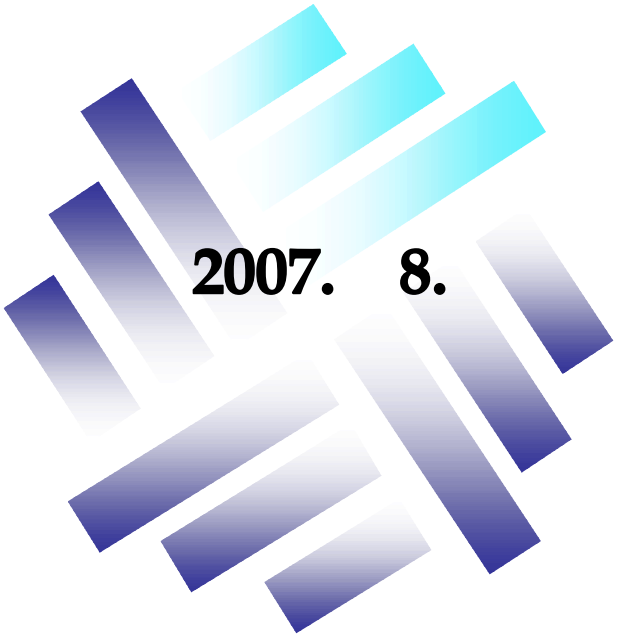
# 2007 Joint Statistical Meeting 참가 결과 보고



2007. 8.

통계개발원 연구기획실

# 2007년 미국 Joint Statistical Meeting 참가 결과 보고

## Ⅰ. 학술대회 참가 개요

□ 학술대회 일자 및 장소 : 2007. 7. 29~8. 2. 미국 Salt Lake

□ 학술대회명 : 2007 Joint Statistical Meeting (JSM)

□ 학술대회 목적

○ 경제, 사회, 생명과학, 조사통계 등 통계 전반적인 분야의 이론가들과 실무자들의 모여 최신 통계기법에 대한 정보를 공유하는 모임
○ 이론과 실용통계 분야 간의 다양한 의견 교환 및 선진 기법 동향을 파악할 수 있는 기회 제공
○ 국가 통계에 대한 응용적 측면을 포괄적으로 다루고 있어, 다양한 조사통계기관의 실무자들에게 통계이론에 대한 실용성을 검토 및 확인할 수 있는 장을 마련

□ 대회규모: 549개 Session에서 약 3500여편 이상의 논문 발표로 세계 통계학회중 내용과 규모면에서 가장 큰 학술대회임

□ 참석자 : 4급 김현중 (통계개발원 연구기획실장),
5급 김서영 (통계개발원 연구기획실)
6급 김황대 (통계개발원 연구기획실)

□ 출장 배경 및 목적
○ 조사통계분야에서 널리 활용되는 선진 기법에 대한 동향 파악
○ 통계개발원과 통계청에서 최근 관심이 높아지고 있는 분야에 집중적으로 연구동향을 파악하고자 함
○ 주요 분야로 조사기법, 소지역 추정, 인구센서스 방향 등을 고려함

## Ⅱ. 선진연구 동향 보고 내용

□ 2007JSM에서 발표된 내용을 크게 세 분야로 나누어 보고함

1. CARI (Computer Audio-Recorded Interview)조사기법
2. 소지역 추정 방법
3. 2011 캐나다 인구센서스 방향

□ CARI 방법 요약

- CAPI와 CATI의 대안으로 부각되고 있는 방법
- 자세한 내용 - <별첨 1>

□ 소지역 추정 기법

- 이론과 응용 측면에서 다양한 내용이 발표됨
- 선진 통계 국가일수록 소지역 추정에 대한 연구가 활발하게 이루어고 있는 것으로 파악
- 장기간에 걸친 연구 성과로 실제적으로 많은 조사분야에서 적용되고 있음
- 자세한 내용 - <별첨 2>

□ 2011 캐나다 인구센서스 방향

○ 비용절감을 위한 노력이 돋보임

- 인터넷 조사를 대폭적으로 확대 적용 계획
- paper 조사를 축소

○ 조사 및 공표에 있어서 시의성 고려

○ 2011 질문지 배송 3단계 전략

① 70%의 가구에 invitation letter 우송

- 주소 목록이 정확하고 인터넷 coverage가 좋은 지역에 배송
- 인터넷 사이트와 접속 코드를 알려줌

② 20%의 가구에 질문지 봉투 우송
- 주소 목록은 정확하나 인터넷 coverage가 좋지 않은 지역에 배송
③ 10%의 가구는 질문지 배송하지 않음
- 주소 목록과 인터넷 coverage 모두 좋지 않은 지역

○ 센서스 day 7일 후에 무응답자에게 질문지 배송
○ 센서스 day 이후 약 4주 후에 무응답 팔로업 시작
○ Contact: dave.dolson@statcan.ca
○ 자세한 내용 - <별첨 3>

## Ⅲ. 참가소감 및 시사점

□ 인터넷이나 학술논문을 통해서 획득할 수 없는 선진동향 기법 파악이 가능한 대회라고 판단됨
○ 이와 같은 대회에 지속적으로 연구 및 실무 인력들이 참가함으로써 국가 통계 발전에 기여할 수 것으로 기대함

□ 새로운 선진통계기법을 확보함으로써 국내 통계가 도약할 수 있는 계기가 될 것으로 기대
○ 청내 관련자들에게 새로운 통계기법에 대한 정보 제공 가능

□ 통계청의 연구결과를 세계 대회에 공표함으로써 통계청의 연구역량을 드높일 수 있을 것임
○ 선진통계국가의 경우 통계 연구에 대한 투자가 장기간에 걸쳐 꾸준히 이루어지고 있고, 오히려 국가 기관이 앞장서 연구에 투자함으로써 통계방법론 개발 및 활용성 검토가 진행되고 있음
○ 이처럼 통계기관의 연구역량 강화를 위해서 지속적으로 예산과 시간을 투입하고, 연구인력 확보를 위한 노력이 필요하다고 판단됨

**[별첨 1]**

# Computer Audio-Recorded Interviewing(CARI)

## Hosehold Wellness Study(HWS) Field Test

## □ 주요 내용

### ○ 연구의 목적

- 미국 Census Bureau가 CARI 방법의 적용가능성을 평가
- Household Wellness Study 현장 조사를 통해 검증
- 다양한 조사, 검정, 분석을 통한 사전 연구의 성격이 강함
- 이 연구의 핵심 목적은 미국 센서스의 모든 CAPI[1] 서베이를 CARI[2] 방법으로 바꾸기 위한 것임
- CARI는 현장재현과 응답자들간의 의사소통 내용 기록에 대한 평가를 가능케 함

### ○ 사전조사

- CARI는 조사 수행에 유용한 믿을만한 시스템
- 적절한 오디오 품질은 현장재현과 응답자간의 의사소통을 명확하게 기록함
- 3개의 30초간의 오디오 파일 리뷰는 인터뷰를 모니터하기에 충분함
- CARI는 Blaise(Computer assisted interviewing)와 견줄만함
- CARI 파일은 56K이상의 모뎀 사양을 요구함

### ○ CARI HWS Field Test 평가

- 3개 city를 대상으로 시험 조사 실시
  (필라델피아, 디트로이트, 캔사스)
- 인터뷰 기간: 2006.2.1-4.11.

---

1) CAPI: Computer Assisted Personal Interview
2) CARI: Computer Audio-Recorded Interview

- 두 개 오디오 파일 테이프
  - CARICON: Recorded consent question
  - Snippet: 30초 인터뷰 분량

## ○ 평가

○ 데이터 품질 상에서 CARI 효과,
- CARI 파일은 데이터 생산에 대한 변조가 발생되지 않음
- CARI 변조율 0 퍼센트
- 반응과 거부율 측면에서 적절한 수준
  : 총 응답율-81.4%, 거부율-15.3%

○ CARI 기록에 대한 오디오 품질
- High quality 오디오 파일 비율= 85.6%

○ 시스템 수행능력상의 CARI의 효과
- 하드웨어 수행능력 불량률 1.8%(447건 중 8건)
- recorder detection rate 0 퍼센트, 기록 불량은 없었음

○ CARI에 대한 응답자와 현장재현 반응
- 총 467응답 중 완전협조율-88.7%, 부분협조율-3%, 거부율-8.4%

## ○ 본 연구의 한계점

- 3개 city에 한정된 조사라는 점
- 스크린 rule이 없었음
- 면접자 training 에 대한 이슈
- 긴급성 측면에서 덜 민감

## ○ 본 연구의 결과

- 데이터 생산에 있어서 오디오 파일 변조가 없음
- 면접자 훈련이 체계적이지 않았기 때문에 오디오 품질에 다소 문제
- CARI는 CAPI방법에 비해 기술적 문제를 일으키지 않음
- 응답자는 CARI방법에 매우 긍정적 반응
- 현장재현 측면에서는 다양한 반응 -> 추후 연구가 더 필요

## ○ 미국 Census Bureau의 CARI에 대한 기대효과

- CARI는 보다 효과적인 샘플링 방법을 가능케 할 것임
- 또한 Census Bureau의 표준적인 data quality를 유지하기 위한 re-interview 프로그램으로서의 역할 수행 가능

## ○ CARI 적용 및 평가

- CARI 방법을 적용하기 위해서는 응답자의 동의를 얻어야 함
- 면접자와 응답자와의 의사교환을 통한 인터뷰 절차를 있는 그대로 효과적으로 기록함
- CAPI의 대안으로 CARI를 검토하기 위해 다음 사항을 체크함
  - ○ data 품질 측면에서 CARI 효과 평가
  - ○ 면접이 진행되는 동안 내용을 기록할 오디오 품질 측정
  - ○ 시스템 수행능력 측면에서 CARI 효과 결정
  - ○ CARI 사용에 대한 응답자의 반응 확인

## ○ CARI에 대한 전반적인 평가

- ○ 대체적으로 CARI 사용에 대한 긍정적인 평가를 내림
- ○ CAPI의 대안으로 사용가능할 것으로 판단
- ○ 특히, 현장조사에 대한 관리가 가능하고, 현장에서 면접을 수행하는 면접자 및 면접과정을 확인하고 그 결과에 대해 피드백이 가능하다는 측면에서 매우 우호적인 결과가 나옴
- ○ 그러나 이번 CARI적용은 미국 3개 주에만 제한적으로 시행한 결과로서 일반화화 시키기에는 한계가 있음
- ○ 따라서 미국 Census Bureau는 CARI의 현실가능성을 보다 구체적으로 검토하기 위해 본 pre-study 개념의 연구를 확대시켜 계속 연구해 나갈 계획임

## ○ 요약 및 시사점

- CARI 방법은 CAPI의 대안으로서 현장조사시 나타나는 문제점을 해결하는데 도움이 될 것으로 기대됨
- 단지, 현장에 투입되는 면접자를 관리하는 차원을 넘어서, 면접자 및 응답자가 느끼는 조사표의 상황, 면접과정의 어려움 등을 피드백함으로써 개선 방향을 찾을 수 있을 것으로 기대함
- 현장조사를 통해 많은 자료를 수집하는 통계청의 경우, 현장조사 개선을 통한 데이터 품질 향상을 위해 많은 노력을 기울이고 있는 상황임
- 이와 관련해서 2007년 통계개발원에서는 상반기 3개 토픽을 선정하여 현장조사 개선과 관련하여 연구가 진행되었거나 진행 중에 있음
- CARI의 국내의 적용가능성 검토는 비용 측면이나 현장조사 개선을 통한 데이터 품질 향상 측면에서 매우 긍정적인 방향으로 사료됨

**[별첨 1- 추가설명자료]**

# CARI 개발 배경 및 기본 내용

[RTI : Research Triangle Institute]

## 1. Introduction

### ○ CARI 기본 배경

- RTI 컴퓨터학자들에 의해 개발된 laptop computer 소프트웨어
- 이는 면접자가 CAPI 조사표(questionnaire)를 작성할 때 정교한 tape recorder의 역할을 함
- CARI는 면접이 진행되는 동안 면접자와 응답자간의 의사소통 내용을 솔직하게 기록함
- 이 시스템은 이미 시스템 내에 내장되어 있는 장치에 의해 작동하는 전적으로 소프트웨어 제어 하에 있고, 기록은 모든 면접자 또는 랜덤하게 선택한 인터뷰에 대해 switch on/off 가능
- 이 기능은 전화 인터뷰 내용을 모니터링하기 위해 사용되는 CATI[3]와 유사함

### ○ CARI 사용 범위

- 인터뷰 내용 위조와 에러를 찾아내기 위해,
- 면접자 능력을 평가하고 이것을 면접자에게 feedback 하기 위해, 그리고
- 조사표의 문제점을 찾고, 면접자와 응답자와 상호작용 체계에 대해 오디오에 기반한 정보를 수집하기 위한 수단으로 사용됨

3) CATI: Computer Assisted Telephone Interview

○ CARI 연구 목적

- RTI와 Census Bureau가 CATI 대안으로 CARI의 실용가능성 검증 연구 착수
- 다음과 같은 측면에서 CARI 실용성을 검토함

ㅇ 다양한 CARI 적용성에 대한 오디오 품질
ㅇ 현장 면접자들의 CARI에 대한 반응
ㅇ 비용측면
ㅇ 조사 응답자들에 의한 CARI 반응
ㅇ 현장조사 수행능력을 평가하고 면접자들에게 이 결과를 피드백하기 위한 수행능력 모니터링 절차

## 2. CARI 실용성 평가

- 다양한 연구절차와 검증 방법을 통해 CARI의 실용성 평가

### □ CARI 오디오 품질 평가

○ 소리 품질 평가 결과 기술적으로 실용 가능한 것으로 판단됨

### □ CARI에 대한 면접자 반응 평가

○ 조사 면접 환경을 평가하는 것이 중요함
○ 이에 대해 면접자 보고용 질문지를 개발하여 측정
○ 대체적으로 우호적인 반응이었으나, 경험이 많은 면접자의 경우 다소 비우호적임

### □ CARI의 일반 조작상의 실용성 평가

○ CARI 조작은 큰 어려움 없이 매우 성공적임

- 랩탑 하드 드라이버 공간 소모,
- 서버 저장 공간 소모,
- 조작상의 파일 변질 등의 측면에서 평가함

□ CARI 수행 비용

○ 전통적인 다른 방법에 비해 저렴함

□ CARI에 대한 응답자 반응 평가

○ 응답자들은 CARI에 대해 매우 긍정적인 반응

○ CARI는 응답자들의 응답에 별로 영향을 주지 않음

- 응답자 의견으로부터 어느 정도 데이터 품질을 향상시킬 가능성이 큼

□ CARI 모니터링으로 면접자 수행능력에 대한 피드백 제공

○ 다양한 연구 디자인을 통해 검토

- 충분히 면접자들의 면접과정 및 능력을 모니터링 할 수 있음

## 3. 요약 및 기대효과

○ CARI는 조사 품질 향상 목적으로 사용 가능

- 면접 위조, 면접자 능력 모니터링 등을 포함

○ CARI 음성 품질은 다른 테이프 리코더와 유사하거나 더 나음

○ 면접자들은 면접과정을 모니터하는 목적으로 CARI를 사용하는 것에 우호적이거나 최소한 중립적인 반응을 보임

- 그러나 경험이 많은 면접자들은 CARI에 대해 비우호적임

○ 비용분석 결과 다른 방법들에 비해 저렴함

- 그러나 면접자 능력과 피드백에 대한 비용은 포함하지 않음

○ Supervisor는 면접과정 동안 면접자의 수행에 대한 정보를 얻는 것에 관심이 많은 것이 일반적임

- CARI는 다른 방법에서는 체크할 수 없는 면접자 능력에 대한 정보를

제공할 수 있음
- 그러나 오디오 파일 제공에 있어서 우편 시스템을 사용하기 때문에 전화 면접보다 피드백이 다소 늦은 감이 있음

○ CARI Key point
- **CARI는 면접 내용 위조 방지를 위한 모니터링, 면접자 능력 모니터링 측면에서 CAPI 서베이보다 훨씬 유용함**

※ Contact address:

○ Census Bureau, Dr. Sherry
Email: sherry.e.thorpe@census.gov

○ RTI International
http://www.rti.org/index.cfm

**[별첨 2]**

# Small Area Estimation

## Techniques and Applications

## □ 보고 내용 및 방법

○ 2007년 미국 JSM 에서 많은 관심을 받은 분야에 해당

○ 이론 개발과 실제 적용 측면에서 다양한 내용이 발표됨

○ 특히, 새롭게 개발된 소지역 추정 이론 및 방법론은 서베이 자료를 통해 그 활용성을 검증하였고, 국가 서베이 기관의 소지역 연구에의 참여가 두드러졌음

○ 응용적 측면에서 연구·발표된 내용들은 대부분이 이미 개발된 이론을 서베이 목적에 맞게 성공적으로 적용함으로써 정확한 소지역 또는 영역별로 추정치를 산정한 결과가 발표되었음

○ 이론적 기법 측면에서는 최근 Issue가 되고 있는 추정치에 대한 분산 추정에 있어서 신뢰구간을 사용하는 topic이 주를 이루었음
- 대표적으로 JSM의 초청 Session에서 발표된 내용을 소개함 (이분야의 권위자인 Dr. Rao, Dr. Lahiri의 발표 내용임)
- 최근 Issue가 되고 있는 소지역 추정에서의 정도(precision)추정에 대한 구간추정 방법의 유용성 소개가 핵심

○ 본 보고서는 JSM에서 발표된 소지역 추정에 관한 발표 내용 중 핵심적인 몇 개의 논문을 주제별로 간추려 보고함
- 신뢰구간 추정에 관한 토픽은 별첨 특별주제로 다룸

# Part Ⅰ. 실제 조사 자료에의 적용

## 1. 성인문맹 정도의 계층적 베이즈에 의한 소지역 추정

### ○ 성인 문맹에 대한 국가 평가

- NAAL[4] 은 미국 성인의 영어의 읽고 쓰는 능력을 측정하기 위해 미국 교육통계 센터에 의해서 설계되었음
- NAAL 데이터를 사용하여 국가와 주요 subdomain 단위에 대해 정확한 추정치가 계산됨
- 그러나 정책 입안자나 연구자 또는 사업가들은 더 작은 단위(county, state)별로 국민 문맹정도에 대한 정보를 요구하고 있음
- 그러나 NAAL 데이터는 이들 소지역에 대해 믿을 만한 정보를 제공할 만큼 샘플이 크지 않음
- 따라서 소지역 추정 기법을 사용하여 국가 내 state/county 단위별로 문맹수준에 대한 추정치를 사용하여 정보를 제공하고자 함

### ○ 추정기법

- Hierarchical Bayesian(HB) 추정기법을 사용함

  → county와 state-level 에서 single small area-level model 유도
- sampled county와 nonsampled states에서 sampled county 와 nonsampled county들에 대한 소지역 추정치를 계산하기 위해 MCMC 방법을 사용함
- 소지역 추정치들에 대한 정도(Precision)을 측정하기 위해 신뢰구간을 사용함

### ○ 결론

- 현재 NCES는 추정치들을 검토 중에 있으며 NCES web site를 통해 공식적 release 여부를 고려하고 있음

---

4) NAAL: National Assessment of Adult Literacy

## 2. 심리적 고통의 소지역 추정을 위한 누적 분포함수 조정방법 적용

### ○ 개념 및 방법

- NSDUH[5] 는 약 67,500 표본 크기를 갖는 연방국가 조사임
- 이는 소지역 추정방법을 사용하여 state-level 추정치를 제공하고 있음
- 심각한 심적 고통은 NSDUH에서 k6 척도(정신상태 측정을 위한 질문지 형태)에 의해 측정되었음
- 2003년에는 K6 질문에 앞서 정신건강 질문이 선행되었음
- 2005년에는 K6 질문지만 사용되었음
- 2004년에는 샘플이 2003년과 2005년의 long and short module로 분리되어 사용되었음
- 2004년 결과는 심적 고통 유병률(prevalence)에 대해 두 모듈간에 큰 차이를 보였음

### ○ 추정방법

- NSDUH 소지역 방법은 2년간의 결합 데이터를 필요로 하기 때문에 누적분포함수 조정방법이 고안되었음

→ 이 방법으로부터 2004년 한 모듈에 의한 심적 고통 추정치는 다른 모듈에 의한 심적 고통 추정치에 의해 조정될 수 있음.

- 이 방법은 state-level 추정치를 얻기 위해 2003/2004, 2004/2005 데이터를 combine 을 허용함

5) NSDUH: National Survey on Drug Use and Health

## 3. Iowa주에서 고등학생 서베이에 대한 소지역 추정

### ○ 개념 및 방법

- Iowa주 교육위원회는 층화된 다단계 샘플 서베이를 실시
  → 취업 준비과정과 그 과정에 Iowa주 public 고등학교 학생들의 참여 정도의 유효성 연구가 목적임
- 주어진 예산 내에서 small sample 이 설계되었고, 이는 직접추정치를 사용할 경우 높은 변동의 원인이 됨

### ○ 추정기법

- 계층적 베이지안 (HB) 분석이 사용되었음
- 이는 변동에 대한 안정성을 증가시키기 위함
- 층의 수가 증가할 때 샘플 디자인은 모든 층에 표본을 적절하게 할당할 수 없게 되고, 이 경우 HB 기법은 적극 추천되는 방법임
- 추정에 대한 향상을 위해서 보조변수들과의 관계도 고려되어야 함

# Part Ⅱ. 이론 기법 측면

## □ 주요내용

- Resampling methods in Small-Area Estimation

## 1. 잭나이프 방법과 붓스트랩 방법: 응용성

### ○ 내용 및 기법

- 잭나이프 방법과 붓스트랩 방법은 최근 소지역 추정 연구에서 자주 등장하고 있음
- 특히, 평균제곱분산(mean squared error:mse)추정과 신뢰구간 측면에서 강조되고 있음

- Dr. Rao는 resampling 기법들에 대해 전반적으로 그 활용성과 응용성 측면에서 소고함
- 추가적으로, 비선형 소지역 모형을 위한 잭나이프 방법에 대한 결과를 소개하고 특히, 지역 특성을 고려한 mse 추정량들을 소개함

○ 본 연구는 아직 연구단계에 있는 관계로 추후 다시 소개하기로 함

## 2. 소지역 추정 신뢰구간 문제

### ○ 내용 및 기법

- 경험적 최량선형 불편추정량(Empirical best linear unbiased prediction: EBLUP)은 다양한 다른 정보 소스로부터 정보를 결합하기 위해 선형 혼합 모형(linear mixed model)을 사용함
- 이 방법은 특히 소지역 문제에 유용함
- EBLUP변동은 평균제곱예측오차(mean squared prediction error:mspe)에 의해 추정되었고, mspe 추정치를 사용하여 구간 추정치가 계산됨
- 그러나 일반적인 이런 방법은 under coverage, over coverage, 해석상의 부족 등 단점을 포함함
- 재표본(resampling) 방법은 이러한 문제를 극복할 수 있는 대안으로 떠오르고 있음

○ 본 주제는 실제 응용측면에서 매우 중요한 기법으로 <특별주제>로 자세하게 설명하였음

## □ 요약 및 시사점

### ○ 최근 다양한 조사 분야에서 소지역 추정 이용이 증가하고 있음

- 이는 제한된 예산 범위 내에서 조사 목적을 달성하기 위해서는 small sample을 사용해야 함.
- 그러나, small sample 조사에 의해 직접 추정치를 사용하는 것은 정확

성 측면에서 다소 신뢰도가 떨어지는 문제가 발생함
- 따라서 이러한 문제를 극복하기 위한 대안으로 소지역 추정치를 사용하게 됨

**○ 소지역 추정치를 사용할 경우, 다양한 문제가 산재되어 있음**
- 첫째, 가장 기본적인 것은 어떤 추정기법을 쓸 것인가
- 둘째, 어떤 보조변수를 쓸 것인가
- 셋째, 가중치 조정을 어떻게 할 것인가
- 넷째, small area 또는 domain의 특성을 얼마나 고려할 수 있는가

=> 이 모든 Issue는 small area 추정치에 대한 신뢰도를 높일 수 있는 핵심 연구 분야에 해당됨

**○ 선진통계 국가일수록 소지역 추정기법 연구 활발**
- 소지역 추정은 조사여건 등을 고려할 때 대세가 될 전망
- 이미 선진 통계 국가들은 다양한 조사 분야에서 분야의 특수성 및 지역 특수성을 고려한 분야별 지역 모형을 개발하고 있음
- 선진통계국가의 경우, 소지역 추정연구는 국가통계기관을 중심으로 수년간에 걸쳐 학문적 연구를 바탕으로 실용성을 검토하는 방식으로 이루어진 성과임

**○ 국내의 경우, 소지역 추정 기법에 관한 연구 착수 및 진행**
- 계속적인 전문 인력 육성에 대한 노력이 다소 미흡
- 시간, 예산, 인력을 집중 투입하여 단계적으로 소지역 추정에 대한 준비를 할 필요가 있음

## Reference


○ Hierarchical Bayes Small Area Estimates of Adult Literacy Using Unmatched Sampling and Linking models, Leyla Mohadjer, J.N.K. Rao et al., 2007 JSM presentation.

○ Resampling methods in Small-Area Estimation, Annals of Applied statistics, S. Chatterjee, P. Lahiri, H. Li (2007).

**[특별주제]**

# 소지역에서의 예측 구간 문제

by Snigdhansu Chatterjee, Parthasarathi Lahiri, Huilin Li

## 요 약

EBLUP 방법은 서로 다른 정보를 결합한 Linear mixed model을 이용한다. 이 방법은 특히 소지역 문제에서 유용하다. EBLUP의 변이는 평균제곱예측오차(mean squared prediction error:MSPE)에 의해 측정되어왔고, 구간 추정은 일반적으로 MSPE 추정량을 이용하여 구해졌다. 이런 방법들은 과소포함, excessive length, 해석가능성의 부족(lack of interpretability)과 같은 단점을 가지고 있다. 본 연구는 재샘플링(resampling) 접근 방법을 제안하고 $d$가 모수의 개수이고 $n$이 관측치의 수일 때, $O(d^3 n^{-3/2})$의 포함 정확성(coverage accuracy)을 얻는다. 모의실험은 제안된 방법이 기존의 방법들에 비해 coverage 면에서 우수한 결과를 보였다.

## 1. Introduction

○ 소지역 추정을 위해서는 다른 행정자료나 센서스 자료로부터 얻어지는 관련 자료를 보충해야 할 필요가 있음

○ 많은 소지역 적용에서 혼합선형모형(mixed linear model)들은 기계적으로 다양한 출처의 그리고 오차의 원인을 설명할 수 있는 정보들을 결합하여 이용되고 있음

○ 이러한 모형들은 “소지역 사이의 변이”를 설명하는 특정 지역 random effect를 포함하지만 모형의 fixed effect에 대한 부분은 설명하지 않음

○ 소지역 연구에 대한 리뷰 문헌 : Rao(1986, 1999, 2001, 2004), Ghosh and Rao(1994), Pfeffermann(2002), Marker(1999), Rao(2003) 등이 있음

○ EBLUP을 이용한 점 예측(point prediction)과 관련된 평균제곱예측오차(MSPE) 추정은 소지역 문헌에서 광범위하게 논의되어져 왔지만, 구간 예측

문제는 그만큼의 진전이 없었음

○ 유효한 방법 안에서 점 예측과 가설 검정 개념의 통합 이래로 구간 예측은 유용한 자료 분석 툴임.

○ 예측 구간은 소지역 연구에서 많은 부분에서 유용함

- 예를 들면 서로 다른 county가 비슷한 자원과 필요를 갖거나 혹은 다른 인종 혹은 다른 소집단 그룹들이 특정한 질병에 동등하게 노출되었다든지 하는 경우를 입증하는데 예측 구간이 도움을 줄 수 있음

○ 소지역 문헌들에서는 때때로 예측구간은

표준적인 $EBLUP \pm z_{\alpha/2}\sqrt{mspe}$ 규칙($mspe$가 EBLUP의 실제 MSPE의 추정값)을 이용하여 생성

- 이러한 예측구간은 포함 확률이 큰 표본 사이즈 $n$에 대해서 $1-\alpha$로 근사적으로 수렴한다는 관점에서는 정확.
- 그러나 그들이 특정한 MSPE 추정량의 선택에 의존하는 small $n$에 대하여 과소포함 혹은 과대 포함 문제를 가지고 있다는 관점에 대해서는 충분하지 않음.
- 통계적인 관점에서 그런 구간의 포함오차는 order $O(n^{-1})$,
- 대부분이 small $n$을 가지는 소지역 연구에 적용하기에는 충분하지 않은 정확성임.
- 만약 모형의 모수 추정값에 포함되어 있는 불확실성에 대한 설명을 하지 않는 naive $mspe$가 사용된다면 그 결과로 나온 예측 구간은 과소포함 문제가 발생
- 반면 Prasad-Rao(1990)의 second-order unbiased MSPE추정량이 사용되어진다면 구간은 과대포함과 over-lengthening 문제 발생

○ 일반혼합선형모형을 위하여, Jeske와 Harville(1988)은 혼합효과(mixed effect)에 대한 예측 구간을 제안

- 그러나 이 방법은 분산 성분의 추정에 의해 야기되는 복잡성(complexity)을 처리하지 않음.
- 특히 그들이 제안한 구간의 포함 오차의 정확성과 관련하여 추정된 미지의 분산 성분의 효과에 대하여 연구하지 않음

○ 위에서 인용되어진 논문들 외에도, 소지역 예측 구간에 관한 연구는 잘 공표되어진 혼합회귀 모형인 Fay-Herriot 모형의 몇 가지 특별한 경우를 제외하면 제한적임

○ 베이지안의 다양성(variety)과 경험적 베이지안 방법은 그러한 구간을 위한 Fay-Herriot 모형에서 사용되어짐

○ Fay-Herriot 모형과 그것의 특별한 경우들의 구간추정의 광범위한 사용 때문에 이 모형의 구간 추정에 대한 다른 접근들을 리뷰함.

○ 소지역을 위한 일반선형혼합모형 : $\boldsymbol{Y}=\boldsymbol{X}\beta+\boldsymbol{Z}v+\boldsymbol{e_n}$

- 혼합 ANOVA모형, Fay-Herriot 모형을 포함하는 다시점(longitudinal) 모형, 그리고 nested오차회귀모형은 위의 모형의 특별한 경우에 해당

○ 이 논문은 일반선형혼합모형의 혼합효과의 예측구간을 얻는 문제에 초점을 맞추고 **모수적 붓스트랩 방법**을 이용한다는 것이 핵심

- 이것은 Fay-Herriot 모형과 같은 특별한 경우에만 적용 가능한 기존의 방법에 비해 구간에 대하여 좀 더 높은 차원(order)의 포함 정확성(coverage accuracy)을 가짐
- 외생/또는 내생의 요인에 의존하는 건강, 경제 활동 그리고 다른 인간 복지는 많은 경우 개인 수준에서 측정되어져야 하고 모형에서 혼합되어져야 한다고 알려져 있음
- 이것은 위 혼합모형에서 모수 $\beta$와 $\psi$의 높은 차원으로 해석.
- 소지역 예측의 차원 접근성 면을 다루기 위하여 우리는 모수 차원 $d=p+k$를 표본 크기 $n$으로 늘이고 $O(d^3n^{-3/2})$ 차원의 포함 정확성을 얻는 것으로 참작

○ 전통적인 소지역 구간 추정은 Fay-Herriot 모형이나 그것의 특별한 경우에, 고정된 $p$와 $k$에 대하여, $O(n^{-1})$을 성취

- 그러나 $O(n^{-1})$은 전형적으로 관련 소지역 추정에 대한 후속 적용에 적당하지 않음
- 그러므로 이 구간들에 대한 보정(calibration)과 수정(correction)이 제안되어졌음

○ 제안한 방법과 기술적 조건에 대한 명확함과 향후 적용에 대한 쉬운 이용을 위하여 Fay-Herriiot 모형에 대한 제안되어진 예측 구간을 자세하게 논의

○ 본 논문을 통해서 일반적인 Das, Jiang, Rao 모형(1)에 대한 예측구간 알고리즘을 제시

○ 제안된 방법은 두개의 다른 기존의 방법(Monte Carlo 모의실험을 이용한)과 비교 실험
- naive MSPE 추정량 : 비교 추정량들 중에서 가장 짧은 구간을 제공하지만, 심각한 과소 포함 문제가 있음
- Prasad-Rao 방법은 넓은 구간을 제공하는 대신 과대 포함 문제가 있음
- 제안된 방법은 단지 coverage에 대한 것이며 length의 관점에서 Prasad-Rao 방법보다 나음

## 2. Fay-Herriot 모형

○ 아래의 Level1과 Level2의 결합모형
- Level1: $\theta=(\theta_1,...,\theta_n)^T$가 주어진 하에서 $\boldsymbol{Y}=(Y_1,...,Y_n)^T \sim N(\theta,\boldsymbol{D})$

  $D_{ii}=\sigma_i^2$, $\sigma_i$ : 기지
- Level 2: $\theta \sim N(\boldsymbol{X}\beta,\tau^2)$, $\tau^2$ : 미지

○ 본 연구의 관심 : $\theta_i = x_i^T\beta + v_i$의 예측 구간

① Level1만 이용했을 때 : $Y_i \pm z_{\alpha/2}\sigma_i$
- 특성 : 포함확률은 정확히 $1-\alpha$, 그러나 average length가 받아들이기에 너무 큼

② Level2만 이용했을 때 : 지역특성을 무시하게 됨
- It fails to be relevant to the specific small area under consideration
- 충분한 포함 정확성(coverage accuracy)을 얻는데 실패

따라서, Level1과 Level2가 결합된 모형에 대한 추정이 필요, 그 중 가장 유용한 방법은 경험적 베이지안 방법

## 2.1 Fay-Herriot 모형에서의 경험적 베이즈 구간

○ $\theta_i$의 EB : $\widehat{\theta_i^{EB}} = (1-\widehat{B}_i)y_i + \widehat{B}_i \boldsymbol{x_i^T}\hat{\beta}$

○ $\theta_i$의 예측 구간(Cox(1975)) : $I_i^C(\alpha) : \widehat{\theta_i^{EB}} \pm z_{\alpha/2}\sigma_i(1-\widehat{B}_i)^{1/2}$ ⇒ 포함 오차 $O(n^{-1})$

∵ 초모수 $\beta$와 $\tau^2$ 추정에 의해 발생하는 추가적인 오차 때문.

- Morris(1983) : $\sigma_i = \sigma$ 일 경우 연구
- Morris(1983) : HB에서의 예측구간 연구

○ Basu, Ghosh, Mukerjee(2003) : Morris의 연구가 Cox의 연구를 발전시키지 못했음을 증명하면서 테일러 전개를 이용하여 Morris의 구간을 포함오차 $o(n^{-1})$까지 조정시켰음

- Carlin-Louis(1996)의 구간에 대한 포함오차 $o(n^{-1})$
- $o(n^{-1})$의 포함 편의를 갖는 새로운 예측 구간의 명시적 형태를 알아냄

○ Datta, Ghosh, Smith, Lahiri(2002) : 포함 오차 $O(n^{-3/2})$

## 2.2 경험적 베이즈 구간에서의 붓스트랩의 이용

○ 붓스트랩 방법 : naive EB의 신뢰구간을 개선하기 위해 사용

- 붓스트랩 샘플 생성방법과 수정(correction) 형태가 다름

○ iid(identical independent distribution)인 $Y_i$(Fay-Herriot 모형)에 대해 Laird, Louis(1987)는 세 개의 다른 방법을 제안

- 이들은 포함된 모수에 대한 가정의 degree가 다름.
- 비모수와 준모수 방법과 관련한 문제는 EBLUP의 분포의 붓스트랩 접근이 일반적으로 일치(consistent)하지 않음
- Laird-Louis Type Ⅲ 붓스트랩은 보통의 모수적 붓스트랩
- 좀 더 정확한 포함 오차를 얻기 위해 HB 접근을 모방

○ Calibration of interval은 붓스트랩의 주요 활용법 중의 하나이고 커버리지 정확성에서 주목할 만한 향상을 이끌어냄
○ 편의 수정과 연결하여 pivotal의 사용, 그리고 Edgeworth 수정, calibration으로부터의 rrotjs은 때때로 드라마틱함

○ Hall(2006) : 비모수 붓스트랩 신뢰구간의 적용 제안
○ 조사에서 robustness는 항상 중요한 이슈이고 연구자(practitioner)들은 항상 능률적인 비모수적 방법에 관심이 있음
- 그러나 소지역에서의 부족한 데이터 때문에 비모수 추정량은 때때로 심하게 제 기능을 다하지 못하는 경향이 있음
○ 이것은 비모수적 모형이 일반적으로 합성 모형이나 회귀 모형에 기반을 둔 붓스트랩의 히스토그램의 생성을 허용하고 주어진 데이터에 대한 조건분포의 근사는 허용하지 않기 때문임
○ 결과적으로 비모수적 붓스트랩 예측 구간은 지역의 특정한 데이터를 underweight하는 경향이 있음

○ 지역의 특정한 데이터를 정확히 가중하는 것은 좋은 커버리지 특성을 얻는데 중요함

## 2.3 모수적 붓스트랩 예측 구간

**○ Calibration이나 사전분포 대안으로 붓스트랩을 이용하는 대신, 이 연구는 붓스트랩 히스토그램으로부터 직접 예측구간을 구함**
- 비모수, 준모수적 붓스트랩 방법은 consistency 문제 때문에 사용하지 않음

## 3 일반 선형 혼합 모형을 위한 모수적 붓스트랩 예측 구간

○ 선형혼합모형을 사용하여 다양한 모수 추정
- 신뢰계수에 대한 현실적인 가정 사용
- 소지역 개수가 많은 경우에 대비하여 보다 현실적인 응용성을 고려하여 디자인 됨
- 구체적 수리적 과정은 생략하기로 함(첨부 파일 참고 가능)

## 4 수입과 빈곤 통계에 기반을 둔 모의실험

○ 50개주와 District of Columbia에 대한 income and poverty 추정량을 이용
○ Fay-Herriot 모형 이용
○ 비교 방법
  (a) naive plug-in method
  (b) Prasad-Rao mothod
  (c) 이 연구에서 제안한 방법 : 모수적 붓스트랩(2.3 참조)
○ Coverage and average length 비교

○ **연구에서 제안한 붓스트랩 방법은 구간의 과도한 lengthening이 없으면서 커버리지 정확성이 높아짐**

# Reference


[1] Basu, R., Ghosh, J.K., and Mukerjee, R., (2003), Empirical Bayes prediction intervals in a normal regression model: higher order asymptotics, *Statist. Probab. Lett.*, 63, 197-203.

[2] Dar, K., Jiang, J., AND Rao, J.N.K., (2004), Mean squared error of empirical predictor, *Ann. Statist.* 29, 139-152.

[3] Hall, P. (2006), Discussion of Mixed model prediction and small area estimation, *Test*, 15, 1-96.

[4] Harris, I.R. (1989), Predictive fit for natural exponential families, *Biometrica*, 76, 675-684.

[5] Hill, J.R. (1990), A general framework for model based statistics, *Biometrica*, 77, 115-126.

[6] Rao, J.N.K. (1986), Synthetic estimators, SPREE and the best linear model based predictors, *Proceedings of the Comference on Survey Methods in Agriculture*, 1-16, U.S. Dept. of Agriculture, Washington,D.C.

[7] Rao, J.N.K (1999), Some recent advances in model based small area estimation, *Surv. Meth.* 25, 175-186.

[8] Rao, J.N.K (2001), EB and EBLUP in small area estimation, in *Empirical Bayes and Likelihood inference, Lecture notes in Statistics No.* 148 (ed. Ahmed,S.E.and Reid,N.), Springer, New York.

[9] Rao,J.N.K., (2003), *Small Area Estimation*, Wiley, New York.

**[별첨 3]**

# 2011년 캐나다 인구센서스 전략 방향

## □ 주요 내용

### ○ 연구의 목적

- 캐나다 통계청(Statistics Canada)에서 2006년 캐나다 인구센서스의 방법론을 개선하여 2011년 캐나다 인구센서스에 적용하기 위한 것임

### ○ 2006년 센서스 요약

- 조사방법
  - ○ 우편조사: 주소록이 정비된 70%의 가구는 우편으로 조사표류를 발송하고, 주소록이 정비되지 않은 나머지 30%의 가구에 대해서는 조사원이 직접 방문하여 전달
  - ○ 조사표 회수 시 우편 또는 인터넷 이용
  - ○ 무응답 가구: 센서스 데이 3주후 조사원(27,000명) 투입

- 인터넷 응답
  - ○ 응답자의 23%(전체적으로 18.3%)가 인터넷으로 회신

- 인터넷 응답률이 높은 이유
  - ○ 고속 통신망을 포함한 캐나다의 인터넷 기반이 잘 되어 있음
  - ○ 인터넷을 이용한 소득세 환급 요청, 공과금 납부 및 은행 거래 등 캐나다 국민들의 인터넷 사용 경험
  - ○ 센서스 홍보 시 인터넷을 통한 응답 권장

- 조사결과
  - ○ 인터넷을 통한 응답이 우편 응답보다 질이 높은 것으로 나타남
  - ○ 전수 조사표의 경우 인터넷을 통한 응답 착오율이 우편 응답 착오율 보다 50% 작음
  - ○ 표본 조사표의 경우 인터넷을 통한 응답 착오율이 우편 응답 착오율의 1/6 수준임

## ○ 2011년 센서스의 전략 방향

- 조사비용 절감
- 조사원수 감축
- 조사방법
  - ○ 인터넷 조사: 주소록 정비가 되어있고 고속 인터넷 망이 잘 되어 있는 지역(전체 가구의 70%)
  - ○ 우편조사: 주소록 정비가 되어있지만 고속 인터넷 망이 잘 되어 있지 않은 지역(전체 가구의 20%)
  - ○ 나머지 10% 가구에 대해서는 조사원이 조사표 직접 전달
  - ○ 센서스 데이 1주일 후 무응답 가구에 대해 조사표류 우편 발송
  - ○ 센서스 데이 4주일 후 무응답 가구에 대해 면접조사

## ○ 2011년 센서스 주요 일정

| | |
|---|---|
| - 조사 항목에 관한 의견 수렴 | 2008년 5월 |
| - 시험조사 | 2009년 5월 12일 |
| - 조사항목 확정 | 2010년 4월 |
| - 센서스 용품 제작 | 2010년 5월 |
| - 거처명부 작성 시작 | 2010년 8월 |
| - 조사표 배달 시작 | 2011년 4월 25일 |
| - 2011 센서스[기준일] | 2011년 5월 10일 |
| - 조사표 수집 완료 | 2011년 6월 29일 |
| - 인구 및 거처 수 공표 | 2011년 12월 13일 |

- 주요 결과 공표 2011년 12월~2012년 11월

## ○ 요약 및 시사점

- 2011 캐나다 인구센서스의 가장 큰 흐름은 비용절감에 있음
- 따라서 비용절감을 위해 다양한 방법 및 절차를 고려하고 있음
- 대표적인 경우가 질문지를 이용하는 직접 조사보다는 인터넷 서베이를 확장하고자 하는 노력이 돋보임
- 인터넷 서베이로의 방향 전환은 2006년 우편조사에서 약 23%가 인터넷으로 회신한 결과에 바탕을 둠
- 이와 같은 방향 전환은 현장조사를 위해 우수한 인력확보가 어려운 절박한 상황을 극복하고자 하는 전략이기도 함

- 현장조사의 어려운 상황과 막대한 비용 소요라는 측면에서는 캐나다와 우리의 경우는 비슷한 실정임
- 캐나다의 경우, 인구센서스 조사 방법 전환을 위한 다년간의 연구를 진행해 왔으며 이를 단계적으로 본 센서스에서 적용해 봄으로써 그 효과를 평가하고 점점 어려워지는 조사환경에 대비하고자 노력하고 있음
- 이처럼 미래에 대비한 보다 상세한 인구센서스 전략이 우리에게도 필요하며, 이를 위해 연구, 실무기관간의 적극적인 노력이 필요하다고 생각됨

**[별첨 3-추가설명자료]**

# 2011년 캐나다 인구센서스 전략 방향

[Strategic Directions for the 2011 Canadin Census of Population]

## □ 발표내용

1. 2006년 센서스 방법
2. 2006년 센서스의 교훈
3. 2011년 센서스의 drivers & 목표
4. How
5. Priorities
6. 주요 일정

## 1. 2006년 센서스 요약

○ 1350만 거처에 사는 3160만 인구를 전수/표본 조사
○ 주소록에 근거하여 70%의 거처에 우편 발송
○ 중앙자료처리센터에서 우편으로 회수
○ 인터넷을 통해 자료 수집: 18.3% 회수
○ 무응답 가구에 대해 조사원 투입: 27,000명 필요
○ 중앙통제: Master Control System
○ 내용 검토 및 imputation: 모든 항목에 대해 CANCEIS[6] 적용
○ 전반적으로 전략 성공
○ 사전 센서스를 통한 주소록 개선으로 적절한 거주명부 작성 가능
○ 누락 거주 확인 작업이 효율적이었음: 계획에 반영

6) CANCEIS: CANadian Census Editing and Imputation System

○ 인터넷을 통한 자료수집이 효과적이었음
  - 자료의 질 우수
  - 회수율 아주 좋음
  - 무응답과 무관함
○ 현장 조사: 조사원 투입시기 및 우수한 조사원의 지역적 배분을 고려해야 함
○ 전임 조사원의 채용과 유지를 위해 지역 경제 사정을 고려하여 조사수당의 융통성 있는 차등이 필요함.
○ 조사에 대해 아직은 우호적이지만 비협조가구의 지속적 증가
○ MIS를 통해 정확한 정보를 얻을 수 있지만 조사원의 MIS접근방법에 대한 개선 필요
○ 회수된 조사의 등록 시점 및 조사원 통보의 시점 개선 필요
○ 무응답 가구에 대한 추적에 융통성 있는 계획 필요
  - 센서스 도중에 개발되었고 효율적이었지만 효과적이지 않음
○ 중앙통제시스템의 필요
  - 조사 통제
  - 조사 방법의 다양: 중복 조사 방지
○ 모든 조사 항목에 대해 CANCEIS의 성공적 수정

## 2. 2011년 센서스 고려사항

□ 2011년까지 계속 될 것으로 예상되는 경직된 노동시장으로 인한 조사원 확보의 어려움
  ○ 우편 발송을 90%로 확대: 캐나다 우체국과의 관계
  ○ 보다 효율적인 조사원 모집 및 배치

□ 비용 절감
  ○ Imputation 방법을 개선하여 조사 착오 항목에 사후 조사를 2/3수준으로 줄임
  ○ 인터넷을 통한 자료 수집 비율을 40%로 확대

□ 종이 사용 줄임

○ 70%의 가구에 인터넷을 통한 조사협조 공문 발송

○ 인터넷을 이용하여 현장조사 관리

□ 이용자들의 센서스 결과 조기 공표 기대

○ 캐나다 우체국을 이용한 분권식 조사 관리 및 신속히 조사원에 통보

○ 효과적인 imputation

○ 보급 시스템의 개선

○ 보급 기간을 24개월에서 18개월로 단축

## 3. 2011년 센서스의 목표

- Wave 1(CD[7]-7)
  - ○ 70% 협조 공문 발송
  - ○ 20% 조사표 발송
  - ○ 10% 조사표 직접 전달
- Wave 2(CD-5)
  - ○ 우편 발송 지역: 조사 시작 협조 공문 발송
  - ○ 조사표 직접 전달 지역: Ad mail
- Wave 3(CD+7)
  - ○ 우편 발송 지역: 우편 발송 지역의 무응답 가구에 조사표 우편 발송
- Wave 4(CD+27)
  - ○ 무응답 가구에 대해 조사원 투입 조사 시작

7) CD: Census Day

## 4. 우선순위

- implications of the wave methodology
- Ongoing listing(AR updating)
  - ○ 다른 STC programs와 협력
- 회수된 가구 적시 현장 통보
- 유연한 조사원의 모집 및 보유, MIS
- FEFU-reduction; quality impact
- E&I stream-lining
- 결과의 조기 공표
- 공표 시스템의 개선: 간소화, 웹 레디 아웃풋

## 5. 기타

- 현재 캐나다 통계청은 인구센서스 내용 확정을 위한 "2011 센서스 자문회의"를 진행하기 위한 작업을 2007년 7월에 착수하여 2006년 내용과 대비 변경하고자 하는 내용을 상당부분 잠정 확정하고, 관심 있는 국내외 모든 관련자들의 의견 수렴 중에 있음
- 자세한 내용은 아래의 두 번째 참고문헌을 참고할 수 있음

※ Contact address:


○ David Dolson, Statistics Canada, 15 O RH Coats Building, 100 Tunneys Pasture Driveway, Ottawa, ON K1A 0T6 Canada
- e-mail: ddolson@statcan.ca


○ 2011 census content consultation guide
- web site: www.statcan.ca