조사효율화를 위한 파라데이터 수집 및 활용방법 관련 선진사례 연구

(부처맞춤형과정)

2019. 8.

통 계 청
최재혁, 허창수, 김상우

국외훈련개요

1. 훈련국 : 스페인, 크로아티아

2. 훈련기관명

- 스페인 폼푸파브라대학교 조사방법론 연구전문센터 (RECSM; The Research and Expertise Centre for Survey Methodology, The Pompeu Fabra University)
- 유럽조사연구학회 2019 컨퍼런스 (ESRA2019 Conferences; European Survey Research Association)

3. 훈련분야 : 조사효율화를 위한 파라데이터 수집 및 활용방법 관련 선진사례 연구

4. 훈련기간 : 2019. 7. 7. ~ 7. 21.(15일간)

목 차

Ⅰ. 훈련 개요

1. 훈련개요

□ 훈련 목적

현재 국가통계생산은 지속적인 조사환경 변화의 시대에 직면하고 있다. 과학적 조사표 설계의 발전, 데이터 수집 및 분석기술의 고도화, 인터넷 및 웹 기반의 조사환경 확산 등 긍정적 변화라고 볼 수 있다. 반면, 조사표 복잡성 확대, 조사비용 증가, 응답률 감소 등 조사환경 악화에 대한 문제점도 동시에 발생하고 있다. 이에 통계생산 비용과 품질 측면 최적화를 위해 유연한 조사 관리와 설계가 필요한 시점이 도래하고 있다. 통계조사 효율화를 위한 현장조사 지침 및 조사방법 개선의 일환으로 조사과정 중에 수집되는 파라데이터 활용 필요성이 강하게 대두되고 있어 파라데이터 수집 및 활용방안 연구를 통해 국가통계 품질, 응답률, 비용 효율화 등 조사효율화를 위한 최신기법 검토가 필요한 시점이다. 파라데이터(paradata; 조사과정자료)란 조사시작과 종료시간, 문항별 응답시간 등의 조사과정에서 발생하는 자료를 의미한다.

□ 훈련의 필요성

현재 미주(캐나다, 미국)와 유럽(스페인, 네덜란드, 영국, 독일, 스웨덴 등)을 중심으로 자동화된 시스템에서 파라데이터를 수집하고 이를 조사환경 개선 및 조사설계 단계에서 활용하고 있다. 따라서 파라데이터 수집 및 활용에 대한 유럽의 선진사례를 벤치마킹하여 조사환경 개선 및 조사방법 선진화에 적용할 필요가 있다. 첫 번째로 조사과정 효율화, 응답거절 최소화 방안, 불응 대응논리 마련, 조사담당자 교육 등 조사관리를 위해 파라데이터를 수집하고 활용할 필요가 있으며, 두 번째로 파라데이터를 이용하여 현장조사 지침 및 조사방법을 개선하고

맞춤형 응답자 접촉전략 수립 등을 통해 통계조사를 효율화할 필요가 있다. 외국에서는 파라데이터 수집을 통해 허위·부실조사를 방지하고 조사원 평가 기준 수립, 조사 기술 향상 개발 및 교육 등에 활용하여 조사효율화에 기여하고 있다.

특히 현 시점에서는 고품질 및 저비용의 통계 생산을 위하여 응답자 특성을 반영한 적응적·반응적 조사설계(adaptive and responsive survey design) 방안 연구 및 도입을 검토해야 한다. 적응적·반응적 조사설계란 통계조사에서 오차와 비용은 상충관계에 있으므로 오차 최소화 목적 하에서 최적 비용 수준의 조사를 설계하는 방법으로 주로 파라데이터와 보조자료를 사용하여 설계하는 방법을 의미한다. 파라데이터 및 보조정보 등을 활용한 선진사례 방법으로는 ① 현장조사 지침 개선, ② 맞춤형 응답자 접촉전략 수립, ③ 설계조정방법, 품질 및 비용의 최적화 전략 연구, ④ 적응적·반응적 조사설계 현장 적용 방안 연구, ⑤ 현장 유용성 평가 등이 있으며, 이에 대한 지속적인 연구와 습득이 필요하다.

□ 훈련대상자 인적사항

구분	소속	직급	성명
팀장	통계청 통계개발원 통계방법연구실	통계사무관	최재혁
팀원	통계청 조사기획과	통계주사	허창수
팀원	통계청 조사시스템과	전산주사보	김상우

□ 훈련기관 및 훈련기간

훈련국 및 훈련기관	기간	훈련내용
스페인 (폼푸파브라 대학교, 조사방법론 연구전문센터)	7.7(일)	○ 출국 및 스페인 도착, 훈련준비
	7.8(월)~7.10(수)	<교육프로그램 참가> ○ 사회과학에서 기계학습
	7.11(목)~7.12(금)	<교육프로그램 참가> ○ 고품질 웹조사의 구현

훈련국 및 훈련기관	기간	훈련내용
크로아티아 (ESRA2019, 자그레브 대학교)	7.13(토)	○ 스페인 → 크로아티아 이동
	7.14(일)	○ 훈련 내용 정리 및 준비
	7.15(월)	<컨퍼런스 참가> ○ 반일 교육과정 수강
	7.16(화)~7.19(금)	<컨퍼런스 참가> ○ 논문 발표 세션 참가
	7.20(토)~7.21(일)	○ 귀국

□ 훈련과제 및 내용

(훈련분야) 훈련과제	훈련내용
조사효율화를 위한 파라데이터 수집 및 활용방법 관련 선진사례 연구	**[스페인 폼푸파브라대학교 조사방법론 연구전문센터]** ○ 품질 높은 웹조사를 위한 파라데이터 수집방법 습득 ○ 비용 및 오차 측면의 최적 웹기반 조사방법 탐색 ○ 자동화된 자료 분석방법인 기계학습(machine learning)의 파라데이터 적용방법 습득 * 기계학습: 자동으로 자료의 패턴을 감지 또는 학습 ○ 의사결정나무 등 파라데이터 분석방법 사례 파악 **[크로아티아 유럽조사연구학회 2019 컨퍼런스]** ○ 파라데이터를 활용한 현장조사 지침 개선방안 및 응답자 접촉전략 연구동향 파악 ○ 파라데이터를 이용한 적응·반응설계 기법 및 사례 습득 ○ 현장조사 개선 및 자료수집 효율화 등 조사기획·설계 분야의 최신 연구동향 파악 ○ 조사효율화와 파라데이터와 관련하여 최신 이슈와 연구 동향을 파악하고 참여 연구자들과 적극적인 소통 기회를 갖고 국제적 연구 협력 네트워크 구축

첫째, 향후 국가통계방법론 연구에 반영하기 위해 전문적인 방법론 습득, 최신 이슈와 연구 동향 파악, 국제적 연구 협력 네트워크 구축 등 향후 조사 효율화에 대한 국가통계방법론의 실질적인 연구의 기초 자료를 습득하고자 한다. 둘째, 파라데이터 수집과 조사원 관리 분야에 실질적 적용가능성을 검토하기 위해 파라데이터 수집·분석방법과 현장조사 지침, 응답자 접촉전략, 적응·반응 조사설계, 등 실제 현장 적용 가이드라인 도출을 위한 자료를 수집하고자 한다. 셋째, 한국 통계청의 파라데이터 수집시스템의 장애를 파악하고 해결방안 검토하기 위해 파라데이터 수집·분석 시스템 개선 도출을 위한 자료를 수집하고자 한다.

□ 훈련 세부일정

날짜	훈련국 및 훈련기관	훈련내용
7. 7. (day 1)	스페인	출국 및 캐나다 도착
7. 8. (day 2)	스페인 폼푸파브라 대학교 조사방법론 전문연구센터	사회과학에서 기계학습 교육 1 (기계학습의 기본 개념 등 학습)
7. 9. (day 3)		사회과학에서 기계학습 교육 2 (파라데이터 분석사례 파악)
7.10 (day 4)		사회과학에서 기계학습 교육 3 (파라데이터 분석에서 기계학습 활용사례 파악)
7.11. (day 5)		고품질 웹조사의 구현 교육 1 (웹조사 기획방법 학습)
7.12. (day 6)		고품질 웹조사의 구현 교육 2 (웹조사의 특성에 따른 문제점 해결방법 파악)
7.13. (day 7)	크로아티아	스페인(바르셀로나) → 크로아티아(자그레브) 이동
7.14. (day 8)		훈련내용 정리 및 훈련준비

날짜	훈련국 및 훈련기관	훈련내용
7.15. (day 9)	크로아티아 유럽조사 연구학회 2019 컨퍼런스	반일 교육과정 1 (조사방법론의 본질) 반일 교육과정 2 (스마트폰: 조사부터 센서까지)
7.16. (day 10)		논문발표 세션 1 (응답자 접촉전략 방법론, 허위·부실 조사 해결방안, 변화하는 조사환경에서 조사의 유용성, 조사원-응답자 간의 상호작용 등)
7.17. (day 11)		논문발표 세션 2 (현장 모니터링 프레임워크, 반응적·적응적 조사설계, 응답자 부담 측정 및 감소방안 등)
7.18. (day 12)		논문발표 세션 3 (온라인 프로빙과 대면조사의 비교, 조사원 효과 탐색 및 관리방안 등)
7.19. (day 13)		논문발표 세션 4 (기계학습, 조사원 능력, 무응답 편향 평가 등)
7.20. (day 14)	크로아티아	귀국 및 인천 도착
7.21. (day 15)	한국	

2. 훈련기관 개요

□ 스페인 폼푸파브라 대학교 조사방법론 연구전문센터(RECSM)

명 칭		스페인 폼푸파브라 대학교 조사방법론 연구전문센터 (The Research and Expertise Centre for Survey Methodology, The Pompeu Fabra University)
소 재 지		Campus de la Ciutadella, Ramon Trias Fargas, 25-27 - 08005 Barcelona ※ 홈페이지 주소 : upf.edu/web/survey
설립목적		조사 개발, 자료수집, 자료 품질 및 분석 등 조사방법론 분야의 전문연구기관(연구, 컨설팅, 교육)
조 직		RECSM은 3명은 공동센터장, 4개의 연구파트*, 약 50명의 연구원으로 구성 * Advanced Survey Quality Methods, Behavioral and Experimental Social Science, Public Opinion and Political Behaviour, Statistical Methods Applied to Social Science
주요인사 인적사항		• Prof. Mariano Torcal(공동센터장) - 폼푸파브라 대학교 정치사회학과 교수 - 스페인 ESS(유럽사회조사) 위원장 - ICREA(카탈루니아 주정부 연구기관) 연구원
훈련기관 특성	**주요기능 및 수행업무**	• 조사개발, 조사설계, 조사표, 자료수집, 자료품질 등 조사방법론 분야와 정책결정, 인구사회 통계 분석 등 고급분석방법론 분야의 전문연구센터 • 통계생산자에게 조사방법론 및 고급분석방법론 분야의 컨설팅 • 전 세계의 통계전문가, 통계종사자, 연구원 등을 대상으로 교육프로그램 운영(summer school, 세미나, 컨퍼런스 등 연중) • 논문집(RECSM Working Paper Series) 발간
	훈련과제 수행과의 관련성	• 조사환경 변화에 대비하는 조사효율화 기법을 꾸준히 연구 중임 • 특히, 조사개발, 조사설계, 조사표, 자료수집, 자료품질 등 조사방법론 분야의 수준 높은 연구가 진행 중임 • 유럽연합통계청, 스페인 통계청 등과 조사환경 악화 대비 및 신기술 조사방법에 대한 공동연구를 수행 중임 • 또한, 전 세계 통계전문가, 통계종사자, 연구원 등을 대상으로 운영중인 교육프로그램의 강사진은 세계적이고 권위 있는 석학과 전문가로 구성되어 우수한 교육프로그램으로 알려져 있음 • 올해 교육프로그램은 파라데이터 분석을 위한 기계학습(자동으로 자료의 패턴을 감지), 의사결정분석 등의 방법론과 조사효율화를 위한 웹조사의 고품질 구현, 조사설계, 오차 측정 등의 조사방법론에 대한 전문적 교육을 진행할 예정임
교섭창구		(담당) Tagreed Mustafa (RECSM Coordinator) (e-메일) Tagreed.mustafa@upf.edu (전화) +34-93-542-1558

□ 유럽조사연구학회 2019 컨퍼런스(ESRA2019 Conference)

명 칭		유럽조사연구학회 2019 컨퍼런스 (European Survey Research Association(ESRA) 2019 Conferences)
소 재 지		학회주소 : Southampton University, SO17 1BJ, UK 컨퍼런스 : University of Zagreb, Trg Johna Kennedyja 6, 10000, Croatia ※ 홈페이지 주소 : europeansurveyresearch.org
설립목적		유럽과 다른 지역 조사방법 연구자 간 연계와 소통 (조사연구 질 향상을 통해 학술·정책·상업 연구에서 적절히 사용될 수 있도록 연구원 간의 의사소통 및 상호작용 지원)
조 직		ESRA 위원회는 학회장 포함 15명의 임원으로 구성되어 있고 전 세계 많은 조사방법 연구자들을 회원으로 두고 있음
주요인사 인적사항		• 학회장 : Prof. Bart Meuleman - 벨기에 루벤대학교 사회과학학부 교수 • 컨퍼런스 운영위원장 : Prof. Caroline Roberts - 스위스 로잔대학교 사회과학학부 교수 - 스위스 사회과학 연구재단 위원 - 스위스 국립연구센터 위원
훈련기관 특성	**주요기능 및 수행업무**	• 유럽의 조사 연구자와 세계 다른 지역의 연구자와의 소통 • 조사방법 연구자와 조사를 이용하는 사회과학 연구자 간의 의사소통을 통한 연구 질 향상 • 조사품질 상상을 위한 조사설계, 조사운영, 분석기술 등에 관한 연구 • 세미나, 학술대회, 심포지엄 등을 개최하고 논문집(Survey research Methods) 발간
	훈련과제 수행과의 관련성	• 유럽조사연구학회 컨퍼런스는 조사품질 향상을 목적으로 다양한 분야의 조사방법론 연구자들이 소통하는 대규모 학술대회 • ESRA2019 컨퍼런스는 조사효율화와 파라데이터과 관련된 표본설계, 조사표설계, 자료수집방법, 혼합모드, 모드효과, 자료연계, 행정자료, 빅데이터 등을 주제로 약 220개 세션으로 진행될 예정임 • 조사효율화와 파라데이터와 관련하여 최신 이슈와 연구 동향을 파악하고 참여 연구자들과 소통 기회를 갖고 국제적 연구 협력 네트워크 구축 • 해외 조사방법 연구 경험을 공유, 통계청에서 향후 추진할 과제에 적용하여 연구 방향 설정 및 과제 발굴에 활용 가능
교섭창구		(담당) ESRA 2019 등록담당자 (e-메일) conference@europeansurveyresearch.org (전화) 없음

Ⅱ. 조사방법론 전문 교육프로그램

1. 교육프로그램 개요

스페인 바르셀로나에 소재한 폼푸파브라대학교 조사방법론 연구전문센터에서는 2014년부터 전 세계의 통계전문가, 통계종사자, 연구원 등을 대상으로 최고의 전문가를 강사로 초빙하여 매년 7월 2주간의 교육프로그램(summer school)을 진행해오고 있다. 이번 2019년 교육프로그램은 SPSS와 R 프로그램 선수 과정을 시작으로 다차원 모형의 실제(practical multilevel modeling)와 종단 조사자료에서의 비교분석(analysing comparative longitudinal survey data using multilevel models), 조사표 설계(questionnaire design), 확률추출법(probability sampling method), 비실험자료(nonexperimental data)와 조사실험(survey experiments)의 원인추론(causal inference), 사회과학에서의 응용베이지안 모형(applied bayesian modeling for social science), 사회네트워크(social networks), 사화과학에서의 기계학습(machine learning for social science), 사회미디어 연구(social media research), 빅데이터(bigdata), 고품질 웹 조사의 구현(Implementing high-quality web surveys) 등의 과정이 개설되었다. 본 훈련에서는 사화과학에서의 기계학습과 고품질 웹 조사의 구현의 2개 과정을 수강하여 진행하였다.

2. 사회과학에서의 기계학습

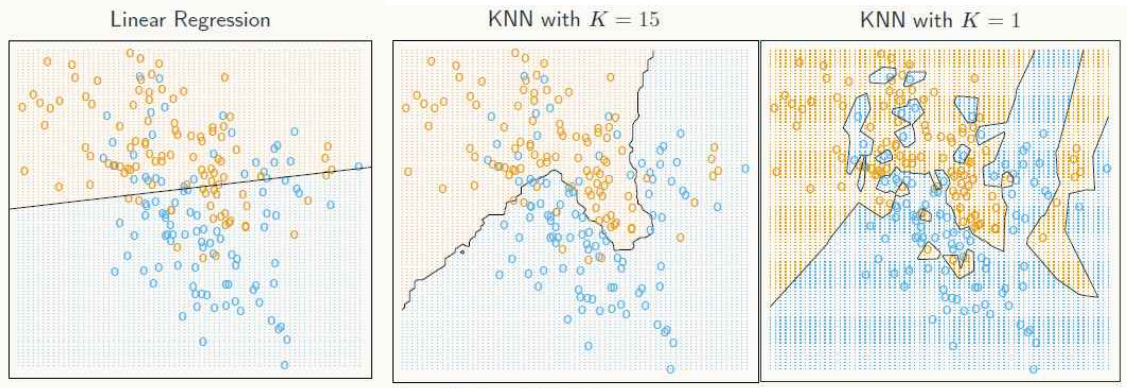
* 본 내용은 교육프로그램 과정의 강의자료를 요약한 것임(Reto Wuest 교수, 제너바대학교 정책과학·국제관계학과, 스위스)

기계학습에서 학습이란 경험을 지식으로 전환시키는 과정으로 기계라는 단어화 합성하여 자동적으로 학습되는 과정을 말하며, 컴퓨터를 프로그래밍해서 그들이 이용할 수 있는 입력으로부터 배울 수 있도록 하고 있다. 여기서 학습 알고리즘에 대한 입력은 훈련 데이터(경험)이고 학습 알고리즘의 출력은 지식으로, 예를 들어 예측, 패턴 감지 등

우리가 어떤 작업을 수행하는 데 사용할 수 있다. 다만 성공적인 학습 알고리즘은 유인적 추론으로 일반화할 수 있어야 한다. 반면 실패한 학습 알고리즘에서 나타나는 것은 학습 메커니즘을 편향시키는 선행 지식의 통합인 유인적 편견이 발생하고 사전 지식(또는 이전 가정)이 강할수록 추가 사례에서 배우는 것이 더 쉽지만 학습 유연성은 떨어진다. 결국 기계 학습 방법의 선정에 대한 논의에서 이러한 문제들을 다시 검토해야만 한다.

이러한 기계학습이 필요한 경우는 당면한 과제를 수행하기 위해 컴퓨터가 직접 프로그래밍하는 것보다 기계 학습에 의존하는 경우, 크고 복잡한 데이터 분석에서 복잡한 태스크가 발생한 경우(예를 들어 전문 지식으로부터 잘 정의된 프로그램을 추출할 만큼 충분히 이해되지 않는 과제), 스팸 탐지, 음성 인식 등 시간에 따라 변경되는 태스크가 발생한 경우(기계 학습 도구는 본래 상호작용하는 환경의 변화에 적응)를 들 수 있다.

일반적으로 파라데이터를 기초적으로 분석하여 활용할 수 있는 것은 조사과정에 발생하는 자료를 토대로 응답자를 어떻게 접촉해야 응답률을 높일 수 있는지 의사결정을 하는 과정으로 볼 수 있다. 이 경우 활용될 수 있는 기계학습 중 감독된 학습방법에 대해 알아보자. 첫 번째로 선형모형과 최소제곱 방법으로 선형 회귀분석에서 조건부 기댓값을 추정할 모형을 지정하고 잔차 합계가 최소화하도록 선택하는 방법이다. 아래의 예제 그림에서와 같이 여러 개의 관찰값이 잘못 분류분석됨을 알 수 있다. 두 번째로 최근방 K 이웃방법으로 유클리드 거리 척도로 가까운 거리에 있는 곳으로 분류하는 방법이다. 아래의 예제 그림에서와 같이 선형모형과 최소제곱 방법에 비해 잘못 분류되는 경우가 더 적지만 의사결정 경계가 더 불규칙적으로 분석된다. 비교하자면 선형모형과 최조제곱 방법은 예측이 안정적이지만 정확하지 않을 수 있는(낮은 분산과 높은 편향) 반면 최근방 K 이웃방법은 국부적으로 일정한 함수에 의한 잘 근사하여 예측은 정확하지만 불안정한 분석결과(낮은 편향과 높은 분산)를 도출할 수 있다.



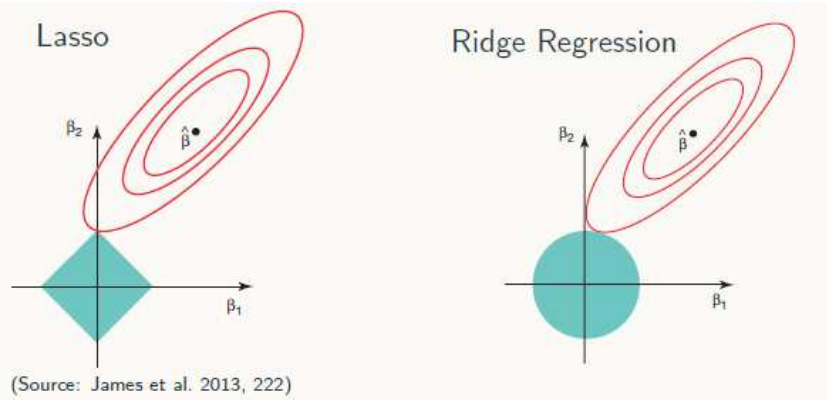
그렇다면 어떠한 방법으로 파라데이터를 분석하는 것이 최적일 것인가에 대한 부분은 모델 정확도 평가를 통해 알 수 있다. 다만 모든 학습 과제를 가장 잘 수행하는 보편적 학습 방법은 없으므로 방법의 일반화 성과에 관심을 두어야 한다. 학습 방법의 일반화 성과는 독립 시험 데이터에 대한 예측 정확도와 관련이 있고 일반화 성과 평가는 학습 과제를 위한 방법 선택을 안내하기 때문에 매우 중요한 부분이다. 예를 들어 선형모형과 최소제곱 방법의 경우 분산과 편향을 동시에 고려할 수 있는 평균제곱오차(MSE)를 이용할 수 있다.

파라데이터를 분석하는 기계학습 방법으로 능형 회귀분석을 이용한 수축방법을 활용할 수 있다. 수축방법은 회귀 모형의 계수 추정값을 0으로 축소하여 편중 증가 비용에 따른 편차 감소로 나타나고 분산 감소가 편향의 증가를 지배할 경우, 이는 시험 오차의 감소를 초래한다. 기본적으로 가장 잘 알려진 회귀 계수를 수축하는 방법은 능형 회귀분석과 라소 방법이다. 능형 회귀분석은 수축값이 증가함에 따라 능형 회귀의 유연성이 감소하여 편향은 증가하지만 분산은 감소한다. 반면 반응변수와 예측변수의 관계가 선형에 가까울 때 최소 제곱 추정값는 편향이 낮지만 분산이 높을 수 있다. 따라서 능형 회귀분석은 최소 제곱 추정값의 분산이 높은 상황에서 가장 효과적이다. 또 다른 의미로 분산이 높다는 의미는 다중공선성이 발생할 가능성이 높으며, 다중공선성을 해결하는 방법 중 하나로 능형 회귀분석을 사용하기도 한다.

또 다른 수축방법인 라소 방법은 능형 회귀분석이 항상 모형에서 p개의 예측변수를 포함한다는 단점을 해결한 방법이다. 능형 회귀분석

과 마찬가지로 라소 방법은 추정값을 0으로 줄이지만 수축값이 충분히 클 경우, 라소 방법은 일부 추정값이 정확히 0과 동일하도록 강제한다. 따라서 라소 방법은 희박한 모델을 산출하는 변수 선택을 수행한다. 능형 회귀분석에서와 같이 조정 매개변수 수축값은 중요한 역할을 한다. 수축값이 0일 경우 라소 방법의 추정값은 선형모형과 최소제곱 방법의 추정값과 동일하지만 수축값이 충분히 커지면, 라소 방법의 추정값는 정확히 0과 동일하게 설정되므로 수측값의 가치에 따라, 라소 방법은 어떠한 개수 예측변수를 포함하는 모형을 만들 수 있다.

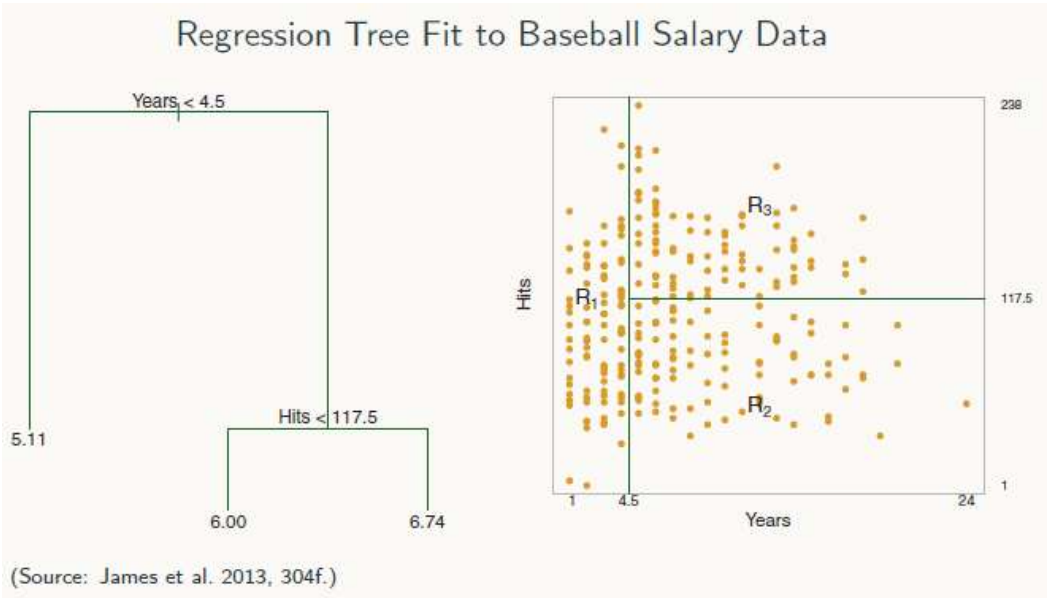
두 방법을 비교하자면 라소 방법은 능형 회귀분석보다 더 단순하여 해석 가능한 모델을 생산하는 장점이 있다. 그러나 어느 방법이 더 나은 예측 정확도로 이어지는지는 또 다른 문제이다. 라소 방법이나 능형 회귀분석 어느 것도 보편적으로 다른 방법모다 무조건 적으로 우수한 방법이 아니다. 라소 방법은 상대적으로 적은 수의 예측변수만이 상당한 계수를 가지고 있을 때 더 나은 성능을 발휘하는 경향이 있고 능형 회귀분석은 예측변수가 많을 때 더 잘 수행되는 경향이 있으며, 모두 거의 동일한 크기의 계수를 가지고 있다.



앞서 설명한 대로 파라데이터를 기초적으로 분석하여 활용할 수 있는 것으로 조사과정에 발생하는 자료를 토대로 응답자를 어떻게 접촉해야 응답률을 높일 수 있는지 의사결정을 하는 과정이고 이를 위해 기계학습방법 중 의사결정나무 분석을 실시할 수 있다. 의사결정나무 기반 방법은 예측변수 공간을 여러 개의 단순한 영역으로 계층화하거나 세분화하여 시험 관찰에 대한 예측을 하기 위해 그것이 속하는 지

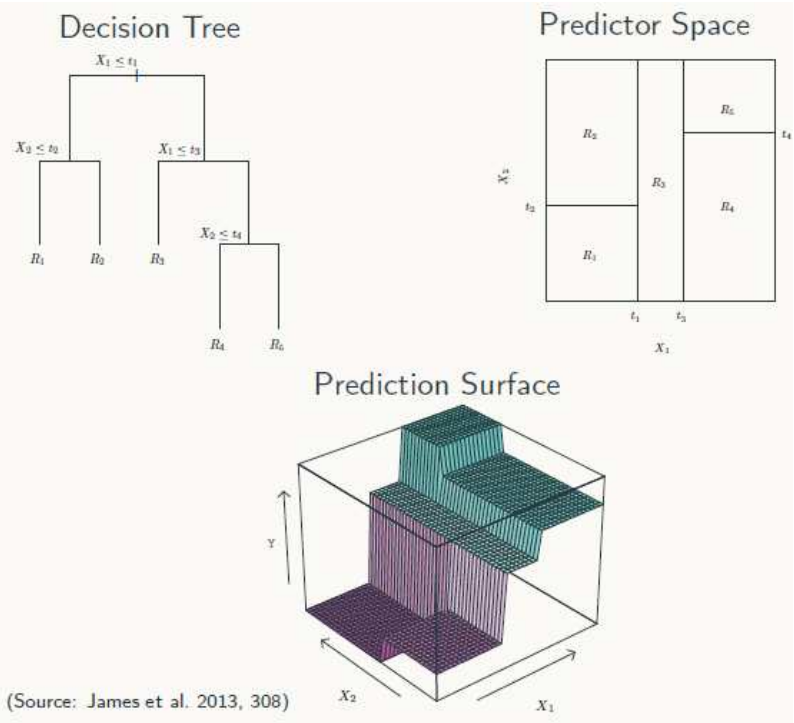
역에서 훈련 관찰의 평균 또는 최빈값을 사용한다. 예측변수 공간을 분할하는 데 사용되는 분할 규칙이 의사결정나무에 요약될 수 있기 때문에 이러한 방법을 의사결정나무 방법이라고 한다. 의사결정나무 방법은 회귀분석과 분류분석 모두에 적용할 수 있다.

예를 들어 회귀나무 방법으로 메이저리그에서 뛴 연수와 전년도 안타 수를 기준으로 야구 선수의 연봉을 예측하는 것이 목표하고 한다면 아래 그림과 같은 분석결과로 나타날 수 있다. 지역 R1, R2 및 R3은 나무의 단자 노드 또는 잎이라고 한다. 예측변수 공간이 분할되는 나무를 따라 있는 점은 내부 노드다(위에서 텍스트 연차 < 4.5 및 안타 수 < 117.5로 표시). 노드를 연결하는 나무 줄기를 분기라고 한다. 결과를 해석하자면 연봉을 결정하는 가장 중요한 요인은 경험이 적은 선수들이 경험이 많은 선수들보다 낮은 연봉을 받는다는 것이다. 경험이 적은 선수 중 안타 수는 연봉에 별 문제가 없으며, 경험이 많은 선수 중 안타가 많은 선수는 연봉이 높은 편이다. 여기서 노드를 분할하는 지표는 잔차제곱합이다.



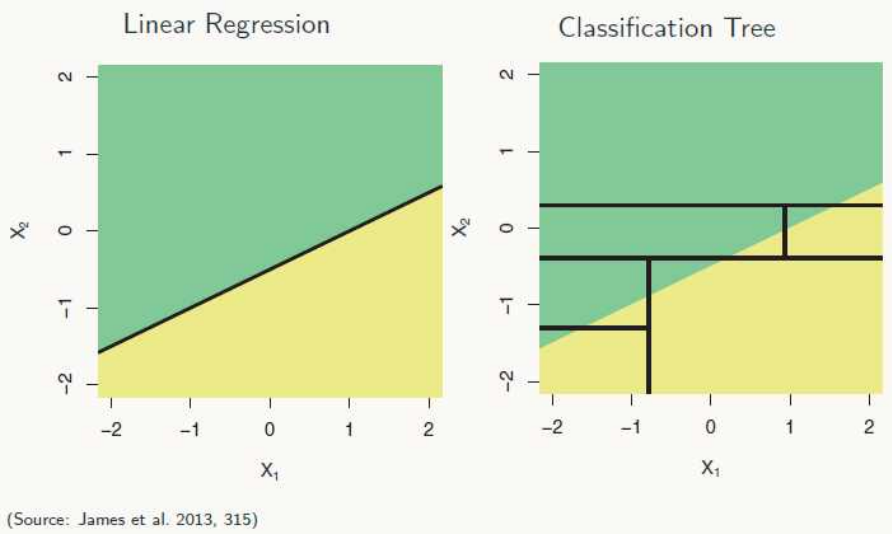
분류나무 방법은 양적 반응변수보다 질적 반응변수를 예측하는 데 이용된다는 점을 제외하고는 회귀나무와 매우 유사하다. 회귀나무의 경우, 관측값에 대한 예측 반응변수는 동일한 단자 노드에 속하는 훈

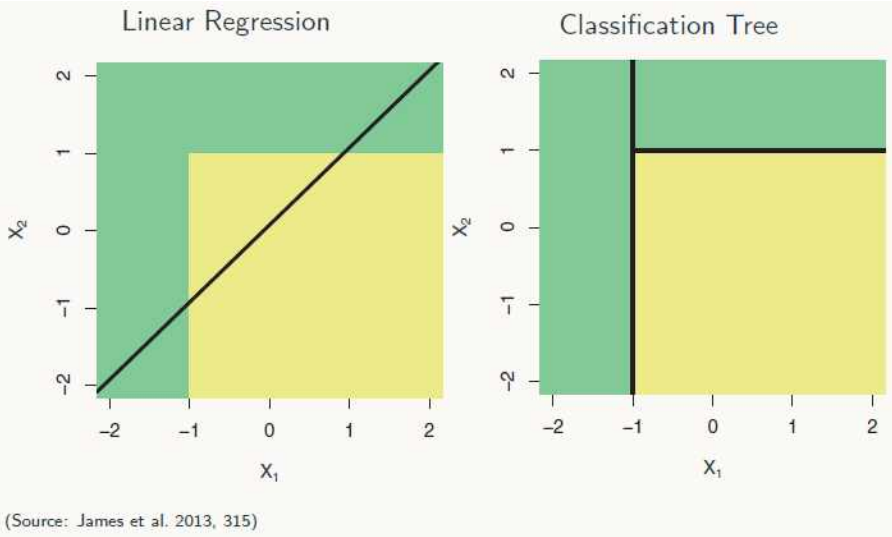
련 관측값의 평균 응답에 의해 주어진다. 분류나무의 경우, 관측값에 대한 예측 반응변수는 동일한 터미널 노드에 속하는 훈련 관측값 중에서 가장 흔히 발생하는 그룹이다. 분류나무 방법은 회귀나무와 마찬가지로 분류나무를 생성하기 위해 재귀 이진 분할을 사용한다. 단, 분류 설정에서는 잔차제곱합을 이진 분할의 기준으로 사용할 수 없고 대신 분류 오류율을 사용할 수 있다. 터미널 노드의 각 관측값을 가장 일반적으로 발생하는 그룹에 할당하므로 분류 오류율은 해당 터미널 노드의 훈련 관찰 비율이다. 그러나 분류 오류율은 나무를 생성하는데 충분히 민감하지 않은 것으로 밝혀져 지니지수와 엔트로피라는 두 가지 다른 방법을 사용하고 있다. 즉, 분류나무 구축에는 분류 오류율보다 더 민감한 지니 지수 또는 엔트로피를 사용하여 특정 분할의 품질을 평가한다. 그러나 노드 수를 결정하는 나무 가지 치기에서는 세 가지 방법 중 어느 것이라도 사용될 수 있지만 최종 트리의 예측 정확도가 목표라면 분류 오류율이 더 좋은 것으로 알려져 있다.



그렇다면 회귀분석을 근간으로 하는 선형방법과 회귀나무를 비교한다면 어떠한 방법이 파라데이터를 분석하기 위한 기계학습 방법으로

생각할 수 있을까? 예측변수와 반응변수의 관계가 선형모형에 의해 잘 근사되면 선형 회귀분석은 성형모형에 잘 근사되지 않는 회귀나무 방법을 능가할 것이다. 예측변수와 반응변수의 관계가 매우 비선형적이고 복잡한 경우 회귀나무 방법은 선형 회귀분석을 능가할 수 있다. 의사결정나무 기반 및 선형 모형의 상대적 성능은 교차검증방법을 사용하여 실험 오류를 추정한다면 그 차이를 평가할 수 있다.





의사결정나무 방법의 장점을 살펴보면 선형모형 방법 보다 설명하기 매우 쉽다는 점과 고전적인 회귀분석 및 분류분석 접근법보다 인간의 의사결정을 더 밀접하게 반영할 수 있다는 점이다. 의사결정나무 방법

의 나무는 그래픽으로 표시할 수 있으며 비전문가(특히 나무가 작을 경우)에서도 해석하기 쉽다. 또한 더미 변수를 생성할 필요 없이 정성적 예측변수를 쉽게 처리할 수 있다.

반면 의사결정나무 방법의 단점을 살펴보면 일반적으로 다른 감독된 학습 방법(예: 수축 방법)과 같은 수준의 예측 정확도를 가지고 있지 않다는 점과 데이터가 조금만 변경되면 최종 추정 나무에 큰 변화가 생길 수도 있어 견고하지 않을 수 있는 방법이라는 점이다. 배킹, 무작위 숲, 부스팅 등 많은 의사결정나무 방법을 사용함으로써 결과의 예측 성능을 실질적으로 향상시킬 수 있다 여기서 배깅, 랜덤 숲, 부스팅 등의 방법은 의사결정나무를 빌딩 블록으로 사용하여 보다 강력한 예측 모형을 구축할 수 있게 한다.

의사결정나무 방법은 기본적으로 높은 분산을 가지므로 훈련 데이터의 작은 변화가 상당히 다른 결과로 나타날 수 있다. 우리가 원하는 분석방법은 분산이 작은 방법이지만 별개의 자료에 반복적으로 방법을 적용하면 분산 수준의 결과가 유사하게 나타난다. 일반적으로 분산을 줄이는 방법은 부스팅, 배깅 방법이다. 특히 배깅은 일반적으로 하나의 의사결정나무 분석결과보다 더 나은 예측 정확도를 가지고 있다. 이는 해석가능성을 줄이면서 발생한다. 즉, 모형을 단일 의사결정나무로 표시하는 것은 더 이상 불가능하고 어떤 변수가 가장 중요한지 더 이상 명확하지 않게 된다. 따라서 회귀나무의 잔차제곱합이나 분류나무의 지니지수(분류)를 사용하여 각 예측 변수의 중요도에 대한 전체 요약을 계산하는 것이 유용할 수 있다. 회귀나무의 경우 모든 나무에서 평균적으로 주어진 예측변수에 대한 분할로 인해 잔체제곱합이 감소하는 총량을 기록할 수 있다. 반면 분류나무의 경우 모든 나무에서 평균적으로 주어진 예측변수에 대한 분할로 인해 지니지수가 감소하는 총량을 기록할 수 있다. 두 경우 모두 큰 값은 중요한 예측변수를 나타낸다.

추가적으로 무작위 숲 방법을 살펴보면 개발된 숲은 나무보다 더 나은 정보을 제공한다. 나무를 장식하는 작은 줄기을 포함하고 배깅과 마찬가지로 부스킹 표본에 많은 의사결정나무를 만들어 낸다. 나무 형성 과정의 각 분할에서 무작위 표본을 분할 대상자로 간주하고 각 분

할에서 새로운 예측변수의 표본을 추출하게 된다. 따라서 의사결정나무의 각 분할에서 알고리즘은 이용 가능한 예측변수의 과반수를 고려하는 것조차 허용되지 않는다. 결과적으로 모든 가지들이 서로 상당히 비슷해 보일 것이므로 의사결정나무들로부터의 예측은 매우 상관관계가 높을 것이다. 고도로 높은 상관관계의 양을 평균화하면 낮은 상관관계의 양의 평균화보다 분산이 더 작아진다. 따라서 배깅은 의사결정나무 하나에 대한 분산의 상당한 감소로 연결되지 않는다. 무작위 숲에서의 의사결정나무들의 평균은 덜 가변적이므로 더 신뢰할 수 있게 된다. 예측변수 부분집합 크기와 예측변수의 수가 같을 경우 배깅과 무작위 숲 방법은 동일한 결과를 가지게 된다.

마지막으로 부스팅 방법은 배깅과 마찬가지로 회귀나무 또는 분류나무를 위한 많은 기계학습 방법에 적용할 수 있는 일반적인 접근 방식이다. 배깅이 원래 훈련 세트에서 여러 부스트랩 훈련 세트를 생성하고 각 부스트랩 훈련 세트에 별도의 의사결정나무를 맞춘 다음, 모든 의사결정나무를 결합하여 단일 예측을 생성하게 된다. 이것은 각 의사결정나무가 다른 의사결정나무들과 독립적으로 부스트랩 표본 위에 생성되었다는 것을 의미한다.

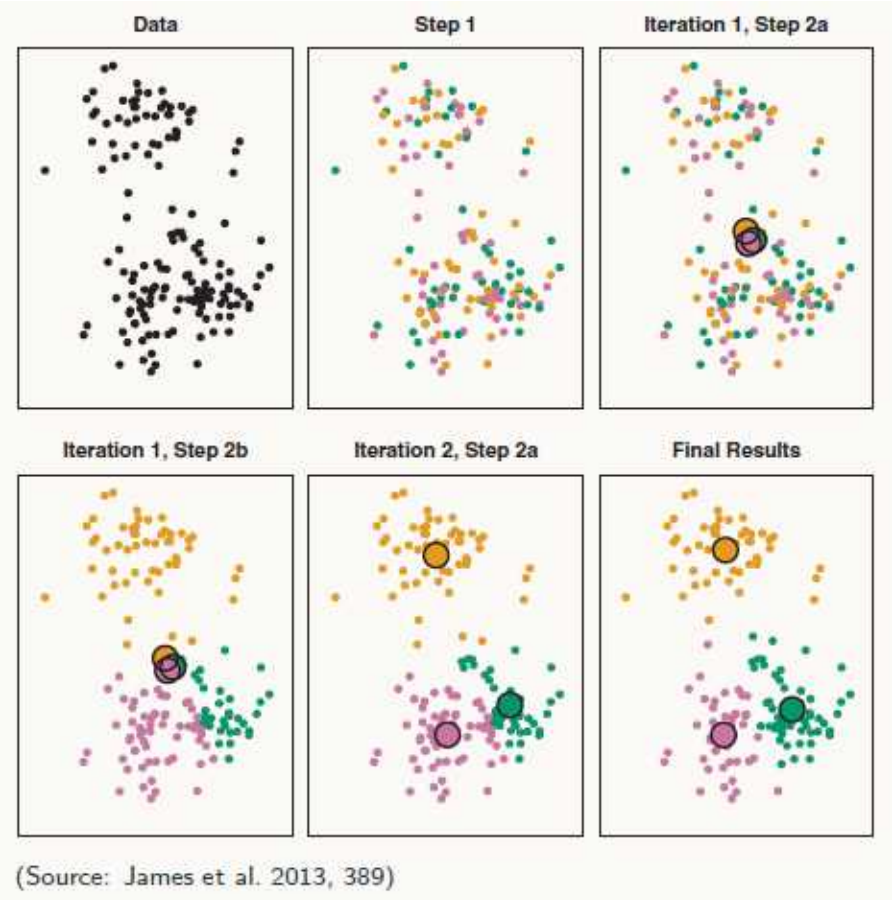
파라데이터 분석을 위한 기계학습 방법에서는 관리되지 않는 학습이 존재한다. 관리되지 않는 학습은 관리되는 학습 기법을 적용하기 전에 데이터 시각화 또는 데이터 사전 처리에 사용하는 주성분분석 방법과 데이터에서 알 수 없는 부분군을 검색하는 데 사용되는 군집분석 방법이 있다. 관리되는 학습에서는 명확한 목표(추정값)을 알고 있고 교차평가, 유효성 검사 등 결과의 품질을 평가하는 방법도 알고 있다. 따라서 관리되는 학습에서는 모형이 적합성에 사용되지 않는 관찰에 대해 얼마나 잘 예측하는지 평가함으로써 우리의 작업을 확인할 수 있다. 반면 관리되지 않는 학습은 예측값에 대한 명확한 목표가 없으므로 대부분 관리되는 학습보다 더 어렵고 보다 주관적이며, 결과를 평가하는 것은 더 어렵다. 즉, 감독되지 않는 학습에서는 명확한 목표값을 모르기 때문에 결과에 대한 평가를 확인할 수 없다는 것이다.

관리되지 않는 기계학습 중 주성분분석의 목표는 가능한 많은 정보를 캡처하는 데이터의 저차원적 표현을 찾는 것이다. 예를 들어 대부

분의 정보를 캡처하는 데이터의 2차원 표현을 찾을 수 있다면, 이 2차원 공간에 관측값를 표시할 수 있다. 다시 말해 주성분분석은 데이터 변동을 가능한 한 많이 포함하는 데이터 세트의 저차원 표현을 찾는 방법이다. 주성분분석의 이면에 있는 아이디어는 다음과 같다. 각 n개의 관측값은 p차원 공간에서 존재하지만 이 모든 p차원이 똑같이 관심이 있는 것이 아니다. 주성분분석은 가능한 한 관심이 있는 소수의 차원을 찾는다는 것이고 여기서 “관심”은 차원 속의 관측값의 규모에 의해 결정되며, 주성분분석에서 결정되는 차수는 p의 선형결합이다. 주성분분석에서 얻은 결과는 변수의 척도에 따라 달라진다. 척도가 조정되지 않은 변수에 대해 주성분분석을 수행하는 경우, 예를 들어 첫 번째 주요 구성 요소 로딩 벡터는 특정 변수에 대해 매우 큰 부하를 가하게 된다면 그 변수의 분산은 작을 것이며, 따라서 첫 번째 주성분 하중 벡터는 그 변수에 대한 값이 매우 작을 것이다. 일반적으로 주성분분석을 수행하기 전에 각 변수의 표준 편차를 1로 조정하여 주요 구성요소가 척도 선택에 의존하지 않도록 한다. 그러나 변수가 같은 단위로 측정될 경우 변수의 크기를 조정하지 않도록 선택할 수 있다.

두 번째로 관리되지 않는 기계학습 중 군집분석은 데이터 세트에서 부분군 또는 군집을 찾기 위한 일련의 기술을 말한다. 군집분석의 목표는 데이터 집합의 관측값을 별개의 그룹으로 분할하여 각 그룹 내의 관측값는 서로 유사한 반면, 다른 그룹의 관측값는 서로 다른 것이다. 군집분석은 데이터 세트를 기반으로 구조(분점 군집)를 발견하려고 하기 때문에 관리되지 않는 기계학습에 속한다. 군집분석과 주성분분석 모두 적은 수의 요약을 통해 데이터를 단순화하려고 하지만 그 메커니즘은 다르다. 주성분분석은 분포의 큰 부분을 설명하는 관측값의 낮은 차원 표현을 찾으려고 하는 반면 군집분석은 관착값 중에서 균일한 하위 그룹을 찾으려고 한다. 파라데이터 분석에 활용할 수 있는 기계학습 중 군집분석은 다양하게 존재하는데 K-평균 군집분석과 계층적 군집분석이 가장 잘 알려진 두 가지 접근방식이다. K-평균 군집분석에서는 관측값을 미리 지정된 수의 군집으로 분할하는 방법을 모색한다. 계층적 군집분석에서는 원하는 군집 수를 미리 알 수 없다. 관측값 중 부분군을 식별하기 위해 형상을 기반으로 관측값을 군집화할 수 있으

며, 관측값을 기반으로 형상을 군집화하여 형상 간의 부분군을 발견할 수 있다. K-평균 군집분석의 알고리즘은 각 관측값에 1부터 K까지의 숫자를 무작위로 할당하고 이것은 관측값에 대한 초기 군집 할당 역할을 한다. 먼저 각 K 군집에 대해 군집 중심점을 계산하는데 k번째 군집 중심은 k번째 군집의 관측값에 대한 p형상 평균의 벡터이다. 그 다음은 중심점이 가장 가까운 군집에 각 관측값를 할당한다(가장 가까운 곳은 유클리드 거리, 즉 두 점 사이의 "직선" 거리를 사용하여 정의된다). 이 두 과정을 군집 할당이 변경되지 않을 때까지 반복하여 진행한다.



(Source: James et al. 2013, 389)

K-평균 군집분석의 잠재적 단점은 K 군집 수를 미리 지정해야 한다는 것이다. 계층적 군집분석은 군집 수에 대한 지정이 필요없는 대안적 접근법이다. 계층적 군집분석은 덴드로그램이라 불리는 관측값의 나무 기반 표현으로 분석하게 된다. 덴드로그램의 각 잎은 관찰값을

나타내고 나무 위로 올라갈 때 잎을 나뭇가지로 다른 가지에 연결한다. 나무 하단의 잎이 서로 비슷한 관측값인 반면 상단에 가깝게 잎으로 표현된 관측값은 다른 의미를 가지고 있다. 관측값을 포함하는 가지가 먼저 병합되는 수직축의 위치를 기준으로 두 관측값의 유사성을 비교할 수 있는 반면 수평축을 따라 근접해 있는 두 관측값의 유사성을 비교할 수 없다. 덴드로그램에 근거하여 군집을 식별하기 위해 덴드로그램을 가로로 절단하여 절단 아래의 관측값 집합은 군집으로 해석할 수 있다. 하나의 덴드로그램으로 다양한 수의 군집을 얻기 위해 사용될 수 있으며 덴드로그램에 대한 절단점의 높이는 K-평균 군집분석에서의 K와 동일한 역할인 군집 수를 제어하게 된다.

계층적 군집분석과 K-평균 군집분석을 비교하면 계층적 군집분석은 주어진 높이에서 절단으로 얻은 군집이 더 큰 높이에서 절단하여 얻은 군집 내에 중첩되기 때문에 계층적 군집분석라고 불린다. 그러나 이러한 계층 구조의 가정은 특정 데이터 집합에 비현실적일 수 있다. 계층적 군집분석(덴드로그램 생성방법) 알고리즘을 살펴보면 먼저 각 관측값의 쌍 사이에 서로 다른 측정값을 정의한다(대부분 유클리드 거리가 사용된다). 덴드로그램 하단에서 시작하여 각 n개 관측치는 자체 군집으로 처리되고 서로 가장 유사한 두 개의 군집이 병합되어 현재 n-1개의 군집이 있게 된다. 다음으로 서로 가장 비슷한 두 개의 군집이 다시 융합되어 n-2개 군집이 남게 된다. 알고리즘은 모든 관측치가 하나의 단일 군집에 속할 때까지 진행한다. 여기서 군집과 개별 관측값, 군집과 군집 등의 측정값을 계산하는 방법이 다양하게 존재하는데 싱글, 평균, 중심점 등의 방법을 사용한다. 싱글 방법은 최소 군집 내 차이점 군집 A의 관측값과 군집 B의 관측값 사이의 모든 쌍별 차이를 계산하고 가장 작은 값을 사용하는 방법으로 단일 관측값이 한 번에 병합되는 확장된 후행 군집을 초래할 수 있다. 평균 방법은 군집 A의 관측값과 군집 B의 관측값 사이의 모든 쌍별 차이를 계산하고 차이의 평균을 이용하는 방법이다. 중심값 방법은 군집 A의 중심점(길이 p의 평균 벡터)과 클러스터 B의 중심점 사이의 차이점을 이용하는 방법이다.

이러한 방법 모두 유클리드 거리를 차이 측정으로 사용하지만 때로는 다른 방법을 선호할 수도 있다. 그 대안으로는 두 개의 관측값이

높은 상관관계가 있는 경우 유사한 것으로 간주하는 상관관계 기반 거리를 사용할 수 있다. 상관관계에 기초한 거리는 관측 프로파일의 크기보다는 모양에 초점을 맞춘게 된다.

실제로 군집분석을 수행하기 위해서는 몇 가지 결정이 내려져야 한다. 관찰 또는 형상을 먼저 표준화할 것인지, 계층적 군집분석의 경우 측정값을 무엇으로 사용할 것인지, 어떤 유형의 연결방법을 사용할 것인지, 군집를 얻기 위해 덴드로그램의 어디를 절단할 것인지, K-평균 군집의 경우 데이터에서 몇 개의 군집으로 수행할 것인지 등이다.

3. 고품질 웹조사의 구현

* 본 내용은 교육프로그램 과정의 강의자료를 요약한 것임(Pamela Campanelli 박사, 조사방법론 컨설턴트, 영국)

어떤 종류의 웹 설문조사를 계획하고 있는가에 대한 문제는 설문지 유형과 응답자 응답방식(조사모드)을 먼저 결정해야 한다. 특히 응답자 응답방식은 테블릿, 컴퓨터, 종이, 스마트폰(그래픽 포함, 그래픽 없음)으로 구분되며, 응답하는 방법으로 키보드, 마우스, 터치스크린 등으로 나눌 수 있다. 대부분 웹 설문 조사는 컴퓨터 기반의 조사방식으로 진행된다고 볼 수 있다.

웹 설문 조사는 시각적 측면이 매우 중요한데 이는 조사표 설계와 매우 밀접한 관계를 가지고 있다. 특히 시각적 배치가 매우 중요하며, 시각적 배치가 잘 된 응답자 친화적 설문지는 자기기입식 설문지에서 무응답 감소 효과를 낼 수 있는 매우 중요한 방법이다. 응답할 가능성이 낮은 집단에서도 설문지의 효과는 나타나고 있는데, 질문의 길이, 어려운 질문의 표현 방식을 통해 항목 무응답의 감소와 데이터 품질의 향상을 가져올 수 있다. 많은 응답자들이 설문지의 전체 내용을 사려깊게 읽지 않으며, 응답자들은 읽어야 할 것과 안전하게 무시될 수 있는 것에 대한 단서를 질문의 배치로부터 얻고 어떤 응답자들은 질문이 자주 나오는 많은 단어들을 생략하는 특성을 가지고 있다. 시각적 배치 요소를 가장 잘 사용하는 방법은 단순성, 규칙성 및 대칭성을 유지하여 대응하고 응답 작업의 용이성이 높은 설문이며, 대부분 연구자들

은 설문지 설계하는 것은 누적된 학습 경험이라고 말한다. 또한 중요한 것은 스타일, 레이아웃, 작업에서 일관성을 유지하는 것으로 전반적으로 동일한 유형의 정보 제공하여 응답 작업 간소화를 통해 설문지 응답자 지원할 수 있다. 시각적 정보를 잘 주는 것이 그 정보를 잘 암시하게 하고 그것을 처리하고 기억하기 더 쉽게 된다. 따라서 각 페이지의 동일한 위치에서 시작하는 웹 질문을 삽입하여 응답 공간을 유지하는 것이 매우 중요하다. 이는 자연스러운 눈 움직임을 따라가기 위해 페이지를 배치하는 것과 같은 효과를 나타낸다.

인터넷 조사에서 나오는 상대적으로 저렴한 데이터의 확산은 웹 표본에 대한 유효한 추론 관심이 증대하고 있는 실정이다. 현재 많이 이용되는 웹 표본은 비확률표본의 일환으로 추정에 대한 정도확보가 어렵다는 단점을 가지고 있다. 따라서 기존의 전통적인 표본의 정보를 이용하여 웹 표본의 추정을 위한 연구가 필요한 시점이다. 웹 표본과 함께 기존의 조사된 표본을 이용하여 웹 표본의 추정을 위한 방안을 제시하였는데 첫 번째로, 두 표본 설정에서 웹 반응성향모형의 모수를 추정하기 위한 "암묵적(implicit)" 로지스틱을 제안하고 제안된 추정량은 웹 표본단위에 대한 지정된 무작위 표본 포함확률의 형태로 추가정보에 의존한다. 두 번째로 추정된 웹반응성향에 근거하여 모집단 평균 추정량을 제안하였다.

<제안된 방법>

○ 가중 로지스틱 회귀(WLR; weight logistic regression) 추정량, with 모수 ϕ

Z-성향: $p_z(x_i) = P(Z=1 | i \in S, x_i)$

$\hat{p}_\delta^{WLR}(x_i) = p_z(x_i, \hat{\phi}) / (1 - p_z(x_i, \hat{\phi}))$

○ 암묵적 로지스틱 회귀(ILR; implicit logistic regression) 추정량

표본지시함수 Z_i, 웹 반응 지시함수: δ_i

$p_Z(x_i) = p_\delta(x_i) / (\pi_{ri}/\pi_{wi}) + p_\delta(x_i))$

ILR은 각 단위에 대한 선택확률(π_{ri}, π_{wi})을 알고 있어야 함

웹 표본단위의 선택확률 π_{ri}를 위해서는 응답자에게 추가정보를 수집해야 하므로 웹 조사 관리자 및 응답자에 부담감 존재

○ 평균 암묵적 로지스틱 회귀(AILR; average implicit logistic regression) 추정량

웹 표본단위에 평균 설계 가중값을 할당하여 산출

제안한 방법에 대해 모의실험을 통해 ILR, AILR, WLR에 의한 추정을 실시하였다. 모형의 모수와 표준오차는 웹 표본 선택 매커니즘에 따라 ILR은 불편추정량이고 AILR과 WLR은 직접추출로 대략적으로 2상 추출이 되도록 설계하였다. WLR은 2상 추출을 설명할 옵션이 없어 편향된 점추정량이 산출되며, AILR 표준오차 추정값이 일반적으로 정확하다면, WLR에 의한 추정값은 점추 정값의 해당 표준편차보다 일관되게 낮게 된다. 모형 모수의 점 추정값에 대해 3가지 추정값 모두 비편향, 효율에 따라 비교하고 ILR과 AILR에 의한 추정된 모형 모수의 표준편차가 가까운 반면 WLR에 의한 추정값은 덜 효율적이다.

Table 1: Simulation design

h	x_5	x_6	N_h	m_h
1	0	0	150,000	0.1
2	0	1	100,000	0.3
3	1	0	50,000	0.5
4	1	1	15,000	0.7

<모의실험 조건>

Model		$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{\phi}_4$
Two-phase	$\hat{\phi}$	-1.0	1.0	-0.5	0.25	0.1
ILR	$\bar{\phi}$, covg	-1.00 , 1.00	1.01 , .96	-.50 , .95	.25 , .95	.10 , .95
	SD, $\widehat{SE}$	0.035 , .076	.091 , .096	.084 , .085	.081 , .081	.080 , .078
AILR		-5.84 , .00	.64 , .00	-.32 , .08	.16 , .62	.06 , 0.88
		0.031 , .051	.053 , .054	.052 , .053	.051 , .052	.050 , .050
WLR		-5.93 , .00	.59 , .00	-.30 , .02	.15 , .29	.06 , .61
		0.033 , .037	.065 , .033	.064 , .034	.061 , .034	.061 , .033
Direct	$\hat{\phi}$	-6.2	1.0	-0.5	0.25	0.1
ILR	$\bar{\phi}$, covg	-6.20 , .98	1.00 , .95	-.50 , .94	.25 , .95	.10 , .95
	SD, $\widehat{SE}$	.050 , .060	.059 , .059	.059 , .057	.056 , .056	.053 , .054
AILR		-6.10 , .61	1.00 , .95	-0.50 , .94	.25 , .95	.10 , .96
		.052 , .058	.057 , .058	.057, .056	.055 , .054	.051 , .053
WLR		-6.22 , .77	1.02 , .50	-.51 , .56	.26 , .58	.10 , .59
		.071 , .045	.097 , .033	.084 , .033	.078 , .033	.074 , .032

<이상추출, 직접추출과 모수 ϕ변화에 따라 모형별 웹 반응모형의 추정량 $\hat{\phi}$의 점추정량, 95% 신뢰구간(covg), 추정값의 표준편차(SD), 추정된 표준오차($\widehat{SE}$)의 결과>

Model		Pop	D_1	D_2
		Two-phase selection		
Direct	RB	0.4	0.4	0.16
ILR	RB , covg	0.0 , .95	0.0 , .96	-.01 , .92
	$100 \times SD$, $\widehat{SE}$	1.69 , 1.75	3.07 , 3.08	2.98 , 3.01
AILR		.01 , .91	0.0 , 0.92	-.02 , 0.86
		1.71 , 1.49	2.96 , 2.62	2.82 , 2.29
WLR		.05 , .82	.04 , .90	0.0 , .86
		2.06 , 1.66	3.28 , 2.92	2.92 , 2.39
		Direct selection		
Direct	RB	0.73	0.72	0.19
ILR	RB , covg	0.0 , .96	0.0 , .95	0.0 , .88
	$100 \times SD$, $\widehat{SE}$	2.04 , 2.06	3.65 , 3.61	3.90 , 3.93
AILR		0.0 , .93	0.0 , .92	-.01 , .86
		1.99 , 1.81	3.46 , 3.18	3.84 , 3.51
WLR		-.01 , .84	-.01 , .91	-.04 , .88
		2.98 , 2.17	4.30 , 3.83	3.98 , 4.30

<모집단 및 영역별 평균 추론: 점추정값의 상대편향(RB), 유한 모집단 평균의 범위에 대한 95% 신뢰구간, 점추정값의 표준편차, 추정된 표준오차 결과>

ILR 추정값은 웹 표본추출의 두 메커니즘에 따라 모집단과 영역별 평균 및 표준편차에 대해 불편성을 만족하고 AILR 점추정값은 사실상 편향이 없으나, 표준오차의 추정값은 과소추정되고 있다. WLR 점추정값의 일부는 편향되었으며, 표준오차의 추정값은 과소추정되고 있다. 적용을 위해서는 제약사항들이 있으나, 제안된 방법을 통해 웹 표본의 추정의 효율성을 볼 수 있다. 모집단을 잘 설명할 수 있고 이용 가능하며 대표성 있는 확률표본조사 데이터가 있어야 한다. 웹 조사의 입력정보를 수집하고 추정에 사용해야 하기 때문에 웹 조사계획 수립의 중요성이 강조되고 있다. 또한, 제안되는 방법은 선택확률과 웹 응답자의 하위 모집단의 전체 크기 등의 추가 입력 정보가 필요하고 기존의 가중 로지스틱 회귀추정량 대신 암묵적 로지스틱 회귀추정량과 평균 암묵적 로지스틱 회귀추정량을 제안하였다. 웹 표본 선택의 모든 경우 비편향 점추정량과 분산추정량은 암묵적 로지스틱 회귀추정량을 통해서 구할 수 있다. 평균 암묵적 로지스틱회귀추정량은 불편추정량을 구할 수 있지만 분산추정량은 일관적으로 과소추정되며, 가중 로지스틱 회귀추정량은 신뢰성이 가장 낮은 편이다. 향후 암묵적 로지스틱 회귀추정량과 웹 응답 결합모형, 결과변수에 대한 최적의 모형선택기법 등 추가적인 연구가 필요하다.

품질관리 수단으로 조사과정의 파라데이터를 이용할 수 있는데 문제가 되는 인터뷰와 대응 감지에 이용한다. 특히, 특정 항목 및 질문모듈 또는 전체에 대한 응답시간은 속도계, 트리거 또는 잠재적으로 위조된 인터뷰와 같은 성능 이상값을 식별하기 위한 정보로 간주된다. 하지만, 현재 조사관련사업에서 이용하는 품질보증기법은 경험을 바탕으로 확립된 것으로 판단된다. RT 측정에 대한 컷오프(cut-off) 임계값을 결정하는 연구와 반응시간 이상값을 검출하는데 미치는 영향들에 대한 연구는 진행되지 않고 있다. 응답시간 이상값을 식별하기 위한 이론적 방법 적용하고 검토할 필요가 있는데 이상값 식별을 위한 다양한 기법을 평가할 수 있다. 그 중 Tukey(1977)가 제안한 방법에 따라 정규화된 RT분포에 응답시간 이상값을 식별을 위한 IQR 컷오프 임계값 결정을 제안하였다. 제안한 방법이 RT 이상값을 식별하는데 보다 강력한 기법임을 확인할 수 있다.

Ⅲ. 유럽조사연구학회 2019 컨퍼런스

1. 컨퍼런스 개요

제8회 유럽조사연구학회 2019 컨퍼런스는 1일간의 반일 단기과정(half day short course)과 4일간에 걸쳐 약 220개의 세션에서 800여개의 주제 및 40여개의 포스터가 발표되었다. 반일 단기과정은 조사방법론의 본질(the essentials of survey methodology), 조사방법론에서의 행정자료(administrative records for survey methodology), R을 이용한 조사자료 시각화(survey data visualisation in R), 오픈형 질문에서의 텍스트마이닝의 응용(text mining applied to open-ended questions), 베이지안통계의 소개(introduction to bayesian statistics), 스마트폰 : 조사부터 센서까지(smartphones: From surveys to sensors), 표본조사에서의 결측값의 처리(handling missing data in sample surveys), 휴대폰 문자를 이용한 조사(text message surveys), 조사에서 기계학습을 통한 연계(integrating machine learning in surveys), 종단자료 분석의 역학(the mechanics of longitudinal data analysis) 등이 개설되었다. 본 훈련에서는 조사방법론의 본질과 스마트폰 : 조사부터 센서까지의 2개의 과정을 수강하여 진행하였다.

주요한 세션으로는 응답자 부담 측정 및 감소방안에 대한 새로운 시각 탐색, 바람직하지 못한 조사원의 행동 및 허위·부실 조사의 예방, 탐지, 개입 등의 해결방안, 응답자 접촉전략 설계, 구현, 평가 등의 방법론 및 고려사항, 종단연구에서 적응적 조사설계 사용에 대한 새로운 개발, 변화하는 조사환경에서 조사의 유용성 테스트, 면접조사에서 조사원 효과 탐색 및 관리방안, 대규모 조사에서 현장 모니터링 프레임워크, 혼합조사모드와 다출처자료 환경에서 조사원의 미래, 대면면접조사, 전화조사, 모바일 문자조사의 조사원-응답자 간의 상호작용, 조사에서 민감한 질문의 범위, 이론, 방법 등 적용방안, 조사 최적화를 위한 반응적·적응적 조사설계 등으로 구성되었다.

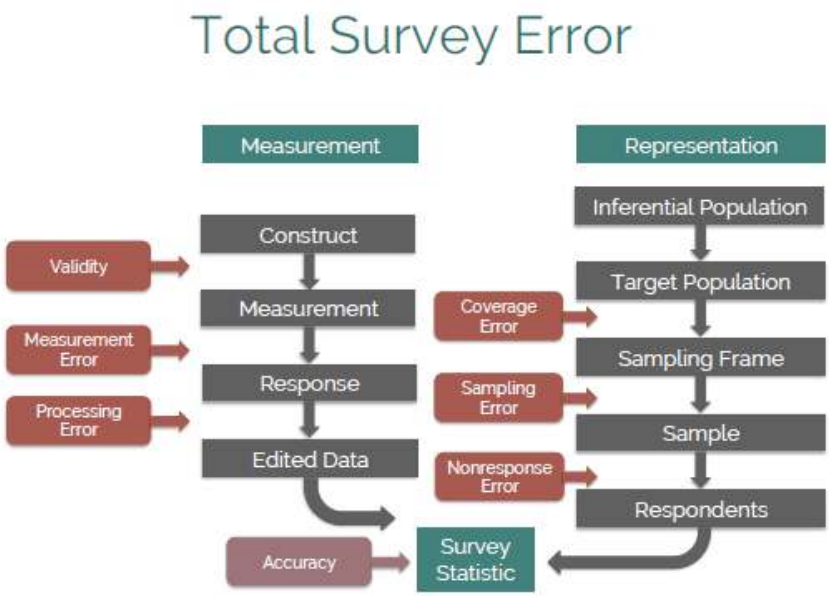
2. 단기과정

□ 조사방법론의 본질

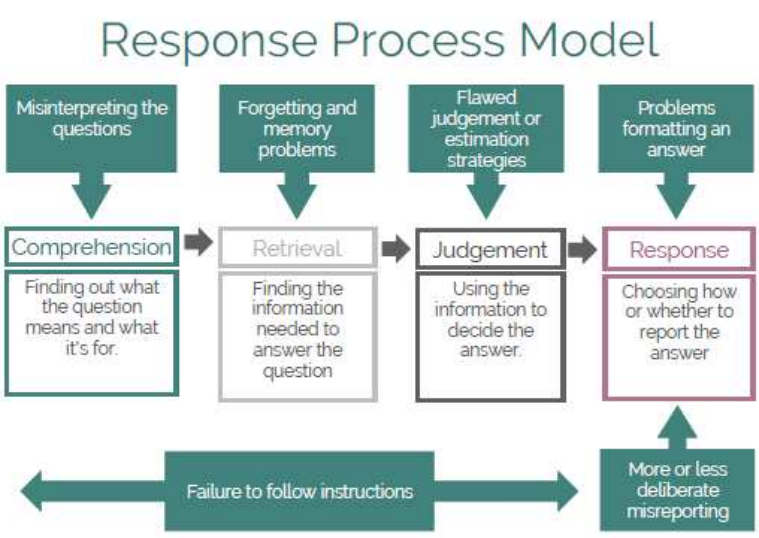
* 본 내용은 단기과정의 강의자료를 요약한 것임(Caroline Roberts 교수, 라우산 대학교, 스위스)

조사방법론에서 활용되는 이론으로 총조사오차 패러다임(Biemer, 2010), 인지적 측면에 대한 응답과정모델(Tourangeau, Rips and Rasinski, 2010), 활용성 이론(Groves, Singer and Corning, 2000), 사회교환이론(Dillman et al., 2000) 등이 있다

총조사오차 패러다임은 전체 조사진행 과정을 세분화하고, 각 단계에서 조사결과의 정확성에 영향을 미치는 원인에 관심을 두는 이론이다.



특히, 측정의 정확성을 높이기 위해 조사표 설계가 중요한데, 조사표를 설계하기 위해 질문에 응답하는 과정과 오류 발생 원인을 이해하는 것이 필요하다. 응답과정 모델에 따르면, 응답과정은 이해-인출-판단-응답으로 이루어지며, 각 단계에서 발생 가능한 오류를 이해하고 확인하는 것이 중요하다.



응답자는 응답을 위한 노력을 줄이려는 경향이 있어 임의로 응답하거나 쉬운 응답을 선택할 수 있는데, 이 때 응답자의 동기, 응답능력, 과제 난이도 등의 요인이 영향을 미칠 수 있다. 응답과정을 평가하는 방법으로 다양한 인지실험 기법(thinking aloud, probes, paraphrasing, confidence ratings, definitions, debriefing)이 있다. 또한 최근 다양한 조사도구가 활용되고 있어 조사도구를 사용하는 응답자 특성과 관련하여 '모드효과'에 대한 고려가 필요하다.

□ 스마트폰 : 조사부터 센서까지

* 본 내용은 단기과정의 강의자료를 요약한 것임(V. Yoepoel 교수 & P. Lugtig 교수, 우트렛 대학교, 네덜란드)

많은 국가에서 스마트폰 사용자가 많아지고 있으며, 미국의 경우 18~29세의 94%가 스마트폰을 보유하고 있다. 따라서 현재 통계조사는 웹조사에서 모바일조사로의 이동하고 있는 실정이다. 혼합기기 조사를 보면 웹조사는 현재 다양한 기기를 이용해 조사되고 있으며 웹 응답자의 5~30%가 모바일을 이용하고 있다. 조사 길이 측면에서는 응답자들은 긴 모바일 조사를 원하지 않으며, 짧은 조사(10분 이내) 또는 조사 분할을 선호하고 있다. 다만 모바일 조사표 설계 측면에서는 응답자들이 다양한 기기로 접속할 수 있어 다양한 스크린 크기에 최적화되어야 한다.

65% of US Smartphone users would not be willing to spend more than 15 minutes completing surveys

MAXIMUM TIME DOING SURVEYS:	COMPUTER	TABLET	SMARTPHONE
5 minutes or less	2%	9%	27%
10 minutes or less	9%	24%	45%
15 minutes or less	19%	42%	65%
20 minutes or less	34%	65%	73%
25 minutes or less	42%	71%	77%
30 minutes or less	65%	81%	85%

RDD, Email, QR code, Text-SMS, App, 위치기반 표지(GPS chip) 등 응답자들은 다양한 방식으로 모바일조사에 접근이 가능하다. 브라우저(Browser)와 에플리케이션(App)을 비교하자면 에플리케이션은 이미지나 스트리밍 비디오 등을 사용하는 것이 가능하나, 설치과정이 필요하지만 만족도는 브라우저를 사용하는 것보다 더 높은 것으로 나타난다. app을 사용하는 경우에 더 높음)

	Mobile app	Mobile web smartphone	Mobile web feature phone
Categorical questions	X	X	X
Multiple responses	X	X	X
Sliders	X	X	
Grid		X	
Long list	X	X	
Open-ended	X	x	
barcodes	X		
GPS	X		
Picture	X		
Video	X		
Clickable image		X	
Ideal length	<10 MIN	<10 MIN	<15 SCREENS

줌 기능이나 방향을 바꾸지 않고 응답할 수 있도록 설계해야 하고,

추가행동이 필요하지 않도록 질문을 설계해야 한다. 큰 글씨크기, 스크린의 가로틀을 고정하여 수평 스크롤링을 하지 않도록 해야 한다. 그리드 질문은 개별 항목으로 분할하고, 시각적 방해물을 최소화할 필요가 있다(이미지, 진행막대, <>). 또한 페이지 로딩 지연을 최소화하고 자동 넘김 기능을 탑재해야 하며, Apple과 Android 기기에서 제시형태가 다를 수 있고, 완성시간이 길고 항목누락 가능성이 높으며, 초두효과가 크게 작용함에 따라 드롭다운 메뉴 사용은 지양되어야 한다.

모바일을 이용한 새로운 프로젝트로 센서 및 웨어러블 기기 등을 통해 온라인 행동이나 앱 사용, 위치 및 움직임 같은 자료를 수집할 수 있게 된다. 회상오류가 감소하고 응답부담이 줄어들 수 있으며, 고품질의 자료를 더 자주 수집할 수 있다는 장점이 있다

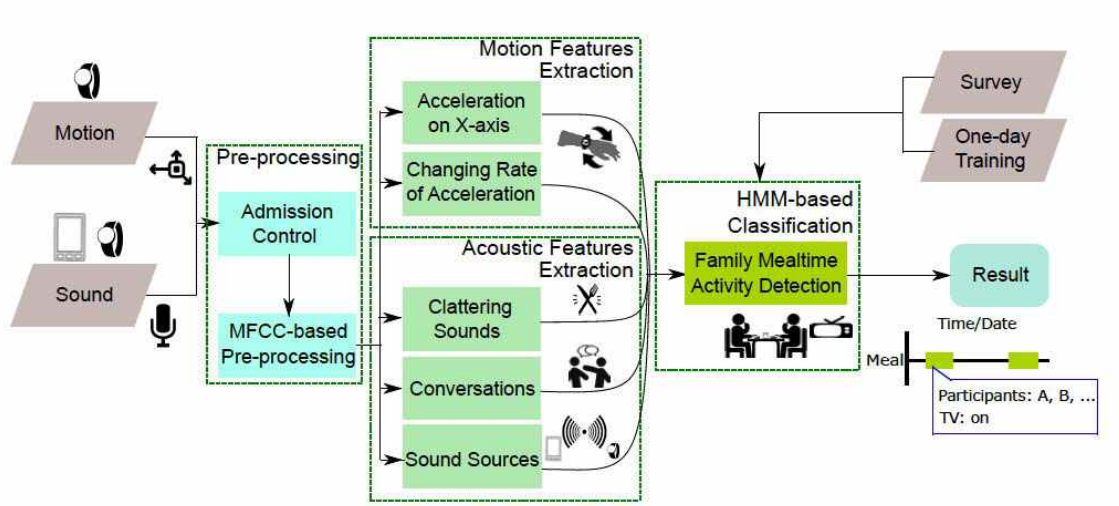


Fig. 1. System overview

3. 발표세션

□ 스마트폰을 통한 자료수집

영수증을 스캔하기 위해 모바일 앱을 사용한 연구에서 참가자와 관련한 제한이 있는 편이다. 예를 들어 참가자 중 앱과 호환되는 기기를 갖고 있지 않거나, 앱을 다운로드 받지 않거나, 불편해하는 경우 등이 있을 수 있다. 따라서 최근 연구에서는 모바일 앱 자료수집 참여를 늘리는 다양한 방법이 검토되고 있다. 특히 다양한 방법을 이용한 자료

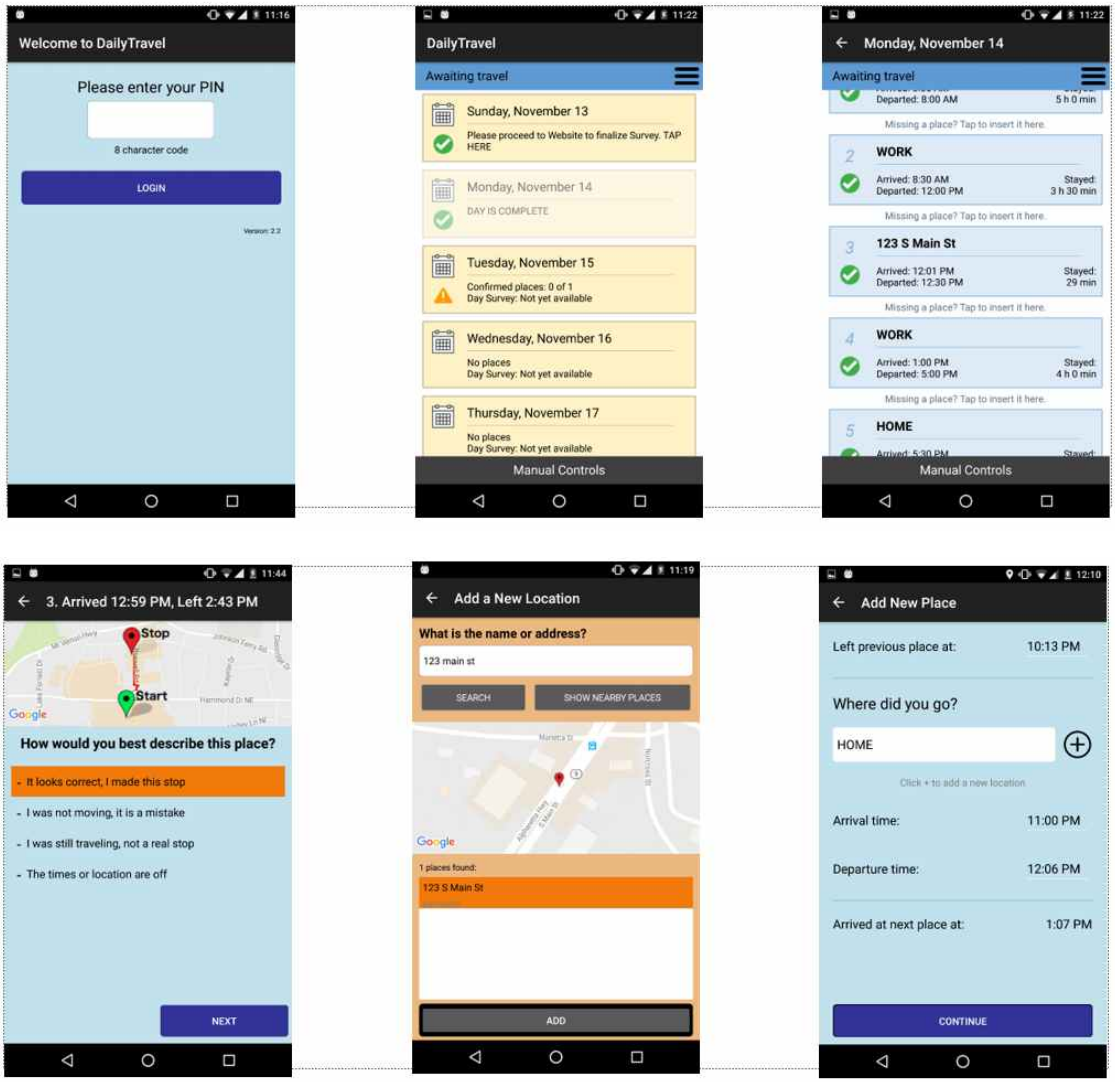
의 참여율, 자료품질, 참가자 편향 등 검토하는 것인데, 면접이나 우편으로 참가자 초대하거나 브라우저 기반 앱의 대안을 제공하는 것으로 검토되고 있다. 이러한 연구결과 면접을 통해 응답자를 앱으로 초대하는 것은 참여를 증가시키며, 브라우저 기반 앱의 대안 제공은 참여를 증가시키는 것으로 나타났다.

또 다른 연구로 스마트폰 센서자료는 응답부담 감소, 정확도 증가 등 유용하지만 응답자가 자료제공 의향이 있어야 한다. 자료제공 의향이 없는 사람들의 자료는 다를 수 있고 결과에 영향을 줄 수 있으므로 이에 대한 고민이 필요하다. 스마트폰 센서자료 수집에 영향을 미치는 과제 및 응답자 특성, 동의하지 않은 경우 편향 분석 등이 필요하며 이를 위해 스마트폰과 테블릿 사용자를 3가지 조건(gain/loss 질문표현, 비밀보장, 수집된 자료에 대한 통제)에 무작위 할당하고, GPS 위치를 공유하고 사진/비디오 자료를 요구하는 방식으로 연구를 진행하였다. 그 결과 통제에 대한 강조는 긍정적 효과 보였지만, 질문표현과 사생활 강조는 효과가 없는 것으로 나타났다. 스마트폰 사용 행동은 센서자료 공유에 영향을 주고 사생활은 GPS 위치 공유에만 영향을 주며, 과거 센서연구 참여경험은 자료 공유경향을 증가시키는 것으로 나타났다. 일반적으로 비동의 편향은 무응답 편향보다 크다고 알려져 있다.

조사방법 분야의 최근 관심이 센서와 웨어러블로 이동하고 있다. 모바일과 웨어러블 기기를 통한 자료수집 가능성이 증가하고 있으며, 이들 기기의 기능과 형태가 다양하게 변화하고 있다. 응답자의 동의를 얻은 후 모바일이나 웨어러블 기기의 다양한 센서자료에 접근할 수 있으며, 위치정보, 비디오, 방향정보 등을 자동 수집할 수 있다. 센서자료를 이용한 사례를 살펴보면 첫 번째로 카메라, 시간-위치 정보를 Household Budget Survey에서 시간-위치 인식, 영수증 촬영 후 결과 확인 과정에 활용하고 있다. 두 번째로는 Healty 자료에서 웨어러블 기기 및 카메라를 활용하여 움직임, 심박동/활동성/노출(CO2, 빛, 소음) 측정하고 있다. 이를 위해 복잡하고 다양한 자료의 유용성 평가를 위해 기준이 필요하며, 윤리적 문제나 사생활과 같은 응답자 관점은 오늘날과 같이 중요하지는 않을 수 있다.

Westat은 Household Travel Survey(HTS)에서 GPS 추적자료를 수집

하고 있다. 이 조사는 스마트폰 앱에서 수집된 방문 정류장의 GPS 자료를 이용하는데, 참가자가 앱 사용을 결정해야 참여할 수 있다. 따라서 모든 가구구성원(13세 이상)에게 참여를 제안하고 이 중 일부만 참여를 선택하였다. HTS에서 스마트폰 앱을 사용하는 참가자와 사용하지 않은 참가자의 인구사회학적 특성을 분석하고 가구 및 가구원의 여행 특성을 분석하였다. 그 결과 스마트폰 앱은 대부분 사용하지만, 더 젊고 교육수준이 높고 부유한 사람들에 편향되는 것으로 나타났다. 일반적으로 스마트론 앱의 사용은 여행에 대한 더 많은 정보를 제공할 수 있다.



□ 면접원과 응답자 간 상호작용

개방형 질문에 대한 응답을 폐쇄형 응답범주로 기록하는 과정에서 오류 발생 가능성 있으며, 이 원인 중 하나는 면접원이 긴 응답범주를 확인하기 어렵기 때문일 수 있다. 이에 응답범주의 길이(수)가 응답분포, 면접원-응답자의 행동, 면접원의 기록 정확성에 미치는 영향을 확인하기 위해 2가지 질문에 대한 응답을 면접원이 코딩하도록 한다. 직업에 대한 응답을 8개 또는 17개의 범주로 코딩하도록 요구하고 사람들이 온라인으로 어떤 활동을 하는지 질문하고 6개 또는 18개의 범주로 코딩하도록 요구하였다. AbySRBI의 Work and Leisure Today 2 Survey(RDD 전화조사) 이용하여 면접내용은 녹음되었고 대화내용을 행동 코딩하였다(27명의 면접원 대상, 2가지 질문 중 하나에 무선할당). 자료처리를 위해 면접원-응답자 행동(Interviewer-Respondent Behaviors) 분석을 위해 898개 사례를 전사한 후, 훈련된 학부생 코더와 숙련자 코더(10%의 subsample)가 코딩하였다(Kappa>.60), 면접원의 정확한 코딩(Interviewer Accuracy Coding) 분석을 위해 석사과정 학생 2명이 코딩하였다(Kappa>.80). 이 두가지 방식의 코딩을 비교한 결과, 응답자들이 다양한 응답을 빠르게 하고 면접원은 응답내용을 응답범주 목록에서 찾아 정확히 기록해야하기 때문에 현장에서 면접원이 코딩하는데 어려움이 있는 것으로 나타났다. 응답범주 길이에 따른 응답분포와 면접원-응답자 행동에 차이가 없었고, 응답범주 수를 줄이는 것이 면접원에게 도움되지는 않은 것으로 분석되었다. 특히 직업질문이 코딩하기에 어려웠고 이는 면접원 교육에서 고려되어야 하는 매우 중요한 부분이다.

컴퓨터를 이용한 조사는 CATI (computer-assisted telephone interview), CAPI(computer assisted personal interview), 자기기입식 web interview 등 세 가지 방법이 있다. 각 조사방법에 따라 사회적 바람직성 편향과 만족도가 다르므로 이를 위해 컴퓨터를 이용한 조사들의 사회적 바람직성 편향과 만족도에 영향을 주는 요인을 확인하였다. 유럽사회조사(European Social Survey)의 혼합모드(60 CATI, 54 CAPI) 실험 결과를 분석한 결과 CATI와 CAPI의 사회적 바람직성과 만족도에 차이를 유

발할 수 있는 쟁점들을 발견하였다. 면접원의 웃음은 CAPI보다 CATI에서 더 일반적으로 나타났고 'sorry'와 같은 사과의 표현은 두 조사모드에서 동일하게 나타났다. 질문 응답하는 과정에서 CAPI보다 CATI에서 더 많은 단어를 사용하며, 응답자는 CATI에서 응답하는 것을 더 어려워하는 것으로 나타났다.

또한 사회가 고령화되고 있으며, 고령화 과정의 이해를 위해 고령자 조사자료의 품질을 평가하고 품질제고 노력이 필요한 시점이다. 유럽 건강 및 은퇴조사(SHARE) 자료의 품질을 평가하고, 노화로 인한 인지기능 변화가 조사품질에 미치는 영향을 평가하기 위해 인지기능 지표(장기기억, 단기기억, 숫자과제 등)가 heaping에 미치는 영향을 분석하였다(종단연구, wave 1/2/4/5/6 자료 활용, CAPI 방식 조사). 여기서 heaping은 the tendency to provide answers that are multiple of a given number이며, heaping에 사용된 변인은 고용과 위험행동 기간(몇 년 동안 흡연했나요?, 가장 최근 일자리에서 몇 년 동안 일했나요?), 건강(몸무게는?, 키는?) 등이 었다. 그 결과, 전반적으로 인지기능이 좋은 고령자는 연령에 관계없이 덜 heaping하지만 인지기능의 영향은 미미하나 다른 요인(교육, 성별, 건강)은 heaping을 야기하는 것으로 나타났다.

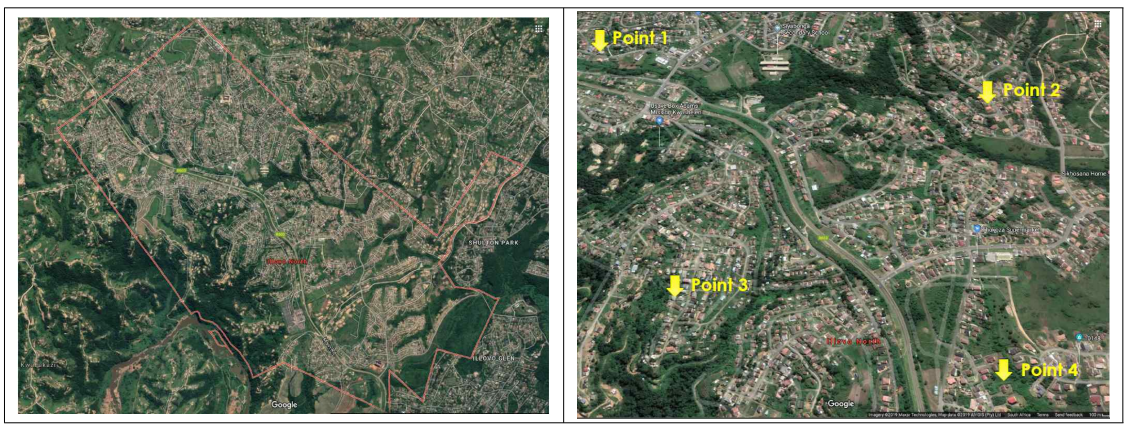
퇴직자 대상 조사(retraite, MOBIlité transnationale et QUALité de vie MOBIQUAL)는 세 가지 모드(PAPI, CAWI, CATI)로 진행되는 조사로 63~97세 응답자 대상 10~15분 길이의 조사를 진행한다. 각 모드에서 고령자의 응답특성을 분석한 결과 CATI와 PAPI는 결과가 유사하고 민감한 질문에 대한 면접원 효과는 없는 것으로 나타났다. CAWI에서 더 높은 긍정 응답 결과가 나타났고 온라인 응답자가 더 높은 긍정 응답 결과로 나타났다. 민감한 주제에 대해 PAPI나 CATI보다 온라인 조사를 더 편하게 느끼는 것으로 분석되었다.

설문조사에서 질문에 적절히 응답하기 위해서는 응답자가 의사소통 능력과 인지능력을 갖추어야 한다. 이는 연령과 관계가 있으며, 연령에 따라 조사에 대한 질문에 다르게 대처한다는 연구 결과가 있다. 이에 응답자의 연령이 면접원-응답자 간 상호작용을 증가시키는지 평가하고, 상호작용이 응답행동에 미치는 영향을 평가하기 위해 유럽사회조사에

서 수집된 자료를 분석하고 7개 연령그룹 참가자의 응답행동(질문에 대한 이해, 확인 질문, 응답을 꺼리는 행동 등) 및 응답시간을 평가하였다. 그 결과 고령 응답자(71세 이상)와의 면접조사는 더 복잡하고 IIC(intra interviewer correlations)를 증가시키는 것으로 나타났다.

□ 파라데이터 분석 및 활용

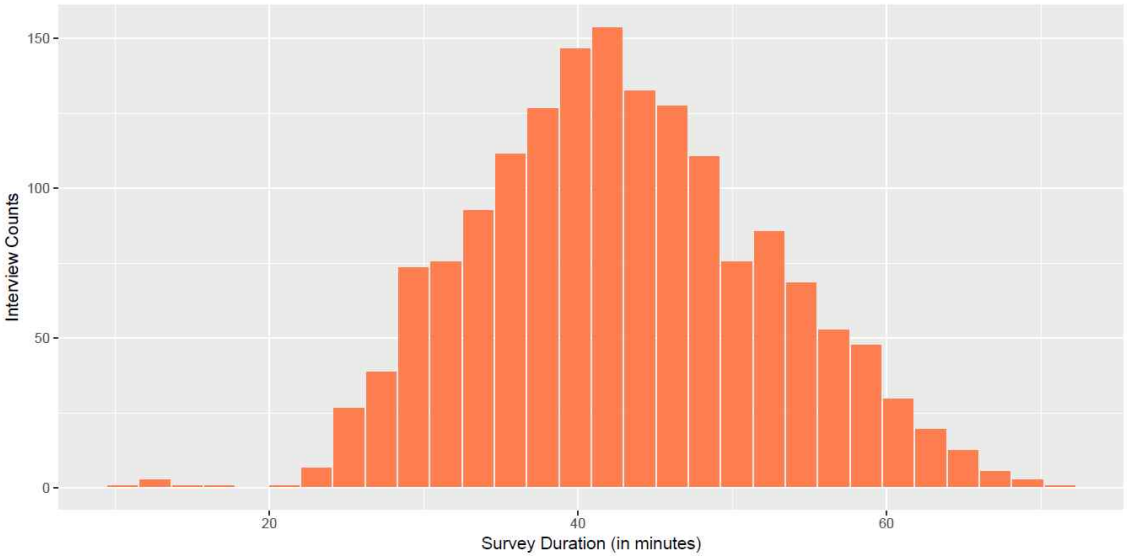
가구접촉 프로토콜 대 현장의 심화된 관행사이의 간격을 절단모서리라고 하는데 이에 대한 실험연구가 지속적으로 진행되고 있다. 실험의 표본설계는 층화무작위집락추출설계로 층내는 확률크기비례(PPS)로 추출하고 1차 표본 추출단위(PSU)는 센서스 조사구(EA)이며, 2차 추출은 PSU당 4개의 무작위 시작지점을 선정하였다.



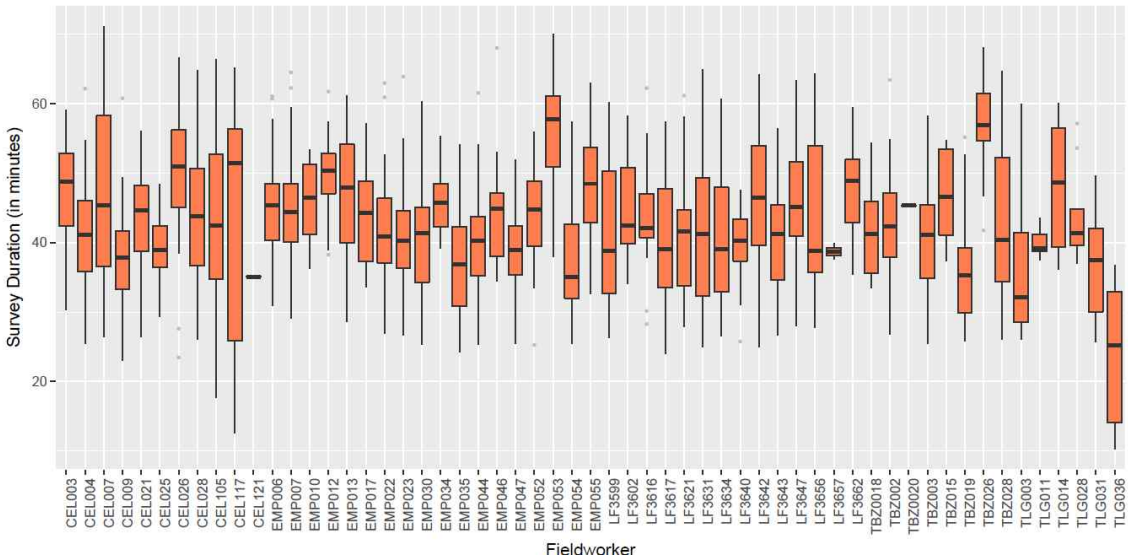
가구접촉 프로토콜은 다음의 그림과 같고 이에 대하여 현장 조사자의 규정 준수는 불일치한다. 발견된 간격(차이)에 대한 목록을 보면 매우 짧은 완료 시간, 건너뛰거나 가구 접촉기록을 남기지 않음, 응답자의 선별적 선발, 무작위 시작지점으로부터의 GPS의 거리, 움직임(교체이동)이 왼쪽 규칙과 일치했는지 여부, 인터뷰 중 GPS 거리, 오디오 녹음의 편차 등이었다. 아래의 그래프들은 이에 대한 분석결과를 나타내고 있다. 현장에서의 표본설계 실현은 연구자와 현장 팀 간의 투명성과 협업이 필요하다. 현장조사자는 실용적이고 다양한 문제 해결이 필요한 일일 과제에 직면하고 있다. 가구접촉전략에서 발생하는 오류와 차이를 이해하기 위해 데이터 수집 중에 현장에서 상당한 시간이 필요

할 것이다. 모든 현장 조사 팀원이 동일한 업무윤리를 가지고 있는 것은 아니므로 모니터링 시스템은 정직한 업무에 보상하고, 신뢰할 수 없는 조사자에게 할당된 업무를 줄어야 한다.

○ 불가능할 정도의 짧은 조사 기간

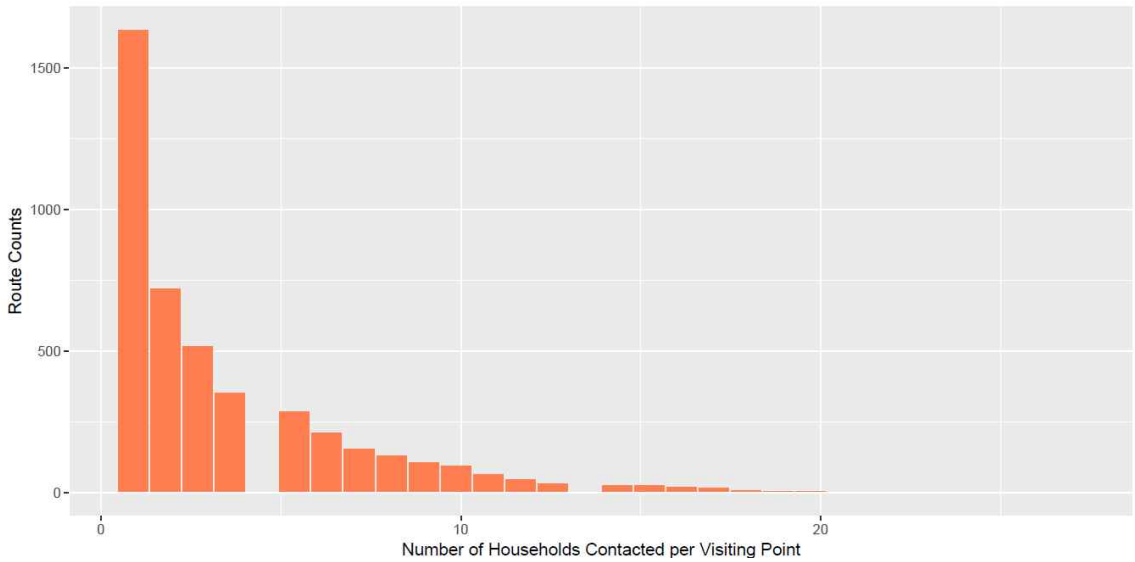


(Y축: 면접 수, X축: 조사 기간(분))

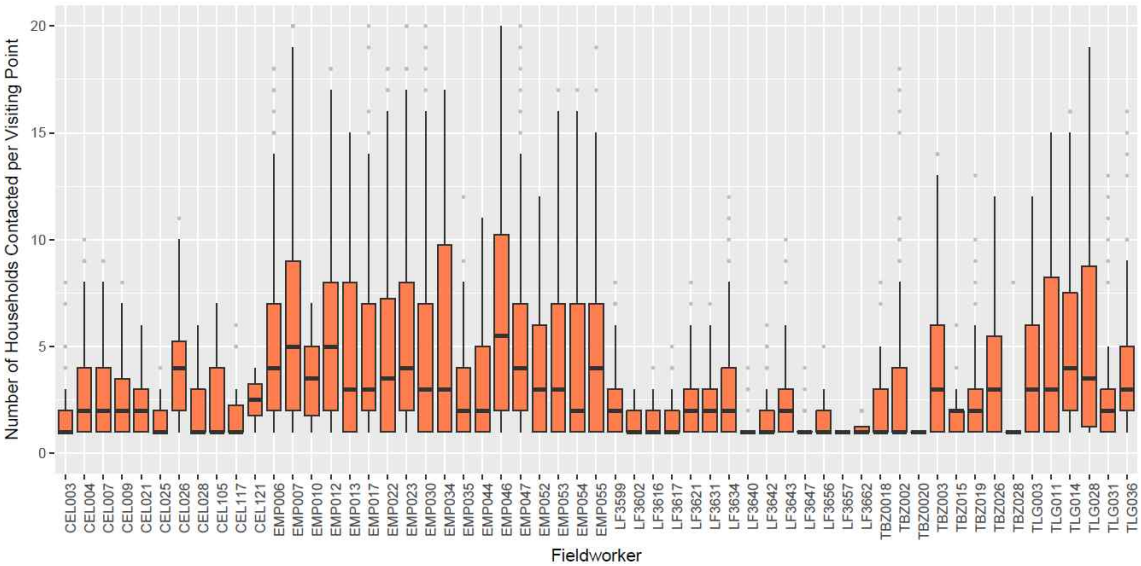


(Y축: 조사 기간(분), X축: 현장조사자 ID)

○ 건너뛰거나 가구접촉기록을 하지 않은 증거

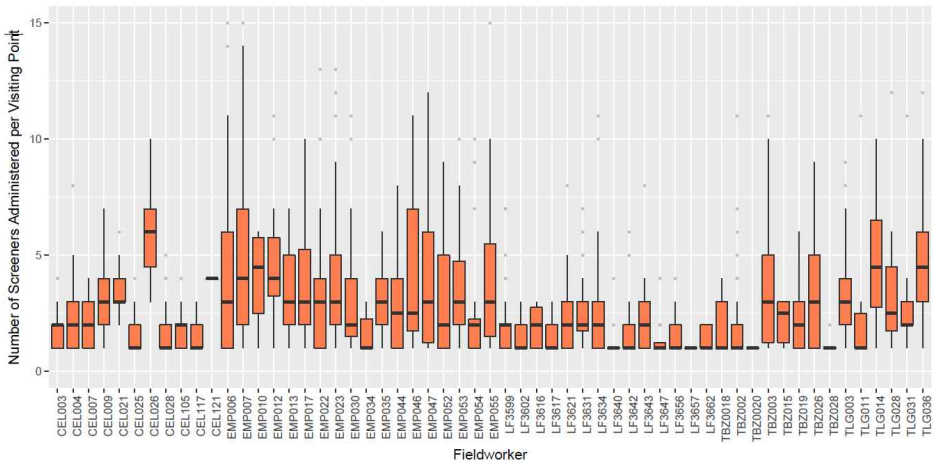


(Y축: 경로(루트) 수, X축: 방문 지점당 가구접촉의 수)



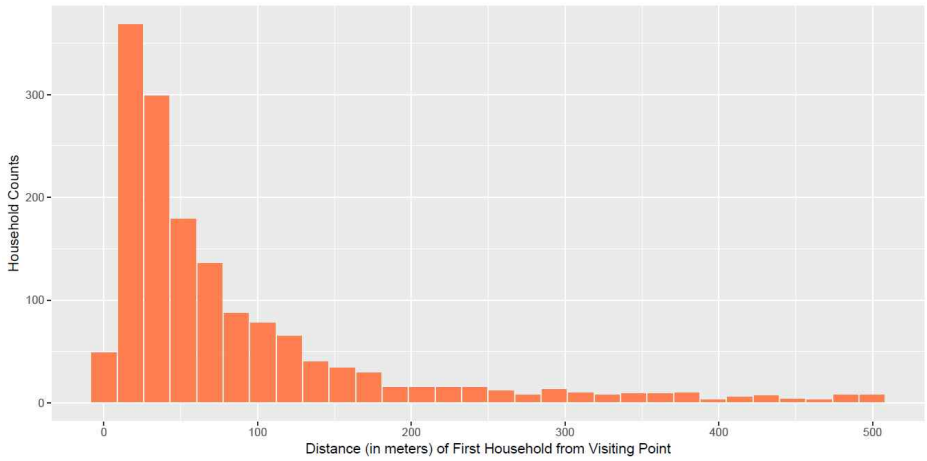
(Y축: 방문 지점당 가구접촉의 수, X축: 현장조사자 ID)

○ 선별적 스크리닝의 증거

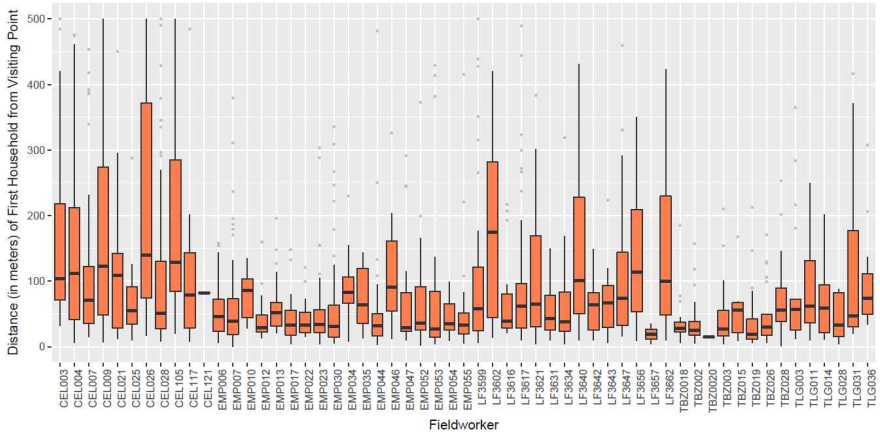


(Y축: 방문지점 당 관리되는 스크리너 수, X축: 현장조사자 ID)

○ 사전 설정된 시작점으로부터의 거리

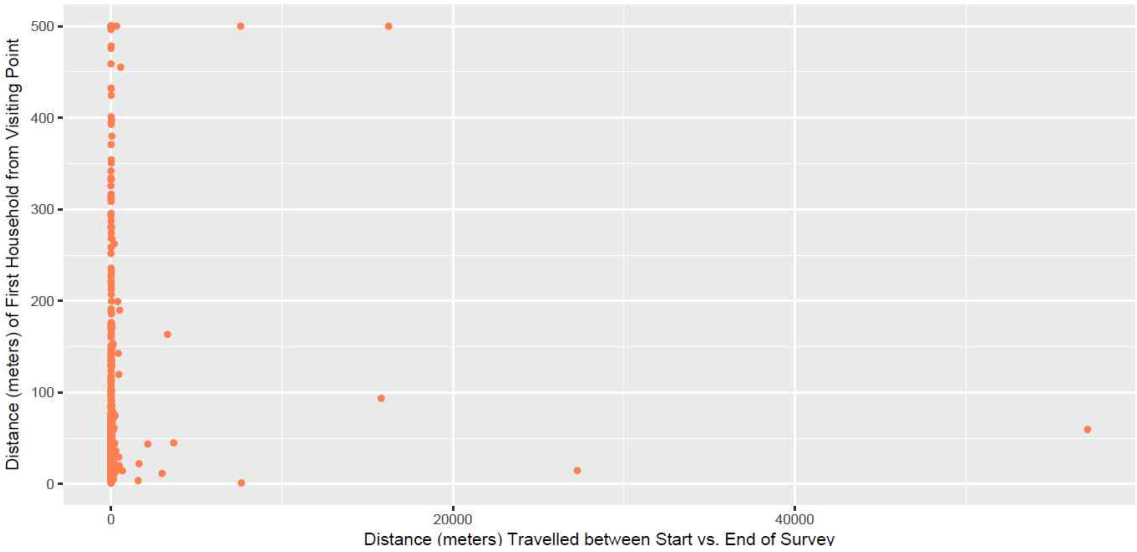


(Y축: 가구 수, X축: 방문지점으로부터의 첫 번째 가구의 거리(미터))



(Y축: 방문지점으로부터의 첫 번째 가구의 거리(미터), X축: 현장조사자 ID)

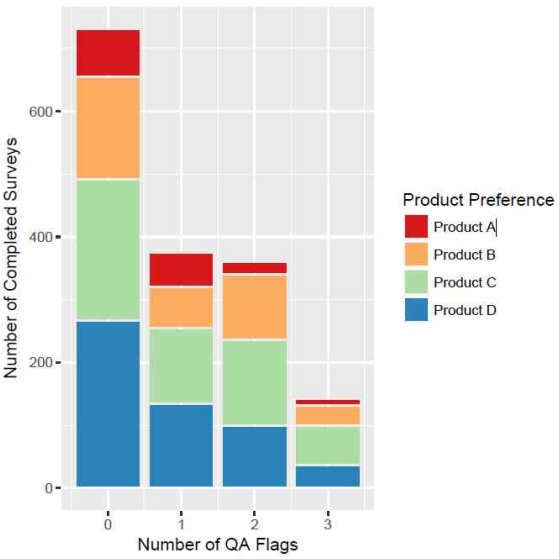
○ 조사시작과 종료사이의 이동거리



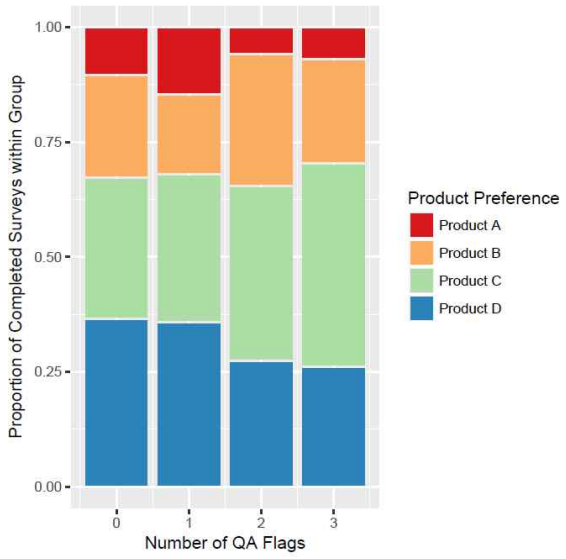
(Y축: 방문지점으로부터의 첫 번째 가구의 거리(미터), X축: 조사시작과 종료사이의 이동거리)

다음의 그래프는 QA 플래그의 범주(classes)별 응답 특성을 비교하여 조사 추정값에 미치는 영향을 알아본 결과이다.

○ QA플래그 간의 응답 선호사항 항목 응답비교

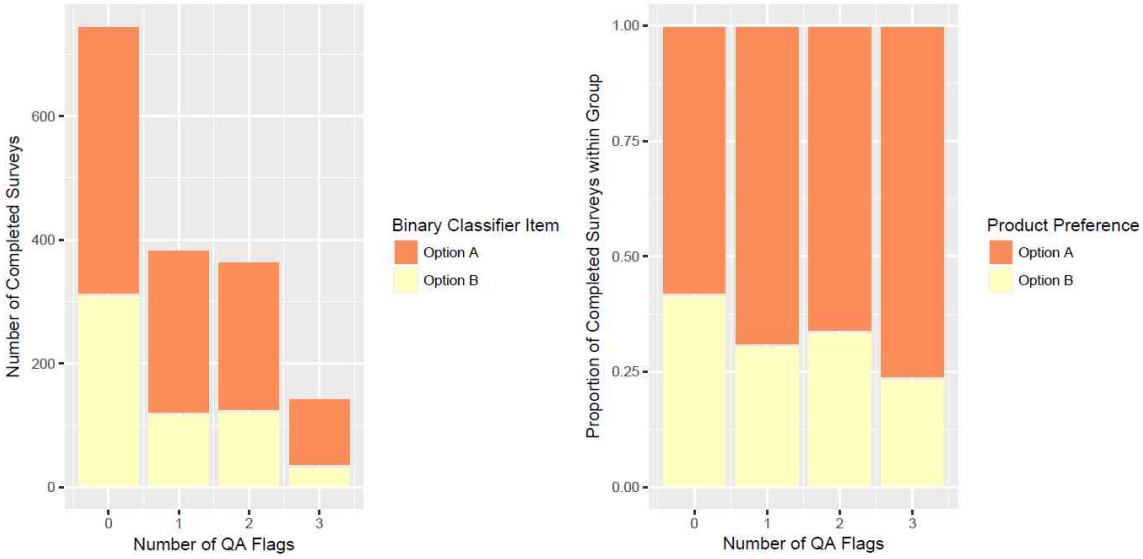


(Y축: 완료된 조사 수)
(X축: QA 플래그의 수)



(Y축: 그룹 내 완료된 조사완료 비율)

○ QA플래그 간의 이항범주의 응답비교



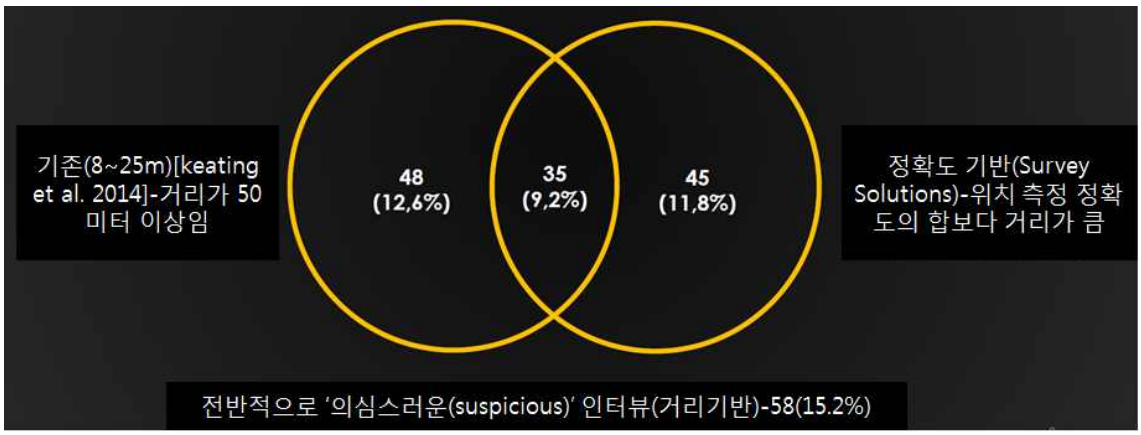
현장 조사자 모니터링을 위한 추가 기회를 제공하는 CAPI의 GPS-파라데이터에 대해 알아보기 위해 개별위치분석(Separate locations analysis)을 살펴보면 인터뷰 시작 및 종료 시 위치를 비교하는 지오펜싱(Geofencing), 인터뷰 장소와 표본추출된 가구의 위치를 비교하거나 응답자가 있는 집의 직접길이(Strand length), 너무 밀집된 인터뷰 그룹 위치를 확인하는 커브스토닝(Curbstoning) 테스트 등이 있다. 이를 토대로 GPS 측정 위치 순서를 경로분석(Route analysis)하기 위해 경로로 인터뷰 장소연결과 조사원 경로분석 등을 실시하였다. 그렇다면 현장 모니터링을 위해 테블릿 CAPI에 GPS-파라데이터를 사용하는 방법은 무엇이 있을까? GPS 위치는 인터뷰 시작과 종료시 테블릿의 위도 및 경도에 대한 정보(Survey Solutions application)와 자동측정인 능동측정(active measurement), GPS 경로인 현장에서 면접관(조사원)의 경로에 대한 정보(GPSLogger 응용프로그램)를 수동측정하는 것이 있다. 추가적으로 면접관(조사원)의 사회 인구통계학적 특성 및 전환 성공에 대한 기대를 할 수 있다.

GPS-파라데이터 품질은 결측데이터(Missing data)와 측정 정확도(GPS Logger, Survey Solutions)로 살펴볼 수 있는데 결측데이터는 인터뷰 시작 또는 종료시 위치 측정 데이터 누락을 종속변수로 하는 이

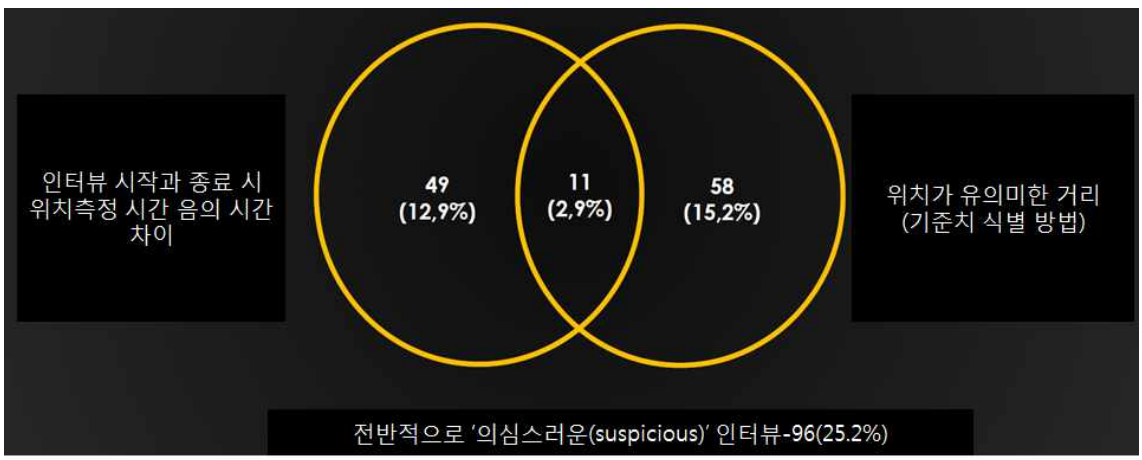
항 로지스틱 회귀분석로 확인할 수 있다. 독립변수로는 태블릿 가용성, 테블릿 자신감, 기대지수, 평균정확도(GPS 로거), 평균 배터리 충전수준, 지역변수 등을 사용한다.

의심스러운(suspicious) 또는 위험한(at risk) 인터뷰 식별은 인터뷰 시작과 종료시 위치를 비교하고 임계값을 선택하며, 위치측정의 시차(Time difference in location measurements)를 통대로 확인할 수 있다.

○ 지오펜싱: 두 가지 기준치(threshold) 식별 방법



○ 위치 측정간 음의 시간 차이



의심스러운 인터뷰의 4가지 지표는 인터뷰 시작과 끝 위치 사이의 유의미한 거리(기존 기준치 식별 48-12.6%) 및 정확도 기반(45-11.8%), 위치 측정 사이의 음의 시차(49-12.9%), 인터뷰 근접성(동일한 HH 구성원 제외) 8미터(90-18.3%) 또는 16미터(132-26.9%), 같은 HH내의 인

터뷰 사이의 상당한 거리(25-5.1%) 등이고 전반적으로 의심스럽거나 위험한 인터뷰는 274-55.8%로 나타났다. GPS-파라데이터 품질은 지역(개발도상국의 품질 저하 및 인터뷰 진행자의 특성(CAPI에 대한 신뢰도)에 따라 달라질 수 있다 CAPI에 대한 경험은 인터뷰 시작과 끝 위치 사이의 유의한 거리와 높은 확률이 있으며(93%), 태블릿 작업에 대한 높은 수준의 비밀보호는 유의한 거리와 낮은 확률이 있다(46%).

GPS-파라데이터 품질 향상과 의심스러운 인터뷰 하위 수준과 연결될 수 있는 CAPI를 사용하면서 면접자 교육에 초점을 맞춰야 한다. 예를 들어 인터뷰 시작과 끝에서와 같이 두 개의 위치 사이의 거리에 대한 기준치 식별로 정확도를 사용한다. GPS데이터 품질은 지역마다 다를 수 있다. GPS-파라데이터는 다른 현장 작업 모니터링 방법과 함께 사용해야 한다. GPS-파라데이터 분석에만 근거한 조작이나 위조에 대한 정확한 가정은 할 수 없다(의도적인 오류 또는 기술적 어려움).

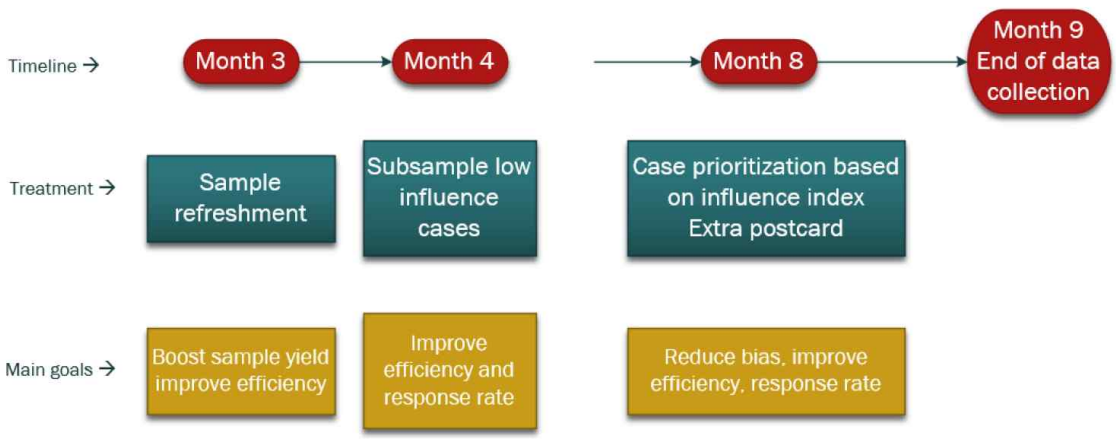
종단면 패널과 횡단면 조사의 경우와 같이 표본설계가 다른 조사별로 GPS-파라데이터 활용이 다를 수 있다. 응답자의 주소를 사용할 수 없고 수동 GPS데이터 수집과 관련하여 일부 추가 소프트웨어가 사용되었는지 여부와 전원이 꺼진 시기를 감지할 수 없는 것이 단점으로 작용하였다. 향후에는 실험용 2차 웨이브 RLMS-HSE CAPI-추가지역, 인터뷰, 응답자 및 데이터 등을 획득하고 웨이브 사이의 위치를 비교하며(패널옵션), 데이터 품질과 정도의 관점에서 애플리케이션(GPS Logger, Survey Solutions)을 비교할 계획이다.

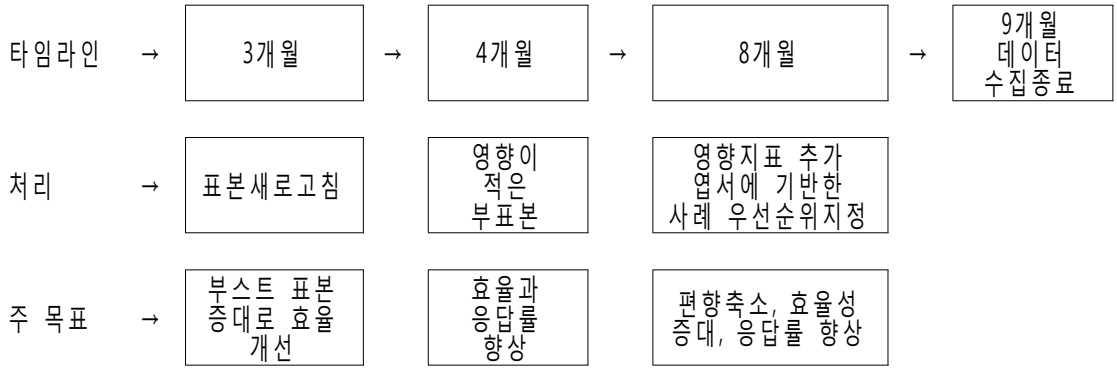
파라데이터를 이용한 적응적 조사 설계 PIAAC 실험 및 국제적 구현을 살펴보면 현재 응답률 감소하고 있어 무응답 편향에 대한 높은 가능성에 대한 민감성이 높아지고 있지만 가장 쉬운 경우에 대해 계속 완료를 통해 편향을 감소시키지 않는 것으로 분석되었다. 따라서 적응적 조사설계(ASD)에 대한 요구가 높아지고 있고 이는 PIAAC ASD 전략으로 수립되고 있다. PIAAC ASD 전략은 표본 산출예측(sample yield projections)으로 잠재적 부족에 대한 적시적 반응과 표본 새로고침(Sample refreshment)으로 고정비용으로 편향을 유발하지 않고 완료 수를 증가하는 방법이다. 매우 낮은 성향 사례를 종결시킨 다음, 폐쇄된 사례와 동일한 예상 횟수를 사용한 것과 같은 크기의 무작위 그

룹을 릴리스 하는 방법으로 볼 수 있다. 우선순위가 낮은 사례의 부표본(subsample)을 통해 편향 없이 이러한 사례에 대한 노력을 줄일 수 있을 것이다. 발견된 초기 사례 중에서 우선순위가 낮은 사례의 1/3 정도를 선택하여 우선순위결정(Case prioritization)을 통해 제약 조건에 따른 편향을 최소화할 수 있다. 우선순위가 높은 사례에 대해서는 추가 엽서(Extra postcard)를 활용할 것이다.

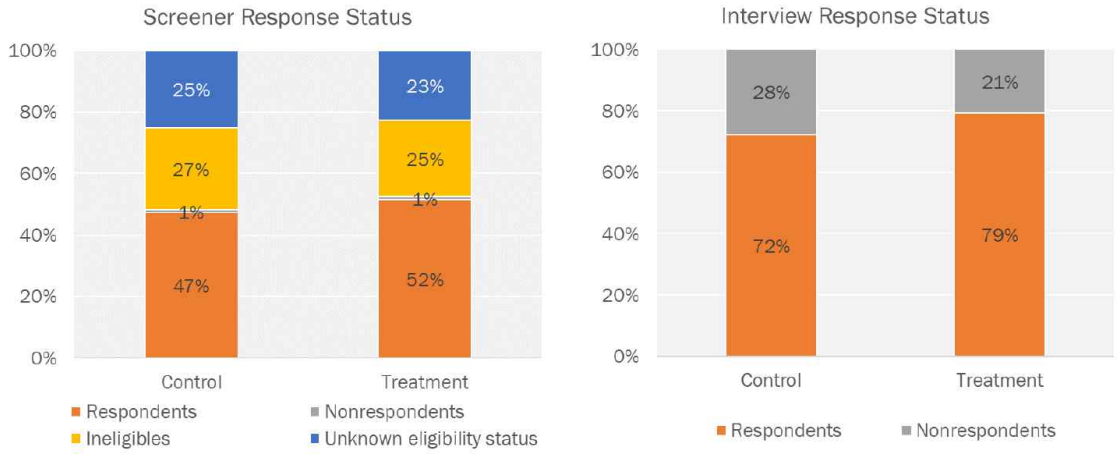
PIAAC 사례에서 보면 우선순위설정은 이상적으로 무응답 편향을 최소화하는 것을 목표로 목표 표본크기, 목표 응답률 등을 달성하고 예산을 유지하기 위해 일정 수준의 접촉 시도(CA; contact attempts)를 유지하는 것이도 진행된다. 실제로는 -표본추출된 각 사례에 대한 예측 테스트 점수를 통해 공개 사례가 결과에 미치는 잠재적 영향 파악하게 된다. 이를 확인하기 위해 우선순위 시뮬레이션의 결과를 보면 영향력 측정을 사용하여 우선순위를 설정할 때 다음과 같은 경우에 편향 및 평균제곱오차의 잠재적 감소를 살펴보게 된다. 그것은 우선 순위화가 없거나 낮은 응답성향에 따른 우선순위설정을 진행하게 된다. 무작위로 선택된 통제(control) 및 처리(treatment) 그룹으로 분할하고, 할당은 확실한 1차 추출단위를 위한 2차 추출단위, 불확실한 1차 추출단위를 위한 2차 추출단위로 설계하고 각 그룹에는 약 1,800명의 응답자가 있으며 통제그룹은 표준 접촉프로토콜 사용하는 방법으로 실험이 설계되었다.

○ 처리의 타임라인 및 목표

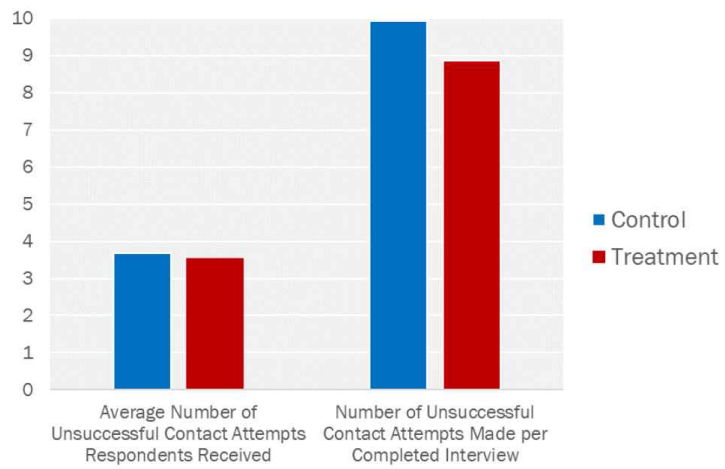




○ 긍정적 응답상태 표시(지표)



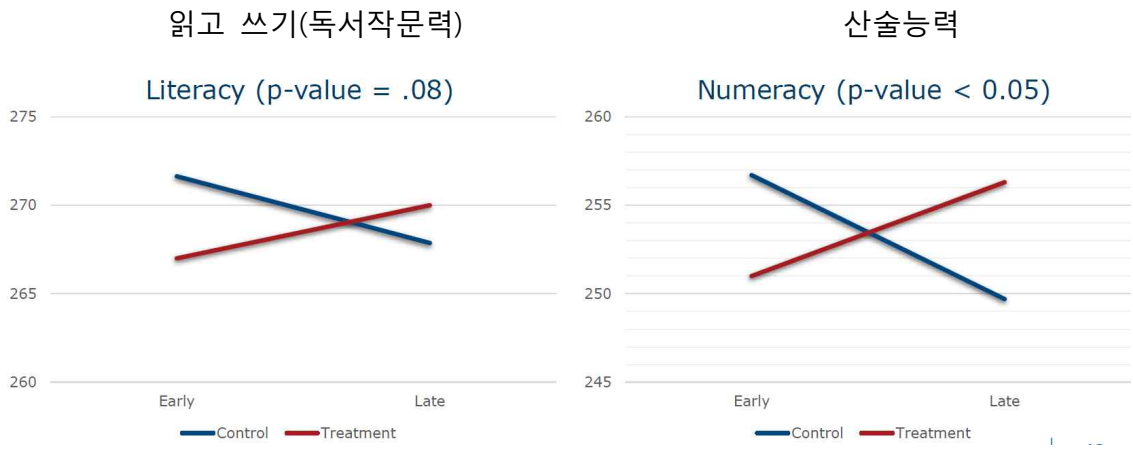
○ 효율성 향상 표시(지표) : 그룹별 접촉시도 실패



PIAAC에서 약 3,600회 시도 횟수를 줄일 수 있음
(완료된 인터뷰 당 실패한 횟수는 약 1회)

응답자가 받은 평균 접촉 시도 실패횟수	완료된 인터뷰당 실패한 접촉시도 횟수	

○ 편향 감소 가능성 표시(지표) : 그룹의 교호작용과 응답시기



ASD 실험 결과를 살펴보면 처리로 인해 높은 응답률이 나타나고 있다. 영향력이 적은 경우의 부표본추출 및 추가적 엽서로 처리하였다. 처리의 조합으로 인해 약간 더 비용이 효율적으로 나타탔다. 사례 우선순위에 의해 편향이 잠재적으로 감소하였고 사례 우선순위를 위한 최적화 프레임워크에서 영향 측정이 도출되었으며, 모의실험과 실제실험에서 좋은 결과로 나타났다. 사례 우선순위 시스템은 무응답 편향 감소를 목표로 하여 진행되었고 실험에 의해 영향을 받았고 시스템을 위한 일반화된 접근법이 사용되었다. 면접원이 우선순위가 높은 경우를 완료하는데 집중할 수 있도록 현장 관리 직원에게 지침이 제공되었고 접촉시도 순서로 처음 몇 개의 우선순위 그룹은 표준 접촉시도 프로토콜을 충족하는지 여부로부터 도출되었으며, 통계적 순서화로 접촉 프로토콜을 충족하는 경우 수집된 파라데이터를 사용하여 각 경우에 대해 통계 기반 순서가 도출되었다. 이것이 국가에서 우선순위를 가져오는 옵션으로 적용되었다.

미국의 전국 가족성장조사(NSFG; National Survey of Family Growth)을 통해 기계학습모델을 사용한 반응 설계 프레임워크에서 대안적 설계에 따른 비용 예측을 살펴보았다. 이 조사는 미시간대학교 사회연구원과 미국 보건인적자원부의 여러 기관으로부터 예산을 지원받아 미국 질병통계예방센터(CDC)의 국립보건통계센터(NCHS; National Center for Health Statistics)에 의해 실시된다. 파라데이터와 같이 현장(field)에서 입수되는 데이터 모니터링으로 실제로 반응 조사 설계가 활

용되고 있다. 이를 위해 대안 설계 단계에서 비용 예측이 필요하여 다중회귀모형, 베이지안 가법 회귀 나무(BART; Bayesian Additive Regression Tree) 방법을 통해 진행하였다.

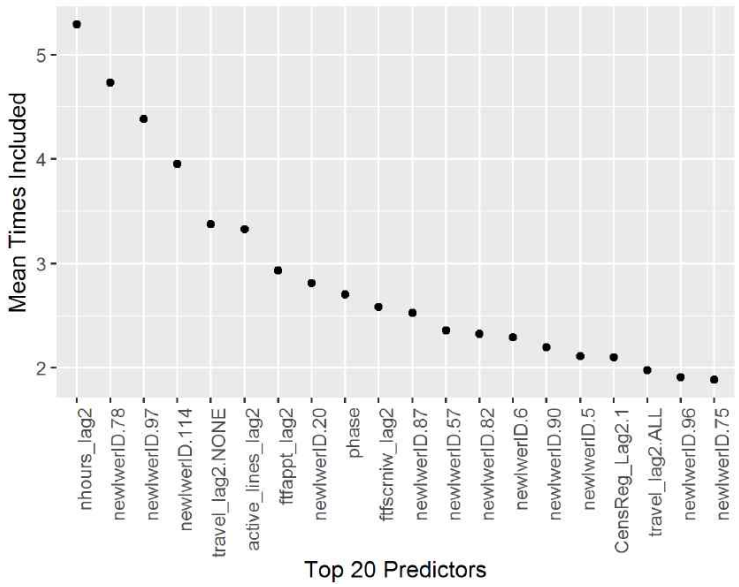
반응 조사설계를 자세히 살펴보면 불확실성이 조사 설계에 이슈가 되고 있어 파라데이터와 같은 현장에서 들어오는 데이터를 사용하여 반응 조사설계는 이러한 불확실성을 해결하는 방법으로 활용되고 있다. 이를 위해 비용 및 오차 지표를 개발하고 비용 증가 또는 오차 안정화, 오차 증가에 대해 계획된 개입을 통해 진행한다. 해결하지 못한 원론적인 문제가 있는데 그것은 비용과 오차는 시간에 따라 다르므로 반응 조사설계에서는 대리(proxy) 지표를 사용한다는 점이다. 비용은 통화시도, 무응답오차는 안정화된 추정값과 단계 수용력(phase capacity)을 지표로 사용하기 때문에 잘못된 지표로 인해 설계가 비효율적일 수 있다는 점을 가정하고 있다. 그렇다면 비용 예측의 정확성을 개선할 수 있는가? 이를 위해 데이터를 다음과 같이 수집하였다.

매 분기마다 새로운 표본 출현하므로 연속적으로 데이터를 두 가지 방법으로 수집한다. 첫 번째는 2단(two-stage) 수집방법으로 1단은 적격자 확인을 위한 선별(Screener) 인터뷰이고 2단은 선택된 사람의 주요(Main) 인터뷰를 통해 획득한다. 두 번째는 2상(two phases) 수집방법으로 1상은 10주 동안 자료를 수집하고 2상은 나머지 경우의 부표본(Subsample remaining cases)을 통해 획득한다. 이는 인터뷰 사례 및 적격 또는 적격 가능한 사례에 대한 과대표본 가능성이 존재하고 조사원 작업량의 2/3가 감소하며, 감사표시 추가, 인터뷰 진행자 행동 변경 등 데이터 수집 모형 변경된다. 가중값을 사용하여 2상의 데이터 및 응답률을 통합하여 사용한다.

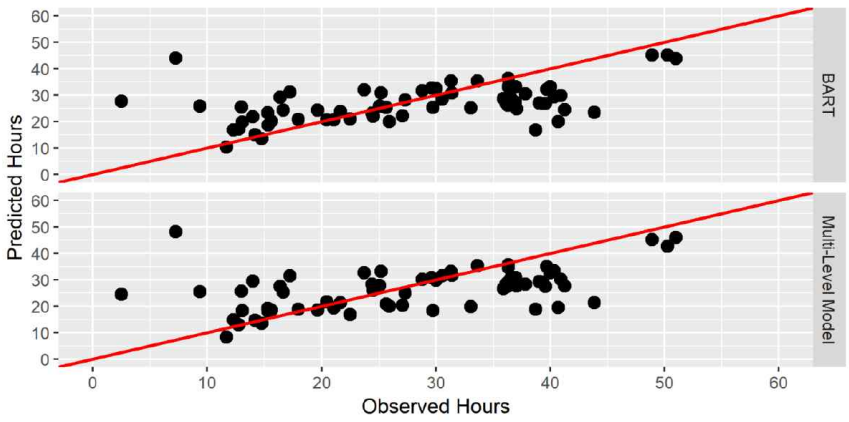
예측 시 사용 가능한 데이터를 사용하기 위해 NSFG 데이터를 사용하고 인터뷰 시간과 높은 상관관계 있는 파라데이터를 이용하였다. 예측하고자 하는 미래의 시점 기간에 대한 표본의 파라데이터와 다른 특성은 예측 시점에 이용할 수 없다. 예측 기간 2주 전의 값인 지연된 값(lagged value) 사용한다.

두 가지 예측방법을 사용하였는데 첫 번째는 다수준 회귀모형으로 각 면접자별 무작위 절편(intercept)을 사용하고 가지고 있는 모든 공변

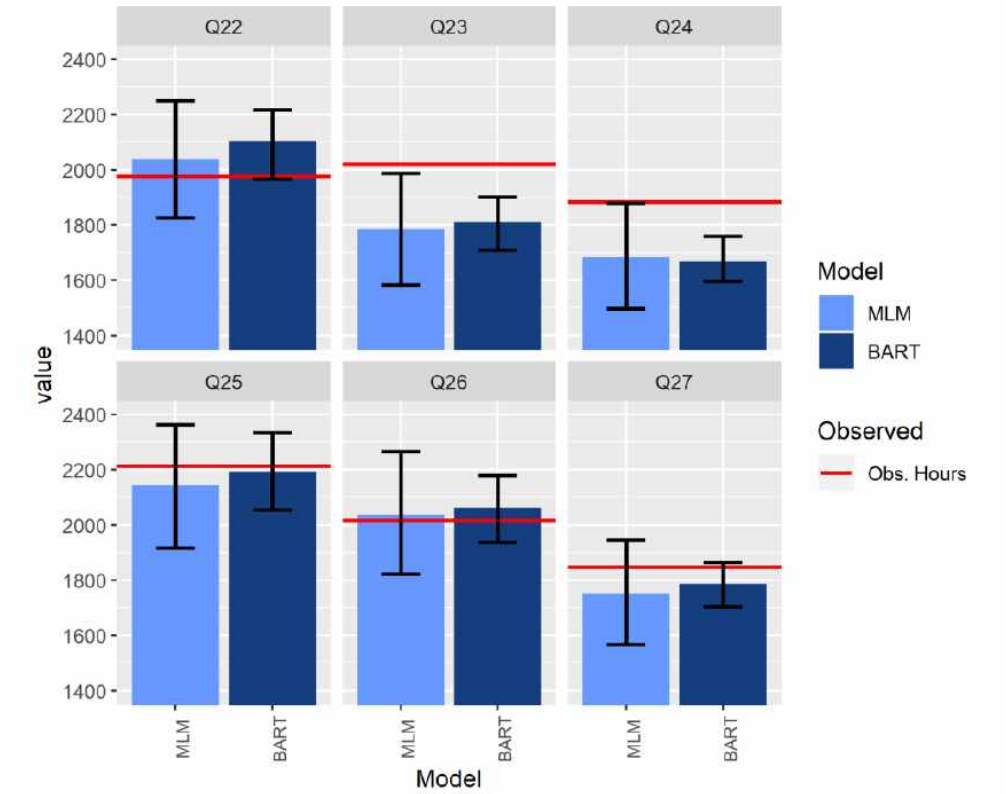
량을 모형에 포함하여 분석하였다. 두 번째는 베이지안 가법 회귀 나무로 의사결정나무의 나무의 합계 방법을 사용하고 예측변수의 포함을 사전에 제한하며, 각 예측변수가 얼마나 자주 포함되는지를 조사할 수 있다. 두 방법을 비교하기 위해 2단계의 총 예상 대 관찰 시간과 2단계의 조사원 수준 예측 대 관찰 시간을 비교하였다. 그 결과 조사원(면접원)은 일관성이 존재하는 것으로 나타났다. 결론적으로 살펴보면 비용 예측은 흥미로운 문제로 기존 예측 방법이 사용 가능하다. 이 문제에서 조사원이 일관된 행동을 하기 때문에 매우 유용할 것이며 이 문제에 대해 MLM과 BART모형은 비슷한 예측 정확도를 보이고 있다.



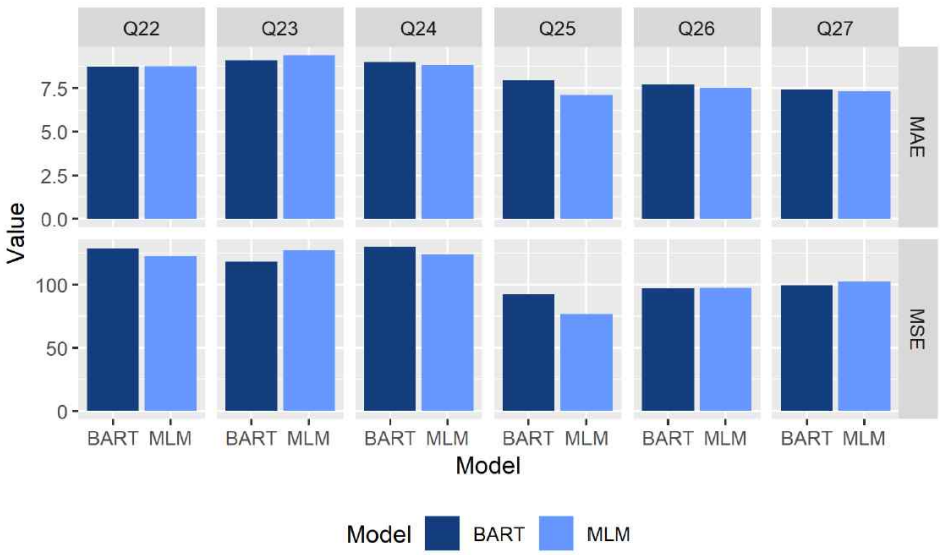
○ 예측 정확도-조사원(인터뷰어)



○ 예측 정확도-총 2단계 시간



○ 예측 정확도-조사원(인터뷰어)(2)



Ⅴ. 시사점 및 정책 제안

현재 국가통계생산은 지속적인 조사환경 변화의 시대에 직면하고 있다. 과학적 조사표 설계의 발전, 데이터 수집 및 분석기술의 고도화, 인터넷 및 웹 기반의 조사환경 확산 등 긍정적 변화라고 볼 수 있다. 반면, 조사표 복잡성 확대, 조사비용 증가, 응답률 감소 등 조사환경 악화에 대한 문제점도 동시에 발생하고 있다. 이에 통계생산 비용과 품질 측면 최적화를 위해 유연한 조사 관리와 설계가 필요한 시점이 도래하고 있다. 통계조사 효율화를 위한 현장조사 지침 및 조사방법 개선의 일환으로 조사과정 중에 수집되는 파라데이터 활용 필요성이 강하게 대두되고 있어 파라데이터 수집 및 활용방안 연구를 통해 국가통계 품질, 응답률, 비용 효율화 등 조사효율화를 위한 최신기법 검토가 필요한 시점이다.

현재 외국에서는 자동화된 시스템에서 파라데이터를 수집하고 이를 조사환경 개선 및 조사설계 단계에서 활용하고 있다. 따라서 파라데이터 수집 및 활용에 대한 유럽의 선진사례를 벤치마킹하여 조사환경 개선 및 조사방법 선진화에 적용할 필요가 있다. 첫 번째로 조사과정 효율화, 응답거절 최소화 방안, 불응 대응논리 마련, 조사담당자 교육 등 조사관리를 위해 파라데이터를 수집하고 활용할 필요가 있으며, 두 번째로 파라데이터를 이용하여 현장조사 지침 및 조사방법을 개선하고 맞춤형 응답자 접촉전략 수립 등을 통해 통계조사를 효율화할 필요가 있다. 외국에서는 파라데이터 수집을 통해 허위·부실조사를 방지하고 조사원 평가 기준 수립, 조사 기술 향상 개발 및 교육 등에 활용하여 조사효율화에 기여하고 있다.

특히 현 시점에서는 고품질 및 저비용의 통계 생산을 위하여 응답자 특성을 반영한 적응적·반응적 조사설계 방안 연구 및 도입을 검토해야 한다. 적응적·반응적 조사설계란 통계조사에서 오차와 비용은 상충관계에 있으므로 오차 최소화 목적 하에서 최적 비용 수준의 조사를 설계하는 방법으로 주로 파라데이터와 보조자료를 사용하여 설계하는

방법을 의미한다. 파라데이터 및 보조정보 등을 활용한 선진사례 방법으로는 ① 현장조사 지침 개선, ② 맞춤형 응답자 접촉전략 수립, ③ 설계조정방법, 품질 및 비용의 최적화 전략 연구, ④ 적응적·반응적 조사설계 현장 적용 방안 연구, ⑤ 현장 유용성 평가 등이 있으며, 이에 대한 지속적인 연구와 습득이 필요하다.

현재 세계 주요나라에서는 조사 디바이스의 사용빈도 증가와 기술향상, 응답자 부담 가중 등 급격한 조사환경 변화에 따라 대응하는 조사방식과 응답자 접촉전략, 응답자 부담을 감안한 조사설계방법 등 다양한 분야에서 연구가 활발하게 진행 중이다. 다만 급격한 변화에 개발, 개선, 연구, 적용 등의 과정이 시의적으로 대응하지 못하고 있는 실정이다. 특히 스마트폰의 사용량의 급격한 증가와 응답자 부담으로 인한 무응답의 가파른 증가는 조사방법론 자체를 흔들 수 있을 정도의 파급력을 가지고 있는 반면, 이에 대한 연구는 많이 부족한 실정이다. 본 연구에서는 현재 세계의 조사설계방법 연구의 동향을 온라인 조사, 스마트폰 활용, 응답자 접촉전략, 적응 및 반응의 조사설계, 비확률추출 기반의 표본 활용, 응답자 부담 경감을 위한 국가통계 재설계 등을 중심으로 살펴보았다. 그 중 온라인 조사, 적응 및 반응의 조사설계, 응답자 부담 경감을 위한 국가통계 재설계 등은 상당한 연구 성과가 나타나 있고 실전에 적용 가능한 수준으로 발전해 있는 분야도 많았다. 그러나 스마트폰 활용, 응답자 접촉전략, 비확률추출 기반의 표본 활용 등은 아직 초창기 연구의 모습을 띠고 있는 것이 사실이었다.

우리나라의 경우는 이와 같은 분야의 연구가 아직 시작도 하지 못한 실정으로 병행, 순차, 적응 등의 혼합조사 설계와 모드효과의 기초연구와 온라인 조사와 응답자 부담 경감을 위한 국가통계 재설계의 개별 통계부문에서 개발, 개선, 적용 등의 연구가 진행되어왔다. 우리나라의 조사환경도 급격하게 변화하고 있어 이에 대한 대응 연구가 신속하게 진행되어야 하고 중장기적으로 개발, 개선, 적용 등을 할 수 있도록 전략을 수립하여 할 것이다. 다만 조사환경은 각 나라마다 다르고 변화하는 모습도 달라서 다른 나라의 우수한 연구사례를 무조건적으로 받아들여서는 곤란하고 우리나라 실정에 맞는 방법을 검토하고 시험조사를 통해 지속적으로 검토할 필요가 있다. 특히 스마트폰 활용, 응답자

접촉전략, 적응적·반응적 조사설계는 우리나라 조사환경에서도 매우 중요한 부분으로 신속한 연구와 검토가 필요할 것이다.

적응적·반응적 조사설계에는 설계 특징이 다양하므로 조사 목적에 맞는 조합을 찾기 위해서는 충분한 시험조사나 검토가 필요해 보인다. 또한 업무 부담을 완화하고 효율적인 현장조사 운영을 위해서는 조사원이 개별적으로 결정하기보다 조사원 위치, 업무장소, 최선의 접촉 시간을 고려해서 조사원들에게 배정하는 방법에 대한 고민이 필요하다. 마지막으로 앞에서 언급한 것처럼 조사원 및 응답자의 부담감소를 위해 파라데이터를 시의 적절하게 활용하기 위해서는 시스템을 통해 바로 수집하고 확인할 수 있는 여건이 마련되어야 할 것이다.