핵심 사용자의 세부특성별 관광 이동량 분석
- 온라인 콘텐츠 핵심 사용자의 세부특성별 관광 소비지출 분석
- 온라인 콘텐츠 핵심 사용자의 세부특성별 오프라인 콘텐츠 이용행태 분석
- 온라인 콘텐츠 핵심 사용자의 세부특성별 온라인 콘텐츠 이용행태 분석
- 구독서비스 핵심 사용자의 세부특성별 온라인 콘텐츠 이용행태 분석
- 구독서비스 핵심 사용자의 세부특성별 관광 이동량 분석
- 구독서비스 핵심 사용자의 세부특성별 관광 소비지출 분석
- 구독서비스 핵심 사용자의 세부특성별 오프라인 콘텐츠 이용행태 분석
- 구독서비스 핵심 사용자의 세부특성별 온라인 콘텐츠 이용행태 분석

강점 3: 패널데이터를 활용한 종단 및 패널 분석

개인정보보호법 개정 전, 익명화된 집계 데이터로만 활용되던 민간데이터를 개인 단위 데이터로 활용할 수 있게 되는 경우 많은 표본 수 확보와 더불어 시의성 높은 대규모 개인 패널 데이터를 구축할 수 있게 됨을 확인하였다. 이에 분석할 수 있는 다음의 종단 연구가 가능할 것으로 예상된다.

패널 데이터를 이용하여 가능한 종단 및 패널 분석 주제(안)

- 코로나 19 전후 관광 이동량 감소율 종단 분석
- 코로나 19 전후 관광 이동량 증가 및 감소 집단의 특성 비교 종단 분석
- 코로나 19 전후 구독서비스 신규 가입자의 특성 종단 분석
- 코로나 19 전후 개인 특성별 온라인 콘텐츠 지출액 변화 종단 분석
- 코로나 19 전후 개인 특성별 관광 소비 지출액 변화 종단 분석
- 코로나 19 전후 개인 특성별 온오프라인 콘텐츠 이용량 변화 종단 분석
- 코로나 19 전후 개인 특성별 관광 이동량 변화 종단 분석
- 구독서비스 구독기간에 미치는 영향 분석(예:Duration Analysis)
- 개인 및 가구 특성이 관광 소비지출에 미치는 영향 분석(예:RE, FE Model)
- 개인 및 가구 특성이 온라인 콘텐츠 이용량에 미치는 영향 분석(예:RE, FE Model)
- 개인 및 가구 특성에 따른 구독서비스 이용 결정 요인 분석(예:Panel Probi Model)
- 거리두기에 정책의 이동량 감소 효과 추정(예. DID)
- 재난지원금 지급 정책에 따른 관광 및 콘텐츠 소비 효과 추정(예. DID)

이러한 강점을 바탕으로 SK텔레콤과 신한카드 지출액 '가명처리 데이터'를 활용하여 분석할 주제는 다음 〈표 4-15〉과 같이 선정하였다.

〈표 4-15〉 가명처리데이터의 문화관광 부문 활용방안 도출을 위한 주제 선정

'가명처리 데이터'를 활용한 관광부문 활용방안 도출을 위한 실증분석 주제 선정

개인단위 원시자료 활용	유효표본 수 확대	패널 데이터 구축
▪ 1인 평균 관광 이동행태 분석 ▪ 1인 평균 관광 관련 지출 추세 분석	▪ 1인 가구의 세부특성별 관광 이동량 분석 ▪ 1인 가구의 세부특성별 관광 소비지출 분석	▪ 코로나 19 전후 관광 이동량 감소율 종단 분석 ▪ 코로나 19 전후 관광 이동량 증가 및 감소 집단의 특성 비교 종단 분석

'가명처리 데이터'를 활용한 문화콘텐츠 부문 활용방안 도출을 위한 실증분석 주제 선정

개인단위 원시자료 활용	유효표본 수 확대	패널 데이터 구축
▪ 1인 평균 온라인 콘텐츠 이용률 분석 ▪ 1인 평균 온라인 콘텐츠 소비지출 분석	▪ MZ세대의 세부특성별 오프라인 콘텐츠 이용률 분석 ▪ MZ세대의 세부특성별 온라인 콘텐츠 지출액 분석	▪ 코로나 19 전후 구독서비스 이용 변화 집단의 특성 종단 분석 ▪ 코로나 19 전후 개인 특성별 온라인 콘텐츠 지출액 변화 종단 분석

제4절 관광 부문 활용방안

기존 총량 위주의 통계를 생산하던 익명화된 집계 데이터와 달리 개인정보를 가명처리 하게 되면 개인단위 원시데이터를 활용할 수 있어 개인 단위 통계 생산이 가능하다 점을 확인한 바 있다. 이에 본 절에서는 SK텔레콤 가명처리 데이터와, 신한카드 가명처리 데이터를 이종을 각각 활용하여 가명처리 전후의 분석 결과를 비교하여 가명처리 데이터의 특징을 실증분석을 통해 살펴보고자 한다.

1. 개인단위 원시자료 활용을 통한 현황 분석

같은 시점에 반출한 SK텔레콤 '익명처리 집계데이터'와 '가명처리 데이터'를 이용하여 휴일 이동량 분석결과를 비교하여 개인정보를 가명처리한 개인단위 원시자료 활용을 통해 얻을 수 있는 이점을 실증적으로 살펴보고자 한다.

가. 1인 평균 관광 이동행태 분석

다음 〈표 4-16〉은 가명처리 데이터 전 후 휴일 이동량[42]을 분석한 비교한 결과이다. 휴일에 발생한 모든 이동이 관광목적으로 이루어졌다고 보기는 어렵지만, 평일보다는 비교적 여가 및 여행 목적의 이동발생이 높다는 점을 고려하여 휴일 이동을 기준으로 분석하였다.

분석 결과를 살펴보면 기존의 SK텔레콤 통신데이터는 익명화된 집계데이터는 각 특성별 개인의 표본 수를 파악할 수 없기 때문에 총량 단위의 통계를 생산하였다. 따라서 이를 기반으로 개인별 행태 보다는 해당 분야의 추세를 파악하는 수준에 머물러 있었다.

42) 휴일은 주말과 공휴일을 포함한다. 또한 이동량은 한달 간 총 이동횟수를 의미한다, 기지국 단위로 30분 이상 체류한 경우를 유의미한 체류로 판단하고, 유의미한 체류에서 1을 뺀 값이 이동횟수로 산출된다. 예를 들어 하루 동안 집->회사->외식->집 의 이동을 보였다면 이동횟수는 3회가 된다.

또한 가입자 수에 대한 정보를 추가적으로 제공받으면 전체의 1인당 평균 이동량 수준까지는 산출할 수 있으나 성별, 연령별 등 특성그룹별 가입자 수에 대한 정보는 제공하지 않기 때문에 각 특성별 평균값을 산출하기 어렵다. 하지만 개인정보를 가명처리 함에 따라 개인단위 원시자료를 활용할 수 있어 전체뿐만 아니라 세부 특성에 따른 1인당 휴일 총 이동량을 다음 〈표 4-16〉과 같이 산출할 수 있게 된다. 이를 통해 가명처리 전후 모두 추세 분석이 가능하며 가명처리 후 1인당 평균 이동량을 산출할 수 있어 추가적인 정보를 파악할 수 있다는 이점이 있다.

또한 1인당 평균 휴일 이동량을 산출함으로 써 집단별 평균적인 차이를 비교 분석할 수 있게 된다. 예를 들어 성별에 따라 휴일 이동총량을 비교하는 경우, 남성과 여성의 규모를 알 수 없기 때문에 남성이 여성보다 높다고 해서 남성의 이동량이 많다고 해석하기에는 무리가 있다. 이는 절대적으로 남성의 비중이 높기 때문에 이동량 합계가 높을 수 있기 때문이다. 하지만 개인별 평균을 산출할 경우 여성이 남성보다 평균적으로 이동횟수가 더 높다는 해석이 가능해 진다고 하겠다. 이러한 점은 개인의 특성을 반영한 관광 참여 활성화 정책 수립 시 유의미한 기초자료로 활용될 수 있을 것으로 예상된다.

〈표 4-16〉 개인단위 원시 자료 활용을 통한 관광 이동량 분석 결과 비교

(익명집계데이터 활용 시) 한달 간 이동총량

(대상: SK텔레콤 가입자, 단위: 전체/백만 회)

구분		1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
전체	'19	-	-	-	-	-	-	-	263.3	263.3	351.0	272.6	333.4
	'20	344.6	238.3	150.4	282.3	384.5	216.6	252.2	268.6	114.2	346.8	243.5	206.6
	'21	203.0	255.4	236.0	218.8	419.6	275.9	-	-	-	-	-	-

자료: SK텔레콤 익명집계데이터

(가명처리데이터 활용 시) 한달 간 이동총량의 1인당 평균

(대상: SK텔레콤 가입자, 단위: 1인/회)

구분		1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
전체	'19	-	-	-	-	-	-	-	10.9	10.9	14.5	11.2	13.8
	'20	14.2	9.8	6.2	11.6	15.9	8.9	10.4	11.1	4.7	14.3	10.0	8.5
	'21	8.4	10.5	9.7	9.0	17.3	11.4	-	-	-	-	-	-
남성	'19	-	-	-	-	-	-	-	9.9	9.9	13.2	10.2	12.6
	'20	12.9	8.2	4.9	9.8	13.8	7.7	9.1	9.6	3.8	12.6	8.7	7
	'21	7.0	9.2	8.5	7.9	15.5	10.1	-	-	-	-	-	-
여성	'19	-	-	-	-	-	-	-	11.8	11.8	15.7	12.2	14.8
	'20	15.4	11.3	7.4	13.3	17.8	10.1	11.6	12.5	5.6	15.9	11.2	9.9
	'21	9.6	11.8	10.9	10.1	19.0	12.5	-	-	-	-	-	-

자료: SK텔레콤 가명처리데이터

나. 1인 평균 관광 관련 지출 추세 분석

다음으로 동일 시점에서 반출한 신한카드의 익명화된 집계 데이터와 가명처리 데이터를 이용하여 월별 관광부문[43] 소비지출액을 분석 결과를 비교하고자 한다. 기존에 본 연구원에서는 신한카드의 익명화된 집계 데이터를 이용하여 매월 관광 지출 동향을 산출하고 있다. 이는 관광 지출액 총액의 추세를 분석함으로써 코로나19와 같은 사회 변화가 해당 분야에 미치는 영향을 분석하는 정책적으로 유의미한 자료로 활용되고 있다. 하지만 카드 데이터도 통신 데이터와 마찬가지로 기존의 익명 데이터의 경우 유효표본 수나 각 세부 특성별 표본을 파악할 수 없기 때문에 1인당 평균 지출액을 산출하기 어렵다. 따라서 일부 통계에서는 카드 이용카드 수를 활용하여 대략적인 1인당 평균 소비지출액이 아닌 카드 1개당 평균적인 소비지출액을 산출하기도 하였다.

이에 가명처리 데이터를 통해서 개인단위의 원시데이터를 확보함에 따라 1인당 소비지출액을 분석할 수 있는 기반이 마련됨에 따라 1인당 평균적인 관광소비지출액을 파악할 수 있게 되었다. 다음 〈표 4-17〉는 익명화된 집계데이터와 개인정보 가명처리 데이터의 관광부문 지출액을 분석한 결과를 보여준다. 기존의 총금액 위주의 통계에서 1인당 평균 소비액을 분석할 수 있게 됨에 따라 추세 분석 외에도 특성별 비교 분석이 가능해지게 된다. 예를 들어 총량을 통해 성별 관광 소비지출 규모에 대해 절대적인 비교만 가능하였다면, 가명처리 데이터를 이용하여 1인당 평균을 산출함으로 해서 여성이 남성보다 비교적 관광관련 소비 지출이 더 크다는 상대적 비교가 가능해짐을 확인할 수 있다.

〈표 4-17〉 개인단위 원시 자료 활용을 통한 관광 카드지출액 분석 결과 비교

(익명집계데이터 활용 시) 한달 간 총 관광소비지출액

(대상 전체*, 단위: 억원)

구분		1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
전체	'19	44,094	40,339	45,039	46,422	48,386	45,721	48,860	47,984	43,657	48,986	44,499	47,536
	'20	42,274	28,470	25,224	29,367	36,846	33,729	36,207	33,646	28,076	36,868	32,599	22,785
	'21	22,795	27,058	33,650	34,659	37,908	36,650	34,763	33,896	35,596	-	-	-

자료: 신한카드 익명집계데이터

참고: 신한카드 익명화된 집계데이터는 신한카드 시장 점유율을 활용하여 전체 시장 취급액으로 추정

43) 현재 연구원에서 신한카드 데이터를 활용하여 매월 생산하고 있는 「COVID19 문화관광콘텐츠영향」과 「관광레저소비지출경제동향」의 관광업종 14개 대분류 43개 소분류를 적용하였다.

(가명처리데이터 활용 시) 한달 간 총 관광소비지출액의 1인당 평균

(대상: 신한카드 가입자, 단위: 1인/만원)

구분		1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
전체	'19	-	-	-	-	-	-	-	16.5	14.4	15.9	15.3	13.8
	'20	16.2	14.5	11.3	10.3	11.7	14.6	13.2	14.0	13.3	11.2	13.8	12.5
	'21	9.0	10.1	12.4	12.8	13.8	13.3	-	-	-	-	-	-
남성	'19	-	-	-	-	-	-	-	13.9	12.1	13.4	12.7	13.5
	'20	12.2	9.1	7.9	9.2	10.9	10.1	11.2	10.5	8.7	10.9	9.9	7.0
	'21	7.1	8.1	9.9	10.2	11.0	10.7	-	-	-	-	-	-
여성	'19	-	-	-	-	-	-	-	18.9	16.6	18.3	17.7	18.6
	'20	16.7	13.4	12.6	14.1	18.0	16.0	16.6	15.9	13.6	16.4	14.9	11.0
	'21	9.6	11.8	10.9	10.1	19.0	12.5	-	-	-	-	-	-

자료: 신한카드 가명처리데이터

2. 패널데이터를 활용한 현황 변화 분석

가명처리 데이터가 가지는 핵심적 강점은 시의성 높은 대규모 패널데이터를 구축할 수 있다는 점이다. 민간사업체는 고객으로 등록되어있는 기간 내 정보를 시계열로 수집·보관하고 있기 때문에 민간데이터를 개인 단위 데이터로 활용하게 되면 시의성 높은 대규모 개인 패널 데이터를 구축할 수 있게 된다.

가. 코로나 19 전후 관광 이동량 감소율 종단 분석

다음 〈표 4-18〉는 SK텔레콤 익명데이터와 가명처리데이터를 이용하여 이동량의 변화에 대하여 비교 분석한 결과이다. 이를 통해 가명처리 데이터를 활용하는 경우 단순 추세 분석에서 확장하여 종단적 분석이 가능함을 확인할 수 있다. 본 분석에서는 2019년 11월 ~ 2020년 1월(3개월) 코로나19 발생 전으로 정의하고 2020년 2월 ~2021년 4월까지(3개월) 코로나19 발생 후로 정의하였다. 이에 기존의 익명화된 집계 데이터를 활용할 경우 코로나19 전후 이동량에 대한 추세를 파악하는 수준에 머물렀다면, 가명처리데이터를 활용할 수 있게 됨에 따라 코로나19 전후 각 집단별 이동량이 감소한 사람의 비중이 얼마인지 보다 더 세부적인 정보를 파악할 수 있게 된다.

〈표 4-18〉 패널화된 가명처리데이터를 활용한 관광 카드지출액 분석 결과 비교

(익명집계데이터 활용 시) 코로나 전후 한달 간 이동총량 변화 비교 분석

(대상: SK텔레콤 가입자, 단위: 백만 회, %)

구분	한달 간 이동 총량		증감률
	2019년 11월 ~ 2020년 1월	2020년 2월 ~ 2021년 4월	
전체	316.9	223.7	-29.4
남성	138.3	89.1	-35.6
여성	178.6	134.6	-24.7
20대 이하	7.6	3.8	-50.5
30대	61.5	9.9	-35.2
40대	79.8	54.0	-32.3
50대	87.6	62.5	-28.7
60대	58.6	46.6	-20.5
70대 이상	21.9	17.0	-22.2

자료: SK텔레콤 익명집계데이터

(가명처리데이터 활용 시) 코로나 전 후 집단별 이동량 감소율 비교 분석

(대상: SK텔레콤 가입자, 단위: %)

구분	(비교1) 2019년 11월 ~ 2020년 1월 vs 2020년 2월 ~ 2020년 4월 까지		(비교2) 2019년 11월 ~ 2020년 1월 vs 2020년 11월 ~ 2021년 1월 까지	
	이동량 감소	이동량 유지 / 증가	이동량 감소	이동량 유지 / 증가
전체	68.8	31.2	70.1	29.9
남성	71.3	28.7	71.3	28.7
여성	66.5	33.5	68.9	31.1
20대 이하	68.5	31.5	65.5	34.5
30대	67.0	33.0	66.0	34.0
40대	71.6	28.4	71.9	28.1
50대	70.6	29.4	72.1	27.9
60대	66.7	33.3	70.1	29.9
70대 이상	64.9	35.1	68.7	31.3

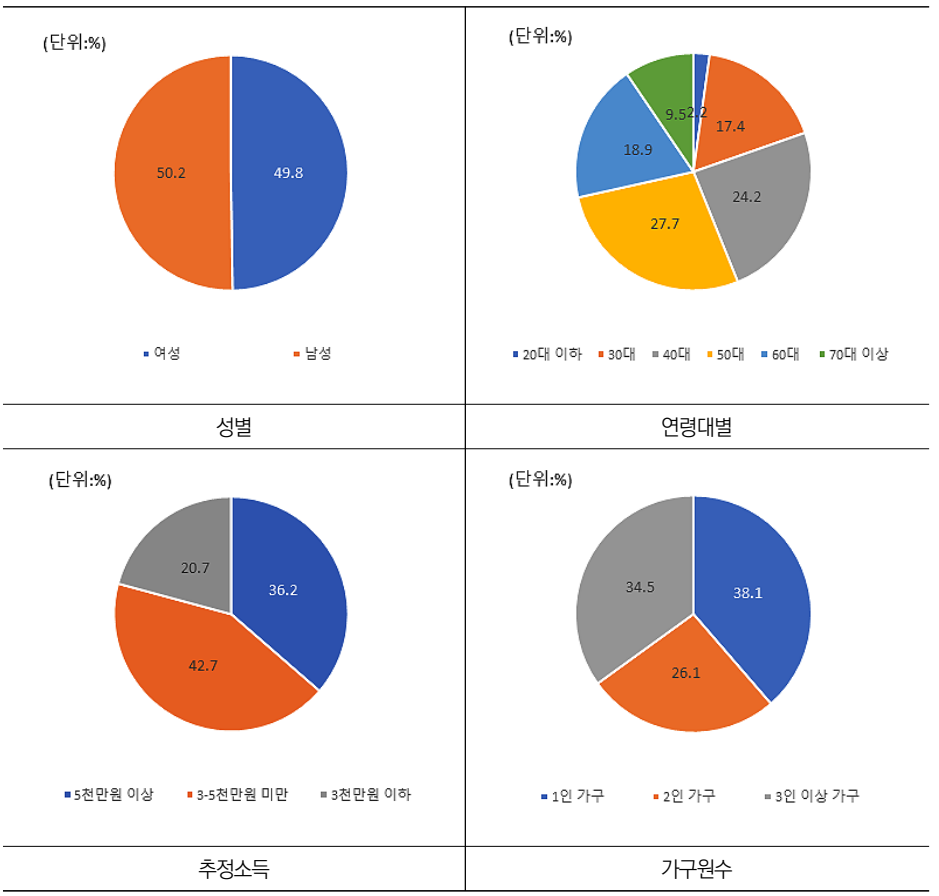
자료: SK텔레콤 가명처리데이터

나. 코로나 19 전후 관광 이동량 증가 및 감소 집단의 특성 비교 종단 분석

뿐만 아니라 가명처리 데이터를 활용하게 되면 각 개인별로 코로나 19 이후 이동량이 감소했는지 유지 혹은 증가했는지 파악할 수 있다. 즉, 개인의 변화에 대해 추적관찰이 가능해서 개인의 변화를 기반으로 다양한 분석을 진행할 수 있다. 다음 [그림 4-2]는 이동량이 감소한 집단과 증가한 집단의 인구통계학적 분포를 산출한 결과이다. 이를 통해, 코로나 19 이후 이동량이 감소한 집단은 남성, 40~50대, 연평균 추정소득 3~5천만원. 1인 가구의 비중이 높은 것을 확인할 수 있다.

[그림 4-2] 패널화된 가명처리데이터를 활용한 코로나19 이후 이동량 감소 집단의 특성 분포

(대상: SK텔레콤 가입자, 단위 %)



자료: SK텔레콤 가명처리데이터

3. 유효표본 수 확대에 따른 세부특성별 현황 분석

개인정보를 가명처리한 데이터를 활용하는 경우 얻을 수 있는 또 하나의 장점은 민간 데이터를 개인단위 원시데이터로 확보할 수 있게 됨에 따라 유효표본수가 확대되어 개인의 다양한 특성을 고려한 분석이 '세부특성별 분석'이 가능해 진다는 점이다. 가명처리를 통해 현재보다 많은 유효표본을 확보할 가능성이 높다는 점과 현재 대부분의 개인단위 데이터가 조사를 기반으로 생산된다는 사실에 기반하여 통계청의 인구총조사와 문화체육관광부의 조사통계 중 표본 규모가 가능 큰 국민여행조사와 SK텔레콤 가명처리 데이터의 표본수를 비교해 보면 다음 〈표 4-19〉과 같다. 즉, 민간 데이터를 가명처리할 경우 기존의 조사 통계보다 확보할 수 있는 표본 수가 매우 크기 때문에 개인의 다양한 특성을 함께 고려할 수 있어 맞춤형 정책 수립과 관련된 자료로 활용성이 높아질 것으로 기대된다.

〈표 4-19〉 가명처리 후 유효표본 수 확대에 따른 확보 가능한 개인 특성별 유효표본 수 비교

(단위: 천명)

2020 인구총조사		SK 가명처리 데이터		2020 국민여행조사	
전체 표본 수	50,710	전체 표본 수	24,240	전체 표본 수	51
남성	25,251	남성	****	남성	25
여성	25,459	여성	****	여성	25
청년층	12,782	청년층	****	청년층	13
중년층	14,246	중년층	****	중년층	14
장년층	13,363	장년층	****	장년층	13
노년층	8,641	노년층	****	노년층	9

가. 1인 가구의 세부특성별 관광 이동량 분석

다음 〈표 4-20〉은 실제 1인 가구를 대상으로 국민여행조사와 통신 가명처리 데이터를 활용하여 세부 특성별 분석 결과를 비교한 결과이다. 같은 1인 가구라 하더라도 청년층 1인 가구와 노년층 1인 가구는 전혀 다른 라이프 스타일을 지니고 있다고 보기 어렵기 때문에 1인 가구의 세부 특성을 고려한 다양한 연구 및 통계를 생산할 필요가 있다.

가명처리 데이터 활용 전 국민여행 조사 기준으로 분석하면 1인 가구를 대상 성별·세대별 분석까지 비교적 분석 가능한 표본이 확대되어 통계 생산이 가능하다. 하지만 1인 가구의

성별·세대별·소득별 분석을 진행할 경우 표본 수가 비교적 적어 대표성 있는 결과를 산출할 수 없다. 하지만 가명처리 된 민간데이터를 활용할 경우에 1인 가구를 대상 성별·세대별 분석에 소득특성를 추가하여 확장할 수 있어 보다 세부 특성을 반영한 분석이 가능하다는 점을 확인할 수 있다. 44)

〈표 4-20〉 가명처리 후 유효표본 수 확대에 따른 1인 가구 특성별 여행 행태 분석

(기존 조사데이터 활용 시) 1인 가구의 성별 X 세대별 월평균 총 여행일 수

(대상 전체*, 단위: 1인/일, %)

구분		18년	20년	증감률('18년 대비)
남성	청년층	1.2	1.0	-15.2
	중년층	1.2	0.9	-29.0
	장년층	1.1	0.6	-47.8
	노년층	0.6	0.3	-43.9
여성	청년층	1.1	1.1	3.9
	중년층	1.1	0.9	-14.6
	장년층	1.0	0.5	-50.6
	노년층	0.5	0.2	-56.0

참고: 표본설계를 통해 만15세 이상 전 국민을 대상으로 함
자료: 2018-2020 국민여행조사

(가명처리데이터 활용 시) 1인 가구의 성별 X 세대별 X소득별 월평균 총 이동횟수

(대상: SK텔레콤 가입자, 단위: 1인/회)

구분		2020년 상반기			2021년 상반기			증감률('20년 상반기 대비)		
		3천미만	3-5천	5천 이상	3천미만	3-5천	5천 이상	3천미만	3-5천	5천 이상
남성	청년층	10.4	12.4	13.9	9.4	11.4	12.8	-9.8	-8.5	-7.4
	중년층	14.2	14.9	15.8	13.0	13.6	14.6	-8.3	-8.4	-7.5
	장년층	13.4	14.1	14.4	12.2	12.9	13.5	-9.1	-8.3	-6.1
	노년층	10.8	14.1	15.0	10.1	13.2	14.1	-6.0	-6.9	-5.6
여성	청년층	8.0	9.3	10.1	7.1	8.5	9.5	-10.3	-8.1	-5.6
	중년층	11.3	11.8	12.7	10.3	10.9	11.7	-9.0	-8.1	-7.7
	장년층	10.5	10.7	11.1	9.8	10.2	10.7	-6.9	-5.0	-3.3
	노년층	10.2	11.7	12.0	10.1	11.6	11.9	-1.4	-1.4	-0.6

자료: SK텔레콤 가명처리데이터

44) 다만, 국민여행조사의 여행일 수, 통신데이터의 이동 총량은 정의가 다르기 때문에 절대적인 값을 비교해선 안된다. 또한 국민여행조사의 모집단은 만15세 이상의 국민인 반면 통신 데이터는 내부 가입자를 기준으로 하기 때문에 분석 대상도 상이하다. 따라서 각 자료의 산출 결과 값을 절대적으로 비교해서는 안돼며, 가명처리 데이터 활용 전후 데이터의 세부특성별 분석이 가능하다는 점을 확인 하는 방법론에 초점을 맞춰야 한다.

나. 1인 가구의 세부특성별 관광 소비지출 분석

비슷한 방식으로 신한카드 가명처리 데이터와 통계청의 인구주택 총조사, 문화체육관광부 국민여행조사의 표본수를 비교하면 다음 〈표 4-21〉과 같다. 신한카드사의 경우 SK텔레콤보다는 비교적 가입자가 적은 수준임에도 불구하고 대략 850만 고객의 정보를 가지고 있기 때문에 세부분석 진행에 무리가 없다.

〈표 4-21〉 가명처리 후 유효표본 수 확대에 따른 확보 가능한 개인 특성별 유효표본 수 비교

(단위: 명)

2020 인구총조사		신한 가명처리 데이터		2020 국민여행조사	
전체 표본 수	50,710	전체 표본 수	8,454	전체 표본 수	51
남성	25,251	남성	****	남성	25
여성	25,459	여성	****	여성	25
청년층	12,782	청년층	****	청년층	13
중년층	14,246	중년층	****	중년층	14
장년층	13,363	장년층	****	장년층	13
노년층	8,641	노년층	****	노년층	9

마찬가지로 신한카드의 개인정보 가명처리 데이터를 활용하면 국민여행조사 기준으로 분석하기 어려운 1인 가구의 성별·세대별·소득별 분석을 진행할 수 있다. 다음 〈표 4-22〉은 가명처리 데이터 활용전 국민여행조사로 산출한 1인당 평균 연간 여행지출액과 신한카드 가명처리 데이터를 활용한 1인당 평균 관광부문 연간지출액 분석 결과이다 이를 통해 가명처리 데이터 활용 시 기존의 개인 단위로 생산되는 조사통계 보다 세부적 특성별 분석의 확장이 가능하다는 것을 확인할 수 있다.[45]

45) 국민여행조사의 여행과 카드 지출액에서 정의하는 관광 부문의 정의가 일치하지 않기 때문에 절대적인 값을 비교해선 안된다. 또한 개인정보를 가명처리하는 문제와 표본의 대표성은 별개의 문제이다. 따라서 민간 데이터 내 개인정보를 가명처리해서 사용하더라도 여전히 민간데이터가 가지는 표본의 대표성 문제가 사라지지 않기 때문에 해석에 유의해야한다.

〈표 4-22〉 가명처리 후 유효표본 수 확대에 따른 1인 가구 특성별 지출액 분석

(기존 조사데이터 활용 시) 1인 가구의 성별 X 세대별 월평균 여행비용

(대상 전체*, 단위: 1인/만원, %)

구분		18년	20년	증감률('18년 대비)
남성	청년층	9.2	7.5	-19.3
	중년층	10.9	7.3	-33.1
	장년층	9.0	5.2	-42.2
	노년층	3.7	2.1	-42.0
여성	청년층	8.8	9.3	5.4
	중년층	10.8	8.6	-20.5
	장년층	7.5	3.8	-49.4
	노년층	3.3	1.3	-61.5

참고: 표본설계를 통해 만15세 이상 전 국민을 대상으로 함

자료: 2018-2020 국민여행조사

(가명처리데이터 활용 시) 1인 가구의 성별 X 세대별 X소득별 월평균 관광관련 카드지출액

(대상: 신한카드 가입자, 단위: 1인/만원)

구분		19년			20년			증감률		
		3천미만	3-5천	5천 이상	3천미만	3-5천	5천 이상	3천미만	3-5천	5천 이상
남성	청년층	11.5	16.8	18.3	10.8	15.3	17.1	-6.3	-9.1	-6.3
	중년층	12.1	14.9	18.1	11.0	13.7	16.9	-9.1	-8.0	-6.3
	장년층	11.3	13.8	20.2	11.0	13.0	18.9	-2.5	-6.2	-6.3
	노년층	7.8	11.3	21.5	7.2	10.7	20.1	-8.3	-5.5	-6.5
여성	청년층	9.2	12.7	14.6	8.9	12.0	14.2	-2.9	-5.3	-2.7
	중년층	8.2	10.3	14.3	7.5	9.6	13.6	-8.0	-6.2	-4.9
	장년층	7.2	9.4	15.0	6.6	8.8	14.5	-8.0	-6.4	-3.5
	노년층	6.1	9.0	15.9	5.8	8.4	15.2	-5.9	-6.4	-4.4

자료: 신한카드 가명처리데이터

제5절 문화콘텐츠 부문 활용방안

본 절에서는 제4절에서 관광 부문의 가명처리 데이터 활용방안을 살펴 살펴본 것과 동일한 방식으로 문화콘텐츠 부문에서 가명처리 데이터 활용 시 가질 수 있는 장점을 중심으로 데이터 분석결과를 살펴보고자 한다.

1. 개인단위 원시자료 활용을 통한 현황 분석

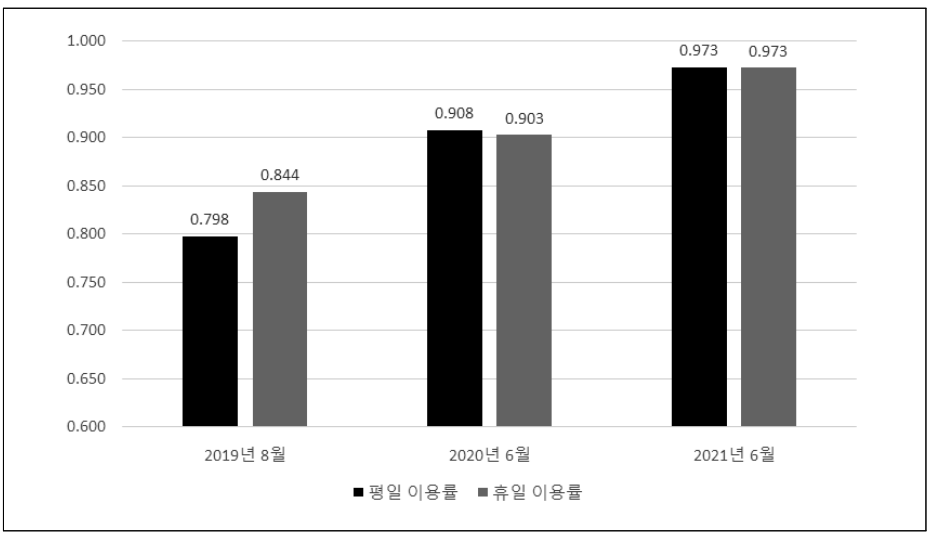
특정 기관이 보유하고 있는 개인정보를 외부기관에서 공공·연구목적으로 사용하고자 할 경우, 개인정보를 취급과 관련된 인적·물적 인프라가 구축되어 있지 않은 경우, 해당 인프라가 구축되어 있다고 하더라도 개인정보 보유기관에서 외부로 자료제공에 소극적인 경우 등에 해당되는데 이 경우 해당 개인정보를 직접 활용할 수 없기 때문에 개인정보를 익명처리한 뒤 제공받게 된다. 시계열(Time-series)을 가진 개인정보를 익명처리하게 되면 매 시점마다 특정인을 식별하기 어려워지기 때문에 데이터를 활용·분석하는 연구자의 경우에도 각 개인단위 분석이 불가능해 진다. 이러한 이유로 민간·공공분야에서 보유하고 있는 개인정보를 활용한 현황 분석에서는 총량 중심의 통계를 중심으로 자료를 작성하고 있다. 본 연구에서 활용하는 자료는 개인정보를 가명처리 한 월 단위 시계열 자료로 개인단위 원시자료에 해당한다. 가명처리 데이터는 개인정보를 가명처리 하였기 때문에 해당 데이터가 어떤 사람의 것인지는 확인할 수 없지만 가명처리 과정에서 개인 ID를 부여하기 때문에 서로 다른 시점에서 동일인을 구별해 낼 수 있다는 장점을 가진다.

가. 1인 평균 온라인 콘텐츠 이용률 분석

아래 그림은 연구를 수행하는 과정에서 이동통신사가 제공한 온라인 콘텐츠 이용률에 대한 익명화된 집계데이터이다. 익명처리 데이터는 앞서 언급한 것처럼 개인을 식별할 수 없기 때문에 아래 그림과 같이 가입자 전체의 온라인 콘텐츠 이용률 추이를 살펴보는 용도로 활용할 수 있다. 하지만 이동통신의 경우 가입자의 개인적인 사유에 의해 해지되거나, 신규 유입이 되는 등 가입자 정보가 매일 변화하게 된다. 따라서 아래 그림은 매 시점마다 변화하는 가입자 집단의 온라인 콘텐츠 이용률 추이로 보는 것이 정확하다.

[그림 4-3] **(익명화된 집계데이터 활용 시)** 온라인 콘텐츠 이용률 추이

(대상: SK텔레콤 가입자, 단위 %)



자료: SK텔레콤 익명집계데이터

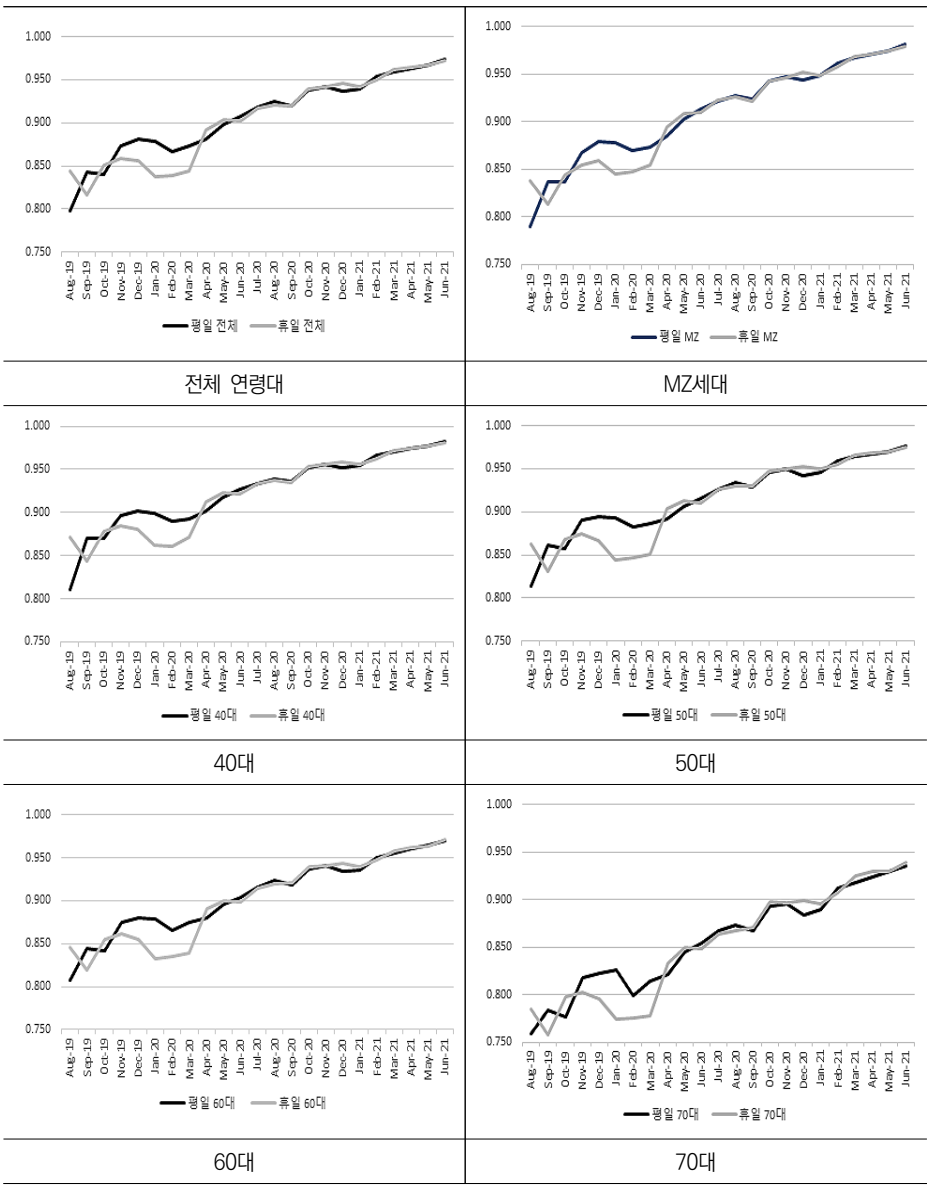
반면에 가명처리 데이터는 가명처리 과정에서 부여된 개인ID로 인해 익명처리 집계데이터 대비 개인의 인구통계학적 특성을 활용할 수 있다는 점, 1인 기준의 평균적인 특성을 도출할 수 있다는 점 등에서 장점을 가진다. 익명처리 집계데이터도 집계시점별 관측치수를 제공받는다면 1인 기준의 평균적인 수치는 뽑아낼 수 있으나 개인의 인구통계학적 특징 등을 활용할 수 없다.

다음 그림은 평일과 휴일의 연령대별 온라인 콘텐츠 이용률의 변화추이를 보여준다. 이용률은 가명처리 SK텔레콤 데이터를 연령에 따라 구분하였고, 각 연령대에 속한 사람

들 중 온라인 콘텐츠 이용을 하는 사람과 이용하지 않는 사람을 구분한 뒤 각 시점별로 비율을 산출하였다. 가명 처리된 데이터가 동일한 사람들로 23개월 간 구성되어 있기 때문에 이들의 시점별 온라인 콘텐츠 이용행태를 추적·분석해볼 수 있다.

[그림 4-4] **(가명처리 데이터 활용 시)** 평일, 휴일 연령대별 온라인 콘텐츠 이용현황 추이

(대상: SK텔레콤 가입자, 단위 %)

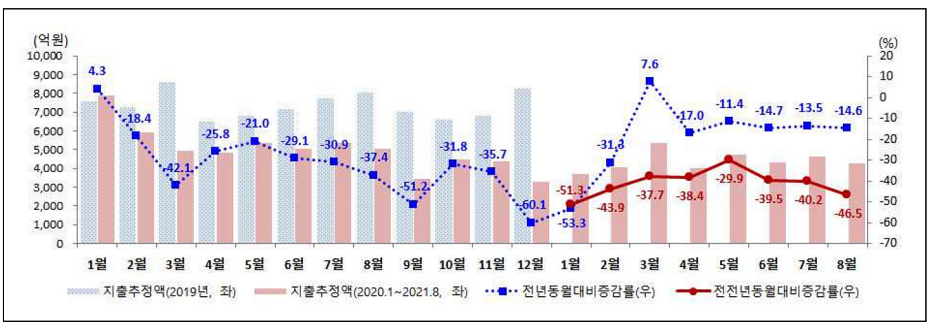


자료: SK텔레콤 가명처리데이터

나. 1인 평균 온라인 콘텐츠 소비 지출 분석

다음으로 신용카드사에서 제공받은 익명집계 데이터와 가명처리 데이터를 이용하여 월별 문화콘텐츠 부문의 온라인 콘텐츠 지출액의 분석 결과를 비교하고자 한다. 현재 연구원에서는 해당 카드사의 익명화된 집계 데이터를 제공받아 매월 '콘텐츠 소비지출 동향' 자료를 생산하고 있다. 해당 자료를 통해 코로나19와 같은 팬데믹 상황에서 업종별로 영향을 받고 있는지를 파악하고 있다는 점에서 정책을 기획하고, 실행하는데 근거 자료로 활용되고 있다. 이 경우에도 앞서 온라인 콘텐츠 이용률 분석에서 본 것과 같이 익명집계데이터이기 때문에 유효표본이 어느 정도 규모인지, 어떤 특성을 가진 사람들이 구성되어 있는지 여부를 파악할 수 없기 때문에 1인당 평균 지출액을 산출하기 어렵다.

[그림 4-5] (익명화된 집계데이터 활용 시) 콘텐츠 소비지출 동향



자료: 이용관(2021.10.06.), '콘텐츠 소비지출 동향 2021년 9월호.

다음에 연속한 표는 동일시점에 대해 신용카드사가 제공한 온라인 콘텐츠 지출액 총량의 익명처리 된 집계 데이터와 가명처리 데이터를 활용해 1인 기준으로 분류한 콘텐츠 지출액 변화 추이를 보여준다. 익명처리 된 집계 데이터를 활용한 경우에는 카드사 가입자의 월별 온라인 콘텐츠 지출액 총량 추이를 파악하는데 활용할 수 있으나 현재 성별로 구분된 지출액을 비교할 경우 남성가입자가 여성가입자보다 평균적으로 지출액이 많은지 적은지 확인할 수 없다. 하지만 가명 처리된 데이터를 활용할 경우 각 집단별 1인당 평균적인 지출액을 산출할 수 있을 뿐만 아니라 여성과 남성, 그리고 연령과 성별 등을 조합해 각 집단별로 지출액을 산출할 수 있다.

〈표 4-23〉 개인단위 원시 자료 활용을 통한 온라인 콘텐츠 지출액 분석 결과 비교

(익명집계데이터 활용 시) 온라인 콘텐츠 월별 지출액 총량 추이

(대상: 신한카드 가입자, 단위: 백만원)

구분		1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
전체	'19	·	·	·	·	·	·	·	6,547	7,346	6,425	6,502	7,021
	'20	9,194	7,933	8,655	7,733	8,082	8,317	8,682	8,843	9,450	9,511	8,156	9,791
	'21	4,553	4,097	5,085	4,577	4,903	4,656	·	·	·	·	·	·

자료: 신한카드 익명데이터,

(가명처리데이터 활용 시) 온라인 콘텐츠 1인당 평균 월별 지출액

(대상: 신한카드 가입자, 단위: 1인/원)

구분		1월	2월	3월	4월	5월	6월	7월	8월	9월	10월	11월	12월
전체	'19	·	·	·	·	·	·	·	1,898	2,130	1,863	1,885	2,035
	'20	2,665	2,300	2,509	2,242	2,343	2,411	2,517	2,563	2,739	2,757	2,364	2,838
	'21	1,320	1,188	1,474	1,327	1,421	1,350	·	·	·	·	·	·
남성	'19	·	·	·	·	·	·	·	2,697	3,038	2,585	2,624	2,787
	'20	3,802	3,198	3,459	3,006	3,107	3,257	3,375	3,448	3,775	3,711	3,123	3,808
	'21	1,323	1,198	1,522	1,373	1,465	1,391	·	·	·	·	·	·
여성	'19	·	·	·	·	·	·	·	1,032	1,145	1,080	1,084	1,220
	'20	1,433	1,326	1,479	1,413	1,515	1,495	1,588	1,604	1,617	1,724	1,542	1,787
	'21	1,316	1,176	1,423	1,277	1,374	1,305	·	·	·	·	·	·

자료: 신한카드 가명처리데이터,

2. 패널데이터를 활용한 종단분석

앞서 가명처리 데이터는 '개인단위 원시데이터'로 총량 중심의 통계작성을 1인 기준으로 특성을 반영해 분석할 수 있다는 장점을 제시하였다. 이와 함께 가명처리데이터는 막대한 비용과 시간을 요구하는 조사로 구축된 기존의 패널데이터를 대체하거나 보완할 수 있다는 장점을 가진다. 통신, 금융 등 주요 생활밀착형 서비스를 제공하는 기업이 보유한 개인정보는 본 연구에서 활용한 것처럼 문화, 관광 등 모든 분야에서 활용 가능할 뿐만 아니라 개인의 행동을 실측한 빅 데이터라는 점에서 장점을 가진다.

예를 들어 국민의 미디어 이용행태를 분석하기 위해 정보통신정책연구원의 한국미디어패널조사는 연간 5,000가구와 이들 가구에 속한 12,000명의 가구원에 대해 전국단위 조

사를 실시하고 있으며 거대한 패널을 유지하고 이들에 대한 조사를 위해 많은 예산을 투입하고 있다. 또한 이 조사에서는 12,000명의 가구원이 3일 간 어떠한 미디어를 이용하고, 얼마나 많은 시간을 이용하였는지 조사대상자가 15분 단위로 직접 기입하는 '미디어 다이어리' 조사를 병행하고 있는데 조사대상자가 미디어 이용에 대한 사항을 직접 기입해야하는 구조적인 문제로 인해 실제 이용행태를 정확하게 담지 못한다는 한계점이 존재한다.

〈표 4-24〉 2020년도 한국미디어패널조사 개요

구 분	가 구	가구원
조사대상	10차년도(2019년) 기준 최종 유지 및 구축된 전국 4,537개 통합 패널 가구 (KMPS11 3,510 + KMPS19 1,027) 및 해당 가구의 만 6세 이상 가구원 모두 (약 10,800명) ※ 이탈 패널에 대한 지속 접촉 병행	
조사예산	730,000,000원	
조사지역	전국	
조사도구	구조화된 설문지(가구용)	구조화된 설문지(개인용) + Media Diary
표본규모	통합 패널의 94.8%로 4,299가구 이상 (KMPS11(N=3,510)의 95.5%, KMPS19(N=1,027)의 92.1% 이상 유지	가구 내 만 6세 이상 가구원 전체로약 10,800표본 (가구원 변동에 따라 유동적)
조사방법	면접원 방문에 의한 1:1 타계식 면접조사 ※ 7차년도부터 CAPI 진행	방문 면접 후 설문 유치 (자기기입식)

자료: 정보통신정책연구원(2020.12.), 2020 한국미디어패널조사 p.36 〈표 1-1〉 재구성.

본 연구에서 활용하고 있는 가명처리 데이터는 생활필수 서비스인 통신, 신용카드 서비스의 이용자가 서비스 이용과정에서 온라인 콘텐츠와 관련되어 활동한 실제 기록이라는 점에서 앞서 살펴본 한국미디어패널조사의 미디어다이어리 조사와 같은 기록형 조사를 보완할 수 있는 방안이 될 수 있다. 또한 이들 서비스를 사용하고 있는 이용자 규모는 100만 명 단위 이상으로 매우 클 뿐만 아니라 이들 서비스가 생활필수 서비스이기 때문에 서비스 이용이 매우 비탄력적이기 때문에 장기간 패널을 유지하기에도 용이하다는 장점을 가지고 있다.

가. 코로나 19 전후 구독서비스 이용 변화 집단의 특성 종단 분석

본 연구에서 활용하는 가명처리데이터는 코로나19기간을 포함한 시계열 데이터이며, 매 시점마다 동일한 사람들로 구성된 균형패널데이터(Balanced Panel Data)이다. 온라인 콘텐츠 구독서비스 이용행태에 대한 분석은 과거 단건 단위로 구매하는 방식에서 다수의 콘텐츠를 월단위로 구독하는 방식으로 콘텐츠 유통/소비 행태가 변화하면서 매우 중요하다. 본 연구에서는 인구통계학적인 특성을 기준으로 온라인 콘텐츠 구독서비스 이용률이 어떻게 변화하였는지에 대한 분석을 다양한 조건에 따라 아래와 같이 진행하였고, 아래와 같이 집단별로 온라인 콘텐츠 구독서비스 이용에 차이점이 존재하는 것을 확인할 수 있었다.

〈표 4-25〉 **(가명처리 데이터 활용 시)** 구독서비스 이용 변화 집단의 특성 분포

(대상: SK텔레콤 가입자, 단위: %)

구 분	성별		연령		가구원수		
	여성	남성	MZ	그외	1인	2인	3인
(변화1) 분석기간 내 지속적으로 구독 서비스를 이용한 집단	10.2%	10.4%	15.1%	6.6%	12.0%	9.8%	10.3%
(변화2) 분석기간 내 한 번도 구독 서비스를 이용하지 않은 집단	43.8%	44.0%	34.0%	51.6%	41.7%	46.0%	44.9%
(변화3) 코로나 19 이후 이용자 특성: 2019년 부터 2020년 1월까지 사용 안하다가 2020년 2월부터 1회 이상 사용한 집단	24.6%	24.7%	24.4%	24.8%	22.8%	23.5%	23.9%
(변화4)코로나 19 이후 이용자 특성: 2019년 부터 2020년 1월까지 사용 안하다가 2020년 2월부터 지속적으로 사용한 사람	0.4%	0.5%	0.6%	0.4%	0.5%	0.5%	0.5%

자료: SK텔레콤 가명처리데이터

나. 코로나 19 전후 개인 특성별 온라인 콘텐츠 지출액 변화 종단 분석

다음으로 신용카드사의 가명처리 정보를 활용해 각 집단별 1인당 온라인 콘텐츠 지출액이 어떻게 변화하고 있는지 살펴보았다. 아래 표에서 확인할 수 있듯이 1인당 온라인 콘텐츠 지출액은 지속적으로 증가하다가 2021년도 2월에는 감소하는 공통점을 보였다. 성별로 구분해 살펴보면 남성의 온라인 콘텐츠에 관한 지출액이 여성보다 2019년 8월부터 2020년 8월까지 높았으나, 2021년 2월에는 비슷한 수준으로 감소한 것을 확인할

수 있다. 연령을 기준으로 분류한 비교표에서 최근 모든 분야의 소비를 주도하고 있는 MZ세대의 온라인 콘텐츠 지출액이 MZ세대가 아닌 세대의 지출액보다 월등히 높은 것을 확인할 수 있다. 연간소득을 기준으로 온라인 콘텐츠 지출액을 분석한 결과에서도 소득과 온라인 콘텐츠 지출액은 비례하는 추이를 확인할 수 있다.

〈표 4-26〉 (가명처리 데이터 활용 시) 인구특성별 온라인 콘텐츠 1인당 평균 월별 지출액

(대상 신한카드 가입자, 단위: 1인/원)

구 분		2019년 8월	2020년 2월	2020년 8월	2021년 2월	추 이
성별	여성	1,032	1,326	1,604	1,176	
	남성	2,697	3,198	3,448	1,198	
연령	MZ	2,817	3,851	4,605	1,812	
	그외	882	971	1,151	709	
연간 소득	3천만원 미만	1,525	1,732	1,949	1,080	
	3천만원 이상 5천만원 미만	1,900	2,263	2,465	1,123	
	5천만원 이상	2,540	3,263	3,767	1,560	

자료: 신한카드 가명처리데이터,

3. 유효표본 수 확대에 따른 세부특성별 현황 분석

개인정보를 가명 처리한 데이터를 활용하게 되면 기존 조사통계 등에서 제공하고 있는 표본의 규모보다 월등히 큰 규모의 샘플을 활용할 수 있다는 장점을 갖고 있다. 앞서 살펴본 한국미디어패널조사는 조사대상 1인당 약 6만원의 비용을 투입해 연간 7.3억 원의 예산을 활용해 5,000가구, 12,000명에 대한 설문조사를 수행하고 있다. 반면에 본 연구에서 활용하는 통신사업자의 데이터는 통신사마다 차이가 있겠지만 1,000만 명에서 2,400만 명 규모의 가입자 규모를 가지고 있기 때문에 조사를 통해 유지/확보하고 있는 패널의 규모보다 월등히 규모가 크다.

설문조사 등을 활용한 조사의 경우 지역, 연령, 성별 등 인구통계학적인 기준으로 전국단위 패널데이터를 구축하더라도 장기간 패널조사를 이어갈수록 이탈하는 샘플규모가 지속적으로 증가하고, 샘플 이탈에 따른 분포상의 문제점이 발생한다는 단점이 존재한다

(오미애, 2019). 반면에 대규모 가명처리 데이터는 설문조사를 통해 얻는 정보보다 시계열을 세분화 할 수 있을 뿐만 아니라 분석대상 규모가 설문조사기반 패널조사보다 월등히 크기 때문에 샘플이 이탈하더라도 분포상의 문제점이 발생할 확률이 낮다.

아래 표는 2020년도 통계청 인구총조사와 KISDI 한국미디어패널조사, 본 연구에서 활용한 이동통신 가명처리 데이터 간 관측치수를 비교한 표이다. 데이터의 규모는 약 2,400배의 차이가 있으며, 각 유형별로 보더라도 수백 배 이상의 규모차이를 확인할 수 있다.

〈표 4-27〉 가명처리 후 유효표본 수 확대에 따른 확보 가능한 개인 특성별 유효표본 수 비교

(단위: 천명)

2020 인구총조사		**SK 가명처리 데이터**		2020 미디어패널조사	
전체표본수	50,710	전체표본수	24,240	전체표본수	10
남성	25,251	남성	*****	남성	5
여성	25,459	여성	*****	여성	5
20대이하	8,704	20대이하	*****	20대이하	3
30대	7,031	30대	*****	30대	2
40대	7,107	40대	*****	40대	2
50대	8,232	50대	*****	50대	2
60대	8,583	60대	*****	60대	1
70대이상	6,619	70대이상	*****	70대이상	1

이렇게 데이터의 규모가 크면 개인이 가진 세부특성 정보를 추가적으로 활용해 분석할 때 매우 유용하다. 예를 들어 한국미디어패널조사에서 소득구간 4,800만원 이상부터 6,000만원 미만에 해당하는 표본 수는 327명이고, 해당 소득구간에 속한 30대 남성들에 속한 특성분석을 실행하게 되면 표본 수는 기하급수적으로 감소해 100샘플 미만의 숫자를 가질 수 있다. 반면에 이동통신사업자의 가명처리 데이터를 분석에 활용한다면 여러 조건에 의해 분석대상 수가 감소하더라도 만 명 단위의 샘플을 가져갈 수 있을 것으로 보인다. 일반적으로 패널 데이터 분석에 많이 활용되는 고정효과(Fixed Effect) 모형은 관측된 데이터에서 각 개인별 평균을 빼주는 방식으로 고정효과를 추정하게 되는데 분석자료의 표본 수가 작거나, 불성실한 응답이 많아 결측치가 많을 경우 고정효과 모형을 적용하기 어렵다. 가명처리 데이터는 불성실한 응답을 회피할 수 있고, 표본수가 충분히 높아 기존 실증분석 모형을 적용할 때에도 조사대비 장점을 가진다. 하지만 가명처리 데이터가 가진 한계점도 분명하다. 설문조사 방식의 패널조사는 매년 공통질문을

제외한 질문들에 현재 시점의 이슈 또는 연구자가 궁금한 사항을 포함하여 원하는 응답을 얻을 수 있다는 장점이 존재하고, 가명처리 데이터는 이용자가 이용한 기록과 이들이 서비스 이용 시 등록한 정보를 기반으로 구축된 자료이기 때문에 이미 기록된 정보 이외에 연구자가 추가적으로 정보를 얻기 어렵다. 본 연구에서는 유형을 세분화하더라도 충분한 관측치를 가질 수 있다는 장점을 활용해 다양한 유형을 세분화하여 살펴본 온라인 콘텐츠 이용률과 지출액에 대한 분석을 실시하였다.

가. MZ세대의 세부특성별 오프라인 콘텐츠 이용률 분석

다음 표는 인구통계학적 특성을 활용해 온라인 콘텐츠 이용자를 세분화하고, 이들의 분석기간 동안의 온라인콘텐츠 이용률을 분석한 표이다. 세분화된 이용자는 결혼을 하지 않은 싱글이면서 MZ세대에 해당하고, 연간 소득이 아래와 같은 구간에 해당되는 사람들로 범주화되어 있다. 이러한 세부특성으로 온라인 콘텐츠 이용자를 좁은 범위로 범주화하더라도 아래 표에서 볼 수 있듯이 연 평균 관측치 수는 최소 48.7만 명에서 최대 275.1만 명의 샘플을 확보할 수 있다. 주요 결과를 살펴보면 연간소득 규모 순서로 온라인 이용률이 큰 것을 확인할 수 있고, 다만 소득이 낮은 구간의 싱글인 MZ세대의 분석기간 동안의 이용률 증가가 소득이 높은 구간에 있는 집단보다 높은 것도 확인할 수 있다.

〈표 4-28〉 (가명처리 데이터 활용 시) 인구특성별 온라인 콘텐츠 이용율

(대상: SK텔레콤 가입자, 단위: 명, %)

구 분 (결혼여부/세대/소득구간)	연평균 관측치수	2019.8	2020.2	2020.6	2021.6	기간 증감
싱글*MZ세대*5000-7000미만	511,951	80.0%	89.7%	93.2%	98.6%	18.7%p
싱글*MZ세대*4000-5000미만	897,909	80.7%	89.4%	93.0%	98.7%	18.0%p
싱글*MZ세대*3000-4000미만	1,547,329	79.9%	88.6%	92.5%	98.5%	18.6%p
싱글*MZ세대*2000-3000미만	2,751,256	78.3%	85.3%	90.3%	97.9%	19.6%p
싱글*MZ세대*1000-2000미만	487,876	77.8%	83.2%	88.4%	97.0%	19.2%p

자료: SK텔레콤 가명처리데이터

나. MZ세대의 세부특성별 온라인 콘텐츠 지출액 분석

다음은 신용카드 가명처리 데이터를 활용해 가입자를 인구통계학적 특성을 기준으로 구분하고 이들의 온라인 콘텐츠 지출액을 분석한 표이다. 세분화된 이용자는 여성이면서 MZ세대이고, 각 소득구간에 해당되며, 구독서비스를 이용하는지 여부로 구분하였다. 앞서 살펴본 온라인 콘텐츠 이용률 추이는 2,400만 규모의 이동통신 데이터를 활용해 이용하였고, 지출액 추이 분석에는 845.4만 명 규모의 신용카드 데이터를 활용하였다. 해당 분석을 통해 살펴 본 MZ세대 여성은 2019년 8월 기준 1인 당 1,200원에서 2,000원 수준으로 온라인 콘텐츠에 대한 소비활동을 하였으며, 2021년 6월 현재 1인 당 1,700원에서 2,200원 수준으로 온라인 콘텐츠에 대한 소비를 이어가고 있다는 점을 확인할 수 있다. 이와 함께 구독서비스를 이용하지 않는 MZ세대 여성 중 연간 소득이 3,000만원 미만인 집단의 1인당 지출액이 분석기간 동안 약 50.9% 증가하였고, 연간 소득이 3,000만원 이상 5,000만원 미만인 MZ세대 여성 중 구독서비스를 이용하는 여성의 지출액 증가도 약 28%로 높은 증가율을 보였다. 이처럼 가명처리 데이터는 분석대상을 세분화하여도 충분한 관측치수를 확보할 수 있기 때문에 기업이 시장을 세분화하는 전략을 수립하는 데에도 활용될 수 있다.

〈표 4-29〉 (가명처리 데이터 활용 시) 인구특성별 1인당 월평균 온라인 콘텐츠 관련 카드 지출액

(대상: 신한카드 가입자, 단위: 명, 1인/원)

구 분 (성별/세대/소득구간/구독서비스 이용여부)	연평균 관측치수	2019.8	2020.2	2020.6	2021.6	기간 증감
여성/MZ/3,000만원 미만/이용안함	15,446	1,272	1,630	2,039	1,920	50.90
여성/MZ/3,000만원 미만/이용함	40,113	1,543	1,365	1,725	1,580	2.41
여성/MZ/3,000만원 이상 5,000만원 미만/이용안함	71,772	1,838	2,297	2,432	2,002	8.91
여성/MZ/3,000만원 이상 5,000만원 미만/이용함	179,015	1,377	1,853	2,017	1,764	28.09
여성/MZ/5,000만원 이상/이용안함	89,683	1,811	2,373	2,720	2,105	16.19
여성/MZ/5,000만원 이상/이용함	226,498	1,979	2,467	2,817	2,208	11.57

제6절 소결

4장에서는 통신 및 카드 가명처리 데이터를 실제 분석에 활용함으로써 가명처리 데이터 활용을 위해 필요한 행정과 데이터 컬럼 구성에 필요한 가명처리 절차, 분석을 위한 데이터 구조 설계방안을 제시하였다. 또한 실제 구성된 가명처리 데이터를 활용하여 관광과 문화콘텐츠 분야의 분석 가능한 주제를 검토한 후, 실제 몇 가지 주제를 점검함으로써 가명처리 데이터가 가지는 강점을 실증적으로 살펴보았다. 이에 가명처리 데이터를 구성하기위한 실제 업무 절차도는 다음과 같이 도식화 할 수 있다.

〈표 4-30〉 가명처리 데이터 활용을 위한 업무 절차도

구분	주체
1단계: 데이터 활용 관련 업무 협의 데이터의 활용을 위한 데이터 제공기관과 이용기관 간, 업무 협의	제공기관, 이용기관
⇩	
2단계: 가명처리 데이터 분석환경 조성 가명정보 활용 관련 내부관리계획 수립 및 점검 가명정보 활용 분석 환경 점검	이용기관
⇩	
3단계: 가명처리 데이터 구성 (*상향식 접근(bottom-up)방식) 가명처리 데이터 활용주제 선정(이용기관) 활용 가능한 컬럼 보유 여부 확인(이용기관, 제공기관) 선정된 컬럼의 활용 관련 범무적 검토(제공기관) 최종컬럼 확정 후 가명 및 익명처리(제공기관)	제공기관, 이용기관
⇩	
4단계: 가명처리 활용 및 반출 적정성 평가 반출 신청서 작성(이용기관) 가명처리 데이터 반출 및 활용 적정성 평가위원회 개최(제공기관) 비밀보호협약서 작성(이용기관, 제공기관)	제공기관, 이용기관
⇩	
5단계: 데이터 반출 및 문석 개인정보 보호를 위해 암호화 전송(제공기관) 지정된 분석 환경 내 업무 진행(이용기관)	이용기관

앞서 살펴본 바와 같이, 실제 가명처리 데이터 활용을 위해서는 제공기간과 이용기관 간의 지속적인 업무 협의가 필요하며 가명정보 활용을 위한 다양한 행정적 절차와, 데이터 활용을 위한 물리적 환경 마련 등의 여러 절차가 필요하다. 또한 아직까지 제공기관이 보유한 정보를 공개하는 것이 활발하게 이루어지는 단계가 아니기 때문에 데이터 구성 시 업무의 비효율이 발생하게 된다는 단점이 있다. 데이터 구성 수준 또한 개인정보 보호를 위한 가명화 처리 과정에서 많은 부분의 정보 손실이 발생함을 확인할 수 있다.

그럼에도 불구하고 가명처리 데이터를 활용하는 경우 기존에 집계데이터로만 활용되던 민간데이터를 개인단위 원시단위 데이터로 접근하여 분석할 수 있게 되고, 누적된 정보를 활용하여 패널 데이터를 구성할 수 있으며, 확보할 수 있는 유효표본수가 크게 확대될 수 있는 가능성을 실증분석을 통해 확인할 수 있었다. 이에 향후 가명처리 데이터를 분석하기 위한 업무 절차의 효율성 및 연구 목적에 따라 가명화 수준의 정도가 조정 가능하다면 향후 가명처리 데이터의 활용성은 더욱 더 증대 될 것으로 예상된다.

제5장

[데이터 결합]
문화관광 부문 활용방안

제1절 데이터 결합 절차도

다음은 가명처리한 SK텔레콤과 신한카드의 가명처리 데이터를 결합한 데이터를 활용하여 기존과 차별화된 정책적 가치를 실증적으로 살펴보고자 한다. 데이터 결합 절차는 제 2장에서 언급한 가명정보 결합절차(참고: 〈표 2-13〉)를 기반으로 진행하되, 제공기관과 결합기관의 요구 사항에 따라 추가적인 행정절차를 진행하였으며 다음 〈표 5-1〉과 같이 도식화 할 수 있다.

〈표 5-1〉 데이터 결합 업무 절차도

구분	주체
1단계: 데이터 결합 관련 업무 협의 데이터의 안전한 활용을 위해 제공기관과 이용기관 간, 협약체결	제공기관-이용기관
⇩	
2단계: 가명정보 활용 내부관리 계획 및 분석 환경 점검 가명정보 활용 관련 내부관리계획 수립 및 점검 가명정보 활용 분석 환경 점검	이용기관
⇩	
3단계: 데이터 결합 신청 가명정보 관리 환경 및 이행 확약서 작성(이용기관) 결합 대상 데이터 작성(제공기관) 데이터 결합 적정성 평가(결합기관)	제공기관, 이용기관, 결합기관
⇩	
4단계: 데이터 결합 및 적정성 평가 적적성 평가위원회 개최, 데이터 결합 및 이용에 대한 적정성 평가(결합기관)	제공기관, 이용기관, 결합기관
⇩	
5단계: 데이터 반출 및 문석	이용기관

1. 데이터 결합 관련 업무 협의

SK텔레콤와 신한카드사의 데이터를 결합하기 위해서는 앞서 2장에 언급한 바와 같이 신용정보법에 따른 금융분야 데이터 전문기관을 통해 진행된다. 이에 본 연구에서는 SK텔레콤과 신한카드사는 결합 대상 데이터를 제공하는 '제공기관', 연구원은 결합된 데이터를 이용하는 '이용기관', 금융보안원은 결합 대상 데이터를 결합하는 '결합기관'으로 정의하였으며 각 기관의 역할은 다음과 같이 구분된다.

〈표 5-2〉 데이터 결합을 위한 결합 데이터 제공기관, 이용기관, 결합기관 주요 업무 내용

데이터 제공기관 -SK텔레콤, 신한카드-	데이터 이용기관 -한국문화관광연구원-	데이터 결합 기관 -금융보안원-
- 정보물 결합 신청 - 데이터 제공 - 데이터 설명 및 가명처리 내용 작성 - 각 변수별 분포 현황 작성 - 평가위원회 참석	- 정보물 결합 신청 - 가명정보에 활용에 대한 필수 조치 사항 구비 - 평가위원회 참석 - 데이터 분석	- 데이터 결합 - 가명 및 익명처리 - 적정성 평가(평가위원회개최) - 심의결과에 따라 결합 결과물 제공(암호화 전송)

2. 결합 데이터 분석환경 조성

결합 데이터 또한 이종의 가명처리 데이터를 활용하는 것이기 때문에 가명처리 데이터 분석환경을 조성하면 결합 데이터도 동시에 다룰 수 있게 된다. 마찬가지로 기 수립된 가명정보 활용 관련 내부관리계획을 점검하고, 가명정보 활용 분석환경을 가명처리 데이터와 동일하게 갖추었다.(참고: 제 4장 1절 가명처리 데이터 분석환경 조성, 70p)

3. 데이터 결합 신청

결합하고자 하는 데이터는 개인의 신용정보를 포함하고 있어 신용정보법 하 결합전문기관인 금융보안원에 SK 신한 데이터 결합을 신청하였다. SK텔레콤과 신한카드사는 데이터 제공기관, 연구원은 이용이관으로 지정하였으며 이를 위해 데이터 이용기관으로써의 결합 신청서와 금융보안원에서 요구하는 가명정보에 대한 필수 조치 이행 확약서를 작성하여 제출하였다.

가. 정보집합물 결합 신청서 작성

금융보안원 데이터전문기관에서 데이터 결합을 진행하는 경우, 금융보안원이 운영하는 포털(https://data.fsec.or.kr/)[46] 통해서 진행된다.

신청서에 주기적/반복적 업무처리 여부, 결합의뢰기관 정보, 결합상대기관 정보, 결합정보를 작성하였으며, 이는 이용이관 뿐만 아니라 데이터 제공기관도 동일하게 적용된다.

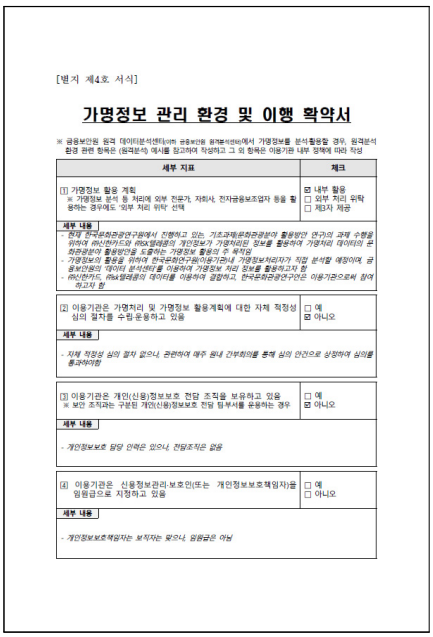
[그림 5-1] 금융보안원 데이터전문기관 - 정보물 결합 신청(예시)

나. 가명정보에 대한 필수 조치 이행확약서 제출[47]

정보집합물 결합을 위해서는 제공기관과 이용기관 모두 가명정보에 대한 필수 조치 이행확약서를 제출해야한다. 확약서는 가명정보 보호를 위한 내부관리계획 수립여부, 개인정보와 가명정보의 분리 저장, 가명정보 활용 계획에 대한 자체 적정성 심의 절차 수립 및 운용 여부 등 총 17개의 지표로 구성되어 있으며, 해당 지표를 바탕으로 심의절차를 거쳐 가명처리 수준을 결정하는 참고자료로 활용된다. 이에 본 연구원과 제공기관은 각 사의 가명처리 내부관리계획을 바탕으로 이행확약서를 작성 제출하였다.

46) 신청과 결합데이터 이용, 비용정산 관련 업무를 포털에서 진행하고 있음
47) 참고: [부록1] 가명정보 환경 및 이행 확약서 주요 내용, 193p

[그림 5-2] 한국문화관광연구원 가명정보 관리 환경 및 이행 확약서 (예시)

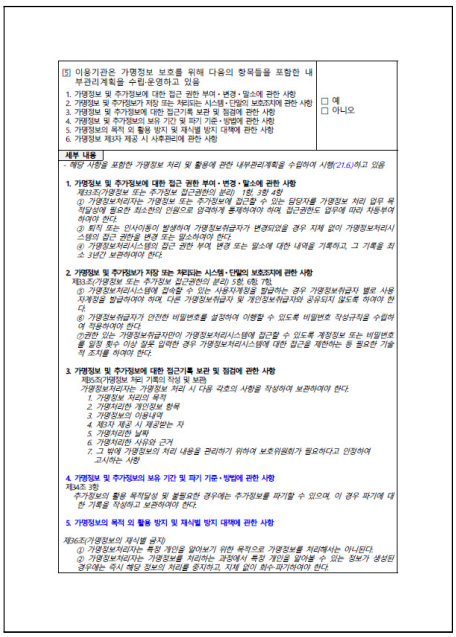


[별지 제4호 서식]

가명정보 관리 환경 및 이행 확약서

※ 융합반장은 원격 데이터분석센터(이하 융합보안실 원격분석센터)에서 가명정보를 분석활용할 경우, 원격분석 환경 관련 항목은 (원격분석) 예시를 참고하여 작성하고 그 외 항목은 이용기관 내부 정책에 따라 작성

세부 지표	체크
[1] 가명정보 활용 계획 ※ 가명정보 분석 등 처리에 외부 전문가, 자회사, 전자금융보조업자 등을 활용하는 경우에도 외부 처리 위탁 선택 세부 내용 - 현재 한국문화관광연구원에서 진행하고 있는 기초과제(문화관광분야 활용방안 연구)의 과제 수행을 위하여 KNU빅데이터와 IRiss빌데이터 개인정보보호 가명처리된 정보를 활용하여 가명처리 데이터의 분석 환경관련 활용방안을 도출하는 가명정보 활용의 주 목적임 - 가명정보의 활용을 위하여 한국문화연구원(이용기관)내 가명정보처리자가 직접 분석할 예정이며, 금융보안원의 데이터 분석센터를 이용하여 가명정보 처리 정보를 활용하고자 함 - KNU빅데이터, IRiss빌데이터의 데이터를 이용하여 결합하고, 한국문화관광연구원은 이용기관으로써 참여하고자 함	☑ 내부 활용 ☐ 외부 처리 위탁 ☐ 제3자 제공
[2] 이용기관은 가명처리 및 가명정보 활용계획에 대한 자체 적정성 심의 절차를 수립·운용하고 있음 세부 내용 - 자체 적정성 심의 절차 없으나, 관련하여 매주 원내 간부회의를 통해 심의 안건으로 상정하여 심의를 통과하여함	☐ 예 ☑ 아니오
[3] 이용기관은 개인(신용)정보보호 전담 조직을 보유하고 있음 ※ 보안 조직과는 구분된 개인(신용)정보보호 전담 팀부서를 운용하는 경우 세부 내용 - 개인정보보호 담당 인력은 있으나, 전담조직은 없음	☐ 예 ☑ 아니오
[4] 이용기관은 신용정보관리·보호인(또는 개인정보보호책임자)을 임원급으로 지정하고 있음 세부 내용 - 개인정보보호책임자는 보직자는 맞으나, 임원급은 아님	☐ 예 ☐ 아니오



[5] 이용기관은 가명정보 보호를 위해 다음의 항목들을 포함한 내부관리계획을 수립·운영하고 있음 1. 가명정보 및 추가정보에 대한 접근 권한 부여·변경·말소에 관한 사항 2. 가명정보 및 추가정보가 저장 또는 처리되는 시스템·단말의 보호조치에 관한 사항 3. 가명정보 및 추가정보에 대한 접근기록 보관 및 점검에 관한 사항 4. 가명정보 및 추가정보의 보유 기간 및 파기 기준·방법에 관한 사항 5. 가명정보의 목적 외 활용 방지 및 재식별 방지 대책에 관한 사항 6. 가명정보 제3자 제공 시 사후관리에 관한 사항	☐ 예 ☐ 아니오
세부 내용 - 해당 사항을 포함한 가명정보 처리 및 활용에 관한 내부관리계획을 수립하여 시행(21.6.)하고 있음 **1. 가명정보 및 추가정보에 대한 접근 권한 부여·변경·말소에 관한 사항** 제23조(가명정보 또는 추가정보 접근권한의 분리) 1항, 3항, 4항 ① 가명정보처리자는 가명정보 또는 추가정보에 접근할 수 있는 담당자를 가명정보 처리 업무 목적달성에 필요한 최소한의 인원으로 엄격하게 통제하여야 하며, 접근권한도 업무에 따라 차등부여하여야 한다. ② 퇴직 또는 인사이동이 발생하여 가명정보취급자가 변경되었을 경우 지체 없이 가명정보처리시스템의 접근 권한을 변경 또는 말소하여야 한다. ④ 가명정보처리시스템의 접근 권한 부여, 변경 또는 말소에 대한 내역을 기록하고, 그 기록을 최소 3년간 보관하여야 한다. **2. 가명정보 및 추가정보가 저장 또는 처리되는 시스템·단말의 보호조치에 관한 사항** 제33조(가명정보 또는 추가정보 접근권한의 분리) 5항, 6항, 7항. ① 가명정보처리시스템에 접속할 수 있는 사용자계정을 발급하는 경우 가명정보취급자 별로 사용자계정을 발급하여야 하며, 다른 가명정보취급자 및 개인정보취급자와 공유되지 않도록 하여야 한다. ⑥ 가명정보취급자가 안전한 비밀번호를 설정하여 이행할 수 있도록 비밀번호 작성규칙을 수립하여 적용하여야 한다. ⑦권한 있는 가명정보취급자만이 가명정보처리시스템에 접근할 수 있도록 계정정보 또는 비밀번호를 일정 횟수 이상 잘못 입력한 경우 가명정보처리시스템에 대한 접근을 제한하는 등 필요한 기술적 조치를 하여야 한다. **3. 가명정보 및 추가정보에 대한 접근기록 보관 및 점검에 관한 사항** 제35조(가명정보 처리 기록의 작성 및 보관) 가명정보처리자는 가명정보 처리 시 다음 각호의 사항을 작성하여 보관하여야 한다. 1. 가명정보 파일의 생성 2. 가명처리한 개인정보 항목 3. 가명정보의 이용내역 4. 제3자 제공 시 제공받는 자 5. 가명처리한 날짜 6. 가명처리한 사유와 근거 7. 그 밖에 가명정보의 처리 내용을 관리하기 위하여 보호위원회가 필요하다고 인정하여 고시하는 사항 **4. 가명정보 및 추가정보의 보유 기간 및 파기 기준·방법에 관한 사항** 제34조 3항 추가정보의 활용 목적달성 및 불필요한 경우에는 추가정보를 파기할 수 있으며, 이 경우 파기에 대한 기록을 작성하고 보관하여야 한다. **5. 가명정보의 목적 외 활용 방지 및 재식별 방지 대책에 관한 사항** 제36조(가명정보의 재식별 금지) ① 가명정보처리자는 특정 개인을 알아보기 위한 목적으로 가명정보를 처리해서는 아니된다. ② 가명정보처리자는 가명정보를 처리하는 과정에서 특정 개인을 알아볼 수 있는 정보가 생성된 경우에는 즉시 해당 정보의 처리를 중지하고, 지체 없이 회수·파기하여야 한다.	

다. 정보집합물 데이터 명세서 제출

추가적으로 데이터 제공기관은 가명정보에 대한 필수 조치 이행 확약서 외, 정보집합물 데이터 명세서를 제출해야한다. 명세서에는 데이터 명세, 가명처리 요약사항, 범주형 변수 분포 현황, 수치형 변수 분포 현황을 제출해야한다. 이를 바탕으로 평가위원회를 통해 가명처리 수준 공개 범위 등이 결정되게 된다.

[그림 5-3] 데이터 제공기관 데이터 명세서(양식)

1. 데이터명세

1-1) 개요

항목	내용
■ 의뢰기관명	A카드
■ 데이터 크기 (용량)	20.5 GB
■ 데이터 레코드 수	25,000,000
■ 데이터 컬럼 수	14
■ 데이터 생성 방법	18년~20년 A카드 이용자에 대한 카드상품 신청/이용/탈퇴 등의 내부정보 및 외부 조회 정보
■ 결합데이터 활용목적	B은행의 신규 신용카드 결합상품 개발시 분석하기 위한 통계작성
■ 결합데이터 이용기관	B은행

1-2) 정보 개요

No.	정보영역	항목건수	비중
1	KEY	1	7.1%
2	신청정보	6	42.9%
3	결제정보	2	14.3%
4	대출정보	2	14.3%
5	기타정보	3	21.4%
	Total	14	100.0%

1-3) 정보항목별 상세

No.	정보원천	자료형태	정보영역	컬럼명	항목명(국문)	항목 설명	분포현황 작성	제출 데이터 예시	비고 (데이터 출처)
1	KEY	-	결합키	A_key		작성 불요		작성 불요	작성 불요
2	내부정보	수치형	신청정보	column2	연령			90	
3	내부정보	범주형	신청정보	column3	성별			남	
4	내부정보	범주형	신청정보	column4	직업코드	○○기관의 직업코드 활용	○	20201	
5	내부정보	수치형	신청정보	column5	신청일자			20180101	
6	내부정보	수치형	신청정보	column6	계좌개설일자			20180101	
7	내부정보	수치형	신청정보	column7	개설상품 건수		○	14	
8	내부정보	수치형	결제정보	column8	KK카드상품 결제금액	A카드의 KK카드상품	○	70000	
9	내부정보	수치형	결제정보	column9	MM카드상품 결제금액	A카드의 MM카드상품	○	70000	
10	외부정보	범주형	대출정보	column10	카드대출 연체액 존재	연체 카드대출상품 존재여부 1:존재,0:미존재	○	7	○○신용평가회사 데이터 조회
11	외부정보	수치형	대출정보	column11	카드대출 상환액	전체 카드대출상품	○	500000	○○신용평가회사 데이터 조회
12	외부정보	수치형	기타정보	column12	소득금액		○	500000	○○청 데이터 조회
13	외부정보	수치형	기타정보	column13	건강보험료납부액		○	500000	○○공단 데이터 조회
14	외부정보	수치형	기타정보	column14	직원생산성지수		○	500000	○○원 데이터 조회

4. 데이터 결합 및 적정성 평가

데이터를 결합하기에 앞서 본 연구원이 가명처리 데이터 이용기관으로 충분한 보안장치와 분석환경을 마련하고 있는지, 그리고 활용할 가명처리 데이터에 개인정보를 침해할 중대한 사항이 포함되어있는 지 등을 종합적으로 평가하기 위한 '평가위원회'를 개최하였다.

평가위원회는 약 3명의 평가위원과 데이터 이용, 제공 및 결합기관 관계자가 모두 참석하였다. 본 연구원은 이용기관으로써 연구원이 기 제출한 가명정보 관리 환경 및 이행확약서에 대하여 간단하게 브리핑 한 후 평가위원의 질의응답에 대응하는 방식으로 진행하였다. 다만 이행확약서에 연구원의 대외비가 있을 수 있어 SK텔레콤과 신한카드사는 퇴장한 후 다시 평가에 참여하였다. 그 외 SK텔레콤과 신한카드는 제공하는 데이터에 대한 설명과 가명처리 과정에 대하여 설명하고 질의응답을 받는 방식으로 진행하였다.

평가 완료 후 승인이 완료되면 데이터 결합이 진행된다. 본 결합은 SK텔레콤과 신한카드 동시가입자를 추출하는 이너조인(Inner Join)으로 350만명이 결합되었으며 Key 변수는 "이름"과 "주민번호"를 사용하여 다음과 같이 구성하였다.

〈표 5-3〉 SK T, 신한카드 가명처리 데이터 결합 결과(예시)

(가명처리 데이터 결합 전)

〈기업1(SK텔레콤)〉

주민번호	전화번호	이동량	구독 서비스
860602	010-245-6789	34	Y
780302	010-457-9876	20	N
180523	010-789-0123	55	Y

〈기업2(신한카드)〉

주민번호	전화번호	관광 지출	콘텐츠 지출
860602	010-245-6789	15만원	10만원
780302	010-457-9876	5만원	6만원
180523	010-789-0123	50만원	8만원

(가명처리 데이터 결합 후)

결합 → SK텔레콤·신한카드

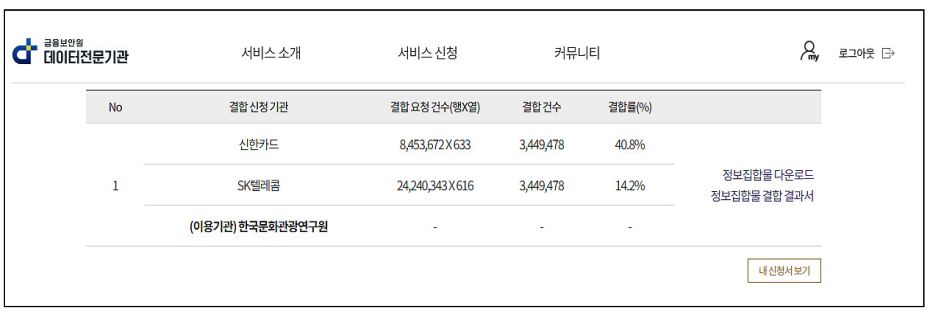
키값	이동량	구독서비스	관광 지출	콘텐츠지출
B0001	34	Y	15만원	10만원
B0002	20	N	5만원	6만원
B0003	55	Y	50만원	8만원

5. 결합 데이터 반출 및 분석

결합 데이터의 분석을 위하서는 데이터를 직접 반출하여 원내 분석환경을 활용하는 방안과 금융보안원 내 원격데이터 분석 시스템 이용 두 가지 방안이 있다.

첫째, 데이터를 직접 반출하고자 하는 경우 금융보안원(https://data.fsec.or.kr/) 포털을 통해서 다운로드 가능하다. 다만 직접 반출하여 분석하는 경우, 개인정보위원회 '가명정보처리 가이드라인'에 따라 가명정보를 분석할 수 있는 물리적·논리적 분리가 필요하다. 이에 본 연구원에서는 가명정보 활용을 위한 분석 환경 공간을 별도로 마련하였으며, 분석시스템 또한 금융보안원 내 포털이용 외 타 사이트 이용을 차단할 수 있도록 방화벽을 설정하여 직접 반출하여 분석하였다.

[그림 5-4] 금융보안원 데이터 반출 예시



둘째, 금융보안원 내 원격 데이터 분석 시스템을 이용하는 경우, 미리 원격시스템에 적제되어있는 결합데이터를 이용하여 분석한다. 권한을 부여받은 가명정보 취급자만 결합 데이터에 접근할 수 있도록, 금융보안원에서는 원격 데이터 분석 시스템 전용 ID/PW 를 이용기관의 가명정보 취급자에게 따로 발급한다. 또한 VDI 원격 데스크톱의 이벤트 로그를 자동으로 기록·보관하고 있다.

해당 분석 시스템은 분석가가 원하는 CPU와 메모리 수준을 지정할 수 있다는 장점이 있다. 그 외 분석가가 업무를 진행하는데 있어서 필요한 프로그램 코드나 명세서 등을 반입이 비교적 자유롭다. 다만, 원격 데이터 분석 시스템에 접속하기 위해서는 본인 인증을 2회 거쳐야 하며 반입된 파일도 암호화하여 개인의 핸드폰으로 암호가 전송되어 철저한 보안 하에 진행되고 있다.

[그림 5-5] 금융보안원 원격 데이터 분석 시스템 화면(예시)

제2절 최종 결합 데이터 구성

1. 컬럼 선정

결합 데이터는 기존의 SK텔레콤과 신한카드 가명처리데이터를 결합하였기 때문에 앞서 선택한 컬럼과 동일한 구조를 가지며 다음 〈표 5-4〉와 같다.[48]

〈표 5-4〉 [가명처리 데이터] 통신 데이터 최종 컬럼 구성

컬럼	기간	세부내용
ⅰ. 개인 특성		
성별	'19.1~'21.6	시군구 코드
연령대	'19.1~'21.6	1인 가구, 2인 가구, 3인 가구, 4인 가구, 5인 가구 이상
세대구분 1	'19.1~'21.6	M세대(1981~96년생), Z세대(1997~2010년생), 그 외 세대
세대구분 2	'19.1~'21.6	청년(18~34세), 중년(35~49세), 장년(50~64세), 노년(65세 이상)
가구원수	'20.1~'21.1	1인 가구, 2인 가구, 3인 가구, 4인 가구, 5인 가구 이상
거주지역	'19.1~'21.6	시군구 코드
연간소득 ** 8개 소득 구간*	'19.1~'21.6	1억이상, 7천-1억미만, 5천-7천미만, 4천-5천미만 3천-4천미만, 2천-3천미만, 1천-2천미만, 1천 미만
ⅱ. 관광 부문: 이동횟수		
평일 이동 횟수	'19.8~'21.6	평일 중에서 법정공휴일을 제외한 날 한 달 간 평일 총 이동 횟수
휴일 이동 횟수	'19.8~'21.6	2개 요일(토일) 또는 법정공휴일 한 달간 휴일 총 이동횟수
ⅲ. 문화 부문: 온라인 콘텐츠 이용		
평일 온라인콘텐츠이용량	'19.8~'21.6	전체, 동영상, 음악, 게임, 도서, 웹툰, 커뮤니티 등
휴일 온라인콘텐츠이용량	'19.8~'21.6	전체, 동영상, 음악, 게임, 도서, 웹툰, 커뮤니티 등
ⅲ. 문화 관련 온라인 콘텐츠 관련:		
구독서비스 정보	'19.8~'21.6	구독서비스 명
구독기간 이용 기간	'19.8~'21.6	개월
ⅱ. 오프라인 카드 지출액		
전체 월별	'19.1~'21.6	전체 발생한 신용카드 지출의 월별 총 금액(단위: 원)
관광 월별	'19.1~'21.6	관광 분야의 신용카드 지출의 월별 총 금액(단위: 원)
콘텐츠 월별	'19.1~'21.6	오프라인콘텐츠 분야의 신용카드 지출의 월별 총 금액(단위: 원)
ⅲ. 온라인 콘텐츠 지출액		
영상 월별 카드 지출	'19.1~'21.6	영상관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)
음악 월별 카드 지출	'19.1~'21.6	음악관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)
게임 월별 카드 지출	'19.1~'21.6	게임관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)
출판 월별 카드 지출	'19.1~'21.6	출판관련 온라인 콘텐츠 지출의 월별 총 금액(단위: 원)

48) 보다 자세한 컬럼 정보 및 가명처리 내용은 [부록2] 가명처리데이터 설명서 203 참고

2. 결합데이터 구성

앞선 4장의 가명처리 데이터를 주민등록번호와 핸드폰 번호를 key변수로 하여 최종 결합 데이터를 구성하였다. 결합 결과 SK텔레콤과 신한카드 동시가입자를 추출하는 이너조인(Inner Join)으로 350만명이 결합되었으며 신한카드 기준 결합률은 40.8%이다. (참고: 〈표 5-5〉)

〈표 5-5〉 통신 및 카드 데이터 최종 결합 결과

데이터 제공기관	결합 요청 건수(행X열)	결합 건수	결합률(%)
SK텔레콤	24,240,343 X 616	3,449,478	14.2%
신한카드	8,453,672 X 633		40.8%

결합 데이터 또한 가명처리 데이터와 마찬가지로 개인단위의 월 패널 데이터로 구조를 변환하여 사용하였다.

〈표 5-6〉 관광 분야 실증 분석을 데이터 구성 예시

ID	Time	휴일 이동횟수	구독서비스 이용	관광소비지출액	온라인소비지출액
A	19년 6월	30	이용	478000	608000
A	19년 7월	30	이용	39000	24000
A	19년 8월	35	이용	247000	856000
……	……	……	……	……	……
A	21년 5월	20	이용	460000	96000
A	21년 6월	25	이용	241000	374000
B	19년 6월	60	이용안함	1354000	856000
B	19년 7월	68	이용안함	383000	1452000
B	19년 8월	82	이용함	929000	508000
……	……	……	……	……	……
B	21년 5월	76	이용함	71000	104000
C	19년 8월	23	이용함	2053000	696000
……	……	……	……	……	……
C	21년 5월	22	이용안함	2026000	1465000
C	21년 6월	15	이용안함	1121000	615000

제3절 실증분석 주제 선정

다음은 구성한 SK텔레콤과 신한카드 지출액 가명처리 데이터 이종을 결합한 자료를 활용하여, 이종의 결합 데이터가 가지는 강점에 따라 도출 가능한 관광과 문화콘텐츠 분야실증분석 주제를 제시하고, 이 중 특정 주제를 선정하여 실제 실증분석을 진행하고자 한다. 이를 통해 이론적으로 살펴본 결합 데이터 활용 시 기대되는 강점을 실제 결합 데이터를 이용한 실증분석을 토대로 재 확인해 봄으로써 그 가치를 확인해 보고자 한다.

강점 1: 이종 데이터 결합을 통한 활용 가능한 정보의 확대

결합합분석은 가명정보 도입으로 가장 기대되는 이점 중 하나이다. 결합전문기관을 통해 결합한 가명정보는 데이터의 가치가 더욱 높아지고 결합정보를 활용해 기존에는 할 수 없었던 연구를 수행하여 새로운 지식과 가치를 발견하여 향후 정책 수립에 기여할 수 있을 것으로 기대되고 있다. 이에 통신데이터와 카드 데이터 결합 시 각각의 데이터로 보기 어려웠던 다음과 같은 분석을 진행할 수 있을 것으로 보인다.

이종 데이터 결합에 따른 분석 가능 주제(안)

- 관광 이동량에 따른 관광 소비지출 현황 및 상관관계 분석
- 관광 이동량에 따른 관광 세부업종별 소비지출 상관관계 분석
- 관광 이동량에 따른 오프라인 콘텐츠 소비지출 현황 및 상관관계 분석
- 관광 이동량에 따른 온라인 콘텐츠 소비지출 현황 및 상관관계 분석
- 온라인 콘텐츠 이용 유무별 관광 소비지출 현황 및 상관관계 분석
- 온라인 콘텐츠 이용 유무별 관광 세부업종별 소비지출 상관관계분석
- 온라인 콘텐츠 이용 유무별 오프라인 콘텐츠 소비지출 현황 및 상관관계 분석
- 온라인 콘텐츠 이용 유무별 온라인 콘텐츠 소비지출 현황 및 상관관계 분석
- 구독서비스 이용 유무별 관광 소비지출 현황 및 상관관계 분석
- 구독서비스 이용 유무별 관광 세부업종별 소비지출 상관관계분석
- 구독서비스 이용 유무별 오프라인 콘텐츠 소비지출 현황 및 상관관계 분석
- 구독서비스 이용 유무별 온라인 콘텐츠 소비지출 현황 및 상관관계 분석

- 지역별 이동량에 따른 관광 소비지출 분석
- 지역별 이동량에 따른 온오프라인 콘텐츠 소비지출 분석
- 축제 또는 MICE 행사의 파급효과 분석(방문객 산출 및 방문객의 소비지출 분석)
- 출퇴근 특성에 따른 온라인 콘텐츠 소비 지출 현황
- 특정 관광지의 코로나19 전후 소비와 이동성 변화 분석

장점 2: 유효표본 수 확대에 따른 세부특성별 분석

기존 가명처리 데이터 활용 시, 익명화된 집계데이터를 형태로 제공되던 민간데이터를 개인단위 원시데이터 자료로 확보할 수 있어 유효표본수가 확대됨에 따라 개인 세부특성분석이 가능하다는 점을 확인하였다. 이는 가명처리 데이터 이종을 결합하여도 세부특성별 분석이나 소규모 지역 추정 시 필요한 충분한 유효표본을 확보할 수 있다. 앞서 SK텔레콤과 통신 데이터의 결합 결과 총 350만개의 표본을 확보하였다. 이 정도의 표본 규모라면 개인의 특성에 따라 세부적으로 구분하여 분석하는 것이 가능한 수준으로 다음과 같은 분석이 가능하다.

유효표본 확대에 따른 분석 가능 주제(안)

- 1인 가구의 세부특성별, 관광이동량별 관광 소비지출 분석
- 1인 가구의 세부특성별, 이동량별 오프라인 콘텐츠 소비지출 분석
- 1인 가구의 세부특성별, 이동량별 온라인 콘텐츠 소비지출 분석
- MZ세대의 세부특성별, 관광이동량별 관광 소비지출 분석
- MZ세대의 세부특성별, 이동량별 오프라인 콘텐츠 소비지출 분석
- MZ세대의 세부특성별, 이동량별 온라인 콘텐츠 소비지출 분석
- 시군구 단위의 관광이동량별 관광 소비지출 분석
- 시군구 단위의 이동량별 오프라인 콘텐츠 소비지출 분석
- 시군구 단위의 이동량별 온라인 콘텐츠 소비지출 분석
- 온라인 콘텐츠 핵심 사용자의 세부특성별, 관광이동량별 관광 소비지출 분석
- 온라인 콘텐츠 핵심 사용자의 세부특성별, 이동량별 오프라인 콘텐츠 이용행태 분석
- 온라인 콘텐츠 핵심 사용자의 세부특성별, 이동량별 온라인 콘텐츠 이용행태 분석
- 구독서비스 핵심 사용자의 세부특성별, 관광이동량별 관광 소비지출 분석
- 구독서비스 핵심 사용자의 세부특성별, 이동량별 오프라인 콘텐츠 이용행태 분석
- 구독서비스 핵심 사용자의 세부특성별, 이동량별 온라인 콘텐츠 이용행태 분석

이러한 강점을 바탕으로 SK텔레콤과 신한카드 지출액의 '결합 데이터'를 활용하여 분석할 주제는 다음 〈표 5-7〉과 같이 선정하였다.

〈표 5-7〉 결합데이터 문화관광 부문 활용방안 도출을 위한 주제 선정

'결합데이터'를 활용한 관광 부문 활용방안 도출을 위한 실증분석 주제 선정	
활용 가능 정보 확대	**유효표본 수 확대**
▪ 관광 이동량에 따른 관광 소비지출 현황 및 상관관계 ▪ 관광 이동량에 따른 관광 세부업종별 소비지출 상관관계	▪ MZ세대의 세부특성별, 관광이동량별 관광 소비지출 분석 ▪ 시군구별 관광 이동량에 따른 관광소비지출 현황

'결합데이터'를 활용한 문화콘텐츠 부문 활용방안 도출을 위한 실증분석 주제 선정	
활용 가능 정보 확대	**유효표본 수 확대**
▪ 이동량에 따른 콘텐츠(온라인, 오프라인)소비 추이 분석 ▪ 지역별 이동량에 따른 온라인 콘텐츠 소비 추이 분석	▪ MZ세대의 세부특성별, 이동량별 온라인 콘텐츠 소비지출 분석 ▪ 시군구 단위의 이동량별 온라인 콘텐츠 소비지출 분석

제4절 관광 부문 활용방안

본 절에서는 SK텔레콤 가명처리 데이터와, 신한카드 가명처리 데이터를 결합한 데이터를 이용하여 결합데이터가 가지는 강점을 활용한 관광분야 분석을 진행하였다. 이를 통해 결합데이터가 가지는 강점을 실증적으로 이해하고 향후 관광 분야 통계 생산 시 기존과 어떠한 차별성이 있는지 비교 분석하고자 한다.

1. 이종 간 결합 정보를 활용한 현황 분석

이종데이터를 결합하게 되면 활용가능한 정보가 확대되어 새로운 결과를 도출할 수 있는 가능성이 높아진다. 본 연구에서는 이러한 강점을 보여줄 수 있는 주제로 SK텔레콤 통신데이터와 신한카드 지출액 데이터를 결합하여 휴일 이동량에 따른 소비지출 현황과 인과관계를 분석하였다.[49]

가. 관광 이동량에 따른 관광 소비지출 현황 및 상관관계

다음은 각각의 가명처리 데이터를 활용하여 이동량과 관광 소비지출 현황을 분석한 결과와 결합데이터 활용 후 휴일 이동량별 관광 소비지출 현황을 분석한 결과이다.[50]여기서 이동량에 수준에 따른 차이를 살펴보기 위하여 이동량 구간은 3분위 배율로 나누어 집단을 구분하여 이동량이 작은 집단은 1분위, 중간 집단은 2분위, 많은 집단은 3분위로 구분하였다. 결합 데이터를 활용하기 전에는 통신 데이터를 이용하여 휴일 이동 총량,

49) 휴일 이동량은 통신 데이터, 소비지출은 카드 데이터의 정보를 각각 활용하게 된다.

50) 여기서 휴일에 발생한 모든 이동이 관광 목적으로 이루어졌다고 보기는 어렵지만 휴일이 평일에 비해 비교적 여가 및 여행 목적의 이동 발생이 높다는 점을 반영하여 휴일 이동량을 기준으로 분석하였다. 또한 관광 소비 지출은 앞서 가명처리 데이터 구성 시 정의한 14개 업종 분야의 지출액을 의미한다.(참고: 〈표 4-4〉, 76p)

카드 데이터를 이용하여 카드사용총액을 각각 산출하였기 때문에 이동과 관광소비의 상관관계를 판단하기 어려웠다. 하지만 결합 데이터를 활용하게 되면 이동량 수준에 따른 관광소비 지출액 현황과 상관관계를 분석할 수 있어 보다 심층적인 시사점을 도출 할 수 있다. 다음 〈표 5-8〉 분석결과를 살펴보면 이동량이 많은 집단에서 전체 소비 특히, 관광 소비지출액이 비교적 크며 이 둘은 양의 상관관계가 있음을 보여준다. 이는 국민들의 관광 참여 증가는 소비지출에 긍정적 상관성을 보이며, 이는 관광활성화가 지역경제에 긍정적 영향을 줄수 있다는 점을 시사한다.[51]

〈표 5-8〉 이종 간 결합 정보를 활용한 이동량과 관광소비지출 분석 결과 비교

(단일 통신 데이터) 1인당 분기 총 휴일 이동량

대상: SKT가입자, 단위: 1인/회)

이동량	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
휴일 이동량	39	30	36	26	33	29	38

자료: SK텔레콤 가명처리 데이터

(단일 카드 데이터) 1인당 분기 총 관광부문 카드지출

(대상: 신한카드 가입자, 단위: 1인/원)

구분	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
관광 지출액	474,310	361,698	394,944	385,762	353,729	314,907	398,888

자료: 신한카드 가명처리 데이터

(통신 + 카드 결합 데이터) 1인당 이동량에 따른 분기 총 관광부문 카드지출

(대상 SK텔레콤, 신한 동시가입자 단위: 1인/원)

이동량	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
1분위(적음)	391,136	285,700	296,840	280,475	262,971	233,391	300,266
2분위(중간)	502,885	396,054	411,345	399,322	376,480	336,582	414,852
3분위(많음)	622,452	509,620	540,874	540,602	510,905	459,255	550,020

자료: SK텔레콤, 신한 결합데이터

51) 본 결과는 이동량과 관광 소비의 교차 분석과 및 상관관계를 분석한 것으로, 인과관계로 해석하기에 무리가 있으므로 이에 유의해야한다.

〈표 5-9〉 **(통신 + 카드 결합 데이터)** 1인당 이동량 별 관광소비지출 상관관계 분석

(대상: SK텔레콤, 신한 동시가입자, 피어슨상관계수)

구분	이동량	전체지출액
이동량	1.0000	
전체지출액	0.0599*	1.0000
전체		

구분	이동량	관광지출액
이동량	1.0000	
관광지출액	0.1793*	1.0000
관광부문 지출액		

자료: SK텔레콤, 신한 결합데이터

나. 관광 이동량에 따른 관광 세부업종별 소비지출 상관관계

결합 데이터를 이용하면 관광 업종 중 이동량에 비교적 크게 영향을 받는 업종과 적게 받는 업종에 대한 보다 세부적인 파악이 가능하다. 이러한 분석은 카드 데이터는 업종별로 세부적인 카드 지출액 정보를 제공하기 때문에 가능하다고 하겠다. 다음은 코로나19의 확산으로 가장 큰 지출 감소를 보이는 여행사, 면세점, 관광숙박업, 항공사 지출액을 이동량 수준에 따라 구분하여 분석한 결과이다. 결합데이터 활용 전에는 각 세부업종의 코로나19 확산 시 거리두기 시행 기간 동안의 지출액 변화를 통해 해당 산업의 피해 규모를 간접적으로 파악하였다. 하지만 결합데이터를 이용하게 되면 이동과 각 세부 업종별 상관성을 각각 추정할 수 있어 각 세부 업종에 대한 거리두기 정책의 영향 정도를 보다 세부적으로 파악할 수 있게 될 것으로 기대된다.

〈표 5-10〉 (통신 + 카드 결합 데이터) 1인당 이동량에 따른 분기 총 여행사 카드지출

(대상 SK텔레콤, 신한 동시가입자 단위: 1인/원)

이동량	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
1분위(적음)	8,325	3,977	633	833	664	512	668
2분위(중간)	7,841	3,835	546	588	474	430	522
3분위(많음)	7,139	3,685	377	461	413	318	429

자료: SK텔레콤, 신한 결합데이터

〈표 5-11〉 (통신 + 카드 결합 데이터) 1인당 이동량에 따른 분기 총 면세점 카드지출

(대상 SK텔레콤, 신한 동시가입자 단위: 1인/원)

이동량	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
1분위(적음)	10,823	5,739	3,508	5,018	6,278	5,377	7,043
2분위(중간)	10,930	5,501	2,087	3,024	3,730	3,056	4,233
3분위(많음)	10,356	5,245	1,195	1,666	2,050	1,712	2,431

자료: SK텔레콤, 신한 결합데이터

〈표 5-12〉 **(통신 + 카드 결합 데이터)** 1인당 이동량에 따른 분기 총 항공사 카드지출

(대상 SK텔레콤, 신한 동시가입자 단위: 1인/원)

이동량	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
1분위(적음)	22,174	13,565	5,665	6,964	5,998	6,577	8,651
2분위(중간)	18,955	11,672	4,620	5,342	4,412	4,820	6,917
3분위(많음)	16,577	10,246	3,312	3,762	3,080	3,198	4,935

자료: SK텔레콤, 신한 결합데이터

〈표 5-13〉 **(통신 + 카드 결합 데이터)** 1인당 이동량에 따른 분기 총 관광숙박업 카드지출

(자료: SK텔레콤, 신한 결합데이터, 대상 SK텔레콤, 신한 동시가입자 단위: 원)

이동량	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
1분위(적음)	17,606	12,629	10,565	14,911	14,117	12,781	16,983
2분위(중간)	12,909	8,988	7,577	12,075	10,035	8,898	12,447
3분위(많음)	8,447	5,614	4,664	7,169	5,917	5,123	7,755

자료: SK텔레콤, 신한 결합데이터

〈표 5-14〉 **(결합 데이터 활용 후)** 이동량과 주요 관광업종 카드지출 상관관계 분석

(대상 SK텔레콤, 신한 동시가입자 단위: 피어슨상관계수)

구분	이동량	지출액
이동량	1.0000	
지출액	0.0065*	1.0000

여행사

구분	이동량	지출액
이동량	1.0000	
지출액	0.0370*	1.0000

면세점

구분	이동량	지출액
이동량		
지출액	0.0182*	

항공사

구분	이동량	지출액
이동량	1.0000	
지출액	0.0333*	1.0000

관광숙박업

참고: * 유의수준 10%

2. 유효표본 수 확대에 따른 현황 분석

개인정보를 가명처리한 데이터를 활용하는 경우 민간 데이터를 개인단위 원시데이터로 확보할 수 있게 됨에 따라 유효표본수가 확대되어 개인의 다양한 특성을 고려한 분석이 '세부특성별 분석'아 가능해짐을 확인한 바 있다. 이러한 강점은 결합 데이터에서도 가능해 진다. 두 개의 대규모 표본을 가진 데이터를 결합하는 경우 동시 가입자만 분석 가능하여 기존 보다는 비교적 표본 수가 줄어들지만 세부특성별 분석을 진행하는데 충분한 표본수를 확보할 수 있다. 다음 〈표 5-15〉은 통계청의 인구총조사 인구 현황과 문체부 승인통계 국민여행조사와 SK텔레콤과 신한카드 결합데이터의 표본수를 비교한 결과이다. 기존 신한카드 가입자 850만명 대비하여 350만명으로 전체 표본이 줄어들었으나, 여전히 많은 규모의 표본 확보가 가능하다는 것을 알 수 있다.

〈표 5-15〉 데이터 결합 후 유효표본 수 확대에 따른 확보 가능한 개인 특성별 유효표본 수 비교

(단위: 천명)

2020 인구총조사

전체 표본 수	50,710
남성	25,251
여성	25,459
MZ세대	18,855
그외세대	28,988

sk, 신한 결합데이터

전체 표본 수	3,449
남성	1,655
여성	1,794
MZ세대	1,497
그외세대	1,952
1인 가구	1,113
2인 가구	740
3인 이상 가구	1,049
남성X1인가구XMZ세대	324
남성X1인가구X그외세대	298
여성X1인가구XMZ세대	269
여성X1인가구X그외세대	223

2020 국민여행조사

전체 표본 수	51
남성	25
여성	25
MZ세대	19
그외세대	32
1인 가구	7
2인 가구	13
3인 이상 가구	31
남성X1인가구XMZ세대	2
남성X1인가구X그외세대	2
여성X1인가구XMZ세대	1
여성X1인가구X그외세대	3

가. MZ세대의 세부특성별, 관광이동량별 관광 소비지출 분석

이러한 강점을 바탕으로 본 연구에서는 관광 이동량별, 개인 세부특성별 관광소비지출 차이를 비교분석하였다. 우선 MZ세대의 경우 총 카드지출액은 다른 세대에 비해 낮지만 전체 소비 중 관광소비지출 비중은 다른 세대에 비해 비교적 높은 것으로 나타났다. (참고: 〈표 5-16〉))

〈표 5-16〉 MZ 세대와 그 외 세대 대상, 전체 지출액 중 관광 분기 총 지출액 및 비중

(단위: 1인/만원, %)

구분		2019년	2020년				2021년	
		4분기	1분기	2분기	3분기	4분기	1분기	2분기
MZ세대	전체카드지출	250	235	239	249	252	249	270
	관광키드지출	42 (15.3%)	34 (12.6%)	36 (13.6%)	35 (12.7%)	32 (11.7%)	29 (10.6%)	37 (12.6%)
그 외 세대	전체카드지출	357	328	331	347	342	332	353
	관광키드지출	52 (14.4%)	38 (11.7%)	42 (12.8%)	41 (11.8%)	38 (11.1%)	33 (10.0%)	42 (12.0%)

자료: SK텔레콤, 신한 결합데이터
참고: ()는 전체 지출액 중 관광 소비지출액이 차지하는 비중

이에 MZ세대를 관심대상으로 설정하였으며, 이들의 세부 성별·소득별·이동량에 따른 관광 소비지출을 분석한 결과는 다음 〈표 5-17〉 과 같다. 분석결과를 살펴보면 남성, 연간 소득 5천 이상, 이동량이 많은 경우 전체 소비 중 관광이 차지하는 비중이 상대적으로 높은 것으로 나타났다. 반면 여성, 연간 소득 3천만원 미만의 이동량이 적은 집단에서 관광 소비 지출 비중이 가장 낮은 것이 특징이다. 이처럼 세부특성분석이 가능해짐에 따라 보다 다양한 관광 소비 트렌드 분석이 가능해짐을 확인할 수 있다.

〈표 5-17〉 MZ 세대의 성별, 소득별, 이동량 별 - 관광 소비지출 비중(단위: %)

성별	소득	이동량	2019년	2020년				2021년	
			4분기	1분기	2분기	3분기	4분기	1분기	2분기
남성	3천만원 미만	1분위(적음)	13.3	10.9	11.9	10.5	9.5	8.7	10.4
		2분위(중간)	14.2	12.3	13.2	12.2	11.0	10.0	11.9
		3분위(많음)	14.8	13.5	14.1	13.2	12.2	11.2	13.2
	3천-5천 미만	1분위(적음)	16.5	14.2	15.1	13.6	12.6	11.4	13.2
		2분위(중간)	18.6	16.7	17.0	16.2	15.0	13.7	15.5
		3분위(많음)	19.8	18.2	18.5	17.9	16.6	15.4	17.1

성별	소득	이동량	2019년	2020년				2021년	
			4분기	1분기	2분기	3분기	4분기	1분기	2분기
	5천 이상	1분위(적음)	19.5	17.4	18.1	16.7	15.4	14.6	16.4
		2분위(중간)	22.0	20.4	20.6	19.7	18.1	17.5	18.9
		3분위(많음)	23.0	21.7	21.4	21.1	19.4	18.9	20.2
여성	3천만원 미만	1분위(적음)	10.9	8.6	9.0	8.4	7.4	7.0	8.6
		2분위(중간)	12.0	10.4	10.6	10.2	9.2	8.3	10.0
		3분위(많음)	11.8	10.8	11.3	10.8	9.8	9.6	11.2
	3천-5천 미만	1분위(적음)	12.9	10.4	10.9	10.1	9.0	8.2	9.9
		2분위(중간)	15.3	13.4	13.5	12.9	11.4	10.6	12.4
		3분위(많음)	16.3	14.6	14.5	14.3	13.0	12.4	14.1
	5천 이상	1분위(적음)	15.8	13.5	14.2	13.0	11.8	11.2	13.2
		2분위(중간)	18.1	16.3	16.4	15.8	14.2	13.9	15.6
		3분위(많음)	18.9	17.6	17.3	16.8	15.4	15.2	16.8

나. 시군구별 관광 이동량에 따른 관광소비지출 현황

결합데이터의 대규모 유효표본을 활용할 경우 또 하나의 강점은 시군구 단위의 통계를 생산할 수 있다는 점이다. 이에 통계청 2020년 인구총조사와 SK텔레콤과 신한카드 결합데이터의 시군구 단위의 표본 수를 비교하면 〈표 5-18〉와 같다. 이에 결합 데이터를 활용하면 충분히 소지역 단위의 통계도 생산이 가능함을 확인할 수 있다.

〈표 5-18〉 데이터 결합 후 확보 가능한 지역 단위 유효표본 수 비교

2020 인구총조사(단위: 천명)

순위	행정지역	표본수
1	경기도 화성시	881
2	경기도 부천시	833
3	경기도 남양주시	696
4	제주특별자치시	671
5	서울특별시 송파구	643
...	....	....
246	강원도 양구군	21
247	전라북도 장수군	21
248	인천 옹진군	19
249	경상북도 영양군	16
250	경상북도 울릉군	8

sk, 신한 결합데이터(단위: 명)

순위	행정지역	표본수
1	경기도 화성시	64,487
2	서울특별시 송파구	62,392
3	서울특별시 강남구	59,300
4	서울특별시 강서구	56,797
5	경기도 부천시	54,718
...	....	....
246	전라북도 장수군	427
247	전라북도 진안군	408
248	경상북도 영양군	397
249	경상북도 군위군	394
250	경상북도 울릉군	339

이러한 강점을 실증적으로 확인하기 위하여 표본수가 적은 경상북도 울릉도 군 거주자 대상으로 이동 수준에 따라 구분하면 다음 〈표 5-19〉과 같이 유효표본을 확보할 수 있다. 이를 통해 비교적 분석을 위해 비교적 안정적인 유효표본을 확보 할 수 있음을 알 수 있다,

〈표 5-19〉 경상북도 울릉군 거주자의 이동량 별 유효표본 수

(대상 SK텔레콤, 신한 동시가입자 단위: 명)

구분	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
전체	313	321	327	330	331	331	339
1분위(적음)	87	74	88	89	82	73	78
2분위(중간)	94	87	104	103	118	128	118
3분위(많음)	132	160	136	138	132	130	142

자료: SK텔레콤, 신한 결합데이터

이어서 이를 바탕으로 을릉군 거주자의 이동량별 관광지출액과 같은 통계를 다음과 같이 산출할 수 있게 된다. 다만, 결합데이터 활용 시 충분한 유효표본을 확보하기 위해서는 내부 고객의 정보가 많고 이탈이 적은 민간 데이터일수록 유리하다. 또한 두 기관의 성격이 달라 중복되는 고객이 없을 경우 이러한 장점을 확보하기 어렵다. 따라서 결합 전 충분한 유효표본 확보가 가능한지에 대한 선행적 검토가 필요하다고 하겠다.

〈표 5-20〉 경상북도 울릉군 거주자의 이동량 별 관광소비 지출액

(대상 SK텔레콤, 신한 동시가입자 단위: 1인/만원)

구분	2019년	2020년				2021년	
	4분기	1분기	2분기	3분기	4분기	1분기	2분기
1분위(적음)	56	43	41	39	39	36	40
2분위(중간)	36	42	41	39	35	38	40
3분위(많음)	32	30	26	29	31	21	26

자료: SK텔레콤, 신한 결합데이터

제5절 문화콘텐츠 부문 활용방안

1. 이종 간 결합 정보를 활용한 현황 분석

문화콘텐츠 분야에서는 단일 가명처리 데이터를 기반으로 얼마나 많은 사람들이 콘텐츠를 이용하고 있는지, 콘텐츠 이용 시 얼마나 많은 금액을 소비하고 있는지를 각 분석하였다. 이동통신사의 가명처리 데이터는 '가입자의 콘텐츠의 이용량[52]', '가입자의 이동량'이 주요 변수이고, 신용카드사의 주요변수는 '소비액', '연간소득 등 신용정보'가 해당된다. 두 사업자의 이종 가명처리 데이터를 결합으로 온라인 콘텐츠 소비행태에 대한 심층적인 분석이 가능해졌다. 예를 들어 이동통신사의 가명처리 정보만을 활용한다면 온라인 콘텐츠 이용량, 이용률은 추적 조사가 가능하지만 이용량과 콘텐츠 공급자 입장에서 중요한 요소 중 하나인 소비액 간의 상관성을 확인할 수 없다. 반대로 신용카드사의 가명처리 데이터만 활용할 경우 온라인 콘텐츠 소비액 추이는 파악할 수 있으나 온라인 콘텐츠에 대한 소비를 하더라도 실제로 이용량이 많은지 적은지 여부를 파악할 수 없다. 본 절에서는 이동통신사의 가명처리 데이터와 신용카드사의 가명처리 데이터를 결합해 단일 가명처리 데이터로 확인할 수 없는 현황에 대한 분석을 시도해보고자 한다.

가. 이동량에 따른 콘텐츠(온라인, 오프라인)[53] 소비 추이 분석

본 절에서는 이동통신 가명처리 데이터의 이동량 데이터와 신용카드 가명처리 데이터를 결합한 데이터를 활용하여 온라인 및 오프라인 콘텐츠 소비액 데이터를 활용해 이동

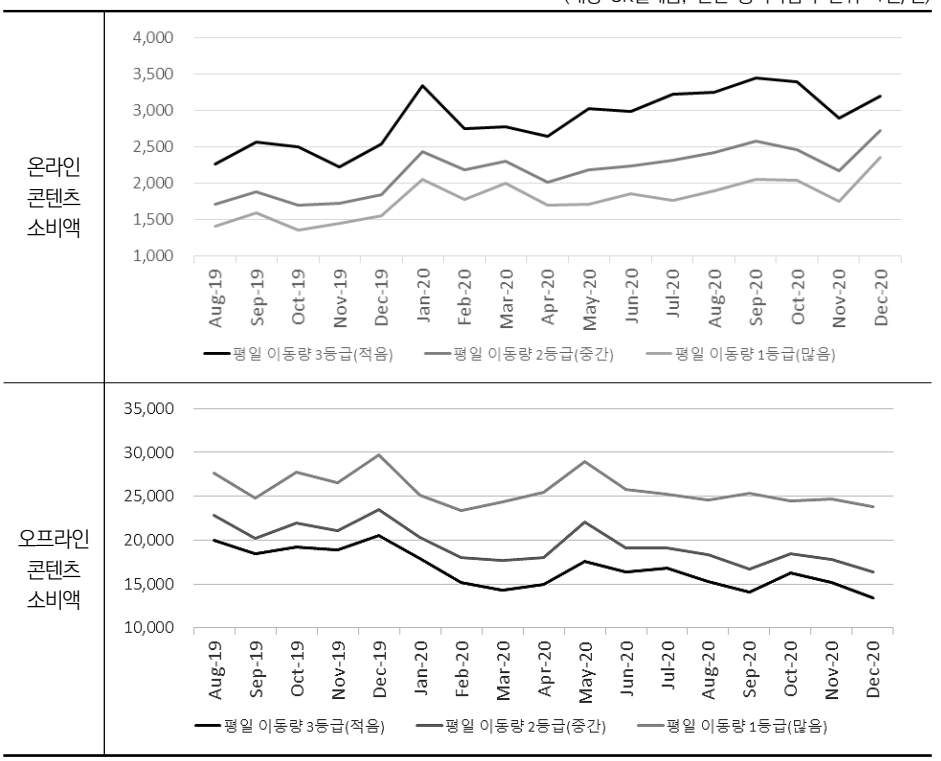
52) 본 연구에서 활용하고 있는 이종결합 가명처리데이터는 결합된 데이터의 개인식별 가능성이 높아진다는 이유로 이를 최소화하기 위해 결합과정에서 많은 정보들에 대해 범주화, 표준화 등의 절차가 진행되었다. 이 과정에서 가명처리 데이터 중 이용량 변수가 이용량의 변화 추이를 확인할 수 없는 형태로 표준화되어 본 연구에서는 이용률을 주요 변수로 활용하게 되었다.

53)여기서 오프라인 콘텐츠소비 정의는 제4장, 〈표 4-7〉, 온라인 콘텐츠 소비 정의는 〈표 4-9〉 참고

량과 콘텐츠 소비액 간 상관성을 파악해 보고자 한다.

[그림 5-6] 평일 이동량에 따른 온라인·오프라인 콘텐츠 소비액 추이(2019.8-2020.12.)

(대상 SK텔레콤, 신한 동시가입자 단위: 1인/원)



자료: SK텔레콤, 신한 결합데이터.

위 그림은 평일 이동량에 따라[54]에 따른 1인 당 온라인·오프라인 콘텐츠 업종 소비액 추이를 보여준다. 검정색 선은 평일에 이동량이 많은 집단이고, 진회색 선은 이동량이 보통 수준에 해당하는 집단이다. 마지막으로 옅은 회색 선은 평일에 이동량이 적은 집단에 해당한다. 본 분석은 이동량이 많을수록 모바일 기기를 활용한 온라인 콘텐츠 이용할 확률이 높아져 관련 소비액이 높을 것이라는 가설을 수립하여 진행하였다. 위 그림을 통해 살펴보면 가설과 다르게 이동량이 낮은 집단 순서로 온라인 콘텐츠 소비액이 높은 것을 확인할 수 있다. 2019년 8월 이동량이 가장 낮은 집단은 평균적으로 월 2,260원 수준의 지출액 추이를 보였으나, 2020년 12월 기준 월 3,100원 이상으로 증가하였다.

54) 여기서 이동량에 수준에 따른 차이를 살펴보기 위하여 이동량 구간은 3분위 배율로 나누어 집단을 구분하여 이동량이 작은 집단은 1분위, 중간 집단은 2분위, 많은 집단은 3분위로 구분

평일에 이동량이 많은 집단은 2019년 8월에 월 1,400원 수준이었으나 2020년 12월에 월 2,400원 수준으로 소비액이 가장 높은 비율(67.6%↑)로 증가하였다. 또한 2019년 연말부터 증가추세에 있었던 온라인 콘텐츠 소비액은 코로나19 기간에도 소비액이 꾸준히 증가하는 추이를 보이고 있는 것을 확인할 수 있다.

다음은 평일 이동량에 따른 오프라인 콘텐츠 소비액 추이를 살펴본다. 오프라인 콘텐츠는 영화관, 노래방, 공연장 등 콘텐츠산업 내 주요 오프라인 업종이 해당된다. 위 그림을 보면 오프라인 콘텐츠 지출액 추이는 온라인 콘텐츠 지출액 추이와 반대의 추세를 나타내는 것을 확인할 수 있다. 오프라인 콘텐츠 업종에 대한 소비액은 평일 이동량이 많은 집단에서 가장 높게 나타났고, 평일 이동량이 낮은 집단의 오프라인 콘텐츠 업종에 대한 소비액 규모가 가장 낮았다. 또한 전반적인 오프라인 콘텐츠에 대한 소비가 감소한 것을 확인할 수 있다. 이러한 현상은 오프라인 콘텐츠 소비를 줄여가고 있는 추이라기보다 코로나19로 인해 소비자의 이동이 제약되고, 오프라인 업종이 분석대상기간동안 '사회적 거리두기'조치에 따라 축소 운영되거나, 영업이 일정기간 정지되었기 때문으로 해석하는 것이 적절해 보인다. 후속 연구에서는 사회적 거리두기 조치 등을 통제한 실증분석 모형을 활용해 이동량과 콘텐츠 소비액 간 순수효과를 추정해 보는 것도 필요하다.

〈표 5-21〉 기간별 평일 이동량 등급에 따른 온라인·오프라인 콘텐츠 소비액 추이 및 증감

(대상 SK텔레콤, 신한 동시가입자 단위: 1인/원)

구 분		2019.08	2019.12	2020.02	2020.06	2020.12	기간증감 ('19.8 - '20.12.)
온라인	1분위(적음)	2,257	2,543	2,749	2,982	3,199	41.7
	2분위(중간)	1,716	1,846	2,187	2,235	2,721	58.6
	3분위(많음)	1,407	1,559	1,770	1,858	2,357	67.6
오프라인	1분위(적음)	19,989	20,512	15,198	16,355	13,387	-33.0
	2분위(중간)	22,799	23,442	17,973	19,135	16,359	-28.2
	3분위(많음)	27,655	29,700	23,392	25,756	23,843	-13.8

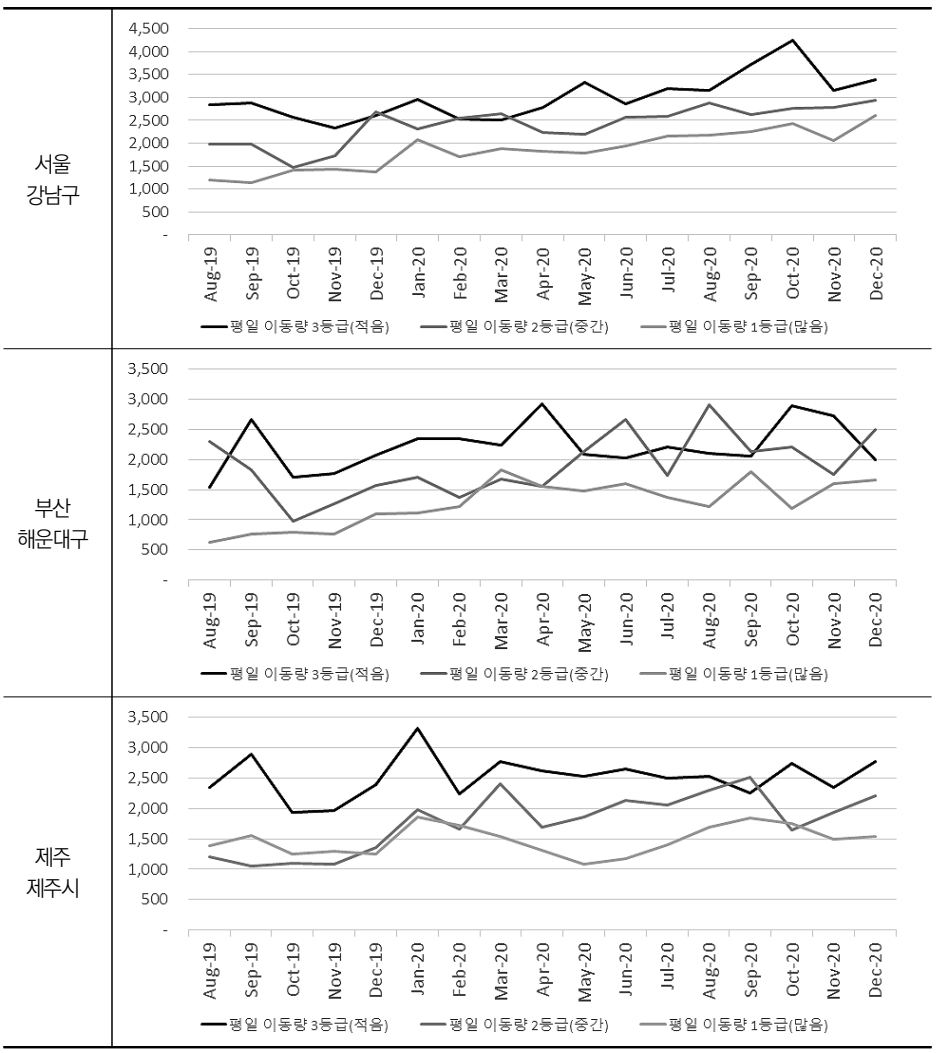
자료: SK텔레콤, 신한 결합데이터

나. 지역별 이동량에 따른 온라인 콘텐츠 소비 추이 분석

이동량에 따른 온라인 콘텐츠 소비에 대한 분석을 시·군·구 단위로 지역의 범위를 좁혀볼 수 있다. 지역구분은 국내에서 평균소득이 가장 높은 서울 강남구, 지방에서 평균소득이 높은 부산 해운대구, 그리고 섬이라는 특징을 가진 제주시로 선정하였고, 해당 지역 거주자를 대상으로 이동량에 따른 온라인 콘텐츠 소비액 추이를 살펴보았다.

[그림 5-7] 지역별 평일 이동량에 따른 온라인 콘텐츠업종 소비액 추이

(대상 SK텔레콤, 신한 동시가입자 단위: 1인/원)



자료: SK텔레콤, 신한 결합데이터

지역별로 구분해 이동량과 온라인 콘텐츠 소비액 추이를 살펴보아도 앞서 살펴본 것과 같이 이동량이 낮은 그룹의 온라인 콘텐츠 소비액 규모가 가장 높은 것을 알 수 있다. 하지만 서울 강남구의 경우 모든 집단의 온라인 콘텐츠 소비액 규모가 지속적으로 상승하는 것을 확인할 수 있는데, 부산 해운대구와 제주 제주시는 소비액 규모가 일정 수준에서 유지되고 있는 것을 확인할 수 있다. 이와 함께 각 지역마다 분석기간 동안 이동량 등급에 따른 소비액 증감에서 차이가 존재했다. 서울 강남구에서는 이동량이 가장 높은 그룹의 온라인 콘텐츠 소비액이 115.8% 증가하였고, 부산 해운대구에서도 이동량이 가장 높은 그룹의 온라인 콘텐츠 소비액이 162.6%로 증가하여 다른 집단대비 월등히 높은 증가율을 보였다. 반면에 제주도의 경우에는 이동량이 보통수준에 해당하는 집단의 온라인 콘텐츠 소비액이 84.2% 증가해 다른 집단 대비 월등히 높은 것을 확인할 수 있었다.

〈표 5-22〉 기간별 평일 이동량 등급에 따른 온라인 콘텐츠 소비액 추이 및 증감

(자료: SK텔레콤, 신한 결합데이터, 대상 SK텔레콤, 신한 동시가입자 단위: 원,%)

구 분		2019.08	2019.12	2020.02	2020.06	2020.12	기간증감 ('19.8 – '20.12.)
서울 강남구	1분위(적음)	2,851	2,601	2,528	2,852	3,393	19.0
	2분위(중간)	1,978	2,691	2,547	2,576	2,935	48.4
	3분위(많음)	1,205	1,374	1,706	1,947	2,600	115.8
부산 해운대	1분위(적음)	1,539	2,077	2,345	2,030	1,997	29.8
	2분위(중간)	2,298	1,575	1,365	2,668	2,500	8.8
	3분위(많음)	1,391	1,245	1,720	1,171	1,534	10.3
제주 제주시	1분위(적음)	2,347	2,388	2,233	2,656	2,777	18.3
	2분위(중간)	1,199	1,353	1,662	2,137	2,208	84.2
	3분위(많음)	2,347	2,388	2,233	2,656	2,777	18.3

2. 유효표본 수 확대에 따른 현황 분석

이종 결합 데이터를 결합 활용하는 것은 독립적인 데이터 셋에서 활용하지 못하는 꼭 필요한 변수를 이용할 수 있다는 점에서 장점을 가진다. 하지만 이종 결합 데이터 간 결합은 사용가능한 데이터의 규모가 축소된다는 단점도 가지고 있다. 본 연구에서는 이동통신 가입자 약 2,400만 명에 대한 개인단위 가명처리 데이터와 신용카드 가입자 약

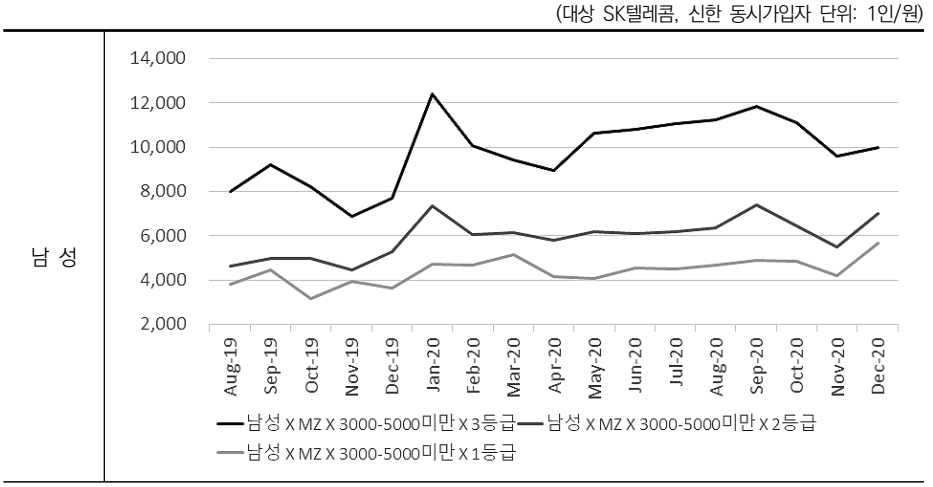
840만 명에 대한 개인단위 가명처리 데이터를 결합하였다. 이종 가명처리데이터 간 결합으로 이동통신가입자의 온라인 콘텐츠 이용정보와 신용카드사의 온라인 콘텐츠 소비액과 같은 정보를 동시에 변수로 활용할 수 있게 되었으나, 두 가지 서비스를 공통적으로 이용하고 있는 사람이 많지 않기 때문에 이동통신사 데이터 기준으로 85.8%가 감소하였고, 신용카드사 데이터 기준으로 59.2%가 감소하였다. 이렇게 결합 후 데이터 규모가 크게 감소했음에도 불구하고 본 연구에서는 349.9만 명의 가명정보 결합 데이터를 확보할 수 있었다. 본 절에서는 이 데이터를 활용해 이동통신 가명처리 데이터와 신용카드 가명처리 데이터로 세분화된 집단의 온라인 콘텐츠 소비액 추이를 살펴보고자 한다.

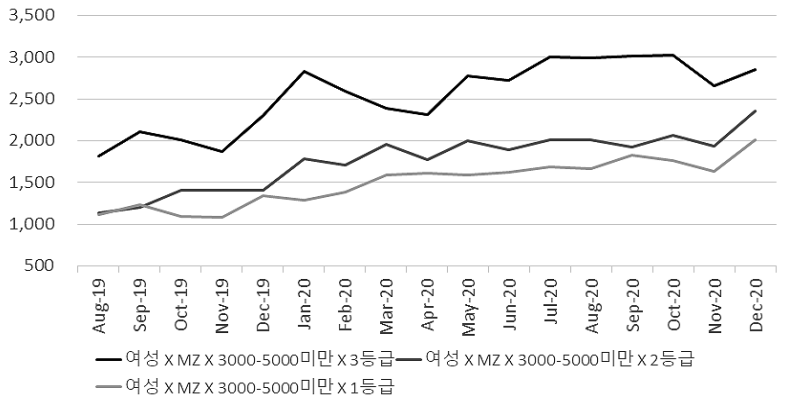
가. MZ세대의 세부특성별, 이동량별 온라인 콘텐츠 소비지출 분석

먼저 MZ세대 남성에 대해 이동통신 가명처리 데이터와 신용카드 가명처리 데이터를 활용해 조금 더 세분화하여 이들의 온라인 콘텐츠 소비액을 살펴본다. 여기에 사용된 정보는 신용카드 가명처리 데이터에 포함되어 있는 연간소득과 온라인 콘텐츠 소비액과 이동통신 가명처리 데이터에 포함되어 있는 이동량 등급이다. 이러한 정보를 활용해 '연소득 3천만원 이상 5천만원 미만 구간에 속한 MZ세대 남성'을 이동량 등급으로 구분하고, 이들의 온라인 콘텐츠 소비액 규모와 추이를 살펴보았다.

[그림 5-8] '연소득 3천만원 이상 5천만원 미만 MZ세대'의 평일 이동량 등급에 따른 온라인 콘텐츠 소비액 추이

(대상 SK텔레콤, 신한 동시가입자 단위: 1인/원)



여 성	

자료: SK텔레콤, 신한카드 결합데이터.

이 조건에 해당하는 남성의 온라인 콘텐츠 소비액은 이동량이 가장 낮은 집단의 1인당 온라인 콘텐츠 소비액 규모가 다른 두 집단 대비 월등히 높았다. 이들은 2019년 8월 당시 월평균 약 8천원 규모로 온라인 콘텐츠를 소비하였고, 2020년 12월 기준 월평균 약 1만원 규모의 소비를 하는 것으로 나타나 분석기간 동안 약 25%의 소비액 규모 증가를 나타냈다. 보통수준의 이동량을 보였던 해당 집단의 남성들은 2019년 8월에 월평균 약 4.6천원 수준의 온라인 콘텐츠 소비액 규모를 보였는데, 2020년 12월 월평균 약 7천원 규모로 약 52% 증가해 모든 집단에서 가장 높은 증가율을 나타냈다. 가장 높은 수준의 이동량을 보였던 집단의 남성들은 2019년 8월에 월평균 약 3.8천원 수준의 온라인 콘텐츠 소비액 규모를 보였는데, 2020년 12월 월평균 약 5.7천원 규모로 약 49% 증가한 수치를 나타냈다. 이처럼 동일한 집단에 대해 추가적인 세부정보로 집단을 세분화하면 집단 간 차이점을 확인할 수 있다.

다음으로 '연소득 3천만원 이상 5천만원 미만 구간에 속한 MZ세대 여성'을 이동량 등급으로 구분하고, 이들의 온라인 콘텐츠 소비액 규모와 추이를 살펴보았다. 이 조건에 해당하는 여성의 온라인 콘텐츠 소비액은 이동량이 가장 낮은 집단의 1인당 온라인 콘텐츠 소비액 규모가 다른 두 집단 대비 월등히 높았고 다른 두 집단의 온라인 콘텐츠 소비액 규모는 격차가 작았다. 이동량이 가장 낮은 집단은 2019년 8월 당시 월평균 약 1.8천원 규모로 온라인 콘텐츠를 소비하였고, 2020년 12월 기준 월평균 약 2.8천원 규모의 소비를 하는 것으로 나타나 분석기간 동안 약 57.1%의 소비액 규모 증가를 나타냈다. 보통수준의 이동량을 보였던 해당 집단의 여성들은 2019년 8월에 월평균 약 1.1천원

수준의 온라인 콘텐츠 소비액 규모를 보였는데, 2020년 12월 월평균 약 2.4천원 규모로 약 106.6% 증가해 모든 집단에서 가장 높은 증가율을 나타냈다. 가장 높은 수준의 이동량을 보였던 집단의 여성들은 2019년 8월에 월평균 약 1.1천원 수준의 온라인 콘텐츠 소비액 규모를 보였는데, 2020년 12월 월평균 약 2천원 규모로 약 80% 증가한 수치를 나타냈다. '연소득 3천만원 이상 5천만원 미만 구간에 속한 MZ세대 여성'들은 이동량으로 구분한 집단 모두에서 1인당 온라인 콘텐츠 소비액 규모가 증가했다. 또한 앞에서 분석했던 동일기준의 남성대비 1인당 온라인 콘텐츠 소비액 규모가 낮으나, 소비액 증가율은 월등히 높다는 점을 확인할 수 있었다.

〈표 5-23〉 연소득 3천만원 이상 5천만원 미만 MZ세대'의 평일 이동량 등급에 따른 온라인 콘텐츠 소비액 추이 및 증감

(대상 SK텔레콤, 신한 동시가입자 단위: 원,%)

구 분		2019.08	2019.12	2020.02	2020.06	2020.12	기간증감 ('19.8 – '20.12.)
남성	1분위(적음)	8,017	7,718	10,083	10,810	10,004	24.8
	2분위(중간)	4,616	5,283	6,050	6,096	7,015	52.0
	3분위(많음)	3,807	3,658	4,666	4,530	5,676	49.1
여성	1분위(적음)	1,818	2,301	2,594	2,726	2,855	57.1
	2분위(중간)	1,139	1,406	1,710	1,896	2,354	106.6
	3분위(많음)	1,116	1,337	1,389	1,627	2,009	80.0

자료: SK텔레콤, 신한 결합데이터

나. 시군구 단위의 이동량별 온라인 콘텐츠 소비지출 분석

다음으로 앞에서 살펴본 연소득 3천만원 이상 5천만원 미만의 MZ세대의 평일 이동량 등급에 따른 온라인 콘텐츠 소비액을 앞에서 구분한 서울 강남구, 부산 해운대구, 제주 제주시로 구분하여 동일한 집단에서 지역 간 어떤 차이가 있는지 살펴본다. 아래 표는 데이터를 지역, 성별, 소득구간으로 구분한 지역별 연소득 3천만원 이상 5천 만원 미만 MZ세대의 수[55]를 보여준다. 지역마다 성별의 구성 등에서 차이가 있는 것을 확인할 수 있고, 집단을 세분화 하더라도 수천 명의 분석대상을 확보할 수 있다.

55) 이종 서비스의 가명처리 데이터를 결합한 데이터이기 때문에 해석 시 해당 이동통신서비스와 신용카드 서비스를 동시에 사용하고 있는 사람들 중 지역별 연소득 3천만원 이상 5천 만원 미만 MZ세대 수라는 점에 유의해야 한다.

〈표 5-24〉 지역별 '연소득 3천만원 이상 5천만원 미만 MZ세대'의 평일 이동량 등급에 따른 유효표본 수

(자료: SK텔레콤, 신한 결합데이터, 대상 SK텔레콤, 신한 동시가입자 단위: 명)

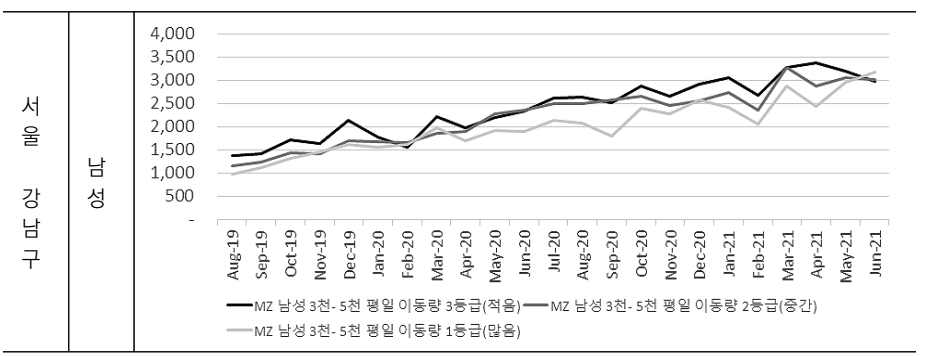
구 분		서울 강남구	부산 해운대구	제주 제주시
남성	1등급	1,031	607	1,036
	2등급	809	712	990
	3등급	706	800	1,651
	계	**2,546**	**2,119**	**3,677**
여성	1등급	2,123	871	987
	2등급	2,081	620	757
	3등급	763	272	562
	계	**4,967**	**1,763**	**2,306**

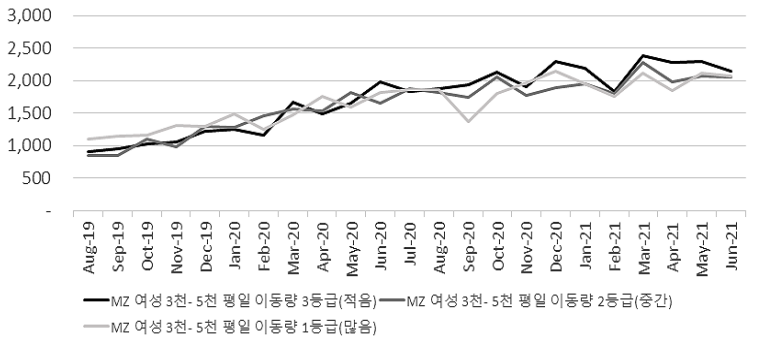
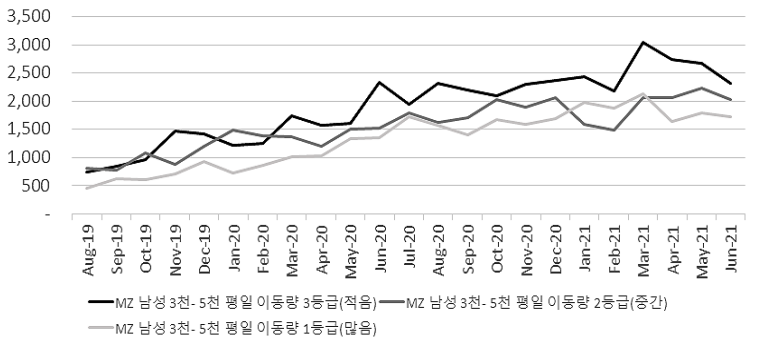
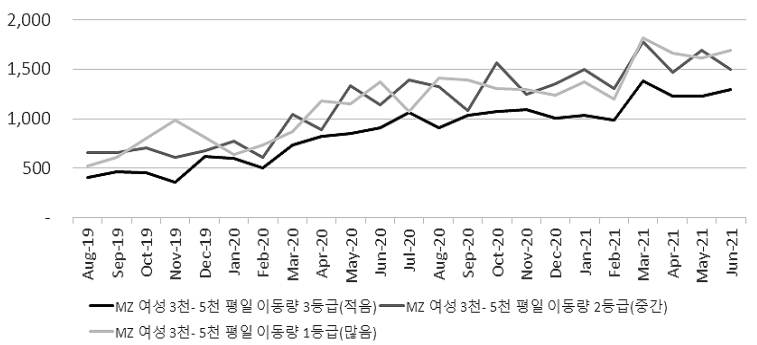
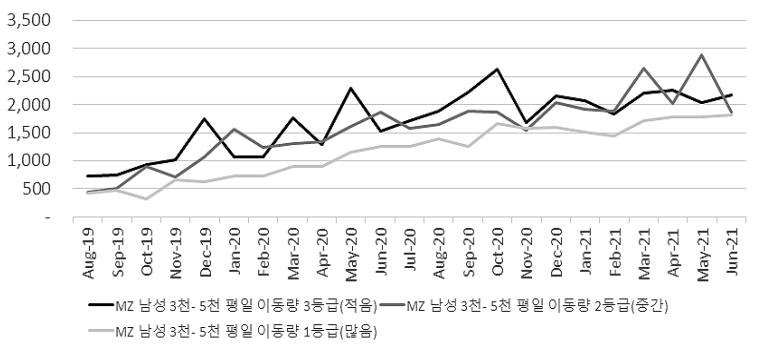
자료: SK텔레콤, 신한 결합데이터

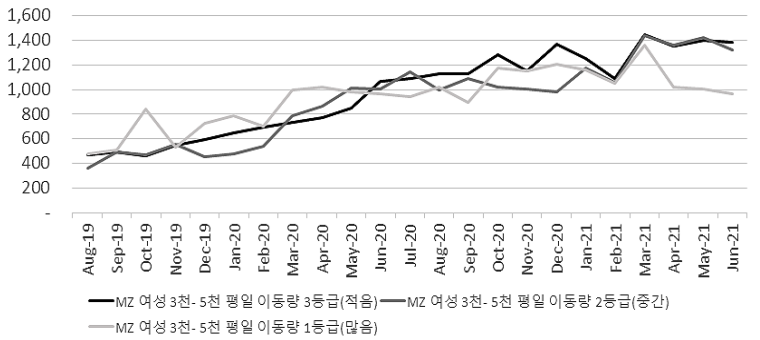
지역별로 해당 소득구간에 속한 MZ세대 남성과 여성의 온라인 콘텐츠 소비액 추이를 분석해보면 아래와 같다. 서울 강남구의 MZ세대 남성은 이동량에 따른 온라인 콘텐츠 소비액이 큰 차이가 없고, 같은 지역 MZ세대 여성도 유사한 패턴을 보였다. 하지만 부산 해운대구 지역의 MZ세대 남성과 여성은 소비 패턴이 달랐다. 이 지역 남성의 경우 서울과 마찬가지로 이동량이 적은 집단의 온라인 콘텐츠 소비액이 높게 나타났으나, 여성의 경우에는 이동량이 적은 집단에서 온라인 콘텐츠 소비액이 낮게 나타났다. 제주도 제주시의 경우에는 서울과 같이 모든 집단에서 이동량이 낮은 집단의 온라인 콘텐츠 소비액이 높게 나타났으나, 집단 간 편차가 남성에서 더 크게 나타났다. 이처럼 거대 가명처리 데이터를 활용하면 집단을 세분화하더라도 유효 표본수가 충분히 많기 때문에 집단 별 특성을 파악하는 분석이 가능하다는 장점이 있다.

[그림 5-9] 지역별 '연소득 3천만원 이상 5천만원 미만 MZ세대'의 평일 이동량 등급에 따른 온라인 콘텐츠 소비액 추이(2019.8-2020.12.)

(대상 SK텔레콤, 신한 동시가입자 단위: 원)



서울 강남구	여성	
부산 해운대	남성	
	여성	
제주 제주시	남성	

제주 제주시	여성	

자료: SK텔레콤, 신한 결합데이터

제6절 소결

5장에서는 가명처리된 통신 및 카드를 결합하여 데이터를 구성, 이를 활용한 실증분석을 통해서 가명처리 데이터 결합 시 필요한 행정절차 및 분석을 위한 데이터 구조 설계에 대한 절차도를 제시하였다. 또한 실제 구성된 결합 데이터 또한 앞선 가명처리 데이터와 마찬가지로 의 정보를 바탕으로 관광과 문화콘텐츠 분야의 분석 가능한 주제를 검토한 후, 실제 몇 가지 주제를 검토하여 데이터 결합을 통해 얻을 수 있는 강점을 실증적으로 살펴보았다. 이에 데이터 결합 구성 절차도는 다음과 같이 도식화 할 수 있다.

〈표 5-25〉 데이터 결합 업무 절차도

구분	주체
1단계: 데이터 결합 관련 업무 협의 데이터의 안전한 활용을 위해 제공기관과 이용기관 간, 협약체결	제공기관-이용기관
⇩	
2단계: 가명정보 활용 내부관리 계획 및 분석 환경 점검 가명정보 활용 관련 내부관리계획 수립 및 점검 가명정보 활용 분석 환경 점검	이용기관
⇩	
3단계: 데이터 결합 신청 가명정보 관리 환경 및 이행 확약서 작성(이용기관) 결합 대상 데이터 작성(제공기관) 데이터 결합 적정성 평가(결합기관)	제공기관, 이용기관, 결합기관
⇩	
4단계: 데이터 결합 및 적정성 평가 적적성 평가위원회 개최, 데이터 결합 및 이용에 대한 적정성 평가(결합기관)	제공기관, 이용기관, 결합기관
⇩	
5단계: 데이터 반출 및 문석	이용기관

이처럼 가명처리 데이터를 결합해서 사용하기 위해서는 가명처리 단계에서의 절차 외 추가적인 행정절차와 가명정보 활용을 위한 환경 구성이 필요하다. 또한 데이터 결합 또한 가명처리된 데이터를 활용하기 때문에 데이터 구성 및 가명처리 단계에서 데이터 값의 손실(총계화, 범주화 등)이 불가피하게 발생하여 분석에 한계가 존재한다. 그 외에도 데이터 결합 후 데이터 클리닝을 통해 데이터 값을 수정하는 작업이 불가능하기 때문에 결합 데이터에서 발견되는 논리적 오류들을 분석 시 어떻게 처리할 것인지에 대한 연구자의 판단이 필요하게 된다.

이러한 한계점이 존재하지만 데이터 결합을 통해 기존 단일 데이터로 파악할 수 없는 다양한 분석이 가능하다는 점과 내부결합(Inner Join)을 하더라도 충분히 기존보다 많은 표본수를 확보할 가능성이 있다는 것을 확인할 수 있었다. 이에 데이터 결합 시 발생할 수 있는 데이터 손실 및 신뢰성 문제, 업무 효율성 개선 등이 이루어진다면 향후 결합된 데이터의 정책적 활용가치가 증가할 것으로 판단된다.

제6장

결론 및 시사점

제1절 결론

본 연구는 데이터 기반 행정의 제도적 기반 마련 및 데이터 3법(데이터 규제 완화 3법)[56]의 개정안 통과를 통해 가명정보를 통한 활용 가능한 데이터의 범주가 확대됨에 따라 선제적으로 문화관광분야에서 가명처리데이터 활용에 필요한 일련의 프로세스를 직접 검토하고 분석하는 절차를 제공함으로써 향후 관련분야 데이터 활용 및 후속연구를 위한 기초적 내용을 제공하기 위해 진행되었다. 위와 같은 목적을 달성하기 위해 연구과정을 설계(design), 분석(analysis), 진단(diagnosis)이라는 틀 안에서 구성하였으며 이를 통해 연구목적 달성을 위한 세부사항을 검토하였다.

설계 단계에서는 우선적으로 가명정보 활용가능성에 대해 검토하였으며 크게 두 가지 관점에서 접근하였다. 첫 번째는 단일 가명데이터만을 활용한 방안이며 두 번째는 이종 간 가명데이터를 매칭키를 통해 결합한 후 활용하는 방안이다. 단일 가명데이터 활용을 위해 기존에 한국문화관광연구원과 가명데이터 활용 관련 양해각서(MOU)를 체결한 SK텔레콤과 신한카드 데이터를 협조 받아 분석을 실시하는 방안을 검토하였다. 가명정보 기반의 데이터 활용은 개인정보라는 민감한 부분을 포함하고 있기 때문에 실제적으로 원활한 협조체계 구축이 매우 어려운 상황이다. 본 연구에서도 수차례 데이터 제공기관(SK텔레콤과 신한카드)과의 협의를 통해 데이터 제공절차 및 방법에 대해 논의하였으며 제공기관의 분석팀, 데이터관리팀, 법무팀 등 데이터 제공기관 내 다양한 검토를 거쳐 최종적으로 활용 가능한 데이터를 협조 받았다. 이러한 일련의 절차들을 보고서에 제시함으로써 향후 가명데이터 활용을 위한 기반을 마련하였으며 후속 연구에서는 이를 바탕으로 보다 원활하게 데이터를 활용할 수 있는 기반이 마련될 수 있을 것이다.

두 번째는 각각의 데이터 제공기관으로부터 입수한 SK텔레콤의 통신데이터와 신한카드의 카드데이터를 가명처리 기반의 이종 간 결합을 통해 다차원적으로 활용하는 방안

56) 데이터 이용을 활성화하는 「개인정보 보호법」, 「정보통신망 이용촉진 및 정보보호 등에 관한 법률(약칭: 정보통신망법)」, 「신용정보의 이용 및 보호에 관한 법률(약칭: 신용정보법)」등 3가지 법률을 통칭

을 검토하였다. 통신사와 카드사의 경우 최초 가입 시 제출하는 주민등록번호와 전화번호가 있다. 이러한 가명정보를 주요변수(key variable)로 활용하여 통신가입자와 카드가입자의 정보를 매칭하고 각기 다른 형태의 데이터를 분석 가능한 하나의 데이터 형태로 구성하는 '데이터 연계'를 실제 연구를 통해 수행하였으며 이를 바탕으로 통신사에서 제공하는 컬럼과 카드사에서 제공하는 컬럼을 병합하여 최종적으로 분석에 활용할 수 있는 기반을 마련하였다. 최종적으로 결합을 통해 약 350만명의 개인데이터를 확보하였으며 이를 통해 기존에 단일 데이터로 분석할 수 있는 내용과 차별화된 분석이 가능하다는 결론과 더불어 보다 의미 있는 결과 도출을 위한 다차원적인 분석이 가능하다는 결론을 도출하였다. 다만 아쉬운 점은 설계단계에서 검토한 데이터 제공기관의 컬럼에 대한 부분이다. 통신데이터와 카드데이터에는 본 연구에서 확인하지 못한 수많은 컬럼 등이 데이터로 저장되고 있다. 그러나 너무 방대한 양이기 때문에 최종적으로 제공받는 데이터는 문화와 관광과 관련된 기본적인 컬럼으로 구성되어져 있으며 제공기관의 설정한 이상의 컬럼을 확보하기는 어려운 상황이다. 향후 보다 많은 컬럼의 정보가 확보된다면 설계 단계에서 문화관광분야 분석에 필요한 컬럼을 보다 많이 확보하여 분석의 다양화를 모색할 수 있을 것으로 판단된다.

분석 단계에서는 앞서 설계한 통신사와 카드사의 가명처리 데이터와 가명처리된 데이터를 매칭키로 결합한 결합데이터를 기반으로 문화관광분야 활용컬럼 도출 및 트렌드 분석을 실시하였다. 데이터 활용 및 분석 주제 도출은 기본적으로 가명데이터가 가진 강점을 보여줄 수 있는 차원에서 설정하였으며 이를 통해 기존 데이터에서 보여줄 수 없는 가명데이터 및 결합데이터만의 강점을 부각하였다. 가명처리 데이터의의 강점은 첫째, 빅데이터 기반의 활용 가능한 유효표본 수 확대이다. 기존에 문화관광과 관련된 분석을 위해 다양한 데이터가 검토되고 활용되었지만 전 국민 50%가 가입되어 활용 가능한 통신데이터와 전 국민 20%가 가입하고 활용하고 있는 카드데이터를 능가할만한 규모의 데이터는 없다. 본 연구에서는 이러한 대용량 데이터 기반의 분석을 통해 기존에 낮은 표본수로 인해 분석하지 못했던 부분까지 분석이 가능함을 검토하였으며 실증분석을 통해 향후 활용가능성을 검토하였다. 분석에서는 기존의 조사에서 표본수의 부족으로 분석이 어려웠던 내용들을 검토하고 유효표본수 확대에 따른 분석형태 및 결과를 제시하였다.

둘째, 가명처리 데이터 기반의 개인단위 데이터 활용이다. 기존에 활용된 통신 및 카드데이터의 경우 가명처리가 되지 않았기 때문에 개인정보 활용의 문제 등으로 인해 성,

연령, 지역 등의 인구통계학적 특성을 재설정하여 그룹화된 컬럼을 통한 분석만 가능한 상황이었다. 본 연구에서는 통신데이터 및 카드데이터의 개인정보를 가명처리한 후 개인별 특성데이터를 활용할 수 있는 기반을 검토하고 실제 활용가능성을 검토함으로써 향후 개인화된 데이터 기반의 분석이 가능함을 제시하였다. 가명처리가 안된 데이터의 경우 통신데이터를 통해 전체 값의 이동총량 정보만을 확보하였다면 가명처리 기반의 개인단위 데이터의 경우 개인 특성별 이동총량에 대한 도출 및 개인별 평균 산출 등의 가능성을 분석하였다. 또한 개인별 평균 산출이 가능해짐에 따라 통신가입자의 개인 특성별 평균 이동량 등의 분석도 실시하였다.

셋째, 이종 간 데이터 결합을 통한 활용가능 정보의 확대이다. 데이터 제공기관을 통해 전달받은 단일 데이터가 가지고 있는 활용가치도 크지만, 데이터 결합을 통해 확보할 수 있는 분석결과는 훨씬 다양해진다. 예를 들어 데이터 결합 전에 특정 축제의 성과평가를 위해 축제에 방문한 방문객 수는 통신데이터를 추정하여 활용하고 해당 지역에 소비지출한 데이터는 카드데이터를 통해 추정하여 각각의 상관성에 대한 검토 없이 활용하였다면 데이터를 결합하여 활용할 경우 축제방문객의 방문객 수와 추정과 더불어 동일집단의 소비지출 패턴을 동시에 분석할 수 있는 분석구조가 가능해지기 때문에 활용할 수 있는 정보는 더욱 방대해질 것이다. 이와 같은 가명데이터의 특징을 정확히 보여주기 위해 1차적으로 데이터 제공기관으로부터 확보한 여러 가지 컬럼 가운데 문화관광 분야 분석에 필요한 컬럼을 도출하였으며 이를 기반으로 기존 분석에서는 도출하기 어려웠던 세분화된 분석결과 도출을 시도하였다. 실증분석에서는 데이터 결합을 통해 확보된 통신데이터의 이동량 변수(x)와 카드데이터의 지출액 변수(y)를 기준으로 데이터 변수간의 상관관계 및 회귀모형 설정을 통해 보다 다양한 분석을 실시하였다. 이를 통해 기존 단일 가명처리데이터로는 확인할 수 없는 확대된 분석의 가능성을 검토하였다.

진단 단계에서는 본 연구의 설계단계, 분석단계 등을 종합적으로 검토하여 향후 가명데이터 기반의 연구 수행 시 필요한 주요 내용들을 정리하고 추후 검토 및 보완해야 할 부분들에 대해 제시하였다. 첫 번째 가명데이터 효율적 활용을 위한 가이드라인 제시이며, 두 번째는 가명데이터 원활한 활용을 위한 제도적 개선방안 제시이다. 진단에 대한 구체적인 내용은 6장 제2절부터 제시하였다. 가명데이터를 통해 기존에 수행할 수 없었던 다양한 분석 및 결과 도출이 가능하다는 긍정적인 특징에도 불구하고 아직 해결해야 할 문제들도 분명히 있는 것이 현실이다. 이번 연구를 통해 가명데이터 및 데이터 결합에 대한 현황

및 개선방안 도출을 통해 향후 보다 발전적인 가명처리 데이터 활용방안을 위한 지속적인 관심과 시도가 필요할 것이다.

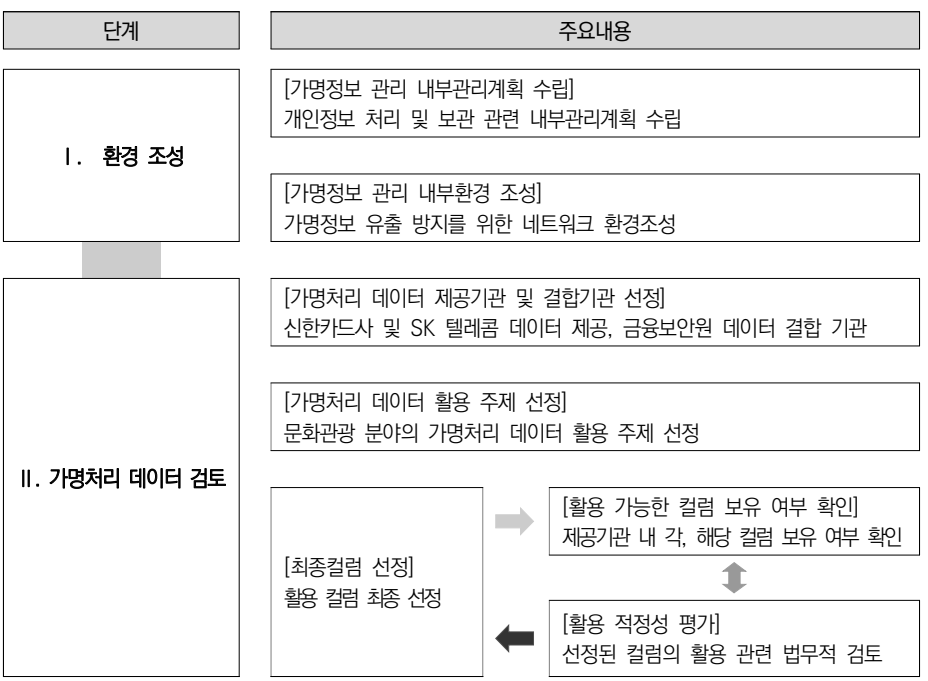
본 연구의 목적에 따른 자료 검토와 실증분석을 통해 도출한 최종 결론은 다음과 같다. 첫째는 '가명처리 데이터를 통한 분석의 다양화'이다. 가명처리 데이터는 개인정보 등의 민감한 사항 때문에 기존에 원활하게 사용할 수 없는 상황이었지만 데이터 활용을 위한 여러 가지 제도개선 등이 수반되면서 활용을 위한 접근이 가능해졌다. 이로 인해 기존의 빅데이터 분석에서 볼 수 없었던 보다 다양한 분석 및 결과 도출이 가능해졌다. 다만 데이터 확보를 위한 여러 가지 절차 및 데이터 제공기관과의 유기적인 업무협의는 해결해야 할 숙제로 남았다. 둘째, '시의성 있는 데이터 분석 가능'이다. 본 연구에서 활용한 데이터의 경우 실시간 데이터를 확보하고 분석할 수 있다는 강점이 있다. 심지에 일 단위를 넘어 시간단위까지도 분석이 가능한 데이터로 활용이 가능하다. 본 연구에서는 데이터 용량에 따른 분석 시간 등을 고려하여 월 단위 개인 집계데이터로 분석을 시도했지만 향후 보다 디테일한 분석을 위해서는 충분히 시의성 있는 데이터 확보를 통한 분석이 가능해질 것이다. 특히나 최근 코로나와 같은 사회·환경적 변화를 보다 실시간 파악하고 대응하기 위해서는 향후 시의성 있는 데이터의 활용성을 보다 극대화될 것이라 판단된다. 셋째, '데이터 결합을 통한 확장성'이다. 기존에 물리적 결합이 어려웠던 이종 간의 데이터를 결합하여 사용할 경우 분석을 위한 확장성을 엄청나게 커진다. 또한 한 개의 데이터로 설명하지 못했던 데이터 기반의 분석결과를 데이터 결합을 통해 보완함으로써 결과의 활용 및 확장성은 계속해서 커질 것이다. 본 연구에서는 통신데이터와 카드데이터의 결합을 통해 이종 간 데이터 결합 및 분석의 가능성을 제시하였지만 향후 문화관광분야에서 유의미하게 사용될 수 있는 공공데이터(기상데이터 등) 결합을 통한 활용방안을 모색한다면 다차원적인 분석 및 정책수립을 위한 시사점 도출의 기초자료로 활용될 수 있을 것이다. 이번 연구를 통한 시도가 향후 효율적인 데이터 결합 및 원활한 분석을 위한 기초자료로 활용되기를 기대해본다.

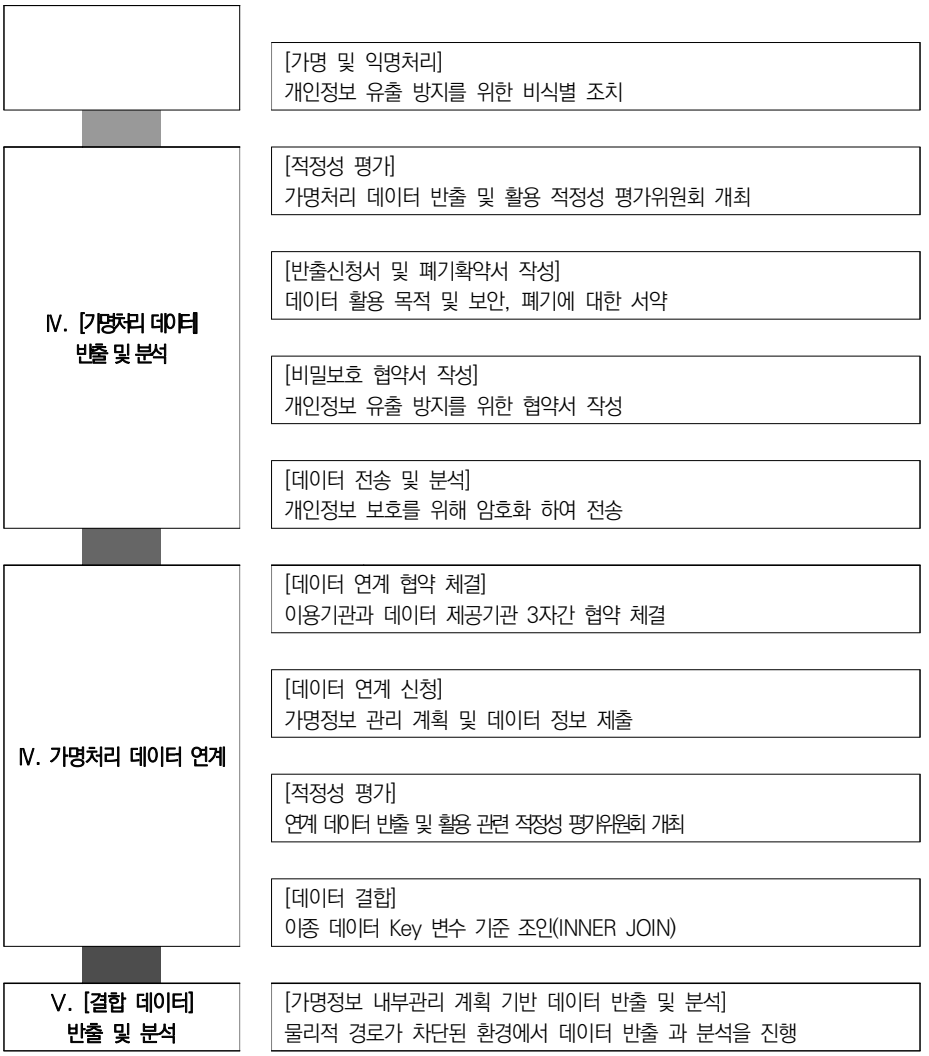
제2절 진단(diagnosis)

1. 가이드라인

본 연구는 기존에 시도되지 않았던 가명처리 데이터를 활용하여 문화관광분야 적용 가능성을 검토한 연구이다. 가명처리 데이터를 활용하기 위해 단계별 검토를 통한 활용 가능성 및 실증분석 결과를 도출하였다. 데이터 활용에 대한 전반적인 내용을 정리하여 공유하는 것을 통해 향후 관련 데이터 활용에 있어서 동일한 오류를 범하지 않고 보다 효율적으로 연구할 수 있는 기반마련이 가능해질 것이다. 이에 본 장에서는 가명처리 데이터의 활용과 가명처리 데이터를 통한 데이터 연계까지 연구를 통해 정리하고 검토한 내용을 제시하고자 한다.

[그림 6-1] 가이드라인





가. 가명처리 데이터 분석환경 조성

가명처리 데이터를 활용하기 위해서는 가명처리 데이터에 대한 기본적인 분석환경 조성이 필요하다. 분석환경 조성은 크게 가명정보 관리 내부관리계획 수립과 가명정보 관리 내부 환경 조성이 필요하다. 가명처리 데이터를 활용하고자 하는 기관에서는 가명정보를 관리하기 위한 사전 절차 마련이 필요하며 이를 문서화하여 보관하고 있어야 한다. 연구원에서는 별도의 가명정보 관리계획이 수립되어 있지 않기 때문에 개인정보 관리 내부계획 내에 가명정보 관리 조항을 신설하여 계획수립을 대체하였다.

또한 개인정보 같은 민감한 내용의 데이터를 다루기 때문에 데이터 분석기관에서는 기본적으로 가명정리 처리를 위한 물리적 환경마련이 필요하다. 가명처리 데이터 활용접근이 가능한 연구진을 사전에 등록하고 해당 직원 외에는 가명정보 데이터를 분석할 수 없게 된다. 또한 가명처리 데이터를 관련기관에서 내려 받기 위해서는 기관내 방화벽 설정을 통한 네트워크 차단 및 가명정보 저장장소 접근 제한 설정이 필요하다. 본 연구를 수행하기 위해 연구진에서는 정책정보센터 내 정보사업팀의 도움을 받아 방화벽 설정을 통한 데이터 확보방안을 마련하였으며 원내 5층에 별도의 분석실을 지정하여 데이터 분석기간 동안에는 외부인의 출입을 차단하고 분석결과를 도출할 수 있는 환경을 마련하였다. 위에 제시된 두 건의 환경 조성은 가명처리 데이터 활용을 위해 필수적으로 갖추어야할 조건이기 때문에 반드시 사전검토가 필요하다.

나. 가명처리 데이터 검토

가명처리 데이터는 명확한 분석 목적 및 활용 가능성을 기반으로 설정해야 한다. 수많은 민간·공공 데이터가 존재하지만 연구목적에 따라 사용되어지는 데이터의 종류 및 컬럼이 달라지기 때문에 사전에 명확한 데이터 검토가 수반될 필요가 있다. 본 연구에서는 문화관광분야의 데이터 활용도를 위해 기존에 관련분야 연구에 빈번히 사용된 SK텔레콤의 통신데이터와 신한카드의 카드지출 데이터를 선정하였으며 데이터 제공가능성에 대한 수차례 논의를 통해 최종적으로 데이터 활용을 득하였으며 이를 통해 데이터 제공기관은 SK텔레콤과 신한카드이며 데이터 이용기관은 한국문화관광연구원으로 설정되었다.

데이터 제공에 있어서는 SK텔레콤과 신한카드와의 수많은 협의가 있었으며 데이터 보유팀, 데이터분석팀, 제공기관 법무팀 등 데이터와 연관된 부서의 의견 및 결정사항들이 다르기 때문에 어떤 데이터를 어떤 방식으로 제공받을지에 대한 사전 논의에 상당기간 이루어졌다. 특히나 개인정보라는 민감한 데이터를 제공하기 때문에 보다 신중한 접근이 이루어졌다. 최종적으로는 SK텔레콤과 신한카드를 통해 기존에 제공받지 못한 개인화된 데이터 제공받았으며 개인정보를 통해 통신데이터와 카드데이터의 매칭을 통해 이종 간 데이터 결합을 실시하였다. 앞선 연구목적에서도 제시한 바와 같이 본 연구는 가명처리 데이터 자체로의 분석과 가명처리 데이터의 결합을 통한 결합데이터의 분석을 동시에 기획했기 때문에 데이터 결합은 연구진행에 있어 매우 중요한 포인트다. 양사를

통해 제공받은 데이터를 데이터 결합기관인 금융보안원을 통해 결합심사(참고: [부록1] 93p)를 받았으며 최종적으로 결합데이터 활용을 득하였다. 데이터 결합기관은 여러 기관이 존재하지만 본 연구에서 활용된 데이터 가운데 신한카드데이터가 포함되어 있기 때문에 신용정보법 내에서 카드데이터 결합이 가능한 금융보안원이 최종 결합기관이 되었다. 향후 활용데이터 종류에 따라 가명처리 데이터의 결합시 결합기관이 달라질 수 있을 것이다.

다. 가명처리 데이터 구성

본 연구에 활용된 통신데이터와 카드데이터는 수많은 컬럼(통신: 616개, 카드 633개으로 구성되어 있으며 연구의 목적에 따라 선별적으로 컬럼을 정비하여 데이터를 최소화할 필요가 있다. 특히 가명 처리된 빅데이터의 경우 용량이 매우 크기 때문에 분석목적을 기반으로 데이터의 삭제, 축소 등의 작업이 선행적으로 이루어져야 한다. 본 연구에서는 코로나 시대의 문화관광의 행태변화를 분석하기 위해 2019년 8월부터 2021년 6월까지 데이터를 확보하였으며 코로나 전인 2019년부터 최근까지 개인화된 통신 및 카드데이터를 통해 분석 가능한 주제 및 칼럼을 도출하였다. 최종 활용을 위해 도출된 칼럼은 데이터 제공기관 담당자와 협의를 통해 제공가능성 및 제공형태에 대해 논의를 거쳤다. 논의 과정에서 여러 차례 의견 조율이 있었으며 최종적으로는 제공기관 법무팀의 검토를 통해 활용 가능한 컬럼이 제공되었다. 이 과정에서 주목해야할 부분이 있다. 데이터 제공기관에서 최종적으로 제공받는 데이터의 경우 raw data 그대로 제공하는 경우가 있는 반면 개인정보 유출 방지를 위한 비식별화 조치(데이터 리코드)를 통해 제공되는 경우가 있다. 데이터 재설정(리코드)을 통해 제공되는 데이터의 경우 원래 분석 목적을 달성하기 어려운 구조로 구성되어 있기 때문에 이 부분에 대한 명확한 확인이 필요하다. 예를 들어 통신가입자의 온라인콘텐츠 이용량 컬럼을 기준으로 코로나 전후 증가 또는 감소에 대한 분석을 실시할 경우 기본적인 이용량 정보가 필요한데 제공되는 데이터의 경우 온라인콘텐츠 이용량의 평균을 기준으로 리코드된 증감 수치가 제공되었다. 이러한 부분은 사전에 데이터 제공기관과의 면밀한 협의를 통해 필요한 데이터 구성에 대한 협의가 필요할 것이다.

라. 가명처리 데이터 활용검토

데이터 제공기관과 가명처리 데이터 구성에 대한 논의 및 개인정보의 가명처리 절차가 끝나면 가명처리 데이터 반출을 위한 단계에 진입한다. 데이터 반출에 대한 결정은 적정성 평가위원회를 통해 진행되며 반출신청 및 폐기 확약서 작성, 비밀보호 협약서 작성 등의 문서작업을 실시하고 데이터 제공기관이 데이터를 암호화 하여 이용기관, 즉 한국문화관광연구원에서 전송한다. 데이터 이용기관에서는 암호화된 가명처리 데이터를 기반으로 사전에 구축한 연구 설계를 기반으로 연구목적 달성을 위한 분석을 수행할 수 있다.

마. 데이터 연계 및 반출

데이터 제공기관에서 전달받은 가명처리 데이터를 기반으로 이종간 데이터 연계를 통한 다차원적 분석을 수행할 수 있다. 데이터 연계의 경우 데이터 제공기관과 이용기관간의 다자간 데이터 활용 관련 협약을 체결해야 한다. 협약 체결후 데이터 제공기관은 가명처리 데이터 컬럼 및 데이터 정보에 대해 작성하며 이용기관은 가명정보 관리 내부계획서를 작성한다. 최종적으로는 데이터 결합기관인 금융보안원의 적정성평가위원회를 통해 데이터 결합의 심의를 받는다. 본 연구에서는 SK텔레콤과 신한카드, 연구원과의 3자가 데이터 협약을 체결하였으며 금융보안원 전문가들이 포함된 적정성평가에 온라인으로 참석하여 데이터 활용 및 보안과 관련된 질의응답 과정을 통해 최종적으로 데이터 결합 및 활용에 대한 승인을 득하였다. 이로 인해 결합된 데이터를 활용할 수 있는 기반을 구축하였다. 또한 연구원 내부의 분석환경 구축에 대한 설명을 통해 연구원에서 직접 데이터를 내려 받아 분석할 수 있게 되었다. 보안과 관련된 분석환경 구축이 준비되지 않을 경우 직접 금융보안원의 데이터분석센터에 방문하여 분석해야 하는 번거로움이 발생될 수 있기 때문에 적정성평가위원회 개최 시 분석환경 구축에 대한 명확한 어필을 통해 결합데이터를 보다 효율적으로 분석할 수 있는 사전 검토방안 마련이 필요할 것이다.

바. 데이터 분석

최종적으로는 데이터 연계를 통해 반출된 데이터를 기반으로 분석을 수행한다. 앞서 설명한 바와 같이 분석의 경우 데이터 결합기관이 제공하는 데이터분석센터 또는 데이터 이용기관 내부에서 수행할 수 있다. 데이터 분석의 경우 사전에 검토한 컬럼을 기준으로 코로나 전후 문화관광분야 행태분석을 파악하기 위해 시도되었으며 연구의 목적에 부합하게 분석결과에 보다는 분석과정 및 분석 가능성에 대한 검토와 더불어 가명처리 데이터가 가진 강점을 보여줄 수 있는 차원에서 실증분석이 진행되었다. 제공받은 데이터는 특성상 일 단위, 혹은 시간단위까지 구분한 세부적인 분석이 가능하지만 본 연구에서는 개인화된 데이터를 월단위로 재설정하여 분석에 활용하였다. 데이터 분석을 위한 통계패키지는 STATA(Software for Statistics and Data Science)와 파이썬(python)을 활용하였으며 이를 통해 개인화된 데이터와 더불어 이종 간 데이터 결합을 통해 구성된 데이터를 분석하고 시사점을 도출하였다.

2. 개선방안 및 향후과제

가. 개선방안

1) 활용 컬럼의 명확한 정의 검토

빅데이터 및 이를 통한 가명데이터 혹은 결합데이터를 활용하기 위해서는 기본적으로 데이터 도출에 대한 구조적 이해 및 제공되는 컬럼에 대한 명확한 확인이 필요하다. 현재 활용되는 대부분의 빅데이터(통신 또는 카드)의 경우 데이터를 제공하는 기관에서 설정한 기준을 그대로 준용하여 활용하고 있으며 이에 대한 검증도 이루어지지 않고 있다. 이에 보다 효율적인 데이터 활용을 위해서는 제공되는 데이터에 대한 생산 방식에 대한 이해 및 구조화된 컬럼에 대한 명확한 기준마련이 필요하다.

첫 번째 데이터 생산방식에 대한 이해이다. 예를 들어 모바일 데이터의 인구이동 자료를 기준으로 각 개인의 이동동선을 받았다고 한다면, 모바일 데이터는 GPS위치처럼 정확한 위치를 산정하는 것이 아닌 기지국의 위치를 산정하여 위치를 파악하는 것이기 때문에 기지국의 전파 영역 등에 대한 이해가 필요하다. 이를 정확히 파악하지 못하고 분

석한다면 분석에 오류가 발생 할 수 있다. 또한 통신데이터가 기지국에 접속한 이용자를 기지국의 전파영역의 50m×50m라는 P-Cell로 통계적으로 할당을 하는 방식(SK텔레콤 기준)으로 집계함을 고려할 때, 1㎞ 단위를 재택과 외출의 기준으로 하고, 이 거리를 누적하여 일간 이동거리로 추정하는 방식은 다소 모호할 수 있다. 카드데이터의 경우에도 업종에 대한 기준이 명확하지 않을 수 있다. 보통 산업 분류로 활용하는 한국표준산업 분류(KSIC) 혹은 업종별 특수분류 체계를 준용하지 않고 데이터 제공기관에서 설정한 업종분류 코드를 활용하기 때문에 타조사와의 비교가 어려울 뿐만 아니라 도출된 결과를 해석함에 있어서도 문제가 생길 수 있다. 또한 카드결제 시 예약결제, 본사매출 집계 문제 등은 반드시 확인하고 활용데이터에 어떻게 적용할 것인지에 대한 검토가 필요하다. 가명데이터를 통한 데이터 결합에서도 통신가입자의 가입자 연령과 카드가입자의 가입자 연령의 차이가 있기 때문에 이로 인해 발생되는 데이터 누락상황에 대해서도 충분한 검토 후 반영해야 한다. 빅데이터는 방대한 양과 정보로 활용도가 매우 높은 반면 아직까지 일관성 있는 정제된 데이터 도출에는 한계점이 있기 때문에 데이터를 활용할 경우 반드시 데이터의 특성을 알고 해당데이터의 생성원리나 추출 방법에 대한 이해가 수반되는 것이 필수 요건이라 할 수 있다.

두 번째 구조화된 컬럼에 대한 이해이다. 현재 사용되는 대부분의 빅데이터 컬럼의 경우 연구자 및 사업목적에 따라 동일한 데이터를 다른 기준으로 활용되고 있다. 이는 분석의 결과값 비교에 심각한 차이를 발생시킴과 동시에 조작적 정의를 통해 수치가 변화하기 때문에 일관성 있는 결과 도출을 저해한다. 예를 들어 통신데이터를 통한 관광의 정의를 규정할 때 물리적 이동기준을 시군구로 볼 것인가? 혹은 광역단위로 볼 것인가?에 따라 이동량은 크게 달라진다. 마찬가지로 방문지에 대한 시간적 체류 기준을 30분 기준으로 볼것인가? 혹은 1시간 기준으로 볼 것인가?에 따라서도 이동횟수, 이동량 등이 크게 달라진다. 이러한 문제는 컬럼 곳곳에 내포되어 있기 때문에 데이터 활용에 있어서 정확한 기준과 근거를 바탕으로 동일한 기준의 데이터를 생산하고 활용할 수 있는 기반 마련이 필요하다. 이러한 구조화된 컬럼의 기준은 데이터를 활용하는 연구자가 여러 가지 사례 및 기준을 통해 설정하여 비교할 수 있지만 데이터 제공기관에서 가명화 작업에서 적용되는 컬럼에 대해서는 반드시 사전 협의 및 활용도 제고를 위한 검토가 필요하다. 예를 들어 연구자는 통신데이터에서 제공하는 온라인콘텐츠 사용량(접속량)을 로데이터로 확보하여 코로나 전후 온라인콘텐츠 사용량의 변화를 분석하고 싶은데 통신

데이터의 가명데이터 처리과정에서 개인정보의 가명뿐만 아니라 컬럼의 정보까지 재결합되어 제공되는 상황이 발생될 수 있기 때문에 이러한 부분은 가명데이터 결합 전에 반드시 확인하고 연구자가 필요한 데이터를 정확하게 받을 수 있는 협의를 진행해야 할 것이다. 민간에서 raw data형식으로 데이터를 수집하고 결합하여 개인단위 데이터 분석에 활용하는 것은 여러 가지 복합적인 분석이 가능하기 필요한 컬럼 및 컬럼의 구조에 대한 협의를 통해 필요한 데이터를 확보하고 분석할 때 보다 질 높고 다양한 차원의 결과값 산출이 가능할 것이다.

2) 시의성 있는 데이터 활용방안 마련

개인화된 가명처리 데이터의 강점은 다차원적인 활용뿐만 아니라 시의성 있는 데이터 생산 및 활용이다. 기존에 조사를 통해 생산되는 통계의 시간적인 문제를 해결하고 최근 코로나19 상황과 같이 급변하는 환경변화를 실시간 체크하고 분석할 수 있는 강점을 빅데이터는 가지고 있다. 본 연구에서 활용된 통신 및 카드 이용자들이 실시간 이동하며 소비하는 패턴을 데이터로 저장하고 분석할 수 있다는 점은 향후 디지털 기반의 정책수립에 있어서 필수불가결한 원칙이 될 것이다. 현재 하루 기준 통신 데이터 약 280테라바이트, 카드 데이터 약 100테라바이트의 디지털 정보량으로 사물정보, 스트리밍 정보 등 실시간 데이터의 생성, 이동, 처리, 분석 등이 가능해졌기 때문에 시의성 있는 데이터 확보 및 활용을 위한 기반마련이 필요하다.

시의성 있는 실시간 데이터 확보는 단순히 빠른 분석 및 결과 도출을 넘어서 일별, 시간대별 데이터 분석이 가능해지기 때문에 디테일한 분석을 할 수 있다. 예를 들어 시간대별로 확보된 통신데이터의 OTT 활용현황의 컬럼을 통해 통신가입자의 출퇴근 시간의 OTT 활용비율 확보가 가능해질 수 있으며 이를 통해 직장인들의 OTT활용 증대를 위한 전략수립이 가능해질 수 있으며, 카드데이터 확보를 통해 시간대별 관광호텔 지출액 파악을 통해 대략적인 일일 점유율 추정이 가능해질 수도 있다. 이처럼 데이터의 확보 시간을 단축하고 세분화하면 기존에 들여다보지 못했던 새로운 차원의 문화관광 현상을 바라볼 수 있는 시스템을 만들 수 있다. 다만 이러한 시스템을 누가 설계하고 어떠한 목적으로 어떤 주제를 분석할 것인지에 대한 심도 있는 고민이 필요할 것이다.

3) 표본설계를 통한 빅데이터 신뢰성 확보

기존의 빅데이터 연구에서 조사결과의 추정을 위해 통신 및 카드데이터 가입자 정보 및 비율을 통해 보정함으로써 모수추정 값에 대한 신뢰가 항상 문제점으로 대두되었다. 이러한 문제를 해결하고 가명처리된 빅데이터를 보다 효율적으로 활용하기 위해서는 기존의 단순 비례 보정방식이 아닌 모집단을 대상으로 한 정교한 표본설계를 통한 모수추정 방안 마련이 필요하다. 기존 연구에서도 빅데이터의 활용을 위해 표본설계를 정교화하고 모수추정을 위한 보정방안을 제시함으로써, 향후 장기적으로 객관적 품질검증이 가능할 수 있는 방안을 마련하였다[57]. 기존 연구에서 제시한 바와 같이 개인정보를 활용한 빅데이터의 대표성을 확보하기 위해서는 통신 및 신용카드 데이터에서 가입자의 주거지를 도출하여 표본으로 추출하는 방안 마련이 필요하다. 다만 주거지를 판단하는 기준이 단순히 청구지 기준인지 보다 정확하게 기존 데이터를 기반으로 특정기간 동안의 야간 체류지를 기준으로 할 것인지에 대한 판단을 필요하다. 또한 단일 빅데이터 아닌 이종간 데이터 결합을 기반으로 표본설계를 진행할 경우 특정 연령대가 배제되는 상황 등을 고려한 정교한 표본설계가 이루어져야 한다. 예를 들어 통신가입자의 경우 보통 초등학생부터의 연령대 확보가 가능하지만 카드데이터는 성인 이상이 되어야 발급 가능하기 때문에 10대의 정보 확보가 어렵기 때문이다. 최종적으로는 단일데이터 혹은 이종간 결합데이터 기반으로 추출된 표본을 통계청의 인구조사통계 집계구 혹은 조사기 기준으로 지역별, 성별, 연령에 따른 인구구조와 매칭하여 모수를 추정할 수 있다. 하였다. 특히 카드 데이터의 경우, 업종에 대한 분류기준이 모호하여 전체 관광을 포괄할 수 없는 분류체계의 문제가 있었다. 이에 본 연구는 관광분야의 특수분류체계를 활용하여 업종분류를 재정립하고, 세부업종별 지출액 구조를 파악함으로써 대표성을 확보하였다. 표본설계를 통해 샘플링된 데이터를 활용할 경우 기존의 총 가입자 정보에서 상당부분이 삭제될 수 있지만 그럼에도 불구하고 전체 모집단 규모가 워낙 크기 때문에 표본 집단을 활용하더라도 보다 세분화된 분석이 가능해질 것이다.[58]

---

57) 관광분야 빅데이터 활용체계 및 실증분석 연구, 2017, 한국문화관광연구원(권태일·이충희)

58) 데이터 결합을 통한 표본설계를 진행할 경우 SK텔레콤 50%, 신한카드 20-25% 가입자 점유율을 기준으로 약 350만명의 모집단 정보 확보가 가능하며 표본설계를 통한 샘플링을 진행하여 최종 표본으로 활용하더라도 조사데이터에 비해 매우 큰 규모의 분석 데이터 확보가 가능함

4) 가명데이터 활용을 위한 표준화 방안 마련

본 연구에서는 가명데이터의 효율적인 활용을 위한 가이드라인을 제시하였다. 가명데이터 활용 및 데이터 결합과 관련한 첫 시도인 만큼 연구진들이 데이터 확보 및 검토, 집계, 분석, 활용 등에서 발생될 수 있는 일련의 내용들에 대해 정리함으로써 향후 연구자들의 접근을 보다 용이하게 하는데 초점을 맞춰 작성하였다. 다만 해당 가이드라인이 통신데이터와 카드데이터인 민간데이터를 기준으로 작성하다 보니 향후 민간+공공, 공공+공공 간의 가명데이터 및 데이터 결합시 발생될 수 있는 전반전인 내용을 포괄하지 못한 한계점은 있다. 향후에는 보다 다양한 데이터를 기반으로 한 사례연구를 통해 가명데이터 활용을 위한 표준화 방안이 마련되어야 하며 데이터 종류에 따라 산발적으로 데이터 결합을 하는 것 보다는 명확한 주제를 기반으로 한 관광분야, 문화분야, 인구분야, 복지분야 등 분야별 데이터의 표준을 정립하는 것도 필요할 것이다. 또한 raw data라고 해도 실제 데이터를 제공하는 각 기관별로 표준 데이터를 정립하는 방식과 데이터를 산출하는 기준 등이 상이가기 때문에 명확한 기준을 바탕으로 데이터를 산출하고 표준화하여 제공하는 방안에 대한 상호간의 협업이 매우 필요할 것이다.

현재 통계청에서는 공식통계의 경우(예, 국립암센터의 암 생존율 통계 작성) 임용기관에서 주민번호를 제공하고 다른 행정자료의 결합을 요청하면 자료를 작성하여 서비스하고 있는 상황이다. 관광통계의 경우에도 국가공식통계 틀에서 이러한 제도를 이용하거나, 아니면 금년 말부터 통계청에서 빅데이터를 활용한 통계작성을 위한 실험통계 제도를 도입할 예정인바, 이를 적극적으로 활용하여 민간데이터 외에 통계청에서 보관 중인 각종 행정자료와 등록 센서스 자료를 활용한 관광통계를 작성할 것을 검토해야 할 것이다.

나. 향후과제

1) 설문조사데이터의 대체방안 검토

가명처리 데이터의 활용도 강화에 따라 기존 설문조사를 통해 도출 가능한 주요 지표들에 대한 대체가 가능해졌다. 그럼에도 불구하고 선제적으로 대체하지 못하는 이유는 가명처리 데이터를 통해 도출 가능한 컬럼들이 주로 정량적 내용들로만 구성되어 있기 때문이다. 이에 다른 조사에서는 빅데이터와 설문조사를 병행하여 조사목적을 달성하기

위한 방안을 마련하고 현실화하고 있다. 교통연구원의 경우 교통실태조사에 빅데이터와 설문조사를 연계함으로써 정량적인 내용과 정성적인 내용을 결합하여 시너지 효과를 모색하고 있으며 국립공원관리공단의 국립공원탐방객 조사에서도 통신 빅데이터를 기반으로 국립공원탐방객수를 추정하고 탐방객을 대상으로 URL을 통한 모바일조사를 병행함으로써 현장에 방문한 유효표본에 대한 방문행태를 보다 정확하게 파악할 수 있는 방안을 마련하였다.

가명처리 데이터의 활용도가 높아졌다고 해도 아직은 현실적으로 빅데이터가 조사데이터를 완벽하게 대체하기는 어려운 상황이다. 다만 통계청[59] 등에서도 빅데이터를 활용한 통계생산 방안 등을 검토하고 있는 만큼 연구원 차원에서도 이에 대한 선제적 검토를 통해 데이터 확보의 효율화 방안을 모색해야 할 것이다. 관광분야에서도 인구 모수를 모바일 자료로 추정하고, 이를 모수로 소규모 관광 표본조사를 실시한다거나 특정 지점을 방문한 집단을 모바일로 추정하고 해당 집단을 대상으로 관광행태 및 만족도 조사를 실시한다면 보다 다차원적인 결과도출 및 활용이 가능해질 것이다. 다만 기존 빅데이터와 조사통계를 융복합 사례에서 설문조사의 응답률이 3% 정도임을 감안하면 향후 모바일 설문조사의 응답률을 어떻게 하면 높일 수 있을 것인가에 대한 검토도 필요할 것이다.

2) 공공데이터의 활용방안 마련

본 연구에서는 가명데이터 활용을 위해 민간데이터 2종을 활용하였다. 데이터 선정 기준은 크게 두 가지이다. 첫 번째는 문화관광분야에서의 데이터 활용도가 얼마나 높은지가 기준이었으며 두 번째는 가명데이터에 활용에 대한 경험을 통해 데이터 원활한 데이터 제공이 가능한지에 대한 부분이다. 이에 기존에 연구원과 MOU를 통해 가명데이터 처리 경험이 있는 SK텔레콤과 신한카드에서 제공하는 민간데이터를 활용하였으며 이를 통해 가명데이터 활용 및 결합분석에 대한 실증분석을 실시하였다.

본 연구에서는 민간데이터 기반의 연구를 수행하였지만 공공에서 제공하고 있는 많은 데이터 검토를 통해 공공데이터 활용기반 마련을 위한 노력이 필요할 것이다. 공공데이터포털(DATA.GO.KR)에서는 현재 많은 양의 공공데이터를 제공하고 있지만 해당 데이

59) 통계청에서는 SK텔레콤의 빅데이터 등 새로운 통계의 활성화를 위해 「통신 모바일 인구이동량 통계」를 실험적 통계 1호로 서비스('21.9.16)하고 코로나19 발생전후의 통계수요 도출 검토

터를 통해 제공되는 컬럼은 무엇인지? 혹은 해당 데이터를 통해 이종간 데이터의 결합은 가능한지? 등에 대한 종합적인 검토는 이루어지지 않은 상황이다. 공공에서 생산되는 많은 양의 데이터를 어떻게 활용하는지에 따라 보다 다양한 문화관광분야의 정책수립의 기초자료 생산이 가능해질 것이다. 데이터 종류, 데이터 구조, 데이터 컬럼 등 공공데이터가 가지고 있는 종합적인 정보를 정리하고 향후 데이터 활용을 위한 기반을 마련한다면 보다 폭넓은 분석결과 도출이 가능해질 것이다.

3) 빅데이터 패널 구축을 통한 종단 연구 및 예측분석 모델 마련

본 연구의 가장 큰 장점은 연구 초반에서도 언급한 바와 같이 빅데이터를 기반으로 한 가명처리 데이터를 활용할 경우 개인화 데이터 기반의 패널 구축 및 종단분석이 가능해진다. 이는 가명 처리된 개인화 데이터 활용의 가장 큰 장점이며 이를 통해 대상의 시계열적 변화 분석 및 예측까지도 가능해진다. 앞서 제시된 제공된 컬럼의 명확한 구조 파악 및 표본서례 등 선행적으로 검토하고 확인해야 할 부분들이 있지만 빅데이터 패널 구축을 통한 대표성 있고 신뢰성 있는 데이터 도출방안 및 정교화된 모형이 제시가 된다면, 실시간 수요 예측이 가능할 것이다. 다만 빅데이터를 활용한 예측시스템을 구축하기 위해서는 관련 기반투자와 통합적으로 관리할 수 있는 조직이 전제되어야 할 것이다. 한국문화관광연구원의 경우 별도의 데이터분석팀을 신설하여 이러한 데이터 분석을 위한 인적 기반을 구축하고 있기 때문에 원활한 데이터 수급이 가능하다면 종단분석 연구 및 예측연구 등 보다 다차원적인 가명처리데이터 활용방안이 마련될 것이다.

가명데이터 기반의 패널 유지가 용이하게 된다면 인구 특성별 여행의 라이프사이클 등에 대한 시계열예측 가능할 것이며 이를 AI가 예측할 수 있는 학습데이터로 축적하여 빅데이터 분석기법(클러스터링 및 유동인구예측)을 이용한 여행객 규모예측 등이 가능해질 것이다. 또한 유동인구데이터를 딥러닝 모델(Mask R-CNN모델) 또는 t-SNE/UMAP 등으로 예측하여 연령별・성별・시간대별 유동인구데이터의 평균값뿐만 아니라, 유동인구의 갑작스런 증감행태 등이 파악 가능한 표준편차, 중간값, 1/2/3/4분위 이동량 등을 분석하여 정책수립의 기초자료로 활용할 수 있을 것으로 판단된다.

4) 빅데이터 활용 변화에 대한 지속적 검토

현재 문화관광분야에서 활용하고 있는 빅데이터(가명데이터 포함)의 경우 데이터 도출을 위한 기술적 환경변화에 따라 데이터 도출값이 달라질 수 있는 우려사항이 있다. 예를 들어 SK텔레콤의 통신데이터의 경우 데이터 집계를 위한 방식이 기술의 발전에 따라 기존방식에서 보다 업그레이된 방식으로 변화할 경우 기존 데이터와의 시계열적 유지를 통한 데이터 활용이 가능한지 등에 대한 검토가 필요할 것이다. 최근 데이터 집계방식으로 삼각측량, 서브기지국 활용, 5G 변화에 따른 집계 수준 변화, WIFI 데이터 활용 등 다양한 변화를 모색하고 있기 때문에 이러한 기술이 적용될 경우 데이터를 어떤 방식으로 활용할 것인지에 대한 검토도 필요할 것이다. 또한 지속적으로 논의되고 있는 통신3사간의 데이터 결합 등의 변화도 눈여겨볼 필요가 있다. 지금은 데이터 이용기관의 선택에 따라 SK텔레콤과 KT의 데이터를 각각 구매하여 기관별로 다른 기준으로 분석 및 활용하고 있는 실정이다. 통신3사간의 데이터 결합 논의에 따라 향후 이용 가능한 샘플수는 확대될 것으로 판단되나 분석을 위한 컬럼별 기준설정은 재검토가 필요할 것이다. 이러한 변화를 빨리 파악하고 선제적으로 분석설계 및 환경을 구축하는 것도 필요한 시점이다.

5) N종간 데이터 연계방안 검토

가명처리데이터의 핵심은 다양한 이종 데이터의 결합 가능성에 있다. 결합에 의해서 사회현상에 대한 데이터 기반 분석이 시간과 공간 측면에서 획기적으로 확대될 수 있기 때문이다. 본 연구에서는 통신데이터와 카드데이터의 이종간 데이터를 매칭키를 통해 결합하는 방안을 검토 및 실행하였으며 실제 결합된 데이터를 기반으로 결합데이터 분석을 실시하였다. 해당 연구에서는 문화관광분야에 기본적으로 가장 많이 활용되는 데이터를 기반으로 연구를 수행하였지만 향후에는 보다 확대된 차원의 분석 및 시사점 도출을 위해 N종간 데이터 검토도 필요할 것이다. 이를 위해서는 가명처리된 개인화된 결합 가능 데이터들이 어떠한 칼럼들을 지니고 있으며 어떻게 상호 결합될 수 있는지에 대해서 다수의 연구자들과 산업관계자들이 함께 검토하고 필요한 데이터에 대한 확보를 요구할 수 있는 시스템이 구축되어야 할 것이다. 가명처리를 통한 결합데이터의 본격적인 산업화는 이러한 시스템의 구축 여부에 달려 있고, 이러한 시스템 구축을 위해서는 결합데이

터에 대한 다양한 분석과 요구가 발생할 수 있는 생태계가 구축되어야 한다. 가명처리데이터 관련 산업 생태계의 시작을 위해서 일종의 데이터 세트별로 코드북(컬럼에 대한 정보 포함)이 공개되어 다수의 연구자들과 산업관계자들의 원활한 접근을 통해 활용 가능성에 대한 지속적인 검토가 필요하며 이를 통합적으로 컨트롤 할 수 있는 컨트롤타워로써의 연구원의 역할도 생각해봐야 할 시점이다.

6) 가명정보 활용을 위한 통합기반 마련

① 가명정보 처리 전담부서 설치

가명처리 관련 업무 총괄·관리 및 의사결정을 위한 전담부서를 지정하여 가명처리 목적 적합성 검토, 가명처리, 가명처리 적정성 검토, 가명정보취급자에 대한 관리·감독 등을 수행하여야 한다. 기관 내 개인정보보호를 전담하고 있는 부서에서 가명처리 담당자를 지정하거나, 별도의 가명정보 전담부서를 설치할 수 있으나 향후 결합전문기관으로의 역할 확장에 대한 고려를 통해 현재 결합전문기관의 조직 구성[60] 등을 참고하여 전담부서 설치를 검토해야 할 것이다.

② 데이터 통합 관리 기반 마련

데이터의 효율적 관리 및 활용을 위해서는 현재 기관에서 보유중인 데이터 및 향후 확보가 예상되는 데이터에 대한 관리가 체계적으로 이루어져야 한다. 향후 연구원에서 직접 가명정보 자체결합[61] 등을 수행할 경우 각 데이터 제공 기관 등에 파편화되어 있는 정보에 대한 면밀한 분석이 필요하며, 보다 다차원적인 분석을 위해서는 결합대상 데이터의 소재를 파악하기 위해 공공 및 민간의 보유 데이터에 대한 조사 등이 필요할 것이다.

결합전문기관의 경우 가명정보 처리 전담부서가 아니더라도 동 본부 내 통계생산 등을 위한 데이터의 입수·정비 및 관리, 기관 간 정보 교류 등을 위한 부서[62]가 있는 것으로 파악되며 한국문화관광연구원에서도 정책정보센터 내 통계관리팀과 데이터분석팀 등의 데이터 관리 및 활용을 위한 자체기반 마련이 필요할 것이다.

---

60) 통계청(통계데이터허브국 통계데이터기획과), 삼성SDS(데이터결합센터), 건강보험심사평가원(빅데이터실 데이터결합부), 한국보건산업진흥원(미래정책지원본부 보건의료빅데이터단), 한국교육학술정보원(개인정보보호부) 등

61) 원칙적으로 결합전문기관은 자체 결합이 불가하나 공공 결합전문기관은 자체 결합을 허용하고 있음(공공기관의 가명정보 결합 및 반출 등에 관한 고시, 2020.12.2.)

62) 통계청(통계데이터허브국 빅데이터통계과), 건강보험심사평가원(빅데이터실 빅데이터운영부) 등

③ 보호조치가 되어있는 가명처리·분석 환경 구축

기관 내 데이터를 활용한 가명정보 결합을 활성화하고자 한다면 가명정보 처리 및 분석 등을 위한 환경이 마련되어야 한다. '불법스팸 실태연구'의 경우, 기존에 구축된 보호조치가 되어있는 환경에서 수행하였기 때문에 결합 데이터 반출을 승인받아 기관 내에서 상세 분석을 할 수 있었으나, 이러한 경우가 아니라면 각 가명정보를 활용하고자 하는 부서에서 가명처리 및 분석 환경을 별도 구축하기에는 현실적으로 제약사항이 많다. 본 연구를 위해 해당과제 연구진들은 가명처리데이터 분석을 위해 방화벽을 통해 특정 PC에 데이터를 내려 받고 외부 접근이 차단된 회의실에서 분석을 수행하였다.

기관에서 가명처리 및 분석 환경을 구축하고자 할 경우, 「개인정보 보호법」에서 요구하는 가명정보 보호조치 수준이 개인정보 보호조치 수준과 유사하므로 개인정보처리시스템 보호조치 항목에 준하여 환경을 구축하여야 한다. 인터넷 망까지 별도 구축한다면 네트워크 단에서의 보안을 위해 방화벽, 침입차단시스템 등 장비까지 도입 필요하나 공공기관이라면 국가정보자원관리원의 자원을 활용하는 방법을 검토하는 것도 가능하다.

참고로 KISA가 중소기업 및 스타트업 기업의 가명정보 활용을 지원하기 위해 구축하여 운영 중인 '가명정보 기술지원 테스트베드'는 가명정보 관련 기술 테스트 등을 수행할 수 있도록 안전한 분석 공간과 안전 조치된 PC 환경(폐쇄망 운영, 데이터 유출방지 솔루션(DLP, Data Loss Prevention) 설치 등)가명/익명처리 솔루션 등을 지원하고 있기 때문에 향후 KISA와의 긴밀한 협의를 통해 가명처리 분석환경 구축의 기반을 마련해야 할 것이다. 또한 의무대상은 아니나, 시스템 구축 시 개인정보 영향평가를 수행하여 설계단계에서부터 개인정보 침해위험성을 검토하고 개선함으로써 시스템 구축·운영 시 발생할 수 있는 침해위협 최소화 및 효과적인 대응책 마련을 위한 노력이 필요할 것이다

참고문헌

〈연구보고서/단행본〉

- 개인정보보호위원회a(2020), 「가명정보 처리 가이드라인」.
- 개인정보보호위원회b(2020), 「가명정보 결합 추진 현황 및 시범사례」.
- 금융보안원(2021), 가명정보 분석환경 가이드라인.
- 통계청(2019), 영리법인통계 통계정보보고서.
- 한국문화관광연구원(2021), 콘텐츠소비지출동향 9월호.
- 한국인터넷진흥원(2018), 개인정보 비식별 조치 가이드라인.
- 한국정보화진흥원(2019), 개인정보 비식별화에 대한 적정성 자율평가 안내서.

- 권태일·이충희(2017), 관광분야 빅데이터 활용체계 및 실증연구, 한국문화관광연구원.
- 박근화(2018), 문화·체육·관광 데이터 연계를 통한 빅데이터 생산 및 활용방안 연구, 한국문화관광연구원.

- WP29(2014), 「Opinion 05/2014 on Anonymisation Techniques」.
- ENISA(2019), 「Pseudonymisation techniques and best practices」.

〈논문〉

- 심원섭·최승묵·심창섭(2018), 관. 빅데이터 분석의 주요쟁점, 관광연구논총, 30(3) 3-22
- Ark, T. K., Kesselring, S., Hills, B., & McGrail, K. (2019). Population Data BC: Supporting population data science in British Columbia.International Journal of Population Data Science, 4(2).
- Schull, M. J., Azimaee, M., Marra, M., Cartagena, R. G., Vermeulen, M. J., Ho, M. M., & Guttmann, A. (2019). ICES: Data, Discovery, Better Health. International Journal of Population Data Science, 4(2).

• 김상광·김선경(2020). 개인정보 규제수준과 데이터 결합이 빅데이터 활용에 미치는 영향. 기술혁신학회지, 23(2), 305-323.
• 김상광(2020). 개인정보 규제요인과 빅데이터 활용간의 관계에서 가명정보 결합의 매개효과 및 조절효과. 정보화정책, 27(3), 82-111.

〈발표자료〉

• 과학기술정보통신부(2020.11.30.), 관계부처 합동 '제2회 가명정보 결합체계 협의회' 개최, 보도자료.
• 금융보안원(2019.6.3.), '금융분야 데이터 주요 인프라 구축방향', 발표자료.
• 현대카드(2020.5.25.), '현대카드 결제데이터로 살펴본 확 달라진 디지털 콘텐츠 라이프'
• 서울특별시(2019.04.16.), 서울시, 빅데이터로 '통근·통학인구 데이터' 개발, 보도자료.
• 통계청(2020.12.17.), 유동인구 및 주거, 직장인구 서비스 알림, 공지사항.
• SK텔레콤(2020.06.01.), 2020 지하철 리포트 '통신 데이터를 활용한 지하철 이용 패턴 및 코로나 19 영향 분석'.
• 아시아경제(2020.08.07.), 카드정보·통신정보 합쳐 상권 입체분석…'데이터 합종연횡' 가속.
• 아이뉴스24(2021.05.13.), 정부 "미개방 핵심데이터 '사업자등록번호' 공개한다"
• 서울경제(2021.09.22.), [단독] KT·야놀자 '신사업 동맹'···여행·데이터 결합 가속도.
• 제4차산업혁명위원회(2021.07.06.), 가명정보 활용 촉진 대책(안).
• 개인정보보호위원회(2021.05.28.), 국립암센터,'가명정보 결합 시범사례'첫 성과발표.
• 개인정보보호위원회(2021.06.04.), 국립암센터, '가명정보 결합 시범사례' 두번째 성과발표.
• 개인정보보호위원회(2021.06.25.), 불법스팸 실태 분석을 위한 가명정보 결합 시범사례 결과 발표.
• 관계부처합동(2021.07.28.), 가명정보 활용성과 및 확산 방안.

〈웹사이트〉

• Statistics Canada, Social Data Linkage Environment. (SDLE). Retrieved from https://www.statcan.gc.ca/eng/sdle/overview/
• TourAPI3.0 https://api.visitkorea.or.kr/main.do
• 관광지식정보시스템 https://www.tour.go.kr/

- 문화데이터광장 https://www.culture.go.kr/
- 문화빅데이터플랫폼 https://www.bigdata-culture.kr/
- 문화셈터 https://stat.mcst.go.kr/
- 문화예술지식정보시스템 https://policydb.kcti.re.kr/#/
- 통계청 통계데이터센터, 통신모바일 빅데이터로 본 유동인구 지도서비스 https://giraf.SK텔레콤elecom.com/cartoweb/kostat/index.html
- 한국관광 데이터랩 https://datalab.visitkorea.or.kr/
- 통합데이터 포털: https://www.data.go.kr/
- 금융데이터거래소: https://kdx.kr/

〈법률〉

- 「공공데이터의 제공 및 이용 활성화에 관한 법률」
- 「개인정보 보호법」
- 「신용정보의 이용 및 보호에 관한 법률」
- 정보통신망 이용촉진 및 정보보호 등에 관한 법률」

ABSTRACT

A Study on the Utilization of Pseudonymized Data in Culture and Tourism

Taeil Gwon·Jeongyeon Song

This study tries to provide elemental contents for data utilization and follow-up research by providing a direct procedure to review and analyze a series of processes necessary for the use of the pseudonymized data in the field of culture and tourism. It is because the category of data has been expanded by the provision of an institutional basis for data-based administration and the passage of the revision of the Data 3 Act (Data deregulation 3 Act). The research was processed within the framework of design, analysis, and diagnosis so that the details for achieving the research objectives were reviewed.

The final conclusions of this study are as follows. First, it is 'diversification of analysis through the pseudonymized data'. Since the pseudonymized data includes the personal information, it was difficult to access in previous studies. However, once the legal and institutional foundation was established, it could be used for research purposes, enabling various analyses that could not be done in existing data analysis. Various procedures and tasks for securing data and collaborative business consultations with data providers are still remained.

Second, it is 'possible to analyze data in a timely manner'. When using the pseudonymized data, the real-time data can be secured and analyzed in individual units so that it can be analyzed by units of hours rather than days. In particular, the demand for timely appropriate data will be escalated to understand and respond to social and environmental changes such as the COVID-19.

Third, it is 'scalability through data combination'. When using heterogeneous types of data that was previously difficult to combine physically, the scalability for analysis is greatly increased. In addition, the utilization of the combined data results will be further expanded by supplementing the analysis results that were difficult to explain with one data.

This study suggests the possibility of data combination and different types of analyses was presented by communication and card data combination. It will be used as basic data for multidimensional analysis and policy establishment and thus, it will be significantly utilized for public data (weather data, etc.) in the field of culture and tourism in the future.

Keywords

Data, Pseudonymized Data, Culture, Tourism, Data deregulation 3 Act, personal information, Data Linkage

부록

가명처리 데이터 설명서

SK텔레콤 통신데이터 설명서

컬럼 항목명	설명	활용 기간(월 단위)	가명처리	
			1차	2차
성별	M/F	2019. 8 ~ 2021. 6	-	-
연령대	20대 이상, 20대, 30대, 40대, 50대, 60대, 70대 이상	2019. 8 ~ 2021. 6	범주화	-
세대별1	M세대(1981~96년생), Z세대(1997~2010년생), MZ세대(1981~2010년), 그 외 세대	2019. 8 ~ 2021. 6	범주화	-
세대별2	청년층(만 18 ~ 34세), 중년층(만 35세 ~ 49세), 장년층(만 50세~64세), 노년층(만65세 이상)	2019. 8 ~ 2021. 6	범주화	-
거주지 주소	시군구	2019. 8 ~ 2021. 6	범주화	-
직장 주소	시군구	2019. 8 ~ 2021. 6	범주화	-
추정소득	*실제 소득이 아닌 모델링으로 추정한 소득 (1) 1억이상 (2) 7000-1억미만 (3) 5000-7000미만 (4) 4000-5000미만 (5) 3000-4000미만 (6) 2000-3000미만(7) 1000-2000미만 (8) 1000수준	2019. 8 ~ 2021. 6	범주화	-
라이프스테이지	* 연령,성별, 카드이용업종, 금액등을 변수로 모델링 (1) 싱글 (2) 신혼 (3) 영유아 어린이 자녀 가족 (4) 청소년 자녀 가족 (5) 성인자녀 가족 (6) 실버	2019. 8 ~ 2021. 6	범주화	-

컬럼 항목명	설명	활용 기간(월 단위)	가명처리	
			1차	2차
가구원수	*카드이용업종, 건수, 금액 등을 변수로 모델링 (1) 1인 가구 (2) 2인 가구 (3) 3인 가구 (4) 4인 가구 (5)5인 이상 가구	2020. 1 ~ 2021. 6	다자녀(3명 이상) 기준, 5이상 상단 코딩	-
평일총이동횟수	*한달 평일(월-금, 공휴일 제외) 총 이동 횟수(단위: 횟수)	2019. 8 ~ 2021. 6	상위 1% 이하 5단위 구간화, 상위 1% 초과 상단 코딩	
휴일총이동횟수	*한달 휴일(토,일 공휴일 포함)의 총 이동 횟수(단위: 횟수)	2019. 8 ~ 2021. 6	상위 1% 이하 5단위 구간화, 상위 1% 초과 상단 코딩	
구독서비스종류	구독서비스 사용 종류 이력	2019. 8 ~ 2021. 6	9명이하 구독서비스 조합 제외	
구독서비스구독기간	구독 서비스 구독 기간(단위: 개월)	2019. 8 ~ 2021. 6	구독 기간 월평균 값, 9명이하 구독서비스 조합 제외	
평일 전체 온라인 콘텐츠지수	한달 간 평일 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
평일 동영상 온라인 콘텐츠지수	한달 간 평일 동영상 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리

컬럼 항목명	설명	활용 기간(월 단위)	가명처리	
			1차	2차
				반올림
평일 음악 온라인 콘텐츠지수	한달 간 평일 음악 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
평일 게임 온라인 콘텐츠지수	한달 간 평일 게임 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
평일 도서 온라인 콘텐츠지수	한달 간 평일 도서 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
평일 웹툰 온라인 콘텐츠지수	한달 간 평일 웹툰온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
평일 커뮤니티 온라인 콘텐츠지수	한달 간 평일 커뮤니티 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
평일 라디오 온라인 콘텐츠지수	한달 간 평일 라디오 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
평일 SNS 온라인 콘텐츠지수	한달 간 평일 SNS 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
평일 블로그 온라인 콘텐츠지수	한달 간 평일 블로그 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
휴일 전체 온라인 콘텐츠지수	한달 간 휴일 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림

컬럼 항목명	설명	활용 기간(월 단위)	가명처리	
			1차	2차
휴일 동영상 온라인 콘텐츠지수	한달 간 휴일 동영상 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
휴일 음악 온라인 콘텐츠지수	한달 간 휴일 음악 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
휴일 게임 온라인 콘텐츠지수	한달 간 휴일 게임 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
휴일 도서 온라인 콘텐츠지수	한달 간 휴일 도서 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
휴일 웹툰 온라인 콘텐츠지수	한달 간 휴일 웹툰온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
휴일 커뮤니티 온라인 콘텐츠지수	한달 간 휴일 커뮤니티 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
휴일 라디오 온라인 콘텐츠지수	한달 간 휴일 라디오 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
휴일 SNS 온라인 콘텐츠지수	한달 간 휴일 SNS 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림
휴일 블로그 온라인 콘텐츠지수	한달 간 휴일 블로그 온라인 콘텐츠 이용량(단위: Packet)	2019. 8 ~ 2021. 6	LOG10 상용로그 적용	지수화(z-score), 소수점 이하 2자리 반올림

신한카드 지출액데이터 설명서

컬럼 항목명	설명	활용 기간(월 단위)	가명처리	
			1차	2차
성별	M/F	2019. 8 ~ 2021. 6	-	-
연령대	20대 이상, 20대, 30대, 40대, 50대, 60대, 70대 이상	2019. 8 ~ 2021. 6	범주화	-
세대별1	M세대(1981~96년생), Z세대(1997~2010년생), MZ세대(1981~2010년), 그 외 세대	2019. 8 ~ 2021. 6	범주화	-
세대별2	청년층(만 18 ~ 34세), 중년층(만 35세 ~ 49세), 장년층(만 50세~64세), 노년층(만65세 이상)	2019. 8 ~ 2021. 6	범주화	-
거주지 주소	시군구	2019. 8 ~ 2021. 6	범주화	-
직장 주소	시군구	2019. 8 ~ 2021. 6	범주화	-
추정소득	*8개 구간화 (1) 1억이상 (2) 7000-1억미만 (3) 5000-7000미만 (4) 4000-5000미만 (5) 3000-4000미만 (6) 2000-3000미만(7) 1000-2000미만 (8) 1000수준	2019. 8 ~ 2021. 6	범주화	-

컬럼 항목명	설명	활용 기간(월 단위)	가명처리	
			1차	2차
라이프스테이지	* 연령,성별, 카드이용업종, 금액등을 변수로 모델링 (1) 싱글 (2) 신혼 (3) 영유아 어린이 자녀 가족 (4) 청소년 자녀 가족 (5) 성인자녀 가족 (6) 실버	2019. 8 ~ 2021. 6	범주화	-
가구원수	*카드이용업종, 건수, 금액 등을 변수로 모델링 (1) 1인 가구 (2) 2인 가구 (3) 3인 가구 4)4인 가구 (5)5인 이상 가구	2020. 1 ~ 2021. 6	다자녀(3명 이상) 기준, 5이상 상단 코딩	-
취미1순위	*카드이용업종, 건수, 금액 등을 변수로 모델링 (00) 없음 (01) 건강 (02) 게임 (03) 골프 (04) 독서 (05) 레져 (06) 뷰티 (07) 쇼핑 (08) 애완동물 (09) 여행 (10) 영화 (11) 예술 (12) 온라인쇼핑 (13) 육아 (14) 식도락 (15) 자동차 (16) 힐링	2019. 8 ~ 2021. 6	범주화	-
앱구매등급	*카드이용업종, 건수, 금액 등을 변수로 모델링 (0) 미사용, (1) 상위 66%~100% (2) 상위 33%~66%, (3) 상위 0%~33%	2019. 8 ~ 2021. 6	범주화	-
디지털음악이용	*카드이용업종, 건수, 금액 등을 변수로 모델링 (0) 미사용 (1) 상위 66%~100% (2) 상위 33%~66% (3) 상위 0%~33%	2019. 8 ~ 2021. 6	범주화	-
게임이용등급	*카드이용업종, 건수, 금액 등을 변수로 모델링 (0) 미사용 (1) 상위 66%~100% (2) 상위 33%~66% (3) 상위 0%~33%	2019. 8 ~ 2021. 6	범주화	-

컬럼 항목명	설명	활용 기간(월 단위)	가명처리	
			1차	2차
전체 신용카드 지출 금액	전체 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
관광 관련 신용카드 지출 금액	관광 관련 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
콘텐츠 신용카드 지출 금액	콘텐츠 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
문화예술 신용카드 지출 금액	문화예술 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
관광여행사 신용카드 지출 금액	관광여행사 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
특급호텔 신용카드 지출 금액	특급호텔 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
1급호텔 신용카드 지출 금액(	1급호텔 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
2급호텔 신용카드 지출 금액	2급호텔 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
콘도 미니엄 신용카드 지출 금액	콘도 미니엄 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
모텔, 여관, 기타숙박 신용카드 지출금액	모텔, 여관, 기타숙박 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거

컬럼 항목명	설명	활용 기간(월 단위)	가명처리	
			1차	2차
카지노 신용카드 지출금액	카지노 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
면세점 신용카드 지출 금액	면세점 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
항공사 신용카드 지출 금액	항공사 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
실외골프장 신용카드 지출 금액	실외골프장 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
실내골프장 신용카드 지출 금액	실내골프장 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
공연장 및 극장 신용카드 지출 금액	공연장 및 극장 신용카드 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
온라인 영상콘텐츠 신용카드 지출 금액	해당 사업체의 카드 지 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
온라인 음악콘텐츠 신용카드 지출금액	해당 사업체의 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
온라인 게임콘텐츠 신용카드 지출금액	해당 사업체의 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거
온라인 도서콘텐츠 신용카드 지출금액	해당 사업체의 월간 총 지출액 (단위: 원)	2019. 8 ~ 2021. 6	천원단위 반올림	이상치 제거

집필내역

연구책임

권태일 한국문화관광연구원 연구위원: 제1장, 제2장, 제6장 연구총괄

송정연 한국문화관광연구원 차석전문원: 제2장, 제3장, 제4장, 제5장

연구진

김성준 한국문화관광연구원 연구원: 제3장, 제4장, 제5장

문화·관광 분야 가명처리 데이터 활용방안 연구

발 행 인 김 대 관

발 행 처 한국문화관광연구원
서울시 강서구 금낭화로 154
전화 02-2669-9800 팩스 02-2669-9880
http://www.kcti.re.kr

인 쇄 일 2021년 12월 15일

발 행 일 2021년 12월 15일

인 쇄 인 (사)한국장애인이워크협회 일자리사업장

I S B N 978-89-6035-903-1 93300

DOI https://doi.org/10.16937/kcti.rep.2021.e46

이 연구보고서를 인용하실 때는 다음과 같은 사항을 기재해 주십시오.
권태일·송정연(2021), 문화·관광 분야 가명처리 데이터 활용방안 연구, 한국문화관광연구원

한국문화관광연구원
서울특별시 강서구 금낭화로 154
전화 02-2669-9800
팩스 02-2669-9880

www.kcti.re.kr

아래의 DOI 또는 QR코드를 통해
이 보고서를 무료로 다운로드할 수 있습니다.
https://doi.org/10.16937/kcti.rep.2021.e46

93300
9 788960 359031
ISBN 978-89-6035-903-1