제2장

사업체 조사에서의 자동 오류위치포착의 적용방안

이의규 · 심규호

제1절 서 론

통계자료 에디팅(statistical data editing)은 자료 수집 및 처리 단계에서 오류를 찾아내고 이를 수정하는 과정을 말한다. 통계청에서는 이러한 에디팅 과정을 내용검토(또는 줄여서 내검)라 부르며 이를 수행한다. 에디팅은 작업 방법에 따라 수작업 에디팅과 에디팅의 과정을 컴퓨터가 대신하는 자동 에디팅(automatic editing)으로 구분하곤 하는데, 에디팅은 컴퓨터의 발전과 더불어 점차적으로 사람의 힘에 덜 의존하는 쪽으로 발전하고 있다.

에디팅(editing)은 흔히 임퓨테이션(imputation)과 혼돈하기 쉽다. 주로 응답된 자료의 오류를 찾아 수정하는 것을 에디팅이라 말하는 반면에 임퓨테이션은 응답하지 않은 항목 값을 확률적인 방법을 통해 대체하는 것을 의미한다. 그러나 응답한 값이 무시될 정도로 의미가 없어 무응답으로 여겨지는 경우, 에디팅은 임퓨테이션 과정을 수반하게 되어 이 둘을 정확히 구분하기 어렵다. 또한 임퓨테이션을 수행하기 위해서는 이상치에 대한 검토가 선행되어야 하며 임퓨테이션 후에는 다시 조사담당자에 의해 설정된 내검규칙(edits)을 만족하도록 해야 하는 등 에디팅과 임퓨테이션은 서로 밀접하게 연관되어 있다. 일반적으로 통계자료 에디팅은 임퓨테이션을 포함하는 자료처리 과정의 포괄적 의미로 해석할 수 있다.

현실의 통계조사에서는 무응답은 흔히 일어난다. 또한 응답한 자료라도 부정확한 응답으로 종종 오류를 갖게 된다. 따라서 분석에 이용되기 전에 모든 결측과 오류는 검출될 필요가 있고 적절한 방법으로 수정될 필요가 있다. 이때 결측치에 대해 재조사하거나, 오류를 수정하기 위해 재접촉 할 수 있다. 그러나 비용과 시간의 제약으로 재조사가 항상 가능할 수는 없으며 게다가 응답자는 조사에 협조하지 않을 수 있다.

한편, 사업체대상 조사는 금액에 관한 정보가 특히 중요하다. 금액에 관한 정보는 민감한 정보 중 하나로 정확한 답변을 얻어내기 어렵다. 따라서 응답한 자료라 하더라도 어떤 특정 사유가 없이 비합리적인 값이 나타난다면 표식 후 적절한 전략에 의거한 수정이 불가피할 경우가 있다. 대부분의 조사는 총계 수준에서 추정값을 제공하는 목적을 갖기 때문에 모든 응답이 세밀하게 에디팅된 자료를 필요로 하지 않는다. 오히려 합리적인 전략에 의해 수정된 자료는 총계 추정치를 바람직하게 개선할 수 있을 것이다.

무엇보다도 통계자료 사용자는 통계자료가 완전하고 일관적일 것을 기대한다. 그러나 하나의 레코드(사업체)에서 항목 간 수리적 관계(내검규칙)로 볼 때 응답한 값이 무시될 정도로 의미가 없다면 상당한 혼란에 빠질 수 있다. 이는 통계작성기관의 신뢰의 문제이기도 하다. 따라서 납득하기 어려운 레코드는 수정되어야 할 필요성이 있으며, 이때 모든 내검규칙을 만족하기 위해 그 실패 레코드에서 어떤 변수(들)를 수정해야 할지를 결정하는 것이 오류위치포착(error localization)의 문제이다.

1. 연구 배경 및 현황

모든 오류자료나 무응답자료를 재접촉을 통해 완벽하게 고치고 채우려고 하는 것이 과연 바람직한 것인가? 고친다고 과연 그 고친 값이 얼마나 정확하다고 보증할 수 있는가? 재접촉으로 인한 응답자의 자료는 얼마나 믿을 수 있는가? 국외 연구에 의하면 그리 긍정적이지 않다. 경제자료의 에디팅으로 만들어진 가장 큰 변경의 10~15%가 전체 변경의 90%를 차지한다고 한다. 또한 가장 큰 변경의 5~10%는 최종 전체 추정치의 1%내의 추정치를 초래하며 재접촉의 단지 20~30%가 변경된 값을

가져온다고 하고, 품질 향상은 미미하거나 없거나 심지어 떨어지며 많은 종류의 심각한 규칙적인(systematic) 오류들은 에디팅에 의해 식별될 수 없다는 연구결과가 보고된 바 있다(Granquist & Kovar 1997, Werking & Clayton 1988, Christianson & Tortora 1995, Linacre & Trewin 1989).

캐나다, 미국, 네덜란드 등 통계 선진국에서는 에디팅 과정을 가능한 자동화하고 있으며 이를 통해 효율적인 에디팅을 수행하고 있다. 이는 시간과 비용을 줄이고 응답자의 부담을 더는 데 그 목적이 있으나 한편으로는 에디팅의 모든 과정이 투명하고 에디팅 수행 후에도 원래의 자료로 다시 복원할 수 있기 때문이기도 하다. 물론 외국의 경우는 재질문이나 재방문을 가능한 금하고 있기에 조사 자료에서 오류나 무응답을 포함하는 경우가 많다. 이러한 조사환경에서 자동 에디팅은 그 역할과 중요성이 매우 크다.

통계청에서는 각 조사마다 입력·내검 프로그램을 통해 조사표를 입력하게 되는데 담당자가 내검규칙을 설정하고 그 규칙에 어긋나는 경우 해당하는 에러코드를 자동으로 나타나게 한다. 이후 자료 내용을 에러코드에 따라 조사표 확인이나 재접촉을 통해 수정 또는 오류의 사유를 기재하여 입력하게 된다. 따라서 오류를 찾는 것은 컴퓨터의 도움으로 자동화된 방식으로 진행되나 수정은 전적으로 조사원 또는 내검요원 등 사람의 힘에 의해 수행되고 있다.

조사원이 현장에서 오류수정을 수행하는 것은 수집 초기단계의 정보에 근거해서 수정을 할 수 있다는 장점이 있다. 그러나 적지 않은 내검 작업량과 촉박한 일정은 재접촉의 질을 보증할 수 없을 가능성이 있으며 응답자가 불응하거나 여러 상황으로 인해 확인이 불가한 경우가 있을 수 있다. 또한 조사원 간의 편차가 발생할 수 있고 주관이 개입될 수 있어 왜곡될 수 있다. 더욱이 재접촉 또는 재방문으로 인한 응답자의 부담은 가중될 것이며 조사원의 조사부담도 역시 클 것이다.

오류 또는 무응답이 발생한 응답자에 대해 재접촉/재조사가 힘들거나 이를 통해서도 해결되지 않을 때에는 최종적으로 내검규칙을 만족시키지 못한 항목 값은 수정되어야 한다. 정보의 손실을 막기 위해 가능한 한 최소의 항목을 수정함으로써 모든 내검규칙을 만족시키는 전략이 필요하다. 통계청의 입력·내검 프로그램은 수작업의 용이한 수정을 위해

위배된 규칙에 대응되는 오류코드를 알려주는 것이지만 자동 에디팅 시스템은 최소한의 항목값 수정을 통해 내검규칙을 만족하도록 자동으로 오류위치를 포착하여 수정한다는 점에서 차별된다.

2. 자동 에디팅의 필요성

재접촉/재방문의 어려움 발생 등 조사환경이 악화될 경우에는 자동 에디팅 시스템의 중요성은 커진다. 또한 행정자료를 활용하거나 행정자료를 기반으로 하는 조사에서 그 역할은 더욱 크다. 이는 주어진 자료에서만 오류를 찾아내고 주어진 정보를 통해서만 자료를 수정하여야 하므로 자동 에디팅은 필수적이다.

자동 에디팅에 대한 우려 중 하나는 각 응답자마다 처해진 상황이 다른데 어떻게 똑같은 잣대로 자동 수정을 하겠느냐는 것이다. 물론 오류자료/무응답자료는 할 수 있다면 재접촉/재방문을 통해 수정하는 것이 자료의 품질을 좋게 할 것이다. 그러나 앞에서도 언급하였듯이 재접촉/재방문에 드는 시간, 비용, 조사부담 문제는 제쳐두고 실질적으로 재접촉/재방문으로 인한 변경량은 얼마나 되고 있으며 그 결과가 얼마나 만족스럽고 전체결과에 얼마나 영향력이 있는 것인가에 대한 대답은 그리 명쾌하지 못하다.

오류자료/무응답자료를 조사원이 재접촉/재방문을 통해서도 해결하지 못하는 경우 자신의 경험적 판단으로 수정하거나 응답을 대체할 수도 있다. 이러한 최악의 경우에 체계적이고 일관된 전략을 가지고 자동화하는 것이 더 효율적이고 위험이 작을 수 있다. 어떻게 보면 자동 에디팅은 조사원이 경험적으로 수정하는 방식을 자동화하는 것일 수 있으며 조사원간의 편차가 없이 일관된 원칙에 의해 수행되는 방식으로 볼 수 있다.

자동 에디팅은 사람이 세밀하게 작업하는 부분을 인정하지 않는 것이 아니라 컴퓨터를 이용하여 가능한 수고를 덜자는 것이다. 이는 옷감이 손상되기 쉬운 세탁물은 손으로 세탁하고 그리 옷감에 신경을 쓰지 않아도 될 세탁물은 자동으로 세탁하는 것에 비유할 수 있다. 또한 더러움이 심한 와이셔츠의 깃이나 소매를 사람이 세탁한 후 이를 자동세탁

기로 처리하여 상당한 손세탁의 수고를 덜어 주는 것과 같다.

3. 연구목적, 범위 및 한계

자동내검기법 연구(이의규와 심규호, 2007)에서는 자료의 자동 오류위치포착 및 수정의 근거가 되는 Fellegi-Holt(F-H) 방법의 원리와 절차를 소개하고 자동내검기법의 적용 가능성을 검토한 바 있다. 본 연구는 자동내검기법 연구의 후속연구로서 외국의 자동 에디팅 시스템과 F-H 기법의 이행 알고리즘을 살펴보고 실제 사업체 조사 자료에 프로그램을 작성·적용하여 자동내검기법이 실질적으로 어떻게 이행될 수 있는지를 검증하고자 한다. 궁극적으로는 현재의 조사/내검 현황을 고려하여 자동내검 시스템에 관한 연구의 기본방향을 도출하고자 한다.

본 연구에서는 내검규칙에 어긋난 레코드가 모든 내검규칙에 만족하기 위해 어떤 항목을 바꾸어야 하는지에 대한 문제를 주로 다룬다. 즉 구체적인 어떤 값으로 바꾸어야 하는 것은 논외로 한다. 다시 말해 본 연구에서는 수정되어야할 위치를 포착하는 문제에 국한하기로 한다. 본 보고서의 작성된 프로그램 역시 자동 수정 또는 임퓨테이션을 위해 수정되어야 할 위치를 찾아주는 데 초점을 두고 있다. F-H 자동 수정의 알고리즘을 통해 실제 자료를 가지고 자동 수정을 위해 수정되어야 할 변수의 위치를 포착하는 것을 목적으로 한다. 어떤 방법으로 대체해야 할지는 각 조사의 성격에 따라 세밀하게 다루어져야 할 것이다.

현재는 조사표 입력 시 입력·내검 프로그램의 운영과 더불어 조사원들의 수작업으로 수정된 자료가 본청으로 집계되어 최종단계에서는 자동 에디팅이 수행될 여지가 없다는 한계를 가진다. 그러나 최종자료에는 에러(반드시 수정해야만 하는 에러보다 유연한 에러)로 나타났으나 내용검토 후 허용된 자료가 존재하므로 이러한 자료를 이용하여 유용성을 검증하고자 한다. 자동 오류위치포착기법은 자료입력·내검 단계에서 수정될 오류의 위치를 참고할 수 있는 효과를 기대할 수 있다.

제2절 해외의 자동 에디팅 시스템

통계자료의 에디팅을 자동화하기 위해서는 크게 두 가지 과정이 요구된다. 첫 단계는 오류위치를 포착(error localization)하는 단계이다. 이 단계에서 자료의 오류들이 검출되는데, 통계자료 분석자나 담당자의 지식에 근거하여 설정된 내검규칙을 사용하여 레코드가 규칙에 위배되는지 아닌지를 결정한다. 만약 레코드가 오류를 갖는다면 이 단계에서 레코드에서 수정되어야 할 오류 변수를 식별한다. 두 번째 단계는 대체(imputation)단계이다. 오류자료는 더 정확한 자료로 대체되어지고 결측자료 역시 대체된다.

오류위치포착은 최소개의 변수의 수정을 통해 모든 내검규칙을 만족하도록 수정될 변수를 결정하는 전략인 펠레지-홀트(Fellegi & Holt, 1976)의 이론을 근거로 한다. 이러한 F-H 기법은 수학적인 최적화 문제로 범주형 자료나 연속형 자료에 모두 적용되는 기법이다.

이 절에서는 외국의 대표적인 자동 에디팅 관련 소프트웨어를 조사하고 시스템의 주요 내용을 살펴보고자 한다. 여기서는 사업체 대상 조사에서 나타나는 연속형 자료에 대한 자동 에디팅 시스템을 살펴본다. 캐나다 외의 많은 국가들이 사업체 자료에 F-H 기법을 적용한 자동 에디팅 시스템을 활용하고 있다. 캐나다 통계청의 GEIS(Generalized Edit and Imputation System), 미국 센서스 국의 SPEER(Structured Programs for Economic Editing and Referrals)와 네덜란드 통계청의 CherryPi가 대표적이다. 이 중 캐나다 통계청의 GEIS는 최근 9개의 SAS 프로시저(procedure)로 구성된 Banff로 발전되었다. 캐나다 통계청은 자동 에디팅 시스템에 있어서 다른 국가에 비해 오랜 경험을 가지고 있어 주목할 만하다. 이 시스템에 대한 더 자세한 내용은 Kozak(2005)의 논문을 참조하기 바란다.

1. Banff

Banff는 정량적인(quantitative) 조사 자료의 자동 에디팅 및 대체를 위해 캐나다 통계청에서 최근에 개발한 자료처리 시스템이다. 1980년 중

반에 프로젝트가 착수되어 1980년대 후반부터 캐나다 통계청에서 사용되었던 일반화된 자료처리시스템인 GEIS(Generalized Edit and Imputation System)로부터 발전된 것이 Banff이다. GEIS는 Fellegi-Holt 이론을 기초로 하고 있으며, 내검규칙 분석(edit analysis)과 오류위치포착(error localization), 대체(imputation)의 세 부분으로 이루어져 있다. Banff 역시 GEIS와 같은 구조로 이루어지고 있으나 Banff는 <표 2-1>에 요약된 바와 같이 독립적으로 또는 필요에 따라 조합하여 운용할 수 있다. 9개의 SAS 프로시저(procedure)로 구성되어 사용자가 사용하기 더 편리한 것으로 평가받고 있다.

〈표 2-1〉 Banff의 SAS 프로시저(procedure)

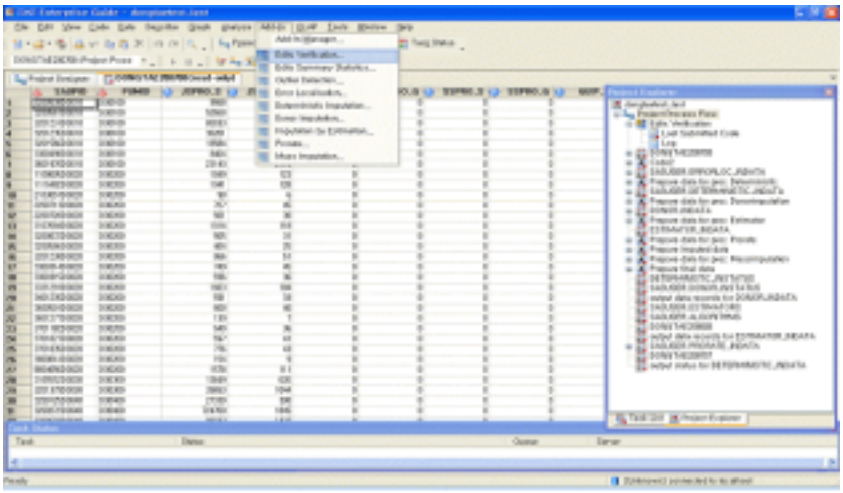
프로시저(Procedures)	주요 처리내용
Proc Verifyedits	내검규칙 설정 및 체크, 유도된 추가규칙 생성
Proc Editstats	각 규칙의 합격/실패 레코드 수 등 5개표 산출
Proc Outlier	대체될 이상치, 대체에 사용불가 이상치 판단
Proc Errorloc	실패 레코드의 수정해야 할 필드 결정
Proc Deterministic	설정된 내검규칙에 의해 결정되는 값 대체
Proc Donorimputation	가장 유사한 레코드를 찾아 대체
Proc Estimator	다양한 대체추정법을 이용 대체
Proc Prorate	소계가 더해져 총계가 됨을 보증
Proc Massimputation	대체될 필드가 알려져 있고 동일할 때 대체

한편, GEIS는 데이터베이스로서 오라클(Oracle)을 사용하는 반면 Banff는 SAS에 기초한다. GEIS의 모듈들은 서로 연결되어 있는 반면 Banff의 프로시저는 서로 독립적이라는 점이 다르다. 또한 GEIS와 Banff 모두 메인프레임과 UNIX 환경에서 사용 가능하나 Banff만이 PC의 Windows 환경에서 사용이 가능하다(<표 2-2> 참조).

〈표 2-2〉 GEIS와 Banff의 비교

	GEIS	Banff
환경	UNIX	Unix, PC Window
데이터베이스	Oracle	SAS
운용방법	각 모듈이 서로 연결	각 프로시저가 독립적

Banff 사용자는 자료처리과정에서 프로시저의 일부 또는 모두를, 어떤 순서에 상관없이 적용할 수 있다. 중요한 것은 하나의 프로시저의 출력자료가 다른 프로시저의 입력 자료로 사용될 수 있다는 것이다. 또한 SAS에서 'BY' 문을 효율적으로 활용할 수 있다. 참고로 Banff의 메인 윈도우가 [그림 2-1]에 주어져 있다. SAS Enterprise에 각 모듈이 add-in 되어 사용자가 편리하게 사용할 수 있도록 구성되어 있다.



[그림 2-1] Banff 메인 윈도우

2. SPEER

SPEER(Structured Programs for Economic Editing and Referrals) 시스템은 연속형 경제자료의 에디팅을 위한 목적으로 설계되었는데 Brian Greenburg(1984)에 의해 첫 번째 버전이 개발되었다. 이 시스템은 자동 에디팅에 대한 방법론으로 Fellegi-Holt 이론을 적용하였다. SPEER는 두 개의 프로그램 gb3.for와 spr3.for로 이루어져 있으며 FORTRAN으로 작성되었다. 첫 번째 프로그램은 명시적 내검규칙으로부터 유도된 내재적 내검규칙을 생성하고 또한 모든 내검규칙의 논리적 일치성을 체크한다. 내검규칙은 두 변수의 비에 대한 하한과 상한 값으로 주어진다. 두 번째

프로그램은 오류위치포착(error localization)을 수행하고 대체(imputation)를 수행하는데 결과물은 요약 통계와 각 수정된 레코드의 상세 정보를 주는 파일로 구성된다. [그림 2-2]는 내검규칙과 각 내검실패 레코드의 대체 결과를 보여준다.

[그림 2-2]에서 1번 레코드는 APR2 변수와 QPR3 변수의 비가 일정한 범위 안에 있어야 하는 규칙을 위배한 경우이다. 이 경우 수정해야 할 변수로 QPR3가 선택되고 13에서 5.714로 수정되었음을 나타낸다. 이어서 5번 레코드는 EMP1/QPR3와 APR2/QPR3의 비에 대한 내검규칙을 만족하지 못한 경우이다. 이때 수정해야 할 변수로 QPR3 변수 하나가 선택되고 4에서 11.429로 수정되어 모든 내검규칙을 만족함을 보여주고 있다.

```
 Record #   1

Failed edits:
        1.8964540 <  APR2  /  QPR3  <       5.9863030

Deleted fields:    3. QPR3

Imputation range for  QPR3 :  Lo =          3.3410    Up =         10.5460
   QPR3  imputed using  QPR3 / EMP1  ratio

    Fields     Revised    Reported       Lower       Upper
    ------     -------    --------       -----       -----
     EMP1         1.000       1.000        .425       1.422
     APR2        20.000      20.000      14.062      34.207
     QPR3         5.714      13.000       3.341      10.546
     FBR4         3.000       3.000        .406       5.435


 Record #   5

Failed edits:
         .0602807 <  EMP1  /  QPR3  <        .4257010
        1.8964540 <  APR2  /  QPR3  <       5.9863030

Deleted fields:    3. QPR3

Imputation range for  QPR3 :  Lo =          6.6819    Up =         21.0920
   QPR3  imputed using  QPR3 / EMP1  ratio

    Fields     Revised    Reported       Lower       Upper
    ------     -------    --------       -----       -----
     EMP1         2.000       2.000        .850       2.845
     APR2        40.000      40.000      28.124      68.415
     QPR3        11.429       4.000       6.682      21.092
     FBR4         6.000       6.000        .812      10.870
```

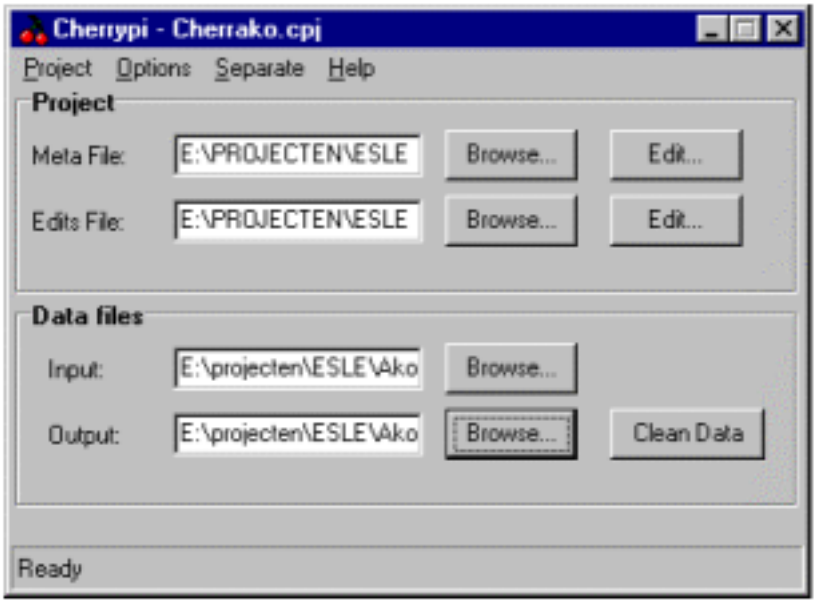
[그림 2-2] 주요 결과물 중 내검규칙을 위반한 레코드의 예

3. CherryPi

네덜란드 노동비용조사(Dutch Labour Cost Survey)는 네덜란드 통계청에서 4년마다 수행되는 경제조사이다. 1992년 기준 조사의 에디팅은 주로 수작업으로 이루어졌으나 1996년 기준 조사는 CherryPi라고 불리는 특별한 소프트웨어를 사용하여 선택적 에디팅과 자동 에디팅의 도입으로 에디팅이 이루어졌다. 이 소프트웨어는 경제자료의 수작업 에디팅의 많은 부분을 자동 에디팅으로 대체하기 위해 네덜란드 통계청에 의해 개발되었다. 이렇게 함으로써 비용을 절감하면서도 품질을 유지할 수 있었다.

통계 소프트웨어 패키지 CherryPi는 윈도우(Windows) 환경에서 볼랜드(Borland) 델파이(Delphi) 3.0으로 작성되었다. 이 소프트웨어는 경제자료를 자동으로 에디팅하고 수정하는 일반화된 시스템이다. CherryPi는 선형 내검규칙과 비(ratio) 내검규칙 모두를 다룰 수 있다. CherryPi의 오류검출방법은 Fellegi-Holt 방법론에 기초한다. 즉 모든 내검규칙을 만족할 수 있도록 각 레코드에서 대체될 최소개의 필드를 찾는 것이다. 캐나다 통계청의 GEIS와 마찬가지로 레코드마다 최소개의 필드를 결정하기 위해 Chernikova의 수정 버전을 사용한다.

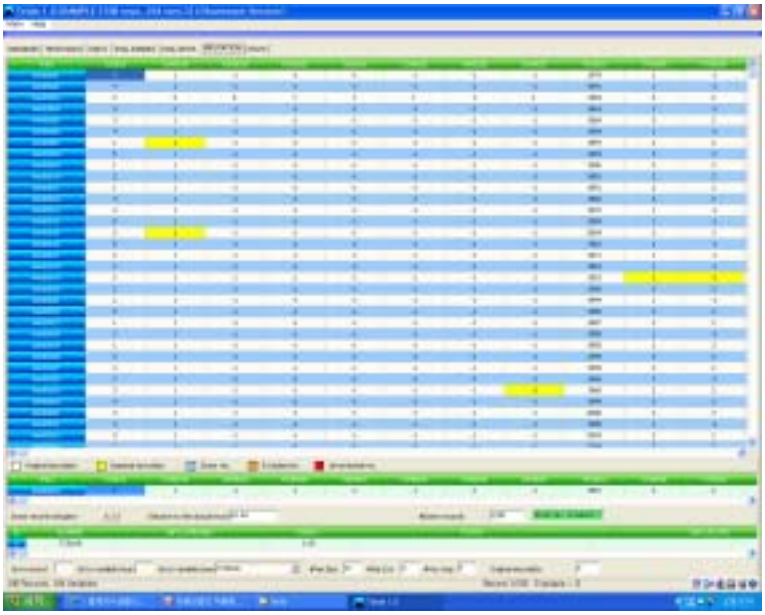
CherryPi의 개선된 버전과 그래픽 매크로 에디팅(graphical macroediting)과 핫덱 임퓨테이션(hotdeck imputation)같은 방법들은 SLICE(Statistical Localisation, Imputation and Correction of Errors)라 불리는 소프트웨어에서 이행된다. CherryPi 프로그램은 오류위치포착(localisation), 선택(selection), 임퓨테이션(imputation), 수정(modification)의 4가지 부분으로 구성된다. 참고로 CherryPi 소프트웨어의 메인 윈도우는 아래 [그림 2-3]과 같다(Nordholt and De Waal, 1999).



[그림 2-3] CherryPi 소프트웨어의 메인 윈도우

4. TEIDE

TEIDE(스페인어로 Techniques for the Editing-and-Imputation of Statistical Data를 의미함)는 스페인의 La Laguna 대학의 Salazar-Gonzalez 교수에 의해 C++로 작성되었다(Salazar-Gonzalez, 2006). 이 프로그램은 MS Windows 환경에서 작동되며 MS Access 데이터베이스로 입출력된다. 비상업적인 공개 에디팅 소프트웨어라고는 하지만 100개까지의 레코드를 쓸 수 있을 정도의 시험 소프트웨어만을 공개하고 있다. 오류위치포착은 F-H 기법을 통해 결정되며 임퓨테이션은 donor 임퓨테이션과 회귀 대체가 선택적으로 사용된다. [그림 2-4]는 TEIDE 소프트웨어를 가지고 예제자료를 실행한 결과이다.



[그림 2-4] TEIDE 소프트웨어의 윈도우

제3절 오류위치포착을 위한 주요 알고리즘

비 F-H 에디팅 시스템은 내검규칙을 모아서 관리한다. 그러나 이 규칙들이 서로 모순될 수 있다는 것을 알아내기는 힘들며 일일이 오류레코드마다 어떤 변수를 수정하여야 할지를 결정하는 것은 그리 쉽지 않다. 레코드 내의 에디팅을 다루는 마이크로 에디팅의 이론적 기반은 펠레지-홀트(Fellegi 와 Holt, 1976)에 의한다. 이들은 모든 내검규칙을 만족하도록 어떤 필드의 값이 수정되어야 할지를 이론적으로 증명하였다.

F-H기법을 이행하기 위해서는 Chernikova의 알고리즘(Chernikova, 1964, 1965)이 핵심이다. 그러나 이 방법의 문제는 많은 계산을 요구한다는 것이다. 그동안 계산 속도를 높이기 위한 수치적 방법을 고안하려는 노력이 있었다. Sande(1997), Garfinkel et al(1986), Schiopu-Kratina와 Kovar(1989), Filion과 Schiopu-Kratina(1993), Winkler(1998), Chen(1998), Quere(2000), Winkler와 Chen(2002), De Waal(2003), De Waal과

Quere(2003) 등에서 찾아볼 수 있다. 이에 따라 캐나다 통계청의 Banff와 미국 센서스국의 SPEER 그리고 네덜란드 통계청의 CherryPi가 개발되었으며 여러 조사에 적용되고 있다(Winkler, 2006).

오류위치포착을 위한 방법에는 크게 3가지의 방법이 있다. 첫 번째 방법은 가장 느린 방법으로 branch-and-bound와 같은 integer programming 방법이다. 두 번째는 변형 Chernikova의 알고리즘(Rubin 1975, Schopiu-Kratina and Kovar 1989, Filion and Schopiu-Kratina 1993)이다. 이 방법이 캐나다 통계청의 GEIS와 네덜란드 통계청의 CherryPi에 사용된다. 이 방법은 다소 느리기 때문에 네덜란드 통계청은 변형 Fourier-Motzkin 소거법을 이용한다. 사업체조사의 자동 에디팅을 위해 필요한 알고리즘에 대한 자세한 정보는 De Waal(2003)의 논문을 참조하기 바란다.

1. 내재적 내검규칙의 생성과 오류위치포착

조사 자료에 오류가 있는지를 판단하기 위해서는 어떤 내검규칙이 있어야 한다. 이는 흔히 조사 담당자의 경험에 의해 설정되는데 자료의 형태가 범주형 자료(categorical data)인 경우에는 논리적 내검규칙(logical edits)에 의해 그 오류 여부를 판단하게 되며, 연속형 자료인 경우에는 산술적인 내검규칙에 의해 판단된다.

오류위치포착(error localization)은 레코드가 모든 내검규칙을 만족하도록 대체되어질 변수를 식별하는 문제이다. 결측치는 당연히 대체가 필요하다. 그러나 레코드의 모든 항목에 값이 기입은 되어 있지만 내검규칙에 어긋나면 몇 개의 변수는 부정확한 것으로 식별되어야 한다. 그런 후 적절한 값으로 대체되어야 한다. 그런데 어떤 값이 부정확하고 대체되어야 하는지의 결정은 그리 단순하지 않다. 대부분의 조사는 대체해야 할 변수를 결정하기 위해 각 설문지를 다시 검토할 여유가 없다.

이때 어떤 변수를 수정해야 할지를 결정하는 자동화 전략이 필요한데 그것이 F-H에 의한 전략이다. 주어진 정보를 최대한 보존하면서 모든 내검규칙을 만족하게 하는 최소개의 수정할 변수를 찾아내자는 것이다. 물론 레코드가 모든 내검규칙을 만족한다면 오류위치포착은 그 레코드에 필요하지 않다. 그러나 적어도 하나의 내검규칙이 만족되지 않

는 경우에는 수정이 요구되는 값을 식별하기 위한 오류위치포착이 필요하다.

오류위치포착에 대한 이해를 돕기 위해 예를 들어 본다. 자산이나 비용과 같은 양적 자료가 나타나게 되는 사업체 조사의 경우에는 계량산술적 내검규칙(quantitative arithmetic edits)이 주어지게 된다. 다음과 같은 명시적 내검규칙(explicit edits)이 주어졌다고 하자(Waal과 Coutinho, 2005).

$$E1: \quad X_1 - X_2 + X_3 + X_4 \geq 0$$
$$E2: \; -X_1 + 2X_2 - 3X_3 \geq 0$$

이제 하나의 레코드가 (3, 4, 6, 1)로 코딩되었다고 하자. 위 내검규칙에 따르면

$$E1: \quad 3 - 4 + 6 + 1 = 6$$
$$E2: \; -3 + 2(4) - 3(6) = -13$$

이다.

이 레코드는 두 번째 규칙을 위반한 레코드이다. 문제는 레코드가 모든 규칙을 만족하는 조건에서 어떤 필드를 수정하여야 최대한 정보를 유지할 수 있는가이다. 일반적으로 경제자료는 비음(nonnegative)이므로 이 레코드에서 X_1만을 바꾸어서는 성립이 안 된다. 역시 X_2만을 바꾸어서는 모든 내검규칙을 만족할 수 없다. 그러나 X_3를 1로 바꾼다면 모두 만족한다. 물론 두 개 이상의 변수를 모두 바꾸어서 성립이 가능할 수 있으나 최대한 자료를 보존한다는 원칙에서 X_3 하나만을 바꾸는 것이 합리적이라는 것이다.

이는 주어진 내검규칙으로부터 각 변수의 소거를 통해 다음과 같은 식으로부터 도출된 결과이다.

$$E3: \quad X_2 - 2X_3 + X_4 \geq 0$$
$$E4: \quad X_1 - X_3 + 2X_4 \geq 0$$
$$E5: \quad 2X_1 - X_2 + 3X_4 \geq 0$$

위의 E3, E4와 E5를 내재적 내검규칙(implicit edits)이라 한다. 다시 레코드의 각 값을 각 규칙에 대입하면

$$
\begin{aligned}
E3: &\quad 4 - 2(6) + 1 = -9 \\
E4: &\quad 3 - 6 + 2(1) = -1 \\
E5: &\quad 2(3) - 4 + 3(1) = 5
\end{aligned}
$$

으로 주어진 레코드는 E3와 E4의 내검규칙을 만족하지 못하고 있음을 알 수 있다. 따라서 전체 위배된 내검규칙 E2, E3, E4에 각 포함된 변수 행렬 <표 2-3>을 얻을 수 있다. 그런데 <표 2-3>에서 X_3는 위배된 모든 내검규칙에 포함되어 있음을 볼 수 있다. 즉 명시된 내검규칙으로부터는 어떤 변수를 바꾸어 주어야 할지가 명확하지 않으나 이처럼 추가된 내검규칙을 이용하면 자료의 오류위치를 효율적으로 판단할 수 있다.

〈표 2-3〉 위배된 내검규칙 행렬

	X1	X2	X3	X4
E2	1	1	1	
E3		1	1	1
E4	1		1	1

더 나아가 X_3값을 미지수로 놓고 나머지 주어진 값을 조건식에 대입하여 풀면 $0 \leqq X_3 \leqq 5/3$ 일 때 모든 규칙을 만족하게 된다. 즉, $X_3 = 1$이 가능한 대체값이 될 수 있다.

2. 선형계획법과 Chernikova 알고리즘

내검규칙은 개별변수 또는 변수들 간에 논리적인 관계와 제약조건으로 정의된다. 특히 사업체대상 조사의 수치적인 내검규칙은 변수들의 부등식 또는 등식으로써 표현될 수 있다. 흔히 이때의 모든 변수는 비음(nonnegative)으로 제한된다. 따라서 n개의 변수를 포함하는 내검규칙군

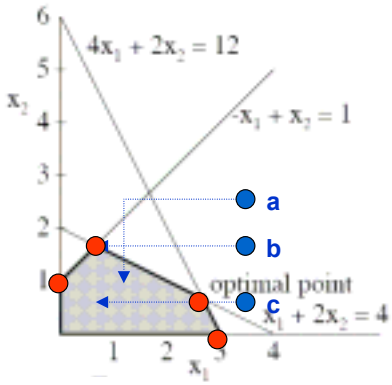
은 *n*-차원 공간에서 가능해 영역(feasible region) 또는 채택영역(acceptance region)을 정의한다. 레코드가 모든 내검규칙을 만족하게 되면 가능해 영역(feasible region)안에 위치하게 되고 그렇지 않으면 이 영역 밖에 위치하게 된다. 다시 말하자면 레코드가 모든 내검규칙을 만족한다면 선형에디팅시스템에 대해 합격이고 그렇지 않으면 실패한 레코드이다.

선형에디팅시스템을 만족하는 모든 점은 시스템의 가능해 영역(feasible area)을 구성한다. 내검규칙을 만족한 레코드는 가능해이다. 따라서 선형 에디팅 시스템은 가능해 영역에 의해 완전히 기술된다. 이와 같이 선형 에디팅 문제는 일반적으로 선형계획법의 솔루션과 관계가 있으며 이는 가능해 영역(feasible area)의 모든 극값(extremal points)을 찾음으로써 풀 수 있다. Chernikova의 알고리즘(Chernikova, 1964, 1965)은 비음 변수를 갖는 선형 시스템의 모든 극값을 찾는 데 사용된다.

예를 들어 다음과 같이 두 개의 비음(nonnegative) 변수와 세 개의 선형제약식이 있다고 가정하자. 이때 $x_1 + x_2$ 합을 최대화하는 선형계획법의 솔루션을 구한다 하자(www.math.ucla.edu/~tom/LP.pdf).

$$\begin{aligned} x_1 + 2x_2 &\leq 4 \\ 4x_1 + 2x_2 &\leq 12 \\ -x_1 + x_2 &\leq 1 \\ x_1 \geq 0,\ x_2 &\geq 0 \end{aligned}$$

이 경우는 변수가 2개이므로 도식화가 가능하다. [그림 2-5]의 음영처리된 영역이 가능해 영역이다. 그리고 (8/3, 2/3)이 최적값(optimal point)이 된다. 이때 [그림 2-5]의 a, b, c 점은 가능해 영역 밖에 있다. 그런데 점 c는 x_2를 변화시켜서는 가능해 영역으로 들어오지 않는다. x_1을 변화시켜야 한다. 점 a는 x_1과 x_2 모두 변화시켜야 하며 점 b는 단지 x_2만을 변화시키면 가능해 영역 안으로 들어갈 수 있다. 즉 주어진 각 점마다 가능해 영역으로 진입하기 위해서는 어떤 변수가 바뀌어야 할지를 결정할 필요가 있는데 이때 다각형의 꼭지값(vertex)이 중요한 역할을 하게 된다. 이들 꼭지값은 Chernikova 방법의 알고리즘에 의해 결정된다(Banff Support Team, 2007).



[그림 2-5] 선형제약식과 가능해 영역

이상과 같이 F-H 기반의 오류레코드의 수정위치 포착 문제는 선형계획법(linear programming)의 해를 구하는 문제와 유관하다. 일반적으로 수치형 조사 자료의 내검은 다음과 같은 선형 내검규칙 군에 의해 정의된다.

$$E_i : a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{\in}x_n \quad \leq \quad b_i , \qquad i = 1,2,\ldots,m$$

$$x_j \quad \geq \quad 0 , \qquad j = 1,2,\ldots,n$$

위 식을 다시 행렬로 바꾸어 표현하면 다음과 같다.

$$\mathbf{A}\mathbf{x} \leq \mathbf{b}$$

$$\mathbf{x} \geq \mathbf{0}$$

여기서 **A**는 $m \times n$ 내검규칙 계수행렬이고 **b**은 $m \times 1$ 벡터이다. 그리고 $\mathbf{x}=(x_1,x_2,\ldots,x_n)'$는 $n \times 1$ 레코드 벡터이다. 따라서 오류위치포착(error localization) 문제해결에 대한 선형계획접근법은 다음과 같이 표현할 수 있다.

$$\begin{aligned} &\max \ \boldsymbol{d}'\mathbf{x} \\ \text{subject to } &\mathbf{A}\mathbf{x} \leq \mathbf{b}, \\ &\mathbf{x} \geq \mathbf{0}, \\ &|\mathbf{x}|^{+} \leq \eta, \end{aligned}$$

여기서 **x**와 **d**는 $n \times 1$벡터이다. **A**는 $m \times n$ 행렬이고 **b**는 $m \times 1$ 벡터이며 $|\mathbf{x}|+$ 는 **x**의 카디날리티(cardinality)이다 그리고 $\mathbf{x}=(x_1,x_2,....,x_n)'$는 레코드 벡터이다. η는 m보다 작은 양의 정수이다. 선형계획법은 $|\mathbf{x}|^{+} \leq \eta$를 만족하는 가능해 영역 $G=\{\mathbf{x}|\mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$의 극값을 직접적으로 준다. 그리고 최적의 극값이 결정된다. 이러한 접근방법이 GEIS와 CherryPi에서 이행된다. 다양한 내검규칙 분석은 벡터 **d**의 값과 목적함수를 적절하게 선택하고 변화시킴으로써 수행되어진다(Weng 2002, Giles 1988).

제4절 자동 오류위치포착의 적용

이 절에서는 2003년 기준 산업총조사 중 4인 이하의 광업·제조업 통계조사 자료에 자동 오류위치포착기법을 응용한 프로그램을 작성·적용하여 그 결과를 분석하고자 한다. 4인 이하의 사업체 자료는 소규모 사업체의 경제자료로서 영향력이 비교적 작다. 또한 응답정보가 중대사업체보다 부정확할 것으로 판단되어 이 자료를 택하였다.

사업체대상 조사의 자동내검기법연구(이의규와 심규호, 2007)의 보고서에서 제시한 종사자 4인 이하 광업제조업 통계조사의 4번 항목(연간생산비 소계), 5번 항목(연간 제품 출하액 합계), 6번 항목(연간 임가공수입액 합계), 7번 항목(유형자산 연말잔액)의 항목 간 수량적 연관규칙을 가지고 자동내검기법을 적용한다.

적용 자료는 이미 내검과 임퓨테이션이 완료된 자료이므로 실제적으로 규칙에 어긋난 자료는 많지 않다. 그러나 이용 자료의 제약으로, 먼저 주어진 내검규칙으로 다시 한 번 에러 체크를 한 후 규칙에서 벗어난 자료가 있는지 살펴보고, 이 자료에 대하여 자동기법을 적용하여 그 의미를 살피고자 한다.

1. 광업·제조업 통계조사의 개요

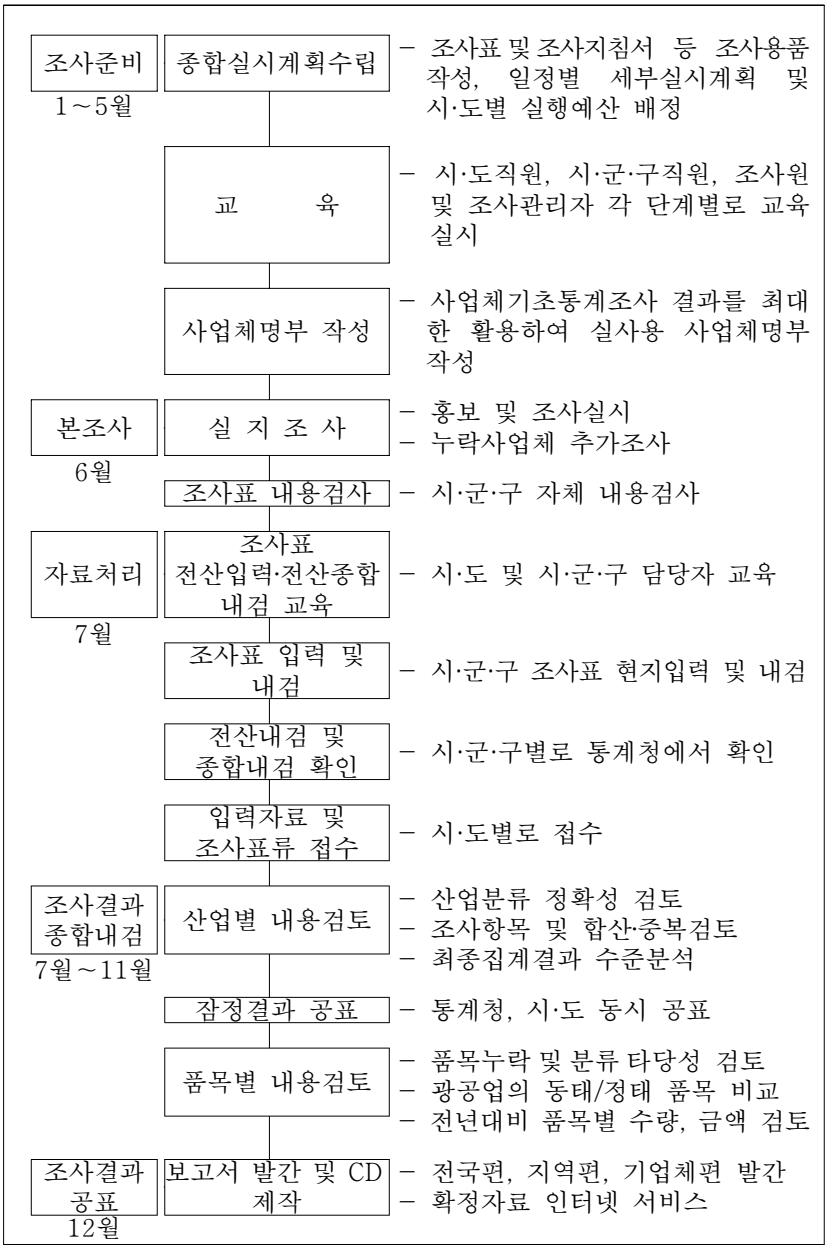
광업·제조업을 영위하는 국내소재 사업체로서 조사기준년도 12월 31일 현재 종사자 수나 월평균 종사자 수가 종사자 5인 이상 광업·제조업 사업체를 대상으로 하는 광업·제조업 통계조사는 광업·제조업 부문에 대한 구조와 분포를 파악하고 산업활동 실태 파악을 목적으로 매년 실시하고 있다. 이 조사는 정부의 각종 경제정책, 민간기업의 경영계획 수립, 그리고 주요 경제통계 작성을 위한 기초자료로 사용되고 있으며 광업·제조업 부문과 관련한 각종 표본조사의 표집틀로 활용된다.

최근 조사의 효율성을 높이고자 조사대상기준을 현행 5인 이상에서 10인 이상으로 변경하였다. 이 조사는 전년도의 1년간의 실적을 매년 6월에 임시조사원이 사업체를 방문하여 조사한다. 통계청↔시·도↔시·군·구↔조사원↔사업체 형태의 조사체계를 갖는다. 잠정결과는 10월에 공표하고 최종결과는 12월에 인터넷(KOSIS)을 통해 제공한다.

1969년부터 2000년까지는 시·도, 시·군·구, 읍·면·동을 통해 조사된 조사표를 본청에 가지고와 직접 입력한 후 전산내검 및 수준점검 등의 자료처리를 하였으나 2001년부터 델파이프로그램(입력, 전산내검, 자료집계 기능)을 이용하여 시·도, 시·군·구에서 직접 조사표를 입력한 후 자료를 취합하여 자료를 처리하고 있다. 자료처리의 흐름은 <표 2-4>와 같다.

한편 광업·제조업은 물론 전기·가스 및 수도 사업을 포함하는 산업총조사는 1955년 한국은행에서 최초로 실시한 이래 2~3년 주기로 한국산업은행에서 실시하여 오다가 1973년부터 통계청에서 인수하여 5년 주기(끝자리가 4, 9인 연도 실시)로 실시하고 있으며 2004년도에 실시한 2003년 기준 산업총조사는 13회째이다. 3, 8자 기준년도에 광업·제조업을 포함하여 종사자 1인 이상의 모든 사업체를 대상으로 조사한다. 광업제조업 통계조사와 산업총조사의 비교는 <표 2-5>에 나타나 있다.

〈표 2-4〉 광업·제조업 사업체조사의 업무흐름도



자료: 통계청, 「통계행정편람」, 2007.

〈표 2-5〉 광업·제조업 통계조사와 산업총조사의 비교

	광업·제조업 통계조사	산 업 총 조 사
조사목적	- 광업 및 제조업부문의 구조와 분포, 경영활동 실태를 파악하여, 경제정책수립 및 시책효과 측정 - 광업, 제조업부문의 표본조사의 모집단자료 제공	- 광업, 제조업, 전기·가스·수도사업부문의 구조와 분포, 경영활동 실태를 파악하여, 경제정책수립 및 시책효과 측정 - 광업, 제조업부문의 표본조사의 모집단자료 제공
실시근거	- 통계법에 의한 지정통계 10109호	- 통계법에 의한 지정통계 10105호
조사주기	- 매년 (끝자리가 4, 9인 연도 제외)	- 5년 (끝자리가 4, 9인 연도 실시)
조사대상 및 규모	- 전국 종사자 5인 이상인 『광업』 및 『제조업』을 영위하는 모든 사업체 약 12만개	- 전국 『광업』 및 『제조업』 『전기업, 가스업, 수도사업』을 영위하는 모든 사업체 약 34만개
조사항목	- 종사자 수, 급여액, 출하액, 재고액, 생산비 및 유형자산 등 16개 항목	- 광업·제조업 통계조사 항목에 『기술연구개발비』, 『연료사용량』 등을 추가 →18개 항목 (종사자 4인 이하 사업체→9개 항목)
조사체계	- 시·도를 통해 임시조사원이 직접 조사대상 사업체 방문조사	- 좌 동 (단, 전기업 · 가스업 · 수도사업은 우편조사)
기준시점 대상기간 실시기간	- 조사 기준년 12. 31. 현재 - 조사 기준년 1년간 - 매년 6월	좌 동 좌 동 좌 동
결과공표	- 매년 10월	좌 동
보고서 발간	- 매년 12월	좌 동

자료: 통계청, 「통계행정편람」, 2007.

2. 4인 이하의 광업·제조업 통계조사의 내용검토

종사자 4인 이하 광업·제조업 사업체조사의 조사표를 살펴보기로 한다. 종사자 수, 출하액 등 8개 항목으로 이루어져 있다(<표 2-6>참조). 조사가 완료된 후 조사표에 대한 내검을 실시한다(통계청, 2004a). 즉, 입력자료의 각 항목 또는 연관된 항목에 대해 입력·내검 프로그램을 운용하여 검사한다. OK에러는 에러사유를 기재하면 내검사항에서 해제되는 에러이며, 필수에러는 나타나면 안 되는 에러이므로 필히 수정을 해야 하는 에러이다.

<표 2-6>의 조사표에서 얻어진 자료는 <표 2-7>에서 보는 바와 같은 내검사항을 검토하고 있다. 각 에러코드에서 밑줄이 있는 코드는 필수에러를 나타내며 밑줄이 없는 코드는 OK에러이다. 종사자 4인 이하 광업·제조업 사업체조사의 조사표 중 4항, 5항, 6항의 연관 내검규칙은 OK에러코드 EA와 EB로, 또한 5항, 6항, 7항과 관련된 내검규칙은 OK에러코드 RD로 내용검토를 하고 있다. 2003년 기준 종사자 4인 이하의 조사표 접수결과 사업체수는 약 19만개로 집계되었다.

〈표 2-6〉 종사자 4인 이하 광업·제조업 사업체조사의 조사표 일부(편집)

4 연간 생산비

* 제품을 생산하는데 투입된 비용을 유형별로 기입

	백억	십억	억	천만	백만원
① 원(부·보조)재료비(미사용분 제외)					
② 연 료 · 전 력 · 용 수 비					
③ 외 주 가 공 비 · 수 선 비					
소 계 (①+②+③)					

5 연간 제품출하액 내역

* 직접 생산한 제품은 물론 타사업체에 원재료를 제공하여 위탁생산한 제품의 매출액도 포함하여 기입
* 부가가치세·특별소비세·주세 등 내국소비세가 제외된 금액

일련번호	★ 품 목 분 류 부 호 (산업 및 품목분류표 참조)	제 품 명	출 하 액				
			백억	십억	억	천만	백만원
01							
02							
03							
99	합 계						

6 연간 임가공(수탁제조)수입액의 품목별 내역

* 원재료(중간제품)를 공급받아 제조·가공한 대가로 받은 금액을 제품별로 기입

일련번호	★ 품 목 분 류 부 호 (산업 및 품목분류표 참조)	임가공품명	임가공수입액				
			백억	십억	억	천만	백만원
01							
02							
99	합 계						

7 유형자산

* 당해 공장의 자산을 기입(임차사용분 제외)

	연 말 잔 액				
	백억	십억	억	천만	백만원
① 토 지					
② 건 물 및 구 축 물					
③ 기 계 장 치 · 차 량 · 기 타					
합 계 (①+②+③)					

〈표 2-7〉 에러코드와 내검사항

error code	전 산 내 검 사 항	착오사항 처리요령
CA	2항 - 2항의 (1)12월말 현재 종사자 수 합계와 (2)월 평균 종사자 수 합계는 4인 이하이어야 함	- 사업체에 확인한 결과 (1)의 합계와 (2)의 합계 모두 5인 이상인 경우는 별도의 조사표로 조사(5인이상 조사표: Ⅰ-1)
CB	2항 - 2항의 ③생산직 종사자 및 ④ 사무 및 기타종사자 1인당 월평균 급여액은 500,000원 이상 400만 원 이하이어야 함	- 2항의 생산직과 사무 및 기타 종사자 수의 급여액이 바뀌어 기입되지 않았는지 확인하여 착오 기입부분을 수정 - 월평균 급여액이 과대한 경우는 연간 급여액이 동일기업 내 타 공장과 합산되었을 가능성이 있으므로 이를 확인하여 수정함 - 배당금, 실적금이 있는 조합원(무급종사자)의 경우 무급가족종사자에 포함 - 확인결과 타당성이 확인되면 사유를 기재한 후 OK 처리함
CD	2항, 4항 - 2항 연간급여액 합계는 4항 ④급여 총액과 일치되어야 함	- 2항의 연간급여액과 4항의 ④급여총액을 재확인하여 수정
EA EB	4항, 5항, 6항 - 조사표상의 수입부문의 합계는 비용부문의 합계보다 커야 함 【수입부문】 (A) 5항 합계, (B) 6항의 합계 【비용부문】 (C) 4항의 소계 (①+…+③) **(A) + (B) < 1.2 × (C)** **(A) > 10 × (C)**	- 다음의 경우는 OK 처리함 · 비고란에 합리적인 적자사유가 반드시 기재된 경우 · 비료제조업, 연탄생산업 등과 같이 정부보조로 적자를 보전하는 경우 - 4항의 ①원재료비에 당해연도의 제품출하를 위해 실제 사용된 원재료비가 아니라 원재료의 재고액까지 합산된 누계액이 조사되었는지 확인함 ※ 대개의 경우 원재료비는 출하액 대비 30~40% 정도임
ED	2항, 5항, 6항 - 종사자 수별로 매출액이 과다인 경우 · 종사자 수 1인 : 매출액이 종사자 수×10억 이상인 경우	- 종사자 규모에 비해 매출액인 큰 경우 제조 이외(상품, 건설, 임대)의 매출 및 타 공장의 매출이 포함되었는지 확인

〈표 2-7〉 에러코드와 내검사항(계속)

	전 산 내 검 사 항	착오사항 처리요령
EC	4항, 5항 - 5항의 제품출하액 합계가 있으면 4항 ①원재료비는 있어야 함	- 대부분의 임가공사업체는 원재료비가 없음(보조재료는 있을 수 있음). 따라서 사업체에 임가공업체 여부를 확인한 후 임가공업체로 확인된 경우 5항 품목별 내역을 삭제한 후 6항에 그 내역을 기입함 - 외주가공비에 원재료비가 포함되었으면 분리 기재함
HA	3항, 7항 - 3항 부지 및 건물연면적 ①자기소유가 있으면 7항 연말잔액①토지, ②건물및 구축물 내역이 있어야 하고, - 3항 부지 및 건물연면적이 ②임차사용만 조사되었으면, 7항 연말잔액 ①토지, ②건물 및 구축물 내역은 없어야 함	사업체에 직접 확인한 결과 - 자기소유의 부지 및 건물연면적이 있는 경우 7항의 ①,②에 대한 내역을 확인 후 수정함 - 자기소유의 부지 및 건물연면적이 없는 경우 3항의 ①자기소유에 대한 내역을 삭제한 후 ②임차사용에 기입함
HD	2항, 4항, 7항 - 4항 ③외주가공비·수선비가 없는 경우 2항 ②생산직 종사자와 7항 ③기계장치 연말잔액과 또는 ⑥기타 연말잔액이 있어야 함	- 사업체에 직접 확인하여 2항, 7항 중 조사누락 부분을 수정 - 외주가공비가 없어도 ·자영업주, 무급(가족)종사자로 이루어진 경우 생산직 종사자가 없을 수 있으므로 비고란에 그 사유를 기재 ·수작업 또는 임차기계를 사용한 경우 7항 ③기계장치·차량기타 연말잔액이 없을 수 있으므로 비고란에 그 사유를 기재
RD	5항, 6항, 7항 - 5항 제품출하액 + 6항 임가공수입액 / 조업월수 × 12개월 < 0.1 × 유형자산 연말잔액	-사업체에 타 공장 및 타 산업의 자산이 포함되었는지 확인하여 수정
QD	5항, 6항 -소규모 사업체에서 생산하기 어려운 품목이 조사된 경우(자동차, 선박, 컴퓨터 등) (34, 35, 30)	-부품을 생산하는 경우가 아닌지 확인
SS	2항 - 세항목 간 가로합은 합계와 일치되어야 함	- 이 경우는 세항목 간의 합과 합계가 불일치 할 경우 재확인 후 착오부분을 수정
TT	2항~7항 - 2항~7항의 세로합은 합계와 일치되어야 함	

자료: 통계청, 「입력·내검 프로그램 운영요령서」, 2004.

3. 내재적 내검규칙의 생성과 프로그램 수행

내검을 통과하지 못한 레코드에서 수정되어질 최소개의 변수를 찾기 위해 담당자에 의해 설정된 명시적 내검규칙(explicit edits)과 내재적 내검규칙(implied edits)이 필요하다. 내재적 내검규칙은 명시적 내검규칙에 의해 논리적으로 유도되는 규칙이다. 내재적 내검규칙의 생성은 Fellegi-Holt의 논문(1976)의 3번째 Theorem에서 주어지는데 내재적 내검규칙 생성의 핵심은 변수(필드)의 소거이다.

이와 같이 주어진 내검규칙과 이로부터 생성된 새로운 내검규칙(내재적 내검규칙)을 이용하여 모든 내검규칙을 만족하게 하는 최소개 변수(항목)의 위치를 자동포착하고자 한다. 이를 위해 먼저 각 레코드가 내검규칙에 대해 검토된다. 실패한 각 레코드는 하나 이상의 내검규칙을 만족하지 않는다. 어떤 레코드는 두 개 이상의 내검규칙에 어긋날 수 있으며 실패한 내검규칙의 변수를 모두 커버(cover)하는 변수군을 가질 수 있다. 이 정보는 오류포착단계에서 중요하다. 본 연구에서는 이상과 같이 내재적 내검규칙의 생성과 공통으로 커버하는 변수군을 찾아내어 오류위치를 포착하고자 한다.

종사자 4인 이하의 광업・제조업 통계조사 자료에 자동내검기법을 적용하여 자동내검의 유용성을 검토한다. 사업체대상 조사의 자동내검기법연구(이의규와 심규호, 2007)에서 제시한 바와 같이 4인 이하 사업체조사의 해당 항목 간 수량적 연관규칙에 대해 적용한다. 아래와 같이 각 항목을 기호화 한다.

- 수입부문 : 5항 합계 = 연간 직접생산, 위탁생산 제품출하액 (X_1)
  6항 합계 = 연간 임가공(수탁제조) 수입액 (X_2)
- 비용부문 : 4항 소계 = 연간 원재료비 (X_3)
- 유형자산 : 7항 합계 = 공장의 자산 (X_4)

이제 EA, EB, RD의 각 내검사항을 수식화하면 다음과 같다.

- 수입부문의 합계는 비용부문의 합계보다 커야 함.

$$X_1 \quad + \quad X_2 \quad \geq \quad 1.2 \; X_3$$

- 원재료비는 출하액의 10%보다는 커야 함.

$$X_1 \qquad \leq \quad 10 \; X_3$$

- 수입부문의 합계가 대략 유형자산 연말잔액의 10%보다는 커야 함.

$$X_1 \quad + \quad X_2 \quad \geq \quad 0.1 \; X_4$$

즉, 주어진 명시된 내검규칙(explicit edits)은 다음과 같이 정리된다.

$$\begin{aligned}
&\mathrm{E}1: \quad X_1 + X_2 - 1.2\,X_3 && \geq 0\\
&\mathrm{E}2: -X_1 \qquad + 10\,X_3 && \geq 0\\
&\mathrm{E}3: \quad X_1 + X_2 \qquad - 0.1\,X_4 && \geq 0
\end{aligned}$$

이로부터 유도되는 내재적 내검규칙(implicit edits)은

$$\mathrm{E}4: \; X_2 + 10\,X_3 - 0.1\,X_4 \quad \geq 0$$

이다. 유도되는 내재적 내검규칙은 이 외에도 존재하나 실질적으로 필요치 않아 생략하였다.

위에서 3번째 내검규칙은 조업월수가 12개월을 가정한 것이다. 만약 조업월수를 고려하면 이 규칙은 다음과 같이 표현할 수 있다.

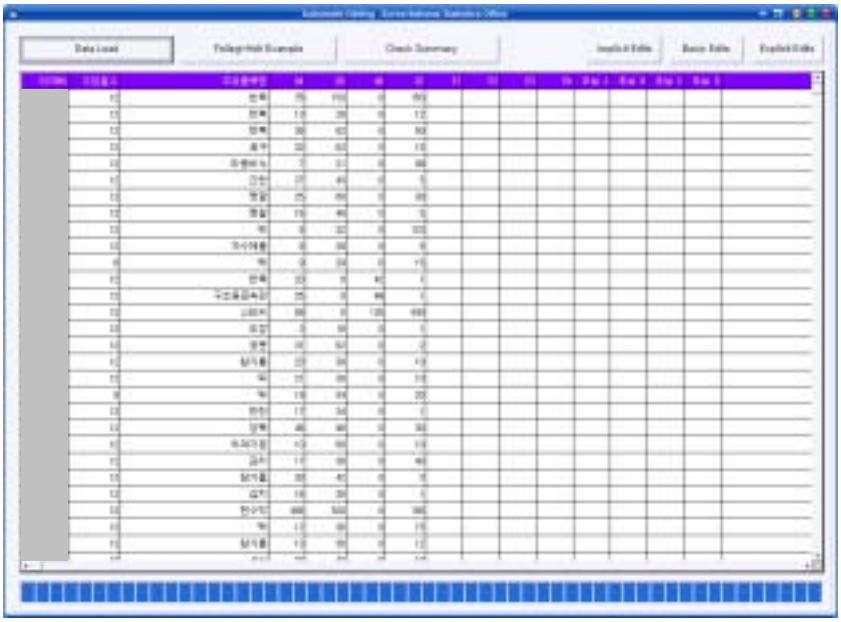
$$\mathrm{E}3': \quad \frac{12}{a}(X_1 + X_2) - 0.1\,X_4 \quad \geq 0$$

여기서 a는 해당 레코드의 조업월수이다. 즉 조업월수가 1개월이면 해당 레코드의 수입부문에 12배를 해 준 값이 유형자산액의 10%보다 커야 함을 의미한다. 따라서 이 식에 따른 내재적 규칙 E4는 다음과 같이 다시 표현된다.

$$\mathrm{E}4': \; \frac{12}{a}(X_2 + 10\,X_3) - 0.1\,X_4 \quad \geq 0$$

4인 이하의 광업·제조업 사업체조사 자료에 자동내검기법을 적용하기 위해 Microsoft Visual Basic 6.0으로 프로그램을 작성하였다. [그림

2-6]은 “Data Load” 버튼을 눌러 데이터를 불러온 상태의 화면이다. [그림 2-7]는 내검규칙을 검토 수행한 화면이다. “FALSE”가 발생한 레코드에 대해서는 알아보기 쉽도록 화면상에 적색으로 표시하였다. 전체자료 화면에서 오류레코드를 탐색하기 쉽도록 “Check summary” 버튼을 누름으로써 오류레코드만을 요약한다. 오류레코드에 내재적 내검규칙과 위배된 내검규칙의 공통변수를 이용하여 최종적으로 오류위치를 나타내었다([그림 2-8]).



[그림 2-6] 데이터 불러오기

[그림 2-7] 내검 수행

[그림 2-8] 오류레코드의 요약 및 오류위치포착

4. 적용 결과

분석결과 <표 2-8>는 내검규칙에 실패한 레코드의 위배규칙별 건수를 나타낸다. 총 189,424건 중 2,175건이 내검규칙에 위배된 것으로 나타나 전체 자료의 약 1.1%를 차지한다. 앞에서도 언급하였듯이 자료이용의 한계로 인해 내검이 완료된 자료를 이용하였다. 여기서의 에러는 OK에러로서 확인이 필요하고 사유가 있으면 허용되는 에러이다. 에러에 걸린 자료에서도 대부분은 소금(천일염) 등과 같이 주요생산비가 거의 없으면서도 출하액이 높은 특정 품목을 생산하는 사업체가 많다. 즉 OK에러에 걸린 자료의 대부분은 허용이 인정된 사업체로 해석에 유의할 필요가 있다. 본 연구에서는 OK 내검규칙에 실패한 레코드의 일부분으로부터 자동오류위치포착 및 수정의 의미를 찾고자 한다.

한편, 각 실패 레코드에서 하나의 명시된 내검규칙을 위배한 경우도 있으나 두 개의 명시된 내검규칙을 위배한 경우도 있다. 또한 어떤 레코드들은 명시된 내검규칙과 유도된 내검규칙을 위배하고 있다. 각 경우에 따라서 어떤 레코드는 내검규칙을 만족시키기 위해 하나의 변수가 결정되기도 하지만 어떤 레코드는 두 개 또는 세 개의 변수 중 하나를 선택해야 할 경우가 발생한다. 다시 말해 유일하게 결정되지 않을 수 있다. 다음에서는 각 레코드의 위배된 경우에 따라서 오류위치포착 문제를 구체적으로 살펴본다.

〈표 2-8〉 실패 레코드의 위배규칙별 건수

위배규칙	건수	위배규칙	건수	위배규칙	건수
E1	594	E1, E3	66	E1, E3, E4	2
E2	824	E2, E4	109	E2, E3, E4	4
E3	498	E3, E4	78	합계	2,175

가. E1을 위배한 경우

수입부문의 합계는 주요생산비의 1.2배보다 커야한다는 내검규칙 E1에 위배되는 건수가 594건으로 나타났다. 다시 말해 이들 레코드는 출하액과 임가공수입액을 합한 금액이 주요생산비 수준보다도 적은 레코

드를 말한다. 이 규칙에 위배된 레코드의 유형은 <표 2-9>에서 보듯이 수입부문보다 비용부문이 상대적으로 큰 사업체이다.

〈표 2-9〉 내검규칙 E1을 위배한 사업체 일부

(단위: 백만 원)

사업체 고유번호	조업 월수	주요품목	출하액	임가공 수입액	생산비 -소계-	유형 자산
513***	12	디지탈 AVR 제어보드	130	0	305 (13, 108)	50
266***	12	떡	6	0	54 (2, 5)	15
165***	12	방충망 샷시	12	0	31 (2, 10)	10
145***	12	산업기계제작	12	0	80 (2, 10)	20
346***	12	의료기기 제조 (맥진기)	10	0	70 (1, 8)	20
164***	12	잡지	3	0	200 (1, 2)	0
146***	12	화장품	5	0	51 (1, 4)	10

규칙에 어긋나는 특별한 사유가 기재되면 내검에 통과되나 만약 재접촉이 불가능한 자료에 대해서는 자동내검 및 수정을 고려할 필요가 있다. 당연하게 수입부문이나 비용부문 둘 중 하나를 선택하여 수정한다. 그러나 수입부문이 유형자산의 10%보다 커야한다는 조건(E3)은 성립하므로 수입부문과 유형자산은 일관성을 유지한다. 따라서 주요생산비를 검토하는 전략이 적절하다. <표 2-9>에서 생산비의 괄호 속에 내검규칙을 만족하게 하는 가능한 범위를 나타내었다. 이는 생산비를 미지수로 놓고 나머지 항목 값을 각 내검규칙에 대입하여 얻은 결과이다.

나. E2를 위배한 경우

주요생산비는 출하액의 10%보다 커야한다는 내검규칙 E2에 위배되는 경우가 824건으로 나타났다. 이 조건을 만족하지 못한 레코드는 출하액이 주요생산비의 10배보다도 많은 레코드로서 주요생산비에 비해

과다한 출하액을 보이는 사업체이다. 이들 사업체들 중에서 소금(천일염) 사업체가 627개로 대부분을 차지한다. 따라서 이들 품목을 생산하는 사업체의 내검기준이 완화될 필요가 있다.

<표 2-10>은 내검규칙 E2를 위배한 사업체의 일부를 나타낸다. 이 가운데 주요생산비가 0(백만 원)으로 기입된 경우는 출하액이 1(백만 원)이라도 규칙을 위반하게 되므로 주요생산비가 0이면서 출하액이 작은 영세사업체는 내검에서 제외되어야 할 것이다. 이외 특별한 사유가 있으면 사유코드를 기재하도록 하여 내검을 통과하게 한다. 만약 재접촉이 불가능한 실패 레코드는 자동내검 및 대체를 고려한다. 주요생산비와 유형자산은 일관성을 띠므로 특정품목과 영세사업체를 제외하고는 출하액을 우선 검토하는 것이 바람직하다(또는 임가공업체로서 임가공수입액에 기재되어야 할 가능성도 있음). <표 2-10>에서 출하액의 괄호 안에 가능한 범위 값을 나타내었다.

〈표 2-10〉 내검규칙 E2를 위배한 사업체 일부

(단위: 백만 원)

사업체 고유번호	조업 월수	주요 품목	출하액	임가공 수입액	생산비 -소계-	유형 자산
182***	12	도장(목)	2	0	0	0
183***	12	도장	2	0	0	0
257***	12	인장	2	0	0	0
257***	12	도자기	1	0	0	0
258***	1	조화, 꽃장식소품	1	0	0	0
258***	12	도장	1	0	0	0
259***	1	반응도 계산기	10	0	0	0
376***	1	장식용목공예	1	0	0	0
385***	1	쌀도정	4	4	0	8
395***	12	도장	1	0	0	0
434***	12	도장	1	0	0	0
502***	2	정기간행물	30	0	0	0
512***	3	떡, 참기름, 들기름, 배즙	1	0	0	0
253***	12	주조형금형	200	0	0	0

194***	12	산업기계제조부품	118 (3, 10)	0	1	30
200***	12	산업용기계부품 절삭가공	155 (7, 10)	0	1	70
215***	12	산업기계부품	120 (7, 10)	0	1	70
217***	12	산업기계	100 (1, 10)	0	1	10
233***	12	철판절단	150 (10, 10)	0	1	100
165***	12	노트북가방	389 (1, 20)	0	2	6
166***	12	기성복	250 (2, 20)	0	2	20
515***	12	발파석	1600 (3, 30)	0	3	30

다. E3를 위배한 경우

수입부문의 합계가 유형자산 연말잔액의 10%보다는 커야한다는 내검규칙 E3에 위배되는 건수가 498건으로 나타났다. <표 2-11>은 내검규칙 E3를 위배한 사업체의 일부이다. 조건을 만족하지 못한 레코드는 유형자산에 비해 수입부문의 금액이 작은 경우에 해당한다. 이 규칙에 위배된 레코드의 유형을 살펴보면, 많은 유형자산을 요구하는 쌀(벼)(도정), 떡, 고추(고춧가루), 참・들기름, 건강보조용 즙(액) 등 곡물 등을 빻고, 찧고, 짜내는 업종이 대부분을 차지한다.

이들 품목과 관련된 레코드에 대해서는 사전에 내검에 걸리지 않도록 내검기준을 완화할 필요가 있다. 또한 실제로 유형자산에 비해 기준 이하의 낮은 출하나 임가공수입이 있을 경우 사유코드를 부여하여 자동내검 및 수정에서 제외되도록 조치한다. 재접촉이 불가한 특이 자료에 대해서는 자동 내검 및 대체를 실시한다.

여기서는 수입부문이나 유형자산 중 하나를 수정한다. 그러나 주요생산비가 함께 낮게 조사되어 적은 주요생산비에 수입부문의 금액이 작은 것은 일관되어 보인다(E2 성립). 따라서 이 규칙을 위배하는 품목유

형과 관련된 사업체 이외에는 유형자산에 대한 검토가 우선적으로 고려되어야 할 것이다. 역시 내검기준을 만족하기 위한 <표 2-11>에 유형자산의 허용되는 최대값을 괄호 속에 나타냈다.

〈표 2-11〉 내검규칙 E3를 위배한 사업체 일부

(단위: 백만 원)

사업체 고유번호	조업 월수	품목	출하액	임가공 수입액	생산비 -소계-	유형 자산
165***	12	떡, 참기름, 고춧가루	50	0	21	1,100 (500)
327***	12	쌀	0	100	42	3,812 (1000)
164***	12	여성정장	100	0	37	2,610 (1000)
167***	12	여성정장	150	0	47	4,011 (1500)
438***	12	유기질 비료	70	0	28	1,439 (700)
118***	12	철강 절단품	80	0	31	1,600 (800)
201***	12	콘크리트 블록	130	0	72	2,701 (1300)
386***	12	콘테이너 제조	60	0	47	1,870 (600)
357***	9	타월 제조	50	0	24	1,058 (667)
201***	12	플라스틱 성형제품	0	12	7	290 (120)
138***	12	한약기세정제	60	0	42	1,653 (600)

이처럼 하나의 내검규칙을 위반한 경우에는 수정하여할 변수가 유일하게 결정되지 않는다. 하나의 해결 방법은 위에서 언급한 바와 같이 다른 항목이나 주로 위배하는 품목정보를 이용할 수 있으며 각 항목의 응답 신뢰도를 고려하여 수정해야할 변수를 선택하는 것이다.

라. E1, E3를 위배한 경우

수입부문의 합계가 주요생산비의 1.2배보다 커야한다는 내검규칙 E1과 수입부문의 합계가 유형자산 연말잔액의 10%보다는 커야한다는 내검규칙 E3를 동시에 위배한 건수는 66건으로 나타났다. 역시 재접촉 불가시 자동내검 및 대체를 고려할 수 있다. 그런데 이 경우 주요생산비 하나를 수정할 때에는 다른 내검규칙을 만족시킬 수 없으며 유형자산을 수정하더라도 다른 한 쪽 내검에 걸릴 수 있다. 물론 두 개의 항목을 수정하면 모든 내검규칙을 만족할 수 있으나 공통으로 들어간 수입부문을 수정하는 것이 가능한 정보를 보존한다는 입장에서 합리적이다(<표 2-12> 참조).

〈표 2-12〉 E1과 E3에 위배된 내검규칙 행렬

	X1	X2	X3	X4
E1	1	1	1	
E3	1	1		1

<표 2-13>에 내검규칙 E1과 E3를 위배한 사업체들 중 몇 개를 제시하였다. 컴퓨터 부품잉크를 생산하는 사업체는 (1, 0, 22, 267)로 코딩되었다. 즉 2억 6천7백만 원의 유형자산을 갖는 사업체가 주요생산비 2천2백만 원을 들여 출하액이 1백만 원인 경우이다. 이때 출하액이 주요생산비와 유형자산에 비해 매우 작은 자료로 어떤 특정사유가 없다면 잘못 보고되거나 기재될 가능성이 크므로 수정되는 것이 바람직할 것이다. 즉 조사 시 해당사유가 기재된다면 문제가 없으나 아무 사유가 없이 입력된 자료는 이 경우 적어도 출하액이 조정되어야 할 가능성이 매우 높다. 각 내검규칙에 출하액을 미지수로 놓고 레코드의 나머지 값을 대입하면 출하액은 2천7백만 원 이상 2억 2천만 원 이하의 출하액 범위를 갖게 되고 이 범위 값 안에서 모든 내검규칙을 만족한다.

또한 멸치액젓을 제조하는 사업체는 (20, 0, 90, 1050)으로 코딩되었다. 즉 10억 5천만 원의 유형자산을 소유한 이 업체가 주요생산비 9천

만원을 들여 출하한 금액이 2천만 원이므로 이때 출하액 하나만을 수정하는 것이 타당할 것이다. 유니폼 사업체는 (1, 0, 20, 130)의 필드 값을 갖는다. 즉 1억 3천만 원의 유형자산을 소유한 업체가 2천만 원의 주요생산비에 백만 원의 출하액을 보고한 자료로 출하액 수정이 합리적이다.

〈표 2-13〉 내검규칙 E1, E3를 위배한 사업체 일부

(단위: 백만 원)

사업체 고유번호	조업 월수	품목	출하액	임가공 수입액	생산비 -소계-	유형 자산
360***	12	컴퓨터 부품잉크	1 (27, 220)	0	22	267
515***	12	멸치액젓	20 (108, 900)	0	90	1,050
506***	12	유니폼	1 (24, 200)	0	20	130

마. E3, E4를 위배한 경우

수입부문의 합계가 유형자산 연말잔액의 10%보다는 커야한다는 내검규칙 E3와 내검규칙 E4(E2와 E3에서 유도)에 동시 위배된 건수는 78건으로 나타났다. 이 내검규칙에 어긋난 업체의 유형에는 수입부문과 주요생산비가 유형자산에 비해 특이하게 작은 경우가 많다. 특히 출하액과 주요생산비가 모두 0인 경우가 많은데 주요생산비가 0이어도 되는 임가공 업체나 유형자산 역시 작은 경우에는 내검조건을 완화할 필요가 있다. <표 14>에서 보듯이 임가공 수입액 또는 유형자산 연말잔액이 두 규칙을 커버함으로써 임가공수입액이나 유형자산 연말잔액을 수정하는 것이 합리적이다. 그런데 해당 레코드에 어떠한 사유가 없고 출하액과 주요생산비가 작은 경우에는 유형자산이 하향 조정되는 것이 설득력 있다.

〈표 2-14〉 E3와 E4에 위배된 내검규칙 행렬

	X1	X2	X3	X4
E3	1	1		1
E4		1	1	1

또한 임가공수입이 없는 사업체인 경우에는 유형자산 연말잔액 항목을 수정함이 바람직하다. <표 2-15>를 참조하면, 50***0 레코드는 알루미늄 절단 사업체로서 (0, 9, 2, 1040)으로 코딩되었으며 이는 유형자산이 10억 4천만 원이고 임가공수입액이 9백만 원 주요생산비가 2백만 원으로서 유형자산에 비해 주요생산비와 임가공수입액이 매우 작은 경우이다. 사유가 기재되지 않고 재접촉이 불가하다면 자동내검 및 수정이 불가피하다. 50***7 레코드는 절삭가공 업체로 (0, 45, 1, 2721)으로 코딩되었다. 즉 유형자산이 27억 2천백만 원이고 임가공수입액이 4천5백 만 원 그리고 주요생산비가 백만 원으로 유형자산이 상대적으로 매우 크다. 또한 164*** 레코드는 호박, 매실즙 업체로서 10억 2천2백만 원의 사업체로서 임가공수입액은 천오백만 원이고 주요생산비는 2백만 원이다. 513*** 레코드(사출제품 업체)는 주요생산비 2억 천9백만 원과 출하액 4억 7천만 원에 비해 유형자산이 386억 5천 1백만 원으로 매우 크다. 앞에서와 같은 방법으로 제약식으로부터 유형자산의 범위를 구할 수 있다. 각각의 유형자산은 <표 2-15>에서와 같이 괄호 안에 있는 값이 내검조건을 만족하는 최대값이다.

〈표 2-15〉 내검규칙 E3, E4를 위배한 사업체 일부

(단위: 백만 원)

사업체 고유번호	조업 월수	품목	출하액	임가공 수입액	생산비 -소계-	유형 자산
51***0	12	알루미늄 절단	0	9	2	1,040 (290)
50***7	12	절삭가공	0	45	1	2,721 (550)
164***	12	호박, 매실즙	0	15	2	1,022 (350)
513***	12	사출제품	470	0	219	38,651 (4,700)

바. E2, E4를 위배한 경우

주요생산비가 출하액의 10%보다는 커야한다는 E2 규칙과 유도된 내검규칙 E4가 동시에 위배된 경우는 109건이다. 이 규칙들에 위배된 자료는 주요생산비가 출하액이나 유형자산에 비해 매우 작은 경우(없거나 4백만 원 이하)로 나타났다. 이 경우 유형자산이나 출하액이 작으면서 주요생산비가 0이면 허용되도록 내검기준을 완화할 필요가 있다.

그 밖에 허용되는 사유가 있으면 자동내검에서 제외되도록 조치한다. 특히 소금과 짚신은 원재료비가 거의 없고 인력에 의해 제조되므로 주요생산비가 적어도 허용되도록 한다. 만약 허용사유도 없고 재접촉 불가시에는 수정이 필요한데 이 경우는 <표 2-16>에서와 같이 주요생산비가 공통으로 두 규칙을 커버하므로 주요생산비를 수정하는 것이 바람직하다. 왜냐하면 출하액을 수정하더라도 다른 하나의 내검을 위배하고 유형자산을 수정하더라도 다른 한쪽 내검을 만족하지 못하기 때문이다.

〈표 2-16〉 E2와 E4에 위배된 내검규칙 행렬

	X1	X2	X3	X4
E2	1		1	
E4		1	1	1

<표 2-17>에서 원격조명릴 사업체(513*** 레코드)는 (189, 0, 0, 350)으로 3억 5천만 원의 유형자산, 1억 8천9백만 원의 매출액에 비해 주요생산비는 0이다. 앞에서와 같은 방법으로 주요생산비가 1천9백만 원 이상, 1억 5천8백만 원 이하로 수정되면 모든 내검규칙을 만족한다. 재단판 사업체(198*** 레코드)는 (86, 0, 4, 605)로 6억 5백만 원의 유형자산, 8천6백만 원의 판매액에 비해 주요생산비는 4백만 원이다(주요생산비는 최소 9백만 원 이상 7천2백만 원 이하가 되어야 함). 또한 치아제조 사업체(373*** 레코드)는 (27, 0, 1, 110)으로 유형자산 1억 천만 원, 매출액 2천7백만 원에 비해 주요생산비는 1백만 원으로 매우 작다.

〈표 2-17〉 내검규칙 E2와 E4를 위배한 사업체 일부

(단위: 백만 원)

사업체 고유번호	조업 월수	품목	출하액	임가공 수입액	생산비 -소계-	유형 자산
513***	12	원격조명릴	189	0	0 (19, 157)	350
198***	12	재단판	86	0	4 (9, 72)	605
373***	12	치아제조	27	0	1 (3, 22)	110

사. E2, E3, E4를 위배한 경우

E2, E3, E4에 위배된 레코드는 모두 4건이다. 이때는 하나의 변수가 모든 내검규칙을 커버하지 못하기 때문에 최소한 두 개를 바꾸어야 한다. 그러나 X2와 X4 두 개를 바꾸어도 모든 내검규칙을 커버하지 못하므로 이 쌍은 제외된다(<표 2-18>참조).

〈표 2-18〉 E2, E3, E4에 위배된 내검규칙 행렬

	X1	X2	X3	X4
E2	1		1	
E3	1	1		1
E4		1	1	1

이들 규칙을 위배한 레코드 중 하나인 자동차 소음방지 부품업체(366***)는 (20, 0, 1, 1325)로 유형자산 13억 2천5백만 원, 매출액 2천만 원, 주요생산비가 1백만 원이다. 이때는 유형자산을 수정해도 주요생산비와 매출액 간 내검규칙을 만족할 수 없고 매출액을 수정해도 주요생산비와 유형자산 간 내검규칙을 만족할 수 없으며 주요생산비를 수정해도 수입부문과 유형자산 간 내검규칙을 만족할 수 없다. 따라서 이 경우는 주요생산비와 매출액, 주요생산비와 유형자산, 매출액과 유형자산 중 하나의 변수쌍을 수정하여야 모든 내검규칙을 만족하게 된다(생산비와 유형자산을 택하였을 경우 생산비는 2와 16 사이, 유형자산은 200보다

작아야 함). 이는 물론 어떤 특정사유가 기재되지 않고 또한 재접촉이 불가한 경우에 필요한 조치이다.

아. E1, E3, E4를 위배한 경우

E1, E3, E4에 위배된 레코드는 모두 2건이다. 이때는 <표 2-19>에서와 같이 임가공수입액이 존재하면 임가공 수입액을 수정하는 것이 정보의 손실을 최소화한다. 그러나 출하액만 존재한다면 역시 두 개의 변수를 수정하여야 한다.

〈표 2-19〉 E1, E3, E4에 위배된 내검규칙 행렬

	X1	X2	X3	X4
E1	1	1	1	
E3	1	1		1
E4		1	1	1

예를 들면 이 규칙들을 위배한 남성복 맞춤 사업체 (403***)는 (1, 0, 1, 234)으로 유형자산 234백만 원, 출하액이 1백만 원, 주요생산비가 1백만 원이다. 이때 출하액이 유형자산의 10%보다 크도록 출하액을 2와 10 사이의 값, 그리고 유형자산을 100보다 작은 값으로 수정하면 모든 내검규칙을 만족하게 된다. 역시 재접촉이 불가하고 필요한 경우에만 수정조치를 취할 수 있다.

이상의 결과를 요약하면 <표 2-20>과 같다. 하나의 내검규칙만을 위배한 경우는 변수의 수정위치를 선택해야 한다. 이 경우 해당 레코드에서 위배된 규칙에서의 항목 이외의 항목들을 검토하거나 주요 품목을 고려하여 자동화 전략을 세울 수 있을 것이다. 두 개 이상의 내검규칙에 위배된 경우는 공통변수를 찾음으로써 최소의 정보손실로 모든 내검규칙을 만족할 수 있다.

앞에서도 언급하였듯이 적용자료는 이미 내검이 완료된 자료이다. 특히 여기서의 에러는 OK에러로 사유가 있으면 허용되는 에러이므로 내검에 걸린 모든 자료가 오류를 의미하는 것은 아니다. 그러나 오류일 가능성이 높은 자료에 자동기법을 적용하여 보았다.

실제로 내검이 수행되기 전 자료는 많은 오류가 발생될 수 있고 이때 자동오류위치포착기법을 적용한다면 수작업 에디팅 시 어떤 항목을 우선 검토해야 하는 지의 정보를 줄 수 있다. 한편 자동 에디팅을 고려할 때는 내검조건을 이보다 좀 더 강하게 부여하여 매우 불합리한 항목간 연관성을 갖는 레코드만을 탐지하고, 수정이 불가피할 시 일관되고 합리적인 전략으로 대처할 수 있을 것이다.

〈표 2-20〉 위배규칙에 따른 오류수정변수

위배규칙 (x1, x2, x3, x4)	건수	오류수정변수 x1:출하액 x2:임가공수입액 x3:주요생산비 x4:유형자산
E1 (1 1 1 0)	594	x1 또는 x2 또는 x3(x3 우선 검토) (x1=0 → x2 또는 x3, x2=0 → x1 또는 x3)
E2 (1 0 1 0)	824	x1 또는 x3(x1 우선 검토)
E3 (1 1 0 1)	498	x1 또는 x2 또는 x4(x4 우선 검토) (x1=0 → x2 또는 x4, x2=0 → x1 또는 x4)
E1, E3 $\begin{pmatrix}1&1&1&0\\1&1&0&1\end{pmatrix}$	66	x1 또는 x2 (x1=0 → x2, x2=0 → x1)
E2, E4 $\begin{pmatrix}1&0&1&0\\0&1&1&1\end{pmatrix}$	109	x3
E3, E4 $\begin{pmatrix}1&1&0&1\\0&1&1&1\end{pmatrix}$	78	x2 또는 x4 (x2=0 → x4)
E1, E3, E4 $\begin{pmatrix}1&1&1&0\\1&1&0&1\\0&1&1&1\end{pmatrix}$	2	x2 {x2=0 → (x1, x3) 또는 (x1, x4) 또는 (x3, x4)}
E2, E3, E4 $\begin{pmatrix}1&0&1&0\\1&1&0&1\\0&1&1&1\end{pmatrix}$	4	(x2, x4)를 제외한 가능한 쌍 {x1=0 → (x2, x3) 또는 (x3, x4), x2=0 → (x1, x3) 또는 (x1, x4) 또는 (x3, x4)}
계	2,175	

제5절 결론

1. 시사점

개인정보보호 인식, 조사원의 조사노력, 응답자의 조사협조 등 우리의 조사환경은 북미·유럽과는 차이가 있다. 또한 우리의 최종자료는 많은 내검과정을 수행하기 때문에 자동 에디팅 시스템에 대한 필요성이 상대적으로 작으며 자동 수정에 대한 우려가 큰 것으로 판단된다. 그러나 철저한 내검에도 불구하고 단위착오 등으로 인한 잘못된 정보가 포함될 가능성은 여전히 존재하므로 수정이 요구될 수 있다. 더욱이 우리나라의 조사환경이 서구의 조사환경 방향으로 가지 않는다고 말할 수는 없다. 따라서 조사환경 변화에 대비하기 위한 일반화된 자동 에디팅 시스템에 대한 연구가 필요하다.

캐나다 통계청의 Banff는 주목할 만하다. 이 시스템은 윈도우 환경에서 9개의 각 SAS 프로시저가 독립적으로 수행가능하다. 자료의 입출력 처리가 SAS로 이루어져 사용자 친화적이며 자료처리과정이 매우 유연하다. 예를 들면 내검규칙에 대한 설정분석, 자료의 오류분석, 이상치분석, 응답대체분석 등이 따로 따로 적용이 가능하여 여러 사업체 조사에 여러 용도로 사용할 수 있다. 이는 마치 PC SAS를 분석하여 통계분석을 하는 것과 유사하다. 또한 내검규칙의 삭제, 변경, 추가가 유연하여 내검규칙에 서로 모순이 없는지, 중복된 것은 없는지, 결정되는 값은 없는지 등 내검규칙의 분석도 가능하고 이상치나 응답대체분석에도 사용자가 대체방법을 사용자 설정 조건으로 시행할 수 있다는 장점이 있다.

캐나다 통계청의 일반화된 에디팅시스템은 각 조사마다 다른 내검프로그램과 임퓨테이션을 수행하는 것이 아니라 하나의 시스템을 공통적인 절차로 이용하는 것이다. 단지 크게 가구조사와 사업체조사에서 쓰이는 시스템으로 대별될 뿐이다. 공통적인 시스템은 논리와 절차가 유사하므로 실무담당자가 이해하기가 매우 쉽다는 장점이 있다. 우리는 캐나다를 비롯한 해외의 자동 에디팅 시스템의 사례로부터 다음과 같은 시사점을 얻을 수 있다.

가. 오류자료와 결측자료의 분석

데이터 에디팅과 임퓨테이션은 자료의 품질을 보증하기 위한 중요한 과정이다. 그러나 무엇보다도 오류나 무응답이 없도록 하는 것이 근본적으로 중요하므로 오류나 무응답을 방지할 수 있도록 하는 노력이 필요하다. 응답이 이상하거나 응답하기 꺼리는 항목에 대한 개선이 중요할 것이다. 즉 에디팅 과정은 자료의 오류나 결측치를 줄이는 방향으로 나아가야 할 것이다. 이러한 점에서 오류자료와 결측자료는 매우 중요한 정보이다. 따라서 에디팅과 임퓨테이션 시스템은 오류자료와 결측치의 원인을 분석할 수 있도록 다양한 분석이 반드시 포함되어야 할 것이다.

나. 일련의 에디팅과정을 포함하는 하나의 시스템

자동 에디팅에서는 무응답뿐만 아니라 내검규칙에 매우 어긋난 이상한 값들도 모두 대체해야 할 대상으로 간주한다. 따라서 이상치에 대한 검토가 필요하다. 이상치를 포함한 분석은 포함하지 않을 때와는 다른 결과가 도출될 수 있으므로 이상치에 대한 기준과 검출이 요구된다. 이는 분석의 첫 단계일 뿐 아니라 무응답 대체 시 이상치가 사용되어서는 안 되기 때문이다. 따라서 에디팅 시스템은 기본적인 자료정제, 이상치 점검, 내검규칙분석, 오류위치포착, 임퓨테이션이 서로 연관되어 있는 하나의 시스템으로 구성되어야 한다.

다. 원자료의 보존

응답자의 원 자료는 매우 중요하다. 이를 위해서는 이상치나 무응답 자료가 불가피할 때 이들을 허용해야 할 필요가 있다. 수정 및 대체가 있을 때에는 그것에 대한 표식을 달아줄 필요가 있다. 자료의 수정 기록을 남겨야 후속연구가 가능하다.

라. 내검규칙의 개발

인간의 섬세한 작업을 기계가 대신하는 것에는 어느 정도의 한계가 있으나 자동화될 수 있는 수작업은 가능한 자동화할 필요가 있다. 예를 들면 각 산업분류마다, 품목마다, 지역마다 다른 내검규칙을 개발하거나 레코드마다 해당 레코드의 상황이나 정보를 판단할 수 있는 추가정보가 이용될 수 있다면 좀 더 세밀한 자동화가 이루어져 자동화로 인한 위험은 작아지고 효율은 높아질 것이다.

마. 최신 에디팅기법 의한 보완

영향력이 작은 조사대상은 자동 수정을 통해 조사 자료의 완전성을 기함과 동시에 시의성, 경제성의 효과를 거둘 수 있다. 그러나 자동 에디팅만으로는 좋은 품질의 데이터를 얻는 데 충분하지 않을 수 있다. 영향력이 큰 레코드는 수작업으로 검토되어야 하는 등 최신 내검기법과 결합되어야 한다. 통계 선진국에서도 매우 중요한 사업체에 대해서는 철저하게 수작업으로 점검한다. 다시 말해 사업체의 규모가 큰 사업체는 수작업으로 내용검토가 이루어지고 다만 중소사업체에 대해 자동내검을 시행하여 내검의 효율성을 제고한다.

2. 기대효과 및 향후과제

담당자에 의해 주어진 내검규칙을 벗어나는 자료가 실제로 맞는 응답일 수도 있으나 이러한 경우는 매우 드물다. 위험을 최소화하기 위해서는 좀 더 강한 내검규칙을 부여하고 이를 위배한 경우에만 자동 수정을 고려할 수 있을 것이다. 현재로서는 통계청의 각 입력·내검 프로그램에 이러한 오류위치포착 기능을 추가하여 입력단계에서 오류발생 시 오류수정의 수작업을 용이하게 할 수 있다. 즉 자료 입력 시 에러코드가 발생했을 때 발생된 에러코드에 따라 어떤 변수를 수정해야 최소의 항목으로 수정되는지를 참고할 수 있다.

그러나 실제 자료분석 결과에서도 알 수 있듯이 수정해야할 변수가

유일하게 결정되지 않을 수 있다. 또한 내검규칙이 항목 간 연관관계에서 필수적인 조건식이 아니면 수용하기 어렵다는 단점이 있다. 현재의 내검규칙은 조사표 내검을 위해 설정된 선택적인(soft) 내검규칙으로서 자동내검을 위한 필수적인(hard) 내검규칙에 이를 적용하는 것은 다소 무리인 부분이 있다. 그러나 선택적인 내검규칙의 조건들을 더욱 관대하게 재설정함으로써 필수적인 내검규칙으로 간주할 수 있고, 이로써 자동내검기법을 적용할 수 있을 것이다. 한 가지 간과해서는 안 될 것은 내검규칙의 설정이 매우 중요하며 내검규칙의 설정을 위해 이상치분석 등 선행 연구가 필요하다는 것이다.

본 연구에서는 명시된 내검규칙으로부터 파생되는 내검규칙을 유도하고 이들을 이용해 위배된 내검규칙의 변수를 공통적으로 커버하는 변수를 찾아 이용하는 단순한 방법을 적용하였지만, 추후 선형계획법을 이용한 더욱 정교한 오류위치포착의 방법론 연구가 진행될 필요가 있다.

현재로서는 사업체조사의 최종자료에 오류자료나 무응답항목이 적어 지금 당장 자동 에디팅 시스템을 적용하거나 도입 효과를 바로 기대하기는 어려울 것이다. 이는 향후 조사환경의 변화에 대비하여 장기적인 안목에서 연구를 진행할 필요가 있을 것이다. 내검의 자동화를 위해서는 내검규칙 분석, 이상치 검출, 수정해야 할 변수위치 포착, 응답대체 절차 등의 관련부문별 기초연구가 필요하며 궁극적으로 각종 사업체 대상 조사의 자료에 일반적으로 확대·적용할 수 있는 일반화된 자동 에디팅 시스템을 목표로 해야 할 것이다.

이를 위하여 우선 시스템개발기반을 위한 실무적용 중심의 관련부문 개별연구를 진행한 후, 이 기초연구와 성과를 토대로 자동내검·대체 시스템 개발을 추진하는 것이 효율적이라 판단된다.

첫 단계는 자동내검 관련부문별 연구 및 적용 등 개발기반 준비단계로 다음과 같은 연구 및 적용을 생각할 수 있다.

- 내검규칙의 모순방지, 중복 방지, 숨겨진 등식 식별, 내재적 규칙 생성 등 내검규칙의 분석방법에 관한 연구 및 적용
- 오류자료의 결과분석방법 연구 및 적용
- 응답자료로 볼 수 없는 자료, 즉 대체되어야 할 극단 이상치 검출 및

응답대체 시 사용될 수 없는 자료의 검출방법 연구 및 적용

- 모든 수량적 내검규칙을 만족하도록 수정변수위치를 자동으로 결정하는 방법연구 및 적용
- 오류항목이나 무응답항목의 자동 수정 방법 연구 및 적용
- 연구용 국외 소프트웨어의 도입, 적용 및 그 유용성 연구

두 번째 단계는 이상과 같은 연구결과를 종합하고 이를 기반으로 자동 에디팅 시스템을 개발하는 단계이다. 마지막으로 오류자료나 항목무응답이 많은 사업체대상 조사 자료에 시범 적용하고, 발생하는 문제점 및 해결방안 연구를 수행한 후, 다양한 사업체조사에 확대·적용하는 단계별 추진이 바람직하다고 판단된다.

내검이 재조사의 성격을 띠는 것은 효율적인 조사라 할 수 없을 것이다. 첫 조사에서 철저하게 이루어지고 그 입력과정이 투명하다면 오히려 그 실상을 확실히 파악할 수 있을 것이다. 점차 재접촉에 대한 불만, 응답자와 조사원의 부담 등 조사환경이 어려워지고, 행정자료의 활용 등 통계조사는 새로운 변화에 직면하고 있다. 통계청의 선진화를 위해서는 노동력 대비 효율을 극대화하고 자동화의 범위를 넓히는 것도 선진화 방편의 하나일 것이다.

참고문헌

이의규와 심규호(2007), “사업체대상 조사의 자동내검기법”, 「국가통계 발전을 위한 통계기법의 개선」, 통계개발원.

통계청(2004a), 「2003년 기준 산업총조사 입력·내검 프로그램 운영 요령서」, 내부자료.

통계청(2004b), “2003년 기준 산업총조사 중간보고 및 종합내용검토 계획”, 내부자료.

Banff Support Team(2007), “Functional Description of the Banff System for Edit and Imputation”, Generalized System Methods Section, Business Survey Methods Division.

Chernikova, N.V.(1964), “Algorithm for Finding a General Formula for the Nonnegative Solutions of a System of Linear Equations”, U.S.S.R. Computational Mathematics and Mathematical Physics 4, 151-158.

Chernikova, N.V.(1965), “Algorithm for Finding a General Formula for the Nonnegative Solutions of a System of Linear Inequalities”, U.S.S.R. Computational Mathematics and Mathematical Physics 5, 228-233.

De Waal, T. and W. Coutinho(2005), “Automatic Editing for Business Surveys: An Assessment of Selected Algorithms”, *International Statistical Review*, 73, 1, pp.73-102.

De Wall, T.(2003), “Solving the Error Localization Problem by Means of Vertex Generation”, *Survey Methodology*, Vol. 29, No. 1, 71-79, Statistics Canada.

De Wall, T.(2003), “Processing of Erroneous and Unsafe Data”, Ph. D. Thesis, Erasmus University Rotterdam.

Fellegi, I.P. and D. Holt(1976), “A Systematic Approach to Automatic Edit and Imputation”, *Journal of American Statistical Association*, 71, pp.17-35.

Giles, P.(1988), “A Model for Generalized Edit and Imputation of Survey Data”, *The Canadian Journal of Statistics*, Vol. 16, 57-73.

Granquist, L.(1997), “The New View on Editing”, *International Statistical*

Review, 65, 3, pp.381-387.

Greenberg, B.(1986), “The Use of Implied Edits and Set Covering in Automated Data Editing”, Bureau of the Census, Statistical Research Division Report Series SRD Research Report Number: Census/SRD/RR-86/02.

Kozak, R.(2005), “The Banff System for Automated Editing and Imputation”, Proceedings of the Survey Methods Section, SSC Annual Meeting.

Nordholt, E.S. and T. De Waal(1999), "Automatic Editing in the Dutch Labour Cost Survey Using CherryPi", UN Statistical Commission and Economic Commission for Europe, Working Paper No.7.

Rubin, D.S.,(1975), “Vertex Generation and Cadinality Constrained Linear Programs”, *Operations Research*, 23, 555-565.

Salazar-Gonzalez, J.J.(2006), “TEIDE: A New Software for Data Editing”, SDE, UNECE.

Schiopu-Kratina, I. and J.G. Kovar(1989), “Use of Chernikova's algorithm in the Generalized Edit and Imputation System”, Statistics Canada, Methodological Branch Working Paper No. BSMD-89-001E.

Weng, S.S.(2002), “Elimination in Linear Editing and Error Localization”, Section on Survey Research Methods, Joint Statistical Meeting.

Winkler, W.E. and B. Chen(2001), “Extending the Fellegi-Holt Model of Statistical Data Editing”, *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9.

Winkler, W.E. and L.R. Draper(1997), "The SPEER Edit System", *Statistical Data Editing*, *Volume Ⅱ*, UN Economic Commission for Europe, pp.51-55.