

빅데이터의 경제통계 활용 현황 및 시사점

정보화 시대에 빅데이터에 대한 경제적·사회적 관심이 높아지고 통계 환경이 변화됨에 따라 빅데이터의 공식통계 활용 방안에 관한 논의가 국제기구 및 주요국 통계기관을 중심으로 시작되었다. 현재 한국은행에서도 주요 경제통계 작성에 빅데이터를 활용하는 방안을 모색하고 있다. 이와 관련하여 빅데이터가 경제통계에 활용된 해외 사례를 살펴보고 빅데이터의 통계적 활용 가능성을 평가해보고자 한다.

빅데이터의 활용 사례를 살펴본 결과 스캐너 데이터, 웹 스크래핑 데이터, 지급결제 데이터 등 비교적 정형화된 빅데이터가 물가, 소비지출 등 경제통계 작성에 부분적으로 이용되고 있다. 소셜미디어, 뉴스 등 텍스트 데이터와 인터넷 검색 데이터도 경제 및 금융 관련 지표 분석에 널리 활용되고 있다. 이러한 사례들은 빅데이터의 특성에 기인한 기초자료로서의 한계에도 불구하고 향후 빅데이터가 경제통계 작성 및 분석에 유용하게 활용될 가능성이 있음을 시사한다.

따라서 빅데이터가 쉽게 적용 가능한 경제통계 영역을 발굴하여 시험편제를 실시해보고, 기존 활용 사례들을 벤치마킹하여 유용성을 검증할 필요가 있다. 또한 빅데이터의 정제·처리·분석을 위한 전문적 지식을 습득하고 빅데이터에 적합한 통계 작성기법에 대한 장기적인 연구도 필요하다.

I. 검토배경

II. 빅데이터의 경제통계 활용 현황

1. 물가지표
2. 심리지표
3. GDP 관련 지표
4. 기타

III. 시사점 및 향후 과제

I. 검토배경

최근 인터넷 및 모바일 기기의 확산으로 빅데이터(big data)에 대한 사회적 관심이 높다. 빅데이터는 데이터의 규모(volume), 생성속도(velocity), 형태(variety) 측면에서 기존의 데이터와 상당한 차이가 있다. 따라서 빅데이터를 처리·분석할 수 있는 관련 기술 개발, 의미 있는 정보 추출 및 새로운 가치 창출에 대한 기대가 확산되고 있다.

이처럼 빅데이터의 출현으로 “데이터”에 대한 중요성과 기대가 높아진 반면 기존 통계 작성 환경은 점점 열악해지고 있다. 1인 가구, 맞벌이 가구 등의 증가로 가구 대상 서베이 조사의 애로사항이 많아지고, 정보보호 요구 및 업무상 비협조 등으로 사업체 대상 조사의 응답거부율도 높아지고 있다. 반면 디지털 경제 하에 신규로 포착해야 할 통계의 수요는 커지고 있다. 이러한 한계를 극복하기 위해 통계 작성기관들이 행정자료, 온라인 자료 등 새로운 형태의 데이터에 주목하게 되었다.

빅데이터의 공식통계 활용 방안에 대한 본격적인 논의는 UN, Eurostat, OECD 등 국제기구를 중심으로 진행되어 왔다. UN 통계위원회(UN Statistical Commission)는 2013년 제45차 회의에서 빅데이터를 정식의제로 다루었고 글로벌 워킹그룹(Global Working Group, GWG)을 구성¹⁾하여 빅데이터 프로젝트 등 관련 활동을 지원하기 시작하였다. 한편 UN 유럽경제위원회(UN Economic Commission for Europe, UNECE)는 고위 전문가 그룹(High-Level Group for the Moderation of statistical production and services, HLG)이 주축이 되어 빅데이터의 개념 및 과제 등에 관한 가이드라인을 제시하고 국제협력 프로젝트에 참여²⁾하고 있다. Eurostat, OECD 등도 주요국의 통계 작성기관과 함께 빅데이터 활용을 위한 국제적 논의와 프로젝트에 동참하고 있다.

이와 같이 빅데이터의 통계적 활용에 대한 관심이 높아짐에 따라 한국은행에서도 주요 경제통계 작성에 빅데이터를 활용하는 방안을 모색하고 있다. 빅데이터는 빠르게 변화하는 정보화 사회를 잘 반영하는 유용한 자료이지만 전통적 방식의 통계작성에 실제로 활용되려면 극복해야 할 한계와 과제들이 많다. 따라서 빅데이터의 경제통계 활용 사례 연구를 통해 빅데이터의 통계적 활용 가능성과 잠재가치를 점검해 볼 필요가 있다. II장에서는 빅데이터가 경제통계 작성 및 분석에 활용된 구체적 사례를, III장에서는 빅데이터의 통계적 활용 가치와 시사점을 살펴보고자 한다.

1) <http://unstats.un.org/bigdata>

2) UNECE Big Data Inventory 참고

II . 빅데이터의 경제통계 활용 현황

빅데이터는 자료의 형태에 따라 일정한 규칙을 갖고 체계적으로 정리된 수치형태의 정형적(structured) 데이터와 텍스트, 사진, 동영상 등의 비정형적(unstructured) 데이터로 구분된다. 정형적 데이터는 그 자체로 의미 해석이 가능하고 통계로 활용하기 용이하나 비정형적 데이터는 데이터의 구조가 복잡하여 전문적인 처리 및 분석 과정을 거친 후에 의미 있는 통계로 활용 가능하다.

UNECE는 <표 1>과 같이 빅데이터를 정보소스의 유형에 따라 사회관계망(social networks), 거래내역(traditional business systems) 및 사물인터넷(internet of things) 자료로 구분하였다. 이 중 공공기관이나 기업의 사업과정을 매개로 하는(process-mediated) 거래내역 자료와 센서, 컴퓨터 등 기계장치에 의해 생성되는(machine-generated) 사물인터넷 자료는 정형적 데이터에 가깝다. 반면 사회관계망 자료는 인간의 사회활동 과정에 생성되는 정보(human-sourced information)로 주로 비정형적인 형태를 가진다. 이하에서는 이러한 다양한 유형의 빅데이터가 주요 경제통계에 활용된 사례를 살펴보도록 한다.

<표 1>

빅데이터의 분류

사회관계망 자료	거래내역 자료	사물인터넷 자료
<ul style="list-style-type: none"> - 소셜미디어(페이스북, 트위터 등) - 블로그, 코멘트 - 개인문서 - 사진 - 동영상(유튜브 등) - 인터넷 검색 - 모바일 데이터 - 사용자 생성 지도 - 이메일 	<ul style="list-style-type: none"> • 공공기관 생성자료 <ul style="list-style-type: none"> - 의료기록 • 기업 생성자료 <ul style="list-style-type: none"> - 상업적 거래 - 은행/증권 기록 - 전자상거래 - 신용카드 	<ul style="list-style-type: none"> • 센서 데이터 (고정센서) <ul style="list-style-type: none"> - 홈자동화 - 기후/오염 센서 - 교통센서/웹캠 - 과학센서 - 보안/감시용 비디오/사진 (이동센서) <ul style="list-style-type: none"> - 휴대폰 위치 - 자동차 - 위성사진 • 컴퓨터 시스템 데이터 <ul style="list-style-type: none"> - 로그 - 웹로그

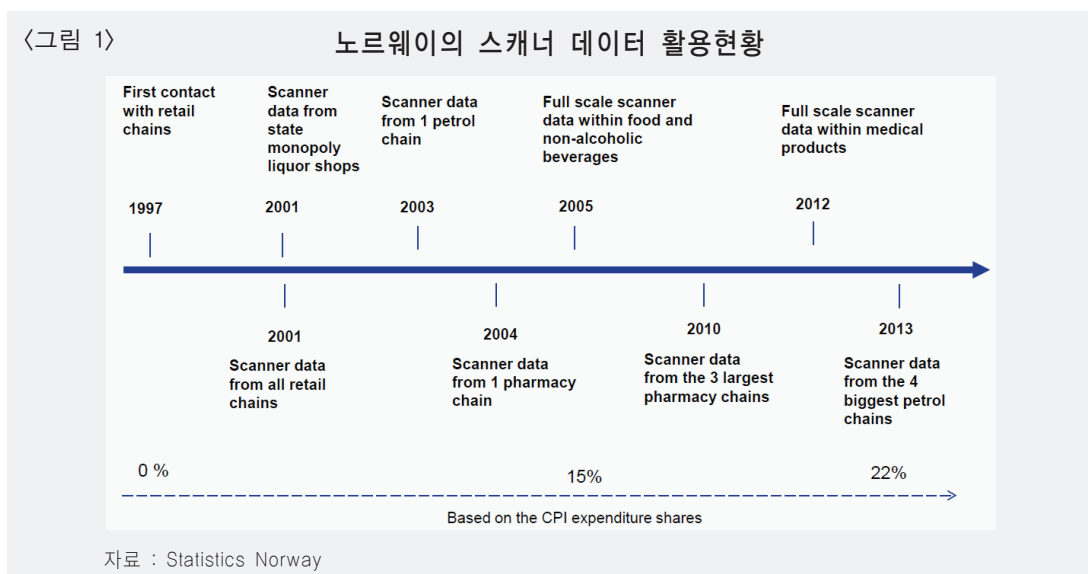
출처 : UNECE (2013) : Classification of Types of Big Data

1. 물가지표

소비자, 생산자, 수출입 물가지수 등 대표적인 공식 물가지수는 주로 거래 규모가 큰 대표 품목을 표본조사하여 작성한다. 그러나 품목의 종류 및 거래 채널이 다양해지고 제품의 생성주기 및 수명도 단축됨에 따라 서베이 기반 물가지수를 보완하는 데 빅데이터를 이용하는 방법들이 제안되고 있다.

가. 스캐너 데이터(scanner data)

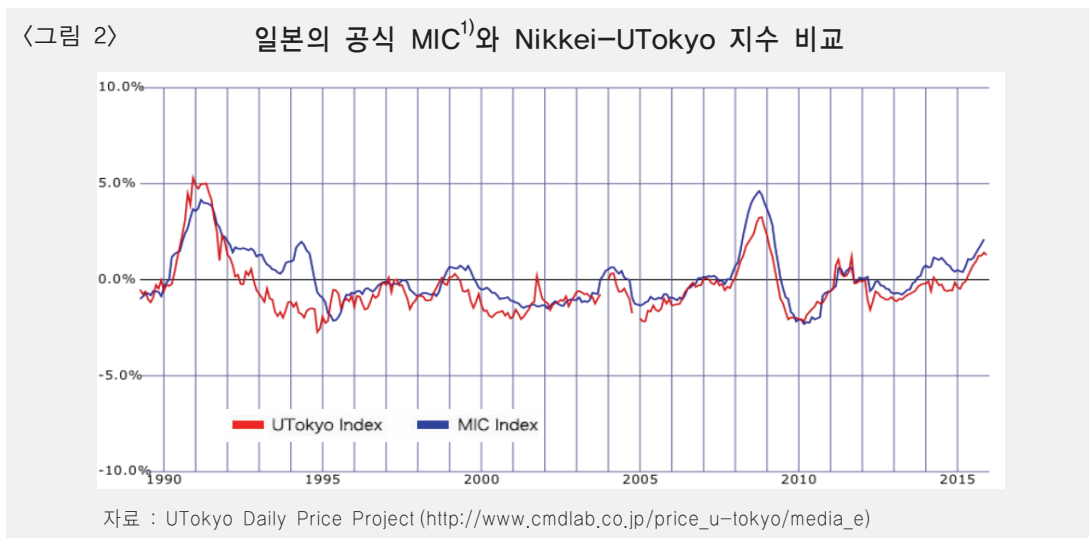
스캐너 데이터는 제품에 부착된 바코드(bar code)가 판매시점에 스캐너에 의해 읽혀지면서 입력되는 데이터로서 POS(Point of Sale) 데이터라고도 한다. 최근 소규모 상점에도 판매 시스템이 구비됨에 따라 스캐너 데이터에 포착되는 품목의 범위가 규모가 확대되고 있다. 스캐너 데이터는 실시간으로 수집되는 실제 거래내역 자료이므로 속보성과 신뢰성이 높고, 판매 가격뿐만 아니라 수량 및 기타 세부 정보도 확보 가능하다. 이러한 장점으로 노르웨이, 네덜란드, 스위스 등 유럽 국가들은 일찍부터 소비자물가지수 편제에 스캐너 데이터를 사용³⁾하고 있다. <그림 1>에서 보는 바와 같이 노르웨이 통계청은 이미 2000년대 초반부터 스캐너 데이터를 활용하기 시작해서 적용 품목을 점점 확대하고 있다.



3) 이들 국가의 통계청에서 스캐너 데이터를 소비자물가지수 편제에 활용하는 구체적인 방법은 Rodriguez and Haraldsen (2006), Randi (2016), Müller(2010), Van der Grient and Hann (2010) 등을 참고하기 바란다.

한편 Eurostat은 통합소비자물가지수(Harmonized Index of Consumer Prices, HICP)⁴⁾의 정확성을 유지하기 위하여 스캐너 데이터의 입수, 처리, 이용 등 일련과정에 관한 가이드라인을 제시하고 회원국에도 스캐너 자료의 활용을 권고하고 있다.

도쿄 대학의 Tsutomu Watanabe 교수는 Nikkei Inc.의 일별 POS 데이터를 이용하여 일본의 Nikkei-UTokyo 일별지수를 개발⁵⁾하였다. <그림 2>에서 2015년까지 Nikkei-UTokyo 지수⁶⁾의 움직임을 보면 일본 총무성(Ministry of Internal Affairs and Communications)의 공식 식료품물가지수(MIC Index for grocery component)와 비슷한 추이를 보이고 있다.



이외에도 Watanabe 교수는 전체 일본인의 40% 이상이 보유하고 있는 T 포인트 카드의 제휴사(슈퍼마켓, 약국, 편의점, 외식 체인 등)를 통해 수집되는 구매가격 데이터를 이용하여 T-point Index(TPI)를 개발하였다. TPI는 식품, 생활용품 이외에도 패션, 주거, 영상, 음악 등 소비의 다양한 측면을 포괄할 수 있고, 구매자의 속성을 파악하여 성별·연령별 물가지수 산출이 가능한 장점⁷⁾이 있다.

한편 관세청의 수출입 통관자료에도 수출입되는 모든 제품들의 가격과 수량 정보가 기록되므로 스캐너 데이터의 일종으로 볼 수 있다. 현재 일부 품목의 통관자료가 수출입 물가지수 편제에 활용되고 있으나 그 활용범위를 확대하기 위해서는 통관자료의 동일 HS 코

4) EU 회원국들의 소비자물가지수를 가중평균하여 산출한 물가지수이다.

5) Watanabe and Watanabe (2014) 참고

6) 2016년 1월부터는 Nikkei CPINow로 명칭을 변경하여 Nikkei.Inc와 Nowcast Inc.가 공동으로 제공한다.

7) TPI의 대상 품목 수는 총무성 CPI 품목의 약 23.4%에 해당한다.(www.cccmk.co.jp/tpi/)

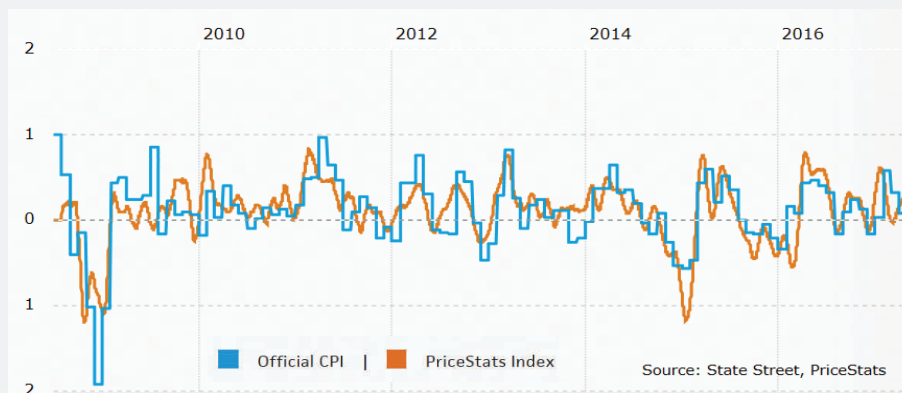
드내에 혼재되어 있는 다양한 품목들에 대한 정확한 품질 식별 방안을 강구할 필요가 있다. 한국은행은 외부연구용역 사업(강규호 외 (2015), 장영재 외 (2017) 참고)을 통해 수출입 물가지수 작성 관련 연구를 실시하였으며 향후 통관자료의 정제·처리 및 활용 방법을 구체적인 모색할 필요가 있다.

나. 웹 스크래핑 데이터(web scraping data)

MIT의 BPP(Billion Price Project, Cavallo and Rigobon (2016) 참고)는 온라인 가격정보를 이용하여 물가지수를 개발한 대표적인 사례이다. 동 방식에서는 자동화된 웹 스크래핑 소프트웨어가 온라인 소매업체 웹의 HTML 코드를 분석하여 수많은 상품의 가격정보를 수집한다. 이러한 온라인 가격자료는 서베이 자료보다 저렴하게 수집할 수 있고, 상품의 브랜드, 사이즈 등에 관한 부가 정보 및 신제품에 대한 정보를 신속히 입수할 수 있는 장점이 있다. 또한 일별 물가지수로 산출 가능하므로 시의성 높은 정보를 제공할 수 있다. 반면 온라인에 존재하지 않는 품목의 가격정보를 수집할 수 없고 온라인 가격과 오프라인 가격 차이 또는 실거래 가격과의 괴리가 발생할 수 있는 단점이 있다.

현재 BPP는 약 70개국 1,000개 이상의 온라인 소매업체로부터 수집한 가격정보를 토대로 22개국⁸⁾의 일별 물가지수를 PriceStats을 통해 공표하고 있다. <그림 3>에서 미국의 BPP 지수를 보면 공식 CPI에 약간 선행하는 움직임을 보이고 있다.

<그림 3> 미국의 공식 CPI와 BPP 지수 비교¹⁾



자료 : PriceStats (www.pricestats.com)

주 : 1) food & beverages; furnishing & household products; recreation & culture; clothing & footwear; housing, electricity & fuel; health 부분의 합성지수

8) 아르헨티나, 호주, 브라질, 캐나다, 칠레, 중국, 콜롬비아, 프랑스, 독일, 그리스, 아일랜드, 이탈리아, 일본, 한국, 네덜란드, 러시아, 남아프리카공화국, 스페인, 터키, 영국, 미국, 우루과이 등

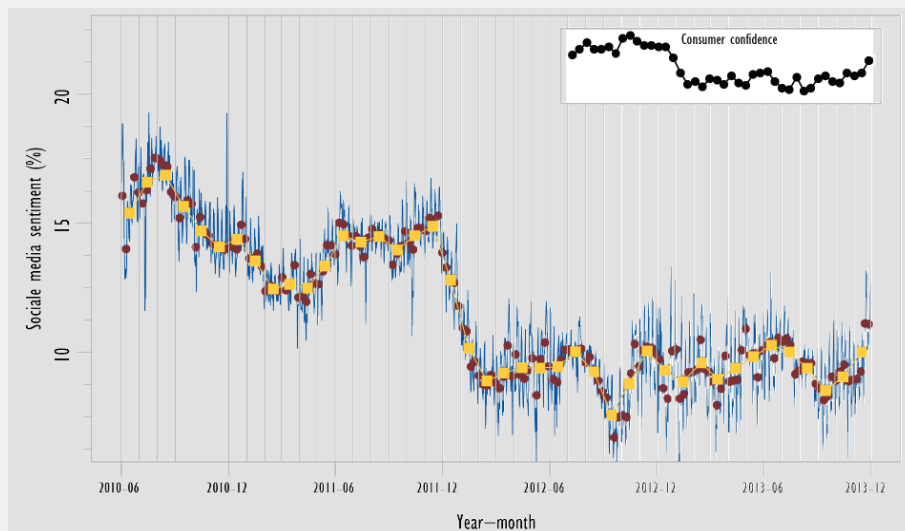
료가 경제주체들의 심리를 추출하거나 경제 및 금융 상황 모니터링을 위한 새로운 정보소스로 활용되고 있다.

가. 소셜미디어 데이터

네덜란드 통계청에서는 페이스북, 트위터 등 소셜미디어 메시지에 나타난 감성(sentiment)을 이용하여 소셜미디어지수(social media index, SMI)를 작성하였다. 네덜란드의 소셜미디어 모니터링 전문업체 Coosto는 각종 소셜미디어에 나타난 메시지들을 수집하여 문장단위로 분석한 후 해당 메시지의 전반적인 의견을 긍정, 중립, 부정으로 분류¹⁰⁾하는 서비스를 제공한다. 네덜란드 통계청은 Coosto에서 구입한 소셜미디어 메시지 감성분류 결과의 긍정의견 비중에서 부정의견 비중을 차감하여 SMI를 산출하였다. <그림 5>에서 SMI는 네덜란드의 공식 심리지표인 소비자신뢰지수(consumer confidence index, CCI)와 유사한 움직임을 보이며 0.78의 상관관계를 가진 것으로 나타났다. 자료 취득원별로 보면 페이스북과 트위터의 SMI가 언론기사, 블로그 등의 SMI보다 CCI와 더 유의미한 관계를 보이는 것으로 나타났다. 보다 자세한 연구결과는 Daas and Puts (2014), Daas et al. (2015), Van den Brakel et al. (2016) 등을 참고하기 바란다.

<그림 5>

네덜란드의 소셜미디어지수와 소비자신뢰지수 비교



자료 : Daas and Puts (2014)

10) 예를 들어 “Despite the high unemployment, the economy is doing well.”는 부정적 측면(high unemployment)도 포함하고 있으나 전반적인 경기인식은 긍정적(the economy is doing well)이므로 긍정 의견으로 분류한다.

네덜란드의 SMI를 벤치마킹하여 한국은행이 우리나라의 소셜미디어 데이터를 이용한 경제심리지표를 내부적으로 시산해 본 결과 공식 심리지표인 한국은행의 소비자심리지수와 유사한 움직임을 보이고 대표 실물경제지표인 GDP에 선행하는 것으로 나타났다. 이와 같이 소셜미디어 데이터를 이용한 심리지표는 속보성 및 시의성이 높은 심리 정보를 제공하고, 경제주체의 판단, 의도, 기대 등의 심리뿐만 아니라 그 원인에 대한 정보도 제공 가능하여 기존 경제심리지표의 보완지표 또는 정책결정을 위한 속보성 있는 모니터링 지표로 활용 가능하다. 그러나 소셜미디어를 적극적으로 활용하는 특정 집단의 의견이 더 많이 반영될 우려가 있고, 텍스트 마이닝(text mining)¹¹⁾의 기술적 제약으로 인해 경제주체의 심리를 완벽하게 추출해 내기 어려운 한계도 있다.

나. 뉴스 및 신문기사

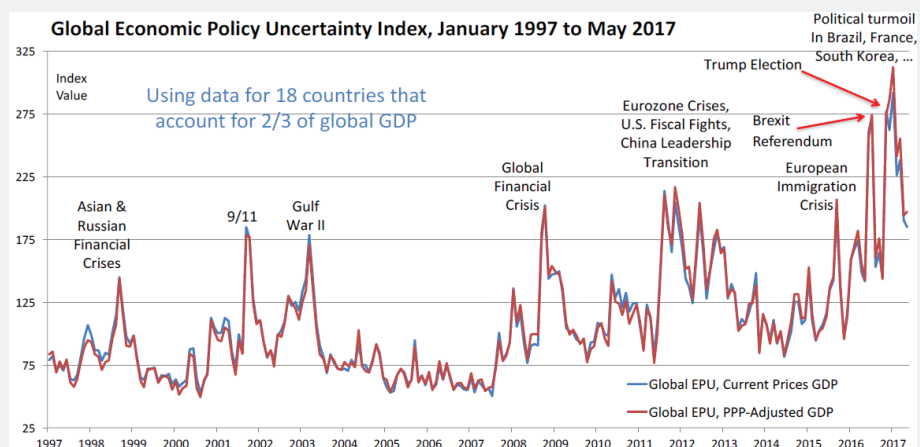
Baker et al. (2015)는 주요 신문사의 기사를 이용하여 경제정책 불확실성 지수(economic policy uncertainty index, EPU)를 개발하였다. 구체적으로 경제(economic 또는 economy), 불확실(uncertain 또는 uncertainty), 정책관련 용어(congress, deficit, Federal Reserve, legislation, regulation, 또는 White House) 등 세 가지 범주의 단어를 모두 포함하는 기사의 빈도수를 전체 기사의 개수로 나누어 EPU 지수를 산출하였다. 현재 미국 등 18개국의 개별 EPU 지수와 이들을 합성한 글로벌 EPU 지수¹²⁾가 웹사이트에 공개되고 있다. <그림 6>을 보면 2016년말 한국의 정치적 불안이 글로벌 EPU 지수에 반영된 것을 확인할 수 있다. 이와 같이 EPU 지수는 정보의 신뢰성, 정확성, 일관성 측면에서 잠재적 한계가 있으나 경제정책 관련 동향을 신속하게 포착하고 국제적 비교가 용이한 장점이 있다.

11) 텍스트 데이터를 계량화하여 분석하기 위해 사용되는 통계적 방법들과 정보처리 방법을 통칭한다.

12) 1997년~2015년 기간을 표준화구간으로 하여 각국의 EPU 지수를 표준화한 다음 IMF의 World Economic Outlook의 GDP를 가중치로 하여 합성지수를 산출한다.

〈그림 6〉

글로벌 경제정책 불확실성 지수



자료 : www.policyuncertainty.com

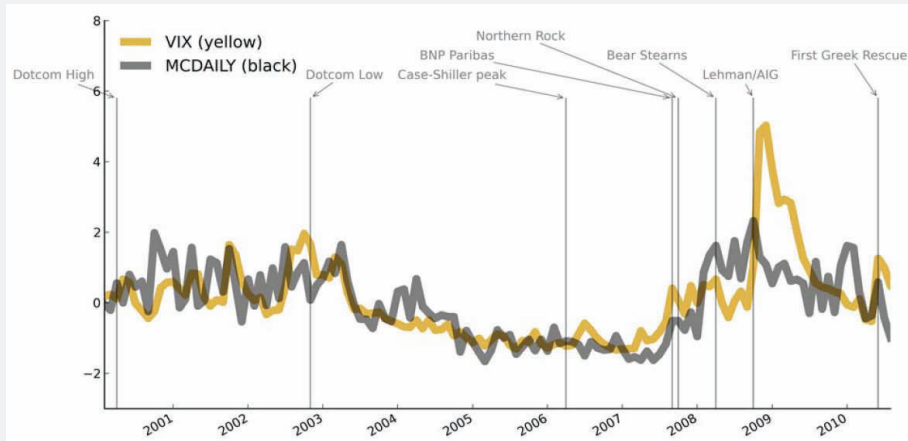
한편 샌프란시스코 연방준비은행의 Economic Letter(2017.4월)¹³⁾는 경제에 관한 뉴스 기사나 사실이 기사나 인터뷰 응답자들에 대한 감정을 전달하여 공식 자료의 단순제공 이상의 역할을 한다고 평가하고 경제·금융 관련 뉴스 기사를 이용하여 심리지표를 작성한 Shapiro et al. (2017)의 연구를 소개하였다. 동 심리지표는 현재의 경제상황을 잘 반영하는 것으로 나타났고, 미래 경제상황 예측력 측면에서 미시간 대학의 소비자심리지수(consumer sentiment index), Conference Board의 소비자신뢰지수(consumer confidence index) 등 전통적 서베이 방식의 심리지표보다 우월한 것으로 평가되었다.

영란은행(Bank of England)은 금융시장 심리지표를 산출하는 공동연구를 런던대학교와 함께 실시하였다. Nyman et al. (2016)은 영란은행 내부 보고서(internal market commentary), 브로커 보고서(broker reports), 로이터통신 뉴스(Reuters news archive) 등의 텍스트를 분석하여 금융시장 관련 심리를 기대(excitement)와 불안(anxiety)의 두 가지 범주로 구분하였다. <그림 7>은 영란은행 내부 보고서에서 추출한 불안지수(MCDAILY)가 금융시장이 불안했던 시기에 높았음을 보여준다. 미국 월가의 ‘공포지수’로 불리는 시카고 옵션거래소(CBOE)의 변동성 지수(Volatility Index, VIX)와는 0.65의 상관관계를 보이는 것으로 나타났다.

13) What's in the News? A New Economic Indicator.

〈그림 7〉

금융시장 불안지수(MCDAILY)와 VIX 비교



자료 : 영란은행, 시카고옵션거래소

이외에도 금융 및 경제 뉴스에 나타난 텍스트를 대상으로 투자자들의 심리를 분석하여 기업의 주가를 예측하거나, 기업의 부실위험을 예측하는 등 뉴스 정보를 이용한 다양한 연구 사례가 많다. 이는 뉴스에서 추출한 심리가 경제 및 금융 상황에 대한 유의미한 정보를 제공할 수 있음을 시사한다.

3. GDP 관련 지표

물가지표 편제에 대표적으로 활용되고 있는 스캐너 데이터는 품목별 소매동향 등 소비 지출 관련 정보로도 활용 가능하다. 실제로 미국 경제분석국(BEA)은 가계소비지출 통계 작성에 스캐너 데이터를 참고한다. 한국은행은 현재 신한카드와의 공동연구를 통해 가계소비 지출, 서비스업 생산 등 국민소득 구성항목 가운데 신용카드 빅데이터로 추정할 수 있는 항목을 발굴해 GDP 추계에 활용하는 방안을 모색하고 있다. Galbraith and Tkacz (2015)는 신용카드뿐만 아니라 직불카드, 수표 거래 등을 통해 발생하는 전자결제(electronic payment) 빅데이터를 이용하여 캐나다의 GDP를 추정(nowcast) 하였다.

한편 국제금융 및 외환거래를 위한 국제금융통신망 SWIFT(Society for Worldwide Interbank Financial Telecommunication)에도 결제 관련 빅데이터가 축적된다. SWIFT 전문 양식(FIN message)에 금융기관간 거래정보 확인, 지급지시, 이체 확인 등 각 메시지 종류에 따라 거래유형, 결제일, 계좌정보 등의 다양한 로그정보가 기록된다. SWIFT는 고객자금 이체와 관련된 메시지(MI 1031⁴⁾) 중 실제 경제활동(real economic activity)과 관련된 메시지들만 추출,

월별 규모를 측정하여 SWIFT Index를 작성¹⁵⁾한다. <표 2>는 SWIFT 통신 메시지 분석 빅데이터를 이용한 GDP 추정(nowcast) 결과인데 공식 GDP 성장률과 다소 차이가 있다. 그러나 SWIFT 빅데이터는 실제 금융거래에 기반한 속도성 있는 경제 모니터링 지표로 활용 가능해 보인다.

<표 2> SWIFT 빅데이터를 이용한 GDP 성장률 추정

Region/ Country	Q3-2013 vs. Q3-2012 (Year-on-Year %)	Q4-2013 vs. Q4-2012 (Year-on-Year %)	Q1-2014 vs. Q1-2013 (Year-on-Year %)	Forecast Q1-2014 Trend	
	GDP Actual ⁽¹⁾ (published by OECD)	GDP Nowcast	GDP Forecast	Direction ⁽²⁾	Rate of change ⁽³⁾
OECD	1.1%	1.4%	1.5%	Growing	Faster
EU27	0.1%	0.7%	0.9%	Growing	Faster
US	1.6%	2.1%	2.2%	Growing	Faster
UK	1.5%	2.3%	2.6%	Growing	Faster
Germany	0.6%	1.4%	1.6%	Growing	Faster

자료 : OECD(2013.11.12일), "SWIFT Index anticipates a strong start to 2014 for the UK and US economies."

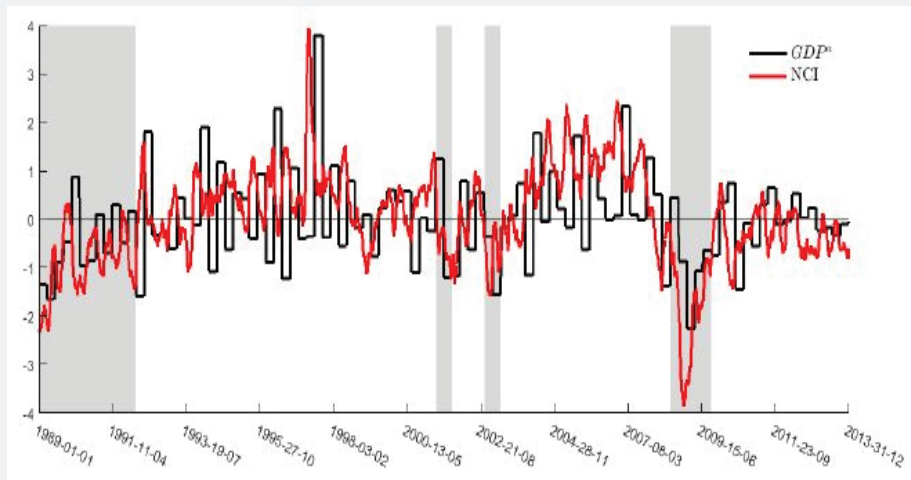
이외에도 신문기사로부터 추출한 텍스트 정보도 경기순환지수 및 분기 GDP 성장률을 추정하는 데 이용 가능하다. 예를 들어 Thorsrud는 노르웨이의 주요 일간 경제신문인 Dagen Naringsliv(DN)의 뉴스 기사를 Latent Dirichlet Allocation(LDA) 모형을 이용하여 기사에 나타나는 단어들을 바탕으로 신문 기사를 여러 주제(topic)들로 분해하고 각 주제가 기사에 언급된 빈도수를 계산하였다. 그리고 하버드 IV-4 심리사전에 정의된 긍정/부정 단어 목록을 이용하여 기사에서 각 주제에 대해 사용된 단어들을 긍정단어와 부정단어로 분류하고 그 차이를 계산하였다. 이러한 과정을 통해 신문기사의 텍스트 자료는 최종적으로 어떤 주제에 대한 기사가 많았는지, 각 주제에 대한 기사의 전반적인 분위기는 어떠한지가 반영된 시계열 자료로 변환되었다.

14) SWIFT는 전문 양식마다 "MT+세 자리 숫자" 형태의 명칭을 부여하는데 이중 MT 103 메시지는 고객자금 이체에 이용되는 전문이다.

15) 현재 200개 이상의 국가 11,000여개 금융기관이 SWIFT의 서비스를 이용하고 있으며 SWIFT Index는 매일 작성되어 SWIFT 기관회원에게 제공되고 있다.

Thorsrud (2016a)는 이와 같이 시계열 자료로 표현된 텍스트 정보를 동적인자모형(Dynamic Factor Model, DFM)에 적용하여 일별 경기동행지수(Newly Coincident Index of Business Cycles, NCI)를 추정하였다. 그 결과 NCI는 일반적으로 사용되는 경기순환지수 대비 정확성과 시의성이 우수하고 특히 2000년대 초반의 경기침체를 주식이나 채권과 같은 시장지표들보다 더 잘 예측한 것으로 나타났다.

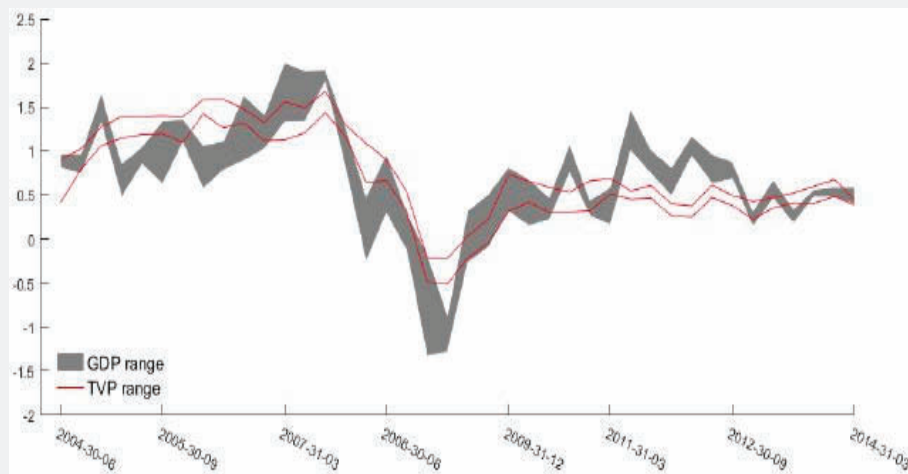
<그림 8> NCI와 GDP 비교¹⁾



자료 : Thorsrud (2016a)
 주 : 1) 회색 음영은 경기침체기

한편 Thorsrud (2016b)는 신문기사에서 추출한 텍스트 정보 시계열과 추세제거와 표준화를 거친 보정된 실시간 GDP 성장률을 DFM, 시변모수(Time-Varying Parameter, TVP) 리스케일링(rescaling) 모형을 이용하여 분기 GDP 성장률을 추정(nowcast)하였다. <그림 9>에서 공식 GDP 성장률(속보치 및 잠정치)의 범위(최대치 및 최소치)와 TVP 리스케일링 모형을 통한 GDP 성장률 예측치 범위를 비교한 결과를 보면 신문기사를 활용한 모형이 전반적으로 경제상황을 잘 예측한 것으로 나타났다. 또한 동 모형을 통한 GDP 예측결과는 노르웨이 중앙은행의 나우캐스팅 시스템이나 여타 벤치마크 모형들과 비슷한 수준의 예측오차를 가지나 경기 전환점 포착 측면에서 더 우수한 것으로 평가되었다.

<그림 9> GDP 성장률과 TVP 모형의 GDP 성장률 예측결과 비교



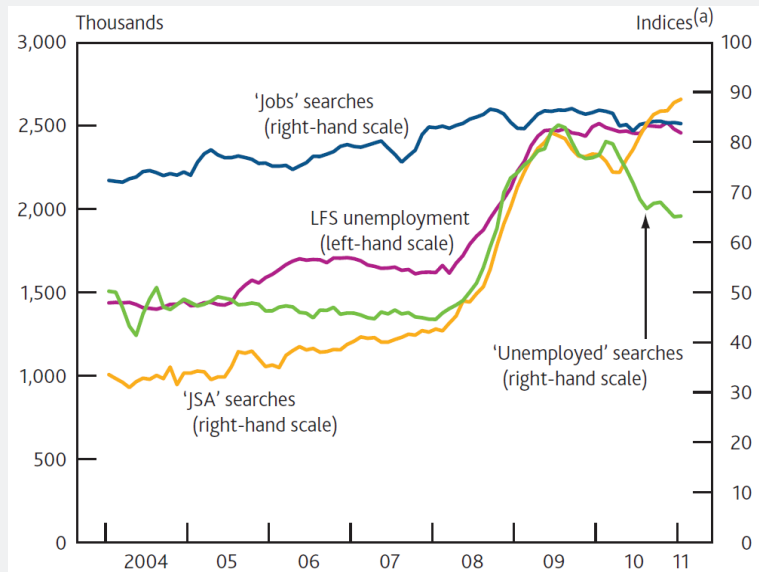
자료 : Thorsrud (2016b)

4. 기타

마지막으로 소개할 인터넷 검색 관련 빅데이터도 고용 시장 연구에 많이 활용되어 있다. Google의 실업 관련 검색 데이터를 이용하여 Choi and Varian (2009)은 미국의 신규 실업수당 청구건수를, Askitas and Zimmerman (2009)과 D'Amuri (2009)는 각각 독일과 이탈리아의 실업률을 예측하였다. <그림 10>은 영란은행의 Quality Bulletin에 게재된 영국 통계청의 노동 서베이(Labor Force Survey, LFS)에 의한 실업자 수(좌축)와 실업 관련 인터넷 검색 데이터 지수(우축)의 비교 자료이다. McLaren and Shanbhogue (2011)는 이러한 실업 관련 검색 빅데이터가 영국의 실업자 수 변화를 예측하는 데 유용하다고 평가하였다. 한국은행에서도 외부연구용역 사업(이궁희 외 (2016))을 통해 네이버 고용 검색지표를 개발하여 고용 관련 공식통계와 비교분석해 보았다.

〈그림 10〉

실업 관련 서베이 데이터와 인터넷 검색 데이터 비교



자료 : Google, ONS, Bank of England

주 : (a) 주별 검색 데이터를 표준화한 지수를 평균하여 산출한 월별 지수

이외에도 인터넷 검색 데이터는 주택가격 등 부동산 동향을 예측하는 데에 활용되기도 하였다. 이와 같이 Google Trends¹⁶⁾ 등을 통한 검색 데이터는 특정 관심 주제에 대한 시간·지역별 추이 관찰이 용이하고, 고정된 질문을 사용하는 서베이에 비해 특정 사건 전후의 변화를 신속·유연하게 분석할 수 있는 장점이 있다. 그러나 검색 데이터도 소셜미디어 데이터와 같이 모집단 대표성이 부족하고 검색 단어의 선택에 민감하며 잡음(noise)이 많다는 단점이 있으므로 의미 있는 정보를 추출하기 위한 정교한 작업이 필요하다.

16) www.google.co.kr/trends/

III. 시사점 및 향후 과제

빅데이터는 디지털 경제 하에서 새로운 경제현상을 포착하고 변화하는 통계환경에 대응하기 위한 수단으로 제안되고 있다. 무작위로 수집된 빅데이터를 사후적으로 의미 있는 통계로 가공하는 방식은 사전적 설계와 조사과정을 거쳐야 하는 전통적 서베이 방식에 비해 저렴하고 응답부담이 없으며 다양한 부가정보 확보가 가능한 장점이 있다. 또한 빅데이터는 실시간으로 생성·축적되므로 자동화된 데이터 처리과정을 통해 시의성 있는 정보를 신속하게 제공할 수 있다.

반면 빅데이터는 관심의 대상이 되는 목표 모집단에 대한 대표성을 보장할 수 없고 지속적이고 안정적인 자료 확보가 어려운 단점이 있다. 또한 빅데이터의 상당수가 민감한 개인정보를 포함하고 데이터의 소유권도 논란의 여지가 많아 빅데이터의 공유 및 활용에 대한 원칙과 제도적 장치 마련이 필요한 상황이다. 한편 빅데이터는 자료의 양이 방대하여 통계분석 수행시 엄청나게 많은 양의 계산이 요구될 뿐만 아니라 작업시간도 매우 길다. 따라서 빅데이터 처리·분석 관련 전문성을 확보하고 IT 시스템을 구축할 필요가 있다.

이와 같은 현실로 인해 단기간에 빅데이터가 새로운 공식통계를 개발하거나 기존 공식통계를 대체하는 데에 활용되는 것을 기대하기는 쉽지 않다. 그러나 현재 스캐너 데이터, 웹 스크래핑 데이터, 지급결제 데이터 등 비교적 정형화된 빅데이터가 물가, 소비지출 등 경제통계 작성에 부분적으로 이용되고 있다. 또한 소셜미디어, 뉴스 등 텍스트 데이터와 인터넷 검색 데이터도 경제 및 금융 관련 지표 분석에 널리 활용되고 있다. 이러한 사례들은 빅데이터의 단점과 한계에도 불구하고 향후 빅데이터가 경제통계 작성 및 분석에 유용하게 활용될 가능성이 있음을 시사한다.

따라서 빅데이터가 쉽게 적용 가능한 경제통계 영역을 발굴하여 시험편제를 실시해보고 기존 활용 사례들을 벤치마킹하여 유용성을 검증해 볼 필요가 있다. 또한 빅데이터의 정제·처리·분석을 위한 전문적 지식을 습득하고 빅데이터에 적합한 통계 작성기법에 대한 조사 연구도 필요하다.

<경제통계국 국민소득총괄팀 과장 문혜정, 빅데이터통계연구반 과장 이해영>

참고문헌

- 강규호·김민수·김성은 (2015), “스캐너타입 자료를 활용한 물가지수 작성방법 연구,” 국민계정리뷰, 2015년 제4호.
- 이궁희·김용대·황희진 (2016), “빅데이터를 이용한 고용지표 개발,” 국민계정리뷰, 2016년 제1호.
- 장영재·박종문·김민수 (2017), “통관자료를 이용한 수출입물가지수 표본추출기법 연구,” 국민계정리뷰, 2017년 제2호.
- Armah, N. (2013), “Big Data Analysis: The Next Frontier,” Bank of Canada Review, Summer, pp. 32-39.
- Askitas, N. and K. Zimmermann (2009), “Google Econometrics and Unemployment Forecasting,” *Applied Economics Quarterly*, Vol. 55(2), pp. 107-120.
- Baker, S., N. Bloom and S. Davis (2015), “Measuring Economic Policy Uncertainty,” NBER Working Paper, No. 21633.
- Cavallo, A. and R. Rigobon (2016), “The Billion Prices Project: Using Online Prices for Measurement and Research,” *Journal of Economic Perspectives*, Vol. 30(2), pp. 151-178.
- Choi, H. and H. Varian, (2009), “Predicting Initial Claims for Unemployment Benefits,” available at: <http://research.google.com/archive/papers/initialclaimsUS.pdf>.
- Dass, P. and M. Puts (2014), “Social Media Sentiment and Consumer Confidence,” European Central Bank Statistics Paper Series, No. 5.
- Daas, P., M. Puts, B. Buelens and P. van den Hurk (2015), “Big Data as a Source for Official Statistics,” *Journal of Official Statistics*, Vol. 31(2), pp. 249-262.
- D’Amuri, F. (2009), “Predicting unemployment in short samples with internet job search query data,” Bank of Italy Research Department, available at: http://mpra.ub.uni-muenchen.de/18403/1/MPRA_paper_18403.pdf.
- Galbraith, J. and G. Tkacz (2015), “Nowcasting GDP with electronic payments data,” European Central Bank Statistics Paper Series, No. 10.
- McLaren, N. and R. Shanbhogue (2011), “Using internet search data as economic indicators,” Bank of England Quarterly Bulletin (Q2), pp. 134-140.
- Müller R. (2010), “Scanner data in the Swiss CPI: An alternative to price collection in the field,” Swiss Federal Statistical Office, available at: <http://www.unece.org/fileadmin/DAM/>

stats/documents/ece/ces/ge.22/2010/zip.10.e.pdf.

- Nyman, R., D. Gregory, S. Kapadia, P. Ormerod, D. Tuckett and R. Smith (2016), “News and narratives in financial systems: Exploiting big data for systemic risk assessment,” September 2016, available at: <http://www.norges-bank.no/contentassets/49b4dce839a7410b9a7f66578da8cf74/papers/smith.pdf>.
- Office for National Statistics (2016), “Research indices using web scraped price data: clustering large datasets into price indices (CLIP),” available at: <http://www.ons.gov.uk/economy/inflationandpriceindices/articles/researchindicesusingwebscrapedpricedata/clusteringlargedatasetsintopriceindicesclip>.
- Randi, J. (2016), “Scanner data in CPI/HICP,” Statistics Norway, available at: http://ec.europa.eu/eurostat/cros/content/use-scanner-data-norwegian-cpi_en.
- Rodriguez, J. and F. Haraldsen (2006), “The use of scanner data in the Norwegian CPI: The ‘new’ index for food and non-alcoholic beverages,” *Economic Survey*, Vol. 4, pp. 21-28.
- Shapiro, A. and D. Wilson (2017), “What’s in the News? A New Economic Indicator,” Federal Reserve Bank of San Francisco Economic Letter, 2017-10.
- Shapiro, A., M. Sudhof and D. Wilson (2017), “Measuring News Sentiment,” Federal Reserve Bank of San Francisco Working Paper, 2017-01.
- SWIFT (2016), “SWIFT Index Service Description,” October, available at: http://www2.swift.com/uhbonline/books/a2z/swift_index.htm.
- Thorsrud, L. A. (2016a), “Words are the new numbers: A newsy coincident index of business cycles,” Working Paper 44, Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School.
- Thorsrud, L. A. (2016b), “Nowcasting using news topics. Big Data versus big bank,” Working Paper 46, Centre for Applied Macro- and Petroleum economics(CAMP), BI Norwegian Business School.
- Van den Brakel, J., E. Söhler, P. Daas and B. Buelens (2016), “Social media as a data source for official statistics; the Dutch Consumer Confidence Index,” Statistics Netherlands Discussion paper, 2016-01.
- Van der Grient, H. and J. Hann (2010), “The use of supermarket scanner data in the Dutch CPI,” Statistics Netherland.
- Watanabe, K. and T. Watanabe (2014), “Estimating Daily Inflation Using Scanner Data: A Progress Report,” CARF Working Paper, No. 37.