빅데이터 활용 통계생산방법론 연구용역
결과보고서

2016. 10. 31.

주관연구기관 한국조사연구학회

빅데이터 활용 통계생산방법론 연구용역
결과보고서

2016. 10. 31.

주관연구기관 한국조사연구학회

제 출 문

통계청장 귀하

본 연구결과보고서를 “빅데이터 활용 통계생산방법론 연구용역”의 최종 연구보고서로 제출합니다.

2016년 10월 31일

주관연구기관 : 한국조사연구학회 (인)

연구 책임자 : 김영원

참여 연구원 : 박민규
유승동
김주영
서우석
이기홍

요 약 문

Ⅰ. 빅데이터 활용 통계생산 방법론

ICT 분야의 발전과 더불어 등장한 『빅데이터』를 활용한 민간분야의 분석 및 그 결과의 활용이 광범위하게 이루어지고 있는 시점에 본 연구에서는 빅데이터를 활용한 공공통계 작성의 가능성에 대한 문제를 살펴보았다. 민간분야 혹은 공공분야에서 고려되고 있는 빅데이터 활용 방안은 크게 데이터 증대(data augmentation)과 빅데이터 분석(big-data analytics)으로 구분할 수 있다. 일반적으로 머신러닝, 인공지능, 신경만 분석 그리고 딥러닝과 같은 빅데이터 분석 방안들이 많이 논의되며 언급되고 있으나 자료 대체(imputation), 자료 연결(data linkage) 그리고 통계적 매칭(statistical matching) 등의 방법들을 통한 데이터 증대 역시 일종의 빅데이터 분석 방안으로 고려될 수 있다. 특별히 공공분야에서 각 기관에서 보유하고 있는 행정자료의 매칭을 통한 보다 다양한 정보를 포함한 대규모의 데이터를 생산하고 이를 이용한 분석을 실시하는 것은 빅데이터 활용 통계 생간을 위한 첫 걸음으로 생각할 수 있다. 본 연구에서는 언급된 빅데이터 활용 방안을 간단히 소개하고 이에 대응하는 국내외 사례들을 살펴보았다. 국외 사례로는 네덜란드의 사례들을 고려하였고 국내 사례들로는 특별히 자료 증대 측면에서 2015년 수행된 인구주택총조사(등록센서스)를 소개하였다.

더불어서 본 연구에서는 공공 분야에서 생산되는 통계나 그 분석결과가 민간분야와는 달리 그 신뢰도에 대한 논의가 매우 중요한 점을 감안하여 이를 검증하기 위한 절차에 대한 논의를 수행하였다. 이는 통계나 분석 결과를 제공하는 국가 혹은 공공기관의 신뢰도가 생산되는 통계의 품질과 직접적으로 연결될 뿐 아니라 정책 결정을 위해 이러한 자료들이 직접적으로 사용되기 때문이다. 이러한 통계 및 분석 결과의 신뢰도를 고려하여 통계청에서는 전통적인 자료수집 방법을 통해 생산되는 통계의 작성을 위해 승인 기관을 지정하고 또한 매우 까다로운 기준을 정하여 이를 만족하는 통계를 국가 승인 통계로 인정하여 공표를 할 수 있도록 하고 있다. 기본적으로 기존 통계의 승인을 위해서 고려되는 요소들은 작성된 통계 혹은 그 결과에 초점을 맞추고 있으며 그러한 각 평가 요소들을 본 연구에서는 검토하여 제공하고 있다.

빅데이터의 품질검증은 기존의 검증방법과는 차별적으로 정의되고 구축되어야 하는데 이는 빅데이터가 가지고 있는 수집 목적, 수집 방법, 자료의 형태 및 사용자의 다양성에 기인한다. 전통적인 자료 수집 방안과는 달리 빅데이터의

경우에는 자료 수집이 이루어지는 단계에 연구자나 사용자가 관여할 수 없기 때문에 직접적으로 자료 수집 과정의 조정을 통한 자료의 신뢰도를 높일 수 없고 따라서 그 과정을 면밀히 검토하고 사용 가능성을 검토하여야 한다. 본 연구에서는 이러한 빅데이터의 특성을 감안하여 빅데이터 품질 검증을 위한 기본적인 틀을 제공하였다. 이를 위해 빅데이터 분석과정을 크게 input, throughput 그리고 output의 세 단계로 구분하고 각 단계별로 평가가 이루어져야 하는 요소들을 정의하고 이에 부합하는 질문들의 예를 제공하였다. 특별히 빅데이터가 가지고 있는 여러 측면에서의 다양성을 감안하여 검증 체계를 작성함에 있어 목적 부합성, 일반성 및 유연성 그리고 효율성을 그 기본원리로 고려하였다.

아직까지는 초보적인 단계에 머물러 있어 빅데이터의 공공분야에서의 활용 및 이를 통한 국가승인통계작성은 가능하지 않으나 서베이를 통한 자료 수집이 어려워지는 현 시점에 빅데이터의 활용을 본격화하기 위해서는 빅데이터 분석 자체의 한계와 문제점을 극복함과 동시에 그 결과에 대한 평가와 품질 검증에 대한 도구 역시 체계적으로 개발되어야 할 것이다. 빅데이터의 활용 방안을 다양화하고 그 결과의 신뢰도를 높이기 위해서는 본 연구에서 제공하고 있는 품질진단에 대한 기본적인 틀을 바탕으로 각 빅데이터 분석 별 검증 체계를 구체화하고 이를 적용하여 사용 목적에 적합한 빅데이터 활용 통계를 생산하는 과정을 구축해 나가야 할 것이다.

Ⅱ. 빅데이터 활용을 위한 제도적 장애요인 검토 및 개선 방안

우리 주변의 사회현상이나 자연현상에 관해 수집한 자료를 이해하기 쉬운 형태로 정리·요약할 뿐 아니라, 이를 분석·해석하여 현상을 기술·설명·예측하는 데 활용되는 통계는 오늘날 정부나 기업과 같은 조직뿐만 아니라 개인에 이르기까지 객관적이고 합리적인 의사결정을 하는 데 중요한 정보를 제공하는 전통적 역할에서 나아가 일상적인 차원에서 효과적인 의사소통수단으로 각광받기에 합당한 것으로 평가되고 있다.

아울러 국가·사회 전반에 걸친 정보화의 물결은 통계와 관련한 새로운 국면을 열어가고 있다. 이미 몇몇 분야에서 실제 구현중인 정보시스템을 통한 통계의 자동작성을 위시하여 특히 '빅데이터(Big Data)' 기술을 필두로 새롭게 부각되는 정보처리기술은 기존 통계영역을 넘어선 새로운 통계의 작성은 물론 기존 통계자료에 대한 새로운 방식의 접근을 가능케 하고 있는 것이다. 그렇지만 이러한 새로운 통계의 생산은 과거에는 문제되지 않았던 새로운 문제들 특히 개인정보보호의 영역과 밀접한 관련을 맺는 문제점들을 초래할 우려가 없지 않다. 그럼에도 불구하고 그동안 이러한 통계에 대한 법적인 의미부여와 법제도적 측면에서의 발전적 방향의 모색은 거의 진행되지 못했던 것도 부인할 수 없는 사실이다.

이에 2장에서는 향후의 본격적인 논의의 기초를 마련한다는 측면에서 빅데이터 환경 하에서 새로운 통계의 생산을 위한 통계법제의 적절성을 검토하는 가운데 아울러 이러한 새로운 통계생산 체제와 현행 개인정보 보호법제와의 조화가능성을 검토해 보았다.

우선 통계제도의 헌법적 근거로는 헌법 제127조 제1항, 제2항 등을 확인하였고, 통계제도의 전반적인 운용양상을 확인하는 가운데 통계작성을 위한 자료조사의 법적 성격을 행정조사의 범주로 규정하여 통계법 및 행정조사기본법의 관련 조항들이 활용될 수 있음을 검토하였다. 아울러 현행 개인정보 보호법제들은 개인정보 보호론 쪽에 무게를 두고 있음에도 통계작성을 위한 근거를 충실히 마련하고 있음을 확인하였지만, 추후 개인정보 보호론의 측면에서 제기될 수 있는 헌법적 문제제기에 사전적으로 대응하는 차원에서라도 헌법상 기본권 침해여부 심사기준이라 할 수 있는 법률유보 및 과잉금지원칙의 준수를 위한 기본적인 틀을 제시하였다.

Ⅲ. 가계부채 관련 현황조사

최근 가계부채에 대한 이슈가 부각됨에 따라 관련 통계정보의 생산, 축적, 활용, 분석 및 개선 혹은 관리 및 기존 통계의 개선 그리고 신규통계의 개발에 대한 관심이 증가하고 있다. 현재 우리나라에서 가계부채 현황파악은 금융안정성 측면에 초점을 두고 있으므로 시장정보에 조사 측면에서 개선의 여지가 있다. 기존 가계부채관련 시장정보는 대부분 금융기관 중심의 가계부채에 대한 정보이므로 우리나라 가계부채를 과소추정하고 있을 가능성이 있다. 가계부채의 규모는 공식적 금융기관이 보유하고 있는 자산(가계의 입장에서 부채)과 더불어 비공식적 부채를 포함하는 광의의 개념으로 볼 수 있다.

미시적인 가계의 관점에서 가계 부채에 대한 정보의 관리가 필요하다. 가계수준의 정보는 다양한 유형의 정책대상자(즉 저소득층, 고령층, 주거복지 수혜대상자 등)에 대한 분석의 한계를 극복할 필요가 있다. 가계부채와 관련된 정책수립과 관련된 그리고 금융정책 수혜대상 계층이 다양하게 존재한다. 금융정책 측면에서 뿐만 아니라 복지정책, 주거정책 등 다양한 분야 그리고

다양한 계층에서 가계부채 관련 정보에 보강의 필요성이 제기되고 있는 상황이다. 이와 같은 상황에서 감독기관에서는 입수가능성을 해결할 수 있지만, 통계작성에 따른 이해관계의 이슈를 적절하게 해결하기 어려울 수 있다. 금융기관 안정성의 관점은 시장자료 분석이 용이할 수 있지만, 가계 안전성과 다양한 정책적 관점에서 미시자료에 대한 분석이 요구되고 있다.

금융관련 정부정책을 평가하기 위하여 (정책성) 공공기관 대출과 일반 시중 금융기관 대출에 대한 분류체계가 요구된다. 기존 통계조사들에 대한 검토에 따르면, 우리나라에서는 가계부채 정보는 공공과 민간의 구분이 모호하다. 따라서 이를 정확히 분류하는 체계의 개선이 요구된다. 공공성격이 있는 금융기관에서 대출을 받은 가계의 경우는 엄밀한 범위에서 금융기관에서 대출을 받았지만, 정부지원으로 민간에서 대출을 받았음으로 이에 대한 새로운 분류체계가 필요하다. 이를 통하여 정부에서 보유하고 있는 금융정책에 대한 수단에 대한 관리와 평가에 기초자료를 확보할 수 있을 것이다.

본 연구에서는 통계청에서 보유하고 있는 가구에 속한 가구원 정보를 활용하여 정부부처에서 제공한 소득정보, 민간기관의 부채정보를 활용하여 가구의 총부채원리금상환부담 통계의 개발을 제안한다. 2015년 12월부터 금융위원회와 은행연합회는 여신심사선진화가이드라인을 시행하고 있다. 가이드라인의 핵심은 차주의 총 금융부채 상환부담을 평가하는 시스템 즉 DSR(Debt Service Ratio)을 도입하는 것이다. 전국은행연합회에 따르면 차주의 기타 금융부채의 원금을 상환하는 부담도 고려하는 총 금융부채 상환부담 평가지표 도입이 필요하여 신규 주택담보대출에 대해 DSR 지표를 통해 차주의 총 금융부채 상환부담을 평가할 수 있다고 한다. 현재 통계청에서는 가구소득을 추정하는 작업을 진행하고 있는 것으로 알려져 있다. 금융권에서는 대출을 이용하는 고객의 증빙소득 혹은 금융기관의 인정소득에 근거한 정보의 제공여부에 따라 일부 소득정보만 보유하고 있을 수 있다. 그러나 통계청에서는 객관적 소득정보를 가구단위로 추정할 수 있는 것이다. 본 연구에서 제안하고 있는 정보는 민간 금융기관이 보유하고 있는 소득정보와 차별적이다. 이들은 (통계청의 가구정보를 활용하지 않는 경우) 차주를 중심으로 소득과 부채정보를 파악할 수밖에 없는 한계가 있다. 통계청에서 작성하고 있는 소득정보는 신규로 대출을 이용하고 있는 금융소비자 뿐만 아니라 기존 대출과 더불어, 부채가 없는 가구의 소득정보도 포함한다. 특히 기존 금융기관의 대출정보는 개인별 정보이다.

통계청에서 생산 가능한 DSR의 경우 다른 기관들에서 생산 가능한 DSR과 세 가지 차별성이 존재한다. 첫 번째로 가계의 재무정보에 대한 추정이다. 기존 소득 혹은 부채정보는 개인 즉 차주중심으로 집계되고 있지만, 통계청의 가구정보를 활용하는 경우 가계부채 정보로 전환할 수 있다. 두 번째로 기존 가계부채 정보는 부채를 보유한 개인에 대한 정보만 존재하지만, 가계정보를 통하여 부채를 보유하고 있는 가계, 부채를 보유하고 있지 않은 가계, 그리고 새롭게 부채를 보유하는 가계에 대한 정보로 구분할 수 있다. 세 번째로 통계청의 경우 객관적 소득정보를 활용할 수 있다.

시장요구의 증가로 최근에는 상대적으로 정확하게 파악하고 있으므로, 가계부채 관련 정보의 축적 혹은 보완은 장기적인 호흡으로 준비하는 것이 필요하다. 동시에 가계부채 관련 신규 통계의 개발이 이루어지는 경우 단기적으로 완전한 시장정보를 보다는 중장기적으로 완전한 시장정보를 구축하는 방향이 바람직하다. 기존 사례에 대한 분석을 통하여 가계부채 관련 통계를 개발하는 경우 중장기적인 보완과 개선을 전제로 작업을 추진하는 것이 바람직하다.

Ⅳ. 빅데이터 활용 국내 사례

국내에서도 빅데이터를 활용하기 위한 움직임은 공공/민간 모든 부문에서 활발하다. 공공분야에서는 교통 빅데이터 사용이 가장 두드러진다. 교통분야 빅데이터 관련 법제 정비와 더불어 다양한 데이터 생산주체들의 협업과정은 타 분야의 빅데이터 활용에도 참고할만한 사례이다. 보건·의료분야는 전통적으로 진료기록, 의료보험정보 등의 대용량 데이터를 보유하고 있어 왔으나, 이의 활용에 있어서는 개인정보 보호와 관련하여 다소 보수적이다. 하지만 최근 들어 일부 데이터가 공개되거나, 지방자치단체와 함께 쓰이는 등 활용 가능성은 점차 증대되고 있는 실정이다. 마지막으로 상권분석 서비스는 소상공인 지원정책의 일환으로서 중소기업청, 서울시 등에서 실시하고 있는 빅데이터 서비스이다. 사용자는 행정자료와 유동인구데이터, 카드매출 데이터 등을 종합하여 사용자가 조회한 지역의 업종별 매출 추이 등을 파악할 수 있다.

국내 지방자치단체들도 빅데이터 활용을 위한 노력을 꾸준히 하고 있다. 상당수 지방자치단체들이 빅데이터 관련 사업을 추진해 왔으며, 관광관련 사업이 상당수를 차지하였다. 서울, 경기도와 같은 수도권 일대의 지자체의 빅데이터 인력, 조직이 상대적으로 더 충실히 갖춰져 있는 경향을 보였다.

서울시는 빅데이터 활용 움직임이 가장 활발한 지방자치단체이다. 주요 사업으로는 올빼미 버스(심야버스) 노선 신설, 노인여가복지시설 입지 선정, 우리 마을 가게 상권 분석 서비스 등이 있다. 서울시의 빅데이터 활용 사례를

통해 빅데이터를 이용한 통계가 정책 수립에 효과적이라는 점을 알 수 있으며, 과거의 공공통계와 달리 정형화된 단위로 공표되기 힘들다는 점을 파악할 수 있다. 부산시는 서비스인구통계를 제시함으로써 주간인구 통계를 국내 최초로 생산하였다. 이는 국가승인통계로도 지정되었다. 부산도시서비스분석시스템은 도시의 실시간 유동인구를 파악함으로써 행정 인구 데이터와 괴리되는 실제 행정 수요를 파악할 수 있다는 의의가 있다. 하지만, 데이터 생산에 있어 이동통신사의 데이터를 구매하는 점은 안정적인 데이터 생산이 어려울 수도 있음을 시사한다.

민간분야에서는 카드결제 데이터의 이용이 활발하다. 각 민간 신용카드사는 결제 데이터베이스를 이용하여 마케팅 보고서를 작성하거나, 상권분석 보고서를 제공하기도 하며, 신규 카드 상품 개발에 참고하는 경향을 보인다. 이동통신사들의 빅데이터 이용이 활발한데, SKT의 경우 SKT지오비전이라는 자회사를 설립하여 상권분석 서비스를 제공하고 있다.

본 연구에서는 빅데이터를 활용한 신규과제로 모바일 폰 데이터 랩과 체감경기통계, 이상의 두 가지 신규과제를 제안하였다. 모바일 폰 데이터는 현재 가장 널리 쓰이고 있는 데이터로 단독으로 쓰이기보다는 관광, 상권분석 등의 타 부문과 연계하여 쓰이는 경우가 많다. 이는 모바일 폰 데이터의 인프라적 성격을 보여주는 것으로 여겨진다. 따라서 모바일 폰 데이터는 빅데이터 중 하나가 아니라, 빅데이터 활용의 가장 기초적인 자료로서 인식되어야 한다. 이러한 모바일 폰 데이터의 중요성에도 불구하고, 데이터의 생산과정 및 품질관리 부문은 부각되지 못하였다. 데이터의 특성상 프라이버시 보호가 필수적이며, 전면적 공개가 어려운 현실에서 모바일 폰 데이터를 통계청이라는 제한된 공간에서 데이터 연계를 해나가는 것이 현실적인 방안이다. 이를 통하여 통계 생산 비용을 절감할 수 있을 뿐 아니라 통계청이 통계 생산의 중심 접점 역할을 수행할 수 있게 될 것이다.

이미 통계청은 생산/소비/투자/경기 동향을 종합하여 산업활동동향을 매월 발표하고 있으며, 지역경제의 생산/소비/고용/물가 등을 포괄하는 지역경제동향 역시 작성하고 있다. 이러한 경제통계는 광범위한 정보를 종합적으로 전달해 준다는 장점이 있으나, 자료들이 추상화된 만큼 국민 개개인에게는 체감이 어렵다는 단점이 있다. 따라서 신용카드 결제 데이터를 활용하여 국민의 체감경기에 부합하는 데이터 생산이 필요하다. 이를 통하여 경제통계에 대한 관심을 제고하고, 통계 활용을 촉진시킬 수 있을 것으로 기대된다. 또한 지역경기 활성화에 필요한 기초 자료로서 정책수립 및 효과 측정에 유용한 도구가 될 수 있을 것이다.

Ⅴ. 빅데이터 활용 해외 공공통계 사례

여러 나라들의 정부에서 공공 목적을 위한 빅데이터 활용을 위해 학계, 기업 등의 비정부 부문과 협력하고 있다.

미국의 경우, 오픈 데이터(open data)로 요약되는 공개 지향성과 프라이버시 보호를 위한 노력을 두 가지의 명시적 목표로 삼고 있다. 단, 연방 정부와 지자체의 대립, 사적 부문의 협조 여부 및 정도가 정부의 빅데이터 관련 노력의 결과에 영향을 미칠 중요한 변수로 보인다. 미국은 DATA.GOV를 통해 다양한 데이터를 적극적으로 공개하는 편이다. RISE Colorado와 같은 교육 서비스, 장차 종합 헬스케어 서비스로 편입될 수 있는 약품 관련 필박스(Pillbox), 범죄에 대한 정보를 구체적으로 거의 실시간으로 시각화하여 볼 수 있는 범죄지도 서비스 등은 정부, 지자체, 민간의 협동을 통해 구현된 빅데이터 기반의 공공 통계 서비스로서 장차 한국에서도 시도할 만한 것이다.

그 외에도 여러 해외 사례를 통해 공공 목적을 위해 다양한 포맷의 행정 데이터를 통합하거나, 공공 통계의 질을 높이기 위해 새로운 빅데이터를 통해 보강, 검증하려는 노력을 확인할 수 있다. UN의 통계 관련 부서에서 지난 60여 년 동안 축적한 각종 데이터를 공개한다. 영국은 국민 식별 번호가 없으므로 가장 포괄적이라고 여겨지는 의료 데이터를 중심으로 한 매칭을 통해 개인에 관련된 정보를 통합하여 행정 자료의 활용도를 높이려 한다. 일본의 경우 빅데이터 활용에 비교적 소극적이나, 채권/채무 리스크 분석 등에 제한적으로 활용하려는 시도를 하고 있다. 올해부터 시행된 마이 남바(일종의 국민 식별 번호) 제도가 정착되고, 그것을 중심으로 여러 데이터가 통합되면 행정/빅 데이터의 통합적 활용에 큰 활력소로 작용할 수 있다. 기타 네덜란드, 아일랜드와 같은 소국 또는 EU같은 국가 연합체에서 시도하는 다양한 빅데이터 관련 노력을 참고하여, 한국이 어떻게 향후에 빅데이터를 활용하여 공공 통계를 확장할 수 있을지에 대한 고민은 계속해야 할 것이다.

신규 과제로서는 범죄 관련 빅데이터를 유관 행정 데이터 등과 결합하여 생성할 수 있는 세이프 코리아 지수와 빅데이터 통계 활용 네트워크 구축을 제안한다.

최근 한국에서는 범죄와 관련된 일반의 경계심이 강화되고 있으나, 현재 한국에는 세부 범죄, 세부 지역에 대해 거의 실시간으로 파악할 수 있는, 몇 국가에서 서비스되고 있는 것과 같은 범죄 전문 지도 서비스가 없다. 생활안전지도는 범죄 외의 정보를 함께 제공하는데, 서울 기준으로 구 수준에서, 임의로 5등급으로 재정리한 내용을 보여주는 정도이다.

현재 부분적으로 제공되는 범죄 관련 데이터와, KICS와 같은 범죄 관련 원천 데이터,

지리 정보를 결합하면, 상당히 구체적인 수준에서 특정 지역, 특정 범죄에 대한 현황을 파악할 수 있으며, 역으로 특정 지역에서, 특정 범죄에 대해 어느 정도 안전한지를 수치로 표현할 수 있다. 이러한 포괄적 서비스가 가능하다면 국민의 행복도 증진 및 과학적 범죄 수사 및 예측에 큰 도움이 될 것으로 보인다.

통계청 중심의 빅데이터 활용 네트워크 구축도 또 하나의 중요한 과제로 보인다. 누구든지 정부가 확보해 둔 행정 데이터 외에 빅데이터의 수집을 정부에 요청할 수 있다면, 그러한 요청 자체가 일종의 빅데이터로 기능한다. 통계청과 같은 중앙 정부 기관이 빅데이터 활용에 대한 제안을 상시 접수하고, 좋은 제안을 바탕으로 추가 빅데이터를 수집하여 국민이 활용할 수 있도록 하는 것이다. 이러한 작업을 상시 업무로 수행한다면 어떠한 빅데이터를 공공 통계를 확장할 때 활용해야 하는지 역시 지속적으로 파악 가능할 것이다.

Ⅵ. 빅데이터 통계 신규과제 제시

과제명	개념
가구단위 총부채원리금상환부담 통계	개인 단위가 아닌 가구 단위 부채 정보 취합
모바일 폰 데이터 랩	이동통신사 유동인구정보 결합 및 이용의 중심 센터 구축
체감경기통계	신용카드 사용 데이터, 결제 데이터를 종합한 실물경제 데이터
세이프 코리아 지수 개발	범죄 데이터와 지리데이터를 결합하여 지역별 안전수치 표현
빅데이터 통계 활용 네트워크 구축	정부가 보유한 행정데이터의 활용 중심지 구축

본 보고서에서 제시하는 신규 과제는 상기 표에 제시된 5가지이다.

가구단위 총부채원리금상환부담통계는 기본 가계부채 자료와 달리 가구단위로 부채 정보를 종합한 새로운 통계이다. 가계부채는 설문과 같은 전통적인 방식으로 조사하기가 점차 어려워지고 있는데, 이는 부채에 대한 통일된 정의가 부족할 뿐 아니라 응답자가 본인의 모든 부채를 스스로 기억하지 못하는 경우도 많기 때문이다. 가계부채는 가계 경제 뿐 아니라 국가 전체적으로도 중요한데, 가계의 부채는 곧 금융기관의 자산이기도하기 때문이다. 따라서 이의 명확한 파악이 필수적이며, 이러한 작업은 통계청에서만 시도 가능한 것으로 평가되는데, 민간 금융기관 및 국세청은 개인 단위의 부채정보를 보유하고 있기 때문이다. 통계청은 가구와 개인 간의 연결을 시도할 수 있는 기관으로서, 이를 통해 가구단위 부채정보를 생산할 수 있을 것으로 보인다.

모바일 폰 데이터 랩은 통계청의 통계 생산기관으로서의 전문성과 정부기관으로서의 중립성, 신뢰성에 기반을 둔 신규 과제이다. 최근 널리 쓰이고 있는 모바일 폰 데이터는 대부분 위치기반 정보로서, 이는 그 자체로 의미 있는 데이터라기보다는 다른 데이터들과의 연계를 통해(신용카드 결제, 교통카드 환승정보 등) 그 유용성이 극대화 되는 데이터이다. 따라서 모바일 폰 데이터와 기타 자료들 간의 중간 매개자 역할을 하는 곳이 반드시 필요하며, 통계청은 이에 가장 적합한 기관이다. 통계청이 중심 매개자 역할을 함으로써 통계 생산 비용이 절감됨과 동시에 모바일 폰 데이터의 활용 가능성도 높아질 것이다.

체감경기통계는 기존에 통계청에서 생산해 왔던 산업활동동향이나 지역경제동향을 보완하는 새로운 통계이다. 상기 언급한 통계들은 생산/소비/투자/경기 뿐 아니라 고용과 물가 등을 포괄하는 중요한 통계이나 그 구성이 업종별, 지역별로 추상화되어 제시되기 때문에 일반 국민들에게는 크게 와 닿지 못하고 있다. 따라서 일반 대중이 상시적으로 이용하는 품목에 대한 경기 지수나 자영업자들을 대상으로 하는 업종별 소비 동향 등을 제시함으로써 국민들의 통계에 대한 관심을 높일 필요가 있다. 이는 신용카드 결제 정보를 바탕으로 생산할 수 있으며, 종래의 통계 보고서와는 달리 더 잦은 주기로 데이터를 생산할 수 있을 것으로 기대된다.

세이프 코리아 지수는 최근 범죄에 대한 염려가 높아지는 시점에 적합한 통계이다. 미국 등 몇 국가에서는 지역별로 범죄에 대한 상세한 정보를 제공하는 범죄전문 지도 서비스가 제공되고 있으나, 아직 한국은 이러한 서비스가 불충분한 상태이다. 현재 지역별 범죄 데이터는 조회가 가능하나 이를 시각화 하여 지역별로 조회할 수 있는 서비스는 없는 실정이다. 따라서 현재 부분적으로 제공되는 범죄 관련 데이터와, KICS와 같은 범죄 관련 원천 데이터, 지리 정보를 결합하면, 상당히 구체적인 수준에서 특정 지역, 특정 범죄에 대한 현황을 파악할 수 있다. 이러한 포괄적 서비스가 가능하다면 국민의 행복도 증진과 과학적 범죄 수사 및 예측에 큰 도움이 될 것으로 보인다.

빅데이터 통계 활용 네트워크 구축은 통계청을 중심으로 데이터 조회/처리를 하는 환경을 구성하는 것을 목표로 하는 과제이다. 정부는 거대한 행정 데이터를 보유하고 있지만, 이의 활용은 각 부처별로 나뉘어져 있어 그 활용성과 접근성이 떨어진다. 따라서 이의 활용을 위해 통계청이 빅데이터 활용에 대한 제안을 상시 접수하고, 좋은 제안을 바탕으로 추가 빅데이터를 수집하여 국민이 활용할 수 있도록 제공하는 것을 고려해 볼 수 있다. 이러한 작업을 상시 업무로 수행한다면 어떠한 빅데이터를 공공 통계를 확장할 때 활용해야 하는지 파악 가능할 것이다.

목 차

표 목 차

그 림 목 차

Ⅰ. 서론

김영원(숙명여자대학교 통계학과)

Ⅰ. 서론

빅데이터(Big Data)는 우리 사회의 산업, 과학, 예술분야와 같이 여러 분야에서 크게 주목받고 있으며, 공공 부문과 민간 부문의 모든 영역에서도 변화를 야기하고 있다. 민간부분에서는 빅데이터를 활용한 분석방법론에 대한 다양한 연구 결과와 적용 사례를 국내외에서 쉽게 찾을 수 있는 편이다. 하지만 공공분야에서는 빅데이터를 활용한 통계생산 방법론에 대한 성숙도가 매우 낮은 것이 현실이다. 이런 이유로 아직까지 우리나라 통계청뿐만 아니라 해외 국가통계 작성기관에서도 빅데이터를 기반으로 한 성공적인 국가승인통계 작성 사례를 아직은 찾아보기 어렵다.

최근 정보사회의 패러다임 변화(PC 시대→인터넷/모바일 시대→스마트 시대)에 따른 데이터의 급격한 증가, 소셜 네트워크 서비스의 발달에 따른 비정형 데이터의 폭증, 그림자 정보(위치정보, 검색패턴, 접속기록)의 증가 등과 함께 데이터 저장/처리 비용의 하락(클라우드 컴퓨팅)이나 대용량, 초고속 유무선 네트워크의 보편화 등으로 국가통계 생산이란 관점에서도 활용이 가능한 다양한 형식의 빅데이터가 출현하고 있으며, 이런 흐름 속에서 새로운 국가통계 개발이나 기존 국가통계의 신뢰성 제고 차원에서 빅데이터를 활용한 국가통계 생산방법론에 대한 연구는 매우 중요한 과제라고 볼 수 있다.

통계청은 빅데이터를 활용한 통계생산의 필요성을 파악하여 2015년에 조직 개편을 시행하였다. 통계청은 이를 통해 빅데이터 통계사업을 다각적으로 추진하고 있으며, 빅데이터 기반 온라인물가작성시스템을 개발하는 등 빅데이터를 활용한 신규 통계개발 사업을 진행하고 있다. 이런 추세를 뒷받침하기 위해 2016년에는 통계청 추진 주요 업무로 부채·신용 DB 및 창업지원 DB 구축 등 빅데이터 활용기반을 갖추기 위해 사업을 추진 중이다. 또한 통계청이 보유하고 있는 자료와 외부 행정자료를 포함한 다양한 빅데이터를 연계한 신규 통계 개발을 위해 기관 간 자료 공유를 목적으로 빅데이터 기반 가계 부채 통계 작성을 위한 KCB와 MOU 체결, 빅데이터 기반 사회 예측 시스템 공공 연구를 위한 네이버와 MOU 체결 등 다각적인 노력을 하고 있다.

이런 흐름 속에서 빅데이터를 활용한 통계생산 기법 및 빅데이터 품질검증 방법에 대한 연구를 비롯해 빅데이터 활용 환경 조성을 위한 법적·제도적 현황 파악과 개선점 도출, 가계부채 관련 통계 생산을 포함한 행정자료나 빅데이터를 기반으로 한 신규 통계 개발 필요성 등이 부각됨에 따라 통계청의 요청에 따라 본 연구에서는 다음과 같은 주제들에 대한 연구가 수행되었다.

첫째, 빅데이터를 활용한 통계생산 방법론 관점에서 빅데이터 활용을 위해 광범위하게 활용되는 다양한 기법들을 데이터 증대(data augmentation)와 빅데이터 분석(big-data analytics)으로 구분해 정리한다. 또한 빅데이터를 활용해 국가통계를 생산하게 되는 경우 우선적으로 고려해야 할 신뢰성 확보를 위한 빅데이터 품질검증을 위한 기본적인 틀을 제시하고자 한다.

기존의 국가통계 품질진단과는 달리 빅데이터의 경우에는 자료 수집이 이루어지는 단계에 연구자나 사용자가 관여할 수 없기 때문에 자료 수집 과정을 통계 생산자가 직접 제어할 수 없기 때문에 자료의 신뢰도를 확보하는 것이 용이하지 않게 된다. 따라서 자교 생성 과정을 사후적으로 면밀히 검토하여 사용 가능성을 검증하여야 할 것이다. 따라서 빅데이터의 품질검증은 기존의 국가승인통계 품질진단 방법과는 차별적으로 정의되고 구축되어야 할 것이다.

둘째, 향후의 본격적인 논의의 기초를 마련한다는 측면에서 빅데이터 환경하에서 새로운 통계의 생산을 위한 통계 관련 법제의 적절성을 검토하는 동시에 빅데이터를 기반으로 하는 새로운 통계생산 체제와 현행 개인정보 보호법제와의 조화가능성을 검토해 보고자 한다.

최근 국가·사회 전반에 걸친 정보화의 물결은 통계와 관련한 새로운 국면을 열어가고 있다. 이미 몇몇 분야에서 실제 구현중인 정보시스템을 통한 통계의 자동작성을 위시하여 특히 '빅데이터(Big Data)' 기술을 필두로 새롭게 부각되는 정보처리기술은 기존 통계영역을 넘어선 새로운 통계의 작성은 물론 기존 통계자료에 대한 새로운 방식의 접근을 가능케 하고 있다. 그렇지만 이러한 새로운 통계의 생산은 과거에는 문제되지 않았던 새로운 문제들 특히 개인정보보호의 영역과 밀접한 관련을 맺는 문제점들을 초래할 우려가 있다. 그럼에도 불구하고 그동안 이러한 통계에 대한 법적인 의미 부여와 법제도적 측면에서의 발전적 방향의 모색은 거의 진행되지 못했던 것도 부인할 수 없는 사실이다. 이런 관점에서 본 연구에서 다루게 될 빅데이터 활용과 관련된 법적·제도적 장애요인을 중심으로 한 현황 파악과 개선 방향 도출을 위한 제언은 향후 관련 연구를 추진하는 원동력이 될 수 있을 것이다.

셋째, 다양한 기관이 보유하고 있는 행정자료를 포함한 빅데이터를 활용한 신규 통계를 개발한다는 관점에서 최근 중요하게 부각되고 있는 가계부채 관련 통계 생산 방안에 대해 검토해 보고자 한다.

최근 정부 경제정책 전반에 걸쳐 가계부채에 대한 통계의 필요성이 중요한 이슈로 부각됨에 따라 관련 통계정보의 생산, 축적, 활용, 분석 및 개선 혹은

관리 및 기존 통계의 개선 그리고 신규통계의 개발에 대한 관심이 증가하고 있다. 현재 우리나라에서 가계부채 현황파악은 금융안정성 측면에 초점을 두고 있어 시장정보 조사 측면에서 개선의 여지가 많다고 보인다. 기존 가계부채관련 시장정보는 대부분 금융기관 중심의 가계부채에 대한 정보이기 때문에 우리나라 가계부채를 과소추정하고 있을 가능성이 크다.

본 연구에서는 통계청에서 보유하고 있는 가구 및 가구원 자료에 정부부처에서 제공한 소득정보, 민간기관의 부채정보를 연계하여 가구의 총부채원리금상환부담 통계의 개발을 제안하고자 한다. 현재 통계청에서는 가구소득을 추정하는 작업을 진행하고 있는 것으로 알려져 있으며, 금융권에서는 대출을 이용하는 고객의 증빙소득 혹은 금융기관의 인정소득에 근거한 정보의 제공여부에 따라 일부 소득정보만 보유하고 있는 것이 현실이다. 만약 이런 자료들을 연계해 활용할 수 있다면 통계청에서는 객관적 소득정보를 가구단위로 추정할 수 있는 것이다. 본 연구에서 제안하게 될 가계부채 관련 통계는 신규로 대출을 이용하고 있는 금융소비자 뿐만 아니라 기존 대출과 더불어, 부채가 없는 가구의 소득정보도 포함한다는 관점에서 특히 기존 금융기관의 대출정보나 민간 금융기관이 보유하고 있는 소득정보와 차별화된 정보를 제공할 수 있을 것이다.

넷째, 국내에서도 빅데이터를 활용하기 위한 움직임은 공공부문이나 민간부문 모두에서 활발하다. 국내 빅데이터 활용 동향을 파악하기 위해 현재 다양한 분야에서 진행되고 있는 국내 빅데이터 활용 사례를 살펴보고자 한다.

국내 빅데이터 활용을 현황을 살펴보면 우선 교통 빅데이터 사용이 가장 두드러진 것으로 보이며, 보건·의료분야는 전통적으로 진료기록, 의료보험정보 등을 기초로 한 대용량 데이터를 보유하고 있어 왔으나, 이의 활용에 있어서는 개인정보 보호와 관련된 이슈가 중요하게 부각되고 있는 분야이다. 지방자치단체들의 경우에도 빅데이터 활용을 위한 노력을 꾸준히 하고 있는 것으로 알려져 있다. 여기서는 다각도로 진행되고 있는 국내 빅데이터 활용 현황을 활용 데이터의 유형을 기준으로 구분해 체계적으로 정리하고자 한다. 아울러 기존 국내 빅데이터 활용 사례를 참고하여 통계청 주관으로 작성 가능할 것으로 판단되는 빅데이터를 기반으로 한 국가통계 작성 방안을 구체적으로 제시하고자 한다.

다섯째, 해외에서 진행되고 있는 공공 목적을 위한 빅데이터 활용 사례를 체계적으로 정리하고, 관련 사례를 국내 관련 분야 실정에 맞추어 보고 해당 빅데이터 활용 기법을 국내에 적용 가능한지 검토해 보고자 한다.

여기서는 DATA.GOV를 통한 데이터 공개를 적극적으로 추진하고 있는 미국을 비롯해 영국, 일본, 네덜란드, UN 등에서 최근 진행되고 있는 빅데이터를 활용한 통계 작성과 관련된 학계나 관련 정부 부서의 연구 동향과 해외에서 성공적인 것으로 평가되고 있는 빅데이터 활용 사례들을 구체적으로 살펴보고자 한다. 아울러 우리나라 통계청에서 도입 가능할 것으로 판단되는 빅데이터 활용 사례들을 검토해 보고자 한다.

Ⅱ. 빅데이터 활용 통계생산방법론

박민규(고려대학교 통계학과)

Ⅱ. 빅데이터 활용 통계생산 방법론

1. 개요

『빅데이터』는 최근 산업, 과학, 예술분야를 포함한 모든 분야에서 장차 의사결정 방식의 변화를 주도할 것으로 주목 받고 있는 키워드이다. 하지만 높은 관심에도 불구하고 빅데이터를 활용한 통계생산 방법론에 대한 숙의는 매우 낮은 실정이다. 또한 빅데이터라는 용어가 아직은 모두가 동의할 수 있는 명확한 정의를 가지고 있지 않기 때문에 빅데이터를 활용한 통계라는 개념 역시 정의되지 않은 상황이다. 따라서 빅데이터를 활용하여 국민의 요구에 부합하고 시의성 있는 공공서비스를 제공하기 위해서는 빅데이터 활용 통계생산 분석기법 및 품질검증 방법의 연구가 국가 통계를 생산하는 국가 기관을 통해 이루어지는 것이 바람직하다. 본 연구에서는 이러한 상황에 맞추어서 공공분야에서 우선적으로 검토되어야 하는 빅데이터 활용 통계에 대한 개념, 방향성 그리고 이의 품질 평가 방안에 대한 내용을 살펴보고자 한다.

빅데이터는 협의적으로는 “기존의 관리 및 분석체계로는 감당할 수 없을 정도의 거대한 데이터의 집합”으로 정의된다. 광의적으로는 “기존 데이터베이스 관리도구의 데이터 수집·저장·관리·분석 역량을 넘어서는 대량의 정형 또는 비정형 데이터 세트 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술”으로 정의할 수 있다. 이러한 빅데이터의 출현은 제반 환경의 변화로 촉발된 것이다. 정보사회의 패러다임 변화[1]에 따른 데이터의 급격한 증가(2015년 데이터 생산량 : 7.9ZB), 소셜 네트워크 서비스의 발달에 따른 비정형 데이터의 폭증, 데이터 저장/처리 비용의 하락(크라우드 컴퓨팅), 그림자 정보(위치정보, 검색패턴, 접속기록)의 증가, 대용량, 초고속 유무선 네트워크의 보편화 그리고 사물 정보통신망 확산에 따른 센서 저변 확대 등이 빅데이터의 등장을 이끌어 낸 배경이다. 통상적으로 빅데이터의 개념은 데이터의 규모(Volume), 다양성(Variety) 그리고 속도(Velocity) 이상의 세 가지 특징을 통해 정의된다.

본 연구에서는 우리 주변에 존재하는 빅데이터를 정부 및 공공기관에서 생산하고 관리하는 행정자료와 민간에서 생산되는 민간자료로 구분하여 살펴본다. 그리고 이를 바탕으로 빅데이터를 활용한 통계 생산의 개념을 크게 두 방향에서 살펴본다. 3장에서는 기존의 이용 가능한 공공 데이터 및 이를 활용한 자료 증대(data augmentation) 측면에서의 통계 생산과 축적된 빅데이터의 분석적 측면(big data analysis)에서의 통계 생산 개념을 연구한다.

1) PC 시대 → 인터넷/모바일 시대 → 스마트 시대

또한 이러한 빅데이터 분석방안 중 대표적인 여러 기법들을 간략히 소개한다. 자료 증대를 통해 생성된 빅데이터의 활용과 빅데이터의 분석을 위해서는 사용되거나 혹은 제공될 데이터의 품질에 대한 평가가 우선 시행되어야 한다. 이를 위해서 기존의 데이터 혹은 통계에 대한 품질 평가 방안들을 살펴보고 공공 및 민간 데이터를 활용한 빅데이터 분석에 이를 적용 확대할 수 있는 방안들을 살펴본다. 4장에서는 기존의 데이터 및 통계 품질의 평가를 위한 요소들을 살펴보고 빅데이터를 활용하여 생산된 통계의 품질 검증 방안을 연구한다. 마지막으로 5장에서는 논의된 내용을 바탕으로 빅데이터 활용 통계 생산 방법론 및 검증체계에 대한 방향에 대한 논의를 서술하였다.

2. 빅데이터 통계 및 분석 현황

본 장에서는 우리 주변에 존재하는 빅데이터의 현황을 국가에서 생산 관리하는 행정자료와 민간에서 생산되는 민간 자료로 구분하여 서술한다.

제1절 행정자료

가. 국가통계포털(KOSIS)에 등록된 공공데이터[2)]

1) 분류

국가통계포털에 등록된 통계는 통계 생산을 위해 사용된 자료 수집 방안, 통계 생산 처리 단계 그리고 승인 여부에 따라 아래와 같이 분류할 수 있다.

(1) 조사통계와 보고통계

조사통계란 통계의 작성을 주목적으로 조사를 실시하여 얻어진 통계를 말하며 제1의 통계라고도 한다. 조사통계는 조사대상 집단(모집단)의 모든 단위를 조사하는 전수조사와 모집단의 일부를 나타내는 표본을 추출하여 조사하고 이를 근거로 모집단에 추론을 수행하는 표본조사로 구분할 수 있다. 모집단의 기본적 구조, 특성, 지역적 세부상황 등을 파악하기 위한 통계는 총조사, 대규모조사(주로 전수조사)에 의하여 작성되며, 경상적인 동향, 추이를 나타내는 통계는 표본조사에 의하여 주로 작성된다.

보고통계는 법령에 의한 개인, 단체의 신고, 보고, 신청, 인·허가 등과 같이 다른 행정 업무에 수반하여 수집된 자료로부터 통계를 작성한 것을 말하며, 제2의 통계라고도 한다. 통계조사의 실시에는 예산, 조사원의 확보, 조사객체의 비협조 등 사실상 어려움이 많고, 최근 들어 이러한 어려움은 점점 더 심해져 가고 있다. 따라서 보고통계는 이와 같은 어려움이 적고 또 대상 집단을 전수로 파악하는 것이므로 세부 소지역에 관한 통계작성도 가능하다는 장점이 있으나 신고율, 신고내용의 정확성 등에 따라 통계의 질이 좌우되는 근본적인 문제가 있다.

2) 데이터베이스, 전자화된 파일 등 공공 기관이 법령 등에서 정하는 목적을 위하여 생성 또는 취득하여 관리하고 있는 전자적 방식으로 처리된 자료 또는 정보로 개별 공공기관이 일상적 업무수행의 결과물로 생성 또는 수집 취득한 다양한 형태(텍스트, 수치, 이미지, 동영상, 오디오 등)의 모든 자료 또는 정보가 공공데이터에 해당된다.

(2) 1차 통계와 가공통계(2차 통계)

1차 통계란 집단에 속하는 개체의 수 또는 개체의 특성을 총체적으로 나타내는 통계를 말한다. 일반적으로 통계조사를 실시하여 그 결과에서 직접 얻어진 통계가 이에 해당한다. 가공통계(2차 통계)는 1차 통계에 어떠한 연산을 하여 얻어진 통계로서 1차 통계에 비하여 해석적 특성이 있는 통계이다. 가공통계에서는 집단 특성치의 평균, 산포도, 지수, 상관계수 등뿐만 아니라 국민소득통계와 같은 추계에 의한 통계도 있다.

(3) 지정통계와 일반통계

지정통계란 중앙행정기관이나 지방자치단체 또는 지정기관이 작성하는 통계로서 통계청장이 지정·고시하는 통계를 말하며, 국가 또는 지방자치단체의 주요정책수립 및 평가 등을 위하여 널리 활용되는 통계 중에서 지정된다. 따라서 지정통계에는 통계법상 자료제출명령, 실지조사 등 일정한 권한이 부여되므로 이를 지정할 경우에는 반드시 고시하여 국민이 알 수 있도록 하여야 한다. 일반통계는 중앙행정기관이나 지방자치단체 또는 지정기관이 작성하는 통계로서 지정통계 이외의 통계를 말한다.

2) 국가통계포털(KOSIS)에 등록된 공공데이터 생산 현황

2016년 9월 13일 기준 정부승인 통계는 총 976종이며 지정통계가 93종, 일반통계는 883종이다. 작성 방법별로는 조사통계가 432종, 보고통계 453종, 가공통계 91종이다. 또한, 작성 기관을 정부기관과 지정기관으로 구분하여 볼 때 정부기관에 의해 작성되고 있는 통계는 803종이며, 지정기관에서 작성하고 있는 통계는 173종이다.

<표 1> 기관별 정부승인통계 작성 현황(2016년 9월 13일 현재, 단위: 기관, 종)

구분	작성기관수	작성통계수	종류별		작성방법		
			지정통계	일반통계	조사통계	보고통계	가공통계
계	397	976	93	883	432	453	91
◎정부기관	**303**	**803**	**75**	**728**	**329**	**402**	**72**
-중앙행정기관	43	355	58	297	180	141	34
통계청	1	59	39	20	40	2	17
이외기관	42	296	19	277	140	139	17
- 지방자치단체	260	448	17	431	149	261	38
◎ 지정기관	**94**	**173**	**18**	**155**	**103**	**51**	**19**
- 금융기관	8	24	10	14	10	6	8
- 공사/공단	27	50	0	50	21	27	2
- 연구기관	20	37	2	35	31	4	2
- 협회/조합	22	32	4	28	27	3	2
- 기타기관	17	30	2	28	14	11	5

<표 2> 부문별 정부승인통계 작성 현황(2016년 9월 13일 현재, 단위: 종, %)

구분	작성통계수		종류별		작성방법		
	통계수	구성비	지정	일반	조사통계	보고통계	가공통계
계	976	100	93	883	432	453	91
인구	40	4.1	4	36	4	25	11
고용.임금	38	3.9	8	30	30	7	1
물가.가계소비(소득)	16	1.6	9	7	15	1	0
보건.사회.복지	216	22.1	6	210	154	53	9
환경	28	2.9	3	25	12	13	3
농림.수산	55	5.6	9	46	35	17	3
광공업.에너지	32	3.3	3	29	21	8	3
건설.주택.토지	40	4.1	2	38	14	20	6
교통.정보통신	49	5	4	45	22	24	3
도소매.서비스	16	1.6	6	10	14	2	0
경기.기업경영	92	9.4	26	66	64	5	23
국민계정.지역계정	21	2.2	5	16	0	0	21
재정.금융	19	1.9	2	17	3	16	0
무역.외환.국제수지	11	1.1	3	8	0	4	7
교육.문화.과학	58	5.9	3	55	42	15	1
기타(시도기본통계포함)	245	25.1	0	245	2	243	0

3) 국가통계포털(KOSIS)에 등록된 공공데이터 주요 생산 기관

(1) 정부 통계기관

현재 우리나라의 정부 통계기관으로는 국가통계행정을 종합적으로 관장하는 통계청이 있으며, 고용노동부, 보건복지부, 교육부, 국토교통부 및 환경부 등 각급 중앙행정기관에서 소관 업무와 관련된 통계를 작성하고 있다. 그리고 각 시·도에는 기획관리실 산하에 정책기획관 또는 법무 통계담당관·정보화담당관을 두고 있다. 시·군·구에서는 통계업무를 담당하는 “계” 단위 조직을 두고 주민등록인구, 통계연보 등 자체 계획 수립 등에 필요한 통계업무 수행과 동시에 중앙행정 기관에서 실시하는 대규모 통계조사의 현지 조사업무 또는 자료수집 업무 등을 지원하고 있다. 정부 통계기관의 2016년 9월 13일 기준 현황은 이하와 같다.

<표 3> 정부 통계기관 목록

구분	기관명칭
정부통계기관 (34개 기관)	통계청, 기획재정부, 통일부, 환경부, 국가보훈처, 행정자치부, 법무부, 교육부, 문화체육관광부, 농림축산식품부, 산업통상자원부, 국토교통부, 보건복지부, 고용노동부, 미래창조과학부, 조달청, 경찰청, 국세청, 관세청, 검찰청, 산림청, 특허청, 기상청, 중소기업청, 농촌진흥청, 병무청, 식품의약품안전처, 해양수산부, 문화재청, 공정거래위원회, 여성가족부, 방송통신위원회, 인사혁신처, 국민안전처

(2) 통계작성 지정기관

통계작성 지정기관은 정부통계기관 이외에 통계를 작성하는 기관으로서 민법이나 그 밖의 다른 법률에 따라 설립된 법인이고 통계의 작성과 보급에 필요한 조직 및 인력과 예산을 충분히 가지고 있으며 구체적이고 실현할 수 있는 통계의 작성/보급에 관한 계획을 갖고 있는 기관을 지정한다(통계법 제15조).

<표 4> 통계작성 지정기관 목록

구분	기관명칭
공사 / 공단 (26개 기관)	소상공인시장진흥공단, 한국전력공사, 한국농어촌공사, 한국도로공사, 한국관광공사, 한국토지주택공사, 한국석유공사, 한국주택금융공사, 국민연금공단, 축산물품질평가원, 한국시설안전공단, 한국에너지공단, 국민건강보험공단, 국립공원관리공단, 한국철도공사, 한국산업안전보건공단, 한국공항공사, 한국장애인고용공단, 한국광물자원공사, 한국산업인력공단, 한국환경공단, 한국산업단지공단, 주택도시보증공사, 한국전기안전공사, 한국방송광고진흥공사, 교통안전공단
금융기관 (8개 기관)	한국은행, 중소기업은행, 수산업협동조합중앙회, 한국거래소, 한국예탁결제원, 금융감독원, 보험개발원, 한국산업은행
연구기관 (20개 기관)	한국보건사회연구원, 한국교육개발원, 한국노동연구원, 한국여성정책연구원, 에너지경제연구원, 한국조세재정연구원, 한국보건산업진흥원, 한국전기산업연구원, 한국직업능력개발원, 국토연구원, 과학기술정책연구원, 한국건설기술연구원, 중소기업기술정보진흥원, 한국청소년정책연구원, 한국형사정책연구원, 정보통신정책연구원, 한국환경산업기술원, 한국행정연구원, 한국한의학연구원, 한국지질자원연구원
협회 / 조합 (20개 기관)	한국금융투자협회, 해외건설협회, 중소기업중앙회, 대한시설물유지관리협회, 한국무역협회, 한국철강협회, 대한건설협회, 대한전문건설협회, 생명보험협회, 대한설비건설협회, 한국전기공사협회, 한국정보통신공사협회, 한국엔지니어링협회, 한국기계산업진흥회, 한국정보통신진흥협회, 한국소프트웨어산업협회, 대한측량협회, 한국건설기술관리협회, 한국여성경제인협회, 한국로봇산업협회
기타 (15개 기관)	한국고용정보원, 한국인터넷진흥원, 한국생산성본부, 건강보험심사평가원, 정보통신산업진흥원, 한국전력거래소, 한국감정원, 한국방위산업진흥회, 국립중앙의료원, (재)한국정보통신산업연구원, 한국언론진흥재단, 한국공정거래조정원, 한국사학진흥재단, 국가평생교육진흥원, 한국기상산업진흥원

나. 국가통계포털(KOSIS)이외 활용 가능한 공공데이터

1) 국민건강정보 데이터

① 내용 : 국민건강보험자격자(전 국민)의 약 2%에 해당하는 100만 명에 대한 2002년부터 2013년까지의 진료내역정보, 약품처방내역정보, 건강검진정보

② 데이터 제공기관 : 국민건강보험공단

③ 데이터 종류 : 진료내역정보(서식코드, 진료과목코드, 주상병코드, 요양일수, 입원내원 일수 등), 약품처방 내역정보(약품일반성분명코드, 1회 투약량, 1일투약량, 총 투여일수, 단가, 금액 등), 건강검진정보(신장, 체중, 허리둘레, 혈압, 혈당, 콜레스테롤, 시력, 구강, 흡연, 음주 등) 등 약 4억 건

④ 활용 시 기대효과

가) 법률에 근거한 국민의 공공데이터 이용권 보장 및 주요 보건 의료 정보에 대한 국민의 알권리를 충족

나) 특정개인과 유사한 집단의 건강상태를 비교할 수 있는 콘텐츠, 지역(시도)별 및 연령대별 건강상태 정보 제공 콘텐츠 등 다양한 콘텐츠 생산 및 제공 가능

다) 국민건강정보데이터를 활용한 건강정보 관련 산업계의 새로운 형태의 창업에 기여

라) 진료내역, 의약품처방정보를 일반국민에게 개방함으로서 의료기관의 고품질 의료 서비스 제공 유도

2) 건축정보 데이터

① 내용 : 건축물의 기획부터 소멸에 이르는 건축물 생애 관련(허가→착공→사용승인→유지관리→철거 등)하여 행정업무 전반에서 발생하는 정보로서 국민의 주거 및 경제활동, 생활에 가장 기초가 되는 데이터

② 데이터 제공기관 : 국토교통부

③ 데이터 종류 : 건축물대장(기본개요, 주택가격, 부속지번 등), 폐쇄말소대장(기본개요, 층 별개요, 부속지번 등), 건축인허가(도로명 대장, 부설주차장 등), 주택인허가(대지위치, 부대시설, 부설주차장 등), 건물에너지(수용가별 에너지사용량), 건물 유지점검(점검접수보고, 점검결과, 유지관리건축물관리대장) 등 약 7억 건

④ 활용 시 기대효과

가) 소상공인, 데이터 유통, 금융, 부동산 매입 및 컨설팅 등 다양한 분야의 민간기업과 일반국민으로 사용자가 확대되고 건축설계, 건축 유지·점검, 건축 콘텐츠 유통 등 건축 관련 산업 활성화에 기여

나) 건축물 기반의 데이터 통합 및 활용을 통해 건물관리(사용점검, 에너지관리 등), 최적 주거지 찾기 등에 활용하여 소요시간 및 비용 절감

다) 건축물 정보를 민간기업 및 개인이 활용하여 다양한 창조적 융합서비스를 창출하게 되면 다양한 사회적 비용의 절감에 기여

3) 상권정보 데이터

① 내용 : 전국 약 200만 상가업소에 대한 정보로서, 창업 시 입지선정, 업종전환, 점포이력, 경쟁분석, 마케팅 등 창업과 경영을 위한 상권정보서비스의 핵심 정보

② 데이터 제공기관 : 소상공인시장진흥공단

③ 데이터 종류 : 상가업소정보(상호 명, 지점 명, 주소, 도로명, 신 우편번호, 상권번호, 표준산업분류코드 등), 상가업종정보(대분류, 중분류, 소분류) 등 약 200만 건

④ 활용 시 기대효과

가) 상가정보를 활용하여 자신의 소비 유형에 맞는 맞춤형 상점 추천서비스 등 다양한 서비스의 이용 가능

나) 상가정보와 연계된 유사한 정보와의 융·복합을 통해 컨설팅과 중개시장의 새로운 서비스 출현 및 창업 코디네이터와 같은 유사한 직업의 출현을 통해 창업 컨설팅 비용 감소

다) 업소 인허가 데이터, 휴폐업 이력, 지역별 분포정보와 민간데이터인 매출액, 유동인구를 융합한 상권분석서비스 활용으로 유망 창업 아이템 선정에 기여

라) 지역 맞춤형으로 특화된 마케팅 전략을 통해 전통시장 및 상권 활성화 기여

4) 실시간 수도정보 데이터

① 내용 : 상수도의 수돗물 생산 및 공급시설의 운영을 위해 설치한 각종 감지기를 통해 실시간으로 수집되어지는 정보

② 데이터 제공기관 : 한국수자원공사

③ 데이터 종류 : 취수장(유량, 압력), 정수장(수질, 유량, 압력, 수위), 가압장(유량, 압력), 배수지(수위)의 실시간(1시간 주기) 정보 등 약 1,200만 건

④ 활용 시 기대효과

가) 환경, 수질, 물 관련 정보를 제공하는 정보서비스 업체를 통해 회원들의 거주지에 공급되는 수질정보 제공 등 대국민의 실생활에 중요한 정보를 제공함으로써 안전한 물 인식 전환을 통한 음용율 향상, 물 절약 유도 등 유용한 부가가치의 창출에 활용

나) 수질, 물 산업 관련 협회, 학계, 연구기관 등에서는 광역상수도에서 공급되는 수질, 유량, 압력 등의 정보를 활용 관심연구 분야의 데이터를 시계열 분석 등 체계적으로 분석할 수 있어 관련 연구 분야가 활성화될 수 있음

다) 물 산업 업계는 개방된 데이터를 기업의 제품성능 정보와 접목하여 제품 성능 향상

및 고객의 눈높이에 맞는 마케팅 전략을 수립하는데 활용

5) 농수축산 경락 및 조사가격정보 데이터

① 내용 : 35개 공영도매시장(105개 도매시장법인)과 한국농수산식품유통공사, 축산물품질평가원, 농협중앙회, 수협중앙회 등 4개 조사기관에서 20여 년 동안 축적한 경락 및 조사가격 데이터로 국민 경제활동에 매우 밀접하고 국내 산업 전반에 영향을 줄 수 있는 중요한 데이터임

② 데이터 제공기관 : 농림축산식품부

③ 데이터 종류 : 경락가격정보(경락 일자, 경매시간 등), 조사가격정보(조사일자, 품목명 등), 산지공판장 경락가격정보(경락 일자, 경매시간, 공판장명 등), 산지 위판장 경락가격정보(경락 일자, 경매시간, 위판장명 등), 신·구표준매핑정보(품목, 시장, 단위, 포장, 크기, 등급, 산지), 조사가격매핑정보(품목, 시장, 단위), 국제표준매핑정보(GPC, HSK) 등 약 10억 건

④ 활용 시 기대효과

가) 생산자는 최적의 출하시기를 결정할 수 있어 소득 증가에 기여

나) 유통인은 원천데이터 수준의 가격·물량 정보를 활용하여 신 유통모델을 발굴 가능

다) 소비자는 실시간 농수축산물 가격·물량정보를 활용한 가격비교 등을 통해 구매 규모별 스마트 소비 실현

6) 법령정보 데이터

① 내용 : 우리나라의 법령, 행정규칙, 자치법규, 판례, 법령 해석례, 행정 심판례, 헌재 결정례 등 우리나라의 모든 법령정보

② 데이터 제공기관 : 법제처

③ 데이터 종류 : 현행법령, 조약, 영문법령, 행정규칙, 자치법규, 법령해석례, 판례, 행정심판례, 헌재결정례, 별표/서식, 법령용어 등 약 300만 건

④ 활용 시 기대효과

가) 법령정보 활용으로 사회적 편익이 증대될 것으로 기대

나) 법령정보를 통해 평소에 몰라서 그냥 지나쳤던 문제들을 해결할 수 있으며, 자신이 사는 지자체의 불합리한 조례 사항을 쉽게 확인하고 개선 요구 가능

다) 법령정보를 활용한 법률상담서비스나 타 분야 정보와 기술들을 융합하여 부가가치를 만들어 내는 서비스 등으로 창업 아이템을 선정 가능

라) 오픈 API로 제공되는 법령정보를 각 기관이나 개인 누구나 손쉽게 새로운

법령 관련 앱/웹 서비스 제작이 가능

7) 식의약품종합정보 데이터

① 내용 : 식품, 의약품, 의료기기 및 화장품의 기준규격, 인허가정보, 업체정보, 부적합정보 등 실생활과 밀접한 데이터

② 데이터 제공기관 : 식품의약품안전처, 식품의약품안전평가원

③ 데이터 종류 : 현의약품정보(업체, 품목, 회수, 마약류 정보 등), 의료기기(재평가, 광고심의 추적관리대상, GMP 등), 식품(인허가, 회수폐기, 건강기능식품, 행정처분, 광고, 식중독정보 등) 화학물질독성정보, 화장품(업체, 기능성화장품), 임상시험 정보, 잔류물질, 생약종합정보, 어린이 급식관리, 영양성분 등 약 400만 건

④ 활용 시 기대효과

가) 실생활과 밀접한 의약품, 화장품, 의료기기 등에 대한 데이터를 활용하여 유해 식의약품의 품목, 음식점 위생 점검 결과, 식단정보 등을 확인하여 일상생활에서 이용

나) 식의약품 관련 업체 및 창업을 준비하는 업체는 인허가 정보, 해외 정보, 부적합 내역 등을 참고하여 신규 일자리 창출 및 부정 및 불량 식의약품이 없는 건강한 제품 생산에 도움

다) 식품을 제조·가공·조리하는 식품원료, 관리기준 등 식품정보를 제공함으로써 안전한 식품관리 제고

8) 노동보험정보 데이터

① 내용 : 1인 이상 근로자를 고용하고 있는 노동보험 가입 사업장에 대한 고용·산재보험의 가입 및 납부, 산재근로자의 요양·보상, 재활사업, 임금채권 보장사업 등 노동보험 업무처리 시스템을 통해 생성되는 데이터

② 데이터 제공기관 : 근로복지공단

③ 데이터 종류 : 보험가입 사업장 정보, 보험사무대행기관 관련정보, 산재보험지정의료기관 정보, 산재재활기관 정보, 고용산재보험 관련 각종 통계 정보, 산재보험 최초요양신청승인 정보, 행정심판 및 심사결정 관련 정보 등 약 400만 건

④ 활용 시 기대효과

가) 안전 또는 보험과 관련된 다양한 사업을 위한 기초자료로 쓰이거나 학술·연구자료 등으로 활용

나) 국민이나 산재근로자, 노무·법률 등 전문직 종사자들이 이용하는 서비스와

노동보험 DB에서 제공하는 서비스가 결합되어 새로운 부가가치를 창출할 수 있을 것으로 기대

9) 교육행정정보 데이터

① 내용 : 전국 초·중등학교의 학사, 인사, 예산 등의 교육행정정보와 학생·교원현황, 학교시설 등 학교 전반의 데이터

② 데이터 제공기관 : 교육부

③ 데이터 종류 : 대학진학률(졸업생의 진로현황), 수업시수, 학교시설(교사현황, 학생교육활동에 필요한 지원시설 현황 등), 교원현황, 수업공개 계획, 교육특색사업, 학생현황, 동아리활동, 교복구매 유형 및 단가 등

④ 활용 시 기대효과

가) 학교 전반의 주요정보를 객관적이고 투명하게 개방함으로써 국민의 알 권리를 보장할 수 있으며, 학교의 교육 실태를 정확하게 파악하여 학교 교육의 경쟁력을 높일 수 있음

나) 교육행정정보·학교 알림이 정보를 통해 지역간·학교간 교육격차 해소하는 데 활용

제2절 민간자료

공공주체에 의해 수집되지 않은 데이터로 신용카드 거래, 전자상거래 등의 금융데이터, 각종 센서에 의해 자동 수집되는 사물정보통신 데이터, GPS와 핸드폰에 의해 수집되어지는 위치 데이터 그리고 온라인 및 SNS 데이터 등이 민간 데이터에 해당한다.

가. 데이터의 형태에 따른 분류

데이터의 정형화 정도에 따라 DB에 저장된 정형 데이터, 웹문서와 같은 반정형 데이터, 오디오, 이미지, 동영상 등과 같은 비정형 데이터의 유형으로 분류할 수 있다. 정형 데이터는 잘 정리되어 있어 분석이 용이하고 따라서 직접적인 추가분석이 가능한 형태이다. 반면, 비정형 데이터의 경우에는 데이터의 형태가 다양하며 따라서 표준화된 데이터의 분석을 위해 작성된 프로그램을 통해 분석하기 매우 어렵다. 흔히, 표현하는 데이터웨어하우스(Data Warehouse)에 기록하고 저장된 데이터를 정형 데이터라고 할 수 있으며 정형 데이터 이외의 모든 복잡하고 다양한 형태의 데이터를 비정형 데이터라고 통칭한다. 예를 들어, 기업 또는 기관에서 저장하고 있는 고객 정보와 매출 정보 그리고 주문 정보와 직원 정보 등 전통적으로 오랫동안 축적되고 관리하는 데이터를 정형 데이터로 볼 수 있으며, 소셜 데이터와 디지털 매체와 온라인 서비스 등을 통해 빠른 속도로 축적되는 음성, 영상, 이미지 등의 다양하고 복잡한 형태의 데이터를 대표적인 비정형 데이터로 구분할 수 있다. 특히, 민간부분의 빅데이터는 구조화되지 않은 비정형데이터가 90% 이상을 차지하고 있다.

나. 데이터의 출처에 따른 분류

출처에 따른 구분 방법으로는 내부 데이터(Internal Data), 그리고 외부 데이터(External Data)로 분류한다. 내부 데이터는 기업이 보유하고 있는 영업 데이터와 고객 데이터 그리고 거래 정보 또는 매출 기록 등을 말한다. 이러한 데이터는 외부로 공개되는 것이 거의 불가능하며 상당한 수준의 보안이 요구된다. 외부 데이터는 내부 데이터의 반대 개념으로 인터넷에서 접할 수 있는 소셜 데이터와 온라인 뉴스 및 블로그 등이 대표적인 외부 데이터이다.

다. 데이터의 생성 주체에 따른 분류

빅데이터는 생성 주체에 따라 어플리케이션, 센서 등을 통하여 수집되는 기계데이터, 외부의 불특정 다수로부터 생성되는 트위터, 블로그 등에 올린 사람과 사람 사이의 상호작용으로 생성되는 소셜 데이터 그리고 개체간의 관계를 통하여 생성되는 관계 데이터 등으로 구분할 수 있다.

3. 빅데이터 활용 통계 생산 방법론

본 장에서는 2장에서 살펴본 여러 형태의 빅데이터를 활용한 가능한 통계 생산 방안들에 대해서 살펴본다.

제1절 데이터 증대(Data Augmentation)

가. 무응답 대체(Imputation)

자료 수집 단계에서는 보편적으로 무응답(nonresponse)이 발생한다. 최근 개인정보보호에 대한 인식이 바뀌면서 조사를 통한 자료 수집이 어려워지고 있다. 조사 거절이나 부재로 인한 응답률 저하는 매우 심각한 수준이다. 또한 조사 과정에서 민감한 문항이나 소득과 같은 정보 제공을 꺼리는 항목에 대한 무응답 역시 증가하고 있다. 무응답은 그 형태에 따라 전체 조사에 참여하지 않아 발생하는 개체 무응답(unit nonresponse)과 일부 항목에 대한 무응답을 나타내는 항목 무응답(item nonresponse)으로 구분할 수 있다.

결측 혹은 무응답은 대부분의 조사에서 여러 가지 이유로 인하여 발생하게 되는데 일반적으로 무응답 자료는 응답 자료와는 그 성격이 다르므로 응답 자료만으로는 모집단을 대표하지 못한다. 따라서 무응답을 제거 또는 무시하고 완전히 관찰된 자료만을 이용하여 기존의 통계분석을 적용하게 되면 그 결과에 편향이 발생할 수 있다. 무응답을 줄이기 위해서 연구계획 및 설계 단계부터 노력을 기울여야 하지만 이러한 노력에도 불구하고 자료에 무응답이 발생하는 경우에는 무응답을 고려한 적절한 통계적 분석방법을 적용하여 연구결과의 신뢰성을 높이는 노력을 하여야 한다. 일반적으로 개체무응답의 통계적 처리를 위해서는 흔히 가중치 조정 방안이 사용된다. 항목 무응답의 처리를 위해서는 결측이 발생한 항목의 값을 적절한 값으로 대체하는 무응답 대체가 사용된다. 가중치 조정을 통한 무응답 조정 방안에 대한 연구는 오랜 기간 동안 많은 연구자에 의하여 수행되어 왔다[3].

무응답 대체의 방안은 두 가지로 나눌 수 있다. 한 가지는 결측의 대체를 위하여 한 번의 대체를 통해 하나의 대체된 데이터 셋을 만드는 방안인 단일 대체(single imputation)이다. 나머지 하나는 Rubin(1987)에 의하여 소개된

3) 다양한 연구결과에 대한 내용에 대해서는 Fuller(2009)와 Sarndal, Swensson and Wretman(1991)을 참고

것으로 하나의 결측을 가능한 여러 값으로 대체하는 다중 대체(multiple imputation) 방식이다. 단일 대체방법은 대체된 값을 마치 실제로 관찰된 값으로 생각하고 자료를 분석하게 되므로 대체로 인하여 발생하는 불확실성(uncertainty), 즉 결측으로 발생하는 불확실성을 고려하지 못하여 추정량의 표준오차 추정량이 실제 오차를 과소추정(underestimate)하는 문제가 발생하게 된다. 다중대체는 이러한 단일 대체를 통해 생성되는 표본 오차의 과소 추정 문제를 해결하고 또한 점추정량의 통계적 효율성을 높이기 위하여 제시된 방안이다.

빅데이터를 활용할 경우에 통상적으로 사용되는 무응답 대체의 효율성을 높일 수 있을 것으로 기대할 수 있다. 예를 들어, 기존의 무응답 대체 방안들이 동일 조사로부터 얻은 정보를 이용한 대체, 일종의 Hot-deck 방안을 고려했다면 관련된 과거 및 현재의 여러 자료를 활용한 무응답 대체 방안을 사용할 경우에는 결측에 대한 보다 많은 정보를 활용할 수 있고 이를 바탕으로 보다 타당한 대체 값을 구할 수 있을 것으로 기대할 수 있다. 이러한 무응답 대체 방안의 확대는 언급되는 데이터 매칭과 직접적으로 연결된다.

나. 데이터 매칭(Data Matching)

데이터 매칭이란 보유하고 있는 데이터 파일에 필요한 변수가 없거나, 결측값이 존재할 경우 다른 원천 데이터로부터 모아진 자료와 정보를 통합하는 것이다. 데이터 매칭을 통하여 데이터의 질을 상당히 높일 수 있으며 재조사로 인한 시간과 비용을 줄일 수 있는 장점이 있다. 또한 많은 조사항목으로 인한 조사응답자의 부담을 경감시켜 조사항목에 대한 무응답률이 낮아지고 응답의 정확성이 높아져 새로운 조사를 통해 얻은 자료보다 더 좋은 자료를 얻을 수 있다. 원천 데이터의 다양성, 단일 자료의 불충분성, 부서간의 자료 공유의 부족으로 인하여 하나의 데이터에서 분석에 필요한 모든 정보를 얻는 것은 매우 어려운 일이다. 이러한 문제를 데이터 매칭을 통해 상당 부분 보완할 수 있다. 일반적인 조사 데이터에 공통적으로 포함하고 있는 요소들이 완전히 일치하지는 않지만 특성이 유사한 사람이나 집단끼리의 정보를 활용하여 데이터를 매칭하고 이를 통해 보다 다양한 분석이 가능한 일종의 빅데이터를 구성할 수 있다.

데이터 매칭은 정확 매칭(Exact Matching), 통계적 매칭(Statistical Matching), 판단 매칭(Judgmental Matching) 등으로 분류할 수 있다. 정확 매칭은 주민등록번호, 국가보험번호, 사회보장번호와 같이 각 개체를 식별할 수 있는 변수가 공통으로 있을 경우에 동일한 값을 갖는 개체들을 결합하는 방법이다.

이는 같은 사람, 같은 물건을 정확하게 결합할 수 있다는 장점이 있다. 측정 오차가 없다면 가장 이상적인 데이터 결합의 형태이다. 이러한 정확 매칭은 데이터 연계(data linkage)로 이해되기도 한다. 2장에서 언급한 자료 중 공공데이터의 경우에는 조사 자료 및 보고 자료에 대부분 각 개인, 가구 혹은 사업체에 대한 공통적으로 고유 식별번호가 포함되어 있다. 따라서 이를 이용하여 정확매칭을 시도할 수 있으며, 보다 다양한 분석을 실시할 수 있는 데이터를 확보할 수 있다.

정확매칭을 통한 통계 생산의 대표적인 사례로는 2015 인구주택총조사(혹은 등록센서스)가 있다. 기존의 인구주택총조사가 직접 방문조사를 통해 이루어진 것과는 달리 과거 인구주택총조사의 전수항목(short form)에 해당되는 문항의 통계 작성을 위해서 2015년에는 행정자료의 매칭을 통해서 작성된 빅데이터를 활용한 등록센서스 방안을 채택하였다. 등록센서스를 통한 통계 생산을 위해서는 11개 기관(국토교통부, 행정자치부, 법무부, 대법원, 보건복지부, 국방부, 외교부, 경찰청, 국민안전처, 한국전력공사, 교육부)에서 관리하고 있는 데이터가 사용되었다. 정확 매칭을 위해서 사용되는 개체의 식별번호의 상용을 위해서는 많은 주의가 필요하다. 특별히 개인 정보 보호에 대한 법률에 위반되지 않도록 자료를 관리하는 방안이 우선 모색되어야 한다.

통계적 매칭은 각 개체에 대한 고유 번호가 모든 데이터에 존재하지 않는 경우에 고려할 수 있는 방식이다. 통계적 매칭은 매칭을 수행할 각 자료 내의 공통된 변수를 기준으로 각 변수간 거리를 구하는 것으로 시작된다. 이 중 상대적으로 더 가까운 변수들을 서로 매칭시킴으로써 데이터를 구축한다. 따라서 통계적 매칭을 위해서는 개체 간의 유사성을 측정하기 위한 거리함수가 정의되어야 한다. 이를 위해서는 공통변수의 통계적 특성, 공통변수의 측정을 위해 사용된 측정 단위와 동일 거리의 처리 방안 등 많은 사항들을 고려하여야 한다. 통계적 매칭의 가장 큰 약점은 이를 위해 통계 이론적으로 요구되는 조건부 독립 가정이다. 이는 매칭을 위한 자료에 공통적으로 포함된 변수가 주어졌을 때 나머지 변수들이 서로 독립이라는 가정이다. 따라서 매칭이 된 자료를 분석할 때 매우 큰 장애 요소로 작용할 수 있다. 이러한 통계적 매칭의 성질을 포함하여 통계적 매칭에 대한 다양한 연구가 이루어지고 있다[4].

다. 보조정보(auxiliary information) 확보

데이터 증대와 관련된 빅데이터 활용으로는 조사 자료 혹은 보고 자료의

4) D'Orazio, Di Zio and Scanu(2006)은 매칭에 대한 전반적인 내용을 설명하고 있다.

분석을 위해 필요한 모집단에 대한 보조정보를 획득하는 것을 고려할 수 있다. 예를 들어, 조사 자료의 분석을 위해서 작성되는 가중치를 산출하기 위해서는 마지막 단계에서 모집단 정보를 활용한 벤치마킹이 이루어진다. 대표적인 벤치마킹 방안으로는 사후층화와 레이킹 비 방안 등이 있다. 이를 위해 필요한 모집단 정보는 주로 통계청을 통해 획득된다. 그러나 인구수나 가구수와 같은 모집단 통계는 주로 추계자료이며 실제 전수조사를 위해 산출된 결과는 현재까지 매 5년 마다 제공되고 있다. 또한 실제 사용가능한 정보 혹은 변수의 수 역시 매우 적다. 실제 최근에는 전통적인 지역, 성 그리고 연령을 기준으로 한 모집단 대표성의 한계가 인정됨에 따라 보다 다양한 기준을 이용한 모집단 대표성의 평가가 요구되고 있다. 이러한 측면에서 빅데이터를 이용한 모집단 정보의 확보는 필요한 것으로 판단된다. 다만 이러한 정보의 타당성 및 정확성은 그 사용에 앞서 적절한 검증이 이루어져야 할 것이다.

제2절 빅데이터 분석(Big Data Analytics)

가. 머신 러닝(Machine Learning)

머신 러닝은 컴퓨터 스스로 데이터를 수집하고 분석해 미래를 예측하는 기술을 말한다. 이는 알고리즘 기반으로 컴퓨터 혹은 기계를 학습시킨 뒤 새로운 데이터를 입력해 결과를 예측하게 함으로써 이루어진다. 컴퓨터는 학습한 내용을 기반으로 방대한 양의 빅데이터를 분석해 앞으로의 행동이나 가능성을 판단하게 된다. 사람이 특정분야를 공부한 것과 같은 통찰력을 컴퓨터가 갖게 되기 때문에 머신 러닝, 우리말로 기계학습이라고 한다. 머신러닝의 역사는 50년대부터 시작되어 80~90년대에 정체기를 보였다. 이후, 2000년대 중반에 들어와서 컴퓨팅 기술의 발달과 함께 머신 러닝 분야에 현저한 발전이 이루어졌다. 소셜 네트워크 서비스의 발달, 초고속 네트워크가 보편화와 사물인터넷이 활성화되면서 엄청난 양의 데이터, 즉 빅데이터가 출현하게 된 것이다. 이러한 빅데이터를 이용하여 학습할 데이터들을 사전 처리하여 최적화함으로써 학습효과의 극대화 및 실용화가 가능한 기계학습 결과가 도출되었다. 이러한 발달의 배경에는 인터넷에 축적된 방대한 데이터와, 이의 처리를 가능하게 한 컴퓨팅 능력 향상이 있었다. 머신 러닝은 이미 우리의 일상생활에 깊숙이 들어와 있다. 일기예보, 교통신호등, 비행기 스케줄, 온라인 쇼핑, 주식시세 등은 물론이고 인터넷에서 음악을 듣거나 영화를 볼 때, 은행에서 업무를 보거나 친구와 메시지를 주고받을 때도 머신 러닝의 도움을 받는다. 특정 문장을 번역한 뒤

오역된 부분을 사용자가 직접 수정하면 번역 알고리즘이 이를 학습해 다음 번역 시에는 더 정교한 결과를 보여주는 구글의 번역 서비스나 포털 사이트 네이버에서 쓰는 '검색어 자동 완성' 기능 등도 머신 러닝을 활용한 것이다[5].

나. 딥러닝(Deep Learning)

딥러닝(Deep Learning)이란 사람의 뇌가 사물을 구분하는 것처럼 컴퓨터가 사물을 분류하도록 훈련시키는 머신러닝의 일종으로 사물이나 데이터를 분류하거나 군집화(clustering)하는 데 주로 사용하는 기술을 말한다. 1942년 한 미국 의대 교수의 아이디어에서 시작된 '딥 러닝'은 1980년대 본격적으로 개발되었다. 하지만 컴퓨터의 성능이 복잡한 계산을 처리하기엔 턱없이 부족했고 처리할 데이터도 많지 않아 사장될 뻔했었다. 그러나 컴퓨터 하드웨어의 비약적인 발달과 빅데이터의 등장으로 인공지능이 학습할 수 있는 정보가 축적되면서 2000년대 들어 각광받고 있다. 원래 물체의 특징을 인식하는 능력은 인간만이 가지고 있는 것이었다. 하지만 기계에 대량의 정보를 입력하고 이를 바탕으로 사물을 학습하고, 인식할 수 있게 됨으로써 방대한 데이터는 곧 더 정교한 인식으로 이어질 수 있게 되었다.

컴퓨터의 학습 방식은 지도 학습(Supervised Learning)과 비지도 학습(Unsupervised Learning)으로 나뉜다. 지도 학습은 먼저 컴퓨터에 분류 기준이나 정보를 미리 입력하는 방식이고 비지도 학습은 분류 기준 없이 데이터를 입력하면 컴퓨터가 스스로 분류하는 방식으로 컴퓨터는 스스로 비슷한 군집을 찾아 데이터를 분류하게 되는데 지도학습과 비교해 진보한 기술이며, 컴퓨터의 높은 연산 능력이 요구된다. 통계적으로는 지도학습은 판별분석(classification)에 해당하며 기존의 기계학습 알고리즘은 대부분 지도 학습 방식으로 데이터를 분류해왔다. 그러나 통계적 방법 중 군집분석(cluster analysis)에 해당하는 딥러닝은 보다 진보한 기술인 비지도 학습(Unsupervised Learning)을 통해 컴퓨터 스스로 데이터를 분류한다. 딥러닝의 주요 활용분야는 음성인식이나 이미지인식 분야로써 현재 구글, 페이스북 등을 비롯한 IT업계의 거대 기업들은 딥러닝을 활용한 음성인식과 이미지 인식의 비약적인 정밀도 향상을 바탕으로 목소리 인증 기술, 얼굴 인식 기술 그리고 자동차의 자율주행 기술의 개발에 박차를 가하고 있다. 실제로 구글은 2009년부터 자율주행 자동차를 이용해 385만 km를 시험주행 하였고 페이스북은 2014년 사람의 얼굴을 97.25%의

5) 머신러닝에 대해서는 Hastie, T., Tibshirani, R. and Friedman, J. H. (2001) 과 Mohri, M., Rostamizadeh, A., and Talwalkar, A.(2012)를 참조.

정확도로 알아내는 '딥 페이스'라는 기술을 개발하였다. 또한 의료 분야에서도 뢴트겐이나 CT스캔의 검사 결과를 기계에 입력하여 질병을 자동 검출하는 시스템 구축도 추진하고 있다[6].

다. 인공 신경망(Artificial Nural Network)

인공신경망은 생물학의 신경망에서 영감을 얻은 통계적 학습 알고리즘으로 인간이 뇌를 통해 문제를 처리하는 방법과 비슷한 방법으로 문제를 해결하기 위해 컴퓨터에서 채택하고 있는 알고리즘이다. 인간의 뇌가 기본 구조 조직인 뉴런(neuron)이 상호 연결되어 일을 처리하는 것처럼, 수학적 모델로서의 뉴런이 상호 연결되어 네트워크를 형성할 때 이를 신경망이라 한다. 이를 생물학적인 신경망과 구별하여 특히 인공 신경망(artificial neuralnet work)이라고 한다. 신경망은 각 뉴런이 독립적으로 동작하는 처리기의 역할을 감당하기 때문에 병렬성(parallelism)이 뛰어나고, 많은 연결선에 정보가 분산되어 있기 때문에 몇몇 뉴런에 문제가 발생하더라도 전체 시스템에 큰 영향을 주지 않으므로 결함 허용(fault-tolerance) 능력이 있으며, 주어진 환경에 대한 학습 능력이 탁월하다. 이와 같은 특성 때문에 인공 지능 분야의 문제 해결에 이용 되고 있으며, 문자 인식, 화상 처리, 자연 언어 처리, 음성 인식 등 여러 분야에서 이용되고 있다[7].

라. 마이닝(Mining)

1) 데이터 마이닝(Data Mining)

대용량의 데이터, 데이터베이스 등에서 감춰진 지식, 기대하지 못했던 경향, 새로운 규칙 등의 유용한 정보를 발견하는 과정으로 데이터 마이닝을 통해 정보의 연관성(순차 패턴, 유사성 등)을 파악함으로써 가치 있는 정보를 만들어 의사결정에 적용하는 기법이다.

2) 텍스트 마이닝

자연어로 구성된 비정형 텍스트 데이터에서 패턴 또는 관계를 추출하여

6) 딥러닝에 대해서는 Bengio, Y., LeCun, Y. and Hinton, G.(2015)과 Deng, L. and Yu, D.(2014)참조.
7) 인공신경망에 대한 개론은 Schmidhuber, J.(2015)을 참조.

가치와 의미 있는 정보를 찾아내는 마이닝 기법으로 텍스트 마이닝은 사람들이 말하는 언어를 이해할 수 있는 자연어처리 기술에 기반하고 있다.

3) 웹 마이닝

인터넷 상에서 수집된 정보를 데이터 마이닝 방법으로 분석·통합하는 기법으로 웹 마이닝은 콘텐츠 마이닝(웹 검색, 수집 데이터), 구조 마이닝(웹 사이트 구조), 활용 마이닝(사용자 이용형태) 등으로 세분화할 수 있다.

4) 소셜 마이닝

소셜 미디어에 올라오는 글과 사용자를 분석해 소비자의 흐름이나 패턴 등을 분석하고, 판매나 홍보에 적용할 수 있다. 또한 마케팅 분야뿐만 아니라 사회의 흐름과 트렌드, 여론 변화 추이를 읽어내는 소셜 미디어 시대의 새로운 마이닝 기법으로 각광 받고 있다.

5) 현실 마이닝

사람들의 행동패턴을 예측하기 위해 사회적 행동과 관련된 정보를 휴대폰 등 모바일 기기를 통해 얻고 분석하는 기법으로 현실에서 발생하는 정보를 기반으로 인간관계와 행동 등을 추론한다.

마. 분류모형(Classification)

연속형 종속변수를 대상으로 하는 예측모형에 비해, 종속변수가 범주형일 경우에 사용하는 모형 중에 한가지로 분류모형을 고려할 수 있다. 예측모형의 결과가 대상에 대한 예측값인 반면, 분류모형은 대상에 대한 모델링을 통해 각 대상을 분류해내는 분류 예측값을 결과로 가진다. 분류모형으로는 기존에 로지스틱 회귀모형, 판별분석 등이 많이 활용되었으나, 근래에는 데이터마이닝 분야의 발전에 따라 의사결정나무, 뉴럴 네트워크 등의 방법이 많이 활용된다. 로지스틱 회귀모형은 설명변수와 범주형 반응변수 사이의 관계를 나타내는 모형을 적용하기 위하여 범주형 변수의 기댓값에 로지스틱 함수를 적용한 모형으로 각 설명변수들의 값에 따라 각 개체가 특정 범주에 속할 확률이 달라지는 모형이다. 판별분석은 연속형 독립변수들의 선형 결합식을 통해

개체를 분류하는 모형이다. 의사결정나무는 의사결정 규칙을 도출하여 관심 대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석 방법이다. 각 분류모형을 구축하면, 각 대상에 대해 각 분류에 속할 확률이 계산되고 이를 대상의 참값과 비교하여 올바르게 예측한 정도를 산출하여 모형의 품질을 평가하는데, 모형의 적용 상황에 따라 데이터 일부를 선택하여 평가기준으로 활용한다.

바. 군집 분석

군집분석은 주어진 자료를 이용하여 개체들 사이의 거리 또는 유사성을 계산하고 이를 통해 전체 개체를 몇 개의 집단으로 분할하는 방법이다. 이를 통해 각 집단의 성격을 파악함으로써 데이터 전체에 대한 이해를 얻고, 전체 집단을 세분화하는 분석방법이다. 민간에서는 군집분석을 고객 세분화를 위한 빅데이터 활용 방법으로 많이 활용한다. 즉, 전체 고객을 유사한 특성을 가진 고객군으로 분류하고, 이를 다른 분석을 위한 기초자료로 활용하거나 세분화된 고객에 대해 타깃 마케팅을 시도하거나, 타깃 상품을 개발하는 등에 활용한다.

사. 연관성 분석

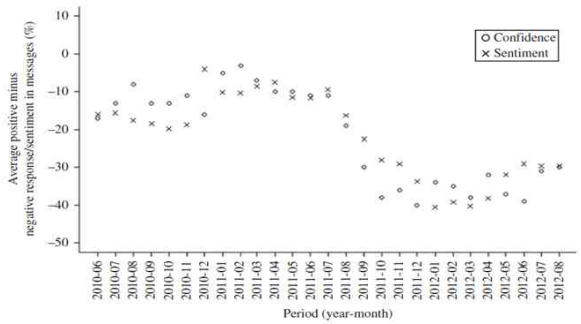
연관성 분석이란 데이터 안에 존재하는 항목간의 연관규칙을 발견하는 분석이다. 상품을 구매하거나 서비스를 받는 등의 일련의 거래나 사건들의 연관성에 대한 규칙을 찾는 과정을 일컫는다. 마케팅 분야에서 손님의 장바구니에 들어있는 품목간의 관계를 알아본다는 의미에서 장바구니 분석이라고 불리기도 한다. 민간에서는 연관성 분석을 삼품간의 판매성향 분석을 위한 빅데이터 활용방법으로 많이 활용한다. 즉, 전체 상품을 판매 관점에서 연관성이 높은 규칙을 찾아내서 이를 연관성이 높은 상품을 포괄하는 신상품이나 묶음상품을 개발하거나, 연관성이 높은 삼품들을 서로 가까이 배치하거나 연관성이 높은 상품과 판촉대상 상품을 묶는 등의 판축 활동 등에 활용한다. 또한 대용량 처리에 적합한 기법이어서 합병증 발생의 징후 탐지나 사기사건 징후 포착의 분석방법으로도 활용된다. 연관성 분석은 통계생산 과정보다는 생산된 통계의 원인 분석에 더 적합하다. 빅데이터 수집 후 대용량 자료에서 항목들 간의 관계를 자동으로 찾아내는 분야에 활용이 가능하다.

제3절 빅데이터를 공공데이터로 활용한 해외사례

가. 소비자 신뢰지수와 SNS 감성경기 지수 비교분석(네덜란드)[8]

소비자 신뢰지수란 국가의 경제 상태를 나타내는 경기선행지수로, 통화정책을 결정할 때 통화 당국자들이 관심을 두는 경제지표들 가운데 하나이다. 지수는 전체 인구 중 매월 1,000가구를 추출하여 조사를 실시하여 도출된다. 조사 문항은 현재의 지역 경제 상황과 고용 상태, 6개월 후의 지역경제, 고용 및 가계 수입에 대한 전망 등에 대한 인식 등을 5개 문항으로 구성되어 있으며, 설문 조사하여 긍정적인 답변을 한 비율과 부정적인 답변을 한 비율의 차를 이용하여 지수를 산출한다. 네덜란드의 경우 매월 20일이 포함된 1주일 동안 설문조사를 실시하여 소비자 신뢰지수를 산출한다.

SNS 감성경기 지수는 일반 국민 대다수가 활동하는 SNS 데이터를 활용하여 앞으로 국가 경제 상황을 예측하는 것이다. 실제 네덜란드 전체 인구 중 70% 이상이 SNS 활동을 하고 있어 이를 이용한 지수 생산의 실험적 연구를 진행하고 있다. 네덜란드에서는 Coosto라는 소셜미디어 수집 전문 업체를 통하여 2010년 6월부터 2012년 8월까지 Twitter, Facebook 등의 소셜미디어에서 매월 6,000만 건(총 16억 건)의 메시지를 수집하였다. 수집된 메시지는 텍스트 마이닝을 통해 긍정, 부정, 중립의 범주로 자동 분류한 후, 긍정과 부정 메시지의 개수를 비교하여 SNS 감성 경기지수를 산출하였다. SNS 감성 경기 지수는 데이터 수집과 분석 시스템을 한번 구축해 놓으면 일일단위 또는 수시로 많은 비용을 들이지 않고 지수 산출이 가능하다.



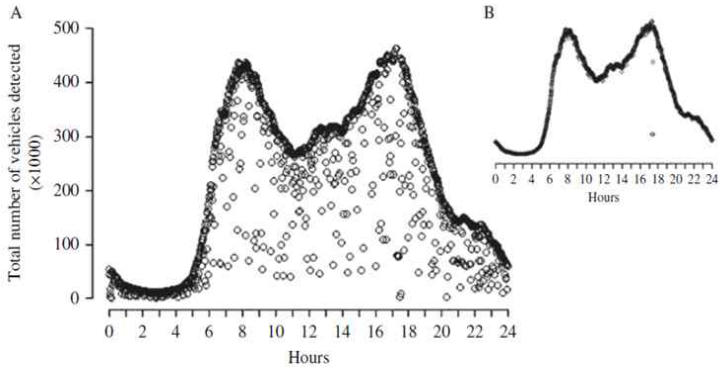
[그림 1] 소비자 신뢰지수와 SNS 감성 경기 지수 비교

8) Daas, P. J. H., Puts, M. J.(2014). Social Media Sentiment and Consumer Confidence. Statistical. European Central Bank Statistical Paper Series

상기 그림에서 ○로 표시된 것은 소비자 신뢰지수를 나타내고, ×로 표시된 것은 SNS 감성 경기 지수를 나타낸다. 그래프에서 보듯이 두 지수가 상당히 유사하게 나타남을 확인할 수 있으며 실제로 두 지수 간의 상관계수가 0.88로 상당한 연관성이 있다는 것을 알 수 있다. SNS 감성 경기 지수가 기존의 소비자 신뢰지수를 대체할 수 있다는 가능성을 확인할 수 있다. SNS 감성경기 지수는 지수 산출이 더 저렴할 뿐 아니라 보다 자주 지수를 산출할수 있다는 점에서 유용한 도구로 여겨진다.

나. 도로센서 데이터 분석을 통한 차량 통행량 예측[9)]

네덜란드에서는 도로상에 설치된 12,622개의 과속을 단속하는 센서를 활용하여 일일단위 차량의 속도, 길이 등 7,600만개의 데이터를 수집하여 분 단위로 도로의 교통 통행량을 분석하였다.

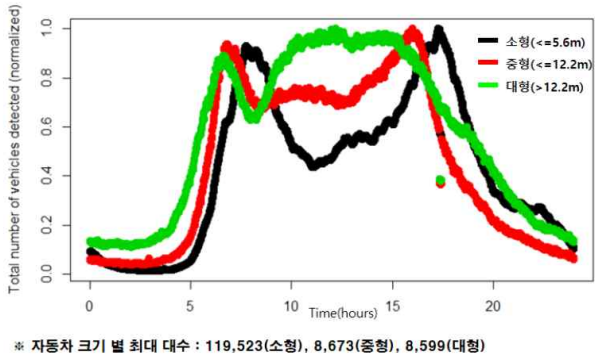


[그림 2] 네덜란드 시간대별 교통 통행량

분석결과 새벽시간대에 교통량이 없다가 출근시간과 퇴근시간대에 교통량이 많이 증가하는 것과 실제 일일 교통 통행량이 얼마나 되는지 빅데이터 분석을 통해 알 수 있었다. 위 그림의 A 그래프를 보면 센서 탐지 오류, 데이터 전송 오류 등 다양한 이유로 결측 및 이상치가 발생하여 분산이 커지게 되고 데이터로서의 가치가 떨어지게 된다. 따라서 통계적 방법을 통해 결측과

9) Daas, P. J. H., Puts, M. J., Buelens, B., and van den Hurk, P. A. (2015). Big Data as a source for official statistics. Journal of Official Statistics, 31(2), 249-262.

이상치를 보정하는 데이터 정제 과정을 거쳐서 차량 통행량을 표기한 B 그래프가 작성되었다. 위의 두 그래프에서 보듯이 빅데이터를 분석할 때 데이터 수집과정에서 다양한 오류와 노이즈가 발생하게 되는데 유의한 분석 결과를 만들어 내기 위해서는 반드시 수집된 데이터를 정제하는 과정을 거쳐야 한다. 이때 데이터를 정제하기 위해 다양한 통계적 기법들을 적용할 수 있다.



[그림 3] 시간에 따른 차량 크기별 교통 통행량(네덜란드)

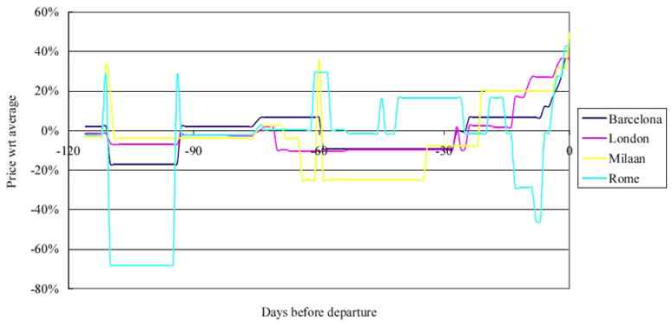
상기 그림은 도로에 설치된 센서를 활용한 차량 크기별 시간에 따른 통행량을 나타내고 있다. 분석 결과 소형차는 혼잡 시간대에 최대 통행량을 나타내는 반면 중형 및 대형 화물차는 혼잡 시간대가 아닌 시간대에 최대 통행량을 나타내는 것을 확인 할 수 있다. 국민들에게 요일별, 도로별, 시간대별 평균 교통통행량 그래프를 제공함으로써 화물 배송, 이동시 최적의 시간대를 이용할 수 있도록 유도하는 등 국민 편의 증진에 기여할 수 있다.

다. 인터넷 데이터 수집 로봇을 활용한 온라인 항공권 가격 분석[10)]

인터넷 데이터 수집 로봇이란 사람이 웹사이트에 방문하여 정보를 얻는 것과 같이 웹사이트에서 필요로 하는 정보를 자동으로 수집하는 프로그램으로 Crawlers, Spiders, Scrapers, Bots 등이 있다. 네덜란드는 수도인 암스테르담에서 출발하는 4개의 노선에 대한 출발 116일 전부터의 항공권 온라인 가격 정보를

10) Olav Bosch, Dick Windmeijer(2014). On the Use of Internet Robots for Official Statistics. Meeting on the Management of Statistical Information System(MSIS 2014).

인터넷 로봇을 활용하여 수집하였다.



[그림 4] 날짜별 항공권 온라인 가격(네덜란드)

위 그림은 4개의 도착지 별 항공권 가격의 변화를 보여주고 있다. [그림 4]에서 보는 바와 같이 날짜별 항공권의 가격 변동이 크기 때문에 소비자 물가지수(CPI) 방식과 같이 한 달에 한번 특정 시점에 가격정보를 수집하여 항공권 가격을 공표하는 것은 의미가 없다고 볼 수 있다. 인터넷의 데이터 자동수집 로봇을 활용하여 항공권 가격 정보를 수집하는 시스템을 구축한다면, 많은 비용을 들이지 않고 일일 단위 또는 수시로 데이터를 수집하여 분석하는 것이 가능하다. 따라서 항공권과 같이 가격 변동이 시간에 따라 큰 정보들은 일정기간에 수동으로 가격정보를 수집하는 것이 아니라 인터넷 로봇을 이용하여 데이터를 수시로 수집/분석 하여 국민들에게 정보를 제공하는 것이 바람직할 것이다.

4. 빅데이터 활용 통계생산 품질 검증

제1절 기존의 통계, 데이터 품질검증 기준

가. 통계품질관리

통계품질에 대한 정의는 통계가 작성되는 국가 또는 지역의 상황에 따라 다르며 시대에 따라 그 내용도 달라질 수 있다. 예를 들어, 품질은 전통적으로 단지 오래 쓰고, 질기고, 튼튼한 것 등 대상의 물리적·객관적 성질을 강조하면서 단순히 제품의 좋고 나쁜 것을 의미하는 개념으로 사용되었다. 역시 전통적 의미에서 품질 좋은 통계란 “정확하고 신속한 혹은 시의성을 갖춘 통계”라고 강조되어 왔다. 그러나 산업사회의 발전과 함께 경영자들이 점차 “고객의 욕구를 충족시킨다.”라고 하는 전략적 품질의 개념에 관심을 갖게 됨으로써 품질에 대한 주관적 요소들이 부각되었다. 따라서 오늘날 통계의 품질은 단순히 통계의 정확성, 신속성만 강조하는 것이 아니라 “통계가 얼마나 이용자에게 이용하기 적합하게 작성 및 제공되고 있는가?”를 고려한 개념을 포함해야 할 것이다. 이러한 통계품질의 개념에는 통계의 정확성, 관련성, 시의성, 접근성, 비교성, 효율성 등의 요소가 내포되어 있는데, 이 중 통계의 현실반영 정도를 나타내는 정확성은 가장 중요한 요소이며, 통계가 작성되는 모든 과정과 연관되어 있다. 이 외의 요소들도 최근 고객 지향적 정부행정이 강조되는 추세에 따라 모두 중요시 되고 있다. 이러한 의미에서 통계품질관리(Quality Management for Statistics) 체계의 개념을 정리하여 보면 “통계 이용자들에게 최대의 만족감을 주면서 동시에 가장 경제적이고 정확하며 시의성을 갖춘 통계를 생산하기 위한 모든 수단을 통합한 체계” 라고 할 수 있다.

나. 통계품질 결정요소

현대의 많은 국가에서는 최근 통계품질을 “이용자 적합성(Fitness for User)” 측면에서 정의하며 다루고 있다. 이는 기존의 품질의 평가 요소들이 이용자 적합성의 큰 범주에 포함될 수 있기 때문인 것으로 파악된다. 각 국의 통계 작성기관과 통계관련 국제기구들이 고려하고 있는 통계의 품질 결정 요소들을 보면 조금씩 다른 양상을 띠고는 있으나 정확성(Accuracy), 관련성(Relevance), 시의성(Timeliness), 접근가능성(Accessibility) 등은 대체로 공통적이다. 이 4 가지 통계품질 결정요소 외에도 통계품질 전문가나

통계작성기관에 따라 통계의 비교가능성(Comparability), 일관성(Coherence), 완결성(Completeness) 등도 품질을 결정하는 차원으로 제시되고 있다. 이러한 모든 요소들은 실제 통계 사용자들이 그 목적에 맞게 사용할 수 있는지를 판단할 수 있는 기본적인 항목들로 간주할 수 있다. 언급된 품질 검증에 대한 각 세부 요소는 이하와 같다.

1) 정확성

통계자료 및 생산된 통계의 정확성은 통계학자들과 조사방법론 학자들이 오랫동안 중요시하면서 가시적으로 평가할 수 있다고 생각해 온 품질의 결정요소이다. 정확성은 추정량과 추론의 목표로 하고 있는 모집단의 특성을 나타내는 모수의 참값이 얼마나 일치하고 있느냐를 의미한다. 정확성은 여러 가지 측면에서 정의할 수 있고 실제로 자료의 정확성을 측정하는 유일한 지표는 존재할 수가 없다. 표본조사인 경우에는 표본오차의 측정을 통하여 파악할 수 있으며 광의의 정확성에는 비표본오차까지 포함된다. 비표본오차는 커버리지 오차, 측정 오차, 자료처리 오차 등이 해당된다.

2) 관련성

통계의 관련성은 자료 이용자에게 얼마나 의미가 있고 유용한 통계를 작성하여 제공하고 있는가를 평가하는 요소이다. 즉, 통계자료가 기여하는 가치를 질적으로 평가하는 것으로 통계의 작성목적, 곧 이용자가 추구하는 목적을 어느 정도 충족시키는 가에 의하여 측정할 수 있다. 예를 들면, 20년 전부터 계속되어 온 표본조사에서 맨 처음 적용한 개념이 오늘날 사회에 적용되지 않을 수도 있다. 이럴 경우의 통계자료는 통계 이용자에게 더 이상 관련성이 없다고 할 수 있다. 통계조사에 사용되는 개념과 정의의 관련성 여부는 조사 직원, 응답자, 통계 이용자, 연구소 전문가 또는 관련 위원회 등으로부터의 전문적인 의견을 통해 판단할 수 있기 때문에 그 평가에 상당한 시간과 노력이 요구된다.

3) 시의성

통계의 시의성은 몇 가지 개념으로 이해할 수 있다. 우선 자료수집 및 생산에 소요되는 시간 즉, 자료수집에서부터 집계결과 자료를 처음 대외적으로 공표하는 시점에 이르기까지 소모되는 시간을 의미한다. 최종 이용자는 누구나

예외 없이 신속하게 자료를 받아 보길 원하기 때문에 시의성도 통계품질의 중요한 결정요소라고 할 수 있다. 또한 통계의 시의성은 자료수집 빈도를 의미하기도 한다. 시간의 경과에 따른 사회나 경제 상황의 변화를 파악하기 위해서는 일정 기간 동안 주기를 갖고 통계가 생산되는 것이 바람직하기 때문에 빈도 측면의 시의성 역기 매우 중요한 요소이다. 결국 통계의 시의성은 이용자의 필요성과 기대정도를 기준으로 판단될 수밖에 없다. 또, 시의성은 정확성과 상반관계(trade-off)에 있다고 할 수 있다. 통계의 시의성을 강조하게 되면 정확성을 확보하기 위한 시간의 부족으로 통계의 정확성 면에서 어느 정도 희생이 불가피하기 때문이다. 역으로 정확성에 치중하게 되면 적절한 시간 내에 결과자료를 생산하기 곤란한 경우가 발생할 수 있다.

4) 접근가능성

접근가능성은 통계이용자들이 통계조사 결과를 용이하게 얻어낼 수 있는지 여부와 그 정도를 의미한다. 통계조사 결과는 최종 이용자들이 쉽게 그리고 그들이 원하는 양식과 포맷으로 이용 가능할 때 가장 가치가 있다. 예를 들면, 통계결과를 다양한 매체를 통하여 이용자 편의에 맞게 작성된 통계, 중요한 조사항목에 의하여 작성된 통계표 또는 분석적이고 기술(記述)적인 분석보고서 등으로 제공함으로 다양한 통계 사용자의 접근을 용이하게 하여야 한다. 다른 의미에서 접근가능성은 조사결과가 적절히 해석되도록 통계결과 자료에 맞는 참고 및 해설 자료를 함께 제공하는지 여부에 따라 평가된다. 한편 통계의 접근 가능성은 통계 생산자가 자문, 교육과정 등을 통하여 이용자로 하여금 자료를 이용하고 해석하는데 실질적인 기술을 습득할 수 있도록 가능한 지원을 제공하는지를 측정함으로 평가될 수 있다.

5) 비교가능성

통계자료의 비교가능성은 시간 또는 공간이 서로 다른 자료 간에 신뢰할 만한 비교가 가능한 지를 평가하는 요소이다. 예를 들어, 통계자료의 시계열 자료가 제공되고 있는 지와 시계열 자료에 대한 충분한 설명 자료가 제공되는지를 통하여 비교가능성을 평가할 수도 있다. 또한 공간적으로 시·도, 구·시·군 또는 동·읍·면 단위의 자료를 제공하여 상호 비교 가능한 지와 나아가 국제적으로 비교 가능한 지도 비교가능성을 평가하기 위한 항목에 포함될 수 있다.

6) 일관성

통계품질 결정요소로서 일관성은 여러 출처에서 수집된 자료가 개념의 정의, 분류 및 방법론적인 공통기준을 근거로 집계 또는 분석되고 있는지를 평가하는 요소이다. 즉, 동일한 작성과정을 통해 생산된 통계자료들 간 또는 각기 다른 과정을 통해 작성된 자료 간에 서로 논리적으로 연결되어 있고 타당성이 있는지를 일관성은 의미한다. 통계자료는 자료 자체 내에서 또는 시계열상에서 논리적 일관성을 유지하여야 한다. 한편 다른 유사한 개념 또는 모집단 등과도 사용 용어의 정의에서 구별 가능하여야 한다.

7) 해석가능성

통계자료의 해석가능성은 통계이용자가 자료를 쉽게 이해하고 활용하며 분석할 수 있는지를 의미한다. 따라서 해당 통계와 관련된 개념, 모집단, 변수, 관련 용어 등에 대한 정의가 적절한 지와 자료가 지니고 있는 한계에 대한 정보가 함께 제공되고 있느냐를 평가하게 된다.

8) 완결성

통계자료의 완결성은 통계 이용자 집단에서 파악하려 하는 모든 영역에 대한 통계를 제공하기 위하여 통계작성 절차가 완벽하게 이루어지고 있는가를 평가하는 요소이다.

이와 같이 통계자료의 평가를 위해서는 연관 관계를 가지고 있는 여러 가지 품질요소들이 고려되어야 한다. 비록 모든 요소들이 모두 중요하지만 실제로는 모든 요소를 동시에 일정 수준 이상 만족시킬 수 없는 경우도 있다. 따라서 고정된 비용(또는 예산)범위 내에서 통계의 품질을 균형적으로 유지할 수 있는 합리적인 방안이 고려되어야 할 것이다.

다. 데이터 품질기준(한국 데이터 베이스 진흥원)[11]

이하에서는 한국데이터베이스 진흥원에서 제공하고 있는 데이터 품질 평가를 위한 요소들을 설명한다. 데이터베이스 진흥원에서는 데이터 품질의 평가를 위해 가장 중요한 판단 요소로 '해당 데이터가 목적에 부합이 되는가' 여부를 간주하고 있다. 한국데이터베이스 진흥원은 데이터가 목적에 부합되기 위해서는 다음과 같은 두 가지 조건을 만족해야 한다고 언급하고 있다. 첫째로는 제공 데이터는 목적에 유효하며 신뢰할 수 있어야 하며 둘째로는 데이터를 필요한 시점에 손쉽게 활용할 수 있어야 한다는 것이다. 따라서 아무리 유효한 데이터라 하더라도 사용하는데 제약사항 등이 있어 활용하기에 어려움이 있다면 데이터의 품질이 우수하다고 볼 수 없다는 기준을 가지고 있다. 아래 표는 한국데이터베이스 진흥원에서 제공하고 있는 지표들에 대한 설명을 요약하고 있으며 각 지표에 대한 한국데이터베이스 진흥원의 데이터 품질 기준은 다음과 같다.

<표 5> 데이터 품질 기준

지표			정의
유효성	정확성	사실성	데이터가 실세계의 사실과 동일한 값을 가지고 있어야 함
		적합성	데이터 값이 정해진 유효 범위 충족하고 있음
		필수성	반드시 필요한 필수 항목에 데이터의 누락이 발생하지 않음
		연관성	연관 관계 갖는 데이터 항목간에 논리상의 오류가 없음
	일관성	정합성	기능, 의미, 성격이 동일한 데이터가 상호 동일한 용어와 형태로 정의
		일치성	동일한 용어를 동일한 용도로 사용
		무결성	데이터 처리의 선후 관계가 명확하게 준수되고 있음
	대표성	대표성	수집된 데이터가 전체 모집단을 대표할 수 있어야 함
활용성	유용성	충분성	사용자의 요구 사항을 충분히 충족시킬 수 있음
		유연성	사용자의 요구사항을 수용할 수 있는 유연한 구조임
		사용성	실제 공급되는 데이터가 현장에서 유용하게 사용되고 있음
		추적성	데이터의 변경 내역이 관리되고 있음
	접근성	접근성	사용자가 원하는 데이터를 손쉽게 이용할 수 있어야 함
	적시성	적시성	사용자가 원하는 데이터를 원하는 시간에 제공 받을 수 있어야 함
	보안성	보호성	훼손, 변조, 유출 등의 다양한 형태의 위협으로부터 데이터를 안전하게 보호하고 있어야 함
		책임성	사용자 접근 권한과 책임을 명확히 부여하고 있어야 함
		안전성	시스템의 에러나 장애를 사전에 차단하고 에러나 장애 발생 시 중단 및 지연을 최소화할 수 있는 체제로 운영해야 함

11) https://www.dqc.or.kr/main/index.html

1) 정확성

정확성은 세계에 존재하는 객체(사건, 사물, 개념 등)의 값이 오류 없이 저장되어 있음을 의미한다. 객체들의 데이터가 정확성을 확보하기 위해서는 객체를 표현하기에 필수적으로 필요한 정보가 누락되어서는 안 된다. 이러한 정보는 사전에 정의한 규칙과 형태대로 저장되고 관리되어야 하며 이러한 측면에서 정확성은 사실성, 적합성, 필수성, 연관성과 같은 품질 기준을 통해 평가될 수 있다.

(1) 사실성

사실성은 데이터가 관측 대상의 참값과 동일한 값을 가지고 있음을 의미한다. 일반적으로 사실성 오류는 데이터의 원천 오류, 입력 오류, 입력 프로세스 문제 등으로 인한 데이터 자체의 오류를 포함하고 있다. 또 다른 형태는 데이터가 나타내는 현상이나 모수는 바뀌거나 변했는데도 데이터는 이를 반영하지 못하는 경우에 사실성은 훼손되게 된다.

사실성 오류를 줄이기 위해서는 두 가지 개선점이 있다. 첫 번째 개선점으로는 데이터 소유권 관리와 입력 프로세스 개선 등이다. 그리고 두 번째는 모집단의 변화를 정확하게 관리할 수 있는 데이터 구조와 프로세스 등의 지원이다.

(2) 적합성

적합성은 데이터 값이 정해진 데이터 유효 범위를 충족하고 있음을 의미한다. 예컨대 적합성 오류는 데이터가 표준 코드 값 또는 표준 도메인 값에 위배될 경우 발생한다. 따라서 적합성 향상을 위해서는 표준 도메인 및 코드에 대한 개선이 필요하다. 적합성을 충족하기 위해서는 전시된 데이터 구조에 사용된 데이터 항목이 표준 도메인, 표준 코드의 범위 값을 준수하여야 한다.

(3) 필수성

필수성은 조직의 업무 지원을 위해 반드시 필요한 필수 항목에 데이터의 누락이 발생하지 않음을 의미한다. 필수성 오류는 데이터 수집 당시의 확인 부족이나 데이터베이스상의 'Not Null' 체크 조건의 누락 등이 원인이

되어 발생한다. 데이터를 설계하는 단계에서 해당 데이터 항목의 필수성에 대한 정의를 확인하고 이를 확인하는 것이 일반적이다. 가능하면 설계 단계에서의 필수성 정의를 데이터베이스 생성 때에 준용하게 해야 한다. 또한 데이터별로 중요도를 선정하여 관리할 필요가 있다. 만약, 필수성 조건이 설계 단계에서 누락되어 실제 데이터베이스에 'Not Null' 체크 조건이 누락된 채로 데이터가 생성된다면 향후에 해당 데이터에 대한 재현이 불가능한 상황을 초래할 수 있다.

(4) 연관성

연관성은 연관 관계를 갖는 데이터 항목 간에 논리상의 오류가 없음을 의미한다. 예컨대 '해지 일자'가 '가입 일자'보다 앞서 있는 경우가 여기에 해당된다. 이와 같이 데이터는 서로 업무적인 연관성을 가지는 경우가 있는데 업무적 연관성에 오류가 없어야 한다.

2) 일관성

일관성은 정보시스템 내의 동일한 데이터 간에 불일치가 발생하지 않음을 의미한다. 일관성 오류는 데이터에 대한 정의가 정확히 이루어지지 않거나, 데이터 참조 무결성(referential integrity)이 불분명한 경우, 개별 시스템 단위로 설계하고 관리되어 전사 관점의 접근이 부족한 경우 등의 원인에 의해 발생한다. 예를 들면, 조직 내의 구매부와 판매부에서 관리하는 '고객' 데이터에 대한 정의는 부서별 업무 특성에 따라 달라질 수 있다. 구매부에서 관리하는 '고객' 데이터는 주로 조직 또는 기관이 해당된다. 반면에 판매부에서 관리하는 '고객' 데이터는 주로 개인 또는 조직이 해당된다. 각 부서별로 관리하고자 하는 고객에 대한 데이터 항목이 다를 수도 있고, 심지어는 각각 다른 형태로 고객 데이터를 정의하여 관리할 수도 있다. 이러한 상황에서는 같은 고객이라고 하더라도 각 부서별로 데이터가 일치하지 않고 일관성이 지켜지지 않을 수도 있다. 따라서 일관성을 확보하기 위해서는 전사(轉寫) 관점의 데이터 용어 표준화, 데이터 구조의 전사적인 조율, 중복 데이터에 대한 관리 등을 통해 데이터 간의 통일성을 유지하는 것이 필요하다. 일관성은 정합성, 일치성, 무결성과 같은 품질 기준으로 구성된다.

(1) 정합성

정합성은 기능, 의미, 성격이 동일한 데이터가 동일한 값으로 정의되어 있음을 의미한다. 특히, 다양한 이유로 중복 관리되고 있는 데이터 간의 값이 동일해야 한다. 원천 데이터를 가공하여 새로운 데이터를 생성하였더라도 반드시 원천 데이터와 동일한 값을 유지하도록 하는 것은 데이터 품질에서 매우 중요하다.

(2) 일치성

일치성은 동일한 용어를 다르게 정의하여 사용하지 않도록 함을 의미한다. 한 예로 남녀의 성별을 구분하는 용어인 성별 구분이 단위 시스템별로 '남녀구분', '성별구분', '성별코드', '남녀코드'로 용어가 서로 다르게 사용되면 전사 관점에서 데이터 통합 시에 시스템 자원의 낭비를 가져오고 데이터 정확성의 확보에도 영향을 미친다.

(3) 무결성

무결성은 데이터 처리의 선후 관계가 명확하게 준수되고 있음을 의미한다. 데이터 구조를 설계할 때에는 데이터의 생성 순서를 배제하고 구조를 설계하게 된다. 하지만, 실제 데이터가 생성되고 변경될 경우에는 선후관계가 존재할 수 있다. 예를 들어 '고객' 데이터가 생성되고 '판매' 데이터, '반품' 데이터가 생성되어야 한다는 데이터의 선후 관계를 나타낸다. '고객' 데이터가 삭제되면 '판매' 데이터도 함께 삭제되어야 하는 경우가 여기에 해당된다. 무결성은 참조무결성을 데이터 구조 설계의 과정에서 반영하고 데이터베이스나 프로그램을 통해 이를 구현함으로써 확보할 수 있다.

3) 유용성

유용성은 조직이 요구하는 데이터의 범위와 상세화 정도를 충족시킬 수 있음을 의미한다. 유용성 확보를 위해서는 사용자의 요건에 대한 관리가 필요하다. 요건 관리는 사용자들의 추상적인 요구를 구체화하여 관리하는 제반 활동을 의미한다. 이를 위해서는 도출된 요건을 정해진 기준과 원칙에 따라 요구사항 수행 범위와 우선순위를 결정하여 반영하고 일련의 결과를 문서화하여

관리해야 한다. 유용성은 충분성, 유연성, 사용성, 추적성과 같은 품질 기준으로 구성된다.

(1) 충분성

충분성은 제공 데이터가 사용자의 요구사항을 충분히 충족시킬 수 있음을 의미한다. 충분성의 확보를 위해서는 사용자의 현재 요구사항 뿐만 아니라, 가까운 미래의 변화나 확장 가능성에 대해서도 데이터 구조에 최대한 반영하는 것이 필요하다.

(2) 유연성

유연성은 데이터가 사용자의 다양한 요구사항을 수용할 수 있는 유연한 구조를 가지고 있음을 의미한다. 즉, 데이터 구조는 다양한 요구사항과 업무를 수용할 수 있어야 하며 변경사항 발생 시에도 유연하게 대처할 수 있어야 한다. 데이터 구조의 유연성을 위해 데이터 집합을 최대한 통합하는 것이 필요하다. 심지어 업무적으로는 다소 이질적인 데이터라고 하더라도 형태상이나 활용상 동질성이 있다면 통합할 수도 있다.

(3) 사용성

사용성은 실제 공급되는 데이터가 현장에서 유용하게 사용될 수 있어야 함을 의미한다. 사용성 향상을 위해서는 데이터 사용 현황 및 사용 빈도를 분석하여 활용 빈도가 높은 데이터에 시스템 자원을 우선 배분하는 등의 활동이 필요하다. 만약, 일정 기간 동안 사용 빈도가 없다면 해당 데이터에 대해서는 원인을 분석하고 사용성 향상을 위한 방안을 강구하거나 관리 대상 제외 등의 활동을 수행해야 한다.

(4) 추적성

추적성은 데이터의 변경 내역이 관리되고 있음을 의미한다. 모든 데이터에 대해서 변경 이력을 관리할 필요는 없다. 하지만, 변경이력 정보가 꼭 필요한 데이터에 대해서는 반드시 추적성에 대한 방안이 있어야 한다. 추적성은 업무 상황의 변화에 따라 최초 설계 시에는 필요 없다고 판단되었던 것이 나중에 그 필요성이 제기되는 경우도 있다. 데이터에 대한 추적성 확보를 위한 데이터 이력 관리는 일반적인 데이터 저장과 비교해서 비용이 많이 들어간다. 그렇기 때문에

상시적으로 각 데이터에 대한 추적성 확보 필요성이 파악되고 반영되어야 한다.

4) 접근성

데이터에 대한 사용자 만족도를 충족시키기 위해서는 데이터 자체의 품질은 물론, 데이터를 효과적이고 효율적으로 제공하는데 필요한 인터페이스 등 제반 서비스 품질을 확보해야 한다. 접근성은 사용자가 원하는 데이터를 손쉽게 이용할 수 있음을 의미한다. 접근성은 사용의 용이성 관점과 검색의 용이성 관점으로 나누어 살펴 볼 수 있다. 사용의 용이성은 정보시스템에서 인터페이스, 도움말, 고객지원 등이 사용자가 데이터를 이용하는데 불편함이 없도록 제공되고 있음을 의미한다. 검색 용이성은 정보시스템에서 제공하는 검색 관련 제반 기능이 사용자가 원하는 데이터를 손쉽고 편리하게 추출하여 활용할 수 있도록 지원되고 있음을 의미한다.

5) 적시성

적시성은 응답시간과 같은 비기능적 요구사항 및 데이터의 최신성 유지와 같은 품질요건에 얼마나 잘 대처하고 있는지를 의미한다. 적시성은 특히 운영시스템의 데이터를 분석 데이터로 변환하여 조직 내부의 의사 결정 등에 활용할 경우 그 중요성이 더욱 커진다. 이는 적시성이나 최신성이 확보되지 못한 데이터로 조직의 중요한 의사 결정을 내리게 될 경우 치명적인 결과를 가져올 수 있기 때문이다. 특히 관리해야 할 데이터의 양이 기하급수적으로 증가하면서 데이터베이스에 대한 지속적인 최적화 작업이 필수적인 데이터 품질 요건으로 여겨지고 있다. 적시성을 향상시키기 위해서는 데이터베이스 튜닝을 통한 성능 개선, 최적의 데이터 구조 유지, 자원의 효율성 확보를 위한 정보의 생명주기 관리 등을 지속적으로 수행해야 한다.

6) 보안성

보안성은 외부 및 내부 요인으로부터 데이터를 적절하게 보호하고 있는지 여부를 의미한다. 여기에는 외부 침입이나 재난 등의 위험으로부터의 데이터 보호, 시스템의 오류나 장애 원인의 사전 차단, 오류나 장애 발생 시 데이터 사용의 중단 및 지연 최소화, 명확한 사용자 권한 정의 등이 포함된다. 보안성은 보호성, 책임성, 안정성과 같은 데이터 품질 기준으로 구성된다.

(1) 보호성

보호성은 훼손, 변조, 유출 등의 다양한 형태의 위협으로부터 데이터를 안전하게 보호하고 있음을 의미한다.

(2) 책임성

책임성은 사용자 접근 권한과 책임을 명확히 부여하고 있음을 의미한다. 데이터 보호 및 사용자 권한 접근 관리에서 중요한 것은 보호 대상에 대해 정확한 정의를 내리고 이에 대한 관리 체계를 구성하는 것이다. 보안 활동은 관리적 보안 활동과 기술, 물리적 보안 활동으로 구분할 수 있다. 관리적 보안 활동에서는 사람, 문서, 재해 복구 계획 등을 주요 관리 대상으로 선정해 이에 대한 조직 보안 규정 및 지침을 수립하여 시행해야 한다. 기술 및 물리적 보안 활동에 물리적 데이터베이스 저장소에 대한 출입 통제 관리와 외부에 의한 해킹 피해, 바이러스에 의한 피해로부터 보안 활동을 수행해야 한다.

(3) 안전성

안전성은 시스템의 오류나 장애를 사전에 차단하고 오류나 장애 발생 시 중단 및 지연을 최소화할 수 있는 체제로 운영함을 의미한다. 데이터 소실 사고는 다양한 요인으로 인해 언제 어디서나 일어날 수 있기 때문에, 오류 발생 시 데이터를 효율적으로 빠른 시간 안에 복구할 수 있는 절차가 필요하다.

이상으로 살펴본 통계 품질 및 데이터 품질의 평가 및 관리를 위해 고려되는 요소들은 상당히 유사하다. 통계 품질의 경우에는 보다 전통적인 방법, 즉 조사나 보고 등을 통한 통계 작성 시 그 과정과 최종 결과인 통계의 제공을 위해 고려할 수 있는 품질의 평가 요소들이 검토되었다. 데이터 품질 측면에서도 품질 유지를 위한 자료 수집상에서 요소들도 고려되고 있으나 그 보다는 데이터와 시스템의 관리 측면을 더욱 강조한 품질의 평가 및 관리를 고려하고 있다. 다음 절에서는 본 절에서 언급된 내용과 더불어 빅데이터의 품질관리를 위해 필요한 요소들과 가능한 측정 방안을 위한 틀(frame)을 대략적으로 제시한다.

제2절 빅데이터 통계 품질 검증

가. 개요

다양한 형태의 데이터를 다루는 공공기관 및 국가 통계 기관에게 빅데이터 시대의 도래는 각 기관들이 다루어야 할 데이터의 출처가 확대 및 다양화 되었으며, 처리해야 할 데이터의 양의 폭발적으로 증가하였다는 것을 의미한다. 국가 통계 기관에서 생산하는 통계는 정책 결정을 위해 사용될 뿐 아니라 공식적으로 인정되는 국가의 상태를 나타내는 수치로써 매우 중요하고 따라서 이에 맞는 품질 기준에 따른 검증이 필수적이다. 기존의 승인 통계의 경우에는 그 형태에 따라 적절한 품질 기준을 정하고 정기적, 비정기적인 품질검증이 이루어지고 있다. 빅데이터를 이용하여 작성된 통계를 기존의 공공기관 및 국가 통계 기관 작성 통계와 동일한 품질 수준으로 유지하는 것은 그 복잡성과 실제 통계 생산과정 전체를 관리할 수 없다는 측면에서 불가능할 것이다. 따라서 빅데이터를 이용한 통계 생산 및 분석은 일정 기간의 실험 연구를 통해 안전성이 확보된 이후에 그 사용 및 자세한 검증과정에 대한 논의가 이루어져야 할 것이다.

이러한 상황에서 비록 빅데이터를 활용한 통계의 실제적 활용이 매우 제한적일 수밖에 없으나 이 후 이를 활용한 통계 생산이 보편화 될 것이 예상되기 때문에 빅데이터 활용 통계의 검증을 위한 품질평가의 대략적인 틀을 마련하는 것이 현 시점에서 요구되고 있다. 빅데이터는 그 출처, 형태 그리고 환경 등의 요소에 따라 매우 다양하고 따라서 각 경우에 대응하는 품질 평가에 대한 기준들과 절차들이 별도로 마련되어야 하나 이러한 맞춤형 품질 진단 과정은 각 빅데이터 활용 통계 작성 시 함께 논의되어야 할 것이다. 본 연구에서는 이를 위해 기본적으로 요구되는 품질 진단의 큰 틀을 제안한다. 이를 위해서 본 연구에서는 빅데이터 품질에 대한 UN 유럽 경제 위원회(United Nations Economic Commission for Europe)의 논의를 바탕으로 품질 평가를 위한 대략적인 틀을 제안한다.

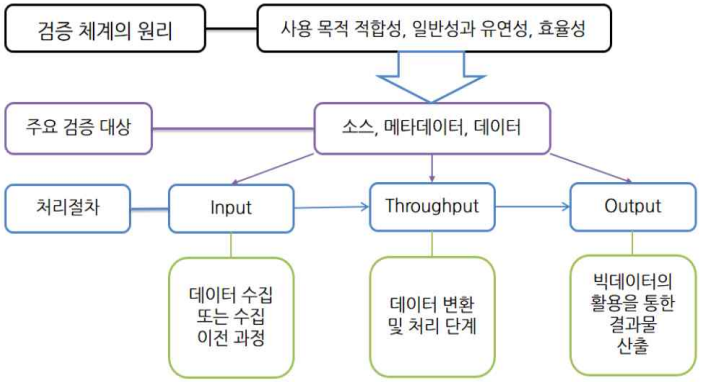
기존의 공공 데이터를 활용한 통계는 데이터를 수집하여 분석한 결과물의 품질 검증에 주로 중점을 두었다. 이는 데이터의 수집 단계에 공공기관이 직접 관여하고 또한 이를 관리하기 때문이다. 특별히 통계청이 통계작성기관으로 지정한 기관에서 작성된 통계의 경우에는 데이터 수집단계부터 통계 생산 단계까지 매우 까다로운 검증이 이루어지고 있다. 공공데이터 및 이를 이용한 통계의 품질을 검증하는 과정과 빅데이터의 검증과정이 차별적으로 정의되어야

하는 이유는 기본적으로 빅데이터를 이용한 분석 혹은 데이터 증대의 방안들이 매우 다양하기 때문이지만, 보다 근본적인 이유는 데이터의 수집 단계에서 나타나는 차이 때문이다. 승인통계 작성을 위해 수행되는 서베이의 경우에는 조사를 통해 작성되어야하는 통계가 정의되고 이의 작성을 위한 서베이 과정의 모든 절차가 결정된다. 즉, 자료 수집 도구, 측정 방안, 조사 도구, 표본설계, 실사 과정 그리고 마지막 통계 생산 단계까지의 모든 서베이 과정이 통계 작성 목적에 맞추어서 기획되고 수행된다.

반면에 현재까지 대부분의 빅데이터를 활용한 통계 분석의 경우에는 데이터 수집 목적이 서로 다른 자료들을 수집하게 된다. SNS와 같은 소셜 미디어 자료의 경우에는 다양한 개인이 제공하는 정보를 수집하게 되고 자료를 제공하는 각 개인의 목적은 매우 다양하다. 따라서 실제 빅데이터의 분석을 위해 수집된 자료는 기본적으로 분석자에 의하여 모든 자료가 정제되어야 하고 그 이전에 우선 자료의 타당성 여부가 결정되어야 한다. 이는 데이터 수집 단계에서 연구자나 사용자가 개입할 수 없는 빅데이터 분석의 구조상의 문제로 볼 수 있다. 따라서 일반적으로 생산된 통계 혹은 연구자가 직접 개입하여 설계를 조정할 수 있는 데이터 수집 단계에 대한 검증이 이루어지는 전통적인 검증방안과 빅데이터 검증방안은 차별적으로 정의되고 적용되어야 한다. 또한 표준화된 자료 수집 및 분석 절차를 밟게 되는 공공데이터의 경우와는 달리 빅데이터 분석을 위해서는 공공데이터 뿐 아니라 민간데이터가 함께 고려되고 언급한 여러 다양한 형태의 자료 수집 과정 및 다양한 분석방안을 통한 결과가 제공되기 때문에 보다 복잡한 형태의 품질검증 절차를 고려하여야 할 것이다.

나. 빅데이터 품질 검증 체계의 틀

빅데이터를 활용한 통계 생산 과정의 품질검증은 우선 통계 생산 절차 혹은 흐름에 맞추어 접근하는 것이 상식적인 것으로 판단된다. 큰 틀에서 이 과정은 데이터의 수집을 포함한 입력(Input)단계, 입력된 데이터를 가공 및 처리하여 실제 분석 가능한 형태로 변환하는 처리(Throughput)단계 그리고 정제된 자료를 활용하여 통계를 생산하거나 분석결과를 제공하는 결과(Output)단계로 정의할 수 있다.



[그림 5] 빅데이터 품질 검증 체계

상기 그림은 도식화된 품질검증 체계를 보여주고 있다. 언급한 바와 같이 빅데이터 활용 통계의 품질 검증을 위한 틀은 현행 국가 통계 작성 기관에서 작성하는 통계에 대한 품질과정과 유사하나 보다 다양한 관점에서의 검증이 이루어져야 한다. 이를 위해 각 단계 별 검증이 이루어져야 하며 그 구체적인 내용은 빅데이터의 형태에 맞추어서 정의되어야 할 것이다.

각 단계에서 공통적으로 고려되어야 하는 검증의 대상이자 도구는 크게 세 가지로 나눌 수 있다. 이는 각기 데이터 출처(source)[12], 메타데이터[13]이며 마지막으로는 수집된 데이터 자체이다.

언급한 처리절차에 따라 각 대상의 검증을 위한 검증 체계 방안 수립을 위한 기본적인 원리 1) 사용 목적 적합성, 2) 일반성과 유연성, 3) 비용 대비 효율성을 고려할 수 있다. 실제 수립된 검증체계는 기본적으로 기존의 품질검증 체계와의 일관성을 유지해야 하며, 빅데이터 검증과 관련된 다양한 측면들이 포함되어야 하고 또한 현실에서 이를 수행하기 위해서는 가능한 한 단순한 방안이 되어야 할 것이다. 이러한 측면에서 빅데이터 품질 검증 체계에 요구되는 3가지 원리가 제안되었고 각각에 대한 부차적인 설명은 다음과 같다.

- 사용목적적합성: 기존의 통계 검증 체계에 있어서도 가장 중요한 요소로서 검증체계 자체의 사용목적 적합성과 더불어 실제 빅데이터 및 이를 활용해 생산된 데이터의 품질평가에 있어서 가장 기본적인 요소이다.
- 일반성과 유연성: 기존의 검증체계가 정형화 혹은 표준화된 데이터를 그 대상으로 하기 때문에 상대적으로 일반성과 유연성이 중요하지 않은 요소였다. 하지만 빅데이터의 경우 그 형식과 분석 방안이 다양하기 때문에 모든 가능한 빅데이터 활용 통계 방안에 적용할 수 있는, 유연성과 일반성을 갖춘 검증체계를 작성하는 것이 요구된다.
- 비용대비 효율성: 빅데이터 및 이를 활용한 통계의 품질 검증을 위해서는 기존의 정형화된 절차에 대한 검과 비교하여 많은 비용과 시간이 필요하다. 따라서 결과의 사용 목적과 요구되는 통계적 정도와 균형을 맞춘 검증 절차 체계가 필요하다.

언급한 3가지 기본적인 원리를 바탕으로 빅데이터 품질 검증 체계를 수립함에 있어 실제 빅데이터 혹은 이를 통해 생산된 통계의 품질을 평가하기 위해 각 단계에서 고려되는 차원(dimension)은 다음과 같다.

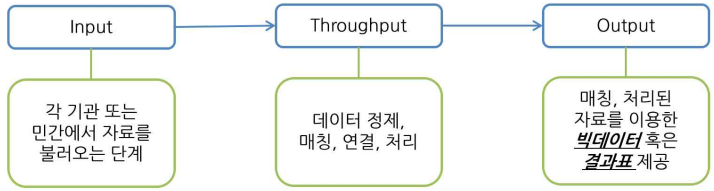
- 기관 및 사업체 환경: 자료의 제공 혹은 자료의 처리를 담당하는 기관이나 사업체 환경은 실제 생성되는 빅데이터와 이를 활용한 통계의 품질에 매우 중요한 영향을 미치게 된다. 따라서 품질 검증을 위한 중요한 차원으로 고려하여야 한다.
- 개인정보 관련 보안: 실제 제공받게 되는 자료와 공표되는 자료에 대한 법적 문제들을 고려하고 필요한 경우에는 비식별화와 같은 과학적인 방안들이 사전에 고려될 필요가 있다. 이 부분이 자료 입력부터 최종 결과 단계까지 적절하게 이루어지고 있는지에 대한 검증이 요구된다.
- 복잡성: 일반적으로 수집되는 자료에 적용되는 차원으로 자료구조의 복잡성, 자료 형식의 복잡성, 자료 출처의 복잡성 그리고 계층구조의 복잡성을 의미한다. 이러한 복잡성은 실제 자료 처리 단계에서 소요되는 시간과 비용에 영향을 미치고 경우에 따라서는 빅데이터 분석의 비용 대비 효율성을 매우 저하시키는 요인이 될 수 있기 때문에 분석 이전에 충분한 검토가 이루어져야 한다.
- 완전성: 제공된 자료에 대한 충분한 정보가 제공되었는지를 판단하기 위한 차원으로 데이터에 대한 설명과 그 내용이 포함된 메타데이터의 완전성을 충분히 검토되어야 한다.
- 유용성: 빅데이터는 그 정의상 매우 다양한 형태와 구조를 가지고 있고 따라서 이의 사용을 위해서는 상당한 수준의 전문성이 요구될 수 있다. 즉, 자료 입력 단계로부터 통계 생산 단계까지의 모든 과정에서 결과

12) 데이터의 형태와 데이터를 수집한 출처의 특성과 관련이 있는 요소로서
13) 데이터의 수집과정 및 데이터 자체를 설명하는 요소로서

제공을 위한 추가적인 비용이나 인원이 필요한지에 대한 검토 및 이의 유용성에 대한 사전 검증이 요구된다.

- 시간적인 요소: 검증을 위한 다른 요소에 있어서 빅데이터는 상대적으로 기존의 방안들보다 약점이 많은 반면 시간 및 공간과 관련된 요소에서의 장점이 많다. 따라서 시의성, 짧은 주기의 반복적인 결과 제공, 대상 지역 포함률 확보 등에 대한 검증이 이루어져야 한다.
- 정확성: 기존 자료에 요구되는 품질 요소와 동일한 것으로 판단할 수 있으며 통계적인 측정이 가능한 랜덤 오차 및 응답률과 같은 참고자료로 측정이 가능한 커버리지 오차, 무응답 오차 그리고 측정 오차에 대한 검증이 필요하다.
- 대표성: 빅데이터 활용에 있어서 가장 문제시 되고 있는 차원으로 이를 활용한 분석 혹은 통계가 가지는 선택 편향을 고려해야 한다는 것이다. 즉 빅데이터를 활용한 통계가 모집단을 대표하는 특성치의 추정량으로 고려되기 위해서는 이에 대한 대표성 검증이 반드시 선행되어야 한다.
- 일관성: 빅데이터 내 및 외부 자료와의 일관성을 갖추고 있는지에 대한 검증이 요구된다. 이는 자료의 표준화 여부, 타 자료와의 연계 가능성 여부 그리고 시간 및 공간 변화에 따른 타당성 및 일치성을 확보하고 있는지에 대한 검증을 포함하고 있다.
- 타당성: 본 검증 차원은 측정 오차와 밀접한 관계가 있다. 대부분의 빅데이터, 특히 소셜 빅데이터는 자료의 측정을 위한 체계적인 연구 없이 수집된다. 따라서 이의 활용을 위해서는 실제 사용자가 원하는 데이터와 빅데이터에서 측정된 수치와의 체계적인 차이 여부를 검증하고 사용할 필요가 있다.
- 접근성 및 명확성: 빅데이터의 접근 용이성 그리고 빅데이터에 대한 명확한 정보의 제공이 이의 바른 활용을 위하여 요구된다.
- 관련성: 빅데이터 활용 통계 혹은 분석 결과가 이의 사용자의 목적에 부합하는지를 평가하는 요소로서 빅데이터 활용 측면에서 그 결과물의 평가에 있어서 매우 중요한 요소이다.

이러한 빅데이터 품질 검증 체계를 빅데이터를 활용한 통계생산 방법론으로 본 연구에서 분류한 데이터 증대와 빅데이터 분석에 적용하면 아래와 같다. 다음의 [그림 6]은 데이터 증대 측면에서 생산된 통계에 대한 빅데이터 품질 검증 체계 적용 방안 예이다.



[그림 6] 데이터 증대에 대한 빅데이터 품질 검증 체계

Input단계는 각 기관 또는 민간에서 자료를 불러오는 단계를 의미하고 Throughput단계는 불러온 데이터를 정제하고 적절히 매칭, 연결하여 처리하는 단계를 의미한다. 마지막으로 Output 단계는 매칭, 처리된 자료를 이용하여 빅데이터를 생산하고 생산된 통계를 활용하여 결과표를 제공하는 단계를 말한다.

예를 들어, 2015년 수행된 등록센서스의 경우에 Input 단계는 11개 기관으로부터 자료를 수집하는 과정으로 이해할 수 있다. 이 경우 자료가 구축된 기관들이 모두 국가기관으로 보고 혹은 신고 형식으로 관리되는 전수 자료이며 따라서 빅데이터 생성을 위해 사용하는 기초자료로써의 신뢰도는 민간자료보다 높다고 판단할 수 있다. 이 후 Throughput 단계에서는 불러들인 자료를 표준화하고 정제하는 작업이 이루어진다. 예를 들어, 기존 거주 기반 인구주택총조사와의 일관성 유지를 위해서 이 과정에서는 등록 거주지와 실제 거주지의 차이가 있는 부분, 주소 오류가 있는 부분, 주택 유형이 서로 다른 경우들을 찾아내고 서로 다른 기관에서 제공된 데이터를 이용하여 수정 보완하여 생성될 자료와 기존 자료와의 일관성을 확보하는 작업이 이루어진다. 마지막 Output 단계에서는 구축된 전수 자료를 바탕으로 2015년 시점 대한민국의 전체 인구수 및 가구수를 포함하여 각 지역 별 인구 및 가구 통계가 작성된다.

다음으로 빅데이터 분석 측면에서 생산된 통계에 대한 빅데이터 품질 검증 체계 적용 방안으로 앞서 제시한 도로 센서 데이터 분석을 통한 통행량 분석 예시를 활용하여 살펴보면 다음의 [그림 7]과 같다.

[그림 7] 빅데이터 분석 측면에서 생산된 통계에 대한 빅데이터 품질 검증 체계

Input단계는 도로에 설치된 센서로부터 데이터를 수집하는 단계로서 자동적으로 데이터가 축적되는 단계이다. 이 단계에서는 센서를 통해 측정된 수치가 그대로 전송되기 때문에 측정 알고리즘에 입력된 방식대로 자료를 처리하게 된다. 가장 기본적인 원시 알고리즘이 사용된 경우에는 이상치나 여러 잡음(noise)이 포함된 내용이 전송되고 축적된다. Throughput단계는 데이터 수집과정에서 센서 오류로 발생한 결측치 및 오작동으로 인한 이상치 등을 정제하고 처리하는 단계를 의미하며 이 과정을 통해 구축된 데이터는 실제 통계를 생산하거나 분석을 실시하기 위한 최종 자료의 형태를 갖추게 된다. 마지막으로 Output 단계는 처리된 데이터 분석을 통하여 시간대별, 차량 크기별 통행량 그래프를 생산하여 제공하는 단계를 말한다.

다. 단계 별 품질검증지표 개발을 위한 문항 사례

빅데이터를 활용한 통계 혹은 이를 이용한 분석결과에 대한 품질검증의 최종 목표는 언급한 각 평가 차원 별 혹은 그 상위 범주에서의 평가지표를 작성하는 것이다. 이러한 목표를 달성하기 위한 첫 단계는 지표 작성을 위한 평가 문항을 개발하는 것이다. 평가 문항은 단순한 “예/아니요”의 답변을 요구하는 질문부터 리커트 척도를 답변으로 활용한 질문 그리고 질적 평가를 위한 서술형 답안을 요구하는 질문으로 다양하게 구성할 수 있다. 본 연구에서는 각 단계에서 고려할 수 있는 질문 항목의 예들을 살펴본다.

1) Input 단계에서의 질문 항목의 예

기존의 검증과 비교하여 빅데이터 품질 검증의 특별한 부분이자 매우 중요한 단계는 입력단계에 대한 검증이다. 일반적으로 빅데이터 활용 혹은 분석을 위한 원 데이터는 통계 작성 기관이나 분석 주체 기관이 아닌 곳에서 수집되기

때문에 자료 입력 단계에서의 품질 검증은 매우 중요하다. 이러한 측면에서 본 절에서는 다양한 데이터에 적용할 수 있는 일반적인 질문들을 제안하였다.

- 기관 및 사업체 환경

① 데이터 제공 기관이 국가통계기관이 요청하는 품질 기준을 충족하지 못할 경우 발생하는 위험은?
② 사업체가 추 후 빅데이터를 제공할 수 없을 때 발생하는 위험은? 데이터를 제공받지 못할 때 대체 가능한 빅데이터 출처가 존재하는가?
③ 자료가 짧은 기간 동안 제공된다면, 이는 주어진 기간 내에 사용 목적에 부합하다고 할 수 있는가?
④ 자료가 사용목적에 부합하다고 여기기 위해서는 얼마나 오랫동안 제공되어야 하는가?
⑤ 빅데이터 출처 혹은 그 기술이 노후화 되었을 때 유사한 (혹은 차세대의) 데이터나 기술로 쉽게 대체할 수 있는가?

- 개인정보 보안

① 국가통계기관은 자료 획득에 있어 법적 권한을 부여받았는가?
② 데이터를 사용하는데 있어서 법적 제약이나 규제가 존재하는가?
③ 법적 이슈가 발생하였을 때 자료 제공자와 국가통계기관은 협상을 통해 해결할 의사가 있는가?
④ 자료 수집 시 관련된 개인정보법을 준수하였는가?
⑤ 국가통계기관의 보안 정책이 자료의 유용성을 제한하는가?
⑥ 이해관계자(민간 및 공공 영역 등)들이 국가통계기관에서 데이터를 사용하고자 하는 목적에 부정적인 반응을 보일 가능성이 높은가?

- 복잡성

① 구조: 빅데이터를 사용 가능한 구조로 얼마나 쉽게 변형할 수 있는가?
② 형식: 데이터가 표준형식(XLS, XML)을 준수하는가? 데이터에 몇 개의 다른 형태가 포함되는가? 데이터의 변수들을 사용 가능한 형식으로 얼마나 쉽게 변환할 수 있는가?
③ 자료: 데이터 표기에 몇 개의 서로 다른 표준이 사용되었는가? 통일되지 않는 비표준 코드가 사용되었는가? 얼마나 많은 다른 코드가 데이터 표기에

사용되었는가?
④ 계층: 레코드나 변수 간에 계층적 구조가 존재하는가?

• 완전성

① 질적 평가 (예. 입력 단계 시 메타데이터의 완전성 평가, 0: 설명 부재, 1: 설명 불충분, 2: 설명 충분)
② 설명 부재/불충분 시 자료 활용성에 있어 어떤 문제점/결점이 존재하는가?
③ 개체 단위가 명시되었는가?
④ 변수가 명확하게 정의되었는가?
⑤ 메타데이터의 완전성 및 명확성에 대한 질적 평가
⑥ 불분명/모호한 설명의 경우 자료 활용성에 있어 어떤 결과/결점이 존재하는가?

• 유용성

① 데이터 사용 및 분석을 위해 국가통계기관에서 새로운 기술을 습득해야 하는가?
② 데이터의 정제 및 처리 과정에 얼마나 많은 자원이 소요되는가?
③ 데이터가 얼마나 큰가?
④ 자료 전송: 자료 전송 시 어떠한 추가 조치가 필요하며, 이 경우 국가통계기관에서 해당 기준을 충족할 수 있는가?
⑤ IT 요구사항: 자료를 저장하는 데 하드웨어 및 소프트웨어 수준에서의 요구사항이 필요한가? 이 경우 특정 IT 인프라의 구축이 요구되는가?

• 시간적인 요소

① 데이터 수령(receipt of data)과 수집 사이의 소요 시간은?
② 자료가 언제 수집되었는가? 자료의 조사기간은?
③ 주기적으로 자료의 수집 및 제공이 가능한가?
④ 개념이나 방법에서의 차이가 과거 자료의 향후 사용을 저해할 수 있는가?

• 일관성

① 다른 데이터 파일과의 자료 통합 시 사용될 수 있는 연결 변수가 존재하는가?
② 빅데이터 및 다른 자료 출처에서 연결되거나 연결되지 않는 비율은 어느 정도인가?

③ 관심 사항과 연관된 변수들을 어떻게 평가하는가?
④ 사용되는 정의들이 국가통계기관의 기준과 부합하는가?
⑤ 자료의 이상치들이 향후 사용을 저해할 수 있는 중요한 오차들을 나타내는가?

• 타당성

① 제공되는 메타데이터가 사용된 방법의 안정성을 평가하기에 충분한가?
② 자료의 향후 사용을 저해할 수 있는 심각한 결점이 프로세스 내에 존재하는가?

• 정확성

① 포함오차(coverage error)의 수준은 어느 정도인가? 예를 들어, 빅데이터 모집단과 대상 모집단 (target population) 간의 거리는 어느 정도인가? (예. Kolmogorov-Smirnov Index, Index of dissimilarity)
② 파일 내에 중복된 자료가 존재하는가?
③ 자료의 값들이 허용범위 내에 존재하는가?
④ 빅데이터 출처 기준으로 과대/과소하게 기술되거나 아예 누락된 집단은 없는가? (질적 평가 포함)
⑤ 측정 도구의 타당성 및 관측의 정확성에 대한 평가

2) Throughput 단계

Throughput 단계는 데이터의 획득과 최종 분석 혹은 결과치의 산출치 사이의 모든 중간 단계를 가리킨다. Throughput단계 품질 검증 체계는 데이터의 형태와 사용 목적에 따라 다르게 적용되므로 구체적인 검증체계의 정의는 제한적이기 때문에 구체적이며 또한 일반적인 질문 항목을 제안하는 것은 불가능하다. 따라서 본 연구에서는 Throughput단계에서 데이터의 품질 검증에 대한 기본적인 원리들을 설명한다.

• 시스템 독립선(System Independence)

데이터의 변환과 분석 결과는 사용하는 시스템에 의해 좌우되는 것이 아니라 이론적인 원리에 의거해 진행되어야 한다. 따라서 데이터의 분석은 시스템과 독립적인 관계이어야 한다. 예를 들어, 회귀분석에서의 잔차는 회귀분석을

수행하는 분석 시스템에 관계없이 항상 같아야 한다.

- 처리 절차의 안정성

데이터는 현재 상태에서 더 나아가서 처리, 분석, 전송되기도 하고 또는 다른 곳에서 생산된 데이터와 합쳐 질수도 있다. 안정성은 데이터 분석과정에서 이를 바탕으로 처리 절차의 개선을 통한 데이터 및 결과에 대한 품질의 향상을 위해 기본적으로 필요한 요소이다. 또한 이를 위해서는 데이터 처리에 대한 매뉴얼(메타데이터 포함)이 존재해야 하고 이를 통해 업무 담당자가 이를 통해 처리 절차의 안정성 및 이를 통해 생산된 통계의 품질을 관리할 수 있어야 한다.

- 품질 평가 기준 관리

빅데이터 처리과정에서 데이터의 품질을 평가할 수 있는 요소들을 정하고 이 요소들을 평가할 수 있는 체크포인트를 마련하여야 한다. 이를 위해서는 데이터의 품질을 평가하기 위한 처리 과정상의 평가 측정도구와 그 진단 시점이 정의되어야 한다. 이 관리 체계는 생산라인의 관리와 유사하다고 판단할 수 있다. 이를 위해 결정되거나 사전 논의가 이루어져야 하는 항목들은 다음과 같다

① 배치 : 비즈니스 프로세스에서 품질평가가 이루어져야 하는 지점
② 측정 : 품질을 평가하는데 사용할 측정방법
③ 역할 : 품질평가의 책임자
④ 허용 한계 : 사전에 수용 가능한 품질 기준
⑤ 행동 : 품질 평가가 이루어지지 못할 때의 대책
⑥ 평가 : 평가기준에 대한 자체 모니터링

3) Output 단계에서의 질문 항목의 예

이단계의 품질 검증은 Input단계 품질 검증의 구조와 비슷하나 생산된 통계의 품질에 더욱 중점을 두고 이루어진다. Output 품질의 지표들은 이전의 Input과 Throughput에 비하여 포괄적인 경향이 있으며 따라서 빅데이터 Output 품질의 특정한 지표는 모든 경우에 언제나 타당하거나 유용한 것은 아니다. 또한 여기에 제시된 요소들과 지표들은 빅데이터에 중점을 둔 것으로써 일반적으로 생산된 통계의 품질을 평가하기 위해 적용된 품질검증 관련 지표들 또한 빅데이터의

품질검증에 적용되어야 한다. 이 부분은 앞에서 논의되었기 때문에 여기서는 따로 언급하지 않았다. input 단계에서 논의된 항목들과 중복되는 부분이 있으나 output 단계에서 고려할 수 있는 지표 개발을 위한 항목들을 살펴보면 이하와 같다.

- 기관 및 사업체 환경
① 어떤 기관들이 어떠한 합의 하에 자료 및 그 작성에 기여했나?

- 개인정보 보안
① 국가통계기관은 자료 획득 및 공표에 있어 법적 권한을 부여받는가?
② 자료 수집 시 관련된 개인정보 법을 준수하였는가?

- 복잡성
① 데이터 전체에 통일되고 일관성 있는 메타데이터 표준과 분류 방식이 적용되었는가?

- 완전성
① 데이터의 적용 범위는 전체 모집단을 포괄하는가?
② 사용가능한 추정치의 종류와 데이터 사용의 제한사항이나 관련성 문제는 없는가?

- 유용성
① 접근 비용
② 지원 문서의 유무

- 시간적인 요소
① 자료 수령(receipt of data)과 수집 사이의 소요 시간은 얼마나 걸리는가?
② 자료 수집과 자료 조사기간 사이의 소요 시간은 얼마나 걸리는가?
③ 자료의 수집 및 제공이 주기적으로 이루어졌었는가?

- 일관성 및 타당성
① 빅데이터에서 도출한 지표와 모수 간의 상관관계는?
② 도출 시 확실하고 투명한 방법론이 사용되었나?
③ 관심범위 내 변수 상의 변화나 추세의 예측은 가능하며 타당한가?

정확성
① 빅데이터 모집단과 목표 모집단 (target population) 간의 차이는?

② 빅데이터 출처 기준으로 과대/과소하게 기술되거나 아예 누락된 표본 집단이 존재하는가? (질적 평가 포함)

③ 측정 도구의 공간적 분포도 및 관측의 주기성에 대한 평가

제3절 온라인 물가지수 품질 검증

가. 온라인 물가지수

온라인 물가지수란 온라인에서 판매되는 소비자 물가지수의 산출을 위해 고려하는 품목의 가능한 범위 내의 온라인 가격을 조사해 산출되는 지수이다. 최근 온라인 소비 트렌드가 확산되고 있으며, 온라인 시장 규모 역시 확대되는 등 소비자의 구매 행태가 온라인으로 크게 옮겨가고 있다. 따라서 기존 오프라인 중심의 소비자 물가지수를 보완할 수 있는 또 다른 물가지수로서 온라인 물가지수가 고려되고 있다. 가격 변동을 빠르게 반영할 수 있는 일 단위 물가 변동 지표인 온라인 물가지수를 개발함으로써 월 단위 물가변동을 반영하는 소비자 물가지수를 보완할 수 있다. 이를 통해 물가 변동에 대응하는 정책의 시차를 줄일 수 있다. 또한 대표 상품 위주의 기존 소비자 물가지수 외에 소비자의 다양한 소비 패턴에 따른 많은 상품의 가격 변화를 반영할 수 있는 보조적 지표로도 활용이 가능하다.

통계청에서는 이마트, 홈플러스, 11번가 등 소비자들이 많이 이용하는 온라인 사이트 6곳에서 쌀, 밀가루 등 284개 품목에 대한 가격정보를 일일 200만 건 이상 수집하여 지수를 시험적으로 산출하고 있다. 2015년 11월 1일을 지수 기준일로 정해서 품목 내 모든 상품 가격의 산출시점까지의 전일비에 대한 기하평균을 이용하여 품목지수를 산출한 후 소비자 물가지수 품목 가중치를 적용하여 최종 품목군별 지수를 산출한다. 소비자 물가지수는 서비스 152개 품목의 가격을 포함하여 지수를 산출하나 온라인 물가지수는 온라인에서 가격을 수집하므로 서비스 품목에 대한 조사의 제한사항이 있다. 또한 소비자 물가지수는 상품 변동 시 품질조정을 반영하여 지수를 산출하는 반면, 온라인 물가지수는 동일상품의 가격 변화만을 연결하므로 품질 조정을 반영하지 못하는 문제가 발생한다.

즉, 온라인 물가지수와 기존의 소비자 물가 지수 사이에는 상호 보완적인 부분이 존재한다. 기존의 소비자 물가의 산출을 위한 상품이나 품목이 제한적이며 일 별 변화를 반영하지 못하는 반면 온라인 물가 지수는 일 단위 통계작성이 가능하며 또한 고려할 수 있는 상품 혹은 품목의 수를 충분히 고려할 수 있다. 그러나 온라인 물가지수의 경우, 매일 산출되는 수치를 위한 상품 혹은 품목의

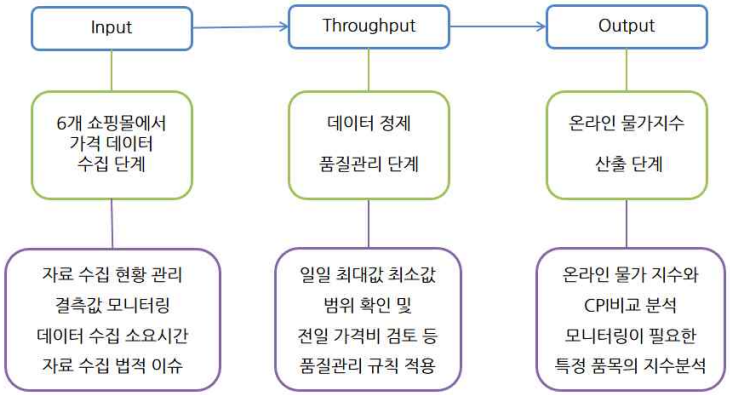
상대 중요성을 산정한 계산이 매우 어렵고 또한 온라인 사에서 발생하는 비정기적인 할인 그리고 상품이나 품목 변화를 적절히 반영하지 못하는 문제점을 안고 있다. 이하의 표는 두 물가지수를 비교한 결과를 제시한 것이다. 본 절에서는 앞에서 논의된 품질검증 체계를 온라인 물가지수 산출을 위한 빅데이터 분석과정에 적용하여 실제 검증 사례를 작성한다.

<표 6> 온라인 물가지수와 소비자 물가지수 비교

		온라인 물가지수	소비자 물가지수
포괄범위	품목	서비스 품목 가격 파악 곤란	상품 및 서비스 ※ 서비스 152개 품목
	상품	품목 내 모든 상품(전수조사)	품목 내 대표 상품(표본조사)
품질조정		미반영	반영
지수산식		상품의 가격비를 기하평균	대상처별 가격 산출평균
		소비자 물가지수 품목 가중치로 산술평균	
작성주기		일 단위	월 단위
지역물가		없음	지역 및 전국단위

나. 품질 검증

온라인상에서 판매되는 상품의 가격 정보를 활용하여 생산된 온라인 물가지수에 앞서 제시한 빅데이터 품질 검증 체계를 적용하면 다음의 [그림 8]과 같이 표현할 수 있다.



[그림 8] 온라인 물가지수의 품질검증 체계

먼저 Input 단계는 데이터 수집 및 수집 이전의 분석 단계로 온라인 물가지수 품질 검증 체계에서는 홈플러스, 이마트 등 6개 쇼핑몰에서 가격 정보를 일일 단위로 수집하는 단계이다. 이 단계에서는 자료 수집의 안정성을 확보하기 위해 일일 단위 수집 사이트별, 품목별 자료 수집 현황을 모니터링 하여 품목 누락을 예방하고 수집 대상 웹사이트 변경(구조나 웹 진열 내용 등)을 조기에 파악하여 품목과 상품 범주에 대한 매핑의 누락을 방지하여야 한다. 또한 데이터 수집에 소요되는 시간을 파악하여 이상 여부를 확인하고 자료를 수집하는데 있어서 법적인 문제가 없는지 확인하여야 한다.

Throughput 단계는 데이터의 획득과 분석 사이의 중간 단계이다. 온라인 물가 지수 품질 검증 체계에서는 데이터 정제 및 품질관리(Quality Control) 단계가 여기에 해당된다. 데이터 품질관리는 총 4단계로 구성이 되어 있는데 1단계는 가격이 품목별 가격이 지정한 최소값과 최대값의 범위를 벗어나는지 검사하는 단계이다. 2단계는 변동비 검사로 가격비(금일가격/전일가격)가 품목별로 최소 변동비(0.4)와 최대 변동비(2.5)의 범위를 벗어나는지 검사한다. 3단계로 수집가격과 평균가격의 차이가 심한지 검사하는 집단 내 평균검사를 실시하고 4단계로 과거 유지된 가격에 비해 급격하게 변화한 경우가 있는지 검사하는 기간 내 평균검사를 실시한다.

마지막으로 Output 단계는 수집된 데이터를 정제, 처리한 후 온라인 물가지수를 산출하는 단계이다. 이 단계에서는 생산된 온라인 물가지수와 소비자 물가지수를 비교 분석하고 가격 폭등 품목 등 모니터링이 필요한 특정 품목에 대한

지수 분석을 실시한다. 그러나 아직까지 온라인 물가지수는 국가승인 통계가 아니므로 생산된 온라인 물가지수의 활용 방안에 대하여 추가적인 연구가 필요하다. 온라인 물가지수의 품질검증을 위해 적용할 수 있는 평가항목과 그에 따른 평가 결과는 아래와 같다.

<표 7> 온라인 물가지수와 소비자 물가지수 비교

품질검증 요소	평가 항목	평가 결과
기관 및 사업체 환경	데이터 및 데이터 제공기관이 품질 기준을 충족하지 못하였을 경우 발생하는 위험은?	품질 기준을 만족하지 못할 경우 소비자가 체험하는 물가와 다른 결과 도출 가능
	데이터 소스가 가용하지 않다면 대체 가능한 빅데이터 소스가 존재하는가?	현재 6개 사이트에서 자료 수집중이며 가용하지 않을시 다른 쇼핑 사이트와 협조
개인정보보호/보안	자료 획득에 있어서 법적 권한을 부여받았는가?	6개 사이트 업체와 협의가 이루어짐
	데이터를 사용하는데 있어서 법적 제약이나 규제가 존재하는가?	없는 것으로 판단됨
복잡성	빅데이터 소스를 사용 가능한 구조로 쉽게 변형할 수 있는가?	자료 수집 시 텍스트 파일로 수집하여 하둡에서 처리 가능한 파일의 형태로 변환함
	계층구조의 유무	품목군-품목-상품 순으로 계층구조 존재
유용성	데이터 사용 및 분석을 위해 새로운 기술을 습득해야 하는가?	대용량 자료 처리를 위한 하둡 기술 습득
	데이터의 저장 및 처리 과정에 얼마나 많은 자원과 시간이 소요되는가?	서버 10대, 90~120분 ※한 달 동안 수집된 데이터 분석 시 일주일 이상 소요
	자료를 저장하고 처리하는데 추가적인 하드웨어 및 소프트웨어 요구사항은?	한 달 또는 일 년 자료 분석 시 오류 발생, 하드웨어 및 소프트웨어 오류 개선 필요
시간관련 요소	데이터가 수집되는데 소요되는 시간은?	90~120분
	주기적으로 자료 수집 및 제공이 가능한가?	일일 단위 자료 수집
연결가능성	다른 데이터 파일과 자료 통합 시 사용될 수 있는 연결 변수가 존재하는가?	소비자 물가 지수 항목과 연결 가능성이 있음
일관성	통계를 생산하는데 사용되는 기준들이 현재 기준과 부합하는가?	소비자 물가지수와 동일한 품목 가중치 사용
타당성	자료의 향후 사용을 저해할 수 있는 결점이 처리과정 내에 존재하는가?	QC과정 재검토 필요

품질검증 요소	평가 항목	평가 결과
정확성	포함오차의 수준은 어느 정도인가? (모집단과 대상 집단 간의 거리)	조사 대상 온라인 매장의 전 항목을 가져오기 때문에 표본오차를 논의하기는 어려움. 또한 온라인 매장의 선택 역시 표본오차를 산출 할 수 있는 방안을 통해 이루어지지 않음
	자료 값들이 허용범위 내에 존재하는가?	QC를 통해 허용범위 외의 자료는 제거함

온라인 물가지수의 품질검증 결과 기관 및 사업체 환경, 개인정보보호/보안, 복잡성, 시간관련 요소 및 일관성 측면에서는 품질이 일정 수준이상으로 판단할 수 있으나 유용성, 연결 가능성, 타당성 측면에서는 추가적인 보완이 필요한 것으로 판단된다. 유용성 측면에서 대용량의 데이터를 처리하기 위한 시간과 자원이 많이 소요되고 특히, 과거의 일부 지수 산출이 잘못되어 수정 보완하기 위해 과거의 데이터까지 포함하여 누적 처리하는데 일주일 이상의 시간이 소요되고 처리과정에서 하드웨어 및 소프트웨어의 오류로 인하여 처리되지 않는 경우도 발생하여 개선이 요구된다.

가격 수집상의 제약 또한 개선되어야 할 부분이다. 쇼핑몰 사이트 담당자가 품목 업로드 시 온라인 물가지수 작성상의 정해진 범주가 아닌 임의의 범주에 상품을 등록(현미 범주에 일반 쌀을 등록하는 등)하거나 인터넷 사이트 URL 주소가 변경되어 자료가 수집되지 않는 등 가격 수집상의 제약이 발생할 때 일일이 사람이 직접 가격 수집을 위해 조작을 하거나 상시 모니터링 해왔다. 이의 개선을 위해서는 앞선 빅데이터 활용의 예에서 살펴보았듯이 다른 범주에 잘못 등록된 상품의 경우 빅데이터를 활용한 Supervised Machine Learning 기술을 통해 잘못 분류된 상품을 충분히 분류할 수 있고 인터넷 가격 수집 로봇을 활용하여 변경된 URL을 실시간 탐지하여 가격 정보를 수집할 수 있을 것이다.

요구되는 또 다른 개선 사항은 데이터 품질관리 단계이다. 앞에서 설명한 것과 같이 온라인 물가지수를 산출하기 위해 6개 대형 온라인 쇼핑몰에서 웹 스크래핑[14] 기법으로 데이터를 수집한 후 데이터의 오류와 노이즈를 제거하기 위한 데이터 품질관리(QC) 단계를 거친다. 그러나 이러한 데이터 품질관리 단계에도 다양한 문제점이 제기될 수 있어 수정 보완이 요구된다. 데이터 품질관리 4단계를 예로 들면 이마트, 홈플러스 등 대형 마트의 경우 단기간 폭탄 세일과 같은 고객 홍보용 여러 이벤트를 실시하는데 이러한 경우 과거에 유지되어온 가격에 비해 급격하게 변화된 가격으로 상품을

14) 자동으로 시스템에 접속해 데이터를 화면에 나타나게 한 후 필요한 자료만을 추출해 가져오는 기술이다. 웹사이트에 있는 정보를 끄집어내 다른 사이트나 데이터베이스에 저장하기 때문에 스크린 스크래핑(Screen Scraping)이라고도 한다.

판매하기 때문에 데이터 품질관리 4단계에서 마트 세일가격이 온라인 물가지수 산출시 제거된다. 그러나 실제 소비자가 구매하는 가격은 세일가격으로 구매하는 것이기 때문에 실제 소비자가 체감하는 물가와 온라인 물가지수는 차이를 보일 수밖에 없으며 전일비 기준 통계의 성질로 인하여 소비자 물가지수에 비하여 상당히 가파른 물가 변동의 결과를 초래할 수 있다. 따라서 이러한 데이터 품질관리 단계에서 데이터의 오류와 노이즈를 제거하기 위해 단계별 명확한 기준과 그 이유를 제시할 수 있도록 재검토하여 다시 설정할 필요가 있다.

5. 소결

ICT 분야의 발전과 더불어 등장한 『빅데이터』를 활용한 민간분야의 분석 및 그 결과의 활용이 광범위하게 이루어지고 있는 시점에 본 연구에서는 빅데이터를 활용한 공공통계 작성의 가능성에 대한 문제를 살펴보았다. 우선은 빅데이터 분석에 대한 관점을 자료 증대의 측면과 빅데이터 분석의 측면으로 구분하고 기존의 통계 작성 방안 및 분석 방안들을 소개하고 그 사례들을 살펴보았다. 공공분야에서 제공되는 통계나 그 분석결과는 민간분야와는 달리 그 신뢰도에 대한 검증이 매우 중요하다. 이는 이를 제공하는 국가 혹은 공공기관의 신뢰도와 직접적으로 연결될 뿐 아니라 정책 결정을 위해 이러한 자료들이 직접적으로 사용되기 때문이다. 따라서 이러한 통계 및 분석 결과의 신뢰도를 고려하여 통계청에서는 전통적인 자료수집 방법을 통해 생산되는 통계의 작성을 위해 승인 기관을 지정하고 또한 매우 까다로운 기준을 정하여 이를 만족하는 통계를 국가 승인 통계로 인정하여 공표를 할 수 있도록 하고 있다.

현재까지는 빅데이터를 활용한 통계 혹은 분석 결과에 대한 신뢰도는 그리 높지 않다. 빅데이터의 분석 사례로 흔히 소개되었던 구글 자료를 활용한 감기 예측에 대한 분석 역시 그 한계와 문제점이 드러났고 또한 여러 매체를 통해 전달되는 소셜 네트워크 분석 결과 등은 아직 초보단계에 머무르고 있다. 아직은 초보단계에 머물러 있는 빅데이터 활용 통계 방법론은 높은 수준의 신뢰도를 요구하는 공공 데이터의 특성과 맞물려서 국가승인통계로 빅데이터를 활용하는 것은 아직 조심스럽다. 하지만 빅데이터 활용의 여러 문제에도 불구하고 서베이를 통한 자료 수집이 어려워지는 현 시점에 빅데이터의 활용에 대한 적극적인 논의가 이루어지는 것은 바람직하다. 빅데이터의 활용 방안을 다양화하고 그 결과의 신뢰도를 높이기 위해서는 본 연구에서 제공하고 있는 품질진단에 대한 기본적인 틀을 바탕으로 각 빅데이터 분석 별 검증 체계를 구체화하고 이를 적용하여 사용 목적에 적합한 빅데이터 활용 통계를 생산하는 과정을 구축해 나가야 할 것이다. 아직은 국가승인통계의 생산을 위해 빅데이터를 활용하는 것은 언급된 문제들로 인하여 가능하지 않으나 추후 구축될 통계 검증 체계를 통한 검증 및 다양한 논의를 통해서 그 가능성을 높여 갈 수 있으리라 기대된다.

참 고 문 헌

2016년도 공공데이터 개방·품질관리 실무교육 자료, 행정자치부/한국정보화진흥원(2016)

공공데이터 관리지침. 안전행정부(2014).

공공데이터의 제공 및 이용 활성화 기본계획(2013). 관계부처 합동

공공정보 품질관리 매뉴얼(2012). 한국정보화진흥원

김기환(2011). 공공부문 빅데이터의 활용성과 위험성. 서울과학기술대학교

빅데이터 활용 국민체감형 통계생산체계 구축방안 수립 최종보고서(2014). 통계청

새로운 미래를 여는 빅데이터 시대(2013). 한국정보화진흥원

송주원·안형진(2009). 무응답 자료 처리 및 분석, 대전 : 통계교육원

알기쉬운 공공부문 빅데이터 분석·활용 가이드 v1.0(2012), 한국정보화진흥원

이영섭·김선웅·안홍엽·임경은·김희경(2009). 통계조사자료와 행정자료 간의 통계적 매칭기법에 관한 연구. 동국대학교

정부 3.0 실현을 위한 빅데이터 활용방안(2013). 한국행정연구원

정영기(2014). 공공개방데이터 품질진단 모델에 관한 연구. 숭실대학교

통계청(2013). 통계행정편람.

Bengio, Y., LeCun, Y. and Hinton, G(2015). Deep Learning. Nature. 521: 436 - 444.

D'Orazio, Di Zio and Scanu(2006). Statistical Matching: Theory and Practice. Wiley.

Daas, P. J. H., M. Roos, M. van de Ven, and J. Neroni(2012). Twitter as a Potential Data Source for Statistics. Discussion paper 201221, The Hague/Heerlen: Statistics Netherlands.

Daas, P. J. H., Puts, M. J.(2014). Social Media Sentiment and Consumer Confidence. Statistical. European Central Bank Statistical Paper Series

Daas, P. J. H., Puts, M. J., Buelens, B., and van den Hurk, P. A. (2015). Big Data as a source for official statistics. Journal of Official Statistics, 31(2), 249-262.

Deng, L.; Yu, D.(2014). Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing. 7 (3-4): 1 - 199.

Fuller, W. A.(2009). Sampling Statistics. Wiley.

Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012). Foundations of Machine Learning, The MIT Press.

Olav Bosch, Dick Windmeijer(2014). On the Use of Internet Robots for Official Statistics. Meeting on the Management of Statistical Information System(MSIS 2014).

Puts, M. J. H., Tennekes, M., Dass, P. J. H., Blois, C.(2016). Using Huge

Amounts of Road Sensor Data for Official Statistics. European Conference on Quality in Official Statistics(2016)
Rubin, D. B.(1987). Multiple Imputation for Nonresponse in Surveys. Wiley.
Sarndal, C.E., Swensson, B. and Wretman, J.(1991). Model Assisted Survey Sampling. Springer.
Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. Neural Networks. 61: 85 - 117.
Trevor Hastie, Robert Tibshirani and Jerome H. Friedman (2001). The Elements of Statistical Learning, Springer.
UNECE(2014). A Suggested Framework for National Statistical Offices for assessing the Quality of Big Data. NTTS 2014 (New Techniques and Technologies for Statistics) Conference.
http://kosis.kr/
http://meta.narastat.kr/
http://www.bicdata.com/bbs/board.php?bo_table=business_archive&wr_id=1796
http://www.dqc.or.kr/main/index.html
https://www.data.go.kr/main.jsp#/L21haW4
https://www.data.go.kr/main.jsp#/L21haW4

Ⅲ. 빅데이터 활용을 위한 제도적 장애요인 검토 및 개선 방안

김주영(명지대학교 법과대학)

Ⅲ. 빅데이터 활용을 위한 제도적 장애요인 검토 및 개선 방안

주지하다시피 우리는 일상생활 속에서 물가상승률을 비롯하여 실업률이나, 경제성장률, TV 프로그램 시청률에 이르는 수많은 통계(statistic)를 접하고 산다. 실로 "오늘날 우리는 온통 통계, 그리고 통계학에 둘러싸여 있다."는 표현이 무색하지 않을 정도이다.[15]

이러한 통계들을 학문적으로 다루는 통계학(statistics)은 일반적으로 판단, 선택, 결정 그리고 예측에 도움이 되도록 자료를 의미 있는 정보로 처리하는 방법과 그 결과를 해석하는 방법을 연구하는 학문[16]으로, 우리 주변의 사회현상이나 자연현상에 관해 수집한 자료를 이해하기 쉬운 형태로 정리·요약할 뿐 아니라, 이를 분석·해석하여 현상을 기술·설명·예측하는 데 활용된다. 이러한 통계학은 정부나 기업과 같은 조직뿐만 아니라 개인에 이르기까지 직관이나 주관적 판단에서 올 수 있는 시행착오나 불확실성을 최소화하면서 객관적이고 합리적인 의사결정을 하는 데 중요한 정보를 제공해 왔다.[17] 이러한 맥락에서 통계학은 체계적인 자료의 수집과 분석을 바탕으로 **우리가 바람직한 의사결정을 수행할 수 있도록 도와주는 학문**[18]으로 기능한다. 학문적 관점에서 볼 때 통계학은 이미 공학, 생물학, 의학, 물리학과 같은 자연과학뿐만 아니라 경제학, 경영학, 사회학, 심리학, 행정학 등의 사회과학 분야에서도 과학적인 연구나 조사를 위해 반드시 필요한 도구가 되었다고 할 수 있다.[19] 나아가 오늘날의 통계는 일상적인 차원에서도 효과적인 소통수단으로 각광받고 있기도 하다. 특히 다원주의 사회로 넘어오면서 동일한 현상에 대한 다양한 해석이 일반화됨에 따라, 서로 다른 의견을 조율할 수 있는 효과적인 수단으로 통계가 적극 활용되고 있는데, 이는 실제 현실의 많은 논쟁들은 사실관계 확인이 제대로 되지 않아 발생하는 경우가 많기 때문에 이러한 사실관계의 확인을 위해 각종 사회통계(social statistics)가 널리 사용되고 있는 실정이다.[20]

한편 현대의 가장 주요한 변화로 손꼽히는 국가·사회 전반에 걸친 정보화는 통계와 관련한 새로운 국면을 열어가고 있다. 발전된 정보기술은 정보시스템을

15) Ian Hacking, *The Taming of Chance*, 정혜경 역, 『우연을 길들이다: 통계는 어떻게 우연을 과학으로 만들었는가?』(서울: 바다출판사, 2012), 7면의 옮긴이의 말.
16) 한성안, 『(인문학으로 풀어보는) 통계학』(서울: 청람, 2013), 6면. 한성안은 통계학이 **"어떤 학문보다도 실용적"**이라고까지 언급하고 있다. 같은 책, 5면.
17) 황인창·이대용·이청호, 『(알기쉬운) 통계학(제3판)』(서울: 비엔엠북스, 2015), 3면.
18) 고길곤, 『통계학의 이해와 활용』(고양: 문우사, 2014), 5면.
19) 황인창·이대용·이청호, 앞의 책, 7면.
20) 고길곤, 앞의 책, 4면. 한편 이러한 통계의 힘으로 인해 의사소통 과정에서 (통계에 관한) 조작이 이루어지기도 한다. 같은 책, 5면.

통한 통계의 자동작성을 가능하게 하고 있으며,[21] 특히 빅데이터 기술을 필두로 새롭게 부각되는 정보처리기술들은 기존 통계영역을 넘어선 새로운 통계의 작성을 가능하게 할 것으로 기대되지만, 반면 기존 통계 및 통계자료에 대한 새로운 접근을 통해 과거에는 문제되지 않았던 새로운 문제들, 특히 개인정보보호의 영역과 밀접한 관련을 맺는 문제점들을 초래할 우려도 없지 않다.

그럼에도 불구하고 이러한 통계에 대한 법적인 의미의 부여와 법제도적 측면에서의 발전적 방향의 모색은 부족한 편이라 할 수 있다.[22] 이에 이하에서는 향후의 본격적인 논의의 기반을 마련한다는 측면에서, 통계에 대한 개괄적인 검토를 바탕으로 통계의 생산·보급을 위한 체제(통계제도)를 규율하고 있는 현행 통계법제의 적절성을 특히 헌법적 측면에서의 정당화 가능성을 중심으로 검토하는 가운데 아울러 특히 새로운 빅데이터 환경 하에서 통계생산 체제와 현행 개인정보 보호법제와의 조화가능성을 검토해 보기로 한다.

1. 통계 일반론

통계에 대한 관심은 서양의 경우 이미 15세기로 거슬러 올라가 피렌체와 베네치아 같은 도시국가들은 상업적 문화의 맥락 안에서 수치로 표시된 정보가 쓸모 있다는 점을 일찍부터 깨달았던 것으로 알려져 있으며, 17세기에 들어서면서 영국과 프랑스 같은 국가도 각각 '통치산술(political arithmetic)'과 '정치적 계산(calculs politiques)'이라는 명칭으로 수량자료에 관심을 가지기 시작하였다고 한다.[23] 한편, 혹자는 인류 최초의 통계로 BC 1440년경 모세에 의해 기록된 것으로 알려진 구약성서 중 민수기(Numbers)에 수록된 인구 조사 내용을 꼽기도 한다.[24] 한편 통계학의 경우는 "(다들 잘 알다시피) 통계학은 근대에 생긴 근대적인 학문"이란 평가가 있듯이, 서구를 기준으로 할 때 19세기 이후부터 시작된 것으로 알려져 있다.[25]

21) 일례(一例)로 '행정구역(시군구)별 총인구, 남자, 여자 인구수'는 행정자치부 자치제도정책관 주민과에서 「주민등록법」에 의한 주민등록인구 및 세대현황에 대하여 전국단위의 기관별(시도, 시군구, 읍면동), 연령별 현황통계를 행정자치부의 **주민등록정보시스템으로 자동 집계하여 작성**한다. 통계청의 '행정구역(시군구)별 총인구, 남자, 여자 인구수'에 관한 보다 자세한 내용은 "국가통계포털"의 '주제별통계' 중 '인구·가구' 항목 '인구총조사' http://kosis.kr/statisticsList/statisticsList_01List.jsp?vwcd=MT_ZTITLE&parentId=A#SubCont (2016.9.18.최종확인)를 참조.

22) 최승필, "통계(通計)의 공법적 의미와 과제", 『공법학연구』 제8권 제2호(2007), 396면.

23) 한성안, 앞의 책, 11~12면의 내용을 일부 수정.

24) 김달호 외 7인 공저, 『통계로 세상보기』 (파주: 자유아카데미, 2012), 5면.

25) Stephen M. Stigler, *History of statistics*, 조재근 역, 『통계학의 역사』 (파주: 한길사, 2005), 16~17면.

제1절 통계의 의의

보통 '통계'라 함은 '특정의 대상에 대하여 객관적으로 표현한 수량적 지표'[26]로 행위기준의 수립 및 평가 등에 활용할 목적으로 작성된다고 한다. 그렇지만, 실제 언어용례에 있어서 통계라는 용어는 보다 다양하게 사용된다. 우선 통계(通計)의 사전적 의미를 살펴보면,

① 한데 몰아서 어림잡아 계산함.
② 어떤 현상을 종합적으로 한눈에 알아보기 쉽게 일정한 체계에 따라 숫자로 나타냄. 또는 그런 짓.
③ <수학> 집단적 현상이나 수집된 자료의 내용에 관한 수량적인 기술. 대상이 되는 집단을 일정한 시점에서 파악하는 것을 정태 통계, 일정한 기간에서 파악하는 것을 동태 통계라 하며, 사회나 자연 현상을 정리·분석하는 수단으로 쓰기도 한다.

정도로 정리된다.[27] 이를 통해서 '통계'라는 단어가 실생활에 활용됨에 있어 수학적 혹은 통계학인 맥락에서의 의미 이외에도 보다 일반적인 의미를 가질 수도 있다는 점을 일단 유념할 필요가 있겠다.

한편 통계학에 있어서 통계의 의미는 첫째, 가장 많이 쓰이는 것으로 '**통계자료, 통계수치**'를 가리키고, 둘째, 자료를 수집하고 분석하는 방법론을 다루는 과학으로서의 '**통계학**'이라는 학문체계를 가리키며, 셋째, 전문적인 용어로 쓰이는 것으로 표본으로부터 산출된 값을 뜻하는 '**통계량(statistic의 복수형)**'을 의미하는데, 흔히 '통계'라고 하면 첫째의 의미로 사용되고 이때의 통계자료란 '집단에 관한 수량적이고 객관적인 정보'를 의미한다.[28] 여기에서의 '집단'은 '실제와 관련된 집합'을 말하며, '집합'은 '다른 것과 구별될 수 있는 것들의 모임'을 의미하므로, 통계수치가 의미를 가지려면 그 통계의 대상범위인 집단이 명확하게 정의되어야 한다. 궁극적으로 통계는 **전체 집단 또는 부분 집단에 관한 사실을 객관적으로 나타내는 것**이지, **집단을 구성하는 개체를 특정하는 특정개체의 개별적인 정보를 나타내는 것이 아니다**(강조는 필자).[29] 아울러 통계가 다루는 집단은 사회현상에 국한되지 않으며 심리 및 자연현상을 포함하여 '집단'으로 구획가능한 모든 대상이 다루어질 수 있다.[30]

26) 최승필, 앞의 글, 395면.

27) "통계", 국립국어원 편, 『표준국어대사전』(http://stdweb2.korean.go.kr/main.jsp).

28) 장치성 외 4인공저, 『국가통계이해』 (대전: 통계교육원, 2015), 9면.

29) 장치성 외 4인공저, 같은 곳. 이러한 통계의 본질적인 특성은 향후 '특정한 개인에 관한 식별가능성'을 중심으로 이루어지는 개인정보 관련 논의에 있어서 특히 중요한 위치를 차지하게 된다 할 수 있을 것이다.

30) 예를 들면 「기상법」(법률 제13162호) 제2조 제8호 라목이 규정하고 있는 "기상현상 및 기후에 관

이러한 통계의 의미를 실정법적인 관점에서 살펴보자면, 「통계법」(법률 제13818호)은 통계를 "**통계작성기관**이 **정부정책의 수립·평가** 또는 **경제·사회현상의 연구·분석** 등에 활용할 목적으로 산업·물가·인구·주택·문화·환경 등 **특정의 집단이나 대상 등**에 관하여 직접 또는 다른 기관이나 법인 또는 단체 등(이하 "기관 등"이라 한다)에 위임·위탁하여 작성하는 **수량적 정보**"로 규정하면서, 이때 "통계작성기관이 내부적으로 사용할 목적으로 작성하는 수량적 정보 등 대통령령으로 정하는 수량적 정보"는 통계에서 제외한다고 한다(동법 제3조 제1호).[31]

아울러 통계작성기관이 통계의 작성을 위하여 수집·취득 또는 사용한 자료(데이터베이스 등 전산자료를 포함한다)는 '통계자료'(법 제3조 제4호), 공공기관이 직무상 작성·취득하여 관리하고 있는 문서·대장 및 도면과 데이터베이스 등 전산자료 가운데 통계자료를 제외한 것은 특별히 '행정자료'(법 제3조 제7호)로 규정하여 '통계'와 구분하고 있다. 이와 관련하여 통계청은 통계청이 제공하는 공공용(公共用)데이터를 자료의 수준에 따라 다음과 같이 분류하고 있기도 하다.[32]

한 통계 정보"도 얼마든지 가능하다.

31) 이처럼 「통계법」의 적용대상에서 제외되는 수량적 정보는 「통계법시행령」 제2조가 정하는 바, 동조에 따르면 그러한 정보는 ① 통계작성기관이 대외적인 공개를 목적으로 하지 아니하고 업무 추진을 위하여 내부적으로 사용할 목적으로 작성하는 수량적 정보 ② 통계작성기관이 통계를 원활하게 작성하기 위한 사전 준비 또는 사후 확인과정에서 통계작성대상이나 절차 또는 방법 등의 적합성 및 타당성, 오차의 발생여부 등을 확인·점검하기 위하여 시험적으로 작성하는 수량적 정보 ③ 통계작성기관에 소속된 직원이 개인적인 학술연구의 목적으로 연구 논문이나 보고서 등에 수록하기 위하여 작성하는 수량적 정보 ④ 통계작성기관이 일상적인 업무수행의 추진 및 관리·감독을 위하여 하부조직, 소속기관, 산하기관 또는 관계기관으로부터 보고받거나 제출받은 현황, 실적 등의 자료를 집계하여 작성하는 수량적 정보 ⑤ 통계작성기관이 그 소속 직원이나 회원, 이해관계자, 서비스 이용자 등을 대상으로 업무 추진성과나 계획에 관한 만족도 등 주관적인 인식이나 의식 또는 의견을 조사하여 작성하는 수량적 정보 ⑥ 그 밖에 통계작성기관이 정부정책의 수립·평가나 경제·사회현상의 연구·분석 등 **사회공공의 이익을 목적으로 작성한다고 보기 어려운 수량적 정보**를 말한다.

32) 국가통계마이크로데이터서비스(http://mdis.kostat.go.kr)의 용어설명 참고.

<표 8> 통계청의 데이터 분류 1. 자료 수준별

구분	내용	비고
원자료 (Raw Data)	• 통계조사 자료에서 최초 입력한 전산화일 자료 • 입력오류, 조사오류 등이 걸러지기 이전 단계의 자료	*메타 데이터(meta data): 통계자료 이용자의 이해를 돕기 위한 통계에 대한 설명자료
마이크로 데이터 (micro data) (별칭: "통계원시자료")	• '원자료'에서 입력오류 등을 제거하여 공표통계표 작성 등 데이터 가공의 기초자료로 사용되는 자료 • 공표통계표에서 얻을 수 없는 심층적인 분석을 원하는 다양한 계층의 이용자에게 제공	
- 공공이용 마이크로데이터 (public use microdata)	• 일반 이용자에게 제공하기 위해 응답자가 식별, 노출되지 않도록 자료 처리를 한 자료	
- 승인된 마이크로데이터 (licensed microdata)	• 자료관리기관의 승인을 얻어야 이용 가능한 자료 • 집계의 정도에 따라 세분된 자료에서 통합된 자료까지 다양하게 제공	
매크로 데이터 (macro data)	• 마이크로 데이터를 임의의 기준에 따라 집계한 자료 • 집계의 정도에 따라 세분된 자료에서 통합된 자료까지 다양하게 제공	

이러한 통계의 특성으로는 **익명성**, **비교성·객관성**, **정확성** 등이 꼽힌다.[33] 우선 익명성은 통계가 집단에 관한 어떠한 정보를 전달하는 숫자로 구성되므로 집단을 구성하는 개체를 특정할 수 있는 고유한 정보가 제거되어 개체를 식별하는 정보가 포함되어 있지 않다는 것인데, 이러한 익명성이 실제의 통계조사나 자료수집 과정에서 개체의 고유명칭 등에 대한 조사의 필요성을 배제하는 것은 아님에 주의해야 한다. 비교성·객관성은 통계가 명확히 규정된 집단으로부터 얻어져야하기 때문에, 집단이나 표지(標識)에 대한 규정 또는 제한이 객관적이어서 모든 관계자들에게 동일하게 이해될 수 있도록 설정되어야 하며, 나아가 이것에 의해 통계의 분석·이용이나 상호비교가 가능해야 한다는 것이다. 마지막으로 정확성은 통계가 가져야 할 핵심적인 속성으로서 파악하고자 하는 통계집단의 진실한 값에 최대한 접근해야 한다는 것이다.

법적인 측면에서 파악해 보자면, 이러한 자료로서의 통계, 특히 국가통계 가운데 일정 부류는 헌법상의 '국가표준'의 하나로 인정될 수 있을 것이다. 현행 「국가표준기본법」(법률 제13747호)에 의하면 국가표준은 "국가사회의 모든 분야에서 정확성, 합리성 및 국제성을 높이기 위하여 국가적으로 공인된 과학적·기술적 공공기준으로서 측정표준·참조표준·성문표준 등 이 법에서 규정하는 모든 표준"을 의미하는데(법 제3조 제1호), 이 가운데 특히 "참조표준"이 "측정데이터 및 정보의 정확도와 신뢰도를 과학적으로 분석·평가하여 공인된 것으로서 국가사회의 모든 분야에서 널리 지속적으로 사용되거나 반복 사용할

33) 통계개발원, 『국가통계제도의 발전(국제공동연구보고서)』 (한국: 통계개발원, 2008), 9~10면.

수 있도록 마련된 물리화학적 상수, 물성값, **과학기술적 통계** 등"을 말한다(법 제3조 제6호)고 명시적으로 규정하고 있다. 아울러 현행「통계법」제22조는 통계청장에게 통계작성기관이 동일한 기준에 따라 통계를 작성할 수 있도록 국제표준분류를 기준으로 산업, 직업, 질병·사인(死因) 등에 관한 표준분류를 작성·고시하도록 하여(동조 제1항), 통계청이 통계작성에 관한 범위 내에서는 국가적인 표준의 확립에 관여할 수 있도록 하는 규정을 두고 있음은 어느 정도까지는 이러한 맥락에서 이해할 수 있는 부분이기도 하다.[34] 그렇지만 모든 국가통계가 국가표준으로 인정되는 것은 아닐 것이기에, 국가표준으로까지 인정되지 못하는 대부분의 국가통계는 그 자체로 '공공정보'[35] 내지 '행정정보'[36]에 속하는 것으로 볼 수 있을 것이다.[37]

제2절 통계의 종류

앞에서 이미 일부 살펴본 바 있지만, 일반적으로 논의되는 통계의 분류는 다음과 같다.[38]

34) 현재의「국가표준기본법」은 기본적으로 산업통상자원부가 주관하는 법률이다. 그렇지만 전자정부의 본격적인 실현 특히 빅데이터 환경 하에서 향후 행정자료의 원활한 공동이용을 위해서는, 특정 행정기관의 관할을 넘어서는 행정기관 전체의 차원에서 행정자료의 표준화에 대한 보다 실질적인 조항을 마련하는 것이 바람직할 것으로 생각된다. 이는 국가표준으로서의 통계의 가치를 전제로 할 때 특히 중요하다 할 수 있을 것인데 기본적으로 통계생산을 위해서는 행정자료의 서식·용어·분류·항목 등이 표준화되어 있어야 통계생산을 위한 행정자료 연계가 수월할 것이기 때문이다.

35) '공공정보'라는 용어는 우리 법제의 일부에서 사용하고 있는 표현이다. 예를 들어,「콘텐츠산업 진흥법」(법률 제13821호) 제11조. ① 국가, 지방자치단체, 그 밖에 대통령령으로 정하는 공공기관의 장(이하 "공공기관의 장"이라 한다)은 그 **공공기관이 보유·관리하는 정보 중 「공공기관의 정보공개에 관한 법률」 제9조에 따른 비공개대상정보를 제외한 정보(이하 "공공정보"라 한다)**를 공개하는 경우에는 콘텐츠사업자로 하여금 해당 정보를 콘텐츠제작 등에 이용하도록 할 수 있다.「이러닝(전자학습)산업 발전 및 이러닝 활용 촉진에 관한 법률」(법률 제14111호) 제22조 역시 유사한 내용을 규정하고 있다. 한편「공공기관의 정보공개에 관한 법률(법률 제14185호)」제2조 제1호는 "정보"란 공공기관이 직무상 작성 또는 취득하여 관리하고 있는 문서(전자문서를 포함한다. 이하 같다)·도면·사진·필름·테이프·슬라이드 및 그 밖에 이에 준하는 매체 등에 기록된 사항을 말한다고 규정하고 있다.

36)「전자정부법」(법률 제13459호) 제2조 제6호 "행정정보"란 행정기관등이 직무상 작성하거나 취득하여 관리하고 있는 자료로서 전자적 방식으로 처리되어 부호, 문자, 음성, 음향, 영상 등으로 표현된 것을 말한다.

37) 행정법학적 측면에서 굳이 분류해 보자면 이러한 통계자료는 행정법상 법률요건 가운데 행정법상의 용태(容態)의 '대상(객체)'으로서 파악될 될 가능성이 높지만 이러한 논의의 실익은 거의 없을 것으로 본다. 주지하다시피 행정법상의 용태(정신적 사실)란 사람의 정신작용을 요소로 하는 행정법상의 법률사실을 말한다. 관련 논의는 김남진·김연태,『행정법I(제18판)』(파주: 법문사, 2014), 128~129면 등을 참고.

38) 이재형,『국가통계시스템 발전방안』, 한국개발연구원 연구보고서(2004), 15면.

<표 9> 통계청의 데이터 분류 2. 법률·생산방법 및 주체별 분류

법률적 분류	승인통계	지정통계	통계작성기관이 작성하는 통계 가운데 통계청장이 지정하여 고시하는 통계
		일반통계	통계작성기관이 작성하는 통계 가운데 지정통계 이외의 통계
	기타통계		통계작성기관이 아닌 곳에서 작성하는 통계로서 통계법상의 통계에 포함되지 않음
조사방법에 의한 분류	조사통계		조사대상에 대한 실지조사를 통하여 얻어진 통계: 전수조사와 표본조사로 구분
	보고통계 (업무통계)		국가기관의 행정업무에 수반되어 수집된 자료로부터 작성된 통계
작성방법에 의한 분류	1차 통계		통계조사를 실시하여 그 결과에서 직접 얻어진 통계
	2차 통계 (가공통계)		1차 통계를 통하여 얻어진 자료 등을 이용하여 일정한 연산을 가하여 얻어진 통계
생산주체에 의한 분류	국가통계		정부기관이 생산하는 통계
	민간통계		민간기관이 조사·생산하는 통계

이하에서는 국가통계와 민간통계를 중심으로 살펴보기로 한다.

가. 국가통계

보통 국가통계(national statistics)란 보통명사로서의 통계가 아니라 국가에 의해서 인정되는 공식통계(official statistics)를 의미한다.[39] 이러한 국가통계는 "국가의 제도적 틀을 설정·유지하고 국민적 합의를 도출하는 하나의 지표"[40]로 평가되며, 구체적으로는 행정행위에 있어서 판단의 기초가 되며, 아울러 재량판단이 적정하였는가가 문제가 될 경우 행정소송의 과정에서 재판상 판단기준으로 이용된다.[41] 이처럼 국가의 행정의 대상이 되는 상황 및 객체에 대한 행위기준과 향후 정책방향을 설정하기 위한 광범위한 정보의 수집 작용이 바로 **통계조사 및 작성**이라 할 수 있다. 즉, 국가통계는 통계법의 대상이 되는 통계로서 사회·경제적 변화를 진단하고 과학적인 정책을 수립하기 위한 필수적인 **공공재**(public goods)이다.

이러한 통계가 갖는 국가적 중요성 및 통계생산의 속성에 비추어 볼 때 정부의 역할은 매우 중요하다 할 수 있다. 무엇보다도 통계는 이미 언급한 바와 같이 공공재로서, 외부효과(externality)가 존재하기 때문에, 민간의 이윤동기에 입각한 시장기능이 적절하게 작용하기 어려워 소위 '시장실패(market failure)'가

39) 특히 공법상 대상이 되는 통계는 민간에서 사적 목적으로 작성하는 민간통계를 제외한 공공통계를 의미한다고 한다. 최승필, 앞의 글, 395면.

40) 최승필, 위의 글, 395면.

41) 최승필, 위의 글, 396면.

발생할 우려가 큰 영역이라 할 수 있으며, 통계의 생산과정에서 개인이나 기업을 상대로 조사를 진행함에 있어서 경우에 따라 조사대상자에게 어느 정도의 강제가 필요할 수도 있는 영역이다 보니 공권력을 가진 국가가 가장 효과적인 통계작성주체로 기능할 수 있게 된다는 것이다.[42]

이미 살펴본 바 있는 「통계법」상의 '통계'는 기본적으로 이러한 국가통계를 의미한다고 볼 수 있다. 다만 「통계법」상 통계 작성주체라 할 수 있는 '통계작성기관'은 '중앙행정기관·지방자치단체 및 「통계법」 제15조에 따라 지정을 받은 통계작성지정기관'을 말하는데(동법 제3조 제3호), 통계청장이 통계의 작성·보급 및 이용을 촉진하기 위하여 **정부정책의 수립·평가** 또는 **경제·사회현상의 연구·분석** 등에 이용되는 수량적 정보를 작성하고 있거나 작성하고자 하는 기관 등의 **신청이 있는 경우**[43] 지정 가능한(동법 제15조 제1항) 통계작성지정기관이 모두 국가기관인 것은 아닐 수 있기 때문에 다소간의 주의를 요한다.

한편 「통계법」은 '지정통계'와 '승인통계'의 구분도 하고 있다. 지정통계는 '「통계법」 제17조에 따라 통계청장이 지정·고시하는 통계'(동법 제3조 제2호)를 말하는데, 「통계법」 제17조는 통계청장이 통계작성기관의 장의 신청에 따라 정부의 각종 정책의 수립·평가 또는 다른 통계의 작성 등에 널리 활용되는 통계로서 법이 규정하고 있는 일정 범주에 해당하는 통계[44]를 지정통계로 지정하게끔 규정하고 있다. 이러한 **지정통계**의 경우 **작성기관은 피응답자에게 작성의무를 부과할 수 있으며, 위반시 과태료 등을 부과할 수도 있다**(동법 제12조, 제25조).

승인통계는 통계작성기관이 작성하고자 하는 통계로서 그 명칭, 종류, 목적, 조사대상, 조사방법, 통계표 서식, 조사사항의 성별구분 등 대통령령으로 정하는 사항에 관하여 미리 통계청장의 승인을 받은 것을 의미하는데, 이러한 승인은 새로운 통계를 작성할 때는 물론, 이미 승인을 받은 사항을 변경하거나 승인을 받은 통계의 작성을 중지할 때에도 필요하다(동법 제18조). 보통 승인통계는 지정통계를 포함하는 개념으로 이해되며, 승인통계 가운데 지정통계를 제외한 통계를 '일반통계'로 지칭하기도 함은 본절의 서두의 표에서 언급한 바 있다.

42) 이재형, 앞의 보고서, 23~24면. 이재형은 그 밖에도 통계의 품질 유지를 위한 절차 준수의 측면 및 통계작성상 중립성의 관점에서도 국가의 통계작성주체로서의 중요성을 강조하고 있다.

43) 단 통계청장은 정부정책의 수립·평가 또는 경제·사회현상의 연구·분석 등에 이용되는 수량적 정보를 작성하고 있는 기관 등이 제1항에 따른 지정신청을 하지 아니하는 경우에는 상당한 기간을 정하여 지정신청을 하도록 **권고**할 수 있다(「통계법」 제15조 제2항).

44) 그 범주에는 ① 전국을 대상으로 작성하는 통계 ② 지역발전을 위한 정책수립 및 평가의 기초자료가 되는 통계 ③ 다른 통계의 모집단자료로 활용 가능한 통계 ④ 국제연합 등 국제기구에서 권고하는 통일된 기준 및 작성방법에 따라 작성하는 통계 ⑤ 그 밖에 지정통계로 지정할 필요가 있다고 통계청장이 인정하는 통계가 규정되어 있다(「통계법」 제17조 제1항).

나. 민간통계

본장의 서두에서 간략히 언급한 바 있듯이 통계학은 우리 주변의 사회현상이나 자연현상에 관해 수집한 자료를 이해하기 쉬운 형태로 정리·요약할 뿐 아니라, 이를 분석·해석하여 현상을 기술·설명·예측하는 데 활용되는 바, 의사결정과정에 있어서 직관이나 주관적 판단에서 올 수 있는 시행착오나 불확실성을 최소화하면서 객관적이고 합리적인 의사결정을 하는 데 중요한 정보를 제공해준다.[45] 이에 국가영역에서뿐만 아니라 민간에서도 연구 혹은 기업의 정책수립을 위시한 다양한 목적하에서 다수의 통계가 작성되고 있다. 이미 살펴본 바와 같이 이러한 통계를 법률적인 측면에서는 '기타통계'로 구분할 수 있을 것이다.

다만 민간에서 작성되는 통계는 철저하게 피응답자의 **자발적인 협조를 전제로** 진행되게 된다. 이에 따라서 통계청에서 작성되는 통계는 인구 및 경제·사회 다양한 영역을 포괄하며, 표본의 대표성 및 신뢰성을 확보하고 있다는 장점을 가지는 반면 하나의 자료 안에 수록하고 있는 정보는 상대적으로 제한적인데 대하여, 민간의 외부 연구기관에서 수행되는 패널조사의 경우는 표본 수는 매우 적은 반면 수록정보는 매우 방대하다는 특성을 갖는다.[46]

제3절 현행법상의 통계제도

「통계법」은 그 목적을 규정하면서 통계의 신뢰성과 함께 '통계제도' 운용의 효율성을 확보함을 제시하고 있다. 일반적으로 한 나라가 국가통계를 생산·보급하는 체제를 '통계제도(statistical system)'라고 하는데, 국가통계의 작성기능을 국가기관 간에 어떻게 배분하며, 각각의 통계를 어떠한 절차나 체계를 통해 작성하며, 국가통계기능에 필요한 인적·물적 자원을 어떻게 관리하며, 어떠한 경로를 통해 통계를 보급하는가 하는 문제가 통계제도의 영역 내에서 논의된다고 한다.[47] 다만 '통계제도'가 국가통계를 중심으로 논의되는 것은 분명한 사실이지만, 이미 살펴본 바와 같이 민간통계의 존재를 고려한다면, 통계제도를 국가통계에 국한시킬 이유는 없다고 본다. 즉 통계제도는 **통계를 생산·보급하는 체제**의 의미로 이해하는 것이 바람직하며 그래야만 민간영역에서의 통계작성과 관련한 다양한 이슈들도 함께 고려하는 것이 가능해 질 것이다.[48] 다만,

45) 황인창·이대용·이청호, 앞의 책, 3면.

46) 심규호·박시내, "통계이용 활성화를 위한 2차 자료 생산 활용 방안 연구", 『(통계개발원) 2010년 하반기 연구보고서 제II권』 (한국: 통계개발원, 2010), 236~237면.

47) 통계개발원, 앞의 보고서, 1면.

48) 아울러 국가통계와 민간통계와 관련하여 굳이 통계제도의 구분이 필요할 경우에는 국가통계제도, 민간통계제도의 용어를 사용할 수 있을 것이다.

이후의 논의는 기존의 논의들을 최대한 고려하는 차원에서 국가통계를 중심으로 살펴보면서, 필요한 범위내에서 민간통계에 관한 논의를 부가하도록 하겠다.

제4절 통계제도의 법적 근거

입헌주의(立憲主義) 및 법치주의(法治主義)의 원칙상 국가작용은 헌법 및 법률에 근거하여 운용되어야 함은 이론(異論)의 여지가 없다. 그렇지만 통계제도의 헌법적 근거에 관한 논의는 지금까지 그다지 심도 있게 진행되지 못한 상태라 할 수 있다. 다만 개별적인 영역에서 통계 작성을 규정하는 법률들에 대해서는 헌법상 각각의 규정, 예컨대 교육관련 통계는 제31조, 노동통계는 제32조, 사회복지관련 통계는 제34조, 보건관련 통계는 제36조, 경제의 경우는 제9장 특히 제119조 2항에서 그 근거를 찾을 수 있다는 견해가 제시된 바 있다.[49] 이러한 점을 염두에 두는 가운데, 이하에서는 먼저 통계에 대한 헌법적 규율양상을 검토한 후 이어 법률적 규율양상을 검토해 보기로 한다.

가. 헌법적 규율: 통계제도의 헌법적 정당화 가능성

우리나라의 공식적인 국가통계제도는 1948년 정부수립 당시 「정부조직법」 제32조 공보처 소관 사무에 통계업무를 명시하고, 동년 11월 4일 대통령령 제15호로 공보처 직제가 공포됨으로써 설치된, 통계관실과 서무, 기획, 국세조사 및 인구조사의 1실 4개과로 구성된 공보처 통계국이 중앙통계기관으로서 역할을 담당한 것으로부터 시작되어, 동년 12월 13일 대통령령 제39호로 공포된 「제1회 총인구조사 시행령」이 최초의 통계시책으로 자리매김한 이래, 여러 차례의 편제상의 변동을 거쳐 1990년 통계청으로 승격 개편된 이후 오늘에 이르고 있다.[50] 그렇지만 현행 헌법은 물론 역대 헌법을 모두 살펴보더라도 '통계'가 명시적으로 헌법에 규정된 적은 없었다.

한편 외국헌법 가운데에는 통계에 관한 명시적인 규율을 하고 있는 헌법이 없지 않은 바, 대표적인 예로는 통계에 관한 독자적인 조항을 두고 있는 스위스연방 헌법과[51] 2006년 개정을 통해 '국가통계지리정보시스템'의 도입을

49) 최승필, 앞의 글, 396면.

50) 우리나라 정부수립 이후 통계제도의 역사에 관해서는 특히, 통계청, 『한국통계발전사: 위대한 숫자의 역사 - 시대사』(대전: 통계청, 2015), 163면 이하를 참조.

51) 스위스연방 헌법(1999년 4월 18일) 제65조(통계) ① 연방은 스위스의 인구, 경제, 사회, 교육, 연구, 국토 및 환경에 관한 현황과 추이에 관한 통계자료를 수집한다.
② 연방은 자료수집 비용을 줄이기 위하여 공식 등록의 조화 및 관리에 관한 법률을 제정할 수 있다.

헌법상 명시한 멕시코 헌법(동헌법 제26조)[52]을 들 수 있다. 그 밖에도 통계에 관한 내용을 (연방)국가의 권한(스페인 헌법 제149조 제1항 31.ª, 오스트리아연방헌법 제10조 제1항 제13호, 이라크 헌법 제110조 제9항 등) 혹은 입법권의 권한(러시아헌법 제71조 p항, 독일연방헌법 제73조 제1항 제11호, 이탈리아공화국 헌법 제117조 r호, 캐나다 헌법 제91조 제6호, 호주연방헌법 제51조 제11호 등) 가운데 하나로 적시하고 있는 국가들이 존재한다. 이미 살펴본 바와 같이 익명성, 비교성·객관성, 정확성을 요구받는 통계의 국가·사회적 중요성과 함께, 이처럼 통계에 관한 명시적인 규정을 두고 있는 일부 국가들의 입법례를 감안하면 **우리 헌법에서도 통계에 관한 조항을 마련하는 방안도 장기적으로 검토해 볼 필요가 있을 것**이다.

비록 우리 헌법이 통계에 관한 명시적인 규정을 두고 있지는 않지만, 그렇다고 해서 현행 헌법에서 통계제도의 헌법적 근거를 전혀 찾을 수 없는 것은 아니다. 특히 국가의 '정보의 개발의무'를 부여하고 국가표준제도의 확립을 명하는 헌법 제127조가 통계제도의 근거로 활용될 수 있을 것이다. 즉, 헌법 제127조는 "① 국가는 과학기술의 혁신과 정보 및 인력의 개발을 통하여 국민경제의 발전에 노력하여야 한다. ② 국가는 국가표준제도를 확립한다. ③ 대통령은 제1항의 목적을 달성하기 위하여 필요한 자문기구를 둘 수 있다."라고 규정하고 있는데, 동조 제1항에서의 "정보의 개발"의 의미는 곧 '정보의 전달가능성의 제고(전달능력의 개발)'와 '의미 있는 정보의 증가(개발)'를 의미하는 것으로 이해한다면,[53] 국가·사회의 의사결정의 토대를 마련하는 통계의 작성은 가장 기본적인 의미 있는 정보의 증가를 위한 작업이라 할 수 있을 것이다. 또 작성된 통계의 온라인 공개 및 활용 방안의 확대·개선 역시 정보의 전달가능성

52) 멕시코헌법(1857년 제정, 2013년 개정) 제26조
A. 국가는 국가의 독립과 정치, 사회, 문화적 민주화를 위하여 견고하고 역동적이며 경쟁력 있고, 지속적이며 공정한 경제성장을 보장하는 국가발전을 계획할 수 있는 국가 민주주의적 계획제도를 수립하여야 한다(2013년 6월 5일 본항 개정). (본항 하략)
B. 국가는 공식적인 데이터로 간주되는 국가통계지리정보시스템을 갖추어야 한다. 연방, 주, 연방직할시 및 지방자치단체는 법률이 정하는 바에 따라 이 시스템에 포함된 데이터를 사용하여야 한다. 국가통계지리정보시스템은 생성된 정보의 수집, 처리 및 공표를 규제하고 이를 집행하는데 필요한 권한이 있고, 기술 및 관리상 자치권, 법인격 및 자체 자산을 보유하는 기관이 규제하고 조정하여야 한다. 해당 기관은 5인으로 구성된 운영위원회를 두고, 위원 중 1인이 운영위원회와 기구의 장이 된다. 5인의 구성원은 상원 또는 휴회 중인 때에는 상임위원회의 동의를 받아 멕시코 대통령이 임명한다. 국가통계지리정보시스템의 조직 및 역할과 운영위원회 구성원의 자격요건, 임기 및 승진에 관한 사항은 정보에 대한 접근성, 투명성, 객관성 및 독립성의 원칙에 따라 법률로 정한다. 운영위원회의 구성원은 중대한 사유가 있는 경우에 한하여 해임될 수 있고, 다른 직무를 겸임할 수 없다. 다만, 무보수의 교육, 과학, 문화 또는 자선단체는 제외하고, 이 경우에는 헌법 제4편의 규정을 적용한다. (1983년 2월 3일, 2006년 4월 7일 본조 개정) * 동헌법의 번역문은 국회도서관 법률정보실 편, 『세계의 헌법: 35개국 헌법 전문 1·2』(서울: 국회도서관 법률자료과, 2013)의 내용을 활용하였음.

53) 이러한 법학적인 관점에서의 정보의 의미 및 그를 활용한 정보개발의무의 의미에 관한 논의로는 김주영, 『정보시장과 균형: 헌법사회학적 접근』(서울: 경인문화사, 2013), 47~54면을 참조.

제고의 측면에서 충분히 정당화 가능할 것으로 생각된다. 아울러 동조 제2항의 경우 국가표준제도의 확립을 규정하고 있는데, 이러한 국가표준 가운데 하나로 통계가 포함될 수 있을 것임은 이미 언급한 바 있다.[54]

다만 동 조항의 연혁을 살펴보자면, 동조항의 전신은 1962년 헌법 제118조로 당시 '경제·과학심의회의'의 필수적 설치를 규정하면서 헌법에 처음 도입되었고,[55] 1972년 헌법에서 제123조로 조문이동하면서 과학기술의 창달, 진흥 필요성의 선언과 함께 관련 자문기구의 임의적 설치를 규정하는 방식으로 개정되었다가,[56] 1980년 헌법 제128조에서 현행 헌법 조항의 기본적인 내용 및 형태를 마련하고,[57] 1987년 현행 헌법에서 '정보 및 인력의 개발'이 추가되는 과정을 겪은 바 있기에, 역사적 해석방법의 측면에서 이러한 연혁을 고려하면 동 조항을 1948년부터 존재해 온 통계제도의 근거로 활용하기에는 다소 적절치 못한 감이 없지 않을 것이다. 그렇지만, 목적론적 해석의 측면에서는 물론이거니와, 동 조항이 정보화의 추세를 반영하여 1987년 헌법에 도입된 점을 감안한다면, 특히 현재의 새로운 정보기술 환경에서의 통계제도의 논거로 활용하기에는 역사적 해석의 측면에서도 부족함이 없으리라 생각한다.

그 밖에도 특히 학술적인 측면이나 기업경영 등과 관련한 경제적인 측면에서 주로 이루어지는 민간통계제도의 경우에는 학문의 자유를 규정한 헌법 제22조[58]와 영업의 자유를 포함하는 것으로서의 직업(선택)의 자유를 규정한 헌법 제15조 역시 주요한 헌법적 근거로 활용할 수 있을 것이다.

54) 예를 들어 김일환, "헌법 제127조", 법제처 편, 『헌법주석서IV: 법원 등에 관한 장(제101조부터 제130조까지)』 (서울: 법제처, 2010), 576면은 정부가 산업과학기술과 정보화 사회에 필요한 '참조표준'을 제정·평가하고 보급해야 한다면서 이러한 참조표준을 '측정데이터 및 정보의 정확도와 신뢰도를 과학적으로 분석·평가하여 공인된 것으로서 국가사회의 모든 분야에서 널리 지속적으로 사용되거나 반복사용할 수 있도록 마련된 물리화학적 상수, 물성(物性)값, 과학기술적 **통계**'로 정의하고 있다(강조는 필자).

55) 1962년 헌법 제118조 ①국민경제의 발전과 이를 위/47한 과학진흥에 관련되는 중요한 정책수립에 관하여 국무회의의 심의에 앞서 대통령의 자문에 응하기 위하여 경제·과학심의회의를 둔다.
②경제·과학심의회의는 대통령이 주재한다.
③경제·과학심의회의의 조직·직무범위 기타 필요한 사항은 법률로 정한다.

56) 1972년 헌법 제123조 ①국민경제의 발전과 이를 위한 과학기술은 창달·진흥되어야 한다.
②대통령은 경제·과학기술의 창달·진흥을 위하여 필요한 자문기구를 둘 수 있다.

57) 1980년 헌법 제128조 ①국가는 국민경제의 발전에 노력하고 과학기술을 창달·진흥하여야 한다.
②국가는 국가표준제도를 확립한다.
③대통령은 제1항의 목적을 달성하기 위하여 필요한 자문기구를 둘 수 있다.
한편 이러한 과학기술 정책에 대한 대통령의 자문을 위한 기관으로는 1989년 6월 한시적으로 국가과학기술자문회의가 설치·운영되다가 현재는 1991년 제정된 「국가과학기술자문회의법」에 따라서 '국가과학기술자문회의'가 설치되어 있다(의장은 대통령). 동 기관의 공식 홈페이지는 https://www.pacst.go.kr 이다.

58) 학문의 자유는 주관적 권리의 성격과 함께 객관적 질서의 성격을 모두 지니고 있다. 학문의 자유는 연구와 교수 및 연구결과의 발표나 학문 활동을 위한 집회·결사에 있어서 국가의 간섭이나 침해에 대한 방어권이라는 주관적 성격을 갖는다. 또한 이와 더불어 학문의 자유는 사회전체의 지적 수준을 향상시키고 문화국가질서를 형성하는 객관적 질서의 성격을 포함한다. 명재진, "헌법 제22조", (사)한국헌법학회 편, 『헌법주석1』 (서울: 박영사, 2013), 786면.

나. 법률적 규율 양상

세부적으로 우리나라에서 국가가 통계자료를 수집, 편제하는 법적 근거를 살펴보자면, 가장 기본적인 법률로서 「통계법」(법률 제13818호)을 들 수 있다. 이 「통계법」은 통계의 작성·보급 및 이용과 그 기반구축 등에 관하여 필요한 사항을 정함으로써 통계의 신뢰성과 통계제도 운용의 효율성을 확보함을 목적(동법 제1조)으로 제시하는 가운데 통계에 있어서의 국가 등의 책무를 규정하고(동법 제4조)[59] 또 국가통계 발전을 위한 중장기 정책목표 및 추진방향을 설정하여 그에 따른 국가통계 발전 기본계획을 수립·추진하도록 규정하고 있으며(동법 제5조의4),[60] 통계의 작성·보급 및 이용, 그리고 통계응답자의 의무 및 보호에 걸친 다양한 규정들을 마련하고 있다.[61] 특히 "통계의 작성·보급 및 이용에 관하여 **다른 법률에 특별한 규정이 있는 경우를 제외하고는** 이 법으로 정하는 바에 따르도록" 규정하고 있으므로(동법 제5조), 「통계법」은 통계에 관한 일반법으로 자리매김하고 있다 할 수 있을 것이다. 그렇지만 우리나라에서 작성하는 개별적인 통계들은 동조항이 예정하고 있듯이 「통계법」 이외에도 다수의 각 사업 내지 행정목적별 법률에 자료수집 및 작성의 근거를 두고 있는 까닭에, 법적용의 일반원칙 가운데 하나인 '특별법 우선의 원리("특별법은 일반법에 우선한다.; *Lex specialis derogat legi generali.*")'에 따라 개별적인 통계의 법적 규율양상은 보다 다양하게 나타날 가능성이 크다.

「통계법」 이외의 법적 규율하의 통계의 운용양상들을 좀 더 구체적으로 살펴보면 우선적으로 「통계법」과의 관련 속에서, 통계작성에 있어서 '통계청장과의 협의'를 요구하는 경우가 있다. 예를 들자면 「지능형전력망의 구축 및 이용촉진에 관한 법률」(법률 제12154호)[62]은 통계작성에 있어서 통계청장과의 협의의무를

59) 「통계법」 제4조(국가 등의 책무) ① 국가 및 지방자치단체는 이 법의 목적과 기본이념을 구현하기 위하여 필요한 정책을 수립·시행하여야 한다.
② 통계청장은 통계가 사회발전에 이바지할 수 있도록 통계에 관한 사항을 종합적으로 조정·정비하고, 통계의 작성·보급 및 이용을 확대할 수 있는 조치를 강구하여야 한다.
③ 통계작성기관의 장은 통계의 작성을 위하여 질문을 받거나 자료제출 등의 요청을 받고 답변을 하거나 자료제출 등을 하는 개인이나 법인 또는 단체 등(이하 "통계응답자"라 한다)의 부담을 최소화하고, 비밀이 보호되도록 노력하여야 한다.
④ 통계작성기관의 장은 통계종사자의 교류, 통계작성기법의 공동연구와 개발 및 통계자료의 공유 등을 위하여 서로 협력하여야 한다.

60) 2013년 10월 국가통계위원회의 심의·의결을 거친 제1차 국가통계발전('13~'17)기본계획의 내용은 http://kostat.go.kr/portal/korea/kor_pi/6/1/index.board?bmode=read&bSeq=&aSeq=309050&pageNo=1&rowNum=10&navCount=10&currPg=&sTarget=title&sTxt (2016.9.18. 최종확인)에서 확인할 수 있다.

61) 아울러 기타 세부적인 사항을 규율하기 위해 「통계법시행령」(대통령령 제26531호), 「통계법시행규칙」(기획재정부령 제497호)이 마련되어 있다.

62) 「지능형전력망의 구축 및 이용촉진에 관한 법률」 제8조(지능형전력망 통계의 작성 및 공개) ① 산

규정하고 있다. 다음으로 「통계법」의 준용을 규정하고 있는 법률로 「정보통신산업 진흥법」(법률 제13015호)[63]의 경우에는 통계작성에 있어서 통계청장과의 협의의무뿐만 아니라 통계작성과정에 있어서 「통계법」의 준용까지 규정하고 있다. 그 밖의 「통계법」의 준용을 규정하고 있는 법률로 「전자문서 및 전자거래 기본법」(법률 제13768호)[64]의 경우 통계작성은 「통계법」을 준용하도록 규정하고 있기는 하지만, 통계의 실태조사는 대통령령을 통해 독자적인 근거를 마련하고 있고, 「지능형 로봇 개발 및 보급 촉진법」(법률 제13744호)[65]도 이와 유사한 입법태도를 보이고 있다.

그렇지만 규정양상만을 놓고 본다면 독자적인 통계운용이 오히려 더 일반적이라고까지 말할 수 있을 정도이다. 예를 들어 「고용정책 기본법」(법률 제13262호)[66], 「문화산업진흥 기본법」(법률 제13448호)[67], 「사회보장기본법」

업통상자원부장관은 지능형전력망에 관한 계획을 효율적으로 수립·시행하기 위하여 **통계청장과 협의**하여 지능형전력망에 관한 통계를 작성·관리하여야 한다.

② 산업통상자원부장관은 지능형전력망의 이용을 촉진하기 위하여 제1항에 따른 통계를 공개하여야 한다. 다만, 「공공기관의 정보공개에 관한 법률」 제9조에 따른 비공개대상정보는 그러하지 아니하다.

63) 「정보통신산업 진흥법」 제6조(통계의 작성) ① 미래창조과학부장관은 진흥계획의 효율적인 수립·시행을 위하여 **통계청장과 협의**하여 정보통신산업에 대한 통계를 작성·관리하여야 한다.

② 제1항에 따른 통계는 **「통계법」을 준용하여 작성**하되, 조사 대상 및 범위 등에 관하여는 미래창조과학부령으로 정한다.

64) 「전자문서 및 전자거래 기본법」 제28조(전자문서 및 전자거래 통계의 실태조사) ① 미래창조과학부장관은 전자문서·전자거래촉진정책의 효과적인 수립·시행을 위하여 전자문서 및 전자거래 통계의 실태조사를 실시할 수 있다. 이 경우 전자문서 및 전자거래 통계의 작성에 관하여는 「통계법」을 준용한다.

② 미래창조과학부장관은 제1항에 따른 전자문서 및 전자거래 통계의 실태조사를 위하여 필요한 경우에는 다음 각 호의 어느 하나에 해당하는 자에 대하여 자료의 제출이나 의견의 진술 등을 요구할 수 있다.

1. 국가기관 등
2. 전자거래사업자
3. 전자문서 또는 전자거래 관련 법인·단체

③ 제2항에 따라 자료의 제출 등을 요구받은 자는 이에 협조하여야 한다.

④ 전자문서 및 전자거래 통계의 실태조사에 필요한 사항은 대통령령으로 정한다.

65) 「지능형 로봇 개발 및 보급 촉진법」 제7조(산업통계 및 실태조사) ① 정부는 지능형 로봇의 효율적인 기술개발과 보급·확산을 위하여 지능형 로봇산업의 분류체계를 구축하고 분류체계에 따른 산업통계를 확보하여야 한다. 이 경우 산업통계를 작성함에 있어서는 「통계법」을 준용한다.

② 산업통상자원부장관은 지능형 로봇산업 관련 정책의 효과적인 수립·시행과 제1항의 산업통계 확보를 위하여 매년 지능형 로봇산업 전반에 걸친 실태조사를 실시하여야 한다.

③ 산업통상자원부장관은 제2항에 따른 실태조사를 위하여 필요한 경우에는 지능형 로봇 관련 사업자 또는 지능형 로봇 관련 법인·단체에 대하여 자료의 제출이나 의견의 진술 등을 요구할 수 있다. 이 경우 자료의 제출이나 의견의 진술 등을 요구받은 지능형 로봇 관련 사업자 또는 지능형 로봇 관련 법인·단체는 특별한 사유가 없는 한 이에 협조하여야 한다.

④ 제1항에 따른 산업통계 작성대상의 범위 및 제2항에 따른 실태조사대상 등에 관하여 필요한 사항은 대통령령으로 정한다.

66) 「고용정책 기본법」 제17조(고용 관련 통계의 작성·보급 등) ① 고용노동부장관은 고용정책의 효율적 수립·시행을 위하여 산업별·직업별·지역별 고용구조 및 인력수요 등에 관한 통계를 작성·공표하여 국민들이 이용할 수 있도록 하여야 한다.

② 고용노동부장관은 제1항에 따라 작성된 통계를 국민들이 편리하게 이용할 수 있도록 데이터베이스를 구축하는 등 필요한 조치를 하여야 한다.

(법률 제13650호)[68], 「조달사업에 관한 법률」(법률 제13817호)[69], 「평생교육법」(법률 제13945호)[70], 「폐기물관리법」(법률 제13411호)[71], 「환경보건법」(법률 제13893호)[72], 「한국은행법」(법률 제14101호)[73] 「공공데이터의 제공 및 이용 활성화에 관한 법률」(법률 제13723호)[74] 등이 「통계법」과의 관련성을 그다지 고려하지 않은 채 독자적으로 통계에 관한 내용들을 규율하고 있다.

다만 주의 깊게 살펴볼 필요가 있는 것으로는 **법률이 아닌 형식의 근거를 가진 통계조사**들이다. 예를 들어 국토교통부장관은 「국토기본법」(법률 제12738호)

67) 「문화산업진흥 기본법」 제30조의3(문화산업통계의 조사) ① 문화체육관광부장관은 중·장기기본계획을 효과적으로 수립·시행하고 문화산업에 활용하는 것을 촉진하기 위하여 국내외의 실태조사를 통한 문화산업통계를 작성할 수 있다.

② 문화산업통계의 작성·관리에 필요한 사항은 대통령령으로 정한다.

68) 「사회보장기본법」 제32조(사회보장통계) ① 국가와 지방자치단체는 효과적인 사회보장정책의 수립·시행을 위하여 사회보장에 관한 통계(이하 "사회보장통계"라 한다)를 작성·관리하여야 한다.

② 관계 중앙행정기관의 장과 지방자치단체의 장은 소관 사회보장통계를 대통령령으로 정하는 바에 따라 보건복지부장관에게 제출하여야 한다.

③ 보건복지부장관은 제2항에 따라 제출된 사회보장통계를 종합하여 위원회에 제출하여야 한다.

④ 사회보장통계의 작성·관리에 필요한 사항은 대통령령으로 정한다.

한편 이 조항은 2012년 1월 26일에 전부개정된 「사회보장기본법」(법률 제11238호)에 의해 처음으로 도입된 것이다.

69) 「조달사업에 관한 법률」 제3조의7(조달통계) ① 조달청장은 공공조달의 현황을 파악하고 효과적인 조달정책을 수립·시행하기 위하여 국가기관, 지방자치단체 및 그 밖에 대통령령으로 정하는 기관(이하 이 조에서 "국가기관등"이라 한다)의 장이 체결한 계약에 관한 통계를 작성하여야 한다. 이 경우 조달청장은 필요한 자료를 국가기관등(조달청장이 체결한 계약에 관하여는 수요기관을 말한다)에 요구할 수 있다.

② 제1항에 따라 자료의 제출을 요구받은 국가기관등은 정당한 사유가 없으면 이에 따라야 한다.

③ 제1항에 따른 통계작성의 대상·방법 및 절차에 관하여 필요한 사항은 대통령령으로 정한다.

70) 「평생교육법」 제18조(평생교육 통계조사 등) ① 교육부장관 및 시·도지사는 평생교육의 실시 및 지원에 관한 현황 등 기초자료를 조사하고 이와 관련된 통계를 공개하여야 한다.

② 평생교육과 관련된 업무 담당자 및 평생교육기관 운영자 등은 제1항의 조사에 협조하여야 한다.

71) 「폐기물관리법」 제11조(폐기물 통계 조사) ① 환경부장관, 시·도지사 또는 시장·군수·구청장은 폐기물 정책의 수립에 필요한 기초자료를 확보하기 위하여 폐기물 종류별 발생·처리현황, 폐기물처리업 등 관련 산업 현황, 폐기물 재활용률 등 자원생산성 향상에 관한 사항 등을 조사하여야 한다.

② 제1항에 따른 조사의 항목, 시기 및 방법에 관한 사항은 환경부령으로 정한다.

72) 「환경보건법」 제22조(환경보건 정보와 통계의 관리) ① 환경부장관은 환경보건에 관한 정보와 통계를 수집하고 관리하여 국민건강 피해의 예방과 관리에 활용할 수 있도록 필요한 시책을 세우고 시행하여야 한다.

② 환경부장관은 제1항에 따른 환경보건에 관한 정보와 통계를 널리 보급하기 위하여 필요한 시책을 마련하여야 한다.

73) 「한국은행법」 제86조(통계자료의 수집·작성 등) 한국은행은 통화신용정책의 수립에 필요한 통화·은행업무·재정·물가·임금·생산·국제수지 또는 그 밖의 경제 일반에 관한 통계자료의 수집·작성과 경제에 관한 조사를 할 수 있으며, 이를 위하여 필요한 자료와 정보를 정부기관이나 법인 또는 개인에게 요구할 수 있다.

74) 「공공데이터의 제공 및 이용 활성화에 관한 법률」 제13조(공공데이터활용지원센터) ① 공공데이터의 효율적인 제공 및 이용 활성화 지원을 위하여 공공데이터활용지원센터(이하 "활용지원센터"라 한다)를 「국가정보화 기본법」 제14조에 따른 한국정보화진흥원에 설치·운영한다.

② 활용지원센터는 다음 각 호의 업무를 수행한다.

1.(생략)

2. 공공데이터의 제공 및 이용과 관련한 통계의 조사·분석

3.(후략)

제25조에 의거하여 국토에 관한 계획 또는 정책의 수립, 「국가공간정보 기본법」 제32조제2항에 따른 공간정보의 제작, 연차보고서의 작성 등을 위하여 필요할 때에는 미리 인구, 경제, 사회, 문화, 교통, 환경, 토지이용, 그 밖에 대통령령으로 정하는 사항에 대하여 조사할 수 있는데('국토조사'), 「국토기본법 시행령」(대통령령 제26922호) 제10조의2는 국토조사 성과의 효율적인 관리 및 활용을 위하여 국토조사를 이용한 **국토통계지도**의 구축, 유지·관리 및 활용 업무를 수행할 것을 규정하고 있다(동시행령 동조 제3호). 한편 광업통계 작성사업의 경우는 「광업법」(법률 제12738호) 제86조 제1항 제4호에 의해 산업통상자원부장관이 광업의 발전을 위하여 예산의 범위 내에서 지원할 수 있는 국가 지질 또는 광물 자원의 조사 사업 및 광물자원 관련 기술 개발사업 가운데 하나로 규정되어 있다(「광업법 시행령」(대통령령 제26703호) 제62조 제4항 제4호). 이와 유사하게 산업통상자원부 산하 지역발전지원단의 업무 가운데 '지역발전 교육 및 **통계** 구축, 연차보고서 작성에 관한 업무'가 「국가균형발전 특별법 시행령」(대통령령 제25249호)에 의해 부여되어 있고(동시행령 제31조 제1항), 해양수산부장관 및 지방자치단체장이 낚시 및 낚시 관련 산업을 지원·육성하기 위해 실시하는 사업 가운데, **낚시정책 수립**을 위한 실태조사 및 **통계조사** 사업이 포함되며(「낚시 관리 및 육성법 시행령」 (대통령령 제26774호) 제1항 제1호), 법무부장관이 작성·관리하는 **법무행정 관련 인권침해 사건의 조사 결과 등에 관한 통계** 역시 법무부령(「인권침해 사건 조사·처리 및 구금·보호시설의 실태조사에 관한 규칙」(법무부령 제745호) 제26조)에 근거하고 있다. 그 밖에도 국민의 삶의 질 향상과 지역 간 문화 격차의 해소를 통한 국민의 문화 향유권의 확대를 위하여 문화 향유와 관련한 실태조사와 관련 조사·연구를 시행하기 위해서 국가와 지방자치단체가 대통령령으로 정하는 바에 따라 문화정책을 전문적으로 조사·연구·개발하는 전담기관과 이를 지원하는 문화정보화 전담기관을 지정·운영할 수 있는데(「문화기본법」(법률 제12134호) 제11조), 이러한 문화정보화 전담기관의 업무 가운데 **'통계시스템을 통한 문화 분야 각종 조사 결과와 통계정보의 통합관리'**(「문화기본법 시행령」(대통령령 제25268호) 제7조 제3항 제7호)가 규정되어 있기도 하다.

그렇지만 이렇게 법률이 아닌 형식의 근거를 가지는 통계조사들은 우선 이미 언급한 바 있는 「통계법」의 '다른 **법률**과의 관계' 조항(동법 제5조)의 해석상 다소간의 문제를 지적받을 수도 있을 것인데 「통계법」은 다른 **법률**의 특별한 규정이 없는 한 일반적으로 적용되도록 규정되어 있기 때문에, 법률의 형식이 아닌 근거를 가지는 통계조사들의 경우에는 문리해석상 「통계법」의 규정들이 우선적으로 적용된다고 할 것이다. 아울러 이러한 법률이 아닌 형식의 근거를

가지는 통계조사들은 후술하듯이 **'법률유보(Vorbehalt des Gesetzes)'**의 관점에서 볼 때 상당한 문제점을 야기할 수 있음을 염두에 두어야 한다.

제5절 통계작성주체

거듭 이야기하게 되지만 현대 국가는 정책의 수립과 이행, 평가과정에서 다양한 통계를 필요로 하다 보니, 각 국가는 국가통계의 최대 생산자이자 이용자인 정부가 필요로 하는 통계를 효율적으로 생산하고 활용하기 위해 각국의 역사적 배경과 정치적 여건에 부합하는 통계제도를 운영하고 있다. 이러한 통계제도는 국가통계 생산구조에 따라 크게 집중형과 분산형으로 구분되곤 한다.

집중형 통계제도는 하나의 전문화된 통계작성기관이 국가정책에 필요한 대부분의 통계를 작성·공급하는 제도로 캐나다, 스웨덴, 호주, 네덜란드, 인도네시아 등이 이 제도를 채택하고 있다. 반면 우리나라나 미국, 일본, 영국 등은 국가통계를 생산하는 데 있어 각 기관의 고유 업무 수행을 위해 필요한 통계를 개별 기관 책임아래 작성하고 있는 분산형 통계제도를 채택하고 있다. 국가규모가 크고 정부기구가 방대한 국가일수록 특정 통계생산기관이 국가통계의 모든 수요를 파악하기 어렵고 또 각 정부기구가 독자적 통계를 생산할 자원과 능력을 보유하고 있기 마련이므로 분산형 통계제도를 채택하는 경향이 강하다고는 하나,[75] 실제 대부분의 국가에서는 어느 한 형태만을 전적으로 채택하지는 않고 양자가 혼재하고 있기 때문에 어느 쪽에 더 근접해 있느냐에 따라 구분된다고 할 수 있다.[76] 두 제도의 차이점은 일반적으로 다음과 같이 정리된다.[77]

75) 이재형, 앞의 보고서, 25면.

76) 김기환 연구책임, 『국내외 통계제도 및 통계작성현황 비교분석 연구용역(최종보고서)』 (한국: 통계개발원, 2009), 5면. 주요 선진국가의 통계제도에 대한 보다 상세한 내용은 동 보고서, 59~280면을 참조.

77) 이재형, 앞의 보고서, 26면의 <표2-4>를 바탕으로 작성.

<표 10> 집중형 통계제도와 분산형 통계제도의 특징통계청의 데이터 분류

	집중형	분산형
주요 특징	• 국가기본통계를 단일화된 통계전문기관에서 작성하여 각 이용자에게 제공 • 부처간 통계연결기구의 설치 필요	• 부처별로 필요한 통계를 자체 작성 활용 • 통계조정기관 설치 필요
장점	• 통계의 균형적 개발과 체계화 용이 • 통계의 객관성과 신뢰성 확보 • 통계전문인력과 장비의 효율적 활용	• 업무분야의 전문지식을 통계 작성에 활용 가능 • 통계수요에 신속히 대응
단점	• 행정업무분야의 전문지식활용 곤란 • 통계수요에 대한 신속한 대응 곤란	• 통계작성의 중복과 불일치로 예산과 인력의 낭비 초래 • 통계전문요원과 장비의 집중활용 곤란으로 비경제적
예	• 캐나다, 독일, 스웨덴, 호주, 네덜란드 등	• 미국, 일본, 영국, 한국, 대만 등

우리나라의 통계제도는 미국과 일본의 영향을 받아 원칙적으로 분산형 통계제도로 출발하였으나 미국이나 일본보다는 상대적으로 집중형 성격이 강하다고 평가된다. 1962년 제정된 「통계법」은 정부를 위시한 각종 통계기관에서 독자적인 통계활동을 수행할 수 있는 분산형 통계제도를 반영하고 있었다. 그렇지만 분산형 통계제도에서는 통계활동의 중복으로 인한 자원의 낭비와 국민의 응답부담가중, 관련 통계 상호간의 비교성 결여 등의 이유로 국가통계의 질적 수준 저하 등과 같은 문제가 발생할 가능성이 높다. 따라서 국가중앙통계기관인 통계청[78]은 이러한 문제들을 최소화하고 국가통계의 체계적인 발전을 위해 통계생산기관으로서의 역할 뿐만 아니라 국가통계조정기관으로서의 기능을 동시에 수행하고 있다. 또한 국가통계포털과 같은 집중형 통계서비스를 통해 분산형 제도로 인해 발생할 수 있는 이용자의 불편사항을 최소화 하는데 노력하고 있다.

세부적으로 우리나라는 분산형 통계제도에 따라 380여 개의 기관이 약 900종의 국가통계를 생산하고 있으며 통계작성기관은 중앙행정기관, 지방자치단체 및 통계작성 지정기관(민간기관) 등으로 나뉜다. 이 중, 중앙행정기관은 통계청을 포함한 약 40개의 개별 부처로 구성되어 있으며 부처별 고유 정책에 필요한 통계를 생산한다. 지방자치단체인 광역 및 기초자치단체 등 약 260개의 기관도

78) 본절의 서두에서 잠시 언급하기도 하였지만, 통계청은 연혁적으로 1948년 11월 공보처 통계국으로 출발하여 1955년 내무부 통계국, 1961년 경제기획원 통계국, 1963년 경제기획원 조사통계국을 거쳐 1990년 비로소 통계청으로 승격·개편된 후 오늘에 이르고 있다. 현재 기획재정부 산하에 국세청, 관세청, 조달청과 함께 규정된 특허청은 통계의 기준설정과 인구조사 및 각종 통계에 관한 사무를 관장하기 위하여 기획재정부장관 소속으로 설치되어 있으며, 통계청에 청장 1명과 차장 1명을 두되, 청장은 정무직으로 하고, 차장은 고위공무원단에 속하는 일반직공무원으로 보하게끔 되어 있다((「정부조직법」 제27조 제9항, 제10항). 세부적인 사항은 「통계청과 그 소속기관 직제」(대통령령 제27147호) 및 「통계청과 그 소속기관 직제 시행규칙」(기획재정부령 제562호)이 규정하고 있으며, 「국가통계위원회 규정」(대통령령 제27129호)이 마련되어 있다.

각 지역에 필요한 통계를 직접 생산하고 있다. 한편 이미 언급한 바와 같이 통계작성 지정기관은 중앙행정기관이나 지방자치단체 이외의 법인으로 국가통계를 작성할 능력이 있는 기관이 통계청장의 승인 후 통계작성 지정기관이 될 수 있는데 금융, 공사 및 공단, 연구기관, 협회 및 조합 등이 해당되며 전장에서 살펴본 바와 같이 현재 약 80개의 기관으로 구성되어 있다.[79]

제6절 통계의 운용

가. 국가통계 기본원칙

국가통계를 생산하는 통계작성기관은 국민의 신뢰를 얻고 고객을 만족시키기 위해 노력하여야 하기 때문에 국가통계작성기관 및 그 종사자들의 자율성을 보장하고 책임성을 강화하기 위하여 통계청은 2011년 9월 국가통계 기본원칙을 정립하여 발표한 바 있다. 그 내용은 다음과 같다.[80]

<표 11> 국가통계 기본원칙

	내용	실천방안
중립성 보장	국가통계는 공익적 가치를 가진 공공재로서 중립성이 보장되어야 한다.	• 통계 작성 및 공표와 관련하여 통계작성기관은 정책기관, 이익단체 등의 영향으로부터 자유로워야 함 • 통계의 작성 방법 및 절차, 공표의 내용 및 일정 등을 독립적으로 결정할 수 있어야 함
신뢰성 제고	국가통계는 객관적이고 과학적인 방법을 사용하여 정확하고 신뢰할 수 있도록 작성되어야 한다.	• 통계적으로 널리 활용되는 과학적인 작성기법을 사용하여야 함 • 조사기획, 자료수집에서 공표에 이르는 적절한 통계작성절차를 수립하고 이를 준수하여야 함 • 통계의 품질을 관리하고, 이를 통해 나타난 문제점을 개선하는 등 품질 제고를 위해 지속적으로 노력하여야 함 • 미리 예정된 결과를 상정하고 통계를 작성하지 않도록 함
효율성 제고	국가통계 작성을 위한 비용, 응답 및 조사부담 등을 고려한 계획을 수립하여 효율적인 조사가 이루어지도록 한다.	• 유기적인 계획수립, 통계작성기법의 적용, 정보통신기술을 활용하여 정확하고 효율적인 조사가 이루어지도록 함 • 작성하고 있는 통계자료를 공유하여 통계의 중복작성을 방지하고 조사 부담을 최소화 하여야 함 • 이용 가능한 행정자료를 활용하여 응답부담을 경감하여야 함

79) 자세한 사항은 국가통계포털. http://kosis.kr/serviceInfo/serviceInfo_0202List.jsp (2016.9.18. 최종확인)
80) 통계청, "국가통계 기본원칙 전문" http://kostat.go.kr/portal/korea/kor_ko/8/index.static (2016.9.18. 최종확인)을 바탕으로 표로 작성.

	내용	실천방안
비교 가능성	국가통계는 다른 통계와 비교하여 사용할 수 있도록 비교가능한 개념, 분류, 방법 등을 사용하여야 한다.	• 통계는 국·내외 기준 등을 고려한 표준화된 방법을 적용하여 비교할 수 있도록 작성하여야 함 • 작성체계 또는 자료수집·분류 방법 등이 변경된 경우 신구통계간 일관성을 유지할 수 있도록 노력하여야 함 • 다른 통계와의 비교를 위하여 개념, 정의, 모집단 구성, 표본추출방법, 분류기준, 통계작성방법 등을 기술하여 공표하여야 함 • 통계의 가치증진을 위하여 국내는 물론 국가간, 국제기구와 협력을 강화하고 지원 및 기술습득을 할 수 있도록 노력하여야 함
비밀 보호	개인이나 법인 또는 단체 등의 비밀에 속하는 자료는 통계목적으로만 사용되어야 하고 엄격히 보호되어야 한다.	• 통계작성과정에서 알게 된 개별정보를 통계목적 외로 사용하거나 타인에게 제공하지 않도록 함 • 통계작성과정에서 수집된 개별정보가 공개되지 않도록 관리하고 자료제공시 비밀을 보호하기 위해 노력하여야 함
인프라 확충	국가통계 작성에 필요한 인력, 예산, 전산 장비 및 프로그램 등을 충분히 확보하여야 한다.	• 통계를 전담하는 조직을 운영하고 전문성을 갖춘 인력을 확보할 수 있도록 노력하여야 함 • 표본설계, 자료수집, 분석 등에 대한 연구 및 교육을 지속적으로 실시하여 전문성을 유지하도록 하여야 함 • 사회경제적 변화에 따른 이용자들의 수요에 부응한 통계 개발 및 개선에 필요한 예산 등을 확보하여야 함
이용자 참여	국가통계의 실용성을 향상시키고, 공익적 가치를 극대화하기 위하여 이용자들을 효과적으로 참여시켜야 한다.	• 통계자료의 활용성을 높이기 위하여 관련기관 및 전문가들과 긴밀하고 지속적인 관계를 발전시켜야 함 • 전문가 회의, 이용자만족도 조사 등을 통해 소관 통계에 대한 개발·개선 요구, 기타 수요에 대한 사항을 반영할 수 있도록 노력하여야 함
서비스 향상	국가통계는 모든 이용자들이 쉽고 편리하게 접근하여 활용할 수 있어야 한다.	• 모든 이용자들이 동일한 시간에 동등한 권한으로 공표자료에 접근할 수 있어야 함 • 이용자들이 통계의 공표일정을 미리 알 수 있도록 사전에 예고하고 그 일정을 준수하여야 함 • 통계자료는 기준시점과 발표시점의 시차를 최소화하여 즉시 공표하여야 함 • 언론 보도, 보고서 발간, 인터넷 등 다양한 방법으로 통계 결과물을 이용자들에게 제공하여야 함 • 통계자료 서비스에 대한 이용자들의 요구사항을 반영할 수 있도록 노력하여야 함

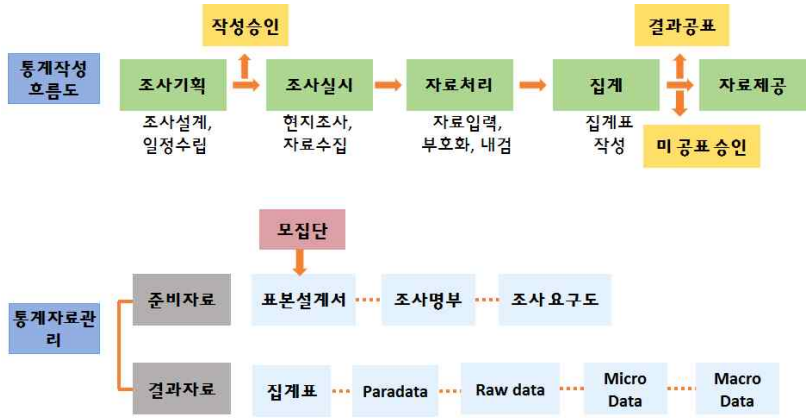
이 국가통계 기본원칙은 이미 살펴 본 바 있는 통계의 특성-**익명성, 비교성·객관성, 정확성**-을 확보하기 위해 필요한 세부적인 원칙들을 잘 설시하고 있는 것으로 통계작성에 있어서 반드시 준수하고 또 지향해야 할 원칙들이라 할 수 있을 것이다. 다만 이 가운데 '비밀보호' 부분은 전통적인 우리 헌법상의 기본적 권리로 평가되어 온 사생활의 비밀과 보호(헌법 제17조) 등에 기반을 둔 것으로 그 가치를 평가할 수 있을 것이나, 오늘날 개인정보 자기결정권의 보호대상으로서의 개인정보가 개인의 내밀한 영역에 속하는 정보에 국한되지 않고 공적 생활에서 형성되었거나 이미 공개된 개인정보까지 포함하는[81] 것을 감안한다면, 비밀보호는

81) "개인정보 자기결정권의 보호대상이 되는 개인정보는 개인의 신체, 신념, 사회적 지위, 신분 등과 같이 인격주체성을 특징짓는 사항으로 서 개인의 동일성을 식별할 수 있게 하는 일체의 정보를 의미하며, 반드시 개인의 내밀한 영역에 속하는 정보에 국한되지 않고 공적 생활에서 형성되었거나 이미 공개된 개인정보까지도 포함한다." 대법 2014.07.24. 선고 2012다49933; 대법 2016.3.10. 2012다105482. 이 판결과 관련한 법적인 논의는 박훤일, "정부기관의 정보열람 요구와 ISP의 협조의무", 『

물론 개인정보의 보호를 포함할 수 있는 확대된 범주의 원칙설정이 필요할 것으로 판단된다.

나. 통계작성과정의 법적 성격

실제 통계는 다양한 과정을 거쳐 생산된다. 일반적으로 국가통계에 있어서 통계작성의 흐름 및 통계자료의 관리방식은 다음의 [그림 9]과 같이 정리해 볼 수 있다.[82]



[그림 9] 통계작성 관리 흐름도

그 동안 통계의 운용에 있어서 구체적인 논의의 출발점이 될 수 있을 통계작성과정의 법적 성격에 관한 기존 논의는 거의 존재하지 않았다. 그렇지만 통계작성의 실제에 있어서 개인의 개인정보자기결정권 및 사생활의 비밀을 비롯한 기본권 침해문제 및 경제적 가치가 있는 자료의 입수과정에서의 논란 등의 실질적인 문제가 가장 빈번하게 발생할 수 있을 부분이 바로 통계, 특히 통계자료를 만드는 과정에 있어서의 통계조사인 만큼, 통계조사의 법적 성격에 대한 논의는 필수적이라 할 수 있을 것이다.

생각건대 민간통계의 일환으로 이루어지는 학술통계 및 경영통계의 자료조사 및 통계작성 과정은 원칙적으로 개인적·사적 행위에 해당되는 것이라 하지 않을 수 없겠지만,[83] 국가통계조사의 제과정 중 특히 민간부문의 자료에 대한

경희법학』, 제51권 제2호(2016)를 참조.

82) 김두만, "국가통계작성 기획 및 승인관리", 제5회 국가통계방법론 심포지엄(2015).

83) 본고에서 엄밀하게 논의하기는 곤란하지만, 이러한 민간통계는 헌법상 학술행위, 경영행위 기타

통계작성을 위한 '조사의 실시' 작업은 일정부분 '행정조사'의 성격을 가지는 것으로 보아야 할 것으로 본다.[84] 이와 관련하여 현행 법체계에 있어 행정조사의 일반법이라 할 수 있는[85] 「행정조사기본법」(법률 제11690호)은 행정조사를 실시하고자 하는 행정기관의 장은 출석요구서, 보고요구서·자료제출요구서 및 현장출입조사서(이하 "출석요구서등"이라 한다)를 조사개시 7일 전까지 조사대상자에게 서면으로 통지하여야 하나, 「통계법」 제3조제2호에 따른 지정통계의 작성을 위하여 조사하는 경우에는 행정조사의 개시와 동시에 출석요구서등을 조사대상자에게 제시하거나 행정조사의 목적 등을 조사대상자에게 구두로 통지할 수 있도록 규정하고 있어(동법 제17조 제1항 제2호), 「통계법」상 지정통계의 작성을 위한 조사는 행정조사의 일종임을 분명히 하고 있다. 그렇지만 「통계법」상 지정통계가 아닌 일반통계의 경우에도 행정조사의 성격을 부정하긴 어려울 것이다. 다만 이러한 「행정조사기본법」의 입법태도를 감안하면 국가통계(승인통계) 가운데 지정통계와 지정통계가 아닌 일반통계는 다소 취급이 달라진다는 점을 확인할 수 있는데, 지정통계의 경우에는 「통계법」과 함께 「행정조사기본법」 제17조 제1항 제2호 단서가 적용되고, 지정통계가 아닌 일반통계에는 「통계법」과 함께 「행정조사기본법」 제17조 제1항 제2호 본문이 적용되게 될 것이다.[86] 이는 통계자료의 조사 대상인 정보의 주체로서의 국민과의 관계에 있어서 특히 중요성을 가지게 된다.

다. 통계작성과 법률유보

일반적 행동자유권에 의해 보호를 받게 될 것이며, 명시적인 적용법규는 존재한다고 하기 어렵겠지만, 「통계법」상 일부 규정의 적용을 적극적으로 검토해볼 필요가 있을 것이다.

84) 최승필, 앞의 논문, 398면도 결국은 같은 입장이다. 행정법학의 경우, 통계의 성격에 대한 본격적인 논의가 이루어진 것으로 보기는 어렵지만, 통계조사를 행정조사의 하나로 파악하는 입장이 적지 않다. 예를 들어, 정하중, 『행정법개론(제8판)』 (파주:법문사, 2014), 493면; 김동희, 『행정법I(제20판)』 (서울: 박영사, 2014), 500면은 "통계법에 의한 통계조사"를 행정조사 가운데 '일반(적) 조사'의 예로 거론하고 있고, 홍정선, 『행정법원론(상)(제22판)』(서울: 박영사, 2014), 677면은 좀 더 구체적으로 "통계법상 국세조사"를 그 예로 들고 있다.

85) 「행정조사기본법」 역시 행정조사에 관하여 다른 법률에 특별한 규정이 있는 경우를 제외하고는 이 법으로 정하는 바에 따른다는 규정을 두어 행정조사에 관한 일반법으로서의 성격을 분명히 하고 있으면서도(법 제3조 제1항), 「행정조사기본법」에 의해 동법의 적용이 제외되는 일부 예외사항(동조 제2항)에 대해서도 법 제4조(행정조사의 기본원칙), 법 제5조(행정조사의 근거) 및 법 제28조(정보통신수단을 통한 행정조사)는 적용한다는 규정을 두어(동조 제3항) 일반적인 일반법(lex generali)보다는 더 개선된 규율능력을 보유하고 있다. 이러한 입법형식은 통계의 일반법인 「통계법」에도 적지 않은 시사점을 준다고 할 수 있을 것이다.

86) 이러한 차이점은 결국, 「통계법」 제3조제2호에 따른 지정통계의 작성을 위한 조사의 경우, 행정조사의 개시와 동시에 출석요구서 등을 조사대상자에게 제시하거나 행정조사의 목적 등을 조사대상자에게 **구두**로 통지하는 것도 가능하지만, 그 외의 통계의 경우는 통계자료 작성을 위한 조사시 작성기관의 장이 출석요구서, 보고요구서·자료제출 요구서 및 현장출입조사서를 조사개시 7일 전까지 조사대상자에게 **서면**으로 통지해야만 하는 것으로 나타난다.

법률유보(Vorbehalt des Gesetzes)의 원칙은 국가작용 특히 일정한 행정권의 발동은 법률에 근거하여 이루어져야 한다는 공법상 원칙으로, 헌법상의 "법치주의(rule of law)" 원칙의 행정법 영역에서의 구체적 표현이라 파악되기도 한다.[87] 헌법재판소 역시 국민주권주의, 권력분립주의 및 법치주의를 기본원리로 채택하고 있는 우리 헌법상 국민의 헌법상 기본권 및 기본의무와 관련된 중요한 사항 내지 본질적인 내용에 대한 정책형성기능은 원칙적으로 주권자인 국민에 의하여 선출된 대표자들로 구성되는 입법부가 담당하여 법률의 형식으로써 이를 수행하여야 하고, 이와 같이 입법화된 정책을 집행하거나 적용함을 임무로 하는 행정부나 사법부에 그 기능을 넘겨서는 아니 된다면서 이러한 법률유보의 원칙을 설시한 바 있다.[88] 따라서 이러한 법률유보의 원칙은 특히 국가통계 작성 과정에는 당연히 적용되는 원리라 할 수 있다.

일반적으로 행정조사는 행정기관이 행정결정을 위하여 필요한 정보를 수집하는 일체의 행정활동으로서 보통의 경우에는 실력행사를 수반하지 않으며, 그 실효성 확보는 통상 행정벌의 형태에 의해 담보되는 행정작용으로 이해된다.[89] 현행 「행정조사기본법」은 "행정기관이 정책을 결정하거나 직무를 수행하는 데 필요한 정보나 자료를 수집하기 위하여 현장조사·문서열람·시료채취 등을 하거나 조사대상자에게 보고요구·자료제출요구 및 출석·진술요구를 행하는 활동"으로 정의하는 한편(법 제2조 제1호), 행정조사의 기본원칙에 있어 "법령 등의 위반에 대한 처벌보다는 법령등을 준수하도록 유도하는 데 중점을 두어" 행정조사를 운용하도록 규정하고 있다(법 제4조 제4항).

이론적인 측면에서 보자면, 권력적 행정조사는 국민의 신체나 재산에 침해를 가져오는 것이므로 헌법 제37조 제2항에 근거하여 법률의 근거를 반드시 필요로 하지만 비권력적 행정조사는 피조사자의 자발적인 협조에 의해 이루어지는 것이므로 법적 근거를 필수적으로 요구하지는 않는다고 할 수 있다. 따라서 원칙적으로 행정작용의 근거규정은 조직법상의 권한 범위 내에서는 비권력적 행정조사의 권능까지 포함하고 있는 것으로 보고 있다.[90] 그렇지만 특정의 조사대상을 대상으로 하지 아니하는 행정조사라도 '사인의 기본권에 대한 침해를 수반하는' 행정조사는 헌법 제37조 제2항에 근거하여 반드시 법률의 근거가 있어야 한다.[91] 현행 「행정조사기본법」 역시 행정기관은 법령 등에서

87) 법률유보가 필요한 행정작용 영역에 관한 견해로는 전부유보설 / 침해유보설 / 급부행정유보설 / 중요(본질)사항유보설 등이 대립하고 있으나, 어느 설에 따르건, 침해행정은 반드시 법률의 유보가 있어야 한다. 법률유보의 원리와 관련한 논의는 정하중, 앞의 책, 29~32면 등을 참조.

88) 헌재 1999. 1. 28. 97헌가8, 판례집 11-1, 1, 7; 헌재 2000. 1. 27. 98헌가9, 판례집 12-1, 1, 8 등.

89) 최승필, 앞의 논문, 397~398면.

90) 홍정선, 앞의 책, 675면.

91) 홍정선, 위의 책, 675~676면.

행정조사를 규정하고 있는 경우에 한하여 행정조사를 실시할 수 있도록 규정하면서 다만 조사대상자의 자발적인 협조를 얻어 실시하는 행정조사의 경우에는 예외를 두고 있다(법 제5조). 결국 행정조사의 성격을 띠는 국가통계작성을 위한 자료조사 역시 같은 법 원리의 지배하에 놓이게 되는 바, 개인정보자기결정권을 비롯한 국민의 기본권과 관련되는 정보가 통계작성 대상일 경우에는 반드시 법률의 근거가 필요함을 유념할 필요가 있다.[92]

이러한 측면은 특히 이미 살펴본 바 있는 「통계법」과 독자적으로 운용되는 상당수의 통계의 법적 근거를 면밀히 검토해야 할 필요성으로 귀결된다 할 것이다. 다만 헌법재판소에 의하면, 기본권 제한에 관한 법률유보원칙은 '법률에 의한 규율'을 요청하는 것이 아니라 '법률에 근거한 규율'을 요청하는 것이므로, 기본권 제한에는 법률의 근거가 필요할 뿐이고 기본권 제한의 형식이 반드시 법률의 형식일 필요는 없으므로, 법규명령, 규칙, 조례 등 실질적 의미의 법률을 통해서도 기본권 제한이 가능하다고 할 수 있으므로,[93] 실질적으로는 통계작성의 법적 근거가 법률에 근거하는지 여부가 주요한 검토의 대상이 될 것이다. 아울러 국가영역에서 이루어지는 여러 통계들이 반드시 통계청장의 협의 하에 진행될 것까지는 요구되기 어렵겠으나, 적어도 「통계법」상의 원칙 및 절차들을 적극적으로 활용하는 형태로 진행될 필요가 있다 하겠다.

라. 통계자료의 수집의 실제

이러한 맥락에서 민간부문의 자료에 대해 진행되는, 「통계법」을 위시한 각종 법률에 근거한 통계작성을 위한 통계자료의 수집은 원칙적으로 행정조사의 맥락에서 운용이 가능하다 할 것이다. 따라서 「행정조사기본법」이 정하고 있는 행정조사에 관한 여러 규정들은 통계작성자들이 반드시 숙지해야 할 사항들이라 할 수 있다. 이 가운데 무엇보다 중요한 사항은 역시 동법 제4조가 선언하고 있는 행정조사의 기본원칙이라 할 수 있을 것이다.[94] 이 원칙들 가운데 특히 통계작성과

92) 현행 「개인정보보호법」이 통계에 관한 예외조항을 두고 있음(동법 제58조 제1항 제1호 등)은 통계와 개인정보자기결정권간의 연관성의 방증(傍證)이라 할 수 있을 것이다.

93) 헌재 2013. 7. 25. 2012헌마167, 관례집 25-2상, 296, 301 등.

94) 「행정조사기본법」 제4조(행정조사의 기본원칙) ① 행정조사는 조사목적을 달성하는데 필요한 최소한의 범위 안에서 실시하여야 하며, 다른 목적 등을 위하여 조사권을 남용하여서는 아니 된다.
② 행정기관은 조사목적에 적합하도록 조사대상자를 선정하여 행정조사를 실시하여야 한다.
③ 행정기관은 유사하거나 동일한 사안에 대하여는 공동조사 등을 실시함으로써 행정조사가 중복되지 아니하도록 하여야 한다.
④ 행정조사는 법령등의 위반에 대한 처벌보다는 법령등을 준수하도록 유도하는 데 중점을 두어야 한다.
⑤ 다른 법률에 따르지 아니하고는 행정조사의 대상자 또는 행정조사의 내용을 공표하거나 직무상 알게 된 비밀을 누설하여서는 아니된다.
⑥ 행정기관은 행정조사를 통하여 알게 된 정보를 다른 법률에 따라 내부에서 이용하거나 다른

직접적인 관련이 있는 것들로는 행정조사의 목적에 필요한 최소한의 범위 안에서의 실시(동조 제1항), 행정조사의 목적에 적합하도록 조사대상자의 선정(동조 제2항), 행정조사의 중복 방지(동조 제3항), 행정조사의 대상자 또는 내용의 공표 금지 및 비밀누설 금지(동조 제5항) 및 행정조사를 통해 얻은 정보의 조사 목적외 이용 및 제3자 제공 금지(동조 제6항)를 들 수 있을 것이다. 아울러 법이 마련하고 있는 출석 · 진술요구(법 제9조)[95], 보고·자료제출요구(법 제10조)[96], 현장조사(법 제11조)[97], 시료채취(법 제12조)[98], 자료 등의 영치(법 제13조)[99],

기관에 제공하는 경우를 제외하고는 원래의 조사목적 이외의 용도로 이용하거나 타인에게 제공하여서는 아니 된다.

95) 「행정조사기본법」 제9조(출석 · 진술 요구) ① 행정기관의 장이 조사대상자의 출석·진술을 요구하는 때에는 다음 각 호의 사항이 기재된 출석요구서를 발송하여야 한다.
1. 일시와 장소
2. 출석요구의 취지
3. 출석하여 진술하여야 하는 내용
4. 제출자료
5. 출석거부에 대한 제재(근거 법령 및 조항 포함)
6. 그 밖에 당해 행정조사와 관련하여 필요한 사항
② 조사대상자는 지정된 출석일시에 출석하는 경우 업무 또는 생활에 지장이 있는 때에는 행정기관의 장에게 출석일시를 변경하여 줄 것을 신청할 수 있으며, 변경신청을 받은 행정기관의 장은 행정조사의 목적을 달성할 수 있는 범위 안에서 출석일시를 변경할 수 있다.
③ 출석한 조사대상자가 제1항에 따른 출석요구서에 기재된 내용을 이행하지 아니하여 행정조사의 목적을 달성할 수 없는 경우를 제외하고는 조사원은 조사대상자의 1회 출석으로 당해 조사를 종결하여야 한다.

96) 「행정조사기본법」 제10조(보고요구와 자료제출의 요구) ① 행정기관의 장은 조사대상자에게 조사사항에 대하여 보고를 요구하는 때에는 다음 각 호의 사항이 포함된 보고요구서를 발송하여야 한다.
1. 일시와 장소
2. 조사의 목적과 범위
3. 보고하여야 하는 내용
4. 보고거부에 대한 제재(근거법령 및 조항 포함)
5. 그 밖에 당해 행정조사와 관련하여 필요한 사항
② 행정기관의 장은 조사대상자에게 장부·서류나 그 밖의 자료를 제출하도록 요구하는 때에는 다음 각 호의 사항이 기재된 자료제출요구서를 발송하여야 한다.
1. 제출기간
2. 제출요청사유
3. 제출서류
4. 제출서류의 반환 여부
5. 제출거부에 대한 제재(근거 법령 및 조항 포함)
6. 그 밖에 당해 행정조사와 관련하여 필요한 사항

97) 「행정조사기본법」 제11조(현장조사) ① 조사원이 가택·사무실 또는 사업장 등에 출입하여 현장조사를 실시하는 경우에는 행정기관의 장은 다음 각 호의 사항이 기재된 현장출입조사서 또는 법령등에서 현장조사시 제시하도록 규정하고 있는 문서를 조사대상자에게 발송하여야 한다.
1. 조사목적
2. 조사기간과 장소
3. 조사원의 성명과 직위
4. 조사범위와 내용
5. 제출자료
6. 조사거부에 대한 제재(근거 법령 및 조항 포함)
7. 그 밖에 당해 행정조사와 관련하여 필요한 사항
② 제1항에 따른 현장조사는 해가 뜨기 전이나 해가 진 뒤에는 할 수 없다. 다만, 다음 각 호의 어느 하나에 해당하는 경우에는 그러하지 아니하다.

공동조사(법 제14조)[100], 중복조사의 제한(법 제15조)[101] 규정들은 통계작성에서도 유용하게 활용할 수 있을 것이다. 그렇지만「통계법」상 규정된 조항들 역시 통계작성 과정에서 중요한 의미를 가질 수 있을 것인 바, 특히 자료제출명령(법 제25조)[102] 및 실지조사(법 제26조)[103]를 통계작성을 위한 자료수집에

1. 조사대상자(대리인 및 관리책임이 있는 자를 포함한다)가 동의한 경우
2. 사무실 또는 사업장 등의 업무시간에 행정조사를 실시하는 경우
3. 해가 뜬 후부터 해가 지기 전까지 행정조사를 실시하는 경우에는 조사목적의 달성이 불가능하거나 증거인멸로 인하여 조사대상자의 법령등의 위반 여부를 확인할 수 없는 경우

③ 제1항 및 제2항에 따라 현장조사를 하는 조사원은 그 권한을 나타내는 증표를 지니고 이를 조사대상자에게 내보여야 한다.

98)「행정조사기본법」제12조(시료채취) ① 조사원이 조사목적의 달성을 위하여 시료채취를 하는 경우에는 그 시료의 소유자 및 관리자의 정상적인 경제활동을 방해하지 아니하는 범위 안에서 최소한도로 하여야 한다.

② 행정기관의 장은 제1항에 따른 시료채취로 조사대상자에게 손실을 입힌 때에는 대통령령으로 정하는 절차와 방법에 따라 그 손실을 보상하여야 한다.

99)「행정조사기본법」제13조(자료등의 영치) ① 조사원이 현장조사 중에 자료·서류·물건 등(이하 이 조에서 "자료등"이라 한다)을 영치하는 때에는 조사대상자 또는 그 대리인을 입회시켜야 한다.

② 조사원이 제1항에 따라 자료등을 영치하는 경우에 조사대상자의 생활이나 영업이 사실상 불가능하게 될 우려가 있는 때에는 조사원은 자료등을 사진으로 촬영하거나 사본을 작성하는 등의 방법으로 영치에 갈음할 수 있다. 다만, 증거인멸의 우려가 있는 자료등을 영치하는 경우에는 그러하지 아니하다.

③ 조사원이 영치를 완료한 때에는 영치조서 2부를 작성하여 입회인과 함께 서명날인하고 그중 1부를 입회인에게 교부하여야 한다.

④ 행정기관의 장은 영치한 자료등이 다음 각 호의 어느 하나에 해당하는 경우에는 이를 즉시 반환하여야 한다.

1. 영치한 자료등을 검토한 결과 당해 행정조사와 관련이 없다고 인정되는 경우
2. 당해 행정조사의 목적의 달성 등으로 자료등에 대한 영치의 필요성이 없게 된 경우

100)「행정조사기본법」제14조(공동조사) ① 행정기관의 장은 다음 각 호의 어느 하나에 해당하는 행정조사를 하는 경우에는 공동조사를 하여야 한다.

1. 당해 행정기관 내의 2 이상의 부서가 동일하거나 유사한 업무분야에 대하여 동일한 조사대상자에게 행정조사를 실시하는 경우
2. 서로 다른 행정기관이 대통령령으로 정하는 분야에 대하여 동일한 조사대상자에게 행정조사를 실시하는 경우

② 제1항 각 호에 따른 사항에 대하여 행정조사의 사전통지를 받은 조사대상자는 관계 행정기관의 장에게 공동조사를 실시하여 줄 것을 신청할 수 있다. 이 경우 조사대상자는 신청인의 성명·조사일시·신청이유 등이 기재된 공동조사신청서를 관계 행정기관의 장에게 제출하여야 한다.

③ 제2항에 따라 공동조사를 요청받은 행정기관의 장은 이에 응하여야 한다.

④ 국무조정실장은 행정기관의 장이 제6조에 따라 제출한 행정조사운영계획의 내용을 검토한 후 관계 부처의 장에게 공동조사의 실시를 요청할 수 있다.

⑤ 그 밖에 공동조사에 관하여 필요한 사항은 대통령령으로 정한다.

101)「행정조사기본법」제15조(중복조사의 제한) ① 제7조에 따라 정기조사 또는 수시조사를 실시한 행정기관의 장은 동일한 사안에 대하여 동일한 조사대상자를 재조사 하여서는 아니 된다. 다만, 당해 행정기관이 이미 조사를 받은 조사대상자에 대하여 위법행위가 의심되는 새로운 증거를 확보한 경우에는 그러하지 아니하다.

② 행정조사를 실시할 행정기관의 장은 행정조사를 실시하기 전에 다른 행정기관에서 동일한 조사대상자에게 동일하거나 유사한 사안에 대하여 행정조사를 실시하였는지 여부를 확인할 수 있다.

③ 행정조사를 실시할 행정기관의 장이 제2항에 따른 사실을 확인하기 위하여 행정조사의 결과에 대한 자료를 요청하는 경우 요청받은 행정기관의 장은 특별한 사유가 없는 한 관련 자료를 제공하여야 한다.

102)「통계법」제25조(자료제출명령) ① 중앙행정기관의 장 또는 지방자치단체의 장은 지정통계의 작성을 위하여 필요하다고 인정되는 경우에는 개인이나 법인 또는 단체 등에 관계 자료의 제출을 명할 수 있다.

적극적으로 활용할 수 있을 것이다.

한편 대국민관계가 아닌 정부기관 내에서의 통계자료 혹은 행정자료의 수집과정은 본질적으로 행정조사와는 다른 양상을 띠게 된다. 그렇지만 특히 빅데이터 환경 하에서의 새로운 정보기술을 활용한 통계작성에 있어서는 통계작성을 위해 수집한 전통적인 '통계자료' 뿐만 아니라 통계와 직접적인 관련성을 가지지 않은 채로 기(旣)수집된 '행정자료'에 기반을 둔 통계작성 역시 중요한 의미를 가질 수 있을 것이기 때문에 통계작성을 위한 행정자료의 수집 역시 검토가 필요할 것이다. 이에 대하여 현행「통계법」은 중앙행정기관의 장 또는 지방자치단체의 장에게 행정자료의 요청 권한을 규정하고(법 제24조 제1항),[104] 행정자료의 제공을 요청받은 공공기관의 장에게는 정당한 사유가 없는 한 이에 응하도록 규정하고 있어(동조 제2항) 원칙적으로 행정자료의 원활한 수집을 위한 근거를 마련하고 있기는 하다.[105] 그렇지만, 현재의 규정은 행정자료의 요청권한을 중앙행정기관의 장 또는 지방자치단체의 장으로 한정하고 있을 뿐만 아니라, 행정자료 제공 요청을 받은 공공기관의 장이

② 통계청장은 통계작성지정기관이 요청하는 경우로서 지정통계의 작성을 위하여 필요하다고 인정되는 경우에는 제1항에 따른 명령을 할 수 있다.

③ 제1항 및 제2항에 따른 자료의 제출명령을 받은 자는 정당한 사유가 없는 한 이에 응하여야 한다.

④ 제1항 및 제2항에 따른 자료제출명령의 절차 및 방법 등에 관하여 필요한 사항은 대통령령으로 정한다.

103)「통계법」제26조(실지조사) ① 통계의 작성에 관한 사무에 종사하는 자는 통계의 작성을 위한 조사 또는 확인을 위하여 제18조에 따라 통계청장의 승인을 받은 사항에 관하여 관계인에게 관계 자료의 제출을 요구하거나 질문을 할 수 있다.

② 지정통계의 작성을 위한 조사 또는 확인에 있어 제1항에 따른 관계 자료의 제출을 요구받거나 질문을 받은 자는 정당한 사유가 없는 한 이에 응하여야 한다.

③ 제1항에 따른 직무를 행하는 자는 그 권한을 나타내는 증표를 지니고 이를 관계인에게 내보여야 한다.

104)「통계법 시행령」(대통령령 제27373호) 제38조(행정자료의 요청 및 제공 등) ① 중앙행정기관의 장 또는 지방자치단체의 장은 법 제24조제1항에 따라 공공기관의 장에게 행정자료 제공을 요청하려면 다음 각 호의 사항을 문서(전자문서를 포함한다)에 적어 요청하여야 한다.

1. 요청기관의 명칭과 주소
2. 행정자료의 명칭
3. 행정자료의 사용 목적
4. 행정자료의 내용(성별로 구분되는 행정자료의 경우에는 성별로 구분된 내용을 말한다)과 범위
5. 행정자료에 포함되어 있는 개인이나 법인 또는 단체 등의 정보를 보호하기 위한 조치
6. 행정자료의 제공방법

② 공공기관의 장은 제1항의 행정자료 제공 요청에 따라 행정자료를 제공하기로 결정한 경우에는 요청을 받은 날부터 30일 이내에 행정자료를 제공하여야 한다.

③ 공공기관의 장은 제2항에도 불구하고 요청 받은 행정자료의 가공 및 처리에 상당한 기간이 걸리는 경우 등 부득이한 사유로 30일 이내에 제공할 수 없으면 요청한 중앙행정기관의 장 또는 지방자치단체의 장과 협의하여 제공기간을 10일 이내의 범위에서 연장할 수 있다.

105) 아울러「통계법」은 통계작성을 위한 행정자료의 수집과 관련하여 '정보보호조치'(법 제24조 제3항) 및 통계작성 목적내사용 및 제3자 제공 금지(동조 제4항) 등의 보호장치를 마련하고 있다.

동조 제2항이 예정하고 있는 '대통령령으로 정하는 정당한 사유'[106]의 존재를 빌미로 제대로 응하지 않을 경우에 대한 조정과정이 마련되어 있지 않은 것은 문제점으로 지적될 수 있을 것이다.[107]

그 밖에도 통계작성을 위한 행정자료의 수집과 관련한 법률로는「전자정부법」(법률 제13459호)을 고려할 수 있을 것이다. 이 법은 행정업무의 전자적 처리를 위한 기본원칙, 절차 및 추진방법 등을 규정함으로써 전자정부를 효율적으로 구현하고, 행정의 생산성, 투명성 및 민주성을 높여 국민의 삶의 질을 향상시키는 것을 목적으로 하는 가운데(법 제1조), 행정기관 등 및 공무원에게 정보통신망의 연계 및 행정정보의 공동이용에 적극 협력할 의무를 부과하고(법 제3조), 전자정부의 원칙 가운데 행정정보의 공개 및 공동이용의 확대(법 제4조 제1항 제5호) 등을 규정하고 있으므로 통계작성을 위한 행정자료의 수집에도 이 법의 조항들이 원용될 여지는 충분하다 할 것이다. 특히 동법이 규정하고 있는 전자정부서비스의 제공과 이용촉진(법 제2절) 관련 조항들은 무엇보다도 빅데이터 환경 하에서의 새로운 통계작성의 적극적인 근거조항으로 기능할 수 있을 것으로 기대해 볼 수 있을 것이다.[108]

제7절 통계의 제 분야

현행 통계부문별 작성현황은 전장(前章)에서 이미 살펴본 바 있지만, 구체적인 측면에서 국가영역 내에서의 대표적인 통계의 사례를 몇 가지 꼽아보기로 하겠다.

106)「통계법 시행령」(대통령령 제27373호) 제39조(행정자료 제공의 예외사유) 법 제24조제2항에서 "대통령령으로 정하는 정당한 사유"란 다음 각 호의 어느 하나를 말한다.

1. 국가안전보장 · 국방 · 통일 · 외교관계 등에 중대한 영향을 미치는 국가기밀에 관한 행정자료로서 통계의 작성을 위하여 제공되면 국가의 중대한 이익을 현저히 침해할 우려가 있다고 인정할만한 상당한 이유가 있는 경우
2. 진행 중인 재판에 관련되거나 범죄의 예방, 수사, 공소의 제기 및 유지, 형의 집행, 교정, 보안처분에 관한 행정자료로서 통계의 작성을 위하여 제공되면 직무수행을 현저히 곤란하게 하거나 형사피고인의 공정한 재판을 받을 권리를 침해한다고 인정할 만한 상당한 이유가 있는 경우
3. 개인이나 기업의 신제품 개발, 신기술 연구 또는 상당한 노력에 의하여 비밀로 유지·관리되고 있는 생산방법이나 판매방법에 관한 행정자료로서 통계의 작성을 위하여 제공되면 개인이나 기업의 중대한 영업상의 비밀을 현저히 침해할 우려가 있다고 인정할 만한 상당한 이유가 있는 경우
4. 개인의 정치적, 종교적 또는 성적 성향이나 생활에 관한 행정자료로서 통계의 작성을 위하여 제공되면 개인의 생명이나 신체, 재산의 보호에 현저한 지장을 줄 우려가 있다고 인정할 만한 상당한 이유가 있는 경우

107) 이와 관련해서는 이미 언급한 바 있는 자료제출명령(「통계법」 제25조)처럼 통계작성지정기관의 요청에 의한 통계청장의 명령권한을 보다 적극적으로 부여하는 것을 검토해 볼 가치가 있을 것이다.

108) 특히 법 제16조(전자정부서비스 개발 · 제공), 법 제18조(유비쿼터스 기반의 전자정부서비스 도입 · 활용), 법 제21조(전자정부서비스의 민간 참여 및 활용) 조항들이 직접적인 관련성이 높은 조문들이라 할 수 있을 것이다.

우선「행정절차법」(법률 제12923호) 제46조의2(행정예고 통계 작성 및 공고) 행정청은 매년 자신이 행한 행정예고의 실시 현황과 그 결과에 관한 통계를 작성하고, 이를 관보·공보 또는 인터넷 등의 방법으로 널리 공고하여야 한다.[109] 사법부의 경우에도 법원의 권한에 속하는 사건 및 사무의 유형을 정확·신속히 파악하여 이를 숫자에 의하여 합리적으로 정리·표현함으로써 법원 운영의 자료를 제공하고 사건처리의 행정적 감독에 기여하고자 사법통계를 작성하고 있으며, 이를 위해「법원통계규칙」(대법원규칙 제2156호)을 마련하고 있다.

한편 대표적인 보고통계의 예로서「국가공무원법」(법률 제13618호)는 국회사무총장, 법원행정처장, 헌법재판소사무처장, 중앙선거관리위원회사무총장 또는 인사혁신처장은 국회·법원·헌법재판소·선거관리위원회 또는 행정 각 기관의 인사에 관한 통계보고 제도를 정하여 실시하고 정기 또는 수시로 필요한 보고를 받을 수 있게끔 규정하고 있다(동법 제18조). 이를 위해「공무원 인사기록·통계 및 인사사무 처리 규정」(대통령령 제26566호)이 마련되어 있다.[110]

경제분야와 관련된 통계로 산업통상자원부 장관은 외국인투자에 관한 통계자료의 수집·작성 권한을 가지며(「외국인투자 촉진법」(법률 제13426호) 제24조), 농림축산식품부장관은 김치산업의 진흥에 필요한 정책을 효율적으로 수립하고 김치 및 김치재료의 원활한 수급을 위하여 김치산업과 관련된 생산·유통·소비 등에 관하여 통계조사를 실시할 수 있다(「김치산업 진흥법」(법률 제13402호) 제9조).[111] 농림축산식품부장관은 인삼산업의 진흥에 필요한 정책을 효율적으로 수립하기 위하여 인삼산업과 관련된 생산·유통·소비 등에 관하여 통계조사를 실시할 수 있다(「인삼산업법」(법률 제13360호) 제20조의3). 그 밖에도 「통계법」 제17조에 따라 지정통계로 지정된 농작물생산조사 및 가축동향조사의 실시에 필요한 사항을 규정한「농업통계조사 규칙」(기획재정부령 제509호),「소재·부품통계조사규칙」(산업통상자원부령 제126호) 등이 있다.

사회분야에서는 실로 다양한 통계조사가 실시되고 있다. 우선 '건축기본조사'에

109)「행정절차법 시행규칙」(행정자치부령 제1호) 제13조(행정예고 통계의 공고) ① 행정청은 법 제46조의2에 따라 다음 각 호의 사항이 포함된 전년도 행정예고 통계를 다음 연도 3월말까지 공고하여야 한다.

1. 총 예고 건수
2. 고시, 훈령, 예규 등 예고 대상별 건수
3. 관보·공보, 인터넷, 신문·방송 등 예고 매체별 건수
4. 예고 기간별 건수

② 제1항에 따른 행정예고 통계의 공고는 별지 제22호서식을 참고하여 행정기관의 장이 정한 서식에 의한다.

110) 아울러 이와 유사한 제도가 지방공무원 및 교육감 소속 지방공무원에도 적용되어,「지방공무원 인사기록·통계 및 인사사무 처리 규칙」(행정자치부령 제1호) 및「교육감 소속 지방공무원 인사기록·통계 및 인사사무 처리 규칙」(교육부령 제96호)이 마련되어 있다.

111) 이 경우 통계의 작성에 관하여 「통계법」의 관련 규정을 준용하면서도, 필요한 자료 및 정보 제공요청권한 및 요청받은 자의 자료 및 정보제공의무는 별도로 규정하고 있다(동법 동조 제2항, 제3항). 아울러 세부적인 사항은「김치산업 진흥법 시행규칙」(농림축산식품부령 제139호) 제2조가 규율하고 있다.

건축에 관한 각종 통계가 포함되고(「건축기본법」(법률 제13470호) 제16조 제1항), 건축물의 착공통계조사(「건축물착공통계조사시행규칙」(국토교통부령 제1호))는 지정통계로 지정되어 있다. 교통행정기관 등의 교통사고조사와 관련된 자료·통계 또는 정보 보관·정리의무가 「교통안전법」(법률 제13426호) 제51조에 의해 부여되어 있으며, 재난 등과 관련한 '안전정보' 역시 국민안전처장관은 재난이나 그 밖의 각종 사고에 관한 통계, 지리정보, 안전정책 등에 관한 정보를 수집하여 체계적으로 관리해야 한다(「재난 및 안전관리 기본법」(법률 제13440호) 제66조의 7 제1항). 최근에는 성인지 통계(「양성평등기본법」(법률 제13369호) 제17조)[112] 도 작성되고 있다. 그 밖에 '정부'가 생산·유통 및 활용해야 할 인적자원개발관련 정보 가운데 "개발된 지표에 따른 관련 통계"가 포함되고(「인적자원개발 기본법」(법률 제12215호) 제11조 제1항), 역시 '정부'가 구축해야 할 온실가스 종합정보관리체계 속에도 온실가스 관련 통계가 포함되어 있다(「저탄소 녹색성장 기본법」(법률 제11965호) 제45조[113] 제1항). 마지막으로 의료분야에서도 결핵통계사업(「결핵예방법」(법률 제13323호) 제6조[114])과 암등록통계사업(「암관리법」(법률 제13654호) 제14조) 등이 수행되고 있다.

제8절 통계의 활용

일반적으로 논의되는 통계학의 목적이 첫째, 자료의 정리와 요약으로 통계학은

112) 「양성평등기본법」 제17조(성인지 통계) ① 국가와 지방자치단체는 인적(人的) 통계를 작성하는 경우 성별 상황과 특성을 알 수 있도록 성별로 구분한 통계(이하 이 조에서 "성인지 통계"라 한다)를 산출하고, 이를 관련 기관에 보급하여야 한다.
② 여성가족부장관은 통계청장 등 관계 기관의 장과 협의하여 성인지 통계의 개발, 산출, 자문 및 교육훈련 등 필요한 사항을 지원할 수 있다.

113) 「저탄소 녹색성장 기본법」 제45조(온실가스 종합정보관리체계의 구축) ① 정부는 국가 온실가스 배출량·흡수량, 배출·흡수 계수(係數), 온실가스 관련 각종 정보 및 통계를 개발·검증·관리하는 온실가스 종합정보관리체계를 구축하여야 한다.
② 관계 중앙행정기관의 장은 제1항에 따른 종합정보관리체계가 원활히 운영될 수 있도록 에너지·산업공정·농업·폐기물·산림 등 부문별 소관 분야의 정보 및 통계를 작성하여 제공하는 등 적극 협력하여야 한다.
③ 정부는 제1항에 따른 각종 정보 및 통계를 작성·관리하거나 종합정보관리체계를 구축함에 있어 국제기준을 최대한 반영하여 전문성·투명성 및 신뢰성을 제고하여야 한다.
④ 정부는 제1항에 따른 각종 정보 및 통계를 분석·검증하여 그 결과를 매년 공표하여야 한다.
⑤ 제1항부터 제4항까지에서 규정한 사항 외에 세부적인 정보 및 통계 관리방법, 관리기관 및 방법 등은 대통령령으로 정한다.

114) 「결핵법」 제6조(결핵통계사업) ① 보건복지부장관은 결핵의 발생과 관리실태에 대한 자료를 지속적이고 체계적으로 수집·분석하여 통계를 산출하는 사업(이하 "결핵통계사업"이라 한다)을 실시하여야 한다. 이 경우 통계자료의 수집 및 통계의 작성 등에 관하여는 「통계법」을 준용한다.
② 보건복지부장관은 결핵환자등과 잠복결핵감염자를 진단·치료하는 의료인 또는 의료기관, 「국민건강보험법」에 따른 국민건강보험공단과 건강보험심사평가원 및 그 밖에 결핵에 관한 사업을 하는 법인·기관·단체에 보건복지부령으로 정하는 바에 따라 결핵통계사업에 필요한 자료 제출이나 의견 진술 등을 요구할 수 있다. 이 경우 자료 제출을 요구받은 자는 특별한 사유가 없으면 이에 따라야 한다.

사회현상이나 자연현상에 관해 수집한 자료를 단순하고 이해하기 쉽도록 표나 그래프 또는 간단한 수치의 형태로 정리하고 요약하는 데 활용되며, 둘째, 현상의 기술(description)로 사회현상이나 자연현상에 관한 특성이나 발생빈도, 또는 자료들 간의 관계(relationship) 등을 파악하는 것을 말하며, 셋째, 현상의 설명(explanation)으로, 단순히 두 자료들 간의 관계를 기술하는 차원을 넘어 인과관계(causality), 즉 어떤 현상이 발생한 원인이 무엇인지를 알아보는 것이며, 넷째, 미래 상황의 예측(prediction)으로 요약되듯이[115] 통계는 다양한 영역에서 이러한 목적들을 위해 활용될 수 있음은 물론이다.

구체적으로 살펴보자면 국가영역에 있어서 우선 인구관련 통계들이 「공직선거법」(법률 제14073호)상 선거사무관리의 기준이 되는 인구는 「주민등록법」에 따른 주민등록표에 따라 조사한 국민의 최근 인구통계에 의하고,[116] 「통계법」 제3조의 규정에 의하여 통계청장이 매년 고시하는 **전국소비자물가변동률**이 선거비용제한액을 산정하는 기준으로 활용된다(동법 제121조). 「정치자금법」(법률 제14074호) 역시 정당에 대한 보조금의 계상에 있어 **선거권자 총수 및 보조금 계상단가**가 활용되는데, 보조금 계상단가는 전년도 보조금 계상단가 역시 「통계법」 제3조에 따라 통계청장이 매년 고시하는 전전년도와 대비한 전년도 **전국소비자물가변동률**이 적용된다(법 제25조).

사회영역에서도 위에서 언급한 바 있는 **전국소비자물가변동률**이 「기초연금법」(법률 제13988호)상 기초연금액의 산정에 활용되고(법 제5조 제2항), 또 「장애인연금법」(법률 제12620호)상 기초급여액의 산정에도 활용된다(법 제6조 제1항). 그 외 「재건축초과이익 환수에 관한 법률」(법률 제12989호)상 정상주택가격상승분을 산정함에 있어서 활용하는 '평균주택가격상승률'은 「주택법」 제87조의 규정에 따라 국토교통부장관의 위탁을 받아 기금수탁자가 통계청 승인을 받아서 작성한 **주택가격 통계**를 이용하여 산정한다(법 제10조 제2항). 한편 「통계법」 제22조제1항의 규정에 따라 통계청장이 고시하는 **한국표준산업분류**가 조세특례관련 업종분류 기준으로 활용되기도 한다(「조세특례제한법」 제2조 제3항 등)

통계의 활용에 있어서 특기할 만한 사항으로는 2012년 도입된 '통계기반정책평가'를 들 수 있다. 즉 중앙행정기관의 장은 ① 국가안보에 관한 사항, ② 행정절차, 행정조직에 관한 사항, ③ 민사·상사·형사, 소송절차, 재판 및 형의 집행에 관한 사항,

115) 황인창·이대용·이청호, 앞의 책, 4면.

116) 「공직선거법」(법률 제14073호) 제4조(인구의 기준) 이 법에서 선거사무관리의 기준이 되는 인구는 「주민등록법」에 따른 주민등록표에 따라 조사한 국민의 최근 인구통계에 의한다. 이 경우 지방자치단체의 의회의원 및 장의 선거에서는 제15조제2항제3호에 따라 선거권이 있는 외국인의 수를 포함한다.

④ 그 밖에 통계기반정책평가가 적절하지 아니하다고 통계청장이 정하는 사항소관을 제외하고는, 법령의 제정 또는 개정을 통하여 새로운 정책과 제도를 도입하거나 종전의 정책과 제도의 중요 사항을 변경하는 경우에는 그 도입·변경되는 정책과 제도의 집행·평가에 적합한 통계의 구비 여부 등에 대한 평가(이하 "통계기반정책평가"라 한다)를 통계청장에게 요청하도록 의무화 하였고, 통계청장은 이러한 요청에 따라 정책 및 제도가 통계에 기반하고 있는지를 평가하고 그 정책 및 제도 관련 통계의 적합성 여부, 통계 개발·개선 계획 등을 포함한 의견을 해당 중앙행정기관의 장에게 통보하면, 중앙행정기관의 장이 해당 법령안을 국무회의에 상정할 때 이러한 통계청장의 평가의견을 함께 제출하게끔 규정하였다(「통계법」 제12조의 2).

2. 빅데이터 환경에서의 통계생산의 법적 문제 - 개인정보보호의 측면에서

이미 국가통계 기본원칙을 서술하면서 잠시 언급한 바 있지만, 현행 국가통계제도는 기본적으로 헌법상의 전통적 기본권 가운데 하나인 프라이버시(privacy)권[117]의 맥락에서 사생활의 비밀과 보호에 입각한 '비밀보호'에 중점을 두고 있지만, 오늘날 활발히 전개되고 있는 국가·사회의 정보화의 흐름 속에서는 전통적인 비밀의 영역보다 더 광범위한 개인정보의 보호가, 특히 새로운 헌법상 기본권으로 인정되고 있는 개인정보자기결정권의 맥락에서 진행되고 있기 때문에 빅데이터 환경 하에서의 통계생산에 있어서는 이러한 개인정보 보호의 측면을 특히 주의 깊게 살펴볼 필요성이 크다 할 것이다.

주지하다시피 개인정보자기결정권은 자신에 관한 정보가 언제 누구에게 어느 범위까지 알려지고 또 이용되도록 할 것인지를 그 정보주체가 스스로 결정할 수 있는 헌법에 의해 보장되는 권리로서, 일반적으로는 헌법 제10조 제1문에서 도출되는 일반적 인격권 및 헌법 제17조의 사생활의 비밀과 자유에 근거하여 보장되는 권리로 인정되고 있다. 헌법재판소에 의하면,

"개인정보의 공개와 이용에 관하여 정보주체 스스로가 결정할 권리인 개인정보자기결정권의 보호대상이 되는 개인정보는 개인의 신체, 신념, 사회적 지위, 신분 등과 같이 개인의 인격주체성을 특징짓는 사항으로서 그 개인의 동일성을 식별할 수 있게 하는 일체의 정보라고 할 수 있다. 또한 그러한 개인정보를 대상으로 한 조사·수집·보관·처리·이용 등의 행위는 모두 원칙적으로 개인정보자기결정권에 대한 제한에 해당한다.[118]"

117) 주지하다시피 프라이버시권에 관한 헌법적 보호는 우리나라의 경우에는 1948년 제헌헌법에서 주거의 자유 및 통신의 비밀의 침해 금지 정도의 보장이 이루어지다가 1980년 개정 헌법에서 사생활의 비밀과 자유의 불가침이 최초로 명문화되었지만, 헌법에 명시되기 이전에도 사인간의 법률관계에서 인격권의 법리에 의해 규율되었으며, 불법행위이론에 의하여 손해배상 등이 인정되어 왔다. 한편 미국의 경우 1890년에 워렌과 브랜다이즈가 프라이버시에 관한 논문을 발표한 이래로 헌법적인 관심을 받기 시작하였고, 독일에서는 일반적인 인격권의 일환으로 학설상 인정되어 오다가, 1974년 미국의 프라이버시법(Privacy Act), 1979년의 독일의 연방정보보호법(Datenshutzgesetz) 등과 같은 특별법의 제정에 의하여 보호받기 시작한 바 있다. 사생활의 비밀과 자유에 관한 개괄적인 내용은 특히 이성환, "제17조", (사)한국헌법학회 편, 『헌법주석(I): 전문, 제1조~제39조』(서울: 박영사, 2013), 568~573면을 참조.

아울러 헌법재판소는 개인정보자기결정권과 사생활의 비밀과 자유가 경합하는 경우에는 특별한 사정이 없는 이상 개인정보자기결정권에 대한 침해 여부를 판단함으로써 사생활의 비밀과 자유의 침해 여부에 대한 판단이 함께 이루어지는 것으로 볼 수 있어 그 침해 여부를 별도로 다룰 필요는 없다고 판시한 바 있어,[119] 양자가 경합하는 것으로 볼 수 있는 경우에는 사생활의 비밀과 자유보다 우선적으로 개인정보자기결정권에 관한 논의를 전개하는 것이 특히 소송경제상 적합하다 할 것이다.

제1절 현행 개인정보보호체계[120]상 통계

일반적으로 개인정보와 관련된 체제(system)의 중점은 개인정보의 보호 혹은 개인정보의 활용 사이에 놓이게 마련인데, 현행 「개인정보보호법」(법률 제14107호)은 "개인정보의 처리 및 보호에 관한 사항을 정함으로써 개인의 자유와 권리를 보호하고, 나아가 개인의 존엄과 가치를 구현함"을 목적으로 선언하고 있어(법 제1조), 개인정보의 활용보다는 보호를 강조하고 있다는 평가가 우세하다[121]. 즉 「개인정보보호법」은 원칙적으로 모든 분야의 개인정보의 처리 행위를 규율할 수 있도록 공공부문과 민간부문에 공통으로 적용되는 개인정보 보호 처리기준을 확립하고, 국제 수준에 부합하는 개인정보 처리원칙 등을 규정함으로써 개인정보의 무분별한 수집, 유출, 오용, 남용 등으로부터 국민의 권리와 이익을 보호하고

118) 헌재 2005. 7. 21. 2003헌마282, 판례집 17-2, 81, 90.

119) 헌재 2005. 5. 26. 99헌마513, 판례집 17-1, 668, 683-684.

120) 현재의 개인정보보호 체계와 관련하여, 개인정보 보호에 관한 일반법인 「개인정보 보호법」이 2011년 제정되었음에도 불구하고 개별적인 개인정보 보호 법제가 시행되어 실제 운용상의 난맥상이 존재하는 실정이다. 즉 개인정보 보호의 일반법으로서의 「개인정보 보호법」 외에도 개별적인 개인정보 보호 법제로 「정보통신망 이용촉진 및 정보보호 등에 관한 법률(약칭: 정보통신망법)」(법률 제14080호), 「신용정보의 이용 및 보호에 관한 법률(약칭: 신용정보법)」(법률 제14122호), 「위치정보의 보호 및 이용 등에 관한 법률(약칭: 위치정보법)」(법률 제13540호) 등이 존재하며 그 외에도 다수의 법령에 개인정보 관련 조항들이 마련되어 있다.

121) 2011년 3월 29일 제정 당시의 「개인정보 보호법」(법률 제10465호)이 "개인정보의 수집·유출·오용·남용으로부터 사생활의 비밀 등을 보호함으로써 국민의 권리와 이익을 증진하고, 나아가 개인의 존엄과 가치를 구현하기 위하여 개인정보 처리에 관한 사항을 규정함"을 목적으로 규정함에 따라(법 제1조), 우리나라는 "(개인의 권리 및 이익 보호, 경제적·공익적 이용과 활용 혹은 양자의 조화와 균형 사이에서 선택가능한 개인정보 보호법의 목적 가운데에서) 개인의 권리와 이익 보호를 선택한 것"이라는 평가[이창범, 『개인정보 보호법』 (파주: 법문사, 2012), 3~4면]는 2014년 3월 24일 개정(법률 제12504호)을 통해 현재와 같이 개정된 후에도 크게 다를 바 없다고 할 수 있을 것이다.

피해를 구제하는 것이 법의 궁극적인 목적이라 할 수 있으므로, 개인정보의 활용에 따른 경제·사회적 이익 측면보다는 개인정보에 관한 정보주체의 권익 보호를 더 중시하고 있다고 할 수 있으며, 이것은 무엇보다 「개인정보 **보호법**」이라는 법명이 잘 설명해주고 있다는 것이다.[122] 그렇지만 개인정보 보호라는 과제의 본질은 일종의 "리스크 관리(risk management)"로서 이해하는 것이 바람직할 것이기에,[123] 개인정보의 적정한 보호가 이루어지는 범위 내에서의 최선의 활용이 가장 바람직한 방향이라 할 것이다.

현행 「개인정보 보호법」상 개인정보는 "살아 있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보(해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것을 포함)"를 의미하는바(법 제2조 제1호), 이에 따르면 개인정보의 주요 지표로는 생존중인 자연인 관련성(국적불문), 정보의 무정형성, 식별가능성을 들 수 있다. 이러한 개인정보의 종류로는 직접/간접 식별정보, 고유/일반식별정보, 민감/비민감정보, 수집/생성/생산정보, 공개/비공개 정보 등의 구분이 논의되고 있다.

「개인정보보호법」은 개인정보 보호의 원칙으로 ① 개인정보의 처리 목적을 명확화 및 목적에 필요한 범위내 최소한의 개인정보만을 적법하고 정당하게 수집 ② 개인정보의 처리 목적 외의 용도 활용금지 ③ 개인정보의 정확성, 완전성 및 최신성의 보장 ④ 개인정보의 처리 방법 및 종류 등에 따라 정보주체의 권리가 침해받을 가능성과 그 위험 정도를 고려하여 개인정보를 안전하게 관리 ⑤ 개인정보 처리방침 등 개인정보의 처리에 관한 사항의 공개 및 열람청구권 등 정보주체의 권리 보장 ⑥ 정보주체의 사생활 침해의 최소화 ⑦ 개인정보의 익명처리가 가능한 경우에는 익명처리 ⑧ 이 법 및 관계 법령상 책임과 의무를 준수, 실천함으로써 정보주체의 신뢰를 얻기 위하여 노력할 것을 선언하고(법 제3조), 아울러 정보주체의 권리로, ① 개인정보의 처리에 관한 정보를 제공받을 권리 ② 개인정보의 처리에 관한 동의 여부, 동의 범위 등을 선택, 결정할 권리 ③ 개인정보의 처리 여부를 확인하고 개인정보에 대하여 열람(사본의 발급을 포함한다. 이하 같다)을 요구할 권리 ④ 개인정보의 처리 정지, 정정·삭제 및 파기를 요구할 권리 ⑤ 개인정보의 처리로 인하여 발생한 피해를 신속하고 공정한 절차에 따라 구제받을 권리를 인정하고 있다(법 제4조). 이러한 「개인정보 보호법」은 개인정보보호의 일반법으로서

122) 성선제, 『개인정보보호법』 (서울: 서울경제경영, 2014), 4~5면.

123) 관련 논의는 이희정, "개인정보 보호법과 다른 법과의 관계 및 규제기관 사이의 관계 - 정보통신망 이용촉진 및 정보보호등에 관한 법률과의 관계를 중심으로", 고학수 편, 『개인정보 보호의 법과 정책』 (서울: 박영사, 2014), 132~134면을 참조.

개인정보 보호에 관하여 다른 법률에 특별한 규정이 있는 경우를 제외하고는 이 법에서 정하는 바에 따르도록 규정하고 있다(법 제6조).

이러한 개인정보보호법제와 통계의 관련에 있어서, 이미 언급한 바와 같이 통계는 전체 집단 또는 부분 집단에 관한 사실을 객관적으로 나타내는 것이지, 집단을 구성하는 특정개체의 개별적인 정보를 나타내는 것이 아니다 보니,[124] 논자에 따라서는 "특정 개인을 알아볼 수 없도록 가공되었거나 **통계적으로 변환된 경우**에는 특정 개인과의 관련성이 없고 식별이 어려우므로 개인정보에 해당하지 않는다"[125]는 견해도 없지는 않지만, 암등록통계사업(「암관리법」(법률 제13654호) 제14조)의 경우에서 보듯,[126] 통계작성의 제(諸)과정에서 활용되는 다양한 정보(자료)의 개인정보 해당성 자체를 부정하기는 어려울 것이며, 아울러 후술하듯이 새로운 정보처리기술의 발전에 기반을 둔 '재식별화(re-identification)'의 가능성을 염두에 둔다면 통계적으로 변환된 경우라 하더라도 사후적으로 '(재)식별가능성'이 생길 우려가 없지 않으므로, 통계자료로 처리되는 (개인)정보의 개인정보 해당성을 검토할 필요성은 적지 않다고 할 것이다.

그렇지만 현행 「개인정보보호법」은 **"공공기관이 처리하는 개인정보 중 「통계법」에 따라 수집되는 개인정보"**에 관하여는 개인정보보호의 구체적인 내용들을 다루는 제3장부터 제7장까지의 규정들을 적용하지 않도록 명시적으로 규정하고 있다(법 제58조 제1항 제1호). 이 조항을 이 조항을 공공기관이 처리하는 개인정보의 「통계법」에 따른 수집의 법적 근거(적극적 허용)로 보려는 입장도 없지 않지만, 기본적으로 통계와 관련해서는 개인정보 보호보다 통계작성을 우선시(소극적 허용)하는 입법자의 의사로 해석하는 것이 바람직할 것이다. 왜냐하면 동 조항에 의하더라도 여전히 「개인정보 보호법」상 개인정보 보호원칙(법 제3조) 및 정보주체의 권리(법 제4조) 등은 「통계법」에 따른 개인정보의 수집에 대해서도 적용이 배제되지 않고 있기 때문에, 「개인정보 보호법」은 여전히 개인정보의 보호의 관점에서 이해하는 것이 자연스러울 뿐 아니라, 나아가 통계를 작성할 때 개인정보 수집의 적극적인 근거는 당연히 「통계법」의 조항들을 원용하는 것이 바람직할 것이기 때문이다. 아울러 이 경우에도

124) 장치성 외 4인공저, 같은 곳.

125) 성선제, 앞의 책, 9~10면. 개인정보의 정의요소로 '개인식별성'을 강조하는 이창범, 앞의 책, 166면도 같은 입장이다.

126) 「암관리법」 제14조(암등록통계사업) ① 보건복지부장관은 암 발생 위험 요인과 암의 발생 및 치료에 관한 자료를 지속적이고 체계적으로 수집·분석하여 암 발생률, 생존율 등의 통계를 산출하기 위한 등록·관리·조사사업(이하 "암등록통계사업"이라 한다)을 시행하여야 하는데, 이 경우 통계자료의 수집 및 통계의 작성 등에 관하여는 「통계법」을 준용하며, **통계의 산출을 위하여 처리되는 개인정보**는 「개인정보 보호법」 제58조제1항에 따라 같은 법이 적용되지 아니하는 개인정보로 본다.

개인정보처리자는 그 목적을 위하여 필요한 범위에서 최소한의 기간에 최소한의 개인정보만을 처리하여야 하며, 개인정보의 안전한 관리를 위하여 필요한 기술적·관리적 및 물리적 보호조치, 개인정보의 처리에 관한 고충처리, 그 밖에 개인정보의 적절한 처리를 위하여 필요한 조치를 마련하여야 한다는 점은 (법 제58조 제4항) 개인정보보호를 목적으로 하는 법의 취지가 여전히 살아있는 부분이라 할 것이다.

그렇지만 이러한 예외적인 내용은 기본적으로 「통계법」에 따른 통계에만 적용되기 때문에, 민간통계는 물론 「통계법」의 준용 없이 독자적으로 작성되는 통계에 대해서는 문언상 동조항의 적용이 어렵다고 보아야 할 것이다. 그렇지만 이에 대해서도 「개인정보보호법」 제18조가 개인정보의 목적 외 이용·제공을 원칙적으로 제한하고 있으면서도, 이에 대한 약간의 예외를 두는 가운데 정보주체로부터 별도의 동의를 받은 경우(동조 제2항 제2호), 다른 법률에 특별한 규정이 있는 경우(동조 제2항 제3호) 등과 함께 "통계작성 및 학술연구 등의 목적을 위하여 필요한 경우로서 **특정 개인을 알아볼 수 없는 형태로 개인정보를 제공하는 경우**(동조 제2항 제4호)"를 규정하고 있으므로 민간통계 등에 대해서도 개인정보를 활용한 통계작성 자체가 상당한 정도까지 가능하리라는 점을 예상해 볼 수 있을 것이다. 물론 이러한 경우에 개인정보를 목적 외의 용도로 제3자에게 제공하는 개인정보처리자는 개인정보를 제공받는 자에게 이용 목적, 이용 방법, 그 밖에 필요한 사항에 대하여 제한을 하거나, 개인정보의 안전성 확보를 위하여 필요한 조치를 마련하도록 요청하여야 하고, 이 경우 요청을 받은 자는 개인정보의 안전성 확보를 위하여 필요한 조치를 하여야 한다(동조 제5항). 한편 이처럼 개인정보를 목적 외의 용도로 이용하거나 이를 제3자에게 제공하는 자가 공공기관일 경우에는 그 이용 또는 제공의 법적 근거, 목적 및 범위 등에 관하여 필요한 사항을 행정자치부령으로 정하는 바에 따라 관보 또는 인터넷 홈페이지 등에 게재하여야 한다(동조 제4항).

그 밖에도 국내 주요 개인정보법제 가운데 하나인 「위치정보의 보호 및 이용 등에 관한 법률」(법률 제13540호)은 「개인정보보호법」과 마찬가지로 **"통계작성, 학술연구 또는 시장조사를 위하여 특정 개인을 알아볼 수 없는 형태로 가공하여 제공하는 경우"**를 개인위치정보의 이용 또는 제3자 제공 허용사유로 규정하고 있다(법 제21조).[127] 아울러 이러한 통계에 대한 예외 조항은 「주민등록법 시행령」(대통령령 제27103호)[128]이나, 「국세기본법」(법률 제13552호)[129], 「초·중등교육법」(법률 제12943호)[130] 등에도 마련되어 있다.

그렇지만 또 다른 주요 개인정보법제 가운데 하나인 현행 「정보통신망 이용촉진 및 정보보호 등에 관한 법률(약칭: 정보통신망법)」(법률 제14080호)은 정보보호의 측면에 있어서는 통계와 관련한 별도의 규정을 두고 있지 않고 있는데,[131] 연혁적으로 볼 때, 과거 2007년 초반까지의 「정보통신망법」(2007.1.5. 시행, 법률 제8030호, 2006.10.4.일부개정)에는 현행 「개인정보보호법」의 규정과 유사한 조항이 있었으나,[132] 2007년 1월의 정보통신망법 개정(2007.7.27.시행, 법률 제8289호)을 통해[133] 그러한 조항이 삭제되게 된 바 있다.

127) 그 밖에도 「위치정보의 보호 및 이용 등에 관한 법률」은 위치정보사업자에게 일정사항(법 제29조제7항에 따른 경보발송 및 제30조제1항의 규정에 의한 **개인위치정보의 제공)에 관한 통계자료**를 매 반기별로 방송통신위원회에 제출하도록 규정하고(**법 제32조),** 위반시 500만원 이하의 과태료를 부과할 수 있게 하고 있다(법 제43조 제3항 제2호).

128) 동시행령은 「주민등록법」이 규정하고 있는 본인이나 세대원 이외의 자가 주민등록표의 열람이나 등·초본의 교부신청을 할 수 있도록 예외를 규정하면서 적시한 "그 밖에 공익상 필요한 경우"(법 제29조제2항제7호)에 "의료·연구 또는 **통계 목적의 달성**을 위하여 필요한 경우로서 행정자치부장관이 인정하는 경우"를 포함하고 있다(동시행령 제47조 제4항 제2호).

129) 동법은 세무공무원의 '비밀유지' 의무를 규정하면서, "**통계청장이 국가통계작성 목적으로 과세정보를 요구하는 경우**"를 그 목적에 맞는 범위에서 납세자의 과세정보를 제공할 수 있는 예외가운데 하나로 규정하고 있고(법 제81조의13 제1항 제5호) 이에 따라 과세정보를 알게 된 사람에게는 이를 타인에게 제공 또는 누설하거나 그 목적 외의 용도로 사용해서는 안될 의무를 부여하고 있다(동조 제4항). 한편, 동법은 과세자료의 제출과 그 수집에 있어서 **국가기관, 지방자치단체, 금융회사 등 또는 전자계산·정보처리시설을 보유한 자는 과세에 관계되는 자료 또는 통계를 수집하거나 작성하였을 때에는 국세청장에게 통보하도록 규정하고 있기도 하다(법 제85조 제2항).**

130) 동법은 학교의 장에게 학생 관련 자료 제공 제한 의무를 부과하면서 예외적으로 "통계작성 및 학술연구 등의 목적을 위한 것으로서 자료의 당사자가 누구인지 알아볼 수 없는 형태로 제공하는 경우"를 허용사유 가운데 하나로 규정하고 있다(제30조의6 제1항 제3호) 이 경우 학교의 장은 예외적으로 자료를 제3자에게 제공하는 경우에는 그 자료를 받은 자에게 사용목적, 사용방법, 그 밖에 필요한 사항에 대하여 제한을 하거나 그 자료의 안전성 확보를 위하여 필요한 조치를 하도록 요청할 수 있고(동조 제2항) 이에 따라 자료를 받은 자는 자료를 받은 본래 목적 외의 용도로 자료를 이용하여서는 아니 된다(동조 제3항).

131) 다만 「정보통신망법」은 정보통신망의 고도화(정보통신망의 구축·개선 및 관리에 관한 사항을 제외한다)와 안전한 이용 촉진 및 방송통신과 관련한 국제협력·국외진출 지원을 효율적으로 추진하기 위하여 설립한 한국인터넷진흥원(이하 "인터넷진흥원"이라 한다)의 사업 가운데 "**정보통신망의 이용 및 보호와 관련한 통계의 조사·분석"**을 포함시키고 있고(정보통신망법 제52조 제3항 제2호), 정보통신망 침해사고 발생시, 주요정보통신서비스 제공자나 집적정보통신시설 사업자 등의 정보통신망 운영자는 대통령령으로 정하는 바에 따라 침해사고의 유형별 통계, 해당 정보통신망의 소통량 통계 및 접속경로별 이용 통계 등 침해사고 관련 정보를 미래창조과학부장관이나 한국인터넷진흥원에 제공하도록 규정하고 있다(동법 제48조의2 제2항). 미래창조과학부장관은 이에 따라 정보를 제공하여야 하는 사업자가 정당한 사유 없이 정보의 제공을 거부하거나 거짓 정보를 제공하면 상당한 기간을 정하여 그 사업자에게 시정명령을 내릴 수 있고(동조 제4항), 이렇게 정보를 제공받은 미래창조과학부장관이나 한국인터넷진흥원에게는 침해사고의 대응을 위하여 필요한 범위에서만 정보를 정당하게 사용하여야 할 의무가 부과되어 있다(동조 제5항).

132) 구 「정보통신망법」(2007.1.5. 시행, 법률 제8030호, 2006.10.4.일부개정) 제24조(개인정보의 이용 및 제공 등) ① 정보통신서비스제공자는 당해 이용자의 동의가 있거나 다음 각호의 1에 해당하는 경우를 제외하고는 개인정보를 제22조제2항의 규정에 의한 고지의 범위 또는 정보통신서비스이용약관에 명시한 범위를 넘어 이용하거나 제3자에게 제공하여서는 아니된다.
1. 정보통신서비스의 제공에 따른 요금정산을 위하여 필요한 경우
2. 통계작성·학술연구 또는 시장조사를 위하여 필요한 경우로서 특정 개인을 알아볼 수 없는 형태로 가공하여 제공하는 경우
3. 다른 법률에 특별한 규정이 있는 경우 (이하 동조 제2항 이하의 내용은 생략)

133) 동 개정은 정보통신망을 이용한 신규서비스의 보급 및 이용 확산 등 정보통신환경의 변화에 따라 새롭게 등장하는 개인정보침해 문제에 적극 대처하기 위하여 개인정보의 수집·이용·제공 등에 관한 절차를 강화하고, 정보통신망의 특성상 익명성 등에 따라 발생하는 역기능 현상에 대한 예방책으로 사회적 영향력이 큰 정보통신서비스제공자와 공공기관의 책임성을 확보·강화하기 위하여

그렇지만 동법 제24조가 개인정보의 이용 제한을 규정하면서 제22조 제2항 각 호의 목적에 따른 이용을 허용하고 있으며, 제22조 제2항의 개인정보의 수집·이용동의의 예외사항에 '이 법 또는 다른 법률에 특별한 규정이 있는 경우'를 규정하고 있기 때문에(동조 동항 제3호) 「정보통신망법」의 경우에는 여전히 「개인정보보호법」의 통계관련 조항이 적용될 것으로 생각해 볼 수 있을 것이다.

또 다른 주요한 개인정보보호법제라 할 수 있는 「신용정보의 이용 및 보호에 관한 법률(약칭: 신용정보법)」(법률 제13216호)도 통계에 관한 규정을 전혀 두고 있지 않은 점은 특기할 만하다. 그렇지만 「신용정보법」 역시 동법 제3조의2 제2항에 의해서[134] 통계작성과 관련한 부분은 「개인정보보호법」의 관련규정이 적용될 가능성이 크다 할 것이다.

제2절 새로운 변수: 빅데이터 환경

국가·사회의 발전과 함께 특히 정보기술의 발전에 맞물려, 통계제도 역시 나름의 발전을 거듭해오고 있다.[135] 그렇지만, 피할 수 없는 "데이터 전쟁"[136]의 시대로 꼽히는 현대 사회에서 특히 '빅데이터 환경'이 통계작성과 관련하여 주요한 논의의 대상으로 거론된다. 이미 앞 장에서도 일부 언급한 바 있으나, 일반적으로 생산자 중심에서 소비자 중심으로 패러다임의 전환, 기업 전산화에 따른 데이터의 축적, 분석에 기반한 의사결정의 과학화 등의

제한적인 본인확인제도를 도입하며, 권리를 침해받은 자의 삭제 요청이 있는 경우 그 피해확산을 방지하기 위하여 이용자의 접근을 정보통신서비스제공자가 임시적으로 차단할 수 있도록 하는 임시조치제도를 도입하기로 하고, 친북게시물과 같은 불법통신이 정보통신망에서 유통되었을 때 사회적 영향력이 크다는 점 등을 고려하여 불법통신과 관련된 이행명령 대상을 확대하고 불법통신물의 삭제 절차 등을 보완하기로 하며, 자료제출요구권 등의 행사요건을 명확히 하며 행사방법 및 절차 등 적법절차 규정을 신설하여 관련 공무원에 의한 불합리하고 과도한 업무개입을 차단함으로써 규제의 투명성과 예측가능성을 확보하고자 이루어진 것이다. 이러한 가운데 통계 관련 규정은 구체적으로 **'개인정보의 수집·이용·제공에 대한 고지 및 동의제도 개선·보완(법 제22조제1항, 법 제24조의2 신설)'의 차원**에서, 개인정보를 수집하는 경우에 고지하고 동의를 얻어야 할 사항을 수집·이용 목적, 수집하는 개인정보의 항목 등 3가지로 명확하게 구분하고, 정보통신서비스제공자가 이용자의 개인정보를 제3자에게 제공할 경우에도 그에 따른 별도의 고지를 하고 동의를 얻도록 함과 아울러 제공받은 자가 개인정보를 이용·제공할 수 있는 범위를 명확하게 규정하기 위해 개정이 이루어진 것이다.

134) 「신용정보법」 제3조의2(다른 법률과의 관계) ① 신용정보의 이용 및 보호에 관하여 다른 법률에 특별한 규정이 있는 경우를 제외하고는 이 법에서 정하는 바에 따른다. ② 개인정보의 보호에 관하여 이 법에 특별한 규정이 있는 경우를 제외하고는 「개인정보 보호법」에서 정하는 바에 따른다.

135) 예를 들어, 조사표 중심의 전통적인 통계조사가 행정목적의 자료 활용을 통한 통계작성으로의 변화양상에 관한 논의로는 전광희, "선진국 공식통계의 패러다임 변용에 관한 연구 - 인구센서스와 경제센서스를 중심으로", 『(충남대학교 사회과학연구소) 사회과학연구』 제22권 제3호(2011)을 참조.

136) 박형준, 『빅데이터 전쟁: 글로벌 빅데이터 경쟁에서 살아남는 법』 (서울: 세종서적, 2015), 15면. 박형준은 특히 기업분야에서의 이러한 상황의 구체적인 양상으로, 고객층 확보와 외형적 성장을 통해 데이터를 독점하고자 하는 '플랫폼 전쟁'과 선순환 유지와 지속적 성장을 위한 경쟁력을 갖추고자 하는 '데이터 분석 전쟁'의 두 분야를 강조하고 있다. 같은 책, 17면 이하.

조건에 힘입어 등장한 개념[137]인 빅데이터(Big Data)란 문자 그대로 '방대한 양의 데이터'[138]를 가리키는 말로 일반적으로 "향상된 통찰과 의사결정과 정보처리과정의 자동화를 가능하게 하는 비용-효율적이고, 혁신적인 형태의 정보처리를 요구하는 거대한 규모와 엄청난 속도, 그리고/또는 극도로 다양한 종류의 정보자산(Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.)"으로 정의[139]되곤 하지만 실제에 있어서는 보다 복잡한 양상을 띤 채 활용되고 있다. 예를 들어 "디지털로 저장한 다양한 정보와 기록을 분석해 그동안 풀지 못했던 많은 문제들을 해결하는 것을 말한다. 다만 분석하는 데이터양이 어마어마하다는 특징이 있다"[140]는 언급이나('처리 기술'에 중점을 둔 정의), "과거 아날로그 환경에서 생성되던 데이터에 비하면 그 규모가 방대하고, 생성주기도 짧고, 형태도 수치 데이터뿐 아니라 문자와 영상 데이터를 포함하는 대규모 데이터"[141]와 같은 언급('속성'에 중점을 둔 정의)들이 각기 빅데이터의 주요 측면들을 보여주고 있는 것이다. 이러한 다양한 논의들 속에서 빅데이터의 특징은 '3V'로 요약하는 것이 일반적이다. 즉 데이터의 양(volume), 데이터의 생성속도(velocity), 형태의 다양성(variety)을 의미하는데, 최근에는 여기에 가치(Value)나 복잡성(Complexity)을 덧붙이기도 한다.[142]

이러한 빅데이터 활용의 선두 주자는 기업으로, 특히 구글(google)을 위시한 검색 포털들과 다양한 전자상거래 기업들은 방대한 고객데이터를 분석해 다양한 마케팅 활동을 하고 있는 실정이다.[143] 그렇지만, 공공부문 역시 보안 및 위험관리시스템, 탈세 등 부정행위 방지, 정부의 투명성을 높이고 국민의 알 권리를 향상시키며 시간과 자원을 절감하기 위한 **공공데이터 공개정책(open data)**, 특정 이슈에 대한 시민의 의견을 분석해 대응책을 마련하는 **오피니언 마이닝(opinion mining)** 등 빅데이터를 활용하기 위한 다양한 노력을 기울이고 있다.[144] 우리 정부 역시 "정부 3.0"의 모토 하에 빅데이터를

137) 정용찬, 『빅데이터』 (서울: 커뮤니케이션북스, 2013), ix면.

138) 박형준, 앞의 책, 259면. 박형준은 이 데이터를 활용해 성과를 내는 것이 중요할 뿐, 빅데이터의 실체가 무엇인가는 전혀 중요하지 않다고까지 언급하고 있다.

139) http://www.gartner.com/it-glossary/big-data (2016.9.18.최종확인)

140) 매일경제 기획팀·서울대 빅데이터 센터, 『빅데이터 세상: 당신의 숨겨진 욕망까지 읽어드립니다』 (서울: 매경출판, 2014), 43면. 같은 곳에서 빅데이터의 단위를 '테라(1조)', '페타(1000조)'와 같은 점두사가 붙은 데이터를 분석한다고 적고 있다.

141) 정용찬, 앞의 책, 2면.

142) 정용찬, 위의 책, 4~5면.

143) 기업의 빅데이터 활용에 관한 현황에 대한 개괄적인 논의로는 곽관훈, "기업의 빅데이터(Big Data) 활용과 개인정보 보호의 조화", 『일감법학』 제27호(2014), 133~136면 참조.

144) 정용찬, 앞의 책, 12~19면. 그렇지만 현재로서는 빅데이터를 직접적으로 규율하는 법령은 존재하지 않으며, 다만 「전자정부법」, 「국가정보화 기본법」, 「통계법」, 「기상산업진흥법」, 「공간정보산업

활용한 과학행정의 구현을 추진하고 있다.[145]

이런 흐름 속에서 데이터의 수집, 저장·관리, 처리·분석, 정보 시각화, 배포·유통, 보안 및 노출 조절기법의 6가지로 구성되는[146] 빅데이터 기술은 이미 다양한 경로로 이루어진 거대한 자료의 축적(물론 필요하다면 새로운 자료의 거대한 축적 역시 가능함)을 바탕으로, 일정한 의도에 따른 자료의 분석을 통해, 과거에는 큰 의미가 없다고 판단되던 사소한 데이터들의 더미 속에서 의미 있는 데이터의 추출을 가능케 함으로써 기존 통계영역을 넘어선 새로운 통계의 작성은 물론, 기존 통계자료에 대한 새로운 접근방식을 열어 과거와는 다른 통계작성의 가능성을 열어주었다. 이러한 통계작성 환경의 변화는 자료수집 비용의 관점에서 새로운 분석 시스템의 도입을 필요한 비용증가 요인 등으로도 기능하겠지만, 모집단 대표성이 약하기 쉬운 빅데이터 자체의 한계 및 전통적 통계생산 방식과는 현저히 다른 빅데이터 분석기법으로 인해 빅데이터를 통계생산 용도로 활용하기 위한 통계작성기법의 개발 등이 우선적으로 요구되는 실정이기에,[147] 본 보고서에서는 전장(前章)에서 이미 빅데이터를 활용한 통계의 품질평가 방안을 살펴본 바 있기도 하다.

법제적인 측면에서 이러한 빅데이터를 활용한 통계생산에 있어서는 우선적으로 빅데이터에 대한 명확한 개념규정의 필요성을 검토해 볼 필요가 있을 것이다. 이는 그 동안의 빅데이터에 대한 법제적인 측면에서의 대응양상을 살펴볼 때, 아직까지도 빅데이터에 대한 명확한 법률수준의 정의조항 조차 마련하지 못하고 있는 실정이기 때문이기도 하다. 다만 행정명령 수준의 일부 조직법제에서 빅데이터를 업무분장내역에 포함하는 행정부서들을 설치하고 있는데,[148]

진흥법」 등의 일부 규정을 통해 간접적으로 빅데이터와의 관련성을 검토하는 견해가 있다. 김지훈, "빅데이터와 개인정보보호", 『법제연구』 제46호(2014), 126~127면.

145) 손영화, "빅데이터 시대의 개인정보 보호방안", 『기업법연구』 제28권 제3호(통권 제58호)(2014), 356면.

146) 상세한 내용은 이지영, 『빅데이터의 국가통계 활용을 위한 기초연구』 통계개발원 2015 상반기 연구보고서(2015), 54~72면 참조.

147) 기존 논의에 있어서도 빅데이터를 활용한 통계생산과 관련하여, 빅데이터 자체는 기존 승인통계에 비교하여 대표성과 방법론에 이슈가 많고, 통계의 중요성에 있어서도 행정자료, 거래자료에 비해 상대적으로 떨어지는 단점을 보유하고 있음을 고려하지 않을 수 없다는 점이 지적되고 있다. 즉 보다 신뢰도 높은 연구의 수행과 정책수립을 위해서는 충분한 자료의 확보가 매우 중요한데 하나의 자료에서 분석에 필요한 충분한 정보를 얻는다는 것은 매우 어려운 일임에 주의해야 하는 것이다. 심규호·박시내, 앞의 보고서, 237면. 그렇지만 동 보고서는 이러한 문제(단일정보의 불충분성)는 데이터 매칭(data matching) 또는 데이터 통합(data fusion)을 통해 상당 부분 보완이 가능한 바, 일반적으로 조사된 데이터에는 가구 식별번호 및 개인 식별번호, 나이, 성별 등 공통적으로 포함된 항목이 있으며, 이러한 공통 항목을 통해 다수의 데이터를 통합하면 보다 많은 정보를 얻을 수 있다고 주장하는 바(같은 곳), 이러한 주장은 전형적으로 추후 살펴 볼 개인정보 활용론에 입각한 것으로 평가해 볼 수 있을 것이다.

148) 예를 들어, 「기상청과 그 소속기관 직제 시행규칙」은 2015년 1월 개정을 통해 기상기술융합팀장의 업무분장내역에 기상기후 빅데이터 활용에 대한 정책 및 기본계획의 수립·종합·조정 등을 규정한 바 있고(동규칙 제9조 제7항 제1호 등), 「국토교통부와 그 소속기관 직제 시행규칙」(제5조 등), 「통계청과 그 소속기관 직제」(제7조의2 등) 등도 빅데이터 관련 업무를 규정하고 있다.

이러한 가운데 「미래창조과학부와 그 소속기관 직제」의 경우는 국민소통실의 분장업무 가운데 "빅데이터 기반 온라인 여론분석 및 대응"을 포함시키면서 빅데이터를 **"초(超) 대용량의 정형 또는 비정형의 데이터세트"**로 규정하여(동령 제17조 제3항 제32호) 간략하게나마 빅데이터에 대한 정의조항을 마련하고 있다.[149] 한편 지방법규의 관점에서는 일부나마 빅데이터에 대한 보다 적극적인 정의조항을 마련하고 있기도 하다. 예를 들어 2016년 3월 제정된 「서울특별시 데이터의 제공 및 이용 활성화에 관한 조례」(서울특별시조례 제6162호)는 빅데이터를 "**아날로그 또는 디지털 환경에서** 생성되는 정형 또는 비정형의 수치, 문자, 영상 등의 대량 데이터의 집합과 이로부터 추출한 가치, 결과를 분석하는 **기술과 환경**을 말한다"고 규정하고 있으며(제2조 제4호), 같은 해 7월 제정된 「경기도 빅데이터 활용에 관한 조례」(경기도조례 제5276호)는 빅데이터를 "**디지털환경**에서 생성되는 정형 또는 비정형의 수치, 문자, 영상 등의 대량 데이터의 집합 및 이로부터 가치를 추출하고 결과를 분석하는 **기술**을 말한다"고 규정하고 있다. 생각건대 빅데이터 관련 사항에 대한 "명확한" 법적 규율이 이루어진다면 향후 빅데이터 관련 정책수립 및 집행에 상당한 편의를 제공해 줄 것으로 일응 기대할 수 있을 것이나, 위에서 살펴본 바와 같이 빅데이터에 대해서는 아직 기술적인 차원에서의 명확한 정의조차 곤란한 실정이라 할 수 있다. 따라서 명확한 개념을 얻기 어려운 현재의 시점에서 빅데이터에 대한 법적 규정은 특히 법적 안정성(Rechtssicherheit)의 측면에서 오히려 불필요한 혼란만을 초래할 우려가 크다 하지 않을 수 없다.

이보다는 오히려 빅데이터를 활용한 통계생산을 위한 공공부문 및 민간부문에 있어서의 자료수집을 둘러 싼 법적문제를 검토하는 것이 보다 실질적인 의미를 가질 수 있을 것으로 생각되는 바, 이에 관한 논의는 이미 앞에서 간략하게나마 언급한 바 있다. 아울러 법제적 관점에서는 무엇보다도 데이터의 수집·활용과 상충되기 쉬운 개인정보보호의 측면에서의 논점들이 중요한 위치를 차지하게 될 것이다. 이미 언급한 바와 같이 현행법상 개인의 "식별가능성"이 주요한 개인정보의 개념지표로 인정됨에 따라, 일견 단독으로는 개인의 식별가능성이 없는 정보라 할지라도 빅데이터 기술 하에서 다른 정보와의 결합 및 분석과정을 통해 특정인의 '(재)식별가능성'이 생기게 된다면 개인정보로 보게 될 가능성이 없지 않으므로,[150] 아무리 경제적, 기술적인 관점에서 빅데이터를 활용한 통계생산이 가능하다 하더라도 헌법상 보장되는

149) 그 밖에도 「행정자치부와 그 소속기관 직제 시행규칙」은 창조정부조직실의 분장업무를 규정하면서, **"행정정보 관련 대용량의 정형 또는 비정형의 데이터세트"**를 빅데이터로 규정하고 있다(동령 제9조 제7항 제21호).

150) 곽관훈, 앞의 논문, 140면; 손영화, 앞의 글, 378~379면.

개인정보자기결정권의 “침해”로 귀결될 경우에는 통계생산 자체가 법적으로 허용되기 어렵기 때문이다.

제3절 빅데이터 환경 하에서의 통계생산의 법적 정당화 가능성

가. 개관: 개인정보 활용론 vs. 개인정보 보호론

이미 개인정보보호법제의 목적과 관련해서 언급한 바와 같이 개인정보의 보호와 관련해서 전통적으로 개인정보 보호론과 개인정보 활용론이 맞서고 있는데, 비록 아직까지는 이에 관한 본격적인 논의가 이루어진 바는 없다고 할 수 있지만, 이러한 흐름은 빅데이터 환경 하에서의 통계와 개인정보 보호와 관련한 논의에 있어서도 그대로 유지될 수 있을 것이다. 즉, 빅데이터 환경 하에서 통계작성을 위해 개인정보를 적극적으로 활용하자는 주장과 개인정보 보호를 위해 통계작성에 대해서도 어느 정도의 제한을 가할 수밖에 없다는 주장이 맞서게 되는 것이다.

사실 빅데이터를 활용한 통계작성에 관하여 개인정보 활용론적인 측면에서의 논의는 이미 어느 정도 진행 중이라 할 수 있다. 물론 이러한 논의가 개인정보보호 분야에서 본격적으로 진행되고 있지는 않으며, 빅데이터를 실제 통계작성에 활용하고자 하는 시도들 속에서 부수적으로 진행되고 있는 실정이다. 즉 통계청에 의해 2013년 10월 마련된 제1차 국가통계발전('13~'17) 기본계획 중에는 “통계 생산방식 선진화를 통한 효율성 제고”의 측면에서 ‘빅데이터 활용 기반 마련’을 모색하여, 실제로 빅데이터를 활용한 통계 생산방안을 검토·추진 중에 있다.[151] 그렇지만 결국 이러한 빅데이터를 활용한 사업에서는 방대하고 다양한 종류의 정보의 수집과 이용을 전제로 할 수밖에 없고, 그와 같이 수집되는 정보에는 개인과 관련된 정보도 포함될 수밖에 없을 것이다.[152] 이에 이처럼 빅데이터를 통계생산에 활용하는데 있어서의 주된 제약사항으로 민간의 정보 제공 거부는 말할 것도 없이 행정자료의 경우조차 부처간 자료 공유의 부담이 존재한다는 점을 비롯하여 결정적으로는 개인정보 보호론에 입각하여 진행될 사회적 반발에 대한 우려가 꼽히고 있다.

이제 빅데이터를 활용한 통계작성을 개인정보 보호론의 입장에서 살펴보자면, 빅데이터를 분석·활용하는 기술 자체가 진정한 ‘빅브라더(Big Brother)’의

151) 통계청은 2016년 현재 우선적으로 도·소매 업체들의 가격 정보를 활용하여 소비자물가지수 보완 여부를 검토하고 있으며, 데이터 정제, 품질 점검 후 대외적으로 공개함으로써 민간의 활용을 지원할 계획이다.

152) 장주봉, “개인정보의 의미와 규제범위”, 고학수 편, 『개인정보 보호의 법과 정책』 (서울: 박영사, 2014), 87면.

출현을 가져올지도 모르는 위험한 측면을 보유하고 있다는 평가가 가능할 정도이다. 왜냐하면 빅데이터의 맥락에서 수집되어 온 특정인(들)에 관련된 위치정보를 위시한 다양한 유형의 데이터가 정교한 통계처리기술과 결합하게 될 경우에는 지금까지의 기술적 기반 하에서는 예측하기 어려운 양상의 개인정보(혹은 보다 넓은 맥락에서 사생활의 자유 등)에 대한 침해가 발생할 가능성을 배제하기 어렵기 때문이다.

다만 이미 살펴본 바와 같이 현행 개인정보 보호법제가 통계에 대한 명시적인 적용예외조항들을 두고 있기 때문에 현행 실정법체계상 빅데이터 환경 하에서의 통계 작성은 개인정보보호의 측면에서 그다지 큰 문제를 발생시키지 않을 것으로 예상해 볼 수도 있을 것이다. 그렇지만 추후 빅데이터 기술 및 환경과 관련한 본격적인 논의가 전개되게 된다면, 실질적 법치주의의 맥락에서 현행 법제의 정당성에 관한 의문도 본격적으로 제기될 가능성이 없지 않다.[153] 결국 이러한 맥락에서 빅데이터 환경하에서의 통계생산의 헌법적 적합성을 검토하는 작업의 중요성은 현행 법제하에서도 충분한 의미를 띠게 될 것이다.

나. 빅데이터 환경하에서의 통계작성의 합헌성 심사

이미 언급한 바와 같이 개인정보자기결정권은 헌법상 보장된 기본권이긴 하지만, 헌법상 보장된 기본권 역시 일정한 요건 하에서는 그에 대한 제한이 적법하게 이루어질 수 있다. 즉, 현행 헌법 제37조 제2항이 “국민의 모든 자유와 권리는 국가안전보장·질서유지 또는 공공복리를 위하여 필요한 경우에 한하여 법률로써 제한할 수 있으며, 제한하는 경우에도 자유와 권리의 본질적인 내용을 침해할 수 없다”고 규정한 바에 따라서 국민의 모든 자유와 권리는 국가안전보장·질서유지 또는 공공복리를 목적으로 필요한 경우에 한하여 법률의 형식에 의해서 제한할 수 있는 바, 이 조항을 바탕으로 일반적인 기본권 침해 여부의 심사기준인 법률유보의 원리 및 과잉금지(過剩禁止; Übermaßverbot)의 원칙이 성립한다. 이 가운데 법률유보의 문제에 대해서는 통계자료의 수집 특히 행정조사와 관련한 부분에서 이미 살펴본 바 있으므로, 여기에서는 과잉금지의 원칙을 중심으로 살펴보기로 한다.

‘비례(比例)의 원칙’이라고도 불리는 과잉금지의 원칙은 기본적으로 국가가

153) 실제로, 2016년 6월 발간된 최근의 논의에 따르면, 「통계법」에 따라 수집하고 있는 개인정보에 대해 정보주체의 권리보장을 적용하지 않고 있는 현행 「개인정보보호법」의 문제점을 지적하면서, 「통계법」에 따라 수집되는 개인정보에 대해서도 통계작성의 공익적 성격과 통계자료 수집 및 이용의 원활화를 고려하는 가운데 수집된 개인정보의 안전한 관리와 개인정보의 열람 및 정정청구권 등 정보주체의 권리를 보장할 수 있는 방안을 검토하고 있다. 자세한 내용은 배상호·신제수·전삼현·정현수, “통계처리를 위해 수집된 개인정보에 대한 개인정보보호 개선방안에 관한 연구”, 『중소기업융합학회 논문지』 제6권 제2호(2016), 28~29면을 참조.

국민의 기본권을 제한하는 내용의 입법 활동을 함에 있어서, 준수하여야 할 기본원칙 내지 입법활동의 한계를 의미하는 것으로서 국민의 기본권을 제한하려는 입법의 목적이 헌법 및 법률의 체제상 그 정당성이 인정되어야 하고(목적의 정당성), 그 목적의 달성을 위하여 그 방법이 효과적이고 적절하여야 하며(방법의 적절성), 입법권자가 선택한 기본권 제한의 조치가 입법목적달성을 위하여 설사 적절하다 할지라도 보다 완화된 형태나 방법을 모색함으로써 기본권의 제한은 필요한 최소한도에 그치도록 하여야 하며(피해의 최소성), 그 입법에 의하여 보호하려는 공익과 침해되는 사익을 비교 형량할 때 보호되는 공익이 더 커야 한다(법익의 균형성)는 헌법상의 원칙이다.[154] 헌법재판소는 이러한 과잉금지 원칙이 충족될 때 국가의 입법작용에 비로소 정당성이 인정되고 그에 따라 국민의 수인(受忍)의무가 생겨나는 것으로서, 이러한 요구는 오늘날 법치국가의 원리에서 당연히 추출되는 확고한 원칙으로서 부동의 위치를 점하고 있으며, 헌법 제37조 제2항에서도 이러한 취지의 규정을 두고 있는 것이라 설시한 바 있다.[155] 아울러 이러한 과잉금지 원칙은 입법작용 뿐만 아니라 기본권을 제한하는 국가작용 전반에 걸쳐 반드시 준수되어야 할 원칙으로 인정됨에 따라, 일반적으로 기본권 제한의 합헌성 심사기준으로 활용되고 있다.

구체적으로 헌법상 보장된 자유와 권리를 제한하는 법률의 목적의 합헌성을 심사함에 있어서는 명문으로 제시된 목적 외에 당해 법률규정이 사실상 의도하는 목적도 모두 고려하여야 한다. 그리하여 일견 타당하거나 정당한 것으로 보이는 목적도 그것만으로는 합헌으로 선언될 수 없고, 그것이 제37조 제2항에 설시된 목적들 중 적어도 어느 하나에 해당되는 것으로 판단되는 경우에 비로소 제37조 제2항에 합치하게 된다.[156] 한편 헌법재판소에 의하면 개인정보의 종류 및 성격, 수집목적, 이용형태, 정보처리방식 등에 따라 개인정보자기결정권의 제한이 인격권 또는 사생활의 자유에 미치는 영향이나 침해의 정도는 달라지므로 개인정보자기결정권의 제한이 정당한지 여부를 판단함에 있어서는 위와 같은 요소들과 추구하는 공익의 중요성을 헤아려야 한다.[157]

이러한 맥락에서 현행 개인정보 보호법제 내의 통계 관련 규정들은 개인정보자기결정권의 제한과 관련한 과잉금지원리에 입각한 합헌성 여부를 간략하게나마 검토해 보자면, 국가통계의 국가·사회적 중요성을 감안하면 '일반적인' 국가통계작성의 목적의 정당성 · 수단의 적절성 및 법익의 균형성은

154) 헌재 1990. 9. 3. 89헌가95, 판례집 2, 245, 260; 1994. 12. 29. 94헌마201, 판례집 6-2, 510, 524-525; 1998. 5. 28. 95헌바18, 판례집 10-1, 583, 595; 2000. 2. 24. 98헌바38등, 판례집 12-1, 188, 224-225; 2000. 6. 1. 99헌가11등, 판례집 12-1, 575, 583 등.
155) 헌재 1990. 9. 3. 89헌가95, 판례집 2, 245, 260면.
156) 김대환, "헌법 제37조", (사)한국헌법학회 편, 『헌법주석1』 (서울: 박영사, 2013), 1184면.
157) 헌재 2005. 7. 21. 2003헌마282등, 판례집 17-2, 81, 93-94.

어렵지 않게 인정될 수 있을 것으로 생각되며, 관련 정보의 목적 내 활용의무 및 안정성 확보 조치(「통계법」 제24조, 「개인정보보호법」 제18조 제5항 등), 비밀유지의무(「통계법」 제33조, 「개인정보보호법」 제18조 제5항 등) 등을 규정하고 있는 현재의 통계작성과정은 침해의 최소성의 관점에서도 어느 정도 이상의 정당성을 인정하는데 큰 어려움은 없을 것으로 판단된다. 다만 거듭 이야기하지만 빅데이터 환경 하에서 새로운 정보처리 기술이 발전함에 따라 현재의 판단이 언제까지 유지될 수 있을지는 알 수 없음은 유념할 필요가 있을 것이다.

결국, 현재로서는 이러한 통계작성과 관련한 법률들이 일반적인 합헌성이 인정된다 하더라도 그 법에 의해 작성되는 개별적·구체적 통계작성이 합헌성을 유지할 수 있도록 사전에 면밀히 신경 쓰는 것이 필요할 것이다. 이를 위해서 빅데이터를 활용한 통계작성에 있어서는 본격적인 통계작성에 앞서 과잉금지 원칙의 세부적인 심사 원리에 입각하여 그 헌법적 정당성을 확보하는 절차가 반드시 필요할 것이다. 즉, 목적의 정당성(; 기본권을 제한하는 국가작용의 목적은 정당한가?)의 측면에서는 통계작성의 목적 및 기대효과의 적시(摘示)가 필요할 것이고, 수단의 적절성(; 그 목적을 달성하기에 적절한 수단을 활용하고 있는가?)의 측면에서는 그 통계작성 방법의 유효성 및 적절성을 면밀히 검토할 필요성이 있을 것이며, 침해의 최소성(; 국가작용에 의해 제한되는 기본권의 피해를 최소화하고 있는가?)의 측면에서는 통계작성으로 인한 개인정보침해의 최소화 대책 - 예를 들면 비식별화 및 재식별화의 문제 -, 통계작성과정에서 입수한 비밀 및 개인정보 보호 준수와 더불어 통계작성 시스템의 보안 및 안전성을 확보 대책까지 점검할 필요가 있을 것이다. 끝으로 법익의 균형성(; 국가작용에 의해 달성되는 공익이 침해되는 기본권 보다 큰가?)의 관점에서는 통계작성으로 달성될 수 있을 공익과 통계작성으로 인해 침해가 우려되는 기본권간의 이익형량(balance of interests) 역시 필요한 범위 내에서 수행할 필요성이 있다고 할 것이다. 이러한 과정은 대체적으로 다음과 같은 체크리스트를 활용하여 점검 및 수행되는 것을 생각해 볼 수 있을 할 것이다.

대범주	소범주	내용	비교
목적의 정당성	통계작성의 규범적 근거	• 적극적 근거: 통계작성의 근거 법령 • 소극적 근거: 통계작성에 장애가 될 만한 법령의 검토	법률유보
	통계작성의 구체적 목적/필요성		
	통계작성 목적의 범주	국가안전보장( )/질서유지( )/공공복리( )	
	기대효과		
수단의 적절성	적용가능한 방법론		
	적용예정 통계작성 방법		
	적용예정방법의 효과성/효율성		
	평가 및 검증방법		
침해의 최소성	비식별화 수준 및 방법		개인정보 침해의 최소화 대책
	재식별기법		
	비밀 및 개인정보 보호의무 숙지여부		
	시스템 보안수준		
	시스템 안정성 수준		
법익의 균형성	예상되는 침해수준		
	기대효과와의 비교		

<표 12> 통계작성으로 인하여 초래될 수 있는 기본권 침해와 공익간의 균형성 파악을 위한 체크 리스트

구체적인 측면에서 볼 때, 현행 빅데이터 환경과 관련하여 개인정보 보호법제 내에서 가장 논란이 되고 있는 부분인 '비식별화' 문제는 통계작성에 있어서도 중요한 논점이라 할 수 있을 것이다. 즉 이미 언급한 바와 같이 방대하고 다양한 종류의 정보의 수집과 이용을 전제로 한 빅데이터 환경에 있어서 수집되는 정보에는 개인과 관련된 정보도 포함될 수밖에 없을 것이기에 빅데이터 기술을 활용하기 위한 기본적인 전제 가운데 하나로 논의되던 비식별화(非識別化; de-identification) 논의는 통계작성에서 특히 중요한 논점으로 부각될 가능성이 높은데, 이는 기본적으로 이미 살펴본 바와 같이 「개인정보보호법」 제18조가 개인정보의 목적 외 이용·제공 제한의 예외로 "통계작성 및 학술연구 등의 목적을 위하여 필요한 경우로서 특정 개인을 알아볼 수 없는 형태로 개인정보를 제공하는 경우(동조 제2항 제4호)"를 규정하고 있기 때문에, 여기에서의 "특정 개인을 알아볼 수 없는 형태"에 대한 논의가 동조의 해석에 필수적인 부분을 차지하고 있기 때문이라 할 수 있다.

이와 관련하여 개인정보처리자가 자신이 보유하고 있는 개인정보의 식별성을 제거한다는 것은 그리 쉬운 일이 아니라면서, 개인정보취급자와 연구자가

구분되어 있고 내부 지침이나 계약에 의해서 연구자가 원본 데이터에 접근하지 못하도록 되어 있다면, 어느 정도의 식별성이 존재한다고 하더라도 통계처리 및 학술연구를 위한 이용·제공을 허용하여야 한다는 주장이 있다.[158] 즉 이에 의하면 통계작성 및 학술연구를 위해 개인정보를 이용하는 경우 대개 이름, 주소, 전화번호 등을 제거하고 이용 또는 제공하게 되는데, 이름과 연락처가 제거되었다고 해서 법률상 식별성이 완전히 없어지는 것은 아닌데, 원본 데이터에서 이름과 연락처를 완전히 삭제하지 않는 한 법률상 식별성은 여전히 존재하게 되다 보니 결국 이 경우에 원본 데이터까지 파기해야 하는 불이익을 감수해야하기 때문이라고 한다.

이러한 견해는 기본적으로 식별성 제거의 수준을 '익명화(匿名化; anonymizing; anonymization)'를 전제하면서 전개하는 주장이라 할 수 있는데, 이러한 익명화는 개인의 사생활을 보호하기 위해 데이터에서 특정인을 완전히 식별하지 못하도록 정보를 가공하는 것으로 어떠한 수단을 활용하더라도 원본과의 대조를 불가능하게 만드는, 즉 재식별화가 **불가능한** 수준에 이른 것을 말한다.[159] 이러한 익명화 수준의 비식별화는 개인정보 보호의 차원에서는 매우 이상적이라 하지 않을 수 없겠지만,[160] 다른 한편으로 이러한 완전한 수준의 비식별화 조치는 통계생산을 위한 개인정보의 이용을 대단히 위축시킬 우려가 있는 것도 사실이다. 이는 특히 빅데이터 환경하에서의 새로운 정보처리 기술은 과거에는 예상하지 못했던 부분적인 정보들 간의 결합을 통해 특정인에 대한 식별성을 이끌어낼 가능성이 엄연히 존재하기 때문이다. 결국, 익명화 수준의 비식별화 조치는 통계작성의 측면에서 전면적으로 수용하기는 어려운 수준이 아닐까 생각하기에, "특정 개인을 알아볼 수 없는 형태"의 해석에 있어서는 익명화 수준에 이르지 않은 적절한 '비식별화(de-identification)' 수준에 대한 사회적인 합의를 도출해 내어야 할 과제가 남게 된다고 할 수 있다.

한편, 이와 관련하여 2016년 6월 30일, 빅데이터, IoT(Internet of Things; 사물인터넷) 등 IT 융합기술 발전으로 데이터 이용 수요가 급증함에 따른 데이터 산업 활성화를 위한 정책의 일환으로, 빅데이터 활용에 필요한 비식별 조치 기준 · 절차 · 방법 등을 구체적으로 안내하여 안전한 빅데이터 활용기반 마련과 개인정보 보호 강화를 도모하고자, 국무조정실, 행정자치부, 방송통신위원회,

158) 이창범, 앞의 책, 166~167면.

159) 익명화된 정보로부터는 인기 검색어의 순위와 같이 전체적인 경향을 분석하는 것은 가능하지만 특정인의 기호를 바탕으로 한 맞춤형 서비스는 제공할 수 없다. 정용찬, 앞의 책, 89면.

160) 실제로 「국세기본법」(법률 제13552호)은 "국세청장에게 조세정책의 수립 및 평가 등에 활용하기 위하여 과세정보를 분석·가공한 통계자료(이하 "통계자료"라 한다)를 작성·관리할 의무를 부과하고 있는데, 이 경우 통계자료는 **납세자의 과세정보를 직접적 방법 또는 간접적인 방법으로 확인할 수 없도록** 작성되어야 하며(동법 제85조의6 제1항)"라고 규정하여 보다 엄격한 비식별화를 요구하고 있다.

금융위원회, 미래창조과학부, 보건복지부 등 관계부처가 합동으로 「개인정보 비식별조치 가이드라인」을 제정하여 기존 개인정보 비식별 조치 관련 지침 등을 일괄 폐지하고 2016년 7월 1일부터는 동 가이드라인을 적용하도록 함으로써 개인정보의 보호와 활용을 동시에 모색하는 세계적인 정책변화에 적극적으로 대응하기 위한 조치를 취한 바 있다. 동 가이드라인은 개인정보를 비식별 조치하여 이용 또는 제공하려는 사업자 등이 준수해야 할 조치 기준으로, '사전검토(개인정보 해당여부 판단하여, 개인정보가 아닌 것이 명백한 경우 법적 규제없이 자유롭게 활용) - 비식별조치(; 정보집합물(데이터 셋)에서 개인을 식별할 수 있는 요소를 전부 또는 일부 삭제하거나 대체하는 등의 방법을 활용, 개인을 알아볼 수 없도록 하는 조치) - 적정성 평가(다른 정보와 쉽게 결합하여 개인을 식별할 수 있는지를 '비식별조치 적정성 평가단'[161]을 통해 평가) - 사후관리(비식별 조치 안전조치, 재식별 가능성 모니터링 등 비식별 정보 활용과정에서 재식별 방지를 위해 필요한 조치 수행)'의 단계별 조치사항을 통해 제시하였다. 이 가운데 특히 비식별조치로는 가명처리, 총계처리, 데이터 삭제, 데이터 범주화, 데이터 마스킹 등 여러 가지 기법을 단독 또는 복합적으로 활용할 것을 제시하면서, '가명처리' 기법만 단독으로 활용된 경우는 충분한 비식별조치로 보기 어렵다고 규정하였다. 동 가이드라인이 예시한 비식별조치 방법들은 다음과 같다.

<표 13> 비식별화조치 일반적 기법

대범주	소범주	내용
가명처리 (Pseudonymization)	· 홍길동, 35세, 서울거주, 한국대 대학 → 임꺽정, 30대, 서울 거주, 국제대 재학	① 휴리스틱 가명화 ② 암호화 ③ 교환방법
총계처리 (Aggregation)	· 임꺽정 180cm, 홍길동 170cm, 이콩쥐 160cm, 김팥쥐 150cm → 물리학과 학생 키 합: 660cm, 평균키 165cm	④ 총계처리 ⑤ 부분총계 ⑥ 라운딩 ⑦ 재배열
데이터 삭제 (Data Reduction)	· 주민등록번호 901206-1234567 → 90년대 생, 남자 · 개인과 관련된 날짜정보(합격일 등)는 연단위로 처리	⑧ 식별자 삭제 ⑨ 식별자 부분삭제 ⑩ 레코드 삭제 ⑪ 식별요소 전부삭제
데이터 범주화 (Data Suppression)	· 홍길동, 35세 → 홍씨, 30~40세	⑫ 감추기 ⑬ 랜덤 라운딩 ⑭ 범위 방법 ⑮ 제어 라운딩
데이터 마스킹 (Data Masking)	· 홍길동, 35세, 서울 거주, 한국대 대학 → 홍○○, 35세, 서울 거주, ○○대학 재학	⑯ 임의 잡음 추가 ⑰ 공백과 대체

161) 동 가이드라인은 비식별조치 적정성 평가단의 구성에 있어서, **해당기관**의 개인정보 보호책임자가 3명 이상의 관련 분야 전문가로 구성하되, 외부전문가를 과반수 이상으로 위촉하도록 규정하고 있다.

아울러 동 가이드라인은 비식별화 조치가 이루어진 데이터에 대한 재식별화 가능성을 검토하기 위해 다음과 같은 재식별가능성 검토기법을 제시하고 있다.

<표 14> 재식별가능성 검토기법

기법	의미	적용례
k-익명성	공개된 데이터에 대한 연결공격(linkage attack) 등 취약점을 방어하기 위해 제안된 프라이버시 보호 모델 특정인임을 추론할 수 있는지 여부를 검토, 일정 확률수준 이상 비식별 되도록 함	동일한 값을 가진 레코드를 k개 이상으로 함. 이 경우 특정 개인을 식별할 확률은 1/k임
l-다양성	k-익명성에 대한 두 가지 공격, 즉 동질성 공격(homogeneity attack) 및 배경지식에 의한 공격(Background knowledge attack)을 방어하기 위한 모델 특정인 추론이 안된다고 해도 민감한 정보의 다양성을 높여 추론 가능성을 낮추는 기법	각 레코드는 최소 l개 이상의 다양성을 가지도록 하여 동질성 또는 배경지식 등에 의한 추론 방지
t-근접성	l-다양성의 취약점 - 쏠림 공격(skewness attack), 유사성 공격(similarity attack) -을 보완하기 위한 모델 l-다양성 뿐만 아니라, 민감한 정보의 분포를 낮추어 추론 가능성을 더욱 낮추는 기법	전체 데이터 집합의 정보 분포와 특정 정보의 분포 차이를 t 이하로 하여 추론 방지

그렇지만 이 가이드라인 역시 「통계법」 등 관련법령에 다라 개인정보를 수집·이용하는 경우에는 당해 법령에 따라 처리하도록 규정하고 있어,[162] 여전히 통계작성과 관련한 비식별화조치의 수준에 대한 추가적인 논의의 필요성은 남아 있다고 할 수 있다.

3. 소결

지금까지 통계일반에 관한 법적고찰을 바탕으로, 개인정보보호의 측면에서 통계의 법적문제를 개괄적으로나마 살펴보았다. 연구의 주요한 성과는 다음과 같이 정리해 볼 수 있겠다.

- 통계제도는 기본적으로 국가통계제도가 중심이 될 수밖에 없겠지만, 민간통계의 존재도 무시해서는 안 될 것이다.
- 통계제도 역시 입헌주의 및 법치주의의 원리 하에서 운용되어야 함에 따라, 법률유보 및 과잉금지원칙의 준수는 필수적이다.
- 통계제도의 운용에 있어서 전통적으로 사생활의 비밀과 자유의 보호에 중점을 두고 있었지만, 오늘날에는 개인정보의 보호에 좀 더 중점을 둘 필요가 있다.

162) 국무조정실·행정자치부·방송통신위원회·금융위원회·미래창조과학부·보건복지부, 『개인정보 비식별조치 가이드라인 - 비식별 조치 기준 및 지원·관리체계 안내』(2016), 3면.

- 통계작성을 위한 자료의 수집과 관련하여 특히 민간부분의 자료수집에 있어서는 기본적으로 행정조사로서의 통계의 본질을 고려할 필요가 있으며, 공공부분의 자료수집에 있어서는 「통계법」 및 「전자정부법」 등의 관련조항들을 적극적으로 활용할 필요가 있다.
- 통계작성에 있어서 개인정보 보호와 관련하여 현행 개인정보보호 법제는 통계작성에 관한 예외조항을 마련하고 있어 실정법상 결정적인 장애요인은 발견되지 않는다 하겠지만, 추후 발생할 수도 있을 헌법적 측면의 논란에 사전적으로 대비하기 위해서라도 적극적으로 통계작성에 있어서 법률유보 및 과잉금지의 원칙을 준수할 필요가 있다.

이처럼 현재의 법체계 내에서는 빅데이터를 활용한 통계작성에 있어서 결정적인 장애라고 할 수 있을만한 부분은 크게 드러나지 않는 것으로 확인된 바 있으며, 오히려 통계의 국가·사회적 중요성 및 학문적 가치 등을 고려한다면, 일정 수준의 품질이 확보된다는 전제하에 개인정보 보호를 완전히 도외시하지 않는 범위 내에서라면, 통계작성을 위해서 보다 적극적으로 정보의 활용을 검토하는 것이 바람직할 것으로 생각된다. 이를 위해서는 충분한 사회적인 합의를 바탕으로 한 개인정보의 비식별화 수준을 토대로, 개인정보의 통계작성 등의 목적 내 활용의무의 준수와 비밀유지 및 안정성 확보에 보다 많은 노력을 경주할 필요성이 크지 않을까 싶다. 이는 무엇보다 개인정보의 이용에는 정보주체의 권리를 침해하는 측면이 존재하는 동시에 정보주체, 정보처리자 및 공공의 이익을 증진하는 측면도 존재하므로 일방적이고 절대적으로 정보주체의 권리만을 보호하는 것은 현대 사회에서 가능하지 않을 뿐만 아니라 바람직하지도 않다고 할 것이기 때문이다.[163]

다만 본 장의 서두에서도 언급한 바 있듯이 통계에 대한 법적인 측면에서의 논의가 많지 않은 현 상황에서 본고의 주장들은 어느 정도까지는 시론(試論)의 성격을 넘어서기가 어려울 것이다. 향후 좀 더 많은 법학자들과 법 실무가들의 적극적인 관심이 필요한 부분이라 할 것이기에 보다 적극적인 문제제기와 후속적인 논의가 필수적이라 하겠다.

특히 통계제도에 대한 법제도적 관점에서 우선적으로 추가적인 논의가 필요한 논점들로는 다음과 같은 것들을 들 수 있을 것이다.

- 통계제도의 헌법적 수용이 필요한가에 대한 논의는 언제가 될지는 모르겠으나 향후 진행될 가능성이 존재하는 헌법개정의 국면 이전에 마무리가 될 필요가 있다.
- 위 논점과 관련하여 국가통계제도의 기본구조에 대한 본격적인 고찰 역시 통계의 객관성·정확성 등을 바탕으로 한 공정성 확보 및 행정부분의 통계작성을 위한 자료수집 등의 관점에서 적극적으로 검토해 볼 가치가 있다.
- 법률유보 나아가 의회유보(Parlamentsvorbehalt)의 관점에서 「통계법」의 위상에 대한 검토와 함께 현행 통계관련 법령의 전체적인 점검을 수행할 필요성이 있다.

아울러 현대의 정보화된 사회에서 특히 빅데이터 환경에서 추가적으로 논의해 볼 주제들로는 다음과 같은 것들을 생각해 볼 수 있을 것이다.

- 현행 개인정보보호법제, 특히 최근의 「개인정보 비식별조치 가이드라인」에 있어서도 통계와 관련한 개인정보 보호 조치를 「통계법」 등에 일임하고 있기 때문에, 통계의 작성 및 운용에 있어서 비식별조치 및 재식별가능성 검토기법 및 운용수준은 「통계법」의 측면에서 적극적으로 설정될 필요가 있다.
- 국경에 구애되지 않는 사이버 공간은 통계작성을 위한 정보수집에 있어서 정보요구권한의 실효성의 문제 및 정보의 국가간 이전을 비롯한 다수의 중요한 문제를 낳게 된다. 예를 들어, 정치적인 사유 등으로 인해 자국 내 서버에서 인터넷을 자유롭게 사용하지 못할 경우 이메일이나 블로그와 같은 인터넷 서비스의 주 사용무대를 국내법의 효력이 미치지 못하는 해외 서버로 옮기는 행위를 가리키는 '사이버 망명(cyber asylum)'[164]은 우리에게도 낯설지 않은 이슈라 할 수 있는데, 외국의 정보보유주체에게 통계작성을 위한 정보요구가 어디까지 가능한지에 대한 검토가 필요하게 될 것이다.
- 통계작성을 위해 축적된 정보들에 대한 개인의 프라이버시의 문제, 특히 잊힐 권리(the right to be forgotten)와 같은 측면에서의 문제제기 역시 의미 있게 고려해 볼 필요가 있을 것이다.

163) 장주봉, 앞의 글, 63면.

164) 정용찬, 앞의 책, 89면.

참 고 문 헌

(사)한국헌법학회 편,『헌법주석1』(서울: 박영사, 2013).

Ian Hacking, The Taming of Chance, 정혜경 역,『우연을 길들이다: 통계는 어떻게 우연을 과학으로 만들었는가?』(서울: 바다출판사, 2012).

Stephen M. Stigler, History of statistics, 조재근 역,『통계학의 역사』(파주: 한길사, 2005).

고길곤,『통계학의 이해와 활용』(고양: 문우사, 2014).

고학수 편,『개인정보 보호의 법과 정책』(서울: 박영사, 2014).

곽관훈, “기업의 빅데이터(Big Data) 활용과 개인정보 보호의 조화”,『일감법학』제27호(2014).

국무조정실·행정자치부·방송통신위원회·금융위원회·미래창조과학부·보건복지부,『개인정보 비식별 조치 가이드라인 - 비식별 조치 기준 및 지원·관리체계 안내』(2016).

김기환 연구책임,『국내외 통계제도 및 통계작성현황 비교분석 연구용역(최종보고서)』(한국: 통계개발원, 2009).

김남진·김연태,『행정법I(제18판)』(파주: 법문사, 2014).

김달호 외 7인 공저,『통계로 세상보기』(파주: 자유아카데미, 2012).

김동희,『행정법I(제20판)』(서울: 박영사, 2014).

김두만, “국가통계작성 기획 및 승인관리”, 제5회 국가통계방법론 심포지엄(2015).

김주영,『정보시장과 균형: 헌법사회학적 접근』(서울: 경인문화사, 2013).

김지훈, “빅데이터와 개인정보보호”,『법제연구』제46호(2014).

매일경제 기획팀·서울대 빅데이터 센터,『빅데이터 세상: 당신의 숨겨진 욕망까지 읽어드립니다』(서울: 매경출판, 2014).

박형준,『빅데이터 전쟁: 글로벌 빅데이터 경쟁에서 살아남는 법』(서울: 세종서적, 2015).

박훤일, “정부기관의 정보열람 요구와 ISP의 협조의무”,『경희법학』, 제51권 제2호(2016).

배상호·신제수·전삼현·정현수, “통계처리를 위해 수집된 개인정보에 대한 개인정보보호 개선방안에 관한 연구”,『중소기업융합학회 논문지』제6권 제2호(2016).

법제처 편,『헌법주석서IV: 법원 등에 관한 장(제101조부터 제130조까지)』(서울: 법제처, 2010).

성선제,『개인정보보호법』(서울: 서울경제경영, 2014).

손영화, “빅데이터 시대의 개인정보 보호방안”,『기업법연구』제28권 제3호(통권 제58호)(2014).

심규호·박시내, “통계이용 활성화를 위한 2차 자료 생산 활용 방안 연구”,『(통계개발원) 2010년 하반기 연구보고서 제II권』(한국: 통계개발원, 2010).

이재형,『국가통계시스템 발전방안』, 한국개발연구원 연구보고서(2004).

이지영,『빅데이터의 국가통계 활용을 위한 기초연구』통계개발원 2015 상반기 연구보고서(2015).

이창범,『개인정보 보호법』(파주: 법문사, 2012).

장치성 외 4인공저,『국가통계이해』(대전: 통계교육원, 2015).

전광희, “선진국 공식통계의 패러다임 변용에 관한 연구 - 인구센서스와 경제센서스를 중심으로”,『(충남대학교 사회과학연구소) 사회과학연구』제22권 제3호(2011).

정용찬,『빅데이터』(서울: 커뮤니케이션북스, 2013).

정하중,『행정법개론(제8판)』(파주:법문사, 2014).

최승필, “통계(通計)의 공법적 의미와 과제”,『공법학연구』제8권 제2호(2007).

통계개발원,『국가통계제도의 발전(국제공동연구보고서)』(한국: 통계개발원, 2008).

통계청,『한국통계발전사: 위대한 숫자의 역사 - 시대사』(대전: 통계청, 2015).

한성안,『(인문학으로 풀어보는) 통계학』(서울: 청람, 2013).

홍정선,『행정법원론(상)(제22판)』(서울: 박영사, 2014).

황인창·이대용·이청호,『(알기쉬운) 통계학(제3판)』(서울: 비앤엠북스, 2015).

Ⅳ. 가계부채 관련 현황조사

유승동(상명대학교 금융경제학과)

Ⅳ. 가계부채 관련 현황조사

1. 서론

최근 가계부채의 증가로 (경제의 외부에서부터) 기대하지 않은 부정적 충격이 발생하는 경우, 경제에 대한 효과가 증폭될 수 있다는 우려의 목소리가 증가하고 있다. 통계청은 가계부채와 관련하여 2016년 정부업무보고에서 “금융자료와 같은 빅데이터를 활용해서 국민수요에 부응하는 통계를 적시에 제공”할 것을 요청받았다. 금융자료는 금융공기업 등 공공 부문에서 제공하는 자료와 은행, 보험 등 민간 부문에서 제공하는 자료로 분류할 수 있다. 그리고 금융부분에서 정형화된 데이터는 “기존 데이터베이스 관리도구의 데이터 수집·저장·관리·분석 역량을 넘어서는 데이터”의 형태로 보관 및 관리되는 경우가 빈번하다. 따라서 본 연구에서는 금융자료 가운데에서도 가계부채를 대상으로 하는 (정형화된) 빅데이터와 가계부채 관련된 기존 데이터, 통계, 그리고 조사 및 설문의 개선방안과 이를 근거로 개발 가능한 신규 통계에 대한 논의를 진행한다.

빅데이터를 연구하고 있는 한국정보화진흥원(2014)에서는 가계소비 및 소득과 관련되어 가계대출 동향을 최우선 신규지표 후보로 제시하였다. 가계대출과 관련된 통계의 중요성이 매우 높다고 제시하고 있으며, 가계부채와 관련된 정보는 사회, 경제, 정치적으로 관심이 높아 빠르게 개선되고 있는 상황이다.[165] 예를 들어 2016년 종합신용정보집중기관으로 한국신용정보원이 출범하였다. 기존 연구에 따르면 가계부채와 관련된 빅데이터는 확보의 용이성과 빅데이터 기술의 확장성이 보통인 수준으로 분류하고 있다.[166] 따라서 본 연구는 2016년 6월 기준 가계부채와 관련된 빅데이터 환경과 중장기적으로 데이터에 대한 확보의 용이성을 개선하고 이를 확장할 수 있는 방안에 대해 검토한다.

금융감독기관 및 중앙은행에서는 다양한 채널을 활용하여 가계부채와 관련된 통계정보를 제공하고 있다. 기존 가계부채 관련 시장자료는 금융기관의 안정성에 초점을 두고 있으며, 금융기관의 자산정보를 기반으로 가계부채에 대한 정보를 제공하고 있다. 정책적으로 가계의 안정성에 대한 이슈도 가계부채의 다른 중요한 한 측면이므로 이에 대한 정보의 보강과 이를 통한 시장분석이 필요하다.

165) 한국정보화진흥원(2014)의 경우 가계부채에 대한 원자료 보유기관은 “은행연합회”라고 인지하고 있다. 은행연합회 회원사인 은행뿐만 아니라 제2금융권과 보험사 등도 가계에 대한 대출을 실행하고 있다. 따라서 기존 자료와 관련 문헌에서는 은행연합회를 가계부채 관련 원천정보를 보유하고 있다고 주장하고 있다.

166) 데이터의 확보는 그 중요성이 매우 높은(“중앙행정기관에서 활용하여 국민체감도와 관련된 통계분야”) 것으로 분류하고 있다. 자료확보의 용이성의 측면에서“협의를 통하여 자료 확보 노력이 필요”한 수준인 보통으로 분류하며, 빅데이터의 기술 확장성 측면에서도 “차후 확장성이 제약”이 존재하고 있는 보통수준이다.

기존 가계부채와 관련된 통계자료는 금융기관 중심의 거시적 시장자료가 대부분이며, 가계부채의 중요한 다른 한 측면인 가계의 미시적 자료는 기대에 부흥하지 못하고 있다(유승동, 2015). 재언하자면 기존 통계자료는 금융기관의 자산 건전성과 관련된 자료이며, 가계부채에 대한 자료와는 차별성이 있을 수 있다. 본 연구에서는 빅데이터 환경에서 통계청에서 향후 진행할 업무에 대한 발굴을 진행한다. 통계청에서 진행하고 있는 미시적 가계통계와 관련된 발전방안을 제시한다. 마지막으로 본 연구에서는 빅데이터를 활용하여 우리나라 가계부채 시장을 진단할 수 있는 신규통계의 개발방안을 논의한다.

2. 가계의 재무정보

제1절 가계의 재무정보 분류

가계의 재무관련 정보는 두 가지 측면에서 접근할 수 있다. 첫 번째로 일종의 재무제표의 측면에서 가계의 재무적 관점의 저량으로 자산, 부채 그리고 순자산이 있다. 가계의 재고(stock) 측면의 재무정보로 특정시점의 수량이나 금액을 의미한다. 두 번째로 손익계산서의 측면에서 가계 자산의 유량(flow)을 나타내는 현금의 유입과 현금의 유출이 있다.

본 연구에서는 가계의 대차대조표상에 부채정보에 관심을 가지고 진행한다. 이하 표에서는 가계의 재무제표를 구성하는 요소들을 제시하고 있다. 본 연구는 자산, 부채, 그리고 순자산으로 구성된 재무제표 가운데 부채를 중심으로 논의한다.

<표 15> 가계의 재무제표

자산	부채
	순자산

가계자산은 실물자산, 금융자산, 기타자산 그리고 인적자산으로 구분될 수 있다. 자산 가운데 실물자산은 주택 등 부동산을 포함한 자산을 의미하고, 금융자산은 저축, 주식, 채권 등을 의미하며 기타 자산이 있다. 인적자산의 경우 정확한 자산규모 측정이 어려워 일반적으로 재무제표에 반영하지 못한다. 인적자산의 경우 (현재와) 미래의 가계에 현금의 유입·유출과 높은 상관관계를 보일 수 있을 것이다.

가계부채는 대출기관을 중심으로 금융기관에 대한 부채, 비금융기관에 대한 부채가 있다. 이와 같은 분류체계를 근거로 하는 경우 한국은행에서 집계하고 있는 가계부채에 대한 정보는 금융기관 중심의 가계부채에 대한 정보이다. 다시 말해

가계부채는 금융기관에게는 자산이 되며, 가계부실이 발생하는 경우 이는 금융기관이 보유하고 있는 자산부실을 의미한다. 따라서 현재의 가계부채에 대한 논의는 금융기관의 자산관점에 기반을 두고 있다. 총량적 관점에서 금융기관에서 취급한 가계부채에 대한 관리는 주로 금융시장의 안정성 측면에서 접근하고 있는 것이다. 기존 가계부채와 관련된 논의들의 대부분은 가계부채에 대한 건전성 보다는 금융기관 자산의 건전성에 초점을 두고 있다(유승동 2015). 순자산의 경우 가계자산에서 가계부채를 제외한 금액으로 자산과 부채에 대한 평가를 통하여 간접적으로 산출할 수 있다. 최근 순자산의 활용에 대한 관심이 증가하고 있다. 동시에 최근 유럽 등 선진국에서는 순자산을 복지제도에 활용하는 정책개발에 논의가 추진되고 있다.

가계는 소비자이며 기업에 생산요소를 제공하고 이에 대한 소득을 창출한다. 동시에 가계는 기업의 소유자가 될 수도 있다. 가계가 보유하고 있는 상가정보, 상가임대차 정보의 경우 경제적 분류와 달리 관행적으로 기업정보로 분류되는 경우가 빈번하다.[167] 예를 들어 상업용 부동산담보대출의 많은 경우 가 기업대출로 분류되고 있다. 참고로 2016년 2월부터 감독기관에서는 금융기관에게 상업용 부동산담보대출관련 정보제공을 요구하고 있는 것으로 알려져 있다.

가계의 손익계산서의 측면에서 현금의 유입과 현금의 유출로 구분할 수 있음을 다음의 <표 16>에서 확인할 수 있다. 현금의 유입은 소득과 기타 현금의 유입이 있다. 현금의 유출은 현재의 소비와 미래를 준비하는 저축 혹은 투자 그리고 부채 즉 현재의 활용을 위하여 가계가 금융기관 등에서 부채로 조달한 자금에 대한 상환이다. 일정기간 중에 발생 혹은 감소한 수량이나 금액을 나타내는 유량 정보를 의미한다. 가계부채가 증가하는 경우 가계의 소비와 저축의 변화를 촉발하며, 금융기관의 자산 및 부채에 대한 변화를 유발할 수 있다. 소비와 저축은 이미 경제학에서 매우 전통적이고 중요한 연구의 대상이다. 최근에는 가계부채의 이슈에 대한 부각으로 신규 대출 및 대출금 상환에 대한 측면에 관심이 증가하고 있다. 본 연구에서도 가계부채 측면에서 현금의 유출입과 이에 대한 부채의 상환능력을 뒷받침할 수 있는 소득과 관련된 논의를 진행한다.

<표 16> 가계의 손익계산서

현금의 유입	현금의 유출
소득 기타 유입	소비 저축 및 투자 대출금 상환

167) 법률적으로 사업자등록증을 발급을 통하여 상가 등을 운영하는 경우 기업으로 간주될 수 있다.

제2절 자산정보에 대한 논의와 자료구축 방향

2000년대 들어 금융기관의 경쟁증가로 가계부채에 대한 이슈가 부각되기 시작하였다(You, 2009). 과거 가계부채에 대한 관심이 현재와 같지 않던 상황에서 가계자산과 연관된 정보는 사회·경제의 주요한 관심의 대상이었다. 가계금융과 관련된 기존 다수의 학술연구는 가계자산을 연구하였고, 이와 관련된 시장정보 축적작업은 소기의 성과를 달성하였다고 평가할 수 있다. 가계자산 정보와 관련된 체계적인 연구가 필요하며 본 연구에서는 가계가 보유하고 있는 실물자산 그리고 금융자산에 대한 논의를 간략하게 진행한다.[168]

<표 17> 가계금융 복지조사에서 자산 유형별 가구당 보유액 및 구성비

구 분	자산	금융자산						실물자산				
			저축액				전·월세 보증금		부동산			기타 실물자산
				적립식	예치식				거주주택	거주주택 이외1)		
평균	2014년	33,539	9,013	6,676	3,681	2,481	2,338	24,526	22,678	12,364	10,314	1,848
	2015년	34,246	9,087	6,740	3,844	2,411	2,346	25,159	23,345	13,179	10,166	1,815
	증감률	2.1	0.8	1.0	4.4	-2.8	0.4	2.6	2.9	6.6	-1.4	-1.8
구성비	2014년	100.0	26.9	19.9	11.0	7.4	7.0	73.1	67.6	36.9	30.8	5.5
	2015년	100.0	26.5	19.7	11.2	7.0	6.9	73.5	68.2	38.5	29.7	5.3
	전년차	-	-0.3	-0.2	0.2	-0.4	-0.1	0.3	0.6	1.6	-1.1	-0.2

(단위 : 만원, %, %p)

주 : 1) '거주주택 이외'에는 '계약금 및 중도금'이 포함됨

자료: 통계청. 2015. "2015년 가계금융·복지조사 결과." 2015년 12월 21일 보도자료.

앞의 표에서 확인할 수 있듯이 가계금융복지조사에 따르면 2015년 현재 우리나라 가계자산의 약 27%는 금융자산이고, 73%는 실물자산이다. 실물자산 가운데 거주주택이 약 37% 그리고 거주주택 이외의 부동산이 약 31%로 부동산자산이 가계자산의 약 68%에 이른다. 금융자산 가운데 전·월세 보증금이 약 7%를 차지하고 있음으로 부동산 자산과 직·간접적으로 관련된 자산은 가계자산의 약 75%에 달한다.

168) 기타자산 즉 계불입금, 빌려준 돈 등의 정보에 대한 논의는 제외한다.

<표 18> 자산관련 현황

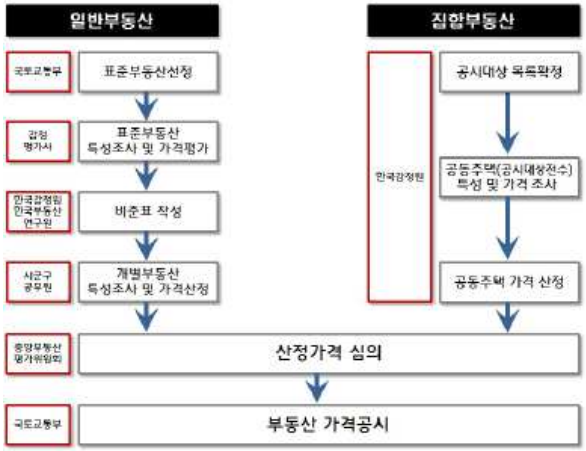
구분		자료명	보유기관	관련법률
금융자산		개인별 예·적금	은행 등	은행법 금융실명제법
		보험 및 연금	보험회사	보험업법 금융실명제법
		주식 및 채권	증권회사 등 한국예탁결제원	자본시장과 금융투자업에 관한 법률 금융실명제법
		주택·상가 임차(전세)보증금	대법원, 국토부 국세청, 지자체	주택임대차보호법, 부가가치세법 부동산등기법, 임대주택법
실물자산	부동산	등기부등본 (토지, 건물)	대법원	부동산등기법
		토지현황 및 시가표준액	행자부, 지자체 국토부	지방세기본법 공간정보의 구축 관리 등에 관한 법률
		주택현황 및 시가표준액		
	기타	자동차등록 현황 및 과세표준액	행자부, 지자체	지방세기본법
		선박등록 현황 및 과세표준액		
		항공기등록 현황 및 과세표준액		
		건설기계등록 현황 및 과세표준액		
		골프회원권		
		승마회원권		
		콘도회원권		
		요트회원권		
		종합체육시설 이용회원권		

자료: 통계청 내부자료.

가계자산 가운데 약 25%가 (보증금을 제외한) 순수 금융자산으로 볼 수 있다. 기획재정위원회에 의하면 금융자산 정보에 대한 접근성이 금융실명법 제4조와 관련법 (은행법, 보험업법, 자본시장과 금융투자업에 관한 법률)에 따라 제한된다. 자산 유형별 자료를 보유하고 있는 기관과 관련 법률은 앞의 표에서 확인할 수 있다. 국세청은 통계청에게 일부의 과세자료(원천징수, 종합소득 등)를 제공하고 있어, 이를 기반으로 원천징수 금융자산의 규모를 추정할 수도 있다. 그러나 이는 자산규모를 추정하는데 있어 일부 한계가 존재할 수 있다. 한국정보화진흥원(2014)이 금융자료는 자료확보의 용이성의 측면에서 "협의를 통하여 자료 확보 노력이 필요"하다고 한다. 그러나 협의(혹은 간접적 방식)로 금융자산 정보에 대한

접근성을 개선하고자 하는 노력은 비용이 과다할 수도 있다[169].

2015년 현재 가계가 보유하고 있는 자산 약 75%는 부동산과 관련된 자산이다[170]. 부동산의 가격 (혹은 가치)관련 정보는 정보의 보유기관 성격에 따라 공공기관과 민간기간으로 구분할 수 있다. 그리고 현재 공공기관에서 부동산 가격과 관련된 정보는 크게 공시가격과 실제거래 가격으로 양분할 수 있다.



[그림 10] 현행 부동산가격 공시 프로세스

자료: 방송희. 2015. 부동산 가격공시제도를 둘러싼 이슈, HF-이슈 리프트, 2015-10. p. 5.

공시가격은 부동산가격공시 및 감정평가에 관한 법률에 근거한 부동산가격 공시제도에 기반을 두고 있다. 토지에 대한 가격공시 제도는 1989년 그리고 건축물에 대한 가격의 공시제도는 2005년에 도입되었다. 적정가격으로 공시가격의 경우 국세 및 지방세의 과세기준 등에 활용되고 있으며, 자세한 활용범위는 아래의 <표 19>를 참조할 수 있다. 그러나 공시가격의 경우 적정가격이란 개념은 모호할 수 있다는 한계가 있다(방송희, 2015). "표준 부동산 평가금액은 시장가치의 80-90% 수준"에서 결정되는 것으로 알려져 있다. 그리고 "유형간, 지역간, 용도간, 가격대별 공시가격의 현실화 수준이 차이"를

169) 가계의 금융자산에 대한 정보에 대한 접근성은 관련법의 개정을 통하여 확대되는 것이 효과적으로 보인다. 가계의 금융자산 현황은 해당기관에서 원천자료를 보유하고 있어 접근성을 확대하기 위하여 제도적 보강이 필요하다.

170) 기타 실물자산이 차지하고 있는 비중은 5.5%에 불과함으로 자동차, 선박, 항공기와 같은 실물자산은 차후 연구과제로 남긴다.

보이고 있다. 따라서 공시정보를 활용하여 가계의 자산규모를 추정하는 것은 체계적인 오류에 노출될 가능성을 배제할 수 없다. 그러나 공시가격은 전국의 토지와 건물에 대한 적정가격을 포함하고 있다. 공시가격은 국토부에서는 1년 단위로 조사하여 발표하고 있어 가계가 소유하고 있는 부동산의 가격정보를 주기적으로 파악할 수 있는 장점이 있다.

부동산 실거래 가격의 경우 2006년 부동산거래를 투명화하기 위하여 "부동산 실거래 가격 신고 의무화제도"가 시행되고 있다. 토지 혹은 건축물을 매매한 경우 30일 이내에 실제거래 가격을 시·군·구청에 신고하여야 한다 (이를 위반시 과태료가 매도자, 매수자 그리고 중개업자에게 부과된다). 부동산 실거래 가격정보의 경우 국토교통부의 부동산거래관리시스템(Real Estate Trade Management System : RTMS)에 축적되고 있다. 동 실거래 가격을 활용하여 "부동산시장을 실시간 모니터링하여 적시에 효과적이고 예측가능한 부동산정책을 수립"하고 있다.

민간기관은 KB 국민은행, r114, 닥터 아파트 등이 대표적인 정보제공 업체로 이들은 주택가격 정보를 관리하고 있다. 예를 들어 KB 국민은행의 경우 1986년 주택은행 시절부터 가격지수를 만들어, 우리나라에서 가장 오래된 KB 국민은행 주택가격지수란 시계열 자료를 공표하고 있다. 민간기관에서 제공하고 있는 부동산 가격정보는 실제가격이 아닌 호가 즉 집주인이 받고자 하는 가격 혹은 부동산 거래업체에서 일반적으로 거래가 될 것이라고 믿는 가격이다. 따라서 공공기관에서 이와 같은 정보를 활용하여 통계를 작성하는 경우 신뢰성이 저해될 수 있다. 호가정보의 경우 '평가가격에 기초한 지수'에서 흔히 나타나는 평활화(smoothing) 현상을 갖고 있다는 비판도 존재한다(이용만, 2008).

중장기적으로 가계의 자산, 부채, 그리고 순자산 정보에 대한 정부의 부처간 공유와 협력체계에 대한 구축이 필요하다. 금융실명제법, 주택임대차보호법, 부동산 등기법 등 다양한 제도적 제약이 존재할 수 있다. 그럼에도 불구하고 최근 사회·경제적인 논의가 확대되고 있는 가계의 건전성 관리와 위험제거를 위한 정책개발을 위하여, 자료 공유와 협력체계의 구성은 긴 호흡을 유지하며 보완될 필요가 있다. 과거 약 10여 년 전에는 우리나라 가계부채의 규모에 대한 정보가 부재하였다는 상황을 감안할 필요가 있다. 따라서 이를 보완하기 위한 추가 작업이 지속적으로 진행되어야 할 것이다. 부동산의 가격정보의 경우만 하더라도 다양한 지표가 존재할 수 있음으로 자료구축의 목적과 용도를 확인하고, 이를 고려한 자료구축 전략이 추진되어야 할 것이다.

<표 19> 개별공시지가 활용범위

제 도	적 용 범 위	적 용 근 거
- 국 세		
· 양도소득세	- 양도가액 산정을 위한 기준시가	「소득세법」 제99조제1항
· 증여세	- 증여재산가액 산정을 위한 기준시가	「상속세 및 증여세법」 제61조제1항
· 상속세	- 상속재산가액 산정을 위한 기준시가	「상속세 및 증여세법」 제61조제1항
· 종합부동산세	- 과세표준액 결정자료	「종합부동산세법」 제13조
- 지방세		
· 재산세	- 과세표준액 결정자료	「지방세법」 제187조제1항
· 취득세	- 과세표준액 결정자료	「지방세법」 제111조제2항
· 등록세	- 과세표준액 결정자료	「지방세법」 제130조제2항
- 기 타		
· 개발부담금	- 개발사업 개시시점지가의 산정	「개발이익환수에 관한 법률」 제10조 제3항
· 개발제한구역훼손부담금	- 개발제한구역훼손부담금 산정기준	「개발제한구역의 지정 및 관리에 관한 특별조치법」 제24조제1항
· 개발제한구역내 토지 매수	- 개발제한구역내 매수대상토지 판정기준	「개발제한구역의 지정 및 관리에 관한 특별조치법」 시행령 제28조
· 국·공유재산의 대부료·사용료	- 대부료·사용료 산정을 위한 토지가액	「국유재산법」 시행령 제26조 제2항

자료: 국토해양부, 《2010년도 적용 개별공시지가 조사·산정지침》, p. 7

3. 가계부채관련 통계현황

제1절 가계부채의 규모관련 통계

가. 통화금융통계(한국은행의 가계신용)

우리나라 가계부채 통계는 대표적으로 한국은행 경제통계국에서 생산하고 있는 국가승인통계인 통화금융통계 내에 존재하는 가계신용이다. 가계신용은 가계부채 규모로 대용변수로 활용되고 있고, 이는 가계대출과 판매신용으로 구성되고 있다. 그러나 가계부채의 경우 가계신용을 포괄하는 광의의 개념으로 볼 수 있다.

<표 20> 가계신용, 가계대출, 판매신용

가계신용	일반가정이 은행 등 금융기관에서 빌린 돈이나 외상으로 물품을 구입하고 진 빚의 합. 단, 개인간의 거래인 사채는 제외
가계대출	가계일반자금대출(은행 등에서 빌린 일반대출금, 신용카드회사의 현금서비스 및 카드론 포함)과 가계주택자금대출(주택은행 등에서 집을 사기 위해 빌린 돈)로 구분
판매신용	신용카드로 물품을 구입하거나 자동차 가전제품 기타상품을 할부로 구입한 금액

자료: 통계청 홈페이지(2016년 10월 현재).

가계신용은 주기적으로 공표되고 있으며, 한국은행의 통계정보시스템에서 2003년 12월부터 월별 가계신용 잔액을 확인할 수 있다. 월별로 가계대출 그리고 분기별로 가계신용에 대한 정보를 보도자료 등을 통하여 공표도 병행하고 있다. 2015년 4/4분기 가계신용의 경우 1,207조원이며, 이는 언론에서 중요한 경제기사로 취급되었다.

가계대출 그리고 신용과 관련된 정보는 지속적으로 개선되고 있다. 가계부채에 대한 논의의 증가와 더불어 가계신용에 대한 통계정보에 대한 개선요구가 증가하였다. 이를 반영하듯 2007년 12월부터 지역별 가계대출 규모를 확인할 수 있다. 가계신용을 월별 주택담보대출, 기타대출로 구분한 통계도 공표되고 있다. 과거 단순 잔액기준의 정보에서 벗어나 최근에는 신규 취급액에 대한 정보도 집계하고 있다.[171] 한국은행에서 발표하고 있는 가계대출의 경우 예금취급기관(예금은행과 비은행)과 기타 금융기관(보험기관, 연금기금, 여신전문회사, 공적금융기관, 기타금융기관 및 기타)의 가계대출로 구성되어 있다. 따라서 (기준시점) 공식적인 금융기관이 가계에게 대출한 공식 잔액이며, 비공식 잔액을 포함하고 있다.

171) 저자는 과거 10년 전인 2005년 주택담보대출의 시장의 규모를 경제모형으로 추정하였다. 최근 들어 가계부채와 관련된 시장정보의 축적과 발전은 매우 빠른 속도로 진행되고 있음을 실감하고 있다.

<표 21> 통화금융통계

목적	통화당국 및 금융기관의 대차대조표 등을 기초로 각종 통화금융통계를 분석, 편제하여 통화신용정책 및 여타 정책수립의 기초자료로 활용
통계종류	지정통계
작성기관	한국은행
작성형태	보고통계
통계분야	금융
승인일자	1976.7
공표주기	월
공표방법	언론(보도자료)+전산망(인터넷)+간행물 조사통계월보, 통화금융, 경제통계연보 보도자료 : http://ecos.bok.or.kr/ KOSIS : 통화금융통계
참고사항 (주요용어)	가계신용:가계대출 및 판매신용 가계대출: 금융기관이 가계에 공여한 대출 판매신용: 재화(물품)의 판매(생산)자나 서비스 제공자가 제공하는 외상(신용)거래를 포괄

자료: 통계청 홈페이지.

<표 22> 가계신용 잔액 (단위: 조 원)

	2014		2015			
	3/4	4/4	1/4	2/4	3/4	4/4
가 계 신 용	1,056.4	1,085.3	1,098.3	1,131.5	1,165.9	1,207.0
	(20.6)	(28.8)	(13.0)	(33.2)	(34.4)	(41.1)
	<62.8>	<66.2>	<75.9>	<95.6>	<109.5>	<121.7>
가 계 대 출	999.0	1,025.1	1,039.3	1,072.0	1,102.4	1,141.8
	(20.6)	(26.1)	(14.2)	(32.7)	(30.4)	(39.4)
	<60.1>	<64.5>	<74.0>	<93.6>	<103.4>	<116.8>
판 매 신 용	57.4	60.2	59.0	59.5	63.4	65.1
	(-0.1)	(2.8)	(-1.2)	(0.5)	(3.9)	(1.7)
	<2.8>	<1.7>	<1.8>	<2.0>	<6.0>	<5.0>

주: ( )안은 전분기대비 증감액 < >안은 전년동기말대비 증감액

자료: 한국은행. 2016. “2015년 4/4분기중 가계신용.” 2016년 2월 24일 보도자료.

은행업감독업무시행세칙에 근거하여 금융기관들은 금융감독원에 은행업무보고서를 제출한다. 금융기관과의 협의결과 금융감독원에 제출하는 것과 유사한 형태의 보고서를 한국은행에 제출하고 있다고 알려지고 있다. 이를 근거로 한국은행에서는 통화금융통계에서 가계부채 관련 정보를 계산하고 있는 것으로 알려져 있다. 매월 가계대출금 현황보고서를 작성하여 금융감독원에 제출하고 있는 것이다.

다음의 그림은 금융감독원의 은행업무보고 양식의 사례이다.[172]

가계대출금 현황보고서(B2426) [작성주기 : 월]

금융기관 _____ 기 준 월 _____ 전 화 번 호 _____
작성자소속 _____ 작성자직위 _____ 작성자성명 _____
확인자소속 _____ 확인자직위 _____ 확인자성명 _____

(단위 : 백만원, %)

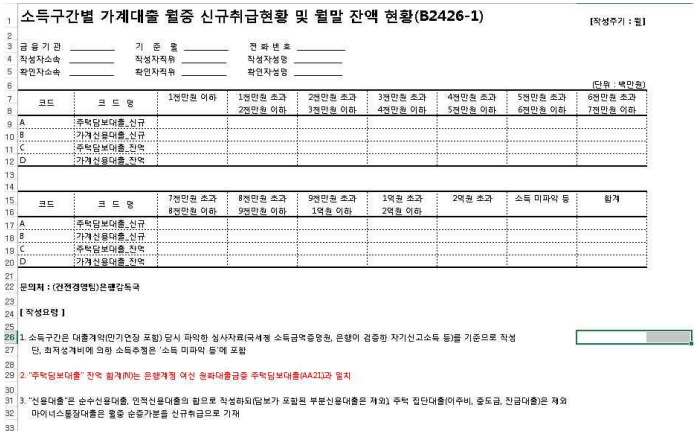
코드	코드명	대출채권(A)	1월이상 원금 연체기준		1개월이상 원리금 연체기준	
			연체대출채권(B)	연체율(C=B/A)	연체대출채권(D)	연체율(E=D/A)
A	가계대출금(A1+A2)					
A1	은행계정(A1=A11+A12)					
A11	가계자금대출금(A11)					
A11a	(주택담보대출)					
A11b	(신용대출)					
A111	5백만원 이상 ~ 1천만원 미만					
A112	5백만원 미만					
A12	가계부문 신용카드채권(A12)					
A121	5백만원 이상 ~ 1천만원 미만					
A122	5백만원 미만					
A2	신탁계정 가계대출금(A2)					
A21	5백만원 이상 ~ 1천만원 미만					
A22	5백만원 미만					

문의처 : (건전경영팀)은행감독국

[작성요령]

1. "대출채권(A)"
 - A11(가계자금대출금) : "여신종별 연체대출채권(B2410)"의 "가계자금대출금(AA2)"와 일치
 - A11a(주택담보대출) : 가계 일반자금대출금 중 주택담보대출 외에도 주택자금대출 등 여타대출에 속한 주택담보대출을 모두 포함
 - A12(가계부문 신용카드채권) : "여신종별 연체대출채권(B2410)"의 "신용카드채권(AF)"중 기업관련 신용카드채권(기업회원카드, 기업구매전용카드 등)을 제외한 금액
 - A2(신탁계정 가계대출금) : "여신종별 연체대출채권(B2410)"의 "가계자금대출금(BA2)"와 일치
2. 1월이상 원금기준 "연체대출채권(B)" : "여신종별 연체대출채권(B2410)"의 작성요령 제4조 참조
3. 1개월 이상 원리금기준 "연체대출채권(D)" : "여신종별 연체대출채권(B2410)"의 작성요령 제5조 참조

[그림 11] 금융감독원의 은행업무보고서 양식 1



소득구간별 가계대출 월중 신규취급현황 및 월말 잔액 현황(B2426-1) [작성주기 : 월]

금융기관 _____ 기 준 월 _____ 전 화 번 호 _____
작성자소속 _____ 작성자직위 _____ 작성자성명 _____
확인자소속 _____ 확인자직위 _____ 확인자성명 _____

(단위 : 백만원)

코드	코 드 명	1천만원 이하	1천만원 초과 2천만원 이하	2천만원 초과 3천만원 이하	3천만원 초과 4천만원 이하	4천만원 초과 5천만원 이하	5천만원 초과 6천만원 이하	6천만원 초과 7천만원 이하
A	주택담보대출_신규							
B	가계신용대출_신규							
C	주택담보대출_잔액							
D	가계신용대출_잔액							

코드	코 드 명	7천만원 초과 8천만원 이하	8천만원 초과 9천만원 이하	9천만원 초과 1억원 이하	1억원 초과 2억원 이하	2억원 초과	소득 미파악 등	합계
A	주택담보대출_신규							
B	가계신용대출_신규							
C	주택담보대출_잔액							
D	가계신용대출_잔액							

문의처 : (건전경영팀)은행감독국

[작성요령]

1. 소득구간은 대출계약(만기연장 포함) 당시 파악한 심사자료(국세청 소득금액증명원, 은행이 검증한 자기신고소득 등)를 기준으로 작성
 단, 최저생계비에 의한 소득추정은 '소득 미파악 등'에 포함
2. "주택담보대출" 잔액 합계(N)는 은행계정 여신 원화대출금중 주택담보대출(AA21)과 일치
3. "신용대출"은 순수신용대출, 인적신용대출의 합으로 작성하되(담보가 포함된 부분신용대출은 제외) 주택 집단대출(이주비, 중도금, 잔금대출)은 제외
 마이너스통장대출은 월중 순증가분을 신규취급으로 기재

[그림 12] 금융감독원의 은행업무보고서 양식 2

자료: 금융감독원 홈페이지(2016년 1월 현재)

금융기관은 과거와 비교하여 최근 대출 잔액을 다양한 기준으로 구분하여 제출하고 있다.

172) 은행업무보고서 양식의 경우 금융감독원에서 은행에게 제공하고 있다.

가계자금 대출금 현황의 경우 은행계정과 신탁계정을 분류 및 주택담보대출과 신용대출을 구분하고 있다. 대출 잔액을 천만 원 기준으로 세분화(1천만 원 이하, 1천만 원-2천만 원,..., 7천만 원 초과 등)하고, 대출의 잔액과 신규 대출 잔액(가계신용과 주택담보대출을 구분하고, 신규금액과 잔액)을 구분하였다. 대출자의 상환능력에 대한 정보가 요구됨에 따라 소득도 천만 원 기준으로 세분화하였다.

가계대출, 주택담보대출, 신용카드 대출의 신규연체와 상각추이(상각, 대환, 정상화에 대한 정보를 포함하여 제공하고 있음)에 대한 정보도 은행은 주기적으로 보고하고 있다. 대출채권의 규모와 더불어 1월 이상 연체채권의 규모와 연체율의 정보도 제공하고 있다. 주택담보대출의 경우 건전성 기준으로 잔액, LTV 별 잔액, 지역별 잔액, 지역별 LTV 잔액, DTI 기준 잔액 등을 포함한다. 언론에서 관심을 보이는 강남3구 등 지역정보를 세분화 하여 보고하는 것은 최근 가계대출과 관련된 다양한 유형의 정보가 필요하다는 것을 의미한다. 그러나 금융기관별로 보유하고 있는 대출 잔액(loan balance)기준으로 집계되고 있어서 대출자 유형, 대출의 특성 등에 대한 정보를 파악하는데 한계가 있다.

제2절 국민주택기금 및 주택분양보증 현황

국민주택기금은 1973년 주택건설촉진법에 근거하여 한국주택은행에서 설치되었으며, 2015년 주택도시기금법에 근거하여 국토교통부장관이 운용 및 관리하고 있다.[173] 주택도시기금은“국민주택채권, 청약저축, 융자금 회수 등으로 자금을 조성하여 국민주택 및 임대주택 건설을 위한 주택사업자와 주택을 구입 또는 임차하고자 하는 개인수요자에게 자금을 지원”하고 있다.

국토교통부에서 운영 중인 주택도시기금과 관련된 통계는 국민주택기금 및 주택분양보증 현황을 통하여 공표하고 있다. 주택도시기금은 주거안정정책을 위하여 기금의 운영실적을 파악하기 위한 목적으로 작성되어, 2006년 국가승인통계로 지정되었다. 주택도시기금의 경우 기금의 운영실적 파악을 통하여 주거안정 정책을 실행하기 위한 목적으로 동 현황자료를 작성하고 있다. 주택자금현황에서 시작된 현황자료는 2009년 11월 국민주택기금 및 주택분양보증 현황에 통합되었다. 최근 가계부채 관련 통계의 개선작업을 통하여 주택도시기금이 가계에게 제공한 대출은 한국은행에 가계신용에 포함되고 있다.

국민주택기금 및 주택분양보증 현황은 국민주택기금의 재원별 조성내역과 운용용도별 운용실적 및 주택기금 대출실적, 주택채권 발행현황 및 주택보증현황을

173) 국민주택기금은 2015년 이후 주택도시기금으로 전환되었으며, 기금의 운용·관리에 관한 사무의 전부 또는 일부를 주택도시보증공사에 위탁하고 있다.

발표한다. 작성대상은 국민주택채권발행, 융자금회수 등 조성내역, 주택건설자금, 구입자금, 전세자금 등 운용내역, 기금 수탁은행 영업점 및 주택도시보증공사 지점 등이다.

<표 23> 국민주택기금 및 주택분양보증 현황

목적	주택건설자금, 주택구입자금 및 주택전세자금 등 무주택서민들의 주거안정을 위해 설치운용 중인 국민주택기금의 조성 및 운용실적을 매년 파악하여 주거복지정책의 기초자료로 활용
통계종류	일반통계
작성기관	국토교통부
작성형태	보고통계
통계분야	주택
승인일자	2006.10
조사주기	연
공표방법	공표방법 : 전산망(인터넷)+간행물 간행물명 : 국토교통통계연보(익년 12월) KOSIS : 국민주택기금및주택분양보증현황

자료: 통계청 홈페이지(2016년 10월 현재).

제3절 주택금융 및 유동화증권 통계

주택금융공사는 2004년 3월 출범하여 보금자리론과 적격대출 공급, 주택보증, 유동화증권 발행 등의 업무를 수행하고 있다. 주택금융공사의 고유 업무인 대출실행, 채권발행과 관련된 통계인 주택금융 및 유동화증권통계를 2005년 국가승인통계로 지정하여 주택금융공사에서 공표하고 있다.

국민주택기금의 사례와 마찬가지로 최근부터 가계부채 관련 통계개선의 일환으로 주택금융공사의 대출도 가계신용에서 집계되고 있다. 주택금융 및 유동화증권통계의 경우 대출시장 정보와 더불어 가계부채와 관련된 채권시장 정보도 제공하고 있다. 가계가 금융기관에게 대출을 받는 시장을 1차 시장이라 분류하며, 주택금융공사의 경우 직접취급하고 있는 가계대출과 관련된 통계를 제공하고 있다. 금융기관은 채권을 통하여 자금을 조달하며 이와 같은 자금조달 시장을 2차 시장이라 한다. 주택금융 및 유동화증권 통계는 2차 시장에서 주택저당증권(MBS) 그리고 주택저당채권(MBB)에 대한 통계를 포함한다.

주택금융 및 유동화증권통계의 보고 항목은 보금자리론 판매 실적 및 현황, 주택유동화증권(MBS · MBB) 발행 실적 및 현황, 주택금융신용보증 공급 및 잔액 현황, 주택금융신용보증기금 출연기준 주택자금대출 규모, 주택연금 공급 및 잔액 현황, 주택구입능력지수(K-HAI), 주택구입기회지수(K-HOI)이다.[174]

보금자리론 실적은 한국주택금융공사에 양도하지 않은 실적을 포함한다.

동 통계는 지역별 소득대비주택가격비율(PIR) 등도 특정대상(예를 들어 보금자리론 이용자)을 중심으로 산출한다. MBS, MBB 발행실적은 보금자리론 이외에 공사가 금융기관으로부터 양수한 주택담보대출을 기초로 발행된 실적을 포함한다. 그리고 주택담보대출을 이용하여 주택의 구입능력을 표시하는 K-HAI 그리고 K-HOI가 포함되어 있다.[175]

<표 24> 주택금융 및 유동화증권 통계

목적	주택금융 및 주택저당증권(MBS) 관련 통계의 제공을 통해 주택금융시장 동향에 관한 대국민 서비스를 제고하고 정부의 주택금융 정책 수립에 기여
통계종류	일반통계
작성기관	한국주택금융공사
작성형태	보고통계
통계분야	금융
승인일자	2005.4
조사주기	월
공표방법	전산망(인터넷)+간행물 주택금융월보(익월말)

자료: 통계청 홈페이지(2016년 10월 현재).

4. 가계부채관련 설문/조사 통계

제1절 한국노동패널조사

한국노동패널조사(이하 노동패널)는 “노동정책의 수립과 실행을 위해서는 신뢰할 수 있는 통계 데이터의 뒷받침이 필수적이지만, 우리나라에서 실시하고 있는 주요 노동관련 통계 조사만으로는 역부족(통계청, 2015)”하여 도입하였다. 노동시장의 변화에 대한 정보와 더불어 가계의 자산 및 부채와 관련된 연구와 정책에서도 활용되고 있다. 우리나라 대표적 패널조사로 인정받고 있으며, 해외에서도 장기적 조사에 대한 호의적인 반응을 보이고 있다.[176]

한국노동연구원에서 1998년 최초로 노동정책에 대한 수요를 뒷받침하기

174) 1988년 설립된 주택금융신용보증기금은 2004년 한국주택금융공사 내 설치되었다.

175) HAI는 “주택가격, 연소득, 대출금리 등 주택구입에 영향을 미치는 다양한 변수를 동시에 고려하여 하나의 지수로 나타낸 것으로 주택구입능력을 파악할 수 있는 구체적인 지표(김다스라, 2011)”이다. “HAI에서는 대출가능금액이 고려되고 있는 것이다. 물론 대출가능금액은 소득, 이자율, 대출기간, LTV 등 주택금융시장 상황에 의해 결정(유승동, 2013)” 된다.

176) 유사한 해외 패널조사는 미국 PSID(Panel Study of Income Dynamics), 영국 BHPS(British Household Panel Study), 독일 GSOEP, 캐나다의 SLID 등이 존재한다.

위하여 노동패널을 실시하였으며, 보다 세부적으로 자세한 현황의 경우 조사연구학회(2016)를 참조할 수 있다. 1998년 1차 조사(wave)의 경우 도시에 거주하고 있는 5,000가구를 대상으로 진행하였다. 그러나 “노동시장 정책의 효과는 일시적이고 단기적인 것도 있지만, 그 효과가 장기적・지속적(통계청, 2015a: 3)”임으로 패널형태의 조사를 결정하였다. 2014년 (17차 조사) 현재 조사를 실시한 가구수는 6,738가구이며, 이중 1998년 원표본가구는 3,451가구로 69.0%의 표본유지율을 보이고 있다(한국노동연구원, 2015). 2014년 조사에 응답한 개인의 경우 13,163명이며, 이 가운데에서도 1차 조사 즉 1998년 표본은 10,757명이다. 2015년까지 18년도 조사를 실시하여 장기간의 시계열 자료를 축적하고 있으며, 다양한 분야에서 학술연구와 정책수립으로 활용되고 있다.

노동패널은 가계의 부채관련 정보를 분류하였다는 측면에서 의의가 있다. 그러나 1차 년도 조사에서 2차 년도의 조사항목이 대폭 변경되었다. 1차 년도의 경우 단순히 부채여부, 총 부채액, 그리고 월평균 부채 상환금을 조사하였다. 그러나 2차 연도부터 부채를 상대적으로 세부적으로 분류하였다. 부채를 금융기관 부채, 비금융기관부채, 개인적으로 빌린 돈, 전세금 임대보증금, 부어야 할 계 그리고 기타로 구분하여 조사하고 있다. 가계부채는 개인적으로 빌린 돈 등 비공식 금융거래까지 포함하고 있음을 확인할 수 있다. 따라서 가계부채의 규모는 앞 절에서 논의한 바와 같이 금융기관 즉 공식적 금융기관의 부채와 더불어 비공식적 부채를 포함하고 있는 광의의 개념이다. 설문지에서는 부채구분을 부채를 이용한 기관(institutions)을 이하와 같이 구분하고 있다.

a. 비금융기관 부채
b. 개인적으로 빌린 돈
c. 계
d. 보증금

따라서 한국은행에서 발표하고 있는 금융기관 부채 및 비금융기관에 국한되어 조사하고 있는 가계신용에 대한 정보와 보완적인 역할을 수행할 수 있다.

<표 25> 한국노동패널조사 부채관련 설문지 1

항 목	부채 유무	잔액	원금과 이자 상환금
사5-1. 금융기관 부채	(1) 예 (2) 아니오	_____만원	월평균 _____만원
사5-2. 비금융기관 부채 (회사를 통해 빌린 돈 등)	(1) 예 (2) 아니오	_____만원	월평균 _____만원
사5-3. 개인적으로 빌린 돈 (사채나 친척/친지에게 빌린 돈 등)	(1) 예 (2) 아니오	_____만원	월평균 _____만원
사5-4. 전세금 받은 것, 임대보증금 받은 것	(1) 예 (2) 아니오	_____만원	월평균 _____만원
사5-5. 미리 타고 앞으로 부어야 할 계	(1) 예 (2) 아니오	_____만원	월평균 _____만원
사5-6. 기타 (_____)	(1) 예 (2) 아니오	_____만원	월평균 _____만원

<표 26> 한국노동패널조사 부채관련 설문지 2

항 목	부채 유무	잔액	원금과 이자 상환금	해당 차수
금융기관 부채	H**2601	H**2602	H**2603	
비금융기관 부채	H**2604	H**2605	H**2606	
개인적으로 빌린 돈	H**2607	H**2608	H**2609	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
전세금/ 임대보증금 받은 것	H**2610	H**2611	H**2612	
미리 타고 앞으로 부어야 할 계	H**2613	H**2614	H**2615	
기타 (_____)	H**2616	H**2617	H**2618	

자료: 한국노동패널조사 설문지.

7차 조사부터 부채를 이용한 원인에 대한 조사를 실시하고 있다. 이는 가계 부채에 대한 원인을 밝히는 데 도움을 줄 수 있다. 그리고 13차부터 가계부채의 중요한 원인으로 지목되고 있는 교육을 고려하기 위하여 1)자녀교육비와 2)자녀이외의 교육비를 구분하여 원인에 대한 조사를 실시하고 있다. 그러나 부채를 이용한 원인이 15가지에 달하고 있으므로, 가계부채의 유형별 규모를 고려하여 일부 조정이 불가피할 수 있다.

<표 27> 한국노동패널조사 중 부채를 이용한 원인 설문 항목

사5-7 부채가 있다면, _____님 댁에서 부채를 지게 된 가장 주된 이유는 무엇입니까?

H**2633 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

(1) 교육비
(2) 주택마련
(3) 전세금 마련
(4) 자동차, 가구, 가전제품 등 내구재 구입비
(5) 생활비
(6) 결혼, 상제비
(7) 질병, 재난
(8) 창업 또는 사업 자금
(9) 재테크(주식 및 부동산 투자)
(10) 보증
(11) 전세금 받은 것, 임대보증금 받은 것, 미리 탄 곗돈
(12) 가족이나 친지, 친구를 도움
(13) 기타 (_____)

자료: 한국노동패널조사 설문지.

노동패널은 노동시장과 관련된 환경변화 및 정책수립에 대한 관심으로 가계대출의 가구별 전체규모만 질의하고 있다. 가계부채의 질적인 내용에 대한 현황 및 정보에 대한 보완이 필요하며, 가계부채의 질적인 수준을 진단하기 위해서는 개선이 필요하다. 가계부채의 유형 즉 주택담보대출 및 그 외 즉, 신용대출 등과 관련된 세부적 분류의 체계성이 높지 않다.

장기적 시계열 자료의 안정성을 고려하는 동시에 가계부채와 관련된 현금흐름 정보에 대한 보강이 필요하다. 부채상환금을 2차 조사부터 질의하고 있지만 유효한 응답이 저조하다. 예를 들어 1-17차 통합 코드북 자료에서 17차 조사에는 15개 그리고 16차 조사에는 26개에 불과하다. 1-17차 통합 코드북에 근거하면 부족한 생활비 마련방법에 대하여 은행대출이나 마이너스 통장, 현금서비스, 친구 친지에게 빌림, 사채 이용에 대한 질의를 진행하지만, 응답 빈도가 거의 없는 것으로 확인된다. 그리고 대출상환에 있어서도 원금과 이자에 대한 구분이 진행될 수 있다. 가계대출의 특성 예를 들어 상환방법, 원금상환액, 이자지급액 그리고 금리조건 등과 관련된 정보에 대한 보완이 필요할 수 있다. 마지막으로 가계부채와 관련된 유량정보를 즉 상환금액 등과 관련된 정보에 대한 보완이 가능하다. 이와 같은 보완 혹은 조정은 조사목적에 따라 장기적인 추세를 고려하여 진행하여야 한다.

제2절 복지패널

한국보건사회연구원에서 복지패널은 2003년부터 저소득·취약계층을 대상으로 설문조사를 실시하였고, 2005년부터는 차상위 계층에 대한 확대조사를 추진하였다. 2006년에는 우리나라 국민의 복지실태에 대한 동태적 변화를 판단하기 위하여 국가승인통계로 지정되었다. “빈곤층, 근로 빈곤층, 차상위층의 가구형태, 소득수준 및 취업형태에 초점”(한국보건사회연구원, 2013:4)을 맞추고 있다. 따라서 저소득층을 과대표집 하여 사회적 약자를 위한 분석을 진행하는 것을 목적으로 하고 있다. 2011년 빈곤실태조사를 기초로 신규표본 1,800가구를 추출하였고, 저소득층의 경제적, 사회적, 인구적 특성에 관심을 두고 있다. 복지패널의 세부현황은 조사연구학회 (2016)를 참조할 수 있다.

가계부채 측면에서 복지패널은 금융기관뿐만 아니라, 사적금융을 이용한 정보에 대한 조사를 진행한다. 저소득층의 경우 고소득층과 비교하여 상대적으로 공식 금융시장에 접근성이 제한적일 가능성이 높고, 사적 금융시장에 의존할 가능성이 높기 때문일 것이다. 금융기관의 대출과 함께 일반사채, 카드빚, 전세보증금, 외상, 곗돈 그리고 기타 부채에 대한 조사를 실시하고 있다는 것은 특징적이다.

노동패널과 달리 복지패널은 상환정보를 이자와 원금으로 구분하여 조사하고 있다. 이자의 경우에도 주거관련 이자와 기타이자를 구분하고 있다. 무엇보다 원금 상환을 통하여 원금의 변화를 추적하고 있어, 가계부채와 관련된 저량과 유량의 정보를 동시에 조사하고 있다. 연체와 관련된 정보를 조사하고 있어, 가계가 부채로 인하여 어려움을 겪고 있을 가능성 등과 관련된 정보를 조사하고 있다.

복지패널은 노동패널과 마찬가지로 가계부채의 특징과 세부 정보에 대한 보완이 바람직하다. 설문 응답자의 경우 연체에 대한 정보를 제공하는 것을 기피할 수도 있다. 그러나 저소득층에 대한 정확한 재무상환을 진단하기 위하여 이를 포함할 수도 있다. 조사의 시간적 일관성을 고려한 개편이 필요할 것이다. 동시에 소득층이 활용하고 있는 부채의 특성이 다른 계층과 차별적일 수 있다. 저소득층과 취약계층의 경우 다른 계층과 달리 금융에 대한 이해력이 낮을 가능성이 높기 때문에 이에 대한 고려가 포함되어야 할 것이다.

<표 28> 복지패널 가계부채 관련 조사항목 1 - 주거

주택구입비용, 보증금 마련 경로(1순위)	1.자기돈(상속인경우포함) 2.무상으로 도움을 받음 3.부모,형제,친척,친구등으로부터 빌림 4.금융기관(회사에서 융자받은 경우 혹은 모기지론도 포함)으로부터 빌림 5.사채
주택구입비용, 보증금 마련 경로(2순위)	위와같음
1년간 총 원금상환액(가)	단위: 만원
12월 31일 기준 주택관련 부채액(나)	단위: 만원
1년간 대출상환액 연체횟수	1.연체한적이없다 2.1회 3.2~3회 4.4회이상

<표 29> 복지패널 가계부채 관련 조사항목 2 - 부채 및 이자

금융기관대출	연간 부채액(단위: 만원)
일반사채	연간 부채액(단위: 만원)
카드빚	연간 부채액(단위: 만원)
전세보증금(받은돈)	연간 부채액(단위: 만원)
외상, 미리탄곗돈	연간 부채액(단위: 만원)
기타부채	연간 부채액(단위: 만원)
주거관련 부채의 이자	2011년 지출한 이자액 (단위: 만원)
기타이자	2011년 지출한 이자액 (단위: 만원)
1년간 총 원금상환액(가)	단위: 만원
12월 31일 기준 주택관련 부채액(나)	단위: 만원
1년간 대출상환액 연체횟수	1.연체한적이 없다 2.1회 3.2~3회 4.4회이상

자료: 한국 복지패널 홈페이지.

제3절 고령화연구패널조사

한국고용정보원에서 중고령 인구의 경제활동에 대한 정확한 실태조사를 통해 향후 고령사회로 변화해 가는 과정에서 개인의 행동을 예측하고 이를 토대로 효과적인 사회경제정책을 실행하기 위하여 고령화연구패널조사를 실시하고 있다. 동 조사는 2006년에 승인통계로 지정되었으며 만45세(패널구축당시인 2006년 기준 1962년 이전 출생한 응답자)이상을 대상 2년 주기로 조사를 진행하고 있다. 세부현황은 조사연구학회 (2016)에서 참조할 수 있다.

<표 30> 고령화연구패널조사의 부채 및 부채의 변화관련 조사사항

F013. 현재 살고 계신 집을 구입하기 위하여 받으신 대출 또는 부채가 있으십니까? 있으시다면 남아있는 대출금이나 부채는 얼마입니까?

F014. 현재 살고 계신 집의 일부를 세를 주고 계십니까? 다음 중 해당하는 사항을 모두 선택해 주십시오.(복수응답선택)
① 전세를 주고 있음
② 보증금 있는 월세를 주고 있음
③ 보증금 없는 월세를 주고 있음
④ 세를 주고 있지 않음 → F057

…

F037. 다음은 부채에 관한 질문입니다. 돌아가신 분께서 남기신 부채가 있습니까?
[면접원: 부채의 종류는 금융기관 대출금, 사채, 부동산 보증금 등이 있습니다.]
① 예
⑤ 아니오 → F044
…

F205. 다음은 부채에 관한 질문입니다. _____님께서 혹시 가지고 계신 부채 항목에 답해 주시기 바랍니
다. 부채에는 금융기관 대출, 신용카드 현금서비스 및 대출, 개인적으로 빌린 돈, 채무 보증, 기타 등이 포함됩니다. 현재 은행, 보험사, 증권사 신용카드사 등 금융기관으로부터 대출 받으신 것이 있습니까?

F212. 앞서 말씀하신 금융기관이나 신용카드사 외에 친척이나 친구 또는 타인으로부터 개인적으로 빌린
부채가 있습니까?

F219. [응답자 성명]께서는 현재 친척, 친구, 타인의 채무에 대한 보증을 서주고 계십니까?
① 예
⑤ 아니오 → F226

F226. 앞서 언급된 부채 항목들 외에 또 다른 부채를 가지고 계십니까?
① 예
⑤ 아니오 → F234

F227. 어떤 종류의 부채입니까?
[면접원: ① 금융기관으로부터 빌린 대출금, ② 친척이나 친구 또는 타인으로부터 개인적으로 빌린 부채, ③ 친척, 친구, 타인의 채무에 대한 보증에 대해 여쭈어 보았습니다. 여기서 응답자가 또 다른 부채로 응답하신 내용이 앞에서 언급한 3가지 항목에 포함되면, F220=⑤로 수정하시고 해당 설문 문항으로 이동하셔서 입력하세요.]

E279. (생활비가 소득보다 많아) 생활비가 부족하였다면, _____님 댁에서는 부족한 생활비를 어떻게 마련하셨습니까? 가장 주된 방법부터 순서대로 3개만 선택해 주십시오. - 보험해약이나 곗돈을 미리 탄 경우 등은 ⑧ 저축이나 예금이나 적금의 해약에 해당

① 은행대출이나 마이너스통장 이용
② 현금서비스 이용
③ 친척이나 친지에게 빌림
④ 친구나 이웃에게 빌림
⑤ 사채이용
⑥ 부동산매각이나 전세금 인상
⑦ 전세나 월세의 규모를 줄임
⑧ 저축이나 예금이나 적금의 해약
⑨ 주식이나 채권을 비롯한 금융자산 매각(파생금융상품(CD, MMF 등) 포함)
⑩ 자동차나 내구재 또는 금/은 등의 귀중품 매각
기타

자료: 고령화연구패널조사 설문지.

동 조사에서 부채는 주택에 대한 부채와 그 외 부채로 구분하고 있다. 주택에 대한 부채의 경우 담보대출과 전월세 보증금에 대한 정보로 구성되어 있으며, 그 외 부채는 a) 금융기관 대출, b) 개인적으로 빌린 돈 그리고 c) 기타 부채로 구성되어 있다. 중고령 인구를 대상으로 함으로 개인보증을

가계부채에 포함하고 있는 것은 흥미로운 점이다. 이는 잠재부채에 대한 고령자들에 대한 부담을 조사에 반영하고 있다. 예를 들어 가계부채를 논의할 때 개인보증도 고려되고 있기 때문이다. 2016년 현재 제2금융권의 연대보증이 철폐되었다(기존 금융기관의 경우 폐지되었다).[177] 보증폐지는 신규보증을 대상으로 하고 있으며, 기존 연대보증의 부담은 여전히 존재하는 것으로 알려져 있다. 그리고 비공식적 금융시장에서도 연대보증은 존재하고 있는 것으로 알려져 있다.

고령화연구패널조사는 다른 조사와 비교하여 부채와 관련하여 상대적으로 간략한 조사를 진행하고 있다. 보증을 고려하고 있으며 추가적으로 사망하신 분의 부채상속과 관련 정보도 조사하고 있다. 따라서 동 조사는 중고령자 들에게 실질적으로 필요한 조사를 실시하고 있는 것으로 보인다.[178] 그러나 향후 중고령 인구의 부채에 대한 상환능력에 대한 조사가 진행되는 것이 바람직하다. 현재 부채의 잔액과 고령자들을 위한 실질적인 부채(보증, 부채상속)에 대한 정보가 있다.

부채증가에 대한 정보를 조사하고 있지만, 부채의 상환에 대한 정보가 개선될 수 있다. 고령층이 활용하고 있는 부채의 특성이 다른 계층과 차별적일 수 있음으로 이에 대한 고려가 진행될 수 있을 것이다. 그리고 저소득층과 마찬가지로 중고령자의 경우에도 다른 계층과 비교하여 금융에 대한 이해력이 낮을 가능성을 배제할 수 없다.

제4절 주택금융 및 보금자리론 수요실태조사

주택금융 및 보금자리론 수요실태 조사의 경우 현재 한국주택금융공사에서 실시하고 있으며, 2010년 국민은행의 주택금융수요실태조사와 통합되었다. 과거 1973년 제1회 조사는 융자주택실태조사로 주택은행 시절 민영주택자금 대출가구를 대상으로 조사되었다. 주택은행을 합병한 국민은행은 2002년부터 조사대상을 대출을 받지 않은 가구까지로 확대하여 명칭을 주택금융수요실태조사로 변경하였다. 그리고 조사대상을 주택은행 거래고객에서 전 가구를 대상으로 확대하였다.

주택금융공사에서는 2015년 현재 주택보유 및 거주가구, 주택금융 이용실태, 그리고 주택금융 인식 및 수요에 대한 인식에 대한 조사를 실시하고 있다.

177) 기본적으로 연대보증이 폐지되었지만, "개인사업자·법인 대출 및 보증보험에 대해서는 최대주주·대주주(30%이상)·대표이사(고용임원제외) 등에 대해서는 연대보증 허용"되고 있다. 중소기업 CEO의 책임경영을 확보하는 차원에서 법인 대표자 1인에게 연대보증 입보 요구되었었지만, 2015년에도 우수기업의 경우 개인보증을 폐지하기로 결정되었다.

178) 다만 향후 금융정책에 있어서 개인보증을 폐지하고 추세를 감안하는 경우 중장기 적으로 개인보증의 규모는 줄어들 것으로 기대된다.

전국의 만 20세-59세 가구주 그리고 그 배우자를 모집단으로 선정하고, 2010년 인구주택 총 조사를 기준으로 샘플을 선정하여 대출을 활용한 가구 그리고 활용하지 않은 일반가구를 포함하여 조사를 진행하고 있다. 공사의 대출상품인 보금자리론을 이용한 가구에 대한 조사를 병행하고 있다(최근 1년 이내 보금자리론을 활용한 가구를 대상으로 조사). 이에 추가로 금융기관에 보금자리론을 담당하고 있는 취급실태에 대한 조사를 병행하고 있다. 상세한 현황은 조사연구학회 (2016)를 참조할 수 있다.

정책금융을 수행하기 위하여 조사의 작성기관이 수립되었으며, 정책목표를 달성하기 위하여 기반이 될 수 있는 시장정보를 조사하고 있다. 이를 반영하듯 조사목적을 “주택금융공사 보금자리론 활성화를 위한 시장정보 도출”로, 주택금융공사의 영업과 밀접하게 관련된 정보에 대한 조사를 진행하고 있다. 조사대상에 공사의 상품을 이용한 고객과 잠재적으로 공사의 상품을 이용할 고객을 구분하여 조사하고 있다. 이를 반영하듯 “주택금융공사의 보금자리론을 선택에 장애요인”등의 질의를 포함하고 있다. 가계부채에 대한 현황에 대한 조사보다는 보금자리론에 대한 수요실태임으로 모집단이 일반 가계라고 보기 어려운 상황이다. 작성기관의 영업활동과 관련된 시장조사를 진행한다는 측면에서 의의가 있다. 따라서 전반적 가계금융 상황과 다소 거리가 있음에 따라 이를 활용한 가계금융 정책결정에 다소 보수적인 접근이 필요하다.

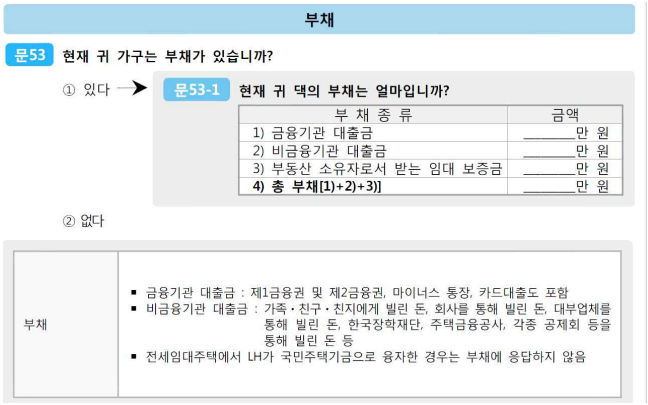
최근 연구에 따르면 “2014년 현재 일반가구 대상의 설문항목이 약 150종, 설문지는 48페이지에 이름”(국토교통부, 2015: 117)으로 건당 평균 38분이 소요되고 있다(통계청, 2015b: 18). 일반인뿐만 아니라 전문가들도 정확히 이해하기 어려운 질의가 포함되어 있다. 동시에 항목에 대한 정확한 구분이 필요한 상황이다(통계청, 2015b: 62). 따라서 조사항목에 대한 대폭적인 축소가 불가피한 상황이다(통계청, 2015b: 60). 조사목적이 공사의 영업확대와 업무방식 개선과 밀접한 연계관계가 있다. 동시에 “표본추출 단계의 정확성이 담보”(국토교통부, 2015: 117; 방송희, 2016)에 대한 지적이 꾸준히 제기되고 있다. 컴퓨터 설문을 통하여 조사비용을 대폭 줄이고 있지만, 이와 관련된 편이가 발생할 수 있을 가능성을 배제할 수 없다.

제5절 주거실태조사

국토교통부에서 “부동산시장 안정화, 주거복지 및 주거평등 실현 등을 위한 정책수립을 지원하기 위하여 국민의 주거생활에 대한 전반적인 실태를 조사하는데 목적”을 두고 주거실태조사를 실시하고 있다. 관련 법령으로 “주택법 제5조 및 동법 시행령 제6조에 의하여 국토교통부장관이 정기조사와 수시조사를 통해 실시”하고 있다. 일반조사가구와 특수조사가구에 대한 조사를 실시하고 있으며, 일반조사가구의 경우 2년 마다 짝수 해에 실시하고 있다. 노인가구, 장애인가구, 임대주택거주가구, 기초생활수급권자, 차상위계층 등 과 같이 소득과 거주특성을 고려하여 특수조사 가구조사도 실시하고 있다. 세부 내역은 조사연구학회 (2016)를 참조할 수 있다.

주거실태조사에서 부채의 경우 금융기관대출금, 비금융기관 대출금 그리고 임대보증금에 대한 조사를 실시하고 있다. 여기에서 금융기관은 제1 및 제2금융권, 마이너스 통장, 카드대출도 포함하고 있다. 그리고 부동산을 소유하고 있는 가계가 임대보증금을 받은 것도 부채로 인식하고 있다. 부동산 세입자 측면에서 임대보증금은 자산이며, 부동산 소유자 측면에서 임대보증금은 부채가 되는 것이다.

주택담보대출의 경우 주택관련 질의에서 주택구입 당시의 부채의 규모에 대한 질의를 하고 있으며, 이는 조사당시의 가계가 보유하고 있는 주택담보대출의 규모를 파악할 수 있는 방향으로 개선될 수 있다. “생활비에서 월평균 총생활비에 대한 질의를 진행하고, 주택부금상환, 빌린 돈 갚은 금액”의 경우 생활비에서 제외하고 있다. 주택과 관련된 조사로 주거관리비를 별도로 조사하고 있는 것은 특징적이다.



부채

문53 현재 귀 가구는 부채가 있습니까?

① 있다 → 문53-1 현재 귀 댁의 부채는 얼마입니까?

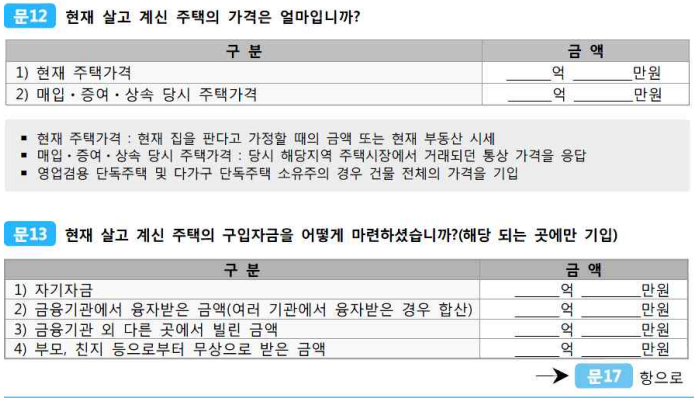
부 채 종 류	금액
1) 금융기관 대출금	_____만 원
2) 비금융기관 대출금	_____만 원
3) 부동산 소유자로서 받는 임대 보증금	_____만 원
4) 총 부채[1)+2)+3)]	_____만 원

② 없다

부채	▪ 금융기관 대출금 : 제1금융권 및 제2금융권, 마이너스 통장, 카드대출도 포함 ▪ 비금융기관 대출금 : 가족·친구·친지에게 빌린 돈, 회사를 통해 빌린 돈, 대부업체를 통해 빌린 돈, 한국장학재단, 주택금융공사, 각종 공제회 등을 통해 빌린 돈 등 ▪ 전세임대주택에서 LH가 국민주택기금으로 융자한 경우는 부채에 응답하지 않음

[그림 13] 주거실태조사 부채관련 질의서

자료: 주거실태조사 설문지.



문12 현재 살고 계신 주택의 가격은 얼마입니까?

구 분	금 액
1) 현재 주택가격	_____억 _____만원
2) 매입·증여·상속 당시 주택가격	_____억 _____만원

▪ 현재 주택가격 : 현재 집을 판다고 가정할 때의 금액 또는 현재 부동산 시세
▪ 매입·증여·상속 당시 주택가격 : 당시 해당지역 주택시장에서 거래되던 통상 가격을 응답
▪ 영업겸용 단독주택 및 다가구 단독주택 소유주의 경우 건물 전체의 가격을 기입

문13 현재 살고 계신 주택의 구입자금을 어떻게 마련하셨습니까?(해당 되는 곳에만 기입)

구 분	금 액
1) 자기자금	_____억 _____만원
2) 금융기관에서 융자받은 금액(여러 기관에서 융자받은 경우 합산)	_____억 _____만원
3) 금융기관 외 다른 곳에서 빌린 금액	_____억 _____만원
4) 부모, 친지 등으로부터 무상으로 받은 금액	_____억 _____만원

→ 문17 항으로

[그림 14] 주거실태조사 주택관련 질의서

자료: 주거실태조사 설문지.

제6절 가계금융복지조사

2012년부터 통계청은 금융감독원 그리고 한국은행이 공동으로 가계금융복지조사를 실시하고 있다. 과거 2006년 가계자산조사로 시작되었으며 5년 주기로 실시되고 있었지만, 2012년부터 매년 조사를 실시하고 패널로 변경하여 조사하고 있다. 조사목적은 “자산, 부채, 소득 등의 규모, 구성 및 분포와 미시적 재무건전성을

파악하여 사회 및 금융관련 정책과 연구에 활용”하기 위함이다. 그리고 “가계생활수준의 정도, 변화, 지속기간, 변화요인 등을 종합적으로 파악하여 재정 및 복지관련 정책과 연구에 활용되고 있다.

<표 31> 가계금융복지조사 항목 분류 체계

대분류	중분류	세분류	세세분류	세부항목
자산액	금융자산	저축액	적립식	입출금이 자유로운 저축 (현금, 당좌수표 포함) 적립식 저축 및 펀드 저축성 보험 만기시 일정금액을 수령하는 보장성 보험
			예치식	예치식 저축 및 펀드 주식, 채권, 선물, 옵션 등
			기타저축	권리금 빌려준 돈, 불입한 곗돈
		전·월세 보증금	전세보증금	전세보증금
			월세보증금	월세보증금
	실물자산	부동산	거주주택	단독, 아파트, 연립 및 다세대, 기타
			거주주택 이외 부동산	단독, 아파트, 연립 및 다세대, 건물, 토지, 해외부동산, 기타 부동산
			계약금·중도금	단독, 아파트, 연립 및 다세대, 건물, 토지, 해외부동산, 기타 부동산
		자동차	자동차	자동차
		기타 실물자산	자동차 이외 기타 실물자산	자영업자 설비 및 재고자산, 건설 및 농업용 장비, 동물 및 식물, 골프회원권, 콘도회원권, 귀금속, 골동품 또는 예술품, 고가 내구재(현재 시가 300만원 이상), 기타(지적재산권, 특허권 등의 무형자산, 오토바이, 기타 회원권 등)
부채액	금융부채	-	담보대출	거주주택 담보, 거주주택 이외 부동산 담보, 예/적금·보험·펀드·채권 담보, 기타(전세권, 보증서, 자동차 등) 담보
			신용대출	마이너스 통장포함
			신용카드 관련대출	현금서비스, 카드론, 대환대출 등
			외상 및 할부	외상, 할부, 카드 선포인트 할부 등 미결제 잔액(일시불 신용카드 미결제액 제외)
			기타부채	곗돈을 탄 후 불입할 금액
	임대보증금	-	거주주택 임대	거주 주택의 일부를 임대
			거주주택이외 임대	거주주택 이외 주택이나 건물, 토지 임대

자료: 통계청. 2015. “2015년 가계금융·복지조사 결과.” 2015년 12월 21일 보도자료.

해당 조사는 2015년까지 4차에 걸쳐 전국 가구의 재무상황을 주기적으로 파악하고 있다. 실제로 조사결과는 최근 가계부채에 대한 이슈의 부각으로 금융정책, 가계부채 정책, 부동산정책, 경제정책 등에 다각도로 정책과 연구에

적극적으로 활용되고 있다. 패널로 전환을 시작한 2012년 횡단과 패널의 분석이 동시에 가능하도록 조사를 재설계하였다. 가계금융복지조사 소개자료에 근거하면 2010년 인구주택총조사에 근거하여 확률비례통계 기법을 활용하여 전국 약 20,000가구를 추출하여 설문을 진행하고 있다. 세부적으로 금융부분 10,000가구 그리고 복지부문 10,000가구로 구성하여 설문을 진행하고 있다.

표본가구에 대하여 조사원이 직접조사하고, 면접자가 기입방식을 원칙으로 하여 가구주 혹은 가구원에 대한 조사하고 있다. 가계금융관련 정보가 명확하지 않거나, 재정상황에 대하여 명확하지 않은 경우 재조사 혹은 인터넷을 통하여 조사를 병행하기도 한다. 가계금융복지조사의 경우 자산과 부채와 관련된 정보의 경우 가구의 응답에 근거하고 있는 것으로 알려져 있다. 자산의 경우 금융자산과 실물자산으로 구분하고 있으며, 부채의 경우 금융부채와 임대보증금을 구분하고 있다.

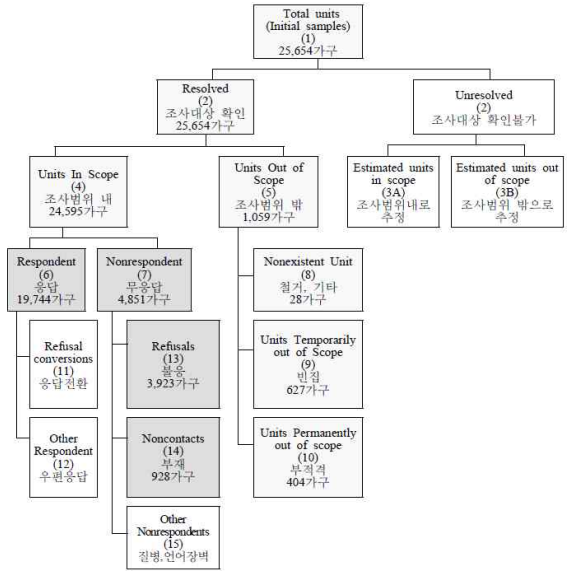
통계청에서는 가계금융복지조사의 표본가구의 소득과 관련된 정보를 관련기관에서 일부분 협조를 받고 있다. 소득 가운데 근로소득, 사업소득, 이전소득, 그리고 이전소득의 경우 국세청 혹은 보건복지부에서 관련정보를 제공받고 있으며, 동 정보를 조만간 활용할 것으로 알려져 있다. 자산 혹은 부채와 관련된 정보의 경우는 금융실명법(제4조)에 근거하여 금융소득은 통계청에서 제공받는데 한계가 있어 금융위원회 혹은 국세청 등과 적극적이고 지속적인 협의를 진행하고 있는 것으로 알려져 있다. 다행인 것은 자산의 경우 조만간 소관부처로 실물자산 특히 부동산과 관련된 정보를 향후 교류할 예정으로 알려져 있다. 그러나 금융자산의 경우에는 금융실명법에 근거하여 활용이 어려운 상황이다. 부채의 경우에도 신용정보법(제32조)에 경우 카드대출, 그리고 신용대출 등에 대한 정보를 소관부처로부터 제공받는데 한계가 있다.

가계금융복지조사에서 타 기관 행정자료를 확보해서 활용하는 것이 바람직하다. 이를 통하여 무응답을 최소화할 수 있을 것이다. 예를 들어 2012년 조사범위 내 24,959 가구 가운데 무응답율은 약 20%인 4,851가구였다. 기존 행정정보를 활용하여 조사를 실시하는 경우 설문조사에 필요한 다양한 유형의 비용절감효과를 기대할 수 있다. 정구현(2014)은 가계금융복지조사의 경우 2013년 평균 3.0회 방문하여 조사시간이 46.3분이 소요된다고 한다. 그리고 조사에 필요한 소요시간이 높을수록 다음해 무응답의 가능성이 높아졌다고 분석한다. 행정자료를 활용하는 경우 정보의 신뢰성을 확보하는 동시에 패널조사의 응답비율을 높일 수 있을 것으로 기대된다.

<표 32> 통계청에서 가계금융복지조사 관련항목 담당 및 제공현황

	항목		입수1)	제약사항	담당부처
소득	근로소득		○		국세청
	사업소득		○		국세청
	금융소득		○ (종합분)		국세청
	이전소득		○		보건복지부
	① 금융소득		X (원천분)	금융실명법 제4조 (금융거래의 비밀보장) 예외조항 미비	금융위원회, 국세청
자산	실물자산		○ (예정)		대법원 국토교통부
	②금융자산	적립식 예치식	X	금융실명법 제4조 (금융거래의 비밀보장) 예외조항 미비 금융자산정보 통합관리기관* 부재	금융위원회
부채	③금융부채	담보대출 신용대출 카드대출	X	신용정보법 제32조 (개인신용정보 제공동의) 예외조항 미비	금융위원회

자료: 기획재정위원회(2015) 내부자료



[그림 15] 가계금융복지조사 응답 및 무응답도표

자료: 변종석, 이석훈, 정구현. 2013. 가계금융 ·복지조사의 무응답 처리를 위한 유용한 보조정보 선정, 《조사연구》 14(1): p.89.

무엇보다 금융수준이 높은 가구의 경우 소요시간이 길었으며, 배우자가 응답가능성이 높아짐에 따라 금융수준이 높은 가구의 경우 오차가 상대적으로 높을 가능성이 존재한다. 따라서 자산, 소득, 부채의 규모에 따라 조사에 비협조적일 가능성이 존재함으로 설문조사의 정확성과 분포의 왜곡을 방지하기 위하여 관련자료 확보와 활용을 지속적으로 추진해야 한다.

5. 신규과제 1: 가계부채 관련 신규통계 개발 방안

통계청에서 보유하고 있는 가구에 속한 가구원 정보를 활용하여 국세청 및 보건복지부의 소득정보, 그리고 신용정보회사의 부채정보를 활용하여 가구의 총부채원리금상환부담(Debt Service Ratio, 이하 DSR) 통계의 개발을 제안한다. 현재 국내에 존재하는 가계부채와 관련된 통계는 가계금융복지조사를 제외하고는 가계(혹은 가구) 수준의 정보가 전무한 상황이다. 이미 살펴본 바와 같이 설문조사를 활용한 가계의 재무정보는 일반적 한계가 존재함으로, 이를 극복하기 위한 방안으로 객관화된 소득 및 부채정보에 활용이 바람직하다.

제1절 배경

2015년 12월부터 금융위원회와 은행연합회는 여신심사선진화가이드라인(이하 가이드라인)을 시행하고 있다. 가이드라인의 핵심은 차주의 총 금융부채 상환부담을 평가하는 시스템 즉 DSR 도입하는 것이 주요 내용이다. 전국은행연합회(2015.12.15.)에 따르면 기존의 "DTI를 활용하여 주택담보대출 차주의 금융부채 상환능력을 평가하고 있으나, 차주의 기타 금융부채의 원금을 상환하는 부담도 고려하는 총 금융부채 상환부담 평가지표 도입이 필요하여 신규 주택담보대출에 대해 DSR 지표를 통해 차주의 총 금융부채 상환부담을 평가"할 수 있다.

가이드라인에서 "주요내용은 차주의 채무상환능력을 정확히 평가하기 위해서는 실제 소득을 명확히 파악하기 위함이며, 대출시 차주의 원천징수영수증 등 객관성이 높은 ①증빙소득 등을 우선 활용하여 소득을 파악하고, 만일 증빙소득으로 확인이 어려운 경우, ②인정소득이나 ③신고소득을 활용하여 소득을 추정하지만 최저생계비 활용은 제한"[179] 한다. 가이드라인은 차입자가 제시하는 인정소득 정보에 근거하고 있다.

179) "① 증빙소득이란 원천징수영수증, 소득금액증명원 등 객관성 있는 자료로 입증된 소득이며, ② 인정소득이란 공공기관 등이 발급한 국민연금, 건강보험료 등을 바탕으로 추정한 소득이고, ③ 신고소득이란 신용카드(체크카드 포함) 사용액, 매출액 · 임대소득, 최저생계비 등으로 추정한 소득이다"

<표 33> 기존 DIT - 신규 DSR 지표간 비교

기존 DTI	신규 DSR*
주담대 원리금상환액 + **기타부채 이자상환액** / 연소득	주담대 원리금상환액 + 기타부채 원리금상환액 / 연소득

* DSR지표는 업권별·대출종류별 평균 만기와 금리 수준을 추정하여 전체 금융부채를 분할상환한다는 가정 하에 차주의 소득 대비 부담정도를 나타내는 지표

자료: 은행연합회. 2015. "앞으로 은행권 가계 주택담보대출에 대해 여신심사 선진화 가이드라인이 시행됩니다." 2015년 12월 14일 보도자료.

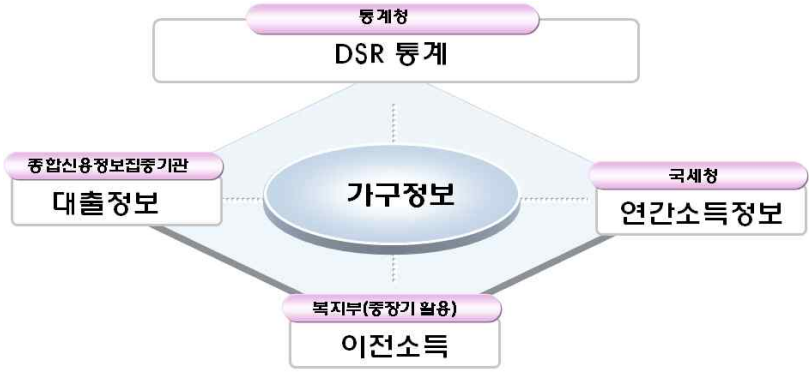
가계부채와 관련한 정보는 제2장에서 언급한 것처럼 저량측면의 재무정보 그리고 유량측면에서 재무정보로 구분할 수 있다. 본 연구에서 제안하는 신규통계는 유량측면에서 재무정보이다. 자산의 저량 측면에 정보에 대한 중장기적으로 추가적인 검토가 필요할 것으로 보인다. 현재 저량 측면에서 자산 그리고 순자산에 대한 정보는 체계적인 접근이 제한적이다. 저량 측면에서 부채정보는 이미 언급한 것처럼 신용정보 제공자 즉 금융기관의 자산정보를 활용하는 것이 적절하다. 현재 통계청에서 확보 및 활용 가능한 부채정보와 그 한계는 이미 언급하였다. 저량 측면에서 가계가 보유하고 있는 자산은 부동산 관련 자산의 비중이 높으며 관련정보는 국토교통부에서 체계적으로 관리하고 있다. 따라서 통계청과 국토교통부의 유기적인 협력관계 구축하는 것이 바람직할 것으로 보인다.

제2절 개발방향 및 기존 통계정보와 차별성

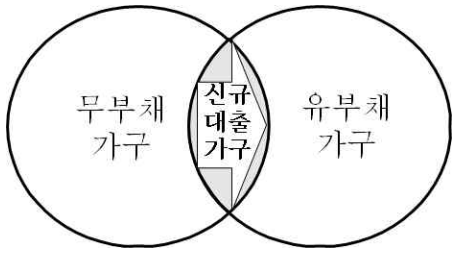
현재 통계청에서는 (국세청의 자료에 근거하여) 가구소득을 추정하는 작업을 진행하고 있는 것으로 알려져 있다. 은행권에서는 대출을 이용하는 고객의 증빙소득 혹은 금융기관의 인정소득에 근거한 정보의 제공여부에 따라 일부 소득정보만 보유하고 있을 수 있다. 그러나 통계청에서는 객관적 소득정보를 가구단위로 추정할 수 있는 것이다. 현재 진행되고 있는 소득정보의 가구단위화 일정을 참조할 수 있다. 그럼에도 불구하고 통계청에 보유하고 있는 가계(혹은 가구)와 관련된 정의에 대한 개선 보완작업도 중장기적으로 지속되어야 한다.[180]

180) 가구 및 주택에 대한 통계적 정의와 통계작성을 위한 코드작성은 통계청과 정부부처의 협력체계를 통하여 작성하여 제공되어야 할 것이다.

본 연구에서 제안하고 있는 정보는 한국신용정보원, 개인신용평가사 등 금융기관이 보유하고 있는 소득정보와 차별적이다. 특히 이들은 (통계청의 가구정보를 활용하지 않는 경우) 차주를 중심으로 소득과 부채정보를 파악할 수밖에 없는 한계가 있다. 국세청의 소득정보는 신규로 대출을 이용하고 있는 금융소비자 뿐만 아니라 기존 대출과 더불어, 부채가 없는 가구의 소득정보도 포함한다. 그리고 금융기관의 대출정보는 개인별 정보이다.



[그림 16] 통계청 가구단위 DSR 개발방안



[그림 17] 부채유형별 가구의 구성

통계청에서 생산 가능한 DSR의 경우 다른 기관들에서 생산 가능한 DSR과 세 가지 차별성이 존재한다. 첫 번째로 가계의 재무정보에 대한 추정이다. 기존 소득 혹은 부채정보는 개인 즉 차주중심으로 집계되고 있지만, 통계청의 가구정보를 활용하는 경우 가계부채 정보로 전환할 수 있다. 두 번째로 기존 가계부채 정보는 부채를 보유한 개인에 대한 정보만 존재하지만, 가계정보를 통하여 a)부채를 보유하고 있는 가계, b) 부채를 보유하고 있지 않은 가계, 그리고 c) 새롭게 부채를 보유하는 가계에 대한 정보로 구분할 수 있다. 세 번째로 통계청의 경우 국세청이 보유하고 있는 객관적 소득정보를 활용할 수 있다. 그러나 기존 정보는 인정소득 혹은 국민건강보험공단의 정보에 근거한 소득에 대한 간접추정이다. 기존의 개인중심의 DSR 혹은 신규로 대출을 이용하는 채무자의 DSR과 차별적으로 다양한 계층을 대상으로 가계의 DSR을 산정할 수 있는 장점이 있다.

DSR의 경우 가계부채와 관련된 유량정보이며, 저량정보에 대한 시장정보 생산도 가능할 것으로 예상된다. 단기적으로는 국세청 소득정보와 대출정보를 활용하여 Loan to Income Ratio의 개발이 가능할 것이다. 통계법에서는 종합신용정보집중기관으로부터 개인별 부채잔액 (Debt Balance) 정보를 제공받을 수 있다. 이를 가구정보를 활용하여 가계의 대출 잔액(Loan Balance)의 산출도 가능할 것으로 보인다.

주택금융시장에서 담보능력으로 간주되는 담보인정비율(Loan to Value Ratio)의 경우 이미 앞에서 언급한 것 같이 주택을 포함한 부동산 정보에 대한 접근성에 한계가 존재한다. 전체 가구를 대상으로 신규통계를 개발하는 경우 개인정보의 이슈와 통계작성의 절차와 시간이 과다할 수 있다. 따라서 우리나라 전체가구를 대표할 수 있는 표본을 대상으로 신규 통계를 생산하고, 중장기적 개선의 관점에서 생산된 통계를 개선하는 작업을 진행하도록 추진될 수 있다.

해외사례를 고려하는 경우 캐나다 통계청에서 DSR를 작성하고 있으며, 관련 자료는 1990년대부터 활용이 가능하다. 캐나다에서는 공급자 중심으로 DSR를 생산하고 있다. 거시적인 관점에서 금융기관의 정보를 기준으로 작성하고 있는 것으로 알려져 있다. 미국에서도 DSR(Debt Service Ratio)을 연방준비위원회(Federal Reserve Board)에서 Mortgage DSR과 Consumer DSR로 구분하여 공표하고 있다. 연방준비위원회에서는 DSR 계산을 위하여 이상적으로 가계정보를 활용해야 한다고 제안한다.[181] 미국에서는 현실적으로 확보하기 어려워 다양한 가정을 토대로 계산되고 있는 상황이다. DSR을 활용하기 위하여 거시자료(대출 규모 및 대출금리) 그리고 서베이 자료를 동시에 활용하여 추정되고 있는 것이다.[182] 우리나라에서도 거시자료에 근거하여 DSR를 계산할 수도 있다.

181) https://www.federalreserve.gov/releases/housedebt/about.htm

182) 미국에서 DSR은 PIRTOTAL = TPAY / INCOME로 추정되는 간접적 방식을 채택하고 있다. TPAY is the total payments on all debt. This includes payments on credit cards, mortgages, home equity loans, home equity lines of credit, other home improvement loans, loans for other residential real estate, vehicle loans, education loans, installment loans, margin loans, loans against the cash value of life insurance, pension loans and other miscellaneous loans. INCOME is the household's gross income for previous calendar year. Includes wages, self-employment and business income, taxable and tax-exempt interest, dividends, realized capital gains, food stamps and other support programs provided by the government, pension income and

그러나 거시자료에 근거한 정보는 이미 앞에서 언급한 바와 같이 가계부채의 다양성을 고려하지 못하는 한계에 노출될 수 있다.

제3절 연관 기관 및 계획(안)

기존 통계생산 노하우는 신규 통계생산에 적절히 활용될 수 있으며, 신규 통계생산을 위한 각종 위험을 제거하고 생산기간을 단축할 수 있을 것으로 기대된다. 국세청 소득정보와 더불어 복건복지부의 이전소득을 활용하여 가계의 정확한 소득을 산출할 수 있다. 이는 중장기적으로 소외계층 등에 대한 DSR 정보를 산출하는데 활용될 수 있다.[183] 이전소득의 경우 복지정책의 수혜가구를 대상으로 진행되고 있기 때문이다. 기타부채(예를 들어 자동차·선박 등의 담보대출 등)에 대한 현황이 파악하기 어려운 경우 주택담보대출을 중심으로 DTI 혹은 DSR을 일차적으로 생산할 수 있다.[184] 동시에 잔액정보를 활용하여 LTI의 생산이 가능할 것으로 예상된다. 이와 함께 통계청에서는 통계적 의미에서 DTI 및 DSR에 대한 정의를 개선함과 더불어 가계단위의 자산에 대해 중장기적으로 정비할 필요가 있다.

한국주택금융공사에서는 “Debt To Income 수준별 대출금액 및 잔액현황”을 공표하고 있다. 그러나 동 통계에서는 공사에서 취급한 담보대출을 대상으로 DTI 관련정보를 계산하고 있지만, 가계부채 시장을 보여주는데 제한적이다.

withdrawals from retirement accounts, Social Security income, alimony and other support payments, and miscellaneous sources of income http://www.bankofcanada.ca/wp-content/uploads/2010/02/wp08-46.pdf 참조

183) 신규 부채보유 가구의 정보를 생산하는 경우, 정치권 혹은 감독기관에서 대출심사에서 국세청 정보를 직접적으로 활용하는 제안이 존재할 수 있다. 그리고 부채정보 보유기관에서 관리 유지하고 있는 부채정보가 신규 부채를 시의적으로 반영하고 있는지에 대한 고민도 필요하다.

184) 중도상환 및 일시상환에 따라 대출상환이 소득과 비교하여 높은 수준에서 발생할 가능성을 배제할 수 없음으로 적정한 임계치(threshold)를 기준을 설정해야 할 것이다.

6. 소결

최근 가계부채에 대한 이슈가 부각됨에 따라 관련 통계정보의 생산, 축적, 활용, 분석 및 개선 혹은 관리 및 기존 통계의 개선 그리고 신규통계의 개발에 대한 관심이 증가하고 있다. 현재 우리나라에서 가계부채 현황파악은 금융안정성 측면에 초점을 두고 있음으로 시장정보에 조사 측면에서 개선의 여지가 있다. 그러나 기존 가계부채관련 시장정보는 대부분 금융기관 중심의 가계부채에 대한 정보이므로 우리나라 가계부채를 과소추정하고 있을 가능성이 있다. 가계부채의 규모는 한국은행에서 발표하는 공식적 금융기관이 보유하고 있는 자산(가계의 입장에서 부채)과 더불어 비공식적 부채를 포함하는 광의의 개념으로 볼 수 있다. 실제로 한 조사에 근거하면 부채구분을 부채를 이용한 기관에 따라 공식 금융기관 이외에 a. 비금융기관 부채, b. 개인적으로 빌린 돈, c. 계 그리고 d. 보증금 등을 포함하기도 한다. 따라서 통계는 금융기관의 자산정보라 볼 수 있다. 우리나라의 가계부채의 실질적 규모를 가계의 입장에서 추정하는 작업이 진행될 여지가 있으며, 금융기관의 자산에 대한 조사가 아니라 가계의 부채에 대한 조사가 진행되어야 함을 의미한다.[185]

기존연구에서는 금융시장 및 거시경제의 안정성 강화의 측면에서 금융기관의 자산정보의 개선과 보완을 제안하고 있다. 그러나 본 연구는 가계부채와 관련하여 미시적 관점에서 가계의 재무상황을 판단할 수 있는 가계재무관련 통계정보의 개선과 보완을 논의하였다. 본 연구의 한계는

- 첫 번째로 금융시장과 거시경제의 안정성 측면에서 가계부채를 바라보는 기존 관점과 다소 차별적일 수 있다. 그러나 기존 관점의 중요성을 간과하는 것이 아님을 밝히고자 한다. 즉 금융기관(혹은 금융시장) 안정성 강화와 가계의 안전성 강화의 적절한 조화를 강조하고 있다.
- 두 번째 한계로 본 연구에서 제시하고 있는 가계부채관련 통계는 가계의 유량중심의 통계이며, 저량중심의 통계에 대한 보완과 개선이 필요하다. 그러나 이는 단기적인 목표달성 보다는 장기적인 체제개선에 초점을 맞추어야 할 것이다.

가계수준의 정보는 다양한 유형의 정책대상자(즉 저소득층, 고령층, 주거복지

185) 그러나 금융기관의 경우 감독기관 등에게 개별대출에 대한 상세 정보를 제공하는 것에 대하여 보수적인 입장을 갖고 있다. 금융기관들이 자신의 영업정보를 제공하기 위하여서는 두 가지 방안이 존재한다. 첫 번째는 관련 법률에 근거하여 의무적으로 제공해야 하며, 두 번째는 금융기관의 영업에 활용되어 금융기관에게 도움이 될 수 있어야 한다.

수혜대상자 등)에 대한 분석의 한계를 극복할 필요가 있다. 가계부채와 관련된 정책수립과 관련된 그리고 금융정책 수혜대상 계층이 다양하게 존재한다. 금융정책 측면에서 뿐만 아니라 복지정책, 주거정책 등 다양한 분야 그리고 다양한 계층에서 가계부채 관련 정보에 보강의 필요성이 제기되고 있는 상황이다. 이와 같은 상황에서 감독기관에서는 입수가능성을 해결할 수 있지만, 통계작성에 따른 이해관계의 이슈를 적절하게 해결하기 어려울 수 있다. 금융기관의 입장에서는 감독기관에게 영업과 관련된 정보제공에 대하여 보수적인 견해를 취할 수 있다. 금융기관 안정성의 관점은 시장자료 분석이 용이할 수 있지만, 가계 안전성과 다양한 정책적 관점에서 미시자료에 대한 분석이 요구되고 있다. 따라서 미시적인 가계의 관점에서 가계 부채에 대한 정보의 관리가 필요하다. 가계부채관련 통계생산을 위하여 “기초통계 입수 가능성, 통계작성 비용대비 활용성의 지속성, 통계품질의 유지, 통계작성의 이해관계 등(김수영, 2016)”을 고려할 필요가 있다.

금융관련 정부정책을 평가하기 위하여 (정책성) 공공기관 대출과 일반 시중 금융기관 대출에 대한 분류체계가 요구된다. 기존 통계조사들에 대한 검토에 따르면, 우리나라에서는 가계부채 정보는 공공과 민간의 구분이 모호하다. 따라서 이를 정확히 분류하는 체계의 개선이 요구된다. 공공성격이 있는 금융기관(예를 들어 주택도시기금, 주택금융공사 등) 에서 대출을 받은 가계의 경우는 엄밀한 범위에서 금융기관에서 대출을 받았지만, 정부지원으로 민간에서 대출을 받았으므로 이에 대한 새로운 분류체계가 필요하다. 이를 통하여 정부에서 보유하고 있는 금융정책에 대한 수단(tool)에 대한 관리와 평가에 기초자료를 확보할 수 있을 것이다.

시장요구의 증가로 최근에는 상대적으로 정확하게 파악하고 있음으로, 가계부채 관련 정보의 축적 혹은 보완은 장기적인 호흡으로 준비하는 것이 필요하다. 예를 들어 우리나라에서 금융기관이 보유하고 있는 가계부채의 경우에도 10년 전에는 그 규모를 파악하는 것도 어려웠다. 하지만 2016년 2월부터 감독기관에서 기업대출의 일종인 상업용 부동산 담보대출의 규모를 조사하고 있다. 따라서 가계부채 사례를 고려하는 경우 이와 관련된 시장정보 축적도 장기간이 소요될 것으로 전망된다. 동시에 가계부채 관련 신규 통계의 개발이 이루어지는 경우 단기적으로 완전한(complete) 시장정보를 보다는 중장기적으로 완전한 시장정보를 구축하는 방향이 바람직하다. 2013년 5월에 개정된 주택담보대출의 리스크 관리기준에 근거하면 “주택담보대출이라 함은 은행이 주택을 담보로 취급하는 가계대출(자산유동화된 대출을 포함한다)을 말하며, 분양 주택에 대한 중도금대출 및 잔금대출 그리고 재건축・재개발 주택에 대한 이주비대출,

추가분담금에 대한 중도금대출 및 잔금대출”으로 개선하였다. 동 기준은 2007년에 처음 도입되었으며, 최근 이주비대출, 추가분담금 등을 추가로 고려하기 시작하였다. 주택담보대출에 대한 정의는 지속적으로 개선되고 있는 상황이다. 다른 사례로는 과거와 달리 주택도시기금에서 개인에게 대출을 하고 있는 실적과 주택금융공사에서 취급하고 있는 보금자리론 통계는 한국은행의 가계신용에 포함되고 있다.[186] 이들은 과거 가계신용에 포함되지 않았으며, 시장요구에 따라 주택담보대출 관련 정의에 대한 보완이 이루어진 상황이다. 기존 사례에 대한 분석을 통하여 가계부채 관련 통계를 개발하는 경우 중장기적인 보완과 개선을 전제로 작업을 추진하는 것이 바람직하다.

향후 연구과제에 대하여 살펴보면,

- 우선적으로 현재의 가계부채와 관련된 통계는 대부분 재고와 관련된 정보이며, 가계중심의 유량에 대한 정보가 보강되어야 한다. 동시에 가계의 안정성 측면에서 시장정보를 구축하고자 하는 경우 대출기준(loan-level)으로 자료를 작성하는 것이 필요하다. 장기적으로 가계가 보유하고 있는 대출의 생애주기(loan life cycle)와 관련된 통계(예를 들어 조기상환, 부분상환, 연체, 부실 등)정보에 대한 축적이 바람직할 것이다.
- 두 번째로 주관적 지표와 객관적 지표의 차이를 보완할 수 있는 방안에 대한 모색이 필요하며, 객관적 지표가 통계작성에 도움을 줄 수 있다는 연구에 대한 진행이 필요하다. 기존 가계의 현황은 조사 및 패널의 미시자료를 활용하고 있어 가계의 재무상황에 대한 적절한 분석에 어려움이 있다(유승동 2014). 실제로 대부분의 조사 혹은 패널의 경우 가계부채의 가격 즉 대출금리에 대한 정보가 보강될 필요가 있다. 대출특성과 관련 정보가 부재하여 가계부채의 현황진단이 어렵다. 동시에 존재한다고 하더라도 제한적 합리성을 보유한 가계의 정보의 경우 그 신뢰성이 높지 않다. 따라서 설문조사에 따른 정보와 실제 객관적 정보의 차이에 대한 검토를 통한 보완이 필요한 것이다. 가계의 재무상황을 파악할 수 있는 행정자료(사례 소득정보)를 활용하여 가계의 정확한 재무상황에 대한 조사가 진행될 필요가 있다. 객관적 지표를 사용하여 조사 혹은 설문이 이루어지는 경우, 중장기적으로 비용절감의 효율성 증진에 기여할 것으로 기대된다.
- 세 번째로 현재 가계부채에 대한 정보는 금융기관 중심의 자산분류 체계가 활용되고 있어, 가계중심의 부채에 대한 체계가 필요하다. 개인 혹은 가구의 의사결정 측면에서는 금융기관 중심에 분류체계와 더불어 사용 용도에 대한 고민 즉 가계입장에서 분류가 필요한 것이다(You, 2016). 비금융기관에서

186) 참고로 현재 국민주택기금 통계의 경우 주택도시기금 통계로 변경승인이 필요하다.

활용하고 있는 부채에 대한 정확한 분류가 필요할 수 있다(일반사채, 카드 빚, 전세보증금, 외상, 미리탄 곗돈, 기타 부채로 분류하고 있는 현재 분류 체계를 가계부채 규모와 목적에 따라 재분류하는 방안을 고민할 수 있음). 예를 들어 주택을 구입용도로 대출을 받은 가계와 투자용도로 대출을 받은 가계의 성과는 동일하지 않다. 단순히 주택을 구입하여 주택담보대출을 이용할 뿐만 아니라, 생활비 목적으로 주택담보대출을 받는 경우도 존재하기 때문이다.

- 네 번째로 단기적으로 각종 조사 및 패널간의 가계부채에 대한 정보의 일관성을 보완하는 작업이 진행될 수 있다. 가계부채에서 중요성을 인식하고 있지만, 이에 대하여 우선적으로 고려하는 요인이 차등적일 수 있다. 이와 같은 가계부채 정보에 대하여 조사 간 일관성과 차별성에 대한 보다 세밀한 연구가 요구된다. 전반적 일관성을 강화할 수 있는 방안을 기반으로 조사별 차등성에 대한 고민이 필요(사례 노령가계를 중심으로 조사가 진행되는 경우 보증의 이슈가 부각될 수 있음)하다. 예를 들어 전문적인 분야에 대한 조사 혹은 패널의 지침 등이 제공되어 조사 간 호환성을 확보할 수 있을 것으로 전망된다.
- 다섯 번째로 일반 조사의 형태로 가계부채 관련 정보가 축적되는 경우 이에 대한 전문성 강화가 요구된다. 일부 조사의 경우 통계 혹은 가계부채에 대한 전문성 보강이 필요하다. 일관성 하락이 우려되고 있으며, 시장을 왜곡할 가능성도 상존하고 있다. 왜곡된 시장정보에 근거하여 가계부채 정책이 수립될 가능성을 배제할 수 없다. 조사기관 담당자들의 전문성에 대한 지원과 더불어 안정적 조사가 진행될 수 있는 방안에 대한 고민이 필요하다.
- 마지막으로 이미 언급한 것처럼 자산과 관련된 정보에 대한 공유와 활용에 대한 지속적 고민과 개선작업이 병행되어야 한다. 그리고 이와 같은 작업은 정부 부처 간 지속적 업무협의 체계구축을 통하여 진행되는 것이 바람직하다. 동시에 자산관련 정보구축 체계 마련에 대한 중장기적 고려가 요구된다.

7. 신규과제의 법적 검토

본 보고서에서 신규과제로 제안하는 가계부채 관련 신규통계 개발 방안에 대한 헌법적 정당성 심사를 간략하게나마 수행해 보고자 한다.

본 장에서 이미 자세히 설시한 바 있듯이 최근 가계부채에 대한 이슈가 부각됨에 따라 관련 통계정보의 생산, 축적, 활용, 분석 및 개선 혹은 관리 및 기존 통계의 개선 그리고 신규통계의 개발에 대한 관심이 증가하고 있다. 이에 이미 2015년 12월부터 금융위원회와 은행연합회를 중심으로 여신심사 선진화 가이드라인을 마련 총부채원리금상환부담(Debt Service Ratio, 이하 DSR)을 도입하기 위한 노력을 경주하고 있으나, 기본적으로 차입자(借入者) 중심으로 파악되는 소득과 부채정보에 기반을 둔 동 DSR로는 완전한 가계부채의 규모를 파악하기가 곤란한 실정이다.

이에 본보고서는 통계청에서 보유하고 있는 가구에 속한 가구원 정보를 바탕으로 국세청 및 보건복지부의 소득정보, 그리고 신용정보회사의 부채정보를 활용하여 가구단위의 DSR 통계를 개발할 것을 제안하고자 하는 것이다. 이는 종합신용정보집중기관의 대출정보와, 국세청의 연간소득정보에 중장기적으로 복지부의 이전소득에 관한 내용까지 포함하여 산출해 낼 수 있는 가계의 정확한 소득을 바탕으로, 기존의 차주(借主) 중심이 아닌 가계 중심의 가계부채정보를 생산해 냄으로써 가계부채의 보다 실질적인 규모를 추정해 낼 뿐만 아니라 가계부채의 다양성을 고려할 수 있게 해 준다. 구체적으로 금융기관 이외에 비금융기관 부채나 개인 간 거래에 의한 부채 등의 다양한 부채양상을 고려할 수 있게끔 해 주고, 부채를 보유하고 있는 가계뿐만 아니라 부채를 보유하고 있지 않거나 혹은 새롭게 부채를 보유하게 되는 가계에 대한 정보까지도 파악이 가능하며, 국세청이 보유하고 있는 객관적 소득정보는 보다 정확한 DSR의 생산을 가능하게 해 줄 것으로 기대해 볼 수 있다.

이러한 측면은 가계부채 관리의 국가경제 중요성에 비추어 볼 때, 본 연구가 제안하는 통계청 중심의 DSR 통계 생산의 목적의 정당성은 특히 공공복리의 측면에서 정당화 가능하며, 경우에 따라서는 사회·경제적 질서유지의 차원에서도 정당화 가능할 것으로 판단된다.

다만 이러한 통계작성 과정에서 민간 금융기관으로부터 부채관련 정보를 입수해야 할 뿐만 아니라, 국가기관인 국세청 등으로부터 관련 정보를 입수해야 하는 문제점이 존재하는데, 현행 관련 법제의 측면에서 살펴볼 때 현행 「신용정보법」[187]이나

187) 「신용정보법」 제23조(공공기관에 대한 신용정보의 제공 요청 등) ⑦ 신용정보회사등은 공공기관의 장이 관계 법령에서 정하는 공무상 목적으로 이용하기 위하여 신용정보의 제공을 문서로 요청한 경우에는 그 신용정보를 제공할 수 있다.

「국세기본법」[188]의 경우 국가기관인 통계청이 문서로써 요구할 경우에 신용정보 및 과세정보를 통계작성에 적법하게 활용할 수 있도록 규정이 이미 마련되어 있다고 볼 수 있을 것이다.

다만 이미 개인정보 관련 부분에서 서술한 바 있듯이, 개인정보 보호론의 측면에서 현행 법제에 대한 위헌성이 제기될 가능성이 없지 않으므로 통계작성의 제 측면에서 엄격한 헌법적 정당성의 준수가 필요할 것이다. 이와 관련하여 앞에서 제시한 헌법적 정당성 심사 체크리스트를 본 제안과제에 적용해 보면 다음과 같이 정리해 볼 수 있을 것이다.

188) 「국세기본법」 제81조의13(비밀 유지) ① 세무공무원은 납세자가 세법에서 정한 납세의무를 이행하기 위하여 제출한 자료나 국세의 부과·징수를 위하여 업무상 취득한 자료 등(이하 "과세정보"라 한다)을 타인에게 제공 또는 누설하거나 목적 외의 용도로 사용해서는 아니 된다. 다만, 다음 각 호의 어느 하나에 해당하는 경우에는 그 사용 목적에 맞는 범위에서 납세자의 과세정보를 제공할 수 있다. <개정 2014.1.1.>
1. 지방자치단체 등이 법률에서 정하는 조세의 부과·징수 등을 위하여 사용할 목적으로 과세정보를 요구하는 경우
2. ~4. (중략)
5. 통계청장이 국가통계작성 목적으로 과세정보를 요구하는 경우
6. (하략)
② 제1항제1호·제2호 및 제5호부터 제8호까지의 규정에 따라 과세정보의 제공을 요구하는 자는 **문서로** 해당 세무관서의 장에게 요구하여야 한다.
③ 세무공무원은 제1항 및 제2항을 위반하여 과세정보의 제공을 요구받으면 그 요구를 거부하여야 한다.
④ 제1항에 따라 과세정보를 알게 된 사람은 이를 타인에게 제공 또는 누설하거나 그 목적 외의 용도로 사용해서는 아니 된다.
⑤ 이 조에 따라 과세정보를 제공받아 알게 된 사람 중 공무원이 아닌 사람은 「형법」이나 그 밖의 법률에 따른 벌칙을 적용할 때에는 공무원으로 본다.

대범주	소범주	내용	비교
목적의 정당성	통계작성의 규범적 근거	• 적극적 근거: 신규통계로서 통계작성 근거가 될만한 기존 규범은 존재하지 않음 - 「통계법」 제17조에 의한 지정통계의 지정을 받거나, 혹은 동법 제18조에 의한 통계작성의 승인이 필요할 것 • 소극적 근거: 「신용정보법」 제23조, 「국세기본법」 제81조의13	법률유보
	통계작성의 구체적 목적/필요성	• 현재 국내의 가계부채 관련 통계는 가계금융복지조사를 제외하고는 가계(혹은 가구) 수준의 정보가 전무한 상황 • 특히 현재의 우리나라의 가계부채 현황파악은 금융안정성 측면에 초점을 맞춘 까닭에 기존 가계부채관련 시장정보는 대부분 금융기관 중심의 가계부채에 대한 정보이므로 우리나라 가계부채를 과소추정하고 있을 가능성이 있음 • 가계부채의 규모는 공식적 금융기관이 보유하고 있는 자산(가계의 입장에서 부채)과 더불어 비공식적 부채를 포함하는 광의의 개념으로 볼 수 있어, 미시적인 가계의 관점에서 가계 부채에 대한 정보의 관리가 필요 / 가계수준의 정보는 다양한 유형의 정책대상자(즉 저소득층, 고령층, 주거복지 수혜대상자 등)에 대한 분석의 한계를 극복할 필요가 있음 / 금융정책 측면에서 뿐만 아니라 복지정책, 주거정책 등 다양한 분야 그리고 다양한 계층에서 가계부채 관련 정보에 보강의 필요성이 제기되고 있는 상황	
	통계작성 목적의 범주	국가안전보장( ) / 질서유지(○) / 공공복리(◎)	
	기대효과	• 정부에서 보유하고 있는 금융정책에 대한 수단에 대한 관리와 평가에 기초자료를 확보	
수단의 적절성	적용가능한 방법론	• 통계청에서 보유하고 있는 가구에 속한 가구원 정보를 바탕으로 국세청 등의 정부부처에서 제공한 소득정보, 민간기관의 부채정보를 활용하여 가구의 총부채원리금상환부담 통계의 개발 • *기타 적용가능한 방법론 추가*	
	적용예정 통계작성 방법	• 생산주체: 통계청 • 주요내용: 총부채원리금상환부담(Debt Service Ratio; DSR) 생산 • *세부적인 방법론 요약*	
	적용예정방법의 효과성/효율성	• 가계의 재무정보에 대한 추정: 기존 소득 혹은 부채정보는 개인 즉 차주중심으로 집계되고 있지만, 통계청의 가구정보를 활용하는 경우 가계부채 정보로 전환가능 • 기존 가계부채 정보는 부채를 보유한 개인에 대한 정보만 존재하지만, 가계정보를 통하여 부채를 보유하고 있는 가계, 부채를 보유하고 있지 않은 가계, 그리고 새롭게 부채를 보유하는 가계에 대한 정보로 구분가능 • 통계청의 경우 객관적 소득정보를 활용할 수 있음	
	평가 및 검증방법	추후 보완	
침해의 최소성	비식별화 수준 및 방법	현 기술 수준내에서 선택가능한 최적의 방법 제시	개인정보 침해의 최소화 대책
	재식별기법	상동	
	비밀 및 개인정보 보호의무 숙지여부	법률의 규정에 따른 엄격한 준수	
	시스템 보안수준	현 기술 수준내에서 선택가능한 최적의 방법 제시	
	시스템 안정성 수준	상동	
법익의 균형성	예상되는 침해수준	개인정보로 파악가능한 통계청에서 보유하고 있는 가구에 속한 가구원 정보, 국세청 및 보건복지부의 소득정보, 그리고 신용정보회사의 부채정보 등의 일정정도의 침해가 예상됨	
	기대효과와의 비교	추정 공익과 추정 비용의 비교형량	

<표 34> 가계부채 관련 현황조사의 합헌성 심사

참 고 문 헌

국토교통부. 2015. 부동산 정책지원 통계 발굴 및 인프라 강화방안 연구.

김다스라, 2011, 주택구입능력지수의 해외사례 분석, 주택금융월보, 한국주택금융공사.

김수영. 2016. 주택통계 개선을 위한 심포지엄 토론자료. 한국주택학회.

방송희, 2015. 부동산 가격공시제도를 둘러싼 이슈, HF-이슈 리프트, 2015-10.

방송희, 2016. 주택금융통계 현황과 개선방안, 한국주택학회 심포지엄자료집, 한국주택학회.

변종석·이석훈·정구현. 2013. 가계금융 ·복지조사의 무응답 처리를 위한 유용한 보조정보 선정, 조사연구 14(1), pp. 69-91.

유승동. 2013. 주택재고를 고려한 주택구입능력 분석의 재고찰, 국토계획, 48(5), pp. 165-176.

유승동. 2014. 주택금융시장의 역할 정립과 발전방안, 국토 395, pp. 34-39.

유승동. 2015. 주택금융시장의 소비자 안심방안, 국토 404: pp. 18-23.

이용만, 2008. 국민은행 주택가격지수의 평활화 현상에 관한 연구, 주택연구 16(4), 27~47.

전국은행연합회, 2015.12, 여신심사 선진화 가이드라인 보도자료.

정구현, 2014. 가계금융 복지조사 발전방안, 국가통계발전포럼 발표자료

통계청, 2015 a. 『한국노동패널조사』 통계정보 보고서.

통계청, 2015 b. 『주택금융및보금자리론수요실태조사』 2015년 정기통계품질진단 결과보고서.

한국노동연구원, 2015, 한국노동패널 1~17차년도 조사자료_유저가이드.

한국보건사회연구원, 2013. 2013 한국복지패널 기초분석 보고서.

한국정보화진흥원. 2014. 빅데이터 활용 국민체감형 통계생산체계 구축방안 수립.

한국조사연구학회, 2016, 빅데이터를 활용통계생산방법론 연구용역, 중간보고.

You, S. D. 2009. H*ousing Finance Mechanisms in the Republic of Korea,* Nairobi, Kenya. UN-HABITAT.

You, S. D. 2016. On A Subtle Balance between Micro- and Macro-Economics Perspectives in the Housing Finance Market, *Housing Finance International Autumn,* pp. 20-24.

고용조사:고령화연구패널조사 survey.keis.or.kr

미국연방준위원회 www.federalreserve.gov

금융감독원 www.fss.or.kr

금융위원회 www.fsc.go.kr

전국은행연합회 www.kfb.or.kr

주거실태조사정보제공시스템 www.hnuri.go.kr

캐나다 중앙은행 www.bankofcanada.ca

통계청 www.kostat.go.kr

한국노동패널조사 www.kli.re.kr/klips

한국은행 www.bok.or.kr

Ⅴ. 빅데이터 활용 국내 사례

서우석(서울시립대학교 도시사회학과)

Ⅴ. 빅데이터 활용 국내 사례

1. 서론

빅데이터에 대한 관심이 늘어남과 동시에 공공분야에서 빅데이터를 활용한 통계 생산의 사례들도 늘어나고 있다. 이러한 변화에는 빅데이터 활용의 제고를 위한 정부의 정책 의지가 크게 작용하고 있다. 정부는 빅데이터 활용을 위한 관계기관 간 협업 활성화 및 공개된 인터넷 데이터의 수집 활용의 근거 확보 등을 위해 2014년 11월 전자정부법을 개정하였으며, 빅데이터 공통기반 플랫폼을 구성하고, 빅데이터 주요 추진기관 협의체를 구성하여 부처 실무자간 협업체계를 마련하였다[189]. 또한 공공데이터 개방 발전을 위해 2013년 공공데이터 전략위원회를 출범하여 공공데이터법을 근거로 공공데이터 개방을 추진한 결과, 공공데이터 개방 건수가 2015년 1만 2889개로 OECD 국가 중 공공데이터 개방 1위를 차지하였다[190]. 또한 빅데이터를 위한 공공서비스 개발의 우수 사례 발굴을 통해 이를 확산시키려는 정부 부처와 지자체의 경쟁적인 사업 전개가 이어지면서 빅데이터를 활용한 통계 생산과 공공 서비스 개발이 빠르게 증가하고 있다.

빅데이터를 활용한 공공서비스 개발에서 빅데이터 기반의 통계가 핵심적인 역할을 하면서도 빅데이터를 이용하여 생산한 통계가 가지는 중요성은 과소평가되고 있다. 빅데이터를 이용한 공공서비스를 개발하는 경우 빅데이터 기반 통계가 공표되지 않는 경우들이 대부분이기 때문이다. 이 경우 빅데이터 기반 통계가 현재의 통계품질관리 제도로부터 벗어나 있다. 하지만 그렇다고 통계의 정확성이 가지는 중요도가 줄어드는 것은 아니다. 빅데이터 기반 통계의 정확성이 뒷받침되지 않을 때 빅데이터를 활용한 공공서비스의 적합성을 확보할 수 없기 때문이다.

본 연구에서는 정부와 지자체의 빅데이터를 활용한 통계 생산의 주요 사례들을 소개하였다. 정부의 사례에서는 빅데이터 활용의 경험 축적이 상대적으로 많은 교통분야와 보건·의료 분야에 초점을 맞추었고, 지자체의 사례에서는 서울과 부산의 주요 사례들을 다루었다. 또한 이러한 공공 분야의 빅데이터 통계 생산에 주로 활용되는 민간의 빅데이터 통계 작성 현황도 소개하였다. 끝으로 이와 같은 현황과 가용한 빅데이터의 종류, 기대효과를 종합적으로 고려하여 신규사례를 제시하였다.

189) 이지영. 2015. 빅데이터의 국가통계 활용을 위한 기초연구. 통계개발원.
190) 한국정보화진흥원. 2015. 2015년 빅데이터산업 10대 뉴스 및 이슈.

2. 현황

제1절. 정부의 빅데이터 활용 통계생산 사례

가. 교통분야 빅데이터 활용

1) 개요

교통분야 빅데이터 활용은 Intelligent Transport Systems(ITS, 지능형교통체계) 추진과 관련하여 비교적 활발하게 추진되어 왔다. ITS는 교통수단 및 교통시설에 전자 제어 및 통신 등 첨단기술을 접목하여 교통정보 및 서비스를 제공함으로써 교통체계의 운영 및 관리를 과학화하고 교통의 효율성과 안정성을 향상시키는 교통체계를 말한다. 교통분야 빅데이터 활용은 개별 사안들에서의 경험 축적을 바탕으로 전반적인 관리 단계에 들어서는 것으로 평가되며 이 과정에서 제시되는 논의들은 다양한 분야 빅데이터 활용의 체계적 관리를 위한 기준 마련에 주요한 선도 사례로서 평가될 만하다.

주요 추진 주체로는 한국교통연구원의 국가교통DB센터를 꼽을 수 있다. 국가교통DB센터는 전국 속도 및 교통량 자료, 수도권 교통카드 및 택시 자료, 영상자료, 교통사고 자료 등의 다양한 교통자료를 수집하는 시스템을 구축하여 기초통계와 응용통계 작업을 수행하고 있다. 최근 지방정부 차원에서도 독자적인 노력을 통해 교통분야 빅데이터를 활용하는 시스템을 구축하고 있다. 대전시가 구축한 대전 교통 데이터웨어하우스가 그 사례이다.

교통분야에서는 다양한 빅데이터가 활용되고 있다. 최근의 이용도 면에서 볼 때 특히 교통카드 데이터 사용 가능성이 주목을 많이 받고 있다.

2) 교통분야 빅데이터 활용 현황

교통 분야 활용되는 빅데이터는 다음과 같은 세 분야로 나눌 수 있다[191]. 첫째, GPS 기반 정보이다. 자가용 위치 정보 (내비게이션), 일반택시 위치정보 (카드결제시스템), 화물차 위치 정보, 휴대폰 위치 정보 등이다. 둘째, 과금 기반 위치 정보로서 교통카드 및 신용카드의 사용정보가 이에 해당한다. 셋째, ITS 센터에서 차량검지기, 영상, DSRC 등을 통해 수집하는 교통정보와 교통사고 정보 등이 있다.

191) 이석주. 2012. 교통 부문에서의 빅데이터 현황 및 활용.

최근에는 좀 더 포괄적이면서도 세분화된 기준을 통해 교통분야에서 활용되는 빅데이터가 제시되었다[192].

<표 35> 교통분야 빅데이터 현황

구분	데이터	수집기관	현재 활용도
공공부문 교통량 속도 돌발상황	교통량, 속도	한국도로공사, 건설기술연구원, 국토관리청, 지자체	보통
	속도, 돌발상황	도로교통공단, 경찰청, 지자체	보통
개인 통행 및 이력	민간 내비게이션 소통정보	현대엠엔소프트, 팅크웨어, Tmap, 김기사 등 민간네비게이션 회사	높음
	민간 핸드폰 통신	SKT, KT, LG	낮음
	운행기록 분석시스템	교통안전공단, 국토교통부	낮음
	자동차 등록관리 시스템	교통안전공단, 국토교통부	낮음
대중교통	교통카드 데이터	한국스마트카드, 유페이먼트, 한페이시스, 카드회사, 은행 등	높음
	통합대중교통데이터 (TAGO)	지자체, 전국고속버스운송사업 조합, 전국 터미널 협회, 인천국제공항공사, 한국철도공사, 코레일 공항 철도, 서울지방항공청, 한국해운조합	낮음
네트워크 및 공간정보	교통망	민간내비게이션 회사, 대중교통관련 기관, KTDB	높음
	POI	국토교통부, 국토지리정보원, 민간 네비게이션 회사, 포털 지도 사이트	높음
	국토공간정보 유통시스템	국토교통부, 국토지리정보원, 한국해양개발, 민간기업	낮음
	집계구 및 기상 데이터	통계청, 기상청	보통
화물데이터	글로벌화물추적시스템	해양수산부	낮음
	화물수송 실적 관리시스템	국토교통부	낮음

교통분야 빅데이터의 한계를 수집, 구축, 유통 단계로 나누어 볼 수 있다. 먼저 수집단계에서는 다음과 같은 한계점들이 있다.

첫째, 데이터 수집범위가 한정적이다. 수집범위가 일부 지역, 시간대, 대상으로 한정되어 범용적 활용이 어렵다.

둘째, 동일 속성 데이터의 수집방법 및 형태가 상이다. 데이터의 일관성을 확보할 수 없어 통합 활용까지 많은 시간과 노력이 필요하다.

셋째, 외부기관에서는 필요 항목이지만, 수집기관에서는 불필요한 항목으로

192) 천승훈 외. 2015. 월간교통 2015.7, 49-57.

간주하여 수집하지 않음으로써 해당 데이터의 활용에 한계가 발생한다.

넷째, 수집단계에서 오류가 발생할 수밖에 없는 데이터인 경우이다. GPS 데이터의 경우 누락되거나 오류가 발생할 수밖에 없는데 이를 보정하기 위해서는 상당한 기술력과 노하우가 필요하다.

데이터 구축과정에서는 수집과정과 관련된 한계들이 발생한다. 커버리지 부족, 데이터 구축 표준지침의 부재, 데이터의 품질관리 문제, 데이터 통합관리 부재와 이력데이터 관리 미흡 등의 한계가 존재한다.

데이터 유통단계에서는 데이터 수집기관 간의 이해관계로 데이터 공유가 잘 이루어지지 않는다. 민간의 데이터에 대한 접근 곤란 또한 어려운 문제이다.

이상의 한계를 극복하기 위한 과제들이 다음과 같이 제시되었다.

- 데이터의 수집범위 확대
- 수집범위 한계 보완을 위한 빅데이터 간의 융복합 활용
- 데이터 형태 표준화 기준 및 지침 마련
- 데이터 품질 기준 및 품질 관리 방안 마련
- 각 데이터 간 상호검증방안 수립
- 각종 사회지표와의 비교를 통한 간접적 검증방안 마련
- 데이터 공유를 위한 민간 공공 협력체계 구축(MOU, 공동연구 등)과 법적 제도적 창치 도입
- 데이터 처리 분석 전문인력 양성
- 분야별 전문가의 융복합 연구 환경 조성
- 통합 DB 관리 시스템 구축

3) 교통카드 데이터 활용

교통안전공단이 국토교통부의 위탁을 받아 2006년부터 대중교통 현황조사를 수행하고 있다. 초기의 대중교통 현황조사는 조사원이 직접 대중교통에 탑승하여 수행하는 탑승조사가 샘플조사로 수행되는 방식이었으며 많은 수의 조사원이 필요하기 때문에 비용이 많이 들고 시간의 제약이 컸다[193].

교통카드 이용자 증가에 따라 교통카드 데이터를 분석하여 대중교통 현황자료를 수집하는 방안을 강구하게 되었다. 서울시의 경우 현재 교통카드 이용률이 99%에 달한다.

193) 강희찬. 2015. 공공을 위한 Real 빅데이터: 교통카드 데이터. The Magazine of the Korean Society of Civil Engineers, 63(6).

교통카드 데이터 이용의 장점은 다음과 같다.

첫째, 교통카드 이용자수 증가에 따라 많은 샘플의 활용이 가능하다.

둘째, 이용한 교통수단, 승하차정류장, 승하차시간, 이용요금, 환승여부 등의 정보가 전산데이터로 저장됨으로써 다양한 분석이 가능하다.

셋째, 교통카드 데이터는 카드요금 정산을 위해 데이터베이스 형태로 관리되기 때문에 별도의 데이터 구축비용이 발생하지 않는다.

넷째, 교통카드가 이용되는 모든 지역의 전체 대중교통 노선 분석이 가능하다.

교통카드 데이터 중 대중교통 현황조사에 활용되는 항목은 이용자별 카드번호, 이용한 교통수단, 환승여부, 노선코드 및 노선 명, 승하차 정류장 ID 및 정류장명, 승하차시간, 승하차요금, 이용자유형(일반, 청소년, 어린이, 기타) 등이다.

교통안전공단은 교통카드 데이터 분석을 수행하면서 2013년에 교통카드 분석시스템을 구축하였다. 이를 통해 조사대상 지자체수가 2011년 92개에서 2013년 127개로 증가하였고, 조사노선수도 1,542개에서 10,290개로 증가하였다. 반면, 조사예산은 2011년 613,303 천원에서 2013년 351,300 천원으로 감소하였다.

교통카드 분석시스템은 대중교통 이용자수, 이용요금, 환승유형, 이용시간·거리, 혼잡도 등의 조사결과를 지역별, 요일별, 수단별, 이용자 유형별, 교통수단 유형별, 시간대별, 주중·주말별, 기상상태별로 제공한다.

교통카드 분석시스템의 정책 활용 사례로는 국토교통부의 광역버스 입석통행 안전성 검토를 위한 시간대별 정류장별 승하차인원, 출근시 평균통행속도를 제공하였으며, 서울시 노선별 승차인원을 제공하여 광고물 설치사업에 활용, 광역시도의 정류장 승하차인원 제공을 통한 WiFi 설치계획 작성을 지원하였다. 또한 교통관련 학술연구를 위한 데이터를 지원한다.

교통카드데이터의 빅데이터로서 활용가치가 높음에도 실제 데이터이용 활성화가 제약되고 있다. 그 이유는 공공정보로서 인식 부족, 관계기관별 역할 정의 미흡, 공공목적을 전제로 하는 교통카드데이터 수집항목의 미정의, 개인정보 및 개인위치정보 침해에 대한 우려 등이다[194].

여기에서 가장 근본적인 문제는 공공정보로서 인식 부족인데, 교통카드데이터 소유권의 문제가 논란이 되고 있다. 교통가드데이터 소유권은 원칙적으로 교통카드 이용자에게 귀속된다. '위치정보의 보호 및 이용에 관한 법률'에 따라 요금 정산의 당초 목적 이외의 목적을 위해서 제3자 제공을 위해서는

194) 이인묵··박선영·민재홍. 2014. 교통카드데이터 공공이용 활성화를 위한 정책방안. 2014년도 한국철도학회 추계학술대회 논문집. 한국철도기술연구원·교통안전공단. 2014. 교통카드 이용 데이터의 공공성 확보 및 이용활성화 방안 연구.

각 이용자로부터 별도의 동의를 구해야 한다. 하지만 연구 및 통계목적의 활용은 예외적으로 가능한 상황이다. 그러나 교통수단운영자, 교통카드사업자 등이 교통카드데이터 생산 투자의 이유로 교통카드데이터의 소유권을 주장함으로써 데이터 활용이 제한되고 있다.

교통카드데이터 공공이용 활성화를 위한 정책방안으로서 공공정보로서의 교통카드데이터 정의, 표준수집항목 정의, 교통카드데이터 수집절차 개선, 교통카드데이터 제공체계 규정, 기술적 시스템 지원 방안 등이 제시되었다. 이 중에서 본 연구와 가장 밀접한 관련을 갖는 것은 공공정보로서의 교통카드데이터 정의에 대한 내용이다. 영리적 목적이 아닌 공익적 목적을 명확히 규정하고 이에 한정하여 필요한 정보(교통수단 이용 위치 및 시각에 관한 정보, 승차 수단 및 노선에 관한 정보, 환승 정보, 경로 정보 등)를 규정한다. 이와 관련하여 교통카드데이터의 이용활성화를 위해서 '대중교통의 육성 및 이용촉진에 관한 법률' 개정이 추진되어 2014년 개정되어 교통카드 데이터의 수집과 활용을 위한 법적 근거가 마련되었다.

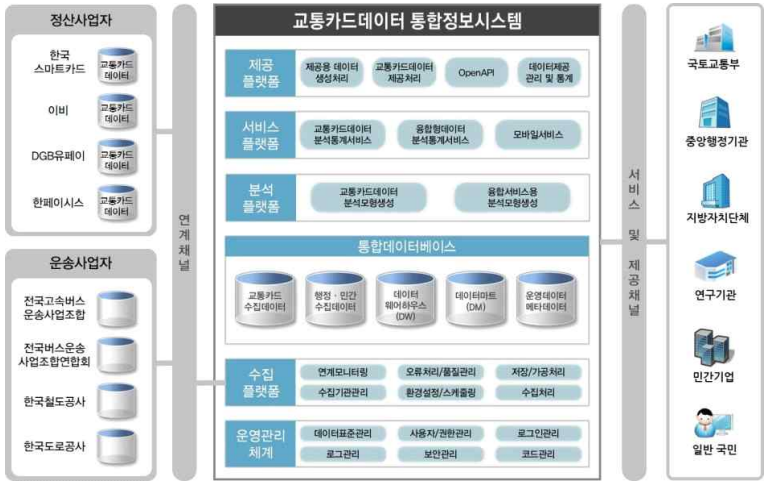
2016년 5월에 국토교통부가 교통카드 빅데이터 활용 방침을 발표하면서 교통카드 빅데이터 통합정보시스템 1단계 구축을 발표하였다[195]. 이에 따르면 데이터 활용을 위한 법적 근거가 없었고 교통카드 정산사업자(한국스마트카드, 이비카드, 코레일 등 8개사) 별로 정보 체계가 달라 효율적인 사용에 제약이 있었으나, 「대중교통의 육성 및 이용촉진에 관한 법률」의 개정으로 교통카드 데이터의 수집과 활용을 위한 법적 근거가 마련되었다[196]. 2016년 1개 교통카드 정산사업자를 대상으로 시스템 표준화 기반을 마련하는 1단계 사업을 추진하고, 2017년에 전체 정산사업자로 확대하여 시스템 구축을 완료한다.

효과로는 기존에 우리나라 전체 대중교통 9천 여 개 노선에 대한 수요조사 방식과 비교하여 조사비용을 약 97% 절감(9억 5천만 원→ 4천7백만 원)하고, 데이터 요청 시 결과 제공까지 걸리던 기간도 기존 45일~ 90일에서 10일 이내로 대폭 단축할 수 있을 것으로 예상한다.

또한 민간에도 관련 데이터를 제공하여 부동산, 통신, 재해·재난, 기상 등 다양한 분야와 연계, 광고입지 분석, 창업 등에 폭넓게 활용될 것을 기대한다.

애로사항으로 선불교통카드업체에서는 정부의 관련 데이터 취합에 협조하기 위해서는 데이터를 가공해야 하고, 이에 필요한 시스템 구축과 인력 투입 비용에 대한 문제를 제기하였다.

195) 하루 2천만 건의 교통카드 빅데이터 활용 시스템을 만든다, 국토교통부 보도자료. 2016.5.25.
196) <첨부1> 참조



[그림 18] 교통카드데이터 통합정보시스템

4) 시사점

빅데이터 활용이 활발한 교통 분야의 사례는 빅데이터 통계 생산을 위해 다음과 같은 시사점을 제기한다.

첫째, 교통카드데이터의 사용의 경우 데이터의 공공재적 성격에 대한 인식을 바탕으로 법안 개정이 이루어져 활용을 위한 기반 구축에 착수함으로써 빅데이터를 활용한 통계 생산에 긍정적인 모델로서 평가된다. 이 과정을 보면 정산사업자들이 반공공적 성격을 갖고 있었음에도 상당한 어려움이 있었고, 실제 지금도 데이터 사용이 어렵다는 전문가들의 보고가 있다. 하지만 빅데이터의 공공적 성격에 대한 법적 인정을 바탕으로 통계 생산의 중요한 모델로서 평가될 만하다.

둘째, 교통 분야 내 다양한 성격의 빅데이터에 대한 포괄적인 검토를 바탕으로 빅데이터 수집, 분석, 유통 과정의 개선점이 도출되었으며 이 과정에서 품질 관리에 대한 논의들이 제시되는 점을 주목할 만하다. 이러한 논의들이 다른 분야에서도 유사하게 제기되는지 파악하고, 각 분야 사이의 공통점을 바탕으로 국가차원에서 빅데이터 통계의 품질에 대한 논의가 가능할 것으로 기대된다.

<첨부1> “대중교통의 육성 및 이용촉진에 관한 법률” 변경 개요

국토교통부는 ‘17년 말까지 진행될 예정인 ‘교통카드빅데이터 통합정보시스템’ 구축사업을 시행 중에 있으며 2016년에는 1개 교통카드 정산사업자를 대상으로 1단계 사업을 추진하여 시스템 표준화 기반을 마련하고, ‘17년에 전체 정산사업자로 확대하여 시스템 구축을 완료할 계획이다.

교통카드 데이터는 법적으로 ‘이용자를 알아볼 수 없는 형태로 가공한 자료’로 규정되어 있으며, 개별 교통카드 정산사업자는 교통카드 정보를 암호화하여 가상번호로 변환한 뒤 이를 통합정보시스템에 제공하게 되므로 데이터 수집단계에서부터 개인정보 보호를 위한 장치가 마련되어 있다.

국토교통부는 ‘교통카드빅데이터 통합정보시스템’을 활용하면 기존에 우리나라 전체 대중교통 9천 여 개 노선에 대한 수요조사방식과 비교하여 조사비용을 약 97% 절감(9억 5천만 원→4천7백만 원)할 수 있으며, 데이터 요청 시 결과 제공까지 걸리던 기간도 기존 45일~ 90일에서 10일 이내로 대폭 단축할 수 있을 것으로 예상하고, 통합정보시스템을 통해서 대중교통 이용자의 통행패턴을 분석하면 노선 신설·조정, 정차 지점 및 배차 간격 최적화 등 정부·지자체·사업자 별로 보다 편리하고 정밀한 교통체계를 만들어 갈 수 있게 되어 대중교통 이용이 활성화되는 효과가 생기게 된다고 밝혔다.

법령 개정 내용 전문

대중교통의 육성 및 이용촉진에 관한 법률 일부를 다음과 같이 개정한다.

제2조에 제6호 및 제7호를 각각 다음과 같이 신설한다.

6. “교통카드”란 교통요금을 전자적으로 지급·결제하는 카드나 그 밖의 매체를 말한다.

7. “교통카드데이터”란 교통카드를 사용하여 대중교통수단을 이용한 전산자료 중 이용자의 통행실태 파악에 필요한 자료로서 이용자를 알아볼 수 없는 형태로 가공한 자료를 말한다.

제10조의2제1항 중 “교통요금을 전자적으로 지불·결제하는 카드나 그 밖의 매체(이하 “교통카드”라 한다)가”를 “교통카드가”로 한다.

제10조의8부터 제10조의11까지를 각각 다음과 같이 신설한다.

제10조의8(교통카드데이터의 수집·관리 및 제출) ① **국토교통부장관은 대중교통수단 이용자의 통행실태를 파악하기 위하여 교통카드데이터를 수집하고 관리하여야 한다.**

② **국토교통부장관은 제1항에 따라 교통카드데이터를 수집하고 관리하기 위하여 대중교통운영자, 「여객자동차 운수사업법」 제53조에 따른 조합 및 같은 법 제59조에 따른 연합회(이하 “대중교통운영자등”이라 한다)와 「전자금융거래법」 제2조제4호에 따른 전자금융업자, 그 밖에 교통요금 정산을 위하여 교통카드 이용자료를 수집하는 자(이하 “교통카드정산사업자등”이라 한다)에게 대통령령으로 정하는 바에 따라 교통카드데이터 제출을 요청할 수 있다.**

③ **제2항에 따라 교통카드데이터 제출을 요청받은 자는 특별한 사유가 없으면 이에 따라야 한다.**

제10조의9(교통카드데이터의 제공) ① 교통카드데이터를 이용하려는 자는 대통령령으로 정하는 바에 따라 국토교통부장관에게 제10조의8에 따라 수집된 교통카드데이터의 제공을 요청할 수 있다.

② 국토교통부장관은 제1항의 요청에 따라 교통카드데이터를 제공하는 경우 집계자료 형태(제10조의8에 따라 제출받은 자료를 분류·합계·변형하는 등 통계처리하여 가공한 형태를 말한다)로 제공하여야 한다. 다만, 국가, 지방자치단체 및 교통 관련 연구기관이나 공공기관으로서 대통령령으로 정하는 기관이 교통 관련 정책수립, 업무수행, 통계작성 및 학술연구 등의 목적으로 요청하는 경우에는 그러하지 아니하다.

③ 제1항 및 제2항에 따라 **교통카드데이터를 제공받은 자**는 다음 각 호의 사항을 지켜야 한다.

1. 제3자의 권리를 침해하거나 범죄 등의 불법행위를 할 목적으로 교통카드데이터를 이용하지 아니할 것

2. 교통카드데이터를 제3자에게 임의로 제공하거나 유출하지 아니할 것

3. 교통카드데이터를 위조하거나 변조하지 아니할 것

4. 교통카드데이터가 분실되거나 도난되지 아니하도록 대통령령으로 정하는 바에 따라 안전성 확보에 필요한 조치를 할 것

④ 국토교통부장관은 제1항 및 제2항에 따라 교통카드데이터를 제공받은 자가 제3항을 위반하거나 제10조의11의 교통카드데이터관리지침을 위반한 경우 교통카드데이터의 제공을 거부하거나 중단할 수 있다.

제10조의10(교통카드데이터 통합정보시스템의 구축·운영 등) ① 국토교통부장관은 제10조의8 및 제10조의9에 따른 교통카드데이터의 수집·관리·제출 및 제공을 위하여 교통카드데이터 통합정보시스템(이하 “통합정보시스템”이라 한다)을 구축·운영할 수 있다.

② 국토교통부장관은 통합정보시스템에서 보유한 정보의 누출, 변조, 훼손 등을 방지하기 위하여 접근 권한자의 지정, 방화벽의 설치, 암호화 소프트웨어의 활용 등 관리적·기술적 보호 조치를 하여야 한다. 이 경우 관리적·기술적 보호 조치의 구체적인 내용은 대통령령으로 정한다.

③ 국토교통부장관은 대중교통운영자등 또는 교통카드정산사업자등에게 제10조의8제2항 및 제3항에 따른 교통카드데이터 제출에 필요한 기술적 지원을 할 수 있다.

④ 국토교통부장관은 대통령령으로 정하는 기관에게 통합정보시스템의 구축·운영 업무를 위탁할 수 있다.

⑤ 제4항에 따라 업무를 위탁받는 기관은 다음 각 호의 사항을 지켜야 한다.

1. 제3자의 권리를 침해하거나 범죄 등의 불법행위를 할 목적으로 교통카드데이터를 이용하지 아니할 것
2. 통합정보시스템 운영 업무를 다른 기관에게 재위탁하지 아니할 것
3. 교통카드데이터를 제3자에게 임의로 제공하거나 유출하지 아니할 것
4. 교통카드데이터를 위조하거나 변조하지 아니할 것
5. 제2항의 관리적·기술적 보호 조치를 따를 것

⑥ 국토교통부장관은 제4항에 따라 업무를 위탁하는 경우 필요한 행정적·재정적 지원을 할 수 있다.

제10조의11(교통카드데이터관리지침) ① 국토교통부장관은 제10조의8부터 제10조의10까지의 규정에 따른 교통카드데이터의 수집·관리·제출·제공 및 통합정보시스템의 구축·운영 등에 관하여 교통카드데이터관리지침을 정하여 고시할 수 있다.

② 제10조의8에 따라 교통카드데이터를 제출하는 자, 제10조의9에 따라 교통카드데이터를 제공받는 자 및 제10조의10에 따라 통합정보시스템의 구축·운영 업무를 위탁받은 자는 제1항에 따른 교통카드데이터관리지침에 따라야 한다.

제16조제1항 각 호 외의 부분 중 “국토교통부령이 정하는 바에 따라 다음 각호의 사항을 조사하여야 한다”를 “매년 다음 각 호의 사항을 조사하고 그 결과를 공표하여야 한다”로 하고, 같은 조에 제4항을 다음과 같이 신설한다.

④ 제1항에 따른 대중교통현황조사 및 결과의 공표에 필요한 사항은 국토교통부령으로 정한다.

나. 보건·의료 분야 빅데이터 활용

1) 개요

보건·의료 분야의 경우 타 분야와 달리 데이터 수집이 비교적 오래 되었으며, 전통적으로 데이터의 규모 또한 크다. 특히 의료보험 이용 및 관리 데이터는 자료의 정형성을 갖추지 못하였으나 전 국민의 의료행위에 대한 상세한 내용을 담고 있다.

주요 데이터 수집 주체별로 데이터의 공공사용에 대해 적극적으로 나서고 있으나, 개인의 진료정보를 담고 있는 데이터이기 때문에 사용에 있어서 제한사항이 많은 편이다. 현재 한국복지패널 데이터를 제외한 기타 빅데이터는 공개가 되지 않고 있거나 제한적으로 공개되고 있다.

2) 보건·의료 분야 빅데이터 활용 현황

공개된 보건·의료 분야의 빅데이터가 주로 활용되는 분야는 학문적 연구가 대부분으로, 주로 의학적 전문분야에서 사용된다. 반면에 상업적인 면으로의 활용은 드물다. 빅데이터를 활용한 주요 사례로는 건강보험 빅데이터를 활용한 국민건강 알람서비스, 건강검진 및 진료정보, 대사증후군맞춤정보, 뇌졸중위험예측프로그램 등이 있다. 또한 지방자치단체에서 의료시설 수요 및 이용패턴 파악을 위해 사용되고 있는데, 광주시 광산구의 사례가 대표적이다.

보건의료분야 주요통계 현황

<표 36> 보건복지분야 국가승인통계 현황(2015년 기준)

통계 종류	분야	통계 명칭
조사(42종)	보건(23종)	건강보험환자 진료비 실태조사, 국민건강영양조사, 국민보건의료실태조사, 근로환경조사, 병원경영실태조사, 아동구강건강 실태조사, 의료기관별 급여 적정성 평가현황, 의료기기 제조/유통 조사, 의약품/의료기기연구개발실태조사, 작업환경실태조사, 전국민장내기생충감염실태조사, 전국출산력 및 가족보건복지실태조사, 정신질환실태조사, 지역사회건강조사, 청소년건강행태 온라인조사, 퇴원손상심층조사, 한국의료패널조사, 한국인인체치수조사, 한방의료이용 및 한약소비실태조사, 한의약산업실태조사, 화장품제조유통조사, 환자조사
	복지(19종)	가정폭력실태조사, 고령화연구패널조사, 국민노후보장패널조사, 기업 및 공공기관의 가족친화수준조사, 남해군 노인실태조사, 노인실태조사, 노후준비실태조사, 보육실태조사, 복지욕구조사, 사회서비스수요/공급실태조사, 생명보험성향조사, 서울특별시 복지실태조사, 성폭력실태조사, 아동종합실태조사, 장애인생활체육실태조사, 장애인실태조사, 장애인편의시설설치현황조사, 한국복지패널조사, 한부모가족실태조사

통계 종류	분야	통계 명칭
보고(34종)	보건(21종)	HIV/AIDS 신고현황, 건강검진통계, 건강보험 주요 수술통계, 건강보험통계, 결핵현황, 공중위생관계업소 실태보고, 근로자건강진단실시상황보고, 급성심장정지조사, 노인장기요양보험통계, 법정감염병발생보고, 보건소 및 보건지소운영현황, 수입식품현황, 식품 및 식품첨가물 생산실적, 식품수거검사실적, 암등록통계, 완제의약품 유통정보통계, 응급의료현황통계, 의료기기생산실적, 전국예방접종률조사, 지역별의료이용통계, 학생건강검사통계보고
	복지(13종)	가정위탁 국내 입양소년소녀 가장현황, 국민기초생활보장 수급자현황, 국민연금통계, 노인복지시설현황, 노인학대현황, 보훈보상금지급현황, 산업재해현황, 산재보험통계, 아동복지시설 보호아동 및 종사자현황보고, 어린이집 및 이용자통계, 요보호 아동현황보고, 장애인현황, 학대피해아동보호현황
가공(5종)	보건(4종)	국민의료비추계 및 국민보건계정, 사망원인통계, 어린이 식생활안전지수, 의약품소비량 및 판매액 통계
	복지(1종)	한국의 사회복지지출

<표 37> 보건의료분야 공공 빅데이터 현황

구분	보유기관	내용	공개여부
건강보험표본코호트DB	국민건강보험공단	자격DB: 건강보험가입자 및 의료수급권자의 성, 연령대, 지역, 사회경제적 변수, 장애, 사망관련 등 진료DB: 요양급여 청구자료로서 진료, 상병, 처방 관련 변수 건강검진DB: 검진 주요결과 및 문진에 의한 생활습관 및 행태관련 자료 요양기관DB: 요양기관 종별, 설립구분, 지역, 시설, 장비, 인력관련 자료	제한적 공개
환자데이터셋	건강보험심사평가원	건강보험 청구자료를 기초로 진료개시일 기준 1년 간 진료받은 환자대상의 표본 데이터	제한적 공개
한국인체자원	질병관리본부	공여자로부터 기증받은 인체유래물(DNA, 조직, 혈액, 뇨 등)과 임상(진단명, 수술명, 병리조직검사결과, 혈액검사 등), 역학(성별, 생년월일, 음주력, 흡연력 등) 및 유전(SNP, CNV, Exome 등)정보	제한적 공개
지역보건의료정보	사회보장정보원	전국 보건기관(보건의료원, 보건소/지소, 보건진료소)의 보건사업 및 행정업무, 전자의무기록 및 진료관련(진료내역 및 검진결과 등) 정보	미공개
지역사회건강조사	질병관리본부	지역 보건의료계획수립 및 보건사업 평가 활용 지표로서, 건강행태, 건강검진 및 예방접종, 질병이환, 의료이용, 사고 및 중독, 활동제한 및 삶의 질, 보건기관 이용, 사회 물리적 환경, 심정지, 교육 및 경제활동 등	공개
국민건강조사	질병관리본부	국민의 건강 및 영양 상태에 관한 현황 및 추이 파악 신체계측, 비만, 고혈압 등 검진조사, 흡연, 음주, 비만 및 체중조절, 신체활동 등 건강설문조사, 식품 및 영양소 섭취현황, 식생활행태, 식이보충제 등 영양조사	공개
한국의료패널	한국보건사회연구원	개인의 건강수준, 의료이용 및 의료비 지출 요인, 건강행태, 의료욕구, 보건의료서비스 수요행태 변화분석 사회경제적 특성, 의약품 구매, 경제활동, 건강수준, 의약품 복용행태, 민간의료보험, 건강기능식품, 건강행태 등	공개

<표 38> 보건의료분야 공공 빅데이터 현황

구분	보유기관	내용	공개여부
사회보장정보	사회보장정보원	각 부처 및 정보보유기관에서 제공하고 있는 복지사업정보 및 지원대상자의 자격정보, 수급이력정보를 통합관리 - 복지대상자 선정·사후관리를 위해 45개기관 552종의 소득·재산자료 및 서비스 이력정보 연계	미공개
복지콜센터 상담 데이터	복지콜129 센터 등	상담이력	미공개
사회서비스 전자바우처	사회보장정보원	정부와 지자체가 사회로부터 도움을 필요로 하는 사람에게 돌봄, 일상생활 지원, 사회적응지원, 문화체험 등의 서비스를 제공하는 것을 전산처리	미공개
보육통합정보	사회보장정보원	보육바우처 운영 및 행정지원 정보	미공개
한국복지패널	한국보건사회연구원	빈곤층, 근로빈곤층, 차상위층 등의 규모 및 생활실태 변화를 동태적으로 파악 인구집단별 생활실태 및 경제활동, 건강·의료, 주거, 소득,복지서비스 이용, 복지욕구	공개

보건의료 영역의 빅데이터 활용이 사회복지 영역과 같은 인접 정책 분야 보다 활발한데, 이는 보건의료부문에서 공개된 데이터가 많기 때문이다. 이러한 공공 빅데이터 활용은 데이터를 보유하고 있는 기관을 중심으로 이뤄지는 경향이 있어서 다방면으로 폭넓게 쓰이지는 못하고 있다. 최근 학술/의학/정책적 목적으로 공공 빅데이터를 활용하려는 시도가 늘어나고 있기는 하지만, 공개된 데이터가 적기 때문에 활용 기회가 제한되는 편이다.

<표 39> 보건복지관련 공공데이터 활용 서비스 현황

서비스	주요내용	데이터 종류
국민건강알림서비스	4단계 건강위험 예보 발령, 개인건강기록 시스템을통한 맞춤형 건강정보 제공	진료내역, 의약품 처방, 건강검진정보 등 (국민건강보험공단)
건강검진 진료정보	검진결과 맞춤건강서비스	건강검진 정보, 의료기관 이용내역 등 (국민건강보험공단)
뇌졸증 위험예측 프로그램	뇌졸중과 관련된 고혈압, 콜레스테롤, 생활습관, 가족력, 환경요인 등을 기초로 10년 이내에 뇌졸중에 걸릴 위험도 평가	건강검진정보, 문진정보 등 (국민건강보험공단)
대사증후군 맞춤정보	대사증후군 요소와 관련한 건강상태 및 위험요인별 맞춤형 처방정보 제공	건강검진정보, 문진정보 등 (국민건강보험공단)
운전면허 발급 간소화 서비스	국가건강검진정보 중 운전면허 적성검사에 필요한 시력 · 청력 정보 공동이용	건강검진정보 (국민건강보험공단)
갑상선암 의사에게 꼭 물어보세요	갑상선암 정보, 종합병원 전문의와 1:1 맞춤형 상담	병원정보 (건강보험심사평가원)

서비스	주요내용	데이터 종류
병원 약국찾기	위치기반 서비스와 연동하여 내 주변의 병원, 약국, 응급의료기관 위치정보 제공	전국 응급의료정보 (국립중앙의료원)
MediMap	서울 소재 병원 규모별, 카테고리별 병원 위치, 연락처 정보 제공	전국병원정보 (공공데이터활용지원센터)
처방약 토탈 검색	약국 위치, 개폐정보	약국정보 (국립중앙의료원)
4th-Life	요양병원 정보 제공	-
옳은 서비스	식단 재료의 합리적 구매기준을 위한 식단정보, 가격정보, 성분정보	식단정보, 가격정보, 성분정보 (식품의약품안전처)
충주 나눔의 집	충주시 소재 장애인 복지시설 관련 정보 및 장애인 등록현황 정보 제공	장애인 등록정보 (충청북도 충주시)
대전당직병원	대전지역 당직병원 응급실정보, 전화번호, 주소, 병원등급 등 정보 제공	당직의료기관정보 (대전광역시)
헬스온스토리	상황인식기술 통한 사용자 맞춤형 건강정보 자동제공	보건기상지수정보 (기상청)

3) 국민건강보험공단 데이터 활용

국민건강보험공단의 데이터는 가입자 및 피부양자의 자격관리, 건강 보험료 부과와 징수, 보험급여 관리, 보험급여 비용의 지급, 영유아를 제외한 국가 건강검진 등을 포함하고 있다. 국민의 다양한 정보들을 포함하는 만큼 빅데이터의 용량(Volume) 기준에는 부합하지만, 대부분 정형데이터이기 때문에 빅데이터의 다양성(Variety)의 기준에는 잘 맞지 많고, 데이터의 생성 및 처리에 구조적인 문제로 상당한 기간을 요구하기 때문에 속도(Velocity)의 기준에는 잘 들어맞지 않는다[197]. 따라서 대용량 데이터(Big-sized data)라 지칭하는 것이 옳을 것이다.[198]

또한 국민건강보험공단의 업무의 특성상 기존의 자료를 정해진 절차대로 활용하는 데에 그치는 경우가 많았기 때문에 새로운 빅데이터 구축이나 활용에 대한 관심은 상대적으로 적었다. 하지만 최근 전체건강보험가입자를 대상으로 100만 명 규모의 표본데이터베이스를 구축하고, 이를 공표하기 시작하였다. 공표되는 데이터는 2002년 자료부터 제공되며, 보험 자격 데이터베이스(성별, 연령, 지역, 가입자 구분, 소득분위 등 대상자의 변수 및 장애, 사망관련)와 진료 데이터베이스(상병 내역, 진료 내역, 진료 명세, 처방전 등), 건강검진 데이터베이스가 제공 된다[199].

197) 오상우. 2015. “건강보험 빅데이터의 의료계 활용” 의료정책포럼 12(3), 18-23.
198) 이연희. 2015. “보건복지분야 공공 빅데이터의 활용과 과제” 보건복지포럼.
199) 오상우. 2015. “건강보험 빅데이터의 의료계 활용” 의료정책포럼, 제12권 3호.

<표 40> 국민건강보험공단에서 관리하는 데이터 목록

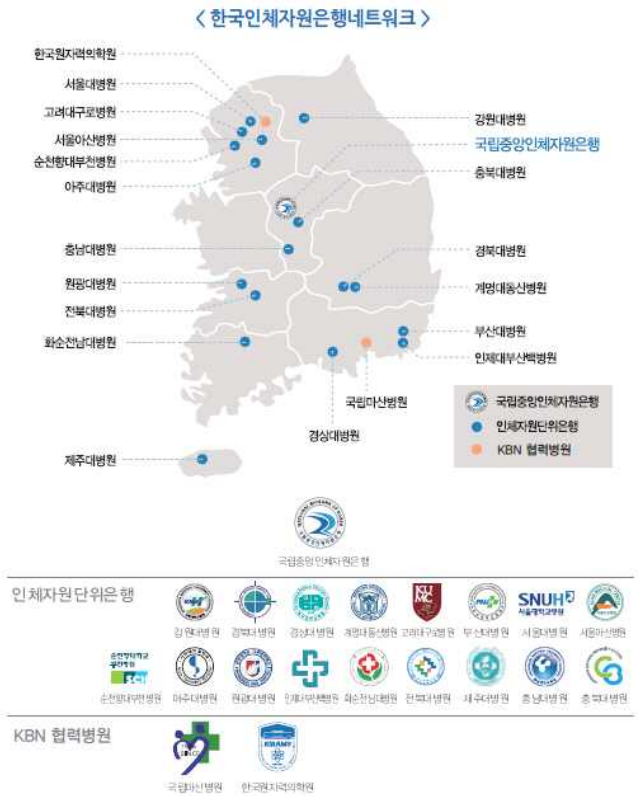
데이터 분류	세부항목
건강보험 자격 및 보험료	건강보험 취득/상실
	성별
	연령
	사업장
	의료급여종별
	장애유형
	보험료 분위
사망자 및 신생아	장제비 지급 관련 사망자 자료
	사망사유
	신생아 자료
진료내역	진료내역
	진료명세서 일반/상세 내역
	진료상병내역
	진료처방내역(의료/보건 기관, 치과/한방, 약국별로 분류됨)
요양기관내역	의료/보건 기관의 인력
	의료/보건 기관의 진료과목
	의료/보건 기관의 장비
	시설의 산재지정 여부
건강검진 기록	일반건강검진
	생애전환기 건강진단
	암검진(위, 대장, 간, 유방, 자궁경부)
	구강검진
	영유아검진
	영유아 구강검진

4) 인체자원은행 데이터 활용

질병관리본부는 2008년부터 국립중앙인체자원은행과 전국 17개 대학병원 소재 인체자원단위은행 및 2개 협력병원(한국원자력의학원, 국립마산병원)과 함께 한국인체자원은행사업(Korea Biobank Project, KBP) 네트워크를 운영하고 있다[200]. 인체자원(Human Bioresource)이란 보건의료 연구에 필요한 모든 자원, 정보를

200) 질병관리본부. 2016. "2016 한국인체자원은행사업 안내자료."

통칭하는 개념이며, 이는 인체유래물[201]과 함께 인체유래물 기증자의 임상, 역학정보 및 유전정보 등을 포괄한다.



[그림 19] 한국인체자원은행 네트워크

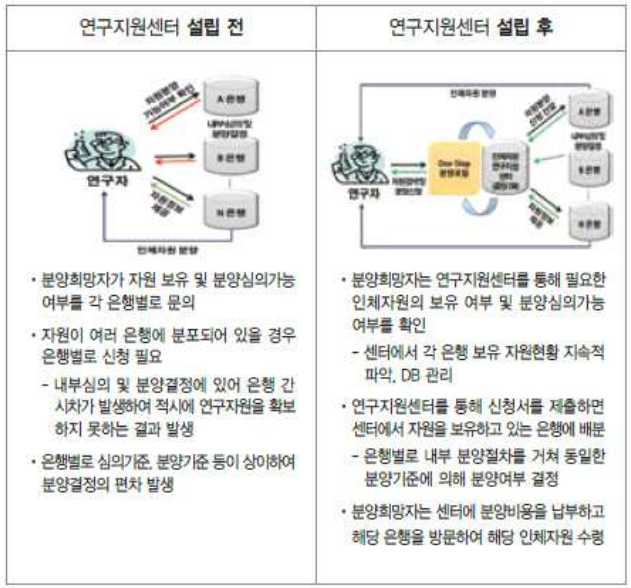
단위은행들을 통하여 수집된 자료의 양을 보면 2015년 기준 누적 수집은 328,489명이고 2016년 3분기 현재 당해 연도 누적 수집은 33,435명이다. 이와 같이 수집된 인체자원은 병원, 대학, 국가 및 민간 연구기관에서 활용된다.

201) 인체유래물이란 인체로부터 수집, 채취한 조직, 세포, 혈액, 체액 등 인체 구성물 또는 이들로부터 분리된 혈청, 혈장, 염색체, DNA, 단백질 등을 말한다. (생명윤리 및 안전에 관한 법률 제2조 제11항)

인체자원은행의 네트워크에서 중심적인 역할을 수행하는 것이 국립중앙인체자원은행이다. 기존에는 인체자원을 필요로 하는 연구자가 인체자원을 보유하고 있는 부서를 찾아 필요 자료를 문의하여 분양이 이루어짐으로써 절차가 번거롭고 많은 시간이 소요되었다.

이에 국립보건연구원이 국립중앙인체자원은행을 설립하여 2013년부터 안내, 분양신청, 상담 업무를 일원화하였고, 온라인을 통한 검색 및 신청 활동을 가능케 되었다. 연구자들에게 공정하고 투명하게 인체자원을 분양하기 위하여 "국립중앙인체자원은행 분양심의위원회"를 구성하여 운영하고 있다.

또한 질병관리본부에 인체자원연구지원센터를 설립하여 연구자의 자료 활용의 편의성을 제고하였다.



[그림 20] 인체자원연구지원센터 설립에 따른 변화

5) DW분석사 자격증 제도

건강보험심사평가원(이하심평원)은 자체 데이터 베이스 관리 및 분석을 위해 사내자격제도로서 DW(DataWereHouse)[202] 분석사를 2004년부터 시행하였다. DW시스템은 심평원에서 구축한 국민의 진료정보 시스템 자료를 일컫는 것으로, 연평균 25만 회 분석을 실시하고 있다. 데이터의 규모는 연간 12억 건에 달한다. DW 분석사는 2010년 3월에 한국산업인력공단에서 '사업내자격'으로 공식 인증 받았다. 2012년부터는 1급 DW 분석사 시험도 수행하고 있다. 2014년까지 사내자격 검정을 통해 자격을 취득한 사람은 모두 440명(1급 15명, 2급 425명)이다.

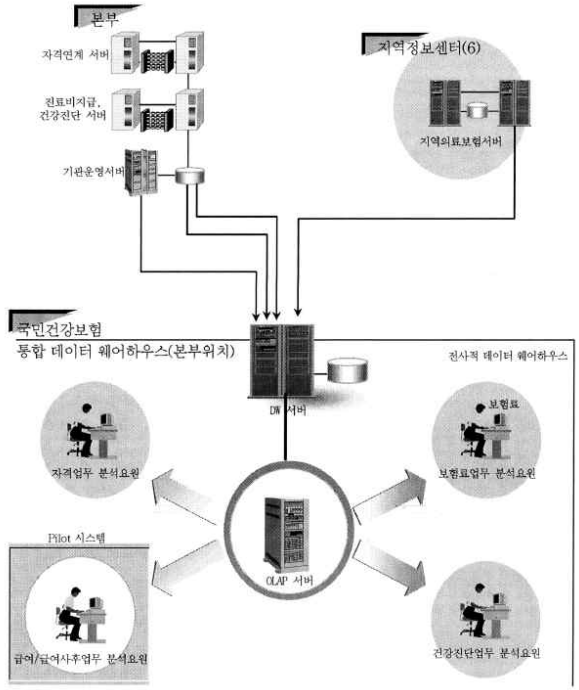
2000년에 지역의료보험공단과 직장의료보험조합, 각종 공단을 통합하여 국민건강보험공단이 출범함으로서 대량의 감사 수요가 발생하였다. 유효한 감사 시스템 구축을 위해 2001년 6월에 감사정보기본시스템을 구축하였고, 분석을 위해 데이터웨어하우스 시스템을 도입하였다.

국민건강보험공단과 심평원에서 보유하고 있는 데이터는 국민 의료정보 중 급여부분에 국한된다. 임의비급여분[203]의 진료 데이터도 있기는 하나, 기본적으로 비급여 진료행위에 대한 정보는 수집되지 않는다. 심평원에서 1년에 진행하는 보험급여 처리 건수는 약 5천억 건에 달한다. 국민건강보험공단과 심평원에서 보유하고 있는 데이터는 대부분 공유하고, 보험료 자료에 대한 것만 심평원에 제공되지 않는다. 이하에서는 DW 구성을 살펴보기 위해 국민건강보험공단의 DW시스템을 정리하였다.

DW시스템은 각 지역의료보험 서버와 연계되어 모든 정보를 통합 관리하는 서버를 말한다. 이 데이터는 분류 기준이 다양할 뿐 아니라 내용이 방대하기 때문에 데이터를 정리할 필요가 있다. 이는 OLAP 서버를 통해 검색/정렬이 가능하다. OLAP(OnLine Anaytical Processing) 서버는 여러 데이터에 대해 일관된 검색기준으로 재정렬해주는 작업을 수행한다.

202) 데이터 웨어하우스(Data warehouse)는 사용자의 의사결정에 도움을 주기 위하여 기간시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 데이터베이스를 일컫는다. 데이터의 효율적 관리를 위해 시계열적 축적과 통합을 중시하는 구조를 보인다. 대부분의 데이터베이스는 그 용량이 수백 기가바이트에서 수 테라바이트에 이르기 때문에 쉽게 구현하기 힘들었으나 최근 병렬서버의 등장과 자기디스크의 발달로 실현이 가능해졌다.

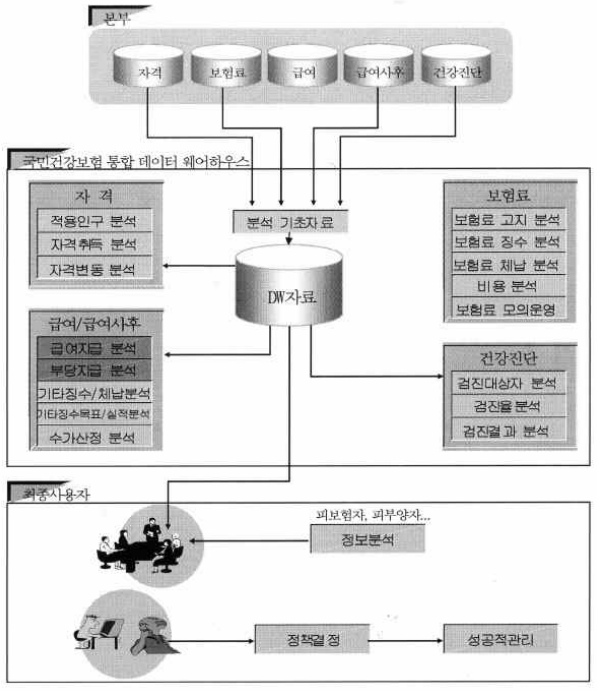
203) 의료의 필요성, 유효성, 안정성이 인정되지 않는 의료서비스는 비급여항목으로 분류되어 의료보험에서 지급이 되지 않지만, 일부 비급여 항목은 국가에서 인정해주기도 한다.



[그림 21] 국민건강보험공단 DW시스템 구성도

자료: 용왕식, 최영식, 조해곤. 2003. 데이터웨어하우스 정보기법을 활용한 감사성과 효율화에 관한 연구, 《감사논집》: p.250.

이용자는 각 서버에 있는 데이터 중 분석 목적에 맞는 데이터를 데이터 웨어하우스에서 추출하여 정렬한다. 정렬한 데이터를 바탕으로 분석을 수행하고, 부당급여 청구 대상의 기준을 정하거나, 적정보험료를 산정하는 업무 등을 수행한다. DW 시스템의 업무 흐름은 다음의 그림과 같이 정리할 수 있다.



[그림 22] DW 시스템의 업무 흐름도

자료: 용왕식, 최영식, 조해곤. 2003. 데이터웨어하우스 정보기법을 활용한 감사성과 효율화에 관한 연구, 《감사논집》: p.251.

6) 시사점

빅데이터 활용이 활발한 보건·의료 분야의 사례는 빅데이터 통계 생산을 위해 다음과 같은 시사점을 제기한다.

첫째, 보건의료 분야의 경우 이미 빅데이터 개념이 본격적으로 등장하기 전부터 건강보험 사업의 운영에서 발생하는 대용량 자료를 이용하여 통계를 작성한 경험을 축적하였다. 건강보험 자료를 바탕으로 이미 국가통계로서 승인을 받은 경우들도 있다. 이와 같은 사례는 대용량의 데이터라 할지라도 정형화된 데이터의 경우 국가승인통계로서 관리하는 것이 가능하다는 점을 시사한다.

둘째, 방대한 규모의 정보를 관리하고 통계를 생산하는 과정에서 필요한

전문성을 확보하기 위하여 이에 맞는 자격증 제도를 운영할 수 있다. 빅데이터를 활용한 통계 생산에 필요한 전문인력을 안정적으로 확보하기 위하여 정부가 주도하는 자격증 제도를 고려하는데 있어서 중요한 참고사례가 된다.

셋째, 빅데이터의 운영과 관리에 필요한 거버넌스 구축의 선도사례로서 참조할 만하다. 인체자원은행의 전국적인 네트워크와 연구지원센터 및 심의위원회 운영을 통해 데이터의 생산과 활용에서 필요한 협력체계의 가능성을 제시하였으며, 윤리적 쟁점을 다루는 제도적 방안을 제시하였다.

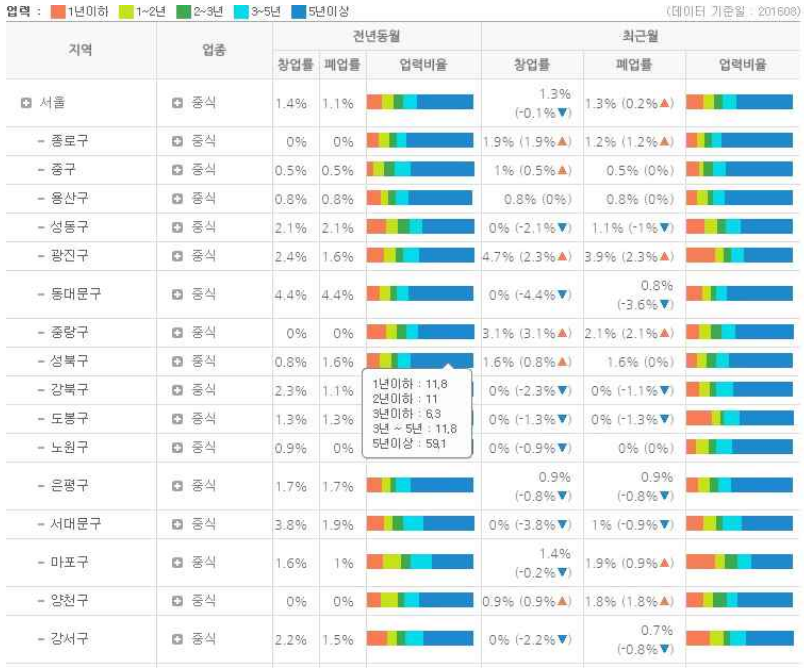
다. 상권분석서비스 (중소기업청, 소상공인진흥공단)

중소기업청과 소상공인진흥공단은 예비창업자, 자영업자를 위해 상권분석서비스를 제공하고 있다(sg.sbiz.or.kr). 홈페이지에서는 상권분석, 상권통계, 상권과밀지수, 점포이력 및 평가 등을 시행할 수 있다. 이러한 분석에 사용된 데이터는 이하 <표 41>과 같다.

<표 41> 상권분석서비스에 사용되는 데이터

데이터	데이터 출처	데이터 내용	업데이트 주기
주거인구	행정자치부 주민등록 인구 통계 및 주거인구를 활용한 추정치	행정구역별 가구수 및 성별/연령대별 인구수 건물단위별 가구수 및 성별/연령대별 인구수 (추정치)	년
직장인구	KCB	소지역별 직장인구 수 및 성별, 연령대별 비율 통계	반기
유동인구	SKT	전국 주요상권 유동량 조사 정보	-
상가 DB	지방자치단체, 자체 조사 데이터	전국 상가/업소의 주소, 전화번호, 업종 데이터	월
매출 DB	카드사	소지역별 업종별 추정 매출 및 요일별/시간대별 매출 통계, 성별/연령대별 이용고객 통계	월
공동주택	국토교통부	전국 아파트 단지별 동별 위치 및 면적/기준시가 정보	년
임대시세	한국감정원 등	전국주요상권 및 시군구 단위 평균 임대시세	분기
기업정보	KCB	대기업, 중소기업, 단체의 주소, 업종, 연락처 정보	년
직업/직종	통계청	시군구별 직업/직종 통계	5년
지하철역별 시간대별 이용인원	도시철도공사	광역시도 지하철 역별 평균 승하차 인원 정보	년
전국주요상권	소상공인시장진흥공단	전국 주요상권 영역 및 업종밀집 정보	년
주요/집객시설	각급기관	공공,금융,의료,교육,유통,문화,숙박,교통 시설위치 및 명칭정보	년
전국학교일람표	교육부	유치원, 초등 중등 고등, 대학교 현황	년

상권정보서비스에서 제공하는 보고서는 이용자가 지역과 업종을 선택하면 이에 해당하는 데이터 셋을 모아서 출력해주는 방식으로 제공된다. 이용자는 한 번에 세 군데까지 상권을 지정할 수 있다. 상권통계 항목에서는 업종별 창업률, 폐업률 및 업력 비율(년도별)을 조회할 수 있다. 이하의 [그림 23]과 같이 구 단위의 데이터 열람이 가능하며 이 외에도 월평균 매출, 임대시세(층별), 지역별 상위 업종 등을 조회할 수 있다.



[그림 23] 상권분석서비스 업력통계 예시

자료: 상권정보시스템 홈페이지(sg.sbiz.or.kr)

상권분석서비스에서 제공하는 정보는 이하의 5개 분류, 20개 중분류, 49종의 분석 데이터로 나눌 수 있다.

<표 42> 상권분석 서비스에서 제공하는 데이터

구분	중분류	제공 정보
개요	개요	상권지도 상권 주요정보 요약 (면적, 가구수, 거주인구수, 직장인수, 주요/집객시설수, 음식/서비스/소매 업종 업소수)

구분	중분류	제공 정보
업종분석	업종 현황(간단)	업소수 추이 업소 정보 및 위치 행정구역별 업소수 추이
	유사업종 현황	업소수 추이 업소 정보 및 위치 전년대비 증가/감소 업종
	중분류 업종 현황	업소수 추이 전년대비 증가/감소 업종
	대분류 업종 현황	업소수 추이 전년대비 증가/감소 업종
	업소 증감 추이	업종별(중분류) 업소수 증감 추이 업종별(소분류) 업소수 증감 추이
	창/폐업률 통계	업종별(중분류) 전국/광역시도/시군구별 창/폐업률 및 업력 통계 업종별(대분류) 전국/광역시도/시군구별 창/폐업률 및 업력 통계
매출 분석	매출 추이	선택업종의 평균매출 및 이용건수 추이 중분류 업종의 평균 매출 및 이용건수 획득
	매출 비교	선택업종의 월평균 매출액 비교 (선택상권/유사상권/인근중심상권) 선택업종의 건당 매출액 비교 (선택상권/유사상권/인근중심상권)
	매출 특성	선택업종의 주말/주중, 요일별 매출 비율 선택업종의 시간대별 매출 비율 선택업종의 성별/연령대별 매출 비율

구분	중분류	제공 정보
지역 분석	주요/집객 시설	주요/집객시설 현황
		주요/집객시설 정보 및 위치
	학교 시설	학교시설 현황
		학교시설 위치
	교통 시설	교통시설 현황
		교통시설 위치
		지하철 이용 현황
	주요 기업	기업체 현황 (전체, 대기업/중.소기업/단체)
		기업체 정보 및 위치
	브랜드 지수	브랜드 지수 추이
	점포임대 시세	지역별 임대료 비교 (제곱미터당 월 임대료)
		인근 주요상권 평균 임대시세 (3.3㎡ 단위 보증금 및 임대료)
		지역별 임대료 추이 (제곱미터당 월 임대료)

2. 지자체의 빅데이터 활용 통계생산 사례

제1절. 국내 지자체 빅데이터 활용 통계생산 동향

가. 개요

한국지역정보개발원은 빅데이터부를 신설하여 지방자치단체의 빅데이터 활용 활성화를 지원하고 있다. 최근 전국의 지방자치단체를 대상으로 설문조사를 실시하여 빅데이터 통계 관련 사업 수행 환경과 애로사항, 사업 결과 등을 설문조사하였다[204].

조사는 전국 지방자치단체(광역 17개, 기초 226개) 빅데이터 담당자를 대상으로 2016년 4월 5일부터 4월 22일까지 2주간 진행되었으며, 질문지 회수는 광역 15개(88%), 기초 152개(66.7%)로 총 167개(68%) 지자체에서 응답하였다. 설문조사와 동시에 '16년 상반기 지역정보화 실무네트워크에 참석한 12개 광역 시·도 업무 담당자들을 대상으로 FGI(focus group interview)를 진행하였다. 이를 통해 업무추진 전반에 대한 애로사항 및 개선방안에 대한 토론 및 질의·응답을 수렴하였다. 주요 결과는 다음과 같다.

나. 지방자치단체의 빅데이터 사업 추진 환경

전국 지방자치단체들 중에서 빅데이터 사업 추진을 위한 조직과 인력이 갖춰진 곳은 광역 5곳, 기초 3곳으로 파악되었다. 광역시·도는 15곳 중 8곳(53%)이 빅데이터 전담 인력을 갖추고 있지만, 기초 시·군·구는 152곳 중 31곳(20%)만이 전담 인력을 보유하고 있어 상대적으로 빅데이터 추진 기반이 약한 것으로 드러났다.

특히 수도권 지자체의 전담 인력이 지방에 비해 더 많은 것으로 파악되었다. 예를 들어, 서울의 경우 10명의 전담 인력이 있는 반면 광주광역시는 2명에 불과하였다.

- 수도권: 서울 10명, 경기도 12명, 인천 4명, (경기)시흥시 4명
- 비수도권: (광주)광산구 3명 / (광주)광역시, 전북, 제주, (전남)여수시, (경남)창원시, (경남)밀양시 각 2명

204) 이하의 내용은 한국지역정보개발원(2016). 『지방자치단체 빅데이터 추진현황과 정책적 시사점』, 2016년 지역정보화 이슈리포트 제2호와 담당자 인터뷰를 통해서 작성

빅데이터 전담인력의 직렬은 전산직이 대부분이며, 행정직과 임기제전문직이 포함되어 있었다.

- 경기도(12명): 전산직 8명, 행정직 2명, 임기제전문직 2명
- 서울시(10명): 전산직 7명, 행정직 2명, 임기제전문직 1명
- 경기 시흥시(4명): 전산직 3명, 임기제전문직 1명

[그림 24] 지방자치단체의 빅데이터 추진 전담 조직 및 인력 현황

자료: 한국지역정보개발원. 2016. 지방자치단체 빅데이터 추진현황과 정책적 시사점, 《2016년 지역정보화 이슈 리포트》 제2호: p.3.

사업 추진을 위한 조례 및 내부규칙이 제정된 곳은 서울특별시, 경기도, 제주특별자치도 이상의 세 군데에 불과하였고, 나머지 모든 지자체에서는 제도적 환경이 조성되어 있지 않았다.

빅데이터 업무를 위한 산·학·연 네트워크는 광역시·도 8군데에만 구성되어 있었고, 기초 시·군·구에는 구성된 곳이 없었다.

<표 43> 빅데이터 산·학·연 협력 네트워크 구성 현황

지자체	산·학·연 네트워크
서울특별시	다음카카오, KT 외 다수
광주광역시	광주전남연구원, 지역대학
세종특별자치시	빅데이터추진협의체
경기도	한국정보화진흥원, 통신사, 카드사 등
강원도	강원창조경제혁신센터
충청북도	충북빅데이터위원회
전라북도	한국국토정보공사
제주특별자치도	지역정보화추진협의회

다. 빅데이터 사업 추진 현황

2016년 5월 현재까지 진행된 빅데이터 관련 사업은 모두 59개로, 광역시도에서 28개, 기초 지방자치단체에서 31개 사업이 진행되었다.

분야별로는 연구용역(7건)이나 인프라(6건)를 제외하고 적용 분야별로 보면 관광관련 사업이 14건으로 가장 많았고, 다음으로 경제(9건), 교통(8건), 복지(3건), 재난·안전(3건) 순이었다. 관광 분야의 사업은 주로 유동인구 분석을 통한 관광객 집계나 관광객의 신용카드 소비 분석 등이었다. 경제 분야에서는 신용카드 분석 등을 통한 상권분석이나 소비양태 분석이 많았다. 교통 분야에서는 빅데이터를 이용한 대중교통체계 개선이나 주정차 시스템 개선이 추진되었고, KTX 개통에 따른 영향 분석 등도 있었다. 복지 분야에서는 응급환자 시스템 개선, 복지시설 입지 분석, 스마트케어 시스템 도입 등 다양했다. 재난·안전 분야에서는 교통사고나 안전관리 현황 분석 등이 추진되었다.

<표 44> 지방자치단체별 빅데이터 추진 사업

지자체	사업명	추진기간	예산 (백만원)	분야
서울특별시	골목상권분석	10개월 (14.12~15.10)	1,401	경제1
	공유활용 플랫폼	8개월(15.10~16.5)	1,562	인프라
	빅데이터 공유 활용	7개월(14.7~15.2)	903	인프라
인천광역시	관광객유동인구분석	3개월(15.12~16.2)	0	관광1
	송도권역 시내버스노선개편	5개월(14.7~14.12)	0	교통1
	지능형교통체계, 교통관제시스템구축	8개월(15.2~15.9)	583	교통2

지자체	사업명	추진기간	예산 (백만원)	분야
광주광역시	교통사고 및 청소년 자살예방	7개월(15.4~15.10)	72	재난·안전1
	시내버스 효율적 운영	3개월(14.8~14.10)	0	교통3
	시민의 소리 분석	2개월(14.1~14.2)	29	민원·여론분석
울산광역시	ict융합 및 빅데이터 활용 마스터플랜 수립	8개월(16.3~16.10)	60	컨설팅·연구용역
경기도	공공데이터 개방	7개월(14.4~14.11)	1,447	인프라
	빅데이터 분석사업	10개월(15.2~15.12)	1,350	컨설팅·연구용역
	빅데이터 전문인력양성	5개월(15.5~15.10)	1,170	인프라
	빅파이센터 구축	3개월(15.12~16.2)	200	인프라
	빅포럼	1개월(15.10)	540	인프라
강원도	전통시장분석	8개월(15.9~16.4)	129	경제2
충청북도	관광행정수요조사	4개월(14.7~4.10)	275	관광2
	소상공인 상권분석	8개월(15.4~15.11)	765	경제3
충청남도	백제문화제 감성분석	2개월(15.3~15.4)	0	관광3
전라북도	KTX개통이 지역에 미치는 영향	6개월(15.9~16.2)	0	교통4
	전주시 교통체계 빅데이터 분석	5개월(15.8~15.12)	0	교통5
	한옥마을 관광 빅데이터 분석	5개월(15.8~15.12)	500	관광4
전라남도	호남선ktx 개통에 따른 이용객 형태분석	3개월(15.11~16.1)	0	교통6
경상남도	응급환자 골든타임 확보 방안	9개월(16.4~16.12)	200	복지1
	중국 관광객유치 분석	9개월(16.4~16.12)	200	관광5
제주특별자치도	관광산업 일자리 미스매치 해소	4개월(15.9~15.12)	171	관광6
	내도 관광객 취향 분석	7개월(14.5~15.1)	296	관광7
	온라인면세점 비즈니스 분석	-	25	경제4
(서울)강동구	빅데이터 테스트 베드 구축	2개월(15.7~15.8)	12	인프라
	서울시 플랫폼 고도화	7개월(15.10~16.5)	0	인프라
(서울)종로구	주정차 개선 방안 모색	4개월(15.8~15.11)	비예산	교통7
	홈페이지 분석	2개월(14.5~14.6)	0	컨설팅·연구용역
(광주)광산구	시립 도서관 입지 분석	16.1~	0	복지2
	안전관리 현황 분석	2개월(16.2~16.3)	0	재난·안전2
	어린이 교통사고 현황 분석	15.9~	0	재난·안전3
(경기)부천시	시민불편사항분석시스템(1차,2차)	8개월 (14.4~9/15.9~12)	무응답	민원·여론분석
	웹소셜(SNS) 모니터링 서비스	7개월(16.2~16.9)	30(이용료)	인프라
(경기)시흥시	소셜 데이터 분석	10개월(16.2~16.12)	20	컨설팅·연구용역

지자체	사업명	추진기간	예산 (백만원)	분야
(경기)안산시	대부도 유동인구 등 빅데이터 융복합 분석	6개월(15.12~16.6)	67	관광8
	안산시 유동인구 및 상권분석	4개월(15.6~15.10)	무응답	경제5
(경기)오산시	오산시빅데이터 분석시스템 구축	5개월(16.5~16.10)	552	인프라
(충북)청주시	교통 빅데이터 분석	9개월(14.3~14.12)	150	교통8
	스마트라이프 케어 서비스	12개월(16.1~16.12)	1,349	복지3
	지능형 기술 활용 빅데이터 분석	4개월(15.9~15.12)	120	컨설팅·연구용역
(충북)충주시	모바일 데이터 활용 서비스 인구 분석	4개월(15.10~16.1)	21	컨설팅·연구용역
	불법주정차분석	10개월(14.3~14.12)	20	교통9
	신용카드분석	3개월(15.12~16.2)	20	경제6
(충북)제천시	스마트관광정보구축	9개월(14.5~15.1)	0	관광9
(전북)전주시	한옥마을관광빅데이터분석(전북도 협업)	5개월(15.8~15.12)	0	관광10
(전북)완주군	시장방문객분석	5개월(16.5~169)	10	경제7
	축제및주요관광지분석	6개월(16.4~16.9)	20	관광11
(경북)창녕군	지역관광 및 축제 활성화	6개월(15.6~15.11)	18	관광12
(경남)창원시	빅데이터컨설팅 및 플랫폼개발	12개월(2015년)	436	인프라
(경남)밀양시	ict융합 빅데이터 분석설계	9개월(16.4~16.12)	100	컨설팅·연구용역
(경남)하동군	축제효과분석	15.5~	3	관광13
(경남)함양군	소비행태분석	2개월(15.7~15.8)	3	경제8
	소비행태분석	2개월(14.7~14.8)	1	경제9
(제주)제주시	지역축제효과분석	7개월(15.3~15.9)	3	관광14

라. 빅데이터 사업 추진 애로사항

대부분의 지방자치단체에서 빅데이터 사업의 추진을 위한 전담 조직과 인력이 부족한 것을 주요한 애로사항으로 꼽았다. 이 외에도 예산 부족, 현업부서의 협조 부족, 데이터 확보의 어려움 등이 주된 애로사항으로 지적되었다.

마. 빅데이터 통계 생산을 위한 시사점

전담 조직과 인력, 예산 부족의 상황에서도 광역지자체 수준에서는 절반 정도가, 기초지자체 수준에서는 20% 정도가 빅데이터를 활용한 사업을 추진하고 있었다. 이는 지자체의 의욕이 현실 여건을 앞설 정도로 크다는 점을 보여준다. 바로 이러한 상황에서 빅데이터의 부적절한 활용이나 부정확한 통계생산이 이루어지지

않도록 정부의 적절한 지원이 필요하다는 점이 나타난다.

지자체들이 갖고 있는 관심이 상당한 공분모를 가진다는 점을 전제할 때 각 분야에서 나타나는 전형적인 사업들이 다른 지자체에서도 반복적으로 추진될 가능성이 매우 큰 것으로 예상된다. 예컨대 유동인구 파악을 통한 관광통계의 생산은 축제나 관광명소를 갖고 있는 모든 지자체의 관심 사업이 될 것이며 또한 공표 통계가 될 가능성이 크다. 이와 같은 전형적인 사업들을 중심으로 통계작성 가이드라인 작성을 시작하는 것이 효과적일 것으로 판단된다.

제2절 서울시의 빅데이터 활용 사례

가. 개요

서울시는 기존의 방식으로 해결하기 어려운 도시문제들이 증가하는 한편 과학적이면서 신뢰성 있는 합리적 의사결정도구는 부재한 상황 속에서 방치되거나 버려지는 데이터의 활용을 통한 문제해결을 추구하기 위해 빅데이터를 활용한 행정혁신을 추진하였다. 이를 위해 빅데이터를 활용한 조직 및 업무 프로세스를 정비하였는데 핵심으로는 “서울형 빅데이터 공유·활용 플랫폼 구축” 사업을 추진하였다. 이 사업에서 주로 다루어진 빅데이터는 서울시가 보유한 교통분야 데이터로서 스마트카드, 택시 승·하차, 센서 정보 등이었다. 또한 빅데이터 전문 인력으로 빅데이터 전략을 제시하고 시스템 구축을 지휘하는 사람으로서 “빅데이터 큐레이터”를 양성하였다.

서울시가 그동안 추진하였거나 현재 진행 중인 사업들 중 가장 대표적인 사례로는 다음 표에 제시된 세 가지 사업을 꼽을 수 있다.

<표 45> 서울시 빅데이터 통계 활용 사업 사례

사업명	시기	사용정보	사업목적
올빼미버스(심야버스)	2013. 09	휴대전화 통화이력	심야버스 노선 최적화
노인여가복지시설 입지 선정	2014. 04	거주인구(행정정보), 시설 이용자 정보 시설물정보	신규 노인여가복지시설 입지 최적화
우리 마을 가게 상권 분석 서비스	2015. 12	사업체 등록정보 신용카드 결제 정보 휴대전화 통화이력 행정정보(인구)	영세 소상인의 창업 입지 선정 보조

나. 올빼미버스

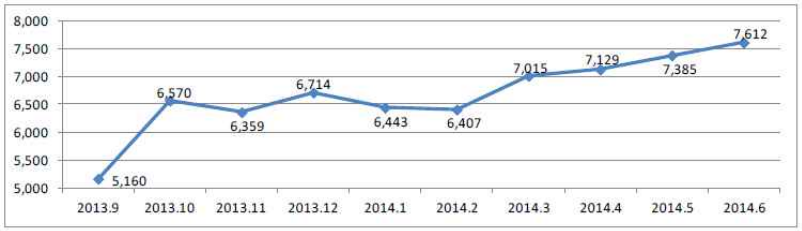
1) 사업 개요

2013년 1월, 서울시는 대중교통 운행이 종료된 심야 시간대의 시민 불편을 줄이기 위해 자정부터 오전 5시 사이에 운행되는 심야 버스를 운행하겠다는 계획을 발표하였다. 동년 4월 19일부터 2개 노선을 3개월간 시범 운영하였고, 동년 9월에 정식 서비스를 개시하였다.

[그림 25] 올빼미버스 노선도

자료: 서울정책아카이브. 2014. Owl Bus Based on Big Data Technology, (https://seoulsolution.kr/en/content/owl-bus-based-big-data-technology)

시범 운영 기간 동안 총 이용자 수는 58,282명을 기록하여 손익 분기점인 2만 명을 초과하였다. 이에 서울시는 노선을 8개로 확장하여 운행하기로 결정하였다. 2014년 6월 현재 일평균 이용자 수는 7,612명이다.



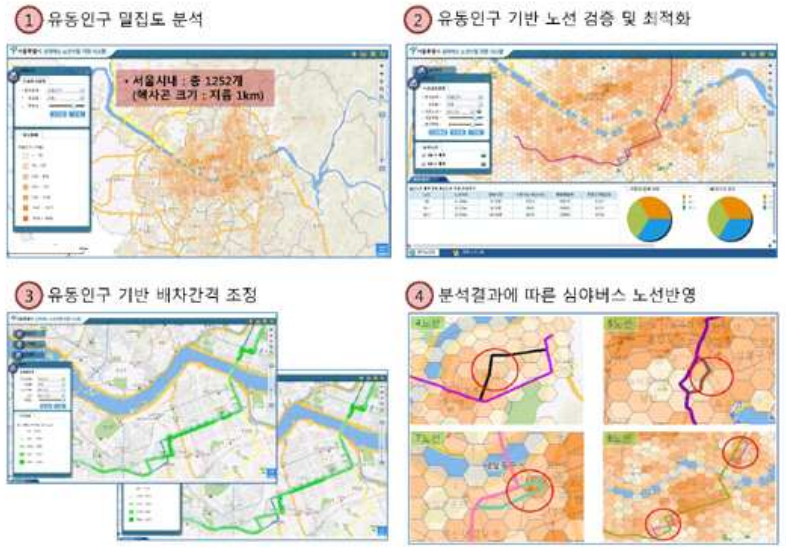
[그림 26] 일평균 올빼미 버스 이용객 수 추이

자료: 서울정책아카이브. 2014. Owl Bus Based on Big Data Technology, (https://seoulsolution.kr/en/content/owl-bus-based-big-data-technology)

2) 빅데이터 활용

시민의 이동 경로는 이동전화의 통화량 데이터 30억 건을 분석하여 계산되었다. 휴대전화 발신 지역을 출발지로 계산하였으며, 도착지는 KT에 등록된 통신요금 납부자의 주소지로 추정하여 분석되었다.

유동인구 분석을 위해 서울시를 반경 1Km 단위의 1,250개 헥사셀로 나누었고, 심야시간 통화량을 통해 각 셀별 심야 유동인구를 추산하였다. 유동인구의 시간/요일별 패턴을 고려함과 동시에 노선 부근의 유동인구에 가중치를 가산하여 각 정류장별 유동인구를 계산, 배차간격을 조절하였다.



[그림 27] 서울시 올빼미버스 노선 도출 과정

자료:
(https://zeronova.kr/2013/08/07/seoul-bus-route-optimization/)

3) 기관 협력

서울시는 심야 유동인구 파악을 위해 KT와 MOU를 체결하여 휴대전화 통화 및 문자메시지 발송 데이터를 입수하였다. 제공된 데이터는 2014년 3월 한달 동안 사용된 휴대전화 통화 기록 데이터이다.

다. 노인여가복지시설 입지 선정

1) 사업 개요

노령인구가 증가함에 따라 노년층을 대상으로 여가, 교육 등을 제공하는 시설의 수요가 증가하고 있다. 하지만 시설의 수요와 공급을 결정하기 위한 데이터가 부족하였는데, 서울시에서는 인구, 교통, 소득 데이터를 통해 신규 시설 입지 후보지를 도출하였다.

연구는 2013년 12월부터 2014년 3월까지 진행되었으며, 노령인구의 시설 이용 현황과 신규 시설 공급 우선순위 및 예상 후보지를 도출하였다. 이 과정에서 서울시는 다음과 같은 통계를 생산하였다.

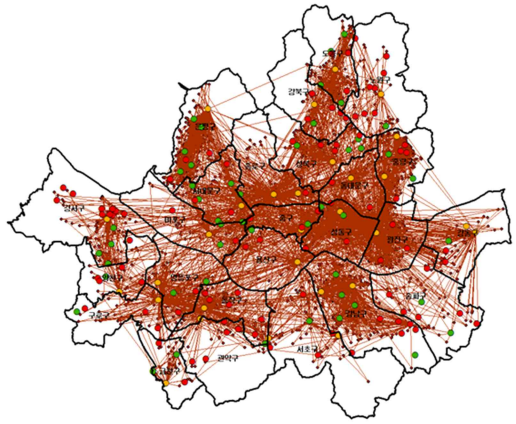
분석 결과 노인 10명 중 9명(89%)은 거주하는 자치구 내 시설을 이용하고 있었고, 서로 다른 두 개 이상의 시설을 이용하는 경우는 4.5%였다. 이에 따라 분석 전에 신규 시설 입지 후보 1위였던 마포구는 강북구와 송파구보다 실수요가 더 적은 것으로 파악되었다.

또한 전체 이용자 중 63.6%가 걸어서 16분~17분 거리에 있는 가까운 시설을 이용하는 것으로 나타났다. 약 20%는 멀어도 좋은 설비를 갖춘 대규모 시설을 선호하는 것으로 나타났다.

시 전체를 작은 구역으로 분할하여 각 구역마다 성별·시간대별·연령별 유동인구, 거주인구, 시설정보를 축적 관리하고 있다.

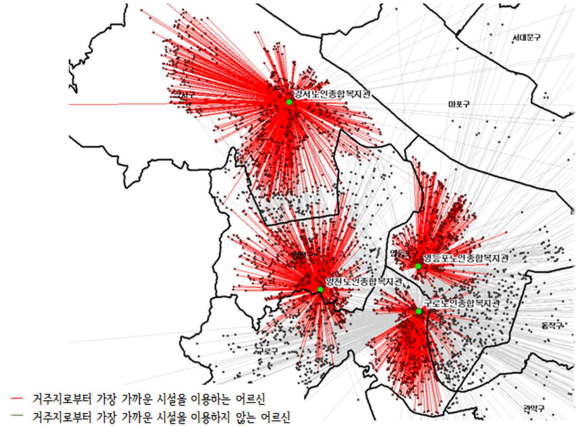
2) 빅데이터 활용

분석 대상은 60세 이상 인구로, 이의 주민등록 분포와 시설 사용 패턴을 정리하여 실제 시설의 수요/공급을 파악하였다. 이동경로는 휴대전화 통신 데이터를 분석하여 파악되었는데, 각 시간/요일별 휴대전화 이용 기록을 분석하여 노령인구의 이동 방향을 추정하였다. 소득은 나이스 평가정보를 기반으로 하여 연소득 평균액을 추정하였다.



[그림 28] 두 개 이상 시설 이용자 분석 결과

자료: 서울시보도자료. 2014. 서울시, "수요가 있는 곳에 정책" 빅데이터 본격 활용



[그림 29] 시설별 이용자 이용현황 분석

자료: 서울시보도자료. 2014. 서울시, "수요가 있는 곳에 정책" 빅데이터 본격 활용

3)기관 협력

올빼미 버스와 마찬가지로 휴대전화 이용 기록 데이터를 KT에 협조를 구해 사용하였다. 이 외에도 노령인구의 소득 수준 추정을 위해 나이스 평가정보의

소득 정보를 사용하였다.

라. 우리마을가게 상권분석 서비스(golmok.seoul.go.kr)

1) 사업 개요

서울시는 영세 소상공인을 위한 창업 위험도 예측 시스템을 2015년 12월부터 운영하였다. 시는 대형 유통 시설이 들어서지 않은 큰 대로변 등의 뒷골목과 같은 영세한 골목상권 1008개를 '서울형 골목상권'으로 규정하고, 해당 지역 내의 중국집, 편의점 등 43개 생활밀착업종의 2,000억개 빅데이터를 기반으로 서비스를 구축하였다.

상권 분석 서비스는 크게 세 가지 서비스를 제공한다. 예비창업자를 대상으로 한 '상권신호등서비스', '맞춤형 상권검색서비스'와 더불어 기존 자영업자를 위한 '내 점포 마케팅 서비스'다.

'상권신호등서비스'는 분기별 상권 데이터를 바탕으로 신규창업 위험도를 4단계의 색깔 표시(주의-파랑, 의심-노랑, 위험-주황, 고위험-빨강)해 해당지역의 창업 위험도와 폐업신고율, 3년 내 폐업신고율, 평균 폐업기간, 점포증감율을 단계별로 확인할 수 있는 서비스다.

[그림 30] 상권 신호등

자료: 우리 마을 가게 상권분석 서비스 홈페이지 (http://golmok.seoul.go.kr/sgmc/main.do)

'맞춤형 상권검색서비스'는 1,008개의 골목상권 중 관심 있는 골목상권의 점포 수, 점포 당 평균 매출액, 일평균 유동인구, 창업생존율, 과밀지수 등 구체적인 상권리포트를 맞춤형으로 검색할 수 있다.

'내 점포 마케팅서비스'는 1,008개 상권에 포함돼 있지 않은 지역이라도 특정 지역의 상권을 임의로 지도상에서 설정(반경 100~1000m 가능)하면 영역 내에서 성별/연령별/요일별/시간대별 유동인구 추이와 주요 집객시설, 아파트 세대 수 등을 분석한 보고서를 제공해준다.

2) 빅데이터 활용

분석대상 업종은 43개 업종[205]으로 서울시내 1,008개 영역이 분석 대상이다. 분석에 사용하는 데이터는 모두 32종이며 이하와 같다.

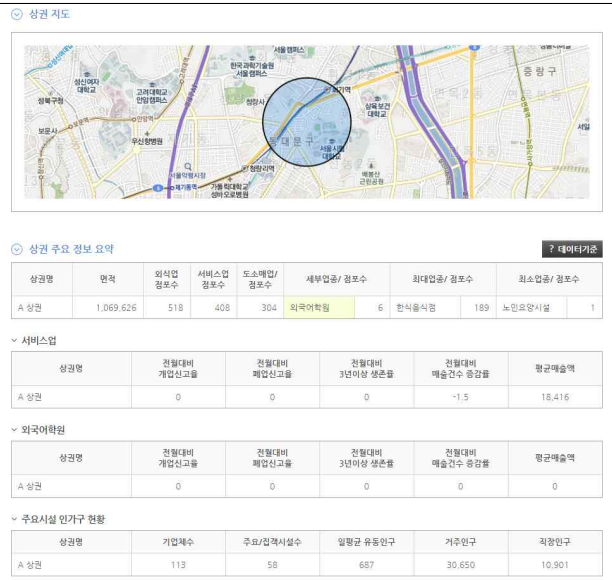
<표 46> 골목상권분석 주요 DB

데이터 명	주요내용	갱신주기	출처	제공 지역
1.상가/업소 정보	업종별 업소정보(업종, 주소, 전화번호)	월	나이스 평가정보	서울시 전역
2.인허가 업소 정보	- 인허가 대상 업종 업소 정보 - 상권별 개폐업률 산출 - 상가 이력정보	월	서울시	서울시 전역
3.사업자 등록(휴폐업) 정보	- 휴, 폐업 신고 사업자 정보(상가/업소,기업DB 보강)	분기	나이스 평가정보	2개 자치구
4.임대시세	- 상가 임대시세 조사 자료 - 행정동별, 자치구별 상가 임대시세 정보	분기	한국감정원	서울시 전역
5.골목상권영역	- 서울시 골목상권 1,000여 개 영역 - 골목상권 단위 정보	년	서울시	서울시 전역
6.발달상권영역	- 서울시 주요 발달상권 - 발달상권 단위 정보	년	중소기업청	서울시 전역
7.매출/소비정보(BC카드)	- 신용카드 매출정보 - 블록별, 상권별 매출정보(매출액, 거래건수) - 성/연령대별, 시간대별, 요일별 거래패턴 정보	월	BC카드	서울시 전역
8.매출/소비정보(신한카드)	- 신용카드 매출정보 - 블록별, 상권별 매출정보(매출액, 거래건수) - 성/연령대별, 시간대별, 요일별 거래패턴 정보	월	신한카드	서울시전역
9.카드비중추정 정보	- 지역별 업종별 신용카드사별 점유비 - 신용카드 정보와 융합하여 추정매출 산출	년	나이스 평가정보	2개 자치구

205) 외식업(10개), 서비스업(22개), 도소매업(11개)

데이터 명	주요내용	갱신주기	출처	제공 지역
10.현금비중추정 정보	- 지역별 업종별 현금영수증 및 현금 비중 - 신용카드 정보와 융합하여 추정매출 산출	년	나이스 평가정보	2개 자치구
11.길단위 추정 유동인구 정보	- 성, 연령, 요일, 시간대별 유동추정 인구 - 통신사 데이터, 집객요인, 교통카드 데이터, 서울시 유동인구 조사자료(2014)를 융합하여 도로 단위의 유동인구 추정 정보	월	서울시	서울시 전역
12.유동인구	- 이동통신 통화량기반 유동인구 정보(50x50 Cell)	월	SKT	서울시 전역
13.택시통행량	- 주요 도로단위 시간대별 택시 통행량	월	서울시	서울시 전역
14.교통카드 정보	- 지하철 및 버스정류장별 승하차, 유동인구수	월	스마트카드사	서울시 전역
15.도보통행량	- 도보 가능 도로별 통행량 추정 정보	년	오픈메이트	서울시 전역
16.기업 DB	- 사업체 정보(규모, 주소, 산업분류, 종업원 수 등)	분기	나이스 평가정보	서울시 전역
17.사업체 통계 DB	- 집계구별 업종별 사업체통계	년	통계청	서울시 전역
18.사업체 조사 DB	- 사업체 총 조사 데이터	년	서울시	서울시 전역
19.주거인구	- 행정구역별 주민등록 통계 데이터를 건물단위별 가구수 및 성별/연령대별 인구수 추정	반기	행정자치부	서울시 전역
20.직장인구	- 50m Cell 단위의 성/연령별 직장인구 정보	반기	나이스 평가정보	서울시 전역
21.직업직종	- 행정구역 단위 업종별 종사자 통계 정보	년	통계청	서울시 전역
22.아파트 DB	- 아파트 단지/동 단위 가구수 정보 - 면적별/기준시가별 가구수 정보	년	오픈메이트	서울시 전역
23.소득데이터	- 블록단위로 가공된 성별 / 연령대별 10분위 기준 소득추정액 정보	반기	나이스 평가정보	서울시 전역
24.소비특성 데이터	- 블록별 소비유형별 비율(식료품, 의류 및 신발,가사용품, 의료, 탈것, 여가, 문화, 교육, 커피, 주류)	반기	나이스 평가정보	서울시 전역
25.블록데이터(블록영역)	- 서울시 6만6천 여개의 블록영역	년	오픈메이트	서울시 전역
26.블록유형화(블록속성데이터)	- 서울시 6만6천 여개의 블록에 대한 배후지 속성 정보	년	오픈메이트	서울시 전역
27.건물 DB	- 건물의 용도, 층수, 면적, 주출입구, 건축일자, 주차장여부, 접도정보, 토지의 형상 등(새주소건물+건축물대장)	반기	서울시	서울시 전역
28.버스정류장	- 시내 버스 정류장 정보(버스노선, 위치정보)	월	서울시	서울시 전역
29.지하철역	- 지하철역사 정보(노선, 위치정보)	년	서울시	서울시 전역
30.주요/집객 시설	- 주요/집객시설 구분 및 위치정보(관공서, 금융기관, 병권, 학교, 유통점, 문화관광 / 영화관, 숙박시설, 교통관련 시설)	반기	각급기관	서울시 전역

데이터 명	주요내용	갱신주기	출처	제공 지역
31.도로명 주소 도로링크	- 도로명 주소 지도 데이터 - 도로구간 데이터	반기	서울시	서울시 전역
32.SNS 트랜드 데이터	- 지역별 업종별 SNS 트랜드 정보 - 업종 및 지역 관련 연관어, 인기 점포	월	다음소프트	서울시 전역



[그림 31] 내 점포 마케팅 리포트 예시

자료: 우리 마을 가게 상권분석 서비스 홈페이지 (http://golmok.seoul.go.kr/sgmc/main.do)

상기 그림의 경우, 동대문구 전농동 주변의 임의의 장소를 원형으로 선택하여 외국어 학원 입지를 파악한 것이다. 해당 상권에 이미 입지해 있는 외국어 학원 수와 지역의 인구 및 직장인 데이터를 파악할 수 있다.

각 업종별 매출액 추이 및 연령별, 성별 매출액 역시 파악가능하며, 이는 신용카드 결제 정보에 바탕을 둔 것이다. 또한 휴대전화 착발신 이력정보를 바탕으로 해당 지역의 시간대별 유동인구 정보 역시 연령별, 성별로 조회할 수 있다.

3)기관 협력

사용한 정보는 종류에 따라 월별, 분기별, 년도별로 갱신되는데, 카드 결제정보는 BC, 신한카드로부터 입수하여 매달 최신화 되며, 점포 정보는 나이스 평가정보로부터, 유동인구는 SKT로부터 매달 제공된다.

주거인구는 행정자치부 데이터를 이용하며, 성별 인구, 직장인 인구 정보는 통계청과 나이스 평가정보로부터 입수한다.

마. 서울시 빅데이터 캠퍼스 사업

1) 개요

서울시는 빅데이터를 공개하여 연관산업의 진흥을 위해 빅데이터 캠퍼스(bigdata.seoul.go.kr)를 설치, 운영할 계획을 발표한 바 있다(시장방침 제342호). 15년 7월에 "16년 시민중심 빅데이터 플랫폼 사업 시행계획 수립"을 시행하였고, 동년 9월에 건강보험심사평가원 등의 타 기관의 데이터 플랫폼을 벤치마킹 하였다. 11월 9일에는 빅데이터 분석 및 상용 SW사용을 위해 국민대, 성균관대, 세종대, SAS코리아 등과 업무협약을 체결하였다. 이를 통해 서울시는 연구소, 시민단체, 개인 등이 함께 참여하여 사회문제를 해결하는 도구로 이용할 것을 천명하였다. 빅데이터 캠퍼스는 2016년 7월에 개소하였다.

[그림 32] 서울시 빅데이터 캠퍼스 비전

자료: 서울특별시 빅데이터 캠퍼스 홈페이지
(https://bigdata.seoul.go.kr/main.do)

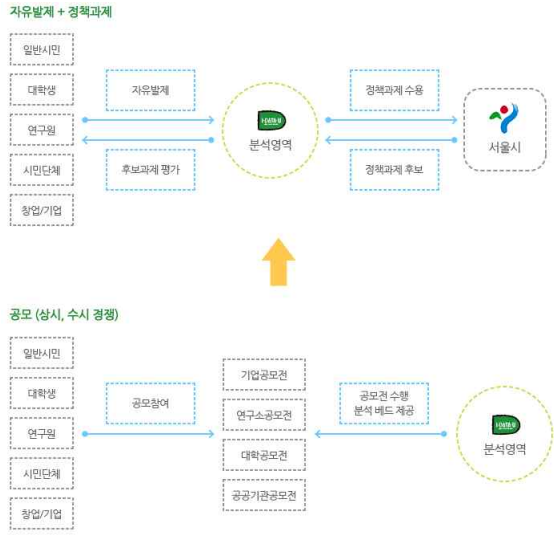
2) 주요 내용

서울시는 협약을 맺은 대학교(성균관대, 세종대, 국민대)에서 빅데이터 강의 및 실습 정규과정을 운영하며, 빅데이터 관련 세미나, 학회 등을 개최하기로 하였다. 이 외에도 업무협약을 맺은 기업(신한카드 등)에 교육을 요청하여 강의나 분석실습을 제공하는 주간 혹은 월간 강의를 개설하는 것도 추진하고 있다.

서울시에서 보유하고 있는 빅데이터를 빅데이터 캠퍼스 내에서 이용가능한 장소를 제공하고 개인 또는 그룹이 이용신청을 통해 빅데이터 캠퍼스에 입주하여 분석 후 결과물만 가지고 떠나는 것을 목표로 한다. 연구 성과는 서울시 정책 수립에 반영한다.

빅데이터 캠퍼스 센터는 분석실을 80석 규모(557m^2)로 갖추고 있으며, 2016년 12월에는 클라우드 기반 서비스를 오픈하는 것을 계획하고 있다. 제공하는 데이터는 열린데이터광장에서 제공하고 있는 데이터와 연계하는 것이 대부분이다.

2016년 8월 현재 빅데이터 캠퍼스 이용 신청 건수는 80여건을 기록하고 있다.



[그림 33] 캠퍼스 협치 모델

자료: 서울특별시 빅데이터 캠퍼스 홈페이지
(https://bigdata.seoul.go.kr/main.do)

바. 빅데이터 통계 생산을 위한 시사점

서울시의 사례를 통해서 빅데이터를 활용한 통계생산이라는 관점에서 다음과 같은 시사점들을 제시할 수 있다.

첫째, 빅데이터를 활용한 통계가 정책 서비스 개발에 매우 효과적으로 활용될 수 있다는 점을 실증하였다. 특히 기존의 조사로는 파악하기 힘들었던 유동인구와 이동 양태를 빅데이터의 활용으로 파악하고, 이를 기존의 통계자료와 연계했을 때 정책 서비스를 위한 매우 강력한 도구가 제공됨을 보여주었다. 빅데이터 통계를 통해서 체감도가 매우 높은 서비스 개발이 이루어졌다.

둘째, 정책 서비스 개발을 위한 빅데이터 활용 과정에서 많은 통계가 생산되지만, 대부분의 통계는 그 자체로서 공표되지 않고, 정책 개발을 위한 인풋으로만 활용된다. 정책 개발의 인풋으로만 존재한다고 해서 통계품질의 중요성이 사라지는 것은 아니다. 하지만 현재와 같은 승인제도의 프로세스 대상에 포함시킬 가능성은 거의 없어 보인다. 이 과정에서 강조되어야 할 통계 품질이 무엇이며 어떻게 접근하는 것이 타당할지에 대한 새로운 모색이 요구된다.

셋째, 빅데이터를 활용한 통계의 생산 단위가 과거의 공공 통계와는 달리 행정구역별 단위 외에 다양한 단위로 생산된다. 행정구역보다 훨씬 작은 단위의 기준으로 통계가 생산되기도 하고, 서비스 이용자의 요구에 따라 그 때 그 때 불특정 공간 단위 기준의 통계가 생산되기도 한다. 따라서 행정통계의 특정한 공표 단위를 정형화하기 힘들다.

넷째, 통계 생산을 위한 빅데이터 활용의 제한점이다. 서울시의 경우 민간 기업의 빅데이터를 무상이나 저렴한 가격으로 공급 받아 사용하였으나 이와 같은 우호적인 여건이 1) 장기적으로 지속될지, 2) 다른 서울시 기관이나 다른 서울시 통계 생산에도 적용될지, 3) 다른 지자체나 정부기관에도 적용될지 현재로서는 판단하기 어렵다. 대체로 반응은 회의적이다. 그렇다면 이러한 문제를 어떻게 해결할 것인가? 결국 빅데이터의 사용권이 독점적으로 민간 업체에 귀속되는 현재의 법적 상황에 대한 근본적인 재고가 없이 안정적인 빅데이터 활용의 확산은 어려우리라 생각된다. 정부가 정부 3.0, 창조경제의 기조에 바탕을 두고 이 문제의 해결에 적극적으로 나서야 한다.

제3절 부산시의 빅데이터 활용 사례

가. 개요

부산시는 2016년 3월부터 '부산 도시서비스분석 정보시스템(sgis.busan.go.kr)'을 구축하여 서비스를 개시하였다. 부산시의 빅데이터 활용 사례가 갖는 의의는 다음과 같은 두 가지 점에서 찾을 수 있다.

첫째, 빅데이터를 활용한 통계 생산 서비스의 종합시스템을 구축하였다. 이전에 서울시가 빅데이터를 활용한 개별 사업들을 수행하였으나 여기에서는 정책 기능에 따라 빅데이터 사용이 분절화 되었다. 이와 달리 부산시의 시스템은 관광객 유입규모, 창업 위치 선정, 기업 마케팅 장소 선정, 버스 노선이나 CCTV 선정 등의 다양한 기능을 한 시스템 안에서 구현한 것이 특징이다. 이와 같이 지역자치단체 차원에서 종합적으로 서비스를 개시한 것은 최초로 평가된다.

둘째, 모바일 빅데이터 활용을 통해서 '서비스 인구'라는 새로운 개념을 개발하였다. 이전에도 서울시 등 여러 사례에서 유동인구 분석 자료를 사용하였으나 유동인구 산출에 대해서 구체적인 정보가 제시되지는 않았다. 부산시와 같이 인구산출 방법을 명시한 것은 최초로 평가된다.

나. 부산 도시서비스분석 정보시스템 구성

추진배경은 관광, 교통, 환경 안전 등 다양한 분야의 행정에서 소지역별로 세분화된 현주인구에 대한 통계 제공을 필요로 한다는 점이다. 여기서 현주인구란 조사시점에 개개인이 위치하고 있던 지역을 기반으로 조사 집계된 인구자료이다. 하지만 상주인구만 집계하는 현재 인구통계는 이러한 정책적 요구나 시민의 필요에 부응하지 못하고 있다.

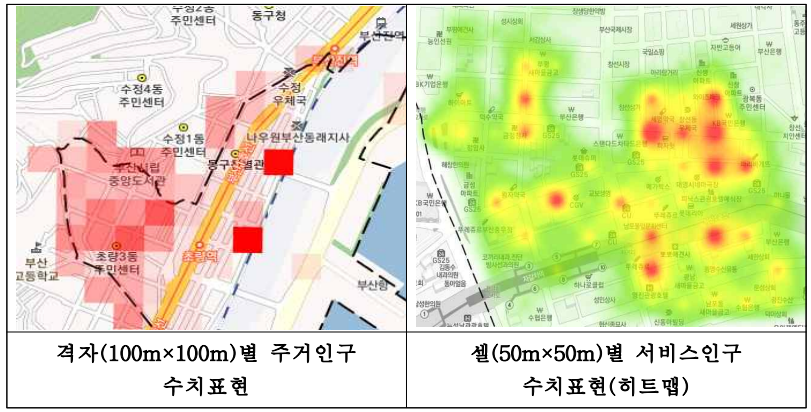
또한 상주인구가 지역의 서비스 수요 공급을 정확하게 반영하지 못할 경우 서비스인구(Service Population)를 작성하라는 UN권고(2008)를 반영하였다.

[그림 34] 부산서비스인구 통계 산출 과정

주요기능으로는 주거인구, 서비스인구, 창업지선정, 축제유입, 버스노선, CCTV 분석 및 시각화 사업을 수행하였다.

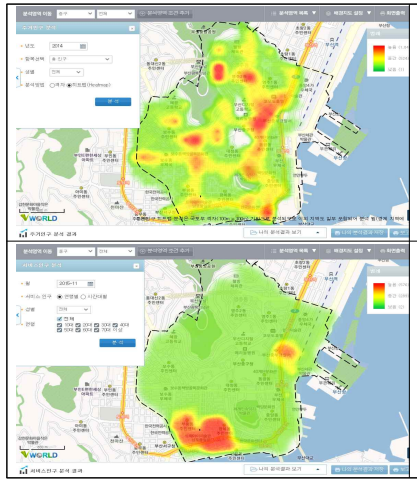
[그림 35] 부산서비스인구통계 이용 구조

모바일 데이터의 사용에 따라 주거인구보다 더 적은 공간 단위에서 유동인구의 집계가 가능해졌다.



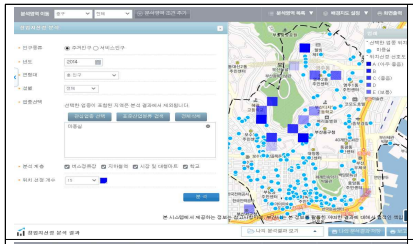
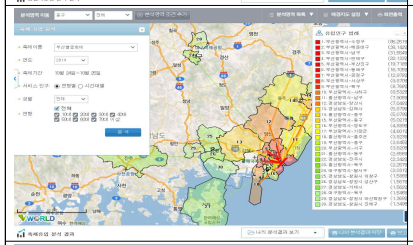
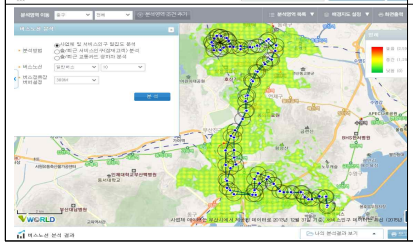
격자(100m×100m)별 주거인구 수치표현	셀(50m×50m)별 서비스인구 수치표현(히트맵)

지도서비스로는 서비스인구(2종), 주거인구(5종), 산업 및 경제(10종), 도시기반(5종), 사회(6종) 등 총 28종의 통계 주제에 대한 통계지도서비스를 제공하고 있다. 주요한 서비스 제공은 다음과 같다.



주거인구
부산에 상주하고 있는 총 인구, 유소년 인구, 생산가능 인구, 고령 인구, 유아 인구, 초등학생 인구, 중학생 인구, 고등학생 인구, 10세별(20대, 30대, 40대, 50대, 60대, 70대) 등 성·연령별 주거인구 자료 제공

서비스인구, 유형별서비스인구
서비스인구(유동인구) 성/연령, 시간대별 데이터를 통한 집중지역 유동인구를 정밀하게 분석하여 대시민에게 홍보 및 마케팅 수단으로 제공

<table>
<tr><td>  </td><td>창업지선정
시민에게 상권입지 선정 컨설팅 비용 부담을 최소화하고 입지선정 도움을 주고자 사용자가 업종 선택에 따른 위치 정보 제공</td></tr>
<tr><td>  </td><td>축제유입
지역별로 축제기간 중 주요 유입지 및 관광코스를 서비스인구로 정밀하게 분석하여 효율적인 홍보 및 마케팅에 활용할 수 있는 자료 제공(교통혼잡 및 주차 공간 부족해소, 관광 마케팅, 관광상품개발 가능)</td></tr>
<tr><td>  </td><td>버스노선 ▸ 내부이용자(공무원) 전용
버스노선, 정류장, 승/하차 정보, 사업체 정보 등 시민을 위한 정책과 제도를 출/퇴근 유동인구 패턴을 통한 정책자료 제공</td></tr>
</table>

다. ‘서비스인구’ 개발

부산시는 모바일 빅데이터를 이용하여 ‘서비스 인구’ 개념을 개발하고 산출 방법을 명시하였다. 서비스 인구는 모바일 빅데이터를 활용해 1시간별로 각 지역의 유동인구를 파악하고, 주민등록상 상주인구와 방문인구를 합친 개념이다. 서비스인구 산출 과정은 다음과 같다.

먼저, 부산지역 각 교환기의 형태를 분석하여 부산지역 인구수를 산정한다. 이 때 부산지역만을 관할하는 교환기의 경우 100% 적용을 하고, 부산 지역 및 타 지역을 중복으로 관할하는 경우 지오비전 시간대별 유동인구를 기준으로 일별, 시간별 부산지역의 비율 값을 구하여 교환기 데이터 비율로 산정한다. 이를 바탕으로 50m^{2}의 pCell 공간데이터를 생성한다.

다음으로 보정계수에 의한 서비스 인구를 산출한다. 그 과정을 보면

1) 교환기 데이터를 바탕으로 부산시 SKT 고객을 산정하고,
2) 총 인구수에 통신사 비율 50.1%를 적용하며,

3) 5세 이하, 85세 이상 인구를 휴대폰 미소지자로 산정하며,
4) 거주 불명등록자를 산정하고,
5) 시간대별 power off 비율을 산정한다.

라. 국가 공식통계 승인

부산서비스인구통계는 2014년 10월 13일에 국가 공식통계(제01402호)로 인정받았다. 이는 빅데이터를 활용한 통계 중 국가 공식통계로 승인된 첫 사례이다. 부산서비스인구통계는 안전행정부의 정부3.0 우수사례로 선정되어 국비 6억 5천만 원을 지원받았으며, 이를 기반으로 서비스인구를 지도상에 소지역 단위(50m X 50m)에 실시간으로 표기하는 것을 목표로 하고 있다. 자료 구축에 활용되는 데이터는 SK텔레콤 지오비전의 데이터이며, 이는 1시간에 4Gb에 달한다.

마. 빅데이터 통계 생산을 위한 시사점

부산시의 사례를 통해서 빅데이터를 활용한 통계생산이라는 관점에서 다음과 같은 시사점들을 제시할 수 있다.

첫째, 부산도시서비스분석시스템은 도시정부의 행정 수요, 도시경제 각 주체의 요구에 부응하는 모바일 빅데이터 활용의 전범으로 볼 수 있다. 서울시 등 이전 사례가 개별 기능에 한정하여 빅데이터를 활용했다면 부산도시서비스분석시스템은 공공과 민간의 다양한 요구를 동시에 충족하는 통합적 시스템을 구축했다는 점에서 진일보한 성과로 평가된다. 도시경제나 도시마케팅 차원에서 필요한 핵심 서비스를 구현했다는 점에서 긍정적이다. 또한 공무원 전용 서비스를 개발하여 통계의 정책 활용 가능성을 높였다는 점도 긍정적이다. 부산시 시스템이 안정적으로 운영될 경우 다른 광역단체에 확산될 가능성이 매우 커 보인다.

둘째, 유동인구 통계의 중요성을 제시하였다. 기존의 통계에 비추어 모바일 빅데이터의 활용 가능성이 가장 두드러진 것은 유동인구 통계이다. 부산시가 UN의 권고안을 배경으로 서비스 인구 개념을 도입하고 이를 산출하는 과정을 제시한 것은 긍정적으로 평가될만하다. 하지만 현재 도입된 보정과정에 대한 통계학적 검토가 보완되어야 할 것으로 보인다.

셋째, 빅데이터 통계 거버넌스의 가능성을 제시하였다. 서울시의 사례는

훨씬 더 많은 민간 데이터를 활용하여 공공 데이터와 연결함으로써 포괄성 측면에서는 우위에 있으나 그 복잡성으로 인해 다른 지자체에 적용하기에는 어려운 점이 있어 보인다. 부산시도 앞으로 카드사와 연계해 매출 발생 현황 정보와 결합하여 상권 분석에 필요한 데이터를 제공할 예정이다. 하지만 부산시의 경우 통신사의 모바일 데이터만으로도 이미 다양한 기능을 제공하고 있어서 효율성이라는 점에서 긍정적이고, 다른 지자체들이 수용하기에도 서울시의 모델 보다 용이할 것으로 보인다. 하지만 SKT와의 협력의 기반이나 재정적인 측면에서의 효과성은 장기적으로 검증되어야 할 것으로 판단된다.

3. 공공통계 관련 민간부문 빅데이터 활용 통계생산 사례

제1절. 신용카드사

가. 개요

신용카드사들은 자사의 결제 데이터 베이스를 바탕으로 개인고객을 대상으로 하는 최적화된 상품을 출시하거나(신한카드 코드 9), 소비 패턴에 맞는 최적 지역/점포를 추천하는 서비스를 제공하고 있다(삼성카드 플레이스 S 등).

<표 47> 카드사별 주요 빅데이터 이용 사례

상품명	시기	주관	사용정보	사업목적
Hyundai card X big data report	2012. 12	현대카드	지역별, 업종별 신용카드 결제 이력	카드 결제 정보를 통한 컨설팅 자료 제공
플레이스 S	2013. 10	삼성카드	소비자 결제 기록 점포별 총 승인 금액 정보	개별 소비자의 소비 패턴에 맞는 추천 점포 정보 제공
코드 9	2014. 05	신한카드	성별, 연령대별 신용카드 사용 내역	성별, 연령별, 소득수준별 최적 카드 상품 제공

카드사의 결제정보 빅데이터 활용 방식을 종합하면, 삼성카드, 현대카드의 경우 소비자의 결제 패턴을 종합하여 특화된 할인 가맹점/사업체를 소개하는 방식에 주력한다. 신한카드는 결제 패턴에 특화된 카드를 발행하여 소비자의 결제를 유도하는 데에 집중하고 있다.

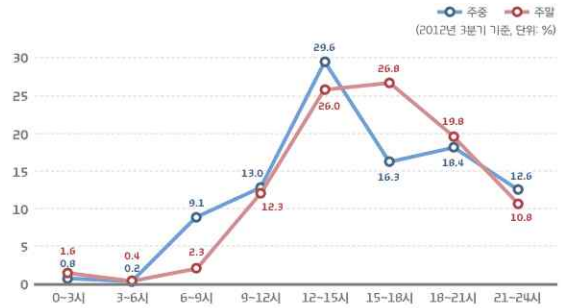
하지만 상기 두 가지 서비스를 대부분의 카드사가 시행하고 있기 때문에 큰 차이는 없으며, 회사별로 주력하는 서비스가 다를 뿐이다.

나. 현대카드

현대카드는 결제 정보를 종합하여 “Hyundai card X big data” 리포트를 발간, 지역별, 성별, 연령별에 따른 지출패턴을 발표하였다. 이는 지역별 점포개설 및 산업별 컨설팅 자료로 활용하기 위한 것으로, 통념과는 다른 소비패턴을 보이는 산업을 집중 조명하는데 주력하였다.

예컨대 통념상 항구도시인 부산에서 일식 소비 비중이 높을 것으로 생각하나 실제 외식 결제 중 양식(10.3%)이 일식(3.2%)보다 더 많이 결제되는 것을 밝히거나[206], 추운 강원도에서 겨울 의류 매출이 전국에서 가장 낮다는 점을

밝히기도 하였다[207]. 커피의 경우, 평일 점심시간대에 남성 회사원의 저가형 커피에 대한 수요가 많음을 밝혀내어 편의점 커피 마케팅에 활용하기도 하는 등 결제 정보를 바탕으로 하여 컨설팅을 제공하는 모습을 보였다.

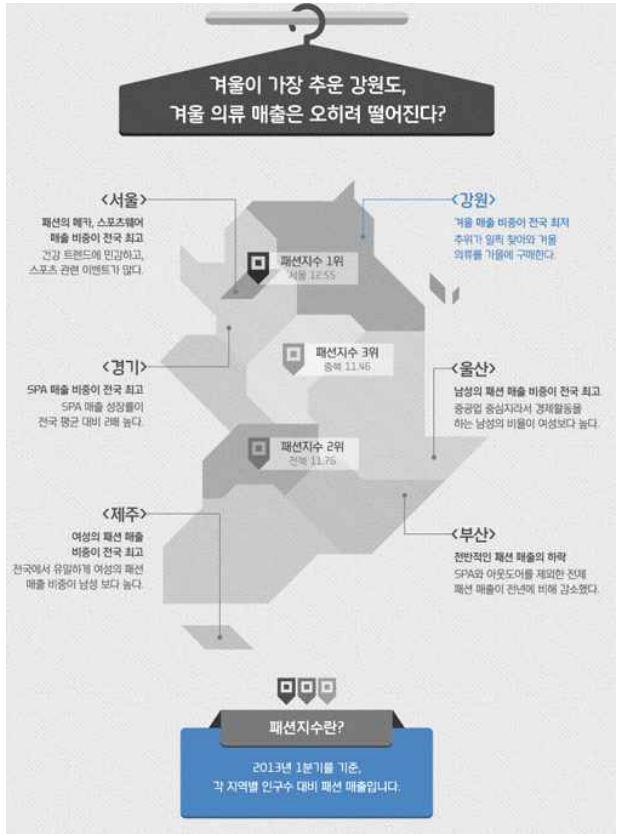


[그림 36] 커피전문점 이용시간대(현대카드)

자료: 현대카드. 금융공학 스토리 커피전문점에는 남성보다 여성이 많다?. 2012. (http://finance.hyundaicardcapital.com/269)

206) http://finance.hyundaicardcapital.com/category/%EC%83%81%ED%92%88%EC%97%90%20%EB%8B%B4%EA%B8%B4%20%EA%B3%BC%ED%95%99/%EC%B9%B4%EB%93%9C%20%EA%B8%88%EC%9C%B5%EC%83%81%ED%92%88?page=11#search

207) http://finance.hyundaicardcapital.com/category/%EC%83%81%ED%92%88%EC%97%90%20%EB%8B%B4%EA%B8%B4%20%EA%B3%BC%ED%95%99/%EC%B9%B4%EB%93%9C%20%EA%B8%88%EC%9C%B5%EC%83%81%ED%92%88?page=11#search



[그림 37] 지역별 겨울 의류 매출(현대카드)

자료: 현대카드. 금융공학 스토리 겨울이 가장 추운 강원도, 겨울 의류 매출은 오히려 떨어진다?. 2013. (http://finance.hyundaicardcapital.com/269)

다. 삼성카드

삼성카드는 소비자의 결제 기록을 토대로 자주 방문하는 지역을 파악하고, 각 지역별로 매출 상위 점포의 정보를 종합하여 “지역 대표 가맹점” 리스트를 소비자에게 전달하였다(플레이스 S 서비스).

[그림 38] 삼성카드 플레이스 S

자료: 삼성카드. 플레이스S 홈페이지 https://www.samsungcard.com/personal/services/place-s/UHPPBE0504M0.jsp?click=gnb_benefit_places)

라. 신한카드

신한카드는 자사 고객 2,200만 명의 소비 패턴을 정리하여 각 세대/계층별로 특화된 카드 상품을 개발하였다(code 9). 남성, 여성별로 각기 9개 카드가 있으며, 이는 각기 사회초년생에서 장년층에 이르는 세대별, 소득수준별로 구성된 것이다.

[그림 39] 신한카드 코드 9 카드

자료: 신한카드 홈페이지 (https://www.shinhancard.com/conts/person/news/1236317_14531.jsp)

마. 시사점

카드사에서의 빅데이터 활용 이슈는 활용 여부 보다는 기술적 문제에 집중되어 있다. 예컨대 “요식업”으로 등록되는 경우 체인점은 체인본부가 가맹점으로 조회된다. 이를 필터링하지 않고 소비자에게 제공할 경우 “비알코리아(주) 00점”과 같은 불명확한 점포 정보가 제공될 뿐이다.

또한 각 지역별로 상위 결제 점포를 선정하여 리스트화 할 경우 거대 기업의 본점이나, B2B서비스를 주로 하는 업체 정보가 노출되기도 한다. 이러한 오류의 필터링을 위해서는 등록 점포를 세분화할 필요가 있다. 일부 카드사에서는 이러한 필터링 작업을 수기로 하는 등 빅데이터의 활용이 비능률적이기도 하다.

이하 그림은 소비자에게 맞춤형 정보를 제공하는 삼성카드의 m포켓 화면이다. 여기서도 볼 수 있듯이 매출 최상위권 업체들이 대기업 본사거나, 체인 본부의 이름으로 제공되어 정보전달의 의미가 약하다.

[그림 40] 삼성카드 m포켓 사례

자료: 삼성카드 m포켓 어플리케이션 캡처

제2절 SKT 지오비전

가. 개요

SK텔레콤 KT LG유플러스 등 국내 이동통신 3사가 미래 성장 동력으로 빅데이터(Big data), 사물인터넷(IoT), 클라우드(Cloud) 등을 아우르는 이른바 'BIC' 사업에 투자를 강화하고 있다. 이동통신사들은 과거 음성통화 위주의 통신시장에 한계를 느끼고 최근 사업 다각화를 추진 중이다.

통신 사업은 스마트폰 등 모바일 기기의 진화로 데이터 중심으로 빠르게 이동하고 있다. 소비자들의 데이터 사용량이 이동통신사의 수익을 좌우하는 시대다. 세계적으로 이동통신사들이 BIC 사업에 투자를 강화하고 있는 것도 이 같은 맥락에서다. 글로벌 BIC 시장은 빠르게 성장하고 있다. 시장조사기관 IDC에 따르면 세계 빅데이터 시장은 지난해 129억달러(약 14조5900억원)에서 2017년 311억달러(약 35조1700억원) 규모로 커질 전망이다. 글로벌 IoT 시장은 작년 6560억달러(약 742조원)에서 2020년에는 1조7000억달러(약 1922조8700억원) 규모로 급성장할 것으로 예상된다.

통신사의 빅데이터를 활용한 서비스 상용화 중 가장 두드러진 것이 상권 분석이다. SK텔레콤은 인터넷 게시글 등 빅데이터를 분석해 기업에 제공하는 '스마트 인사이트' 사업을 벌이고 있다. 스마트 인사이트로 상품과 서비스 등을 분석하면 소비자 호감도와 여론 점유율, 경쟁사 상품 비교 등이 가능하다는 게 회사 측 설명이다.

SK텔레콤은 유동 인구, 부동산 관련 데이터 등을 분석해 제공하는 '지오비전'이란 서비스도 운영하고 있다. 지오비전을 활용하면 인구 이동 트렌드와 업종별 매출 정보 등을 결합한 상권 분석 결과를 얻을 수 있다. 회사 관계자는 "민간뿐 아니라 공공정책 수립 등에도 활용할 수 있는 데이터"라고 소개했다.

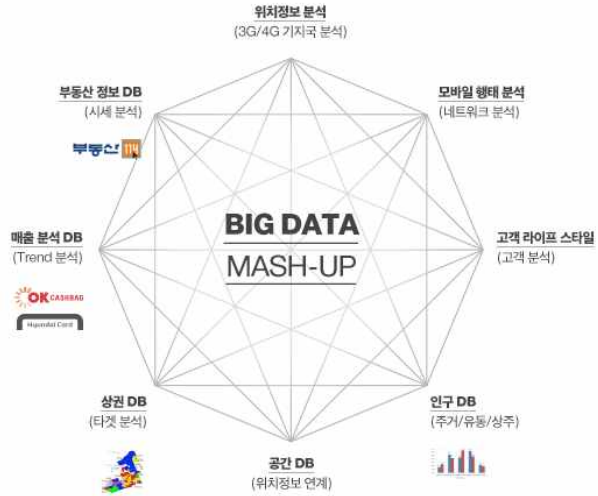
KT는 서울시와 협업해 심야버스 노선도를 설계하는 데 빅데이터를 활용했다. 기지국 데이터 등을 이용해 심야시간에 사람들이 몰리는 곳을 분석한 뒤 이를 반영한 버스노선을 설계했다. KT는 지난해 발생한 조류독감(AI) 농가 데이터를 분석해 AI 확산을 막는 데도 도움을 줬다. 회사 관계자는 "국내 최고의 빅데이터 분석 기업인 KT 자회사 넥스알(NexR) 등과 함께 사업을 확대 중"이라고 설명했다.

LG유플러스는 미디어 분야에서 빅데이터를 활용하고 있다. 'LTE비디오포털' 등에서 시청자에게 최적화된 큐레이션 서비스를 제공하기 위해 빅데이터 분석을 이용 중이다. 개인들의 생활패턴을 분석해 위치와 시간에 따라 다양한 생활정보를 맞춤형으로 제공하는 'U스푼'과 같은 서비스도 운영하고 있다.

이와 같이 통신사들의 다양한 서비스들 중에서 현재 시장에서 가장 큰 영향을 가지는 것은 SK텔레콤과 현대카드 등 9개사가 g-CRM(지역기반 고객관리시스템) 기반 기업 솔루션 서비스로 2010년 출시한 지오비전이다.

나. 데이터 구성 협력 관계

SKT지오비전의 데이터는 현대카드, NICE 신용평가정보, SK마케팅앤컴퍼니, 한국생산성본부, KIS정보통신, 선도소프트, 부동산114, 아이엘엠소프트 등의 8개 파트너사와 SKT가 함께 제작한 것이다.



[그림 41] SKT지오비전이 활용한 데이터 종류

자료: SK텔레콤 지오비전 홈페이지
(http://www.geovision.co.kr/)

지오비전에서 사용하는 데이터는 다음과 같다.

- 유동인구(매월갱신) : 휴대전화 발신/수신 이력과 전화 가입자의 성별, 연령대
- 업종별 매출액(매월갱신): 카드 결재 데이터에 업종별 평균 현금결재비율을 산입하여 계산

<표 48> SKT지오비전 사용 데이터 리스트

대분류	중분류	조건	조건 상세	DB 구성 형태
인구/가구	인구 수	성/연령	성/연령 + 순위/비율	성/연령(5세단위) 별, 주거 인구수
			증감율	증감율
	인구총괄	평균연령	순위/비율	지수형태
	인구총괄	인구밀도		지수형태
	인구총괄	노령화지수		지수형태
	인구총괄	노년부양비		금액
	인구총괄	유년부양비		금액
	인구총괄	총부양비		금액
	종교별 인구수	종교	종교 + 순위/비율	개신/천주/불/유/원불교 등
	성혼인상태별 인구수	혼인상태	혼인상태 + 순위/비율	기혼/미혼/이혼 등, 성별 구분
	가구	가구 수	순위/비율	총 가구수
			증감율	증감율
		가구원 수	순위/비율	평균 가구원 수
		가구소득	건물단위별 추정가구 소득	

대분류	중분류	조건	조건 상세	DB 구성 형태
OK캐시백	라이프스타일	비율	특화지역 여부 or 순위/비율	내근형(SKT)
				정착형(SKT)
				인근이동형(SKT)
				야간이동형(SKT)
				발바리형(SKT)
				교대근무형(SKT)
				규칙적소량이동(SKT)
				규칙적대량이동(SKT)
				불규칙이동형(SKT)
				은둔형(SKT)
				거주지내이동형(SKT)
				시내근교이동형(SKT)
				원거리이동형(SKT)
				브랜드 중시형 비율(OkCashBag)
				경제성 중시형 비율(OkCashBag)
				웰빙형 비율(OkCashBag)
				가족중심형 비율(OkCashBag)
				사교형 비율(OkCashBag)
				출퇴근형 비율(OkCashBag)
				생계형 비율(OkCashBag)
				여가/레저형 비율(OkCashBag)
				자가운전형 비율(OkCashBag)
				CYBER형 비율(OkCashBag)
				Out-Door 비율(OkCashBag)
				위험수용형 비율(OkCashBag)
				아파트/빌라 비율(OkCashBag)
				단독형 비율(OkCashBag)
				자영업종사형 비율(OkCashBag)
	업종별 소비지출액	인구 수	업종 + 성 + 요일 + 금액/비율	소비액 30/50/70/100/150/200/300/500 만원, 성/요일 별
		증감율	업종 + 성 + 요일 + 금액 + 증감율	증감율
	업소별 거래건수(객 수)	시간대	(전체 or 시간대) + 순위/비율	1시간 단위, 24 시간, OCB 가맹점 정보
		증감율	(전체 or 시간대) + 증감율	증감율
	업종별 구매자	성/연령	성/연령 + 업종 + 요일 + 시간대 + 순위/비율	성/연령별 매출액
		증감율	성/연령 + 업종 + 요일 + 시간대 + 순위/비율	증감율
	구매자 유입 지역	행정구역	유입지 + 순위/비율	동일 행정동/타 행정동/타 구/타 시도
	재구매 주기	주기	주기 + 순위/비율	주/월/사분기/기타

대분류	중분류	조건	조건 상세	DB 구성 형태
주택	공동주택 호별 DB	평형	평형 + 순위/비율	점유면적 기준, 평형 타입 포함
		시세	시세 + 순위/비율	시세/전세 상/하한가
		증감율(시세)	증감율	증감율
	공동주택 분양 정보	평형	여부 or 입주예정일 + 세대수 + 금액	총세대수/입주예정일, 분양권/분양정보로 구분
		평형	평형 + 순위/비율	점유면적 기준, 평형 타입 포함
	오피스텔 DB	시세	시세 + 순위/비율	시세/전세 상/하한가
		증감율(시세)	증감율	증감율
	단독 등 기타 주택 DB	주거 여부	밀집지역 여부	주거여부 구분
	주택총괄DB	주택수	순위/비율	총주택수
		증감율	증감율	증감율
	건축년도 별주택수	건축년도	년도 + 순위/비율	10년 단위/2000~2003/2004/2005이후
	교육정도 별인구수	교육수준	교육수준 + 순위/비율	초/중/고/대/대학원
	난방시설 별가구수	난방시설	난방시설 + 순위/비율	중앙/지역/도시가스/기름 등
	세대구성 별가구수	세대구성	세대구성 + 순위/비율	1/2/3/4세대가구/1인가구
	점유형태 별가구수	점유형태	점유형태 + 순위/비율	자가/전세/월세 등
	방거실식당 수별가구수	방 수		1~5개 이상
		거실 수	수 + 순위/비율	없음~2개이상
		식당 수		없음~2개이상
	평수별 주택수	연건평	평형 + 순위/비율	7평미만~70평이상, 10평 단위
	주택유형 별주택수	주택유형	유형 + 순위/비율	단독/다세대/아파트/연립 등

대분류	중분류	조건	조건 상세	DB 구성 형태
주간상주인구	주간상주 인구수	성/연령	성/연령 + 순위/비율	연령/성별 구분
		증감율	증감율	증감율
	라이프스타일	유형	특화지역 여부 or 순위/비율	SKT(구분기준추가가능/검토필요)
	사업체 리스트	업종	업종 + 매출액 + 순위/비율	설립년도/업종, 매출8억원 이상 기업
	사업체수 총괄DB	업종	업종 + 순위/비율	정부 분류 기준
	창업년도 별사업체 수	창업년도	창업년도 + 순위/비율	10년 단위, 2000년 이후 년도 별, ~2010년 까지
	업종별 종사자수	업종	업종 + 종사자수 + 순위/비율	대분류업종
		증감율	증감율	증감율
유동인구	유동인구수	성/연령/시간	성/연령 + 요일 + 시간대 + 순위/비율	성/연령 별, 요일/시간대 별
		증감율	증감율	증감율
	라이프스타일	유형	특화지역 여부 or 순위/비율	SKT(구분기준추가가능/검토필요)
	유입지	행정구역	유입지 + 순위/비율	동일 행정동/타 행정동/타 구/타 시도
	거주지/직장	행정동	거주지/직장 + 순위/비율	유동인구의 거주지/직장 Top5 지역
주요시설	버스정류장	노선	노선 수 + 순위/비율	노선 수
	지하철 역	시간/월	승하차인원수 + 순위/비율	노선 수, 월/시간대 별 승하차인원 수
	고객 유입 효과가 있는 시설	종류	주요시설 유형 + 여부	병원, 극장, 대형쇼핑시설, 체육시설 등
상권	전국 1,000대 상권	지역	여부 or 상권명	상권 영역, 상권명(현장조사결과)
	전국 4,700대 메인 스트리트	주업종	여부 or 주업종 + 여부	메인스트리트 영역, 주업종
	상권 내 시세	유형	회전율 + 권리금 + 임대료 + 순위/비율	입지/면적/주고객/회전율/보증금/임대료/권리금 상/하한
		증감율		증감율
매출	업종별 매출	업종	업종 + 요일 + 시간대 + 증감율	주중/주말, 시간대 별 매출액
		증감율		증감율

대분류	중분류	조건	조건 상세	DB 구성 형태
부동산	점포 임대 정보	유형	여부 or 세대수 + 입주예정일 + 금액	상가정보/층/사진/관리비/임대료/권리금/현업종/추천업종, 3,000건
		유형		상가정보/층/관리비/임대료/권리금/현업종/추천업종, 전체(R114)
	점포 매매/분양 정보	유형		상가정보/층/관리비/임대료/권리금/현업종/추천업종, 전체(R114)
	공동주택 분양 정보	평형		총세대수/입주예정일, 분양권/분양정보로 구분
	택지개발지구	영역	여부	택지지구 영역
	상가분양정보	유형	여부 or 점포수 + 입주예정일 + 금액	입주일/분양일/점포수/층면적(지하 포함)
	상가 정보	유형	여부 or 점포수 + 층면적 + 분양가 + 순위/비율	서비스시작일/점포수/층면적/평당 분양가 상/하한
	건물정보	시세		전국 건물명, 위치, 건물상태, 용도, 층수, 구조, 공시지가
	아파트실거래가격	시세		전국 APT 실거래 정보 총 500만건 위치, 아파트명, 면적, 거래가(단지기준)
	공동주택 호별 DB	평형	평형 + 순위/비율	점유면적 기준, 평형 타입 포함
		시세	시세 + 순위/비율	시세/전세 상/하한가
		증감율(시세)	증감율	증감율
	공동주택 분양 정보	평형	여부 or 입주예정일 + 세대수 + 금액	총세대수/입주예정일, 분양권/분양정보로 구분
		평형	평형 + 순위/비율	점유면적 기준, 평형 타입 포함
	오피스텔 DB	시세	시세 + 순위/비율	시세/전세 상/하한가
		증감율(시세)	증감율	증감율

다. 활용 사례

SKT지오비전에서 제공하는 서비스는 상권분석 서비스로서 SKT지오비전 자체 데이터를 이용하여 컨설팅 서비스를 제공하며, 기업의 내부 정보를 자체개발 툴로써 분석할 수 있는 서비스도 제공 한다

<표 49> SKT지오비전 서비스 리스트

서비스	개요
상권분석서비스	각 지역별 유동인구, 업종별 매출액, 부동산 정보를 종합하여 신규 창업 입지에 도움이 되는 정보 제고
데이터 분석 솔루션 제공 (X-ray Map, Buisness GIS, Alteryx)	기업대상으로 데이터 분석 솔루션을 제공함. 지도, GIS, 분석 툴 등으로 세분하여 서비스 함.

1) 상권분석서비스

상권분석 서비스는 개인 또는 프랜차이즈 본부를 대상으로 제공되는 컨설팅 서비스이다. 상권분석에 사용되는 데이터는 이하와 같다.

<표 50> SKT지오비전 상권분석 서비스에 사용되는 데이터 리스트

데이터 제공사	데이터	데이터 특징	업데이트 주기
SK 텔레콤	유동인구	기지국 통화량 통계분석(시간, 성별, 연령)	월별
	주간상주인구	건물데이터와 통화량 통계 분석	1년
	주거인구	통계청 센서스와 행정안전부 주민등록 통계 활용	1년
SK 플래닛	지도	지도 및 시설정보(POI)	3개월
현대카드	업종 매출	업종별 카드 가맹점 매출 통계	월별
부동산 114	부동산 시세/매물	부동산 시세 및 매물 정보(아파트, 상가)	월별
	개발 예정 정보	개발예정 구역 및 정보	수시
한국창업전략연구소	주석	리포트 항목에 대한 전문 컨설턴트 주석	수시
	전문가 칼럼	창업컨설턴트의 상권정보 및 업종정보	월별
기업정보 제공	외감 기업	지역별 외부감사법인 정보	월별
통계청	인구센서스	인구통계 및 산업세세분류별 매출 규모	5년
행정안전부	주민등록인구 통계	주기인구 통계 활용	1년
도시지하철공사	지하철 노선수	전국 지하철 노선 정보	수시
	지하철 승하차 인원수	지하철역 승하차 인원수	수시

이용자는 특정 지역을 선택하고, 창업할 업종을 고른 뒤 리포트를 받아 이용하는데 리포트에는 지역의 유동인구, 해당업종의 평균 매출, 주변의 인구유입 시설 개수 등을 정리하여 보여준다.

상권분석서비스는 이하의 세 가지 방식으로 제공된다.

1. 일반상권분석보고서(무료): 지역별 상권에 대한 기본 정보 제시

2. 심층상권분석보고서(유료, 15만): 해당 업종의 시간대별 매출 추이, 성별/연령대별 유동인구 변화 제시
3. 상권비교분석보고서(무료): 여러 지역에 걸쳐 업종의 입지를 비교, 일반상권분석보고서와 흡사한 형태

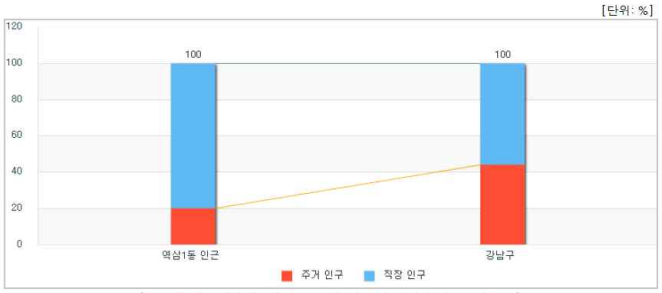
[선택상권 위치 정보]

매출정보	생활/잡화 07월 평균 추정 매출	
	1,707만원	
업종정보	편의점 매장 수	
	12개	
인구정보	상주인구 수	1일 평균 유동인구 수
	61,527명	44,513명
지역정보	인구 유입 시설	
	34개	

[선택상권 요약 정보]

[그림 42] 선택 지역의 상권 정보(일반상권분석보고서)

자료: SK텔레콤 지오비전 상권분석 서비스 홈페이지
(http://bizanalysis.geovision.co.kr:8080/main.do)

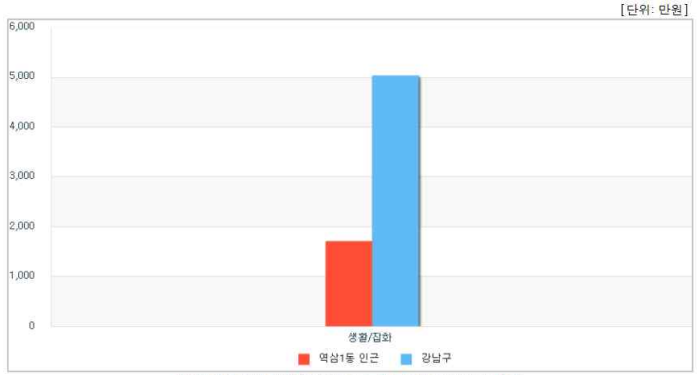


[역삼1동 인근 상권 상주인구 구성비율(인구수) 및 선택구(강남구) 비교]

구분		역삼1동 인근		강남구	
		구성비율	인구수	구성비율	인구수
상주인구	주거 인구	20%	12,170명	44%	487,680명
	직장 인구	80%	49,358명	56%	621,337명
합계		100%	61,527명	100%	1,109,017명

[그림 43] 선택지역의 인구 데이터(일반상권분석보고서)

자료: SK텔레콤 지오비전 상권분석 서비스 홈페이지
(http://bizanalysis.geovision.co.kr:8080/main.do)

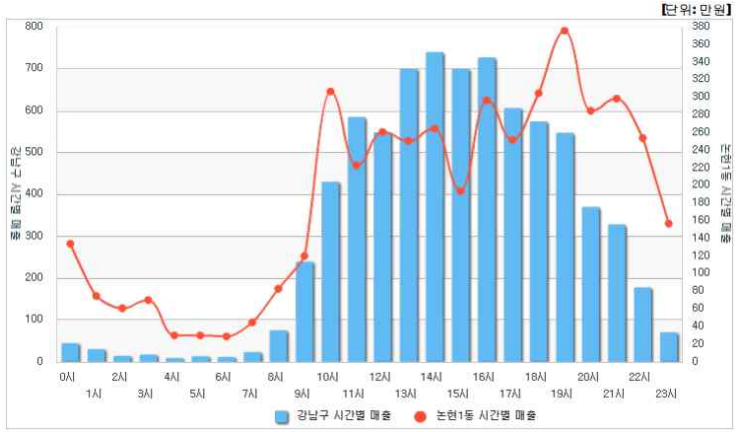


[선택 업종군(생활/잡화) 07월 추정 매출 선택구(강남구) 비교]

선택 상권 내 생활/잡화 추정 매출	07월 평균 추정매출	선택구(강남구) 대비
	1,707만원	-3,321만원

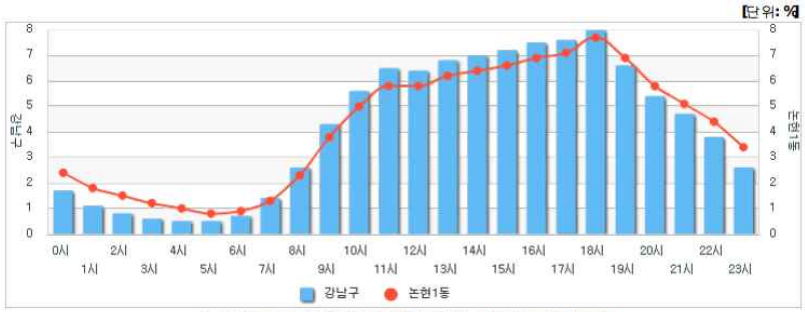
[그림 44] 선택업종의 월 추정 매출액 규모(일반상권분석보고서)

자료: SK텔레콤 지오비전 상권분석 서비스 홈페이지
(http://bizanalysis.geovision.co.kr:8080/main.do)



[그림 45] 선택업종의 시간대별 추정 매출액(심층상권분석보고서)

자료: SK텔레콤 지오비전 상권분석 서비스 홈페이지
(http://bizanalysis.geovision.co.kr:8080/main.do)



[그림 46] 시간대별 유동인구 분포(심층상권분석보고서)

자료: SK텔레콤 지오비전 상권분석 서비스 홈페이지
(http://bizanalysis.geovision.co.kr:8080/main.do)

[그림 47] 상권분석 결과 예시

자료: SK텔레콤 지오비전 상권분석 서비스 홈페이지
(http://bizanalysis.geovision.co.kr:8080/main.do)

2) 데이터 분석 솔루션

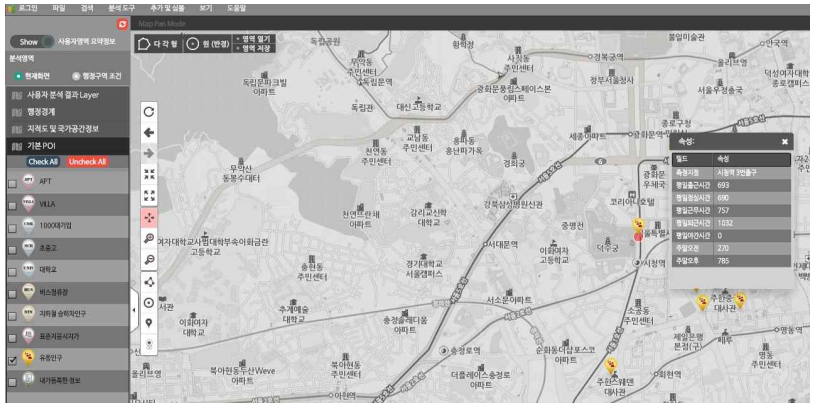
데이터분석 솔루션 제공은 크게 세 가지 서비스로 나뉘어서 제공한다. 세 서비스 모두 사용하는 데이터베이스는 상권분석서비스와 같으나 상권분석서비스의 기능이 경직된 반면에 세 서비스는 이용자가 커스터마이징을 하여 사용한다는 특징이 있다.

<표 51> SKT지오비전 데이터 분석 솔루션 목록

서비스	내용
X-ray Map	지도 기반의 빅데이터 분석 툴이며, SKT지오비전의 내부 데이터도 사용가능함
Buisness GIS	지역별 상권분석을 개별 업종으로 특화하여 지도로 표현
Alteryx	빅데이터 분석 툴

X-ray Map

X-ray Map 서비스는 지도상에 기업이 보유하고 있는 데이터를 업로드하여 분석할 수 있는 툴이다.

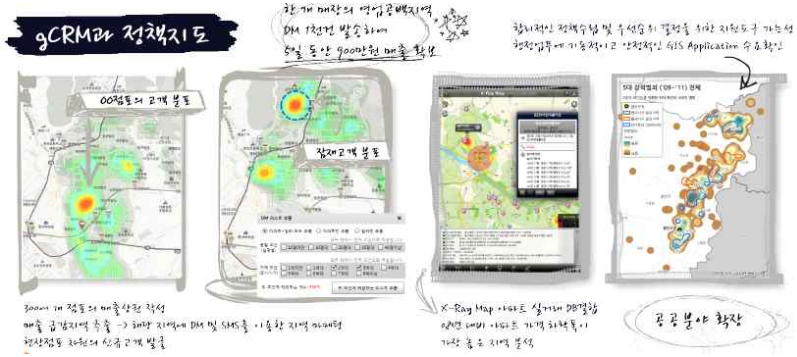


[그림 48] X-ray Map 사용화면

자료: SK텔레콤 지오비전 X-ray Map 서비스 홈페이지
(http://www.biz-gis.com/XRayMap/)

특정 지역을 설정하여 유동인구, 거주인구 및 지역의 직장인 인구, 월평균소득,

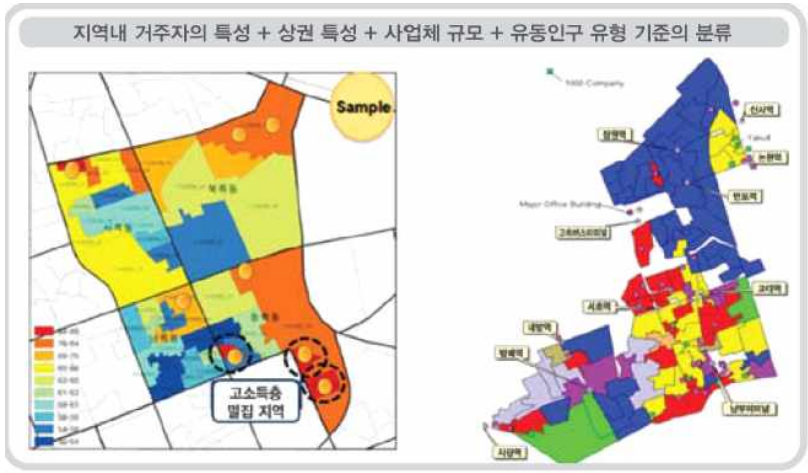
지하철 승하차인구 등을 조회할 수 있으며, 내부정보를 업로드하여 분석할 경우, SKT지오비전 데이터상에 제공된 잠재고객분포와 현재 고객분포를 비교하여 취약지점을 파악할 수 있다.



[그림 49] X-ray Map 사용 예
자료: SK텔레콤 지오비전 X-ray Map 서비스 소개자료
(http://www.geovision.co.kr/html/solution01.html)

Buisness GIS

Buisness GIS는 상권분석에 사용되는 데이터를 바탕으로 개별 고객사에 특화된 분석을 제공하는 서비스로서 도로 형태, 건물위치, 유동인구, 인구의 이동경로를 분석하여 소비자가 매장에 도달하는 범위, 출발지 등을 표현하며, 상기 분석을 이용자의 내부 데이터와 매칭하여 소비자의 실제 분포와 방문객의 실제 분포를 파악하고, 매장의 입지 선정에 참고한다.



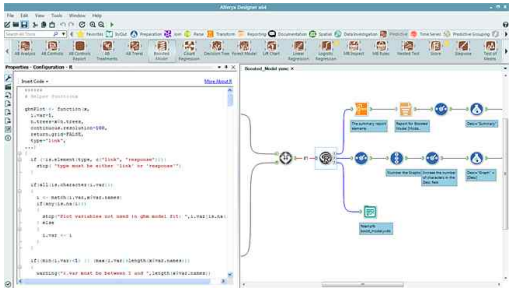
[그림 50] Buisness GIS 사용 예
자료: SK텔레콤 지오비전 Buisness GIS 서비스 소개자료
(http://www.geovision.co.kr/html/solution02.html)

상기 그림에서는 거주자의 소득, 그리고 각 지역별 유동인구의 이용 교통수단을 파악하여 지도상에 표현하였다.

Alteryx

Alteryx는 R기반의 분석 툴로, 미국 캘리포니아 소재의 Alteryx Inc에서 개발/유통하는 프로그램으로서 주요 분석기법의 코딩이 완료된 상태로 제공되기 때문에 데이터 매칭만 하면 바로 분석을 할 수 있다.

분석은 GIS파일에 연동하여 출력할 수 있도록 구성되어 있으며, 부산시서비스인구통계 작성에도 사용되었다.



[그림 51] Ateryx 사용 화면

자료: SK텔레콤 지오비전 Ateryx 서비스 소개자료
(http://www.geovision.co.kr/html/solution03.html)

4. 신규과제 제안

제1절 모바일 폰 데이터 랩

가. 배경 및 필요성

본 연구에서는 빅데이터를 활용한 공공통계의 생산을 위해서 필요한 통계청의 신규 과제로서 모바일 폰 데이터 랩(Mobile Phone Data Lab) 운영을 제안하고자 한다. 모바일 폰 데이터 랩 운영을 통계청이 수행해야 할 필요성으로는 다음과 같은 근거들을 고려할 수 있다.

첫째, 모바일 폰 데이터가 정책 수립의 자료로서 활용 가치가 크다는 점이다. 앞서 현황에서 살펴 본 바와 같이 모바일 폰 데이터는 현재 공공통계나 정책 수립에 가장 활발하게 사용되고 있는데, 상권분석, 공공서비스의 공간 배치, 관광 통계 작성에 활용되고 있다. 국토연구원의 연구진들은 수도권 모바일 폰 데이터를 활용함으로써 국토분야에서는 도시공간구조 파악 및 사회조사 비용 절감, 공간위계별 동적 행정수요 파악, 개인 활동을 고려한 정책수립의 신개념 정책 자료 활용 가능성을 제시하였고, 교통분야에서는 활동인구의 실시간 이동패턴을 이용한 교통계획 활용, 교통량 보완자료 활용, 기종점 통행량 구축의 신뢰도 제고 가능성을 제시하였으며(김종학 외, 2015), 재난대응 분야에서는 미시에서 거시까지 다양한 공간위계의 재난대응에 필요하고, 특히 거주인구로 파악하기 어려운 주간시간 대 도심, 관광지, 대형 상가 등의 재난 발생 시 요구조자

정보 제공의 기초 자료로 활용될 수 있음을 제시하였다(김종학 외, 2016).

둘째, 모바일 폰 데이터가 다른 빅데이터 활용에 기반이 되는 인프라의 성격을 갖는다는 점이다. 지금까지 국내외 사례들을 종합해 볼 때 모바일 폰 데이터를 이용해서 생산할 수 있는 공공통계는 주간인구(daytime population statistics), 이동통계(mobility statistics), 관광통계(tourist statistics) 등이다. 이러한 사실은 모바일 폰 데이터를 이용한 통계 생산이 특정 통계에만 한정되지 않는다는 점을 보여준다. 모바일 폰 데이터를 활용하여 통계를 생산하기 위해서는 어려운 행정적, 기술적, 통계방법론적 쟁점들을 해결해야 하는데 이와 같이 어려운 과정들은 여러 통계 생산의 기본 핵심 과정을 이룬다. 이러한 점에서 모바일 폰 데이터를 이용한 통계 생산은 각 통계별로 별개로 작업할 내용이 아니고, 공통으로 추진하는 것이 더 효율적이라는 점은 EU에서 수행한 파일롯 프로젝트의 제안 사항이다(Positium. 2014).

이는 모바일 폰 데이터가 다른 어떤 형태의 빅데이터에 비해서 보편성을 갖기 때문이다. 예컨대 소셜미디어가 해당 서비스 이용자에 한정되고, 교통이나 보건 빅데이터가 해당 교통수단 이용이나 해당 질병에 한정되는 것과 달리 모바일 폰 데이터는 모바일 폰 소유의 보편성과 휴대성에 의해서 다양한 형태의 인구와 공간 관련 통계 작성과 관련성을 가진다. 모바일 폰 데이터는 이와 같은 보편성 때문에 여러 빅데이터 중 하나가 아니라 인구와 공간 통계의 가장 기본적인 통계 인프라로 다루어져야 한다. 따라서 모바일 폰 데이터가 가지는 인프라적인 성격이 관련 자료 구입에 필요한 예산 배정에도 반영되어야 한다.

셋째, 모바일 폰 데이터를 이용한 통계의 품질관리가 시급하다는 점이다. 그동안 진행되어 온 모바일 폰 데이터의 활용 과정을 보면 통계 품질의 측면은 통신사 자료의 공급이나 혹은 통계 분석 업체에 크게 의존해 온 것으로 보인다. 일부 통계 생산을 정교화시킨 사례가 있으나 모바일 폰 데이터 이용에 따르는 다양한 쟁점들을 투명하게 해결한 것으로 평가하기 어렵다. 이와 같이 모바일 폰 데이터를 이용한 통계생산이 품질관리 측면에서 투명성을 갖추지 못했음에도 불구하고 별로 문제가 부각되지 못한 것은 이러한 작업이 전체적으로 새로운 파일롯 프로젝트로서 환영받는 분위기가 있었기 때문이다. 특히 모바일 폰 데이터를 이용한 통계가 직접 공표되기 보다는 상권분석이나 버스 노선 조정과 같은 목적 수립을 위한 도구적, 보조적 성격을 갖고 있었기 때문에 큰 문제가 되지 않았다. 하지만 이와 같은 모바일 폰 데이터의 사용이 국토계획이나 국민안전계획 수립에 중요한 자료로 사용되려면 품질관리 필요성이 강조될 수밖에 없다.

반면 모바일 폰 데이터를 이용한 통계의 품질관리는 기존의 승인통계와 같은 품질관리 방식으로는 달성하기 어렵다. 모바일 폰 데이터를 이용한 통계 생산에 필요한 기술적, 통계방법론적 역량을 민간 통신업자나 다른 정부기관에 기대하기 어렵기 때문이다. 따라서 이에 대해서는 통계청이 직접 모바일 폰 데이터의 기술적, 통계방법론적 쟁점을 해결하여 기초 통계를 공급하는 것이 실현가능한 방안이다.

넷째, 빅데이터 시대에 필요한 공공통계 생산의 거버넌스를 가시화하는 것이 필요하다. 빅데이터 시대에는 Axciom과 같은 데이터 중개자들이 나타나서 통계청의 역할을 대체해나갈 가능성이 커진다(Kitchin, 2015). 소비자 정보를 가지고 있는 기관이 독자적인 통계를 작성하여 발표하는 것이 하나의 사례이다. 국내에서도 통신사 제공 자료를 공공기관이 그대로 활용하거나 데이터 분석 업체가 통신사로부터 모바일 폰 데이터를 받아서 공공기관에 통계를 공급해 주는 방식들이 보편화되고 있다. 이 과정에서 주간인구의 경우 통계청 통계를 대체할 만한 통계를 생산함으로써 국가 통계를 대체하는 방향으로 발전이 진행되고 있다. 문제는 품질검증이 되지 않은 통계들이 미디어에 그대로 보도됨으로써 공공통계를 대체하는 현상이 발생하고 있다는 점이다. 현재는 주간인구에 초점이 맞추어져 있으나 장기적으로는 상주인구를 포함한 인구통계 전체를 실질적으로 대체될 가능성을 배제하기 어렵다. 공공통계를 생산하는 공공기관들이 통계청과 협력적으로 공공통계를 생산하고 그 틀을 바탕으로 민간 업체들이 적절한 역할을 찾을 수 있도록 거버넌스를 조성하는 것이 필요하다.

다섯째, 빅데이터를 이용한 통계 생산의 과정에서 기존 공공 통계의 다양한 활용이 불가피하다. 하지만 프라이버시 보호의 문제로 공공통계의 개방이 전면적으로 이루어질 수 없는 상황에서 현실적으로 데이터 연계의 과정을 통계청이 제한된 공간에서 진행하는 것이 바람직하다.

나. 과제 내용

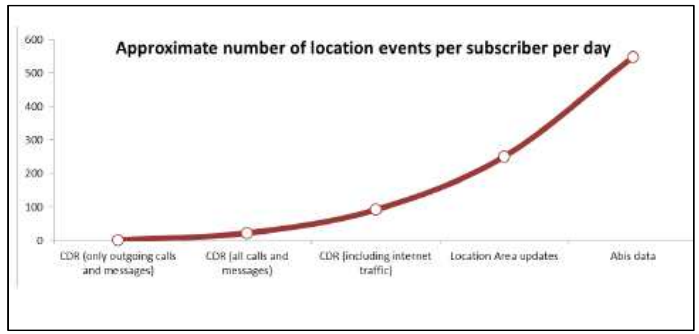
1) 모바일 폰 데이터를 이용한 통계 생산의 연구개발

빅데이터를 통계 목적으로 사용하기 위해서 해결해야 하는 과제들은 다음과 같다(Tiru, 2014). 자료 처리를 위한 기술 과제로서 자료 원천별 특성 이해와 통계별 파악, 과정에 따른 기술적 문제 해결하고, 스케일 선택: 어떠한

지역단위로 과정을 작업할지에 대한 선택이 필요하다. 끝으로 자료 처리를 위한 통계 방법론의 과제로서는 익명화 작업을 통한 프라이버시 보호, 데이터의 부정확한 값을 제거하고, 데이터의 누락 기간 확인, 노이즈를 제거하는 필터링 작업, 통계 작성 분야별 환경에 대한 정의를 내리는 분야별 특화 방법, 레퍼런스 데이터와 추정 등이 수행되어야 한다. 현재 특히 중요하게 부각되는 방법론적 이슈들은 다음과 같다.

첫째, 모바일 폰 데이터의 종류에 따른 특성을 파악하고 개별 데이터 종류가 가지는 한계를 극복하는 것이 필요하다. 모바일 폰 데이터의 가장 중요한 부분은 모바일 기기의 위치를 확인하는 시간과 공간의 자료인 모바일 위치 데이터(Mobile Positioning Data)이다. 모바일 위치 데이터가 수집되는 주체는 통신사인 모바일 네트워크 오퍼레이터(MNO: Mobile Network Operators)와 모바일 어플리케이션 개발자로 나누어진다. 위치 데이터를 수집하는 방법은 능동적 포지셔닝(Active Positioning)과 수동적 포지셔닝(Passive Positioning)으로 구분된다. 수동적 포지셔닝 데이터의 유형은 CDR(Call Detail Records), 위치 정보 업데이트, 다른 네트워크 데이터로 구분된다. 추가적인 데이터로는 지리정보, CRM 정보(인구학적 정보, 폰 사용, 평균 통신 사용 요금 등), 모바일 뱅킹 자료 등이 가능하다.

모바일 폰 데이터의 양은 CDR 데이터와 위치확인 데이터 등 종류에 따라 차이가 크다.



[그림 52] 모바일 폰의 일일 위치 확인 양 (자료: Tiru, 2014)

현재 가장 많이 활용되는 CDR 방식의 경우 통화를 하거나 문자를 사용하는 인구만이 고려된다. 또한 이동인구의 중복 집계 문제: 셀 단위를 이동하는 인구를 어떻게 파악할 것인지의 문제가 발생한다. 모바일 폰 데이터를 이용한

통계 생산의 방법론에 대한 국내의 연구는 이와 같은 CRD 방식의 한계를 극복하는 것에 초점이 맞추어져 있다. 김감영 외(2015)는 VLR(Visitor Location Register), RT(Real Time), CDR(Call Detail Record) 데이터의 특징을 비교하여 개선방안을 제시하였다. 김경태 외(2016)은 통신확률의 개념을 제시하고 통신인구를 바탕으로 전수화 방법을 제시하였다. 이러한 연구들은 현재의 문제점을 제시하고 해결 가능성을 제시한 수준이었다.

둘째, 빅데이터를 활용한 통계 생산에는 모델 기반 추정의 필요성이 나타난다. 빅데이터의 경우 표본추출을 통해 확보한 자료가 아니고 알려진 모수를 바탕으로 생성한 자료가 아니기 때문에 디자인 기반 추정(design based estimation)이 적합한 접근 방법이 아니며, 모델 기반 추정(model based estimation)이 적합한 대안으로 사용하는 것이 필요하다(Buelens et al., 2016). 통계청이 그동안 주로 디자인 기반 추정에 의존해 왔으며 모델 기반 추정을 기피하거나 명시적으로 드러내기를 꺼려했지만 이미 소지역 추정, 월별 실업율(시계열 모델링), 무응답 교정(가중치, 캘리브레이션), 혼합조사방법 조사에서 조사방법 효과의 교정, 시즌 효과 보정 등과 같은 분야에서 모델 기반 추정을 해왔다는 점에서 볼 때, 비표본자료 혹은 비확률표본자료의 사용이 중요한 이슈로 부각되고 있으며 그 중요한 사례가 빅데이터인 상황에서 이를 명시적으로 더욱 적극적으로 사용하기를 권장되고 있다.

이와 같이 표집 없는 추정(inference without sampling)에 대한 논의를 보면, 표집 없는 추정의 대상이 되는 자료들의 생산이 최근 급증하였다. 대량의 자료를 수집함으로써 데이터가 가질 수 있는 편향이 감소될 수 있다는 암묵적인 전제하에 큰 규모의 자료가 수집된다. 또한 이 과정에서 서베이나 직접 질문이 생략되어 자료 수집 과정에서 발생하는 비용을 절감하고자 한다. 표집 없는 추정이 적용되는 자료들은 소셜미디어 연구, 집단 지혜(wisdom of crowds) 연구, 빅데이터 연구 등이다(Baker et al., 2013).

빅데이터가 확률 표본의 대안으로 사용될 수 있다는 전제에서 추진되는 사업이 Longitudinal Employer Household Dynamics (LEHD 프로그램이다. 이 프로그램은 주정부의 노동시장 정보기구와 통계청에 의해서 추진되는데 실업보험 가입자의 자료를 사용한다. 전체 노동자의 정보를 사용하지 못하지만 다른 행정자료의 활용을 바탕으로 노동시장 전체를 추정하는 정보를 생산한다(Baker et al., 2013).

셋째, 모바일 폰 데이터와 기존 공공통계(센서스, 행정자료 등)를 연계시켜

모바일 폰 데이터를 이용한 통계 생산의 정확성을 제고하고 모바일 폰 데이터 통계의 활용가치를 높인다. 빅데이터와 기존 공공통계의 관계, 빅데이터를 공공통계에 통합시키는 방안이 마련되어야 한다(Florescu et al., 2014). 빅데이터와 기존 공공통계는 전면 대체, 부분 대체, 보조적 통계 정보 제공, 추정치 개선, 새로운 통계 정보 제공 등으로 구분할 수 있다. 모바일 폰 데이터의 활용에서도 다른 기존 공공통계와의 관계성들을 명확하게 정리하여 효과를 극대화하는 작업이 필요하다.

2) 통계 생산

모바일 폰 데이터를 이용한 통계는 직접적으로는 인구통계, 이동통계, 관광통계로 구분할 수 있다.

먼저, 인구통계의 경우 주간인구통계가 가장 직접적으로 관련된다. 통계청의 인구, 통근, 통학 통계를 발표하면서 함께 주간인구가 산출되어 발표 된다. 이 통계의 가장 중요한 결과는 수도권과 비수도권으로 나누어 유입과 유출의 인구를 보여준다. 통계청 DB에서 주간인구는 한국도시통계의 영역에서 다루어진다. 한국도시통계는 현재 행정자치부의 지역경제과에서 생산하고 있다. 주간인구지수도 생산하고 있다. 이와 같이 통계청에서는 센서스 발표 시점 이외 그다지 중요하게 취급받지 못하고 있으나 지자체에서는 매우 중요한 통계로 다루고 있다.

현재와 같이 주간인구통계가 가지는 한계는 자명하다. 주간인구는 상주인구에 통근 통학 인구만을 고려하여 작성하기 때문에 쇼핑이나 관광, 혹은 다른 일에 따른 이동을 포착하지 못한다는 점에서 약점을 가진다. 맨해튼의 주간인구에 대한 루빈연구소의 연구 결과를 보면 센서스 기반 주간활동인구는 실제 주간활동인구 보다 25%나 적은 것으로 추정되었다Moss, & Qing, 2012). 또한 현재는 주간인구가 통계청 DB에는 5년 동안 같은 값이 제시된다. 이는 상주인구와는 다른 차원에서 오해의 소지를 낳는다는 점에서 문제를 가진다. 이와 같은 약점에도 불구하고 기존의 주간인구는 대부분의 통계청에서 사용하는 방식으로 생산되어 왔다. 하지만 빅데이터의 활용에 따라 기존에 만들었던 통계의 한계가 더 명확히 부각되고 있다. 모바일 폰 데이터를 활용한 주간인구 통계의 보완은 이러한 문제를 상당히 해결해 줄 것으로 보인다.

통계의 생산은 이론적으로 세분화된 시간단위에서도 가능하지만, 공식통계로는 일간 첨두시간 (예: 오전 9시, 오후 12시 30분, 오후 7시 등)별로 발표할 수 있다.

모바일 폰 데이터를 활용한 주간인구 통계 생산에서 다른 준거 통계 자료를 활용하는 것이 바람직하다. 대중교통의 스마트 카드와 같이 기존의 통계를

활용하는 방안도 있으나, 새롭게 CCTV를 활용한 영상분석, 센서 자료를 통한 집계, 드론이나 인공위성을 이용한 원격 탐사(remote sensing), 패널을 활용한 개별 추적 조사 등을 도입할 수 있다. 또한 현재 시행중인 유동인구 조사 통계와의 비교 분석을 모델 추정의 정확성을 제고하는데 활용할 수 있다. 현재 진행 중인 유동인구조사로서 가장 큰 규모의 조사는 서울시에 의해서 수행되고 있다. 서울 유동인구조사는 2009년에 첫 실시되었고, 이후 2012년부터 재개되어 2015년까지 조사가 진행되었다. 조사 규모는 2009년에 서울시내 1만여 지점을 대상으로 시작하였으나 점차 축소되어 1천여 지점 수준으로 줄어들었다. 조사는 단순 계수로 이뤄지는 유동인구 조사(5분 조사, 10분 검수)와 속성조사(유동인구 중 일부는 표집하여 실시하는 개별면접조사)로 구성되어 있다. 면접조사에서는 보행자의 인적사항(성별, 연령, 거주지 등)과 통행목적, 조사지역의 평균 통행횟수와 직전에 이용한 교통수단 등에 대해 조사하였다. 년도별 조사규모 및 시점은 이하 표와 같다.

<표 52> 서울유동인구조사 년도별 개요

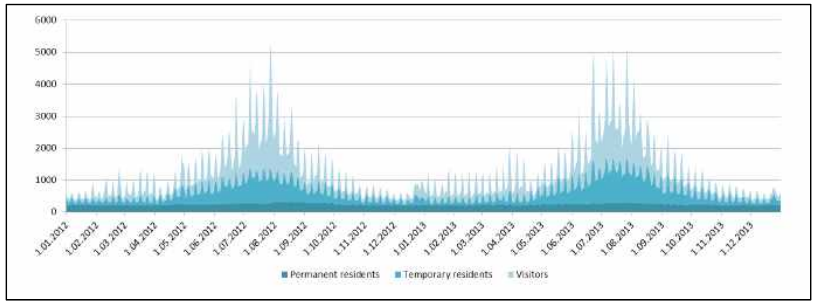
	조사지점	조사시점	조사시각
2009	1만여 지점	2009년 8월~10월	07:30~20:30
2012	2천 지점	2012년 10월~11월	07:00~21:00
2013	1천 지점	2013년 9월~10월	07:30~19:30
2014	1천 지점	2014년 10월~11월	07:30~19:30
2015	1천 지점	2015년 10월	07:30~19:30

국토연구원은 모바일 빅데이터와 통계청의 서울시 주간활동인구의 상관계수가 0.94로 높게 나왔다고 밝혔다. 하지만 상권 분석에서 MP데이터를 실제 사용하는 경우에서는 현재 통신사에서 제공되는 MP데이터를 기반으로 모수 추정을 할 경우 전체 인구보다 훨씬 많은 인구수가 추정되는 문제점이 나타난다는 점이 지적되었다.

모바일 폰 데이터를 이용한 인구통계는 주간인구 통계에만 한정되지 않는다. 위치 자료를 이용하여 거주지, 직장, 학교, 제2 거주지, 국가 내에서 내부 이동, 직장 이동, 그리드형 인구통계(1 km^2), 일시인구(temporary population) 통계, 특정 이벤트 관련 일시인구통계 등과 같은 인구통계가 생산 가능하다. 다음의 그림을 보면 모바일 폰 데이터를 활용하여 거주 인구, 일시 인구, 방문자 통계가 구분되어 생산된 것을 볼 수 있다.

따라서 주간인구에 머물지 않고 상주인구의 측정에도 모바일 폰 데이터가 활용될 수 있다. 모바일 폰을 통해서 오전 2시부터 오전 6시까지 특정 지역에

머무는 사람의 숫자를 파악한다면 이 정보는 상주인구를 파악하는 중요한 원천이 될 것이다. 모바일 폰 데이터가 등록인구와 실 거주인구 사이의 괴리를 추정하고 보정하는 것이 중요한 과제인 상황에서 모바일 폰 데이터가 상주인구 추정에 중요한 자료가 될 것으로 기대된다.



[그림 53] 모바일 위치 자료를 이용한 인구통계의 사례: 에스토니아 농촌 지역의 일일 인구통계 (출처: Tiru, 2014)

다음으로 모바일 폰 데이터를 이용하여 교통계획, ;국토관리, 도시계획에 활용될 수 있는 이동통계의 항목들은 다음과 같다(Tiru, 2014).

- ㉮ 출발지-목적지 매트릭스 (특정일 혹은 평균 산정)
- ㉯ 매일 통근 유형
- ㉰ 세분화된 공간 단위
- ㉱ 인구학적 속성에 따른 통계
- ㉲ 개인당, 일당 평균 이동 횟수
- ㉳ 평균 이동 거리
- ㉴ 평균 이동 횟수

다음으로 모바일 폰 데이터를 이용하여 관광통계에 활용될 수 있는 이동통계의 항목들은 다음과 같다(Tiru, 2014).

- ㉮ 여행 횟수, 여행객 수
- ㉯ 특정 국가 혹은 특정 지역의 방문 기간
- ㉰ 외국관광객의 출신국별 통계
- ㉱ 국가 내 관광객의 거주지역별 비교

㉠ 시기별 비교(일별/주간별/월별)
㉡ 관광의 전체 지속기간
㉢ 관광 활동의 공간범위(국가, 하부 지역단위)
㉣ 관광활동 경로
㉤ 재방문
㉥ 관광목적지, 이차 목적지, 경유지

다. 제도적 운영 방안

모바일 폰 데이터 랩이 효과적으로 운영되기 위해서는 처음부터 생산 통계의 관련 부처들과 협력적으로 과제를 수행하는 것이 필요하다. 인구통계는 행정자치부, 국가안전처, 중소기업청 등과 협력하고, 이동통계는 국토교통부, 관광통계는 문화체육부와 협력한다. 이와 같은 협력과정에서 모바일 폰 데이터와 다양한 조사 자료, 행정 자료가 연계되는 것이 불가피한 상황에서 통계 작성에 프라이버시 문제에 대한 우려가 부각될 가능성이 커진다.

이와 같은 우려를 불식시키기 위해서 통계청의 통계 생산 기구로서의 전문성과 신뢰성 그리고 객관성이 강화되어야 한다. 통계청이 가지는 신뢰가 자산으로서 가지는 중요성이 더 커지는 상황에서 이를 강화하기 위한 제도 정비가 필요하다.

먼저, 통계청이 빅데이터 관리에 필요한 지침을 청장의 훈령으로 제정한다. 빅데이터를 이용하여 공공 통계를 생산하는 모든 사람이 지켜야할 윤리 강령을 제정한다. 윤리 강령의 내용에는 사생활 침해 가능성 관리, 목적 필요성 부합 여부, 메타데이터 생산 및 공표에 대한 지침 등을 포함하게 한다. 빅데이터를 이용한 공공통계 생산에 참여하는 자는 이 지침을 학습하고 이를 준수할 것을 서명하도록 한다.

다음으로 프라이버시 침해 위험을 관리하기 위해 통계청이 운영하는 랩의 필요성이 정당화되기 위해서는 랩의 운영을 중립적으로 감시할 수 있는 감독위원회의 설치가 필요하다. 감독위원회의 위원은 국회에서 추천한 인사들로 구성하며 감독위원회는 랩이 보고한 품질관리 및 윤리강령 준수 내용을 심의하여 공표한다. 이를 통해서 국민들이 가질 수 있는 의구심과 우려를 해소한다. 정보인권과 관련하여 언론이나 NGO가 제기하는 문제제기에 대응할 수 있도록 한다. 이와 같은 유사한 사례로는 영국의 통계감독기구(UK Statistics Authority)를 볼 수 있다.

라. 기대효과

첫째 비용 상의 이점이다. 동일한 데이터를 사용하여 하나 만의 통계를 생산하는 것보다 여러 통계를 생산함으로써 비용효율성을 제고한다. 각 부처마다 별개로 모바일 폰 데이터를 사용하는 것보다 한 데이터를 가지고 여러 부처에서 관계하는 통계를 작성함으로써 예산 사용의 효율화를 기할 뿐만 아니라 부처간 협력의 기반을 강화한다.

둘째, 통계청이 모바일 폰 데이터를 통한 여러 통계를 동시에 생산함으로써 실제로는 분산형 제도와 달리 협렵적인 집중형의 성격을 갖게 된다. 이 과정에서 각 부처들은 통계 생산과 품질 확보 및 승인통계 작성을 위한 부담을 덜게 된다. 통계청 입장에서는 협력적 집중형 통계생산의 구심점으로서 새로운 성격에 부합하는 역량을 키우고 검증받는 계기를 갖게 된다.

셋째, 통계청의 위상 강화는 지방청에서도 나타나게 된다. 현재 지자체들은 통계청과 협력하여 빅데이터와 통계청 자료의 결합을 통한 분석을 원하지만 이에 대한 수행 역량의 부족 혹은 전문 인력의 부족으로 실효를 거두지 못하고 있다. 모바일 폰 데이터 랩에서 작업 경험과 전문성을 쌓은 인력들이 각 지방청에서 지자체와 업무를 같이 수행하고 지역에서 필요로 하는 통계를 생산함으로써 각 지방청의 전문화와 함께 위상 강화가 기대된다.

넷째, 세계적으로도 한국 통계청의 위상을 제고하는데 기여할 것이다. 빅데이터를 활용한 통계 생산에 관계된 연구들을 보면 한국의 사례들이 주요하게 거론된다. IT가 발전한 나라라는 점에서 한국이 빅데이터를 활용한 통계 생산에서 앞서 가리라는 기대를 받고 있다. 모바일 폰 데이터는 우리나라뿐만 아니라 많은 나라에서 공공통계 생산을 위해 주목하고 초점을 맞추는 분야이다. 모바일 폰 데이터의 통계 생산에 집중함으로써 이에 대한 기초 연구를 수행하여 방법론적인 우위를 갖고 각종 통계 생산을 효율적으로 진행함으로써 세계적으로 우위를 갖는 통계청이 되어 궁극적으로는 국제협력이나 ODA에서 선진 통계청이 될 것으로 기대된다.

제2절 체감경기통계

가. 배경 및 필요성

통계청은 개별 통계의 생산 외에도 국민경제와 지역경제의 핵심을 이루는 각 부문별 통계를 종합하여 국민에게 경제동향에 관한 정보를 제공하고 있다. 대표적으로 생산동향, 소비동향, 투자동향, 경기동향을 종합하여 산업활동동향을 매월 발표하고 있다. 또한 국민경제의 여러 부문을 대표하고 경기 대응성이 높은 경제지표들을 선정하여 이를 가공 종합한 경기종합지수를 산출하여 1981년부터 매월 작성 발표하고 있다. 경기종합지수는 시차 정도에 따라 3개 군으로 나누어 선행지수 8개, 동행지수 7개, 후행지수 5개로 구성되어 있다.

지역경제에 대한 정보로는 매분기 지역경제 동향을 작성하여 제시하고 있다. 지역경제동향에 대한 내용은 생산, 소비, 고용, 물가, 건설, 무역 동향 및 국내인구이동으로 구성되어 있다. 지역경제동향은 광역권과 17개 광역시도로 구분하여 제공되고 있다.

이와 같은 경제통계는 광범위한 부문에 대한 정보를 종합적으로 제공한다는 점에서 중요한 의의를 가지나, 통계작성의 수준이 포괄적이고 광범위하여 국민들이 일상적으로 체감하는 경기와는 차이를 보이는 경우들이 많다. 특히 국민들이 실제 경제활동에서 중요하게 느끼는 지역 생활경제권의 경기를 반영하기 어렵게 되어 있다.

따라서 신용카드사용에서 발생하는 빅데이터 정보를 활용하여 통계를 작성함으로써 국민들의 체감경기에 부합하는 경제통계의 작성이 필요하다. 또한 업종별 매출동향을 파악함으로써 국민의 창업지원에 필수적인 정보를 제공하고 생계형 자영업자 지원을 위한 효과적인 대책마련에 기여하는 것이 필요하다.

나. 과제 내용 및 활용 자료

통계발굴을 위한 과제 내용은 여신금융협회 등에서 수집하여 국세청에 제출하는 신용·직불카드 승인자료를 활용하여 개인소비 동향, 자영업 동향, 유류 소비동향에 관한 통계를 작성하는 것이다. 2015년말 현재 지급수단 중 신용·직불카드 비중은 53.8%, 금액기준 55.5%로 파악된다. 따라서 신용·직불카드 사용 자료를 활용함으로써 시의성 있게 전국 및 지자체 수준에서 작성하여 기존의 경기종합지수를 보완하고 지역경제정책 및 지역상권 분석의 기초자료를 제공하는 것이 가능하다.

통계청이 파악한 카드 승인내역은 사업자등록번호, 카드사명, 카드 종류(신용, 체크), 카드구분(개인, 법인, 기타), 승인금액, 거래일시(초 단위), 할부기간이며, 주유 관련은 품목, 수량 판매금액이다.

수집주기는 실시간으로 여신금융협회 등에서 국세청에 제출한다. 신용카드 승인 자료를 활용하여 작성 가능한 통계는 개인소비 동향, 자영업 동향, 유류소비 동향으로 구분되며 통계청이 파악한 바에 따르면 다음과 같은 내용의 통계 생산이 가능하다.

<표 53> 신용카드 승인 자료를 활용하여 작성 가능한 통계

구분	개인 소비동향	자영업 동향	유류 소비동향
내용	금액별 할부기간, 경기민감 업종 소비동향 파악을 통해 개인의 자금사정과 국민의 경기전망 및 인식정보 제공	지역별·업종별 업체수, 매출동향, 업종별 시간대별 매출액 등 자영업 영업동향에 대한 세부적인 자료 제공	자동차용 유류판매량을 통한 산업활동 및 개인의 경제활동 동향 정보 제공
작성 항목	업종별·지역별 소비동향, 금액별 할부기간, 경기민감 업종* 소비동향 등	지역별·업종별 매출동향 및 업체수, 지역별 매출 규모별 매출액, 특정업종 매출동향 등	자동차 용도별 판매량 및 평균 판매금액, 유류소비량 지수
작성 단위	업종별, 전국/지역별, 금액별/할부기간별	업종별, 전국/지역별, 매출·종사자 수 규모별 등	전국/시도별, 용도별

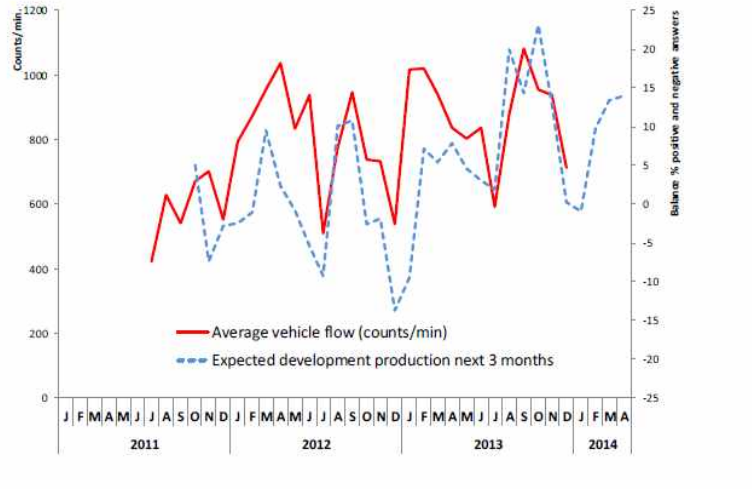
부문별로 작성 가능한 지표항목을 지표체계에 따라 제시하면 다음과 같다.

<표 54> 신용카드 승인 자료를 활용하여 작성 가능한 통계의 지표체계

부문	작성지표	작성항목
개인소비지출동향	지역·업종별	• 업종별 시·군·구별 카드승인액
	소비동향	• 전기·전월 대비 증감률
	개인자금사정 지수	• 지역별로 카드승인액 규모별 할부 건수 및 기간
		• 전기·전월 대비 건수 및 기간 증감률
	지역별 경기민감업종 소비동향	• 남녀정장 소매업 등 경기 민감업종의 평균 승인건수 및 금액
자영업 영업 동향	카드승인액지수	• 업종별 경상·불변·계절조정 지수
	지역·업종별	• 업종별 시·군·구별 카드승인액
	매출액	• 전기·전월 대비 증감률
	지역·업종별	• 업종별 시·군·구별 업체수
	업체수	• 전기·전월 대비 증감률
	지역·업종별	• 업종별, 시·군·구별 승인건수 및 평균 매출액
	평균매출액	• 전기·전월 대비 증감률

	지역·매출	• 시·군·구 단위로 매출규모별 매출액
	규모별 판매액	* 4~5개 구간으로 구분하여 집계
	지역·업종별	• 시·군·구별 업종별 시간대별 월단위 매출액
	시간대별 매출액	
유류소비량	특정업종 매출액	• 사회적 이슈가 되는 특정업종에 대한 업체수 및 매출액
	작성지표	작성항목
	유류 소비량 및	• 시·도 단위의 용도별 유류 소비량 및 평균 판매금액
	평균 판매금액	• 전기·전월 대비 증감률
	유류소비량지수	• 전국단위의 용도별 가격 변동분 반영·미반영 지수
		• 전기·전월 대비 증감률

이와 같이 카드승인내역을 통한 통계 생산이 확립되면 다음 단계로 체감경기 파악을 위해 다른 빅데이터와 융합으로 지수를 개발하는 것이 바람직하다. 우선 앞서 모바일 폰 데이터 랩에서 개발한 인구통계 및 이동통계와의 연계를 통해서 실제 지역경제에 영향을 미칠 인구 이동 현황을 고려한다. 또한 교통이동량을 통한 지역경제활성화 정도의 지표를 파악한 네덜란드 통계청의 사례를 참고하여(van Ruth, 2014), 교통통계를 결합하여 지수를 개발한다.



[그림 54] 아인트호벤의 교통량 통계와 제조업 예상 생산량 통계의 비교 (자료. van Ruth, 2014)

다. 제도적 장애요인 및 개선 방안

납세의무를 이행하기 위해 제출되는 신용·직불 카드의 정보는 원칙적으로 제3자에게 제공이 금지되지만 국가통계목적으로 통계청장이 요구하는 경우 제공이 가능할 수 있다. <과세자료의 제출 및 관리에 관한 법률>의 제11조 비밀유지 의무 규정에 따라서 과세 자료의 제3자 제공은 금지되나, <국세기본법> 제81조의13에 있는 비밀 유지의 예외 조항에 따라 납세자의 과세 정보를 제공할 수 있으며, 5항에 "통계청장이 국가통계작성 목적으로 과세정보를 요구하는 경우"가 예외로 명시되어 있다. 또한 <통계법> 제24조에서는 "통계의 작성을 위하여 필요한 경우에는 공공기관의 장에게 행정자료의 제공을 요청할 수" 있도록 되어 있다.

납세 자료의 통계목적 활용을 위해서는 법적 규정의 가능성 이외에 실제 국민들의 프라이버시에 대한 우려를 해소하기 위하여 비식별화 작업이 철저히 수행되는 것이 필요하다. 이를 위해서는 카드사명 삭제, 사업자 등록번호 비식별화 조치 등이 수반되어야 한다.

라. 추진 방안

추진 방안은 통계 생산과 통계 활용으로 구분하여 파악하는 것이 필요하다.

먼저 통계 생산을 위해서는 국세청 및 여신금융협회와의 협력관계를 바탕으로 원천자료를 정비하고, 자료간 연계를 통해 마이크로데이터를 생성하며 비식별화 및 보안조치를 수행하여 세부 통계지표를 작성한다.

다음으로 통계 활용으로는 세부 통계지표의 발표 이외에 세부 통계지표를 바탕으로 한 체감경기동향 지수를 개발한다. 또한 상권분석을 수행하고 있는 중소기업청 및 지자체들에게 해당 지표 통계를 제공함으로써 통계 생산의 활용성을 높인다.

마. 기대효과

첫째, 국민의 생활경제에 밀접한 경제동향 통계를 작성함으로써 체감경기에 부합하는 통계 작성이 가능해진다. 이를 통해서 국민의 경제통계에 대한 관심이 커지고 통계 활용도가 증진될 것으로 예상된다.

둘째, 경제통계의 시의성 문제를 보완할 수 있다. 속보성 통계 자료를 통해 기존의 경제통계가 가지는 시의성 문제를 보완할 수 있다. 이는 공공의 경제정책 수립뿐만 아니라 민간분야에서 통계청의 경제통계 활용도를 증진시킬 것으로 예상된다.

셋째, 지역상권 분석에 필요한 정보를 보완 제공할 수 있다. 현재 중소기업청이나 지자체가 많은 관심을 기울이고 있는 지역상권 분석을 위한 유용한 정보를 제공할 것으로 기대된다.

넷째, 지역에서 나타나는 다양한 지역경제 살리기 정책의 효과를 측정할 수 있는 구체적인 정보가 제공됨으로써 이와 같은 정책 효과에 대한 과학적 연구를 촉진시킬 것으로 기대된다.

5. 소결

본 장에서는 국내에서 시도되고 있는 빅데이터의 공공 활용 노력과, 민간에서의 이용 시도 들에 대해 정리하였다.

- 서울을 제외한 대부분의 지방자치단체는 빅데이터 활용에 필요한 인력/조직이 갖춰지지 못하였다.
- 관광, 경제 등 대부분의 자방자치단체에서 비슷한 주제의 빅데이터 활용에 관심이 있으며, 이는 특히 통신사나 카드사에서 정제된 형태로 제공받은 데이터에 의존하는 경향이 크다.
- 서울시는 빅데이터 활용이 가장 활발한 곳으로, 심야버스 노선 확충, 노인여가복지시설 입지 분석 등에 빅데이터를 활용하고 있다.

빅데이터가 활발하게 활용되는 분야는 교통분야로, 데이터의 공공재적 성격에 대한 합의가 이뤄지면서 데이터 활용이 탄력을 받았다. 상권분석 서비스도 중소기업청과 서울시에서 각기 제공하고 있으며, 카드결제정보, 휴대전화 위치정보 등을 통합하여 분석하고 있다.

- 결제, 환승정보, 네비게이션 및 카드 단말기 등의 위치 정보를 혼합하여 유동인구에 대한 포괄적인 분석이 가능해졌다.
- 상권분석서비스는 여러 지자체에서 시도하고 있으며, 민간 기업들도 리포트를 제공하는 등 데이터 활용에 대한 관심이 높은 분야이다.

빅데이터를 이용한 공공통계로서 본 장에서는 두 가지 과제를 제시하였다.

- 모바일 폰 데이터 랩: 휴대전화 정보는 단순 상권분석 뿐 아니라 유동인구 분석을 기반으로 한 안전, 교통 등 다방면으로의 활용 가능성이 큰 데이터이다. 따라서 이의 통계품질 관리가 시급하나 현재로서는 통일된 기준이 없는 상태이다. 다른 데이터와의 연계가 필수적인 데이터 성격상 통계청이 중심이 되어 데이터 연계 장소로서 통계청의 입지를 굳건히 하여야 할 것이다. 빅데이터 관리에 필요한 지침, 운영하는 랩의 중립성을 담보하는 조직, 및 규칙 제정이 필요하다.
- 체감경기통계: 기존의 통계청에서 생산해 온 경기종합지수는 포괄적인 정보를 전달한다는 의의는 있으나, 국민들이 체감하기 어려운 형태로 발표되어 온 것이 사실이다. 따라서 일반 국민들의 데이터 활용을 높이고, 통계에 대한 관심을 증진시키기 위해 실생활에서 자주 구매하고, 이용하는 상품을 중심으로 체감경기통계를 개발할 필요가 있다. 이는 국세청에서 수집하는 각종 카드 결제 자료를 활용하여 개인 및 자영업자의 매출 동향을 정리함으로써 가능하다. 일상생활과 밀접한 통계를 발표함으로써 경제통계의 시의성을 제고함과 동시에 지역의 경제활성화 정책 평가의 지표를 제공할 수 있다.

참 고 문 헌

고숙자·정영호. 2012. “국민건강 미래예측 시스템 구축 방안 - 빅데이터를 활용한 건강위험 예측 방안 모색을 중심으로” 보건복지포럼(193), 43-52.

김감영·이건학·정남수·강계화. 2015. 「모바일 폰을 활용한 서비스인구 추정 연구: 대구·경북 지역」. 통계청.

김경태·이인묵·곽호찬·민재홍. 2015. “유동인구 추정 시 통신 자료의 활용에 관한 연구” 서울도시연구, 제34권 제3호. 222-238.

김경태·이인묵·곽호찬·민재홍. 2016. “이동통신 자료 전수화를 통한 존재인구 산정 방안” 대한교통학회지, 제16권 제3호. 177-187.

김종학·고용석·김준기. 2015. “모바일 빅데이터의 국토교통 분야 활용 및 시사점.” 국토정책 Brief(499), 1-6.

김종학·고용석·김준기·박종일. 2016. “모바일 빅데이터를 활용한 재난대응방안.” 국토정책 Brief(563), 1-6.

손재기·신순애·한태화. 2015. “빅데이터를 활용한 라이프케어 동향” 한국통신학회지 32권 11호, 3-7.

송태민. 2013. “보건복지 빅 데이터 효율적 활용방안” 한국컴퓨터정보학회지 21권 1호, 45-53.

송태민. 2015. “소셜 빅데이터 분석과 활용 방안 - 메르스 정보확산과 위험예측 중심으로” 보건복지포럼, 29-49.

신병주·유성준. 2015. “라이프케어 실현을 위한 빅이데이터 활용” 한국통신학회지 제 32호 11권, 8-11.

오미애. 2014. “정부 3.0과 빅데이터: 보건복지 분야 사례를 중심으로” Issue & Focus(230), 1-8.

오상우. 2015. “건강보험 빅데이터의 의료계 활용” 의료정책포럼, 제12권 3호. 18-23.

이연희. 2015. “보건복지분야 공공 빅데이터의 활용과 과제” 보건복지포럼, 5-16.

이지영. 2015.「빅데이터의 국가통계 활용을 위한 기초연구」. 통계개발원.

이지혜·제미경·조명지·손현석. 2014. “보건의료 분야의 빅데이터 활용 동향” 한국통신 32권 1호, 63-75.

한국정보화진흥원. 2015.「2015년 빅데이터산업 10대 뉴스 및 이슈」.

진재현. 2015. “보건복지 분야 국가승인통계 작성 현황과 통계 신뢰도 제고 방안” 보건복지포럼, 51-59.

최준영. 2015. “의료정보시스템 운영에서 생성되는 의료 빅데이터의 활용가치” 한국전자통신학회논문지 10권 12호, 1403-1410.

Baker, R., Brick, J.M., Bates, N.A., Bbattaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. and Tourangeau, R. (2013). Report of the AAPOR Task Force on non-probability sampling. American Association for Public Opinion Research.

Braaksma, B., & Zeelenberg, K. (2015). ''Re-make/Re-model'': Should big data change the modelling paradigm in official statistics?. Statistical Journal of the IAOS, 31(2), 193-202.

Buelens, B., de Wolf, P., and K. Zeelenberg. (2016). Model based estimation at Statistics Netherlands. European Conference on Quality in Official Statistics (Q2016) Madrid, 31 May-3 June 2016

Florescu, D., Karlberg, M., Reis, F., Del Castillo, P. R., Skaliotis, M., & Wirthmann, A. (2014). Will ‘big data’transform official statistics. In Q2014 - European conference on quality in statistics.

Kitchin, R. (2015). “Big Data and Official Statistics: Opportunities, Challenges and Risks.” The Programmable City Working Paper 9.

Moss, M. L., & Qing, C. (2012). The Dynamic Population of Manhattan. Rudin Center for Transportation Policy and Management, Wagner School of Public Service, New York University: New York, New York.

Positium, L. B. S. (2014). Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics. Consolidated Report Eurostat Contract No 30501.2012. 001-2012.452, 31.

Tiru, M. (2014, October). Overview of the sources and challenges of mobile positioning data for statistics. In International Conference on Big Data for Official Statistics, Beijing

van Ruth, F. (2014). Traffic intensity as indicator of regional economic activity. Internal discussion paper, Statistics Netherlands.

Ⅵ. 빅데이터를 활용한 공공 통계의 외국 사례들

이기홍(한림대학교 사회학과)

Ⅵ. 빅데이터를 활용한 공공 통계의 외국 사례들

1. 서론: 외국의 빅데이터 기반의 공공 통계 개요

빅데이터 활용에 대한 열의는 전 세계적인 현상이다. 이 장에서는 외국의 중앙 정부, 지자체 등의 공공 기관이 빅데이터를 활용함으로써 공공 통계의 영역을 어떻게 확장하고, 또 궁극적으로는 공공의 어떻게 이익에 기여하는지를 살핀다. 미국, UN, 영국, 일본, 네덜란드 등을 다루되, 기본적으로 각 사례에서 빅데이터를 인식하는 방식 및 그것을 통해 사용자들에게 혜택을 주는 방식(예: 오픈 데이터)으로 나누어 정리할 것이다. 미국의 경우, 라이즈 콜로라도, 필박스 및 범죄 지도와 같이 공공 기관이 실시하는 빅데이터 통계를 통한 혁신적인 서비스를 보다 상세히 검토한다.

이어, 검토한 해외 사례들로부터 한국에서 실시할 만한 새로운 사업을 두 가지 제안한다. 하나는 미국의 범죄지도 서비스에서 착안하여 구성한 세이프 코리아 지수이다. 한국은 최근 새로운 형태의 범죄 및 그것에 대한 언론 보도 등이 급증하여, 대중들이 범죄에 대해 민감해졌으나, 몇 외국 사례들과는 달리 범죄 또는 안전도와 관련된 정보를 실시간으로 파악할 수 있는 서비스가 없다. 여기에서 제안하는 세이프 코리아 지수 서비스를 개시한다면 그러한 문제를 해소하는 데에 도움이 될 것이다.

또 하나는 '빅데이터 통계 활용 네트워크' 구축 제안이다. 빅데이터는 원래적 속성에 의해 그것의 종류와 (컴퓨터 파일로서의) 형태가 지속적으로 다양화할 수밖에 없다. 그러한 변화에 맞추어 빅데이터 기반의 공공 통계가 확장, 발전하기 위해서는 통계청을 중심으로 하여 전국을 아우르는 빅데이터 통계 활용 네트워크를 구축하여, 빅데이터 기반의 공공 통계 수요를 지속적으로 파악하고, 국가적 지원, 활용 체계를 구축할 것을 제안한다.

2. 본론: 빅데이터 기반 공공 통계 해외 사례

제1절 미국

가. 미 연방 정부의 정책 기조: 개방과 프라이버시 보호

연방 국가, 합중국으로서의 미국의 빅데이터 활용 전략은 Big Data: Seizing Opportunities, Preserving Values(빅데이터: 기회를 활용하여 가치를 지킴)[208]에

잘 요약되어 있다. 오바마 행정부에서는 빅데이터 활용과 관련하여 오픈 데이터(open data)와 프라이버시(privacy)를 두 키워드로 잡아, 정부가 확보한 빅데이터를 최대한 공개하되 빅데이터의 수집 및 활용 과정에서 프라이버시 역시 최대한 지킬 수 있도록 노력할 것이라고 밝혔다. 특히 보건, 교육, 국토 보안(homeland security), 지역별 치안(law enforcement) 등의 분야에 주력할 것이라고 명시하기도 했다.

이러한 연방 정부의 입장에 대해, 지방 자치가 발달한 미국과 같은 큰 나라에서, 시장 사정에 우선적으로 반응할 수밖에 없는 사적 부문이 어느 정도 협력할지 또한 다양한 지자체 법률과 제도가 연방 정부의 빅데이터 정책 주도에 맞추어 어느 정도 변화할지는 가늠하기 어렵다고 보는 것이 현실적이라고 여겨진다. 즉, 당분간은 상당수의 빅데이터 기반의 통계가 연방 정부가 아닌 지자체 또는 특정 지역에서 서비스될 것이라고 전망하는 것이 더욱 정확할 것이다.

따라서 한국의 통계청도, 미국이 연방 차원에서 제공하는 빅데이터 기반의 통계 못지않게, 지자체 또는 일부 지역에서 제공하는 빅데이터 기반의 통계 서비스에 주목해야 할 것이다.

나. DATA.GOV

위에서 밝힌 오바마 행정부의 첫째 기조는 오픈 데이터, 즉 데이터의 공개인데, 그것을 반영하는 형식 및 수단은 "DATA.GOV (https://www.data.gov/)"에 잘 드러나 있다. 즉 현재 미 연방의 빅데이터 활용 결과가 가장 뚜렷이 나타나는 사례가 바로 이 인터넷 서비스이다. 위에 소개한 거시적 전략 및 방침에 따라, 마치 상업용 검색 엔진처럼 관심 키워드의 조합을 입력하면, 미 연방 정부가 확보한 모든 데이터로부터 관련된 결과물을 정리하여 보여주도록 되어 있다. 지속적으로 업데이트되는 중이다.

이 서비스에서는 데이터 분석 결과를 농업, 비즈니스, 기후, 소비자, 환경, 교육, 에너지, 재정, 보건, 지방 정부, 제조업, 바다, 공공 안전, 과학 및 연구 등의 범주로 나누어서 검색할 수 있도록 되어 있기도 하다. 이러한 범주화를 한국 통계청에서 제공하는 서비스에 적용한다면, 한국적 상황에 맞도록 지속적으로

208) https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

변형할 수 있어야 할 것이다.

위의 분류에 따라 다양한 데이터를 시각화할 수도 있고, 모바일 기기에서 접근하기 편하도록 제공하고 있으며, 향후에는 현재는 초보 단계인 자연어(natural language) 검색, 음성 검색도 가능할 것으로 전망된다.

다. 미국 DATA.GOV의 데이터 공개 정책[209]

공개 정부 데이터(open government data)를 누구나 변형된 서비스 또는 앱의 형태로 사용할 것을 적극적으로 권장하고 있으며, 그 데이터는 접근 가능하고(accessible), 검열되어 있고(vetted), 활용 가능하다(available)고 명시되어 있다. 또한 무료로, 사용하기 위한 등록이 불필요한 방식으로 제공되는 것을 원칙으로 한다. 미 연방 정부 산하의 여러 기관들이 이 데이터를 활용하여 특화된 서비스 웹사이트 또는 앱의 형태로 재가공하여 제공하되 특정한 방식을 지지하지는 않는다.

미 정부 기관에서 오픈 데이터를 활용하는 주요 사례는 https://www.data.gov/applications 하단에, 관련 앱 카탈로그는 http://wiki.civicommons.org/App_catalogs/ 페이지에, 관련 모바일 앱은 https://www.usa.gov/mobile-apps 페이지에 정리되어 있다.

DATA.GOV의 데이터를 소프트웨어 개발자들이 잘 활용할 수 있도록 API 관련 정보 등을 제공하고 있으며, 미 정부의 데이터를 접근하기 위한 서비스, 앱 등을 제작하는 데 필요한 도구를 제작하는 데 활용할 수 있는 공개 소스 프로젝트도 있다.[210]

라. 기타 사례

(1) RISE(라이즈) Colorado: 미국 콜로라도주 교육부 통합 자료 시스템 사례

RISE[211]는 'Relevant Information to Strengthen Education(교육 강화 관련 정보)'의 약자이다. 미국 콜로다도주 교육부는 주 전체에 걸쳐, 유치원, 초중고 교육구 및 공립(주립 및 국립) 대학들로부터 학생들의 복지, 소득, 인력 데이터를

209) https://www.data.gov/applications
210) https://github.com/GSA/data.gov
211) http://www.rise-colorado.org/

수집해 하나의 통합 플랫폼에 제공함으로써 학생들의 학업 성취도를 취학 전 시기부터 대학 졸업 전 단계까지 분석할 수 있는 시스템을 마련하였다.

- 함의 -

학생들의 학교 생활 관련 정보를 학생과 학부모가 열람하기, 교직원 채용, 민원 제기, 검정 고시, 방과 후 학교 등의 기능을 중심으로 한 한국의 나이스(NEIS) 대국민서비스[212]와 다른 점과 그에 따른 함의는 아래 두 가지이다.

가) RISE Colorado는 학교 내로부터의 정보 중심인 나이스 대국민서비스와는 달리 학교 외의 정보도 담고 있다. 고등학생 때부터 일을 하는 경우가 많은 미국의 현실을 고려, 그들의 학교 외적 요인이라 할 수 있는 복지, 소득 등까지 고려하여 학업 성취도와의 상관, 인과 관계를 분석할 수 있는 근거를 마련할 수 있다. 현재 나이스 대국민서비스는 학생들의 학교 생활에 대한 내용을 중심으로 이루어지므로 각 학생의 상황을 다차원적으로 분석하기에는 역부족이다. 현재 서비스중인 고등학교 학생들의 학교 외적 정보를 추가하여 빅데이터 통계를 구성할 수 있다면, 고교생들의 학업 성취도 및 진학 관련 사항을 더욱 현실적이고도 입체적으로 분석할 수 있을 것이다.

나) RISE Colorado는 대학생들의 정보도 담고 있다. 대학 진학율이 OECD 국가들 중에서도 높고, 상당수의 청년 취업이 대졸 후에 일어나는 한국의 현실을 고려할 때, 제도 교육 전 과정 동안의 학교 내외로부터의 교육, 경제, 복지 등 빅데이터 기반의 통계를 축적/분석할 수 있는 환경이 조성되어야 할 것이다. 현재의 나이스 대국민서비스를 대학 등으로 확대 실시하고, 학교 외적 정보도 DB화한다면 RISE처럼 제도 교육 기간 동안의 학교 내외로부터의 정보를 종합하여, 학업 성취도와 취업 성공률 등을 가늠할 수 있는 근거 자료가 만들어진다.

다) 결론: [나이스 대국민서비스 + 학교 외 정보 + 대학생의 학교 내외 정보]와 같은 빅데이터 기반의 통계가 새로이 작성/분석된다면, 새로운 데이터 기반의 진학, 취업 지도가 가능해질 것이다.

212) http://www.neis.go.kr/

(2) 필박스(Pillbox)

미 국립보건원(NIH, National Health Institute)에서는 미국에서 제조, 판매된 알약 식별하기를 도와주는 필박스(Pillbox)[213] 서비스를 미 국립의약도서관(NLM, National Library of Medicine) 및 약물정보포털(Drug Information Portal)과 연계하여 무료로 실시한다. 따라서 원래는 약의 정확한 사용을 위해 유료로 유사한 서비스를 써야 했을 소비자에게 이익이 제공된다. 그뿐 아니라, 어느 지역에서 어떤 약에 대한 문의가 들어오는가를 사용자 기반의 빅데이터로 파악함으로써 어느 지역에 어떤 질병에 걸린 이들이 어느 정도 분포되어 있는지를 파악할 수 있는 근거 정보를 축적하고 있다. API를 통해 내부 데이터 베이스를 외부에서 쓸 수 있도록 제공하고 있으며, 필박스 내에 축적된 데이터를 사용하는 인터넷 서비스 또는 앱을 개발하는 데 필요한 각종 소프트웨어 도구도 제공하고 있다.

필박스는 사실상 충분한 의료 보험이 없는 상당수의 고령자들이 병원 치료보다는 약에 의존하여 건강을 유지해야 하는 미국의 실정을 감안할 때, 앞으로 종합 헬스케어 서비스의 일부로 편입될 가능성이 매우 높다고 할 수 있다.

- 함의 -

우리나라의 경우, 대한약사회에서 운영하는 약학정보원[214]과 건강보험심사평가원에서 운영하는 보건의료빅데이터개방시스템[215] 내 의약품 사용 통계가, 약과 관련된 다양한 정보를 파악할 수 있다는 점에서 미국의 필박스와 기능적으로 비슷하다. 단, 이 두 서비스는 일반인이 아닌 의약업계 종사다들을 위한 서비스이자 관련 보험 정보를 파악하는 용도 위주로 개발되었고 사용된다는 점에서 필박스와 다르다.

가) 우리나라도 급속도로 고령화하는 중이다. 앞으로 약 20여년 동안 이 추세는 지속될 것으로 전망되는데, 고령자들의 경우, 각종 약에 대한 의존도가 다른 연령대에 비해 높다. 따라서 필박스와 같은 소비자를 위한 약 관련 빅데이터 서비스가 우리나라에서 론치될 경우, 그 효용은 일반인 및 제약 회사에 매우 클 것으로 전망할 수 있다. 이러한 빅데이터를 기반으로 정부 통계를

213) 웹사이트 주소는 https://pillbox.nlm.nih.gov/ 이며, 이 맥락에서 필박스는 알약통 또는 알약상자 이란 뜻이다.
214) http://www.health.kr/
215) http://opendata.hira.or.kr/home.do

개발한다면, 일반인과 제약 회사에 모두 득이 될 것이다.

나) 앞으로 개방형 API의 발전 등으로 일반인 중심의 종합 헬스케어 서비스가 등장한다면, 약학정보원 및 보건의료빅데이터개방시스템 등이 모두 그러한 새 시스템으로 통합되어 관리될 것으로 전망된다. 우리나라의 경우, 국민 식별 번호가 있으므로, 개인을 중심으로 한 다양한 빅데이터의 합병이 비교적 용이하다. 필요하다면 법과 제도를 정비하여, 국민 보건 상황의 변화에 어떤 변화가 있는지, 어떠한 범주의 국민에게 어떠한 맞춤형 서비스를 제공할 수 있는지 등을 분석하기 위해, 빅데이터 기반의 헬스케어 관련 통계의 구축이 행정부 내에서 필요할 것으로 보인다.

다) 결론: [약학정보원 정보 + 보건의료빅데이터개방시스템 정보 + 의료 정보 + 보험 정보]와 같은 내용을 개인 식별 변수를 통하여 합병할 수 있다면 총체적인 헬스케어 서비스 데이터베이스가 구축되면, 국가 차원에서, 더욱 과학적인 보건 사업의 진행이 가능할 것이다.

(3) 범죄 지도

샌프란시스코 등에서 시작하였으나, 이제는 전 미국으로 확산되어 경찰이 제공하는 정보를 민간 회사가 지리적 프로파일링(geographic profiling)이 가능한 수준으로 바꾸어 서비스하는 방식으로 바뀌어가고 있다. 기본적으로 범죄 발생 일시, 장소, 구체적 유형 등을 세밀하게 분석하여 지도에 시각화하여주는 서비스이다. 경찰 및 지자체의 협조를 통해, 범죄 데이터와 지리 정보 시스템(GIS: geographic information system)을 결합함으로써 쉽게 제공할 수 있는 서비스이다.

최근 미국에서는 CrimeMapping.com[216], CrimeReports[217] 등의 회사가 경찰 및 미 정부의 오픈 데이터 관련 기관의 도움을 받아 운영하고 있다. 사용자가 지정할 경우, 지역, 기간, 유형 등의 조건에 따라 자료를 제공하기도 하며, 범죄 유형, 지역 등을 설정해 두면, 특정 범죄가 어느 지역에서 일어나는지에 대해 통보를 받을 수 있다. 미국의 경우 오픈 API가 발달되어 있는 편이므로,

216) 웹사이트 주소는 http://www.crimemapping.com/ 이다. 현재 샌프란시스코 경찰은 이 서비스로의 링크를 제공하고 있다.
217) https://www.crimereports.com/

같은 원천 데이터를 이용하더라도 더욱 다양한 서비스가 개발될 수 있다. 이러한 서비스는 궁극적으로 데이터 기반의 범죄 예측 시스템으로 발전될 것으로 보인다.

- 함의 -

안전행정부는 일찍이 2013년부터 이와 비슷한 서비스를 구상해 왔으며[218], 현재 2016년 현재 국립재난안전연구원은 '생활안전지도'[219]를 서비스하고 있다. 하지만, 미국의 경우, 범죄 유형이 매우 구체적으로 제시되는 반면, 생활안전지도에서는 범죄와 관련하여 제한된 범주에 대해 정보를 제공한다. 또한, 생활안전지도는 범죄 외에 미세 먼지, 산업 관련 정보 등도 함께 제공하는데, 몇 새로운 서비스의 경우, 전국이 아니라 일부 지역에 한하여만 제공하고 있다. 이와 별도로 정부는 '성범죄자 알림e 서비스'[220]를 제공하고 있다. 따라서 범죄에 대해서만 특화된 빅데이터 기반의 서비스는 현재 한국에 없다.

가) 현재 한국의 생활안전지도는 범죄 외의 다양한 정보를 제공하고 있다. 최근에는 과거에 비해, 묻지마 식의 범죄, 여러 주체로부터의 테러 가능성 등이 크게 보도되어 국민의 범죄 관련 심리가 불안해졌음을 감안하면, 미국의 사례처럼 범죄에 특화된 서비스를 세부 정보와 함께, 사실상 실시간으로 제공하는 것도 납세자의 삶의 질 향상에 도움이 될 것으로 보인다. 또한 이러한 빅데이터 기반의 범죄 통계를 상시 업데이트/모니터할 수 있는 체계를 갖추어, 과학적인 치안 대책의 마련에 기여할 수 있을 것이다.

나) 이러한 범죄에 특화된 서비스를 실시할 경우, 과거 여러 차례 혐오 시설 유치 거부와 같은 님비(NIMBY, Not in my Back Yard)성 지역 운동의 사례에서 나타날 수 있다. 즉, 범죄율이 높다고 판단되는 지역에서 해당 지역의 정보 공개를 반대할 가능성이 있다. 이러한 서비스를 계획할 경우, 데이터 서비스 자체 외에 법 제도의 정비 및 주민 설득 등도 충분히 고려해야 한다.

다) 결론: [생활안전지도 정보 + 성범죄자 알림e 정보 + 포괄적 범죄 정보 + 지리 정보 시스템]과 같은 새로운 빅데이터 기반의 통계가 작성/분석될 수

218) 경향신문 2013.4.11. [지금 논쟁 중:] 범죄 지도 공개 (http://news.khan.co.kr/kh_news/khan_art_view.html?artid=201304112131525)
219) http://www.safemap.go.kr/
220) http://www.sexoffender.go.kr/index.nsc

있으면, 궁극적으로 범죄에 대한 신속한 대처 및 그 예방에 더욱 과학적으로 기여할 수 있을 것이다.

제2절 UN

가. UN의 빅데이터 정책

UN(the United Nations, 국제 연합)은 글로벌 워킹 그룹(global working group)[221]이란 하위 조직을 설치하고 전 세계의 공동 번영을 위해 빅데이터의 활용법에 대해 연구 및 결과물 공유를 제의하였다. 하지만, UN 조직 자체의 실질적 구속력 및 지역과의 네트워킹의 한계 등으로 인해 그 결과가 뚜렷하다고 보기는 어렵다.

이와 관련된 가장 뚜렷한 결과물은 UN Global Pulse로서 UN이 (주로 개발도상국의) 발전과 인도주의적 행동을 위해 조직한 하위 조직이다.

나. UN의 오픈 데이터[222]

UN의 경제사회과(Department of Economic and Social Affairs) 통계부(Statistical Division)에서는 전 세계의 인터넷 사용자들을 위해 UN 데이터베이스를 이용할 수 있는 단일 창구를 마련하였다. 이 일은 지난 60여 년 동안 UN이 축적한 각종 데이터를 공개함으로써 모든 이들이 전 세계의 발전과 평화에 기여하기 위해 연구할 수 있는 자료를 적극 제공하겠다는 취지에 의해서 이뤄진 것이다.

UN에서 제공하는 자료를 UN이 설치한 서비스를 통하지 않고 직접 사용 가능토록 하기 위하여 API 관련 정보도 적극적으로 제공한다.[223]

221) http://unstats.un.org/unsd/bigdata/
222) http://data.un.org/Default.aspx
223) http://data.un.org/Host.aspx?Content=API

제3절 영국

가. 영국의 빅데이터

영국 통계청(Office of National Statistics, UK 또는 ONS UK)에서는 공식적 프로그램을 통해 빅데이터 관련 업무를 확대, 추가하려고 노력 중이다. 영국은 비교적 규모가 작으나, 잉글랜드, 스코틀랜드, 북아일랜드, 웨일즈 등으로 실은 나뉘어져 있는 연방적 나라이므로 빅데이터의 지역 특화적 활용에 대해 참고하기에 적당한 사례일 수도 있다. 현재, 영국 통계청 홈페이지[224]에서 위에 언급한 미국 연방 정부의 데이터 웹사이트와 유사하게 고유의 검색 엔진을 통해 영국 관련 각종 통계 및 그 분석 결과를 용이하게 볼 수 있는 서비스를 제공 중이다.

나. 개인 식별 방식[225]

영국의 경우, 한국의 주민등록번호와 같은 국민 식별 번호(national identification number)가 없다는 점, 통계청이 타 정부 부서의 행정 정보를 이용하기 위해서는 까다로운 절차를 거쳐야 한다는 점 등이 당면 난제이다. 따라서 영국 통계청 빅데이터과에서는 여러 개인 정보를 종합하여 개인을 식별하는 방식을 개발하려고 노력 중이다.

센서스 외 개인 식별 자료 구축을 위해, 영국식 일반주치의(general practitioner) 제도에 따라 거주지에서 가까운 병원에 등록한 개인 정보가 가장 포괄적이라고 판단되어 그 자료를 수집하는 것이 일차적으로 중요하다. 단, 의료 케어가 덜 필요한 건강하고 젊은 남성 등의 경우, 담당보건 기관에 개인 정보를 제공하지 않는 경우가 많고, 영국인들이 이름을 다양한 철자로 쓰는 경우가 많아, 소셜 미디어 자료 등에 나타난 거주, 이주 정보와의 매칭을 통해 개인 식별 방식을 개발하는 것을 당면 목표로 하고 있다. 또한 주거지로도 쓰일 수 있는 캠핑카(caravan home) 집결지를 위성 지도 데이터 등을 활용하여 파악해 센서스 등 주소 기반의 자료를 보충하는 데에 쓴다.

224) http://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/theonsbigdataproject
225) 2016.8.30. 영국 통계청 빅데이터과에서의 회의 내용에 기반하였음.

다. 영국의 오픈 데이터

DATA.GOV.UK를 통해 “정부를 공개한다! (Opening up Government)”라는 표어와 함께 제공하고 있다.

사용자가 웹 브라우저를 통해 공식 사이트에 직접 접속하지 않고도 오픈 데이터를 활용할 수 있는 API 관련 정보를 제공한다.[226]

영국 ONS의 관계자[227]에 따르면, ONS 내 빅데이터 관련 부서는 비교적 작으며, 여러 가능성을 실험하는 단계라고 자평한다.

라. 기타 사례: 소비자 물가 지수(CPI: Consumer Price Index)와 소매 물가 지수(RPI: Retail Price Index)[228]

이 사례는 영국 통계청에서 현재까지 수행한 빅데이터 기반의 통계 과제로서 가장 성공적인 것으로 자평하는 것이다.[229] 소비자 및 소매 물가를 전통적인 오프라인 방식이 아닌 인터넷의 빅데이터를 수집하여 평가하는 방식이다. 기본적으로 머신 러닝(machine learning)이 가능한 웹 스크레이퍼(web scraper: 빅데이터 자동 수집 프로그램)를 이용하여 인터넷에서 물가 자료를 수집하는데, 99% 수준의 정확도를 확보할 수 있었다. 이러한 과제를 통해, 식료품 가격이 CPI와 RPI의 가장 중요한 요소로서, 특히 수퍼마켓의 가격이 중요하다는 결론에 도달하였다. 단, 자주 발견되는 오분류(mis-classification, 예: ‘위스키’를 ‘럼’으로 분류) 사례를 정정하는 작업이 중요하다는 교훈도 얻었으며, 향후 영국 통계청 내부의 웹 스크레이퍼 수집 자료와 마이수퍼마켓[230]과 같은 외부 민간 기관의 자료를 비교하는 것을 기획 중이다.

- 함의 -

영국에서도 인터넷 기반의 데이터 물가 데이터 수집은 전통적인 오프라인 조사에 비해 비용이 적게 들기 때문에, 이루어지는 것으로 보인다. 이러한 영국 사례를 물가 조사에 도입할 경우, [물가 수집 봇이 온라인 몰로부터 직접 긁어 온 데이터 + 민간 온라인 가격 비교 사이트에서 수집한 데이터]와 같은 구조의 데이터를 합병할 수 있으면, 합병한 데이터 내에서 1, 2차 자료를 비교함으로서, 빅데이터 기반의 물가 통계를 구축할 수 있다. 이 경우 1차 자료는 원 자료이고, 온라인 가격 비교 사이트의 자료는 검증 수단이 된다.

226) https://data.gov.uk/data/api/
227) Karen Gask와의 2016년 4월 교신 근거.
228) file:///C:/Users/abc/Downloads/bigdataproject2015qtr1progressreport_tcm77-409604.pdf
229) 2016.8.30. 회의 내용 중
230) https://www.mysupermarket.co.uk/

제4절 일본

가. 일본의 빅데이터

일본 총무성은 2009년 빅데이터의 활용이 중요하다고 공식적으로 언급했다.[231] 단, 그 이후 전국적 후속 조치는 위에 요약된 미국, UN, 영국 등의 사례에 비해 덜 가시적인 적으로 판단된다.

일본의 통계수리연구소(統計修理研究所)[232]는 정부에서 설립하여 현재는 법인화되었으며, 일본의 공공 통계에 대한 연구 및 교육을 담당하는 기관이다. 그 연구소의 관계자[233]에 따르면, 일본에서는 빅데이터를 1) 대용량(예: 100억 사례 이상); 2) 다양한 포맷의 단일구조화 되지 않은 상태의 데이터 등으로 정의하는 등의 방향성이 있다. 단, 현재 중앙 정부에서는 전통적으로 수집한 통계에 비해 빅데이터가 지닌 대표성 등의 문제를 극복하기 어렵다고 판단하여, 빅데이터 기반의 통계를 정책에 적극적으로 사용할 계획이 구체적으로는 없다고 한다.

단, 일본은 지방 자치의 전통이 길고, 서양 국가들에 비해 한국과 제도적/문화적 유사성이 큰 편이므로, 여러 지자체 또는 중앙 정부의 하위 부서가 빅데이터를 공무 운영에 활용한 사례는 지속적으로 관심을 갖고 검토할 만할 것으로 보인다.

2016년 1월부터는 새로이 “마이 남바”(マイナンバー, My Number, 共通番号制度) 제도가 세무 행정을 위해 시행되었는데, 이것이 미래에 국민식별번호로 기능할 계획도 있다고 한다. 이 경우, 다양한 빅데이터를 개인 중심으로 합병하는 방식이 수월해져 앞으로 중앙 정부에서의 빅데이터 기반의 통계 활용 가능성이 높아질 수도 있다.

231) http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h26/html/nc131300.html
232) http://www.ism.ac.jp/
233) 박요성 교수와의 2016년 4월 면담 및 10월 회의 내용 근거.

나. 일본의 오픈 데이터

DATA.GO.JP로 명명되어 있으며 오픈 데이터(オープンデータ)라는 표현과 함께 관련 서비스 제공 사이트를 데이터 카타로그 사이트(データカタログサイト)라고 하기도 한다.

위의 사례로 검토한 서양 나라들의 경우와 달리, API 등을 통한 데이터 제공, 활용에는 적극적이지 않아 보인다.

다. 기타 사례:

(1) 일본 통계수리연구소 리스크해석전략연구센터의 금융 정책에서의 신용 리스크 통계 모델[234]

일본 통계수리연구소 내의 리스크해석전략연구센터(リスク解析戦略研究センター)[235]에서는 기업, 국가, 부동산 관련 대출에 대한 리스크를 아래와 같이 연구하는데 있어서, 빅데이터 기반의 통계를 활용을 모색하고 있다.

예금자의 예금을 은행이 기업에 대출하는 것이 일반적인 은행의 업무이다. 이 경우 기업이 도산하면 은행과 예금자는 피해를 보게 된다. 따라서 기업이 도산할 가능성을 계산하는 것이 은행과 예금자를 위해 매우 중요한 작업이 된다.

이러한 작업을 위해 일본에서는 중소기업청, 통계수리연구소, 일본은행, 전국신용보증협회연합회를 설립하여 '신용리스크데이터베이스(CRD, Credit Risk Database)협회'를 구성하여 여러 기업, 나라들의 데이터베이스를 작성하여 기업별, 나라별 위험률을 계산하여, 회원사에게 통보해 준다.

또한 한국의 장기주택담보대출에 해당하는 아파트론에 대해서도 연구한다. 해당 부동산에 대해 인터넷으로 가격 및 부대 시설 등을 최대한 조사(빅데이터에 해당)하고, 그 부동산을 전문가들에게 직접 방문 조사하도록 해(전통적 조사에 해당) 두 가격을 비교하여, 세입자가 나갈 확률을 계산함으로써, 아파트론에 대한 위험률을 계산하기도 한다.

234) 야마시타 사토시(山下 智志) 일본 통계수리연구소 리스크해석전략연구센터장 제공 자료
235) http://www.ism.ac.jp/risk/

(2) 일본에서의 공적 통계 마이크로 데이터[236] 이용의 현황과 새로운 대응: 공적통계마이크로데이터연구콘소시움 및 온사이트네트워크의 운영[237]

일본 정부의 통계수리연구소 등에서는 여러 조사를 국비 등으로 시행하는데, 그러한 조사 데이터를 2차적으로 최대한 활용하기 위해 '공적통계마이크로데이터연구콘소시움(公的統計ミクロデータ研究コンソーシアム)'[238] 및 '온사이트네트워크(オンサイトネットワーク, Onsite Network)'[239]를 운영한다. 온사이트 네트워크는 보안 관계로 인터넷이 일반적인 방식으로 연결되지 않은 특정 장소에서 데이터를 보고 분석할 수 있도록 하는 시설이다. (그러한 시설들의 인터넷과 같은 네트워크로 연결되어 있다는 뜻은 아니다.)

공적통계마이크로데이터연구콘소시움에서는 연구 계획서를 받아 심사하고, 합격된 경우, 통계수리연구소 또는 지정한 장소에서 신청한 데이터를 열람할 수 있다. 여행이 어려운 경우, 보안 관계로 광학 디스크에 데이터를 전달받을 수 있는데, 이 경우 일정 시간이 지나면 복사본을 폐기하는 것을 원칙으로 한다.

- 함의 -

일본의 경우 아직 중앙 정부 차원에서는 빅데이터의 활용에 대해 적극적이지 않은 듯하다. 다른 나라에서는 다양한 포맷의 행정 데이터를 빅데이터라고 간주하고 그렇게 일컫는 경우가 많은데, 일본의 경우 '공적 통계 마이크로 데이터'라는 전통적 명칭을 선호하는 것으로 보인다. 단, 올해 초부터 시행된 새로운 국민식별 번호 제도는 앞으로 행정, 민간 데이터의 통합을 가속화할 수 있는 좋은 계기가 될 것이다. 따라서 일본이 행정 데이터를 포함한 빅데이터의 활용에 어떻게 대응하는지를 지속적으로 주목할 필요가 있다.

제5절 네덜란드

가. 네덜란드의 빅데이터

네덜란드 통계청(CBS, 네덜란드어 Centraal Bureau voor de Statistiek의 약자)[240]은

236) 일본어의 '公的ミクロデータ'는 한국 등에서 다양한 포맷으로 작성되어 있는 행정 데이터에 해당한다고 할 수 있다.
237) 오카모토 모토이(岡本 基) 일본 통계수리연구소 연구 행정 담당 제공 자료
238) http://www.rois.ac.jp/tric/micro/moc/
239) http://www.rois.ac.jp/tric/micro/moc/onsite.html

빅데이터를 다방면으로 사용하려고 시도 중이다. 대학, 도시, 교통, 휴대폰 및 소셜 미디어 자료 등의 활용을 위해 노력 중인데, 국가 규모를 볼 때 한국과의 친화성이 잠재적으로 클 수 있을 것으로 보인다.

네덜란드 통계청은 2016.9.27. 빅데이터 통계 센터(영어명칭: Center for Big Data Statistics 또는 CBDS)를 출범시켰고, 한국 통계청과도 이미 MOU를 체결하였다.[241]

CBDS 센터장은 빅데이터 기반 통계 작성의 장점에 대해 언급하며, 품질과 지속성을 담보하는 것이 중요하다고 강조했다.[242] 그는 또한 경제 성장, 보안, 건강, 노동 시장 및 지속가능한 발전 목표들(sustainable development goals)에 대해 빅데이터 기반의 통계가 기여할 수 있는 바가 클 것으로 전망하면서, 특히 지속가능한 발전 목표들 관련 프로젝트의 경우 많은 지수들을 새로이 작성해야 하는데, 빅데이터의 활용이 이 부분에 응용될 수 있을 것으로 전망했다.

센터는 네덜란드 내외의 금융계, 학계, 국제기구, 정부의 통계 전문 기관 등 다양한 네트워크를 이미 확보하고 있다. 빅데이터를 사용할 때 자주 등장하는 프라이버시 문제를 검증받기 위해 프라이스워터하우스쿠퍼스(PwC, PricewaterhouseCoopers)와 같은 다국적 회계 감사 기업의 인증을 받기도 한다.

CBDS의 위치는 네덜란드 통계청의 Heerlen과 The Hague 사무실 두 곳으로 분산되어 있다. 이와는 별도로 도시 데이터 센터(urban data center)를 아인트호벤(Eindhoven)과 Heerlen 등에 설치하여, 학계, 스타트업, 기업 등을 연계하는 Brightlands Smart Services Campus 등에서 네덜란드 통계청이 확보한 데이터를 적극 활용할 수 있는 기반도 조성 중이다.

나. 네덜란드의 오픈 데이터[243]

오픈 데이터 부서는 통계청에서 수집한 각종 자료를 웹사이트의 검색 엔진을 통해 제공할 뿐 아니라, 프로그래머들이 어플리케이션을 통해 사용할 수 있는 일종의 API 제공의 확대를 위해 노력 중이다.

네덜란드는 국가 전반의 통계를 네덜란드 정부 데이터포털(Dataportaal van de Nederlandse overheid)[244]에서 17개 분야로 나누어 제공한다. 일반적으로

240) https://www.cbs.nl/
241) https://www.cbs.nl/en-gb/our-services/innovation/nieuwsberichten/big-data/cbs-launching-center-for-big-data-statistics
242) https://www.cbs.nl/en-gb/news/2016/39/cbs-starts-unique-initiative-for-big-data-research
243) https://www.cbs.nl/en-gb/our-services/open-data
244) https://data.overheid.nl/

2차 분석 자료를 제공하나, 일부 데이터는 원자료에 가까운 세부 자료를 열람 또는 다운로드할 수 있도록 되어 있다.

네덜란드 통계청이 오픈 데이터 프로토콜의 하나인 OData(the Open Data Protocol)를 통해 제공하고 있다. 기본적으로 1) 기분류된 범주 또는 키워드 입력 창을 통한 간단한 검색으로 이용할 수 있는 웹 기반의 서비스/앱인 StatLine[245]과, 2) 사용자가 통계청 웹사이트의 데이터를 보다 자율적으로 필터링 기능 등을 구성하여 사용할 수 있는 CBS OData, 두 가지 방식으로 오픈 데이터를 구현하고 있다.

- 함의 -

네덜란드 통계청은 최근 빅데이터 센터를 열었고, 한국 통계청과도 이미 MOU를 맺었다. 다방면에서 개방적이고 창의적인 정책을 보여주는 북서유럽 국가의 전형으로서, 최근 센터 개소를 계기로, 빅데이터의 활용 부문에서도 다른 나라들에 못지않게 적극적일 것으로 전망된다. 강소국, 무역이 발달한 나라라는 점 등이 한국과 비슷하다고 평가되는 바, 그들의 빅데이터 활용은 한국에 앞으로 강한 함의를 줄 것으로 보인다. 특히 이미 오픈 데이터 서비스로 통계청이 확보한 데이터를 외부와 공유하는 데에 적극적인데, 네덜란드에서 현재 제공하는 두 가지 방식 그리고 그들이 앞으로 어떻게 변화할 지에 주목할 필요가 있어 보인다.

제6절 기타 사례

가. EU GDPR(개인정보보호규정, General Data Protection Regulation)[246]

EU는 빅데이터 시대를 맞이하여 이전의 관련 규정을 대체하는 GDPR을 2016.4.27. 채택하고, 그로부터 약 2년 후 2018.5.25. 발효하도록 하였다. 주된 내용은 EU 회원국 국민들의 개인 정보를 갖고 있는 EU 내외의 기업들이 회원국마다의 법적 특이성을 초월하여 그 정보를 관리할 수 있도록 하며, 이를 어길 경우 엄격한 처벌 역시 일관성 있게 내릴 수 있도록 하는 것이다. 각 회원국은 각국의 특성에 맞는 감독 기관(supervisory authority)을 둘 수 있으나, GDPR을 따라야 하므로 EU 전체적으로는 상당한 통일성이 확보되도록 하는 것이다.

GDPR에 따르면, 개인 정보를 활용하는 기업 등은 그 정보를 어떻게 활용할 지에

245) http://statline.cbs.nl/Statweb/
246) 원문은 다음을 참조: http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

대해 개인들에게 명료하게 고지하고 명확한 동의를 받는 등의 과정에서 EU 전체 및 각 회원국의 지침을 따라야하므로, 개인 정보의 오남용에 의한 개인의 피해를 최소화할 수 있는 법적, 제도적 근거가 마련되는 것이다. 해당인이 원할 경우, 개인 정보는 삭제되어야 하며, 한 기관으로부터 다른 기관으로 옮기는 것도 가능한데, 이러한 절차 역시 GDPR이 명시하고 있다.

이와 관련, 몇 가지 쟁점 이미 제기되었다는 의견도 있다. 약간을 예시하면 다음과 같다.

1) EU 차원의 GDPR은 정보보호관(DPO, Data Protection Officer)을 회원국에 두도록 하는데, 이러한 새 제도는 각 나라의 법적/제도적 환경에 적당치 않을 수 있다.

2) GDPR이 기업이 관리하는 개인 정보에 초점을 두어, 각 기업의 피고용인 관련 정보 관리에 대해서는 충분히 다루지 않는다는 비판을 받고 있다.

3) 새 규정에 따르면 개인은 유사시 기업이 지정하는 정보보호기관(DPA, Data Protection Authority)을 접촉하여야 하는데, 유럽 바깥에 기반을 둔 상당수의 기업들은, 언어 장벽 관계로, EU 내 영어 사용국인 영국, 아일랜드에 위치한 정보보호기관을 쓸 것으로 보인다. 이 경우, 정보보호기관에서 영어 외의 서비스를 제공하지 않을 경우, 영어에 익숙치 않은 소비자들은 불편할 수 있다. 여러 DPA들 간의 일관성의 문제 역시 새로운 논란을 일으킬 수 있을 것으로 보인다.

- 함의 -

EU에 진출해 있거나 진출할 한국 기업들이 특히 주의해야 할 부분이다. GDPR이 본격적으로 발효하는 2018년 상반기에는 EU 전체의 GDPR뿐 아니라 각 회원국의 지침 역시 따라야 한다. 단, 한국 기업들의 경우 DPA는 영어권 국가인 영국 또는 아일랜드에 있는 것들을 쓸 가능성이 높다. 더욱이, 영국은 조만간 EU 탈퇴의 가능성도 있어, 향후 수년간 많은 빅데이터의 활용 범위 및 방식에 있어서 변화가 있을 가능성이 매우 높다.

나. 아일랜드 더블린의 대시보드(DublinDashboard)[247]

아일랜드의 국립 메이누스 대학(National University of Ireland Maynooth)에서

247) http://www.dublindashboard.ie/pages/index

운영하는 인터넷 서비스로서, 더블린 시의 도로 교통 상황, 대여 자전거 위치, 부동산 정보, 경제 지표 등 다양한 형식의 실시간 원자료(raw data)를 시 또는 데이터 서비스 기관으로부터 제공받아 일반인들이 보기 쉽게 시각화하여 전달한다.

운영의 효율을 위해 교통 상황과 같은 비디오 자료는 일부 원자료는 시에서 저장하지 않되, 개방형 API 등을 방식을 통해 스트리밍(streaming)함으로써 접근하는 주체가 저장할 수 있도록 배려해 준다. 또한 프라이버시 문제 등을 고려하여 해상도를 적절히 조정함으로써, 정보를 제한적으로 전달(예: 특정 지역에 몇 명이 있는지는 알 수 있으나, 누구인지는 알 수 없음)하기도 한다.

상시 교통난에 시달리는 더블린 시에 크게 기여하는 것으로 긍정적으로 평가되고 있으며, 아일랜드의 대표적인 빅데이터 기반 통계 활용 사례로서 유럽 및 다른 지역의 도시에서 참고할 만한 것으로 알려져 있다.

[그림 55] 더블린 대시보드 예시

- 함의 -

우리나라의 대도시는 유럽의 오랜 도시들에 비해 도로가 넓고 전철/지하철이 더욱 발전되어 있음에도, 상시 교통난을 겪는다. 더블린의 대시보드를 참고하여, 교통뿐 아니라 다양한 정보를 관련 기관이 -2차적으로 정리한 것을 열람시키는 것이 아닌- 스트리밍 및 원자료 제공에 초점을 둔다면, 대도시의 각종 문제 해결을 위한 민간의 빅데이터 활용에 정부/지자체의 입장에서 좋은 환경을 제공하는 것이다.

3. 결론: 연구의 의의, 한계 및 신규 과제 제안

제1절 이 연구의 소결, 의의 및 한계

가. 소결

정보화가 급진전하는 빅데이터 시대를 맞이하여, 여러 나라의 공공 부문에서는 빅데이터 활용을 통한 공공 통계의 확장을 고민하며, 여러 학계, 기업 등의 비정부 부문과 협력하면서 다양한 실험을 진행 중이다.

미국은 정부가 빅데이터 기반의 공공 통계를 활용함에 있어서, 공개 지향성과 프라이버시 보호를 주요 원칙으로 삼고 있다. 미국은 DATA.GOV를 통해 정부가 전통적으로 또 빅데이터 기반으로 새로이 작성한 다양한 데이터를 매우 적극적으로 공개하는 편이다. RISE Colorado와 같은 교육 기록 서비스, 종합 헬스케어 서비스로 발전할 수 있는 필박스(Pillbox), 범죄에 대한 다양한 정보를 실시간으로 볼 수 있는 범죄 지도 서비스 등은 우리나라에서도 적극 참고할 만한 것으로 보인다..

기타 해외 사례를 통해 상이한 포맷의 행정 데이터들을 통합하여 새로운 대형 데이터베이스를 구축하거나, 기존 정부 통계를 효율적으로 검증하기 위하여 빅데이터를 활용하려는 시도도 비록 파편적으로나마 상당히 진행 중이다. UN은 UN이 여러 나라로부터 수집, 통합하여 국제 비교가 가능한 다양한 데이터를 공개한다.

영국은 한국의 주민등록번호와 같은 국민 식별 번호가 없어 중앙 정부에서 또는 지역별로 주민의 수를 정확히 파악하는 것조차 현재 어려운데, 가장 포괄적이라고 여겨지는 의료 데이터를 중심으로 개인에 관련된 정보를 통합함으로써 새로운 행정 데이터를 통합하려고 노력 중이다. 일본은 겉으로 빅데이터 개념을 강조하지는 않으나, 다른 나라들과 마찬가지로 다양한 데이터를 이용하여 공공 통계의 신뢰도와 타당도를 검증하려는 다각도의 시도를 하고 있다. 특히 일본의 경우, 올해부터 시행된 마이남바(일종의 국민 식별 번호) 제도가 정착되고, 그것을 중심으로 현재는 흩어져 있는 행정 데이터들이 통합되면, 몇 년 내에 행정/빅 데이터의 활용이 급증할 수 있다. 그 외, 네덜란드, 아일랜드와 같은 작은 나라들로부터도 통계청의 빅데이터 센터, 도시 정보의 포괄적 제공 방식 등의 사례를 참고할 만하다.

이러한 해외 사례들을 참고하여, 통계청에서 앞으로 할 만한 신규 과제로서는 범죄 관련 빅데이터를 유관 행정 데이터 등과 결합하여 생성할 수 있는 세이프 코리아 지수와 빅데이터 통계 활용 네트워크 구축을 제안한다.

최근 한국에서 범죄와 관련된 일반의 경계심이 강화되고 있으나, 현재 한국에는 세부 범죄, 세부 지역에 대해 거의 실시간으로 파악할 수 있는, 몇 국가에서 서비스되고 있는 것과 같은 범죄 전문 지도 서비스가 없다. 생활안전지도는 범죄 외의 정보를 함께 제공하는데, 서울 기준으로 구 수준에서, 임의로 5등급으로 재정리한 내용을 보여주는 정도이다. 현재 부분적으로 제공되는 범죄 관련 데이터와, KICS와 같은 범죄 관련 원천 데이터, 지리 정보를 결합하면, 상당히 구체적인 수준에서 특정 지역, 특정 범죄에 대한 현황을 파악할 수 있으며, 역으로 특정 지역에서, 특정 범죄에 대해 어느 정도 안전한지를 수치로 표현할 수 있다. 이러한 포괄적 서비스가 가능하다면 국민의 행복도 증진 및 과학적 범죄 수사 및 예측에 큰 도움이 될 것으로 보인다.

통계청 중심의 빅데이터 활용 네트워크 구축도 또 하나의 중요한 과제로 보인다. 누구든지 정부가 확보해 둔 행정 데이터 외에 빅데이터의 수집을 정부에 요청할 수 있다면, 그러한 요청 자체가 일종의 빅데이터로 기능한다. 통계청과 같은 중앙 정부 기관이 빅데이터 활용에 대한 제안을 상시 접수하고, 좋은 제안을 바탕으로 추가 빅데이터를 수집하여 국민이 활용할 수 있도록 하는 것이다. 이러한 작업을 상시 업무로 수행한다면 어떠한 빅데이터를 공공 통계를 확장할 때 활용해야 하는지 역시 지속적으로 파악 가능할 것이다.

나. 의의

빅데이터 기반의 공공 통계 관련 해외 사례들에 대한 이 연구의 의의는 대략 다음으로 정리할 수 있다.

첫째, 미국, UN, 영국, 네덜란드, 일본 등 해외 주요 사례들을 1) 빅데이터 기반의 공공 통계 활용에 대한 기본 방침, 2) 현재 시행중인 오픈 데이터 서비스, 및 3) 관련 주요 사례 정도의 범주로 나누어 살펴보았다.

둘째, 각 나라, 각 기관 및 하위 기관에서 관리하는 데이터는 매우 다양하여, 그것을 통합하여 하나의 DB로 재구축하는 것이 매우 중요한 과제임을 알 수 있었다. 국민식별번호의 존재 여부가 그러한 작업의 난도를 결정하는 주요 요인임을 알 수 있었다.

셋째, 궁극적으로는 빅데이터 기반의 공공 통계를 어떻게 사용자들에게 전달하고(예: 오픈 데이터 정책 및 방식), 그것으로 국민, 사용자들의 삶의 질을 어떻게 높일 수 있느냐가 관건이었음도 확인하였다. 일부 좋은 사례들을

참고하여, 아래에서 신규 사업 둘을 제안하였다.

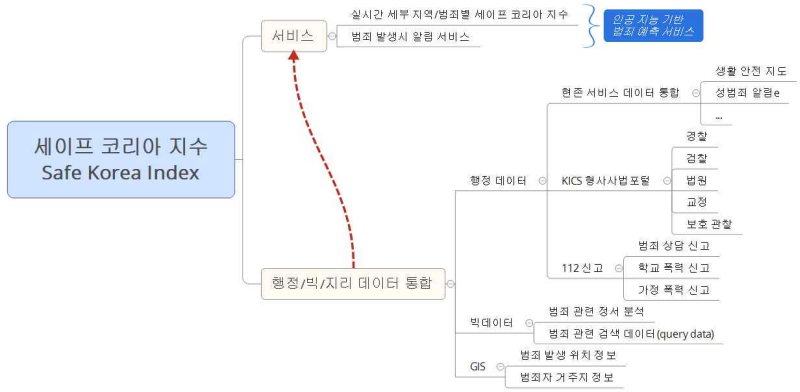
다. 한계

이 연구의 한계는 연구 기간의 짧음 등으로 인해, 해외 사례들을 직접적으로 충분히 검토하지는 못했다는 점이다. 일부 사례는 이차 문헌을 종합적으로 검토하고, 해외의 통계청 등 현지 담당자들과의 교신을 통해 연구할 수밖에 없었는데, 연구 전 및 초기 계획 당시와 보고서를 본격적으로 집필하는 시기 간에, 몇 사례는 변화하여 자료를 다시 수집하고 집필 내용을 수정해야만 하는 경우도 있었다.

앞으로 빅데이터 기반의 공공 통계 관련 해외 사례 연구는 특정 케이스를 장기적으로 추적하는 방향으로 보완될 필요가 있다고 본다. 예컨대, 미국의 범죄 지도 서비스 기관인 CrimeMapping.com[248], CrimeReports[249] 같은 곳이나, 영국의 ADRN, 네덜란드 통계청의 빅데이터 센터, 일본의 공적통계마이크로데이터연구콘소시움과 온사이트네트워크와 같은 사례들은 서비스의 유용성 및 국가의 규모 등을 고려할 때, 앞으로 한국 통계청 실질적 교류를 지속할 만한 기관으로 여겨진다. 빅데이터 기반의 공공 통계는 가변성이 심하므로 일회적으로 관찰하기보다는, 장기적으로 어떠한 유형을 보이며 변화하는지를 이해하고 참고하는 것이 매우 중요할 것으로 보인다.

248) 웹사이트 주소는 http://www.crimemapping.com/ 이다. 현재 샌프란시스코 경찰은 이 서비스로의 링크를 제공하고 있다.
249) https://www.crimereports.com/

제2절 '세이프 코리아 지수(Safe Korea Index)' 서비스[250] 제안



[그림 56] 세이프 코리아 지수 서비스의 구성 예시

가. 필요성

최근 혐오 범죄, 묻지마 범죄, 외국인 범죄 등 과거에는 범주화되지 않았던 새로운 유형의 범죄들의 등장과 더불어 일반인들의 범죄 관련 불안도가 높아졌다. 하지만 CrimeMapping.com[251], CrimeReports[252] 등 외국에서 일부 시행하는 서비스와 같이 일반인이 범죄 관련 세부 데이터를 지리 정보와 함께 실시간으로 열람할 수 있는 서비스는 없다.

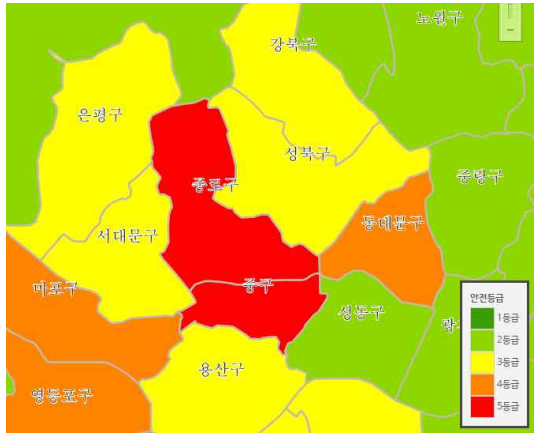
아래 [그림 57]는 미국의 범죄 지도를 예시한 것이다. 거의 블록 단위로 통계를 볼 수 있다.

250) 이 부분은 상당 부분 박순진 교수(대구대 경찰행정학과)와 최형민 박사(한국형사정책연구원)로부터의 자문 내용을 바탕으로 한다.
251) 웹사이트 주소는 http://www.crimemapping.com/ 이다. 현재 샌프란시스코 경찰은 이 서비스로의 링크를 제공하고 있다.
252) https://www.crimereports.com/

[그림 57] 미국 CrimeMapping.com: 샌프란시스코 시내 예시

현재 한국에서는 경찰, 검찰, 형사정책연구원 등 범죄 데이터를 갖고 있거나 분석하는 기관에서 제공하는 범죄 관련 정보는 대부분 2차적으로 분석한 결과이며, 국가 전체의 범죄 관련 원자료(raw data)는 제시하지 않는다. 지역 수준의 자료는 모든 지자체에 대해 통일적으로 제시되는 경우는 드물며, 지역에 관한 자료는 광역별이거나, 대도시의 경우 구 수준에 그치는 경우가 대부분이어서, 일반인이 범죄가 발생한 구체적 위치에 대한 정보를 얻기는 매우 어려운 실정이다.

아래 [그림 58]은 한국에서 현재 제공되는 생활안전지도를 예시한 것이다. 제공 기관 측에서 분류한 5등급으로 현재 서비스 가능한 가장 세부적인 단위인 구 수준에서 보여주는데, 미국의 범죄 지도에 비하면 얻을 수 있는 정보가 그리 구체적이지 못하다.



[그림 58] 현재 제공되는 생활안전지도: 서울시의 구 수준 예시

범죄 정보 관련 인터넷 서비스를 보면, 국립재난안전연구원은 범죄 외의 정보를 함께 제공하는 생활안전지도 서비스[253]는 원활하지 않으며, 범죄에 대한 구체적인 정보가 없고, 시 또는 구 수준의 집단 데이터를 임의로 몇 등급으로 나누어 제공한다. 성범죄 관련으로는 성범죄 알림e 서비스[254]가 따로 있는데, 이것의 경우 생활안전지도에서 제공하는 정보와 중복되는지 또는 배타적인지를 불분명하다.

이러한 상황에 비추어, 중앙 정부에서 범죄 관련 공공 통계 및 범죄 관련 빅데이터를 통합하여 국가 전체 및 지역에 대한 새로운 형태의 안전 정보 및 지수, 즉 가칭 '세이프 코리아 지수(Safe Korea Index)'를 실시간으로 서비스하는 시스템을 만드는 것은 일차적으로 국민들의 범죄에 대한 공포를 덜고, 그들의 행복도를 궁극적으로 증진시키는 데에 기여할 것이다. 나아가 범죄 관련 데이터를 실시간으로 통합하여 관리하는 체제를 구축하여 인공 지능과의 연결을 꾀한다면, 과거의 패턴을 바탕으로 미래의 범죄를 예측 그리고 예방하는 데에도 도움이 될 것이다.

253) http://www.safemap.go.kr/
254) http://www.sexoffender.go.kr/index.nsc

나. 데이터 구성의 개요

전국 및 지역의 안전 정도를 실시간으로 보여줄 수 있는 세이프 코리아 지수 구성을 위해서는 1) 범죄 관련 행정 데이터, 2) 범죄 관련 빅데이터 및 3) 지리 정보를 결합한 새로운 구성의 데이터가 필요하다.

- 범죄 관련 행정 데이터

이미 서비스되고 있는 생활안전지도에서의 범죄 데이터, 성범죄 알림e 등의 정부 데이터를, KICS(Korea Information System of Criminal Justice Services, 형사사법포털), 112 범죄 신고와 결합한다.

KICS는 일선 경찰에서 접수한 범죄에 대해 입력하는 방식으로 최초 데이터가 작성되는데, 앞으로는 이 시스템을 중심으로 경찰, 검찰, 법원, 교정, 보호 관찰 과정을 모든 정보가 통합될 계획이다. 나이, 성별, 거주지 등의 범죄자 신상, 범죄 발생 시각/위치, 범죄의 성격, 범죄의 경위, 관련자 등을 모두 포괄적으로 입력하도록 되어 있다. 단, 현재는, 각 기관별 포맷 및 입력 방식이 다르고, 긴밀히 협조해야 할 기관들 간의 열람 승인, 공유가 매우 제한적으로 또 간헐적으로 이루어지고 있어, 국민의 안녕과 정부 운영의 효율성을 위해서는 범죄 관련 데이터의 통합적 관리가 시급히 이루어져야 할 것으로 보인다. 이를 바탕으로 수사율, 검거율 등의 범죄 관련 통계가 실시간으로 파악될 수 있어야 한다.

또한 112 범죄 신고 및 그것으로 연결/통합된 기타 범죄 신고, 즉 범죄 상담, 학교 폭력, 가정 폭력 등에 대한 신고 데이터를 체계적으로 수집함으로써 행정 데이터에 포함시킨다.

- 범죄 관련 빅데이터

이 데이터는 범죄 관련 정서 분석 및 범죄 관련 검색 데이터(query data)로 구성된다. 범죄 관련 정서 분석은 뉴스 포털, 소셜 미디어 등 인터넷에 제시된 특정 범죄 정보에 대해 네티즌들이 어떻게 반응하느냐를 질적으로 분석하는 것을 뜻한다. 이 작업은 전문 소프트웨어 또는 코딩을 통해 구현할 수 있다. 이 분석 결과를 통해 특정 지역에서 일어난 특정 범죄에 대한 부정적 반응이 큰 경우, 가중치를 적절히 부여함으로써, 안전 지수를 낮출 수 있다.

범죄 관련 검색 데이터는 네이버, 다음, 구글 등의 주요 포털의 협조를 통해 구할 수 있다. 즉 어느 위치에서, 언제, 어떤 범죄 또는 범죄자와 관련된 검색이 일어나는지를 파악함으로써, 지역 주민들의 범죄 관련 관심사 및 지역에서 잠재적으로 발생할 수 있는 범죄에 대한 예측 데이터로 사용할 수 있다.

- 지리 정보 시스템(GIS, geographic information system)

위에 제시한 범죄 관련 행정 데이터와 빅데이터를 지리 정보 시스템과 결합함으로써 언제, 어디에서, 어떤 범죄가, 어느 정도 빈번히 일어났는지를 파악할 수 있는 데이터가 완성된다. GIS 작업을 위해서는 전문 소프트웨어를 쓰거나 통계청의 통계지리정보 서비스[255]를 쓸 수 있을 것이다.

다. 서비스 구성의 개요

위에서 범죄 관련 행정 데이터 및 빅데이터를 구체적 범죄 종류, 범죄 발생 장소의 지리 정보, 범죄 발생 시각, 범죄 발생 빈도 수준으로 체계적으로 그리고 확보 가능한 기간 동안의 모든 내용을 시계열적으로 입력하고, 추가 분석을 통하여 범죄 발생 정도와 반비례하도록 가칭 '세이프 코리아 지수(Safe Korea Index)'를 작성할 수 있다. 이 지수는 적어도 세부 지역별, 세부범죄별, 세부 시간대별로 검색 및 열람이 가능해야 한다.

추가적으로 이동 통신 시스템과 결합함으로써 사용자가 원하는 대로 특정 범죄가 특정 지역에서 발생했을 때, 푸쉬 알림(push notifications)을 통해 공지함으로써 범죄에 대한 경각심을 고양하고, 추가 피해를 막는 데에 기여할 수 있다.

추후 이러한 시스템이 잘 정착되고, 적절한 수준의 인공 지능과 결합되면 범죄 발생 예측 서비스로 이어질 수도 있다.

라. 추가 고려 사항

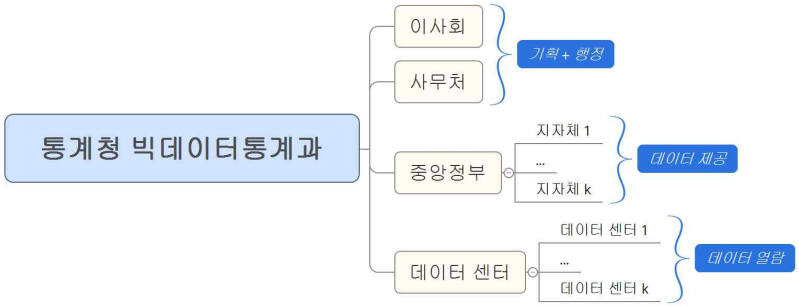
범죄자 관련 정보를 관리할 때, 법이 정하는 대로 개인 정보를 보호하기 위해서는, 여느 다른 빅데이터 관련 업무와 마찬가지로 비식별화 작업이 철저히 진행되어야 한다.

KICS를 중심으로 범죄 관련 기본 정보가 통일되고, 경찰, 검찰, 법원, 교정,

255) http://sgis.kostat.go.kr/view/index

보호 관찰 관련 기관들의 데이터 추가 입력 및 상시 전체 공유가 가능해지면, 세이프 코리아 지수 서비스뿐만 아니라 사법 행정 전반에 대한 과학적 연구 및 개선이 현재에 비해 훨씬 더 수월해진다.

제3절 '빅데이터 통계 활용 네트워크' 구축256) 제안



[그림 59] 빅데이터 통계 활용 네트워크의 구성 예시

가. 필요성

센서스가 표본 조사로 대체되었으므로, 국가 차원에서 자료 기반의 정책 수립을 위한 행정 자료 및 빅데이터를 적극적으로 활용하여 센서스 관련 자료를 보강하고 매칭하는 작업이 더욱 중요해진다.

행정 자료 및 관련 빅데이터는 형식들이 매우 다양하고 가변적이므로, 자료의 합병(merge)과 적절한 분석 방법 역시 정형적으로 유지되기 어려워, 데이터의 형식과 내용을 잘 아는 전문가 집단이 그것들을 공익적으로 이용할 수 있도록 가공, 분석할 수 있는 좋은 환경을 만드는 것이 필요하다.

현재 한국정보화진흥원에서 제공하는 공공데이터포털257), 서울시의 빅데이터캠퍼스258), 경기도의 빅파이센터259)에서 제공하는 공공 데이터는 내용적으로 또는 지역적으로 제한되어 있어, 통계청에서 국가 차원의 데이터

256) 2016.8.30. 영국 Southampton 대학의 Peter Smith 교수와의 면담을 기반으로 한 내용임. Peter Amith 교수의 소개 페이지: http://www.southampton.ac.uk/demography/about/staff/pws.page
257) https://www.data.go.kr/
258) https://bigdata.seoul.go.kr/
259) http://www.gg.go.kr/big-fi-center

서비스를 빅데이터와 함께 제공하는 것이 필요하다.

적어도 다음의 쟁점에 대해 방침을 굳혀야 네트워크가 순조로이 가동할 수 있다.

1) 자료를 배포 전에 어떻게, 어느 정도 비식별화할 것인가?
2) 자료는 사용자에게 어떤 형태로 배포할 것인가? (인터넷 배포는 보안상 위험할 수 있다.)
3) 자료를 열람하는 환경은 어떻게 구성할 것인가?

나. 참고 사례 1: 영국의 행정 자료 연구 센터(ADRN)260)

영국은 연방 구분에 따라 잉글랜드, 스코틀랜드, 웨일즈, 북아일랜드에 행정 자료 연구 센터(ADRC, Administrative Data Research Centre)를 설립하고, 서로 협조하는 네트워크로 기능하는 ADRN을 학계를 중심으로 구축하였다. ADRN은 대학, 정부, 지자체, 국가 통계 관련 기관, 시민 사회, 재정 후원자, 기타 연구자들을 아우르고 있다. ADRN은 심사 위원회(Approvals Panel)를 통해 접수된 연구 계획서를 철저히 검토하여 비식별화된(de-identified) 관련 데이터를 사용할 수 있도록 조치하되, 연구 후 결과까지 검토하는 기능을 수행한다.

내부구조를 개략적으로 살펴보면 다음과 같다.

가. 대외 사무를 총괄하는 행정 데이터 서비스(Administrative Data Service)
나. 학계를 중심으로 한 각 지역(잉글랜드, 북아일랜드, 스코틀랜드, 웨일즈)의 센터
다. 각 지역의 센터와 파트너쉽을 맺고 있는 국가 통계 관련 기관
라. 각종 데이터를 일차적으로 수집, 보관하고 있는 정부와 지자체
마. 영국 경제사회연구재단(ESRC, Economic and Social Research Council)261)
바. ADRN의 이사회를 운영하는 영국 통계원(UK Statistics Authority)

잉글랜드의 경우, ADRC for England(또는 ADRC-E)는 사회 통계/인구학 자원이 꽤 있는 사우샘튼 대학(University of Southampton)의 통계과학연구원(Statistical Sciences Research Institute)의 일부의 형태로 있다. 사우샘튼 대학의 ADRC-E에서는 인터넷을 통해 근방 모처에 저장된 각종 행정 데이터의

260) https://adrn.ac.uk/
261) 우리나라의 연구재단과 비슷한 성격으로 주로 연구비 지원을 담당하기에 '경제사회연구재단'으로 번역하였다.

원본 또는 분석 결과를 열람할 수 있는 랩과 같은 시설을 보유하고 있다. 즉, 보안 등의 이유로, 행정 데이터 자체는 ADRC-E에 보관되지 않고 독립된 데이터 센터에 저장하는 방식을 취하고 있다.

오래된 지방 자치의 전통에 따라 동일한 또는 유사한 내용의 행정 자료도 각 지역에 따라 다른 방식, 포맷으로 작성되어 있는 경우가 많아, 특정 정보를 국가 차원에서 확보하기 위해서는 ADRN 및 각 지역의 센터를 중심으로 한 참여 기관의 자발적 활동과 긴밀한 협조가 매우 중요하다.

다. 참고 사례 2: 일본의 공적통계마이크로데이터연구콘소시움 및 온사이트네트워크

일본의 경우, 빅데이터 활용을 독려하는 기관이 공적통계마이크로데이터연구콘소시움(公的統計ミクロデータ研究コンソーシアム)[262] 및 온사이트네트워크(オンサイトネットワーク, On-site Network)[263]로 이원화(二元化)되어 있는데, 전자는 데이터의 관리, 계획서 검토 등을 맡고, 후자는 데이터를 열람하는 물리적인 공간으로서 역할한다.

이러한 기관을 통해 이차적으로 이용할 수 있는 일본의 공적 통계는 다음과 같다. 현재까지 축적된 것들 중, 일부 년도 자료에 대해 제한적으로 제공하고 있다. 일본 노동후생성(勞働厚生省) 관장하는 (마지막) 국민생활기초조사를 제외하고는, 모두 총무성(總務省)에서 수집한 것이다.

국세조사(國勢調査): 다른 나라의 센서스에 해당하는 조사
노동력조사(勞働力調査)
주택·토지통계조사(住宅·土地統計調査)
전국소비실태조사(全國消費實態調査)
취업구조기본조사(就業構造基本調査)
사회생활기본조사(社會生活基本調査)
국민생활기초조사(國民生活基礎調査)

262) http://www.rois.ac.jp/tric/micro/moc/
263) http://www.rois.ac.jp/tric/micro/moc/onsite.html

라. '빅데이터 통계 활용 네트워크' 구축 제안

필요성: 한국 역시 세계적인 추세와 유사한 방식으로 센서스가 행정 자료로 대체, 보충되므로, 앞으로 행정 데이터 및 행정 데이터의 근거 자료가 될 수 있는 빅데이터의 활용이 더욱 필요할 것이다.

구성 예시: 이원화된 일본과 달리 조직의 밀도가 더 높은 영국의 사례를 바탕으로 하면 아래와 같이 적어도 약 6개의 기관이 유기적으로 연결된 네트워크를 구성한다는 초기 계획이 있어야 할 것이다.

<표 55> 빅데이터 통계 활용 네트워크 조직 구성 예시

기관명(가칭)	역할	비고
사무처	연구 계획서 접수를 포함한 대외 업무 및 내부 기관들 간의 행정적 연결을 담당한다.	행정 기능
데이터 센터	연구진이 행정 데이터를 분석하는 물리적, 가상적 장소로 기능한다.	데이터 열람 및 분석하는 장소
통계청	데이터를 수합, 정선하여 직접 또는 보안 연결을 통해 데이터 센터에 제공한다.	모든 행정, 기획을 총괄한다.
중앙 정부 및 지자체	각종 행정 데이터를 일차적으로 수집하고 업데이트한다.	
재정지원처	네트워크의 운영을 재정적으로 지원한다.	연구재단 또는 기타 정부 내외의 기관
이사회	네트워크의 장단기적 계획을 세우고 감독한다. 정부 및 국회 등과 연락을 담당할 수 있다.	기획 기능

위 네트워크를 통한 공모 주제 예시: 주제는 자유롭게 하되, 상기에서 한국 관련 함의를 제시한 것들을 기획 공모 방식으로 제시할 수도 있다. 계획서의 내용 자체가 어떤 빅데이터를 정부가 기획보한 공공 통계에 더하여 수집해 주기를 원하는 지를 가늠할 수 있는 일종의 빅데이터로서 기능한다. 연구 계획서에 주제어(key word)를 명시하도록 하여, 그 데이터를 시계열적으로 축적하면 언제, 어떤 빅데이터에 대한 요청이 있었는지를 체계적으로 평가할 수 있다.

예 1) 학교 내외의 다양한 데이터 합병을 통한 학생 경력 관리 통계 구축 및 분석 모형 제시: 나이스(NEIS) 대국민서비스에 1) 유치원 및 대학생 시기 추가 및 2) 아르바이트 소득, 가족 등의 학외 정보를 추가함으로써, 유치원부터 대졸까지의 시기 동안 학업 성취도와 최종 취업 간의 관계를 파악할 수 있는 데이터 및 분석 모형을 제시한다.

예 2) 의약 제품 및 서비스, 의료 보험 등 종합 헬스케어 데이터 구축 및 의료비, 건강 상태 등 분석 모형 제시: 약학정보원, 보건의료빅데이터개방시스템, 약 구입 경험, 의료 서비스 경험, 보험 등의 데이터를 기반한 통계 자료를 구축하여 의료비, 건강 상태 등을 예측하는 모형을 제시한다.

참 고 문 헌

http://bigdata.seoul.go.kr
http://data.gov.uk
http://data.overheid.nl
http://data.un.org
http://data.un.org
http://ec.europa.eu
http://opendata.hira.or.kr
http://pillbox.nlm.nih.gov
http://sgis.kostat.go.kr
http://unstats.un.org
http://www.adrn.ac.uk
http://www.cbs.nl
http://www.crimemapping.com
http://www.crimereports.com
http://www.data.go.kr
http://www.dublindashboard.ie
http://www.gg.go.kr/big-fi-center
http://www.github.com
http://www.health.kr
http://www.ism.ac.jp
http://www.mysupermarket.co.uk
http://www.neis.go.kr
http://www.ons.gov.ujk
http://www.rise-colorado.org
http://www.rois.ac.jp
http://www.safemap.go.kr
http://www.sexoffender.go.kr
http://www.soumu.go.jp
http://www.whitehouse.gov

Ⅶ. 결론

김영원(숙명여자대학교 통계학과)

Ⅶ. 결론

본 연구에서는 빅데이터를 활용한 통계생산을 위해서 향후 통계청에서 중점적으로 추진해야 할 과제를 개발하고 이를 실현하기 위한 구체적인 발전 전략을 수립하기 위해 필요한 핵심 사항들을 주제별로 제시하고 있다. 본 연구를 통해 도출된 주요 연구 결과를 정리하면 다음과 같다.

첫째, 빅데이터를 활용한 통계를 생산하는 경우 신뢰성 제고를 위해 필수적인 품질검증을 위한 기본적인 틀을 제안하였다.

빅데이터의 품질검증은 통계청에서 수행하고 있는 기존의 국가통계 품질진단과는 차별적으로 정의되고 구축되어야 하는데, 이는 전통적인 자료 수집 방법과는 달리 빅데이터의 경우에는 자료 수집이 이루어지는 단계에 통계작성을 주관하는 연구자나 사용자가 관여할 수 없기 때문에 자료 수집 과정에서 적절한 통제를 통해 자료의 신뢰도를 높이는 데 한계가 있을 수밖에 없다. 따라서 작성 목적에 맞는 빅데이터 통계 생산을 위해서는 빅데이터 생성 과정을 사후적으로 면밀히 살펴보고 자료의 활용 가능성을 확인하여야 한다.

본 연구에서는 이러한 빅데이터의 특성을 감안하여 빅데이터 품질 검증을 위한 기본적인 틀을 제공하였다. 품질검증을 위해 먼저 빅데이터 통계 작성 과정을 크게 input, throughput 그리고 output의 세 단계로 구분하고 각 단계별로 평가가 이루어져야 하는 요소들을 정의하였고, 이런 요소들을 구체적으로 평가하기 위한 기준을 질문 형식으로 구현해 제시하였다.

아직까지는 초보적인 단계에 머물러 있는 빅데이터의 공공분야에서의 활용 및 이를 통한 국가승인통계 작성을 위해서는 국가통계 작성이라는 관점에서 볼 때 빅데이터가 갖고 있는 근본적인 한계와 문제점을 직시할 필요가 있으며, 이를 위해서는 본 연구에서 제시한 기본적인 틀을 기초로 해서 향후 보다 구체적으로 빅데이터 품질 검증 도구가 체계적으로 개발되어야 한다. 이런 품질검증 과정을 토대로 사용 목적에 적합하면서 신뢰할 수 있는 빅데이터 활용 국가통계를 생산하는 방법을 구축해 나가야 할 것이다.

둘째, 빅데이터 환경하에서 새로운 통계의 생산을 위한 통계법제의 적절성을 파악하는 동시에 이러한 새로운 통계생산 체제와 현행 개인정보 보호법제와의 조화가능성을 검토하였다.

최근 국가·사회 전반에 걸친 정보화의 물결은 통계와 관련한 새로운 국면을 열어가고 있다. 특히 '빅데이터(Big Data)' 기술을 필두로 새롭게 부각되는

정보처리기술은 기존 통계영역을 넘어선 새로운 통계의 작성은 물론 기존 통계자료에 대한 새로운 방식의 접근을 가능케 하고 있다. 그렇지만 이러한 새로운 통계의 생산은 과거에는 문제되지 않았던 새로운 문제들 특히 개인정보보호의 영역과 밀접한 관련을 맺는 문제점들을 초래할 우려가 없지 않다.

여기서는 향후 새로운 통계생산 체제의 구축과 관련된 본격적인 법제적인 논의의 기초를 마련한다는 측면에서 현행 우리나라 통계법제의 특성을 살펴보았다. 우선 통계제도의 헌법적 근거로는 헌법 제127조 제1항, 제2항 등을 확인할 수 있었고, 통계제도의 전반적인 운용양상을 확인하는 가운데 통계작성을 위한 자료조사의 법적 성격을 행정조사의 범주로 규정하여 통계법 및 행정조사기본법의 관련조항들이 활용될 수 있음을 검토하였다. 아울러 현행 개인정보 보호법제들은 개인정보 보호론 쪽에 무게를 두고 있음에도 통계작성을 위한 근거를 충실히 마련하고 있음을 확인하였지만, 추후 개인정보 보호론의 측면에서 제기될 수 있는 헌법적 문제제기에 사전적으로 대응하는 차원에서라도 헌법상 기본권 침해여부 심사기준이라 할 수 있는 법률유보 및 과잉금지원칙의 준수를 위한 기본적인 틀을 제시하였다.

셋째, 통계청 보유 자료와 금융기관 등 관련 기관들이 보유하고 있는 행정자료 및 관련 빅데이터를 연계하여 가계부채를 파악할 수 있는 통계 작성 방안을 제시하였다.

우선 통계청에서 보유하고 있는 가구에 속한 가구원 정보를 활용하여 정부부처에서 제공한 소득정보, 민간기관의 부채정보를 활용하여 가구의 총부채원리금상환부담 통계의 개발을 제안하였다. 2015년 12월부터 금융위원회와 은행연합회는 여신심사선진화가이드라인을 시행하고 있으며, 가이드라인의 핵심은 차주의 총 금융부채 상환부담을 평가하는 시스템 즉 DSR(Debt Service Ratio)을 도입하는 것이다. 한편 통계청에서는 가구소득을 추정하는 작업을 진행하고 있는 것으로 알려져 있으며, 특별히 통계청에서는 객관적 소득정보를 가구단위로 추정할 수 있다는 점에서 이는 민간 금융기관이 보유하고 있는 소득정보와 차별적이다. 다시 말해 통계청에서 작성하고 있는 소득정보는 신규로 대출을 이용하고 있는 금융소비자 뿐만 아니라 기존 대출과 더불어, 부채가 없는 가구의 소득정보도 포함할 수 있으며, 이는 기존 금융기관의 대출정보가 개인별 정보라는 것과 대조된다.

따라서 통계청에서 생산 가능한 DSR의 경우 다른 기관들에서 생산 가능한 DSR과 세 가지 차별성이 존재한다. 첫째로 가계의 재무정보에 대한 추정이다. 기존 소득 혹은 부채정보는 개인 즉 차주중심으로 집계되고 있지만, 통계청의 가구정보를 활용하는 경우 가계부채 정보로 전환할 수 있다. 둘째로 기존

가계부채 정보는 부채를 보유한 개인에 대한 정보만 존재하지만, 가계정보를 통하여 부채를 보유하고 있는 가계, 부채를 보유하고 있지 않은 가계, 그리고 새롭게 부채를 보유하는 가계에 대한 정보로 구분할 수 있다. 셋째로 통계청의 경우 객관적 소득정보를 활용할 수 있다.

가계부채 관련 정보의 축적 혹은 보완은 장기적인 호흡으로 준비하는 것이 필요하며, 동시에 가계부채 관련 신규 통계의 개발이 이루어지는 경우 단기적으로 완전한 시장정보 보다는 중장기적으로 완전한 시장정보를 구축하는 방향이 바람직하다. 기존 사례에 대한 분석을 통하여 가계부채 관련 통계를 개발하는 경우 중장기적인 보완과 개선을 전제로 작업을 추진하는 것이 바람직하다는 점에 특별히 유의할 필요가 있다.

넷째, 국내 빅데이터 활용 사례들을 공공 부문과 민간 부문으로 구분하는 동시에 활용되고 있는 자료의 특성에 따라 정리하고, 각 사례별로 당면하고 있는 문제점과 향후 전망에 대한 논의하였다.

현재 국내 공공분야에서는 교통 빅데이터 사용이 가장 두드러지며, 교통분야 빅데이터 관련 법제 정비와 더불어 다양한 데이터 생산주체들의 협업과정은 타 분야의 빅데이터 활용에도 참고할만한 사례이다. 보건·의료분야는 전통적으로 진료기록, 의료보험정보 등의 대용량 데이터를 보유하고 있어 왔으나, 이의 활용에 있어서는 개인정보 보호와 관련하여 다소 보수적이라는 특징을 갖고 있다. 최근 들어 일부 데이터가 공개되거나, 지방자치단체와 함께 쓰이는 등 활용 가능성은 점차 증대되고 있다. 한편 상권분석 서비스는 소상공인 지원정책의 일환으로서 중소기업청, 서울시 등에서 실시하고 있는 빅데이터 서비스이다.

지방자치단체들의 경우 빅데이터 관련 사업을 추진해 왔으며, 현재 관광 관련 사업이 상당수를 차지하고 있다. 서울, 경기도와 같은 수도권 일대의 지자체의 빅데이터 인력, 조직이 상대적으로 더 충실히 갖춰져 있으며 지자체별로 차이가 매우 크다. 서울시는 빅데이터 활용 움직임이 가장 활발한 지방자치단체이다. 주요 사업으로는 올빼미 버스(심야버스) 노선 신설, 노인여가복지시설 입지 선정, 우리 마을 가게 상권 분석 서비스 등이 있다. 서울시의 빅데이터 활용 사례를 통해 빅데이터를 이용한 통계가 정책 수립에 효과적이라는 점을 알 수 있으며, 과거의 공공통계와 달리 정형화된 단위로 공표되기는 힘들다는 것을 알 수 있다. 부산시는 서비스인구통계를 제시함으로써 주간인구 통계를 국내 최초로 생산하고 있다. 부산도시서비스분석시스템은 도시의 실시간 유동인구를 파악함으로써 행정 인구 데이터와 괴리되는 실제 행정 수요를 파악할 수 있다는 점에서 의의가 크다.

민간분야에서는 카드결제 데이터의 이용이 활발하다. 각 민간 신용카드사는 결제 데이터베이스를 이용하여 마케팅 보고서를 작성하거나, 상권분석 보고서를 제공하기도 하며, 신규 카드 상품 개발에 참고하는 경향을 보인다. 이동통신사들의 빅데이터 이용이 활발한데, SKT의 경우 SKT지오비전이라는 자회사를 설립하여 상권분석 서비스를 제공하고 있는 것으로 파악되었다.

다섯째, 해외에서 공공 목적을 위한 빅데이터 활용 사례들과 연구동향을 정리하였다.

미국의 경우, 오픈 데이터(open data)로 요약되는 공개 지향성과 프라이버시 보호를 위한 노력을 두 가지의 명시적 목표로 삼고 있으며, 미국은 DATA.GOV를 통해 다양한 데이터를 적극적으로 공개하는 편이다. RISE Colorado와 같은 교육 서비스, 장차 종합 헬스케어 서비스로 편입될 수 있는 약품 관련 필박스(Pillbox), 범죄에 대한 정보를 구체적으로 거의 실시간으로 시각화하여 볼 수 있는 범죄 지도 서비스 등은 정부, 지자체, 민간의 협동을 통해 구현된 빅데이터 기반의 공공 통계 서비스로서 장차 한국에서도 시도할 만한 것이다.

그 외에도 여러 해외 사례를 통해 공공 목적을 위해 다양한 포맷의 행정 데이터를 통합하거나, 공공 통계의 질을 높이기 위해 빅데이터를 통해 보강, 검증하려는 노력을 확인할 수 있었다. 영국은 국민 식별 번호가 없으므로 가장 포괄적이라고 여겨지는 의료 데이터를 중심으로 한 매칭을 통해 개인에 관련된 정보를 통합하여 행정 자료의 활용도를 높이려는 노력을 하고 있으며, 일본의 경우 빅데이터 활용에 비교적 소극적이나, 채권/채무 리스크 분석 등에 제한적으로 활용하려는 시도를 하고 있다. 올해부터 시행된 마이남바(일종의 국민 식별 번호) 제도가 정착되고, 그것을 중심으로 여러 데이터가 통합되면 행정/빅 데이터의 통합적 활용에 큰 활력소가 될 것으로 보인다. 기타 네덜란드, 아일랜드와 같은 소국 또는 EU같은 국가 연합체에서 시도하는 다양한 빅데이터 관련 노력을 참고하여, 한국이 어떻게 향후에 빅데이터를 활용하여 공공 통계를 확장할 수 있을지에 대한 고민은 계속해야 할 것이다.

여섯째, 국내외 빅데이터 활용 사례들에 대한 분석 결과를 토대로 국내 실정에 적합한 다음과 같은 몇 가지 빅데이터 활용 신규 과제를 제안하였다.

○ 가계부채관련 신규 통계 개발-총부채원리상환부담(Debt Service Ratio)

기존의 가계부채관련 시장정보는 대부분 금융기관의 자료에 기반한 것이어서 가계부채를 과수추정하고 있을 확률이 높다. 가계부채는 가계 입장에서는 부채이지만, 금융기관에게는 자산이기 때문이다. 하지만 가계금융복지조사를

제외하면 가구 단위의 부채정보가 없는 수준이다. 또한, 기존의 설문조사 방식으로는 가계부채의 전모를 파악하기 힘든 것 또한 사실이다. 따라서 기존의 한계를 극복하기 위한 방안으로 소득 및 부채에 대한 새로운 통계 개발이 필요하다.

본 보고서에서는 통계청에서 보유하고 있는 가구 및 가구원 자료에 정부부처(국세청 등)에서 제공하는 소득정보, 그리고 민간기관의 부채정보를 연계한 가구의 총부채원리금상환부담통계 개발을 제안하였다. 이미 통계청에서는 가구소득을 추정하는 작업을 시행 중이며, 개별 금융기관은 개인의 소득 및 부채에 대해 부분적으로만 정보를 가지고 있기 때문이다. 각 주체별로 보유하고 있는 자료의 장단점은 이하와 같다.

한국신용정보원, 개인신용평가사 등 금융기관이 가지고 있는 소득정보는 개인 단위이기 때문에 차주를 중심으로만 소득과 부채를 파악할 수 밖에 없다. 국세청의 소득정보는 좀 더 포괄적이며 실제 소득을 기록하고 있다는 장점이 있다. 이 외에 각종 금융기관의 자료들도 가용한 정보인데, 이는 개인별로 정보가 수집되어 있다.

통계청은 타 기관과 달리 가계의 재무정보에 대한 추정이 가능하다는 장점이 있다. 또한, 기존 가계부채 정보는 부채가 있는 가구에 대한 정보만 수집 가능하다는 한계가 있지만, 통계청의 DSR은 이를 세분하여 무부채가구, 부채 보유 가구, 신규 부채 가구 등으로 파악할 수 있다.

○ 모바일 폰 데이터 랩

모바일 폰 데이터는 현재 가장 널리 쓰이고 있는 데이터로 주로 관광, 상권분석 등의 타 부문과 연계하여 쓰이는 경우가 많다. 이는 모바일 폰 데이터의 인프라적 성격을 보여주는 것으로 모바일 폰 데이터는 빅데이터 중 하나가 아니라, 빅데이터 활용의 가장 기초적인 자료로서 인식되어야 한다. 이러한 모바일 폰 데이터의 중요성에도 불구하고, 국내에서 아직 데이터의 생산과정 및 품질관리 부문은 부각되지 못하고 있는 것이 현실이다. 데이터의 특성상 프라이버시 보호가 필수적이며, 전면적 공개가 어려운 현실에서 모바일 폰 데이터를 통계청이라는 제한된 공간에서 서로 연계해 활용할 수 있다면 이는 매우 현실적이면서 효율적인 빅데이터 활용 방안이 될 수 있다.

특히 모바일 폰 데이터의 핵심이 연계를 통한 높은 활용 가능성에 있는 만큼 이를 극대화하기 위해서라도 타 데이터들과의 연계 장소는 반드시 필요하다. 통계청은 이러한 중간 연계자 역할을 하기에 매우 적합한 기관이다. 통계 생산기관으로서의 전문성과, 정부기관으로서의 신뢰성, 중립성을 강화한다면 향후

통계 생산에 있어서 중심적인 역할을 수행할 수 있다. 그 뿐 아니라 모바일 데이터를 이용한 통계 개발을 국제적으로도 선도한다는 위상 역시 얻을 수 있을 것으로 기대된다.

○ 체감경기통계

통계청에서는 이미 생산/소비/투자/경기 동향을 종합하여 산업활동동향을 매월 발표하고 있으며, 지역경제의 생산/소비/고용/물가 등을 포괄하는 지역경제동향 역시 작성하고 있다. 이러한 경제통계는 광범위한 정보를 종합적으로 전달해준다는 장점이 있으나, 자료들이 추상화된 만큼 국민 개개인에게는 체감이 어렵다.

또한 자료 생산에 오랜 시간이 걸리며, 즉각적인 경기 동향 변화를 파악하기 힘들다. 따라서 신용카드 결제 데이터를 활용하여 국민의 체감경기에 부합하는 데이터 생산이 필요하다. 이를 통하여 경제통계에 대한 관심을 제고하고, 통계 활용을 촉진시킬 수 있을 것으로 기대된다. 또한 지역경기 활성화에 필요한 기초 자료로서 정책수립 및 효과 측정에 유용한 도구가 될 수 있을 것이다.

○ '세이프 코리아 지수(Safe Korea Index)' 서비스

최근 한국에서는 범죄와 관련된 일반의 경계심이 강화되고 있으나, 현재 국내에는 세부 범죄, 세부 지역에 대해 거의 실시간으로 파악할 수 있는, 몇 국가에서 서비스되고 있는 것과 같은 범죄 전문지도 서비스가 없다. 생활안전지도는 범죄 외의 정보를 함께 제공하게 되는데, 서울 기준으로 구 수준에서, 임의로 5등급으로 재정리한 내용을 보여주는 수준을 고려해 볼 수 있을 것이다.

현재 부분적으로 제공되는 범죄 관련 데이터와, KICS와 같은 범죄 관련 원천 데이터, 지리 정보를 결합하면, 상당히 구체적인 수준에서 특정 지역, 특정 범죄에 대한 현황을 파악할 수 있으며, 역으로 특정 지역에서, 특정 범죄에 대해 어느 정도 안전한지를 수치로 표현할 수 있을 것이다. 이러한 포괄적 서비스가 가능하다면 국민의 행복도 증진과 과학적 범죄 수사 및 예측에 큰 도움이 될 것으로 보인다.

○ '빅데이터 통계 활용 네트워크' 구축

통계청 중심의 빅데이터 활용 네트워크 구축은 또 하나의 중요한 과제로 보인다. 누구든지 정부가 확보해 둔 행정 데이터 외에 빅데이터의 수집을 정부에 요청할 수 있다면, 그러한 요청 자체가 일종의 빅데이터로 기능한다. 통계청과 같은 중앙 정부 기관이 빅데이터 활용에 대한 제안을 상시 접수하고, 좋은 제안을 바탕으로 추가 빅데이터를 수집하여 국민이 활용할 수 있도록 제공하는 것이다. 이러한 작업을 상시 업무로 수행한다면 어떠한 빅데이터를 공공 통계를 확장할 때 활용해야 하는지 역시 지속적으로 파악 가능할 것이다.