제4장

사업체대상 조사의 자동내검기법

이의규 · 심규호

제1절 서 론

국가통계기관은 경제 · 사회현상에 대한 자료의 수집 및 공표가 주요한 업무 중의 하나이다. 따라서 국가통계기관에서는 정확한 자료의 수집이 무엇보다도 요구된다. 또한 조사 자료의 사용자는 그 자료가 정확하고, 완전하고, 항목간 논리적 · 수리적으로 어떤 문제가 없기를 기대한다. 그러나 정확한 자료의 수집을 위한 많은 노력에도 불구하고 현실적으로 수집 및 입력과정에서 여러 가지 요인에 의해 자료의 오류가 발생하게 된다. 이러한 오류는 정확한 자료를 제공해야 하는 기관의 신뢰성에 대한 문제일 뿐만 아니라 이용자의 통계분석에도 많은 문제를 일으키게 된다. 그러므로 조사 자료의 수집·입력과정에서 자료를 검토하고 교정하는 자료편집(data editing)[1]의 절차는 반드시 필요하다.

그런데 자료편집은 자료를 엄밀히 검토·교정하는 섬세한 과정으로서 많은 인력과 시간을 필요로 한다. 특히 자료의 양이 방대한 조사의 경우, 편집에 드는 비용과 시간은 전체 조사비용의 큰 부분을 차지하게 된다. Granquist(1997)의 연구에 의하면, 자료편집을 위한 하드웨어, 소프트웨어, 인건비 및 교육비 등이 조사 예산의 20~40%를 차지한다고 보고하고 있다. 또한 자료편집에 소요되는 막대한 비용이 통계품질 개선이라

1) 통계청에서는 이를 내검(내용검토)이라 지칭한다.

는 명분으로 정당화될 수 없다고 한다. 더 많이 더 세세하게 검토하고 재접촉하는 것이 품질을 더 좋게 한다는 구 패러다임에서 전체조사의 계속적인 개선을 위한 근거를 제공하기 위하여 오류자료를 찾아 모아서 오류의 원인분석에 초점을 맞추어야 한다고 주장한다. 더욱이 재접촉에 의한 응답부담 가중, 공표 지연에 따른 시의성 상실과 관련된 비용을 고려한다면 편집 분야의 새로운 시각이 필요하다고 강조하고 있다.

한편 유럽과 미국, 캐나다 등지에서는 자료의 오류를 찾아내고 수정하는 작업을 부분적으로 자동화하고 있으며 이를 통해 효율적인 편집을 수행하고 있다. 더욱이 유엔유럽경제위원회(UNECE)에서는 통계자료편집(Statistical Data Editing: SDE) 세션에서 각국의 사례와 연구를 매년 소개하고 있으며, 기법연구나 시스템개발이 활발히 진행되고 있다. 반면에 국내는 편집기법 연구기반이 매우 미약하며 사례연구가 거의 없는 상황이다. 편집 분야는 다른 나라와 마찬가지로 국가통계기관인 통계청에서 주요 절차로 인식되고 있으며, 청에서는 입력 내검 프로그램에 의해 입력자료편집이 자동화되어 있으나 자동오류포착 및 자동수정 분야로의 연구가 점진적으로 필요하다고 판단된다.

이에 따라 본 논문에서는 자동자료편집에 대하여 살펴보고자 한다. 자료의 오류위치 자동포착 및 자동수정의 근거가 되는 Fellegi-Holt 방법의 원리와 절차를 연구하고 자동편집시스템의 유용성을 검토하고자 한다. 다음 절에서는 자료 편집의 개요를 살펴볼 것이다. 제3절에서는 국내외 사례현황을 살펴보고 이후 Fellegi-Holt의 기법을 간략하게 소개하고자 한다. 그리고 사업체 대상 조사에 대하여 이 기법의 유용성을 검토하고 본 연구의 결론 및 시사점을 제시한다.

제2절 자료편집의 개요

1. 자료편집

흔히 자료편집(data editing)이라 함은 자료 수집 및 처리단계에서 오류를 찾아내고 이를 수정하는 과정을 말한다. 오류의 원인은 응답자가

질문을 이해하지 못한 경우, 응답자나 조사자가 잘못된 응답에 체크한 경우, 조사자가 응답을 잘못 코딩하거나 잘못 이해한 경우, 응답자가 부정확한 응답을 제공한 경우에 일어날 수 있다. 이때 편집을 어떻게 수행할 것인가는 조사담당자의 경험에 의해 주어지는 편집규칙에 따라 주로 이루어진다. 편집유형은 데이터의 유효성, 중복성, 논리적 일치성, 과거 자료와의 일관성, 분석상의 이상치 유무 등으로 볼 수 있다(캐나다 통계청 홈페이지).

또한 편집은 작업방식에 따라 전통적 방법인 수작업으로, 컴퓨터 프로그래밍에 의하여 자동으로, 또는 수·자동 동시작업으로 수행될 수 있으며 편집단계에 따라 다음과 같이 구분할 수 있다.

- 미시 편집 (micro-editing)
  개별 레코드의 검사를 통하여 자료의 오류를 찾아낸다. 자료가 논리에 맞는지 확인하고 교정하는 것이 목적이다. 이는 입력단계에서 수행되므로 입력자료편집(input data editing)이라고도 한다.
- 거시 편집 (macro-editing)
  전체 자료의 분석을 통하여 문제가 없는지를 확인한다. 이는 출력자료편집(output data editing)이라 하기도 한다.

자료편집은 국내논문에서 데이터편집으로 소개된 바 있다(박진우, 2005). 반면, 통계청에서는 이와 같은 과정을 내검(내용검토)이라 하여 입력단계에서 수작업이나 입력내검 프로그램으로 오류자료를 찾아내고 수정하고 있다. 또한 출력단계에서 총체적으로 문제가 없는지 수행·검토되고 있다. 본 연구에서는 내검과 편집을 같은 의미로 간주하고 혼용하여 사용하기로 한다.

2. 자동편집

전통적인 자료편집은 조사원이 상세하게 모든 자료를 검토하여 오류가 포착되면 조사양식을 다시 참고하거나 응답자를 접촉함으로써 오류를 수정하는 것이다. 자동편집(automatic editing)은 방대한 자료의 편집에 드는 시간과 경비를 줄이기 위해 편집과정에서 조사원들이 수작업으

로 수행하는 단계를 컴퓨터 프로그래밍을 이용하여 자동으로 행하는 것을 말한다.

자료편집을 자동화하기 위해서는 오류위치를 찾는 단계와 수정하는 단계로 나눈다(Winkler & Chen, 2001). 오류위치포착 단계에서는 레코드가 편집규칙에 일치하는지 여부를 판단한다. 만약 일치하지 않는 레코드라면 이제 어떤 변수에서 오류가 있는지를 식별해야 한다. 즉, 오류포착단계는 어떤 변수가 오류인지를 결정하고, 수정단계는 이러한 변수의 실제 값을 결정한다. 본 연구에서는 주로 자동오류포착에 초점을 맞추어 논의하고자 한다. 4절에서 자동오류포착에 대한 방법으로서 수학적 최적문제해결에 기초한 Fellegi-Holt의 알고리즘을 소개한다.

제3절 국내외 사례 및 현황

1. 국내 사례 및 현황

자료편집(data editing) 분야의 현 국내연구는 매우 미약한 상황이다. 일례로 데이터편집을 다룬 연구로는 “주택가격동향조사를 위한 데이터편집 사례연구”(박진우, 2005)를 들 수 있을 정도이다. 이 논문은 자료편집의 사례연구 발표로서 자료편집에 대한 사례를 공론화하였다는 데 큰 의미가 있다.

이 사례논문에서는 자료의 수집, 입력, 처리 등의 단계에서 발생할 수 있는 오류를 방지하기 위한 국민은행의 주택가격동향 자료편집시스템을 소개하고 있다. 조사목적에 맞도록 편집규칙을 정하는 과정과 온라인 입력에서 응답실수를 방지하기 위한 입력시스템 개발의 사례를 보여주고 있다. 특히 Hidiroglou & Berthelot(1986)의 편집규칙을 사용하고 있다. 이는 t시점과 $t+1$시점의 조사가격의 비가 일정범위 안에 드는지의 여부에 따라 재조사를 실시하는데, 이 범위를 결정하는 것은 재조사의 표본수가 전체표본의 5%가 되도록 하고 있으며, 입력시 이상치가 입력되었을 때는 오류메시지를 보낸다거나 입력거부 또는 사유를 기입하여야만 입력이 가능하도록 한다는 것이다.

자료편집 분야는 다른 나라와 마찬가지로 대체로 대규모의 조사가 반복적으로 이루어지는 국가통계기관에서 더욱 필요로 하게 된다. 국내 관련전문가들은 대규모 조사에서 자료편집 분야는 절대적으로 필요하며, 보다 활발한 연구와 국내 자료편집사례 공론화의 필요성을 강조하고 있다.

2. 해외 사례 및 현황

해외에서는 국내와는 달리 편집 분야의 연구 활동이 매우 활발하다. 특히 유엔유럽경제위원회(United Nations Economic Commission for Europe: UNECE)에서는 매년 통계자료편집(Statistical Data Editing: SDE) 세션에서 자국의 경험사례를 공론화하고 있으며, 편집기법 관련 연구를 발표하고 있다. 각국의 자동편집 및 수정 시스템 현황을 개괄적으로 살펴보면 다음과 같다.

가. 네덜란드

네덜란드 통계청은 2000년 3월부터 2003년 2월까지 영국, 핀란드, 스위스, 이탈리아, 덴마크 등과 함께 자료편집 연구개발 프로젝트에 참여하였다(Pannekoek & De Wall, 2003). 이 프로젝트의 주목적은 현재 사용중인 자료편집방법을 평가하고 새로운 방법을 개발하는 것이다. 두 개의 데이터 세트, 영국의 연간경기조사(UK Annual Business Inquiry), 스위스의 환경보호지출조사(Swiss Environmental Protection Expenditures)를 가지고 타 기관의 자료편집과 비교·평가를 하였다. 프로젝트의 결과로서 네덜란드 통계청은 다음과 같은 편집절차를 제시하였다.

- 숫자 단위 착오와 같은 명백한 오류들을 수정
- 레코드의 중요성에 따라 선택적 편집(selective editing)을 적용
- 중요 레코드는 대화식 편집, 중요도가 낮은 레코드는 자동편집 및 대체
- 편집된 총체자료를 거시적으로 검토

네덜란드 통계청은 현재 Blaise라고 불리는 편집시스템을 사용하고 있다. 이 시스템은 미국과 캐나다가 지원하고 네덜란드에서 개발한 CHERRYPI(Automatic Editing and Imputation System)와 SLICE(Statistical Localization, Imputation and Correction of Errors)를 통합하는 조사시스템이다. 가구조사(Household Surveys), 경기 및 경제조사(Business and Economic Surveys), 건강조사(Health Surveys), 노동인력조사(Labor Force Surveys), 에너지/환경/농업조사(Energy/Environment/Agriculture Surveys)와 같은 다양한 분야의 조사에 사용되고 있다. 또한 국제 Blaise 사용자 모임(International Blaise Users Group: IBUG)이 매년 개최되어 세계 각국의 Blaise 시스템 사용자들 간에 많은 의견 교환이 이루어지고 있다. 앞에서 언급한 대로 CHERRYPI는 Blaise 시스템 중의 한 부분으로 Fellegi-Holt 방법을 기초로 한다. 이 시스템은 4년마다 이루어지는 경제조사인 네덜란드 노동비용조사(Dutch Labour Cost Survey)에 활용하였다(Nordholt and De Waal, 1999).

나. 영국

영국통계청(Office for National Statistics)에서는 자동편집(automatic editing)을 도입함으로써 수동편집(manual editing)에 할당된 자원을 줄이고 수동편집의 남는 자원을 다른 조사 프로세스에 재배치하는 것을 목적으로 하는 연구를 하였다(Ceri, 2001). 이 연구에서는 기존의 내용검사 시스템을 재검토하고 자동편집방법을 소개하였다. 특히 조사 결과에 중요한 차이를 가져올 것이라고 추측되는 이상값만을 편집하는 선택적 편집(selective editing) 방법을 소개하였다. 유통서비스부문 월간조사(Monthly Inquiry into Distribution and Services Sector: MIDSS)에 선택적 편집이 처음으로 도입되었다. 2001년 8월에 MIDSS에서 구현된 선택적 편집은 또 다른 월간임금및보수조사(Monthly Wages and Salary Survey)와 연간경기조사(Annual Business Inquiry), 월간생산조사(Monthly Production Inquiry) 등에 적용되었다.

또한 인구센서스조사청(Office of Population Censuses and Surveys: OPCS)에서는 1981년 센서스에 자동편집기법을 사용하였다. Brant &

Chalk(1985)는 1981년 센서스에 사용된 자동편집 시스템을 설명하고 이 시스템의 검사 결과를 요약하였다. 결과는 고용상태만 제외한 모든 항목에 대해 정확한 결과를 준 것으로 보고하고 있다.

다. 미국

미국은 SPEER(Structured Programs for Economic Editing and Referrals) 시스템을 1984년에 개발하였다. SPEER 시스템은 연속형 경제 데이터의 편집을 위해 설계되었다. 이 시스템은 자동편집에 대한 방법론으로 Fellegi-Holt 모델을 적용하였다. SPEER는 두 개의 프로그램으로 이루어져 있는데, 첫 번째 프로그램은 데이터의 편집 한계값들을 생성한다. 두 번째 프로그램은 오류위치포착(error localization)을 수행하고 대체(imputation)를 수행한다. 이 시스템은 연간제조업조사(Annual Survey of Manufactures)와 제조업 및 광업 센서스(Census of Manufactures and Mineral Industry)에 적용되었다(Winkler & Draper, 1997).

라. 캐나다

1985년에 캐나다 통계청에서는 모든 경제조사를 재설계하고자 Business Survey Redesign Project(BSRP)를 수행하였다. 프로젝트의 목적은 경제조사에 적합한 일반화된 소프트웨어의 개발이다. 프로젝트의 결과로 개발된 시스템이 일반화된 편집수정시스템(Generalized Edit and Imputation System: GEIS)이다(Whitridge & Kovar, 1990). GEIS는 Fellegi-Holt 기법에 기반을 두고 있으며, 편집분석(edit analysis)과 오류위치포착(error localization), 대체(imputation)의 세 부분으로 이루어져 있다. 1989년에 '1991년도 농업총조사(1991 Census of Agriculture)'의 데이터 편집과 대체를 적용하기 위한 GEIS를 개발하기 위해 연구팀이 구성되었다. 1986년에 진행된 실제 데이터를 적용한 연구 결과가 매우 긍정적이어서 1991년도 농업총조사에서부터 GEIS 시스템이 적용되었다.

캐나다에서는 다양한 경제조사에서 소득세 자료 목록이 표본 틀로 사용된다. 이전에는 소득세 자료에 대한 대체(imputation)가 이루어지지 않았으나, 소규모 비즈니스 조사를 가능하게 하기 위해 1988년 회계연

도 소득세 자료에 GEIS를 이용한 대체방법을 적용하였는데, 1990년에 소득세 자료에 대한 편집과 대체를 완료하였다. 또한 1990년에는 연간 자동차화물운송조사(Annual Motor Carrier Freight Survey)에도 GEIS를 적용하여 편집 및 대체를 시행하였다.

제4절 Fellegi-Holt 기법

컴퓨터의 출현은 수작업 검사로 해 왔던 오류검색을 자동화하는 큰 변화를 가져왔다. 이와 때를 같이하여 Fellegi & Holt(1976)는 자동편집 문제를 이론적으로 체계화하였으며, 이후 각국에서 데이터편집 시스템 개발 및 연구 발표가 활발히 진행되고 있다. 본 절에서는 자동편집의 근거가 되는 Fellegi-Holt 기법에 대해 간단한 예제를 통하여 그 개념을 살펴보고자 한다. 이 이론의 핵심을 먼저 말하자면, 하나의 레코드(자료)는 가능한 한 최소 항목(필드, 변수)을 수정함으로써 모든 편집규칙을 만족하게 해야 한다는 것이다. 자료의 정보를 수정하는 것은 매우 치명적일 수 있으므로 가능한 한 정보를 보존해야 한다는 원칙에 따른 것이다. Fellegi-Holt의 가장 큰 특징은 오류자료에 대한 수정할 값을 결정할 때 모든 변수가 동시에 고려된다는 것이다(Greenberg, 1986). 특히 주어진 편집규칙으로부터 유도된 내재적 편집규칙(implied edits, implicit edits)이 오류자료의 변경할 변수들을 결정할 때 주요한 역할을 한다.

1. 범주형 자료의 예

조사 자료에 오류가 있는지를 판단하기 위해서는 어떤 편집규칙이 있어야 한다. 이는 흔히 조사 담당자의 경험에 의해 설정되는데 자료의 형태가 범주형 자료(categorical data)인 경우에는 논리적 편집규칙(logical edits)에 의해 그 오류 여부를 판단하게 된다. 다음과 같이 5개의 변수(항목, 필드)와 그에 따른 가능한 코드가 <표 4-1>과 같이 주어졌다고 가정하자(Fellegi & Holt, 1976).

그리고 <표 4-2>의 편집규칙이 조사담당자에 의해 선정되었다 하자.

예를 들면, 편집규칙 E1은 성별이 남자이면서 가족관계가 아내로 되어 있으면 오류로 판단한다는 것을 의미하며 나머지도 같은 방식으로 해석된다. 그런데 여기서 편집규칙 E4는 E2와 E3가 원래 주어지고 난 후 내재적으로 얻어지는 편집규칙이다. 즉, 조사담당자는 E2와 E3를 편집규칙으로 부여하였으나 이로부터 E4의 편집규칙이 유도된 것이다. 이를 내재적 편집규칙(implicit edit)라 한다. 실제로 이 내재적 편집규칙이 오류포착을 위한 주요한 역할을 하게 된다. 이는 만약 하나의 레코드에서 결혼상태가 결측치이고, 나이가 8살, 가족관계가 배우자(아내나 남편)라면 오류 자료로 판단되어야 하지만, E4의 규칙이 없이는 오류를 포착할 수 없게 된다는 점에서 쉽게 알 수 있다. 주어진 편집규칙에서 내재적 편집규칙을 유도하는 과정은 여기서 생략하고 이후에 다시 언급하기로 한다.

〈표 4-1〉 예제: 5개의 변수와 가능한 코드

	성별		나이			결혼상태					가족관계				교육			
			0~14	15~16	17+													

〈표 4-2〉 편집규칙

규칙	성별	나이	결혼상태	가족관계	교육
E1	남자	*	*	아내	*
E2	*	0~14	결혼경험 있음	*	*
E3	*	*	기혼이 아님	배우자	*
E4	*	0~14	*	배우자	*
E5	*	0~16	*	*	대학 이상

주: *는 어떤 코드도 관계없음을 의미함.

이제 하나의 레코드가 <표 4-3>과 같이 코딩이 되었다고 하자. 즉, 레코드는 남자, 0~14, 이혼, 아내, 대학 이상을 나타낸다. 이제 이 표를 이진수로 나타내면 오류를 찾는 알고리즘을 효율적으로 실행할 수 있다. 각 변수에서 해당코드는 1로 하고 나머지는 0으로 표시한다. 또한 <표 4-2>의 편집규칙을 같은 방식으로 표현하면 <표 4-4>의 편집규칙 행렬을 갖는다.

〈표 4-3〉 예제 레코드

	성별		나이			결혼상태					가족관계				교육			
			0~14	15~16	17+													

〈표 4-4〉 편집규칙 행렬

	성별		나이			결혼상태					가족관계				교육			
레코드	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1
E1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1
E2	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
E3	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1
E4	1	1	1	0	0	1	1	1	1	1	1	1	0	0	1	1	1	1
E5	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1

주어진 레코드의 코드가 1이 표시된 코드위치에 각 편집규칙에서 모두 1이 표시되면 그 편집규칙을 위배한다는 것이다. 즉, 현재 레코드의 코드벡터와 각 편집규칙의 코드벡터의 스칼라 곱(scalar product)이 변수의 수와 같으면 해당 편집규칙을 위배한다는 것이다. 따라서 예제의 레코드는 모든 편집규칙을 위배하는 오류자료로 판명된다. 또한 명시적

편집규칙 E2와 E3를 곱하면(두 규칙에서 중복되어 나타나는 결혼상태는 두 개의 코드 중 1이 있으면 1) 내재적 편집규칙 E4를 얻게 된다. 이를 해석하면, 나이가 1~14세이면서 배우자로 코딩되었다면 오류로 판단함을 확인할 수 있다.

이제 이 레코드가 모든 편집규칙을 만족하게 하고자 한다면 하나의 변수만을 바꾸어서는 성립할 수 없다. 따라서 모든 편집규칙을 만족하는 최소개의 변수를 찾는 것이 문제이다. 결론적으로는 나이와 가족관계를 바꾸어 줄 때 모든 편집규칙을 만족하게 하는 최소개의 변수 조합이 된다. 이는 아래 <표 4-5>의 행렬에서 나이와 가족관계의 변수가 위배된 편집규칙에 가장 많이 들어가 있는 것으로부터 짐작할 수 있다.

〈표 4-5〉 위배된 편집규칙 행렬

	성별	나이	결혼상태	가족관계	교육
E1	1	0	0	1	0
E2	0	1	1	0	0
E3	0	0	1	1	0
E4	0	1	0	1	0
E5	0	1	0	0	1

이러한 알고리즘은 If-Then-Else의 구조보다 효율적이고 편집규칙의 수정 또는 변경시 그 관리가 용이하다(Chen et al., 2002). 또한 명시적 편집규칙에서 쉽게 내재적 편집규칙을 유도할 수 있어 유용하다.

2. 연속형 자료의 예

자산이나 비용과 같은 양적 자료가 나타나게 되는 사업체 조사의 경우에는 계량산술적 편집규칙(quantitative arithmetic edits)이 주어지게 된다. 다음과 같은 명시적 편집규칙(explicit edits)이 주어졌다고 하자(De Waal & Coutinho, 2005).

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

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

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

으로 두 번째 편집규칙 E2에 위배되어 이 레코드는 조사 담당자의 경험에 따라 오류자료로 판단한다. 이때 이 자료에 대하여 수작업이 아닌 자동으로 오류위치를 찾아내고 수정하고자 할 때 정보를 최대한 보존하면서 모든 편집규칙을 만족하기 위한 해답은 쉽게 보이지 않는다. 이를 위해 주어진 편집규칙으로부터 각 변수의 소거를 통해 다음과 같은 식을 얻을 수 있다.

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

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

$$E3: \quad 4 - 2(6) + 1 = -9$$
$$E4: \quad 3 - 6 + 2(1) = -1$$
$$E5: \quad 2(3) - 4 + 3(1) = 5$$

으로 주어진 레코드는 E3와 E4의 편집규칙을 만족하지 못하고 있음을 알 수 있다.

이제 우리는 각 위배된 편집규칙에 포함된 변수행렬 <표 4-6>으로부터, X_3가 위배된 모든 편집규칙에 포함되어 있어 이 변수 하나를 바꾸어 준다면 모든 편집규칙을 만족하게 됨을 볼 수 있다. 다시 말하자면,

명시된 편집규칙으로부터는 어떤 변수를 바꾸어 주어야 할지 명확하지 않으나, 이와 같이 내재적 편집규칙을 이용하면 자료의 오류위치를 효율적으로 판단할 수 있다. 더 나아가 X_3값을 변수로 놓고 나머지 주어진 값을 대입하여 다시 방정식을 풀면 $0 \leqq X_3 \leqq 5/3$ 일 때 모든 규칙을 만족하게 된다. 즉, $X_3 = 1$이 가능한 대체값이 될 수 있다.

〈표 4-6〉 위배된 편집규칙 행렬

	X_1	X_2	X_3	X_4
E2	1	1	1	
E3		1	1	1
E4	1		1	1

제5절 사업체대상 조사에 대한 F-H 기법의 유용성 검토

현재 사업체대상 조사에 대한 조사표내검은 입력된 조사표 자료에 대한 내검을 실시한다. 즉, 입력자료의 각 항목 또는 연관된 항목을 입력내검프로그램을 운용하여 검사한다. OK에러는 에러사유를 기재하면 내검사항에서 해제되는 에러이며, 필수에러는 나타나면 안 되는 에러이므로 필히 수정을 하는 에러이다. 이 밖에 전년대비를 통하여 이상치를 체크하고 있다. 이 절에서는 종사자 4인 이하 광업·제조업 사업체조사에 대해 F-H 기법의 적용을 논의하고자 한다.

1. 조사표와 내검사항

먼저 종사자 4인 이하 광업·제조업 사업체조사의 조사표를 살펴보기로 한다. 생산비, 출하액, 임가공수입액 및 자산에 관한 항목으로, 다음 소절의 계량적 연관성 검토를 위해 조사표의 일부를 제시한다(표 4-7 참조).

〈표 4-7〉 종사자 4인 이하 광업·제조업 사업체조사의 조사표(일부)

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 4-7>의 조사표에서 얻어진 자료는 <표 4-8>에서 보는 바와 같이 종사자 4인 이하 광업·제조업 사업체조사의 조사표 중 4항, 5항, 6항의 연관 편집규칙은 에러코드 EA와 EB로, 또한 5항, 6항, 7항과 관련된 편집규칙은 에러코드 RD로 내용검토를 하고 있다.

〈표 4-8〉 에러코드와 내검사항

에러코드	전산내검사항
EA EB	4항, 5항, 6항 - 조사표상의 수입부문의 합계는 비용부문의 합계보다 커야 함 【수입부문】 (A) 5항 합계, (B) 6항의 합계 【비용부문】 (C) 4항의 소계 (①+②+③) (A) + (B) < 1.2 × (C) (A) > 10 × (C)
RD	5항, 6항, 7항 - 5항 제품출하액 + 6항 임가공수입액 / 조업월수 × 12개월 < 0.1 × 유형자산 연말잔액

2. Fellegi-Holt 기법의 적용

우선 F-H 기법의 적용을 위해 각 해당 항목을 다음과 같이 간결하게 기호화하자.

- 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.

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

– 원재료비는 출하액의 10%보다는 커야 함.

$$X_1 \geq 10\ X_3$$

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

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

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

$$\begin{aligned}
&\text{E1}: \quad X_1 + X_2 - 1.2\,X_3 \geq 0\\
&\text{E2}: -X_1 + 10\,X_3 \geq 0\\
&\text{E3}: \quad X_1 + X_2 - 0.1\,X_4 \geq 0
\end{aligned}$$

이로부터 유도되는 내재적 편집규칙(implicit edits)은

$$\begin{aligned}
&\text{E4}: \quad -1.2\,X_3 + 0.1\,X_4 \geq 0\\
&\text{E5}: \quad X_2 + 10\,X_3 - 0.1\,X_4 \geq 0
\end{aligned}$$

이다. 즉, X_3와 X_4의 관계와 X_2, X_3, X_4의 변수간의 관계가 유도되었다.

따라서 앞 절의 예제에서 살펴본 바와 같은 방식으로 주어진 편집규칙으로부터 유도된 편집규칙을 이용하여 오류의 위치를 자동포착할 수 있을 것이다. 그러므로 내검 전 자료에 F-H 알고리즘을 적용하여 4인 이하 사업체 조사의 이와 같은 일부 수량적 연관항목에 수작업이 아닌 자동오류포착, 더 나아가 자동수정하는 방안을 검토하여 볼 수 있을 것이다.

제6절 결론 및 시사점

각 나라의 국가통계기관에서는 통계조사 작성 과정에서 어떤 형태로든지 자료의 편집절차를 거치고 있다. 그러나 주어진 시간과 예산의 제약 속에서 단기간내 최소한의 교정이 필요할 때에는 자동편집이 효과적인 한 방법이 될 수 있다. 여러 선진통계 국가에서는 신속하고 비용절감

적인 방법으로 자료의 오류를 찾아내고 수정하고 있다. 급속한 IT의 발전과 함께 최초 조사단계에서 데이터편집을 수행하며, 여러 향상된 편집기법을 연구하고 활용하여 자료의 오류를 찾아내고 수정하고 있다. 이처럼 편집 분야는 새로운 방향으로 나아가고 있음을 앞에서도 확인하였다. 이에 따라 우리나라도 조사환경 악화에 대비하고 효율적인 자료편집을 위하여 자동편집기법의 연구가 필요하다고 판단된다.

더욱이 Granquist(1997)는 비용과 투자가 조사 자료의 품질을 향상시키기보다는 응답부담 가중과 기회비용의 증가를 가져온다고 보고하고 있다. 철저한 내검을 한다고 해도 조사자가 생각지 못한 오류를 찾아낼 수는 없으며, 어떤 중요한 오류는 전통적인 오류검토로 식별될 수 없으며 오히려 내검 중에 다시 새로운 오류(자의적 대체 오류 등)가 나올 수 있고, 재접촉을 통해 응답자의 부담을 가중시킬 수 있다는 것이다. 무엇보다도 개선해야 할 조사표나 기획이 묻힐 수 있다는 단점을 지적하고 있다.

따라서 고품질 통계는 단지 검사를 많이 하고 재접촉 수를 늘리는 것만으로 확실히 보증될 수 없다. 엄격한 내검을 통해 자료를 정제하는 것 이상으로 중요한 것은 논리에 맞지 않는 항목에 대한 원인 분석 및 피드백을 통해 향후 오류를 방지할 수 있도록 조치하는 것이다. 즉, 응답이 이상하거나 응답하기 꺼리는 항목에 대한 피드백이 중요하다. 재접촉은 자료의 정제 차원에서보다도 어떤 항목에 어려움이 있어서 그러한 오류가 나타났는지 그 원인을 물어 보는 것이 필요하다. 또한 오류는 조사표에 기인한 것이 매우 큰바, 조사표 개선, 조사표 단순화, 조사방법 개선 등으로 이어져야 할 것이며 최초 조사 자료에서 수정된 표지를 남겨 후속연구에 사용될 수 있어야 한다.

그러나 우리나라 사업체대상 조사응답자의 응답환경 등은 선진국과는 상당한 차이가 있고 조사목적이나 조사환경이 다르기 때문에 이를 이해할 필요가 있다. 현 상황으로서는 자동수정보다는 입력내검프로그램을 통한 오류위치의 자동포착에 중점을 두어야 할 것으로 판단된다. 특히 F-H 오류포착의 알고리즘 등을 통한 입력프로그램 개선을 도모할 수 있다. 장기적으로는 자동수정을 포함한 전반적인 자동내검시스템의 검토가 필요할 것으로 본다.

한편 무응답대체(imputation)는 논리적·계량적 규칙에 어긋나는 오류를 포착하는 것을 다루지 않고 있다는 것이 편집(editing)과 구별된다. 오류항목에 대한 대체는 여러 가지의 방법으로 대체할 수 있지만 편집규칙을 만족하도록 최소한의 대체를 강구한다. 특히 반복조사나 불일치 항목 자료가 많이 발생하게 되는 사업체대상 조사인 경우, Fellegi-Holt와 같은 편집기법을 부분적으로 적용해 볼 수 있을 것이다. 또한 매우 큰 비중을 갖는 자료인 경우에는 재조사와 같이 정확 조사를 하고, 그렇지 않은 경우는 편집조건에 위배되지 않도록 선택적으로 자동 대체하는 것도 하나의 방법이 될 수 있다. 사업체대상 조사에서 소사업체와 같이 영향력이 작은 자료는 자동수정으로 하고 대사업체는 매뉴얼수정을 생각해 볼 수 있다.

편집의 자동화는 조사비용 및 시간을 줄일 수 있을 뿐만 아니라 응답부담을 경감시킬 수 있을 것이며, 합리적 수정논리에 의하여 교정된 조사 자료는 통계분석의 적합성을 향상시킬 수 있을 것이다. 따라서 통계자료 제공시 일관된 수정 및 수정논리를 확보할 수 있다는 차원에서나 향후 조사환경 악화에 대비하여 이러한 편집기법에 대한 연구가 요구되며, 국내 편집분야의 공론화를 통한 편집기법 연구가 활성화되어야 할 것이다. 또한 국내의 이론적 연구나 자동편집시스템 개발의 기반이 미약하므로 선진기법을 단기간 내에 습득할 수 있도록 자동편집기법 연수 프로그램 참여나, 관련전문가와 공동으로 국내 편집분야를 활성화시킬 수 있도록 공동연구를 추진하는 것도 하나의 방안이 될 수 있을 것이다.

참고문헌

박진우(2005), “주택가격동향조사를 위한 데이터편집 사례연구”, 「조사연구」, 제6권 제1호.

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

Brant, J.D. and S.M. Chalk(1985), "The Use of Automatic Editing in the 1981 Census", *Journal of Royal Statistical Society*, 148, pp.126-146.

Chen, B., Y. Thibaudeau, and W.E. Winkler(2002), "A Comparison Study of ACS If-Then-Else, NIM, and DISCRETE Edit and Imputation Systems Using ACS Data", *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

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.

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

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.

Hidiroglou, M.A. and J.M. Berthelot(1986), “Statistical Editing and Imputation for Periodic Business Surveys”, *Survey Methodology*, 12, pp.73-84.

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.

Pannekoek, J. and T. De Waal(2003), "Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the Euredit Project", Discussion Paper 03011, Statistics Netherlands, Voorburg.

Underwood, C.(2001), "Implementing Selective Editing in a Monthly Business Survey", *Economic Trends*, 577, pp.41-45.

Whitridge, P. and J. Kovar(1990), "Applications of the Generalized Edit and Imputation System at Statistics Canada", Statistics Canada.

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.