조사연구(2020. 02) 21권 1호: 1−32
DOI http://dx.doi.org/10.20997/SR.21.1.1

연구논문

한국조사연구학회 발표 논문 내용분석, 1999～2019년: 학문분야·방법론·연구대상·데이터분석기법의 지속과 변동*

설동훈** · 고재훈*** · 유승환**** · 이기재*****

이 논문은 한국조사연구학회가 창립된 1999년부터 2019년까지 한국조사연구학회에서 출판·발표된 모든 논문의 전체 텍스트 내용을 분석하여 국내 조사연구의 변동 양상을 포착하는 것을 목표로 한다. 이를 위해, ≪조사연구≫ 게재 논문, 정기학술대회 발표 논문, 특별심포지엄 및 세미나 자료집 수록 논문을 빠짐없이 수집하였고 두 단계의 분석을 진행하였다. 첫째, '고전적 계량 내용분석'을 통해, 학문분야·방법론·연구대상·데이터분석기법의 네 차원에서 총 1,190개의 논문을 분류하였다. 둘째, '컴퓨터를 이용한 자동화된 텍스트 분석'의 하나인 '구조적 토픽모델'(Structural Topic Model: STM)을 1,145개 논문의 말뭉치(text corpus)에 적용하였다. 1999～2019년의 자료를 분석한 결과, 87개 토픽을 발견하였고, 연도별 각 토픽의 비중을 데이터로 삼아 스피어만(Spearman) 순위상관관계 분석을 진행하여 그 순위가 유사한 연도들을 묶어 1999～2004년, 2005～2012년, 2013～2019년의 세 시기로 구분했다. 토픽을 다섯 개의 핵심단어로 요약하여, 네 차원별로 세 시기씩 제시하였다. 학문분야·방법론·연구대상·데이터분석기법의 네 차원에서 연구 초점의 지속과 변동을 설명하였고, 각 토픽의 성장·지속·쇠퇴 형태 분석을 통해, 국내 조사연구의 발전 궤적을 파악하였다. 첫째, 학문분야의 다양성이 증가해왔다. 둘째, 방법론에서는 조사방법 자체에 관한 연구보다는 통계 분석기법을 적용한 논문이 증가해 왔다. 셋째, 연구대상에서는, 최근으로 올수록 연구방법 자체에 관한 연구보다는 통계적 분석을 적용하는 논문이 더 많이 발표되고 있다. 넷째, 데이터분석기법에서는 고급통계기법을 사용한 논문이 점점 늘고 있다.

주제어: 내용분석, 텍스트마이닝, 토픽모델링, 구조적 토픽모델, 한국조사연구학회

* 본 논문의 초고는 '한국조사연구학회 2019년 추계 학술대회'(대한상공회의소, 2019·11·22)에서 발표하였다. 논문의 완성도를 높일 수 있도록 비판과 조언 및 격려를 아끼지 않은 숙명여자대학교 김영원 교수, 한신대학교 변종석 교수, 연세대학교 강정한 교수, 그리고 ≪조사연구≫의 익명 심사자들께 감사드린다. 그렇지만 이 논문에 있을 수 있는 오류는 전적으로 저자들의 몫이다.
** 전북대학교 사회학과 교수(dhseol@jbnu.ac.kr), 주저자.
*** 전북대학교 사회과학연구소 전임연구원(jh.ko@jbnu.ac.kr).
**** 전북대학교 사회학과 박사과정(ysh9671@naver.com).
***** 한국방송통신대학교 정보통계학과 교수(kjlee@knou.ac.kr), 교신저자.

Ⅰ. 서론

한국조사연구학회는 조사연구와 관련이 있는 학계 및 실무 분야 전문가들이 조사연구를 활용하고 연구하는 학문분야 간 공동연구와, 이론을 중시하는 학계와 실무에 주력하는 업계의 협력이 필요하다는 점에 공감한 학자들이 1999년 11월 13일 설립하였다(설동훈·이유경·변종석·강정한 2019). 한국조사연구학회는 지난 20년 동안 정기학술대회 개최, ≪조사연구≫ 학술지 간행, 조사연구의 이론개발 및 보급·응용을 위한 제반 학술활동, 외국 조사연구학회와의 교류 및 공동학술회의 개최, 기타 조사연구에 관련된 활동을 수행해 왔다.

본 연구는 한국조사연구학회가 창립된 1999년부터 2019년까지 한국조사연구학회에서 출판·발표된 모든 논문의 전체 텍스트(full text) 내용을 분석하여 국내 조사연구의 변동 양상을 포착하는 것을 목표로 한다. 구체적으로, 학문분야·방법론·연구대상·데이터분석기법의 네 차원에서 시간 흐름에 따른 발표 논문의 주제와 관심사가 어떻게 바뀌어 왔는지 살펴보기로 한다. 본 연구는 지난 20년 동안의 한국조사연구학회의 학술활동의 의의를 되새기고, 앞으로의 성장의 방향을 모색하는 데 필요한 기초 자료를 제공할 것이다.

Ⅱ. 선행 연구 검토

1. 내용분석

내용분석(content analysis)을 연구방법으로 사용한 것은 20세기 초부터이다. 사회학에서는 일찍부터 내용분석을 이용하여 텍스트를 분석해 왔다(홍두승·설동훈 2012). 내용분석 방법은 '고전적 계량 내용분석'(classical quantitative content analysis)과 '컴퓨터를 이용한 자동화된 텍스트 분석'(computer-assisted or automated text analysis)의 두 가지가 있다(설동훈 2018).

첫째는 '고전적 계량 내용분석'은 인간(human coders)이 코딩 틀(coding scheme)

을 적용하여 텍스트에서 주제들을 식별하는 방법이다. 사람이 문서를 분석하여 코드 또는 주제를 분류하여야 하므로, 핵심 요소는 계량화된 정보 산출 이전에 만들어야 한다. 인간의 의사결정(human decision-making)이 코딩 과정에서 핵심 역할을 한다.

둘째, '컴퓨터 지원 내용분석'은 컴퓨터를 이용하여 텍스트 데이터에서 자연 언어 처리기술을 이용하여 유용한 정보를 추출·가공하는 방법으로, 연구자의 개입 또는 간여를 최소화한다. 예컨대, 여러 학문분야에서 전문 학술지에 게재된 논문 텍스트를 분석하는 연구 방법을 채택하고 있다(Amado, Cortez, Rita, and Moro 2018; Sun and Yin 2017; Prabhakaran, Hamilton, McFarland, and Jurafsky 2016; Blei 2012; Wang and McCallum 2006; Griffiths and Steyvers 2004).

한국에서도 텍스트의 내용을 통계적으로 분석하는 작업이 활발하다. 학술지 논문 초록을 이용한 학문분야의 분포와 연도별 추이 분석, 논문 전문을 이용한 단어 공출현 및 빈도, '개체명 인식'(Named Entity Recognition: NER)을 이용한 소속국가 및 기관 분석, 토픽모델링, 공저자 분석, 학술데이터베이스 *Web of Science*에서 검색한 논문의 초록을 이용한 학문분야 분석, 정보과학 연구 토픽 분석 등 방법론적으로 주목할 만한 연구가 다수 이루어졌다(박자현·송민 2013; Song and Kim 2013; 송혜지·박경수·정혜은·송민 2013; Kim, Song and Song 2015).

신재생에너지연구, 교통공학, 정보통신기술연구, 핀테크(FinTech: Financial Technology) 연구, 산업공학, 산업정보학, 전기전자공학 등 공학 분야에서도 각 학술지 수록 논문을 대상으로 학문분야 분포와 변동을 살피는 연구가 활발하게 이루어져 왔다(김태경·최회련·이홍철 2016; 박주섭·홍순구·김종원 2017). 이는 인문·사회과학 분야도 예외가 아니어서, 토픽모델링 등의 방법을 활용하여 연구동향을 검토하는 연구가 다수 이루어졌다(이재연, 2016; 박진균·김택윤·송민 2017). 최근에는 60년에 달하는 한 학회 활동 전체의 역사적 변동과 주제별 이합집산을 분석하여 연구동향을 살피는 수준에 이르렀다(설동훈·고재훈·유승환 2018).

또한, 연구주제뿐 아니라 연구방법을 분석한 논문도 다수 찾을 수 있다(Neall and Tuckey 2014; Long 2014; Randolph, Griffin, Zeiger, Falbe, Freeman, Taylor, Westbrook, Lico, Starling, Sprull, Holt, Smith, and McAnespie 2011; Edwards and Pedrotti 2008; Hutchinson and Lovell 2004). 이 논문들은 여러 학문분야가 모인 학제적 성격의 학술지에 수록된, 조사 데이터와 통계 분석을 활용한 이론적·경험적 연구를 분석하고 있다. 한국조사연구학회에서도 일찍이 이러한 성격의 외국 학술지의 연구동향을 분석하기도 했다(김규성·이기재 2001; 구자숙·김은미·이현희 2001).

2. 토픽모델링

텍스트마이닝(text mining)은 텍스트를 형태소 또는 연구자가 원하는 수준의 단어로 분류한 후, 여러 계량분석 방법을 적용해 결과를 도출하는 분석방법을 일컫는다. 자동화된 방식으로 텍스트에 등장하는 단어들의 출현 빈도 및 분포 경향을 텍스트가 발표된 시기, 작성자 등의 변수들을 함께 고려하여 분석한다. 이는 기존 분석방법으로는 연구대상으로 삼기 어려웠던 방대한 자료를 분석대상으로 삼을 수 있고, 그것을 정량적 방식으로 분석할 수 있다는 점에서 유용하다. 현재 텍스트마이닝 기법은 텍스트를 포함한 문서 간의 관련성, 특정 대상 또는 주제의 변동 등을 탐색하는 데 적극적으로 활용되고 있다(Evans and Aceves, 2016).

토픽모델링(topic modeling)은 텍스트마이닝 기법 중 하나로, 분석된 텍스트가 갖고 있는 주제들의 분포를 요약하여 제시한다. 토픽모델링은 연구자가 문서의 단위를 정하고 토픽의 개수를 지정하여 분석을 진행할 수 있는 활용도 높은 분석방법이다. 토픽모델링은 텍스트가 가지고 있는 주제들을 어휘들의 집합인 토픽(Topic)으로 요약하여 제시한다. 이를 통해, 연구자는 해당 토픽이 어떠한 내용을 담고 있는지 해석할 수 있고, 시기별 토픽의 분포를 분석하여 특정 주제의 변동을 파악할 수 있다.

블레이·응·조던(Blei, Ng, and Jordan, 2003)이 발표한 '잠재 디리클레 할당'(Latent Dirichlet Allocation: LDA) 모델은 가장 널리 쓰이고 있는 토픽모델링 알고리즘에 해당한다. LDA에서 하나의 문서는 다수의 토픽으로 구성되어 있다고 가정하며, 이러한 토픽은 전체 문서들에 포함된 어휘들의 분포로 나타난다고 가정한다. 전체 텍스트가 어떠한 주제들로 구성되어 있는지를 토픽의 분포를 통해 보여준다. 연구자는 토픽을 구성하는 단어들을 통해 그것이 나타내는 내용이 무엇인지 해석할 수 있다. 또한, 연구자는 유사한 주제들이 언급된 문서들은 비슷한 내용을 공유하는 문서라 해석할 수 있다고 가정하여, 각 문서가 가진 토픽의 분포를 이용하여 여러 가지 분석을 시도할 수 있다. 그리고 동일 범주에 속한 문서들이 보여주는 주제들의 분포를 통해, 토픽의 변동 양상을 설명할 수도 있다.

'구조적 토픽모델'(Structural Topic Model: STM)은 LDA를 기반으로 하여 개발된 것으로(Roberts, Stewart, and Tingley 2014),[1] 텍스트가 가지는 외부 변수를 문서별 토픽의 분포나 토픽을 구성하는 단어의 분포를 추정하는 데 활용할 수 있다

1) STM은 R에서 활용할 수 있는 Package로 개발되었다.

토픽모델링은 학술연구 동향을 살펴보는 데 유용하다. 연간 여러 차례 발간되는 학술지에 게재된 연구논문은 그 분량이 많으므로, 기존 '고전적 계량 내용분석'을 적용하기는 곤란하다. 그렇지만 토픽모델링 분석은 방대한 자료를 분석하는 데 아무 제약이 없을 뿐 아니라, 분석대상에 대한 통시적 관찰을 시도할 수도 있다. 토픽모델링은 주로 분석대상 논문의 전문, 요약문 또는 키워드 등을 분석하여 '활발히 연구되고 있는 주제'와 '그렇지 못한 주제'를 식별할 수 있도록 하므로, 연구주제의 변동 탐구에 쉽게 적용할 수 있는 것이다(설동훈·고재훈·유승환 2018; Hong 2019 참조).

Ⅲ. 자료와 분석방법

연구동향에 관한 기존의 계량적 연구들은 대부분 '고전적 계량 내용분석'을 활용하였다(홍두승·설동훈 2012: 30-31). 한 편의 '논문'을 단위로 하여 논문의 제목, 초록, 저자 등을 고려해 각 논문을 사전에 설정한 연구 분야로 할당하는 방식이 이에 해당한다. 이러한 접근 방식은 연구자가 사전에 설정해 놓은 분야 범주 이외의 주제를 포착하기 어렵다는 점과 복합적인 내용을 담고 있을 한 편의 논문을 하나의 범주로 할당하게 된다는 한계를 가지고 있다. 또한, 비록 상호주관성(inter-subjectivity)을 활용하여 자의적 판단을 최소화하였을지라도 주제들을 분류하면서, 연구자의 주관성을 완전히 배제하기 어렵다는 점 또한 취약점이다(설동훈·고재훈·유승환, 2018 참조). 이러한 한계에도 불구하고, 전통적 방식을 활용한 기존 연구는 특정 학회나 학술지의 연구동향 변동을 설명하는 데 많은 시사점을 주고 있다.

토픽모델링 분석을 통한 연구동향 분석은 기존 '고전적 계량 내용분석'이 가진 여러 가지 한계를 극복했다. 토픽모델링 분석은 문서에 분포된 전체 단어들을 동시에 고려하여 전체 텍스트가 가지고 있는 주제들을 추출하므로, '사전에 설계된 연구주제 범주'를 필요로 하지 않는다. 이는 '사전에 설계된 연구주제 범주'를 분석 대상 연구논문에 적용하는 기존의 연역적 접근 방식과 그 방향이 다르다. 생성된 모델을 통해 자동화된 분석을 하므로, 연구의 신뢰성 문제도 쉽게 극복할 수 있다.

그러나 토픽모델링 분석도 몇 가지 한계를 갖고 있다. 첫째, 생성된 토픽을 해석하는 방법에 관한 쟁점이다. 토픽모델링에서, 생성된 토픽은 어휘들의 집합으로 구성되

어 있고, 연구자는 토픽의 어휘와 그 분포를 통해 특정 토픽이 담고 있는 내용을 해석한다. ‘특정 토픽의 연도별 분포’나 ‘토픽에 특정 어휘가 포함된 확률’은 전체 텍스트에서 토픽이 담고 있는 주제가 어느 정도인지 해석할 수 있는 정보를 제공하지만, 연구자가 토픽을 해석할 때 자신의 주관적 시각을 활용할 수밖에 없다는 점이 문제다.

둘째, 최적 모델 설정에 관한 쟁점이다. 대표적으로, 토픽모델링에서 연구자는 토픽의 개수를 사전에 설정해야 한다. 토픽 개수를 적게 설정하면, 분석결과는 전체 텍스트의 주제들을 큰 갈래로 분류해 보여준다. 반대로, 토픽 개수를 많이 설정하면, 분석결과는 세부 주제들을 보여준다. 분석 텍스트를 가장 잘 설명할 수 있는 모델을 수학적으로 추론하기는 쉽지 않은데, 이는 ‘수학적으로 각 문서의 어휘 분포를 가장 잘 설명해내는 것’과 ‘연구자가 텍스트를 통하여 보고자 하는 정보’ 사이에 차이가 있기 때문이다(남춘호 2016). 텍스트의 전반적 주제 분포뿐 아니라 상세한 내용의 변동을 살피는 경우 그러한 문제가 전면에 대두된다.

‘고전적 계량 내용분석’과 ‘토픽모델링 분석’을 결합하면 각각의 한계를 보완할 수 있다. 고전적 계량 내용분석은 시간과 비용 및 문서 처리의 한계가 뚜렷하지만, ‘사전에 설계된 연구주제 범주’를 이용하여 연구논문의 변동을 파악할 수 있는 장점이 있다. 이 정보는 토픽모델링 분석을 통해 도출된 토픽을 해석하는 데 적용할 수 있다. 특정 범주로 분류된 논문들에 높게 분포하는 토픽이 나타내고 있는 주제는 해당 범주와 관련이 깊은 것으로 볼 수 있기 때문이다. 유사한 맥락에서, 토픽모델링을 통해 모델의 토픽 개수를 설정하는 데 참고할 수 있는 판단 기준을 찾을 수 있다. 이를 통하여, 토픽모델링 분석과 고전적 계량 내용분석을 결합한 분석은, 전통적 방식으로는 분석하기 어려웠던 장기간에 걸친 분석대상을 통시적으로 분석할 수 있을 뿐 아니라, 토픽모델링 분석에서 드러난 문제점을 보완할 수도 있다.

1. 자료

분석대상은 1999~2019년 한국조사연구학회에서 발행한 ≪조사연구≫에 게재된 논문, 정기학술대회, 특별심포지엄 및 세미나 자료집에 수록된 논문이다(변종석·설동훈 편, 2019 참조). 자료 중 디지털화되어 있지 않은 자료들에 대해서는 광학문자판독(Optical Character Recognition: OCR) 작업을 수행해 텍스트를 디지털화하였다. 이 과정에서 발생할 수 있는 잘못 인식된 어휘와 기호는 불용어(不用語, stop words) 처리 및 동의어 처리 등 텍스트 전처리(pre-processing)과정을 통해 분석에서 제외

하였다. 총 1,190편의 연구논문 중, 특히 파워포인트(PowerPoint) 슬라이드로 만들어진 정기학술대회, 특별심포지엄 및 세미나 자료집에 수록된 논문 중 텍스트는 거의 없이, 그래프와 그림으로 이루어진 45개는 토픽모델링 분석에 포함하지 않았다(설동훈·이유경·변종석·강정한 2019 참조). 그 결과, 이 논문의 토픽모델링 분석에 사용한 연구논문 수는 1,145편이다.

2. 분석방법

텍스트 분석방법은 '고전적 계량 내용분석'과 '컴퓨터를 이용한 자동화된 텍스트 분석'(토픽모델링)으로 구분할 수 있다. 이 두 분석방법은 완전히 별개로 이루어지지만, 양쪽에 모두 사용되는 사전 분류 기준을 통해 결과를 비교할 수 있다(설동훈·고재훈·유승환 2018).

1) 고전적 계량 내용분석

'고전적 계량 내용분석'의 절차에 따라 한국조사연구학회의 발표 논문 1,145편을 학문분야·방법론·연구대상·데이터분석기법의 네 차원에서 각각 기준을 세워 일련의 범주로 분류하였다(<표 3> 참조).

2) 토픽모델링

한국조사연구학회에서 출간된 학술연구논문 중 참고문헌을 제외한 모든 텍스트를 토픽모델링 분석에 포함하였다.[2] 참고문헌을 제외한 이유는 참고문헌들이 담고 있는 키워드들은 해당 논문의 주제를 벗어난 분야를 포괄하고 있을 가능성이 있기 때문이다. 텍스트 전처리(pre-processing) 단계에서는 '한국어 형태소 분석기 KOMORAN ver. 3.0'을 활용하였다. 더욱 정교한 단어 사전을 구축하기 위하여 한국통계학회의 통계 용어집과[3] 『사회조사분석』(홍두승·설동훈 2012)의 부록 '통계용어의 비교'를 참고하였으며, 그 외에도 각 논문의 서지 사항에 기록된 논문의 키워드와 저자의

2) 요약문 또는 초록이 아닌 '논문 전체'를 분석대상으로 삼음으로써 요약문이 없는 연구논문까지 분석 대상에 포함할 수 있었다. 연구논문의 주요 내용을 압축적으로 소개하는 요약문이 아닌 전체 텍스트를 분석에 사용하면서 우려되는 사항은, 단지 문장을 구성하는 데 사용되는 어휘들이나 상투적으로 쓰이는 단어들로 구성된 '중요한 의미를 갖지 않은 어휘들의 집합'이 특정 주제를 담고 있는 토픽으로 잘못 해석될 가능성이 있다는 것이다. 이를 보완하기 위하여, 우선, 고전적 계량 내용분석을 통해 전체 연구논문을 분류하여 분석된 토픽을 평가하는 데 활용하였고, 토픽모델링 분석결과를 제시할 때에도 범주별 토픽의 분포 변동을 제시함으로써, 각 범주에서 이루어진 연구내용 변동을 포착할 수 있도록 하였다.

3) https://www.kss.or.kr/bbs/board.php?bo_table=psd_sec

<표 1> 연도별 87개 토픽 계수의 분포를 근거로 산출한 순위상관계수 표

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2013a	2014	2015	2016	2017	2018	2019
1999	1.000	.648**	.503**	.532**	.228*	.338**	.310**	.283**	.428**	.328**	.390**	.403**	.221*	.200	-.090	.321**	.263*	.240*	.166	.375**	.181	.452**
2000		1.000	.606**	.595**	.455**	.463**	.382**	.311**	.359**	.351**	.504**	.278**	.369**	.301**	-.148	.317**	.322**	.346**	.243*	.388**	.155	.390**
2001			1.000	.460**	.410**	.250*	.356**	.423**	.246*	.307**	.308**	.250*	.062	.180	-.290**	.103	.096	.120	.064	.267*	.111	.303**
2002				1.000	.568**	.463**	.477**	.349**	.359**	.212*	.336**	.220*	.254*	.267*	-.115	.231*	.255*	.182	.114	.299**	-.019	.363**
2003					1.000	.398**	.476**	.283**	.234*	.212*	.315**	.237*	.151	.192	-.227*	.119	.281**	.078	.187	.167	.048	.252*
2004						1.000	.387**	.225*	.245*	.188	.259*	.326**	.220*	.269*	.090	.289**	.245*	.204	.090	.245*	.084	.199
2005							1.000	.428**	.329**	.294**	.397**	.375**	.302**	.424**	-.173	.241*	.339**	.194	.298**	.298**	.212*	.424**
2006								1.000	.347**	.429**	.236*	.334**	.087	.162	-.165	.130	.265*	.243*	.240*	.373**	.299**	.356**
2007									1.000	.283**	.260*	.327**	.175	.215*	.073	.401**	.317**	.256*	.367**	.492**	.304**	.420**
2008										1.000	.390**	.188	.085	.157	-.017	.264*	.269*	.147	.152	.236*	.167	.277**
2009											1.000	.208	.230*	.025	-.127	.158	.203	.120	.338**	.256*	.220*	.321**
2010												1.000	.275**	.241*	-.149	.067	.180	.039	.156	.112	.309**	.317**
2011													1.000	.504**	.136	.314**	.195	.192	.353**	.258*	.313**	.227*
2012														1.000	.171	.468**	.328**	.180	.244*	.343**	.171	.186
2013															1.000	n.a.	.186	.146	.194	.173	-.019	.011
2013a																1.000	.433**	.444**	.419**	.603**	.225*	.348**
2014																	1.000	.330**	.471**	.362**	.158	.423**
2015																		1.000	.354**	.495**	.297**	.451**
2016																			1.000	.474**	.326**	.432**
2017																				1.000	.336**	.562**
2018																					1.000	.422**
2019																						1.000

주: 1) 2013a 국제회의 발표논문 제외한 것.

2) * $p < .05$ 양측검정; ** $p < .01$ 양측검정.

3) n.a. 해당사항 없음(not applicable).

성명 등을 어휘 사전에 추가로 등록하고 동의어 사전을 구축했다. 최종적으로 분석에 활용된 어휘는 '2음절 이상의 명사' 중 최소 출현 빈도가 5회 이상인 어휘들의 '단어빈도-역문서 빈도'(Term Frequency-inverse Document Frequency: TF-IDF)의 값을 고려하여 확정하였다.[4)]

<표 2> 한국조사연구학회 발표 논문의 시기별 형태소 분석, 1999~2019년

(단위: 개, %)

구분		논문 수 (A)	단어 종류 (B)	단어 개수 (C)	단어 종류와 개수의 비(B / C)
전 시기 (1999~2019)	전체	1,145	115,137	2,017,155	5.7
	≪조사연구≫	332	23,303	977,351	2.4
	국내학술대회	655	32,378	834,360	3.9
	국제학술대회	86	12,082	128,756	9.4
	특별세미나	72	6,249	76,688	8.1
제1시기 (1999~2004)	전체	248	31,154	479,469	6.5
	≪조사연구≫	60	8,598	140,483	6.1
	국내학술대회	131	13,253	242,860	5.5
	특별세미나	57	9,303	96,126	9.7
제2시기 (2005~2012)	전체	369	38,136	642,984	5.9
	≪조사연구≫	123	14,037	348,566	4.0
	국내학술대회	223	19,655	271,558	7.2
	특별세미나	23	4,444	22,860	19.4
제3시기 (2013~2019)	전체	528	45,847	894,702	5.1
	≪조사연구≫	149	16,449	488,302	3.4
	국내학술대회	301	20,518	319,942	6.4
	국제학술대회	72	6,249	76,688	8.1
	특별세미나	6	2,631	9,770	26.9

주: '단어 종류와 개수의 비'(Type/Token Ratio: TTR)는 텍스트나 코퍼스의 길이에 따라 그 값이 크게 변동한다.
http://www.lexically.net/downloads/version5/HTML/index.html?type_token_ratio_proc.htm.

4) TF-IDF는 단어의 사용빈도는 높지만, 문서 대부분에 분포하여 문서의 특성을 잘 설명하지 못하는 단어를 찾아내는 데 활용된다. 각종 대명사나 특정 어휘들이 거의 모든 문서에 분포하여 높은 출현 빈도를 보이는 경우, 토픽모델링 분석 시 대부분의 토픽에 해당 어휘들이 등장하게 된다. 그 경우 토픽이 나타내는 주제를 파악하는 데, 또는 토픽이 분포하는 문서들을 설명하는 어려움이 발생한다. 한편, OCR 과정에서 생긴 오탈자를 걸러내기 위하여 최소 출현빈도를 고려하였다.

토픽의 개수는 토픽에 포함된 단어들의 내적 의미 일관성(semantic coherence)과 토픽 간의 상호배제성(mutual exclusivity)을 가장 잘 충족하는 값을 기준으로 하여 87개로 선정했다(Roberts, Stewart, and Tingley 2014).[5] 토픽 해석은 토픽에 포함된 상위 어휘들 및 토픽의 시기, 범주별 분포를 고려하여 결정하였다.

분석의 결과는 시기 및 고전적 계량 내용분석 방법에 따라 분류한 학문분야·방법론·연구대상·데이터분석기법별로 산출하여 분석하였다. 연구에서는 연도를 변수로 하였을 때, 토픽모델링을 통해 도출된 '각 문서상에 분포하는 토픽'이 '그해에 발표된 문서들'을 1로 할 때 차지하는 비중(weight)을 데이터로 하여, 스피어만의 순위상관계수(Spearman rank-order correlation coefficient)를 산출하였다. 토픽의 순위가 유사한 것으로 나오는 연도들을 묶어서(표 1 참조), 1999~2004년, 2005~2012년, 2013~2019년의 세 시기로 구분했다.[6]

<표 2>는 언어 형태소 분석을 통해 추출된 전체 단어의 종류와 개수를 시기별로 보여준다. 2013~2019년 논문 수는 528편으로, 1999~2004년의 248편보다 두 배 가까이 늘었고, 단어의 개수도 그 이전 시기보다 두 배 가까이 증가하였다. 다만, 내용의 다양성의 한 지표로 삼을 수 있는 단어의 종류 수는 약 1.5배 정도 상승하여 개수의 증가에는 미치지 못했다. 토픽모델링은 말뭉치에 포함된 단어를 대상으로 하므로, 논문 수의 증가율보다는 단어의 종류 수와 개수에 주목할 필요가 있다.

분석을 통해 추출된 토픽에 포함된 단어들은 비중(weight)이 높은 순으로 제시하였다. 토픽모델링 결과에는 '해석 불가능하거나 의미 없는 토픽'(junk and insignificant topics: J/I topics)도 포함되어 있을 수 있는데(Alsumait, Barbará, Gentle, and Domeniconi 2009), 그 경우는 'J/I토픽'이라 명명하여 잔여범주로 처리하였다. 전체 87개 토픽 중 'J/I토픽'이 1개 포함된 것을 확인하였다.

토픽모델링 분석 결과는 각 분류 차원별 범주에서의 토픽의 분포로 제시하였다. 각 분류영역에서 토픽은 순위에 따라 그 비중이 급격히 감소하는 멱법칙(冪法則, power law) 분포를 보이므로, 상위 토픽을 중심으로 해석하는 것이 필요하다. 우리는 학문

5) 설동훈·고재훈·유승환(2018)은 1964~1979년, 1980~1999년, 2000~2017년의 세 시기별로 각각 토픽모델링 분석을 진행했다. 연구논문 편수가 시기별로 크게 달라 상대적으로 분량이 적은 초기 연구논문이 분석 결과에서 과소평가될 가능성이 있었기 때문이다. 그러나 본 연구에서는 시기별 연구논문 편수가 크게 차이나지 않으므로, 단일 모델로 분석을 수행하였다. 이는 시기에 따른 주제(토픽)별 분포의 변동을 직접 비교하여 파악할 수 있는 장점이 있다.

6) 토픽모델링 분석은 하나의 문서가 다수의 토픽을 가지고 있다고 가정하기 때문에, 각 문서는 서로 다른 비중의 모든 토픽에 대한 값을 가지고 있다. 전체 텍스트의 잠재적인 주제들을 나타내는 토픽의 분포 패턴이 유사한 문서들은 서로 유사한 주제를 가지고 있는 것으로 파악할 수 있다.

분야·방법론·연구대상·데이터분석기법의 범주별로 각 시기의 상위 토픽과 그 기준값이 되는 세타의 평균(average theta: $\overline{\theta_d}$)을 제시하였다. 그 특성을 파악할 수 있도록 상위 토픽의 개수를 고정하지 않았다. 10위까지를 상위권으로 보고, 상위권 내 순위는 차등을 두지 않았고, 단 한 시기라도 상위권에 진입했던 토픽만을 제시하였다.

각 범주로 분류된 각각의 문서에 등장한 특정 토픽의 수가 많을수록 $\overline{\theta_d}$가 비례하여 커진다. 즉, 특정 토픽의 $\overline{\theta_d}$값이 클수록 그 토픽이 비중 있게 다루어졌음을 의미한다. 이를 시기별로 비교하여, 각 토픽의 성장·지속·쇠퇴와 같은 추세를 파악하였다. 상위권 진입 및 탈락과 상위권 외 순위에서 10단계 이상의 증감을 변동으로 규정하였다. '상위유지형'(—), '상승형'(/), '하강 후 상승형'(∨)은 '활발히 진행되는 연구주제'(hot topics)라고 할 수 있고, '상승 후 하강형'(∧), '하강형'(\ )은 '점차 사라지는 연구주제'(cold topics)라고 할 수 있다.

Ⅳ. 분석결과

1. 고전적 계량 내용분석

<표 3>에는 '고전적 계량 내용분석'에 의하여 논문 단위로 학문분야·방법론·연구대상·데이터분석기법의 네 차원에서 각각을 분류한 결과가 제시되어 있다. 전체적으로, 시간 흐름에 따라 전체 연구논문의 수가 증가하였고, 이는 한국조사연구학회의 연구활동이 활발해져 왔음을 의미한다. 학문분야에서는 'A2 기초사회과학'과 'A3 응용사회과학'의 논문 편수 증가에 주목할 수 있다. 조사방법론·통계학은 제1시기와 제2시기에 증가한 이후로 비슷하게 유지되는 반면, 기초사회과학과 응용사회과학 연구는 꾸준히 증가하고 있다. 이는 정립된 조사방법론·통계학을 활용한 사회과학 연구들이 활발히 발표됐음을 말해준다.

방법론의 경우, 'B2 통계적 분석법 적용'의 증가가 뚜렷하다. 한국조사연구학회 연구자들이 정립된 통계적 분석방법을 다양한 분야에 활용한 연구를 수행했음을 뜻한다. 'B1 조사방법'이 꾸준히 가장 높은 빈도를 보이는 것은 한국조사연구학회의 정체성을 반영한다.

연구대상을 살펴보면, ‘C1 사회’의 증가가 크게 눈에 띈다. 이는 다양한 통계적 분석들이 사회의 다양한 분석대상들에 적용됐음을 말해준다. 반면, ‘C2 경제’ 관련 연구대상은 제3시기에 이르러 줄어드는 추세를 보인다.

데이터분석기법을 살펴보면, 최근으로 올수록 분석 데이터가 사용되지 않은 경우가 적으며, ‘D2 기술통계분석’보다는 ‘D3 추론통계분석’을 활용하는 경향이 증가하는 것을 알 수 있다. 이는 한국조사연구학회에서 발표되는 논문들이 더욱 다양한 분석 기법을 활용해 왔음을 가리킨다.

<표 3> 한국조사연구학회 발표 논문의 유형별·시기별 분포, 1999～2019년

구분	편수(편)				구성비율(%)			
	전 시기 (1999 ～ 2019)	제1기 (1999 ～ 2004)	제2기 (2005 ～ 2012)	제3기 (2013 ～ 2019)	전 시기 (1999 ～ 2019)	제1기 (1999 ～ 2004)	제2기 (2005 ～ 2012)	제3기 (2013 ～ 2019)
전체	1,145	248	369	528	100.0	100.0	100.0	100.0
학문분야								
A1 조사방법론·통계학	583	160	211	212	50.9	64.5	57.2	40.2
A2 기초사회과학	386	54	102	230	33.7	21.8	27.6	43.6
A3 응용사회과학	176	34	56	86	15.4	13.7	15.2	16.3
방법론								
B1 조사방법	591	146	232	213	51.6	58.9	62.9	40.3
B2 통계적 분석법 적용	440	60	103	277	38.4	24.2	27.9	52.5
B3 질적 연구	114	42	34	38	10.0	16.9	9.2	7.2
연구대상								
C1 사회	610	87	163	360	53.3	35.1	44.2	68.2
C2 경제	176	44	72	60	15.4	17.7	19.5	11.4
C3 통계 일반	359	117	134	108	31.4	47.2	36.3	20.5
데이터분석기법								
D1 분석 데이터 사용 없음	269	101	102	66	23.5	40.7	27.6	12.5
D2 기술통계분석	82	40	32	10	7.2	16.1	8.7	1.9
D3 추론통계분석	794	107	235	452	69.3	43.1	63.7	85.6

2. 전 분야의 시기별 토픽 분석

<표 4> 한국조사연구학회 발표 연구논문의 시기별 주요 토픽, 1999~2019년

구분	유형	토픽		$\overline{\theta_d}$				순위			
				P_a	P_1	P_2	P3	P_a	P_1	P_2	P_3
전분야	—	T42	인터넷/사용자/온라인/응답/사이트	.046	.067	.051	.033	1	2	1	2
		T76	스트레스/청소년/학교/부모/폭력	.022	.027	.021	.020	5	7	10	7
	/	T64	가구/가구주/표본/웨이브/주택	.020	.019	.023	.019	8	15	8	9
		T47	조사/표본/조사방법/응답/오차	.018	.010	.012	.025	12	39	29	5
		T74	인구/총조사/주택총조사/주택/주민등록	.016	.001	.010	.027	13	75	45	4
		T13	행복/아동/주관적/소득/변수	.015	.001	.019	.020	19	73	14	8
		T14	정책/소통/아동/뉴스/지지	.015	.000	.001	.031	20	78	83	3
		T17	weight/body/study/people/factors	.011	.001	.001	.024	34	74	84	6
	∨	T12	sample/sampling/survey/row/coverage	.033	.012	.004	.063	2	34	75	1
		T69	불평등/계급/조직/계층/관리자	.019	.033	.008	.019	10	5	52	10
		T56	표본/데이터/확률/변수/추정량	.012	.032	.005	.008	28	6	70	51
		T79	사업체/산업/업종/접근법/통계조사	.012	.024	.005	.011	33	9	72	38
	∧	T30	센서스/인구/여성/남성/사망	.023	.019	.044	.009	3	17	2	47
		T54	문항/척도/측정/응답/검사	.022	.024	.028	.018	4	10	4	14
		T50	멘토/여성/정서/네트워크/남성	.018	.022	.024	.011	11	11	7	36
		T72	조사원/조사/가구/응답자/접촉	.015	.009	.027	.010	15	45	6	41
		T4	지지/대통령/장관/정부/리더십	.015	.011	.028	.009	18	37	5	49
		T8	척도/경청/오차/소통/조사	.013	.005	.029	.005	25	58	3	72
		T39	항목/한국인/측정치/군집/응답자	.012	.012	.022	.006	26	28	9	64
	\	T41	후보/선거/지지/지지율/대통령	.021	.036	.020	.015	6	4	13	20
		T37	지표/문화/측정/지수/지표체계	.021	.071	.013	.003	7	1	26	79
		T85	소득/가계/소비/자산/금융	.020	.054	.016	.007	9	3	17	58
		T83	취업/고용/일자리/노동/직업	.011	.024	.011	.004	38	8	35	77

주: P_a: 1999~2019년 전 시기, P_1: 1999~2004년, P_2: 2005~2012년, P_3: 2013~2019년.
—: 상위유지형, /: 상승형, ∨: 하강 후 상승형, ∧: 상승 후 하강형, \: 하강형.

<표 4>는 한국조사연구학회 발표 연구논문의 시기별 주요 토픽을 변동 유형별로 제시한 것이다. 토픽의 키워드는 가장 앞에 있는 것 하나만 언급하며 해석하기로 한다.[7] 상위유지형에 해당하는 'T42 인터넷'은 갈수록 그 영향력이 증대하는 인터넷, 온라인에 관한 토픽이다. 이는 연구논문들이 빈번하게 인터넷을 연구대상으로 삼아 왔고, 온라인 조사 등 조사방법 연구를 충실히 수행해 왔음을 말해준다. 다른 상위유지형인 'T76 스트레스'는 청소년의 교육에 관한 것으로 파악할 수 있는데, 이는 교육분야 연구가 전 시기에 걸쳐 빈번하게 이루어져 왔음을 가리킨다.

'상승형'과 '하강 후 상승형'에 해당하는 토픽은 대체로 계량적 조사방법에 관한 내용으로 해석된다. 이는 한국조사연구학회 연구자들이 조사방법에 관한 연구논문들을 지속해서 다루어 왔음을 보여준다. 또한, 'T13 행복', 'T14 정책', 'T69 불평등', 'T79 사업체' 등 연구분야 및 연구대상에 관한 것으로 해석되는 토픽은 연구방법과 관련하여 활발히 연구되어 온 주제라 할 수 있다.

'상승 후 하강형'과 '하강형'에 해당하는 토픽을 살펴보면, 그중 상당수가 선거에 관한 것이다. 이는 최근으로 올수록 계량분석 방법을 활용한 선거 관련 연구가 줄어들었다기보다는, 다양한 주제들이 다루어짐에 따라 그 상대적 비중이 하락한 것으로 해석하는 것이 타당하다.

3. 학문분야별 주요 토픽: 조사방법론·통계학, 기초사회과학, 응용사회과학

텍스트를 'A1 조사방법론·통계학', 'A2 기초사회과학', 'A3 응용사회과학'으로 분류하여 학문분야별 토픽을 살펴보기로 한다. <표 5>의 'A1 조사방법론·통계학'의 경우, <표 4>의 전 분야를 대상으로 한 결과에서는 볼 수 없었던 'T58 데이터', 'T78 전화', 'T46 출구조사' 등 구체적인 조사방법에 관한 토픽이 꾸준히 상위에 있다. 또, 전 분야에서 하강하는 토픽이었던 'T72 조사원'을 비롯해 'T67 모형', 'T34 무응답'과 같이, 조사의 어려움으로 인한 문제에 대한 통계적 대응책을 다루는 내용이 최근으로 올수록 집중적으로 연구되고 있다. 반면, 전 분야에서 상승하는 토픽이었던 'T42 인터넷'이나 'T79 사업체'는 최근으로 올수록 '조사방법론·통계학'에서 덜 다루어지고 있다.

---

7) 나머지 키워드는 표에서 확인할 수 있다.

<표 5> 학문분야별 연구논문의 주요 토픽, 1999~2019년

구분	유형	토픽		$\overline{\theta_d}$				순위			
				P_a	P_1	P_2	P_3	P_a	P_1	P_2	P_3
A1 조사방법론·통계학	—	T47	조사/표본/조사방법/응답/오차	.066	.078	.064	.059	1	3	1	1
		T58	데이터/품질/작성/기관/시스템	.040	.036	.041	.043	2	8	3	2
		T78	전화/유선/편향/전화조사/여론	.037	.039	.033	.040	4	6	7	3
		T46	출구조사/투표/예측/선거/오차	.036	.042	.034	.034	5	5	6	7
	/	T23	응답/질문/응답자/서베이/면접조사표	.028	.022	.026	.036	8	15	12	5
		T67	모형/변수/추정/측정/모수	.026	.018	.029	.031	10	19	10	8
		T43	응답/응답자/설문/문항/온라인	.026	.017	.032	.028	11	22	8	9
		T34	무응답/응답/표본/응답률/변수	.025	.023	.025	.027	12	14	15	10
		T64	가구/가구주/표본/웨이브/주택	.025	.019	.021	.035	13	17	18	6
		T72	조사원/조사/가구/응답자/접촉	.018	.004	.007	.039	21	40	43	4
	∨	T6	survey/data/quality/response/unit	.020	.049	.004	.015	17	4	56	23
	∧	T12	sample/sampling/survey/row/coverage	.030	.017	.048	.021	7	20	2	16
		T74	인구/총조사/주택총조사/주택/주민등록	.027	.013	.038	.026	9	28	4	13
		T44	가중치/가중/표본/조정/보정	.022	.018	.035	.012	15	18	5	31
		T16	감염/변수/텍스트/함수/판별	.015	.000	.029	.011	27	74	9	33
	\	T86	추정/추정량/표본/분산/모집단	.039	.083	.028	.017	3	2	11	20
		T42	인터넷/사용자/온라인/응답/사이트	.031	.086	.015	.005	6	1	24	48
		T84	표본/조사구/가구/조사/모집단	.021	.037	.019	.009	16	7	19	36
		T29	sample/variance/design/rotation/estimator	.017	.031	.021	.004	22	9	17	50
		T79	사업체/산업/업종/접근법/통계조사	.012	.031	.008	.002	32	10	38	60
A2 기초사회과학	/	T21	media/public/opinion/social/political	.036	.000	.003	.058	2	83	59	1
		T49	여성/자녀/출산/돌봄/가족	.030	.003	.035	.035	3	37	6	5
		T76	스트레스/청소년/학교/부모/폭력	.028	.002	.015	.040	5	45	27	3
		T13	행복/아동/주관적/소득/변수	.025	.000	.014	.035	8	70	30	4
		T38	자본/참여/공동체/연결망/모임	.022	.022	.019	.024	11	18	19	9
		T5	건강/은퇴/노인/우울/식사	.021	.000	.026	.024	14	62	13	8
		T25	level/life/factors/student/income	.016	.000	.000	.027	23	75	79	6
		T24	student/products/study/program/people	.015	.000	.001	.024	27	74	68	7
	∨	T51	정당/지지/유권자/태도/정치	.043	.044	.023	.051	1	6	15	2
		T71	여론/보도/기관/조사/선거	.025	.070	.006	.023	7	2	50	11
		T65	투표/선거/동원/투표율/지방선거	.023	.059	.001	.024	10	5	72	10
		T80	예측/선거/여론/후보/선거여론조사	.011	.041	.001	.009	34	7	69	39
		T46	출구조사/투표/예측/선거/오차	.008	.040	.000	.004	48	8	83	59

		T70	인터넷/미디어/측정/변수/행동	.027	.022	.044	.020	6	17	1	15
		T36	광고/소비자/제품/마케팅/시장	.022	.017	.038	.016	12	23	2	24
		T60	세대/중독/자살/스마트폰/게임	.022	.008	.032	.021	13	29	8	14
	∧	T66	가격/주택/아파트/월세/유동인구	.017	.008	.038	.009	20	28	3	35
		T50	멘토/여성/정서/네트워크/남성	.015	.033	.030	.004	28	11	9	61
		T63	브랜드/지수/산업/주가/가동	.011	.000	.029	.005	37	68	10	52
		T16	감염/변수/텍스트/함수/판별	.010	.007	.035	.000	40	30	5	77
		T75	보행/도시/사업/기금/보행자	.010	.000	.033	.003	41	66	7	65
		T47	조사/표본/조사방법/응답/오차	.030	.065	.036	.020	4	4	4	17
		T54	문항/척도/측정/응답/검사	.019	.039	.024	.012	15	9	14	32
	\	T42	인터넷/사용자/온라인/응답/사이트	.015	.066	.015	.003	26	3	25	62
		T28	지불/입찰/응답자/출발점/금액	.011	.074	.000	.000	38	1	81	79
		T33	기업/조직/사업장/제도/업무	.010	.039	.010	.004	39	10	40	60
	—	T33	기업/조직/사업장/제도/업무	.054	.085	.056	.040	3	2	5	6
		T10	health/study/care/elderly/stroke	.091	.001	.004	.184	1	49	35	1
		T76	스트레스/청소년/학교/부모/폭력	.040	.002	.036	.057	6	40	7	3
	/	T61	집단/잠재계층/변수/직무/점수	.034	.011	.025	.048	10	21	12	5
		T7	media/study/social/internet/people	.030	.021	.032	.032	11	17	8	10
		T11	network/social/retirement/mobile/preparation	.020	.000	.001	.040	16	71	60	7
		T24	student/products/study/program/people	.038	.002	.001	.077	8	37	54	2
		T4	지지/대통령/장관/정부/리더십	.027	.002	.002	.054	12	34	45	4
A3 응용사회과학	∨	T40	만족도/시민/만족/우울/아내	.025	.041	.006	.032	13	10	30	8
		T82	직업/종사자/기관/초과/서비스	.019	.056	.004	.013	17	6	34	21
		T15	factor/language/items/anxiety/class	.017	.003	.002	.032	19	30	47	9
		T36	광고/소비자/제품/마케팅/시장	.066	.035	.175	.007	2	11	1	30
	∧	T47	조사/표본/조사방법/응답/오차	.015	.017	.030	.005	21	19	9	32
		T53	유형/보상/스트레스/문화적응/외국인	.015	.002	.028	.012	22	36	10	22
		T1	학교/학생/학원/학년/수업	.047	.046	.072	.031	4	9	2	12
		T55	부패/공무원/행정/시민/측정	.042	.112	.052	.008	5	1	6	26
		T19	서비스/고객/만족/사원/모델	.039	.068	.069	.008	7	5	3	27
	\	T73	음주/구강/우식/건강/방문	.034	.071	.064	.001	9	4	4	54
		T54	문항/척도/측정/응답/검사	.023	.078	.019	.003	14	3	16	38
		T70	인터넷/미디어/측정/변수/행동	.016	.053	.016	.002	20	8	22	39
		T75	보행/도시/사업/기금/보행자	.014	.055	.004	.004	24	7	36	34

주: P_a: 1999~2019 전 시기, P_1: 1999~2004, P_2: 2005~2012, P_3: 2013~2019,
—: 상위유지형, /: 상승형, ∨: 하강 후 상승형, ∧: 상승 후 하강형, \: 하강형.

‘A2 기초사회과학’의 경우 전 분야에서 상위 토픽이었던 ‘T76 스트레스’와 ‘T13 행복’을 비롯해 ‘T21 미디어’, ‘T49 여성’, ‘T38 자본’, ‘T5 건강’ 등 전 분야에서는 나타나지 않았던 보다 다양한 사회적 쟁점이 발견된다. 지속적인 상위 토픽은 없었고, 각 토픽의 상승과 하강의 변동 폭이 다른 영역에 비해 컸다. 전 분야의 주요 토픽인 ‘T47 조사’의 경우 오히려 최근에는 상위권에서 내려온 것으로 나타났다.

‘A3 응용사회과학’의 경우에도 전 분야에서 상위 토픽이었던 ‘T76 스트레스’와 함께 ‘T33 기업’, ‘T10 건강’, ‘T61 집단’, ‘T7 미디어’, ‘T11 네트워크’ 등 전 분야에서는 나타나지 않았던 보다 전문적이고 다양한 사회적 쟁점이 연구의 관심을 끈다. 경영에 관한 내용인 ‘T33 기업’이 지속적인 상위 토픽으로 나타났고, 또한 전 분야의 주요 토픽인 ‘T47 조사’가 최근에는 ‘응용사회과학’ 연구의 상위권에서 탈락하였다.

세 분류범주에서 ‘T47 조사’의 내용은 어느 시기에든 나타나는 것으로 확인되었다. ‘T46 출구조사’의 내용은 ‘A1 조사방법론·통계학’과 ‘A2 기초사회과학’에서 공통으로 상승하였고, ‘T42 인터넷’, ‘T16 감염’은 공통으로 하강하는 토픽이었다. ‘A2 기초사회과학’과 ‘A3 응용사회과학’은 ‘T24 학생’, ‘T76 스트레스’가 공통으로 상승하였고, ‘T70 인터넷’, ‘T36 광고’ 등이 공통으로 하강하였다.

4. 방법론별 주요 토픽: 조사방법, 통계적 분석법 적용, 질적 연구

연구방법론에 따라 ‘B1 조사방법’, ‘B2 통계적 분석법 적용’, ‘B3 질적연구’로 분류하여 토픽을 살펴보기로 한다. <표 6>의 ‘B1 조사방법’의 경우, ‘T47 조사’, ‘T64 가구’와 같이 <표 4>의 전 분야에서 발견한 토픽 외에도 ‘T78 전화’, ‘T46 출구조사’, ‘T58 데이터’ 등 보다 구체적인 조사방법 토픽이 지속해서 상위권을 차지한다. 그 외에도 최근 상승하는 토픽이나 하강하는 토픽의 경우, 문항 개발이나 척도 구성과 같이 본격적인 조사방법을 사용한 것으로, 앞서 ‘A1 조사방법론·통계학’의 텍스트들의 내용을 분석한 결과와 유사한 경향을 보인다.

‘B2 통계적 분석법 적용’의 경우 전 분야에서 상위 토픽이었던 ‘T76 스트레스’와 ‘T13 행복’을 비롯해 ‘T10 건강’, ‘T51 정당’, ‘T49 여성’, ‘T24 학생’, ‘T21 미디어’, ‘T83 취업’, ‘T38 자본’, ‘T25 수준’ 등 전 분야에서는 나타나지 않았던 보다 다양한 사회적 쟁점이 관심을 끌고 있다. 모델 구성과 시계열 분석 등 통계적 분석법을 다양한 사회현상에 적용하는 연구가 활발히 이루어지고 있음을 뜻한다.

<표 6> 방법론별 연구논문의 주요 토픽, 1999~2019년

구분	유형	토픽		$\overline{\theta_d}$				순위			
				P_a	P_1	P_2	P_3	P_a	P_1	P_2	P_3
B1 조사방법	—	T47	조사/표본/조사방법/응답/오차	.062	.073	.059	.058	1	3	1	1
		T78	전화/유선/편향/전화조사/여론	.040	.038	.033	.048	2	6	6	2
		T46	출구조사/투표/예측/선거/오차	.036	.040	.031	.038	4	5	8	4
		T58	데이터/품질/작성/기관/시스템	.031	.031	.029	.033	6	8	9	8
	/	T67	모형/변수/추정/측정/모수	.029	.021	.028	.035	7	15	11	5
		T43	응답/응답자/설문/문항/온라인	.028	.017	.028	.034	8	21	10	6
		T23	응답/질문/응답자/서베이/면접조사표	.023	.020	.016	.034	12	16	24	7
		T64	가구/가구주/표본/웨이브/주택	.022	.015	.018	.032	15	22	18	9
		T72	조사원/조사/가구/응답자/접촉	.018	.004	.006	.041	19	44	49	3
		T20	data/pr/fit/analysis/Japanese	.017	.000	.018	.028	22	67	19	10
	∨	T6	survey/data/quality/response/unit	.021	.047	.007	.017	16	4	46	19
	∧	T54	문항/척도/측정/응답/검사	.027	.023	.035	.021	9	14	5	14
		T12	sample/sampling/survey/row/coverage	.026	.005	.043	.021	10	41	2	16
		T44	가중치/가중/표본/조정/보정	.023	.019	.033	.014	13	19	7	23
		T74	인구/총조사/주택총조사/주택/주민등록	.022	.008	.040	.013	14	34	3	28
		T16	감염/변수/텍스트/함수/판별	.017	.000	.035	.010	21	70	4	33
	\	T86	추정/추정량/표본/분산/모집단	.037	.083	.026	.017	3	2	12	22
		T42	인터넷/사용자/온라인/응답/사이트	.033	.097	.018	.005	5	1	17	47
		T84	표본/조사구/가구/조사/모집단	.019	.037	.017	.009	18	7	20	35
		T29	sample/variance/design/rotation/estimator	.015	.026	.019	.004	25	10	16	53
		T79	사업체/산업/업종/접근법/통계조사	.011	.029	.007	.002	36	9	47	59
B2 통계적 분석법 적용	/	T10	health/study/care/elderly/stroke	.039	.000	.002	.062	1	70	58	1
		T76	스트레스/청소년/학교/부모/폭력	.038	.003	.025	.050	2	43	12	2
		T51	정당/지지/유권자/태도/정치	.031	.014	.013	.042	3	27	33	5
		T49	여성/자녀/출산/돌봄/가족	.029	.002	.033	.034	4	49	5	6
		T24	student/products/study/program/people	.028	.001	.002	.044	5	56	62	3
		T21	media/public/opinion/social/political	.026	.000	.000	.042	9	86	84	4
		T83	취업/고용/일자리/노동/직업	.024	.008	.032	.024	10	32	7	9
		T13	행복/아동/주관적/소득/변수	.023	.000	.006	.033	11	73	49	7
		T38	자본/참여/공동체/연결망/모임	.021	.020	.015	.024	14	19	28	10
		T25	level/life/factors/student/income	.020	.000	.010	.027	16	68	43	8
	∨	T65	투표/선거/동원/투표율/지방선거	.020	.051	.001	.020	15	4	67	16

	∧	T68	종교/단체/죽음/성장/시민	.021	.024	.029	.018	13	18	9	17
		T47	조사/표본/조사방법/응답/오차	.019	.031	.033	.011	18	12	6	33
		T5	건강/은퇴/노인/우울/식사	.019	.000	.027	.020	19	69	10	15
	\	T33	기업/조직/사업장/제도/업무	.028	.079	.031	.016	6	1	8	22
		T36	광고/소비자/제품/마케팅/시장	.027	.044	.064	.009	7	5	1	40
		T70	인터넷/미디어/측정/변수/행동	.026	.065	.035	.015	8	2	4	27
		T19	서비스/고객/만족/사원/모델	.017	.042	.038	.004	20	6	2	52
		T55	부패/공무원/행정/시민/측정	.017	.064	.019	.006	21	3	20	49
		T82	직업/종사자/기관/초과/서비스	.016	.036	.017	.011	27	8	22	34
		T73	음주/구강/우식/건강/방문	.014	.040	.036	.000	31	7	3	76
		T54	문항/척도/측정/응답/검사	.009	.033	.008	.004	46	10	48	54
		T28	지불/입찰/응답자/출발점/금액	.005	.034	.000	.000	58	9	73	78
	—	T47	조사/표본/조사방법/응답/오차	.068	.097	.054	.049	2	1	6	5
		T58	데이터/품질/작성/기관/시스템	.047	.030	.056	.057	4	7	4	4
	/	T36	광고/소비자/제품/마케팅/시장	.076	.020	.191	.034	1	17	1	10
		T21	media/public/opinion/social/political	.033	.000	.009	.091	6	69	20	1
		T74	인구/총조사/주택총조사/주택/주민등록	.030	.019	.021	.049	8	18	15	6
		T70	인터넷/미디어/측정/변수/행동	.023	.001	.036	.037	14	61	9	9
		T18	대출/시장/주택/금융/영아	.015	.000	.000	.046	22	75	84	7
	∨	T80	예측/선거/여론/후보/선거여론조사	.059	.092	.003	.072	3	2	39	3
		T71	여론/보도/기관/조사/선거	.044	.071	.009	.045	5	3	22	8
B3		T11	network/social/retirement/mobile/preparation	.026	.002	.001	.076	11	50	63	2
		T85	소득/가계/소비/자산/금융	.013	.030	.001	.004	25	8	60	30
질	∧	T23	응답/질문/응답자/서베이/면접조사표	.031	.015	.059	.024	7	21	3	15
		T50	멘토/여성/정서/네트워크/남성	.027	.000	.053	.034	10	67	7	11
적		T27	군중/생애사/생애/실업자/운동	.025	.000	.055	.027	12	79	5	14
		T16	감염/변수/텍스트/함수/판별	.025	.000	.078	.005	13	81	2	26
연		T20	data/pr/fit/analysis/Japanese	.019	.003	.050	.009	16	44	8	22
		T7	media/study/social/internet/people	.017	.001	.030	.024	19	58	10	17
구	\	T42	인터넷/사용자/온라인/응답/사이트	.027	.066	.006	.003	9	4	27	33
		T46	출구조사/투표/예측/선거/오차	.023	.062	.000	.000	15	5	75	68
		T12	sample/sampling/survey/row/coverage	.019	.048	.002	.001	18	6	44	50
		T53	유형/보상/스트레스/문화적응/외국인	.011	.029	.001	.000	28	9	62	72
		T29	sample/variance/design/rotation/estimator	.011	.028	.001	.000	29	10	53	84

주: P_a: 1999~2019 전 시기, P_1: 1999~2004, P_2: 2005~2012, P_3 2013~2019,
—: 상위유지형, /: 상승형, ∨: 하강 후 상승형, ∧: 상승 후 하강형, \: 하강형.

‘B3 질적 연구’의 경우 전 분야에서 상위 토픽이었던 ‘T47 조사’와 ‘T74 인구’ 등이 상승 토픽으로 나타났고, ‘T36 광고’, ‘T21 미디어’, ‘T70 인터넷’, ‘T18 대출’ 등 다른 분류에서 하강하는 토픽이거나 나타나지 않았던 토픽이 상승하는 것으로 나타나 질적 연구의 토픽이 구분됨을 알 수 있다.

세 분류범주의 방법론은 다르지만, ‘T47 조사’는 항상 발견되었고, ‘B1 조사방법’과 ‘B3 질적 연구’에서는 상위 토픽을 유지하였다. ‘T58 데이터’의 내용은 ‘B1 조사방법’과 ‘B3 질적 연구’에서 공통으로 상승하고, ‘T42 인터넷’, ‘T16 감염’, ‘T12 표본’, ‘T29 표본’은 공통으로 하강한다. ‘B2 통계적 분석법 적용’과 ‘B3 질적 연구’는 ‘T21 미디어’가 공통으로 상승하고, ‘T70 인터넷’, ‘T36 광고’도 추세의 방향은 다르지만 두 범주에서 모두 발견된다.

5. 연구대상별 주요 토픽: 사회, 경제, 통계 일반

연구대상을 ‘C1 사회’, ‘C2 경제’, ‘C3 통계 일반’으로 분류하여 토픽을 살펴보기로 한다. <표 7>의 ‘C1 사회’의 경우, ‘T47 조사’, ‘T76 스트레스’, ‘T13 행복’과 같이 <표 4>의 전 분야에서 발견한 토픽 외에도 ‘T46 출구조사’, ‘T51 정당’, ‘T10 건강’, ‘T21 미디어’, ‘T49 여성’, ‘T24 학생’ 등 전 분야에서는 나타나지 않았던 다양한 사회적 쟁점이 포함되었다. 또, 전 분야에서 하강하는 토픽이었던 ‘T41 후보’가 최근 들어 ‘C1 사회’ 분야에서 관심을 끌고 있다. 반면, 전 분야에서 상승 토픽이었던 ‘T42 인터넷’, ‘T74 인구’는 사회 대상 연구에서는 최근으로 올수록 덜 다루어진다.

‘C2 경제’의 경우, 전 분야에서 상승하는 토픽이었던 ‘T79 사업체’를 비롯해, ‘T70 인터넷’, ‘T83 취업’, ‘T24 학생’, ‘T82 직업’, ‘T61 집단’, ‘T25 수준’ 등 전 분야에서 나타나지 않았거나 다른 범주와 분류의 텍스트들에서 볼 수 없었던 경제 토픽이 발견되고 있다. 또, 전 분야에서 하강하는 토픽이었던 ‘T83 취업’과 ‘T85 소득’과 같이 경제 쟁점이 최근 들어 이 분야에서 관심을 끌고 있다. 반면, 전 분야에서 상승하는 토픽이었던 ‘T47 조사’는 경제연구에서는 최근으로 올수록 덜 다루어진다.

‘C3 통계 일반’의 경우 전 분야에서 상승하는 토픽이었던 ‘T47 조사’와 ‘T64 가구’, ‘T56 표본’을 비롯해 ‘T58 데이터’, ‘T23 응답’ 등이 주요 토픽으로 발견되었다.

세 분류범주에서, ‘T47 조사’는 어느 시기에든 발견되고, ‘C1 사회’과 ‘C3 통계 일반’에서는 지속적인 상위 토픽이다. ‘T24 학생’은 ‘C1 사회’과 ‘C2 경제’에서 공통으로 상승하고, 반면 ‘T66 가격’의 내용은 공통으로 하강한다. ‘T42 인터넷’, ‘T16 감

염', 'T74 인구'는 'C1 사회'과 'C3 통계 일반'에서 공통으로 하강하고, 'T86 추정'은 'C2 경제'와 'C3 통계 일반'에서 공통으로 하강한다.

<표 7> 연구대상별 연구논문의 주요 토픽, 1999~2019년

구분	유형	토픽		$\overline{\theta_d}$				순위			
				P_a	P_1	P_2	P_3	P_a	P_1	P_2	P_3
C1 사회	—	T47	조사/표본/조사방법/응답/오차	.039	.065	.046	.030	1	2	1	5
		T46	출구조사/투표/예측/선거/오차	.036	.085	.039	.022	2	1	2	9
	/	T51	정당/지지/유권자/태도/정치	.032	.027	.026	.036	3	14	9	4
		T76	스트레스/청소년/학교/부모/폭력	.029	.002	.021	.040	4	55	18	3
		T10	health/study/care/elderly/stroke	.028	.000	.000	.047	5	68	78	1
		T21	media/public/opinion/social/political	.025	.000	.002	.042	6	86	72	2
		T49	여성/자녀/출산/돌봄/가족	.020	.002	.022	.024	10	52	16	8
		T13	행복/아동/주관적/소득/변수	.018	.000	.011	.026	13	75	38	6
		T24	student/products/study/program/people	.012	.000	.000	.021	30	78	84	10
	∨	T71	여론/보도/기관/조사/선거	.025	.048	.012	.025	7	4	34	7
		T80	예측/선거/여론/후보/선거여론조사	.019	.053	.009	.016	11	3	42	22
		T65	투표/선거/동원/투표율/지방선거	.016	.037	.004	.016	19	7	66	20
	∧	T78	전화/유선/편향/전화조사/여론	.024	.029	.030	.019	8	13	5	12
		T37	지표/문화/측정/지수/지표체계	.018	.002	.028	.018	12	54	7	16
		T74	인구/총조사/주택총조사/주택/주민등록	.017	.010	.031	.012	16	32	3	30
		T1	학교/학생/학원/학년/수업	.015	.011	.031	.008	23	29	4	47
		T70	인터넷/미디어/측정/변수/행동	.014	.014	.026	.009	24	23	8	40
		T66	가격/주택/아파트/월세/유동인구	.011	.003	.024	.007	35	48	10	53
		T16	감염/변수/텍스트/함수/판별	.010	.013	.028	.002	39	26	6	75
	\	T54	문항/척도/측정/응답/검사	.018	.042	.024	.009	15	6	11	45
		T55	부패/공무원/행정/시민/측정	.012	.032	.017	.005	32	10	26	61
		T42	인터넷/사용자/온라인/응답/사이트	.012	.043	.013	.003	33	5	32	64
		T73	음주/구강/우식/건강/방문	.011	.033	.023	.001	34	9	12	79
		T84	표본/조사구/가구/조사/모집단	.006	.036	.004	.001	60	8	67	80
	—	T33	기업/조직/사업장/제도/업무	.056	.094	.051	.033	2	1	4	8
		T70	인터넷/미디어/측정/변수/행동	.029	.041	.026	.025	9	8	10	10
	/	T83	취업/고용/일자리/노동/직업	.054	.010	.044	.098	3	27	6	1
		T24	student/products/study/program/people	.029	.001	.002	.080	10	48	49	2
		T82	직업/종사자/기관/초과/서비스	.028	.009	.034	.035	11	28	7	5

		T61	집단/잠재계층/변수/직무/점수	.022	.008	.023	.031	15	31	13	9
		T25	level/life/factors/student/income	.012	.000	.000	.034	27	71	80	6
		T85	소득/가계/소비/자산/금융	.033	.048	.014	.044	6	5	20	3
	∨	T79	사업체/산업/업종/접근법/통계조사	.032	.063	.011	.033	7	3	24	7
		T52	표본/농가/조사/모집단/조사구	.024	.033	.007	.039	13	11	35	4
		T63	브랜드/지수/산업/주가/가동	.048	.012	.094	.020	4	26	2	14
C2		T19	서비스/고객/만족/사원/모델	.029	.013	.054	.012	8	25	3	31
경	∧	T58	데이터/품질/작성/기관/시스템	.014	.003	.028	.005	21	39	8	42
제		T87	면적/경지/조사/표본/총조사	.011	.000	.028	.000	28	77	9	72
		T36	광고/소비자/제품/마케팅/시장	.093	.053	.179	.019	1	4	1	18
		T47	조사/표본/조사방법/응답/오차	.038	.045	.048	.021	5	6	5	12
	\	T86	추정/추정량/표본/분산/모집단	.026	.042	.025	.015	12	7	12	25
		T28	지불/입찰/응답자/출발점/금액	.023	.091	.001	.001	14	2	57	62
		T66	가격/주택/아파트/월세/유동인구	.022	.039	.017	.016	16	10	16	24
		T32	연금/추계/가입/수급/가입자	.011	.040	.001	.001	30	9	60	59
		T47	조사/표본/조사방법/응답/오차	.062	.077	.059	.048	1	3	2	4
	—	T58	데이터/품질/작성/기관/시스템	.056	.046	.052	.070	2	5	3	1
		T23	응답/질문/응답자/서베이/면접조사표	.039	.029	.040	.048	6	10	7	5
		T67	모형/변수/추정/측정/모수	.038	.024	.036	.056	7	12	9	2
		T20	data/pr/fit/analysis/Japanese	.031	.002	.041	.050	9	47	5	3
	/	T34	무응답/응답/표본/응답률/변수	.031	.028	.034	.030	10	11	11	9
		T64	가구/가구주/표본/웨이브/주택	.025	.012	.025	.040	15	26	18	7
		T62	패널/응답/패널조사/가구/표본	.014	.003	.012	.029	25	39	26	10
C3		T56	표본/데이터/확률/변수/추정량	.035	.034	.028	.043	8	9	16	6
통	∨	T6	survey/data/quality/response/unit	.030	.068	.007	.017	11	4	35	17
계		T78	전화/유선/편향/전화조사/여론	.029	.035	.022	.030	12	8	20	8
일		T12	sample/sampling/survey/row/coverage	.040	.023	.065	.028	5	14	1	13
반		T43	응답/응답자/설문/문항/온라인	.028	.021	.035	.028	13	17	10	12
	∧	T44	가중치/가중/표본/조정/보정	.024	.022	.040	.008	16	15	6	31
		T16	감염/변수/텍스트/함수/판별	.021	.000	.043	.016	18	70	4	18
		T74	인구/총조사/주택총조사/주택/주민등록	.020	.010	.037	.009	20	27	8	27
		T42	인터넷/사용자/온라인/응답/사이트	.046	.117	.017	.004	3	1	23	45
	\	T86	추정/추정량/표본/분산/모집단	.045	.086	.029	.022	4	2	15	16
		T29	sample/variance/design/rotation/estimator	.028	.043	.032	.007	14	6	13	34
		T71	여론/보도/기관/조사/선거	.017	.035	.008	.008	23	7	32	32

주: P_a: 1999~2019 전 시기, P_1: 1999~2004, P_2: 2005~2012, P_3: 2013~2019,
—: 상위유지형, /: 상승형, ∨: 하강 후 상승형, ∧: 상승 후 하강형, \: 하강형.

6. 데이터분석기법별 주요 토픽: 데이터 미사용, 기술통계분석, 추론통계분석

대부분 조사연구가 데이터분석기법을 사용하고 있다. 조사연구 텍스트들이 어떤 데이터분석기법을 사용했는가는 'B 방법론'과 구분되는 측면을 나타낼 뿐 아니라, 일찍이 연구자들이 관심을 기울인 사항이기도 하다(김규성·이기재 2001; 구자숙·김은미·이현희 2001). 텍스트의 데이터분석기법에 따라 'D1 분석데이터 사용하지 않음', 'D2 기술통계분석', 'D3 추론통계분석'으로 분류하여 토픽을 살펴보기로 한다.

<표 8>의 'D1 분석데이터 사용하지 않음'의 경우, 'T47 조사', 'T74 인구'와 같이 <표 4>의 전 분야에서 발견한 토픽뿐 아니라, 'T58 데이터' 등과 같이 상승 경향 토픽이 대부분 조사와 분석에 관련된 것으로, 비록 데이터를 사용하여 분석한 논문은 아니지만, 그 내용은 조사와 데이터, 분석의 내용을 담고 있음을 알 수 있다.

'D2 기술통계분석'의 경우 'T47 조사', 'T42 인터넷', 'T74 인구', 'T64 가구'와 같이 전 분야에서 나타난 토픽뿐 아니라 'T37 지표', 'T1 학교', 'T45 문화' 등과 같이 전 분야에서 나타나지 않았거나, 다른 범주와 분류의 텍스트들에서 볼 수 없었던 다양한 것들이 상승하는 토픽으로 나타났다. 또, 전 분야에서 하강하는 토픽이었던 'T72 조사원'과 'T37 지표' 등이 최근 들어 이 분야에서 관심을 끌고 있다. 이는 대개 통계 관련 분류와 범주에 등장하는 토픽이라는 특성을 가진다.

'D3 추론통계분석'의 경우, 'T47 조사', 'T76 스트레스', 'T13 행복'과 같이 전 분야에서 나타난 토픽뿐 아니라, 'T46 출구조사', 'T51 정당', 'T10 건강', 'T21 미디어', 'T78 전화', 'T49 여성' 등과 같이 사회현상 연구에서 볼 수 있었던 것들이 상승하는 토픽으로 나타났다.

세 분류범주의 각각의 텍스트들은 분석기법이 다르지만, 'T47 조사', 'T42 인터넷', 'T16 감염'이 어느 시기에든 나타났고, 특히 'T47 조사'는 모든 범주에서 지속적인 상위 토픽이기도 하였다.

<표 8> 데이터분석기법별 연구논문의 주요 토픽, 1999～2019년

구분	유형	토픽		$\overline{\theta_d}$				순위			
				Pa	P_1	P_2	P_3	Pa	P_1	P_2	P_3
D1 분석데이터 사용하지 않음	—	T58	데이터/품질/작성/기관/시스템	.073	.056	.071	.100	1	6	2	1
		T47	조사/표본/조사방법/응답/오차	.072	.090	.067	.053	2	2	3	3
	/	T20	data/pr/fit/analysis/Japanese	.037	.002	.057	.059	6	52	4	2
		T67	모형/변수/추정/측정/모수	.029	.017	.028	.050	10	16	7	5
		T74	인구/총조사/주택총조사/주택/주민등록	.025	.009	.031	.041	13	27	6	7
		T37	지표/문화/측정/지수/지표체계	.021	.004	.025	.039	17	44	10	8
		T16	감염/변수/텍스트/함수/판별	.019	.008	.021	.034	18	31	16	9
		T72	조사원/조사/가구/응답자/접촉	.012	.003	.009	.032	21	48	39	10
	∨	T71	여론/보도/기관/조사/선거	.039	.056	.014	.050	4	5	22	4
		T6	survey/data/quality/response/unit	.033	.078	.005	.007	7	3	51	28
		T80	예측/선거/여론/후보/선거여론조사	.027	.042	.001	.044	11	7	67	6
	∧	T36	광고/소비자/제품/마케팅/시장	.037	.015	.081	.003	5	21	1	39
	\	T42	인터넷/사용자/온라인/응답/사이트	.043	.101	.010	.005	3	1	31	29
		T86	추정/추정량/표본/분산/모집단	.030	.067	.012	.001	8	4	25	53
		T29	sample/variance/design/rotation/estimator	.030	.042	.031	.009	9	8	5	26
		T12	sample/sampling/survey/row/coverage	.027	.026	.027	.029	12	9	9	12
		T23	응답/질문/응답자/서베이/면접조사표	.023	.024	.027	.015	14	10	8	21
D2 기술통계분석	—	T47	조사/표본/조사방법/응답/오차	.078	.089	.065	.070	2	2	3	4
	/	T74	인구/총조사/주택총조사/주택/주민등록	.074	.020	.125	.122	3	19	1	1
		T37	지표/문화/측정/지수/지표체계	.018	.001	.021	.071	18	46	17	3
		T1	학교/학생/학원/학년/수업	.014	.002	.004	.092	23	42	46	2
		T45	문화/음악/문화자본/교사/치료	.008	.000	.000	.067	40	74	65	5
	∨	T42	인터넷/사용자/온라인/응답/사이트	.061	.101	.016	.048	4	1	20	9
		T84	표본/조사구/가구/조사/모집단	.042	.066	.007	.060	5	4	34	6
		T71	여론/보도/기관/조사/선거	.024	.028	.009	.053	10	10	31	7
		T64	가구/가구주/표본/웨이브/주택	.022	.032	.006	.038	12	8	37	10
		T82	직업/종사자/기관/초과/서비스	.016	.029	.004	.003	19	9	43	30
		T72	조사원/조사/가구/응답자/접촉	.014	.012	.004	.053	24	24	41	8
	∧	T23	응답/질문/응답자/서베이/면접조사표	.025	.027	.030	.002	9	12	9	35
		T3	응답자/통화/거절/전화번호/무선표집	.023	.023	.031	.001	11	16	7	42
		T16	감염/변수/텍스트/함수/판별	.020	.000	.050	.000	14	73	4	55
		T69	불평등/계급/조직/계층/관리자	.015	.000	.038	.000	21	65	5	52

		T14	정책/소통/아동/뉴스/지지	.011	.000	.027	.002	31	60	10	41
	\	T78	전화/유선/편향/전화조사/여론	.080	.081	.096	.023	1	3	2	14
		T55	부패/공무원/행정/시민/측정	.038	.053	.031	.000	6	5	8	69
		T79	사업체/산업/업종/접근법/통계조사	.035	.051	.026	.000	7	6	12	66
		T46	출구조사/투표/예측/선거/오차	.033	.038	.038	.000	8	7	6	58
D3 추론통계분석	—	T47	조사/표본/조사방법/응답/오차	.034	.037	.042	.029	1	5	1	4
		T46	출구조사/투표/예측/선거/오차	.024	.048	.026	.018	2	2	7	10
	/	T51	정당/지지/유권자/태도/정치	.023	.019	.015	.029	3	18	20	5
		T76	스트레스/청소년/학교/부모/폭력	.022	.001	.014	.032	4	62	28	3
		T10	health/study/care/elderly/stroke	.022	.000	.001	.038	5	76	80	1
		T21	media/public/opinion/social/political	.020	.000	.000	.036	6	86	85	2
		T78	전화/유선/편향/전화조사/여론	.019	.016	.019	.020	10	27	14	9
		T49	여성/자녀/출산/돌봄/가족	.017	.002	.016	.021	13	58	19	7
		T13	행복/아동/주관적/소득/변수	.014	.000	.006	.021	21	80	60	8
	∨	T24	student/products/study/program/people	.016	.001	.001	.027	15	71	83	6
		T65	투표/선거/동원/투표율/지방선거	.012	.029	.002	.013	33	10	79	29
	∧	T67	모형/변수/추정/측정/모수	.019	.029	.024	.015	9	11	8	19
		T44	가중치/가중/표본/조정/보정	.015	.019	.028	.007	17	17	5	56
		T12	sample/sampling/survey/row/coverage	.013	.000	.032	.006	24	75	3	58
		T63	브랜드/지수/산업/주가/가동	.010	.010	.023	.003	49	41	9	77
		T16	감염/변수/텍스트/함수/판별	.010	.003	.030	.001	52	52	4	83
	\	T36	광고/소비자/제품/마케팅/시장	.020	.030	.034	.010	7	9	2	42
		T54	문항/척도/측정/응답/검사	.020	.037	.026	.012	8	3	6	33
		T70	인터넷/미디어/측정/변수/행동	.018	.037	.022	.012	11	6	10	31
		T86	추정/추정량/표본/분산/모집단	.018	.058	.020	.008	12	1	12	50
		T52	표본/농가/조사/모집단/조사구	.011	.031	.013	.005	40	8	35	66
		T42	인터넷/사용자/온라인/응답/사이트	.010	.031	.014	.002	51	7	27	81
		T28	지불/입찰/응답자/출발점/금액	.005	.037	.000	.000	78	4	84	85

주: P_a: 1999~2019 전 시기, P_1: 1999~2004, P_2: 2005~2012, P_3: 2013~2019,
—: 상위유지형, /: 상승형, ∨: 하강 후 상승형, ∧: 상승 후 하강형, \: 하강형.

Ⅴ. 결론

한국조사연구학회에서 지난 20년간 다루어진 연구논문의 내용을 전 분야로 개괄한 이후, 학문분야·방법론·연구대상·통계기법별로 각각을 살펴보았다. 그 각각을 비교하기 위하여, <표 9>에 주요 토픽의 상위유지·상승·하강 등의 추세를 요약하여 제시하였다. 전 분야에서 최고 순위 토픽인 'T47 조사'는 조사방법을 가리킨다. 'T58 데이터', 'T78 전화', 'T64 가구', 'T72 조사원' 등은 데이터 품질과 편향, 조사 관련 사항을 담은 토픽으로, 다양한 분류 기준에서 공통으로 상승하는 경향을 보인다. 이는 한국조사연구학회에서 출판·발표된 논문들이 연구대상이나 학문분야와 관계없이 충실히 조사방법을 충실히 다룸으로써 조사연구의 발전에 기여해 왔음을 보여준다.

<표 9> 주요 토픽별 유형 비교

(단위: 개)

		전체	T_h	T_c	T_h			T_c	
					—	/	∨	∧	\
T47	조사/표본/조사방법/응답/오차	13	9	4	8	1	0	2	2
T42	인터넷/사용자/온라인/응답/사이트	10	2	8	1	0	1	0	8
T16	감염/변수/텍스트/함수/판별	9	1	8	0	1	0	8	0
T74	인구/총조사/주택총조사/주택/주민등록	8	4	4	0	4	0	4	0
T46	출구조사/투표/예측/선거/오차	7	5	2	4	0	1	0	2
T70	인터넷/미디어/측정/변수/행동	7	2	5	1	1	0	2	3
T36	광고/소비자/제품/마케팅/시장	7	1	6	0	1	0	3	3
T12	sample/sampling/survey/row/coverage	7	1	6	0	0	1	4	2
T54	문항/척도/측정/응답/검사	7	0	7	0	0	0	2	5
T24	student/products/study/program/people	6	6	0	0	4	2	0	0
T76	스트레스/청소년/학교/부모/폭력	6	6	0	1	5	0	0	0
T58	데이터/품질/작성/기관/시스템	6	5	1	5	0	0	1	0
T71	여론/보도/기관/조사/선거	6	5	1	0	0	5	0	1
T78	전화/유선/편향/전화조사/여론	6	4	2	2	1	1	1	1

T23	응답/질문/응답자/서베이/면접조사표	6	3	3	1	2	0	2	1
T86	추정/추정량/표본/분산/모집단	6	0	6	0	0	0	0	6
T13	행복/아동/주관적/소득/변수	5	5	0	0	5	0	0	0
T21	media/public/opinion/social/political	5	5	0	0	5	0	0	0
T64	가구/가구주/표본/웨이브/주택	5	5	0	0	4	1	0	0
T67	모형/변수/추정/측정/모수	5	4	1	0	4	0	1	0
T72	조사원/조사/가구/응답자/접촉	5	4	1	0	3	1	1	0
T79	사업체/산업/업종/접근법/통계조사	5	2	3	0	0	2	0	3
T29	sample/variance/design/rotation/estimator	5	0	5	0	0	0	0	5
T6	survey/data/quality/response/unit	4	4	0	0	0	4	0	0
T10	health/study/care/elderly/stroke	4	4	0	0	4	0	0	0
T49	여성/자녀/출산/돌봄/가족	4	4	0	0	4	0	0	0
T51	정당/지지/유권자/태도/정치	4	4	0	0	3	1	0	0
T65	투표/선거/동원/투표율/지방선거	4	4	0	0	0	4	0	0
T80	예측/선거/여론/후보/선거여론조사	4	4	0	0	0	4	0	0
T82	직업/종사자/기관/초과/서비스	4	3	1	0	1	2	0	1
T20	data/pr/fit/analysis/Japanese	4	3	1	0	3	0	1	0
T33	기업/조직/사업장/제도/업무	4	2	2	2	0	0	0	2
T37	지표/문화/측정/지수/지표체계	4	2	2	0	2	0	1	1
T84	표본/조사구/가구/조사/모집단	4	1	3	0	0	1	0	3
T28	지불/입찰/응답자/출발점/금액	4	0	4	0	0	0	0	4
T44	가중치/가중/표본/조정/보정	4	0	4	0	0	0	4	0
T55	부패/공무원/행정/시민/측정	4	0	4	0	0	0	0	4
T25	level/life/factors/student/income	3	3	0	0	3	0	0	0
T43	응답/응답자/설문/문항/온라인	3	2	1	0	2	0	1	0
T83	취업/고용/일자리/노동/직업	3	2	1	0	2	0	0	1
T85	소득/가계/소비/자산/금융	3	2	1	0	0	2	0	1
T1	학교/학생/학원/학년/수업	3	1	2	0	1	0	1	1
T19	서비스/고객/만족/사원/모델	3	0	3	0	0	0	1	2
T50	멘토/여성/정서/네트워크/남성	3	0	3	0	0	0	3	0
T63	브랜드/지수/산업/주가/가동	3	0	3	0	0	0	3	0
T66	가격/주택/아파트/월세/유동인구	3	0	3	0	0	0	2	1
T73	음주/구강/우식/건강/방문	3	0	3	0	0	0	0	3

주: T_h: 활발히 진행되는 연구주제, T_c: 점차 사라지는 연구주제,
—: 상위유지형, /: 상승형, ∨: 하강 후 상승형, ∧: 상승 후 하강형, \: 하강형.

다양한 분류 기준에서 공통되게 상승하는 경향이 높았던 토픽인 ‘T24 학생’, ‘T76 스트레스’, ‘T13 행복’은 청소년의 건강과 교육 등의 내용으로, 다양한 방법론과 통계기법으로 분석되어 왔으며, 최근으로 올수록 더 많이 다루어지고 있다. 역시 다양한 분류 기준에서 공통으로 상승한 ‘T46 출구조사’, ‘T71 여론’, ‘T21 미디어’는 선거나 여론조사와 관련된 내용이고, ‘T67 모형’은 통계 분석에 관한 내용으로 조사연구의 초점이 무엇인가를 보여준다.

학문분야별로는, 조사방법론·통계학에서 조사에 관한 토픽이 꾸준히 증가하고 있다. 이는 조사연구에서 이 분야가 필수 주제임을 뜻한다. 기초사회과학과 응용사회과학에서는 다양한 연구주제들이 새롭게 출현하고 있다. 특히, 건강, 미디어, 정서 등은 최근으로 올수록 더욱 활발히 연구되고 있다.

방법론별로 주목할 만한 사항은 ‘통계적 분석법 적용’에 관한 토픽의 변동이다. 토픽의 분포 변화는 미디어, 건강, 정서 등 다양한 주제들이 더욱 적극적으로 연구되어 왔음을 보여준다. 이는 학문분야 변동에서 기초사회과학과 응용사회과학에서 발견된 변동 양상과 일치한다. 조사방법 자체를 다루는 연구보다는 통계 분석법의 적용 논문이 늘어난 것은, 학회의 초창기에는 조사방법에 대한 기초 연구가 많았지만 점차 고급통계기법 연구로 연구 관심이 옮겨가고 있음을 뜻한다. 질적 연구에서는 미디어, 광고, 인터넷 등 기존의 계량적 연구방법으로는 분석하기 어려운 비정형적 분야를 지속해서 다루어 온 것을 알 수 있다.

연구대상별 특성은 학문분야·방법론의 변동과 유사하다. 사회연구 논문들은 건강, 정서, 여성 등의 주제를 더욱 많이 다루고, 경제연구 논문은 직업, 취업 등을 주로 다룬다. 또한, 통계 일반을 다루는 연구보다는 사회현상 연구로 연구대상이 달라지는 모습을 보인다. 기초 방법론 연구보다는 사회와 경제 현상에 통계방법을 적용하는 응용 연구가 활발해지는 것으로 해석할 수 있다.

데이터분석기법별로는 초창기의 기초적인 통계분석에서 발전하여 점차 고급 통계기법이 적용되고 있음을 확인하였다. 데이터를 사용하지 않는 연구논문에서 다루는 주제는 모형, 조사, 모델, 조사원 등 조사방법에 관한 것이 주류를 이룬다. 기술통계분석을 통해 활발히 연구되어 온 분야는 교육 및 문화 등을 지표를 통해 살펴보는 경우가 많다. 추론통계분석은 다양한 대상에 적용한 연구가 시기에 따라 그 양상을 달리하며 활발히 연구됐다. 그중에서도 ‘T46 출구조사’ 등 선거 관련 주제는 모든 시기에서 상위를 기록하고 있다. 이는 국내 조사연구에서 선거 관련 연구의 위치를 보여준다.

지난 20년 동안, 한국조사연구학회를 통해서 출판·발표된 논문이 축적되면서, 학문분야·방법론·연구대상·데이터분석기법의 면에서도 발전을 거듭해오고 있다. 향후 한국조사연구학회에서 주력해야 할 두 과제를 제시하면서 이 논문을 마무리하기로 한다. 첫째, 질적연구에 대한 관심을 고취하고, 해당 분야 연구자들을 한국조사연구학회 회원으로 적극적으로 유치하여야 한다. 계량연구와 질적 연구의 연구대상이 다르지 않음을 확인한 이상, 다른 방법론을 사용하는 연구자들과의 교류를 활성화할 필요가 있다. 둘째, 융·복합 연구를 지금까지 충실히 수행해 왔고, 앞으로도 더욱 활발하게 추진하여야 한다. 여러 학문분야를 아우르는 연구와 조사연구학계와 조사연구업계 간의 협동 연구가 주는 장점을 살려야 한다. 특히, 최근 조사환경의 변화, 컴퓨터를 이용한 정량적·정성적 분석기법의 발전을 고려할 때, 융·복합 연구의 중요성은 더욱 드러난다.

참고문헌

구자숙·김은미·이현희. 2001. "≪계간여론조사≫에 실린 논문들의 성격과 경향." ≪조사연구≫ 2(2): 1-16.

김규성·이기재. 2001. "≪조사방법론≫에 게재된 논문들의 성격과 경향." ≪조사연구≫ 2(2): 17-26.

김태경·최회련·이홍철. 2016. "토픽모델링을 이용한 핀테크 기술 동향 분석." ≪한국산학기술학회 논문지≫ 17(11): 670-681.

남춘호. 2016. "일기자료 연구에서 토픽모델링 기법의 활용 가능성 검토." ≪비교문화연구≫ 22(1): 89-135.

박자현·송민. 2013. "토픽모델링을 활용한 국내 문헌정보학 연구동향 분석." ≪정보관리학회지≫ 30(1): 7-32.

박주섭·홍순구·김종원. 2017. "토픽모델링을 활용한 과학기술동향 및 예측에 관한 연구." ≪한국산업정보학회논문지≫ 22(4): 19-28.

박진균·김택윤·송민. 2017. "텍스트마이닝을 이용한 윤동주 연구의 개체계량학적 분석." ≪한국비블리아학회지≫ 28(1): 191-207.

변종석·설동훈 편. 2019.『한국조사연구학회 연구논문 목록 1999~2019』. 한국조사연구학회.

설동훈. 2018. "내용분석을 이용한 한국 사회학의 연구동향 연구." ≪KIPA조사포럼≫ 25: 22-29.

설동훈·고재훈·유승환. 2018. “한국사회학회와 사회학연구, 1964~2017년: 한국사회학회 발표 논문의 연구분야별 내용분석” ≪한국사회학≫ 52(1): 153-213.

설동훈·이유경·변종석·강정한. 2019. “한국조사연구학회 학술활동 분석, 1999~2019.” ≪조사연구≫ 20(4): 85-108.

송혜지·박경수·정혜은·송민. 2013. “텍스트마이닝 기법을 활용한 한국의 경제연구 동향 분석.” ≪한국정보관리학회 학술대회 논문집≫ 2013·8: 47-50.

이재연. 2016. “‘생활’과 ‘태도’: 기계가 읽은 ≪개벽≫과 ≪조선문단≫의 작품 비평어와 비평가.” ≪개념과 소통≫ 18: 5-52.

홍두승·설동훈. 2012. ≪사회조사분석≫. 다산출판사.

AlSumait, Loulwah, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. “Topic Significance Ranking of LDA Generative Models.” pp. 67-82 in *Machine Learning and Knowledge Discovery in Databases*, Part I, Wray Buntine, Marko Grobelnik, Dunja Mladenić, and John Shawe-Taylor(eds). Berlin: Springer-Verlag.

Amado, Alexandra, Paulo Cortez, Paulo Rita, and Sérgio Moro. 2018. “Research Trends on Big Data in Marketing: A Text Mining and Topic Modeling Based Literature Analysis.” *European Research on Management and Business Economics* 24(1): 1-7.

Blei, David M. 2012. “Probabilistic Topic Models: Surveying a Suite of Algorithms That Offer a Solution to Managing Large Document Archives.” *Communications of the ACM* 55(4): 77-84.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3(1): 993-1022.

Edwards, Lisa M. and Jennifer Teramoto Pedrotti. 2008. “A Content and Methodological Review of Articles Concerning Multiracial Issues in Six Major Counseling Journals.” *Journal of Counseling Psychology* 55(3): 411-418.

Evans, James A. and Pedro Aceves. 2016. “Machine Translation: Mining Text for Social Theory.” *Annual Review of Sociology* 42: 21-50.

Griffiths, Thomas L. and Mark Steyvers. 2004. “Finding Scientific Topics.” *Proceedings of the National Academy of Sciences* 101(Supplement 1): 5228-5235.

Hong, Doo-Seung. 2019. “Sociology in Korea: Past and Present.” *Sociological Problems* 51(2): 401-409.

Hutchinson, Susan R. and Cheryl D. Lovell. 2004. “A Review of Methodological Characteristics of Research Published in Key Journals in Higher Education: Implications for Graduate Research Training.” *Research in Higher Education*

45(4): 383-403.

Kim, Su Yeon, Sung Jeon Song, and Min Song. 2015. “Investigation of Topic Trends in Computer and Information Science by Text Mining Techniques: From the Perspective of Conferences in DBLP.” *Journal of the Korean Society for Information Management* 32(1): 135-152.

Long, Haiying. 2014. “An Empirical Review of Research Methodologies and Methods in Creativity Studies (2003–2012).” *Creativity Research Journal* 26(4): 427–438.

Neall, Annabelle M. and Michelle R. Tuckey. 2014. “A Methodological Review of Research on the Antecedents and Consequences of Workplace Harassment.” *Journal of Occupational and Organizational Psychology* 87(2): 225-257.

Prabhakaran, Vinodkumar, William L. Hamilton, Dan McFarland, and Dan Jurafsky. 2016. “Predicting the Rise and Fall of Scientific Topics from Trends in Their Rhetorical Framing.” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol. 1. Berlin, Germany. August 7-12, 2016. pp. 1170-1180.

Randolph, Justus J., Andrea E. Griffin, Samara R. Zeiger, Kristina N. Falbe, Noreen A. Freeman, Bridget E. Taylor, Amy F. Westbrook, Cheryl C. Lico, Cristy N. Starling, Nakiesha M. Sprull, Carolyn Holt, Kristie Smith, and McAnespie. 2011. “A Methodological Review of the Articles Published in Georgia Educational Researcher from 2003-2010.” *Georgia Educational Researcher* 10(1): 1-18.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2014. “stm: R Package for Structural Topic Models.” Unpublished Working Paper. Retrieved from https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf.

Song, Min and Su Yeon Kim. 2013. “Detecting the Knowledge Structure of Bioinformatics by Mining Full-Text Collections.” *Scientometrics* 96(1): 183-201.

Sun, Lijun, and Yafeng Yin. 2017. “Discovering Themes and Trends in Transportation Research Using Topic Modeling.” *Transportation Research, Part C: Emerging Technologies* 77(4): 49-66.

Wang, Xuerui and Andrew McCallum. 2006. “Topics Over Time: A Non-Markov Continuous-Time Model of Topical Trends.” *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Philadelphia, PA, USA. August 20–23, 2006. pp. 424-433.

<접수 2020.02.18.; 수정 2020.02.18.; 게재확정 2020.02.19>

Content Analysis of Papers Published and Presented at the Korean Association for Survey Research, 1999-2019: Continuation and Change in Disciplines, Methodologies, Research Objects, and Techniques of Data Analysis

Dong-Hoon Seol
Jae-Hun Ko
Seung-Hwan Yoo
Kee Jae Lee

This paper reviews research outputs in the Korean Association for Survey Research (KASR) 1999-2019, and identifies the topic trend in survey research in Korea. To this end, we collected full text of all the research papers published/presented in its official journal, conference proceedings, and tutorial lectures of the KASR. We conducted the analyses in two steps. First, using classical quantitative content analysis, we catorized the total of 1,190 papers in four dimensions: disciplines, methodologies, research objects, and techniques of data analysis. Second, we also applied computer-assisted content analysis-Structural Topic Model (STM)-to the text corpus of 1,145 papers, which is a large and structured set of texts. The results of the STM analysis revealed 87 topics in 1999-2019. We conducted Spearman rank-order correlation analysis using data of the weights of the 87 topics by each year, and divided into three distinctive periods: 1999-2004, 2005-2012, and 2013-2019. We presented each topic with five key words by the three periods in the four dimensions. We discussed continuation and change in the focus topics at each dimension and identified the trajectory of survey research in Korea. We identified the growth, continuation, and decline patterns of each topic: (1) the diversity of disciplines has increased, (2) in terms of methodology, more papers have been published/ presented applying statistical analysis rather than research on the method itself, (3) in terms of research objects, more research has been conducted on society than on statistics itself, and (4) in terms of data analysis techniques, advanced statistical techniques have been more increased.

Key words: content analysis, text mining, topic modeling, Structural Topic Model(STM), Korean Association for Survey Research(KASR)