

연구노트

경지면적조사에서 분리비추정량의 조정*

Adjustments of the Separate Ratio Estimator in the Survey of Cultivated Land Area

김규성**

Kyu-Seong Kim

이 논문은 경지면적조사에서 분리비추정량을 활용하는 방법을 다루고 있다. 경지면적조사는 해마다 실시되는 일종의 계속조사로서 각 시·도에서 선정된 층화표본으로부터 분리비추정량을 이용하여 경지면적을 추정하고 있다. 분리비추정법을 계속조사인 경지면적조사에 적용하려 할 때 현실적으로 직면하는 문제로는, (i) 해가 지나면서 모집단 조사구수가 바뀌는 문제, (ii) 일부 층에서 보조변수가 0의 값을 취하는 문제, 그리고 (iii) 표본재설계로 인하여 층화 기준이 바뀌는 문제 등을 들 수 있다. 이 논문에서는 경지면적조사에서 발생하는 이와 같은 현실적인 문제들을 고찰한 후, 경지면적조사에 적용 가능한 조정된 분리비추정량을 제시하였다.

주제어: 계속조사, 비형 보정 추정량, 표본재설계, 합성추정량.

This paper deals with the use of the separate ratio estimator in the survey of cultivated land area. As a kind of repeated survey conducted every year, this survey estimates the total cultivated land area by using the separate ratio estimator. When using the separate ratio estimation in the survey, we may face with some practical problems such as (i) changing population size over time, (ii) auxiliary variables taken zero values in some strata and (iii) changing stratification criteria due to sample redesign. In this paper, after investigating practical problems occurred in the survey we presented some adjusted separate ratio estimators to tackle such problems.

Key words: composite estimator, ratio-type-adjusted estimator, repeated survey, sampling redesign.

* 이 논문은 2007년도 서울시립대학교 학술연구조성비에 의하여 연구되었음.

** 서울시립대학교 통계학과 교수 김규성.

E-mail: kskim@uos.ac.kr

I. 서론

국립농산물품질관리원이 발표한 바에 의하면 2006년 말 현재 우리나라의 경지면적은 180만 1천ha로서, 이 중에서 60.2%인 108만 4천ha가 논이고 나머지 39.8%인 71만 6천ha는 밭이다. 우리나라의 경지면적, 논 면적, 밭 면적은 해가 지나면서 조금씩 감소하는 추세에 있고, 밭 면적보다는 논 면적이 더 많이 감소하고 있다(〈표 1〉 참조).

잘 알려진 바와 같이 경지면적 통계는 표본조사를 통하여 얻어지므로 〈표 1〉에 나타난 수치는 표본 단위구에서 측정한 값을 근거로 하여 계산된 추정값이다. 그리고 각 추정값은 분리비추정법을 이용하여 계산되었다(2006년 작물 통계: 7~8). 작물통계 개요 부분에 경지면적 통계는 분리비추정법을 활용하여 계산하였다고 설명되어 있지만, 통상적인 분리비추정법을 곧바로 적용하여 경지면적 통계를 얻을 수는 없다. 경지면적조사는 계속조사이므로 분리비추정량을 경지면적조사에 적용하여 활용하기 위해서는 계속조사의 특성을 분리비추정량에 반영하여야 한다.

분리비추정법에 대해서는 이미 많은 연구결과가 발표되었다(예를 들면 Cochran 1977; 박홍래 2000; Wu 1985). 분리비추정법을 표본조사에 활용하기 위해서는 (i) 층화가 되어 있어야 하고, (ii) 각 층에서 보조변수로 활용하는 자료의 값이 모두 양수이어야 한다. 이와 같은 조건을 만족하고 보조변수가 주변수와 매우 밀접한 연관이 있을 때, 분리비추정법은 높은 효율을 기대할 수 있

〈표 1〉 우리나라의 경지면적

(단위 : 천ha)

구 분	2003년	2004년	2005년	2006년
논 면적	1,127 (61.0%)	1,115 (60.7%)	1,105 (60.6%)	1,084 (60.2%)
밭 면적	719 (39.0%)	721 (39.3%)	719 (39.4%)	716 (39.8%)
경지면적	1,846 (100.0%)	1,836 (100.0%)	1,824 (100.0%)	1,801 (100.0%)

출처: 2006년 경지면적통계(5쪽).

다. 경지면적조사와 같은 계속조사에서는 전회 혹은 전년도 자료를 보조변수로 이용할 수 있기 때문에 분리비추정법은 훌륭한 모수추정법 후보가 된다. 그러나 계속조사에는 시간 흐름에 따른 효과가 있기 때문에, 분리비추정법을 계속조사에 활용하기 위해서는 시간효과로 인하여 발생하는 문제들을 현실적으로 보완하는 보정된 분리비추정법을 찾아야 한다.

계속조사에서 분리비추정법을 적용할 때 고려해야 하는 첫 번째 문제는 조사구수에 관한 것이다. 시간이 지남에 따라 모집단 크기가 변하고 모집단 조사구수가 변하기 때문에 계속조사에서는 이를 감안하여야 한다. 두 번째로 고려해야 하는 점은 층화에 관한 것이다. 주기적으로 표본재설계를 하는 계속조사에서는 표본재설계 시 층화 기준이 바뀌기 때문에 전년도 데이터를 이용해야 하는 분리비추정량의 특성상 이에 대한 보정 방안을 가지고 있어야 한다. 마지막으로 고려해야 할 점은 분리비추정량에서는 양수인 보조정보를 사용해야 하는 점이다. 이 문제는 일회성 조사에서도 문제가 될 수 있는데, 왜냐하면 일부 층에서 보조변수가 0의 값을 갖는 경우가 종종 생기기 때문이다. 이 경우 층화를 다시 하거나 층을 병합하여 보조변수가 0의 값을 갖지 않도록 할 수 있고, 혹은 양수의 값을 갖는 다른 보조변수를 찾아서 분리비추정법을 사용할 수 있다. 그러나 다변량 층화를 하여야 하고 주요 보조변수의 활용이 제한적일 때에는 어쩔 수 없이 세 번째 조건을 충족하지 못하는 경우가 발생한다.

경지면적조사의 조사모집단은 5년 주기로 실시되는 경지총조사에서 조사된 전국의 경지로, 전국의 경지는 지적도, 토지대장 등을 기초로 하여 구획된 약 2ha 크기의 단위구로 구성된다. 2004년 경지총조사 당시 조사모집단 단위구수는 약 93만 개였다. 조사모집단은 시·도별로 층화되며, 층 내에서 지정된 추출률에 의하여 단위구가 단순확률표집된다. 현지조사를 통하여 확정된 표본은 추출률(가중치)을 부여받은 후 자료 입력 시스템에 저장되어 향후 5년간 표본의 역할을 하게 된다. 경지면적 총계는 전년도 데이터를 활용하는 분리비추정값으로 계산된다.

경지면적조사는 고정표본을 사용하고, 표본설계 당시의 추출률과 가중치가 자료입력 시스템에 기억되기 때문에 설계 이후 변동하는 모집단 단위구수는 사후적으로 추정과정에 반영해 주어야 한다. 또한 표본재설계는 이전의 경

지총조사와는 완전히 다른 새로운 경지총조사 자료를 바탕으로 하므로 통계 결과의 시계열 유지를 위해서는 시·군별로 층별 보정을 해야 할 필요성이 있다. 또한 경지면적조사에서도 각 시·도별로 층화를 하고 전년도 층별 논 면적과 밭 면적을 보조변수로 활용하는데 일부 층에서 논 면적 혹은 밭 면적이 0이 되는 경우가 발생하였다. 지역적인 문제 때문에 원천적으로 논 면적 혹은 밭 면적이 언제나 양수의 값을 갖도록 층화를 다시 하기는 어렵다. 예를 들어 서울이나 광역시의 일부 층에서는 논이나 밭 면적이 0일 가능성이 높고 제주도의 일부 층에서는 논 면적이 0일 가능성이 매우 높다. 이와 같이 사전에 보조변수가 0의 값을 갖는 문제를 해결하기 어려운 경우 사후적으로 문제를 해결하는 것이 현실적이다.

이 논문에서는 분리비추정량을 경지면적조사에 활용할 때 부딪치는 현실적인 문제를 고찰하고 이를 해결하는 방안을 제시할 것이다. 이 논문은 다음과 같이 구성되어 있다. 2절에서는 경지면적조사에서 분리비추정량을 사용할 때 발생할 수 있는 문제점을 고찰한 뒤 각각에 대한 해결방안을 모색한다. 2.1절에서는 문제점을 고찰하고, 2.2절에서는 모집단 단위구수가 변하는 경우에 활용 가능한 분리비추정법을 제시하며, 2.3절에서는 보조변수가 0의 값을 갖는 경우를 다룬다. 3절에서는 표본재설계로 인하여 문제가 발생하는 경우를 다룬다. 3.1절에서는 표본재설계 효과를 알아보고 3.2절에서는 비형 보정 추정법을 알아본다. 마지막으로 4절에서는 논문의 내용을 간략하게 요약한다.

II. 분리비추정량의 보정

1. 문제 제기

어떤 표본조사에서 층화 표본을 선정하고 각 표본으로부터 조사변수(y)와 보조변수(x)의 값을 관측했다고 하자. 그리고 각 층에서 두 변수의 표본평균이 모두 양수의 값을 갖고 있고, 또한 각 층에서 보조변수의 총계값을 알고 있다고 하자. 그러면 통상적인 분리비추정량의 형태는 다음과 같다.

$$\hat{Y} = \sum_h \left(\frac{\bar{y}_h}{\bar{x}_h} \right) X_h \quad (1)$$

여기에서 \bar{y}_h (\bar{x}_h)는 조사변수 y (보조변수 x)의 h 층 표본평균이고 X_h 는 h 층의 보조변수 총계이다(Cochran 1977: 164).

분리비추정법의 효율은 추정에 활용되는 보조변수에 크게 의존하기 때문에 추정의 효율을 높이기 위해서는 보조변수를 잘 선택해야 한다. 경지면적조사는 해마다 실시되는 조사이므로 전년도 면적을 보조정보로 활용하면 추정의 효율을 크게 높일 수 있다. 전년도 자료를 보조변수로 활용할 때 분리비추정량의 형태는 아래와 같이 된다.

$$\hat{Y}_t = \sum_h \left(\frac{\bar{y}_{t,h}}{\bar{y}_{t-1,h}} \right) \hat{Y}_{t-1,h} \quad (2)$$

여기에서 y 를 경지면적(혹은 논 면적, 밭 면적)이라고 하면 $\bar{y}_{t,h}$ ($\bar{y}_{t-1,h}$)는 t 시점($t-1$ 시점)의 h 층의 경지면적 표본 평균이고 $\hat{Y}_{t-1,h}$ 는 $(t-1)$ 시점에서 h 층의 경지면적 총계 추정값이다. 단, 첫째의 전년도 추정값으로는 경지총조사 결과를 이용한다. 즉,

$$\hat{Y}_1 = \sum_h \left(\frac{\bar{y}_{1,h}}{\bar{y}_{0,h}} \right) Y_{0,h}$$

여기에서 $Y_{0,h}$ 는 h 층의 경지총조사 면적이고, $\bar{y}_{1,h}$ 는 첫째의 층 표본평균, $\bar{y}_{0,h}$ 는 첫째의 h 층 표본 단위구에 해당하는 경지총조사 자료의 평균이다.

전년도 경지면적을 보조변수로 활용하기 위해서는 전년도 층 평균과 층 총계값이 양수인지 여부를 검토하여야 한다.

- (i) $(t-1)$ 시점에서 각 층의 면적 표본평균($\bar{y}_{t-1,h}$)은 양수인가?
- (ii) $(t-1)$ 시점에서 각 층의 면적 총계 추정값($\hat{Y}_{t-1,h}$)은 양수인가?

또한 시간 변화에 따른 모집단 단위구수의 변동에 대한 고려가 추가로 필요

하다. 식(2)에서 제시된 분리비추정량은 기준년도의 모집단 추출틀을 기준으로 만들어졌지만, 기준년도 이후에도 사용되어야 하므로 모집단 변동을 분리비추정량이 반영하여야 한다.

표본 선정 이후에 바뀐 모집단 변동을 이미 뽑힌 표본이 반영하기는 현실적으로 어렵다. 해가 바뀌면서 새로 추가된 지역의 면적 변동 상황을 기존의 표본만으로 파악할 수는 없고, 또한 기존의 추출틀에는 있더라도 표본에 선정되지 않은 단위구가 해가 바뀌면서 사라지는 경우 표본만으로는 그러한 상황을 파악할 수 없다. 따라서 전년도 경지면적 추정값을 분리비추정량에 보조변수로 사용하기 위해서는 이와 같은 경지면적조사의 특성을 반영하여야 한다.

2. 모집단 단위구수 변동

대부분의 표본설계에서는 기준년도의 모집단 단위구수를 기준으로 표본을 선정하고 추출율과 가중치를 계산한 뒤 추정량을 구한다. 경지면적조사에서는 해마다 전국 곳곳에서 택지 개발, 공공시설 개발 등으로 논밭이 감소하기도 하고 반대로 개간, 간척 등으로 일부 농경지가 새로 생겨나기도 한다. 따라서 표본설계 당시 모집단에 포함되지 않았던 일부 지역이 농경지가 되어 조사모집단에 새로 포함되거나 혹은 설계 당시에는 농경지였으나 그 뒤에 도로 건설 등으로 비 농경지가 되어 조사모집단에서 제외되는 경우 사후적으로 모집단 단위구수 및 표본수를 변경하여 추정에 반영하여야 한다.

그러나 모든 변동에 대하여 매번 찾아 다니면서 변동 사항을 파악할 수는 없으므로 단위구 크기의 평균값을 기준하여 이 값보다 큰 면적의 변동이 발생할 때에만 조사원이 방문하는 등의 규칙을 정할 수 있다. 현행 경지면적조사에서는 경지면적 변동이 2ha 이상 일어나는 지역이 생기면 그 지역을 방문하여 증감을 조사한 후 표본으로 추정한 경지면적에 더하거나 빼서 경지면적을 추정한다(2006년 경지면적통계: 3). 이러한 상황을 식(2)의 분리비추정량에 반영하면 다음과 같은 보정된 분리비추정량을 얻을 수 있다.

$$\hat{Y}_t = \sum_h \left(\frac{\bar{y}_{t,h}}{y_{t-1,h}} \right) \hat{Y}_{t-1,h} + C_{t,h} - D_{t,h} \quad (3)$$

여기서 $C_{t,h}$ 는 t 시점 h 층의 추출틀 이외의 지역에서 경지가 2ha 이상 새로 생긴 지역의 면적이고, $D_{t,h}$ 는 t 시점 h 층의 추출틀에서 표본이 아닌 지역 중 경지가 2ha 이상 감소된 지역의 면적이다. 실제로 2006년에 우리나라에서 모집단의 변동으로 인하여 증가된 경지면적은 4,547ha이고 감소된 면적은 28,116ha라고 보고되었다(〈표 2〉 참조).

식(3)에 제시된 보정된 분리비추정량은 이론적인 최적 추정량이라기보다는 현실적인 대안이다. 따라서 개선의 여지가 여전히 남아 있다. 예를 들면 단위구의 소멸과 생성은 임의로 변동한다고 볼 수 있으므로 보정항 $C_{t,h}$ 와 $D_{t,h}$ 는 확률변수로 다루는 것이 타당할 것이다. 이러한 이론적인 문제들은 향후 연구로 남긴다.

3. 보조변수의 값이 0인 경우

경지면적 추정 시 보조변수가 양수일 조건(즉, $\bar{y}_{t-1,h} > 0$ 과 $\hat{Y}_{t-1,h} > 0$)은 대부분 만족된다. 층 전체에서 경지가 소멸되거나 층 내 표본 단위구의 경지가 모두 사라지는 경우는 극히 예외적이기 때문이다. 따라서 경지면적 추정에서 분리비추정량은 안정적으로 이용할 수 있다. 그러나 경지면적과는 달리 논

〈표 2〉 증감사유별 경지면적

(단위 : ha)

구분		2003년	2004년	2005년	2006년
증 가 사 유	개간	2,349	7,249	1,008	4,159
	간척	-	345	411	16
	기타	897	648	312	372
	합계	3,246	8,242	1,731	4,547
감 소 사 유	건물건축	7,927	6,695	5,209	9,819
	공공시설	4,823	5,742	3,517	6,315
	유휴지	4,406	4,209	3,068	4,462
	기타	2,718	1,956	1,532	7,520
	합계	19,874	18,602	13,326	28,116

면적과 밭 면적 추정에서는 분리비추정량 사용에 주의를 해야 한다. 왜냐하면 일부 층에서 논 면적이거나 밭 면적이 0인 경우가 종종 발생하기 때문이다. 그 원인은 논과 밭의 지역 특성과, 이러한 특성을 층화에 활용한 층화 기준에서 찾을 수 있다. 우리나라 경지는 전국적으로 논이 과반수이상을 차지하고 있으나, 지역 특성에 따라 일부 지역에서는 특정 밭작물이 경지면적의 대부분을 차지하는 경우도 있다. 따라서 경지면적 추정과 작물재배면적 추정의 효율을 높이는 방법으로 논 비율과 밭작물 종류가 층화에 사용되었다.

현행 경지면적조사에서 활용하는 표본설계의 층화 기준을 보면, 논 비율과 주요 작물을 층화 변수로 하여 만일 단위구의 논 비율이 100%이면 층1, 논 비율이 50%이상이고 100% 미만이면 층2, 그리고 논 비율이 50% 미만인 단위구는 주로 재배되는 작물을 중심으로 층이 구성되어 있다. 이때 사용된 재배 작물은 시설작물, 벼, 맥류, 두류, 서류, 채소, 조미채소, 특용작물 과수이다 (2006년 작물통계 7쪽, <표 3> 참조).

<표 3> 표본설계 층화 기준(충남, 전북, 전남, 경북의 경우)

층 번호	층 이름	층 특성
1	논벼 층	논벼가 100%인 단위구
2	논벼 층	논벼가 50% 이상, 100% 미만인 단위구
3	맥류 층	맥류가 주인 단위구
4	시설작물 층	시설작물이 주인 단위구
5	식량작물 층	충남, 전북, 경북, 경남: 두류, 잡곡, 서류가 주인 단위구 전남: 두류, 서류가 주인 단위구
6	채소류 층	채소류(엽채, 과채, 근채, 기타 일반채소)가 주인 단위구
7	조미채소 층	충남: 고추, 마늘, 기타조미채소가 주인 단위구 전북, 전남, 경북, 경남: 고추, 마늘, 양파, 기타조미채소가 주인 단위구
8	특약용 층	특약용이 주인 단위구
9	과수 층	과수가 주인 단위구
10	기타작물 층	기타작물이 주인 단위구
11	기타 층	나머지 단위구

현행 총화 기준에 의하면 총1에 속하는 단위구의 밭 면적이 0이다. 표본설계는 경지총조사의 논, 밭 자료를 바탕으로 하기 때문에 설계 당시에는 총화 기준이 잘 지켜진다. 그러나 시간이 흐르면서 논과 밭의 비율이 바뀌고 작물 재배 패턴이 바뀌기 때문에 총화 효과는 점점 감소하게 된다. 그렇다고 총화의 효과를 유지하기 위하여 해마다 모집단을 정비하고 총을 보정하기에는 총화 효과에 비하여 너무 많은 비용이 필요하므로 현실적인 대안이 되기 어렵다. 다만, 큰 변동이 있는 지역에 한해서는 부분적인 보정이 필요할 것이다.

이와 같이 시간이 흐름에 따라 총화 효과가 감소하는 현상은 자연스럽다 하더라도, 시간이 지남에 따라 총의 성격이 바뀌는 현상은 경지면적조사에 분리비추정량을 적용하는 데 치명적일 수 있다. 특히 논 면적과 밭 면적을 추정할 때 문제가 발생한다. 예를 들어 경지총조사에서 논 비율이 100%인 단위구를 고려하자. 현행 총화 기준에 의하면 이 단위구는 총1에 속하게 된다. 그리고 밭 면적은 0이다. 그런데 다음해에 총1의 일부 표본 단위구에서 밭 면적이 측정되었다고 하자. 표본 단위구의 일부가 논에서 밭으로 변경되는 경우 총1에서 밭 면적이 측정될 수도 있다. 그런데 분리비추정법에 의하면 전년도에 총1에서 밭 면적이 0이면 $\hat{Y}_{t-1,1} = 0$, $\bar{y}_{t-1,1} = 0$ 이므로 금년도에는 표본 단위구에서 밭 면적이 양수라 하더라도 분리비추정법을 직접 적용할 수 없다.

이에 대한 대안으로는 선형추정량을 고려할 수 있을 것이다. 선형추정량은 과거의 보조 정보를 필요로 하지 않기 때문에 전년도 상황과 무관하게 당해연도의 총 총계를 추정할 수 있는 장점이 있다. 따라서 전년도 면적이 양수이면 분리비추정량을 사용하고 그렇지 않으면 선형추정량을 사용하는 것이 이러한 문제를 해결하는 하나의 방법이 될 것이다. 보정된 분리비추정량은 다음과 같이 표현될 수 있다.

$$\begin{aligned} \hat{Y}_{t,h} &= \left(\frac{\bar{y}_{t,h}}{\bar{y}_{t-1,h}} \right) \hat{Y}_{t-1,h} + C_{t,h} - D_{t,h}, \quad \text{만일 } \hat{Y}_{t-1,h} > 0 \\ &= N_h \bar{y}_{t,h} + C_{t,h} - D_{t,h}, \quad \text{만일 } \hat{Y}_{t-1,h} = 0 \end{aligned} \quad (4)$$

아직까지 식(4)에 제시된 보정된 분리비추정량은 완전하지 않다. 왜냐하면 흔치는 않지만 식(4)에서 총 총계 추정치가 음수가 나올 수 있기 때문이다. 예를 들어 전년도에 어떤 층의 밭 면적이 0이어서 당해연도에 선형추정량을 사용하여 밭 면적을 추정하였는데, 표본으로 계산한 경지면적 총계값보다 그 층에서 감소된 경지면적이 더 크게 조사되면 총 총계 추정치는 음수가 된다. 모집단에서는 어떤 층에서든 면적이 음수가 될 수는 없으나 식(4)와 같은 추정량에서는 추정치가 음수가 되는 경우가 발생할 수 있는 것이다. 따라서 이러한 경우가 발생하면 강제적으로 면적을 0으로 치환하여 사용하는 것이 현실적이다. 즉, 식(4)에 다음을 추가한다.

$$\hat{Y}_{t,h} \equiv 0 \text{ 만일 } \hat{Y}_{t,h} < 0$$

III. 표본재설계 후 분리비추정량의 보정

1. 표본재설계 효과

계속조사에서 당해연도의 모총계 추정을 위해서는 부분교체표본을 사용하는 것이 동일표본을 사용하는 것보다 추정의 효율 면에서 우수하다(Cochran 1977: 345). 그럼에도 불구하고 동일표본이 선호되는 이유는 다음과 같다. 첫째, 해가 바뀔에 따라 생기는 변화량 때문이다. 계속조사에서 해가 바뀌면서 발생하는 변화량을 추정하는 데는 동일표본을 사용하는 것이 부분교체표본을 사용하는 것보다 유리하다. 두 번째는 현실적인 이유인데, 동일표본을 사용하는 것이 부분교체표본을 사용하는 것보다 비용이 저렴하고 조사하기가 쉽기 때문이다. 조사변수가 적고 조사내용이 간단한 조사에서는 어떤 표본을 사용하든 큰 문제가 되지 않으나, 조사내용이 많고 복잡한 조사에서는 부분교체표본을 사용하면 조사관리 비용이 증가하므로 상대적으로 조사비용이 적은 동일표본이 선호되는 것이다. 우리나라의 경지총조사는 10년에 한 번씩 이루어지고 있고, 이 때 모집단 단위구를 새로 확정하면서 행정구역의 변경과 같은 모집단 변경사항을 반영하게 된다. 그리고 경지총조사 자료를 기초로 한 경지면적조사 표본재설계에서는 모집단 환경에 맞게 층화를 개선하고 목표 표준오차

에 맞추어 표본수를 조정한 후 표본을 선정하게 된다. 이렇게 선정된 표본은 다음 설계 때까지 반복하여 사용된다.

동일표본을 사용하여 식(4)의 분리비추정량을 사용하면 전년도 결과가 금년도 추정에 영향을 주므로 전년도 통계와 매우 다른 통계가 생산될 가능성은 크지 않다. 그러나 표본재설계 이후 첫 해에는 기존의 표본과는 무관한 독립표본을 사용하여 모수를 추정해야 하므로 동일표본을 사용할 때와 같은 시계열상의 이점은 없다. 만일 재설계 후 추정치가 전년도 추정치와 큰 차이를 보이면, 그 이유는 모집단 변동 효과와 표본 재설계 효과 때문이지만 통계 이용자에게는 표본 개편으로 인한 재설계 효과가 더 크게 부각될 수 있고, 결과적으로 통계의 신뢰도를 떨어뜨리는 원인이 될 수 있다. 따라서 추정 과정에서 이러한 재설계 효과를 제거하는 방법을 찾아보아야 한다.

2. 비형 보장 추정량

문제를 분명히 하기 위하여 경지면적조사에서 시점에 따른 추정량의 종류를 고려해 보자. 현 시점을 t 시점이라고 할 때 현 시점까지는 기존의 표본을 이용하여 모총계를 추정한다고 하자. 그리고 다음 시점인 $(t+1)$ 시점부터는 표본재설계를 하여 뽑은 신규표본을 활용하여 총계를 추정한다고 하자. 그리고 경지총조사는 $(t-1)$ 시점에 실시되었고 표본재설계는 $(t-1)$ 시점에서 조사한 경지총조사 자료를 바탕으로 한다고 하자. 또한 본격적인 재설계표본을 활용하기에 앞서 재설계표본으로 시험조사를 실시한다고 하면 시험조사는 t 시점에 실시될 것이다. 이 같은 상황을 정리하면 <표 4>와 같이 된다. 이 표에서 알 수 있듯이 t 시점에서는 기존 추정량 \hat{Y}_t 과 신규 추정량 \hat{Y}_{t_0} 가 얻어진다.

<표 4>에 나타난 6개의 추정량과 경지총조사 면적값 X_{t-1} 을 계산 알고리즘의 측면에서 살펴보자. 먼저 기존 설계를 이용하여 구한 \hat{Y}_{t-2} , \hat{Y}_{t-1} , \hat{Y}_t 는 연차적으로 구하는 것이므로 계산상의 문제는 없다. 또한 경지총조사에서 구한 X_{t-1} 도 집계를 하는 것이므로 경지총조사 자료만으로 독립적으로 구할 수 있다. 다음으로 표본재설계에서 얻어야 하는 추정치를 생각해 보자. 시

〈표 4〉 표본재설계 시점을 전후한 추정량의 종류

구분	시점				
	t-2	t-1	t	t+1	t+2
본조사(기존 설계)	\hat{Y}_{t-2}	\hat{Y}_{t-1}	\hat{Y}_t		
경지총조사		X_{t-1}			
시험조사(재설계)			\hat{Y}_{t0}		
본조사(재설계)				\hat{Y}_{t+1}	\hat{Y}_{t+2}

험조사에서 구해야 하는 \hat{Y}_{t0} 는 전년도 총 총계와 총 표본평균이 없으므로 전년도 총 총계와 총 표본평균을 대신할 수 있는 활용 가능한 다른 자료를 찾아야 한다. 이 상황에 부합하는 자료로 경지총조사 자료를 떠올리는 것은 매우 자연스러운 일이다. 따라서 \hat{Y}_{t0} 의 전년도 자료로 경지총조사 자료를 이용하고, \hat{Y}_{t+1} 의 전년도 자료로는 시험조사 자료를 이용하면 된다. 이와 같이 하면 표본재설계 후에도 모총계 추정을 체계적으로 계속할 수 있으므로 계산 알고리즘 측면에서 보면 아무런 문제가 없다.

그런데 이렇게 모총계 추정을 한다면 기존의 설계와 재설계는 아무런 연결 고리가 없음을 발견할 수 있다. 즉 (\hat{Y}_{t-2} , \hat{Y}_{t-1} , \hat{Y}_t)는 기존설계의 동일표본을 사용하는 반면, (\hat{Y}_{t0} , \hat{Y}_{t+1} , \hat{Y}_{t+2})는 재설계의 동일표본을 사용한다. 그리고 두 설계를 연결하는 시점에서 만들어지는 X_{t-1} 는

\hat{Y}_{t0} 에 보조정보를 제공하므로 재설계와 연관이 있으나 기존설계와는 무관하다. 따라서 표본재설계 후 처음 발표되는 추정값 \hat{Y}_{t+1} 과 기존 설계의 마지막 추정값 \hat{Y}_t 는 서로 독립적으로 만들어진다.

이제 두 추정치가 적지 않은 차이를 보이는 경우를 가정하자. 각각의 추정량의 관점에서 보면 두 추정량은 독립적으로 만들어지고 또한 모집단 총계에 대한 일치추정량이므로 두 추정값이 비록 차이를 보이나 하더라도 각각의 표

준오차가 크지 않으면 사용가능한 수치일 것이다. 그리고 실제로는 추정치의 오차보다는 모집단 변동이 심해서 두 추정치가 큰 차이를 보일 수도 있다. 그렇다 하더라도 앞서 언급한 대로 통계이용자는 두 추정치의 차이를 근거로 통계조사의 신뢰도를 문제 삼을 수 있다. 두 추정치가 시점이 다르므로 현실적으로 문제가 제기되는 시점은 (t+1)시점에 \hat{Y}_{t+1} 를 발표하기 직전이 대부분이다. t시점에 이미 \hat{Y}_t 는 발표가 되었고, (t+1)시점에 \hat{Y}_{t+1} 를 발표해야 하는데 두 추정치가 큰 차이를 보이면 통계작성기관은 발표 직전에 고민에 빠지게 되는 것이다. 두 추정량을 동등하게 비교한다면 \hat{Y}_{t+1} 의 신뢰도가 \hat{Y}_t 의 신뢰도보다 더 높을 것이다. 왜냐하면 \hat{Y}_{t+1} 가 기초로 하는 모집단 프레임이 \hat{Y}_t 가 기초로 하는 모집단 프레임보다 모집단 상황을 더 잘 반영하기 때문이다. 그러나 현실적으로는 \hat{Y}_t 는 이미 발표가 되었으므로 두 추정량을 양자택일해야 하는 상황은 발생하지 않는다. 대신 (t+1)시점의 추정량 \hat{Y}_{t+1} 을 보정하는 쪽에서 해결방안을 찾는 것이 현실적이다.

한 가지 대안은 합성추정량(composite estimator)을 이용하는 것이다 (예, Cochran 12.13절). 만일 \hat{Y}_{t+1} '이 기존설계에서 계산된 총계 추정량이라고 한다면 (t+1)시점에서 총계에 대한 합성추정량은 다음과 같이 표현될 수 있다.

$$\hat{Y}_{t+1}'' = w \hat{Y}_{t+1} + (1-w) \hat{Y}_{t+1}' \quad (5)$$

여기에서 w 는 0과 1 사이의 가중값으로 추정량 \hat{Y}_{t+1}'' 의 분산을 최소화하는 값을 찾아야 한다. 그리고 기존설계에서는 (t+1)시점의 추정치를 제공하지 않으므로 t시점까지 얻은 추정값을 기초로 하여 (t+1)시점의 모총계 예측값을 구하여 사용해야 한다.

또 다른 방법으로는 표본재설계 각 층에서 보조변수로 이용된 시험조사의 층 총계 추정량 $\hat{Y}_{t0,h}$ 를 보정하는 방법을 생각할 수 있다. 각 층별로 다음과 같이 t시점의 층 총계를 기존설계의 값을 활용하여 비형(ratio type) 보정을

한다.

$$\hat{Y}_{t,h}' = \hat{Y}_{t0,h} \left(\frac{\hat{Y}_t}{\hat{Y}_{t0}} \right) \quad (6)$$

이 층별 추정량을 더하면 기존설계의 추정량 \hat{Y}_t 가 되므로 보조변수로 활용하는 층별 총계에 시험조사값과 기존설계의 값을 동시에 이용하는 효과가 있다. 이제 (t+1)시점의 층 총계 추정량은 다음과 같이 보정된다.

$$\hat{Y}_{t+1,h}' = \left(\frac{\bar{y}_{t+1,h}}{y_{t,h}} \right) \hat{Y}_{t,h}' + C_{t+1,h} - D_{t+1,h} \quad (7)$$

(t+1)시점의 총계는 모든 층에서 위의 층별 총계 추정량을 더하여 얻는다.

식(5)의 합성추정량과 식(7)의 비형 보정 추정량은 모두 일치추정량이므로 (t+1)시점의 모총계 추정에 사용 가능한 추정량이다. 합성추정량을 사용하기 위해서는 가중치 w 와 예측치 \hat{Y}_{t+1}' 를 추가로 구해야 하는 반면, 비형 보정 추정량에서는 층별 총계 추정량 $\hat{Y}_{t,h}'$ 만 추가로 구하면 되므로 계산 과정만 비교하면 비형 보정 추정량을 구하는 것이 더 쉽다. 그렇지만 두 추정량을 엄밀히 비교하기 위해서는 두 추정량의 표준오차를 구하는 등의 이론적인 접근이나, 모의실험 등을 통한 경험적인 접근을 통하여 발표하는 추정값에 대한 확신을 높일 필요가 있다.

경지면적조사에서도 경지면적과 논 면적, 밭 면적을 추정할 때에 두 추정량을 사용하면 표본재설계로 인한 추정치의 차이를 어느 정도 방지할 수 있을 것이다. 그러나 여전히 문제는 남는다. 어느 추정량을 사용해야 하는가? 두 추정량의 차이는 보정을 하는 단계에서 차이를 보인다. 합성추정량은 발표 단위에 보정을 하는 것이고, 비형 보정 추정량은 층별로 보정을 하는 것이다. 따라서 발표 단위에서의 보정으로 충분하면 두 추정량은 모두 사용 가능한 반면, 층별 보정이 필요하면 합성추정량은 층별 보정을 하지 않기 때문에 사용하기 어렵고 비형 보정 추정량만이 사용 가능하다. 그런데 경지면적조사에서는 층별 보정이 요구된다. 왜냐하면 경지면적 표본은 작물재배면적조사에도 활용되는 다목적 표본이고, 작물재배면적은 각 층에서 경지면적을 보조변수로 하여

분리비추정을 하기 때문에 각 층별로 경지면적이 필요하다. 따라서 경지면적 조사에서는 비형 보정 추정량을 사용할 수밖에 없다.

IV. 결론

정부기관이나 민간기관에서 작성하는 통계 중 중요통계는 해마다 반복하는 계속조사인 경우가 많다. 이러한 계속조사에서는 시계열 추세가 중요한 요소이기 때문에 동일표본을 다년간 사용하게 된다. 그리고 일정 기간에 한 번씩 표본재설계를 하여 표본을 교체함으로써 동일표본의 대표성이 감소하는 것을 방지한다.

경지면적조사도 계속조사의 일종으로 분리비추정법을 사용하여 면적을 추정한다. 이 논문에서는 경지면적조사에서 분리비추정법을 사용할 때 발생하는 문제점을 알아보고, 각각에 대한 대안을 제시하였다. 이 논문에서 제시한 대안들은 이론적인 고찰과 함께 실제 데이터를 이용한 경험적인 검토를 통하여 경지면적조사에 활용될 수 있을 것이다. 향후 이 논문에서 제시한 방법이 경지면적조사의 품질을 향상시키는 데 일조할 수 있기를 기대한다.

참고문헌

- 국립농산물품질관리원. 2007. 《2006년 경지면적통계》.
- 국립농산물품질관리원. 2007. 《2006년 작물통계》.
- 박홍래. 2000. 《통계조사론》 개정판. 영지문화사.
- Cochran. W. G. 1977. *Sampling Techniques*. John Wiley & Sons.
- Wu, C.F.J. 1985. "Variance Estimation for the Combined Ratio and Combined Regression Estimators." *Journal of the Royal Statistical Society B47*: 147-154.

[접수 2008/5/2, 1차수정 2008/7/9, 2차수정 2008/7/14,
게재확정 2008/7/15]