데이터 과학에 기반한 납세자 행동 예측모델 도입방안 연구

2023년 5월

국 세 청
김 태 수

국외훈련 개요

1. 훈련국 : 미국 (United States of America)

2. 훈련기관명 : 메사추세츠 대학교 애머스트 (University of Massachusetts Amherst)

3. 훈련분야 : Master of Statistics

4. 훈련기간 : 2021.8.4 ~ 2023.6.3

훈련기관 개요

1. 명칭 및 소재지

훈련기관 명칭은 메사추세츠 대학교 애머스트(University of Massachusetts Amherst)이며, 메사추세츠주(州) 중서부에 위치한 인구 4만의 작은 도시인 애머스트(Amherst)에 소재하고 있다.

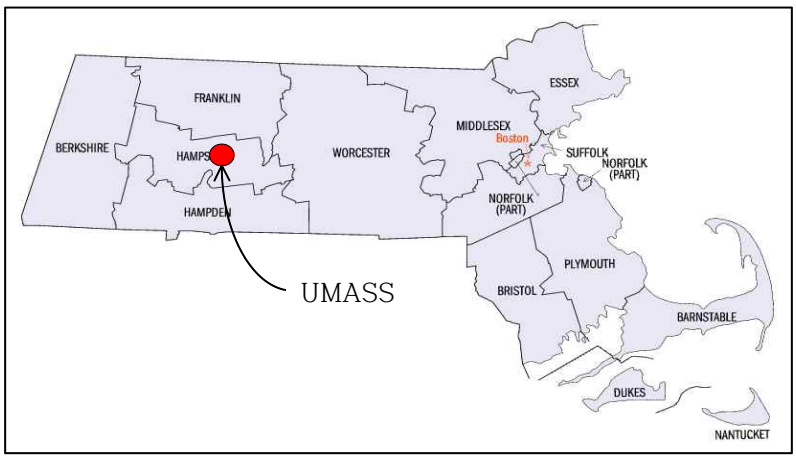


그림 1 메사추세츠주(州) 내 UMASS 위치

2. 주소 및 전화번호

메사추세츠 대학교 애머스트는 캠퍼스가 광활하고 건물이 많아 주소를 특정하기 어려우나, Main Office의 주소와 전화번호는 아래와 같다.

주 소	120 Tillson Farm Road, Amherst, MA 01003-9346
전화번호	413-545-2488

한편, 수학 및 통계학과의 주소 및 전화번호는 아래와 같다.

주 소	Lederle Graduate Research Tower, 1623D University of Massachusetts Amherst 710 N. Pleasant Street Amherst, MA 01003-9305, USA
전화번호	413-545-2762

3. 대학의 주요 역사

1863년 Morrill Land-Grant Colleges Act에 따라 매사추세츠 농업 대학(Massachusetts Agricultural College)으로 처음 설립되었다. 1867년에 첫 수업이 시작되었으며, 교직원 4명, 목조 건물 4개, 학생 56명으로 현대 농업, 과학, 기술 과정 및 교양 과목을 결합한 교과 과정이 제공되었다. 1931년 확장되는 커리큘럼을 반영하여 교명을 Massachusetts State College로 변경하였고, 1947년에는 종합대학교로 승격하면서 현재의 교명(University of Massachusetts)으로 변경하였다.

제2차 세계대전이 끝나고 퇴역군인들의 고등교육에 대한 수요가 늘면서 교육 프로그램이 증가하여 1954년에는 재학생 수가 4,000명으로 늘어났고, 1964년에는 등록된 학부생이 10,500명으로 급증했다. 1990년까지 매사추세츠 대학교 애머스트는 Lederle 대학원 연구 센터와 Conte National Polymer Research Center의 건설과 함께 주요 연구 시설로 부상했으며 플래그쉽 캠퍼스로 자리매김하였다.

4. 대학의 구성과 규모

매사추세츠 대학은 설립 후 주변 대학을 합병하며 규모를 확장하였다. 그 결과 현재 매사추세츠 대학은 5개의 캠퍼스(Amherst, Boston, Dartmouth, Lowell, Worcester의 의과대학)로 구성되어 있다. 그중에서 애머스트(Amherst) 캠퍼스는 메사추세츠 대학 최초의 캠퍼스이자 가장 규모가 크며, 메인 캠퍼스로서의 역할을 하고 있다.[1] 메사추세츠 대학 애머스트는 2022-2023 US NEWS 대학평가에 따르면 미국 227개의 공립학교 중에서 26위를 차지하고 있으며, NECHE 인증[2]을 받았다.

메사추세츠 대학은 110개의 학사, 79개의 석사, 48개의 박사 프로그램을 제공하고 있다. 2021년 기준으로 학부생은 24,000여명, 대학원생은 7,500여명이 재학하고 있으며, 전임 교수진 약 1,400명이 교육 서비스를 제공하고 있다.

---

1) 보스턴 대학과 하버드 대학에 이어 메사추세츠주에서 세 번째로 큰 대학이다.
2) New England Commission of Higher Education 인증, 뉴잉글랜드 6개 주에 걸쳐 대학의 조직, 운영, 프로그램 등을 종합 평가하는 교육 인증

5. 수학 및 통계학과 소개

메사추세츠 대학의 수학 및 통계학과(Department of Mathematics and Statistics)에서는 아래와 같은 석·박사 프로그램을 제공하고 있다.

① 수학 석사(MS in Mathematics)

② 응용수학 석사(MS in Applied Mathematics)

③ 통계학 석사(MS in Statistics)

④ 수학 박사(PhD in Mathematics)

⑤ 통계학 박사(PhD in Statistics)

2023년 현재 91명의 대학원생이 등록되어 있으며, 27%는 여성이고 50%는 미국 이외의 국가에서 온 학생들이다.

6. 석사 학위 이수 조건

통계학 석사 학위를 이수하기 위해서는 회귀분석(Regression Analysis), 수리 통계 I 및 II(Mathmatical Statistics I and II), 전산 통계(Statistical Computing)를 필수로 이수해야 하며, 이를 포함하여 학과 내·외에서 석사 과정 이상에 해당하는 수업을 30학점 이상 이수해야 한다.

아울러, 교수의 지도를 받아 통계 프로젝트를 완료해야 한다. 프로젝트는 다양한 형태로 진행될 수 있지만, 사전에 통계학과 코디네이터의 승인을 받아야 한다. 또한 응용 통계, 확률, 통계 분야의 기본 시험 중 2개를 통과하거나, 통계 컨설팅 과목을 최소 1학점 이상 이수해야 한다.

훈련 결과 보고서

목 차

Ⅰ 연구의 배경 및 목적

과거에 사후 검증을 강화함으로써 성실신고를 유도하려 했던 여러 세정 측면의 노력들은 2012 ~ 2014년 세수 부족 시기에 납세자의 반발에 부딪히며 한계를 노정하였다. 이후 '자발적 성실신고 지원', '탈세·체납 엄정 대응' 2가지 큰 방향으로 국세행정의 패러다임이 전환되었고 현재까지 국세행정 운영의 근간이 되고 있다.

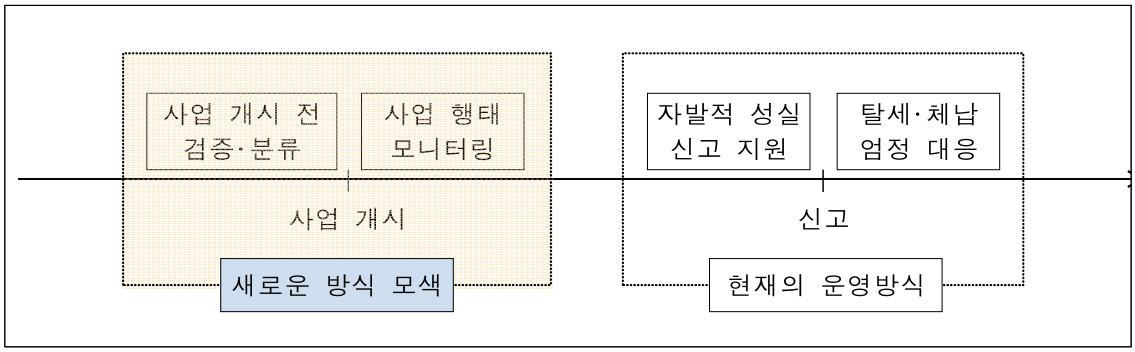
< 국세행정 운영의 2가지 기본 방향 >

구분	자발적 성실신고 지원	탈세·체납 엄정 대응
기본 개념	국세청 내·외부의 자료를 활용하여 신고에 앞서 납세자의 성실신고에 도움이 되는 자료를 최대한 제공	성실신고 지원에도 불구하고 정상적인 궤도를 일탈한 사업자에 대해서는 가용 역량을 활용하여 적극 조치
주요 방안	맞춤형 신고 안내자료 제공, 미리채움·모두채움 서비스 제공 등	체납자에 대한 명단공개 및 감치처분, 변칙적 탈세 혐의자 조사 등

변화된 국세행정은 권력적 행정에 의존하지 않고 납세자의 자발적인 행동변화를 유도하여 세정 순응도를 제고하는 데 크게 기여했지만, 한계점도 여전히 존재하고 있다. 예컨대 명의위장 사업자, 자료상 및 폭탄사업자 등 세수의 누수를 야기하는 불성실한 사업 행태는 여러 노력에도 좀처럼 근절되지 않고 있으며 체납, 탈세 등도 날이 갈수록 고도화·지능화되고 있어 인력·예산·시스템 등의 대폭적인 보강 없이 한정된 행정자원만으로 대응하기는 쉽지 않은 실정이다.

따라서, 성실납세 문화를 실질적으로 정착시키고 행정의 효율성을 제고하기 위해서는 국세행정의 새로운 운영방식을 모색해 볼 필요가 있다. 현재의 패러다임은 '신고 전·후'의 성실도 확보에 중점을 두고 있지만, 그보다 앞서 사업을 개시하고 거래가 이루어지는 단계에서 일상적으로 납세자를 모니터링하고 납세자의 행동을 예측하는 방안을 한층 더 발전시킬 필요가 있다.

< 새로운 국세행정 발전 방향 제안 >



그간 국세청은 납세자가 신고 시 제출하는 자료뿐만 아니라, 「과세자료의 제출 및 관리에 관한 법률」에 따라 외부기관으로부터 다양한 자료들을 수집하고 있으며, 이를 납세자의 자발적 성실신고 지원, 불성실 신고 검증 등 다양한 목적으로 활용해 왔다.

최근에는 지금까지 축적한 대량의 자료들을 통계적·과학적으로 분석하여 국세행정에 활용함으로써 신고 성실도를 높이고 탈세를 사전에 차단하는 데 많은 관심과 노력을 기울이고 있다. 이를 위해 2019년 7월에는 정식으로 빅데이터센터를 출범하였고, 법인 카드 사적 사용 혐의 분석 등의 과제를 수행한 바 있으며 분석된 자료를 맞춤형 신고도움 서비스 등을 통해 납세자들에게 제공하기도 하였다.

하지만 아직 일부 분야에서만 활용되고 있는 빅데이터 분석을 국세행정 전반으로 확산하여 일상적인 세원 관리 분야에도 적용할 수 있는 방안을 더 모색할 필요가 있다. 경제 성장에 따라 사업자 수, 거래 규모도 기하급수적으로 늘어나고 있어, 개인의 경험이 아닌 시스템에 의해 업무가 처리될 수 있는 환경을 조성하는 것이 국세행정의 주요 과제로 떠오르고 있기 때문이다.

이에 민간에서 많이 활용되는 연체 위험도 예측[3], 고객행동 예측 분석[4] 등을 국세행정에 응용할 수 있는 방안을 강구해 보고자 「데이터 과학에 기반한 납세자 행동 예측모델 도입방안 연구」를 연구주제로 선정하였다.

---

3) 대출 신청자의 각종 정보(성별, 소득, 직업, 신용도 등)를 이용하여 원리금 상환을 연체할 확률을 계산하고 대출 승인·거절 판단을 위한 참고자료로 활용한다.

4) 성별, 연령, 구매 이력, 웹 로그, 방문 빈도 등을 통계적으로 해석함으로써 고객의 관심 상품, 구매 시기, 이동 동선 등을 예측하여 마케팅 및 제품 개선 등에 활용한다.

II 주요 개념들의 정리

1. 데이터 과학의 개념

인공지능, 빅데이터 분석의 열풍과 함께 '데이터 과학'이라는 용어도 이제 낯설지 않게 되었다. 하지만 데이터 과학이 무엇인지에 대해 명확하게 답변하기는 여전히 쉽지 않다. 그것은 과학의 한 분야일 수도 있고, 연구 패러다임 또는 연구 방법론일 수도 있으며, 일련의 작업 흐름이나 직업을 의미하는 것일 수도 있기 때문이다.

< 데이터 과학의 개념 분류[5] >

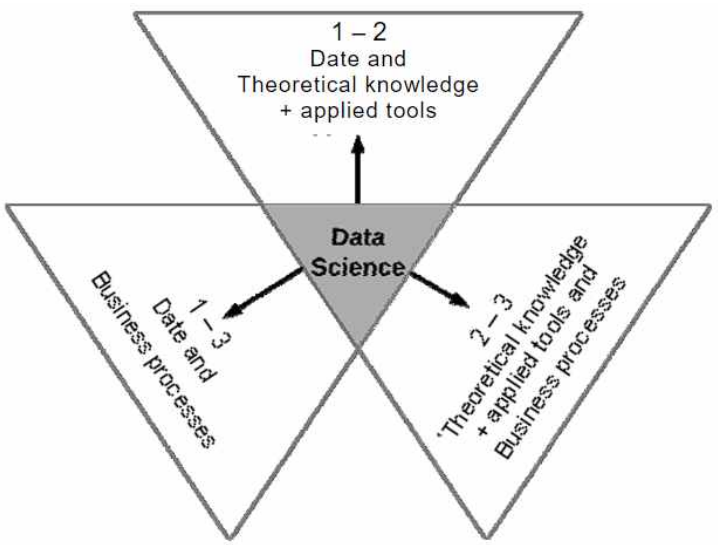
개념 분류	내용
① 과학으로서의 데이터 과학	데이터 과학은 데이터 자체를 천연자원의 관점으로 바라보며 이로부터 가치를 추출하는 방법을 다루는 것이다.
② 연구 패러다임으로서의 데이터 과학	데이터 과학은 실증 과학으로서 데이터 분석을 의미한다. 즉, 과학적 지식을 추론해 내기 위해 미리 정해진 결론 없이 데이터를 수집, 탐색, 발견하는 것을 말한다.
③ 연구 방법론으로서의 데이터 과학	사회과학 등에서 주로 사용되는 연역적 연구 프로세스를 귀납적으로 전환하는 연구 방법을 의미한다.
④ 분야의 하나로서 데이터 과학	컴퓨터 과학, 수학, 통계학, 기타 응용 학문의 지식과 기술의 통합을 의미한다.
⑤ 작업 흐름으로서의 데이터 과학	데이터의 수집, 정제, 체계화, 분석, 해석, 시각화 등 데이터를 처리하는 과정을 의미한다.
⑥ 직업으로서의 데이터 과학	데이터 과학자라는 직업으로부터 도출된 개념으로 주변 세계를 탐색하고 데이터에서 발견하는 과정을 의미한다.

Koby Mike와 Orit Hazzan는 위의 표와 같이 데이터 과학의 개념을 분류하고 정리하였지만, 여전히 각 개념의 경계가 분명한 것인지 의문의 여지가 있어 보인다. 다양한 분야의 학문 지식을 바탕으로 데이터를 탐색하고 의미를 발견함으로써 가치를 창출해 낸다는 점이 위의 개념에서 공통으로 발견되고 있기 때문이다.

5) Koby Mike and Orit Hazzan. 2023. "What Is Data Science?" Communications of the ACM 66 (2): 12–13. doi:10.1145/3575663.

한편, Pavlo Maslianko, Yevhenii Sielskyi는 데이터 과학의 기본 요소들을 도출하고, 그것들의 공통집합으로서 데이터 과학을 이해한다. 이러한 아이디어는 아래의 벤 다이어그램처럼 표현될 수 있다.

< 학제 분야 간 교집합으로서 데이터 과학[6] >



이와 같은 배경 아래에서 이들은 데이터 과학을 '데이터, 정보 및 지식의 분석 및 추출을 위한 활동을 표현하는 분야 간 과학 및 방법론'이라고 정의하고 있다. 다양한 학문 지식의 활용, 데이터로부터 정보와 지식의 추출이라는 개념 표지들이 포함된 것을 알 수 있다.

종합해 보면, 데이터 과학이란 통계학, 수학, 컴퓨터 과학 등 다양한 분야의 지식과 기술을 종합하여 데이터를 수집하고 분석하여, 의미 있는 정보와 지식을 추출하는 과정을 의미한다고 볼 수 있다. 이하에서는 별도로 정의하지 않는 한 이와 같은 관점에서 '데이터 과학'이라는 개념을 사용할 것이다. 국세행정의 관점에서 '데이터 과학'을 바라본다면 그것은 다양한 지식을 활용하여 납세자로부터 데이터를 수집하고 분석하여 유용한 정보를 창출함으로써 국세행정의 효율적이고 효과적인 운영을 도모하는 모든 활동을 의미한다고 볼 수 있을 것이다.

---

6) Pavlo Maslianko, and Yevhenii Sielskyi. 2021. "Data Science – Definition and Structural Representation." Sistemnì Doslìdženâ Ta Informacìjnì Tehnologìi, no. 1 (July). doi:10.20535/SRIT.2308-8893.2021.1.05.

2. 데이터 과학과 빅데이터 분석

근래 빅데이터 분석이 민간뿐만 아니라 공공분야에서도 큰 반향을 불러일으켰다. 공공분야에서 데이터가 갖는 의미는 빅데이터 분석 열풍 이후 크게 달라졌다. 과거에는 행정을 효율적으로 수행하기 위한 자료의 축적과 보존에 초점을 맞추고 있었다. 대량의 데이터는 행정의 비효율과 부담을 초래했기 때문에 적절한 보존기간을 정해 그 총량을 제어하려고 했다.

하지만 인공지능 및 컴퓨팅 기술의 발달로 대량의 자료로부터 유의미한 정보와 가치를 추출할 수 있게 됨에 따라 이를 행정에 활용하는 일에도 관심이 집중되었다. 과거 경력자의 직관과 감각에만 의존해야 했던 일들, 인력과 시간의 부족으로 추진하기 어려웠던 일들을 빅데이터 분석이 해결해 줄 것이라 기대했다.

빅데이터는 정형 또는 비정형 데이터를 포함하는 모든 데이터를 의미한다. 대량의 데이터에 숨겨진 패턴, 미지의 상관관계 등을 밝혀 정보를 추출하고 해석하여 의사결정을 지원하는 것이 빅데이터 분석의 주된 목적이며, 이를 위해서 기계학습, 통계 알고리즘, 데이터 시각화 등의 다양한 기술이 사용된다.

앞서 설명한 데이터 과학과 빅데이터 분석을 비교해 보면 많은 유사점을 발견할 수 있다. 데이터로부터 가치 있는 정보를 찾아낸다는 점, 다양한 분야의 지식과 기술을 활용한다는 점 등이 그러하다. 하지만 데이터 과학은 빅데이터 분석과 달리 대량의 데이터를 필수적 요건으로 하지 않는다. 데이터 과학은 소규모 데이터에서도 의미 있는 패턴과 통찰을 발견하기 위해 가능한 모든 지식과 기술을 활용한다.

컴퓨팅 기술이 비약적으로 발전하였지만, 여전히 실제 현장에서 전체 빅데이터를 다루기는 쉽지 않다. 빅데이터 분석이라고는 하지만 데이터를 처리가 가능한 수준으로 분할하여 분석하는 경우도 많다. 따라서 자료의 양에 구애되지 않는 데이터 과학이라는 표현이 더 적절한 것일 수도 있다. 빅데이터 분석은 데이터의 양적 측면보다는 그동안 비용과 기술의 한계 때문에 애물단지로만 여겨졌던 대량의 데이터로부터 유용한 가치를 찾아내려고 했다는 점에서 의의가 있다고 할 것이다.

3. 납세자 행동의 개념

소비자 행동은 경영학에서 주로 사용하는 개념으로 다양한 연구자들에 의해 다양하게 정의되고 있다. Agarwala, Mishra, Singh는 소비자 행동을 소비자가 소비하는 과정에서 노출하는 태도, 가치 및 행동으로 묘사하고 있으며, Théophile Nassè는 소비자들이 어떤 제품과 서비스를 필요로 하며 어떤 동기로 그것들을 구매하는지, 어떻게 제품과 서비스를 조사하고 평가하고 결정하는지를 의미하는 것이라고 본다.[7)]

종합해 보면, 경영학 분야에서 소비자 행동이란 상품 및 서비스를 구매하는 개인 및 조직의 의사결정 프로세스 및 행동을 의미한다. 소비자 행동을 이해하려는 이유는 시장을 더 잘 이해함으로써 효과적인 마케팅 전략을 개발하고 고객 만족도를 높이기 위함이다. 기업들은 소비자의 행동을 분석함으로써 구매 동기, 구매 결정에 영향을 미치는 요인, 제품 및 서비스를 평가하는 방식에 대한 통찰력을 얻을 수 있다.

이와 유사한 맥락에서 납세자 행동이라는 개념을 정의해 보고자 한다. 납세자 행동은 세금의 신고, 납부와 관련한 납세자의 의사결정과 실제로 이루어진 행위를 의미하며, 납세자 행동은 소득, 재정 상황, 조세에 대한 태도, 정부 정책에 대한 인식, 세법 준수 의식 등 다양한 요인에 의해 영향을 받을 것으로 예상해 볼 수 있다. 따라서 납세자 행동은 세금의 정확한 신고·납부, 세법의 준수처럼 긍정적인 형태로 나타날 수 있지만, 탈세 또는 편법 증여처럼 부정적 형태로 나타날 수도 있다.

납세자 행동을 분석하고 그 동기 및 영향을 미치는 요인을 파악하는 것도 국세행정을 발전시키고 납세 서비스의 품질을 향상시키는데 큰 의미가 있을 것으로 보이지만, 본 보고서의 주제를 벗어나므로 별도로 논의하지 않는다. 다만 여기서는 납세자 행동에 미치는 요인을 탐구하기보다는 현실로 나타나는 납세자 행동을 어떻게 과학적으로 사전에 예측하고 대응할 것인지에 초점을 맞추고자 한다.

---

7) Dr. Theophile Bindeoue Nasse “The Concept of Consumer Behavior: Definitions in a Contemporary Marketing Perspective.” 2021. International Journal of Management & Entrepreneurship Research; Vol. 3 No. 8 (2021); 303-307.

III 데이터 과학의 적용 사례

□ 해외 주요 사례

1. 사기 탐지(Fraud Detection)

미국의 사회보장국(The United States Social Security Administration, SSA)은 은퇴, 장애, 유가족 등을 지원하는 사회보장제도를 관리하는 정부 기관이다. 사회보장국은 장애 혜택 청구의 사기 적발과 예방을 위해 빅데이터 분석을 활용하고 있다. 과거 청구 사례와 알려진 사기 사례 데이터에 기반하여 공통적인 특징과 의미 있는 패턴을 찾아냄으로써 사기 신청이 처리되는 것을 방지하고 있다. 사회보장국 대변인에 따르면 장애 혜택 청구 사기 예방에 1달러를 지출할 경우, 16달러를 절약할 수 있는 효과가 있다고 한다.[8)]

미국의 의료보장센터(The Centers for Medicare & Medicaid Services, CMS)는 65세 이상의 노년층 또는 저소득층을 대상으로 제공되는 공공 의료보험 프로그램을 운영하고 관할하는 정부 기관이다.[9)] 의료보장센터는 비정상적이고 의심스러운 보험 청구 패턴을 식별하기 위해 과거의 대규모 데이터를 분석하여 드러나지 않는 추세 또는 발생할 수 있는 사건의 가능성을 예상하는 예측 분석(Predictive Analyrics)을 사용한다. 지원금 지급에 앞서 사기 방지 시스템으로 심사함으로써 약 8억 2천만 달러의 부적절한 지급을 확인하고 방지할 수 있었다고 한다.[10)]

2. 범죄 예방(Crime Prevention)

미국 시카고 경찰국은 IBM과의 협력을 바탕으로 기계학습 및 예측 분석을 범죄 예방에 활용하고 있다. 범죄 사건, 체포 기록, 날씨 등의

---

8) Stephanie Kanowitz, “Social Security to step up fraud detection with predictive analytics” 2014. https://gcn.com/data-analytics/2014/04/social-security-to-step-up-fraud-detection-with-predictive-analytics/297150/

9) 김지애, “미국 Center for Medicare and Medicaid Services (CMS)의 데이터 이용 지원 사례 고찰” 2015. 건강보험심사평가원(HIRA) 정책동향 9권 5호.

10) Kashif Afzal Khan, “Using Analytics to Reduce Fraud in Public Procurement – Implementing the Fraud Reduction and Data Analytics (FRDA) Act” 2018.

정보를 활용하여 범죄 위험 지역을 표시하고, 경찰 순찰차 배치 등 권장되는 회피 조치를 표시하는 시스템을 도입하여 활용하고 있다. 이러한 정보는 '결정 지원 시스템(decision support system)'에 수집되며 개별 경찰관이 실시간으로 이용할 수도 있다. 영국 맨체스터에서도 이와 같은 예측 분석을 범죄 예방에 활용하고 있으며 차량 도난 및 절도에서 큰 효과를 보이는 것으로 나타나고 있다. 시스템의 권장 조치를 채택한 후 강도는 12%, 절도는 21%, 차량 도난은 32% 감소했다고 한다.[11]

< 능동적 치안이 범죄를 예방하는 방법 >

출처 : Jen Clark, "Facing the threat: Big Data and crime prevention"

미국 로스엔젤레스 경찰국은 과거 범죄 및 체포 데이터를 기반으로 범죄 발생 장소와 미래 범죄자를 예측하고자 노력하고 있다. 로스엔젤레스 경찰국은 과거의 전과 기록을 바탕으로 개인을 점수화한 후, 경찰이 더 자주 집을 방문하고 경찰에게 더 많이 제지될수록 더 많은 점수를 부가한다. 점수가 일정 기준 이상이 되면 만성 범죄자 범주에 포함되어 경찰의 주된 관심의 대상이 된다. 그 밖에도 로스엔젤레스 경찰국은 특정 지역에서 발생한 범죄의 종류, 시간, 장소를 바탕으로 다른 범죄 발생 가능성, 범죄 발생 시점을 예측하는 Predpol이라는 소프트웨어를 사용하고 있다. 이 Predpol은 경찰의 순찰을 강력히 권장하는 핫스팟이 표시된 지도를 매일 업데이트하여 보여준다.[12]

11) Jen Clark. "Facing the threat: Big Data and crime prevention" 2017.
12) Issie Lapowsky. "How the LAPD Uses Data to Predict Crime" 2018.

< 실시간 범죄 분석 센터(RCAC) 내 상황실 >

출처 : Sarah Brayne. 2017. “Big Data Surveillance: The Case of Policing”

3. 교육 지원(Assist in Education)

미 교육부(The United States Department of Education)는 교육 현장에서 데이터 마이닝과 빅데이터 분석을 이용하여 교육의 질을 향상시키고 있다. 예를 들어 온라인 학습을 하는 학생의 키 클릭 패턴을 감지하여 학생이 수업에 집중하지 못하거나 지루함을 느끼는 것을 식별함으로써 학생의 주의를 집중시키거나 교과 과정을 수정하도록 유도한다. 이러한 데이터는 실시간으로 수집되며 지속적인 개선을 위해 여러 경로를 통해 피드백된다. 학생에게는 다음 문제를 위해 즉각적으로 피드백을 제공하고, 교사에게는 다음 날의 수업을 위해 매일, 교장에게는 수업 진행 상황 파악할 수 있도록 매월, 학군의 행정관에게는 학교의 전반적인 개선 상황을 파악할 수 있도록 매년 피드백을 제공한다.[13)]

교육 데이터의 활용 측면에서 가장 혁신적인 대학으로 평가받고 있는 애리조나 주립 대학(Arizona State University)은 교육 데이터를 활용하여 eAdvisor라는 온라인 상담 서비스를 운용하고 있다. 권영옥(2019)은 애리조나 대학의 eAdvisor의 활용 사례를 아래와 같이 설명하고 있다.

13) The United States Department of Education, “Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief” 2012.

> 학생들이 자신의 관심 있는 분야와 전공을 찾을 수 있도록 도움을 주고 수강 신청 시에도 가장 효율적인 수업과 시간을 조언해준다. 전공별 수강 과목의 순서도 제안하여 특정 전공을 위해서 가능한 저학년때 수강해야 하는 기초 과목들도 알려준다. 학생들은 가이드라인으로 주어진 트랙에서 벗어났을 때 바로 도움을 받을 수 있다. eAdvisor를 통해 졸업률은 11.6% 향상 되었으며 학생유지율(Retention)은 84%까지 향상되었다.[14)]

조지아 주립 대학(Georgia State University)도 학생들의 학업 수행을 저해하는 요인을 찾는 일에 빅데이터 분석을 활용하고 있다. 권영옥(2019)에 따르면 조지아 주립 대학은 예측 분석 기법을 적용한 Graduation Progress System(GPS)을 발하여 중도 탈락의 요인을 분석하고 있으며 이를 통해 경제적 문제로 학업을 성공적으로 마치지 못할 위험이 있는 학생에게는 장학금을 지급하도록 하고 있다. 그 밖에도 지도교수가 언제 학생들을 상담해야 하는지 알려주는 조기 알림 시스템을 개발하여 적절한 시기에 학업 및 재정 상담이 이루어지도록 하고 있다.

4. **공공 안전**(Public Safty)

미국 샌프란시스코 시(市) 정부는 도시 내에서 빈번하게 발생하는 교통 충돌 사고를 줄이기 위해 공중 보건부(The Department of Public Health)와 교통부(The Department of Transportation)에 빈번한 교통사고를 줄일 수 있는 대책 마련을 요구하였고, 그에 따라 사고의 원인 파악을 위한 데이터 분석이 실시되었다. 도시 전역의 교통사고를 지도 위에 매핑하여 시각화하는 플랫폼을 개발하였고, 사고가 주로 발생하는 장소를 보여주는 High Injury Network를 개발하였다. 이를 통해서 전체 교차로 중 12%에서 심각한 부상의 70%가 발생한다는 사실을 알아낼 수 있었다. 이를 통해 보호된 교차로(Protected Intersection)[15)], 지하화된 교차로, 보호된 자전거 도로 등의 방안을 도입할 수 있었다.[16)]

---

14) 권영옥. 2019. “교육 데이터와 분석 기법: 사례 연구를 중심으로” 한국데이터학회지 2019년 제4권 제1호, pp.73-81.

15) 네덜란드식 교차로라고도 하며 자전거 타는 사람과 보행자가 자동차와 분리되는 평면 도로 교차로의 한 유형이다.

16) Abhi Nemani. "Data-Driven Policy - San Francisco just showed us how it should work.“ 2016. https://abhinemani.com/essays/2020/08/28/Data-Driven-Policy-San-Francisco-just-showed-us-how-it-should-work/

< 고위험 지역 네트워크(High Injury Network) >

출처 : Abhi Nemani. 2016. "Data-Driven Policy - San Francisco just showed us how it should work.“

□ 국내 주요 사례

1. 범죄 예방

국가정보자원관리원은 경찰청의 임장일지[17] 데이터를 이용하여 여죄 추적 모델을 구현하였다. 그간 누적된 임장일지가 너무 많아 피의자의 여죄를 입증하기 위해 임장일지를 다시 검토하는 데에는 시간과 인력이 많이 소모되었다. 이에 임장일지를 바탕으로 보다 손쉽게 여죄를 추적할 수 있도록 기존의 TF-IDF 알고리즘[18]뿐만 아니라, 구글의 Doc2Vec[19] 등 4개의 알고리즘을 적용하여 최적화된 모델을 개발하였다. 부산지방경찰청에서는 범죄 피의자의 여죄 추적에 이 모델을 활용하였고 3건의 추가 여죄를 입증하는 데 성공한 바 있다. 부산뿐만 아니라 다른 지역의 침입 및 절도사건 등에도 이 모델이 활용되고 있다.[20]

---

17) 사건의 개요 및 범행 수법 등이 상세하게 기술된 수사기록의 일종이다.
18) 다른 문서에는 등장하지 않지만, 특정 문서에서만 자주 등장하는 단어를 찾아내 문서 내 중요한 단어의 가중치를 계산하는 방법이다. 이를 통해 문서 내 비중 있는 단어 또는 단어 묶음을 추출할 수 있다.
19) 구글이 개발한 자연어 처리를 위한 도구로서 문서에서 의미 있는 정보를 추출한 다음 해당 정보를 바탕으로 문서를 식별하고 분류하는 알고리즘이다. 학술 논문 또는 장문의 텍스트에서 요약을 자동으로 생성하거나, SNS 사이트에서 관련 콘텐츠를 찾는 데에도 사용된다.
20) 행정안전부. 2018. “내 삶을 바꾸는 공공 빅데이터!”

서울시 영등포구는 2018년 전국 최초로 ‘여성 안심 빅데이터 셉테드(CPTED[21]) 협업 플랫폼’을 구축하여 활용하고 있다. 셉테드는 범죄를 예방할 수 있도록 도시 환경을 재설계함으로써 삶의 질을 향상시키는 종합적인 범죄 예방 전략이다. 영등포 경찰서, KT 등으로부터 지역 내 범죄 데이터, 야간 여성 유동 인구 데이터, 여성 1인 가구 데이터, 여성 안심 스카우트 경로 데이터, 여성 안심 시설물 정보 등 다양한 정보를 수집하고 기계학습 방법으로 이를 분석하여 중점적인 관리가 필요한 지역과 안전한 지역을 도출한다. 이와 같은 분석 결과를 토대로 여성 안심 귀가 경로, 범죄 예방 순찰 경로, CCTV 설치 지역 등을 최적화함으로써 범죄 예방을 도모한다.

경산 결찰서는 2019년 영남대 경북 빅데이터 센터, SK텔레콤 등과 협업하여 빅데이터 분석을 통해 요일별 및 시간대별로 범죄 발생 가능성이 높은 지역을 예측하는 범죄 예측 모델을 개발하여 운영하고 있다. 범죄 예측 모델은 유동 인구 112 신고 현황, 유흥업소 정보 등을 통해 범죄 발생 가능성이 큰 지역을 예측한다. 이를 통해 경찰력을 효율적으로 운용함으로써 범죄를 예방하고, 현장 출동 시간을 단축하여 국민 안전을 더 효과적으로 보장할 수 있게 되었다.

2. 보건 및 복지 지원

건강보험공단은 2013년부터 ‘국민건강 알람서비스[22]’를 통해 감기, 눈병, 식중독, 천식, 피부염 5개 질병에 대해 위험, 경고, 주의, 관심 4단계로 위험 정도를 제공하고 있다. 건강보험공단의 국민건강정보 DB를 바탕으로 식약처의 식중독 자료, 기상청의 기상자료, 환경부의 환경자료를 연계하고 SNS 정보를 융합하여 질병 발생 예측 모델을 구축하였다. 그 밖에도 건강보험공단은 성별, 신장, 체중, 흡연 및 음주 여부 및 의료 검진 기록 등을 토대로 뇌졸중 위험도, 골다공증성 골절 위험도, 심장 질환 위험도를 예측하는 프로그램을 마련하여 대국민 서비스를 제공하고 있다.[23]

---

21) Crime Prevention Through Environmental Design의 줄임말이다.
22) http://forecast.nhis.or.kr
23) 행정안전부. 2018. “내 삶을 바꾸는 공공 빅데이터!”

< 국민건강 알람서비스 홈페이지 화면 >

출처 : 국민건강 알람서비스(http://forecast.nhis.or.kr)

남양주시와 국민연금공단은 데이터를 공동으로 이용하여 사회취약계층의 취업을 지원하고 있다. 남양주시는 구직·구인 신청 목록, 취업자 목록 등을 제공하였고 국민연금공단은 연금 가입자 및 사업장 정보를 추가하였다. 이를 통해 실직 기간의 교차 분석, 실직자의 세대 특성에 대한 상관분석이 진행되어 실직자의 현황을 파악할 수 있게 되었다. 이와 같은 정보는 직업훈련 계획 등 취약계층의 취업 정책을 수립하는 데 기초가 되었으며 실직자들에게 알선할 사업장을 선정하는 데에도 도움이 되었다. 공공기관 간 데이터 공유를 통해 6개월간 약 100명의 신규 취업이 이루어졌으며 연간 총 18억 원의 지역 경제 활성화 효과가 있었을 것으로 추산하고 있다.[24)]

3. 법 집행의 효율성 제고

고용부는 반복적으로 위법 행위를 하는 사업장을 선별하는데 데이터 분석을 활용하였다. 사업장 자료와 근로자의 신고사건 자료를 바탕으로 근로복지공단의 고용 및 산재 사업장 정보, 건강보험공단의 보험료 체납 정보 등을 종합하여 서면계약, 임금 체불, 최저임금, 근로시간, 약자 보호 5대 취약 유형에 대한 종합 취약지수를 도출하고 우선 감독이 필요한

24) 행정안전부, 2018. "내 삶을 바꾸는 공공 빅데이터!"

사업장을 선별하였다. 2016년부터 취약지수를 사용한 근로감독이 본격적으로 이루어지고 있다. 이를 통해 정부는 3년간 1,461억 원의 임금 체불이 감소하는 효과가 있을 것으로 기대하고 있다.[25)]

국가와 지방자치단체가 소유한 국·공유지는 허가를 받지 않는 한 개인이 무단으로 활용할 수 없다. 하지만 우리나라 국유재산 중 6.8%는 무단 점유되고 있으며 이에 대한 변상금도 제대로 회수되지 못하고 있다. 이에 통영시는 항공사진을 수집하여 행정 시스템 내 국·공유지 관리 기능을 추가하고 인공지능 영상판독 모델을 개발했다. 이를 통해 토지 정비 대상을 쉽게 파악할 수 있게 되었고 무단 점유지와 시설 불법 토지 형질변경 여부도 수월하게 확인할 수 있게 되었다. 이번 모델을 적용하면 국·공유지 내 불법 무단 점유지 관리 업무를 수행하는 데 연간 1.9억 원의 비용을 절감할 수 있을 것으로 전망된다.

□ 주요 시사점

국내·외의 데이터 분석 적용 사례들을 보면 공공부문에도 이미 상당히 다양한 분야에 걸쳐 데이터 분석이 적극적으로 활용되고 있음을 알 수 있었다. 데이터 분석 분야에서는 하루가 다르게 새로운 기법이 계속 소개되고 있어, 최근에는 기존의 데이터 분석 방법을 고도화하는 데에도 많은 역량을 모으고 있는 것으로 보인다. 한편, 분석 유형으로 보면 기존의 방법으로는 좋은 결과를 얻기 어려웠던 위험 예측, 숨은 패턴 발견 등에 데이터 분석이 많이 활용되고 있음을 알 수 있었다.

다만, 미국에서는 광범위한 데이터 과학의 적용에 대한 우려의 시선도 적지 않다. 미국 연방정부는 인공지능 및 기계학습 분야에서 세계적 우위를 유지하기 위해 투자를 확대하면서도, 연방정부 인력들이 인공지능을 윤리적으로 안전하게 사용하도록 하는 법[26)]을 2022년 10월 시행하기도 하였다. 앞으로는 데이터 윤리가 뒷받침되지 않는 한 글로벌 리더십을 유지하기 어렵다는 판단이 깔린 것으로 보인다. 앞다퉈 데이터 분석을 도입했던 우리나라에도 중요한 시사점을 줄 것으로 생각된다.

---

25) 행정안전부, 2018. “내 삶을 바꾸는 공공 빅데이터!”
26) The Artificial Intelligence Training for the Acquisition Workforce Act (S. 2551)

Ⅳ 국세행정에 필요한 예측의 유형

1. 국세행정의 범위

‘국세행정’이라고 하면 보통 세금을 강제 징수하는 것을 떠올리지만, 실제로는 다양한 형태와 의미를 갖는 행정을 포괄하는 개념이다. 크게 보면 내부적으로 효력을 갖는 업무 계획 수립, 예산 운용, 인사 관리, 직원 교육, 업무 평가, 통계 작성 등도 국세행정의 범주에 포함된다고 볼 수 있다. 물론 이와 같은 범주의 행정에도 데이터 과학을 접목하여 효율성을 높일 수 있는 여지가 있고 실제로 긍정적인 효과를 보여주는 사례도 있다. 하지만 납세자에게 직접적인 영향을 미치는 행정이라고 할 수는 없으므로 여기에서는 별도로 논의하지 않는다.

납세자에게 영향을 미치는 행정으로 국한해도 상당히 다양한 활동이 포함된다. 아래는 ‘국세청 사무분장 규정’에 예시되어 있는 국세청 본청 각 부서의 업무 중에서 납세자에게 직접적 영향을 미치는 업무만 간추려 정리해 본 것이다.

< 납세자 직접 관련 국세행정의 종류 >

부서	담당 업무
징 세 법무국	■ 국세 환급금 사무 ■ 국세징수 관계 법규에 관한 민원 처리 ■ 체납정리업무 집행, 고액체납 정리 및 은닉재산 추적조사 ■ 체납자에 대한 행정규제 및 출국규제, 고액·상습체납자 명단공개 ■ 납기연장·징수유예·체납처분유예 등 세정지원 대책 수립
개 인 납세국	■ 부가가치세의 부과, 공제 감면 ■ 부가가치세 신고 관리 및 전자신고 상담창구 운영 ■ 과세 유흥장소의 개별소비세 관련 기획 ■ 종합소득 등에 대한 소득세 신고·환급·감면 ■ 부가가치세 면제 대상 재화·용역의 공급에 대한 수입금액 신고 업무 ■ 개인사업자의 신고성실도 평가 기준 및 정기 조사대상 선정 업무 ■ 신용카드·전자세금계산서 조기경보시스템 개발·분석 및 운영 ■ 전자상거래 세원관리 및 관련 제도 개선

< 납세자 직접 관련 국세행정의 종류(계속) >

부서	담당 업무
법 인 납세국	■ 법인세 신고 및 세적 관리 ■ 감면법인 및 공익법인의 사후 관리 ■ 법인의 신고성실도 평가기준 및 정기조사 대상 법인의 선정기준 제정 ■ 원천세 세원 관리 및 원천징수의무자 세적 관리 ■ 전자기부금영수증 제도 및 시스템 운영에 관한 사항 ■ 주세, 개별소비세 등의 부과·감면 및 관련 전산시스템 개발 운영 ■ 주세법에 의한 주류 제조 및 판매면허와 면허업체 관리·감독 ■ 주류, 유사석유제품 및 농어업용 면세유에 대한 유통과정 조사계획 수립
자 산 과세국	■ 양도소득세 신고·부과·감면 및 관련 전산시스템 개발 운영 ■ 양도소득세 및 부동산 투기 관련 세무조사 대상자 선정 기준 마련 ■ 부동산 거래정보의 수집 분석, 투기대책 수립 ■ 부동산 거래 관련 탈세제보의 처리 ■ 상속·증여세의 신고·부과·감면 및 관련 전산시스템 개발 운영 ■ 상속·증여세 조사 관련 분석자료의 출력 및 관리 ■ 주식 및 파생상품 관련 양도소득세 신고·부과·감면 ■ 주식 변동조사에 관한 총괄·조정·관리
조사국	■ 내국세 세무조사에 관한 기획 및 조정 업무 ■ 조세범칙조사 제도의 운용 및 개선 관련 사항 ■ 조사 관련 납세자 애로·불만사항 해소 등 권익보호 관련 사항 ■ 세금계산서 및 계산서 수수질서 문란행위 분석·관리 ■ 법인, 개인납세자에 대한 실태분석 및 관리 ■ 국제거래 관련 조사 관리 및 지원 ■ 내국세 관련 탈세정보 및 세원동향정보의 수집·분석 및 종합관리 ■ 신종 산업 관련 납세자 및 관련인에 대한 실태분석 및 관리
소 득 지원국	■ 근로장려세제 및 자녀장려세제 홍보, 상담업무 총괄 ■ 근로장려금 및 자녀장려금 신청 관련 전산시스템 개발·운영 ■ 근로장려금 상담센터 운영 ■ 근로장려금 및 자녀장려금 신청자 만족도 조사 ■ 일용근로소득지급명세서 등 과세자료 수집 및 분석 ■ 학자금 상환 업무 ■ 학자금 의무상환액 체납 관련 업무 ■ 장기 미상환자 관리 업무

2. 국세행정의 분류

이처럼 다양한 국세행정을 납세자와의 관계에서 유사한 성격을 갖는 행정들로 묶어보면 납세자로부터 세법에 규정된 세금을 걷는 '세금 확보를 위한 행정', 세금을 내는 데 고충을 겪는 납세자를 도와주거나 법에서 정한 장려금 대상자에게 장려금을 지급하는 '납세자 지원을 위한 행정', 세금과 관련된 각종 증명 또는 서류를 발급해 주는 '민원 발급을 위한 행정'으로 구분해 볼 수 있다. 이들 각각의 범주에 속하는 국세행정을 구체적으로 예시해 보면 아래의 표와 같다.

< 납세자 행동 예측 관점에서 본 국세행정의 분류 >

유형	주요 활동
① 세금 확보를 위한 행정	사업자등록 및 관리, 전자세금계산서 등을 통한 정상적 상거래 여부 확인, 성실신고 안내, 무신고자 및 신고 오류 처리, 신고성실도 분석, 세무조사 대상자 선정, 체납 정리, 고액 체납자 추적 등
② 납세자 지원을 위한 행정	납기연장, 징수유예, 체납처분 유예 등 세정 지원 대상 확정 및 세정 지원 신청 처리, 세금 신고를 도와주는 안내 및 정보 제공, 근로·자녀장려금 신청 및 지급, 공제 대상 여부 사전 심사 등
③ 민원 발급을 위한 행정	홈택스 운영, 민원창구 민원 발급 등

각 범주가 납세자와의 관계에서 유사한 성격을 갖고 있으므로 동일한 유형의 납세자 행동 예측이 필요할 것으로 생각된다. 세금 확보를 위한 행정에서는 정상적인 궤도에서 일탈하여 세금을 확보하는 데 어려움을 겪게 만드는 불성실 사업자를 예측하고 관리하는 데 납세자 행동 예측의 의의가 있을 것이다. 또한 납세자 지원을 위한 행정에서는 세금 신고를 손쉽게 마칠 수 있도록 도와주고, 세금 납부의 어려움이 있는 사업자를 신속하게 찾아내는 데 납세자 행동 예측이 의미를 가질 수 있을 것으로 생각한다. 그리고 민원 발급을 위한 행정에서는 민원인이 얼마나 많이 집중될지, 민원인이 얼마나 기다려야 할지 예측함으로써 민원 서비스의 품질 제고에 납세자 행동 예측이 필요할 것으로 보인다.

Ⅴ 불성실 사업자 관리를 위한 예측

□ 현행 불성실 사업자 관리

1. 사업자등록 단계

사업을 시작하려는 사업자는 「부가가치세법 제8조」에 따라 사업자등록을 해야 한다.[27] 사업자등록은 납세의무를 지는 사업자에 관한 정보를 세무관서의 대장에 수록하는 것을 의미한다. 행정적 측면에서 보면 단순히 사업자를 세무 관리의 대상으로 포함시키는 행위이지만, 실질적으로는 금융계좌의 개설, 세금계산서의 발행 등 기본적인 상행위를 위한 전제조건이 되기 때문에 사업을 시작하려는 자에게는 매우 중요한 절차이다.

사업자등록을 신청하려고 할 때는 사업자의 인적 사항, 신청 사유, 사업 개시일 등을 적은 사업자등록 신청서를 관할 세무서장에게 제출해야 한다.[28] 사업자등록 신청을 받은 세무서장은 신청일로부터 2일 이내에 사업자등록을 발급해야 하나, 사업장 현장 확인 등을 위해 필요한 경우 발급기한을 5일 이내에서 연장하고 조사한 사실에 따라 사업자등록증을 발급할 수 있다.[29] 실제 현장에서는 대부분의 사업자등록은 즉시 발급되고 있으며, 사업장 등에 대한 현장 확인이 필요한 경우에만 시일이 다소 소요되고 있다.

---

27) 부가가치세법 제8조(사업자등록) ① 사업자는 사업장마다 대통령령으로 정하는 바에 따라 사업 개시일부터 20일 이내에 사업장 관할 세무서장에게 사업자등록을 신청하여야 한다. 다만, 신규로 사업을 시작하려는 자는 사업 개시일 이전이라도 사업자등록을 신청할 수 있다.

28) 부가가치세법 시행령 제11조(사업자등록 신청과 사업자등록증 발급) ① 법 제8조제1항에 따라 사업자등록을 하려는 사업자는 사업장마다 다음 각 호의 사항을 적은 사업자등록 신청서를 관할 세무서장이나 그 밖에 신청인의 편의에 따라 선택한 세무서장에게 제출(국세정보통신망에 의한 제출을 포함한다)해야 한다.
1. 사업자의 인적사항
2. 사업자등록 신청 사유
3. 사업 개시 연월일 또는 사업장 설치 착수 연월일
4. 그 밖의 참고 사항

29) 부가가치세법 시행령 제11조(사업자등록 신청과 사업자등록증 발급) ⑤ 제1항이나 제2항의 신청을 받은 사업장 관할 세무서장은 사업자의 인적사항과 그 밖에 필요한 사항을 적은 사업자등록증을 신청일부터 2일 이내(토요일, 「관공서의 공휴일에 관한 규정」 제2조에 따른 공휴일 또는 「근로자의 날 제정에 관한 법률」에 따른 근로자의 날은 산정에서 제외한다. 이하 이 항에서 같다)에 신청자에게 발급하여야 한다. 다만, 사업장시설이나 사업현황을 확인하기 위하여 국세청장이 필요하다고 인정하는 경우에는 발급기한을 5일 이내에서 연장하고 조사한 사실에 따라 사업자등록증을 발급할 수 있다.

신규로 사업을 시작하려는 자가 사업 개시일 전에 사업자등록을 신청하였지만, 사실상 사업을 개시하지 않을 것으로 관할 세무서장이 판단하면 사업자등록을 거부할 수 있다.[30] 아울러 사업자등록을 한 자가 폐업했거나, 사실상 사업을 하고 있지 않다면 사업자등록을 말소하여야 한다.[31] 즉 사업자등록은 사업을 영위하는 자에게만 발급되는 것이며, 사업을 영위하지 않을 때는 폐업·휴업신고를 해야 한다. 사업을 영위하지 않으면서도 폐업·휴업신고를 하지 않을 경우에는 관할 세무서가 직권으로 말소 조치를 할 수 있다.

사업자등록 발급에 있어서 어떤 신청자를 현장 확인 대상으로 분류할 것인지는 오랜 기간 국세행정의 과제였다. 모든 사업자등록 신청자를 현장 확인 대상으로 분류할 수도 있지만, 행정 부담이 커지고 사업자등록에 지나치게 오랜 시간이 소요될 수 있다는 문제가 있다. 2021년 연간 신규 사업자 수가 145만 명에 달한다는 점을 고려해 보면 사업자등록 신청자 모두를 현장 확인 대상으로 분류하는 것이 얼마나 현실성 없는 일인지 짐작할 수 있을 것이다. 반대로 모든 사업자등록을 즉시 발급할 수도 있다. 하지만 실제 사업을 하지 않는 자에게 사업자등록이 발급될 경우 사기, 조세 포탈, 명의 위장 등 각종 불법적 행위에 사업자등록이 악용될 가능성이 있다. 따라서 건전한 경제 활동을 장려하기 위해서는 사업자등록을 면밀하게 관리할 필요가 있다.

이와 같은 딜레마적 상황을 타개하기 위해 최근에는 빅데이터 분석을 활용하고 있다. 지금까지는 사업자등록 신청 시 관할 세무서의 담당자가 인허가, 사업 이력 등 납세자의 과거 정보를 고려하여 현장 확인의 필요성을 판단해야 했으나, 2019년 8월 '사업자등록 예측 모델'을 개발하여 사업자등록이 거부될 확률을 담당자에게 제공해 주고 있다. 시범 운영 결과, 현장 확인 대상자로 선정된 납세자의 수는 전년 대비 1/3로 감소했음에도 불구하고, 사업자등록이 거부된 총 건수는 전년과

30) 부가가치세법 시행령 제11조(사업자등록 신청과 사업자등록증 발급) ⑦ 법 제8조제1항 단서에 따라 사업자등록의 신청을 받은 사업장 관할 세무서장은 신청자가 사업을 사실상 시작하지 아니할 것이라고 인정될 때에는 등록을 거부할 수 있다.

31) 부가가치세법 제8조(사업자등록) ⑨ 사업장 관할 세무서장은 제7항에 따라 등록된 사업자가 다음 각 호의 어느 하나에 해당하면 지체 없이 사업자등록을 말소하여야 한다.
1. 폐업한 경우
2. 제1항 단서에 따라 등록신청을 하고 사실상 사업을 시작하지 아니하게 되는 경우

유사한 수준이었다. 즉시 발급되는 건수가 늘어나면서 납세자의 편의는 향상되었고, 직원들의 현장 확인 업무가 감소하여 업무 효율이 늘어나는 일석이조의 효과가 있었다.

< 사업자등록 신청·정정 절차에 빅데이터 분석 도입 >

출처 : 국세청 보도자료. 2019.11.4. “국세청, 빅데이터로 사업자등록 즉시 발급률 높인다.”

2. 세원 관리 단계

국세청은 사업자등록이 발급된 이후에도 사업자가 세법에 따른 의무를 성실하게 이행하는지 계속 점검하고 관리한다. 신고한 대로 사업을 영위하는지, 신고하는 데 어려움을 겪고 있지 않은지, 신고 내용이 사실과 부합하는지, 체납이 발생하는지 등을 끊임없이 점검한다. 또한 사업자등록을 하지 않고 음성적으로 사업을 영위하는 사업자가 없는지, 세법의 허점을 이용한 편법적 탈세 행위가 확산되고 있지 않은지 등을 파악하기 위해 노력한다. 이러한 행정을 통상 ‘세원 관리’라고 부른다.

세원 관리의 궁극적인 목표는 납세자가 세법에서 정한 정상적인 궤도 내에서 상행위를 하도록 하고 세법상 의무를 성실히 이행하도록 유도하는 데 있다. 상당히 다양한 형태의 행정이 세원 관리라는 개념 내에

포함될 수 있으나, 보통 권력적 행정에 기초하고 있는 탈루 세금의 징수, 체납 세금의 추징 등은 세원 관리의 범위를 넘어서는 것으로 본다.

세원 관리의 범주에 포함되는 활동들이 매우 다양하기 때문에, 이러한 활동을 유기적으로 연계하여 시너지를 만드는 데에는 많은 어려움이 있다. 우선 여러 행정이 서로 다른 부서 또는 담당자에 의해 행해지다 보니 행정의 결과물들을 종합적으로 이해하고 의미를 통찰해 내기가 쉽지 않다. 국세청 내부적으로는 부서 간 협력과 정보·자료의 공유를 통해 이러한 한계를 극복하고자 부단한 노력을 기울이고 있다. 하지만 이와 같은 노력의 성패는 대개 담당자 개인의 역량과 관심에 달려있기 때문에, 대국민 행정의 품질이 담당 직원에 따라 달라지는 결과가 발생하기도 한다.

또한, 사업자 수가 국세 공무원의 수에 비해 압도적으로 많기 때문에 개별 공무원의 관심과 노력만으로는 세원 관리의 효과성을 높이는 데 한계가 있을 수밖에 없다. 사업자 수는 2007년 500만 명에서 2021년 920만 명으로 2배 가까이 증가했다. 하지만 국세 공무원의 규모는 2007년에 2만 명을 넘어선 이후 15년이 지난 시점에도 2만 1천여 명에 그치고 있을 뿐이다. 국세 공무원 1명이 담당해야 하는 사업자 수가 빠르게 증가하고 있기 때문에, 과거처럼 개별 공무원이 담당 구역 내 사업자의 사정을 일일이 파악하고 관리하는 것은 기대할 수 없다.

3. 세무조사 및 체납 징수 단계

국세청은 사업자가 부가가치세를 신고 기간에 신고하지 않았거나 신고 내용에 오류·누락이 있을 경우 「부가가치세법 제57조」에 따라 부가가치세 과세표준과 납부·환급세액을 조사하여 결정·경정한다. 이처럼 신고를 하지 않았거나 신고 내용에 문제가 있을 경우 과세관청이 직접 납부·환급세액을 정하는 절차는 「소득세법[32]」과 「법인세법[33]」에도 규정

32) 소득세법 제80조(결정과 경정) ① 납세지 관할 세무서장 또는 지방국세청장은 제70조, 제70조의2, 제71조 및 제74조에 따른 과세표준확정신고를 하여야 할 자가 그 신고를 하지 아니한 경우에는 해당 거주자의 해당 과세기간 과세표준과 세액을 결정한다.
② 납세지 관할 세무서장 또는 지방국세청장은 제70조, 제70조의2, 제71조 및 제74조에 따른 과세표준확정신고를 한 자(제2호 및 제3호의 경우에는 제73조에 따라 과세표준확정신고를 하지 아니한 자를 포함한다)가 다음 각 호의 어느 하나에 해당하는 경우에는 해당 과세기간의 과세표준과 세액을 경정한다.

되어 있다. 과세표준과 세액을 결정 또는 경정하기 위해 세법에 규정된 권한에 근거하여 질문을 하거나 장부, 서류 또는 그 밖의 물건을 검사·조사하고 그 제출을 명하는 활동을 ‘세무조사’라고 한다.[34]

세무조사의 대상자 선정은 크게 정기선정과 비정기선정으로 구분할 수 있다. 정기선정은 납세자의 신고 내용을 바탕으로 정기적으로 성실도를 분석한 결과 불성실 혐의가 있다고 인정되는 경우, 최근 4 과세기간 이상 같은 세목의 세무조사를 받지 않은 납세자의 신고 내용이 적정한지 검증할 필요가 있는 경우 등에 실시된다. 비정기선정은 세법에서 정한 신고, 성실신고확인서 제출 등의 납세 협력 의무를 이행하지 않은 경우, 무자료, 위장·가공거래 등 거래 내용이 사실과 다른 혐의가 있는 경우, 납세자에 대한 탈세 제보가 있는 경우 등에 실시된다.

세무조사의 대상자 선정은 납세자의 신고 자료뿐만 아니라 국세청이 보유한 다양한 자료들을 비교·검토하여 이루어지는데, 최근에는 세무조사 대상자 선정에도 빅데이터 분석을 활용하고 있다. 구체적인 분석 과정과 모델은 공개되고 있지 않으며 일부 사례들만 보도된 바 있다. 그중 하나가 차명계좌 분석시스템이다. 2020년에 국세청은 세금계산서, 현금영수증, 친인척자료 등의 다양한 과세정보와 차명계좌 입·출금자 정보를 종합하여 사업자가 차명계좌를 이용하여 소득을 속이고 있을 가능성을 판단해 주는 시스템을 개발하였다. 이 시스템을 통해 보고된 탈루 의심 사례들은 세무조사 대상자 선정 등에 활용되었다.

---

1. 신고 내용에 탈루 또는 오류가 있는 경우
2. 제137조, 제137조의2, 제138조, 제143조의4, 제144조의2, 제145조의3 또는 제146조에 따라 소득세를 원천징수한 내용에 탈루 또는 오류가 있는 경우로서 원천징수의무자의 폐업ㆍ행방불명 등으로 원천징수의무자로부터 징수하기 어렵거나 근로소득자의 퇴사로 원천징수의무자의 원천징수 이행이 어렵다고 인정되는 경우

(이하 각호 생략)

33) 법인세법 제66조(결정 및 경정) ① 납세지 관할 세무서장 또는 관할지방국세청장은 내국법인이 제60조에 따른 신고를 하지 아니한 경우에는 그 법인의 각 사업연도의 소득에 대한 법인세의 과세표준과 세액을 결정한다.

② 납세지 관할 세무서장 또는 관할지방국세청장은 제60조에 따른 신고를 한 내국법인이 다음 각 호의 어느 하나에 해당하는 경우에는 그 법인의 각 사업연도의 소득에 대한 법인세의 과세표준과 세액을 경정한다.

1. 신고 내용에 오류 또는 누락이 있는 경우

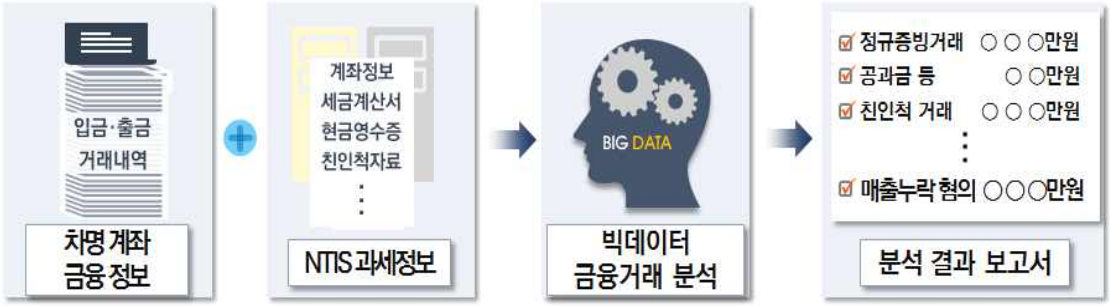
(이하 각호 생략)

34) 국세기본법 제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다.

(1호 ~ 20호 생략)

21. “세무조사”란 국세의 과세표준과 세액을 결정 또는 경정하기 위하여 질문을 하거나 해당 장부ㆍ서류 또는 그 밖의 물건(이하 “장부등”이라 한다)을 검사ㆍ조사하거나 그 제출을 명하는 활동을 말한다.

< 차명계좌 등 금융거래 분석 절차 >



출처 : 국세청 보도자료. 2020.7.2. "국세청 빅데이터센터 지난 1년간 국세행정 혁신을 위한 발판 마련"

세무조사 외에도 권력적 국세행정을 대표하는 사례 중 하나가 체납된 세금의 추징이다. 국세를 지정된 납부 기한까지 납부하지 않은 것을 '체납'이라고 하는데, 체납이 발생하면 「국세징수법」에 따른 강제 징수 절차가 진행되게 된다. 독촉, 압류, 공매, 청산의 강제 징수 절차가 진행되면 납세자의 재산권에 직접적인 피해가 발생하기 때문에 강제 징수 절차 이외에도 체납된 세금의 납부를 간접적으로 압박하는 방법을 사용하기도 한다. 대표적인 예로는 신용정보집중기관에 체납자료 제공, 사업에 관한 허가 등의 제한, 출국금지, 명단공개 등을 들 수 있다.

국세청은 체납자의 실거주지를 파악하여 재산을 압류하고 체납된 세금을 추징하는 데 데이터 분석을 활용하고 있다. 고액 체납자의 체납 행태가 날로 교묘해지면서 일선 세무서에서는 체납 사실을 파악하고도 체납자의 소재를 찾지 못해 체납된 세금을 걷지 못하는 사례들이 많았다. 이와 같은 애로를 해소하기 위해 2019년에 출범한 국세청 빅데이터센터는 다양한 데이터를 활용하여 주민등록상 주소지가 아닌 곳에 거주하는 것으로 분석된 체납자의 실제 거주지를 추정하였다. 이를 통해 체납자 28명에 대한 시범 조사(Pilot Test)에서 24명의 실거주지를 찾아내는 등 상당히 높은 적중률을 보여주기도 하였다.[35)]

35) 국세청. 2021. "2020년 자체평가 결과보고서(주요정책 부문)"

□ 데이터 과학의 적용 방향

현재 데이터 과학이 국세행정에 적용되고 있지만, 사업자등록, 세무조사 대상자 선정 등 분야별로 적용되고 있다. 필요성이 높은 분야부터 우선 데이터 과학을 적용하다 보니 다양한 분석 기법들이 독립적으로 활용되고 있다. 이처럼 세분화된 분야별로 데이터 과학을 적용하면 모델의 조정 및 관리가 용이하다는 장점도 있지만, 개업부터 폐업까지 사업의 생애주기를 유기적으로 다루는 데는 한계가 있을 수밖에 없다. 사업의 특정 시기가 아니라 생애의 관점에서 바라보고 데이터 과학을 적용한다면 더 효과적인 불성실 사업자 관리가 가능할 것이다.

또한, 현재 세무조사 대상 선정, 체납자 실거주지 파악 등에 사용되는 분석은 실제로 사건이 발생한 이후에 문제가 있는 사업자를 찾아내는 데 초점을 맞추고 있다. 세법을 위반한 사업자들을 사후에 적발해 내는 일은 매우 어렵기 때문에, 이러한 분야에 주로 데이터 과학이 적용되고 있다. 사후 적발은 불성실 사업자에 대응하는 최후의 수단으로서 작동하고 있지만, 그것만으로 불성실 사업 행태를 뿌리 뽑는 데는 한계가 있다. 결국에는 사업자들이 법의 테두리 내에서 사업을 영위하고 성실하게 세금을 납부하는 것이 당연한 일로 여겨지도록 스스로 행동양식을 바꿔야 하는데, 이를 위해서는 사전 관리가 매우 중요하다. 과세관청은 사업자의 문제를 미리 탐지하여 불성실한 사업 행태가 당연시되기 전에 안내·경고를 할 수 있어야 한다. 하다. 따라서 데이터 과학은 불성실 사업자의 사전 탐지 측면에서 적용되어야 할 것이다.

아울러, 국세행정은 세법의 근거에 따라 과세하고 집행하는 기관이기 때문에, 납세자들에게 언제나 집행의 이유를 설명할 수 있어야 한다. 불성실 사업자를 선별해 내기는 하지만 어떤 이유로 선별되었는지 시스템이 알려주지 않는다면, 결국 담당 직원이 이를 다시 분석해야 한다. 이처럼 중복적으로 진행되는 업무 프로세스는 데이터 과학이 기대하는 모습이 아니다. 따라서 국세행정에 적용되는 데이터 과학은 불성실 사업자를 효과적으로 탐지할 수 있어야 하겠지만, 불성실 사업자로 선별된 이유도 함께 제공해 줄 수 있어야 한다.

□ 다항 로지스틱 회귀분석을 이용한 불성실 사업자 예측

1. 다항 로지스틱 회귀분석의 개념

회귀분석(Regression)은 하나의 종속변수와 하나 이상의 독립변수 간의 관계를 구체적인 함수로 보여주고 설명하는 통계적 방법이다. 로지스틱 회귀분석(Logistic Regression)은 회귀분석의 하나로 종속변수가 성공 또는 실패 두 가지 값을 가질 수 있는 이진변수(Binary variable)인 경우 그 성공확률과 독립변수 간의 관계를 구체적인 함수로 보여주는 통계적 기법이다. 로지스틱 회귀분석은 어떤 대상을 대상이 가진 특성에 따라 두 개의 범주 중 하나로 분류하는 데 주로 사용된다. 어떤 독립변수의 변화가 성공확률에 어떤 영향을 미치는지 쉽게 설명할 수 있기 때문에, 화이트 박스[36] 모형의 하나로 볼 수도 있다.

다항 로지스틱 회귀분석(Multinomial Logistic Regression)은 종속변수가 2개의 값만 가질 수 있는 로지스틱 회귀분석을 종속변수가 3개 이상의 값을 가질 수 있는 경우로 확장한 것이다. 다항 로지스틱 회귀분석의 모델[37]을 구체적으로 설명해 본다. 종속변수 Y가 J개의 범주를 갖는다고 하자. $\underline{x}$라는 벡터로 표현되는 독립변수들의 값이 주어졌다고 할 때, 확률변수 Y가 j범주에 속할 확률은 아래와 같이 표현할 수 있다.

$$\pi_j(\underline{x}) = P(Y = j \mid \underline{x}),\ j = 1, 2, ..., J$$

다항 로지스틱 회귀분석은 $\pi_j(\underline{x}),\ j = 1, 2, ..., J$, 즉 종속변수가 각 범주에 속할 확률을 독립변수들을 이용하여 설명하는 데 초점을 맞추고 있다. 이를 위해 다항 로지스틱 회귀분석은 아래와 같이 각 종속변수의 범주와 하나의 기저 범주(baseline category)의 쌍으로 모델을 표현한다.

---

36) 화이트 박스 모형(white box model)은 내부 작동 구조가 사용자에게 알려져 있는 모델을 의미한다. 따라서 사용자는 모델이 어떤 과정을 거쳐 예측 또는 결정에 도달하는지 확인할 수 있다. 화이트 모형은 기계학습 분야에서 종종 언급되는 블랙 박스 모형(black box model)에 반대되는 개념이다. 블랙 박스 모형은 내부 작동 구조가 사용자에게 공개되어 있지 않다. 사용자는 모델에 입력되는 데이터와 출력되는 결과만 볼 수 있으며, 어떤 과정을 거쳐 예측 또는 결정에 이르게 되는지 이해하기 어렵거나, 불가능하다. 주로 신경망에 기반한 모형과 복잡한 기계학습 알고리즘이 블랙 박스 모형으로 여겨진다.

37) 종속변수가 가질 수 있는 값이 순서(Order)의 의미를 갖는 경우에도 다항 로지스틱 회귀분석이 사용될 수 있지만, 여기서는 종속변수가 명목 변수(Nominal Variable)인 경우만을 고려한다.

$$\log\frac{\pi_j(x)}{\pi_J(x)} = \alpha_j + \underline{\beta}_j^T \underline{x},\ \ j = 1, 2, ..., J-1,\ \ \sum_{j=1}^{J}\pi_j(x) = 1$$

$\pi_J(x)$는 종속변수가 기저 범주에 속할 확률을 의미한다. 즉, 다항 로지스틱 회귀분석은 기본적으로 기저 범주에 속할 확률에 대한 j범주에 속할 확률의 비(ratio)를 설명하는 모델이다. 물론 위의 연립 방정식을 풀면 $\pi_j(x),\ j = 1, 2, ..., J$의 관점에서 모델을 표현할 수도 있다.

$$\pi_j(x) = \frac{\exp\left(\alpha_j + \underline{\beta}_j^T \underline{x}\right)}{1 + \sum_{h=1}^{J-1}\exp\left(\alpha_h + \underline{\beta}_h^T \underline{x}\right)},\ \ j = 1, 2, ..., J-1$$

다항 로지스틱 회귀분석 모델의 모수는 α_j와 $\underline{\beta}_j^T$이며 이를 추정하기 위해 보통 최대우도추정법(Maximum Likelihood Estimation)[38]을 사용한다. 우도함수의 연립 방정식을 풀어 일반적인 형태를 갖는 최대우도추정량을 구하기는 어려우므로, 로그 우도함수가 오목함수(concave function)라는 점을 이용하여 보통 뉴턴-랩슨 알고리즘(Newton-Raphson method)[39]을 적용한 근사적인 해를 구하여 모수를 추정한다. 이렇게 모수를 추정한 다음에는 종속변수가 각 범주에 속할 확률을 추정할 수 있다. 예를 들어 아래와 같이 모수가 추정되었다고 해 보자.

$$\log\left(\widehat{\pi_1}/\widehat{\pi_J}\right) = -1.55 + 1.46x_1 - 1.66x_2$$

이는 지수함수를 이용하여 변형하면 아래와 같이 나타낼 수도 있다.

$$\widehat{\pi_1}/\widehat{\pi_J} = \exp\left(-1.55 + 1.46x_1 - 1.66x_2\right)$$

---

38) 확률분포로부터 추출된 표본으로부터 확률분포의 모수를 추정하는 방법 중 하나이다. 표본이 추출될 가능도(likelihood)를 가장 높게 만드는 모수를 선택하는 방법이다.

39) $f(x) = 0$을 만족하는 함수의 해를 수치적으로 구하는 방법 중 하나이다. 어떤 점 (x_0, y_0)가 주어졌을 때 $f(x)$의 접선과 x축과의 교점을 (x_1, y_1)이라고 하면 x_1이 x_0 보다 함수의 해에 더 가까워지는 기하학적 특성을 반복적으로 적용하여 함수의 근사적인 해를 구한다.

이는 x_1이 한 단위(unit)만큼 증가할 때 승산(odds)이라고 일컬어지는 확률의 비(ratio) π_1/π_f가 exp(1.46)≈4.31만큼 증가하는 경향을 보인다고 해석할 수 있다. 다른 독립변수가 종속변수에 미치는 영향에 대해서도 이와 유사한 방법으로 해석할 수 있다.

2. 다항 로지스틱 회귀분석의 적용

사업자는 사업자등록 또는 신고를 할 때 국세청에 다양한 정보를 제출한다. 아래의 표는 사업자등록 신청서 및 각종 신고서에 기재해야 하는 정보들을 간추려 본 것이다.

< 각종 신고·신청서에 기재해야 하는 정보(예시) >

신고서	분류	주요 신고내용
사업자등록 신청서	인적 사항	■상호 및 사업장 주소 ■대표자 성명 및 주민등록번호
	사업장 현황	■개업일, 업종, 종업원 수, 인터넷 주소 ■사업장 면적, 임대차 명세 ■허가 사업 여부 ■사업자금 명세
부가가치세 신고서	매출 관련	■세금계산서 발행분 매출액 및 세액 ■신용카드, 현금영수증 발행분 매출 명세
	매입 관련	■세금계산서 수취분 매입액 및 세액 ■매입처별 세금계산서 합계표 ■건물 등 감가상각자산 취득 명세
종합소득세 신고서	소득 관련	■이자·배당소득, 사업소득 등 종합소득 금액 ■비영업대금의 이익
	공제 관련	■인적공제 등 소득공제 명세 ■배당세액 공제 등 세액공제 명세
법인세 신고서	익금·손금	■장부 기재된 익금 및 손금 관련 내용 ex) 법인 신용카드 사용 금액, 기부금 내역,, 접대비 지출액 등
	공제·감면	■각종 법인세 공제·감면 내용 ex) 창업 중소기업 세액 감면액, 연구·인력 개발비 세액 공제액 등

사업자가 각 세법에 따라 제출하는 정보 외에도 국세청이 「과세자료의 제출 및 관리에 관한 법률」(이하 ‘과세자료법’이라 한다.)에 따라 수집하는 정보도 있다. 과세자료법에 따라 과세자료 제출 기관[40]은 법률에 따른 인가, 허가, 특허, 등록, 신고 등에 관한 자료, 조사·검사·감사 등의 결과, 법률에 따라 보고받은 영업·판매·생산·공사 등의 실적 자료 등을 국세청에 제출하여야 한다. 또한 조세 탈루 혐의 확인을 위해 금융거래 관련 정보 또는 자료가 반드시 필요한 경우에는 국세청장이 금융회사 등에 금융거래정보의 제출을 요구할 수도 있다. 과세자료 제출 기관이 제출해야 하는 구체적인 자료의 범위 및 제출시기는 과세자료법 시행령 별표로 규정되어 있는데, 그 일부가 아래의 표와 같다.

< 과세자료의 범위 및 제출 시기 >

번호	과세자료제출기관	과세자료명	받을 기관	제출시기
⋮	⋮	⋮	⋮	⋮
5	법무부	가. 「출입국관리법」 제3조 및 제6조에 따른 국민의 출국심사 및 입국심사에 관한 자료 중 출입국기록 나. 「출입국관리법」 제12조 및 제28조에 따른 외국인의 입국심사 및 출국심사에 관한 자료 중 출입국기록	국세청	매년 1월 31일, 7월 31일
6	법무부	「출입국관리법」 제34조제1항에 따른 등록외국인기록에 관한 자료 및 「재외동포의 출입국과 법적 지위에 관한 법률」 제6조에 따른 외국국적동포의 국내거소신고에 관한 자료	국세청	매일
⋮	⋮	⋮	⋮	⋮

그 밖에도 탈세 제보 자료, 자체적으로 수집하는 정보 등도 가용한 정보 중 하나로서 내부적으로 축적되고 있다.

40) 「과세자료의 제출 및 관리에 관한 법률 제4조(과세자료제출기관의 범위)」에 따르면 국가재정법에 따른 중앙관서와 하급 행정기관 및 보조기관, 지방자치단체, 금융감독원 및 금융회사, 공공기관 및 정부의 출연·보조를 받는 단체, 지방공사, 지방공단, 지방자치단체의 출연·보조를 받는 기관 또는 단체, 민법 외의 다른 법률에 따라 설립되거나 국가 또는 지방자치단체의 지원을 받는 기관이나 단체 중 중앙정부 또는 지방자치단체로부터 감독·감사·검사를 받는 기관이나 단체 등 상당히 광범위한 기관 또는 단체가 과세자료 제출 기관의 범위에 포함된다.

국세청이 수집하여 보유하고 있는 정보들을 이용하면 불성실 사업자를 예측하는 다항 로지스틱 모델을 만들 수 있다. 불성실 사업자를 포함한 사업자의 여러 유형들이 종속변수가 될 것이다. 예컨대 정상 사업자는 $Y=0$, 고액·상습체납자는 $Y=1$, 탈루 혐의자는 $Y=2$로 종속변수 Y를 정의해 볼 수 있을 것이다. 또한 개업일, 업종, 사업장 소재지, 매출액 등 국세청이 보유한 각종 정보는 사업자의 유형을 예측하기 위한 독립변수로 활용될 수 있을 것이다.

실제로 신용 위험을 모델링[41]하거나, 부도 가능성이 있는 은행을 조기 탐지[42]하기 위해 다항 로지스틱 회귀분석을 응용한 연구가 있다. 전자의 경우 개인의 신용한도, 소득, 신용점수, 신용카드 수 등을 독립변수로 이용하였으며, 후자의 경우 자본 적정성 비율, 자본 건전성 비율, 이익, 유동성 등 금융기관의 각종 금융·재무구조 비율을 이용하였다. 컴퓨터 보안 분야에서도 이상 침입을 탐지하기 위해 시스템 공격과 관련된 위험 요소들을 이용하여 다항 로지스틱 회귀분석을 적용한 결과 오분류율이 낮은 모델을 구현할 수 있었다는 연구[43]도 있다.

시뮬레이션을 통해 다항 로지스틱 회귀분석을 이용한 불성실 사업자 예측을 간단히 설명해 본다. 먼저 종속변수(category)는 3개의 범주 중 하나의 값을 갖도록 설정했다. 정상 사업자일 경우 category=0, 악성 체납자일 경우 category=1, 탈루 혐의자일 경우 categoty=2로 정하였고, 각각의 유형이 나올 확률이 0.7, 0.2, 0.1이 되도록 설정하였다. 예측을 위한 독립변수로는 먼저 사업 기간(period)을 사용하였다. 사업 기간이 길수록 사업자 수가 상대적으로 적어지는 점을 고려하여 사업 기간은 λ=1인 지수분포를 따르는 확률변수로 설정하였다. 이윤(profit)은 통상 비대칭적인 분포를 보이므로 평균 1, 표준편차 0.25인 log-normal 분포로 가정하였다. 체납 기간(overdue)은 악성 체납자가 아닌 경우 짧은 기간 체납이 발생할 뿐 장기 체납은 존재하지 않으므로 λ=10인 지수분포로

41) M R Adha, S Nurrohmah, and S Abdullah. 2018. “Multinomial Logistic Regression and Spline Regression for Credit Risk Modelling.” Journal of Physics: Conference Series 1108 (1): 1. doi:10.1088/1742-6596/1108/1/012019.

42) Qurriyani, Tengku, Early Detection of Potential Bank Bankruptcy Through Financial Ratio Analysis: Multinomial Logistic Regression Model (January 18, 2013). Available at SSRN: https://ssrn.com/abstract=2379517 or http://dx.doi.org/10.2139/ssrn.2379517

43) Wang, Yun. 2005. “A Multinomial Logistic Regression Modeling Approach for Anomaly Intrusion Detection.” Computers & Security, November 1.

가정하였고, 악성 체납자의 경우 장기 체납이 많이 존재하므로 꼬리가 더 두터운 λ=1인 지수분포로 가정하였다. 세액(tax)은 이윤에 세율을 곱해 계산되는데, 탈세 혐의자의 경우 세율은 평균 0.08이고 표준편차가 0.01인 정규분포를 따른다고 보았으며, 탈세 혐의자가 아닌 경우 세율은 평균이 0.1이고 표준편차가 0.01인 정규분포를 따른다고 가정하였다.[44]

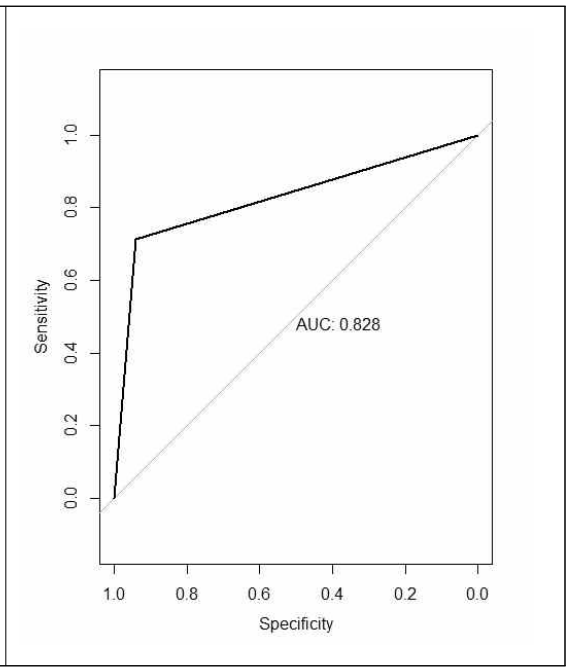
---

44) 이 시뮬레이션의 전체 R 코드는 아래와 같다.

```
# Import library
library("caret")
library("nnet")
library("pROC")

# Generate data
N <- 3000
set.seed(3)
category     <- sample(c(0, 1, 2), N, replace=TRUE, prob=c(0.7, 0.2, 0.1))
period       <- rexp(N, 1)
profit       <- rlnorm(N, 1, 0.25)
overdue      <- ifelse(category == 1, rexp(N, 1), rexp(N, 10))
tax          <- rnorm(N, ifelse(category == 2, 0.08, 0.1), 0.01) * profit
dt <- data.frame(category, period, profit, overdue, tax)

# Split data into train and test set
train.percent <- 0.8
training.rows <- createDataPartition(dt$category, p = train.percent, list = FALSE)
dt.train      <- dt[training.rows, ]
dt.test       <- dt[-training.rows, ]

# Estimate parameters using train set
test <- multinom(category ~ period + profit + overdue + tax, data = dt.train)
coef <- summary(test)$coefficients

# Predict probability of each category using test set
exp1 <- exp(coef[1, 1] + coef[1, c(2:5)] %*% t(dt.test[, c(2:5)]))
exp2 <- exp(coef[2, 1] + coef[2, c(2:5)] %*% t(dt.test[, c(2:5)]))
pp1 <- exp1 / (1 + exp1 + exp2)
pp2 <- exp2 / (1 + exp1 + exp2)

predict1  <- as.numeric(pp1 > 0.3)
category1 <- as.numeric(dt.test$category == 1)
roc(category1 ~ predict1, plot = TRUE, print.auc = TRUE)
confusion1 <- confusionMatrix(as.factor(predict1), as.factor(category1))
print(confusion1)

predict2  <- as.numeric(pp2 > 0.2)
category2 <- as.numeric(dt.test$category == 2)
roc(category2 ~ predict2, plot = TRUE, print.auc = TRUE)
confusion2 <- confusionMatrix(as.factor(predict2), as.factor(category2))
print(confusion2)
```

이와 같은 설정 하에 3,000개의 데이터를 생성하였고, 그중 랜덤하게 선택된 80%의 데이터는 모형 적합에, 20%는 모형 평가에 사용하였다. 적합된 모형과 모형 평가 데이터로 계산한 악성 체납자에 대한 예측 확률이 0.3을 초과하면 악성 체납자로 판단하였고, 탈세 혐의자에 대한 예측 확률이 0.2를 초과하면 탈세 혐의자로 판단하였다. 종합적 결과는 아래와 같다. 왼쪽 Confusion Matrix의 행은 적합된 모형에 의해 예측된 결과를, 열은 실제 결과를 의미한다. 악성 체납자일 것으로 예측한 118명 (= 28명 + 90명) 중 90명이 실제 악성 체납자였다. 그 비율 76%는 음성 예측도(Negative Predictive Value)라고 한다.[45]

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 446  36
         1  28  90

               Accuracy : 0.8933
                 95% CI : (0.8658, 0.9169)
    No Information Rate : 0.79
    P-Value [Acc > NIR] : 1.543e-11

                  Kappa : 0.6708

 Mcnemar's Test P-Value : 0.3816

            Sensitivity : 0.9409
            Specificity : 0.7143
         Pos Pred Value : 0.9253
         Neg Pred Value : 0.7627
             Prevalence : 0.7900
         Detection Rate : 0.7433
   Detection Prevalence : 0.8033
      Balanced Accuracy : 0.8276

       'Positive' Class : 0
```



오른쪽 그림은 ROC 곡선[46]이라고 하며, AUC[47]라는 값과 함께 모형의 성능을 평가하는 지표로 사용된다. ROC 곡선이 대각선에서 멀리 떨어져

45) 왼쪽 표의 수치들은 모형의 성능을 평가하는 지표들이다. 대표적으로 사용되는 몇 가지 지표를 설명해 본다. 민감도(Sensitivity)는 실제 정상 사업자 중에 정상 사업자로 예측된 사업자의 비율을 의미하며, 특이도(Specificity)는 실제 악성 체납자 중에 악성 체납자로 예측된 사업자의 비율을 의미한다. 양성 예측도(Positive Predictive Value)는 정상 사업자로 예측된 사람 중에 실제 정상 사업자의 비율을 의미한다.

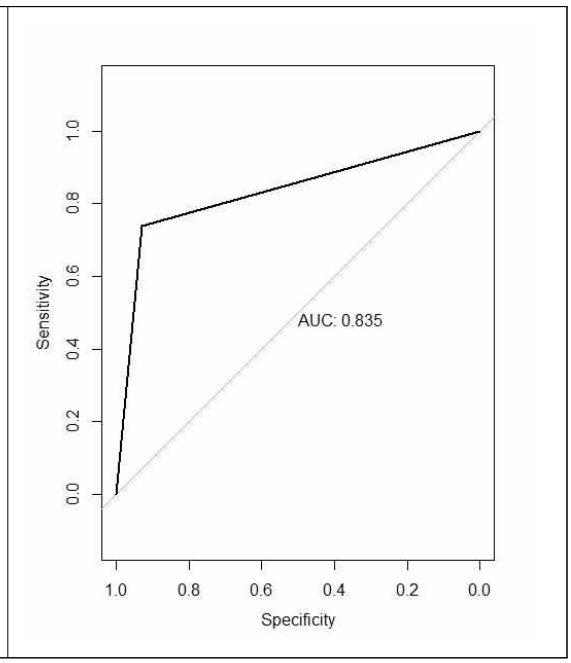
46) ROC는 Receiver Operating Characteristic의 줄임말이며 다양한 임계점(threshold)에 대해 모델의 성능을 특이도(Specificity), 민감도(Sensitivity)로 평가하여 하나의 평면에 표현한 것이다. ROC 곡선이 대각선 위에 있을 때는 동전 던지기와 같은 우연을 이용하여 분류하는 것과 다르지 않은 모델임을 의미하는 것이며, ROC 곡선이 대각선에서 멀어질수록 좋은 성능을 보이는 모델이라고 평가한다.

47) AUC는 Area Under the Curve의 줄임말이며, ROC 곡선의 아래쪽 면적의 크기를 의미한다. 여러 ROC가 교차하면 모델 비교가 쉽지 않은데, AUC를 이용하면 하나의 수치로 모델을 비교할 수 있다.

있고 AUC가 0.83로 상당히 크므로 모형이 악성 체납자를 잘 예측하는 것으로 평가할 수 있다.

아래는 탈세 혐의자에 대한 모형의 예측 결과이다. 총 82명(= 27명 + 45명)을 탈세 혐의자로 예측하였고 그중 45명이 실제 탈세 혐의자였다. 음성 예측도는 0.55이다. ROC 곡선이 대각선에서 멀리 떨어져 있고, AUC 값도 커서 대체로 모형이 잘 작동한다고 판단할 수 있다.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 502  16
         1  37  45

               Accuracy : 0.9117
                 95% CI : (0.886, 0.9331)
    No Information Rate : 0.8983
    P-Value [Acc > NIR] : 0.15526

                  Kappa : 0.5805

 Mcnemar's Test P-Value : 0.00601

            Sensitivity : 0.9314
            Specificity : 0.7377
         Pos Pred Value : 0.9691
         Neg Pred Value : 0.5488
             Prevalence : 0.8983
         Detection Rate : 0.8367
   Detection Prevalence : 0.8633
      Balanced Accuracy : 0.8345

       'Positive' Class : 0
```



이와 같이 불성실 사업자를 예측하는 다항 로지스틱 회귀분석 모델을 구현하면 임의의 사업자가 고액·상습체납자일 확률이 얼마일지, 탈루 혐의자일 확률이 얼마일지 예측해 볼 수 있다. 불성실 사업자일 확률이 일정한 수준 이상인 사업자들을 선별하여 세원 관리 또는 세무조사를 담당하는 직원들에게 제공한다면, 과거 경험 많은 직원의 직관에 의존해야 했던 불성실 사업자 관리 방식을 과학적이고 시스템적인 방식으로 전환하는 중요한 계기가 될 수 있을 것이다.

어떻게 작동하는지 이해하기 어려운 신경망 모델과 달리, 다항 로지스틱 회귀분석은 해당 사업자가 고액·상습체납자 또는 탈세 혐의자일 확률을 높게 평가한 이유를 알려준다. 따라서 담당 직원들이 사업자에게 성실

신고 안내문을 발송하거나 사업장을 방문하여 행정지도를 할 때 어떤 부분을 안내하고 설명해야 하는지 참고할 수 있는 자료를 제공해 줄 수 있다. 이로써 데이터에 기반한 국세행정이 실현될 수 있는 기초가 마련되는 것이다.

3. 고려해야 할 사항

다항 로지스틱 회귀분석은 모수를 추정하기 위해 근사적인 방법을 사용하기 때문에, 전국적인 단위로 모형을 구축하기 위해서는 방대한 자료와 복잡한 모형을 다룰 수 있는 고성능 컴퓨터와 DB가 필요하다. 아울러 데이터를 정제하고 모델을 조정할 수 있는 통계 전문 인력도 필요하다. 결국 상당한 인력, 예산, 시간이 투입될 수밖에 없다. 게다가 이와 같은 모델을 NTIS에 편입시켜 업무 절차의 하나로 통합하려면 많은 노력이 투입되어야 할 것으로 예상된다.

따라서 중앙 집중적으로 모델을 개발하기보다는 세무서 단위로 모델을 구축하는 방법도 고려해 볼 수 있을 것이다. 각 지역의 사정을 감안하여 모델의 독립변수를 추가하거나 조정하면 더 효과적인 모델을 개발할 수 있을 것으로 생각된다. 또한, 일부 세무서를 시범 삼아 불성실 사업자 예측 모델을 개발하고, 효과에 대한 평가와 미비점에 대한 보완을 거쳐 여타 세무서로 확대해 나간다면 대규모 프로젝트의 실패에 따르는 부담도 최소화할 수 있을 것으로 생각한다.

더불어, 다항 로지스틱 회귀분석 모형이 적합(학습)되기 위해서는 먼저 정상 사업자 범주에 어떤 사업자들이 속하는지, 고액·상습체납자 범주에는 어떤 사업자들이 속하는지 등을 명확히 구분한 학습 데이터가 필요하다. 사업자를 임의로 추출하여 세무조사를 실시한 뒤에 어떤 사업자 유형에 속하는지 분류하는 방법을 고려해 볼 수 있다. 이를 위한 법적 근거는 이미 마련되어 있다. 국세기본법 제81조의6 제2항 제3호는 세무조사 정기 선정을 실시하는 경우 중 하나로 ‘무작위추출방식으로 표본조사를 하려는 경우’를 규정하고 있다. 따라서 이에 근거하여 모형을 구축하는 데 필요한 데이터를 얻을 수 있을 것이다.

□ 이상점 분석을 이용한 불성실 사업자 탐지

1. 불성실 사업자 탐지의 어려움

‘정상 궤도를 이탈한 불성실 사업자에 대해서는 세무조사 등 가용한 역량을 모두 활용하여 대응한다.’는 표현은 국세청의 여러 자료에서 찾아볼 수 있다. 세법에서 정한 의무들을 성실히 이행하고 세법에서 정한 세금을 신고·납부하는 것을 정상적인 궤도에 해당한다고 보며, 이를 이행하지 않는 사업자를 불성실 사업자로 보고 있다.

문제는 어떤 사업자가 정상적 궤도를 이탈한 불성실 사업자인지 찾아내기 어렵다는 데 있다. 세무조사 대상을 선정하는 과정에서도, 본격적으로 세무조사에 착수하기에 전에도 사업자에 대해 다각적인 분석을 실시하지만, 실제 세무조사 결과가 사전 분석된 결과와 부합하지 않는 경우들이 종종 발생한다. 불성실 사업자도 정상적인 사업자와 동일한 외양과 행태를 보이는 경우가 많기 때문에, 실제 세무조사를 실시해 보기 전에는 그 문제를 구체적으로 파악하기 쉽지 않다.

사업자들이 정기적으로 제출하는 세금 신고서는 많은 정보들을 담고 있지만, 그것이 진실된 상행위에 기반한 것인지를 보여주지는 못한다. 진실한 신고를 위한 다양한 검증 장치들이 마련되어 있기는 하지만, 탈세를 의도한 사업자에게는 무용지물인 경우가 많다. 예컨대 사업자가 부가가치세 신고를 할 때는 매출·매입처별 세금계산서 합계표를 함께 제출해야 한다.[48] 거래처 간에 매입·매출처별 세금계산서 합계표를 상호 대사함으로써 거래의 누락이 있는지, 부당하게 공제받은 매입세액이 있는지 확인하기 위한 것이다. 하지만 실제 거래를 숨기려는 사업자들은 담합을 통해 세금계산서 발행 없이 현금으로 거래하기도 하고, 실제로 재화와 용역을 제공하지 않은 채 거래를 가장하여 세금계산서를 발행함으로써 부당하게 매입세액을 공제받기도 한다. 결국 이와 같은 탈세 문제는 사업자가 제출한 신고서만으로는 발견하기 어렵다.

---

48) 부가가치세법 제54조(세금계산서합계표의 제출) ① 사업자는 세금계산서 또는 수입세금계산서를 발급하였거나 발급받은 경우에는 다음 각 호의 사항을 적은 매출처별 세금계산서합계표와 매입처별 세금계산서합계표(이하 “매출ㆍ매입처별 세금계산서합계표”라 한다)를 해당 예정신고 또는 확정신고(제48조 제3항 본문이 적용되는 경우는 해당 과세기간의 확정신고를 말한다)를 할 때 함께 제출하여야 한다. (각호 생략)

2. 불성실 사업자 탐지를 위한 아이디어

일찍이 금융기관들은 사기 거래 탐지(fraud detection)을 위해 이상점(outlier) 분석에 기반한 시스템을 도입하여 운영하고 있다. 정상적 범주에서 벗어난 거래가 발생할 경우 사기 거래 탐지 시스템은 이용자에게 직접 경고를 보내거나, 금융기관의 담당자들이 해당 거래를 검토하고 확인할 것을 제안한다. 금융기관뿐만 아니라 인터넷 계정에 로그인할 때도 정상적인 접근인지, 해킹봇(bot) 등에 의한 비정상적인 접근인지 확인하기 위해 이상점 탐지에 기반한 시스템이 활용되고 있으며, 인터넷 광고 분야에서도 봇에 의한 가짜 클릭, 가짜 뷰(view)인지를 식별하기 위하여 사기 거래 탐지를 사용하기도 한다.

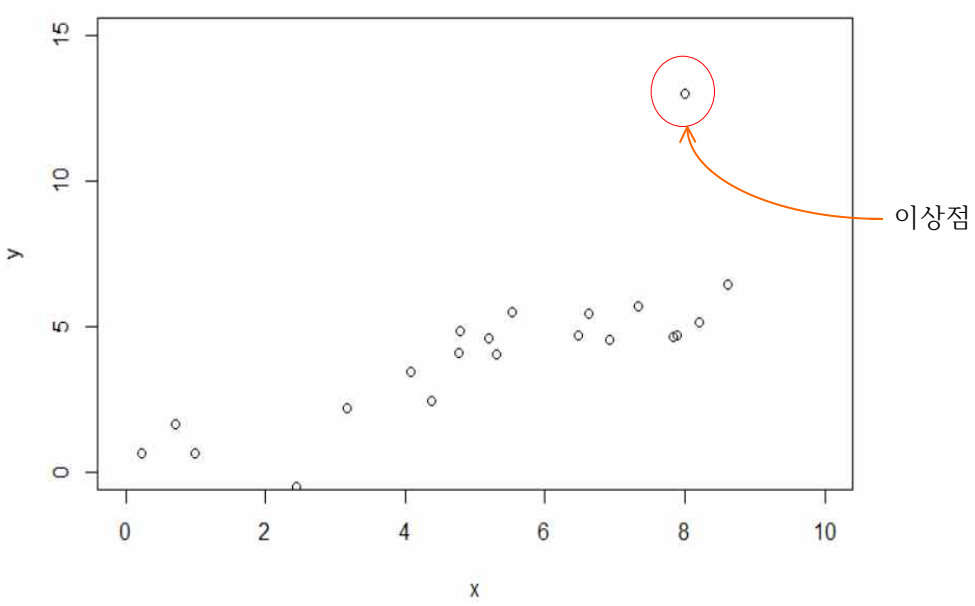
국세행정에서도 이상점 분석을 응용하면 불성실 사업자 탐지가 가능할 것으로 생각된다. 매출, 비용, 고객 수, 전기 사용량 등 사업의 행태를 보여주는 다양한 속성을 이용하여 사업자 간의 거리를 측정하고 그 중 동떨어져 있는 사업자를 찾아냄으로써 정상적 궤도에서 벗어난 불성실 사업자를 탐지할 수 있을 것이다. 다만, 탈세를 의도한 사업자가 탐지 시스템의 허점을 노려 매출, 매입 등 사업의 외양을 조정할 수 있으므로, 효과적으로 시스템이 작동하기 위해서는 사업의 속성을 보여주는 수치 중에서 사업자가 의도한 대로 조정할 수 없는 것들을 선택하는 것이 바람직할 것으로 보인다.

정상적인 범주에서 벗어난 사업자를 빠르게 탐지할 수 있다면 탈세 행위가 확대되고 만연하기 전에 차단할 수 있는 강력한 대응 수단을 얻게 될 것이다. 또한 일상적인 사업자 관리를 통해 탈세 대응을 위한 행정력의 소모도 크게 줄일 수 있을 것으로 보인다.

3. 이상점의 개념

이상점(outlier)에 대한 엄밀한 정의는 없다. 어떤 데이터를 이상점으로 볼 것인지는 상황과 맥락에 따라 달라질 수 있기 때문이다. 보통 다음 그림에서 볼 수 있듯이 관측된 데이터의 추세 또는 범위에서 동떨어져 있는 아주 작은 또는 아주 큰 값을 이상점이라고 한다.

< 이상점의 예시 >



이상점은 데이터 측정 또는 입력의 오류, 극단적 상황의 발생 등에 의해 나타날 수 있다. 이상점은 데이터 분석의 결과를 왜곡하여 잘못된 결론에 도달하도록 만들 수 있다. 예를 들어 이상점은 데이터의 평균과 표준편차에 영향을 미칠 수 있으며, 이러한 영향은 검정, 추정, 회귀분석 등의 통계 분석의 왜곡으로까지 이어지기도 한다. 예컨대 평균은 데이터 분석 과정에서 많이 활용되는 통계량 중 하나인데, 평균은 극단값에 민감한 특성을 갖고 있다. 따라서 이상점은 평균값을 크게 변화시키는 경향이 있다. 결국 이상점의 영향을 받은 평균은 데이터의 중심 위치라는 특성을 왜곡하여 보여주게 될 것이다.

따라서 이상점을 적절하게 식별하고 처리하는 일은 통계 분석에서 매우 중요하다. 이상점을 데이터에서 제거하는 것은 고려해 볼 수 있는 가장 간단한 처리 방법이지만, 이상점이 데이터에 관한 중요한 정보를 가지고 있을 수도 있으므로 개별 데이터를 직접 확인해 보고 왜 이상점이 발생했는지 원인을 찾은 다음에 어떻게 처리할지 결정하는 것이 바람직하다. 데이터에서 이상점을 제거하는 방법 이외에도 데이터를 변환하거나, 이상점에 덜 민감한 통계적 방법론(Robust Statistical Methods)을 사용하는 것도 하나의 대안으로 사용될 수 있다.

4. 이상점을 탐지하는 여러 방법

(1) 회귀분석에서 이상점의 탐색

회귀분석에서 이상점은 다양한 방법으로 탐색할 수 있다. 가장 손쉬운 방법은 앞서 보았던 것처럼 데이터의 산점도를 이용하는 것이다. 하지만 산점도를 이용할 경우 이상점 판단에 연구자의 주관이 개입될 수 있기 때문에, 숫자로 표현되는 값을 이용하여 이상점을 판단하는 방법들도 병행하여 사용하는 경우가 많다.

이와 같은 방법 중에는 잔차(residual)를 이용하는 방식이 주를 이룬다. 잔차는 데이터로부터 추정한 회귀식을 통해 예측한 값과 실제 관측값 사이에 나타나는 차이를 의미한다. 따라서 잔차가 크다는 것은 추정된 회귀식에서 많이 동떨어져 있는 값이라는 의미와 같다. 이러한 관계를 고려한다면 잔차가 매우 큰 값을 찾아 이상점이라고 판단하는 것도 하나의 방법이 될 것처럼 보인다.

하지만 잔차는 종속변수가 가지고 있는 단위의 영향을 받으므로 서로 다른 모델 또는 서로 다른 데이터 간에는 비교가 불가능하다. 따라서 이상점을 탐색할 때는 아래와 같은 방법으로 잔차를 표준화한 '표준화 잔차(statndardized residual)'를 많이 사용한다.

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

여기서 r_i가 표준화 잔차이다. e_i는 일반적인 방법으로 계산된 잔차를 의미한다. MSE는 평균제곱오차(Mean Squared Error)이다. 평균제곱오차는 잔차 제곱합의 평균으로 볼 수 있다.[49] h_{ii}는 i번째 관찰값의 레버리지를 의미한다.[50] 보통 표준화 잔차의 절대값이 2 ~ 3보다 클 때 이상점이라고 판단한다.

---

49) $MSE = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2$, (여기서 k는 독립변수의 개수를 의미한다.)

50) 잔차를 행렬로 표현하면 $e = [I - X(X'X)^{-1}X']y$로 나타낼 수 있는데, $X(X'X)^{-1}X'$의 i번째 대각선 위의 값이 h_{ii}이다. 레버리지는 실제 관측된 종속변수의 값이 예측된 값에 미치는 영향을 의미한다.

이상점 탐지에는 아래와 같이 정의되는 스튜던트화 잔차(Studentized Residual)를 사용하기도 한다. 여기서 MSE_i는 y_i를 제외하고 $n-1$개의 측정값으로부터 계산한 평균제곱오차이다. 표준화 잔차와 마찬가지로 스튜던트화 잔차가 2 ~ 3보다 클 때 이상점이라고 본다.

$$r_i = \frac{e_i}{\sqrt{MSE_i(1-h_{ii})}}$$

그 밖에도 Grubbs 검정, Dixon Q 검정 등 이상점 탐지를 위힌 검정을 실시할 수도 있으며, K 중심 군집화(K-centroid clustering) 및 기계학습을 이상점 탐지에 활용하기도 한다.

(2) 군집화(Clustering)를 이용한 이상점 탐색

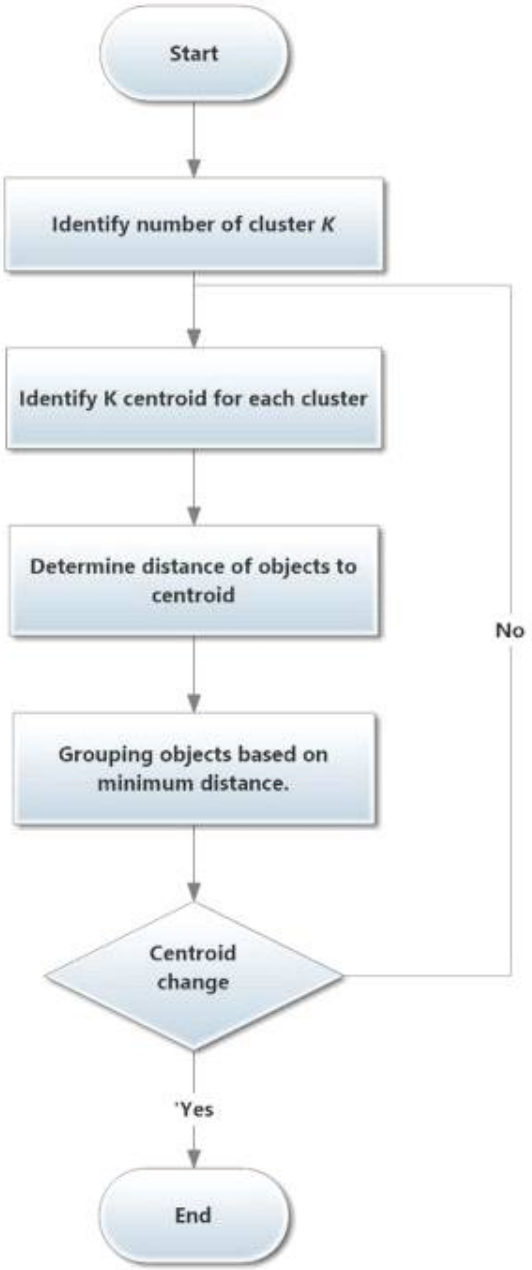
K-평균 군집화 알고리즘(K-means clustering algorithm)은 서로 다른 데이터 포인트들의 유사성을 기반으로 전체 데이터를 K개의 군집으로 구분하는 데 사용되는 알고리즘이다. K-평균 군집화 알고리즘은 아래와 같은 몇 개의 단계를 거쳐 작동한다.

① 군집의 개수(K) 설정 : 가장 먼저 해야 할 일은 군집의 개수를 결정하는 것이다. 군집의 개수에 따라 결과는 크게 달라지며 적절하지 않은 K가 선택될 경우 납득하기 어려운 결과를 보여주기도 한다. 군집의 개수를 결정하는 데는 Rule of thumb, Elbow Method, 정보 기준 접근법(Information Criterion Approach) 등을 사용할 수도 있다.

② 초기 중심점(Centroid) 설정 : 초기 중심점은 데이터의 무게 중심과 같은 역할을 한다. 초기 중심점을 어떻게 설정하는지에 따라 성능이 크게 달라지는데 초기 중심점으로 랜덤한 값을 선택할 수도 있고 어떤 값을 직접 선택할 수도 있지만, 보통 2007년 David Arthur와 Sergei Vassilvitskii는 제안한 K-mean++[51]를 많이 사용한다.

51) 데이터로부터 임의의 데이터를 하나 선택하여 첫 번째 중심점으로 설정한다. K 개의 중심점이 선택될 때까지 다음을 반복한다. ① 각 데이터와 선택된 중심점 중 가장 가까운 중심점과의 거리 D(x)를 계산한다 ② 확률이 $D(x)^2$에 비례하는 편중 확률 분포를 사용하여 임의의 데이터를 선택한 후 n번째 중심으로 선택한다. K 개의 중심점이 선택되면 이를 초기값으로 하여 K 평균 군집화를 수행한다.

< K-평균 군집화 알고리즘의 단계 >



출처 : Reham Arnous, Ali I. El-Desouky, Amany Sarhan and Mahmoud Badawy. "ILFCS: an intelligent learning fuzzy-based channel selection framework for cognitive radio networks"

③ 데이터를 군집에 할당 : 거리상 가장 가까운 중심점으로 데이터를 배정한다.

④ 중심점 재설정 : 데이터의 군집 배정이 완료되면 군집의 중심점을 그 군집에 속하는 데이터의 평균에 해당하는 지점으로 재설정한다.

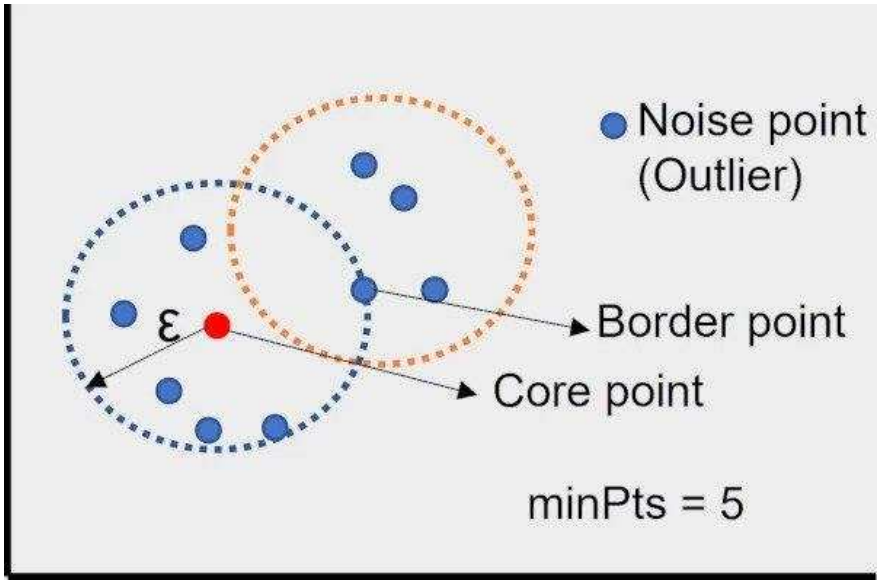
⑤ 데이터를 군집에 재할당 : 데이터를 다시 가까운 중심점으로 배정한다. 이와 같은 과정을 중심점이 이동하지 않을 때까지 반복한다.

오른쪽 순서도는 앞에서 설명한 K-평균 군집화 단계들을 간략하게 보여준다. 이상점 탐색을 위해서는 군집의 수를 2개 또는 3개로 설정한 다음 각 데이터가 어떻게 할당되었는지 확인하면 된다. 중심까지의 거리가 매우 큰 데이터를 이상값으로 간주할 수 있을 것이다. 또는 K를 큰 값으로 설정한 다음 데이터를 거의 포함하고 있지 않은 군집이 존재하는지 검사하는 방법을 고려해 볼 수도 있다. 이러한 군집들은 다른 데이터와 매우 다른 값을 가지는 군집이라고 할 수 있으므로 이상값으로 볼 수 있다. 다만, K-평균 군집화는 본래 이상점 탐지 목적의 알고리즘이 아니므로, 다른 방법들과 병행하여 사용할 때 보다 의미를 가질 수 있을 것이다.

군집화를 이용하는 방법 중에 DBSCAN(density-based spatial clustering application with noise)이라는 알고리즘도 있다. DBSCAN은 밀도의 관점에서 서로 근접한 데이터의 군집을 식별하는 밀도 기반 군집화 알고리즘으로 1996년 KDD'96 데이터 마이닝 컨퍼런스에서 처음 소개되었다. 이 알고리즘에 따르면 군집의 일부가 아닌 데이터를 구분해 낼 수 있는데 이를 이상점으로 간주할 수 있다.

DBSCAN에는 두 가지 중요한 하이퍼 파라메터가 존재한다. 하나는 이웃으로 여겨질 수 있는 두 데이터 사이의 최대 거리(반경)이며, 다른 하나는 군집으로 형성되기 위해 요구되는 최소한의 데이터의 개수(최소 이웃 데이터 수)이다. DBSCAN에서 각 데이터는 다른 데이터와 얼마나 인접해 있는지에 따라 핵심 포인트, 경계 포인트, 노이즈 포인트로 구분된다. 핵심 포인트는 지정된 반경 내에서 최소 이웃 데이터 수를 충족하는 데이터를 의미하며, 경계 포인트는 지정된 반경 내에 인접한 데이터가 존재하지만, 최소 이웃 데이터 수보다 적은 데이터를 의미한다. 노이즈 포인트는 지정된 반경 내에 인접한 데이터가 없는 데이터이다. 노이즈 포인트는 이상점으로 여겨 군집의 일부로 간주하지 않는다.

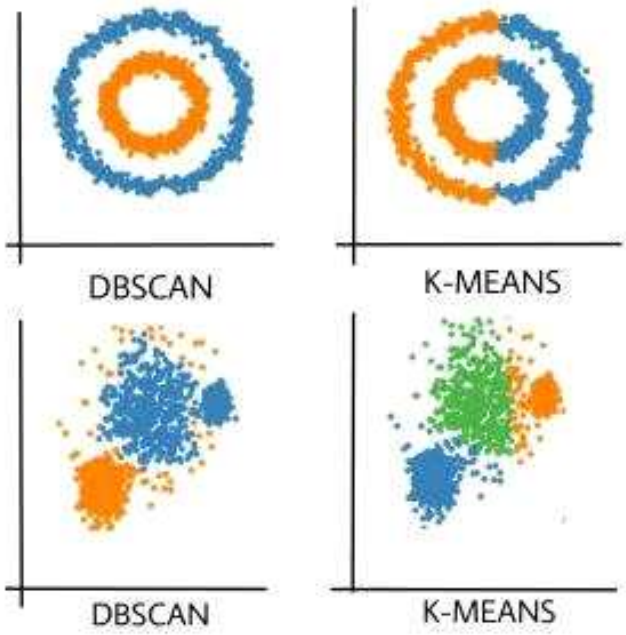
< DBSCAN 알고리즘 동작 원리 >



출처 : Renesh Bedre, "DBSCAN clustering algorithm in Python (with example dataset)"

DBSCAN 알고리즘은 밀도를 기준으로 하여 가까운 데이터를 군집에 포함시키기 때문에, 거리 기반의 K-평균화 군집과 다른 결과를 낳을 수밖에 없다. 어느 방법이 우월하다고 단정 짓기 어려우나 DBSCAN은 K-평균 군집화 알고리즘에 비해 클러스터의 개수를 미리 지정할 필요가 없고 이상점을 효과적으로 제어할 수 있다는 장점이 있다.

< DBSCAN과 K-평균 군집화 비교 >



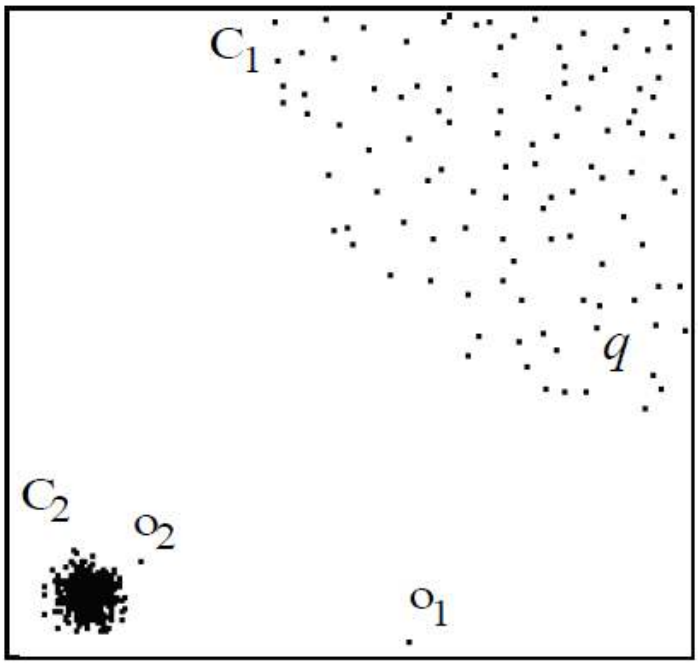
출처 : Abhishek mamidi, "DBSCAN - Density-Based Spatial Clustering of Applications with Noise"

(3) 기계 학습 및 신경망을 이용한 이상점 탐색

의사결정 나무(Decision Tree)를 지속적으로 분기시키면서 데이터의 고립 여부 정도에 따라 이상점인지 판단하는 Isolation Forest를 이상점 탐지에 이용할 수도 있다. 또한 전체 데이터의 분포에서 지역적 밀도(Density)를 고려하여 이상치를 판단하는 Local Outlier Factor를 이상점 탐색에 활용할 수도 있다. 밀도 기반의 이상점 탐지 방법은 밀집도가 다른 두 데이터 집단이 존재할 경우 이상점으로 판단할 임계치 결정이 어렵다는 단점이 있다. 예컨대 아래 그림에서 C1, C2로 구분되는 데이터가 있다고 할 때, 밀도 기반의 이상점 탐지 알고리즘은 o1의 경우 쉽게 이상점으로

탐지하지만 o2는 이상점으로 탐지하지 못하는 경우가 있다. 이를 보완하기 위한 알고리즘이 Local Outlier Factor이다. 그 밖에도 기계학습 알고리즘으로 대표적인 SVM(Support Vector Machine)이 이상점 탐지에 활용될 수 있다.[52)]

< 밀도 기반 이상점 탐지 알고리즘의 문제 >



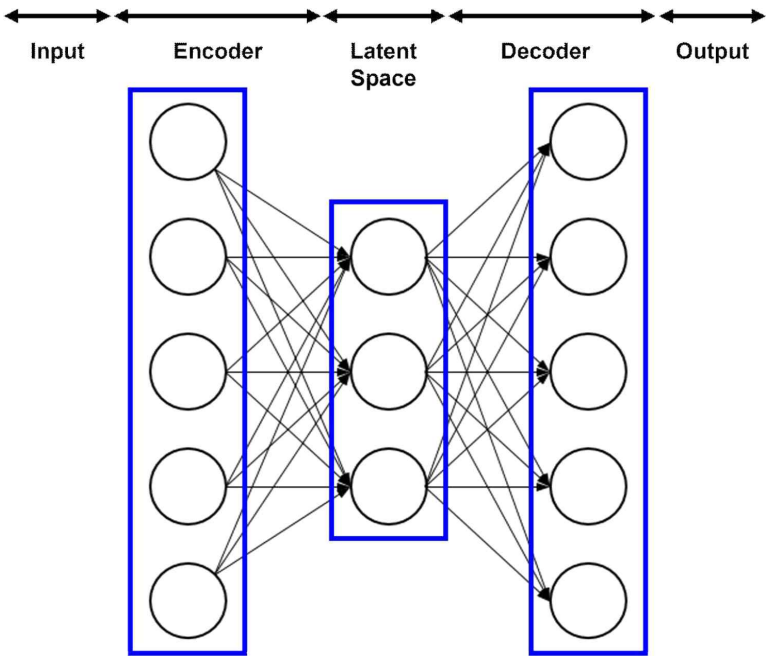
출처 : Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander, "LOF: identifying density-based local outliers"

이상점 탐지에 신경망(Neural Network)을 이용할 수도 있다. 신경망은 인간의 두뇌와 비슷한 계층 구조로 상호 연결된 노드 또는 뉴런을 사용하는 기계학습의 한 유형이다. 신경망을 이용한 이상점 탐지 방법에는 적대적 생성 네트워크(Generative Adversarial Network, GAN), 오토인코더(Autoencoder) 등이 있다. 그중에서 오토인코더에 대해 서재홍, 박준성, 유준우, 박희준은 다음과 같이 설명하고 있다.[53)]

52) Vasudeva Kilaru, 2022. "One Class Classification Using Support Vector Machines"
53) Jaehong Seo, Junsung Park, Joonwoo Yoo, and Heejun Park. 2021. "Anomaly Detection System in Mechanical Facility Equipment: Using Long Short-Term Memory Variational Autoencoder"

아래 그림은 Autoencoder의 구조이다. 인코더와 디코더로 구성되어 있으며, 입력과 출력 시의 노드 개수가 같다. Input layer에서 데이터를 받아 Hidden layer로 갈수록 그 수가 줄어들면서 정보를 압축하고 다시 Output layer로 복원하는 과정을 거쳐 재구축 오류를 생성하는데, 이때 발생하는 이 재구축 오류의 정도를 이상 점수(Anomaly score)로 사용하여 임계값(threshold)과 비교하여 데이터의 정상유무를 판단한다.

< 오토인코더의 구조 >



5. 이상점 탐지 기법의 국세행정 적용

앞서 설명했듯이, 이상점이란 명확하게 규정할 수 있는 개념이 아니다. 따라서 앞에서 설명한 방법들이 모두 동일하게 불성실 사업자를 분류해 내지 못한다. 결국 여러 방법을 병행하여 적용함으로써 서로 보완하는 방식으로 불성실 사업자를 탐지할 수밖에 없다.

매출액, 매입액 등은 사업자들 간에 담합을 통해 은닉하거나 가장할 수 있으므로 이러한 정보들은 불성실 사업자를 탐지하는 기초 자료로 활용하기에 적당하지 않다. 산출세액을 포함하여 전기 및 수도 사용량, 영업 허가, 사업장 면적, 공시지가, 유동 인구, 상권 밀집도 등 사업자가 스스로 통제할 수 없는 요인들을 이용하여 군집을 형성하거나 모형을

적합시킨 다음, 군집 또는 추정된 모형에서 동떨어져 있는 사업자가 존재하는지 확인하면 불성실 사업자를 탐지해 낼 수 있을 것이다.

DBSCAN을 이용하여 어떻게 불성실 사업자를 탐지할 수 있는지 설명하기 위하여 간단한 시뮬레이션을 실시해 보았다. 변수들 중에 사업장 면적(area)은 표준편차 10인 정규분포를 따르는 확률변수로 설정하였다. 다만, 평균은 100, 200 중에 임의로 하나가 선택되도록 하였다. 유동 인구(traffic)는 표준편차 150인 정규분포를 따르는 확률변수이며, 평균은 1000, 1500 중 하나가 임의로 선택되도록 하였다. 전력 사용량(power)은 면적에 100부터 200 사이의 값이 임의로 더해지도록 정하였다. 탈루액(tax_evasion)은 0의 값이 0.97의 확률로, -5000의 값이 0.03의 확률로 부여되도록 정하였다. 0은 정상적으로 신고하는 사업자를, -5000은 조세를 탈루하는 사업자를 의미한다. 산출세액은 아래와 같은 관계에 따라 계산되었다.

$$\text{세액} = 10000 + 15\times\text{면적} + 7\times\text{유동인구} + 6\times\text{전력 사용량} + \text{탈루액}$$

위와 같은 설정에 따라 R을 이용하여 100개의 Data를 임의로 생성하고 DBSCAN에 의해 탈루하는 사업자가 감지될 수 있는지 테스트해 보았다. 앞서 언급한 대로 DBSCAN는 2개의 하이퍼 파라메터를 갖는다. 두 데이터 사이의 최대 거리(epsilon)은 400으로 설정했으며, 최소 이웃 데이터 수는 5로 설정하였다. DBSCAN을 실행한 결과 4개의 군집이 만들어졌으며 100개의 데이터 중에서 14개가 노이즈(이상점)로 분류되었다.

```
DBSCAN clustering for 100 objects.
Parameters: eps = 400, minPts = 5
Using euclidean distances and borderpoints = TRUE
The clustering contains 4 cluster(s) and 14 noise points.

 0  1  2  3  4
14 44 17 16  9

Available fields: cluster, eps, minPts, dist,
                  borderPoints
```

14개의 데이터 중에서 탈루액(tax_evasion)이 -5000인 사례들이 얼마나 포함되었는지 확인해 보았다. 아래의 표가 그것을 보여준다.

```
> table(cluster, tax_evasion)
       tax_evasion
cluster -5000  0
      0     5  9
      1     0 44
      2     0 17
      3     0 16
      4     0  9
```

Cluster 0은 노이즈(이상점)로 분류된 사례들이다. 14개 중에 탈루액이 -5000인 사업자 5명 모두가 포함되어 있는 것을 볼 수 있다. 탈루액이 0인 정상 사업자 중에서 노이즈(이상점)으로 분류된 사례는 없었다. 매우 간단한 사례이지만 이상점 탐지를 통해 불성실 사업자를 어떻게 발견할 수 있는지를 보여주는 예시가 될 것으로 생각한다.[54]

실제 사례는 훨씬 복잡하기 때문에 정확하게 불성실 사업자를 탐지할 수 있도록 데이터를 정제하고 알고리즘을 가다듬는 데 많은 공을 들여야 할 것이다. 하지만 알고리즘이 성공적으로 동작한다면, 지금까지 막연하게 외쳐왔던 정상적인 궤도를 이탈한 불성실 사업자 관리가 실현될 수 있을 것으로 기대한다.

---

54) 이 시뮬레이션의 전체 R 코드는 아래와 같다.

```
# Simulate Data
N <- 100
set.seed(3)
area <- rnorm(N, 0, 10) + sample(c(100, 200), N, replace=TRUE)
traffic <- rnorm(N, 0, 150) + sample(c(1000, 1500), N, replace=TRUE)
power <- area + runif(N, 100, 200)
tax_evasion <- sample(c(0, -5000), 100, replace=TRUE , prob=c(0.97,0.03))
tax <- 15 * area + 7 * traffic + 6 * power + 10000 - tax_evasion
data <- data.frame(tax, area, traffic, power)

# Perform DBSCAN
library("dbscan")
dbscan_res <- dbscan(data, eps = 400, minPts = 5)
dbscan_res
cluster <- dbscan_res$cluster
table(cluster, tax_evasion)
```

Ⅵ 납세자 지원을 위한 예측

□ 납세자 지원을 위한 국세행정

1. 신고 과정에서의 납세자 지원

부가가치세, 소득세, 법인세, 개별소비세 등 대부분의 세목은 신고납세제도를 택하고 있다.[55] 신고납세제도는 납세의무 확정 권한을 1차적으로 납세자에게 부여하고 과세권자의 납세의무 확정 권한은 2차적 또는 보충적 지위로 유보해 두는 제도이다. 납세자가 스스로 납세의무를 확정하기 때문에, 조세저항이 적고 행정력을 절감할 수 있다는 장점이 있다.

하지만, 경제 및 사회 구조가 나날이 고도화되면서 납세자가 신고해야 할 내용도 점차 복잡해지고 있다. 이제 납세자에게 세금 신고는 자발적 납세의무의 확정이라는 의미보다는 잘못 신고할 경우 차후에 세금이 추징될 수도 있는 큰 부담이 되고 있다. 이런 상황에서 세금 신고를 납세자의 몫으로만 맡겨 둘 수 없어, 2010년대 중반부터 국세청은 세금 신고를 도와줄 수 있는 다양한 방안들을 도입하기 시작했다. 대표적인 방안으로는 미리채움서비스와 모두채움서비스를 들 수 있다.

납세자가 세금 신고를 하기 위해 홈택스에 접속하면 모든 또는 일부 항목에 미리 금액이 기재된 세금 신고서를 확인할 수 있다. 과거 신고 검증을 위해 보유하고 있던 자료들을 미리 신고서에 기재해 줌으로써 납세자의 신고 부담, 기재 오류 등을 줄이는 효과가 있었다. 예를 들어 영세사업자의 경우 수입금액, 필요경비, 납부(환급)세액 등 신고서의 모든 항목이 사전에 기재된 채로 신고서가 제공되며, 기재된 금액에 문제가 없는 경우에는 인터넷으로 제출만 하면 신고가 완료된다. 미리채움서비스의 경우에는 신고서의 일부 항목만 미리 기재하여 제공한다. 우편신고제도도 도입되었다. 납세자가 미리 채워진 신고서를 우편으로 받은 후 이상이 없음을 확인하고 회신하면 신고가 완료되는 제도이다.

---

55) 이와 대비되는 납세의무 확정 방식으로 정부부과제도가 있다. 상속세, 증여세, 종합부동산세의 경우 정부의 부과 처분에 의해 납세의무가 확정되며 결정된 통지서(고지서)가 납세자에게 도달하는 시점에 납세의무 확정의 효력이 발생한다.

< 부가기치세 신고 미리채움서비스 제공 항목과 일정(예시) >

No	구분	제공 항목(총30개)	제공일정
1	매출	전자세금계산서(거래처별 명세 포함) 매출 합계	23. 1.12.
2		신용카드 매출	23. 1.12.
3		판매·결제대행자료	23. 1.17.
4		현금영수증 매출	23. 1. 1.
5		내국신용장·구매확인서 전자발급금액	23. 1.15.
6		수출실적 내역(수출신고번호, 선적일, 수출액, 환율)	23. 1.11.
7	매입	전자세금계산서(거래처별 명세 포함) 매입 합계	23. 1.12.
8		수출 중소법인의 수입 부가가치세 납부유예세액	23. 1.14.
9		사업용 신용카드 매입	23. 1.12.
10		화물운전자복지카드 매입	23. 1. 1.
11		현금영수증 매입	23. 1. 1.
12		면세농산물등 매입가액(의제매입세액 공제신고서)	23. 1.14.
13	공제	직전기 재고매입세액	23. 1. 1.
14		재고납부세액	23. 1. 1.
15		신용카드 매출전표 발행세액공제 기공제세액	23. 1. 1.
16		일반과세자 예정신고 미환급세액	23. 1. 1.
17		일반과세자 예정고지세액	23. 1. 1.
18		간이과세자 예정부과세액	23. 1. 1.
19		간이과세자 예정신고세액	23. 1. 1.
20		철스크랩 등 매입자납부특례 기납부세액	23. 1.15.
21		재활용폐자원 의제매입세액공제 신고서상 계산서 금액	23. 1.14.
22		신용카드사를 통한 대리납부 관련 세액공제금액	23. 1.11.
23	기타	부동산임대공급가액명세서 직전기 임차인 명세	23. 1. 1.
24		수정신고.경정청구시 당초 부가세 신고서 및 부속서류	신고마감후
25		전자세금계산서 발급세액 공제액	23. 1.12.
26		전자계산서 매출 합계, 거래처별 명세	23. 1.12.
27		전자계산서 매입 합계, 거래처별 명세	23. 1.12.
28		국고입금 예정세액 정보(세무대리인)	23. 1.15.
29		전자세금계산서 지연 발급.수취.전송 관련 가산세 내역	23. 1.15.
30		음식·숙박업 직전기 사업장현황명세서	23. 1. 1.

출처 : 2022년 2기 부가가치세 확정신고 국세청 보도자료

위의 표는 부가가치세 신고 시 제공되고 있는 미리채움 항목이다. 부가가치세뿐만 아니라 소득세, 법인세에도 다양한 항목들이 미리 채워진 신고서를 납세자에게 제공하여 납세 편의를 도모하고 있다.

2. 납부 과정에서의 납세자 지원

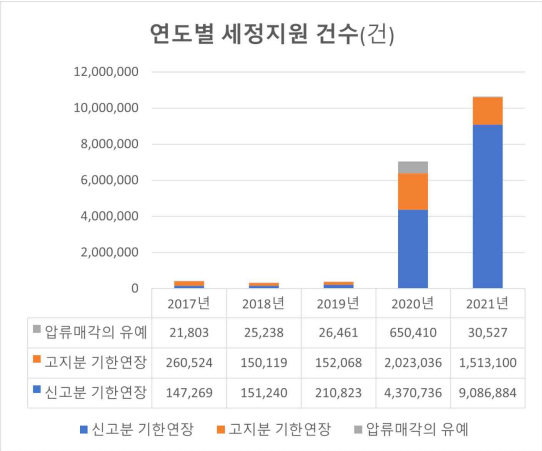
국세청은 인터넷 뱅킹, 간편결제 등 세금을 납부할 수 있는 채널을 다양화하는 데 많은 노력을 기울여 왔다. 이와 같은 납부 시스템의 개선은 세금 납부의 편리함을 높이는 데 목적이 있다. 이와는 다른 관점에서 당장 세금을 납부하기 어려운 납세자를 위한 지원 방안도 존재한다. 「국세징수법」은 재난, 도난, 경영상 현저한 손실 등이 있는 경우 납부기한을 연장하거나 납부 고지를 유예할 수 있도록 규정하고 있다.[56] 그 밖에도 재산의 압류나 압류 재산의 매각을 유예하면 체납자가 사업을 정상적으로 운영할 수 있게 될 것이라고 인정되는 경우에는 압류·매각을 유예할 수도 있다.[57]

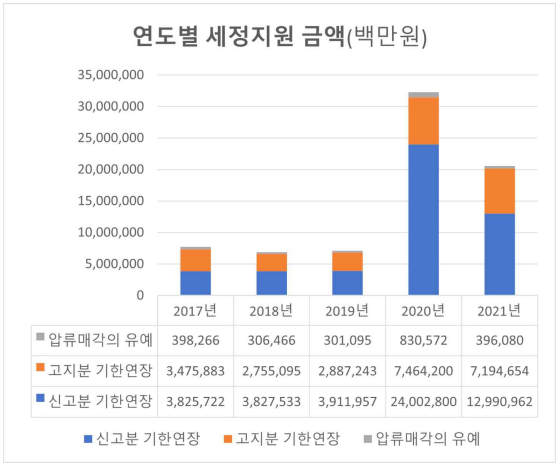
국세청은 납세자의 신청이 있는 경우 「국세징수법」에 규정된 조치들을 취하기도 하고, 자연재해가 발생하거나 경제적으로 어려움을 겪는 지역의 납세자를 대상으로 일률적으로 세정지원을 실시하기도 한다. 예컨대 2017년 경상북도 포항 일대에 발생한 지진으로 피해를 입은 납세자에게 납부기한 연장, 납부 고지 유예, 압류·매각 유예, 세무조사 연기 등의 세정지원을 실시한다고 발표한 바 있다.

---

56) 국세징수법 제13조(재난 등으로 인한 납부기한등의 연장) ① 관할 세무서장은 납세자가 다음 각 호의 어느 하나에 해당하는 사유로 국세를 납부기한 또는 독촉장에서 정하는 기한(이하 이 조, 제15조 및 제16조에서 “납부기한등”이라 한다)까지 납부할 수 없다고 인정되는 경우 대통령령으로 정하는 바에 따라 납부기한등을 연장(세액을 분할하여 납부하도록 하는 것을 포함한다. 이하 같다)할 수 있다.
1. 납세자가 재난 또는 도난으로 재산에 심한 손실을 입은 경우
2. 납세자가 경영하는 사업에 현저한 손실이 발생하거나 부도 또는 도산의 우려가 있는 경우
3. 납세자 또는 그 동거가족이 질병이나 중상해로 6개월 이상의 치료가 필요한 경우 또는 사망하여 상중(喪中)인 경우
4. 그 밖에 납세자가 국세를 납부기한등까지 납부하기 어렵다고 인정되는 경우로서 대통령령으로 정하는 경우
(중략)
제14조(납부고지의 유예) ① 관할 세무서장은 납세자가 제13조제1항 각 호의 어느 하나에 해당하는 사유로 국세를 납부할 수 없다고 인정되는 경우 대통령령으로 정하는 바에 따라 납부고지를 유예(세액을 분할하여 납부고지하는 것을 포함한다. 이하 같다)할 수 있다.
② 납세자는 제13조제1항 각 호의 사유로 납부고지의 유예를 받으려는 경우 대통령령으로 정하는 바에 따라 관할 세무서장에게 신청할 수 있다.
(이하 생략)

57) 국세징수법 제105조(압류ㆍ매각의 유예) ① 관할 세무서장은 체납자가 다음 각 호의 어느 하나에 해당하는 경우 체납자의 신청 또는 직권으로 그 체납액에 대하여 강제징수에 따른 재산의 압류 또는 압류재산의 매각을 대통령령으로 정하는 바에 따라 유예할 수 있다.
1. 국세청장이 성실납세자로 인정하는 기준에 해당하는 경우
2. 재산의 압류나 압류재산의 매각을 유예함으로써 체납자가 사업을 정상적으로 운영할 수 있게 되어 체납액의 징수가 가능하게 될 것이라고 관할 세무서장이 인정하는 경우
(이하 생략)





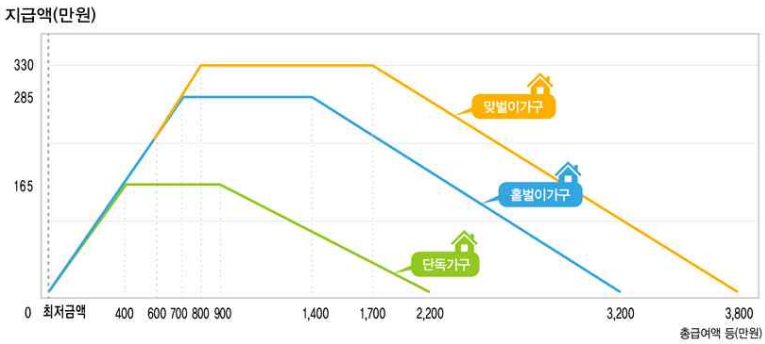
출처 : 국세통계포털(tasis.nts.go.kr)

위의 막대그래프는 최근 5년 간의 연도별 세정지원 실적이다. 건수로나 금액으로나 2020년과 2021년에 크게 증가한 것을 볼 수 있다. 코로나19의 영향으로 경영상의 어려움을 겪은 사업자가 많아 세정지원이 예년보다 많이 실시된 것으로 추측된다.

3. 장려금 지급을 통한 납세자 지원

국세청은 근로장려금, 자녀장려금을 지급하고 있다. 근로장려금은 소득이 적어 생활이 어려운 근로자, 사업자(전문직 제외) 가구에 대하여 가구원 구성과 근로소득, 사업소득 또는 종교인소득에 따라 산정된 장려금을 지급함으로써 근로를 장려하고 실질소득을 지원하는 제도이다.

< 근로장려금 구조 >



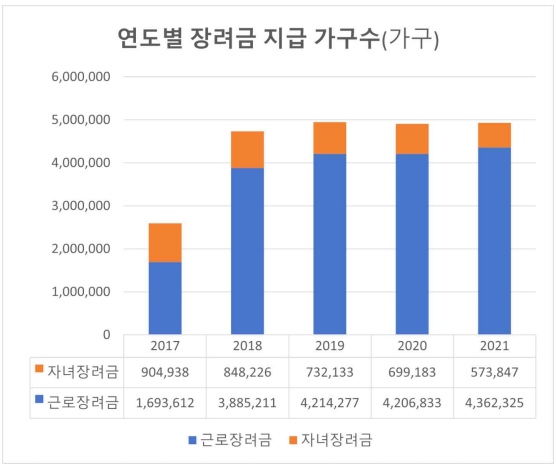
출처 : 국세청 홈페이지(www.nts.go.kr/nts/cm/cntnts/cntntsView.do?mi=2450&cntntsId=7781)

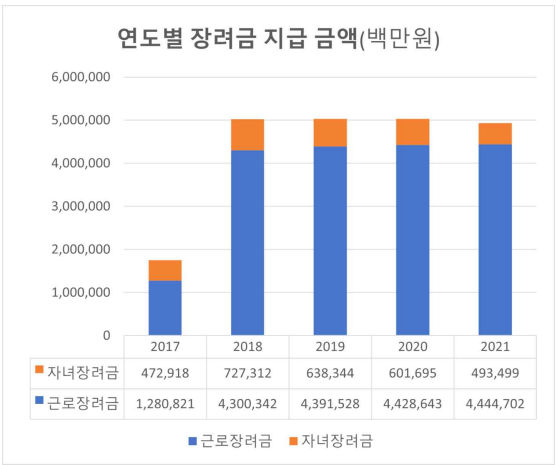
자녀장려금은 저소득 가구의 자녀 양육을 지원하기 위해 총소득(부부 합산) 4,000만원 미만이면서 부양 자녀(18세 미만)가 있는 경우 1인당 최대 80만원(최소 50만원)을 지급하는 제도로 총소득 기준을 제외한 나머지 수급 요건은 근로장려금과 동일하다.

< 자녀장려금 구조 >

국세청 홈페이지(https://www.nts.go.kr/nts/cm/cntnts/cntntsView.do?mi=2451&cntntsId=7782)

아래 그래프에서 보듯이 장려금 지급 가구수와 금액은 2018년도에 크게 증가하였다. 수급 요건이 완화되어 지급 대상이 증가하였고, 지급 금액도 인상되었기 때문이다. 두 제도로 연간 약 5백만 가구에 5조원 가량의 지원이 이루어지고 있다. 이제 근로·자녀장려금은 저소득 근로 계층을 지원하고 자녀 양육을 보조하는 대표적인 제도로 자리매김하였다.





출처 : 국세통계포털(tasis.nts.go.kr)

□ 데이터 과학의 적용 방향

국세청은 사후 검증에서 사전 성실신고 지원으로 국새행정의 패러다임을 전환한 이후 신고 지원 및 편의 제고에 많은 노력을 해 왔다. 그 결과 신고서의 많은 항목들이 자동으로 입력되어 제공되고 있고 홈택스를 통한 전자신고가 많이 간편해졌기 때문에, 과거 납세자가 일일이 자료를 찾아서 입력하고 검증을 받아야 했던 불편함은 많이 사라졌다. 그럼에도 불구하고 사전 성실신고 지원에 데이터 과학이 적용될 여지는 여전히 남아 있다. 시스템을 더욱 발전시켜서 납세자의 신고 오류를 사전에 탐지하고 적정한 값을 납세자에게 추천해 줄 수 있다면 신고의 정확도를 더욱 높일 수 있을 것이다. 이러한 관점에서 데이터 과학의 신고 지원 적용 방안을 모색해 보고자 한다.

한편, 신고 후 납부하는 과정에서도 국세청은 납세자의 부담을 줄이기 위해 많은 노력을 기울여 왔다. 태풍, 지진 등의 자연재해가 발생하면 분초를 다투어 세정지원 방안을 발표하기도 한다. 다만, 자연재해나 대형 사고처럼 모두가 인지할 수 있는 사건이 발생한 경우에는 지원 대상자를 선정하는 데 큰 어려움이 없지만, 어느 지방의 경제가 서서히 침체된다든지, 신도심 개발로 구도심 지역의 경기가 조금씩 하강할 경우에는 이를 민감하게 탐지하기가 여전히 쉽지 않다. 직원들은 대체로 2~3년 주기로 인사이동을 하기 때문에, 그 지역에서 거주하지 않는 한, 특정 지역 또는 업종의 부침을 체감하기가 쉽지 않다. 데이터 과학이 이와 같은 허점을 메워줄 수 있는 도구로 활용될 수 있다면 세정지원 대상을 정밀하게 선정하고 더 일찍 조치를 취할 수 있게 될 것이다.

장려금의 경우에는 신청이 있어야 지급되기 때문에, 빠짐없이 신청을 할 수 있도록 신청기간 전에 많은 홍보를 실시하고 있다. 많은 예산이 투입되고 있지만, 한도에 제한이 있는 만큼 효과적인 홍보 매체를 탐색하고 평가하는 것은 중요하다. 또한 특정한 매체를 사용할 때 홍보의 목표 그룹을 인지하는 것도 중요하다. 그래야만 차별화된 홍보도 가능해지는 것이고, 홍보 효과도 극대화되기 때문이다. 데이터 과학은 홍보 매체를 비교하여 어떤 집단에게, 어떤 매체가 효과적인지 판별하는 데 도움을 주는 수단으로 활용될 수 있을 것이다.

□ 행렬분해를 이용한 납세자 신고서 오류 탐지

1. 행렬분해의 개념

행렬분해(Matrix Factorization)는 영화, 드라마 등의 스트리밍 서비스를 제공하는 넷플릭스(Netflix)의 추천 시스템에 사용되면서 크게 인기를 얻었다. 추천 시스템은 넷플릭스뿐만 아니라 아마존(Amazon)과 같은 인터넷 쇼핑몰에서도 많이 사용되는데, 어떤 특성에 기반하여 추천을 하는지에 따라 컨텐츠 필터링(Contents Filtering)과 협업 필터링(Collaborative Filtering)으로 나누기도 한다. 넷플릭스가 채택한 협업 필터링은 사용자의 행동에 의존하는 방식인데, 사용자와 제품 사이의 상호의존적 관계를 분석하여 사용자-제품(user-item) 간의 관계를 찾고 이를 근거로 추천을 하는 방식이다. 사용자-제품 관계는 유사한 행동 패턴을 보이는 다른 사용자의 선택에 기반하여 추천을 해 주는 이웃 방식(Neighborhood Method)도 있지만, 넷플릭스는 사용자-제품 사이의 관계를 설명하는 숨은 요인을 찾아내어 이를 바탕으로 사용자에게 추천하는 잠재 요인 모델(Latent Factor Model)을 사용하였다. 이러한 잠재 요인을 찾는데 사용된 수학적 기법이 바로 행렬분해이다.

행렬분해는 추천 시스템에도 사용되지만, 이상점을 탐지하거나 이미지 또는 데이터를 압축하는 데도 사용된다. 행렬분해는 기본적으로 행렬로 표현된 데이터를 특정한 구조를 가진 다른 행렬의 곱으로 나타내는 것을 의미한다. 행렬을 분해하는 방법도 상당히 다양한데, 여기서는 많이 사용되는 특이값 분해(Singular Value Decomposition, SVD)를 설명해 본다. 임의의 m×n 차원의 행렬 A는 아래와 같이 분해할 수 있다.

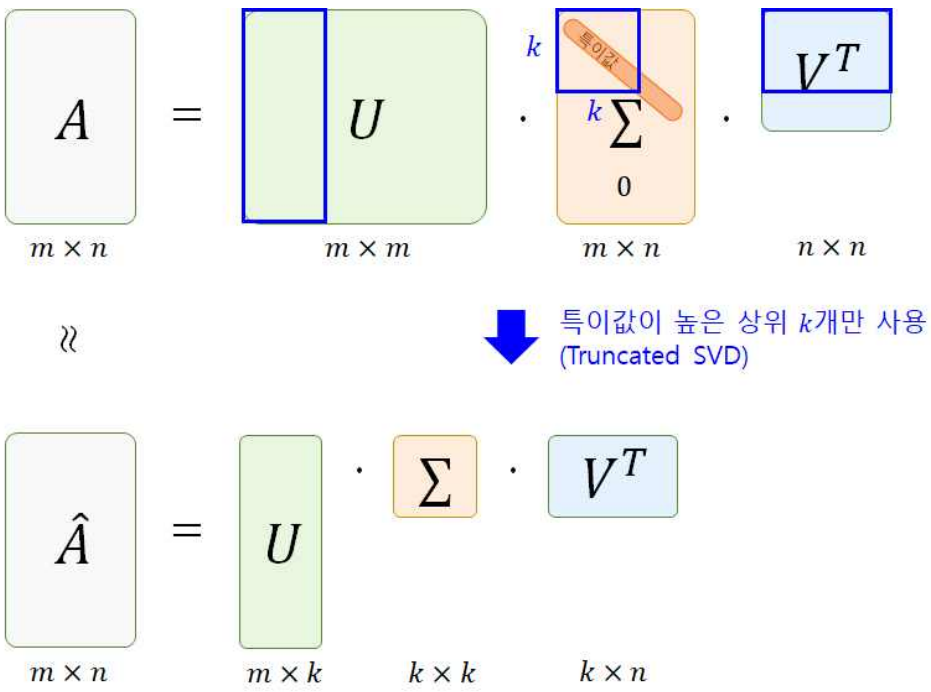
$$A = U\Sigma V^T$$

여기서 U는 m×m인 직교 행렬[58]이고 Σ는 m×n인 대각 행렬[59]이며, V는 $n \times n$ 직교 행렬이다. 수학적으로는 행렬의 고유값과 고유벡터를 이용하여 다소 복잡한 방식으로 구해야 하지만, 파이썬이나 R과 같은

58) X가 $XX^T = X^TX = I$일 때 직교 행렬(Othogonal Matrix)라고 한다.
59) 행과 열의 개수가 같은 정사각행렬이면서 대각선 이외의 값은 모두 0인 행렬을 의미한다.

프로그래밍 언어에서는 특이값 분해를 빠르고 간편하게 계산해 준다. 특이값 분해가 유용한 이유는 본래 행렬 A에서 유용한 정보만을 골라 행렬 A와 유사한 행렬을 쉽게 만들어 준다는 데 있다. Σ을 보면 숫자가 큰 순서대로 대각선 위에 배열되는데, 그 숫자의 크기가 본래 행렬 A에 대한 정보의 양을 의미한다. 따라서 숫자가 큰 일부의 값들만 포함한 축소된 행렬을 만들고, 이와 대응되는 U의 일부 열, V^T의 일부 열을 이용하면 새로운 행렬 A*을 만들 수 있다. 이렇게 만들어진 A*는 본래 행렬 A의 유용한 정보들로 만들어진 근사 행렬이 된다.

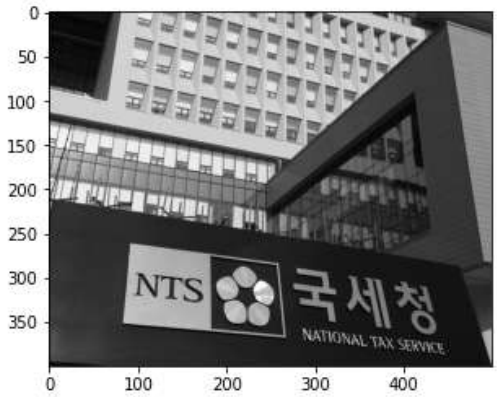
< 특이값 분해와 잘려진 특이값 분해 >



출처 : 책 읽는 성키, "[추천 시스템] Matrix Factorization(SGD)"
sungkee-book.tistory.com/12

특이값 분해를 이용한 이미지 압축을 예로 들어 설명해 보면 이해가 쉬울 것으로 생각한다. 아래의 왼쪽 이미지는 해상도가 400×500인 흑백 이미지이다. 즉 이 이미지 파일은 400×500 행렬로 이루어져 있고 행렬의 각 원소는 흑백의 정도를 나타내는 숫자들로 채워져 있다. 실제 이 이미지를 행렬로 나타낸 것은 오른쪽과 같다.

< 이미지 파일과 행렬로 표현된 이미지 >



$$= \begin{bmatrix} 7 & 8 & 6 & \cdots & 78 \\ 7 & 6 & 8 & \cdots & 75 \\ 9 & 5 & 6 & \cdots & 76 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 19 & 13 & 16 & \cdots & 31 \end{bmatrix}_{400 \times 500}$$

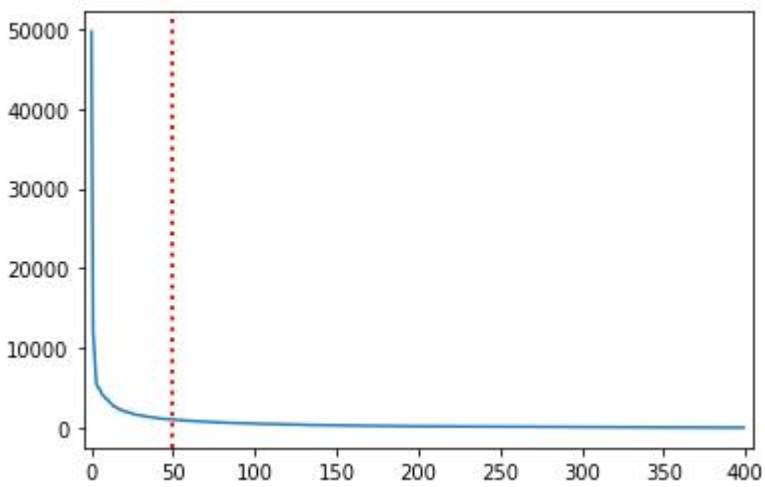
이 행렬을 A라 하면 아래와 같이 특이값 분해를 할 수 있다.

$$\underset{400\times500}{A} = \underset{400\times400}{U} \underset{400\times400}{\Sigma} \underset{400\times500}{V^T}$$

$$\begin{bmatrix} 7 & 8 & 6 & \cdots & 78 \\ 7 & 6 & 8 & \cdots & 75 \\ 9 & 5 & 6 & \cdots & 76 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 19 & 13 & 16 & \cdots & 31 \end{bmatrix} = \begin{bmatrix} -0.06 & 0.07 & \cdots & 0.01 \\ -0.06 & 0.07 & \cdots & -0.02 \\ -0.06 & 0.07 & \cdots & 0.02 \\ \vdots & \vdots & \ddots & \vdots \\ -0.01 & 0.01 & \cdots & -0.05 \end{bmatrix} \begin{bmatrix} 49706 & 0 & \cdots & 0 \\ 0 & 12166 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 7 \end{bmatrix} \begin{bmatrix} -0.04 & -0.04 & \cdots & -0.05 \\ -0.09 & -0.10 & \cdots & -0.03 \\ -0.00 & -0.01 & \cdots & -0.03 \\ \vdots & \vdots & \ddots & \vdots \\ -0.00 & -0.05 & \cdots & 0.03 \end{bmatrix}$$

이미지에 대한 정보의 중요도를 담고 있는 Σ 행렬의 대각선 위의 값을 그래프로 나타내보면 아래와 같다.

< Σ 행렬의 대각선 위의 값의 분포 >



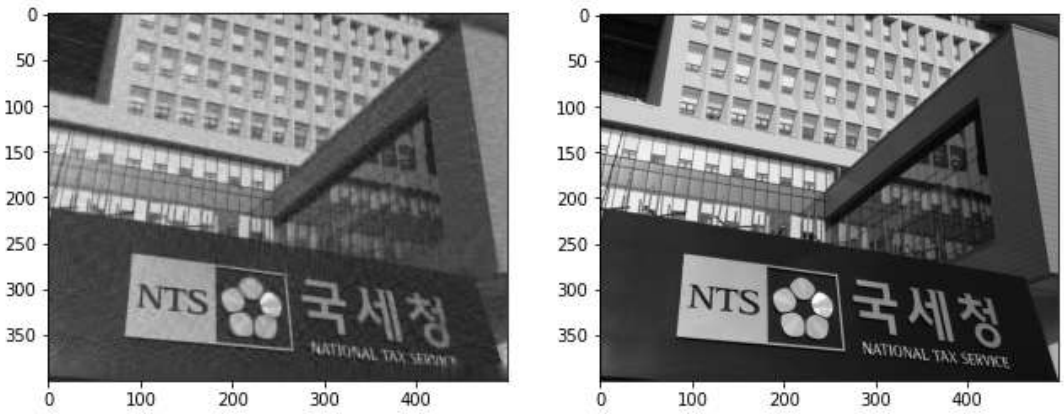
Σ 행렬의 대각성 위의 값 중에서 처음 50개 정도의 값은 비교적 크지만 50개 이후의 값들은 매우 작으므로 처음 50개만 선택해 본다. 그리고 이와 대응되는 U의 일부 열과, V^T의 일부 행을 이용하여 새로운 행렬 A ' 을 아래와 같이 계산해 본다.

$$\underset{400\times50}{U'}\ \underset{50\times50}{\Sigma'}\ \underset{50\times500}{V'^{T}} = \underset{400\times500}{A'}$$

$$\begin{bmatrix} -0.06 & 0.07 & \cdots & 0.13 \\ -0.06 & 0.07 & \cdots & 0.04 \\ -0.06 & 0.07 & \cdots & -0.01 \\ \vdots & \vdots & \ddots & \vdots \\ -0.01 & 0.01 & \cdots & -0.00 \end{bmatrix} \begin{bmatrix} 49706 & 0 & \cdots & 0 \\ 0 & 12166 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1023 \end{bmatrix} \begin{bmatrix} -0.04 & -0.04 & \cdots & -0.05 \\ -0.09 & -0.10 & \cdots & -0.03 \\ -0.00 & -0.01 & \cdots & -0.03 \\ \vdots & \vdots & \ddots & \vdots \\ 0.13 & 0.22 & \cdots & 0.01 \end{bmatrix} = \begin{bmatrix} 7 & 8 & 6 & \cdots & 78 \\ 7 & 6 & 8 & \cdots & 75 \\ 9 & 5 & 6 & \cdots & 76 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 19 & 13 & 16 & \cdots & 31 \end{bmatrix}$$

이렇게 만들어진 행렬 A ' 을 이미지로 다시 변환하면 아래의 왼쪽과 같다. 원본인 아래의 오른쪽 이미지와 비교하면 화질에 다소 손실이 있음이 보이지만, 어떤 이미지인지 인식하는 데는 무리가 없다.

< 특이값 분해를 이용해 압축된 이미지와 원본 이미지 비교 >



압축된 이미지 원본 이미지

이와 같은 유사한 이미지를 만들어 내는데 얼마나 많은 데이터가 필요했을지 계산해 보자. 원본 이미지는 400×500에 해당하는 200,000개의 숫자를 가지고 있다. 반면, 압축된 이미지는 400×50 + 50 + 50×500 = 45,050개의 숫자만 가지고 있다. 즉 원본 데이터의 22.5%를 가지고도 유사한 이미지를 표현할 수 있게 된 것이다.[60)]

2. 두 벡터 사이의 유사도 측정

벡터는 수학에서 배열(array)로 표현되는 여러 숫자의 묶음을 의미한다. 행렬도 차원이 동일한 여러 벡터들이 적층된 것으로 볼 수 있다. 벡터의 유사도를 측정하는 여러 방법이 있는데, 기계학습 등 데이터 과학에서 유용하게 사용된다. 몇 가지만 간단히 소개해 본다.

먼저 유클리디안 거리(Euclidean Distance)는 좌표평면에서 두 벡터가 의미하는 두 점 사이의 최단 거리를 의미한다. 두 벡터가 $(x_1, x_2, ..., x_n)$, $(y_1, y_2, ..., y_n)$으로 주어질 때 유클리디안 거리는 아래와 같이 구한다.

$$D = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

유클리디안 거리 외에도 두 점 간의 거리를 측정하는 방법으로 맨하탄 거리(Manhattan Distance), 체비셰프 거리(Chebyshev Distance)도 있으나, 측정하는 방법이 상이할 뿐, 두 점 사이의 거리를 측정한다는 점에서는 유클리디안 거리와 유사한 유형으로 분류할 수 있다.

---

60) 이미지 압축 예제의 파이썬 코드는 아래와 같다.

```
import cv2
import numpy as np
from scipy.linalg import svd
import matplotlib.pyplot as plt

img = cv2.imread("C:/NTS.jpg")
img_gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
plt.imshow(img_gray, cmap='gray')
plt.show()

U1, s1, Vt1 = svd(img_gray)

plt.xlim(-5, 405)
plt.plot(list(range(400)), s1)
plt.axvline(x = 50, color = 'red', ls=':', lw=2)
plt.show()

U2 = U1[:, :50]
s2 = s1[:50] * np.identity(50)
Vt2 = Vt1[:50, :]
img_compressed = U2 @ s2 @ Vt2

plt.imshow(img_compressed, cmap='gray')
plt.show()
```

그 밖에 코사인 유사도(Cosine Similarity)가 있다. 두 벡터는 좌표평면에서 방향과 크기를 갖는 화살표로 표시될 수도 있는데, 코사인 유사도는 두 화살표 사이의 각도를 통해 유사도를 측정하는 방법이다. 코사인 유사도는 아래와 같이 계산된다.

$$D = \frac{x \cdot y}{\| x \| \ \| y \|}$$

여기서 ·은 두 벡터의 내적(inner product)를 의미하며 ‖ ‖는 벡터의 norm을 의미한다. 코사인 유사도에 따르면 두 점 사이의 거리가 멀어도 두 벡터가 이루는 각도가 좁다면 유사도가 높다고 평가한다. 거리에 기반한 유사도 측정 방법과 각도에 기반한 코사인 유사도 사이의 관계를 그림으로 나타내면 아래와 같다. 기준 벡터( )와 유사한 값으로 측정되는 벡터들이 주황색 점선으로 표시되어 있다.

< 거리 기반 유사도 및 코사인 유사도 >

출처 : anweh. 이것저것 기록. “[Python] 데이터의 유사성 측정 방법” (https://anweh.tistory.com/54)

이와 같은 유사도 측정 방법들 사이에 우열 관계가 있는 것은 아니며, 필요에 따라 적절한 방법을 선택하는 것이 바람직하다.

3. 유사도 측정과 행렬분해를 이용한 입력 오류 탐지

납세자가 입력하는 신고서의 내용을 행렬로 만든다고 생각해 보자. 행은 각각의 납세자가 될 것이며, 열은 납세자가 입력하는 총수입금액, 소득금액 등 각종 신고 항목이 될 것이다. 이제 납세자가 어떤 항목을 잘못 기재한 채로 신고서를 제출하려 한다고 해 보자. 우선 다른 납세자들의 신고 내용과 비교하여 유사도가 적은 경우 납세자에게 경고를 하고 어떤 항목에 문제가 있는지 알려줌으로써 납세자가 어떤 항목을 과대하게 또는 과소하게 작성했는지 탐지할 수 있다. 그리고 행렬분해를 이용하면 잘못 기재된 값의 근사값도 알려줄 수 있다. 즉 유사도 측정과 행렬분해를 이용하면 신고 과정에서 납세자에게 어떤 항목에 오류가 있는지 어떤 값을 입력하면 적절할지 제안할 수 있다는 뜻이다.

물론 현재도 여러 가지 방법으로 신고서 입력 오류를 납세자에게 알려주고 있다. 하지만 이는 납세자의 계산이 부정확하다거나 반드시 입력해야 할 항목을 입력하지 않았을 때 알려주는 것이지, 납세자가 입력한 값이 정확하지 않을 수 있음을 알려주는 것은 아니다. 예컨대 납세자가 과세 매출을 10,000원이라고 기재해 놓고도 부가가치세 매출세액을 1,000원(과세 매출의 10%)이 아닌 값으로 입력하면 경고를 띄우는 일은 지금도 가능하다. 하지만 애초에 납세자가 10,000원으로 입력한 값에 잘못이 있지 않은지 확인해 줄 수는 없다. 하지만 유사도 측정과 행렬분해를 이용하면 10,000으로 입력한 과세 매출이 다른 납세자들의 신고 내용과 비교해 보았을 때 타당한 값인지, 아닌지 확인해 줄 수 있다. 이와 같은 시스템이 마련된다면 신고 과정에서 발생하는 오류를 상당히 줄일 수 있을 것으로 기대된다. 신고 오류가 줄어들면 신고 후 과세관청이 오류를 확인하고 정정하는 과정에서 발생하는 납세자의 불편과 행정력의 낭비도 줄일 수 있을 것이다.[61]

시뮬레이션을 통해 유사도 측정과 행렬분해가 어떻게 신고 오류를 탐지하는지 설명해 본다. 특정 시기에 납세자들이 신고한 자료의 내역이 다음과 같은 행렬로 표현된다고 해 보자. 현재 신고서를 작성하고 있는

61) 유사도 측정과 행렬분해를 역으로 이용하면 불성실 사업자를 탐지하는 데에도 활용할 수 있다. 하지만 불성실 사업자 탐지와 관련해서는 이미 여러 방안을 언급하였으므로 여기서는 더 설명하지 않는다.

납세자는 이 행렬의 가장 마지막 행으로 표현되고 있다. 마지막 행을 보면 매출세액이 4로 빨갛게 기재되어 있다. 이는 납세자8이 본래 45를 기재하려 했으나 오기한 것이다.

< **납세자별 부가가치세 신고 내용**(예시) >

	매출	매출세액	매입	매입세액
납세자1	700	60	700	70
납세자2	100	10	80	8
납세자3	100	10	100	10
납세자4	300	25	170	20
납세자5	30	3	20	2
납세자6	200	19	180	20
납세자7	400	44	300	30
납세자8	450	4	400	40

이처럼 납세자가 신고서에 실수로 지나치게 작은 또는 큰 값을 입력했거나, 기재하지 않은 항목이 있는 경우 이를 탐지하는 방법이 필요하다. 즉 납세자8이 신고한 항목들이 이전에 신고한 납세자들의 것과 유사한지 확인할 수 있어야 한다. 이를 위해서는 코사인 유사도를 사용할 수 있다. 사업자의 규모는 다양하지만 매출과 매출세액, 매입과 매입세액 등의 관계는 대체로 일정하기 때문이다. 시뮬레이션은 간단한 임의의 자료를 이용하므로, 전처리가 필요 없지만, 복잡한 현실에서는 앞에서 설명한 군집화 기법을 미리 사용하면 계산된 코사인 유사도를 평가할 때 더 효과적일 것으로 생각된다.

코사인 유사도는 크기가 큰 값에 많은 영향을 받으므로 동일한 스케일로 맞추기 위해 각 열의 평균과 표준편차를 이용하여 데이터를 정규화하였다. 정규화한 데이터로 코사인 유사도를 구한 결과가 아래와 같다.

< **신고 항목 간의 코사인 유사도** >

	매출	매출세액	매입	매입세액
매출	1.0000	0.7755	0.9762	0.9812
매출세액	0.7755	1.0000	0.7484	0.7562
매입	0.9762	0.7484	1.0000	0.9986
매입세액	0.9812	0.7562	0.9986	1.0000

매출세액의 유사도가 낮게 나타나고 있다는 점을 확인할 수 있다. 즉 납세자8의 매출세액 입력에 문제가 있다는 뜻이다. 이제 납세자8에게 입력된 매출세액에 문제가 있다는 경고를 보낼 수 있다.

이제 납세자8에게 어떤 값이 입력되어야 하는지도 알려줘 보자. 문제를 일으키는 되는 입력 항목을 제거하고 행렬분해를 이용하여 근사치를 구해 본다. 데이터가 없는 항목이 있는 경우 우선 그 항목이 가져야 할 값을 1차적으로 추정하는 과정이 필요하다. 여기에도 다양한 수학적 기법이 존재하지만, 자세히 설명하지 않고 많이 활용되는 확률적 경사 하강법(Stochastic Gradient Descent, SGD)이라는 방법을 사용해 본다. 이렇게 누락이 채워진 행렬을 분해하고 주요 정보만 일부 추출하면[62] 근사 행렬을 얻을 수 있다. 이러한 근사 행렬은 정규화된 상태로 얻어지므로, 평균과 표준편차를 이용하여 복원할 수 있다. 그러한 과정을 통해 얻은 결과가 아래와 같다.

< 정규화된 납세자별 부가가치세 근사 신고 내용 >

	매출	매출세액	매입	매입세액
납세자1	1.9834	1.9711	2.1042	2.0969
납세자2	-0.8973	-0.7782	-0.7631	-0.7811
납세자3	-0.8224	-0.6907	-0.7469	-0.7638
납세자4	-0.1472	-0.1267	-0.1112	-0.1149
납세자5	-1.1615	-0.9869	-1.1575	-1.1755
납세자6	-0.3551	-0.3398	-0.2644	-0.2697
납세자7	0.5172	0.7410	0.4095	0.3883
납세자8	0.7689	0.9040	0.7410	0.7259

⇒ 0.9040 × 19.1002 + 24.4286 ≈ 42 (근사된 매출세액)

근사된 매출세액으로 42가 제안되었다. 납세자 매출액이 450임을 감안하면 크게 어긋나는 숫자는 아니다. 매출세액이 아닌 매출 또는 다른 항목에 이상이 발생하는 경우에도(예컨대 매출세액으로 4500이 입력된 경우) 이와 같은 알고리즘으로 잘못된 입력을 탐지하고 적정한 숫자를 제안할 수 있다.[63]

62) 이 시뮬레이션에서는 대각행렬의 대각선 위의 값 중 3개만을 사용하였다.
63) 이 시뮬레이션의 파이썬 코드는 매우 길어 보고서 말미의 '시뮬레이션 코드 1.'에 수록하였다.

4. 고려해야 할 사항

코사인 유사도 측정과 행렬분해가 제대로 동작하기 위해서는 사전에 적절한 값이 시스템에 입력되어 있어야 한다. 이전의 값들과 비교하여 동떨어진 값인지 확인하고, 잘못 입력된 값이 가져야 할 적정한 값을 이전에 입력된 데이터로부터 추정하는 것이 알고리즘의 동작 방식이기 때문이다. 따라서 다양한 업종, 규모, 지역 등에 걸쳐 정확하게 작성된 신고서가 미리 입력되어야 한다. 반드시 당해 연도의 신고서일 필요는 없으며, 가상으로 만들어진 신고서라도 무방하지만 정확하게 작성되어야 한다. 일반적으로 알고리즘이 좋은 성능을 내기 위해서는 좋은 학습 데이터가 필요한데, 이 알고리즘도 예외는 아니다.

앞서 소개한 행렬분해가 누락된 값의 근사치를 예측할 수 있는 유일한 방법은 아니다. 회귀분석 또는 신경망과 같은 다른 방법을 이용할 수도 있다. 다만, 구글 검색엔진 등이 검색어를 입력하면 바로 예상되는 다음 검색어를 자동으로 보여주듯이, 납세자가 홈택스의 신고 항목에 숫자를 입력했을 때 바로 사용자에게 경고를 할 수 있으려면 정확한 알고리즘 보다는 계산 속도가 빠른 알고리즘이 더욱 적합할 것으로 보인다.

이와 같은 알고리즘은 다양한 방식으로 응용될 수 있다. 국세청 내부에서 세무조사 대상 사업자를 분석할 때 재무제표 등에 알려지지 않은 정보를 추정해 낼 수 있을 것으로 기대되며, 불성실 사업자를 탐지하는 데에도 활용될 수 있을 것으로 보인다. 더 나아가 당장 실현되기는 어렵겠지만, 알고리즘이 정교화되고 근거가 되는 세법이 개정된다면, 장부가 기장되어 있지 않은 사업자의 소득을 추계해야 할 때도 이러한 알고리즘을 적용할 수 있을 것으로 기대한다. 구체적인 장부나 증빙이 아닌 컴퓨터 알고리즘이 세금을 계산하는 근거로 사용될 수 있을지는 앞으로의 세무 환경의 변화에 따라 달라질 것이다.

□ 생존분석을 이용한 업황 분석

1. 생존분석의 개념

생존분석(Survival Analysis)은 어떠한 사건이 발생하기까지 걸리는 시간에 대해 분석하는 통계학의 한 갈래이다. 일반적으로 의료, 제약 등의 분야에서 새로운 치료법 또는 신약이 환자의 생존율을 높이는 데 효과가 있는지 분석하기 위해 사용한다. 하지만 생존분석은 다양한 가정을 필요로 하지 않기 때문에 의료, 제약 분야뿐만 아니라 광범위한 분야에서 응용되고 있다. 예를 들어 어린 나이에 범죄를 저지를 자와 늦은 나이에 범죄를 저지른 자가 교정시설을 나온 이후 재범을 일으킬 때까지 걸린 시간을 생존분석을 이용하여 분석하기도 한다.[64] 나아가 기업의 지식재산 등 내부 정보의 유출 위험을 보다 정확하게 예측하기 위하여 생존분석을 활용하기도 한다.[65]

아래에서 생존분석을 간단히 설명해 본다. 보통 우리가 어떤 대상을 계속 모니터링하면서 사건이 발생할 때까지의 시간을 측정하다 보면 여러 현실적인 문제에 당면하게 된다. 주어진 시간적 또는 비용적 한계 때문에 사건이 일어날 때까지 대상을 추적하지 못할 수도 있다. 또는 추적에 실패하거나, 추적하는 대상이 실험에서 빠지겠다고 할 수도 있다. 이와 같은 불완전한 데이터를 가지고도 생존분석은 의미 있는 결과를 제공해 준다는 데 의의가 있다.

이처럼 여러 가지 이유로 사건이 관측되지 못한 데이터를 절단된 자료(Censored Data)라고 한다. 카플란과 마이어는 1958년에 ‘Journal of the American Statistical’에서 이와 같은 절단된 자료로부터 생존함수(Survival Function)[66]를 추정할 수 있는 비모수적인 방법을 제안했다.

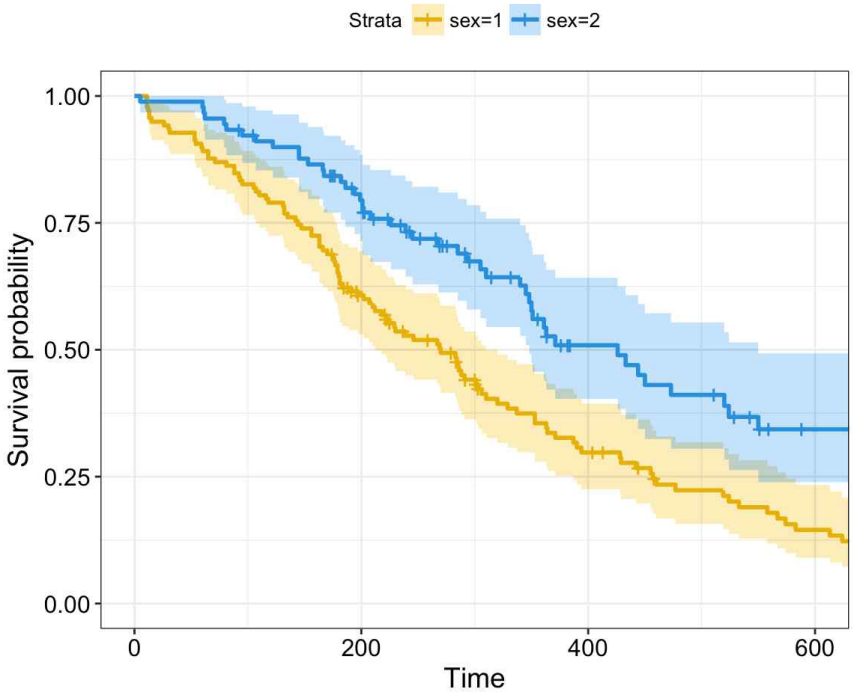
---

64) Benda, Brent B. 2003. “Survival Analysis of Criminal Recidivism of Boot Camp Graduates Using Elements from General and Developmental Explanatory Models.” International Journal of Offender Therapy & Comparative Criminology, February 1. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,sso&db=edsgea&AN=edsgcl.100132070&site=eds-live&scope=site.

65) Alhajjar, Elie, and Taylor Bradley. 2022. “Survival Analysis for Insider Threat; Detecting Insider Threat Incidents Using Survival Analysis Techniques.” Computational & Mathematical Organization Theory 28 (4): 335. doi:10.1007/s10588-021-09341-0.

66) 생존함수는 시간 t의 함수이며, t 시점을 넘어 대상이 생존할 확률을 의미한다.

< 추정된 생존함수의 예시 >



출처 : STHDA, “Survival Analysis Basics” http://www.sthda.com/english/wiki/survival-analysis-basics

위의 그림은 추정된 생존함수의 예시이다. 두 그룹 사이에 생존함수에 차이가 있는지를 확인할 수 있는 검정 방법이 Nathan Mantel에 의해 제안[67]되었는데, 이를 log-rank test 또는 Mantel-Cox test라고 부른다. 이는 범주형 자료 분석에서 사용되는 Cochran-Mantel-Haenszel test를 시간으로 계층화하여 적용한 것으로 볼 수 있다.

생존함수는 위험함수(Cumulative Hazard Funcion)[68]와 일정한 관계를 갖고 있는데, 위험함수는 보통 Nelson-Aalen 추정량을 사용하여 추정한다. 성별, 인종, 체중, 투약된 약물의 종류 등에 따라 위험함수는 달라질 수 있는데, 이와 같이 요인들을 이용하여 위험함수를 설명하는 모형을 Cox-proportional hazards model(Cox-PH model)이라고 부른다. 이를 구체적으로 적어 보면 아래와 같다.

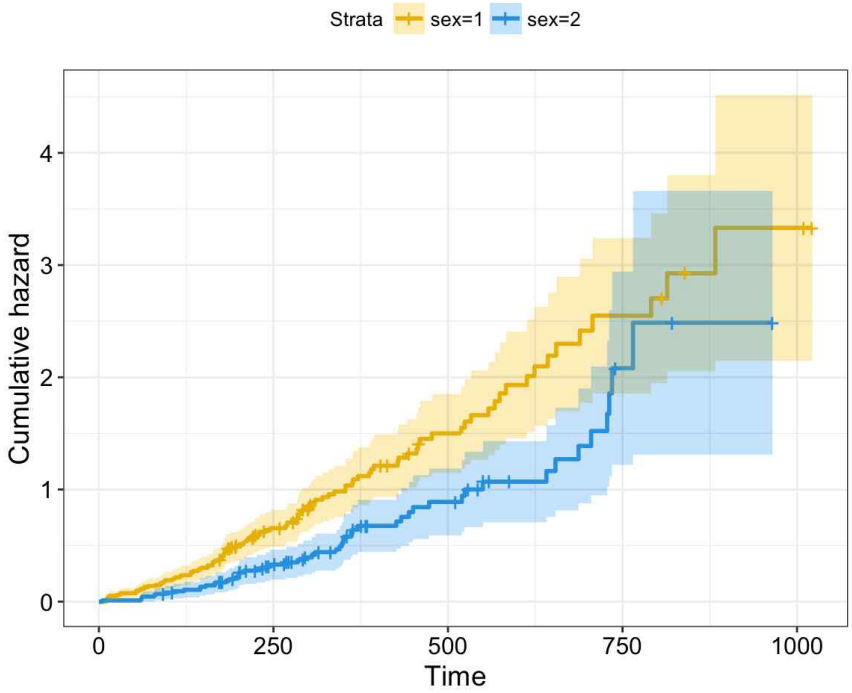
67) Mantel, Nathan. 1966. "Evaluation of survival data and two new rank order statistics arising in its consideration". Cancer Chemotherapy Reports. 50 (3): 163-70. PMID 5910392.

68) 위험함수는 시간 t의 함수이며, t시점까지 생존한 사람이 t시점 직후 사망할 확률을 의미한다.

$$\lambda(t : \vec{Z}) = \lambda_0(t)\exp(\beta\vec{Z}) \Leftrightarrow \log\left(\frac{\lambda(t : \vec{Z})}{\lambda_0(t)}\right) = \beta Z$$

여기서 $\lambda(t : \vec{Z})$는 위험함수이다. 시간의 함수이며, 벡터 $\vec{Z}$로 주어지는 요인들에 의해 영향을 받는다. βZ는 $\beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_p Z_p$의 벡터 표현이다. $Z_1, Z_2, ..., Z_p$는 위험함수에 영향을 미칠 수 있는 각종 요인이며, $\beta_1, \beta_2, ..., \beta_p$는 그 영향의 정도를 나타내는 계수이다. $\lambda_0(t)$는 기저 위험함수(baseline hazard function)라고 부른다. Cox-PH model에서 $\lambda_0(t)$는 시간의 함수라는 가정 외에는 아무런 가정을 부여하지 않는다.[69] 모든 $Z_1, Z_2, ..., Z_p$가 0일 때의 위험률($\lambda(t : Z = 0) = \lambda_0(t)$)은 기저 위험함수와 같아지는데, $Z_1, Z_2, ..., Z_p$가 0인 경우가 특별한 의미를 갖는 데이터일 수도 있지만, 특별한 의미를 찾을 수 없는 데이터일 수도 있다.

< 추정된 위험함수의 예시 >



출처 : STHDA, "Survival Analysis Basics" http://www.sthda.com/english/wiki/survival-analysis-basics

69) 위험함수에 대한 독립변수의 영향에 관해서는 모수적인 가정(Parametric Assumption)을 하지만, 위험함수 λ(t) 자체의 특성에 대해서는 아무 가정을 하지 않기 때문에 Cox-PH model을 Semi-parametric model로 여기기도 한다.

Cox-PH model을 적합시키면 각각의 요인이 위험함수에 미치는 영향을 추정된 계수를 통해 확인할 수 있다. 예를 들어 위의 그래프는 성별에 따른 위험함수의 그래프이다. 양자가 차이가 있는지를 확인하려면 Cox-PH 모델에서 성별의 추정된 계수를 확인하면 된다. 이러한 계수는 지수함수로 변환되어야 위험률로써 의미를 갖기 때문에 유의미한 해석을 위해서는 지수변환이 수반되어야 한다.

2. 생존분석을 이용한 불황 지역·업종 탐지

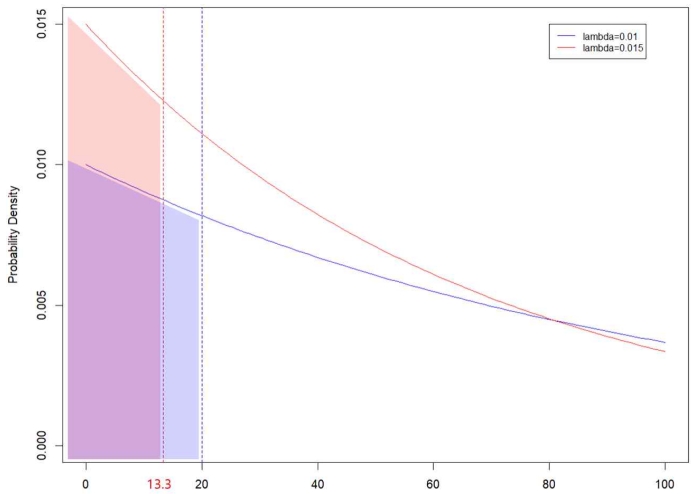
특정한 지역 또는 업종이 불황인지 확인할 수 있다면 해당 지역 또는 업종의 납세자를 대상으로 세정지원 안내를 해 줄 수 있을 것이다. 지금까지 이를 통계적으로 확인하기 위해서는 한국은행, 통계청 등에서 발표하는 각종 경제 지표를 이용할 수밖에 없었다. 하지만 이와 같은 자료는 실시간으로 구할 수 없으며, 조사 시점과 발표 시점에 큰 차이가 있어 시의적절한 지원을 위해 사용하기에는 적절하지 않다. 또한 각 기관이 조사하는 단위가 국세청이 관리하는 단위와 달라 자료가 있어도 활용 가치가 낮다. 예컨대 한국은행은 기업경기실사지수(BSI)를 전국 단위로 발표하고 있으며, 각 지역 상공회의소가 기초지방자치단체별로 기업경기실사지수를 발표하고 있지만, 여전히 세부적인 지역(동 단위), 업종별로 업황을 알기는 어려운 실정이다.

국세청이 보유한 여러 자료도 업황의 영향을 받으므로 외부기관에서 제공하는 자료보다 국세청이 보유한 자료를 적극적으로 이용한다면. 불황을 겪고 있는 지역 또는 업종을 실시간으로 탐지해 낼 수 있을 것이다. 특히 사업자등록 및 폐업 신고 자료를 이용하여 사업 영위 기간에 대해 생존분석을 실시한다면 어떤 지역 또는 업종의 생존률이 낮은지 탐지할 수 있을 것으로 기대된다. 예컨대 예년에 비해 생존률이 낮아진 지역 또는 업종이 있다면 업황이 좋지 않은 것으로 볼 수 있을 것이다. 물론 법인세, 소득세 등의 세금 신고 자료도 업황을 반영하고 있으므로 이를 활용할 수도 있지만, 1년 동안의 실적 자료를 다음 해에 제출하는 것이기 때문에, 이를 이용하여 실시간으로 업황을 탐지하는 데는 근본적으로 한계가 있을 수밖에 없다.

간단한 시뮬레이션을 통해 생존분석이 어떻게 불황을 탐지할 수 있는지 설명해 본다. 예를 들어 OO시 OO동에서 2018년에 개업한 사업자들을 선별하고 이들이 2021년말까지의 기간 중에 언제, 몇 명이 폐업했는지 확인한다고 해 보자. 그리고 같은 지역에서 2019년에 개업한 사업자들을 선별하여 2022년말까지의 기간 중에 언제, 몇 명이 폐업했는지도 확인한다고 해 보자.

2018, 2019년의 개업자 수는 300~500개 중에 임의로 추출되었다. 추출된 결과에 따르면 2018년에는 353개의 사업자가 개업하였고, 2019년에는 374개 사업자가 개업하였다. 개업 시점(open_date)은 1월부터 12월 사이에서 월 단위로 임의로 선택되었다. 사업 영위 기간(period)도 월 단위로 추출되었는데, λ=0.01인 지수분포에서 추출되었다. 개업 시점에 사업 영위 기간을 더하면 폐업 시점(close_date)이 되는데, 폐업 시점이 2022년에 있는 사업자는 당초 임의 추출된 사업 영위 기간과 누적 확률이 동일한 사업 영위 기간을 λ=0.015인 지수분포로부터 계산하여 당초 사업 영위 기간 대신 이용하였다. (아래 그림에서 두 지수분포의 빨간색과 파란색 영역의 넓이가 같아지도록 함으로써 λ=0.01에서 추출된 20이라는 숫자를 λ=0.015 하의 13.3으로 변환하였다는 의미이다.)

< 시뮬레이션 데이터에 불황 반영 >

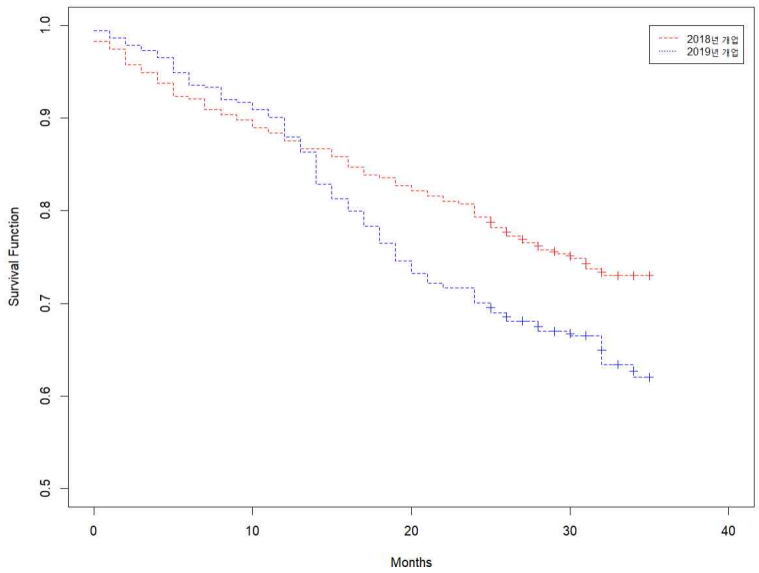


지수분포는 모수의 역수가 평균이므로, 모수가 클수록 더 짧은 사업 영위 기간을 가지는 경향이 있음을 의미한다. 즉, 이와 같은 복잡한 조치는

예년에 비해 2022년에 유난히 불황이 심해졌음을 반영하기 위한 것이다. 관찰 기간이 정해져 있기 때문에, 폐업 시점(close_date)에는 절단(censoring)이 발생하게 된다. 2018년에 개업한 사업자는 폐업 시점이 2021년 이후일 경우 절단하였으며, 2019년에 개업한 사업자는 폐업 시점이 2022년 이후일 경우 절단하였다. 따라서 시뮬레이션에 실제 이용할 수 있는 자료는 관측된 사업 영위 기간(period_obs)과 절단 여부(censoring)가 된다.

2018, 2019년에 개업한 사업자들의 생존함수를 카플란과 마이어스가 제안한 방법으로 추정하여 그래프로 나타내면 앞의 그래프와 같다. 2019년 개업한 사업자의 생존함수가 2018년 개업한 사업자보다 대체로 아래 쪽에 위치한 것을 확인할 수 있다.

< 추정된 생존함수 >



이를 검정하려면 앞서 언급한 log-rank test를 실시해 보면 된다. 하지만 log-rank test는 두 집단 간에 시간에 걸친 위험비(Hazard Ratio)가 일정하다는 proportional hazards이 충족될 때 검정력이 가장 높은데 이때 두 생존함수는 평행한 모습을 보이게 된다. 하지만 위 그래프의 생존함수는 평행하지 않다. 또한 우리는 최근인 2022년 동안에 생존함수에 차이가 발생했는지에 관심이 있다. 따라서 생존함수의 오른쪽 부분에 가중하여 log-rank test를 실시하는 것이 바람직할 것으로 보인다. 생존

함수의 오른쪽 부분에 가중하기 위해서는 Fleming-Harrington의 통계량을 사용하되 ρ=0, γ=1로 설정하면 될 것이다.[70] 이와 같은 설정으로 검정 통계량을 계산하면 아래와 같다. 검정 통계량 값은 3.7250이며, Z 값은 3.0195이다. P-value를 계산하면 약 0.001이다. 따라서 유의수준 0.05 하에서 두 그룹의 생존함수가 동일하다는 가설을 기각할 수 있다. 다시 말하자면, 두 그룹의 생존함수가 동일하지 않다는 유의미한 증거가 있음을 의미하는 것이다.[71]

```
wlrt(Surv(period_obs, censoring) ~ group,
     method="fh", data=dt, rho = 0, gamma = 1)
       u     v_u        z trt_group
3.725014 1.52192 3.019478         2
```

시뮬레이션은 편의를 위해 연 단위로 데이터를 수집했지만 반드시 연 단위로 데이터가 수집되어야 하는 것은 아니다. 데이터만 충분하다면 월별로 비교할 수도 있고 주별로 비교할 수도 있다. 또한 시뮬레이션은 두 그룹을 기준으로 개업 일자를 기준으로 나누었지만, 동일한 기간에 개업한 사람들을 지역별, 업종별, 규모별로 비교할 수도 있다. 하지만 매번 log-rank test를 할 필요 없이 조금 더 손쉽게 어려움을 겪는 집단을 찾아낼 수 있는 방법이 있다. 바로 Cox-proportional hazards model(Cox-PH model)을 이용하는 것이다.

조금 더 복잡한 데이터를 생성해 본다. 지역(area)은 2개의 범주로 분류하였고, 업종(type)은 3개의 범주로 분류하였다. 6개의 범주 각각에 대해 0.01부터 0.1 사이의 임의의 값을 추출하여 지수분포의 모수로 삼고 200 ~ 300 사이의 임의의 정수에 해당하는 만큼의 데이터를 생성하였다. 각 범주에 해당하는 지수분포의 모수와 이로부터 생성된 데이터의 개수는 아래와 같다. 한편, 매출액(sales)은 모수가 0.1인 지수분포로부터 추출한 숫자에 100을 곱해서 데이터를 생성하였다.

---

70) Fleming-Harrington 통계량의 가중치는 $w_j = \left[\hat{S}(t_j)\right]^{\rho}\left[1-\hat{S}(t_j)\right]^{\gamma}$이다. 추정된 생존함수($\hat{S}(t_j)$)는 사건이 발생할 때까지의 시간이 길어짐에 따라 감소하는 모습을 보이는데, ρ=0, γ=1으로 두면 시간이 길어질수록 가중치 w_j가 커지도록 만들 수 있다.

71) 이 시뮬레이션의 R 코드는 매우 길어 보고서 말미의 '시뮬레이션 코드 2.'에 수록하였다.

```
> table(area, type)
     type
area   0   1   2
   0 226 237 257
   1 290 220 289

> round(lambda, 4)
[1] 0.0212 0.0365 0.0620 0.0668 0.0561 0.0555
```

생성된 2개의 지역, 3개의 업종, 매출액 데이터를 이용하여 위험률을 예측하는 Cox-PH model을 적합시킨 결과는 아래와 같다.

```
coxph(formula = Surv(period, censoring) ~ as.factor(area) + as.factor(type) +
    sales, data = dt)

                        coef  exp(coef)   se(coef)      z        p
as.factor(area)1  3.523e-01  1.422e+00  6.120e-02  5.757 8.58e-09
as.factor(type)1  1.434e-01  1.154e+00  7.706e-02  1.861   0.0627
as.factor(type)2  3.994e-01  1.491e+00  7.674e-02  5.204 1.95e-07
sales            -1.129e-05  1.000e+00  3.119e-05 -0.362   0.7174

Likelihood ratio test=68.98  on 4 df, p=3.731e-14
n= 1486, number of events= 1099
```

제일 아래의 Likelihood ratio test = 68.98은 전반적인 모델의 유의성을 검정하기 위한 통계량이다. p-value가 상당히 작은 것으로 보아 모델이 전반적으로 유의하다고 판단할 수 있다. 구체적인 계수를 살펴보면 매출액(sales)를 제외한 모든 변수들의 추정된 계수가 양수인 것을 볼 수 있다. 각 변수의 p-value를 보면 area = 1인 지역의 p-value가 상당히 작다. type = 1인 업종은 유의수준 0.05를 고려하면 p-value가 경계선 근처에 있다고 볼 수 있다. type = 2인 업종은 상당히 작은 p-value를 갖고 있다. 하지만 매출액(sales)는 p-value가 매우 커서 위험률에 유의하지 않은 영향을 미치는 것으로 볼 수 있다.

구체적으로 살펴보자. 범주형 변수들은 기준값(baseline)과 비교하여 해석해야 한다. area = 1인 지역은 area = 0인 지역에 비해 폐업 위험률이 42% = (1.422 - 1) × 100% 가량 높다고 해석할 수 있다. 또한 type = 1인 업종은 type = 0인 업종에 비해 폐업 위험률이 15% = (1.154 - 1) ×

100% 가량 더 높고, type = 2인 업종은 type = 0인 업종에 비해 폐업 위험률이 49% = (1.491 − 1) × 100% 가량 더 높다고 해석할 수 있다. 이와 같은 해석에 따르면 결국 area = 1인 지역에 있는 type = 2의 업종의 폐업 위험률이 높은 편이라고 볼 수 있고, 이러한 지역과 업종의 사업자에게 세정지원 조치가 우선 실시되어야 한다고 결론 내릴 수 있다.[72)]

3. 고려해야 할 사항

생존분석은 세정지원이 필요한 사업자를 탐색하는 분야뿐만 아니라 국세행정의 다양한 분야에서 활용될 수 있다. 예를 들어 사업자가 사업을 개시한 이후 첫 번째 체납이 발생할 때까지의 기간을 생존분석의 대상으로 삼아 어떤 요인이 주로 영향을 미치는지 찾아낼 수 있을 것이다. 예컨대 특정 업종, 특정 규모, 특정 위치가 주요한 요인이라면 이에 해당하는 사업자를 중점적으로 관리하면 체납의 발생을 줄일 수 있을 것이다. 마찬가지의 방법으로 탈세에 대한 분석도 해 볼 수 있다.

본 보고서의 범위에서 다소 벗어나지만, 내부적인 관리에도 생존분석이 활용될 수 있다. 예를 들어 어떤 직원이 인시 이동한 이후 담당하게 된 체납이 정리되기까지의 시간을 이용하면 어떤 직원이 체납을 더 빠르게 정리했는지 찾을 수 있다. 분석의 단위를 세무서로 확대하는 것도 가능하다. 비슷한 방법으로 담당 직원에게 맡겨진 업무가 처리되는 시간을 이용하면 어떤 직원이 빠르게 업무를 처리했는지 확인할 수도 있다.

사업자등록 건수, 탈세 적발 건수, 체납정리 금액 등 국세행정의 많은 업무들이 대개 건수와 금액으로 평가되는 경향이 있다. 하지만 조금만 시선을 달리하여 어떤 사건이 발생하는 데까지 걸린 시간에 주목한다면 생존분석을 통해 이전에는 미처 알지 못했던 많은 통찰을 얻을 수 있을 것이다.

72) 이 시뮬레이션의 R 코드는 매우 길어 보고서 말미의 '시뮬레이션 코드 3.'에 수록하였다.

□ 효과적인 장려금 안내 홍보 매체 평가

1. 범주형 자료 분석의 개념

범주형 자료 분석(Categorical Data Analysis)은 범주로 표현되는 자료를 다루는 통계적 방법이다. 성별, 선호도, 혈액형, 지역, 학력 등 일상적인 생활 속에서 마주하는 많은 데이터들이 범주형 자료에 속한다. 범주형 자료는 간단하게 표의 형태로 나타낼 수도 있는데, 이를 분할표(Contingency Table)라고 한다. 예컨대 특정 집단을 대상으로 성별에 따른 MBTI 성격 유형별 빈도를 아래와 같이 조사해 보았다고 해 보자.

< **MBTI 성격 유형별 빈도**(예시) >

성별	INTJ	INFJ	…	ESFP
남성	10명	5명		3명
여성	15명	7명		6명

남성과 여성 MBTI 성격 유형별 분포에 차이가 있다고 말할 수 있을까? 구체적으로 어떤 유형이 상대적으로 더 많고, 적은지 말할 수 있을까? 이와 같은 질문들에 대답하기 위해서는 범주형 자료 분석이 필요하다. 범주형 자료 분석은 아래와 같은 통계량을 사용하여 성별에 따른 MBTI 성격 유형별 분포가 동일한지 검정한다.

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \widehat{\mu_{ij}}\right)^2}{\widehat{\mu_{ij}}},$$

여기서 n_{ij}는 i행, j열의 빈도수를 의미하며 $\widehat{\mu_{ij}}$는 해당되는 셀의 기대 빈도수로 아래와 같이 계산한다. I, J는 행과 열의 개수를 의미한다.

$$\widehat{\mu_{ij}} = n\widehat{\pi_{i+}}\widehat{\pi_{+j}} = \frac{n_{i+}n_{+j}}{n}$$

X^2를 피어슨 카이제곱 통계량(Pearson X^2)이라고 하는데, 성별에 따른 MBTI 성격 유형별 분포에 차이가 없다면 자유도가 $(I-1)(J-1)$인 카이제곱 분포를 따른다. 성별 간 성격 유형 분포에 차이가 있다고 결론이 난다면 어떤 항목이 더 큰지 또는 작은지 찾아야 하는데, 이를 위해서는 잔차(Residual)을 계산해야 한다. 보통 피어슨 잔차(Pearson Residual)와 표준화 잔차(Standardized Residual)를 많이 이용한다. 잔차의 절대값이 2 ~ 3을 벗어난다면 기대한 것에서 많이 벗어났다는 의미로 이해한다.

$$e_{ij} = \frac{n_{ij} - \widehat{\mu_{ij}}}{\sqrt{\widehat{\mu_{ij}}}} \qquad r_{ij} = \frac{n_{ij} - \widehat{\mu_{ij}}}{\sqrt{\widehat{\mu_{ij}}(1-n_{i+}/n)(1-n_{+j}/n)}}$$

피어슨 잔차 표준화 잔차

2. 장려금 홍보 매체 효과 분석에 활용

장려금은 신청이 있어야만 지급되기 때문에 신청기간에 앞서 광범위한 홍보를 실시한다. TV, 유튜브, 대중교통 등 다양한 광고 매체가 활용되고 있다. 홍보에는 많은 예산이 소요되는 만큼, 홍보 매체별로 정확하게 효과를 평가할 수 있어야, 예산을 절약할 수 있고 파급력 있는 홍보 매체에 집중할 수 있다. 이처럼 대상자별로 효과적으로 전달될 수 있는 홍보 매체를 분석하는 데는 범주형 자료 분석이 적합하다.

장려금 신청자에 대해 어떤 홍보 수단을 통해 신청기간이라는 것을 처음 인식하게 되었는지 설문조사를 실시한 결과가 아래와 같다고 해 보자.

< 장려금 안내를 최초 접촉한 홍보 수단(예시) **>**

성별	연령대	TV	신문	유튜브	대중교통
남성	20대	4명	5명	6명	15명
	30대	5명	9명	8명	12명
	40대	20명	18명	10명	5명
	50대	10명	9명	10명	7명
여성	20대	10명	7명	10명	14명
	30대	7명	8명	4명	7명
	40대	18명	15명	17명	10명
	50대	13명	11명	6명	8명

복잡한 분석이 없어도 표를 통해서 여러 정보를 얻을 수 있다. 20대는 유튜브를 통해 장려금 안내를 많이 접한 것으로 보이며, 연령대가 높아질수록 TV, 신문 등 전통 매체가 유용한 것으로 보인다. 하지만 직관을 이용해서 답변하기에 쉽지 않은 질문들도 있다. 예를 들어 20대 남성과 여성의 홍보 매체별 분포에 차이가 있는가? 차이가 있다면 20대 남성은 주로 어떤 홍보 매체를 통해 장려금 안내를 접하는가? 여성만을 고려하면 연령대별로 홍보 매체별 분포에 차이가 있는가? 이런 질문에 답할 수 있다면 홍보를 차별화하는 데 도움이 될 것이다.

먼저 전반적인 차이에 대해 살펴보자. 성별에 따라 연령대와 접촉 홍보 매체와의 관계가 달라지는지 검정하기 위해 Cochran-Mantel-Haenszel Test[73]를 실시하였다. 아래의 검정 결과를 보면 p-value가 매우 작다는 것을 알 수 있다. 이는 성별에 따라 연령대와 접촉 홍보 매체와의 관계가 달라진다는 것을 의미한다.

```
> mantelhaen.test(dt)

        Cochran-Mantel-Haenszel test

data:  dt
Cochran-Mantel-Haenszel M^2 = 23.051, df = 9, p-value = 0.006082
```

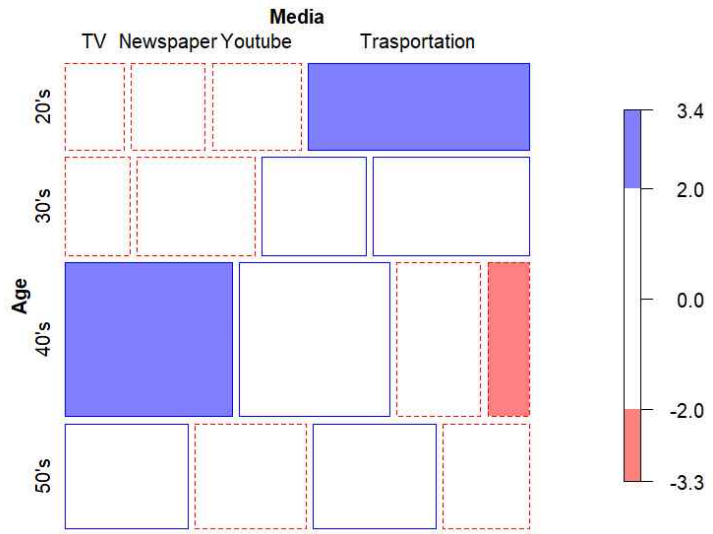
구체적으로 살펴보기 위해 남성, 여성으로 구분하여 피어슨 카이제곱 통계량을 계산해 보면 아래와 같다. 남성은 p-value가 매우 작아 연령별 접촉 홍보 매체에 차이가 있다고 결론을 내릴 수 있지만, 여성은 p-vlaue가 커서 차이가 있다는 결론을 내리기 어렵다.

```
> knitr::kable(df1,"simple")


        statistic   parameter   p.value
-------  ----------  ----------  ------------
Male     23.7814     9           0.004660631
Femail   8.035434    9           0.5305793
```

73) Cochran-Mantel-Haenszel Test는 I × J × K 분할표가 주어졌을 때 조건부 독립성을 검정하기 위한 방법 중 하나이다. Cochran-Mantel-Haenszel Test의 귀무가설은 K의 각 수준에서 다른 두 변수의 모든 조건부 승산비가 1이라는 것이다.($\theta_{XY(1)} = \theta_{XY(2)} = \ldots = \theta_{XY(K)} = 1$)

아래 그림은 남성의 연령 및 홍보 매체별로 계산된 잔차의 모자이크 그림이다. 사각형의 면적은 각 셀에 해당하는 표본의 크기를 의미하며, 파란 음영은 기대한 값보다 실제 관찰된 값이 큰 영역을, 빨간 음영은 기대한 값보다 실제 관찰된 값이 작은 영역을 의미한다.

< 남성의 연령 및 홍보 매체별 잔차의 모자이크 그림 >



위 그림은 20대 남성은 대중교통을 통해 장려금 안내를 주로 접하며, 40대 남성은 TV를 통해 장려금 안내를 주로 접하고 있음을 보여준다. 또한 40대 남성에게 대중교통을 통한 홍보는 효과적이지 않다는 사실도 보여준다. 이러한 정보들을 고려하여 홍보 매체별 타게팅 그룹을 정한다면 더 효과적인 장려금 홍보가 가능해질 것이다.

이번에는 연령대에 따라 성별과 접촉 홍보 매체와의 관계가 달라지는지 살펴보자. 마찬가지로 Cochran-Mantel-Haenszel Test를 실시하였다. p-value가 매우 낮은 것으로 볼 때, 연령대에 따라 성별과 접촉 홍보 매체와의 관계가 달라진다고 보기 어렵다고 결론 내릴 수 있다.

```
> mantelhaen.test(new.dt)

        Cochran-Mantel-Haenszel test

data:  new.dt
Cochran-Mantel-Haenszel M^2 = 0.55701, df = 3, p-value = 0.9062
```

이미 전반적인 검정에서 차이가 없다는 결론에 이르렀으므로 연령대별로 카이제곱 통계량을 살펴보는 것은 큰 의미가 없으나, p-value가 경계선 위에 있을 수도 있어 아래와 같이 출력해 보았다. 하지만 모든 p-value가 상당히 커서 다시 한번 앞선 검정의 결과를 확인할 수 있을 뿐 새로운 정보는 얻을 수 없었다.[74]

```
> knitr::kable(df2,"simple")


       statistic   parameter   p.value
-----  ----------  ----------  ----------
20's   2.289986    3           0.514442
30's   2.010353    3           0.5702608
40's   3.439041    3           0.3287548
50's   1.605089    3           0.6582364
```

3. 고려해야 할 사항

피어슨 카이제곱 검정, Cochran-Mantel-Haenszel Test 모두 근사적인 분포를 이용하므로, 표본의 크기가 너무 작을 경우에는 검정이 어렵다. 따라서 충분한 크기의 표본이 얻어질 수 있도록 설문조사를 설계해야 하며, 불가피하게 특정 셀에 너무 작은 표본이 얻어진 경우에는 주변의 셀을 통합시켜 분석하는 것을 고려해야 한다.

74) 이 시뮬레이션의 R 코드는 매우 길어 보고서 말미의 '시뮬레이션 코드 4.'에 수록하였다.

Ⅶ 민원 발급 편의를 위한 예측

□ 국세증명 민원서류 발급 현황

납세자들은 금융기관, 거래 상대방 등에게 제공하기 위해 다양한 국세증명 민원서류를 발급받고 있다. 통계를 보면 최근 들어 발급 건수가 빠르게 증가하는 모습이다. 2017년 약 3천만 건이었던 발급 건수는 지난 2021년 약 8천만 건으로 증가하였다. 특히 홈택스, 정부24 등을 통한 온라인 발급이 크게 늘어났다. 최근 들어 코로나19의 영향으로 사회 각 분야에 비대면·비접촉 트렌드가 확산되면서 민원 발급에서도 온라인 발급이 선호되고 있는 것으로 보인다.

< 국세증명 민원서류 발급 연도별 현황 >,

발급채널	2017	2018	2019	2020	2021
방문발급	4,880,659	5,298,445	5,446,302	6,349,099	4,600,606
홈택스	20,532,335	26,491,176	38,637,909	60,510,208	63,388,494
어디서나 민원처리	999,485	1,243,677	1,322,995	3,364,573	3,778,814
정부24	889,809	842,265	1,046,958	1,815,125	2,562,929
모바일 홈택스	138,297	347,992	764,134	1,594,388	2,010,950
무인민원 발급기	1,530,779	2,128,274	2,678,106	3,834,727	3,025,427
합계	28,971,364	36,351,829	49,896,404	77,468,120	79,367,220

출처 : 국가통계포털(KOSIS), https://kosis.kr/

방문 민원서류 발급 건수는 해당 기간 동안 늘지 않았으나 여전히 500만건에 달하고 있다. 다른 제도의 변경 등으로 국세증명 민원서류가 증빙으로 요구될 경우에는 방문 건수가 급격히 증가하기도 한다. 많은 납세자들이 세무서 민원실에 집중되지 않도록 홈택스를 통해 세무서 민원실 대기인원 조회 서비스[75]를 제공하고 있으나, 세무서마다 민원실

75) 민원실 대기인수 실시간 조회 https://teht.hometax.go.kr/websquare/websquare.html?w2xPath=/ui/ca/e/c/UTECAECA05.xml&mi=6768

종사 직원의 규모가 다르기 때문에, 납세자 입장에서는 민원서류 발급에 얼마나 시간이 소요될지 예측하기 어려운 상황이다. 세무서 입장에서도 방문 민원인이 많을 것으로 예상된다면 추가 인력의 투입 또는 인근 세무서 분산 안내 등을 통해 원활하게 민원을 처리할 수 있는 방안을 강구해 볼 수 있지만, 현재로서 이와 같은 예측은 민원실 담당 관리자의 경험과 직관에 의존할 수밖에 없는 실정이다.

납세자는 국세증명 발급 외에도 신고기간 중에 신고를 하기 위해서도 세무서를 방문한다. 통상 신고 마감일에는 많은 납세자들이 몰려 혼잡한 상황이 벌어지고는 하는데, 방문자 수를 과학적으로 예상할 수 있다면 보다 효율적으로 신고 업무를 처리할 수 있을 것이다. 시간이 갈수록 세무서의 인력 구조에서 경력자의 비중이 점차 낮아지고 있는 점을 고려하면 경력자가 보유한 경험과 직관을 대체할 수 있는 방안이 시급한 상황이다.

□ 데이터 과학 적용 방향

기상예보처럼 일주일 전 또는 며칠 전에 방문자 수를 예상하기에는 국세청이 가진 전산 자원과 인력의 한계가 명확하다. 최소한 당일의 민원인 방문 추세로부터 일정한 시간이 경과한 후에 얼마나 더 많은 민원인이 방문하게 될지 예상할 수 있다면 그것으로도 적적한 대응 조치를 취하는 데는 큰 도움이 될 것이다.

복잡하고 구성하는데 시간이 오래 걸리는 모델보다는 몇 개의 모수의 조정만으로 다양한 상황을 시뮬레이션해 볼 수 있는 간단한 모형이 더 쓸모가 있을 것으로 보인다. 최소한 방문자 유입과 관련된 모수, 방문자 처리와 관련된 모수, 민원실 직원의 수가 모델에서 사용되어야 할 것으로 생각된다.

또한, 대규모 데이터를 이용하는 모델은 사용하기 어려울 것으로 보인다. 세무서마다 민원실 환경이 다르기 때문에, 각 세무서마다 서로 다른 모델을 구성하고 적합시켜야 하는데, 대규모 데이터까지 사용할 경우 시간과 비용이 지나치게 많이 소요되기 때문이다.

□ M/M/s 대기열 모델을 이용한 방문자 수 예측

1. 마코프 체인의 개념

마코프 체인(Markov Chain)이란 한 상태에서 다른 상태로 전이되는 시스템을 확률적 방식으로 나타내는 수학적 모델의 하나이다. 특정한 상태로 전이될 확률은 시스템의 과거의 상태가 아닌, 현재의 상태에 의해서만 결정된다는 특징을 가지고 있다. 달리 말하자면, 현재와 과거의 상태를 모두 고려했을 때 미래의 특정 상태로 전이될 확률은 현재의 상태만을 고려했을 때 미래의 같은 상태로 전이될 확률과 동일하다는 의미이다. 이를 식으로 표현하면 아래와 같다.

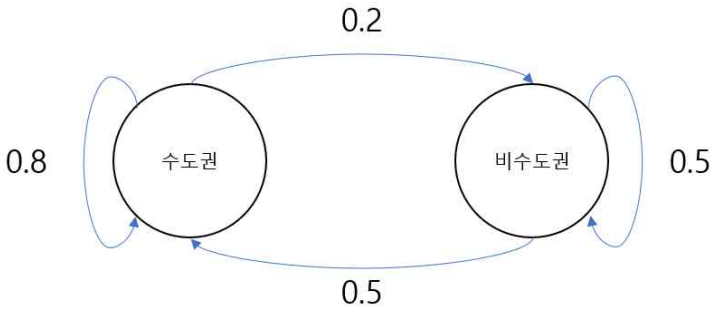
$$P(S_{t+1}|S_t) = P(S_{t+1}|S_t, S_{t-1}, \cdots S_1)$$

아래와 같이 어떤 상태에서 다른 상태로 전이되는 확률을 상태전이확률(State Transition Probability)이라고 한다. 상태전이확률은 행렬로 나타낼 수도 있고 도표로 나타낼 수도 있는데, 행렬로 나타낼 경우 정방행렬이 되며, 행의 확률 합계는 1이 되는 특징이 있다.

$$p_{ij} = P(S_{t+1} = j|S_t = i), \ i, j = 1, 2, \cdots, k$$

예컨대 수도권, 비수도권을 각각 상태로 여길 수 있는데 지역 간의 상태전이확률은 행렬로 나타낼 수도 있고, 아래와 같은 도표로도 나타낼 수 있다. 도표로 나타낸 것은 상태전이 다이어그램이라고 한다.

< 상태전이 다이어그램 >



< 상태전이행렬 >

$$P = \begin{bmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{bmatrix}$$

상태전이행렬을 이용하면 어떤 시점에서 수도권과 비수도권 인구가 얼마나 될지 예측해 볼 수 있다. 예를 들어 t=0인 시점에 수도권의 인구가 1천만명이고, 비수도권의 인구가 4천만명인데, t=3인 시점에 수도권과 비수도권의 인구가 궁금하다면, 아래와 같이 상태전이행렬의 거듭제곱을 이용하여 구하면 된다.

$$[1000\text{만 } 4000\text{만}]\begin{bmatrix}0.8 & 0.2\\ 0.5 & 0.5\end{bmatrix}^3 = [1000\text{만 } 4000\text{만}]\begin{bmatrix}0.722 & 0.278\\ 0.695 & 0.305\end{bmatrix}$$
$$= [3502\text{만 } 1498\text{만}]$$

이를 확장하면 오랜 시간이 지난 후에 어떤 상태에서 값이 변화하지 않고 균형을 이루는지도 찾아볼 수 있다. 이처럼 마코프 체인은 경제, 금융, 화학, 물리학 등의 광범위한 분야에서 확률적 프로세스를 모델링하는데 사용되고 있다. 특히 기계학습 분야, 자연어 처리, 음성 인식 분야 등에서도 많이 활용되고 있다.

2. 마코프 체인을 활용한 대기시간 예측

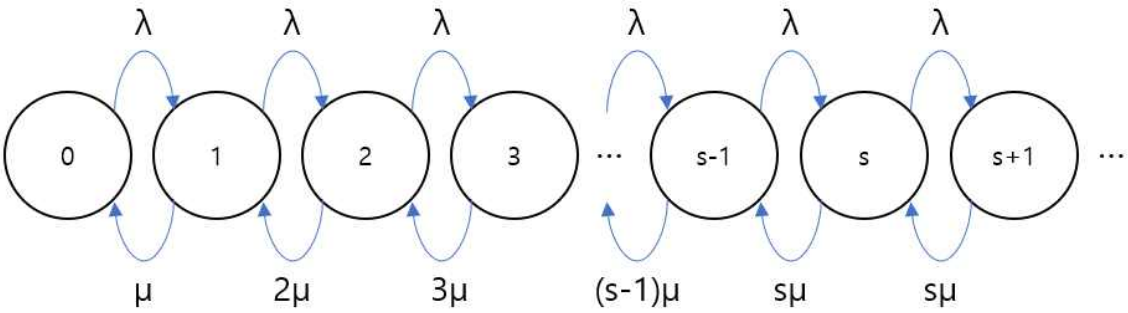
마코프 체인은 고객이 서비스 시설에서 서비스를 제공받기 위해 대기해야 하는 시스템을 모델링하는 데에도 사용될 수 있다. 대기 중인 고객 수와 현재 서비스를 제공하고 있는 서비스 제공자의 수에 해당하는 상태가 있으며, 고객의 유입과 유출에 따라 한 상태에서 다른 상태로 이동하는 확률을 생각해 볼 수 있다.

예를 들어, 모든 서비스 제공자가 업무 중일 때 고객이 시스템이 유입되면 대기열에 포함되고 시스템 내에 고객이 한 명 더 존재하는 새로운 상태로 전이된다. 서비스 제공자가 업무를 완료하면 대기열에 있는 고객 중 한 명에게 서비스가 제공되고 서비스 제공자는 업무 중인 새로운 상태로 전이된다. 이와 같은 모델을 사용하면 평균 대기시간, 대기해야 할 확률, 대기열에 있는 평균 고객 수 등을 예상할 수 있다.

현재 국세청 민원실에서 이루어지는 국세증명 민원발급은 마코프 체인이 모델링하고 있는 상황과 매우 흡사하다. 따라서 이를 그대로 적용할 수

있을 것으로 보인다. 다만 대기열을 모델링하는 방법에도 여러 종류가 있으나, 여기서는 M/M/s 대기열 모델을 적용해 보고자 한다. M/M/s 대기열 모델은 고정된 수의 서비스 제공자를 포함하고 있다는 특징이 있다 이 모델은 포아송 프로세스(Poisson Process)에 따라 민원실을 방문하는 고객을 상정하고 있으며 고객들이 방문하는 시간 사이의 간격은 지수 분포를 따른다고 본다. 이 모델이 고려하는 상황을 상태전이 다이어그램으로 나타내면 아래와 같다.

< M/M/s 모델의 상대전이 다이어그램 >



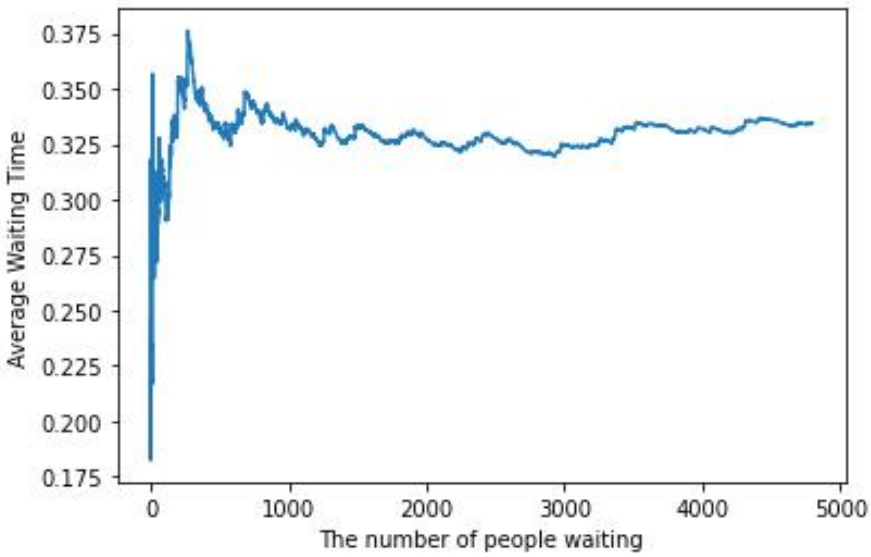
원 안의 숫자는 시스템 내 고객의 숫자를 의미한다. 단위 시간당 고객의 유입률은 λ로 표현된다. 단위 시간당 고객의 유출, 즉 서비스의 완료는 서비스 제공자 1명당 μ의 비율로 발생한다. 하지만 서비스 제공자가 s명만 존재하기 때문에 시스템 내 고객의 수가 s명을 초과하기 전까지는 $s\mu$의 비율로 서비스 제공이 완료되지만, s명을 초과하면 처리 속도는 $s\mu$의 비율로 유지된다. 서비스 제공자의 수(s)를 초과하는 시스템 내 고객의 숫자는 대기하는 고객의 숫자가 된다.

시뮬레이션에서는 고객 유입률(λ)과 서비스 완료율(μ)은 각각 2로 설정하였고 서비스 제공자의 수(s)도 2로 설정하였다. 그리고 10,000단위의 시간까지 변화를 관찰해 보았다. 10,000 단위의 시간이 경과한 후 대기를 해야 했던 고객 1명의 평균 대기 시간[76]은 0.33 단위 시간이었다. 그 역수를 구하면 3.03(= 1 / 0.33)인데, 이는 단위 시간당 대기열에서 대기하고 있는 평균 고객 수이다. 또한 대기 확률[77]은 0.24였다.

76) 고객들이 서비스 제공을 받기 위해 대기해야 했던 총 시간을 대기를 경험한 고객의 수로 나누어 구한다.
77) 서비스가 완료된 고객들 중에 대기를 경험했던 고객들의 비율로 구한다.

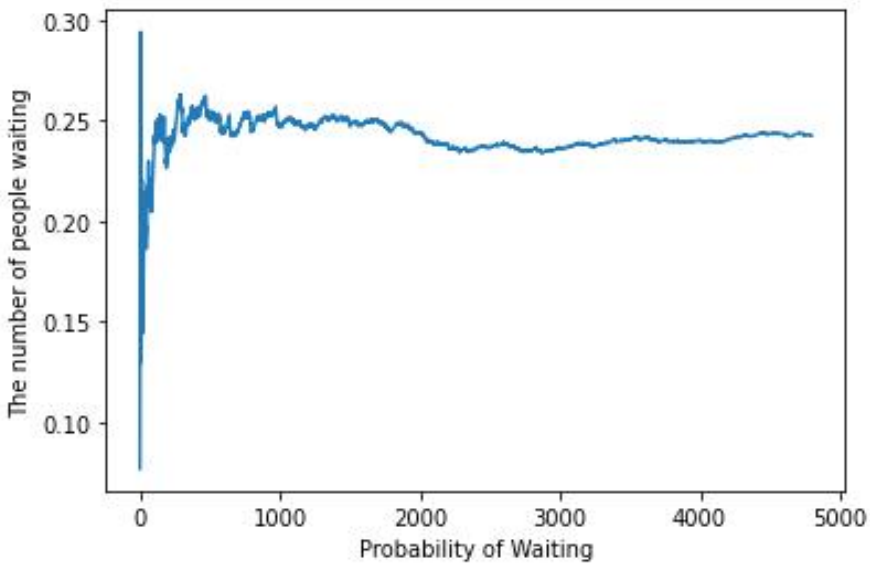
아래의 그래프는 평균 대기 시간이 변화하는 과정을 보여주고 있다. 처음에는 크게 요동치다가 일정한 시간이 지나면 안정된 범위 내에서 변동하는 모습을 볼 수 있다.

< 평균 대기 시간의 변화(λ = 2, μ = 2, s = 2) >



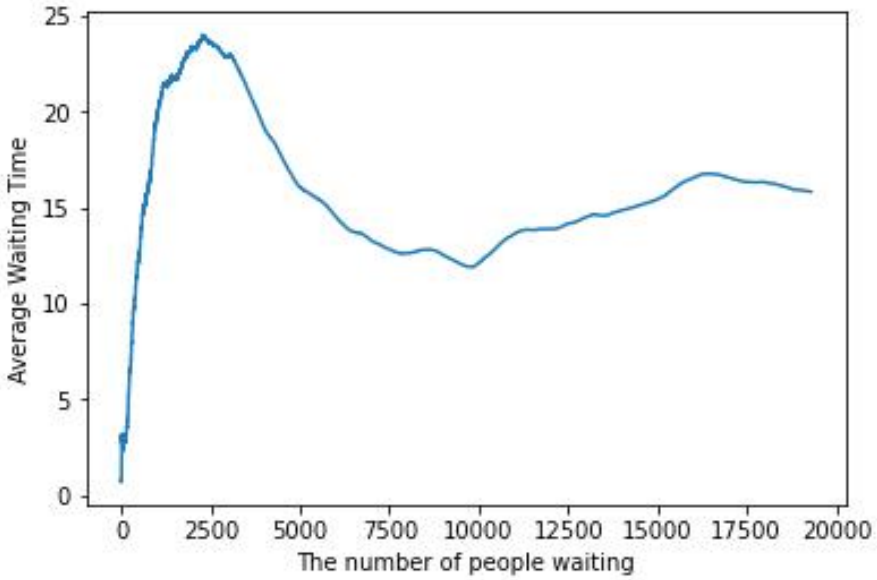
아래의 그래프는 대기 확률이 변화하는 과정을 보여주고 있다. 앞서와 마찬가지로 처음에는 크게 요동치다가 일정한 시간이 지나면 일정한 값으로 수렴하는 모습을 보여준다.

< 대기 확률의 변화(λ = 2, μ = 2, s = 2) >

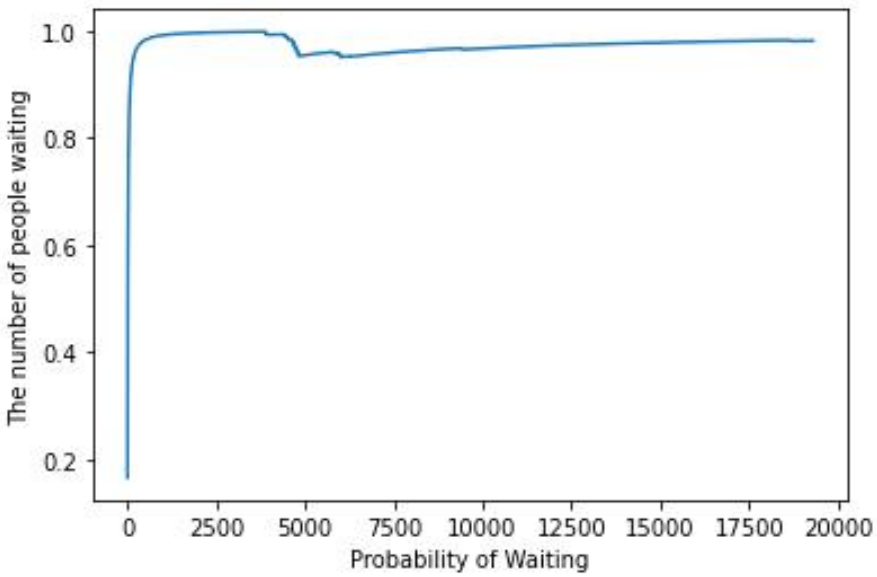


이 모델의 모수 역할을 하는 λ, μ, s가 적절한 균형을 이루고 있기 때문에 평균 대기시간이 수렴하는 모습을 보여주며, 대기 확률도 어느 한쪽으로 치우치지 않고 적정한 값으로 수렴한다. 하지만 λ, μ, s를 변화시키면 전혀 다른 그래프를 얻을 수 있다. 아래는 λ = 2, μ = 1, s = 2인 경우의 그래프이다.

< 평균 대기 시간의 변화(λ = 2, μ = 1, s = 2) >



< 대기 확률의 변화(λ = 2, μ = 1, s = 2) >



민원이 처리되는 속도가 절반으로 줄어들자 이전의 시뮬레이션에 비해 평균 대기시간이 크게 증가하였다. 또한 대기 확률도 거의 1에 수렴하고 있다. 이런 상황에서 민원실을 방문하는 민원인은 거의 확실하게 상당히 오랜 시간 대기를 해야 할 것으로 예측할 수 있다. 따라서 해당 세무서는 인력을 증원하여 대응하거나(s를 크게 만드는 것을 의미한다.), 민원실 직원들이 업무 처리 속도를 높이도록 독려할 수 있을 것이다, (μ를 크게 만드는 것을 의미한다.) 또는 방문하는 민원인이 인접한 다른 세무서를 이용하도록 안내할 수도 있다.(λ를 낮추는 것을 의미한다.)[78]

□ 고려해야 할 사항

매우 간단하고 기본적인 대기열 모델을 이용한 것이므로 실제 현실에서 이용하기 위해서는 보다 다양한 모수들을 포함하는 모형으로 발전시킬 필요가 있을 것이다. 민원실을 방문하였다가 사람이 너무 많으면 민원 발급을 포기하고 세무서를 빠져나가는 사람도 고려해 볼 수 있으며, 직원들의 숙련도에 따라 μ값을 달리 적용할 수도 있다. 분야를 달리하여 콜센터의 상담 전화 대기시간을 알려주거나, 법령에 대한 질의·회신이 처리되는 시간을 예측하는 데에도 같은 모델을 사용할 수 있다.

이 모델을 사용하기 위해서는 λ, μ를 추정할 필요가 있다. μ는 단위 시간당 업무 처리량이므로 내부적인 시스템을 이용하면 쉽게 추정할 수 있다. λ는 단위 시간 내에 방문하는 민원인의 수인데, 어떤 시간대에 민원인이 몇 명 방문했는지 추적할 수 있는 방법이 필요할 것으로 보인다. 간단하게는 놀이공원에서 탑승 인원수를 기록하듯이 일일이 세어보는 방법도 생각해 볼 수 있다. 또는 번호표를 뽑는 숫자를 기준으로 추정해 볼 수도 있다. 매우 어려운 방법으로 추정해야 하는 모수는 아니므로 다양한 방법을 이용할 수 있을 것이다.

78) 이 시뮬레이션의 R 코드는 매우 길어 '시뮬레이션 코드 5.'에 수록하였다.

Ⅷ 결론

□ 새로이 밀려드는 데이터 과학의 물결

지금까지 영화 및 드라마 추천, 신약 효과 검증, 콜센터 대기시간 예측 등 다양한 분야에서 활용되고 있는 데이터 과학 기법들을 불성실 사업자 관리, 납세자 지원, 민원 발급 등의 국세행정에 적용하는 방안을 소개하였다. 그리고 간단한 시뮬레이션을 함께 실시하여 제안한 수학·통계적 기법들이 현실에서도 작동할 수 있다는 가능성을 보여주었으며, 다른 응용 분야 또는 주의·고려해야 할 사항도 덧붙였다.

데이터 과학은 지금껏 사람들이 엄두를 내지 못했던 다양한 문제들을 해결해 줄 수 있지만, 그렇다고 데이터 과학을 모든 것을 해결해 줄 수 있는 만능 열쇠처럼 여기는 것도 위험한 태도이다. 때로는 너무 복잡해서 거대한 애물단지가 되어 버리기도 하고, 때로는 막대한 시간과 예산을 투입하고도 기대에 미치지 못하는 결과들을 보여주기도 한다. 지금까지 알려진 여러 기법은 특정 분야 또는 문제에 적용하기 위해 고안된 것이기 때문에 국세행정에 직접 대입하기는 어렵다. 하지만 민간이든, 공공이든 마주하는 문제의 본질은 대동소이하므로 어떤 분야가 데이터 과학을 적용하기에 적정한지 미리 섬세하게 검토한 후에 적정한 기법을 선택한다면 좋은 결과를 얻을 수 있을 것이라 기대한다. 예컨대 파일럿 데이터로 소규모의 모델을 구성하고 점진적으로 확대해 나가는 방식으로 접근한다면 실패의 위험도 크게 줄일 수 있을 것이다.

더 큰 문제는 실패의 위험을 너무 크게 생각하여 새로운 기법의 도입을 주저하는 것이다. 더 복잡하고 성능이 좋은 분석 기법들이 이미 많이 존재하고 있으며, 하루하루가 다르게 더 새로운 기술이 쏟아져 나오고 있다. 이와 같은 상황에서 아무런 조치도 취하지 않는다면 조금만 시간이 지나도 민간 부문과의 데이터 분석 역량의 격차는 더욱 크게 벌어지고 말 것이다. 민간과의 역량의 간극을 더 이상 좁히기 어려운 지경이 되면 그때는 더 큰 수고와 예산을 들여 민간 기업에게 분석을 의존해야 할 것이며, 데이터 주도권을 민간에게 내어주어야 할 것이다.

2000년을 전후로 인터넷이 보급되고 컴퓨터의 성능이 빠르게 향상됨에 따라 국세청에서는 직원들이 엑셀, 한글 등의 전산 프로그램을 자유자재로 이용할 수 있도록 교육을 실시하고 자격을 부여하기도 했다. 이러한 관심과 노력이 뒷받침되어 정보통신시대의 물결을 완만하게 맞이할 수 있었다. 이제 또 새로운 물결이 다가오고 있다. ChatGPT와 같은 인공지능이 이미 학생들의 과제를 도와주는 데 사용되고 있고, 데이터 과학에 기반하여 흥행이 예상되는 시놉시스를 만들고 이를 영화 또는 드라마로 제작하고 있는 상황이다.

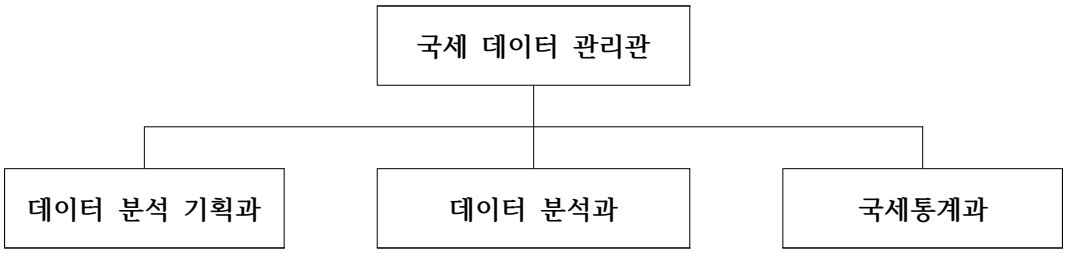
하지만 이전처럼 직원 교육을 통해 새로 밀려오는 물결에 적응하기가 쉽지 않다는 것이 문제이다. 새로운 분석 기법들은 매우 복잡한 수학, 통계학에 기초하고 있어 다루기도, 해석하기도 까다로운 경우가 많다. 장기적으로는 교육을 통해 직원들의 역량을 높이는 것이 바람직한 방향이겠지만, 효과가 나타나기까지는 시일이 많이 소요될 것이다. 따라서 단기적으로 조직의 데이터 분석 역량을 높이는 방법을 모색해야 한다.

□ 정책 제언 : 민간 인력 채용 및 조직 확충 필요

데이터 과학이 공공 부문에서 성공적으로 뿌리내리기 위해서는 행정적 측면의 뒷받침이 필요하다. 현재 국세청에서 데이터 과학 관련 업무를 맡고 있는 부서는 주로 기획조정관실 내의 국세데이터담당관실과 전산정보관리관실 내의 빅데이터 센터이다. 빅데이터 분석이 부상하면서 두 부서에서는 관련된 여러 가지 일들을 시작했는데, 국세청이 보유한 데이터의 규모, 업무의 다양함, 조직의 방대함을 고려하면 턱없이 인력이 부족한 실정이다. 결국 관련 부서의 협조가 필수적인데, 관련 부서는 현재 담당하고 있는 일로도 매우 바쁘기 때문에 당장 효과를 보장할 수 없는 데이터 분석에 시간과 인력을 투입할 여력이 없다.

따라서 데이터 분석을 전담할 별도의 실 단위의 조직이 필요하다. 국세데이터 관리관(가칭)이라고 하면 좋을 것이다. 그 아래에 데이터 분석을 기획하고 관련 업무를 총괄하는 데이터 분석 기획과, 기획된 데이테 분석을 실제 행정에 적용하는 데이터 분석과, 통계를 개발하고 국세통계연보를 발간하는 국세통계과가 설치되면 적절할 것이다.

< 데이터 분석 전담 조직 신설안 >



조직 신설뿐만 아니라 전문 인력의 확충도 병행되어야 한다. 현재 국세데이터담당관실에서는 일부 통계 분야의 민간 인재를 채용하고 있으나, 새로운 데이터 분석 과제를 발굴하고 시행하기에는 역부족이다. 따라서 실 단위의 조직을 본청에 설치하게 되면 통계, 수학, 데이터 과학 분야의 인재를 채용할 필요가 있다. 공무원 선발 직렬에 통계 또는 데이터 과학 직렬은 없기 때문에, 대부분 민간 인력을 채용해야 할 것이다.[79]

이들은 전산정보관리관실의 도움 없이 직접 로우 데이터(raw data)를 다룰 수 있어야 한다. 여러 차례 데이터를 가공하고, 최적의 모델을 찾기 위해서는 여러 차례 학습도 시켜야 하는데, 매번 전산정보관리관실의 도움을 받아야 한다면 업무 효율성이 극히 낮아지게 될 것이다. 다만 이들이 다루는 과세정보, 개인정보가 유출되거나 악용되지 않도록 각별한 보안 대책이 마련되어야 할 것이다.

각 세무서에도 데이터 분석, 통계 등을 전공한 민간 인력을 채용하면 국세청 본청 단위의 대규모 프로젝트로 다루기 어려운, 일선의 실정에 맞는 분석을 시행할 수 있을 것으로 기대한다. 국세청 본청에서는 모든 세무서에서 공통으로 사용할 수 있는 자료를 이용하기 때문에 세무서에서만 사용할 수 있는 소소한 데이터를 이용한 분석은 실시하기 어렵다. 따라서 일선 세무서 직원들의 다양한 수요를 충족해 줄 수 있는 민간 인력을 세무서에 배치할 필요가 있다고 생각한다.

미국의 IRS(Internal Revenue Service)는 IRS Integrated Modernization Business Plan을 2019년부터 실행하고 있다.[80] 노후화된 인프라를 개선

79) 다만, 자료의 의미와 해석을 돕기 위해 일부 세무 직렬 공무원이 함께 근무해야 할 것이다.

하여 납세자의 경험을 개선한다는 데 목표를 두고 있다. 그 구체적인 계획 중 하나는 Modernized IRS Operations인데, AI, 로봇 자동화, 데이터 분석 등 새로운 기술을 이용하여 프로세스를 자동화하는 것이다. 또한 Core Taxpayer Services & Enforcement 계획에서도 데이터 분석 기능을 강화하고 분석 기법을 향상시켜 조세 사기를 적발하고 납세자 문제를 탐지하는 한편, 신고 패턴을 예측한다는 목표를 세우고 있다. 실제로 IRS는 데이터 마이닝을 비롯하여 데이터 관리 응용 프로그램에 대한 새로운 접근 방식을 적용할 인재를 구하는 공고를 내고 있다.[81]

데이터 과학은 공공 부문보다는 민간 부문에 의해 주도되고 있는 분야이다. 많은 교육과 훈련이 필요하기 때문에, 기존의 인력을 교육하여 나날이 발전하는 데이터 과학을 뒤쫓아 가기에는 큰 무리가 따른다. 다행히 각광을 받는 분야이기에 많은 사람들이 도전하고 있다. 따라서 국세청, 그리고 넓게는 정부 조직 전체에 걸쳐 데이터 분석 관련 민간 전문 인력을 적극적으로 채용할 수 있을 것이다. 그리고 이를 주도할 전담 조직을 별도로 구성할 필요가 있다. 전산 담당자의 업무 중 하나로 치부하기에는 이미 거대한 대세적인 흐름이 되고 있다.

행정적 측면에서 뒷받침되지 않는다면 어떤 좋은 아이디어도 실현되기 어렵다. 조직의 체질을 근본적으로 바꾼다는 각오로 구조적인 변화를 단행한다면 '데이터 중심의 국세행정의 구현'도 멀지 않은 일이 될 것이다.

80) https://www.irs.gov/pub/irs-pdf/p5336.pdf
81) https://www.jobs.irs.gov/resources/job-descriptions/all-it-positions

시뮬레이션 코드

1. 코사인 유사도 및 행렬분해 시뮬레이션의 파이썬 코드

```
import numpy as np

def mf(R, k, n_epoch=5000, lr=.0003, l2=.04):[82)]
    tol = .001  # Tolerant loss.
    m, n = R.shape

    # Initialize the embedding weights.
    np.random.seed(13)
    P = np.random.rand(m, k)
    Q = np.random.rand(n, k)
    for epoch in range(n_epoch):

    # Update weights by gradients.
        for u, i in zip(*R.nonzero()):
            err_ui = R[u,i] - P[u,:].dot(Q[i,:])
            for j in range(k):
                P[u][j] += lr * (2 * err_ui * Q[i][j] - l2/2 * P[u][j])
                Q[i][j] += lr * (2 * err_ui * P[u][j] - l2/2 * Q[i][j])

        # compute the loss.
        E = (R - P.dot(Q.T))**2
        obj = E[R.nonzero()].sum() + lr*((P**2).sum() +(Q**2).sum())
        if obj < tol:
            break
    return P, Q

def st(M):
    m, n = M.shape
    R = np.zeros((m, n))
    for i in range(n):
        R[:,i] = (M[:,i] - np.nanmean(M[:,i])) / np.nanstd(M[:,i])
    np.nan_to_num(R, copy=False)
    return R, np.nanmean(M, axis=0), np.nanstd(M, axis=0)
```

82) mf 함수는 아래 사이트의 코드를 그대로 사용하였다.
Kyle Chung. 2019. “Matrix Factorization for Recommender Systems” https://everdark.github.io/k9/notebooks/ml/matrix_factorization/matrix_factorization.nb.html

```
def cos_sim(M):
    m, n = M.shape
    cosine = np.dot(M.T, M) / np.dot(np.linalg.norm(M, axis=0).reshape(n,1), np.linalg.norm(M, axis=0).reshape(1,n))
    return cosine, np.median(cosine, axis=0)

def replace(M, c):
    R = M.copy()
    R[-1, c[0]] = np.nan
    return R

# Original Data, The first row means a tax return in progress.
returns = np.array([[700, 60, 700, 70], [100, 10, 80, 8], [100, 10, 100, 10], [300, 25, 170, 20], [30, 3, 20, 2], [200, 19, 180, 20], [400, 44, 300, 30], [450, 4, 400, 40]], dtype='f')
np.savetxt('file1.csv', returns, fmt='%d', delimiter=',')

# Standardize data and Calculate cosine similarity
returns_st, _, _ = st(returns)
sim_matrix, sim_mean = cos_sim(returns_st)
np.savetxt('file2.csv', np.round(sim_matrix, 4), fmt='%f', delimiter=',')

# Detect an error field
idx = np.where(np.median(sim_mean) - sim_mean > 0.1)
print(idx[0][0] + 1, "번째 필드를 다시 확인해 보세요!")

# Change error value into NaN
returns_nan, nanmean, nanstd = st(replace(returns, idx))

# Predict appropriate value instead of NaN
P, Q = mf(returns_nan, k=3)
predictions = P.dot(Q.T)
np.savetxt('file3.csv', np.round(predictions, 4), fmt='%f', delimiter=',')

# Compare predicted value to the original value
predicted_value = np.round(predictions[-1, idx[0]] * nanstd[idx[0]] + nanmean[idx[0]], 0)
print(predicted_value[0], "정도가 되어야 할 것 같습니다." )
```

2. Weighted log-rank test 시뮬레이션을 위한 R code

```
library(survival)
library(nphRCT)

# Generate data
set.seed(1)
N <- floor(runif(2, 300, 500))
open_date1  <- floor(runif(N[1], 1, 12))
period1     <- floor(rexp(N[1], 0.01))
close_date1 <- open_date1 + period1
censoring1  <- ifelse(close_date1 > 12 * 3, 0, 1)
period_obs1 <- pmin(close_date1, 12 * 3) - open_date1
open_date2  <- floor(runif(N[2], 1, 12))
period2     <- floor(rexp(N[2], 0.01))
close_date2 <- open_date2 + period2
close_date2[which(close_date2 > 24)] <- open_date2[which(close_date2 > 24)] +
floor(qexp(pexp(period2[which(close_date2 > 24)], 0.01), rate=0.015))
censoring2  <- ifelse(close_date2 > 12 * 3, 0, 1)
period_obs2 <- pmin(close_date2, 12 * 3) - open_date2

# Draw plot for explaining transformation of data
x <- seq(0, 100, length=100)
y1 <- dexp(x, rate=0.01)
y2 <- dexp(x, rate=0.015)
plot(x, y1, type='l', lty=1, col="blue", ylim=c(0, 0.015), xlim=c(0, 100), ylab="Probability
Density")
lines(x, y2, lty=1, col="red")
abline(v = 20, col="blue", lty=2)
abline(v = qexp(pexp(20, 0.01), 0.015), col="red", lty=2)
legend(80, 0.015, legend=c("lambda=0.01", "lambda=0.015"), col=c("blue", "red"), lty=1,
cex=0.8)

# Make a data frame
period_obs <- c(period_obs1, period_obs2)
censoring  <- c(censoring1, censoring2)
group      <- c(rep(1, N[1]),rep(2, N[2]))
dt <- data.frame(period_obs, censoring, group)
colnames(dt)<-c("period_obs", "censoring", "group")
```

```
# Fit the model and draw survival function
ft <- survfit(Surv(period_obs, censoring) ~ group, data = dt)
plot(ft, lty = 2, mark.time = TRUE,
      xlab = "Months", ylab = "Survival Function",
      xlim=c(0, 40), ylim=c(0.5, 1), col=c("red", "blue"), conf.int = FALSE)
legend(35, 1,
        legend=c("2018년 개업", "2019년 개업"), col=c("red", "blue"), lty=2:3, cex=0.8)

# Non-weighted log-rank test
survdiff(Surv(period_obs, censoring) ~ group)

# Weighted log-rank test
wlrt(Surv(period_obs, censoring) ~ group, method="fh", data=dt, rho = 0, gamma = 1)
```

3. Cox-PH model 시뮬레이션을 위한 R code

```
library(survival)

# Generate data
set.seed(3)
N <- floor(runif(6, 200, 300))

area <- c(rep(0, N[1]+N[2]+N[3]), rep(1, N[4]+N[5]+N[6]))
type <- c(rep(0, N[1]), rep(1, N[2]), rep(2, N[3]), rep(0, N[4]), rep(1, N[5]), rep(2, N[6]))
table(area, type)

lambda <- runif(6, 0.01, 0.1)

period1 <- runif(N[1], 1, 12) + rexp(N[1], rate=lambda[1])
censoring1 <- ifelse(period1 > 12 * 3, 0, 1)
period_obs1 <- pmin(period1, 12 * 3)

period2 <- runif(N[2], 1, 12) + rexp(N[2], rate=lambda[2])
censoring2 <- ifelse(period2 > 12 * 3, 0, 1)
period_obs2 <- pmin(period2, 12 * 3)

period3 <- runif(N[3], 1, 12) + rexp(N[3], rate=lambda[3])
censoring3 <- ifelse(period3 > 12 * 3, 0, 1)
period_obs3 <- pmin(period3, 12 * 3)
```

```
period4 <- runif(N[4], 1, 12) + rexp(N[4], rate=lambda[4])
censoring4 <- ifelse(period4 > 12 * 3, 0, 1)
period_obs4 <- pmin(period4, 12 * 3)

period5 <- runif(N[5], 1, 12) + rexp(N[5], rate=lambda[5])
censoring5 <- ifelse(period5 > 12 * 3, 0, 1)
period_obs5 <- pmin(period5, 12 * 3)

period6 <- runif(N[6], 1, 12) + rexp(N[6], rate=lambda[6])
censoring6 <- ifelse(period6 > 12 * 3, 0, 1)
period_obs6 <- pmin(period6, 12 * 3)

sales <- 100 * rexp(sum(N), rate=0.1)

# Make a data frame
period <- c(period_obs1, period_obs2, period_obs3, period_obs4, period_obs5, period_obs6)
censoring <- c(censoring1, censoring2, censoring3, censoring4, censoring5, censoring6)
dt <- data.frame(period, censoring, area, type, sales)

# Fit the Cox-PH model
coxph(Surv(period, censoring) ~ as.factor(area) + as.factor(type) + sales, data = dt)
```

4. 범주형 자료 분석을 위한 R code

```
# Input Data
dt <- as.table(array(c(4, 5, 20, 10, 5, 9, 18, 9, 6, 8, 10, 10, 15, 12, 5, 7,
                       10, 7, 18, 13, 7, 8, 15, 11, 10, 4, 17, 6, 14, 7, 10, 8),
                     dim=c(4, 4, 2),
                     dimnames(dt) <- list(Age=c("20's", "30's", "40's", "50's"),
                                          Media=c("TV", "Newspaper", "Youtube", "Trasportation"),
                                          Sex=c("Male", "Female"))))
ftable(. ~ Sex + Age, dt)

# Find association between Age and Media controlling Sex
mantelhaen.test(dt)

s1<-chisq.test(dt[ , ,1])
s2<-chisq.test(dt[ , ,2])
```

```
df1 <- data.frame(rbind(s1[1:3], s2[1:3]))
rownames(df1) <- c("Male", "Femail")
knitr::kable(df1,"simple")

mosaic(dt[,,1], residuals = chisq.test(dt[,,1])$stdres, gp=shading_Friendly)

# Find association between Sex and Media controlling Age
new.dt <- aperm(dt, c(3,2,1))
mantelhaen.test(new.dt)

# Chi-squared tests by age
x1<-chisq.test(dt[ 1, ,])
x2<-chisq.test(dt[ 2, ,])
x3<-chisq.test(dt[ 3, ,])
x4<-chisq.test(dt[ 4, ,])
df2 <- data.frame(rbind(x1[1:3], x2[1:3], x3[1:3], x4[1:3]))
rownames(df2) <- c("20's", "30's", "40's", "50's")
knitr::kable(df2,"simple")
```

5. M/M/s 대기열 모델 시뮬레이션을 위한 파이썬 코드

```
import numpy as np
import matplotlib.pyplot as plt

# Set simulation parameters
lm = 2              # Arrival rate
mu = 2              # Service rate
s = 2               # Number of servers
sim_time = 10000  # Simulation time

# Initialize the system
state = 0               # Number of customers in the system
queue = 0               # Number of customers in the queue
servers = np.zeros(s)  # Status of servers (0 for idle, 1 for busy)

# Initialize simulation clock
time = 0
```

```
# Initialize performance measures
num_arrivals      = []
num_completed     = []
num_wait          = []
total_wait_time = []
Probability_of_waiting = []

# Initialize random seed
i = 777

# Run simulation
while time < sim_time:

  # Set random seed
  np.random.seed(i)

  # Determine next event (arrival or completion)
  next_arrival      = np.random.exponential(1/lm)
  next_completion = np.inf if state == 0 else np.min(np.random.exponential(1/mu, s))

  if next_arrival < next_completion:

    # Handle arrival event
    time += next_arrival
    num_arrivals.append((num_arrivals[-1] if num_arrivals else 0) + 1)
    state += 1

    if state <= s:
      # Customer is served immediately
      servers[np.where(servers == 0)[0][0]] = 1
    else:
      # Customer is queued
      queue += 1
```

```
    else:
        # Handle completion event
        time += next_completion
        num_completed.append((num_completed[-1] if num_completed else 0) + 1)
        state -= 1

        if queue > 0:
            # Serve queued customer
            num_wait.append((num_wait[-1] if num_wait else 0) + 1)
            total_wait_time.append((total_wait_time[-1] if total_wait_time else 0) + queue *
next_completion)
            Probability_of_waiting.append(num_wait[-1] / num_completed[-1])
            queue -= 1

        else:
            # Free server
            servers[np.where(servers == 1)[0][0]] = 0

    # Update random seed
    i += 1

# Print results
print("Average Waiting Time:", round(total_wait_time[-1] / num_wait[-1], 2))
print("Probability of waiting:", round(num_wait[-1] / num_completed[-1], 2))

plt.plot(range(0, len(total_wait_time)), np.array(total_wait_time) / np.array(num_wait))
plt.xlabel('The number of people waiting')
plt.ylabel('Average Waiting Time')
plt.show()

plt.plot(range(0, len(Probability_of_waiting)), Probability_of_waiting)
plt.xlabel('Probability of Waiting')
plt.ylabel('The number of people waiting')
plt.show()
```

기타 참고문헌

Agresti, Alan. 2013. Categorical Data Analysis. Wiley Series in Probability and Statistics. Wiley-Interscience.

Brémaud, Pierre. 2020. Markov Chains : Gibbs Fields, Monte Carlo Sim ulation and Queues. 2nd ed. 2020. Texts in Applied Mathematics: 31. Springer International Publishing.

Cox, D. R., and David (Statistician) Oakes. 1984. Analysis of Survival Data. Monographs on Statistics and Applied Probability. Chapman and Hall.

Fawcett, Tom. 2006. "An Introduction to ROC Analysis." Pattern Recognition Letters 27 (8): 861.

Fox, John, Sanford Weisberg, and John Fox. 2011. An R Companion to Applied Regression. 2nd ed. SAGE Publications.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. The Elements of Statistical Learning. [Electronic Resource] : Data Mining, Inference, and Prediction. Second Edition. Springer Series in Statistics. Springer New York.

Iñaki Ucar, and modified Stefanka Chukova. Queueing Systems M/M/1 and M/M/c. https://homepages.ecs.vuw.ac.nz/~schukova/SCIE201/Lectures/Lecture9_final2018.html

James, Gareth, Trevor Hastie, Robert Tibshirani, and Daniela Witten. 2021. An Introduction to Statistical Learning : With Applications in R. 2nd ed. 2021. Springer Texts in Statistics. Springer US.

Kleinbaum, David G., and Mitchel Klein. 2012. Survival Analysis. [Electronic Resource] : A Self-Learning Text, Third Edition. Statistics for Biology and Health. Springer New York.

Klein, John P., and Melvin L. Moeschberger. 2003. Survival Analysis : Techniques for Censored and Truncated Data. Second edition. Statistics for Biology and Health. Springer.

LEE, Jung Wan. 2020. Big Data Strategies for Government, Society and Policy-Making. The Journal of Asian Finance, Economics and Business 7(7):457-487.

Lin, Haiqun, and Daniel Zelterman. 2002. "Modeling Survival Data: Extending the Cox Model T. M. Therneau P. M. Grambsch." Techno metrics 44 (1): 85-86.

McGee, Daniel L. 2010. "Applied Survival Analysis: Regression Modeling of Time-to-Event Data (2nd Ed.) David W. Hosmer Stanley Lemeshow Susanne May." The American Statistician 64 (2): 191.

R. Jordan Crouser. 2017. SDS 293 - Machine Learning. https://www.science.smith.edu/~jcrouser/SDS293/

Rhys, Hefin. 2020. Machine Learning with R, the tidyverse, and mlr (1st edition.). Manning Publications.

Rick Durrett. 2019. Probability: Theory and Examples. Cambridge UP,

Sandro Sperandei. 2013. Understanding logistic regression analysis. Biochemia Medica 2014;24(1):12-8

Weisberg, Sanford. 2014. Applied Linear Regression. Wiley Series in Probability and Statistics. Wiley.

Wickham, Hadley, and Garrett Grolemund. 2016. R for Data Science : Import, Tidy, Transform, Visualize, and Model Data. First edition. O'Reilly Media.