https://doi.org/10.15813/kmr.2023.24.1.011

벤처기업정밀실태조사와 한국기업혁신조사 데이터를 활용한 통계적 매칭의 타당성 검증[1)]

The Validity Test of Statistical Matching Simulation Using the Data of Korea Venture Firms and Korea Innovation Survey

안경민 (An, Kyungmin) 동국대학교 WISE 캠퍼스[2)]
이영찬 (Lee, Young-Chan) 동국대학교 WISE 캠퍼스[3)]

〈 국문초록 〉

최근 데이터 경제가 가속화되면서 경영학 분야에서는 데이터 매칭이라는 새로운 기법이 주목받고 있다. 데이터 매칭은 모집단이 같지만 서로 다른 표본에서 수집된 데이터셋을 합치는 기법 또는 처리 과정을 의미한다. 그중에서 통계적 매칭은 서로 다른 데이터를 결합하는데 있어서 사업자 번호와 같이 기준이 되는 변수가 없는 경우 통계적 함수를 활용하여 데이터를 매칭하는 방법이다. 선행연구 검토결과 경제학, 교육학, 보건, 의료 등 다양한 분야에서 통계적 매칭이 많이 사용되고 있는데 반해 경영학 분야는 제한적임을 확인할 수 있었다. 본 연구는 기존 경영학 분야에서 충분히 연구되지 않았던 통계적 매칭의 유용성을 검증하고 활용도를 높이는 방안을 연구하고자 한다. 연구목적을 달성하기 위해 본 연구에서는 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사 자료를 활용하여 통계적 매칭 시뮬레이션을 수행하였다. 먼저, 선행연구를 바탕으로 통계적 매칭에 사용되는 변수를 선정하였다. 공통변수는 업종, 종업원수, 지역, 업력, 상장시장, 매출로 설정하였고, 검증을 위한 고유변수와 제공변수는 중소기업 혁신에서 가장 중요한 연구인력 비율과 R&D 비용으로 각각 설정하였다. 사전 검증을 위해 2020 벤처기업정밀실태조사 자료를 수여자 데이터 30%와 기여자 데이터 70%로 분할하였다. 통계적 매칭에는 마할라노비스 거리와 랜덤 핫덱을 결합한 방식을 사용하였고, 성능평가는 수여자 데이터와 원시 데이터의 평균값 비교와 커널 밀도 함수(Kernel Density Estimation)를 통해 데이터 분포를 비교하였다. 검증결과, 수여자 데이터 30%와 기여자 데이터 70%에서 추출된 매칭 데이터의 평균값이 통계적으로 유의한 차이가 없는 것으로 나타나 유사한 데이터가 매칭된다는 것을 확인하였다. 또한, 두 데이터의 커널 밀도 함수로 도출한 데이터 분포 역시 유사한 형태가 나타나는 것을 확인할 수 있었다. 사후 검증에는 2020 벤처기업정밀실태조사에서 임의로 30%를 수여자

---

1) 이 논문은 2021년 대한민국 교육부와 한국연구재단의 인문사회분야 신진연구자지원사업의 지원을 받아 수행된 연구임(NRF-2021S1A5A8061237)
본 논문은 2021년 동국대학교 안경민의 박사학위논문 "안경민 (2021). 통계적 매칭과 머신러닝 앙상블 기법을 활용한 기업혁신 및 경영성과 예측모형 개발" 을 기반으로 작성
본 논문은 과학기술정책연구원(STEPI: Science and Technology Policy Institute)에서 제공한 한국기업혁신조사(KIS: Korean Innovation Survey) 데이터를 이용하여 수행함.
2) 제1저자, snss1212@dongguk.ac.kr
3) 교신저자, chanlee@dongguk.ac.kr

논문접수일 2023년 02월 13일 | 1차 수정 2023년 03월 02일 | 게재확정일 2023년 03월 04일

데이터로 추출하고 2020 한국기업혁신조사 자료를 기여자 데이터로 설정하여 통계적 매칭을 수행하고 검증하였다. 사전 검증과 마찬가지로 공통변수는 업종, 종업원수, 지역, 업력, 상장시장, 매출로 설정하였고, 검증을 위한 고유변수는 연구인력 비율과 R&D 비용으로 정의하였다. 분석 결과, 수여자 데이터의 연구인력 비율의 평균과 기여자 데이터의 평균은 예상과 다르게 통계적으로 차이가 있는 것으로 나타났다. 하지만 커널 밀도 함수에 따른 두 데이터의 분포는 유사한 형태를 보이는 것으로 조사되어 통계적 매칭의 적절성을 확인할 수 있었다. R&D 비용은 통계적 매칭 수행 결과, 수여자 데이터의 R&D 비용 평균과 기여자 데이터의 평균이 통계적으로 차이가 없었고, 커널 밀도 함수도 유사한 분포를 보이는 것으로 조사되었다. 이러한 결과는 모집단은 동일하지만 서로 다른 표본에서 수집된 자료를 통계적으로 결합하여 신뢰할 수 있는 새로운 데이터를 확보할 수 있다는 측면에서 큰 의의가 있다. 또한, 경영학 분야에서 많이 사용되지 않았던 데이터 매칭 방법론을 모의실험을 통해 타당성을 검증함으로써 연구용 데이터 확보와 연구방법론의 확장에 기여했다는 점에서 시사점을 가진다.

주제어: 데이터 매칭, 통계적 매칭, 공공데이터, 중소기업, 기술혁신, 시뮬레이션

1. 서론

최근 데이터 경제 시대가 가속화되면서 데이터를 활용하여 새로운 부가가치를 창출하는 방안이 모색되고 있고(Wiener et al., 2020; Zheng et al., 2022; Limna et al., 2023; 김동성 등, 2017; 최봉 등, 2019), 경영학에서도 통상적인 연구를 넘어 데이터를 활용한 새로운 분석이 시도되고 있다(Eccles et al., 2014; Chang and Shim, 2015; Kwon & Johnson, 2018; 이규엽 등, 2020; 안경민, 이영찬, 2021). 데이터 매칭은 데이터 시대의 도래와 함께 통계 자료의 형태와 수준이 복잡해지면서 중요성이 강조되고 있는 방법으로 기존의 2차 데이터를 재조합하여 새로운 데이터를 창출하는 방법이다(Martín-de Castro et al., 2020; D'Alberto & Raggi, 2021). Chang and Shim (2015)은 가족 기업에서 전문 CEO 체제로의 전환에 따른 기업성과 향상을 탐색하기 위해 데이터 매칭 방법을 적용하였고, Eccles et al. (2014)은 기업의 시계열 데이터를 활용하여 사회적 책임이 주가 및 재무적인 경영성과에 미치는 영향을 탐색하기 위해 서로 다른 2차 데이터를 연계한 방법을 시도하였다. 또한, Holsapple and Wu (2011)은 지식관리가 재무성과 달성을 위해서 중요하게 고려해야 할 요소임을 입증하기 위해 데이터 매칭 방법을 시도하였으며, Nold (2012)은 기업 신뢰와 성과 관계를 분석하기 위해 데이터 매칭 기법으로 비교 기업 데이터를 생성하여 조직문화에 내재되어 있는 상대적 신뢰가 높을수록 우수한 기업이 될 수 있다는 결과를 도출하였다. 따라서 데이터 매칭은 경영학 분야 연구에서 계속 지적되고 있는 데이터 수집의 한계점을 극복하고 연구의 스펙트럼을 넓힐 수 있는 방법론으로 도입할 필요가 있다(Van Der Putten et al., 2002).

데이터 매칭은 기존의 데이터를 활용하여 재조합하는 데 있어서 강점이 있다(안경민, 2021). D'Orazio et al. (2006), Yang and Kim (2020)은 기존 데이터 수집 방법과 활용의 한계점을 지적하며, 데이터 매칭의 필요성을 제시하였다. 첫째, 기존의 데이터 수집 방법은 정보 제공의 적시성에 문제가 있다. 새로운 조사를 수행하기 위해서는 많은 시간이 소요되는데, 시간이 소요됨에 따라 정보는 변화하게 되고 이 정보는 현실과 괴리감으로 인하여 정확한 의사결정을 방해할 수 있다. 즉, 적절한 의사결정을 위해서는 현재 보유하고 있는 데이터를 활용하여 의미 있는 정보를 창출하는

방안이 요구된다. 둘째, 새로운 조사는 경제적 비용을 수반한다. 유의미한 정보를 획득하기 위해서는 조사 방법, 측정지표 개발, 데이터 수집, 데이터 분석, 정보 도출 등의 일련의 과정을 거치게 되고, 이 과정은 경제적 비용을 동반한다. 특히, 복잡한 사회현상을 분석하기 위해 다양한 변수를 고려할 때 데이터 수집 방법에 오류가 있는 경우 경제적 비용이 기하급수적으로 늘어날 수 있다. 데이터 매칭은 기존에 수집된 데이터를 재조합하여 데이터셋을 구축하는 방법으로 이와 같은 오류에서 비교적 자유로우며, 경제적 비용을 절감할 수 있다. 셋째, 설문지의 경우 조사 내용이 길어질수록 응답의 품질이 낮아지며, 무응답의 빈도가 높아지는 상황이 발생한다. 이러한 상황이 발생하면 추가 조사가 요구되며, 추가 조사는 응답에 대한 부담을 증가시키고 무응답을 늘리는 악순환을 발생시킬 수 있다. 즉, 데이터 수집의 오류를 줄이고 데이터의 품질을 높이는 방안에서도 데이터 매칭은 필요하다고 할 수 있다(박희창, 조관현, 2006).

따라서 기존 조사 방법의 한계점과 데이터 품질의 향상, 효율적인 분석, 학술적·정책적 함의를 도출할 수 있는 새로운 방법론으로서 데이터 매칭을 적극 고려할 필요가 있다(안경민, 2021). 데이터 매칭은 데이터의 양과 범위를 넓혀 양질의 데이터를 구성하고, 데이터 수집의 시간과 비용을 줄일 수 있는 방법론으로 사용될 수 있다. 또한, 응답자의 부담 경감, 낮은 응답율 해소, 조사비용 절감에 긍정적인 효과를 창출할 수 있으며, 원천 데이터의 다양성, 단일 자료의 불충분성, 자료 공유의 부족으로 인하여 나타나는 데이터 분석의 문제점을 해소할 수 있다(정용찬 등, 2017).

그러나 이와 같은 데이터 매칭의 활용성에도 불구하고 최근까지 경영학 분야에서는 거리함수를 활용하는 정도로 제한적인 연구가 이뤄지고 있다. 대표적으로 김성호, 조성빈(2005)은 고객관계관리와 관련된 전략을 도출하기 위해 신용카드 데이터를 확보하여 마할라노비스 거리 함수를 적용한 통계적 매칭을 수행하였으며, 성능평가는 변수 간의 상관계수, 평균제곱오차(Mean Squared Error)로 이뤄졌다. Ferrando and Mulier (2015)은 금융위기 기간 동안 유럽 기업의 자금조달 상황의 인식 정도를 분석하기 위해 고워 거리(Gower's distance) 함수를 활용하여 유럽 비영리협회 SAFE의 서베이 데이터와 Bureau van Dijk의 AMADEUS 데이터베이스의 데이터를 통계적으로 결합하였다. 이 방법에서 나타나듯이 이전의 경영학 분야에서 통계적 매칭은 단순한 거리함수를 활용하고 있다는 점을 볼 수 있다. 하지만 최근 사회과학 분야에서는 단계적 매칭(van Pelt, 2001), K-최근접이웃(김희경, 2010), 회귀분석(Ingram et al., 2000), 회귀분석과 K-최근접이웃의 결합(정성석 등, 2004), 랜덤 핫덱(Singh et al., 1990) 등을 조합하여 데이터 매칭의 성능을 높이는 방법이 모색되고 있다.

이 같은 이전 연구의 한계점을 극복하고자 본 연구는 한국에서 조사되고 있는 경영학 분야의 2차 데이터를 수집한 후 최적의 데이터 매칭 방법을 도출하고자 한다. 연구목적을 달성하기 위해 먼저 경영학 분야를 대표할 수 있는 국가 통계자료를 탐색한다. 데이터 매칭은 모집단을 공유하고 있지만 개별적으로 조사된 2개 이상의 데이터를 결합하는 방법론이다. 본 연구에서는 중소벤처기업 분야의 대표적인 데이터인 '2020 벤처기업정밀실태조사'와 '2020 한국기업혁신조사' 결과를 사용하여 데이터 매칭을 수행한다. 다음으로 최적의 데이터 매칭 방법을 탐색하기 위한 시뮬레이션을 수행한다. 데이터 매칭의 우수성은 데이터 매칭 방법을 잘 설계하는 것에서 출발한다. 데이터 매칭의 설계는 선행연구를 바탕으로 데이터 확보, 변수 설정, 매칭 방법 설정, 시뮬레이션을 통해서 확인할 수 있다. 본 연구에서는 2020 벤처기업정밀실태조사 결과를 임

의로 분할하여 통계적으로 매칭하는 시뮬레이션을 수행하고, 데이터 매칭의 적절성을 확인한다. 다음으로 데이터 매칭의 타당성을 확인하기 위해 새로운 데이터와의 결합을 시도한다. 본 연구에서는 벤처기업이라는 데이터의 특성을 반영하여 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사를 통계적으로 매칭하였으며, 각 자료에 존재하는 변수값을 비교함으로써 통계적 매칭의 적용 가능성을 탐색한다.

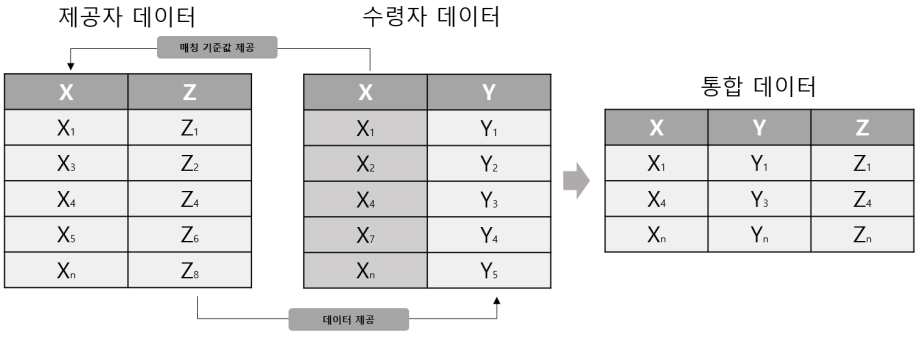
2. 이론적 배경

2.1. 데이터 매칭

데이터 매칭은 모집단이 같지만 서로 다른 표본에서 수집된 데이터셋을 합치는 기법 또는 처리 과정을 의미하며, 데이터 연계(Data Linkage), 데이터 퓨전(Data Fusion), 데이터 결합(Data Integration) 등으로 표현된다. 이 방법론은 통계학, 의학, 경제학, 정치학, 사회학, 법학 등을 포함한 많은 분야에서 인기 있는 추론 방법으로 활용되고 있다(오미애 등, 2015; 박희창, 조광현, 2006; Kwon & Johnson, 2018; Chang & Shim, 2015; Eccles et al., 2014; Holsapple & Wu, 2011; Nold, 2012)

데이터 매칭의 기본 개념은 <그림 1>과 같이 표현되며, 크게 제공자 데이터(Donor Data)와 수령자 데이터(Recipient Data)로 나뉜다. 제공자 데이터는 고유의 정보를 담고 있는 데이터셋을 가지고 있으며, 데이터 매칭 과정에서 수령자 데이터에게 적절한 데이터를 제공하는 역할을 한다. 수령자 데이터는 기준 데이터, 주체 데이터, 수용 데이터라고도 불리며, 정보를 받는 기준의 역할을 한다. 즉 데이터 매칭은 두 개 이상의 데이터 집합에서 데이터 전달이 이루어지는 구조이며, 주요 정보가 관측되는 데이터 집합에서 관심 있는 하나의 데이터 집합으로 핵심 정보를 이전하는 방법을 수행한다. 이 같은 데이터 매칭은 서로 다른 데이터를 연계하여 하나로 통합된 데이터셋을 추출하기 때문에 정보 전달 과정을 결측치 처리(Imputation) 방법이라고 한다. 데이터 이전 과정을 통해 제공자 데이터와 수령자 데이터가 결합이 되면 매칭된 데이터셋이라고 하며, 이 데이터셋은 공통변수인 'X'라는 고유변수를 가지게 된다.

데이터 매칭은 크게 정확 매칭(Exact Matching)과 통계적 매칭(Statistical Matching)으로 구분된다. 정확 매칭은 사업자등록번호, 주민등록번호, 국가보험번호, 사회보장번호와 같이 각 개체를 식별할 수 있는 변수가 공통으로 있을 때 같은 값을 갖는 개체들을 결합하



〈그림 1〉 데이터 매칭 개념도

출처: D'Alberto, R., Zavalloni, M., Raggi, M. and Viaggi, D. (2018). AES impact evaluation with integrated farm data: Combining statistical matching and propensity score matching. Sustainability, 10(11), p.7, Figure 1 수정 후 인용.

는 방법이다. 같은 내용을 정확하게 결합할 수 있다는 장점이 있고 측정 오차가 없다면 가장 이상적인 데이터 결합의 형태라고 볼 수 있다. 반면, 통계적 매칭은 각 개체에 대한 고유 번호가 모든 데이터에 존재하지 않을 때 고려할 수 있는 방식이다. 통계적 매칭은 같은 모집단을 공유하고 있다는 가정에서 시작되며, 매칭 단계에서 각각의 데이터 내에서 공통된 변수를 기준으로 변수 간의 정보를 통계적으로 계산하여 매칭된 값을 도출한다(안경민, 2021).

2.2. 통계적 매칭

2.2.1. 통계적 매칭 개념

통계적 매칭은 수령자 데이터와 제공자 데이터라는 두 가지 이상의 데이터를 결합하는 과정에서 이들을 연계할 공통의 고유정보가 없을 때 사용하는 방법이다. 이 데이터들을 연계하기 위해서는 선행연구를 바탕으로 고유값을 공통변수로 개발하여 설정하고, 통계식을 사용하여 데이터들을 결합하게 된다(오미애, 2015). 대표적인 통계적 매칭 방법은 단계적 매칭(van Pelt, 2001), K-최근접이웃(김희경, 2010), 회귀분석(Ingram et al., 2000), 회귀분석과 K-최근접이웃의 결합(정성석 등, 2004), 랜덤 핫덱(Singh et al., 1990) 등의 방법이 있다. 통계적 매칭에 사용되는 데이터들은 모집단을 공유하고 있지만, 데이터의 수집 방법, 시점, 내용 등이 상이하기 때문에 특정 통계적 매칭 방법만으로 가지고 성능을 판단할 수는 없다. 따라서 통계적 매칭에 적절한 방법론을 찾기 위해서는 선행연구와 시뮬레이션을 통해 적절한 방법을 찾아야 한다.

한편, 통계적 매칭을 수행하기 위해서는 두 가지 가정을 충족하여야 한다(Rässler, 2004, 오미애, 2015). 첫째, 두 데이터에서 나타나는 공통변수의 평균에 유의한 차이가 존재해서는 안된다. 통계적 매칭에 사용되는 데이터는 모집단을 공유하고 있기 때문에 공통변수가 같다라는 가정이 성립되어야 매칭이 진행될 수 있다. 둘째, 공통변수인 'X'가 주어져 있을 때, 수령자 데이터와 제공자 데이터의 고유변수인 'Y'와 'Z'는 조건부 독립(Conditional Independence)을 만족해야 한다. 이 같은 조건부 독립을 가정하는 이유는 수령자 데이터와 제공자 데이터를 결합한(X, Y, Z)의 결합확률분포인 $\hat{f}_{X,Y,Z}(X, Y, Z)$가 추정되어야 하기 때문이다.

$$\begin{aligned}\hat{f}_{X,Y,Z}(X, Y, Z) &= f_{Z,X}(Z, X) f_{Y|X}(Y|X) \\ &= f_{Z|X}(Z, X) f_X(X) f_{Y|X}(Y|X) \\ &= f_{Z|X}(Z|X) f_{Y,X}(Y, X)\end{aligned}$$

	X_1	X_2	Z
1	F	54	22
2	M	21	17
3	F	48	15
4	F	33	14
5	M	63	13
6	F	29	15
7	M	36	19
8	M	55	24
9	F	50	26
10	F	27	18

	X_1	X_2	Y
1	F	27	22
2	M	35	19
3	M	41	47
4	F	61	41
5	F	52	17
6	F	39	26

수령자A	제공자b	X_1^A	X_2^A	X_1^B	X_2^B	Y	Z
1	2	F	27	M	21	22	17
2	8	M	35	M	55	19	24
3	5	M	41	M	63	47	13
4	6	F	61	F	29	41	15
5	4	F	52	F	33	17	14
6	2	F	39	M	21	26	17

〈그림 2〉 랜덤 핫덱 예시

조건부 매칭 분포는 다음과 같은 수식으로 표현된다.

$$\hat{f}_{Z,\,Y|X}(z,y|x) = f_{Z|X}(z|x) f_{Y|X}(y|x)$$

통계적 매칭에는 앞서 제시한 바와 같이 다양한 방법이 사용되고 있지만, 본 연구에서는 랜덤 핫덱(Random Hotdeck)과 마할라노비스(Mahalanobis) 거리 함수를 사용하여 제공자 데이터와 수령자 데이터를 결합하였다. 랜덤 핫덱은 결측 자료를 무작위로 선택하여 매칭하는 방법으로 매칭을 위한 공통변수의 일반적인 공통 특성에 따라 층화한 동질적인 그룹을 형성하고 랜덤하게 데이터를 추출하여 대체한다.

예를 들어, 그림과 같이 수령자 데이터(A)는 6개의 관측치에서 X_1, X_2, Y의 3개 변수를 가지고 있고, 제공자 데이터(B)는 10개의 관측에서 X_1, X_2, Z의 3개 변수가 있다고 가정하면, 이들의 공통변수는 X_1, X_2 이고 각각의 유일변수는 Y, Z이다. 대체군이 없다는 가정에서 랜덤 핫덱은 수용자 데이터(A)가 제공자 데이터(B) 10개로부터 무작위로 데이터를 할당받는다. 이론적인 통계적 매칭의 산출식은 $n_{B^A}^{n}$으로 10^6가지의 조합이 된다. 즉, 제공자 데이터 Z에 대한 경우의 수는 $(n_M^B)^{n_M^A} + (n_F^B)^{n_F^A} = 4^2 + 6^4 = 1312$으로 10^6 가지가 가능한 분포로 나타난다. 하지만 여기서 수령자 데이터(A)와 제공자 데이터(B)의 공통변수 X_1를 대체군으로 정의한다면 제공자 데이터(B)는 수령자 데이터(A)의 공통변수 X_1가 대체군으로 고정되어 랜덤하게 선택되며, 선택이 폭이 급격하게 줄어든다.

한편, 거리함수로 고려된 마할라노비스 거리는 표본의 점과 분포 사이를 측정한 값으로 다차원의 단위 공간으로서 마할라노비스 공간을 정의하고 임의의 대상이 그 공간으로부터 얼마나 떨어져 있는가를 거리로 나타낸다. 마할라노비스 거리는 변수의 표준화뿐만 아니라 변수 사이의 상관관계도 거리하고 있어 비모수 거리를 측정하는데 적절하다. 따라서 마할라노비스 거리의 원리는 데이터를 통계적으로 더 정확하게 추론하는 과정에서 주로 활용된다.

2.2.2. 통계적 매칭 평가 방법

통계적 매칭의 결과를 평가하는 방법은 서로 다른 데이터를 통계적 방법에 따라 매칭하고 결합하는 미관측된 자료를 의미하는 MCAR(Missing Completely At Random)의 결측 매커니즘을 따른다(D'Orazio et al., 2006). Rassler (2002), 변종석 등(2013)은 다양한 출처 자료 처리 및 통계생산 방안을 통해 통계적 매칭의 주요 평가 방법을 정리하였다. 첫째, 결합 데이터는 관측된 데이터와 실측된 데이터의 참값이 일치해야 함을 가정한다. 이 값은 이상적인 수치로 정확한 확인은 불가능하지만, 논리적인 수준에서 평균, 분포, 편차 등을 통해 확인할 수 있다. 둘째, 통계적으로 모든 변수의 결합 분포는 결합 데이터에 반영되어야 한다는 것을 가정한다. 이러한 추정은 관측 데이터인 통계적 매칭 데이터(X, Y, Z)가 실측 데이터인 $f(x, y, z)$의 표본으로 생각할 수 있으며, 표본의 특성을 추론하기 위한 일반적인 유의 표본(General Purposive Sample)으로 사용할 수 있다. 셋째, 변수 상관 구조, 원시 데이터의 모든 변수의 주변 분포와 결합 분포가 통계적 매칭 데이터에서도 유지된다고 가정한다. 이 가정만으로 통계적 매칭 데이터를 이용한 추론의 적절성을 판단하기에는 무리가 있으나, 통계적 매칭 데이터가 어떤 주요 특성이 있는지를 확인할 수는 있다. 따라서 매칭 결과가 원시 데이터의 성질을 그대로 유지하는 대표성 확인을 통해 원시 데이터의 고유변수와 평균, 분산 등의 모수가 같은지를 살펴보거나 변수들의 관계를 측정하는 상관관계, 공분산 및 분포 등을 살펴봄으로써 동일한 속성을 유지하는지 평가할 수 있다. 또한, 서로 다른 조사 자료를 연계해 생성된 통계적 매칭 데

이터는 같은 분포로부터 생성된 표본이 아니기 때문에 매칭 후 실제 데이터와 생성된 매칭 데이터의 기본 모형 간의 불일치 정도를 감소시키는 것이 매우 중요한 과정이다. 즉 통계적 매칭 데이터는 매칭 잡음의 크기가 작아야 추론 성질이 잘 유지된 결과로 평가할 수 있다.

2.3. 통계적 매칭 활용 사례

경영학 분야에서 통계적 매칭은 주로 성향점수매칭(Propensity Score Matching)과 거리함수를 활용하는 방법이 활용되며, 최근에는 이 방법들을 조합하여 통계적 매칭의 성능을 높이는 방법들이 사용되고 있다. 성향점수매칭은 Rosenbaum and Rubin (1983)이 제안한 통계적 매칭 방법으로 이미 관찰된 자료에서 처리집단과 유사한 통제집단 그룹을 찾아 처리집단의 인과효과와 통제집단의 인과효과의 차이를 분석하고자 하는 방법이다. 성향점수는 정규분포, 로지스틱 누적분포, 선형 함수 등을 고려하여 매칭값이 도출되며, 이 값을 통해 처리집단과 통제집단이 구분된다. 이 방법은 이중차분법(Difference-in-Differences), 역확률가중치(Inverse Probability Weighting), 통제집단합성법(Synthetic Control Method) 등과 함께 집단 간 비교연구에서 활용되고 있다. 이중차분법은 처치집단과 통제집단의 특성이 동질하다는 가정이 성립한다면 두 집단 간의 차이를 이용하여 효과의 인과적 관계를 정교하게 분석할 수 있어 경영학 분야에서 가장 많이 사용되고 있는 방법이다(이유진, 2021). 역확률가중치 방법은 각 피험자가 해당 집단에 속할 확률의 역수를 가중치로 활용하여, 가중치를 통해 처치집단과 통제집단의 사전 특성의 분포를 비슷하게 조정해주는 방법이다. 이 방법은 자료의 구조나 변수의 특성에 제한을 적게 받는다는 장점 때문에 광범위하게 활용되고 있다(Curtis et al., 2007). 통제집단합성법은 이중차분법과 매우 유사한 특성을 보이지만 통제집단의 특성을 반영하는 변수에 대한 가중평균을 구하여 합성된 통제집단을 구성한다는 점에서 차이가 있다(김경훈, 2022).

경영학 분야의 연구에서 성향매칭점수를 활용한 연구는 다음과 같다. 이유진(2021)은 수도권과 비수도권에 소재한 기업의 산업단지 입주에 따른 기업의 경제적 편익에 관한 연구에서 성향점수매칭과 이중차분법을 활용하였으며, 산업단지 입주특성에 따른 기업의 경영 및 혁신성과를 분석하였다. 이 연구에서는 로짓모형을 이용해 기업의 산업단지 입주 성향점수를 추정한 후 성향점수가 가장 근접한 개체 간의 매칭인 최인접개체매칭(Nearest Neigibhor Matching)을 적용해 처리집단과 비교집단 내 개체가 일대일로 대응되도록 하였다. 매칭성능을 확인하기 위해 평균과 표준편차를 산출하여 평균차이 검정을 수행하였으며, 이중차분모형으로 산업단지 입주가 기업의 생산성과 고용증가율에 미치는 영향을 도출하였다.

이준원(2019)은 성향점수매칭을 통해 기술금융 혜택을 받은 중소기업의 산업분야, 업력, 종업원 수, 자산 및 자본규모와 유사한 일반 중소기업을 선별하여 기술금융 중소기업과 일반 중소기업의 3년간 경영성과를 비교분석 하였다. 매칭 성능으로 데이터의 균형을 판단하기 위해 Hansen and Bowers (2008)의 검정통계량을 사용하였으며, 카이제곱분포(Chi-square) 검정, 다변량 불균형지표를 통해 데이터의 불균형성을 검증하였다.

한편, 경영학 분야에서는 거리함수나 이를 응용한 통계적 매칭 방법은 사례가 충분하지 않으나 다른 사회과학 분야에서는 적극적으로 수용하고 있는 양상을 보인다. 그 이유는 경영학 분야에서는 주로 경영환경을 둘러싼 내・외부 요인의 적용 여부에 따른 성과창

출을 밝히는 연구에 초점이 맞춰져 있기 때문으로 판단된다. 하지만 거리함수는 비모수 통계까지 적용이 가능하고 높은 정확도, 오류 데이터 문제 최소화, 데이터 가정에 대한 생략 가능으로 통계적 매칭의 우수성이 입증된 방법이다. 거리함수를 활용한 선행연구를 살펴보면 김성호, 조성빈(2005)은 고객관계관리와 관련된 전략을 도출하기 위해 신용카드 데이터를 확보하여 마할라노비스 거리 함수를 적용한 통계적 매칭을 수행하였다. 이 분석에 사용된 신용카드 데이터의 공통변수는 연회비, 현금환불, 신용카드 사용가능 장소, 비상차량지원 무료전화 서비스, 공항 시내 간 리무진 서비스, 24시간 의료/법률상담 서비스였고, 매칭 성능을 검증하기 위해 비공통 변수를 증가시키는 시뮬레이션을 수행하였다. 성능평가 방법에는 변수 간의 상관계수, 평균제곱오차 지표를 활용하였다. 시뮬레이션 결과, 마할라노비스 거리가 공통변수 간에 실재하는 통계적 종속을 모델에 반영함으로써 예측의 성과를 높이는데 유효한 역할을 한다는 결과를 도출하였다.

또한, Ferrando and Mulier (2015)은 금융위기 기간 동안 유럽 기업의 자금조달 상황의 인식 정도를 분석하기 위해 고워 거리 함수를 활용한 통계적 매칭을 수행하였다. 통계적 매칭에는 유럽 비영리협회 SAFE의 서베이 데이터와 Bureau van Dijk의 AMADEUS 데이터베이스가 사용되었다. SAFE는 기업 규모, 사업 부문, 기업 자율성, 이직률, 기업 연령, 소유권 등 기업 수준의 정보가 포함되어 있으며, AMADEUS는 대차대조표 정보가 포함되고 천만 개 이상의 공공, 민간 기업 재무 정보를 포함하고 있다. 통계적 매칭을 위한 공통변수로는 SAFE의 구조적 특성인 국적, 사업 부문, 이직률, 기업 연령을 고려하였다. 고워 거리 함수를 활용한 통계적 매칭결과 31%의 데이터가 완벽하게 매칭되었고 71%의 데이터가 오차범위 0.01에서 매칭되어 고워 거리 함수의 적절성을 확인할 수 있다. 또한, 매칭된 데이터는 기술통계를 통해 평균, 중위값, 최소값, 최대값을 확인하였고, 재무 데이터에 대한 평균차이 검정을 통해 통계적 매칭의 적절성을 확인하였으며, 프로빗 모델을 활용하여 기업의 연도별로 수익성, 유동성, 레버리지, 정보 비대칭의 재무적 특성을 분석하였다.

그 외 통계적 매칭을 활용하고 있는 분야를 살펴보면 다음과 같은 방법을 사용하고 있는 것을 확인할 수 있다. 교육학 분야에서 금종예, 모영민(2022)은 초중고의 사교육비 조사를 위한 통계적 매칭을 수행하였다. 통계적 매칭에는 마할라노비스 거리를 이용한 k-최근접이웃 방법이 사용되었으며, 매칭성능을 평가하기 위해 평균차이 검정, 커널 밀도 분포, 회귀분석을 활용하여 매칭 데이터를 시뮬레이션하였다. 또한, 오미애 등(2014)은 보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안을 제시하면서 맨하튼 거리, 마할라노비스 거리, 정확 거리의 3가지 거리함수와 최근접 이웃 핫덱, 랜덤 핫덱의 2가지 기법을 활용한 통계적 매칭을 수행하였으며, 평균제곱오차를 활용하여 매칭성능을 평가하였다.

이같이 사회과학 연구분야에서 다양한 통계적 매칭이 시도되고 있는 만큼 평가방법에서도 차이가 나타난다. 하지만 앞서 살펴본 선행연구에서 공통적으로 제시하고 있는 중요한 검정은 통계적 매칭 후 만들어진 매칭 데이터가 실제 데이터를 적절하게 반영하고 있는지를 확인하는 것이다. 따라서 본 연구에서는 기준 데이터를 임으로 분리하여 통계적 매칭이 적절하게 진행될 수 있는지 시뮬레이션을 수행하여 알고리즘의 적정성을 판단하고자 한다. 알고리즘의 적정성은 평균차이 검정, 데이터 분포, 회귀분석을 통해 검정을 진행한다. 이 과정이 적정하다고 판단된 후, 기준 데이터와 매칭 데이터를 활용하여 통계적 매칭을

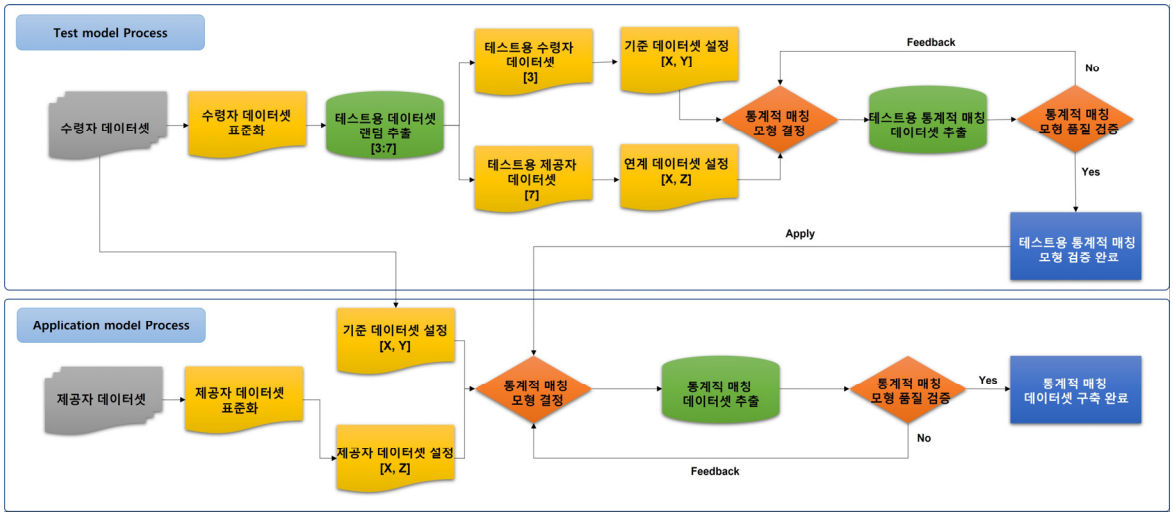
수행하고 매칭되어 새롭게 만들어진 데이터를 검정하는 과정으로 분석을 진행하고자 한다.

3. 연구방법

본 연구에서는 선행연구를 바탕으로 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사를 활용하여 <그림 3>과 같은 절차로 통계적 매칭을 수행하였고, R 프로그래밍의 Hmisc, StatMatch, Matching 라이브러리가 사용되었다. 개괄적으로 분석 과정은 테스트 모형 프로세스와 적용 모형 프로세스로 구분된다. 테스트 모형 프로세스에서는 수여자 데이터를 활용하여 테스트용 통계적 매칭 알고리즘을 구축하고 검증하였으며, 적용 모형 프로세스에서는 검증된 통계적 매칭 모형을 활용하여 수여자 데이터와 제공자 데이터의 통계적 매칭을 수행하였다. 이에 대한 구체적인 테스트 모형 프로세스는 다음과 같다. 첫째, 공통변수 표준화를 수행하였다. 공통변수 표준화는 수령자 데이터와 제공자 데이터가 공통으로 가지고 있는 데이터를 일치시키는 것으로 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사가 공통으로 가지고 있는 지역, 업력, 업종, 상장시장, 매출, 기업규모, R&D 비용을 수령자 데이터인 2020 벤처기업정밀실태조사를 기준으로 <표 1>과 같이 전처리하였다.

둘째, 수여자 데이터를 활용하여 테스트용 통계적 매칭 모형을 구현하였다. 통계적 매칭은 모집단을 공유하는 서로 다른 집단에서 데이터를 수집한다는 것을 가정하고 있으므로 두 집단의 데이터는 유사성이 검증되어야 한다. 데이터의 유사성을 검증하는 방법은 수집된 데이터에서 무작위로 데이터를 추출하여 모형을 구축하고 검증하는 것이다. 이를 위해 수여자 데이터를 테스트용 수여자 데이터 30%(286개)와 테스트용 제공자 데이터 70%(670개)로 무작위 추출하였다. 수여자 데이터와 제공자 데이터의 차이가 발생하는 것은 많은 제공자 데이터에서 적절한 데이터를 선별하는 것이 매칭의 정확도를 높일 수 있기 때문이다. 또한, 수여자 데이터를 바탕으로 테스트용 수여자 데이터의 공통변수 X, 고유변수 Y를 설정하고, 테스트용 제공자 데이터의 공통변수 X, 제공변수 Z를 설정하였다.

셋째, 통계적 매칭을 위한 모형을 결정하였다. 통계



〈그림 3〉 통계적 매칭 프로세스

적 매칭을 수행하는데 있어서 고정값, 변동값, 함수 설정 등은 매칭 성능에 직접적인 영향을 미친다. 본 연구에서는 기업의 규모를 파악할 수 있는 매출과 종업원수를 고정값으로 설정하고, 업력, 업종, 상장시장을 변동값으로 설정하였고, 마할라노비스 거리 함수를 사용하여 통계적 매칭을 수행하였다.

넷째, 통계적 매칭의 모형의 적정성을 검증하기 위하여 커널 밀도 함수 추정과 평균차이 검정을 통해 원시 데이터와 매칭 데이터를 비교분석 하였다. 커널 밀도 함수는 비모수 밀도추정 방법 중 하나로서 커널 함수를 이용하여 히스토그램 방법의 문제점을 개선한 방법이다. 이와 같은 과정을 통해 검증된 테스트용 통계적 매칭 알고리즘을 구현하였다.

이를 바탕으로 적용 모형 프로세스를 정리하면 다음과 같다. 첫째, 수여자 데이터와 제공자 데이터에서 수여자 데이터의 공통변수 X, 고유변수 Y를 설정하고, 제공자 데이터의 공통변수 X, 제공 변수 Z를 설정하였다. 둘째, 앞서 수여자 데이터를 활용하여 구축한 통계적 매칭 모형을 적용하여 수여자 데이터와 제공자 데이터의 통계적 매칭을 수행하였다. 통계적 매칭에서 매출과 종업원수는 고정값으로 사용하였으며, 업력, 업종, 상장시장은 변동값으로 사용하였다. 셋째, 통계적 매칭 모형의 적정성은 커널 밀도 함수 추정과 평균차이 검정을 통해 수여자 데이터의 원시 데이터와 제공자 데이터의 매칭 데이터를 비교분석 하였다. 이러한 과정을 통해 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사의 통계적 매칭 알고리즘을 구현하였다.

3.1. 데이터 전처리

본 연구 분석을 위해 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사 데이터의 공통변수를 확인하였다. 두 자료에서 측정된 공통변수는 지역, 업력, 업종, 상장시장, 매출, 종업원수, 연구인력, R&D 비용으로 나타났다. 하지만 각 조사에서 측정하는 방법은 다소 상이한 부분이 있다. 예를 들어 업종을 측정하는 경우, 2020 벤처기업정밀실태조사에서는 비율을 구분하여 측정하고 있지만 2020 한국기업혁신조사에서는 설립년도로 측정하고 있다. 이 경우 같은 내용을 측정하고 있지만 측정 방법이 상이하여 통계적 매칭을 수행할 수 없다. 따라서 공통변수를 표준화해주는 전처리 과정이 필요하다.

통계적 매칭을 위한 데이터 전처리에서는 수여자 데이터가 기준이 되므로 요인의 대부분을 2020 벤처기업정밀실태조사의 측정 기준에 맞춰 2020 한국기업혁신조사를 일치시키는 방향으로 진행하였다. 업종의 경우, 기여자 데이터에서 확인되지 않거나 제공될 수 없는 데이터는 삭제하여 공통변수를 구성하였다. 상장시장의 경우, 벤처기업 특성상 데이터가 많이 확보될 수 없어 코스피와 코스닥은 통합하였고, 기업 규모에서도 500인 이상은 300인~499인과 통합하였다. 또한, 데이터 전처리 중 측정된 값에 데이터의 유사성이 부족하거나 이상치가 있는 경우 삭제하였다. 그 결과, 최종적으로 분석에 사용된 데이터는 2020 벤처기업정밀실태조사 956개, 2020 한국기업혁신조사 2020 벤처기업정밀실태조사 2,360개가 확보되었다. 일반적으로 기여자 데이터는 수여자 데이터보다 더 많은 수가 제공되어야 매칭이 될 수 있다. 또한, 수여자 데이터와 기여자 데이터에서 명목척도와 서열척도로 측정되고 있는 지역, 업력, 업종, 상장시장, 매출, 종업원수를 직접적으로 비교하는 것은 불가능하다. 따라서 이 변수의 경우 더미변수로 처리하였으며, 더미변수의 평균을 비교분석을 수행하였다. 통상적으로 더미변수는 자체적인 의미를 가지고 있지 않다고 여겨지고 있으나 통계적 매칭의 경우 절대적인 측정방안이 마련되

〈표 1〉 공통변수의 표준화 전처리

구분	2020 벤처기업정밀실태조사	2020 한국기업혁신조사	공통변수 표준화
지역	1. 서울/인천/경기 2. 대전/세종/충청/강원 3. 부산/경남/울산 4. 대구/경북 5. 광주/전라/제주	1. 서울, 2. 인천, 3. 경기, 4. 대전, 5. 세종, 6. 충청, 7. 강원, 8. 부산, 9. 경남, 10. 울산, 11. 대구, 12. 경북, 13. 광주, 14. 전라, 15. 제주	1. 서울/인천/경기 2. 대전/세종/충청/강원 3. 부산/경남/울산 4. 대구/경북 5. 광주/전라/제주
업력	1. 창업 3년 이하 2. 4~10년 3. 11~20년 4. 21년 이상	설립년도	1. 창업 3년 이하 2. 4~10년 3. 11~20년 4. 21년 이상
업종	1. 에너지/화학/정밀 2. 의료/제약 3. 컴퓨터/반도체/전자부품 4. 통신기기/방송기기 5. 기계/자동차/금속 6. 음식료/섬유/비금속/기타제조 7. 소프트웨어개발/IT기반서비스 8. 정보통신/방송서비스 9. 도소매/연구개발서비스/기타서비스 10. 기타	10. 식료품, 11. 음료, 13. 섬유, 14. 의류, 15. 가죽/잡화, 16. 목재/나무, 17. 제지, 18. 인쇄/기록매체, 19. 석유정제, 20. 화학, 21. 제약, 22. 고무/플라스틱, 23. 비금속 광물, 24. 1차 금속, 25. 금속가공, 26. 전자/컴퓨터, 27. 의료/정밀, 28. 전기장비, 29. 기타 기계장비, 30. 자동차, 31. 기타 운송장비, 32. 가구, 33. 기타 제품, 34. 기계/장비 수리업,	1. 에너지/화학/정밀 2. 의료/제약 3. 컴퓨터/반도체/전자부품 4. 기계/자동차/금속 5. 섬유/비금속/기타제조
상장시장	코스피 2. 코스닥 3. 코넥스 4. 해당사항 없음	1. 거래소 상장기업 2. 코스닥 상장기업 3. 코넥스 상장기업 4. 해당사항 없음	1. 코스피/코스닥 2. 코넥스 3. 해당사항 없음
매출	2019년 매출액(백만원)	2019년 매출액(백만원)	2019년 매출액(백만원)
종업원수	1. 10~49인 2. 50~99인 3. 100~299인 4. 300인~499인 5. 500인 이상	1. 10~49인 2. 50~99인 3. 100~299인 4. 300~499인 5. 500인 이상	1. 10~49인 2. 50~99인 3. 100~299인 4. 300 이상
연구인력	2019년 연구개발인력 수	2019년 연구개발인력 비율(%)	2019년 연구개발인력 비율(%)
R&D 비용	2019년 R&D 비용(백만원)	2019년 R&D 비용(백만원)	2019년 R&D 비용(백만원)

어 있지 않기 때문에 전체적인 성능지표를 검정하는 과정에서 보조적 수단으로 활용될 수 있다(금종예, 모영민, 2022).

3.2, 통계적 매칭 시뮬레이션

3.2.1. 사전 시뮬레이션

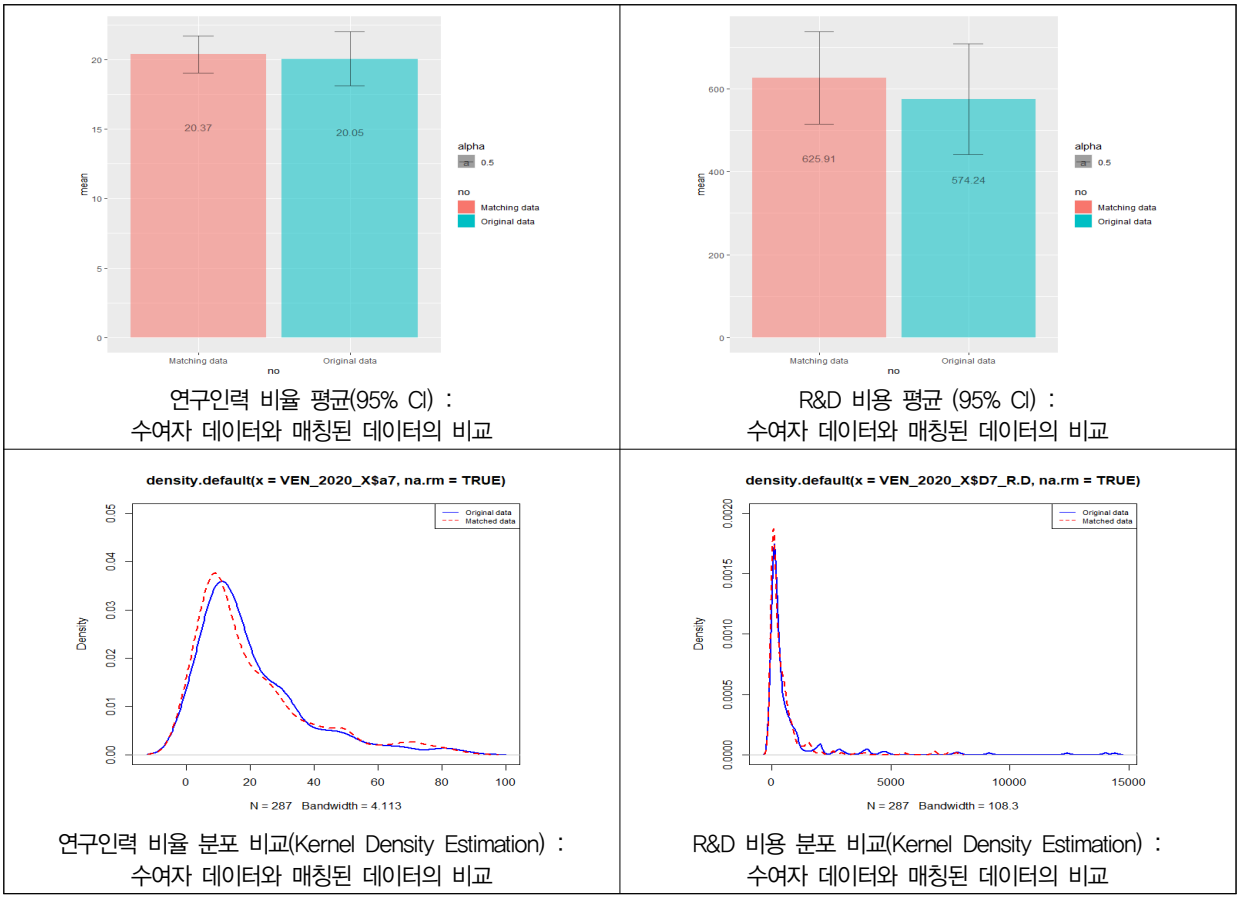
2020 벤처기업정밀실태조사와 2020 한국기업혁신조사의 통계적 매칭을 수행하기 전에 수여자 데이터가 통계적 매칭을 수행하기 적절한가에 대한 시뮬레이션 분석을 수행하였다. 시뮬레이션에는 최종적으로 통계적 매칭을 수행하려고 하는 종속변수인 연구인력과 R&D 비용을 중심으로 살펴보았다. 연구인력과 R&D 비용은 기업혁신의 핵심적인 요소로서 벤처기업의 혁신성과를 창출하는데 중요한 역할을 한다. 수여자 데이터와 기여자 데이터가 통계적 결합을 하기 위해서는 변수 간의 이론적, 현실적 밀접한 관계가 있어야 한다. 본 연구에서 매칭하고자 하는 공통변수인

지역, 업력, 업종, 상장시장, 매출, 종업원수, 연구인력, R&D 비용은 대부분의 기업에서 혁신과 성장을 위한 기초값으로 활용된다.

통계적 매칭 시뮬레이션을 위해서 2020 벤처기업정밀실태조사를 수여자 데이터 30%(286개)와 기여자 데이터 70%(670개)로 분할하여 랜덤 추출하였다. 기여자 데이터를 충분히 확보할수록 통계적 매칭의 성능이 향상될 수 있으므로 기여자 데이터를 수여자 데이터의 2배 이상 확보하였다. 통계적 매칭 과정에서 업종과 종업원수는 완전 일치하도록 분할한 후 지역, 업력, 상장시장, 매출을 공통변수로 하였고 연구인력과 R&D 비용이 매칭되도록 모형을 설계하였다. 통계적 매칭은 랜덤 핫덱과 마할라노비스 거리 함수를 활용하여 분석이 진행되었다. 랜덤 핫덱은 수용자 데이터의 각 관측치에 대해 제공자 데이터의 관측치를 랜덤하게 선택하여 매칭시키는 방법이다. 특히 수용자 데이터와 제공자 데이터의 관측치들은 대개 주어진 일반적인 특성(지형적 특성, 사회적 특성 등)에 따라 동질적인 부분집합으로 그룹화될 수 있다. 따라서 각각의 수용자 관측치에 대해 주어진 지형적 특성 내에서 동일지역의 관측치만이 가능한 제공자로 고려된다. 마할라노비스 거리는 다차원의 단위공간으로서 마할라노비스 공간을 정의하고 임의의 대상이 그 공간으로부터 얼마나 떨어져 있는가를 거리로 나타낸다. 마할라노비스 거리는 변수의 표준화뿐만 아니라 변수 사이의 상관관계도 거리하고 있어 비모수 거리를 측정하는데 적절하다. 이와 같은 방식으로 설정된 수여자 데이터의 공통변수를 기준으로 제공자 데이터인 연구인력과 R&D 비용이 매칭된다. 매칭된 데이터의 품질은 원데이터와 평균, 분포, 다른 변수와의 관계를 비교하여 검증된다.

시뮬레이션 결과는 고유변수 연구인력 비율과 제공변수 R&D 비용으로 검증하였다. 기준 데이터인 수여자 데이터의 연구인력 비율의 평균값 20.05, 새롭게 매칭된 데이터의 값이 20.37이며, 두 값은 통계적으로 유의한 차이가 없는 것으로 나타났다. 또한 수여자 데이터와 매칭된 데이터의 비교에서 연구인력 비율 분포를 커널 밀도 함수로 추정하였을 때도 유사한 형태가 나타나는 것을 확인할 수 있었다. 또한, 기준 데이터의 R&D 비용은 평균 574.2(백만원)이고 매칭된 데이터의 R&D 비용은 625.91(백만원)이며, 두 값은 통계적으로 유의한 차이가 없는 것으로 나타났다. 또한 수여자 데이터와 매칭된 데이터의 비교에서 R&D 비용의 분포를 커널 밀도 함수로 추정하였을 때도 유사한 형태가 나타나는 것을 확인할 수 있었다. 따라서 매칭 데이터의 알고리즘의 성능을 확인하기 위한 선행단계로 2020 벤처기업정밀실태조사를 수여자 데이터 30%와 기여자 데이터 70%로 분할하여 매칭을 진행한 결과, 통계적 매칭이 적절하다는 것을 확인할 수 있었다(<그림 3> 참조).

세부적으로 통계적 매칭의 품질을 판단하기 위해서는 매칭이 된 변수의 분포와 함께 다른 변수의 관계도 살펴봐야 한다. 이는 분석에 포함된 자료는 이론적이고 현실적인 내용을 반영하고 있고 모든 데이터는 통계적 매칭에 활용되기 때문이다. 이러한 맥락에서 지역, 업력, 업종, 상장시장, 매출, 종업원수의 평균값을 비교하였다. 분석결과, 지역의 기준 데이터의 평균은 2.11, 매칭 데이터의 평균은 2.13으로 통계적으로 유의미한 차이를 보이지 않았다. 또한, 업력의 기준 데이터의 평균은 2.72, 매칭 데이터의 평균은 2.75, 업종의 기준 데이터의 평균은 3.60, 매칭 데이터의 평균은 3.66, 상장시장의 기준 데이터의 평균은 3.83, 매칭 데이터의 평균은 3.80, 매출의 기준 데이터의 평균은 21237.91, 매칭 데이터의 평균은 22291.63, 종업원수의 기준 데이터의 평균은 1.54, 매칭 데이터의 평균은 1.57로 모든 값이 통계적으로 유의한 차이를 보이지



연구인력 비율 평균(95% CI) : 수여자 데이터와 매칭된 데이터의 비교

R&D 비용 평균 (95% CI) : 수여자 데이터와 매칭된 데이터의 비교

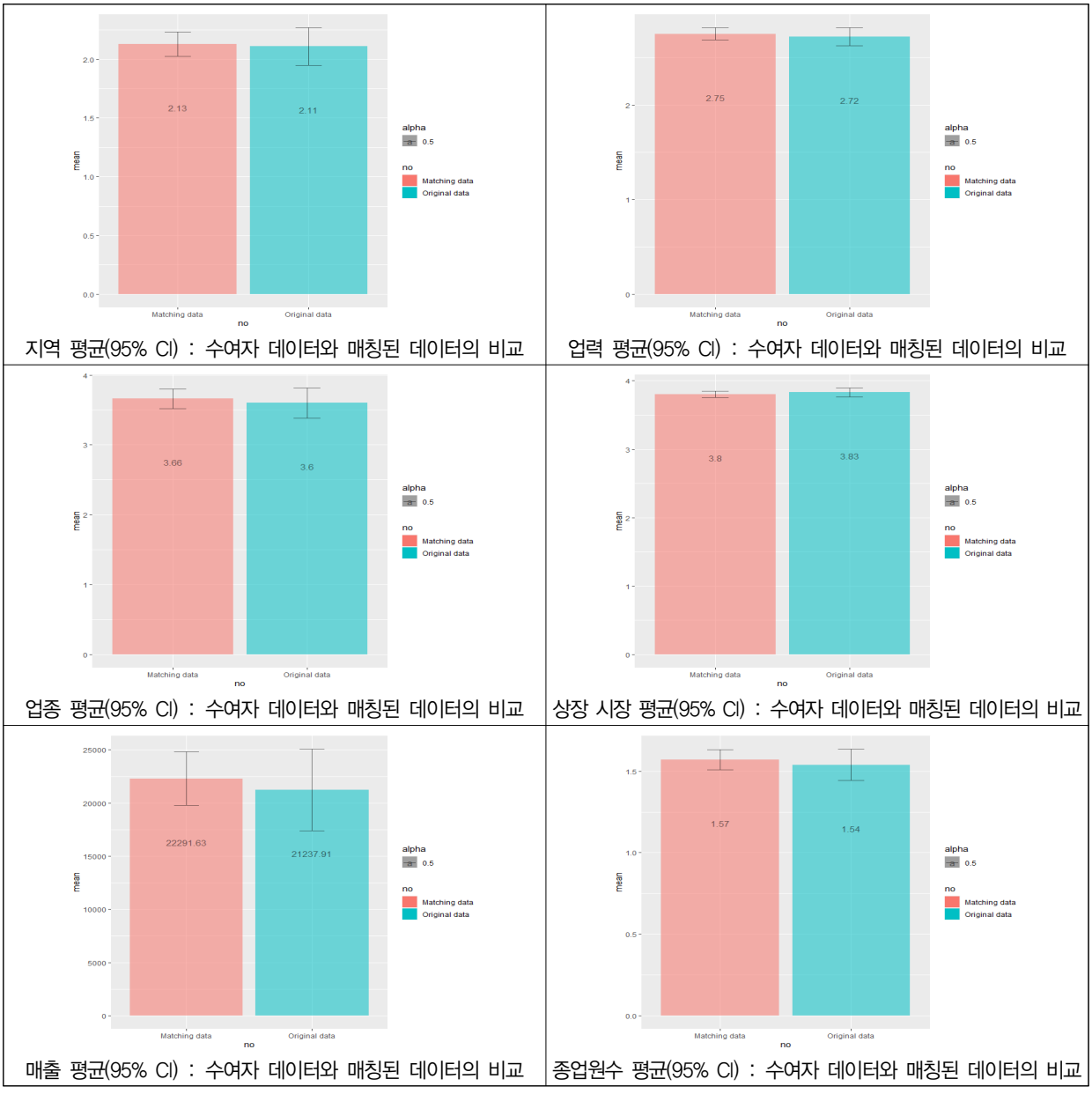
연구인력 비율 분포 비교(Kernel Density Estimation) : 수여자 데이터와 매칭된 데이터의 비교

R&D 비용 분포 비교(Kernel Density Estimation) : 수여자 데이터와 매칭된 데이터의 비교

〈그림 4〉 수여자 데이터를 활용한 통계적 매칭 검증

않았다. 즉, 매칭을 통해 생성된 변수가 분포의 보존이라는 측면에서 전체 수준의 평균에 유의미한 차이가 없을 뿐만 아니라 다른 변수와의 관계 보전이라는 측면에서도 모든 변수와 평균값에 통계적인 차이가 없었다(<그림 4> 참조).

통계적 매칭의 품질을 세밀하게 판단하기 위해서 매칭된 변수의 세부적인 값의 분포와 관계를 살펴보았다. 분석결과, <표 2>와 같이 일부 변수는 통계적으로 유의한 차이가 있었지만, 대부분의 변수는 통계적으로 유의한 차이가 없는 것으로 확인되었다. 지역의 경우, 1. 서울/인천/경기(p>0.429), 2. 대전/세종/충청/강원(p>0.930), 3. 부산/경남/울산(p>0.787), 4. 대구/경북(p>0.780), 5. 광주/전라/제주(p>0.072)에서 모든 변수 차이가 통계적으로 유의하지 않았고, 업력은 1. 창업 3년 이하(p>0.862), 2. 4~10년(p>0.818), 3. 11~20년(p>0.838), 4. 21년 이상(p>0.903)에서 모든 변수 차이가 통계적으로 유의하지 않았다. 또한, 1. 에너지/화학/정밀(p>0.985), 2. 의료/제약(p>0.987), 3. 컴퓨터/반도체/전자부품(p>0.937), 5. 기계/자동차/금속(p>0.982), 6. 섬유/비금속/기타제조(p>0.981)에서 모든 변수의 차이가 통계적으로 유의하지 않았다. 상장시장의 경우, 2. 코넥스(p>0.410)에서는 모든 변수의 차이가 통계적으로 유의한 차이가 없었지만 1. 코스피/코스닥(p<0.05), 3. 해당사항 없음(p<0.05)에서는 통계적으로 유의한 차이가 있었다. 종업원수는 1. 10~49인(p>0.905), 2. 50~99인(p>0.194), 3. 100~299인(p>0.344), 4. 300인



지역 평균(95% CI) : 수여자 데이터와 매칭된 데이터의 비교

업력 평균(95% CI) : 수여자 데이터와 매칭된 데이터의 비교

업종 평균(95% CI) : 수여자 데이터와 매칭된 데이터의 비교

상장 시장 평균(95% CI) : 수여자 데이터와 매칭된 데이터의 비교

매출 평균(95% CI) : 수여자 데이터와 매칭된 데이터의 비교

종업원수 평균(95% CI) : 수여자 데이터와 매칭된 데이터의 비교

〈그림 5〉 수여자 데이터를 활용한 통계적 매칭 상세 검증

이상(p>0.100)에서 모든 변수는 통계적으로 유의한 차이가 없었으며, 매출(p>0.099), 연구인력(p>0.704), 그리고 R&D 비용(p>0.841) 모두 통계적으로 유의한 차이가 없는 것으로 나타났다. 따라서 대부분의 세부 변수에서 기준 데이터와 매칭데이터 간에 통계적으로 유의한 차이가 발견되지 않아 매칭 성능이 우수하다는 것을 확인할 수 있었다.

매칭된 데이터가 다른 변수들과의 관계에서도 기준 데이터와 유사한 결과를 보이는지 확인하기 위해서 기준 데이터의 지역, 업력, 업종, 상장시장, 매출, 종업원수를 독립변수로 하고 연구인력 비율, R&D 비용을 종속변수로 하여 OLS 회귀분석을 실시하였다. 회귀

〈표 2〉 기준 데이터와 매칭 데이터의 평균차이 검정

	구분	기준 데이터	매칭 데이터	F	Sig.
지역	1. 서울/인천/경기	50.00%	46.69%	1.255	0.429
	2. 대전/세종/충청/강원	18.18%	18.47%	0.031	0.930
	3. 부산/경남/울산	12.59%	11.85%	0.292	0.787
	4. 대구/경북	11.54%	10.80%	0.313	0.780
	5. 광주/전라/제주	7.69%	12.20%	13.260	0.072
업력	1. 창업 3년 이하	5.24%	5.57%	0.122	0.862
	2. 4~10년	34.97%	35.89%	0.213	0.818
	3. 11~20년	37.06%	36.24%	0.168	0.838
	4. 21년 이상	22.73%	22.30%	0.060	0.903
업종	1. 에너지/화학/정밀	17.48%	17.42%	0.001	0.985
	2. 의료/제약	13.99%	13.94%	0.001	0.987
	3. 컴퓨터/반도체/전자부품	21.68%	21.95%	0.025	0.937
	5. 기계/자동차/금속	22.73%	22.65%	0.002	0.982
	6. 섬유/비금속/기타제조	24.13%	24.04%	0.002	0.981
상장시장	1. 코스피/코스닥	13.64%	7.67%	22.312	0.020
	2. 코넥스	1.40%	0.70%	2.730	0.410
	3. 해당사항 없음	84.97%	91.64%	25.811	0.013
매출	2019년 매출액(백만원)	25310.490	20831.641	10.178	0.099
종업원수	1. 10~49인	60.49%	60.98%	0.057	0.905
	2. 50~99인	15.38%	19.51%	6.828	0.194
	3. 100~299인	22.38%	19.16%	3.601	0.344
	4. 300인 이상	1.75%	0.35%	11.057	0.100
연구인력	2019년 연구개발인력 비율	19.917	20.460	1.383	0.704
R&D 비용	2019년 R&D 비용(백만원)	695.563	670.700	0.061	0.841

분석은 변수 간의 관계와 영향도를 동시에 살펴볼 수 있는 분석 방법으로 기준 데이터와 매칭 데이터를 종합적으로 검증하기에 적합하다. 여기에서 회귀분석은 독립변수와 종속변수의 영향 관계를 검증보다는 기준 데이터와 매칭 데이터의 결과값이 일관성 있게 도출되고 있는지에 초점이 맞춰진다.

연구인력 비율을 종속변수로 회귀분석 결과는 <표 3>과 같다. 지역의 경우, 2. 대전/세종/충청/강원, 3. 부산/경남/울산, 4. 대구/경북의 회귀분석 결과값이 유사하다는 점을 확인할 수 있었고, 5. 광주/전라/제주는 모형(1)과 모형(2)가 다소 상이한 것으로 나타났다. 업력의 경우, 2. 4~10년, 3. 11~20년, 4. 21년 이상의 모든 회귀분석 결과값이 유사하다는 것을 확인할 수 있었다. 업종의 경우, 2. 의료/제약, 3. 컴퓨터/반도체/전자부품은 유사한 결과가 나타났으나 5. 기계/자동차/금속, 6. 음식료/섬유/비금속/기타제조는 모형(1)과 모형(2)의 차이가 발견되었다. 상장시장의 경우에도 3. 코넥스는 모형(1)과 모형(2)가 통계적으로 유사하나 4. 해당사항 없음의 경우 모형(1)과 모형(2)의 차이가 발견되었다. 종업원수에서는 2. 50~99인, 3. 100~299인은 모형(1)과 모형(2)의 차이가 발견되었고, 4. 300~499인은 모형(1)과 모형(2)의 차이가 없는 것으로 확인되었

〈표 3〉 기준 데이터 모형과 매칭 데이터 모형의 연구인력 비율 회귀분석 비교

		기준 데이터 모형(1)			매칭 데이터 모형(2)		
(D.V. : 연구인력 비율)		Coef.	S.E.	Sig.	Coef.	S.E.	Sig.
	상수	28.689	5.777	0.000	50.570	6.254	0.000
지역	1. 서울/인천/경기(r)						
	2. 대전/세종/충청/강원	2.694	2.604	0.302	−0.631	2.629	0.811
	3. 부산/경남/울산	−1.226	3.041	0.687	−2.108	3.144	0.503
	4. 대구/경북	−3.659	3.326	0.272	−4.531	3.188	0.156
	5. 광주/전라/제주	1.591	3.658	0.664	−8.241	3.025	0.007
업력	1. 창업 3년 이하(r)						
	2. 4~10년	−0.558	4.485	0.901	−4.601	4.310	0.287
	3. 11~20년	−3.501	4.551	0.442	−4.328	4.341	0.320
	4. 21년 이상	−2.968	4.867	0.543	−6.812	4.719	0.150
업종	1. 에너지/화학/정밀						
	2. 의료/제약	0.962	3.516	0.785	−4.933	3.473	0.157
	3. 컴퓨터/반도체/전자부품	4.237	3.040	0.164	−1.055	3.056	0.730
	5. 기계/자동차/금속	−4.549	3.178	0.154	−12.411	3.024	0.000
	6. 음식료/섬유/비금속/기타제조	−7.501	3.049	0.015	−11.788	3.007	0.000
상장시장	2. 코스피/코스닥						
	3. 코넥스	5.180	8.587	0.547	−19.754	11.677	0.092
	4. 해당사항 없음	−0.575	3.264	0.860	−12.565	4.010	0.002
종업원수	1. 10~49인						
	2. 50~99인	−6.481	2.901	0.026	−12.828	2.861	0.000
	3. 100~299인	−9.009	3.452	0.010	−4.793	4.210	0.256
	4. 300~499인	−8.712	7.958	0.275	−24.526	16.430	0.137
매출	2019년 매출액(백만원)	0.000	0.000	0.164	0.000	0.000	0.141
R Square		.191			274		

다. 또한, 2019년 매출액(백만원)은 모형(1)과 모형(2)의 차이가 없는 것으로 나타났다.

R&D 비용을 종속변수로 회귀분석 결과는 <표 4>와 같다. 지역의 경우, 4. 대구/경북은 모형(1)과 모형(2)의 회귀분석 결과값이 유사하게 나타났지만 2. 대전/세종/충청/강원, 3. 부산/경남/울산, 5. 광주/전라/제주의 회귀분석 결과값에는 모형(1)과 모형(2)의 차이가 있는 것으로 확인되었다. 업력의 경우에는 2. 4~10년, 3. 11~20년, 4. 21년 이상의 회귀분석 결과값이 유사하다는 것을 확인하였다. 업종의 경우에는 2. 의료/제약, 5. 기계/자동차/금속, 6. 음식료/섬유/비금속/기타제조의 모형(1)과 모형(2)의 회귀분석 결과값이 유사하게 도출되었지만 3. 컴퓨터/반도체/전자부품은 차이가 있는 것으로 나타났다. 상장시장의 경우 3. 코넥스는 모형(1)과 모형(2)의 회귀분석 결과값이 유사하지만, 4. 해당사항 없음은 상이한 것으로 나타났고, 종업원수의 경우에도 3. 100~299인의 회귀분석 결과값은 모형(1)과 모형(2)가 유사하였지만 2. 50~99인, 4. 300~499인은 차이가 있다는 것을 확인할 수 있었다. 또한, 매출의 경우에도 모형(1)과 모형(2)의 차이가 존재하였다.

〈표 4〉 기준 데이터 모형과 매칭 데이터 모형의 R&D 비용 평균 회귀분석 비교

		기준 데이터 모형(1)			매칭 데이터 모형(2)		
(D.V. : R&D 비용)		Coef.	S.E.	Sig.	Coef.	S.E.	Sig.
	상수	−61.592	454.999	0.892	1907.373	417.327	0.000
지역	1. 서울/인천/경기(r)						
	2. 대전/세종/충청/강원	−164.264	205.084	0.424	−431.114	175.420	0.015
	3. 부산/경남/울산	−202.085	239.509	0.400	−665.241	209.827	0.002
	4. 대구/경북	−121.270	261.959	0.644	−391.151	212.768	0.067
	5. 광주/전라/제주	−347.924	288.163	0.228	−456.881	201.845	0.024
업력	1. 창업 3년 이하(r)						
	2. 4~10년	39.786	353.247	0.910	−172.191	287.630	0.550
	3. 11~20년	−7.984	358.441	0.982	159.170	289.686	0.583
	4. 21년 이상	128.909	383.373	0.737	−61.120	314.906	0.846
업종	1. 에너지/화학/정밀						
	2. 의료/제약	2054.092	276.911	0.000	1112.366	231.787	0.000
	3. 컴퓨터/반도체/전자부품	211.288	239.437	0.378	−448.137	203.948	0.029
	5. 기계/자동차/금속	433.161	250.332	0.085	92.208	201.770	0.648
	6. 음식료/섬유/비금속/기타제조	422.510	240.146	0.080	194.075	200.681	0.334
상장시장	2. 코스피/코스닥						
	3. 코넥스	−1411.153	676.368	0.038	−2504.879	779.241	0.001
	4. 해당사항 없음	−208.056	257.090	0.419	−1435.904	267.600	0.000
종업원수	1. 10~49인						
	2. 50~99인	155.499	228.513	0.497	−71.606	190.921	0.708
	3. 100~299인	1182.061	271.894	0.000	1202.992	280.925	0.000
	4. 300~499인	1916.738	626.819	0.002	1389.218	1096.401	0.206
매출	2019년 매출액(백만원)	0.006	0.003	0.032	0.001	0.004	0.756
R Square		0.427			0.491		

3.2.2. 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사의 통계적 매칭

2020 벤처기업정밀실태조사와 2020 한국기업혁신조사의 통계적 매칭에 앞서 기준 변수인 공통변수 간의 차이가 있는지 평균차이를 검정하였다(<표 5> 참조). 분석 결과, 지역의 경우에는 2. 대전/세종/충청/강원, 4. 대구/경북, 5. 광주/전라/제주는 차이가 없었으나 1. 서울/인천/경기, 3. 부산/경남/울산은 통계적으로 차이가 있었고, 업력은 1. 창업 3년 이하는 차이가 없었으나 2. 4~10년, 3. 11~20년, 4. 21년 이상은 통계적으로 차이가 있는 것으로 확인되었다. 업종의 경우, 6. 섬유/비금속/기타제조는 통계적인 차이가 없는 것으로 확인되었으나 1. 에너지/화학/정밀, 2. 의료/제약, 3. 컴퓨터/반도체/전자부품, 5. 기계/자동차/금속에서 통계적 차이가 있는 것으로 나타났다. 상장시장의 경우에는 1. 코스피/코스닥, 3. 해당사항 없음에서 변수의 평균값이 통계적으로 차이가 없었고, 2. 코넥스는 차이가 있는 것을 확인하였다. 종업원수는 2. 50~99인에서 통계적인 차이가 없었으나 1. 10~49인, 3. 100~299인, 4. 300인 이상에서는 통계적으로 차이가 있다는

〈표 5〉 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사의 통계적 매칭 평균 검정

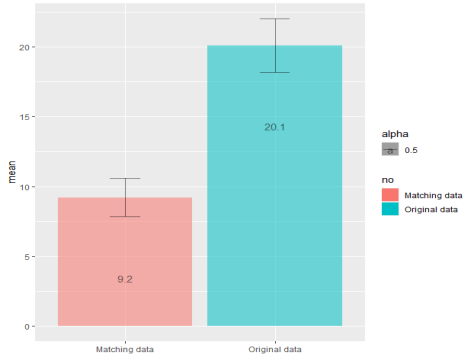
구분		기준 데이터	매칭 데이터	F	Sig.
지역	1. 서울/인천/경기	48.43%	43.87%	12.377	0.017
	2. 대전/세종/충청/강원	19.67%	19.16%	0.441	0.739
	3. 부산/경남/울산	11.92%	16.87%	54.974	0.000
	4. 대구/경북	10.77%	11.06%	0.235	0.809
	5. 광주/전라/제주	9.21%	9.03%	0.102	0.873
업력	1. 창업 3년 이하	5.33%	5.51%	0.164	0.840
	2. 4~10년	35.36%	25.35%	109.528	0.000
	3. 11~20년	39.12%	33.87%	27.573	0.004
	4. 21년 이상	20.19%	35.27%	400.772	0.000
업종	1. 에너지/화학/정밀	16.21%	23.78%	105.111	0.000
	2. 의료/제약	17.05%	4.20%	667.899	0.000
	3. 컴퓨터/반도체/전자부품	20.92%	10.98%	213.079	0.000
	5. 기계/자동차/금속	23.64%	36.46%	267.747	0.000
	6. 섬유/비금속/기타제조	22.18%	24.59%	9.052	0.140
상장시장	1. 코스피/코스닥	9.21%	8.22%	3.334	0.359
	2. 코넥스	1.15%	0.47%	19.345	0.028
	3. 해당사항 없음	89.64%	91.31%	8.928	0.132
매출	2019년 매출액(백만원)	21976.40	44554.47	192.724	0.000
종업원수	1. 10~49인	64.23%	52.40%	182.217	0.000
	2. 50~99인	17.05%	16.24%	1.298	0.567
	3. 100~299인	17.36%	25.43%	115.438	0.000
	4. 300인 이상	1.36%	4.66%	88.324	0.000
연구인력	2019년 연구개발인력 비율	20.29	7.30	390.313	0.000
R&D 비용	2019년 R&D 비용(백만원)	596.41	602.35	4.760	0.910

것을 확인하였다. 또한, 매출과 연구인력 변수의 평균값은 통계적으로 차이가 있으나 R&D 비용은 차이가 없는 것으로 나타났다.

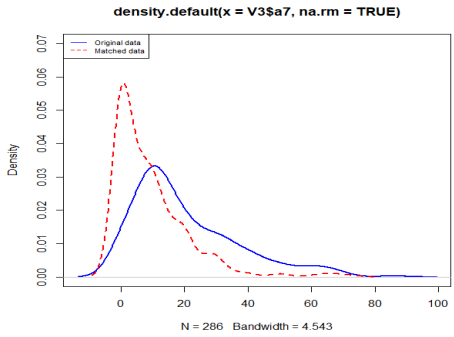
2020 벤처기업정밀실태조사와 2020 한국기업혁신조사의 자료 수집 방법과 시점 등에 차이가 있으므로 두 데이터가 완전히 일치할 수는 없다. 하지만 주요변수에서 통계적인 차이가 없다는 것이 확인되었으므로 통계적 매칭을 수행하는 것에 문제가 없다고 판단된다. 또한, 통계적으로 평균값의 차이가 있는 변수의 경우 완전 일치 알고리즘을 통해 문제를 해소할 수 있다.

구체적으로, 통계적 매칭의 목적인 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사의 인구인력 비율과 R&D 비용이 통계적으로 적절히 매칭되는지 확인하였다. 먼저 연구인력 비율에 대해서 통계적 매칭을 수행하였다. 수행결과, <그림 5>의 하단 표와 같이 기준 데이터의 연구인력 비율의 평균은 20.10, 매칭 데이터의 평균은 9.20으로 나타나 두 변수는 통계적인 차이가 존재하였다. 그러나 커널 밀도 추정 결과를 보면 통계적 매칭된 데이터의 분포가 유사한 형태를 보인다는 점을 확인할 수 있다.

다음으로 R&D 비용에 대해 통계적 매칭을 수행한 결과, <그림 6>의 하단 표와 같이 수여자 데이터의 R&D 비용 평균은 464.06, 기여자 데이터의 평균은 458.74로 나타나 통계적인 차이가 없는 것을 확인하였다. 또한,



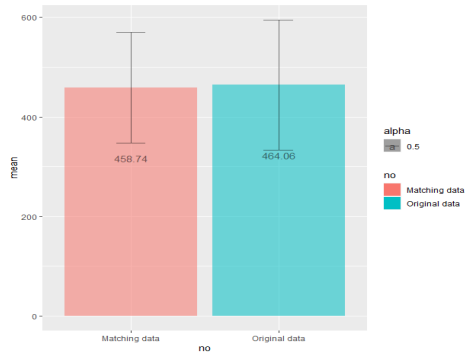
연구인력 비율 평균(95% CI) :
수여자 데이터와 매칭된 데이터의 비교



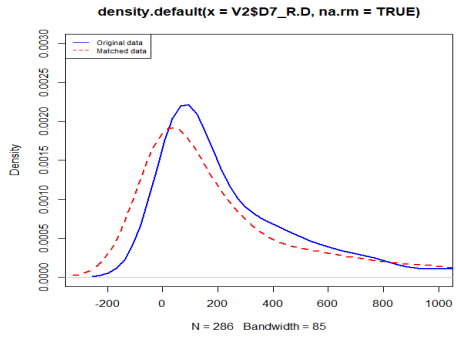
연구인력 비율 분포 비교(Kernel Density Estimation) :
수여자 데이터와 매칭된 데이터의 비교

	기준 데이터	매칭 데이터	F	Sig. (2-tailed)
연구인력 비율	20.10	9.20	34.406	0.000

〈그림 6〉 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사의 연구인력 비율 평균 통계적 매칭 검정



R&D 비용 평균 (95% CI) :
수여자 데이터와 매칭된 데이터의 비교



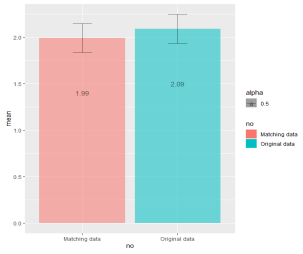
R&D 비용 평균 분포 비교(Kernel Density Estimation) :
수여자 데이터와 매칭된 데이터의 비교

	기준 데이터	매칭 데이터	F	Sig. (2-tailed)
R&D 비용 평균	464.06	458.74	0.486	0.951

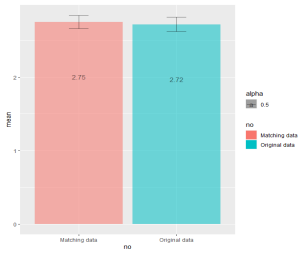
〈그림 7〉 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사의 R&D 비용 평균 통계적 매칭 검정

커널 밀도 추정 결과 통계적 매칭된 데이터의 분포 역시 매우 유사한 형태를 보인다는 점을 확인할 수 있다.

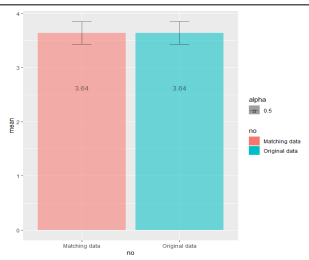
세부적으로 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사의 공통변수에 대한 통계적 매칭을 수행한 결과는 다음과 같다(<그림 7> 참조). 지역의 경우, 수여자 데이터의 평균값은 2.09, 매칭 데이터의 평균값은 1.99로 나타났다. 세부변수의 평균값을 비교한 결과 1. 서울/인천/경기, 2. 대전/세종/충청/강원, 3. 부산/경남/울산, 4. 대구/경북, 5. 광주/전라/제주의 수여자 데이터와 매칭 데이터의 평균값은 통계적으로 차이가 없다는 것을 확인할 수 있었다. 업력의 경우, 수여자 데이터의 평균값은 2.72, 매칭 데이터의 평균값은 2.75로 나타났다. 세부변수의 평균값을 비교한 결과 1. 창업 3년 이하, 2. 4~10년, 3. 11~20년, 4. 21년 이상의 수여자 데이터와 매칭 데이터의 평균값은 통계적으로 차이가 없다는 것을 확인할 수 있었다.



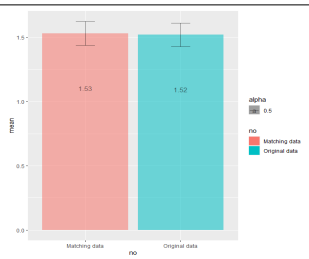
	기준 데이터	매칭 데이터	F	Sig
1. 서울/인천/경기	49.65%	55.24%	3.129	0.181
2. 대전/세종/충청/강원	19.58%	16.08%	4.792	0.275
3. 부산/경남/울산	11.54%	11.19%	0.069	0.895
4. 대구/경북	10.84%	9.09%	1.949	0.486
5. 광주/전라/제주	8.39%	8.39%	0.000	1.000



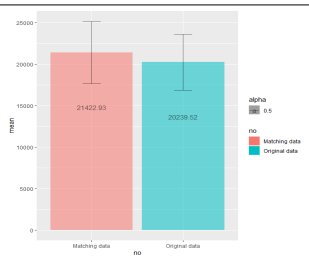
	기준 데이터	매칭 데이터	F	Sig
1. 창업 3년 이하	6.29%	3.85%	7.201	0.183
2. 4~10년	34.97%	34.97%	0.000	1.000
3. 11~20년	39.16%	43.36%	3.949	0.309
4. 21년 이상	19.58%	17.83%	1.148	0.593



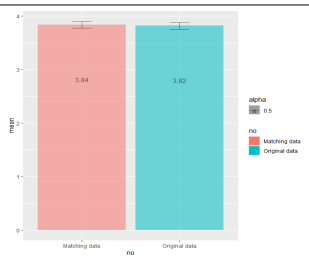
	기준 데이터	매칭 데이터	F	Sig
1. 에너지/화학/정밀	14.69%	14.69%	0.000	1.000
2. 의료/제약	18.88%	18.88%	0.000	1.000
3. 컴퓨터/반도체/전자부품	22.03%	22.03%	0.000	1.000
5. 기계/자동차/금속	20.63%	20.63%	0.000	1.000
6. 섬유/비금속/기타제조	23.78%	23.78%	0.000	1.000



	기준 데이터	매칭 데이터	F	Sig
1. 코스피/코스닥	8.74%	7.34%	1.512	0.539
2. 코넥스	0.70%	1.05%	0.806	0.654
3. 해당사항 없음	90.56%	91.61%	0.773	0.660



	기준 데이터	매칭 데이터	F	Sig.
매출	20239.52	21422.930	0.444	0.644



	기준 데이터	매칭 데이터	F	Sig
1. 10~49인	65.38%	65.38%	0.000	1.000
2. 50~99인	19.23%	16.43%	3.058	0.383
3. 100~299인	13.64%	17.83%	7.667	0.169
4. 300인 이상	1.75%	0.00%	21.026	0.025

〈그림 8〉 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사의 세부내용 통계적 매칭 검정

업종의 경우, 수여자 데이터의 평균값은 3.64, 매칭 데이터의 평균값은 3.64로 나타났다. 세부 변수의 평균값을 비교한 결과 1. 에너지/화학/정밀, 2. 의료/제약, 3. 컴퓨터/반도체/전자부품, 5. 기계/자동차/금속, 6. 섬유/비금속/기타제조의 수여자 데이터와 매칭 데이터의 평균값은 통계적인 차이가 없는 것으로 나타났다. 상장시장의 경우, 수여자 데이터의 평균값은 1.52, 매칭 데이터의 평균값은 1.53으로 나타났다. 세부변수의 평균값을 비교한 결과 1. 코스피/코스닥, 2. 코넥스, 3. 해당사항 없음의 수여자 데이터와 매칭 데이터의 평균값은 통계적인 차이가 없는 것으로 나타났다. 매출의 경우, 수여자 데이터의 평균값은 20239.52, 매칭 데이터의 평균값은 21422.93으로 나타났고 차이가 통계적으로 유의하지 않았다. 마지막으로 종업원수의 경우, 수여자 데이터의 평균값은 3.82, 매칭 데이터의 평균값은 3.84로 나타났다. 세부 변수의 평균값을 비교한 결과 1. 10~49인, 2. 50~99인, 3. 100~299인의 수여자 데이터와 매칭 데이터의 평균값은 통계적인 차이가 없는 것으로 나타났고, 4. 300인 이상의 경우에만 통계적인 차이가 있는 것을 확인할 수 있었다.

4. 결론

본 연구는 경영학 분야에서 새로운 데이터 확보 방법으로 데이터 매칭을 소개하고 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사의 데이터를 활용한 통계적 매칭 시뮬레이션을 통해 그 유용성과 실용성을 검증하였다. 분석결과를 정리하면 다음과 같다.

첫째, 데이터 매칭을 수행하기 위해서는 기본적인 조건에 부합하는 데이터 확보가 요구된다. 데이터 매칭의 기본적인 조건은 모집단을 공유하고 있지만, 서로 다른 표본에서 수집된 데이터라는 점이다. 서로 다른 모집단에서 확보된 데이터를 인위적으로 결합하는 것은 데이터 매칭의 기본적인 조건에 부합되지 않으며, 데이터 매칭의 활용과 결과에 치명적인 오류를 발생시킬 수 있다. 다시 말해, 데이터 매칭은 데이터 활용 목적에 따라 다양한 데이터를 결합하고 매칭하는 방법이 적용될 수 있으나 기본적인 조건이 충족되지 않는다면 적절한 데이터셋을 구현하였다고 볼 수 없다. 본 연구에서 사용된 2020 벤처기업정밀실태조사와 2020 한국기업혁신조사 데이터는 조사 내용은 다르지만, 중소기업이라는 모집단을 공유하고 있으므로 데이터 매칭이 적절하다고 판단할 수 있었다.

둘째, 데이터 매칭에 사용되는 변수와 방법은 선행연구에 근거해야 한다. 데이터 매칭의 방법론적 가치는 정보의 적시성, 경제적 비용 절감, 정보의 다양성 확보 등 기존 조사의 한계점을 극복하는 것에 있다. 무분별한 데이터 매칭은 이러한 가치를 상실시킬 수 있으며, 유의미한 정보를 창출하는 것에 한계가 발생할 수 있다. 따라서 데이터 매칭을 수행하는데 있어 사용되는 변수는 선행연구에 근거해야 하며, 유의미한 변수 간의 결합을 통해 새로운 정보와 가치를 창출할 수 있어야 한다. 본 연구에서는 적절한 변수를 선정하기 위해 보수적인 관점에서 중소기업의 기본적인 인구통계 자료인 지역, 업력, 업종, 상장시장, 매출, 종업원수를 기준으로 기술혁신에서 중요한 요소인 연구인력, R&D 비용을 확인하여 데이터 매칭을 수행하였으므로 데이터 매칭의 변수를 적절하게 선정했다고 볼 수 있다.

셋째, 다양한 시뮬레이션 과정을 통해 데이터 매칭의 우수성이 확인되어야 한다. 선행연구를 살펴보면 지배적인 데이터 매칭 방법은 찾아볼 수 없다. 경영학 분야에서는 성향점수를 주로 활용하고 있으나 다른 사회과학분야나 통계학 분야에서는 단계적 매칭(van Pelt, 2001), K-최근접이웃(김희경, 2010), 회귀분석

(Ingram et al., 2000), 회귀분석과 K-최근접이웃의 결합(정성석 등, 2004) 등을 종합적으로 활용하고 있다. 즉, 성능이 우수한 데이터 매칭을 찾기 위해서 다양한 방법론의 시도와 검증이 요구된다. 본 연구에서는 중소기업 조사자료를 활용한 최적의 데이터 매칭을 수행하기 위해서 선행연구를 통해 마할라노비스와 랜덤 핫덱을 결합한 데이터 매칭 방법론을 선택하였으며, 통계적 매칭 시뮬레이션을 수행한 후 비모수 통계를 고려한 평균차이 검정, 커널 밀도 분포, 회귀분석을 활용하여 데이터 매칭의 적절성과 우수성을 확인하였다.

5. 시사점 및 한계점

본 연구는 기존에 경영학 분야에서 충분히 활용되지 않았던 데이터 매칭이라는 연구방법론을 실험하여 연구방법의 스펙트럼을 확장하고자 하였으며, 이를 통해 다음과 같은 학술적 시사점과 실무적 시사점을 도출하였다. 첫째, 경영학 분야의 이론 검증을 위한 데이터 확보 차원에서 본 연구는 긍정적인 방향성을 제시하고 있다. 최근 경영환경은 더 복잡하고 빠르게 변화하고 있다. 이러한 환경변화를 시의적절하게 이해하기 위해서는 다양한 변수를 활용한 분석이 요구된다(Yang & Kim, 2020). 하지만 분석을 위한 데이터를 확보하기 위해서는 많은 시간과 노력이 필요하며, 경제적 지원이 필요하다(박희창, 조관현, 2006). 이러한 한계점을 극복하는 차원에서 본 연구가 제시하고 있는 데이터 매칭은 실험적인 차원에서 연구를 검증할 수 있는 데이터 확보에 도움을 줄 수 있으며, 시간과 비용을 절감하고, 더 나아가 견고한 연구를 진행하는데 기초 자료를 제공할 수 있다.

둘째, 연구방법의 다양성 확보 측면에서 시사점을 제시하고 있다. 데이터 매칭에는 다양한 방법이 있음에도 불구하고, 경영학 분야의 연구에서는 한정된 데이터 매칭 방법론이 사용되고 있다(금종예, 모영민, 2022). 이처럼 한정된 방법론을 사용하는 것은 연구의 내용과 범위에서 필연적으로 한계를 가지게 된다. 또한, 통계학에서는 특정한 데이터 매칭 방법이 지배적인 성능을 보장할 수 없으며, 데이터 매칭의 우수성을 판단하기 위해서는 다양한 방법론을 적용하고 비교해봐야 한다는 점을 제시하고 있다(D'Orazio et al., 2006; 변종석 등, 2013). 이에 따라 본 연구에서는 경영학 분야의 선행연구의 한계점을 극복하고자 마할라노비스 거리와 랜덤 핫덱을 결합한 데이터 매칭 방법을 활용하였으며, 결과적으로 우수한 성능을 확인하였다.

셋째, 변화하는 경영환경을 이해하기 위한 새로운 데이터 구축의 기초 자료로 활용될 수 있다. 산업현장에서는 시시각각 변화하는 경영환경을 이해하기 위한 데이터 수요가 나타나고 있으며, 데이터에서 의미 있는 정보를 탐색하기 위한 방안을 모색하고 있다. 본 연구에서 제시하고 있는 방법론은 무차별적인 데이터 수집에서 벗어나 각각의 의미 있는 데이터를 통계적으로 조합하여 새로운 데이터를 구축하고 구축된 데이터에서 의미 있는 정보를 찾아내는 것에 목적이 있다. 따라서 본 연구에서 제시한 과정을 통해 구축된 데이터는 기업이 변화하는 환경을 이해하는 것에 도움을 줄 수 있다.

이상의 시사점에도 불구하고 본 연구는 다음과 같은 측면에서 한계점을 가진다. 첫째, 사회과학 분야에서 제시하고 있는 엄격한 통계 기준이 확인되지 않았다. 본 연구는 실험적인 차원에서 연구가 진행되었기 때문에 기존의 사회과학에서 검증하는 변수 간의 관계, 신뢰성, 타당성, 인과성이 충분히 검증되지 않았다. 따라서 후속 연구에서는 실험적인 수준을 넘어 이론을 검증하는 단계까지 진행되는 연구가 필요하다. 둘째, 단편적인 연구방법으로 데이터 매칭의 성능을

판단하고 있기 때문에 보편적인 성능이 제시되어 있지 않다. 본 연구에서는 두 가지의 2차 데이터와 2가지 데이터 매칭을 조합한 방법론만 사용하였기 때문에 향후 다양한 데이터 매칭 방법을 활용하여 성능을 비교하는 후속 연구가 진행될 필요가 있다. 셋째, 2차 데이터의 결측치(Missing Data)에 대한 부분이 충분히 반영되고 있지 않다. 2차 데이터는 조사 내용과 범위가 넓어 많은 부분에서 결측치가 나타난다. 일반적으로 결측치가 있는 표본의 경우 그 표본을 삭제하거나 대응되는 값을 투입하는 방법을 사용한다. 하지만 이 같은 방법은 원천 데이터가 보유하고 있는 고유의 정보가 훼손될 수 있다는 문제가 발생할 수 있다. 이에 후속연구에서는 머신러닝을 활용한 결측치 처리 방안을 모색하여 원천 데이터를 온전히 활용할 수 있는 방안을 마련되어야 한다. 마지막으로 본 연구에서는 기존 경영학 분야에서 적용된 통계적 매칭과 비교분석을 통한 성능의 우수성을 확인하고자 하였으나 직접적인 비교연구가 수행되지 못하였다. 선행연구를 고찰한 결과, 경영학 분야에서 거리함수, 성향점수매칭과 같은 한정적인 방법론이 활용되고 있었으며, 타 학문의 경우 비교연구의 적절한 접점을 찾아낼 수 없었다. 따라서 향후 연구에서는 이번 연구결과를 바탕으로 다양한 알고리즘을 적용한 성능 비교연구가 추진될 필요가 있다.

〈참고문헌〉

[국내 문헌]

1. 금종예, 모영민 (2022). 교육 분야 데이터의 통계적 매칭 적용 가능성 탐색-사교육 변수를 중심으로. **교육연구논총, 43**(4), 43-76.
2. 김경훈 (2022). 통제집단합성법 (Synthetic Control Method)을 사용한 한국의 자본이동관리정책에 대한 효과 분석. **시장경제연구, 51**(1), 29-47.
3. 김동성, 김종우, 이홍주, 강만수 (2017). 공공부문 데이터의 경제적 가치평가 연구: 소상공인 신용보증 데이터 사례. **지식경영연구, 18**(1), 67-81.
4. 김성호, 조성빈 (2005). 마할라노비스 거리를 이용한 자료융합 전략의 성과측정. **경영학연구, 34**(6), 1853-1867.
5. 김희경 (2010). **가중 k-최근접이웃방법을 이용한 통계적 매칭 기법에 관한 연구**. 박사학위논문, 동국대학교 대학원, 서울.
6. 박희창, 조광현 (2006). 통계적 데이터 퓨전을 위한 SAS 매크로. **Journal of the Korean Data Analysis Society, 8**(5), 1927-1937.
7. 벤처기업정밀실태조사(공공용). (2020). doi:10.23333/P.142003.001
8. 변종석, 이석훈, 정구현 (2013). 가계금융, 복지조사의 무응답 처리를 위한 유용한 보조정보 선정. **조사연구, 14**(1), 69-91.
9. 안경민 (2021). **통계적 매칭과 머신러닝 앙상블 기법을 활용한 기업혁신 및 경영성과 예측 모형 개발**. 박사학위논문, 동국대학교 대학원, 서울.
10. 안경민, 이영찬 (2021). 앙상블 학습을 이용한 기업혁신과 경영성과 예측. **정보시스템연구, 30**(4), 247-275.
11. 오미애 (2015). 보건복지분야 데이터 연계 필요성 및 활용방안. **보건복지포럼, 9**, 17-28.
12. 오미애, 최현수, 김수현, 장준혁, 진재현, 천미경 (2017). **기계학습(Machine Learning) 기반 사회보장 빅데이터 분석 및 예측 모형 연구**. 세종: 한국보건사회연구원, pp. 1-183.
13. 오미애, 최현수, 김용대, 이용희, 진재현 (2014). **보건복지통계정보 생산 및 활용 촉진을 위한 마이크로데이터 통합 연계 방안**. 세종: 한국보건사회연구원, pp. 1-206.
14. 이규엽, 박상철, 류성열 (2020). 공공 빅데이터 플랫폼 성과평가 모형. **지식경영연구, 21**(4), 243-263.
15. 이유진 (2021). 산업단지 입주가 기업의 생산성과 고용 증가에 미치는 영향 분석. **산업경제연구, 34**(4), 897-923.
16. 이준원 (2019). 기술금융 중소기업과 일반 중소기업의 경영성과 비교분석-기술신용대출을 받은 기술금융 중소기업을 중심으로. **한국혁신학회지, 14**(1), 279-299.
17. 정성석, 김순영, 김현진 (2004). 데이터 보강을 위한 데이터 통합기법에 관한 연구. **응용통계연구, 17**(3), 605-617.
18. 정용찬, 이원태, 정혁, 김윤화, 유선실, 정부연, 오윤석, 박민규, 권헌영, 오형나 (2017). **조사환경 변화에 대응한 ICT 통계 생산체계 혁신 방안 연구(II) 총괄보고서**. 정보통신정책연구원, pp. 1-237.
19. 최봉, 윤종진, 엄태휘 (2019). 서울시 공공빅데이터 활성화 방안 연구. **지식경영연구, 20**(3), 73-89.

[국외 문헌]

20. Chang, S. J., & Shim, J. (2015). When does transitioning from family to professional management improve firm performance? ***Strategic Management Journal, 36***(9), 1297-1316.
21. Curtis, L. H., Hammill, B. G., Eisenstein, E. L., Kramer, J. M., & Anstrom, K. J. (2007). Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases. ***Medical Care, 45***(10), 103-107.
22. D'Alberto, R., Zavalloni, M., Raggi, M., & Viaggi, D. (2018). AES impact evaluation with integrated farm data: Combining statistical matching and propensity score matching. ***Sustainability, 10***(11), 1-24.
23. D'Alberto, R., & Raggi, M. (2021). How much reliable are the integrated 'live' data? A validation strategy proposal for the non-parametric micro statistical matching. ***Journal of Applied Statistics, 48***(2), 322-348.
24. D'Orazio, M., Di Zio, M., & Scanu, M. (2006). ***Statistical matching: Theory and practice***. John Wiley & Sons.
25. Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). The impact of corporate sustainability on organizational processes and performance. ***Management Science, 60***(11), 2835-2857.
26. Ferrando, A., & Mulier, K. (2015). Firms' financing constraints: Do perceptions match the actual situation? ***The Economic and Social Review, 46***(1), 87.

27. Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. ***Statistical Science, 23***(2), 219-236.

28. Holsapple, C. W., & Wu, J. (2011). An elusive antecedent of superior firm performance: The knowledge management factor. ***Decision Support Systems, 52***(1), 271-283.

29. Ingram, D. D., O'Hare, J., Scheuren, F., & Turek, J. (2000). Statistical matching: A new validation case study. ***In Proceedings of the Survey Research Methods Section***, American Statistical Association, 746-751.

30. Kwon, J., & Johnson, M. E. (2018). Meaningful healthcare security: Does meaningful-use attestation improve information security performance? ***MIS Quarterly, 42***(4), 1043-1068.

31. Limna, P., Kraiwanit, T., & Siripipatthanakul, S. (2023). The growing trend of digital economy: A review article. ***International Journal of Computing Sciences Research, 7***, 1351-1361.

32. Mart n-de Castro, G., L pez-S ez, P., Delgado-Verde, M., Quintane, E., Casselman, R. M., Reiche, B. S., & Nylund, P. A. (2011). Innovation as a knowledge-based outcome. ***Journal of Knowledge Management, 15***(6), 928-947.

33. Nold, H. A. (2012). Linking knowledge processes with firm performance: organizational culture. ***Journal of Intellectual Capital, 13***(1), 16-38.

34. R ssler, S. (2004). Data fusion: identification problems, validity, and multiple imputation. ***Austrian Journal of Statistics, 33***(1/2), 153-171.

35. Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. ***Biometrika, 70***(1), 41-55.

36. Singh, A. C., Mantel, H., Kinack, M., & Rowe, G. (1990). On methods of statistical matching with and without auxiliary information. ***Technical Report SSMD-90-016E***, Methodology Branch, Statistics Canada.

37. Van Der Putten, P., Kok, J. N., & Gupta, A. (2002). Data fusion through statistical matching. ***MIT Sloan School of Management***, 1-13.

38. Van Pelt, X. (2001). ***The fusion factory: A constrained data fusion approach, master's thesis, leiden institute of advanced computer science***. Leiden University, The Netherlands.

39. Wiener, M., Saunders, C., & Marabelli, M. (2020). Big-data business models: A critical literature review and multiperspective research framework. ***Journal of Information Technology, 35***(1), 66-91.

40. Yang, S., & Kim, J. K. (2020). Statistical data integration in survey sampling: A review. ***Japanese Journal of Statistics and Data Science, 3***, 625-650.

41. Zheng, W., Zhou, Y., Liu, S., Tian, J., Yang, B., & Yin, L. (2022). A deep fusion matching network semantic reasoning model. ***Applied Sciences, 12***(7), 3416.

저 자 소 개

안 경 민 (Kyungmin An)

동국대학교 테크노경영협동과정 박사를 졸업하고 인공지능 분야 스타트업, 동국대학교 글로벌융합연구소, 스마트콘텐츠연구소, 교육역량개발원 등에서 근무하였다. 주요 관심 분야는 혁신기술과 기업성과 창출 관련 분야이며 스케일업, 머신러닝과 빅데이터 분석, 플랫폼 비즈니스, 다기준의사결정 등을 연구하고 있다. 지식경영연구, Korea Business Review, e-비즈니스연구, 디지털융복합연구, 인터넷전자상거래연구 등에 다수의 논문을 게재했다.

이 영 찬 (Young-Chan Lee)

현재 동국대학교 WISE캠퍼스 상경대학 경영학부 교수로 재직 중이다. 서강대학교에서 경영학사, 동 대학원에서 경영학 석사 및 박사학위를 취득하였다. 주요 관심분야는 데이터마이닝, 다기준의사결정, 기업성과측정, 지식경영, 시스템 다이내믹스 등이다. 지금까지 International Journal of Bank Marketing, Human Factors and Ergonomics in Manufacturing & Service Industries, Information Systems Management, Expert Systems with Applications 등 주요 학술지에 논문을 발표하였다.

〈 Abstract 〉

The Validity Test of Statistical Matching Simulation Using the Data of Korea Venture Firms and Korea Innovation Survey

An, Kyungmin*, Lee, Young-Chan**

The change to the data economy requires a new analysis beyond ordinary research in the management field. Data matching refers to a technique or processing method that combines data sets collected from different samples with the same population. In this study, statistical matching was performed using random hotdeck and Mahalanobis distance functions using 2020 Survey of Korea Venture Firms and 2020 Korea Innovation Survey datas. Among the variables used for statistical matching simulation, the industry and the number of workers were set to be completely consistent, and region, business power, listed market, and sales were set as common variables. Simulation verification was confirmed by mean test and kernel density. As a result of the analysis, it was confirmed that statistical matching was appropriate because there was a difference in the average test, but a similar pattern was shown in the kernel density. This result attempted to expand the spectrum of the research method by experimenting with a data matching research methodology that has not been sufficiently attempted in the management field, and suggests implications in terms of data utilization and diversity.

Key words: Data matching, Statistical matching, Public data, SMEs research, Innovation research, Simulation

* Dongguk University WISE

** Dongguk University WISE