출장 보고서

일본 총무성 통계국 방문

소지역 추정 연구 관련
출장 결과 보고

2008. 4.

통계개발원 연구기획실

〈 차 례 〉

소지역 추정 연구를 위한 일본 통계국 출장 결과

Ⅰ. 출장 개요

1. 목적 : 일본 통계국 방문을 통해 일본의 노동력조사에 대한 소지역 추정 통계 개발 방법 및 일련의 과정들에 대한 폭넓은 이해를 함으로써, 우리나라 고용통계 소지역 추정 연구의 효과적인 추진을 위해 필요한 현실적 시사점을 얻고자 함

2. 출장자 : 연구기획실 김서영 사무관, 권순필 주무관

3. 기간 : 2008. 4. 7 ~ 4. 12. 일본 동경

4. 출장지 : 일본 동경(통계국, 인구문제연구소)

Ⅱ. 회의 개요

1. 주요 내용 (자세한 내용은 **별첨** 참고)
   - ○ 일본의 노동력조사 소개
     - 목적, 표본설계 방법, 공표 현황 등
   - ○ 노동력조사에 적용된 일본의 소지역 추정 소개
     - 배경, 연구 과정, 연구 결과 등
   - ○ 「고용통계에 대한 일본 통계청의 소지역 추정 관련 질의서」 답변 및 상호 토론
   - ○ 소지역 추정 프로그램 시연 및 알고리즘 설명
   - ○ 현재인구 및 장래인구 추계 방법 소개

2. 참석자

○ 한국측 : 김서영 사무관, 권순필 주무관

○ 일본측

- 통계국 : 노동력인구통계실 과장보좌, 츠카다
  경제통계과 기획계장, 다까베(주 개발자)
  총무과 과장보좌, 다까하시
  히토츠바시 대학 경제연구소 교수, 야마구찌
- 인구문제연구소 : Dr. 사사이 (연구원)

3. 일정

일 시	주 요 일 정	담 당 부 서
4.8(화)	○ 노동력조사 방법 및 표본설계 소개 ○ 소지역 추정 세부 사항에 대한 소개 및 토론	노동력인구 통계실
4.9(수)	○ 소지역 추정시 사용된 알고리즘과 구체적 프로그램 code 및 내용 설명 ○ 통계국장 면담	노동력인구통계실 총무과
4.10(목)	○ 현재인구 및 장래인구 추계 방법 - 일본의 현재인구는 통계국 국세조사과, 장래인구는 인구문제연구소에서 각각 추정	인구문제연구소
4.11(금)	○ 전체 토론	

※ 회의장소 : 통계국 노동력인구통계실(회의실) 및 인구문제연구소

Ⅲ. 출장 결과

[소지역 추정 방법 개발 측면]

1. 일본에서 생각하는 소지역 추정 통계의 중요성

○ 고용통계의 소지역 추정 통계 작성 배경

- 일본의 노동력 통계는 전국 및 10개 지역(표본설계 단위) 결과만 공표하고 있음

(전국 : 월별, 사분기별, 반기별, 연별 / 10개 지역 :사분기, 연별)

- 이에 2000년 전후, 국가의 실업정책에 대한 관심과 더불어 보다 작은 지역단위의 고용통계 자료에 대한 수요가 증가하였음
- 당시 47개 도도부현 단위로 고용통계를 작성키로 하고, 현재 인터넷을 통해 참고자료로 공표하고 있음

※ 일본의 도도부현 인구단위 : 100만 이상 지역이 대부분임

○ 통계 작성 방법으로 조사보다 추정 방법을 택한 이유

- 국가 예산의 절대적 한계로 조사통계는 생각할 수 없음
- 이때, 통계의 중요도, 사용빈도, 신뢰성 측면을 고려하였음
- 조사통계의 신뢰성 측면에서 어느 정도 표본이 확장된다 하더라도 소지역 단위의 조사 오차 커버에 대해 확신할 수 없음

○ 소지역 통계 추정은 통계국 국장의 직접적인 지시로 추진되었음

2. 추정 통계를 바라보는 일본의 시각

○ 일본의 경우, 추정 통계도 하나의 통계 작성 방법으로 자리매김

- 세부 단위 조사에 대한 비표본오차 문제의 심각성을 통계청 내지는 통계 연구자들이 광범위하게 인지하고 있는 것으로 보임

○ 통계의 활용성 측면에서 예산을 고려하여 추정 통계를 봐야 한다는 주장임

3. 일본 노동력통계 소지역별 통계 추정 개요

○ 통계적 방법 : 시계열 모형 (Rotation 표본 영향 반영 가능)

- 표본추가 여부 : 소지역 통계 추정을 위한 표본 추가는 없음
- 보조정보 사용 : 이웃지역의 실업률 정보만 사용

○ 대상 소지역 : 47개 도도부현

○ 공표 : 사분기별(해당 분기 마지막 월), 연별 추정 통계

○ 공표방법 : 통계국 홈페이지를 통한 공표

○ 실사과 사용프로그램 : SAS(자료 핸들링), S-plus(추정)

○ 소지역통계 활용여부 : 지역별 실업자 파악 및 정책에 활용

[소지역 추정 방법 연구 측면]

1. 중장기 연구과제로서 지속성이 중요

○ 일본은 2002년부터 2004년까지 3년 동안 담당자를 지정하여 소지역 추정만을 전담 연구토록 하고 지속적으로 연구가 가능하도록 연구 환경을 마련해 줌

- 실무 담당자였음에도 업무의 90%를 연구에 집중토록 함

○ 인사이동에 의한 시스템·전문지식 등의 단절이 없도록 매뉴얼을 구비하고 연수 등의 기회 부여

2. 효율적 연구체계 확립의 중요

○ 내부 인력의 부족으로 일본 통계국 내에서는 1인이 담당하였음

○ 그러나, 실제로는「고용통계지역추계연구회」의 역할이 매우 컸음

- 다양한 이론적 방법의 검토: 후보 방법 5가지를 검토
- 추정결과에 대한 검토
- 프로그램 작성 코드에 대한 검토

○ 일본은 통계국 내부인력이 프로그램을 작성하고,「고용통계지역추계연구회」에서 실질적인 연구방향과 절차 및 결과에 대한 가이드라인을 제시하는 역할을 하였음

○ 이러한 일본 연구 체계의 장·단점
- 일본의 통계국 내부에 연구를 진행할 만한 인력이 없었기 때문에 외부 위원회를 구성하여 진행상의 주요 역할을 맡겼음
- 따라서 상당한 시간적 소요가 있었을 것으로 보임
- 연구결과의 형평성을 유지하는 데 위원회의 검토 결과를 이용하는 것은 다소 이점이 있어 보임
- 그러므로 내부에서 연구 진행을 담당하고, 외부에서 프로그램 요원을 확보하여 연구 소요 시간을 단축시키고, 다양한 분야의 인사들로 구성된 자문위원회를 통해 연구결과의 신뢰성을 확보하는 것이 바람직해 보임

2. 추정 통계 제공에 대한 이해 확보가 필요

○ 지역별 추정 통계 제공에 따른 부작용(예, 서열화, 오차 무시 경향)을 줄이기 위해, 제공되는 통계가 추정 통계라는 점을 강조하고 해당 통계에 대한 이해 확보

○ 도도부현 단위로 지역 통계 조사를 실시했던 지역들도 추정통계 통계 공표 이후 조사를 중지하고 있는 실정임
(* 지자체의 예산이 조사 중지의 가장 큰 이유라고 함)

3. 추정 결과에 대한 구체적인 품질 평가 기준의 확립

○ 일본은 다양한 모형을 이용하여 소지역에 대한 추정을 시도하였고 이들 모형에 대하여 세부적인 기준을 가지고 그 품질을 평가하였음

○ 기준 : 일반성, 재현성, 간명성, 정부통계로서의 적용성, 추계 결과의 적용성, 실용성 등 6개 항목에 대한 기준을 보유함으로 추정 결과에 대한 설명력 확보

○ 통계국 내·외 연구를 위한 조직 구성

- 외부조직 :「고용통계지역추계연구회」구성
  - 구성원 : 이론부분-통계 분야의 전문가, 시계열해석 전문가
    노동통계-노동통계(노동경제) 전문가
    실무부문-사용자 대표, 프로그램 검토 연구자 등
- 내부조직 : 통계국 및 통계연수소의 간부들로 위원회를 구성
  - 역할 : 통계추정치의 정도 등 전문적인 측면뿐 아니라 일반 이용자들로부터 이해를 받을 수 있을지 등과 같이 정성적인 측면도 고려해서 추정방법을 검토하고 선정

【소지역 추정 환경의 양국 간 차이】

○ 양국의 인구·표본 규모 비교

한/일 비교	한국	일본
소지역 정의 (행정단위일 때)	230여 개 시군구	47개 도도부현
소지역 인구규모	최소 1만명 지역도 포함	최소 100만 이상
표본조사구 규모	시군구당 1개조사구도 있음 (2007년 경활조사 표본규모 기준)	최소 표본조사구 수 16개

○ 일본과 우리의 소지역 현실은 인구베이스와 표본 규모면에서 매우 다름

- 우리의 현실에 맞는 타당한 방법을 적극 찾을 필요가 있음

○ 표본수가 일정 정도(100조사구) 이상인 경우 직접 추정량 발표

【결론 및 시사점】

○ 열악한 조사환경 대비를 위한 대안으로서 추정통계에 대한 새로운 시각이 필요한 시점임

○ 추정통계의 신뢰성 확보를 위해, 지속적 연구와 객관적 평가 필요

○ 장기적으로는 일관성 있는 연구와 실무의 지속성을 위해 추정 방법에 대한 매뉴얼 작성이 있어야 함

○ 소지역 통계 작성에 대해 다른 나라의 사정을 문헌연구만이 아닌 직접 접촉을 통해 해당 방법을 정확히 이해하고, 필요한 정보와 노하우를 끊임없이 확보함으로써 국제적 경향을 따를 수 있어야 함

[별첨 1] 일본 노동력조사의 표본추출 및 추정법 요약

1. 표본추출방법

노동력 조사는 층화 2단계 추출로, 조사구(국세조사구)를 제 1차 추출단위, 각 가구[1]를 제 2차 추출단위로 한다.

(1) 조사구 추출(1차 추출)

○ 1차 추출인 조사구 추출은 각 지역별로 모든 조사구를 국세조사 결과 등에 근거한 특성 층으로 나누고, 각 지역의 각 층별로, 적절한 추출률과 난수번호를 이용해서 계통 추출한다(결국 랜덤추출과 같음).

○ 이 계통추출은 각 조사구 가중치(15세대가 거의 1 가중치)에 근거한 확률비례추출이다.

- 매월 표본조사구수는 약 2,900개 정도.

○ 단, 형무소・구치소가 있는 구역(국세조사 조사구로서의 끝 번호가 5인 조사구), 자위대구역(동 6조사구), 주유군구역(동 7조사구) 및 수면조사구(동 9조사구)에 대해서는 추출하지 않는다.

※ 국세조사구(현의 개수): 홋카이도(1), 토호쿠(6), 남관동(4), 북관동(5), 호쿠리쿠(6), 토카이(4), 킨키(6), 츄코쿠(5), 시코쿠(4), 큐슈(8) 등 10개 지역

(2) 가구 추출 (2차 추출)

○ 2차 가구 추출은 1차에서 추출된 조사구(이하 “표본조사구”)에 포함된 모든 가구에서 1조사구 당 약 15가구가 추출되도록 적절한 추출율(가중치 역수) 및 난수번호를 이용해서 계통 추출한다.

- 추출된 가구에 거주하는 모든 세대(합계 약 4만 가구)가 조사 대상임.

① 월 또는 년 결과의 정도와 매월 및 연간 변화를 보는 경우의 정도 등을 고려하기 위해 한 개의 표본 조사구는 4개월간 조사한다.

- 추출된 가구를 전반(2개월간)과 후반(2개월간)으로 나누어 조사.

② 전년도 결과와의 비교 정도를 높이기 위해 표본 조사구로서 선정된 조사구는 익년 동월에 다시 조사한다[2].

- 매월 표본조사구 가운데 반은 당해 년에 새롭게 조사를 행하는 조사구(따라서, 익년 동월에 다시 조사를 시행하는 조사구. 이하 “1년째 조사구”라

1) 주택 및 기타 건물 각 호(한 세대가 거주할 수 있도록 되어 있는 건물 또는 건물의 한 구획)

2) 각 표본조사구에 대해서 익년까지 없어진 각 호에 주거했던 조사가구는 조사에서 제외된다. 한편, 신설된 가구는 명부에 추가되고, 그 명부에서 가구가 추가 추출되어 그 곳에 주거하는 세대가 조사 세대로 추가된다.

함)가 되고, 나머지 반은 전년 동월에 조사를 행한 조사구(이하 “2년째 조사구”라 함)가 되도록 한다.

③ 이상과 같이 표본을 교체해서 추정치의 표본오차 산출을 위해 표본 조사구는 조사 개시월(A, B, C, D)과 조사회수(1, 2)를 구분하여 다음과 같은 8개의 부표본으로 구성된다. 각 부표본은 각각 동일한 확률을 갖는 전국 확률표본이 되도록 설계되었다.

○ 부표본 구성

부표본 그룹	특성 (개시월 및 햇수)	부표본 그룹	특성 (개시월 및 햇수)
A1	1월, 5월, 9월이 조사개시 1년째	C1	3월, 7월, 11월이 조사개시 1년째
A2	1월, 5월, 9월이 조사개시 2년째	C2	3월, 7월, 11월이 조사개시 2년째
B1	2월, 6월, 10월이 조사개시 1년째	D1	4월, 8월, 12월이 조사개시 1년째
B2	2월, 6월, 10월이 조사개시 2년째	D2	4월, 8월, 12월이 조사개시 2년째

- 부표본 8개 가운데 4개는 조사개시 1년째 조사구, 나머지 4개는 2년째 조사구.
- 결과적으로 매달 부표본 가운데 2개(즉 표본조사구의 1/4)에서는 표본조사구 교체가, 다른 2개 부표본에서는 동일 조사구 내에서 조사가구의 교체가 발생.
- 따라서, 표본 조사구가 교체되는 부표본과 표본조사구 내의 조사 가구가 교체되는 부표본을 합하면 매월 조사 가구의 1/2이 갱신
- 특별 조사표의 조사가구는 2년 2개월째에 해당하는 2번째 부표본(A2 및 C2, B2 및 D2)임. 따라서 상세결과 조사규모는 기본집계의 약 1/4임.

(3) 추출율과 표본 크기

○ 추출율

- 1차단위, 2차 단위 추출이 일정하지 않지만, 두 단위의 곱은 약 1/1000.

○ 표본 크기

표본종류	표본크기
1차 추출단위(조사구)	2,912 조사구
2차 추출단위(가구)	약 40,000 호
조사세대	약 40,000 세대
조사세대 세대원(15세 이상)	약 100,000 명

2. 결과 집계 방법

(1) 결과 추정 (기본 집계)

○ 매월 전국 결과는 대도시부・비대도시부[3], 성별/연령별(15구분[4])에 대하여 국세조사에 근거한 추계인구를 기준인구로 비추정한다.

(2) 추정방법(기본 추계)

○ 전국 결과 산출 순서

step1. 각 표본조사구의 성별/연령계급별 조사 인구에 선형추정용 승수를 곱해서 계산을 하고, 성별/대도시부・비대도시부/연령계급별 인구 선형추정치를 산출

step2. 성별/대도시부・비대도시부/연령계급별 기준인구를 각각 step1에서 산출한 선형추정치에서 빼고, 비추정용 승수를 산출

step3. 각 표본조사구 속성 X를 갖는 성별/연령계급별 조사인구에 선형추정용 승수를 곱해서 필요한 합산을 하고, 거기에 step2에서 산출한 비추정용 승수를 곱해서, 성별/대도시부・비대도시부/연령계급별로 비추정치 $\hat{X}$를 산출한다.

step4. 이 비추정치 $\hat{X}$을 대도시부와 비대도시부에 대해서 합산하거나 성별/연령계급 등에 대해서 합산해서 각종 결과수치를 얻는다.

(참고) 위의 step1-step3을 계산식으로 표현하면 다음과 같다.

$$\hat{X}=\sum_{i=1}^{L}\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{w_i}{w_{ij}}\cdot f_{ij}\cdot x_{ij}\frac{P}{\sum_{i=i}^{L}\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{w_i}{w_{ij}}\cdot f_{ij}\cdot P_{ij}}$$

$$=\sum_{i=1}^{L}\sum_{j=1}^{m_i}x_{ij}\cdot F_i\frac{P}{\sum_{i=1}^{L}\sum_{j=1}^{m_i}P_{ij}\cdot F_i}$$

여기서

i : 10개 지역, 층 번호 ($i=1,2,\cdots,L$)

j : 각 층 내 표본조사구 번호 ($j=1,2,\cdots,m_i$)

x_{ij} : i번째 층, j번째 표본조사구 내의 속성 X를 갖는 조사 인구(성별/연령계급별)

w_{ij} : i번째 층, j번째 표본조사구의 가중치

---

3) 전국을 대도시부(동경도의 구부, 요코하마시, 나고야시, 교토시, 오사카시, 고베시 및 큐슈시)와 그 외로 나눈 것

4) 2007년부터 15구분(15-19세부터 80-84세까지 5세급 및 85세 이상)별 추계로 변경했다. 산출기본식은 다음과 같다
<취업자의 경우> 취업자수=선형추정에 의한 취업자수*(기준인구/선형추정에 의한 인구)

f_{ij} : i번째 층, j번째 표본조사구의 각 가구 추출율의 역수 ($f_{ij} = w_{ij}$)

w_i : i번째 층에 포함된 모든 조사구 가중치의 합계

m_i : i번째 층의 표본조사구 수

F_i : i번째 층의 선형추정용 승률 ($F_i = w_i / m_i$)

P : 기준인구(성별/대도시부·비대도시부/연령계급별)

P_{ij} : i번째 층, j번째 표본조사구내의 조사인구 (성별/연령계급별)

$\frac{P}{\sum_{i=1}^{L}\sum_{j=1}^{m_i} P_{ij} \cdot F_i}$: 비추정용 승률

○ 지역별 결과는 전국결과와 같은 방법으로 각 지역별 결과를 구한 후 10개 지역 합계가 전국결과와 일치하도록 보정해서 산출한다.

(3) 기준인구 및 기준인구 갱신

○ 결과를 산출하기 위한 기준인구는 매월 1일(=전월 말일) 현재 추계인구를 이용한다.

- 현재 추계인구는 국세조사에 의한 인구를 기초로 다른 인구 관련 자료로부터 구한 각 달의 인구 동향을 가감함하여 산출[5](총무성 통계국)

○ 추계인구는 최근 국세조사의 인구를 기초로 하고 있고, 5년마다 새로운 국세조사 확정 인구에 근거한 계산치로 교체된다. 때문에 노동력조사 결과 산출에 이용하는 기준 인구도 5년마다 교체가 일어난다. 과거 기준에 의한 추계인구와 신 기준에 의한 기준인구 간의 차이가 있는 경우, 노동력조사 기준인구의 교체가 일어난 해의 결과에는 통계상의 불일치적 측면이 포함된다.

- 현재 기준인구는 2005년 국세조사 결과를 사용하여 2007년 1월부터 적용

(4) 상세결과 추정

○ 사분기 평균 및 연평균 결과는 해당하는 기간의 월 평균 결과를 단순 평균해서 산출한다.

○ 월 결과는 매월 기준 추계결과의 성별/연령별(10세 계급, 5구분), 취업상태(취업자, 완전실업자, 비노동력 인구)별 인구를 기준인구로 하는 비추정으로 산출한다.

---

5) 2007년부터 추계인구 산출에 이용하는 인구사회 동향에 대해서, 일본인의 출입국자수를 '해외체재 기간 91일 이상 출입국자수'를 이용해서 산출하는 방법으로 변경

○ 산출 기본식

- 예 : 측정조사표 A란 (취업자에 관한 항목)의 항목에 대해

  A란의 추정치=선형추정치에 의한 A란의 값*(기준추계 취업자 수/상세결과 취업자수)

- 선형추정치는 기본추계 결과를 산출할 때 이용한 선형추정용 승률에 의한 추계값

3. 추정치 표본오차

○ 표본오차의 크기는 추정치의 크기 외에 조사항목의 종류나 조사년 또는 월에 따라 다르다. 이것의 측정기준이 되는 표준오차는 부표본을 이용하여 계산

- 세세한 결과는 생략키로 함.

4. 도도부현별 결과(모형 추정 방법)

(1) 경위

○ 도도부현별 결과에 대해서는 2002년부터 참고적으로 비추정치에 의한 연평균 결과(시산치) 공표를 시작하였으며

○ **시계열 회귀모형**에 의한 추계방법을 이용해서 정도 향상을 꾀한 다음 2006년 5월부터 새로운 사분기 평균 결과(모형 추계치)를 공표하고 있다.

(2) 공표 계열

○ 모형 추계치는 1997년 이후부터 **노동력인구, 취업자수, 완전실업자수, 비노동인구수, 완전실업율** 항목에 대해 도도부현별 **사분기 및 연 평균** 결과를 공표하고 있다.

(3) 추정방법

○ 노동력조사의 도도부현별 추계 방법은 다음과 같이 5개 요인을 포함하는 시계열회귀모형을 이용한다.

$$Y(t) = X(t)\beta(t) + T(t) + S(t) + I(t) + e(t)$$

○ 각각의 요소는 다음과 같은 변동을 나타내고 있다.

- 회귀항, $X(t)\beta(t)$: 각 도도부현 움직임과 도도부현이 속한 지역 trend와의 관계
- **Trend**항, $T(t)$: 경제 성장 등에 따른 장기적 변동으로 나타나는 변동과 경기 순환에 동반되는 변동 등 일정한 주기를 갖는 변동

으로 주기가 **12개월을 넘는 순환변동** 등을 합한 변동

예) 경기 후퇴와 회복에 있어서, 완전실업자가 증가·감소하는 경향

- **계절변동**항, $S(t)$: 12개월을 주기로 하는 계절변동

  예) 취업자수는 신년도가 시작되는 3~4월까지 증가하고, 6~7월에 최고가 되며, 연 후반기에 감소하는 움직임이 있음

- 불규칙변동항, $I(t)$: 순환변동, 계절변동 이외의 변동으로 돌발적 상황에 의한 변동 및 경기의 단기적 변동

  예) 과거 석유 파동 및 지진 등 자연재해 등에 의해 일어나는 일시적 움직임

- 표본오차, $e(t)$: 추계 모형에서 가장 중요한 부분으로, 표본오차의 변동 패턴과 변동의 진폭을 나타냄.

[별첨 2] 일본 통계청의 소지역 추정에 대한 질의·응답 자료

☐ 질의 배경

○ 한국은 지금 다른 나라와 마찬가지로 고용통계에 있어서(실업율, 취업률) 소지역 단위 통계에 대한 사용자의 요구가 증가하고 있으며 특히, 지방자치단체의 경우, 지역별 고용통계를 이용하여 실업정책 및 예산배정에 활용하려는 움직임을 보이고 있음. 그러나 현재 한국은 고용통계에 있어서 전국 결과와 16개 시도(대단위지역) 결과만 작성하고 있으며, 따라서 보다 더 작은 230개의 시군구(일본의 현단위보다 면적, 인구규모가 작음)단위로 추정통계를 작성하려고 하고 있음.

○ 일본의 추정 모형과 일련의 과정을 정확하게 이해함으로써 우리의 소지역 통계 추정치 작성을 위해 도움을 얻고자 함

☐ 질의 응답

1. 소지역(small area) 관련

Q1. 47개 도도부현에 대하여 소지역 추정 통계를 제공하고 있는데 시정촌과 같이 더 작은 규모의 지역에 대한 소지역 추정량은 고려하지 않았는가?

A. 시계열 모형을 이용한 사분기별 소지역 추계는 도도부현 수준 밖에 제공하고 있지 않음. 시정촌 수준이 되면 표본수가 작아서 사분기에서의 공표는 곤란. 시정촌에 대해서는 주로 국세조사 결과를 이용하게 되지만, 시의성은 없음. 이처럼 아주 작은 지역 결과를 월별 또는 사분기로 공표하고자 하는 경우에는 추정치의 신뢰성 측면에서 고용보험 등을 이용한 회귀모형으로 일단 추정치를 낼 수 밖에 없다고 생각함.

Q2. 추정 통계 생산 단위(small area)에 대한 인구 및 표본 규모는 어느 정도 되는지?

A. 표본수는 <표 1> 참고. 인구는 하나의 소지역에 수십만~수백만 정도.

Q3. 표본수가 0인 지역이 존재하는지, 존재한다면 추정할 때 이 지역들을 어떻게 처리하였는가? 한국의 경우 추정하고자 하는 소지역 단위에서 표본규모가 0인 지역이 발생하고 있는데, 현 표본수를 늘리지 않고서 이를 극복할 수 있는 방법이 있다면 무엇인지?

A. 표본수가 0인 지역은 없음. 만약 시정촌 수준에서 표본수가 0인 지역을 추정하게 된다면 고용통계 등에 의한 cross-sectional 회귀모형으로 결과를 보간하는 것이 타당한 방법이라고 생각됨.

<표 1> 일본 노동력조사 지역별 표본수

No	도(都),도(道),부(府),현(縣)	표본수*	10개 층
1	北海道 홋카이도	178	① 北海道 홋카이도
2	青森 아오모리	34	② 東北 토호쿠
3	岩手 이와테	35	
4	宮城 미야기	54	
5	秋田 아키타	26	
6	山形 야마가타	30	
7	福島 후쿠시마	55	
8	茨城 이바라키	66	③ 北關東·甲信 북관동·카츠노부
9	栃木 토치기	46	
10	群馬 군마	50	
11	埼玉 사이타마	114	④ 南關東 남관동
12	千葉 치바	96	
13	東京都 도쿄또	220	
14	神奈川 카나가와	148	
15	新潟 니가타	75	⑤ 北陸 호쿠리쿠
16	富山 토야마	30	
17	石川 이시가와	41	
18	福井 후쿠이	23	
19	山梨 야마나시	21	③ 北關東·甲信 북관동·카츠노부
20	長野 나가노	52	
21	岐阜 기후	45	⑥ 東海 토카이
22	靜岡 시즈오카	74	
23	愛知 아이치	148	
24	三重 미에	40	
25	滋賀 시가	24	⑦ 近畿 킨키
26	京都 쿄토	49	
27	大阪 오사카	175	
28	兵庫 효고	106	
29	奈良 나라	28	
30	和歌山 와카야마	20	
31	鳥取 톳토리	16	⑧ 中國 츄코쿠
32	島根 시마네	21	
33	岡山 오카야마	50	
34	廣島 히로시마	77	
35	山口 야마구치	45	
36	德島 토쿠시마	30	⑨ 西國 시코쿠
37	香川 카가와	33	
38	愛媛 에히메	55	
39	高知 코치	33	
40	福岡 후쿠오카	107	⑩ 九州 큐슈
41	佐賀 사가	18	
42	長崎 나가사키	33	
43	熊本 쿠마모토	38	
44	大分 오이타	26	
45	宮崎 미야자키	27	
46	鹿兒島 카코시마	39	
47	沖繩 오키나와	144	
계	1도 1도 2부 43현	2,895	

※ 주 : 표본수는 2000~2002년 평균 조사구 수임

2. 추정 방법 관련

(1) 모형선택

Q4. 현재 일본에서 사용하고 있는 시계열 모형을 선택하는 과정에서 가장 중요하게 고려한 것은 무엇인가? 예를 들면 일본의 표본설계(rotation sample) 구조나, 경제상황 등.

A. 시계열 모형에 대해서는 아래와 같은 장점을 생각해서 채택
- 시스템 구축이 용이
- 추정치의 표준편차를 보거나 국세조사 등과 비교해 보았을 때 좋은 결과를 얻음
- 그래프를 통한 시각적인 분석도 보다 좋은 결과를 보임

Q5. 일본의 소지역 추정 모형은 미국 사례를 벤치마킹한 Time Series 모형으로 알고 있는데, 이 모형이 일본의 노동력 표본설계인 2-10-2 시스템을 잘 반영하고 있는지, 혹시 이 모형을 채택하는 단계에서 표본설계와 관련하여 어떤 문제점은 없었는지? (가장 관심은 rotation sample의 효과를 잘 반영할 수 있을 지 여부임)

A. rotation 시스템을 잘 반영한 미국 소지역 추계 방법을 도입했기 때문에 rotation 효과를 잘 반영하고 있다고 생각됨. 그러나 미국 CPS에서는 BRR법(Resampling에 의한 표본오차 추정법)에 의한 표본오차추정이 가능하도록 표본이 설계되어 있지만 일본은 그처럼 되어있지 않기 때문에 조사구를 단위로 취급해서 표본오차를 추정함. 표본오차의 시계열을 보면 표본 rotation 구축에 의해 lag를 1과 12시점에 어느 정도 자기 상관이 존재한다고 생각되는데, 시계열 모형에 의해 추정된 표본오차의 시계열 자기상관을 보면 lag가 1과 12로 상관이 높아지고 있고 rotation 효과는 잘 반영되었다고 말할 수 있음

Q6. 소지역 추정과 관련하여 국제적으로 많은 연구가 활발하게 진행되고 있고, 다양한 모형들이 사용되고 있다. 각 나라의 특성을 반영하여 모형을 선택했다고 판단되는데, 일본에서는 지금의 시계열 모형을 최종적으로 선택할 당시에 어떤 다른 후보 모형들(예. 영국의 Logit 모형)을 검토했는지 궁금함.

A. 미국의 Fay-Herriot 모형(EBLUP), 영국의 Logit 모형, 캐나다의 Rao-Yu 모형, Stein 추정, 단순 계층적 베이지안 모형을 검토. 이들 많은 모형을 검토한 결과 실제 적용을 위해 아래 5개 방법을 후보로 택했다.
- Stein 방법, 시계열 모형, EBLUP, Cross-Sectional and Time Series, 계층적 베이즈 모형 등.

Q7. 만약 다양한 형태의 모형을 고려하는 과정에서 현재의 시계열 모형으로 결정하게 되었다면 그렇게 결정한 이유는 무엇인가? 예를 들면 추정치의 정도(precision), 표본오차, MSE(Mean Squared Error)등 고려한 것인지? 아니면 다른 기준이 될만한 것이 있었는지 궁금하다. 또한 현재 일본에서 사용하고 있는 모형의 장・단점은 무엇인가?

A. 추정치의 표본편차 분석 및 국세조사 결과와의 비교에서 평가 결과가 좋았다.
- 시계열 모형은 전문가 검토 결과에서도 가장 지지가 많았음
- 다른 방법에서는 고용통계(고용보험)등의 보조정보에 의한 회귀모형을 이용하지 않으면 안됨. 그러나 고용통계의 노동력조사에의 적용은 좋지 않음. 시계열 모형이라면 그런 보조정보를 이용하지 않아도 추정을 할 수가 있다는 장점이 있음.

(2) 보조정보 선택

Q8. 모형에 기반한 소지역 추정 통계 작성에 있어서 보조정보 선택은 가장 중요한 문제라고 생각한다. 보통의 경우 보조정보로 고용보험 자료를 많이 사용하고 있는 것 같다. 일본의 경우 고용보험 자료가 보조정보로 적당하지 않다고 했는데, 그 이유는 무엇인가?

A. 소지역추정 방법 검토를 행하기 전에 고용보험이랑 취업안정통계(유효구인배율)와 노동력조사와의 상관을 살펴본 결과 그 적합은 그다지 좋지 않았음. 특히 시간이 경과함에 따라서 적합이 더 좋지 않은 경향을 보였음

Q9. 한국도 현재 지역별 고용보험 자료의 이용을 고려하고 있는데, 보조정보로서 가치가 있을지는 아직 의문이다. 우리가 이 자료를 검토하는데 있어서 어떤 점을 중점적으로 검토하면 좋겠는가? 일본의 사례에 비추어 설명

A. 일본에서는 고용보험 등 보조정보의 적용은 문제가 있었기 때문에, 이들 보조정보를 가능한 이용하지 않고 추정할 수 있는 방법에 중점을 두게 됨

Q10. 미국의 경우, 취업자 수 추정 모형과 실업자 수 추정 모형에서 사용하는 보조정보가 각각 다르다. 일본에서도 각각 다른 정보를 사용하였는가? 그렇지 않았다면 어떤 특별한 근거라도 있는가? 특성치에 영향을 주지 않는다든지 등.

A. 일본 소지역 추정 모형에서는 고용보험 등 정보는 이용하지 않음. 따라서 지역별 보조정보가 다른 것은 없음. 단 각 소지역 추정값을 구할 때 이보다 큰 범위의 지역으로 추정한 트렌드를 보정정보로 이용하고 있음.

Q11. 모형에서 사용하는 공간 정보로 10개 이웃지역의 자료를 사용하는 것으로 설명되어 있는데 이 10개 이웃지역 선정 기준은 무엇이며 모형 내에서 어떻게 이용되는가? 표본추출 과정과 함께 설명주시면 좋겠다.

A. 노동력조사는 층화이단계표집 방법을 이용하고 있음. 이때 1단계로 조사구를 추출할 때 모집단이 10개의 층으로 이루어져 있음. 10개의 층은 각각 동일 경제권에 들어간다고 생각되는 소지역을 합해서 구성되어 있음. <표 1> 참고

(3) 추정량 선택

Q12. 모형 사용시 추정량은 EBLUP, EB, HB 중 어느 것이며 관련 분산 추정 방식은 어떤 방식을 사용하는가? (이 부분은 프로그램과 연결되어 설명되어진다면 더 좋을 것 같음)

A. 일본 모형은 시계열 해석방법(상태공간모형)을 이용한 것이고, EBLUP, EB, HB라 하는 회귀모형과는 다름. 단, 아래의 2가지 측면,

- 최우추정법에 의한 파라미터 추정치를 이용
- 상태공간모형은 베이즈 통계의 의미로 해석 가능 측면

을 보면, 넓은 의미로 EB 부류에 들어간다고 생각할 수 있음. 또, EBLUP, HB에 대해서는 추정방법의 검토단계에서 Fay-Herriot 모형 및 J.N.K RAO의 Time-Series and Cross Sectional 모형(HB) 방법에 의한 추정치를 계산해 보았음.

(4) 모형평가

Q13. 일본도 다양한 후보 모형을 고려하여 위의 추정량들을 고려했을 거라고 생각하는데, 다양한 모형들을 평가하는데 사용한 방법은 무엇인가? 최근에 신뢰구간, MSE라든가, 전문가 검토 등 다양한 방법 등을 단계적으로 사용하고 있는 것 같은데, 일본은 어떤 검증 단계를 거쳤는지 궁금하다.

A. 다음과 같이 추정치의 좋은 점을 비교

- 추정치의 표본오차
- 동 시기에 행해진 국세조사나 다른 대규모조사에 의한 완전실업율 등의 추정치와 비교
- 마지막으로 6개의 평가기준을 정하고 전문가에 의한 평가 실시

(4) 벤치마킹 문제

Q14. 소지역 추정의 문제점 중 하나가 대지역의 추정량 결과와 소지역의 추정량들의 합이 일치하지 않는다는 것인데 일본의 경우 이러한 불일치를 어떻게 해결하고 있는가? (벤치마킹 수행 여부 및 방법)

A. 당초는 Denton법이나 미국 CPS에서 이용되고 있는 방법에 의해 벤치마킹을 생각했지만, 연평균으로 계산함으로써 12월과 익년 1월 추정치에 단차가 발생할 수 있는 등의 문제점이 있기 때문에 최종적으로 벤치마킹을 보류. 미국에서는 새로운 지역 값에 합해서 벤치마킹 방법이 개발된 것 같은데 이에 대해서도 추정을 해본 결과 추정치가 흔들리는 문제가 있어 채택하지 않음. 이런점에 비추어 소지역 추정량은 참고치로서만 이용하길 권고하고 있음

3. 프로그램 작성

Q15. 소지역 추정량을 계산하기 위한 프로그램은 어떤 소프트웨어를 사용하였는가? 프로그램 작성을 위한 구체적인 알고리즘은 일본의 방식을 이해하는데 많은 도움이 될 것 같다.

A. 계산 속도를 빠르게 하기 위해 S언어와 포트란을 사용. 시스템 인터페어 부분 및 비선형 최적화 루틴은 S를 이용해서 구동을 용이하게 함. 주 계산부분은 포트란을 이용하고 이것을 S에서 불러들여 이용. 미국 추정법에서는 소지역 표본오차나 표본오차 시계열 자기상관을 추정할 필요가 있고 여기에는 매월 10만건 이상의 마이크로데이터를 처리할 필요가 있음. 그때는 대량자료 처리에 강한 SAS를 이용하고 있음.

4. 공표

Q16. 일본에서 발표하고 있는 추정량은 국가통계로서 승인된 통계인가? 승인된 통계라면 추정통계를 승인통계로 인정하는데 여러 가지 문제가 있을 것 같은데, 혹시 어떤 문제들이 있었는가?

A. 소지역 추정량은 참고치로서만 이용하길 권고하고 있음.

Q17. 소지역 모형 추계치는 인터넷을 통해 발표되고 있는 것으로 알고 있는데 그 외(책자 등)의 형태로도 발표되고 있는가?

A. 속보형태

Q18. 일본에서 소지역 추정량을 발표하면서 제시하고 있는 최대 CV 수준은 어느 정도이고, 그렇게 정한 데는 특별한 이유가 있는가?

A. CV 수준에 의해 발표할 지역을 선택한다고 할 수 없고 기본적으로 전체 소지역에 걸쳐 소지역 추정치를 공표하고 있음. 그 때에 표본수가 작은 지역에 대해서는 오차가 크고 이용에 주의할 필요가 있음을 명시하고 있음. 어느 정도 표본수를 확보할 수 있는 지역(표본조사구수 100개 이상)에서는 비추정치를 그대로 공표.

5. 연구기간 및 내용

Q19. 일본의 경우 2002년부터 연구를 시작하여 현재의 결과를 얻은 것으로 알고 있다. 전체적인 소지역 추정 연구 과정 및 내용을 알 수 있는가?

A. 일본 통계국 제공 첨부 파일 참고

6. 활용

Q20. 소지역 통계는 주로 누가 요구하며, 어떤 목적으로 사용되고 있는가? 특히 국가 정책 수립시 어떤 식으로 사용되고 있는지 구체적으로 설명을 해주시면 좋겠다.

A. 구체적인 사례는 아직 파악되고 있지 않지만 소지역 소비자 물가지수와 함께 이용되어 지역별 필립스 곡선을 찾는데 사용될 수 있다.

7. 기타

Q21. 모형 추계치를 공표하기 위하여 표본 수를 증가시키는 등의 특별한 노력이 추가된 것이 있는가? (예 : 캐나다, 영국의 경우 소지역 추정을 위해 표본수를 증가시켰음)

A. 표본수 증가 등의 조치는 없음. 오히려 예산 면에서 이같은 조치가 없었기 때문에 소지역 추계의 필요성이 생겼음

Q22. 모형 추계치의 품질에 대한 검증 과정을 거쳤는가? (예 : 내·외부 전문가들과의 협의체를 통한 검증 등)

A. 이 분야의 유명인사들로 검토회를 만들어 검토를 했음.

- 통계전반 분야의 전문가, 시계열해석 전문가, 노동통계(노동경제) 전문가, 사용자 대표, 각종 추정작업에 관한 어드바이저(주로 프로그램 검토 등)로 박사과정 연구자 등 폭넓은 분야의 전문가와 통계국 및 통계연수소의 간

부들로 위원회를 개최해서 추정치의 정도 등 전문적인 측면뿐 아니라 일반 이용자들로부터 이해를 받을 수 있을지 등과 같이 정성적인 측면도 고려해서 추정방법을 검토하고 선정.

Q23. 인터넷 공표 자료에 의하면 추정 결과에 대한 품질 평가시 일반성, 재현성, 간명성, 정부통계로서의 적용성, 추계 결과의 적용성, 실용성 등 6개 항목을 살펴보았다고 했는데 이 6개 항목들이 의미하는 바에 대한 정확한 정보가 있는가?

A. 각 항목에 대해서 그 장단점에 대해서 정성적인 평가를 했다. 추정방법에 관해서 ①~③, 노동력조사에 관해서 ④~⑥의 관점에서 평가를 했음

① 추정방법이 전문가들에게 알려져 있는가(일반성)
② 추정결과를 재현할 수 있는가(재현성)
③ 추정결과가 복잡해서 이해하기 어렵지 않은가(간명성)
④ 국가 통계인 노동력 조사에 적용하는데 문제는 없는가(적용성1)
⑤ 추정결과가 원 계열의 경향을 반영하고, 도도부현의 특징을 잘 표현하고 있는가(적용성2)
⑥ 매월(또는 사분기) 집계가 필요한 시기에 가능한가(실용성)

Q24. 향후 소지역 추정 연구(개선) 방향으로 설정된 내용이 있는가? 다른 분야(가계 조사 등)에의 적용 여부 등

A. 없음

Q25. 추정 과정에서 부딪힌 실제적인 문제들은 어떤 것이 있으며 이를 극복하기 위해 어떤 노력을 했는가?

A. 추정 시스템에 대해서 당초는 공개 패키지인 R을 사용했으나 이 경우 국가통계로서 공표할 때 문제가 있기 때문에 유사한 스펙을 갖는 S 언어에 의해 프로그램을 작성. 그리고 미국 방법은 표본오차시계열 모형을 결정할 필요가 있는데. 이에 대해서 원 논문에서는 복잡한 ARMA 모형이 이용되었음. 그러나 이는 추정이 상당히 어렵기 때문에 미국 모형 개발자인 R. Tiller에 직접 메일로 질문을 보냈는데 실제 추정에는 보다 간략화한 AR 모형이 이용될 수 있음을 알았음.

Q26. 한국의 소지역 추정의 문제에는 소지역 추정량 발표시 지방자치단체별로 ranking을 매기는 등의 서열화로 인한 문제점이 발생할 수 있을 것으로 우려하고 있음. 이와 관련하여 일본은 어떤 식으로 이 문제에 대처하였는가?

A. 추정량과 공표에 관해 직접 이해를 받을 수 있도록 했음.

Q27. 한국보다 소지역 추정 통계량을 먼저 발표한 나라로써 소지역 추정 통계를 작성하는 단계에서 한국의 연구자들에게 어떤 조언을 해주고 싶은가?

A. 연구회에 대해서 “이 전문가에 의한 검토라면 이후 문제점(혹은 문제제기)은 나오지 않을 것”이라고 할 수 있는 전문가를 선택하는 것이 중요

- 일본의 경우 그 시점에서 가능한 한 폭넓은 소지역 추정방법에 대해서 추정을 했음. 그래서 평가기준을 명확히 한 다음 가능한 많은 후보 가운데 추정방법을 선정.
- 추정결과를 공표할 때에는 이것을 바람직하지 않다고 생각하는 지역도 있을 수 있음. 그때는 그들 지역과의 대화와 설명이 중요.
- 인사이동에 의한 시스템, 전문지식 단절이 없도록 매뉴얼 구비, 연수 등이 중요

[별첨 3] 일본 노동력 조사에 대한 소지역 추정 방법 적용

<자료 설명>

○ 본 자료는 일본의 소지역 추정연구를 하는 과정에서 검토한 방법 중 하나로, 최종적으로 선택된 시계열 모형 검토 방법 및 과정을 번역한 것이다. 따라서 이 방법을 선택하는 과정에서 일본이 가장 고려했던 부분들을 이해하는데 많은 도움이 될 것으로 생각한다.

○ 또한 이 방법은 미국 BLS에서 채택하고 있는 시계열 모형을 응용한 것으로, 일본이 소지역 통계의 작성과정에서 상당 부분 미국의 자료를 참고한 것임을 밝혀둔다.

○ 이론적 부분이 많이 설명되어 있으므로, 이 방법을 적용하는 데 검토되어야 할 이론적 부분을 자세하게 알 수 있을 것이다.

○ 또한 일본이 이 방법을 노동력통계에 적용하면서 고려했던 측면과 적용절차를 쉽게 이해할 수 있을 것이다.

<요 약>

미국 BLS(Bureau of Labor Statistics)는 주(state)별 완전실업율 결과를 매월 공표하고 있다. 이때 시계열회귀모형 및 상태공간모형 방법을 이용한다. 본 자료에서는 미국의 추정 방법을 일본 노동력조사에 적용해 도도부현별 완전실업율을 추정한 결과에 대해서 정리한다. 1장에서는 미국 소지역 추정에 이용된 모형에 대해서 설명한다. 2장에서는 추정에서 중요한 상태공간모형 및 칼만-필터 이론을 소개한다. 3장에서는 이 외의 관심사항에 대해서 다룬다. 4장에서는 일본 노동력 결과에 적용한 결과 등에 대해서 정리한다. 마지막으로 이후 과제에 대해서 서술하고 있다.

Ⅰ. 미국의 소지역 추계 방법

1. 시그널(Signal) + 노이즈(Noise) 모형

소지역 별 관측 자료를 시그널과 노이즈라고 하는 독립적인 두 확률과정의 합으로 표현한다.

$$y(t) = \theta(t) + e(t) \quad (1)$$

여기서, $y(t)$: 관측치, $t = 1, \ldots, T$

$\theta(t)$: 참값에 대한 추정치 (시그널)

$e(t)$: 표본오차에 대한 추정치 (노이즈)

이와 같은 모형을 시그널+노이즈모형이라 부르며, $y(t)$에서 $e(t)$를 빼는 것이 이 방법의 목적이다. $\theta(t)$는 다음과 같이 네 개의 항으로 분해할 수 있다고 가정한다.

$$\theta(t) = M(t) + T(t) + S(t) + I(t) \quad (2)$$

여기서, $M(t)$: 회귀항, $T(t)$: 트렌드항

$S(t)$: 계절항, $I(t)$: 불규칙 변동항

$M(t)$, $T(t)$, $S(t)$, $I(t)$ 및 $e(t)$가 따르는 모형에 대해서 다음과 같이 설명한다. 또 $e(t)$ 모형은 다소 복잡하기 때문에 다음 절에서 자세하게 설명한다.

(1) 회귀항 $M(t)$ 모형

회귀항 $M(t)$는 $\theta(t)$ 가운데 보조정보에 의해서 설명되어지는 부분이다.

$$M(t) = X(t)\beta(t) \quad (3)$$

$$\beta(t) = \beta(t-1) + v_\beta(t) \quad (4)$$

여기서, $X(t)$: 보조정보(설명변수), $v_\beta(t) \sim N(0, \sigma_\beta^2)$

회귀항은 단순회귀를 가정한다. 식(4)는 회귀계수 $\beta(t)$가 시간과 함께 완만하게 (느리게) 변화하는 것을 허용하고 있다.

(2) 트렌드항 $T(t)$ 모형

트렌드 항 $T(t)$는 $\theta(t)$ 가운데 완만한 변동을 찾아내기 위해 삽입된다. 트렌드 항은 각 시점에서 수준을 나타내는 부분 $T(t)$와 기울기를 나타내는 부분 $R(t)$를

이용해서 다음과 같이 표현된다.

$$T(t) = T(t-1) + R(t-1) + v_T(t) \quad (5)$$

$$R(t) = R(t-1) + v_R(t) \quad (6)$$

여기서, $v_T(t) \sim N(0,\sigma_T^2)$, $v_R(t) \sim N(0,\sigma_R^2)$

$\sigma_R^2 = \sigma_T^2 = 0$인 경우에는 기울기 $R(t)$가 일정한 1차 다항식(직선)을 트렌드로 가정하는 셈이 된다. $\sigma_R^2 > 0$ 또는 $\sigma_T^2 > 0$ 인 경우에는 기울기 및 수준이 각각 독립으로 랜덤워크(random walk)에 따라서 변화하는 유연한 모형이 된다.

회귀항에 의해서 트렌드를 충분히 설명할 수 있을 경우에는 트렌드항이 필요없게 되고 자동적으로 모형에서 빠진다. 이 항은 설명변수에서 추려내고 남은 잔차 트렌드(residual trend)를 끄집어내기 위해 도입된 것이다.

(3) 계절항 $S(t)$ 모형

계절항은 시그날 가운데 계절적인 변동을 추출하기 위해 도입된다. 계절적인 변동으로는 시계열 성분 가운데 주기가 일년인 것을 가리킨다. 계절항은 계절주파수(12개월 주기, 6개월 주기, 4개월 주기, 3개월 주기, 2.4개월 주기, 2개월 주기)에 대응하는 6개 삼각계수 상의 합으로서 다음과 같이 표현된다.

$$S(t) = \sum_{j=1}^{6} S_j(t) \quad (7)$$

여기서, $S_j(t) = \cos(\omega_j) S_j(t-1) + \sin(\omega_j) S_j^*(t-1) + v_{S_j}(t)$,

$$S_j^*(t) = -\sin(\omega_j) S_j(t-1) + \cos(\omega_j) S_j^*(t-1) + v_{S_j^*}(t),$$

$$v_{S_j}(t) \sim N(0,\sigma_S^2),\ v_{S_j^*}(t) \sim N(0,\sigma_S^2),\ \omega_j = \frac{2\pi j}{12}$$

$\nu_{s_j}(t)$, $\nu_{s_j^*}(t)$는 모두 상호 독립으로 분산이 동일한(σ_S^2) 정규분포를 따른다. 회귀항에 의해서 계절성을 충분히 설명할 수 있는 경우에는 이 항이 필요없게 되고, 자동적으로 모형에서 빠진다. 이 항은 설명변수로 설명되고 남은 잔차 계절성(residual seasonality)을 끄집어내기 위해 도입된다.

(4) 불규칙 변동항 $I(t)$ 모형

불규칙 변동항은 위의 각 항에 포함되지 않고 남은 변동을 나타낸다. 여기서는 경기의 단기적인 변동 등이 들어갈 가능성이 있다.

$$I(t) = v_I(t) \tag{8}$$

여기서, $v_I(t) \sim NID(0, \sigma_I^2)$

$\theta(t)$가 $M(t), T(t), S(t)$에 의해서 거의 설명될 경우에는 이 항은 필요 없게 되고 자동적으로 모형에서 빠진다.

2. 시그날 모형 일부 변경

앞 절에서 소개한 모형 가운데 $T(t), S(t)$에 대해서는 일본 노동력 조사에 적용하는 과정에서 약간 변경을 더했다. 다음에서 변경된 점을 설명한다.

(1) $T(t)$ 모형 변경

트렌드항 $T(t)$을 식(5), (6)으로 정의할 경우, $T(t)$가 상당히 변동할 것으로 예상된다. 회귀항 $M(t)$ 계수도 시간과 함께 변동한다고 가정하기 때문에 $T(t)$는 더욱 단순한 모형이 좋을 것으로 생각된다. 그래서 $T(t)$를 아래와 같은 1차 트렌드로 가정한다.

$$T(t) = T(t-1) + \nu_T(t) \tag{9}$$

1차 트렌드라도 원계열 트렌드를 충분히 유지할 수 있을 것으로 생각된다.

(2) $S(t)$ 모형 변경

계절항 $S(t)$는 아래의 DECOMP(北川, 1993, 1997)의 정의를 사용한다.

$$S(t) = -\sum_{j=1}^{11} S(t-j) + v_S(t) \tag{10}$$

식(10)에 의한 정의가 모형이 간결하고 나중에 설명할 상태공간모형을 이용할 때에도 설명하기가 간단해진다.

3. $e(t)$ 모형

표본오차 $e(t)$ 모형이 미국 소지역 추정 가운데 가중 중요한 부분이다. 표본에는 로테이션 구조가 있는 경우 표본오차에 계열 상관이 있을 것으로 생각된다. 또

표본오차 분산은 ① 표본수, ② 표본설계 변경, ③ 참값 $\theta(t)$와 같은 3개의 영향을 받아서 시간과 함께 변한다고 생각할 수 있다. 따라서 $e(t)$ 모형에 다음과 같은 가정을 둔다.

- 가정1 : 표본오차항 $e(t)$는 계열 상관을 갖는다.
- 가정2 : 표본오차항 $e(t)$의 분산은 시점 t에 따라 다르다.

표본설계에서 자연히 사용되는 이들 2개의 가정을 $e(t)$ 모형에 넣었기 때문에 표본오차항 $e(t)$는 아래와 같은 승법형태의 모형으로 표현할 수 있다.

$$e(t) = \gamma(t)\, e^{*}(t) \tag{11}$$

여기서, $\gamma(t)$: 표본오차 분산 변화를 나타내는 스칼라 값

$e^{*}(t)$: 로테이션 구조에 의한 계열 상관을 나타내는 AR(13) 과정

$e^{*}(t)$는 표본오차항 $e(t)$의 변동 패턴을 나타내는 AR(13)모형이다. $\gamma(t)$는 표본오차항 $e(t)$ 변동의 진폭을 나타낸다. 이들 움직임이 결합해서 표본오차 $e(t)$가 된다. $e^{*}(t)$와 $\gamma(t)$ 추정방법에 대해서 다음과 같이 설명한다.

(1) $e^{*}(t)$ 추정방법

표본오차는 1차·11차·13차에 있어서 자기상관을 갖는다. 이에 AR(13) 모형으로 $e^{*}(t)$를 표현한다.

$$e^{*}(t) = \sum_{j=1}^{13} \phi_j e^{*}(t-j) + v_e(t) \tag{12}$$

여기서, $v_e(t) \sim N(0, \sigma_{v_e}^2)$

표본오차의 자기상관 $\hat{\rho}(k)$는 제 4장에서 설명할 방법으로 추정할 수 있다. 이에 의해 $e^{*}(t)$ AR 계수 $\phi_i (i = 1, \cdots, 13)$는 아래의 Yule-Walker 방정식을 풀어 추정할 수 있다.

$$\begin{bmatrix} 1 & \hat{\rho}(1) & \cdots & \hat{\rho}(12) \\ \hat{\rho}(1) & 1 & \cdots & \hat{\rho}(11) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}(12) & \hat{\rho}(11) & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{13} \end{bmatrix} = \begin{bmatrix} \hat{\rho}(1) \\ \hat{\rho}(2) \\ \vdots \\ \hat{\rho}(13) \end{bmatrix} \quad \text{(Yule-Walker 방정식)} \tag{13}$$

(2) $\gamma(t)$ 추정방법

각 시점 t에 의한 표본오차 $e(t)$ 분산을 $\sigma_e^2(t)$라 하자. $\sigma_e^2(t)$는 매월 집계자료로 추정할 수 있다. $e(t)$분산이 $\sigma_e^2(t)$가 같게 되도록 조정하기 위한 계수가 $\gamma(t)$이다. 그렇기 하기 위해서는 아래와 같이 $\gamma(t)$를 구성하면 된다.

우선 AR과정을 MA(∞)과정으로 전환시킨다. 그때 AR 가정인 정상성 조건은 만족된다고 가정한다.

$$e^*(t) = \sum_{k=0}^{\infty} g_k v_e(t-k) \tag{14}$$

여기서, g_k 는 응답함수이고,

$$g_k = 1,\ g_k = \sum_{j=1}^{k} \phi_j g_{k-j}, \quad (k > j \text{ 일 때 } \phi_j = 0)$$

이때 $e^*(t)$의 분산 $\sigma_{g^*}^2$은 다음과 같이 계산할 수 있다.

$$\sigma_{e^*}^2 = var(e^*(t)) = var(\sum_{k=0}^{\infty} g_k v_e(t-k)) = \sigma_{v_e}^2 \sum_{k=0}^{\infty} g_k^2 \tag{15}$$

$e^*(t)$의 자기상관이 그다지 크지 않으면 g_k^2는 급속하게 감소하기 때문에 충분히 많은 항수를 넣으면 우변의 무한급수 $\sum_{k=0}^{\infty} g_k^2$ 는 일정한 값으로 수렴하게 된다. $e(t)$의 분산이 $\sigma_e^2(t)$와 같게 되도록 하기 위해서는 식(15)에 의해 $\gamma(t)$를 다음과 같이 정한다.

$$\gamma(t) = \sqrt{\frac{\sigma_e^2(t)}{\sigma_{v_e}^2 \sum_{k=0}^{\infty} g_k^2}} \tag{16}$$

모형으로부터 계산한 $e(t)$의 분산이 개별 자료로부터 계산한 $e(t)$의 분산 $\sigma_e^2(t)$에 임의의 시점 t에서 일치한다는 것은 아래 같은 간단한 계산에 의해 알 수 있다.

$$var(e(t)) = var(\gamma(t)e^*(t)) = \gamma(t)^2 var(e^*(t))$$

$$= \frac{\sigma_e^2(t)}{\sigma_{\nu_e}^2 \sum_{k=0}^{\infty} g_k^2} \times \sigma_{\nu_e}^2 \sum_{k=0}^{\infty} g_k^2 = \sigma_e^2(t)$$

위의 계산에서 $\sigma_{\nu_e}^2$은 약분되어 없어지기 때문에 임의의 값으로 사용하기 좋다. 계산 편의상 $\sigma_{\nu_e}^2 = 1$이라 하자. 이에 의해 $\gamma(t)$의 추정치는 아래와 같다.

$$\gamma(t) = \sqrt{\frac{\sigma_e^2(t)}{\sum_{k=0}^{\infty} g_k^2}} \tag{17}$$

Ⅱ. 상태공간모형에 의한 시그날 추출

1. BLS 모형인 상태공간모형에 의한 표현

상태공간모형은 다음과 같은 2개의 방정식으로 구성된 동적인 시스템이다.

$$y(t) = \boldsymbol{H(t)\alpha(t)} + w(t), w(t) \sim N(0, \sigma_I^2) \ : \text{관측방정식} \tag{18}$$

$$\boldsymbol{\alpha(t)} = \boldsymbol{F\alpha}(t-1) + \boldsymbol{V}(t), \boldsymbol{V}(t) \sim N(0, \boldsymbol{Q}) \ : \text{전이방정식} \tag{19}$$

여기서, $\boldsymbol{F}$: 전이행렬, $\boldsymbol{\alpha}(t)$:상태변수 벡터, $\boldsymbol{H}(t)$: 관측행렬, $w(t)$: 관측 노이즈, $V(t)$: 시스템 노이즈

초기상태 $\alpha(0)$는 정규분포를 따른다고 하자. $\boldsymbol{F}$, $\boldsymbol{H}(t)$는 전부 알려져 있다고 가정한다. $\boldsymbol{Q}$와 σ_I^2에 대해서는 초기상태에서는 알려져 있다고 가정하고, 나중에 추정한다. 상태변수 $\boldsymbol{\alpha(t)}$는 직접 관측할 수 없고 식(18)의 관측방정식을 통해서 관측노이즈 $w(t)$가 부가되어 실제 관측치로 관측된다. 또 $\boldsymbol{\alpha(t)}$는 식(19)의 이동방정식에 의해서 시간과 함께 변한다. 상태공간모형에 의한 분석은 다음과 같은 장점이 있다.

- $\boldsymbol{\alpha(t)}$의 동적인 특성(전이방정식)에 의해 각 상태변수가 시간과 함께 완만하게 변화하는 모습을 잘 반영하도록 추정할 수 있다.

▪ 온라인·리얼타임 처리가 가능하다. 즉 새로운 관측치가 관측될 적에 그 시점에 의한 적절한 추정치를 좋은 효율을 갖도록 구할 수 있다.

$\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{H}(t)$, $\boldsymbol{Q}$를 아래와 같이 설정함에 따라, Ⅰ장에서 설명한 모형을 식(18), (19)와 같은 상태공간모형 형태로 표현할 수 있다.

$$\boldsymbol{\alpha}(t) = \left[\beta(t)|T(t)|S(t)\cdots S(t-10)|e_1(t)\cdots e_{13}(t)\right] \tag{20}$$

$$\boldsymbol{H}(t) = [X(t)|\ 1\ |\ 1\ \ 0\cdots 0|\gamma(t)\quad 0\cdots 0]$$

$$\boldsymbol{F}=\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & -1 & -1 & \cdots & -1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & -1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \phi_1 & \phi_2 & \cdots & \phi_{13} \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & & 1 & 0 \end{bmatrix},\quad \boldsymbol{G}=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix},\quad \boldsymbol{Q}=\begin{bmatrix} \sigma_\beta^2 & 0 & 0 & 0 \\ 0 & \sigma_T^2 & 0 & 0 \\ 0 & 0 & \sigma_S^2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

이때 상태변수 벡터 $\boldsymbol{\alpha}(t)$의 차원은 26이 된다. $\Theta=(\sigma_\beta^2\ \ \sigma_T^2\ \ \sigma_S^2\ \ \sigma_I^2)$은 상태변수의 변화 움직임을 결정하는 모수(parameter)이고 이를 초모수(hyper-parameter)라 부른다.

각 시점에 의한 상태변수 $\boldsymbol{\alpha}(t)$를 추정할 수 있으면 $\boldsymbol{H}(t)$대신에 다음의 $\boldsymbol{H_\theta}(t)$를 이용함으로써 관측치 자료 $y(t)$로부터 표본오차 $e(t)$를 뺄 수 있다.

$$\boldsymbol{H_\theta}(t) = [X(t)|1|1\ \ 0\cdots 0|0\cdots 0] \tag{21}$$

$\boldsymbol{\alpha}(t)$의 효율적인 추정방법에 대해서는 다음 절에서 설명한다.

2. 칼만-필터에 의한 상태변수 추정

관측치로부터 표본오차를 제거하기 위해서는 모형 선택, 모수 추정 및 시계열 분해를 위한 계산이 필요하다. 상태공간모형으로 표현된 시스템에 대해서 칼만-필터에 의해 좋은 효율을 갖도록 계산할 수 있다. 칼만 필터는 관측방정식과 전이방정식을 연결해서 각 시점에 의한 상태변수의 평균제곱오차를 최소로 하는 추정값을 효율 좋게 추정하는 일련의 방정식이다. 사용 가능한 관측값의 정도에 따라 상태변수의 평균을 추정하는 문제는 아래와 같이 구별된다(원문 그림 1 참조).

- 일기선예측 · · · 시점 t-1까지의 관측값이 이용가능한 경우
- 필터 · · · 시점 t까지 관측값이 이용가능한 경우
- 평활화 · · · 시점 T(T>t)까지 관측값이 이용가능한 경우

전이방정식과 관측방정식은 선형이고, 또 $\boldsymbol{\alpha}(0), w(t), \boldsymbol{V(t)}$는 모두 정규분포를 따른다. 이에 의해 관측값이 반영된 때의 상태변수의 조건부분포는 정규분포를 따르고 그것은 평균과 분산에 의해 완전히 결정된다. 따라서 상태변수 추정시에는 평균과 분산 추정만을 생각하면 된다. 칼만 필터를 적용하면 상태변수의 평균과 분산의 예측치 및 필터 값을 순차적으로 높은 효율을 갖도록 추정할 수 있다. 그래서 그들 값을 이용한 역방향 순차적 계산에 의해 평활화 값을 추정할 수 있다.

각 성분의 확률분포 분산이 알려진 경우, 칼만 필터에 의한 상태변수의 평균과 분산 예산측 및 필터의 값은 다음과 같다.

$$\text{일기선예측}\ \begin{cases} \boldsymbol{\alpha}(t|t-1) = \boldsymbol{F}\boldsymbol{\alpha}(t-1|t-1) \\ \boldsymbol{P}(t|t-1) = \boldsymbol{F}\boldsymbol{P}(t-1|t-1)\boldsymbol{F}' + \boldsymbol{G}\boldsymbol{Q}\boldsymbol{G}' \\ y(t|t-1) = \boldsymbol{H}\boldsymbol{\alpha}(t|t-1) \\ \tilde{y}(t) = y(t|t) - y(t|t-1) \quad \cdots \quad \text{일기선 예측 오차} \end{cases} \tag{22}$$

$$\text{필터}\ \begin{cases} f(t|t-1) = \boldsymbol{H}(t)\boldsymbol{P}(t|t-1)\boldsymbol{H}(t)' + \sigma_I^2 \quad \cdots \quad \text{일기선 예측 오차의 분산} \\ \boldsymbol{K}(t) = \boldsymbol{P}(t|t-1)\boldsymbol{H}(t)'/f(t|t-1) \quad \cdots \quad \text{칼만-게인} \\ \boldsymbol{\alpha(t|t)} = \boldsymbol{\alpha}(t|t-1) + \boldsymbol{K}(t)\tilde{y}(t) \\ \boldsymbol{P}(t|t) = (\boldsymbol{I} - \boldsymbol{K}(t)\boldsymbol{H}(t)')\boldsymbol{P}(t|t-1) \end{cases} \tag{23}$$

또 여기서는 $\boldsymbol{t_1}$기까지의 관측치를 얻을 수 있는 경우, $\boldsymbol{t_2}$기에 의한 상태변수의 평균 및 분산공분산 행렬의 추정치를 각각 $\boldsymbol{\alpha}(t_2|t_1)$ 및 $\boldsymbol{p}(t_2|t_1)$으로 나타낼 수 있다. 식의 도출 등의 상세한 내용에 대해서는 北川(1993) 및 Harvey(1989)를 참조하기 바란다.

상태변수의 평균 필터 $\boldsymbol{\alpha}(t|t)$는 1기 전 자료에 기반한 예측 $\boldsymbol{\alpha}(t|t-1)$과 지금 시점에서 관측치 $y(t)$와의 가중합이 된다. 현시점의 관측 자료가 관측되었을 때, 칼만-게인 $\boldsymbol{K}(t)$에 기초해서 1기선예측 $\boldsymbol{\alpha}(t|t-1)$을 수정한 것이지만 $\boldsymbol{\alpha}(t|t)$이라고 해석할 수 있다. 또, 상태변수의 분산공분산행렬의 필터 $\boldsymbol{P}(t|t)$는 현 시점의 관측치가 관측됨에 의해 우변의 $\boldsymbol{K}(t)\boldsymbol{H}(t)'$만 분산이 개선되었다(작아졌다)고 해석할 수 있다.

시점 T까지의 관측치를 이용할 수 있을 경우, 1기선예측과 필터 결과를 이용해서 다음과 같은 후향 순차 계산에 의해 평활화 할 수 있다.

$$\text{평활화} \quad \begin{cases} A(t) = \boldsymbol{P}(t|t)\boldsymbol{F}'\boldsymbol{P}(t+1|t) \\ \boldsymbol{\alpha}(t|T) = \boldsymbol{\alpha}(t|t) + \boldsymbol{A}(t)(\boldsymbol{\alpha}(t+1|T) - \boldsymbol{\alpha}(t+1|t)) \\ \boldsymbol{P}(t|T) = \boldsymbol{P}(t|t) + \boldsymbol{A}(t)(\boldsymbol{P}(t+1|T) - \boldsymbol{P}(t+1|t))\boldsymbol{A}(t)' \end{cases} \tag{24}$$

3. 초모수 및 초기상태 추정

지금까지는 초모수 $\boldsymbol{\Theta}=(\sigma_\beta^2 \;\; \sigma_T^2 \;\; \sigma_S^2 \;\; \sigma_I^2)$가 알려져 있다고 가정하였다. $\boldsymbol{\Theta}$의 취급에 따라서 시계열 분해도 상당히 변하기 때문에 관측치 자료로부터 어떠한 방법으로든 합리적인 $\boldsymbol{\Theta}$를 결정할 필요가 있다. 거기서 관측된 자료를 근거로 해서 대수우도를 계산하고 최우추정법에 의해 $\boldsymbol{\Theta}$를 추정한다. 1기선예측오차 $\tilde{y}(t)$는 평균 0, 분산 $f(t|t-1)$인 정규분포를 따른다. 따라서 기간 $1 \leqq t \leqq T$에 의한 관측치에 대한 대수우도함수 $L(\boldsymbol{\Theta})$는 다음과 같이 계산된다 (Harvey, 1989).

$$L(\boldsymbol{\Theta}) = \log\left[\prod_{t=1}^{T} \frac{1}{\sqrt{2\pi f(t|t-1)}} exp\left(\frac{\tilde{y}(t)^2}{f(t|t-1)}\right)\right] \tag{25}$$

$$= -\frac{T}{2}log2\pi - \frac{T}{2}\sum_{t=1}^{T}\log(f(t|t-1)) - \frac{T}{2}\sum_{t=1}^{T}\frac{\tilde{y}(t)^2}{f(t|t-1)}$$

식(25)는 $L(\boldsymbol{\Theta})$를 최대화 하는 $L(\boldsymbol{\Theta})$가 초모수인 최우추정량이 된다.

칼만-필터 공식과 식(25)을 이용하는 것으로 다양한 $\boldsymbol{\Theta}$에 대한 우도함수를 매우 효율이 좋게 계산할 수 있다. 식(25)는 복잡한 비선형 함수이고 그 최대치를 갖는 모수를 해석적으로 얻기가 어렵기 때문에 수치계산에 의한 비선형최적화를 한다. 비선형최적화에는 다양한 방법이 있는데, 본 분석에서는 R의 비선형최적화 패키지인 'optim'을 이용해서 대수우도함수의 최대화를 얻어냈다. $\boldsymbol{\Theta}$의 각 성분은 분산이기 때문에 양의 값일 필요가 있다. 거기서 각 성분을 (예를 들어 $\sigma_\beta^2 = \exp(\tau)$와 같이) 변환해서 τ 추정을 하기로 한다.

상태공간모형 및 칼만-필터를 이용한 분해를 위해서는 상태변수의 평균 및 분산의 초기치를 설정해야 한다. Tiller(1989)는 최초의 다수 개 시점의 관측치로부터 초기치를 설정했다. 상태변수 벡터 차원이 D인 경우에 $t \leqq D$에 의한 관측치에 대해서 칼만-필터를 적용해서 얻어진 $\boldsymbol{\alpha}(D|D)$ 및 $\boldsymbol{P}(D|D)$를 상태변수 평균 및 분산의 초기치라 하자. $\boldsymbol{P}(0|0)$를 크게 함으로써 초기상태 $\boldsymbol{\alpha}(0|0)$의 영향을 작게 할 수 있을 뿐 아니라, 안정된 초기상태를 얻기 위해서는 충분히 장시간 동안 칼만-필터를 작동시켜야 한다. 그러나 가장 좋은 $\boldsymbol{\alpha}(0|0)$ 값을 결정할 경우에는 비교적 단시간에 초기상태를 안정시킬 수 있다. 이번 분석에서는 $\boldsymbol{\alpha}(0|0)$ 및 $\boldsymbol{P}(0|0)$를 다음과 같이 설정한다. 또 사전 정보가 부족한 상태를 표현하기 위해 분산의 초기치를

크게 설정하고 있다.

$$\boldsymbol{P}(0|0) = 10000 \times \boldsymbol{V}_y \boldsymbol{I} \tag{26}$$
$$\boldsymbol{\alpha}(0|0) = \left[\widetilde{\beta}_1 | \widetilde{\beta}_2 | 0 \ \cdots \ 0\right]$$

여기서, $\boldsymbol{V}_y$: 최초 2년간 자료 $y(t)$로부터 구한 표본 분산

$\boldsymbol{I}$: 상태변수와 같은 차원(D)의 단위행렬

$\widetilde{\beta}_1$, $\widetilde{\beta}_2$: 최초 2년 자료 $y(t)$와 $X(t)$로부터 구한 회귀 계수

Ⅲ. 모형 진단 및 추정치 정도 평가

1. 자기상관성, 정규성 및 분산 동일성 검정

기준화한 1기선예측오차 $\tilde{y}(t)$를 아래와 같이 정의한다.

$$\tilde{y}(t) = \frac{y(t) - y(t|t-1)}{\sqrt{\boldsymbol{H}(t)\boldsymbol{P}(t|t-1)\boldsymbol{H}(t)' + \sigma_I^2}} \tag{27}$$

이것은 칼만-필터를 이용하여 효율이 좋은 계산이 가능하다.

가정한 모형이 자료에 잘 적합하고, 모수 $\boldsymbol{\Theta}$가 알려진 경우에는 $\tilde{y}(t)$는 평균 0, 분산 1인 상호 독립인 정규분포를 따른다고 할 수 있다. 이것을 이용해서 추정한 모형을 진단할 수 있다. 진단방법으로는 자기상관, 정규성, 분산 동일성 검정을 생각할 수 있다(Durbin과 Koopman, 2001).

다음에서 각종 진단방법에 대해서 설명한다. 또 식(28) 가운데 D는 상태변수벡터 차원을 나타내고, 이 기간 상태변수는 초기상태를 구하는데 사용하고 있기 때문에 진단 시에 사용하기로 한다 (Ⅱ절 3 참고).

(1) 자기상관성 검정 (Ljung-Box 통계량)

$\tilde{y}(t)$의 자기상관 검정에는 Ljung-Box 통계량을 이용한다.

$$L.B(m) = (n-D)(n-D+2)\sum_{l=1}^{m} \frac{\hat{\gamma}_l^2}{n-D-l} \tag{28}$$

$$\text{여기서, } \hat{\gamma_l^2} = \frac{\sum_{t=D+1+l}^{T} \tilde{y}(t)\tilde{y}(t-l)}{\sum_{t=D+1}^{k} \tilde{y}(t)^2}$$

예측오차가 자기상관을 갖지 않을 때 위의 검정통계량은 자유도 k인 χ^2 분포를 따른다. 만약 5% 유의수준에서 유의하면 $\tilde{y}(t)$에 자기상관이 있다고 간주한다. 본 분석에서는 자유도의 차수 k로 12와 24를 선택하였다.

(2) 정규성 검정 (Jarque-Bera 통계량)

오차 정규성 검정에는 이하 Jarque-Bera 통계량을 이용한다.

$$J.B = n\left[\frac{S^2}{6} + \frac{(k-3)^2}{24}\right] \tag{29}$$

$$\text{여기서, } S = \frac{\frac{1}{n}\sum_{t=D+1}^{T} \tilde{y}(t)^3}{\left[\sqrt{\frac{1}{n}\sum_{t=D+1}^{T} \tilde{y}(t)^2}\right]^3} \quad , \; K = \frac{\frac{1}{n}\sum_{t=D+1}^{T} \tilde{y}(t)^4}{\left[\frac{1}{n}\sum_{t=D+1}^{T} \tilde{y}(t)^2\right]^2}$$

여기서 S는 표본의 왜도, K는 표본의 첨도를 나타낸다. 예측오차의 분포가 정규분포이면 Jarque-Bera 통계량은 자유도 2인 χ^2분포를 따른다. 유의수준 5%에서 유의하면 $\tilde{y}(t)$는 정규분포를 따른다고 간주한다.

(3) 분산 동일성 검정

F검정 통계량에 의해 분산 동일성 정도를 검정한다.

$$F_{m|m} = \frac{\sum_{t=T-m+1}^{T} \tilde{y}(t)^2}{\sum_{t=D+1}^{m} \tilde{y}(t)^2} \tag{30}$$

$$\text{여기서, } m = \frac{T-D}{3}$$

$\tilde{y}(t)$의 최초 1/3과 최후 1/3을 이용해서 분산 동일성 검정을 행한다. 1기선예측오차의 분산이 동일하다면 검정통계량 H 는 자유도 (m,m)인 F분포를 따른다고 볼 수 있다. 유의수준 5%에서 유의하다면 $\tilde{y}(t)$의 분산은 동일하지 않은 것으

로 간주한다.

2. 평활화 후의 분산에 의한 정도 평가

평활화에 의해 추정한 시그날 $\theta(t)$의 분산은 다음과 같이 추정할 수 있다.

$$var(\theta(t)) = \boldsymbol{H}_\theta(t)\boldsymbol{P}(t|t)\boldsymbol{H}_\theta(t) + \sigma_I^2 \tag{31}$$

여기서, $\boldsymbol{H}_\theta(t) = [X(t)|\ 1\ |\ 1\ 0\cdots 0|0\cdots 0]$

여기서 $\boldsymbol{H}_\theta(t)$는 시그날을 추출하기 위한 관측행렬이다(2절 1 참고).
식(28)로부터 아래와 같이 $\theta(t)$의 오차율 $CV_\theta(t)$를 구할 수 있다.

$$CV_\theta(t) = \sqrt{var(\theta(t))}/\theta(t) = \frac{\sqrt{H_\theta(t)P(t|t)H_\theta(t) + \sigma_I^2}}{\theta(t)} \tag{32}$$

이것을 관측치의 오차율과 비교하면 정도의 향상 정도를 파악할 수 있다.

3. 벤치마킹

BLS에서는 시계열회귀모형에 의해 구한 추정치의 정도를 더욱 높이기 위해 벤치마킹이라는 방법을 이용하고 있다. 이것은 모형에 의한 추정치의 연평균과 원래 관측치의 연평균이 일치하도록 조정하는 방법이다. 관측치를 월차별로 본 경우에는 표본수가 적고 불안정하지만 1년간 분량의 관측치 자료를 전체로 보면 추정의 정도가 증가하고, 보다 정확한 추정치가 될 것으로 생각된다. 따라서 관측치의 연평균에 모형으로 구한 연평균을 맞춰줌으로써 모형 결과에 대한 정도 향상을 기대할 수 있다. Tiller(1989)에서는 Denton-Method(Denton, 1971)라 부르는 방법을 소개하고 있다. 본 분석에서는 이 방법에 의한 벤치마킹을 수행하였다.

모형에 의해 추정된 계열 $\boldsymbol{z}=(z_1,z_2,\cdots,z_{12k})^t$ 변동과 가능한 같은 움직임을 갖는 시계열로, 그 연평균이 관측치 $\boldsymbol{y}=(y_1,y_2,\cdots,y_k)^t$의 연평균과 가깝게 되도록 시계열 $\boldsymbol{x}=(x_1,x_2,\cdots,c_{12k})^t$를 추정하는 문제로서 벤치마킹을 생각할 수 있다.

이 경우 $\boldsymbol{z}$와 같은 움직임을 갖는 시계열 $\boldsymbol{x}$를 수학적으로 어떻게 정의할 것인가가 문제이다. Denton(1971)은 이것을 층차의 제곱합이 최소로 되는 시계열로서 다음과 같이 정의하고 있다.

$$\sum_{i=1}^{12k}(\Delta x_i-\Delta z_i)^2 \quad \left(=\sum_{i=1}^{12k}(\Delta(x_i-z_i))^2\right) \text{ 를 최소로 하는 } x \qquad (33)$$

이때 위의 벤치마킹 문제를 다음과 같이 선형제약조건의 2차형식의 최소화 문제로 바꿀 수 있다.

$$\frac{1}{12}\sum_{i=12T-11}^{12T} x_i=\frac{1}{12}\sum_{i=12T-11}^{12T} 12y_t \left(\Leftrightarrow \sum_{i=12T-11}^{12T} x_i=\sum_{i=12T-11}^{12T} 12y_t\right) \quad (1\leqq T\leqq k) \quad (34)$$

목적함수 : $\mathrm{u}=(\mathrm{x}-\mathrm{z})'\mathrm{D}'\mathrm{D}(\mathrm{x}-\mathrm{z})-2\lambda^{\mathrm{t}}(\mathrm{y}-\mathrm{B}'\mathrm{x})$

여기서, z_i : 최초부터 i월째의 시계열 모형에 의한 추정치 $(1\leqq i\leqq 12k)$

x_i : 최초부터 i월째의 벤치마킹 후 값 $(1\leqq i\leqq 12k)$

$\lambda=(\lambda_1,\lambda_2,\cdots,\lambda_{12k})^t$: Lagrange 미정 승수

$$\mathrm{D}=\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix},\ \mathrm{B}=\begin{bmatrix} \mathrm{j} & \mathrm{0} & \cdots & \mathrm{0} \\ \mathrm{0} & \mathrm{j} & \cdots & \mathrm{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{0} & \mathrm{0} & \cdots & \mathrm{j} \end{bmatrix},$$

$\mathrm{j}=(1,1,\cdots,1),\quad \mathrm{0}=(0,0,\cdots,0)$

식(34)의 결과는 Lagrange 미정 승수법에 의해 구할수 있고, 다음의 식으로 나타낼 수 있다.

$$\begin{bmatrix}\mathrm{x}\\ \lambda\end{bmatrix}=\begin{bmatrix}\mathrm{A} & \mathrm{B}\\ \mathrm{B}' & \mathrm{0}\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{A} & \mathrm{0}\\ \mathrm{B}' & \mathrm{I}\end{bmatrix}\begin{bmatrix}\mathrm{z}\\ \mathrm{y}-\mathrm{B}'\mathrm{z}\end{bmatrix} \qquad (35)$$

여기서, $\mathrm{A}=\mathrm{D}'\mathrm{D}$

이에 의해 벤치마킹 후의 값은 부분 행렬의 역행렬에 관한 공식을 식(35)에 적용하여 아래와 같이 추정할 수 있다.

$$\mathrm{x}=\mathrm{z}+\mathrm{A}^{-1}\mathrm{B}(\mathrm{B}^{\mathrm{t}}\mathrm{A}^{-1}\mathrm{B})^{-1}(\mathrm{y}-\mathrm{B}^{\mathrm{t}}\mathrm{z}) \qquad (36)$$

Ⅵ 노동력 조사에 적용

1. 추정방법 및 결과 비교 방법

BLS의 소지역 추정방법을 노동력 조사에 적용하여 도도부현별 완전실업율을 추정한다. 추정 절차는 다음과 같다.

- 헤세3년(1991) 1월 ~ 헤세15년 12월(2003) 관측 자료를 이용해서 추정
- 도도부현별 표본오차 ψ_i는 4장에서 설명한 방법으로 개별 테이블 자료로부터 추정한 것을 이용
- 최우추정법으로 추정시 비선형 최적화에는 $\hat{\sigma_\nu^2} = \exp(\tau)$로 변환해서 BFGS 공식에 의한 뉴톤 방법을 이용해서 추정한다(1.3절 참고). 이때 우선 Nelder-Mead법에 의해서 추정치를 구하고 그것을 초기치로서 다시 뉴톤법을 적용한다. 이런 일련의 계산은 R의 비선형 최적화 패키지 'optim'을 이용한다.
- 회귀 부분의 보조정보에 대해서는 직업안정업무통계에 의한 도도부현별 월별 유효구인배율을 이용한다.
- 추정한 $\theta(t)$에 대해서 Denton Method에 의한 벤치마킹을 적용해서 조정한다.
- 추정한 결과에 대해서 각종 검정통계를 이용한 분석 등을 수행하고 결과를 고찰한다.

2. 추정대상이 되는 소지역 선정

모든 도도부현에 대해서 분석하는 것은 어렵기 때문에 추정대상이 되는 소지역을 묶기로 한다. 시계열은 표본오차에 크게 영향을 미칠 것으로 생각된다. 또 표본오차는 표본수(표본조사구수)에 영향을 받는다. 이에 의해 2000~2002년(헤세12~헤세14)의 연평균 표본조사구수에 따라서 추정할 현을 선정한다.

표본조사구수를 도도부현별로 보면 상당히 치우친 분포를 하고 있음을 알 수 있다(원문의 그림 2-1, 2-2 참고). 이 가운데 표본수에 따라서 다음 8개 소지역을 선정한다.

- 조사구수가 작은 지역의 대표로서 3개 현을 선정
- 조사구수가 많은 지역의 대표로서 홋카이도, 도쿄도, 오사카부, 오키나와현을 선정
- 히스토그램에 의하면 표본수가 30, 40, 50 정도의 조사구수가 있는 지역이 많다. 이에 의해 이들 대표지역으로서 각각 3개 현을 선정

이들 각 현에 대해서 추정하고 결과를 고찰한다.

3. 완전실업율과 보조정보와의 상관관계

완전실업율과 상관관계가 높다고 생각되는 보조정보로 도도부현별로 월별 이용가능한 통계 가운데 중요한 통계로서 유효구인배율을 생각할 수 있다. 이것은 직업안정사업소에 의한 직업소개업무를 기초로 한 업무통계이고 다음의 식에 의해 정의한다.

$$\text{유효구인배율} = 100 \times (\text{월간 유효구인수} / \text{월간 유효구직자수}) \qquad (37)$$

월간 유효구직자수란 전월부터 반복된 유효구직자수에 당월 신규구직신청서 건수를 더한 것이다. 또 월간 유효구인수란 전월부터 반복된 유효구인수에 당월 신규구인수를 더한 것이다. 유효구인배율의 증가는 완전실업율 감소와 어느 정도 관계가 있을 것으로 생각된다. 이에 의해 유효구인배율은 완전실업율과 음의 상관이 있다고 볼 수 있다.

우선 소지역의 완전실업율 분석에 있어서 각 설명변수와 완전실업율과의 상관에 대해서 정리한다. 앞 절에서 선택한 8개의 소지역에 대해서 완전실업율과 보조정보의 관계를 시계열 그래프 및 산포도로 보면 다음의 사항을 알 수 있다(원문의 그림 3 참고).

- 시계열 그래프를 보면, 유효구인배율은 거의 완전실업율과 반대의 경향을 보이는 것처럼 보인다. 산포도에서도 음의 상관관계를 보인다. 단 결정계수는 전체적으로 낮고 높다고 해도 0.5를 넘는 곳은 없다.
- 1993(헤세5년)년 전후로 양자의 관계가 변한다고 생각되는 지역이 몇 개 있다. 많은 소지역에서는 1993년(헤세5년) 이전의 급격한 하락에 비해서 완전실업율의 상승은 완만하다. 그러나 그 후의 완전실업율의 대폭적인 상승과 비교해서 유효구인배율의 변동은 완만하다. 완전실업율은 빠른 변화를 반영하는 지표로 유효구인배율에 비해 경기에 대한 반응이 약간 늦는 것이 원인이 아닌가 생각해볼 수 있다. 이에 동반하는 상관관계의 변화도 상태공간모형이면 어느 정도 파악할 수 있을 것으로 생각된다.

- 홋카이도에서는 보조정보의 계절성이 상당히 높다. 계절노동자가 많음을 시사한다. 게다가 관측치와의 사이에 세로 방향으로 약간 어긋남이 있어 보인다. 홋카이도에서는 보조정보가 강한 계절성 때문에 회귀부분이 상당히 영향을 받을 것으로 예상된다.

4. 결과 고찰

(1) 시계열 그래프에 의한 비교

초기 상태가 안정적이라고 생각되는 헤세 5년 1월부터 헤세 15년 12월에 걸쳐 각 항의 추정치 및 오차율을 시계열 그래프에 나타내 보면 다음의 사실을 알 수 있다(원문의 그림 4-1~4-8 참고).

- 거의 모든 소지역에서 표본오차가 비교적 잘 제거된 것처럼 보인다
- 도토리현 등 3개 현에서는 벤치마킹에 의해 변경이 크게 나타났다. 사가현의 경우 1998(헤세10년)년의 이상치의 영향이 벤치마킹 후에도 여전히 남아 있음을 알 수 있다.
- 시가현의 이상치라고 생각되는 시점(1994, 1998년)에서는 모형에 의한 추정치의 오차율이 관측치의 오차율을 상회하고 있다. 이 외의 대부분 지역에서는 관측치보다도 모형의 오차율이 상당히 낮은 값을 보이고 있다. 게다가 모형에 의한 추정치의 오차율은 시간에 관계없이 거의 일정하고 추정치는 매우 안정되어 있는 것으로 보여진다.
- 대부분의 소지역에 대해서 회귀항과 트렌드항에 의해 전체적인 경향이 증가하고 있음을 알 수 있다. 단 홋카이도에 대해서는 회귀항 부분에 보조정보의 계절성이 상당히 나타나고 있고 게다가 그 변동이 관측치의 계절성과 미묘하게 어긋나 있다. 그러나 이 편의는 계절항에 의해 잘 반영되는 것 같다.
- 계절항의 움직임은 모든 지역에서 꽤 안정되어 있다.

(2) 각종 통계량에 의한 비교

추정한 값에 대한 각종 통계량을 보면 다음의 사항들을 알 수 있다(원문의 표 2 참고).

- 각종 검정통계량을 이용한 진단을 해보면 진단을 통과하지 못한 지역은 다음의 소지역들이다.
  - 정규성 검정을 통과하지 못한 소지역 : 2개
  - 분산 동일성 검정을 통과하지 못한 소지역: 2개

이 외의 거의 모든 소지역은 진단을 통과했고 그 결과도 좋았다.

- 초모수를 보면 어떤 소지역에서도 계절항 모수가 꽤 낮고 안정적인 계절성을 보이고 있다. 이것은 그래프에서도 살펴볼 수 있다. 이에 대해서 트렌드항의 모수는 높은 값이다. 완전실업율은 추정기간 중에는 상승기조이고 이 기울기를 보조정보로 설명할 수 없다고 보여진다. 사가현에서는 불규칙 변동항 모수

가 큰 값을 나타내고 있다. 이것은 이상치의 영향이라고 생각된다. 이상치가 회귀항, 트렌드항, 계절항으로 설명할 수 없고, 불규칙변동항에 들어갔기 때문이라고 생각할 수 있다. 이 외의 지역에서는 불규칙변동항 모수는 낮아지고 관측치 자료가 각 항에 의해서 잘 설명되고 있음을 알 수 있다.

Ⅴ 향후 과제

1. 지역의 값을 이용한 회귀모형

이후 분석은 유효구인배율을 보조정보로서 시계열회귀모형을 이용해서 도도부현별 월별 완전실업율을 추계했다. 이에 대해서 이웃 지역의 소지역 완전실업율을 보조정보로서 이용하는 방법을 생각할 수 있다. 이웃 소지역의 트렌드는 서로 연관되어 있다고 생각할 수 있기 때문이다. 이 경우에는 이웃이라는 정의가 문제이다. 우선 노동력조사 10개 지역 구분을 우선 생각할 수 있다.

- 이동평균과 전년동월값으로 이상치를 보정한다.
- 이상치를 결측치로 간주하고 1기선예측에 의해 보정한다.

미국 센서스국(1996)에서는 1기선예측오차의 대상치가 개별 자료에 의한 표분오차 추정치의 3배를 넘는 시점을 이상치로 간주한다(3σ 규칙).

2. 비선형, 비정규 모형에 의한 추정

이번에는 선형인 상태공간모형을 이용해서 분석했다. 또 모든 확률변수의 분포로 정규분포를 가정했다. 그러나 이들 가정이 만족되지 않을 경우도 많을 것이다. 이런 경우에 비선형모형 및 비정규모형을 이용하여 보다 정도 높은 추정을 할 수 있을 것이다.

비선형, 비정규모형을 이용한 분석에 대해서는 많은 연구가 있고 그 일부는 Durbin과 Koopman(2001)에 설명되어 있다. 이들은 추정에 대수우도를 분석하는 대신 재표집(Resampling) 등을 이용한 시뮬레이션 방법을 이용했다. 이 방법에 의해 비선형인 시스템을 직접 분석할 수 있다. 도 이상치 및 수준 이동(level shift) 처리도 칼만 필터로 자동적으로 행할 수 있기 때문에 보다 Robust한 추정을 할 수 있을 것으로 기대된다. 단 컴퓨터 용량에 대한 부담이나 계산 시간이 걸릴 것으로 예상된다. 재표집을 많이 하기 위해서 루프와 정렬도 많이 이용한다. 그러므로 이 경우에는 이번 연구처럼 R을 사용하지 않고, 포트란(FORTRAN) 컴파일러를 이용하는 것이 효과적일 것이다.