J. Inf. Technol. Appl. Manag. 24(4): 197~214, December 2017
https://doi.org/10.21219/jitam.2017.24.4.197

ISSN 1598-6284 (Print)
ISSN 2508-1209 (Online)

원천 데이터 품질이 빅데이터 분석결과의 유용성과 활용도에 미치는 영향

박소현* · 이국희** · 이아연***

An Empirical Study on the Effects of Source Data Quality on the Usefulness and Utilization of Big Data Analytics Results

Sohyun Park* · Kukhie Lee** · Ayeon Lee***

Abstract

This study sheds light on the source data quality in big data systems. Previous studies about big data success have called for future research and further examination of the quality factors and the importance of source data. This study extracted the quality factors of source data from the user's viewpoint and empirically tested the effects of source data quality on the usefulness and utilization of big data analytics results. Based on the previous researches and focus group evaluation, four quality factors have been established such as accuracy, completeness, timeliness and consistency. After setting up 11 hypotheses on how the quality of the source data contributes to the usefulness, utilization, and ongoing use of the big data analytics results, e-mail survey was conducted at a level of independent department using big data in domestic firms. The results of the hypothetical review identified the characteristics and impact of the source data quality in the big data systems and drew some meaningful findings about big data characteristics.

Keywords : Big Data, Source Data, Data Quality, Analytics Results

Received : 2017. 11. 01. Revised : 2017. 12. 26. Final Acceptance : 2017. 12. 27.

※ This paper was written as part of Konkuk University's research support program for its faculty on sabbatical leave in 2016.

* Primary Author, Invited Professor, Konkuk University, e-mail : shpark@konkuk.ac.kr
** Corresponding Author, Professor, Dept. of Business Administration, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul, 05029, Korea, Tel : +82-2-450-3631, e-mail : kukhie@konkuk.ac.kr
*** Hanyang University, e-mail : ayeon.lee33@gmail.com

1. 연구 필요성 및 목적

빅데이터는 치열한 경쟁환경에서 기업 경쟁력을 확보할 수 있는 내부역량이며, 앞으로 4차 산업혁명을 선도하는 핵심 기술 중 하나로 거론되고 있다[Schwab, 2016]. 현재 많은 공공조직 및 민간기업들이 혁신적 효과에 대한 기대, 시대적 동향과의 부합, 경쟁사 도입에 의한 압박 등을 이유로 빅데이터를 도입하였거나 계획하고 있다. 현재까지 발표된 여러 조사보고서에 의하면 향후 빅데이터 시장은 매년 20% 이상의 성장률을 보일 것으로 전망되고 있다[Smith, 2015].

그러나 빅데이터에 대한 낙관적 전망 이면에는 재무적 성과나 현업 활용에 대한 의문이 지속적으로 제기되고 있다. Fortune 500 기업 중 80% 이상이 실망스러운 성과를 거두었으며, 18개국 대상 설문조사에서 빅데이터 성공률이 28%에 불과하다는 수치가 발표되었다[가회광 등, 2014]. 대한상공회의소가 국내 기업 500개사를 대상으로 빅데이터 활용 현황을 조사한 결과, 응답 기업의 81.6%가 제대로 활용하지 못한다고 응답했다[김차영 등, 2014]. 국내 기업의 빅데이터 수준은 현재 성공 가능성이 있는 사례를 발굴하는 단계에 머물고 있으며, 간헐적으로 나타나는 실제 성공사례가 타 기업이나 업종에 전파되기 위해서는 상당 기간 원천 데이터 축적 및 맞춤화 과정을 거쳐야 한다는 의견도 있다[차승언 등, 2015].

따라서 빅데이터 분야에서 최근에 부상하는 한가지 연구 주제는 리스크와 도전과제(risk and challenges)이다. 빅데이터 성공에 영향을 미치는 요인으로 분석 플랫폼, 분석 기법, 분석가 역량, 변화지향적 조직 문화, 최고경영층 등이 논의되고 있으며, 이 중 가장 많이 거론되고 있는 리스크가 원천 데이터 품질이다. 한국정보화진흥원 보고서[2012]는 성공적인 빅데이터 활용을 위한 3대 요인 중 하나로 신뢰할 수 있는 데이터 자원의 확보를 꼽고 있고, Halaweh et al.[2015] 연구는 원천 데이터 품질을 빅데이터 구축 성공의 선행요인으로 설명하였다. 장명수[2017]는 저품질 데이터로 인한 의사결정 오류와 경제적 손실 사례를 열거하며 데이터 품질이 보장되지 않은 빅데이터 분석은 무의미하다고 설명하였으며, 조완섭[2017]은 고품질 데이터 확보 및 관리를 목적으로 하는 빅데이터 거버넌스(governance)를 강조하고 있다.

그러나 이러한 중요성에도 불구하고 원천 데이터 품질에 관한 심층 분석과 실증 연구가 아직까지 충분히 이루어지지 않은 실정이다. 우선 빅데이터 역사가 오래 되지 않아서 관련 연구 수가 많지 않고, 그나마 탐색적, 개념적 모델 제시에 머무는 연구가 대부분이다. 품질 관련 연구 동향에서 특히 문제가 되는 것은 원천 데이터 품질체계가 연구마다 상이하다는 점이다. 데이터 품질은 개념적으로 사용 적합성(fitness for use)인데, 이 적합성을 구성하는 항목이 연구마다 다르다. 품질요소간 비중이 다른 정도가 아니라 현저하게 상이한 품질 요소와 측정항목을 채택함으로써 일종의 혼란과 시행착오가 발생하고 있다. 예컨대 원천 데이터는 산출된(output) 정보가 아니라 빅데이터 분석 대상이며 원재료적(input material) 용도임을 고려하지 않은 채 정보시스템 성공모델[DeLone and McLean, 2003]의 성공지표 중 하나인 정보 품질(information quality) 개념을 그대로 사용하는 경우도 있다. 빅데이터 환경에서 원천 데이터 품질이 무엇인지, 저품질 데이터로 인하여 어떤 문제가 발생하는지, 데이터 품질관리 방법이 무엇인지를 이해하기 위해서는 원천 데이터 품질에 관한 이론적 분석과 실증적 검증이 선행되어야 하며, 본 연구의 필요성을 여기서 찾을 수 있다.

이 연구의 목적은 빅데이터 성공에 미치는 원천 데이터 품질의 영향을 실증 분석하는데 있다.

우선 원천 데이터 품질 개념을 정의하고, 빅데이터 환경에서 범용적으로 사용될 수 있는 품질체계를 정립한다. 그리고 원천 데이터 품질이 빅데이터 성공 경로에서 작용하는 제반 영향에 대한 가설을 제시하고, 실증 조사에 의하여 그 타당성을 검증한다. 이 연구는 선행 연구의 다양한 시각을 하나의 틀로 조명하는 시도이며, 다음 3가지 의문(research question)에 접근하는 노력으로 이해될 수 있다. (1) 원천 데이터 품질은 무엇인가? 어떤 품질요소로 구성되어 있는가? (2) 원천 데이터 품질이 빅데이터 성공에 어떤 영향을 미치는가? (3) 빅데이터 환경에서 특히 중요하게 작용하는 품질요소는 무엇인가?

2. 관련 연구 동향

2.1 원천 데이터 품질요소

데이터 품질은 사용자 요구사항의 충족 정도로 측정된다. 사용자 요구사항이 다양하므로 품질은 여러 복합적 요소(attributes)를 지니는 다차원적 개념으로 정의된다. 선행 연구는 연구 목적이나 적용 분야의 특성에 따라 데이터 품질요소를 달리 구성하고 있으며, 그 중에서 빅데이터 환경의 원천 데이터 품질요소와 관련된 5편의 연구를 <Table 1>의 [A]열부터 [E]열까지 살펴볼 수 있다.

<Table1>의 첫 번째 열 [A]에 나타난 MIT의 데이터 품질체계[Wang and Strong, 1996]는 가장 널리 알려져 있는 데이터 품질 모델이다. 이 품질체계를 구성하는 4개 카테고리는 (1)데이터에 내재되어 있는 본질적 품질, (2)사용자와 적용업무 상황에 의하여 결정되는 상황적 품질, (3)데이터 표현적 품질, (4)검색성 관련 품질이다. 그리고 각 카테고리는 다시 여러 품질차원으로 세분화되어 총 15개 차원으로 이루어진다. 이 품질체계의 특징은 특정 분야에 의존하지 않는 범용성을 지향한다는 점, 사용자 대상 중요도 실증조사에 의하여 도출되었다는 점, 정량적 척도

〈Table 1〉 Data Quality Items in Major Prior Studies

[A] 15 dimensions by Wang and Strong [1996]		[B] 6 quality items in Kdata [2010]	[C] 4 big data quality items in Gartner [2011]	[D] 17 quality items in Moges et al. [2013]	[E] 5 public data quality items in Park et al.[2015]
4 category	15 dimensions				
(1) Intrinsic	• accuracy • objectivity • reliability • reputation	(1) accuracy	(1) accuracy	• accuracy • objectivity • reputation	(1) accuracy (2) domain effectiveness
(2) Contextual	• relevancy • value added • completeness • amount • currency	(2) usefulness (requirements fulfillment)	(2) completeness (3) time effectiveness	• completeness • amount • value added • relevancy	(3) completeness (4) timeliness
(3) Representational	• interpretability • understandability • format consistency • conciseness	(3) consistency (data redundancy)	(4) consistency	• interpretability • understandability • format consistency • conciseness • actionability	(5) consistency (data redundancy)
(4) Accessibility	• accessability • security	(4) accessability (5) response time (6) security		• accessability • security • data alignment • traceability	

를 사용할 수 없는 데이터 신뢰성, 평판 등을 적극 수용했다는 점을 들 수 있다. 오랫동안 이론적 품질체계의 효시로 인정받아 왔으며, 여러 후행 연구와 실무조사에 의하여 채택되고 있다[Lee et al., 2002; Moges et al., 2013]. 그러나 이 품질체계를 원천 데이터에 그대로 적용하기에는 일부 적합하지 않는 면이 존재한다. 네 번째 품질 차원인 검색성은 원천 데이터 자체의 품질이 아니라 그것을 편리하고(timely convenient) 안전하게(secure) 이용할 수 있게 하는 소프트웨어 기능적 성격이 강하기 때문이다. 따라서 후술하는 원천 데이터나 공공 데이터 관련 연구는 검색성을 품질요소로 채택하지 않고 있다.

한국데이터진흥원[2010]은 공공부문 및 민간기업을 대상으로 주기적으로 실시한 데이터 품질관리 성숙수준 조사에서 <Table 1>의 [B]에 나타난 6개 항목으로 데이터 품질을 조사하였다. 6개 항목을 살펴보면 (1)정확성은 데이터 값의 오류 없음, (2)유용성은 사용자 요구사항의 충족도, (3)일관성은 데이터 중복 없음, (4)접근성은 사용자의 접근 용이성 및 데이터 공유 수준, (5)적시성은 시스템의 응답속도 및 처리속도, (6)보안은 사용 권한, 복구, 백업, 보안감사 등으로 정의되고 있다. 이 품질체계는 기업 내부 데이터베이스의 산출물 정보 품질에 초점을 맞추고 있으며, 기업 외부의 대용량, 비정형적, 실시간 생성 등의 특징을 지닌 원천 데이터 품질에는 적합하지 않는 부분이 있다.

빅데이터 환경의 원천 데이터 품질요소는 Gartner[2011]가 최초로 공식화하였다. 빅데이터의 특성인 대용량성, IoT 센서 단위의 정밀성, 소유자의 불분명함, 외부 데이터의 비정형성 등을 반영하여 <Table 1>의 [C]에 나타난 것처럼 4개 품질요소를 제시하였다. (1)정확성은 데이터 용도에 따라 요구 수준이 달라질 수 있으며(예 : 고객의 이동 위치 데이터와 신용카드 결제 데이터의 정확도 차이), 빅데이터 분석 과정에서 많은 데이터 값의 집계 정확성, 업무규칙 준수성, 시간적 선후관계 정확성 등을 중시한다. (2)완전성은 현실세계에 존재하는 모든 데이터의 100% 확보가 아니라 주어진 비용이 허용하는 범위 내에서 신뢰할 수 있는 데이터의 안정적 확보를 강조하고 있다. (3)시간적 유효성은 휘발성이 강한 빅데이터의 존속 여부를 결정하는 기준이다. 예컨대 웹로그, 트윗, 교통안내 위치 데이터는 길어야 몇 시간, 짧으면 몇 분 동안만 유효하다. 최신 데이터를 신속하게 수집하고, 유효 데이터를 점검하고, 노후 데이터를 제거하는 노력을 지속적으로 요구하는 품질요소이다. (4)일관성은 다양한 데이터 수집 채널과 필터링 시스템으로 인하여 동일 데이터가 다른 의미로 사용되는 경우가 빈번하므로 의미적(semantics) 일관성을 중시한다. 가트너는 4개 품질요소에 기반하여 기업 내부 데이터뿐만 아니라 소유하지 않은 외부 데이터의 신뢰성을 담보할 수 있는 품질관리 전략을 제안하고 있다.

네번째 열[D]에 나타난 Moges et al.[2013] 연구는 [A]에 나타난 품질항목들의 상대적 중요도를 분석하였다. 여러 국가의 민간기업을 대상으로 실시한 조사에서 산업별로 품질항목의 중요도가 달리 나타나고 있음을 확인하였다. 예컨대 금융기관에서는 데이터 정확성을, 유통업체는 사용자에게 필요한 데이터의 충족 여부를 가장 중요하게 인식하고 있었다. 그리고 여러 품질항목들은 상호배타적이 아니라 상당한 trade-off 관계를 맺고 있음을 발견하였다. 예컨대 정확성 품질이 높아질수록, 완전성 품질이 그만큼 낮아지는 것이다. 특히 이 연구가 새롭게 제시한 품질항목 3개는 빅데이터 품질과 연관되어 주목을 끌고 있다 : (1)별도 가공이나 편집 없이 데이터를 즉각 사용할 수 있는 즉각 사용성(actionability), (2)서로 다른 출처의 데이터를 함께 묶어서 사용

할 수 있는 데이터 연계성(alignment), 그리고 (3)데이터 출처, 시점, 수집방법을 알 수 있는 추적 가능성(traceability).

박고은 등[2015] 연구는 사회적 빅데이터 자원의 큰 축을 이루고 있는 교통, 기상, 보건의료, 산업고용 등 공공부문 개방 데이터의 품질을 공공성, 활용성, 신뢰성, 적합성 4개 영역으로 구분하고, 그 중에서 데이터 자체의 품질에 해당하는 적합성(suitability) 영역에서 [E]열에 나타난 5개 품질항목을 제안하였다. (1)정확성 항목은 데이터 값이 사실과 일치하는지, 문법 규칙을 준수하는지를 의미하고, (2)유효성(effectiveness) 항목은 값의 범위, 형식, 도메인 충족 여부로 설명하고 있다. 그러나 이 유효성 항목은 의미가 추상적이고, 내용적으로는 정확성과 중복된다는 비판이 제기될 수 있다. (3)완전성 항목은 필수 항목의 누락이 없는지, 사용 목적에 맞게끔 충분한 양을 확보하는지로 정의되고, (4)적시성(timeliness) 항목은 주로 데이터 나이의 적합성에 초점을 맞추고 있다. (5)일관성은 전통적 데이터베이스 품질 기준인 데이터 중복으로 설명하고 있다. 이러한 품질항목은 공공 데이터뿐만 아니라 민간 데이터에도 그대로 적용될 수 있을 것으로 판단된다.

2.2 원천 데이터 품질의 중요성

입력 데이터 품질이 낮으면 산출 정보의 품질도 낮다는 GIGO(Garbage in, Garbage out)는 데이터 품질의 중요성을 단적으로 대변한다. 정보시스템 성공모델을 제시한 DeLone과 McLean은 진정한 가치를 창출하는 것은 정보시스템의 기술적 품질이 아니라 정보 자체의 품질임을 강조하였다[Petter et al., 2008]. 데이터 품질은 업무 프로세스와 의사결정 효율성 향상 외에도 이제는 전략적 경쟁우위 요인으로 인식되고 있다[Baskarada, 2011].

빅데이터 분야에서도 원천 데이터 품질의 중요성은 널리 강조되고 있다. 한국정보화진흥원[2012]은 빅데이터 성공을 위한 3대 요인으로서 플랫폼 기술, 분석 역량, 그리고 신뢰할 수 있는(reliable) 데이터 자원을 꼽고 있으며 이러한 입장은 여러 후속 연구에서도 수용되고 있다[이서구, 2015; 차승은, 2015]. 노성여[2016] 연구는 빅데이터 분석에 활용되는 원천 데이터의 보유량이 많을수록(원천 데이터의 충분성 품질이 높을수록) 기업 생산활동 역동성과 매출이 증가한다는 실증 조사 결과를 발표하였다. 한편 김선호 등[2015] 연구는 우리나라의 경제사회적 빅데이터 기반을 형성하는 공공부문 데이터의 개방 및 활용이 미국, 독일, 영국 등 선진국에 비하여 낮은 이유를 낮은 데이터 품질에서 찾았다. 공개 데이터의 활용가치가 낮고 실제 사용률이 12% 수준에 그치는 원인을 여러 출처 사이의 낮은 데이터 호환성과 프라이버시 침해 우려로 인한 데이터 누락에서 찾는 연구도 있다[조완섭, 2017]. 시장조사업체 가트너는 2016년 매직 쿼드런트(Magic Quadrant) 보고서에서 앞으로 원천 데이터 품질의 중요성이 더욱 커질 것으로 전망하고 있다[전자신문 2017. 5. 10].

그러나 빅데이터 분석의 특성상 원천 데이터 품질이 큰 문제가 되지 않는다는 반론도 만만치 않다. 현재 기업이 확보할 수 있는 원천 데이터의 대부분은 상당한 품질 문제를 지니고 있지만 빅데이터 분석과 활용에는 큰 영향을 미치지 않는다는 주장이다. Smith[2015] 연구는 빅데이터 분석은 원시 데이터의 100% 정확성을 요구하는 회계분석과는 다르다고 말한다. 빅데이터 분석가는 원천 데이터가 완전하지 않고, 각종 노이즈로 인하여 정확하지도 않다는 것을 인정하고 있다고 설명한다. 그러므로 작은 데이터 실험에서 찾은 상관관계 패턴을 보다 큰 규모의 데이터로 점진적으로 확장하는 과정에서 다양한 통계 기법을

적용함으로써 결함 있는 원천 데이터에서도 유용한 정보를 찾을 수 있다고 주장한다. Jagadish et al.[2014] 연구는 데이터 노이즈가 많더라도 빅데이터 분석은 일반적인 표본 분석보다 더 유용할 수 있다고 주장한다. 빅데이터의 빈도 패턴과 상관관계에 대한 통계치는 표본 편차 문제를 극복할 수 있으며, 숨겨져 있던 데이터 패턴을 충분히 파악할 수 있다는 것이다. Bottles et al.[2014] 연구는 도메인 지식과 문제해결 방법을 갖춘 전문가가 데이터 분석결과를 보고 합당한 의문을 가지고 통찰력 있는 해석을 함으로써 원천 데이터 품질 문제를 극복할 수 있다고 보고 있다. 즉, 분석결과에 대한 기대수준만 조절한다면 원천 데이터 품질 문제는 대처 가능하다는 것이다. 이러한 반론은 빅데이터 목적의 특성, 분석결과의 용도, 분석 역량과 기법에 대한 믿음, 품질관리 비용 부담 등을 명분으로 앞으로도 계속 제기될 전망이다.

2.3 빅데이터 성공지표

빅데이터 성공을 의사결정 지원, 생산성 향상, 비즈니스 기회 발견 등 전략적 차원에서 찾는 연구도 있지만[가회광 등, 2014], 측정성을 고려해야 하는 실증 연구에서 성공을 의미하는 지표로 채택하고 있는 것은 빅데이터 분석결과(analytics results)의 유용성[김승현 등, 2015], 활용도[박소현 등, 2016], 지속적 활용의사[김정선 등, 2014] 등이다.

Davis[1989]의 기술수용모델(TAM)에서 제시된 유용성 개념은 수많은 연구에서 성공 또는 성공으로 연결되는 매개 변수로 사용되어 왔다. 빅데이터의 경우도 성공이 일차적으로 가시화되는 것은 유용한 분석결과의 도출 시점이다. 그러나 실제 빅데이터 현장에서 관찰할 수 있듯이 유용한 분석결과를 도출하는 것은 결코 쉬운 일이 아니다. 빅데이터 분석에서 찾은 아이디어의 대부분은 이미 과거에 검토되었거나, 타 기업에 존재하고 있거나, 기존 의사결정 기준과 차이가 없거나, 우리 고객이나 당면 문제와는 관련이 없거나, 아이디어를 실현하기에는 비용을 감당할 수 없는 경우에 해당한다. 따라서 현재 빅데이터 분석기법으로 도출하는 상관관계 패턴형 정보가 전문가 경험과 직관에서 얻어진 통찰보다 우수할 수 없다는 비판도 있다[Jagadish et al., 2014]. 빅데이터 분석결과의 유용성은 여러 연구에서 중요한 변수로 논의되고 있는 중이다.

활용도는 빅데이터 성공 경로에서 중요한 변곡점으로 작용한다. 빅데이터 분석결과가 유용하다고 인식되더라도 실제 의사결정이나 현업 프로세스에 활용되지 못하고 사장되는 경우가 많기 때문이다. 활용도가 낮은 이유는 조직구성원의 무관심, 태만, 조용한 반발에서 찾을 수 있다. 대부분 조직 구성원들은 아무리 유용한 아이디어라 할지라도 기존 관행과 다르다는 이유만으로 일단 부정적 자세를 취하며, 조금이라도 미흡한 점을 찾아서 그것을 근거로 활용을 미루는 사례가 많다[박소현 등, 2016]. 특히 원천 데이터 품질이 낮다면 빅데이터 분석결과의 활용을 반대하는 강력한 명분이 될 수 있다. 이외에도 유용한 분석결과가 제대로 활용되지 못하는 이유를 조직문화의 폐쇄성, 직관 의존형 의사결정 관행, 변화를 수용하지 못하는 업무체계의 경직성에서 찾는 시각도 있다[Ross et al., 2013]. 이선우 등[2014]은 빅데이터 도입을 위한 통합모형 연구에서 빅데이터 활용도를 최종 종속변수로 사용하였으며, 빅데이터 성능과 분석결과 유용성이 빅데이터 활용도에 긍정적 영향을 미친다고 설명하고 있다.

빅데이터를 앞으로 계속 활용하겠다는 의사는 일종의 충성도 개념이며 빅데이터와 유사한 혁신기술의 성공 여부를 나타내는 변수로 채택되고

있다. Benlian et al.[2011] 연구는 혁신 기술의 유용성 인식도가 지속적 활용의사(continuance intention)에 긍정적 영향을 미친다고 보았다. 그리고 김정선[2014] 연구는 빅데이터 수용모델을 제시하면서 빅데이터 유용성이 수용의도와 지속적 활용의사에 미치는 영향을 실증적으로 분석하였다. 기술 확산을 중시하는 관점에서는 지속적 활용의사를 중요한 성공지표로 보고 있는 것이다.

3. 연구 모형 및 가설

3.1 연구 모형

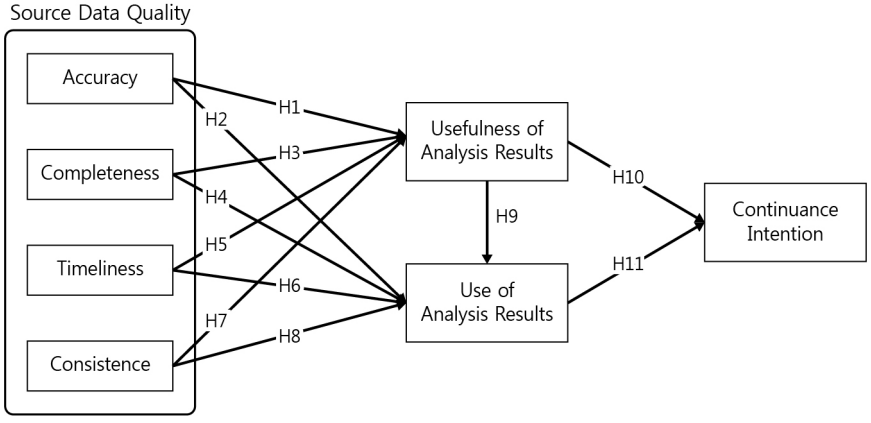
본 연구의 목적은 원천 데이터 품질이 빅데이터 성공에 미치는 영향을 분석하는데 있다. 데이터 품질은 다차원적 개념이므로 그 영향을 주요 품질요소 별로 구분할 필요가 있다. 본 연구는(4장의 연구 변수에서 후술하듯이) 선행 연구와 전문가 검토에 의하여 정확성, 완전성, 적시성, 일관성 4개 품질요소를 도출하였다. 이러한 4개 원천 데이터 품질요소가 빅데이터 성공 경로에서 나타나는 주요 변수인 분석결과 유용성, 활용도, 지속적 활용의사에 미치는 영향을 검증하기 위한 연구 모형을 <Figure 1>과 같이 설정하였다.

3.2 연구 가설

3.2.1 원천 데이터 4개 품질요소에 관한 가설

현재까지 실증 분석이 미흡하거나, 여러 연구결과가 상충하여 논란이 되는 쟁점은 다음 3가지로 압축할 수 있다. (1) 원천 데이터 품질이 빅데이터 분석결과 유용성과 활용도에 어떤 영향을 미치는가? (2) 어떤 품질요소가 영향을 미치는가? 특별히 중요하게 작용하거나 혹은 의미있는 영향을 미치지 못하는 요소가 있는가? (3) 전통적인 데이터 품질에 비하여 빅데이터 환경의 원천 데이터 품질의 특징은 무엇인가? 이러한 쟁점에 접근하기 위하여 본 연구는 4개 품질요소마다 빅데이터 분석결과의 유용성과 활용도에 관한 가설을 2개씩 설정하였다.

H1 : 원천 데이터의 정확성 품질은 분석결과 유용성에 긍정적 영향을 미칠 것이다.
H2 : 원천 데이터의 정확성 품질은 분석결과 활용도에 긍정적 영향을 미칠 것이다.
H3 : 원천 데이터의 완전성 품질은 분석결과 유용성에 긍정적 영향을 미칠 것이다.
H4 : 원천 데이터의 완전성 품질은 분석결과 유용성에 긍정적 영향을 미칠 것이다.
H5 : 원천 데이터의 적시성 품질은 분석결과 유



〈Figure 1〉 Research Model

용성에 긍정적 영향을 미칠 것이다.

H6 : 원천 데이터의 적시성 품질은 분석결과 유용성에 긍정적 영향을 미칠 것이다.

H7 : 원천 데이터의 일관성 품질은 분석결과 유용성에 긍정적 영향을 미칠 것이다.

H8 : 원천 데이터의 일관성 품질은 분석결과 유용성에 긍정적 영향을 미칠 것이다.

3.2.2 분석결과 유용성과 활용도 사이의 상관관계에 관한 가설

일반적으로 산출물(output)과 활용(outcome) 사이에는 정(+)의 상관관계가 존재한다. 정보시스템 산출 정보의 품질이 유용할수록 이용도가 높아진다는 논리는 여러 실증연구에 의하여 입증되었다[Petter et al., 2008]. 기술수용모델도 유용성(usefulness)과 활용도(use) 사이의 상관관계를 전제하고 있다[Davis et al., 1989]. 수많은 선행 연구들이 유용성과 활용도 사이의 상관관계를 입증하였으며 이에 대한 반론은 없다고 보아도 무방하다. 이러한 논리가 빅데이터에도 적용될 것이라 판단되며, 따라서 분석결과의 유용성과 활용도 사이의 긍정적 상관관계에 관한 가설을 설정하였다.

H9 : 분석결과 유용성은 분석결과 활용도에 긍정적 영향을 미칠 것이다.

3.2.3 지속적 활용의사에 관한 가설

본 연구는 빅데이터 성공을 나타내는 최종 변수로 지속적 활용의사를 채택하였다. 빅데이터 최종 성공은 빅데이터 분석결과가 실제 현업 프로세스나 의사결정에 얼마나 지속적으로 활용되는지에 따라 결정되기 때문이다. 정보시스템의 유용성, 활용도, 지속적 활용의사 사이의 정(+)의 상관관계는 수많은 선행 연구에 의하여 입증되었으며, 빅데이터 분야에서도 빅데이터 이용(use)과 지속적 이용의사(continuance intention) 사이의 관계 분석을 시도한 연구도 있다[김정선 등, 2014]. 따라서 본 연구는 빅데이터 성공 경로에서 세 변수 사이의 긍정적 상관관계가 존재할 것으로 판단하며, 다음 2개의 가설을 설정하였다.

H10 : 분석결과 유용성은 빅데이터의 지속적 활용의사에 긍정적 영향을 미칠 것이다.

H11 : 분석결과 활용도는 빅데이터의 지속적 활용의사에 긍정적 영향을 미칠 것이다.

4. 연구 방법

4.1 연구 변수

원천 데이터 품질요소와 측정항목을 도출하기 위하여 문헌조사, 전문가 검토, 사용자 대상 예비조사를 차례로 수행하였다. 국내외 관련 연구 및 기업실무자료 30여 편을 조사하여 다양한 품질항목들을 중복 배제하며 수집하였다. IEEE 1061의 4가지 기준인 (1) 상관관계성(correlation) (2) 실무적 유용성 및 측정 용이성(practical and computable) (3) 값의 일관성(consistency) (4) 품질의 높고 낮음을 차별화할 수 있는 변별력(discriminative power)으로 평가하여 일차적으로 총 28개의 품질항목을 도출하였다.

도출된 항목들의 내용 타당성(content validity)을 검증하고 분류체계를 구축하기 위하여 포커스 그룹 워크샵을 실시하였다. 기업 빅데이터 분석가 2인, 데이터베이스 전문가 1인, 해당 분야 교수 2인으로 구성된 포커스 그룹은 일차 도출된 품질요소와 측정항목들이 빅데이터 분석가의 다양한 요구사항들을 얼마나 충실하게 대표하는지를 평가하였다. 평가 과정에서 유사 항목들을 통합하고, 응답속도, 업무처리속도 등 원천 데이터 품질에 해당되지 않는 항목은 배제하였다. 상호

〈Table 2〉 Measures of Research Variables

Research Variables		Items	Manipulational Definition	Prior Studies
Source Data Quality	(1) Accuracy	accuracy of value	No error values during data collection	Gartner[2011] Jagadish et al.[2014] Moges et al.[2013] Nicolaou et al.[2013] Petter et al.[2012] Wang et al.[1996] Park et al.[2015] Kdata[2010]
		reliability	Consonance with the facts	
		objectivity	Objective record without subjective prejudice	
		source clarity	Who entered the data when	
	(2) Completeness	no missing value	Required data items are not missing	
		data sufficiency	Sufficient data quantity for the purpose	
		relevancy	No unrelated or unnecessary data	
		completeness	Missing level compared to total data	
		actionability	Can be used without any processing	
	(3) Timeliness	update	Updated with the latest data	
		update details	state when the update is made	
		deleting old data	Purge of obsolete data over time	
		data age	Usability of data	
	(4) Consistence	format consistency	Consistency of data representation forms	
		semantic consistency	Data values are clear and meaningful	
		data redundancy	No data redundancy or inconsistency	
		data conciseness	Concise and understandable	
		data compatibility	Data from different sources can be bundled together	
(5) Usefulness of Analysis Results		freshness	Are the results of the analysis novel?	Koch[2015] Lesca[2012] McNeely et al.[2013] Nicolaou et al.[2013] Kim et al.[2015] Kim et al.[2013] Kim et al.[2014] Park et al.[2016]
		plenishment	Diverse, Rich in Content	
		number of results	Number of findings discovered	
		relevancy	Relevancy with real world problems	
		cost adequacy	cost effectiveness	
(6) Use of Analysis Results		intent to use	Intent to utilize the analysis results	
		frequency of use	Use frequency for actual decision making	
		user satisfaction	Satisfaction with the use of analysis results	
		post change	Changes in work as a result of use	
(7) Continuance Intention		intent of continued use	Willingness to continue to use big data	
		recommendation	Recommend Big Data to others	
		extend the usage	Trying to use Big Data in new areas	

배타성과 전체적 완전성 확보를 위하여 여러 차례 논의와 조정이 이루어졌다. 그리고 간단한 예비조사를 실시하여 액면(face) 타당성을 검증하였고, 문장 워딩, 길이, 포맷에 대한 피드백을 반영하였다. 그 결과 <Table 2>에 나타난 바와 같이 정확성, 완전성, 적시성, 일관성 4개 품질요소와 18개 측정항목을 도출하였다.

본 연구에서 채택한 원천 데이터 품질요소와 측정항목의 특징은 다음과 같다. 우선 (1) 품질요소 정확성의 4개 측정항목 중 출처 명확성은 다른 선행 연구에서는 크게 강조되지 않는다. 그러나 빅데이터 환경에서는 여러 출처의 데이터 값이 다를 수 있으며, 이럴 경우 출처의 평판에 따라 데이터 신뢰도를 판단해야 하므로 출처 명확성을 측정항목으로 채택하였다. 때로는 평판이 낮은 출처로부터 놀랄만한 정보가 담긴 데이

터를 확보할 수 있으나 출처가 불분명한 데이터는 왜곡이나 오염의 우려로 인하여 사용하지 않는 것이 바람직하다. (2) 품질요소 완전성은 유용한 분석결과를 구하기 위하여 필요한 원천 데이터 자원을 얼마나 충분히 확보했는지에 초점을 맞추었다. 표면적, 직관적으로 인식하는 부가가치성(value-added) 항목은 배제하였으며, 빅데이터 분석 비용 효율성을 제고할 수 있는 즉각 사용가능성(actionability)을 추가하였다. (3) 품질요소 적시성(timeliness)은 빅데이터 분석 목적, 분석 과정, 원천 데이터의 용도 등을 고려하여 데이터 최신성(currency) 개념에 초점을 맞추었다. 최신 데이터로 업데이트 되어 있는지, 업데이트 시점과 내용을 알 수 있는지, 노후 데이터를 제거하여 불필요한 시행착오를 예방하는지, 데이터 나이(age)가 빅데이터 분석 목적에 적합한지를 측정항목으로 채택하였다. 종전 데이터베이스 시스템의 적시성 품질을 의미하던 검색속도나 응답속도는 원천 데이터 품질에 해당되지 않는다고 판단하였다. (4) 품질요소 일관성은 기업 내부의 정형적 데이터 차원에서 중시하던 데이터 중복과 불일치성 외에도 비정형적 데이터 차원에서 고려해야 할 표현구조와 의미의 일관성을 강조하였다. 선행 연구에서 채택되었던 이해용이성이나 해석용이성(interpretability) 항목은 원천 데이터 사용자가 데이터 해석에 능숙한 빅데이터 분석가라는 점을 고려하여 제외하였다. 그리고 전통적 품질항목으로 중시되던 접근편리성, 보안 등은 원천 데이터 품질보다는 시스템 품질에 가까우므로 본 연구에서는 제외하였다.

분석결과 유용성은 빅데이터 목적을 반영하여 정보 가치 개념을 강조하였다. 정보 가치는 사람들이 얼마나 깜짝 놀라는지에 의해 결정되며, 따라서 참신성을 측정항목으로 채택하였다. 이외에도 선행 연구를 토대로 분석내용 풍부성(richness), 현실문제와의 관련성, 그리고 실행비용의 조달가능성을 측정항목으로 채택하였다. 분석결과 활용도는 현업 의사결정자의 활용의사부터 활용 결과에 대한 만족도, 실제 활용한 빈도, 활용 이후에 발생한 변화에 이르기까지 전 과정에 걸쳐 균형있게 측정 항목을 개발하였다. 지속적 활용의사는 선행 연구에서 신뢰성과 타당성이 검증된 측정 도구를 사용하였으며, 빅데이터 특성을 고려하여 문장 일부를 수정하였다

4.2 데이터 수집

최종 설문지는 측정항목별로 리커트 5점 척도로 조사대상자 인식을 묻는 문항으로 구성하였다. 1점 “아주 미흡하다” 부터 5점 “아주 우수하다”까지 등간격 의미를 부여하였다. 국내 민간기업을 대상으로 설문조사를 실시하였다. 금융, 유통, 식음료, 통신서비스, 제조 등의 업종에서 현재 빅데이터를 활용할 수 있는 규모의 150개 상장회사를 표본으로 선정하였다. 조사 대상 기업은 내부적으로 여러 부서가 독자적으로 별도의 원천 데이터를 확보하여 빅데이터 분석을 실시하고 있으므로(예 : 연구개발, 기획, 생산공정, 품질관리, 시장동향분석, 고객관리, 콜센터 등) 본 연구의 데이터 분석 단위(unit of analysis)를 부서 레벨로 설정하였다. 조사방법은 이메일 설문조사를 채택하였다. 각 기업의 정보화 담당자에게 이메일을 발송하여 빅데이터 활용 부서를 1곳부터 3곳까지 선정하여 해당 부서의 빅데이터 분석가나 현업 의사결정자에게 설문지 작성을 요청하여 줄 것을 의뢰하였다. 2016년 10월에 1차 이메일을 발송하였으며, 미회신 표본을 대상으로 follow up 메일을 추가로 발송하였다. 그 결과 67개 기업에서 115개의 설문지가 회수되었고, 필수 항목이 누락되었거나 불성실한 응답지를 제외한 104개를 유효 응답지로 채택하였다.

4.3 표본 기초 통계

<Table 3>은 응답자에 대한 기초 통계 분석 결과이다. 업종 분포는 이 연구의 임의적 업종 선정과 표본 추출 결과이므로 일반적인 빅데이터 도입 기업의 업종 분포와는 일치하지 않을 수 있다. 빅데이터 활용 업무는 품질-고객-시장에 집중되어 있으며, 이는 단기적, 즉각적인 빅데이터 효과를 기대하고 있는 현실을 반영하고 있다. 빅데이터 활용기간 분포에서 3년 이상 장기간 활용기간과 1년 미만 단기간 활용 기간의 비율이 함께 낮은 현상은 빅데이터 도입 및 확산 추세가 실제로는 미디어에 발표되는 낙관적 전망에는 미치지 못할 수 있음을 의미한다. 응답자 직무 분포에서 중복 응답률이 138%로 나타난 것은 빅데이터 분석가와 활용자가 엄격히 분리되어 있지 않고, 분석과 활용을 병행하는 의사결정자가 상당함을 의미한다. 한편 응답자 성별 및 연령대 분포는 연구 목적과 관련이 없다고 판단하여 생략하였다.

〈Table 3〉 Statistics of Respondents

Classification	Responses	Distribution (n = 104)
Business Domain	Financial Sector	23.1%
	Communications/Portal	7.7%
	Culture/Entertainments	10.6%
	Manufacturing Sector	25.9%
	Distribution/Food	26.9%
	Others	5.8%
Big Data Application (multiple responses)	R&D	12.7%
	Plan	20.2%
	Service quality	45.2%
	Customer	51.9%
	Market	63.5%
	Others	12.5%
Period Use of Big Data	Less than 1 year	18.3%
	1~2 years	27.9%
	2~3 years	25.0%
	3~4 years	17.3%
	More than 4 years	11.5%
Respondent (multiple responses)	Big Data Analyst	73.1%
	Decision Maker	42.3%
	Business Management	22.1%

5. 통계 분석

본 연구는 SPSS 20.0을 이용하여 신뢰도와 타당성을 검증하였고, 가설 검증을 위하여 구조방정식 모델의 대안인 Smart PLS(Partial Least Square; 부분최소제곱) 2.0을 이용하였다.

5.1 신뢰성 및 타당성 분석

이 연구에서 사용한 7개 변수의 신뢰성을 분석한 결과가 <Table 4>의 우측에 나타나 있다. 모든 변수의 Cronbach 계수가 0.7 이상으로 나타났으며 따라서 안정적 신뢰도를 확보한 것으로 판단하였다. 그리고 구성 타당성 분석을 위하여 요인분석을 실시하였다. 각 요인의 추출은 주성분분석을 통한 VARIMAX 회전방법을 사용하였으며, <Table 4>의 좌측 부분은 요인분석 결과를 나타내고 있다. 고유치(eigen value)가 1.0 이상인 요인에 대하여 요인 적재치가 0.5 이상인 것만을 나타내고 있는데 2곳 이상 교차 적재된 측정항목은 없었으며, 예상대로 총 7개의 요인이 추출되었다. 연구변수 (1), (2), (3), (5), (6), (7)에 속하는 측정항목들은 연구 초기의 설정대로 명확하게 적재되었다. 그러나 연구변수 (4)의 측정항목 중 "데이터 간결성"은 7개 요인 중 어느 곳에도 적재되지 않았으므로 추가 분석에서 제외하였다. 그리고 7개 요인의 적재치 값들이 모두 0.5 이상을 만족하고 있고, AVE(Average Variance Extracted; 평균분산추출) 값이 모두 기준값 0.5 이상으로 나타났으므로 구성 개념 간 수렴타당성(convergent validity)이 확보되었다고 판단하였다.

〈Table 4〉 Factor Analysis Results

Research Variables	Items	Loading	eigen value	Cumulation (%)	Cronbach's alpha	AVE
(1) Accuracy	accuracy of value	0.906	2.670	14.24	0.873	0.731
	reliability	0.842				
	objectivity	0.824				
	source clarity	0.795				
(2) Completeness	no missing value	0.889	3.012	21.56	0.847	0.794
	data sufficiency	0.854				
	relevancy	0.790				
	completeness	0.812				
	actionability	0.792				
(3) Timeliness	update	0.889	2.918	30.24	0.845	0.832
	update details	0.825				
	deleting old data	0.780				
	data age	0.818				
(4) Consistence	format consistency	0.780	2.256	39.41	0.768	0.688
	semantic consistency	0.768				
	data redundancy	0.807				
	data conciseness	*				
	data compatibility	0.802				
(5) Usefulness of Analysis Results	freshness	0.856	2.756	47.80	0.841	0.732
	plenishment	0.884				
	number of results	0.819				
	relevancy	0.826				
	cost adequacy	0.862				
(6) Use of Analysis Results	intent to use	0.846	2.398	58.24	0.794	0.780
	frequency of use	0.883				
	user satisfaction	0.805				
	post change	0.785				
(7) Continuance Intention	intent of continued use	0.825	2.734	69.21	0.802	0.76
	recommendation	0.847				
	extend the usage	0.794				

〈Table 5〉 Discrimination Validity

Research Variables	(1) Accuracy	(2) Completeness	(3) Timeliness	(4) Consistence	(5) Usefulness of Analysis	(6) Use of Analysis	(7) Continuance Intention
(1) Accuracy	0.856*						
(2) Completeness	0.575	0.891*					
(3) Timeliness	0.790	0.743	0.912*				
(4) Consistence	0.723	0.776	0.694	0.822*			
(5) Usefulness of Analysis	0.645	0.732	0.795	0.778	0.854*		
(6) Use of Analysis	0.637	0.765	0.806	0.798	0.756	0.884*	
(7) Continuance Intention	0.568	0.790	0.756	0.802	0.778	0.774	0.873*

* = Root square of AVE.

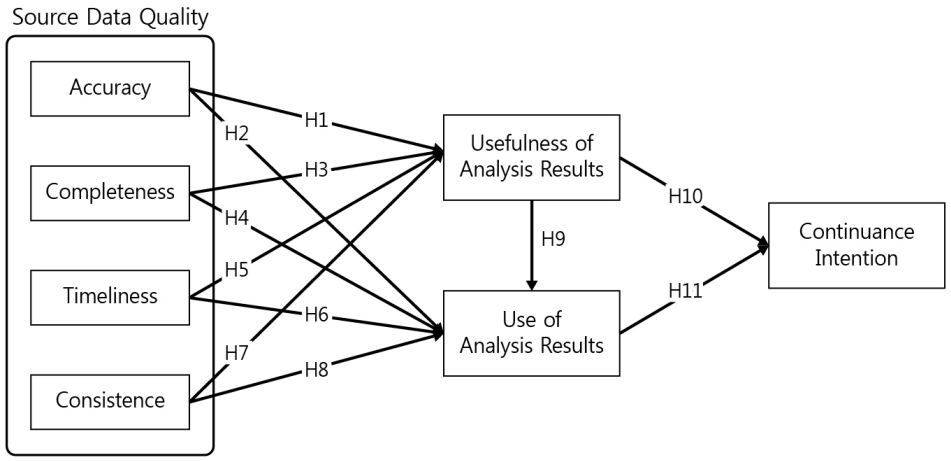
판별 타당성(discriminant validity) 측정을 위해 <Table 5>에 나타난 바와 같이 AVE 값을 사용하였다. 별표(*)로 표시한 값은 AVE 제곱근 값이며 나머지 행렬에서의 값은 각 변수의 상관계수 값을 나타낸다. AVE 제곱근 값이 모두 0.7 이상이고, 인접한 가로축과 세로축의 다른 변수와의 상관계수 값보다 크게 나타나고 있으므로 다중공선성 문제가 발생하지 않고 판별 타당성이 있다고 판단하였다.

5.2 경로 분석 및 가설 검증

본 연구는 PLS 부트스트랩 기법을 적용하여 경로모형 분석에 의하여 연구 모형을 검증하였으며 그 결과가 <Figure 2>에 나타나 있다. 11개 가설 중 8개 가설이 채택되었다. PLS 분석에서는 R^2 값에 의한 종속변수 설명력이 높을수록 우수한 모형으로 평가된다. <Figure 2> 분석결과에서 보는 것처럼 선행변수에 의해 설명되는 최종 종속변수 "지속적 활용의사"의 R^2 값은 0.567이고, 매개변수 "분석결과 유용성"의 R^2 = 0.615, "분석결과 활용도"의 R^2 = 0.627로 나타났다. R^2 값이 일반적으로 인정되는 적정 검정 수준 10% 이상이므로 구조모형에 적합하다고 판단하였다.

<Table 6>은 연구 모형의 경로계수와 11개 가설의 검증 결과를 정리하고 있다. 원천 데이터 품질의 영향에 관한 8개 가설 중에서 5개 가설이 통계적으로 유의한 것으로 나타났으나, 완전성 품질 가설 H4와 일관성 품질 가설 H7과 H8이 기각되었다는 점이 주목을 끈다. 검증 결과에 관한 논의는 다음과 같다.

첫째, 원천 데이터 일관성 품질은 분석결과의 유용성과 활용도에 미치는 영향이 함께 유의하지 않는 것으로 나타났다. 빅데이터 환경에서는 일관성 품질이 낮더라도 다양한 분석 기법을 활용한다면 저품질 문제에 대처할 수 있기 때문으로 판단된다. 예컨대 이미지 해상도가 극히 낮아서 정확성이나 완전성이 부족한 상황은 더 이상의 진행이 불가능하지만, 이미지의 크기, 규격, 색상, 저장형태 등의 일관성 품질이 낮은 데이터는 다양한 기법과 노력을 투입한다면 대처 가능하기 때문이다. 그리고 기업 내부의 정형적 데이터를 관리하는 데이터베이스 시스템과는 달리 빅데이터 환경에서는 표현구조 일관성이나 데이터 중복이 큰 비중을 차지하지 않음을 의미한다. IoT 센서 데이터, 다양한 양식의 문서, 가변 길이 텍스트, 이미지, 영상, 음성 등과 같은 원천 데이터



〈Figure 2〉 Hypothesis Test Results

〈Table 6〉 Hypothesis Testing Results

Hypothesis	Path	Path coefficient	t value	P value	Result
H1	Accuracy → Usefulness of Analysis	0.618	6.454	***	Accepted
H2	Accuracy → Use of Analysis	0.222	2.009	**	Accepted
H3	Completeness → Usefulness of Analysis	0.515	5.804	***	Accepted
H4	Completeness → Use of Analysis	0.061	0.396	0.503	Rejected
H5	Timeliness → Usefulness of Analysis	0.329	3.464	***	Accepted
H6	Timeliness → Use of Analysis	0.227	2.179	**	Accepted
H7	Consistence → Usefulness of Analysis	0.056	0.319	0.467	Rejected
H8	Consistence → Use of Analysis	0.037	0.485	0.616	Rejected
H9	Usefulness → Use of Analysis	0.382	3.557	***	Accepted
H10	Usefulness → Continuance Intention	0.268	3.254	***	Accepted
H11	Use of Analysis → Continuance Intention	0.318	3.112	***	Accepted

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

는 애초부터 일관성 품질이 낮다는 인식도 이러한 결과가 나타나는 한 가지 원인으로 판단된다.

둘째, 완전성 품질과 분석결과 활용도 사이의 가설이 기각된 것은 원천 데이터 품질과 전통적 정보 품질의 가장 현저한 차이로 볼 수 있다. 불완전하더라도 활용하거나, 이와 반대로 완전하더라도 활용하지 않는 경우는 빅데이터 환경에서 흔히 나타나는 현상이다. 빅데이터 분석결과가 실제 현업에 활용되기 위해서는 아이디어 자체의 유용성과 아이디어를 도출한 데이터의 정확성이 중요한 것이지 얼마나 많은 양의 데이터를 검색하였는지는 큰 의미가 없기 때문이다. 외부의 비정형 데이터를 포착, 스크리닝, 필터링하는 과정에서 데이터 누락이나 왜곡이 언제든지 발생할 수 있다는 점을 고려할 때 원천 데이터의 완전성 품질에 대한 새로운 시각이 요구되고 있다.

셋째, 원천 데이터 적시성 품질이 가지는 영향의 통계적 유의성이 확인되었다. 실시간으로 발생하는 대용량 데이터를 처리하는 빅데이터 환경에서는 데이터의 최신성, 나이, 업데이트 문제를 포괄하는 적시성 품질이 중요한 비중을 차지한다. 그러나 실시간 원천 데이터의 경우, 짧은 시간적 유효성으로 인하여 적시성 품질확보 비용이 현실적으로 큰 부담이며 앞으로 원천 데이터 품질관리의 비중 있는 이슈가 될 것이다.

그리고 빅데이터 분석결과 유용성과 활용도 사이의 긍정적 상관관계에 관한 가설 H9가 채택되었다. 정보 품질과 정보 이용 사이에 정(+)의 상관관계를 제시한 정보시스템 성공모델이나 정보 유용성과 활용도 사이의 상관관계를 기반으로 하는 기술수용모델이 빅데이터 환경에서도 유효함을 의미한다. 그러나 빅데이터 분석결과가 아무리 유용하더라도 반드시 현업에서 활용된다는 보장은 없다. 분석결과가 현업이나 의사결정 과정에서 제대로 활용되기 위해서는 최고경영층 지원, 변화지향적 조직문화, 빅데이터 분석가의 역할 등이 함께 작용되어야 하며, 분석결과 유용성은 분석결과 활용성의 충분조건이 아니라 필요조건임을 유의할 필요가 있다. 빅데이터 지속적 활용의사에 관한 가설 H10와 H11도 통계적으로 유의하게 나타났다. 정보시스템의 유용성과 지속적 활용의사 사이의 상관관계를 입증한 여러 선행 연구 결과와 전제가 빅데이터 환경에서도 적용될 수 있음을 의미한다.

6. 결론 및 제언

이 연구는 원천 데이터 품질이 빅데이터 분석결과에 미치는 영향을 검증하였다. 선행 연구와 전문가 평가를 토대로 원천 데이터 품질을 정확성, 완전성, 적시성, 일관성 4개 요소로 구분하고, 각 요소마다 다수의 측정문항을 도출하였다. 원천 데이터 품질의 영향에 관한 11개 가설을 설정하고, 국내 기업 대상 이메일 설문조사에 의해 확보한 104개 유효 응답데이터를 통계 분석한 결과는 다음과 같다. (1) 원천 데이터 품질요소 중 정확성과 적시성이 모두 분석결과 유용성 및 활용도에 정(+)의 영향을 미친다. (2) 그러나 일관성 품질은 어느 쪽에도 유의미한 영향을 미치지 못하였으며, 완전성 품질은 분석결과 유용성에만 긍정적 영향을 미치고 있다. (3) 분석결과 유용성과 활용도 사이에 긍정적 상관관계가 존재하며, 유용성이나 활용도가 높을수록 빅데이터의 지속적 활용의사가 높아지는 것으로 나타났다.

이 연구의 이론적, 실용적 의의는 크게 두 가지이다. 연구 목적과 빅데이터 분야의 특성상 이론적 의의와 실용적 의의를 굳이 구분할 실익이 없으므로 동시에 살펴 본다. 우선 원천 데이터 품질 개념을 전통적 정보 품질 개념과 구분하고, 범용적 품질요소 및 측정항목을 정비하였다. 이 연구에서 제시하는 연구 변수의 조작적 정의는 향후 후행 연구와 기업 품질관리 실무에서 참고 모델로 활용될 수 있을 것이다. 그리고 원천 데이터 품질이 빅데이터 성공에 미치는 영향을 실증적으로 검증하였다. 4개 품질요소 별로 세분화함으로써 어떤 품질요소가 어떤 영향을 미치는지에 대한 깊이 있는 분석을 수행하였다. 이 연구 결과는 선행 연구들이 제기한 원천 데이터 관련 논란과 쟁점에 대하여 의미 있는 정보를 제공하고 있다.

이 연구의 한계는 우선 표본의 선정방법에서 찾을 수 있다. 빅데이터 도입 가능성이 높은 업종과 기업을 임의 선정하였으며, 표본 기업 자체적으로 복수의 응답 부서를 선택하였다. 따라서 응답자 추출의 무작위성이 완전히 확보되지 않았고, 연구결과의 범용성 문제가 제기될 수 있다. 향후 이러한 외적 타당성 한계를 극복하여 이 연구 결과를 재검증할 수 있는 연구 노력이 나타나기를 기대한다. 그리고 이 연구는 산업별 특성이나 품질, 고객, 시장 등 업무분야에 따라 원천 데이터 품질요소와 품질수준이 다를 수 있다는 점을 고려하지 않았다. 기업 내부의 정형적 데이터인지 혹은 기업 외부의 비정형적 데이터인지에 따라 차이가 발생할 수도 있으나 연구 범위에 포함하지 않는 것도 미진한 점이다.

이 연구 결과와 한계점을 토대로 향후 연구 방향을 다음과 같이 제언한다. 빅데이터 성공에 영향을 미치는 변수는 원천 데이터 품질뿐만 아니다. 기술적 차원에서 분석플랫폼, 분석기법 및 도구, 분석가 역량 등이 있고, 조직적 차원에서 기존 정보화 수준, 경영층 리더십, 조직 문화 등이 있으며, 외부 환경적 차원에서 정책적 요인, 경쟁사 도입에 의한 심리적 압박 등 다양한 변수가 존재한다. 이러한 변수들의 영향을 원천 데이터 품질과 함께 분석한다면 빅데이터 분석결과의 유용성, 활용도, 지속적 활용의지에 대한 폭넓고 깊이 있는 통찰력을 제공할 수 있을 것이다. 그리고 기업 규모, 업종, 빅데이터 활용기간 등 일부 변수들의 조절 효과도 체계적으로 규명되기를 기대한다.

References

[1] Baskarada, S., “How spreadsheet applications affect information quality”, *Journal of Computer Information Systems*, Vol. 11,

No. 2, 2011, pp. 77-84.

[2] Benlian, A. and Hess, T., "Opportunities and risks of software-as-a-service : Findings from a survey, of IT executives", *Decision Support Systems*, Vol. 52, No. 3, 2011, pp. 232-246.

[3] Bottles, K. and Begoli, E., "Understanding the Pros and Cons of Big Data Analytics", *Physician Executive Journal*, Vol. 40, No. 4, 2014, pp. 6-10, p. 12.

[4] Cha, S. E. and Joo, H. T., "A study on the Big Data application trends and vitalization", *Industrial Engineering Magazine*, Vol. 22, No. 1, 2015, pp. 41-45.

[5] Cho, W. S., "The trend of Big Data governance and standardization", *OSIA S&TR Journal*, Vol. 30, No. 2, 2017, pp. 26-29.

[6] Davis, F. D., Bogazzi, R. P., and Warshaw, P. R., "User acceptance of customer technology : A comparison of two theoretical models", *Management Science*, Vol. 35, No. 2, 1989, pp. 982-1003.

[7] DeLone, W. H. and McLean, E. R., "The DeLone and McLean Model of Information Systems Success : A Ten-Year Upgrade", *Journal of Management Information Systems*, Vol. 19, No. 4, 2003, pp. 9-30.

[8] E-newspaper, "Developing Big Data quality assessment tools in Korea", 2017. 05. 10.

[9] Ga, H. K. and Kim, J. S., "A study on the influential factors on the intention of Big Data adoption", *in Proceedings of Spring Conferences of KMIS,* 2014, pp. 691-707.

[10] Gartner, *Data Quality for Big Data,* Gartner, 2011.

[11] Halaweh, M. and Massry, A. E., "Conceptual model for successful implementation of big data in organizations", *Journal of International Technology and Information Management*, Vol. 24, No. 2, 2015, pp. 21-34.

[12] Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., and Shahabi, C., "Big data and its technical challenges", *Communications of the ACM*, Vol. 57, No. 7, 2014, pp. 86-94.

[13] Jang, M. S., "Big Data quality management based on AI", in *Proceedings of Spring Conferences of KMIS*, 2017, pp. 168-179.

[14] Kdata, "Top 3 Elements of Success with Big Data : Resources, Technology, and Analysts", *IT & Future Strategy*, No. 3, 2012.

[15] Kdata, *Data Quality Management Maturity Survey Report*, 2010.

[16] Kim, C. Y., Lee, J. W., and Park, C., "Classification and prospect of enterprise Big Data utilization from the perspectives of strategic value", in *Proceedings of Fall Conferences of KMIS,* 2014, pp. 501-510.

[17] Kim, J. S. and Song, T. M., "A study on the initial characteristics of acceptance of data technologies : focused on the regulatory effects of technical users and technology users", *Journal of the Korean Content Association,* Vol. 14, No. 9, 2014, pp. 538-555.

[18] Kim, S. H., Lee, C. S., Jeong, S. H., Kim, H. C., and Lee, C. S., "An organizational maturity assessment model for public data quality Management", *Information Policy,* Vol. 22, No. 1, 2015, pp. 28-46.

[19] Kim, S. H., Park, J. H., Kim, E. H., and Park,

J. S., “A study on the data analysis and utilization and the improvement of decision making quality”, *Journal of Information Technology and Architecture*, Vol. 22, No. 1, 2015, pp.159-170.

[20] Koch, R., “Big data or big empathy?”, *Strategic Finance*, 2015, pp. 62-63.

[21] Lee, S. G., “Marketing approach to Big Data analysis”, *Journal of Korean Management Society,* Vol. 28, No. 1, 2015, pp. 21-35.

[22] Lee, S. W. and Lee, H. S., “A study of the integrated model for implementing Big Data systems”, *Journal of Information Technology Applications and Management*, Vol. 21, No. 4, 2014, pp. 463-482.

[23] Lee, Y. W., Strong, D. M., Kahn, B. K, and Wang, R. Y., “AIMQ : a methodology for information quality assessment”, *Information & Management*, Vol. 40, No. 2, 2002, pp. 133-146.

[24] Lesca, N., Caron-Fasan, M. L., and Falcy, S., “How managers interpret scanning information”, *Information & Management*, Vol. 49, No. 2, 2012, pp. 126-134.

[25] Moges, H. T., Dejaeger, K., Lemahieu, W., and Baesens, B., “A multidimensional analysis of data quality for credit risk management : New insights and challenges”, *Information & Management,* Vol. 50, No. 1, 2013, pp. 43-58.

[26] Nicolaou, A. I., Ibrahim, M., and van Heck, E., “Information quality, trust, and risk perceptions in electronic data exchanges”, *Decision Support Systems*, Vol. 54, No. 2, 2013, pp. 986-996.

[27] Park, G. E. and Kim, C. J., “A study on the quality characteristics of public open data”, *Journal of Digital Convertgence*, Vol. 13, No. 10, 2015, pp. 135-146.

[28] Park, S. H., Goo, B. J., and Lee, K. H., “The impact of CEO leadership on Big Data success”, *Information Systems Review*, Vol. 18, No. 2, 2016, pp. 39-57.

[29] Petter S., DeLone, W., and McLean, E. R., “The past, present, and future of ‘IS Success”, *Journal of the Association for Information Systems*, Vol. 13, No. 5, 2012, pp. 341-362.

[30] Roe, S.Y ., “How Big Data awareness and retailing by small to medium businesses distresses production activities to improve productivity”, *Journal of Startup Business*, Vol. 11, No. 2, 2016, pp. 48-68.

[31] Ross, J. W., Beath, C. M., and Quaadgras, A., “You may not need big data after all”, *Harvard Business Review*, December 2013, pp. 90-98.

[32] Schwab, K., *The Fourth Industrial Revolution,* World Economic Forum, 2016.

[33] Smith, K., “Big data big concerns”, *Best's Review*, 2015, pp. 58-61.

[34] Wang, R. Y. and Strong, D. M., “Beyond accuracy : what data quality means to data consumers”, *Journal of Management Information Systems*, Vol. 12, No. 4, 1996, pp. 5-34.

저자소개

박 소 현

한국외국어대학교에서 경영정보학 석사와 건국대학교에서 경영정보학 박사 학위를 취득하고 현재 건국대학교 경영대학에서 초빙교수로 근무하고 있다. 관심 분야는 정보기술 평가, 빅데이터, 클라우드 컴퓨팅 등이다.

이 국 희

Georgia State University에서 경영정보학 박사학위를 취득하고 현재 건국대학교 경영대학 교수로 재직하고 있다. 관심 분야는 IT평가, IT컨설팅, 클라우드 컴퓨팅, 빅데이터이다.

이 아 연

한양대학교에서 경영학 석사 학위를 받았으며 동 대학원에서 박사과정을 수료하였다. NCS 일학습 병행프로그램 개발과 PMO 제도적용 실태조사 및 개선방안 마련 연구 등에 참여하였다. 관심 분야는 프로젝트 관리, PMO, 서비스경영 등이다.