제2장

자동내검기법의 적용방안
-서비스업조사를 대상으로-

이의규

제1절 서 론

조사 자료는 응답자의 단위착오나 부정확한 응답 등으로 오류를 포함하곤 한다. 예를 들면, 항목간 연관성이 불일치하는 오류가 존재할 때가 있다. 이 경우 조사표를 재확인하거나 재접촉을 통해 해결하는 것이 전통적인 방법이다. 그러나 실제 조사업무에서 한정된 시간 내에 많은 자료를 처리하는 것은 그리 간단하지 않다.

한편 응답되어야할 항목에 응답이 누락된 경우, 우리는 대체기법을 적용하여 적절한 값으로 대체하곤 한다. 그런데 만약 응답은 되어 있지만 무응답으로 여겨질 정도의 항목간 모순된 자료가 존재한다면, 이러한 값의 처리방안 또한 필요하다. 즉, 비록 응답된 항목일지라도 응답의 의미가 없다면 위험을 최소화하는 범위 내에서 합리적인 방법으로 수정하여 제공할 필요가 있다.

현재 통계청에서는 항목 간 연관규칙을 부여하고 이에 어긋날 경우를 전산으로 검토한다. 즉 오류발생시 오류코드들을 화면에 출력하고 이에 따라 내검(editing) 요원은 조사표의 항목들을 확인하거나 응답자를 재접촉하여 처리한다. 따라서 현재의 전산내검은 어떤 항목과 어떤 항목 사이에 오류가 발생하였음을 일괄적으로 알려 주는 것이며 이후의 조치는 내검요원에게 맡겨진다.

반면, 자동내검(automatic editing)이란 규칙에 어긋난 자료를 찾고 수정되어야 할 항목을 결정하며 더 나아가 모든 점검규칙을 만족하는

값으로 자동 수정하는 절차를 말한다. 이러한 자동내검은 크게 두 단계의 절차로 이루어지는데, 첫 단계는 수정 변수를 선택하는 오류위치포착(error localization)의 단계이고, 두 번째 단계는 수정값을 결정하는 수정 또는 대체(correction or imputation)의 단계이다. 따라서 자동내검(automatic editing)은 어떤 항목간에 오류가 발생했다는 것 이외에 어떤 항목을 어떻게 고쳐줄지를 자동으로 알려 준다는 점에서 전산내검과 차별된다.

Granquist(1997)는 내검(editing)이 컴퓨터의 발전과 더불어 점차적으로 사람의 힘에 덜 의존하는 쪽으로 발전하고 있음을 피력하였다. 이를 증명하듯이 국외에서는 이미 사업체 대상 조사의 자동 에디팅 시스템이 개발되어 사용되고 있다. 1984년에 개발된 미국의 SPEER(Winkler 등, 1997), 1985년부터 착수하여 개발된 캐나다의 GEIS(Whitridge등, 1990), 네덜란드의 CherryPi(Nordholt등, 1999)가 대표적이다. 특히 캐나다의 GEIS는 최근 Banff(Kozak, 2005)로 발전되어 사용되고 있다.

국내에서는 그동안 에디팅이 반복적인 일로 치부되거나 수학통계적인 기법 적용의 필요성이 많지 않은 분야로 인식되어왔으며 그 결과, 통계자료의 내검(에디팅)과 관련한 국내 논문은 주택가격동향조사를 위한 데이터 편집 사례연구(박진우등, 2005)외 극소수였다. 그러나 최근에 들어서는 그 관심이 고조되고 있다. 특히 2007년 통계의 날 기념 워크숍에서는 에디팅 관련 사례들을 소개한 바 있으며(변종석등, 2007), 에디팅 품질관리 매뉴얼(김규성, 2008)이 작성된 바 있다.

이와 같은 배경에서 이의규 등(2007, 2008)은 자료의 자동 오류위치 포착 및 수정의 근거가 되는 Fellegi와 Holt(1976)의 자동내검 원리를 광업·제조업조사 자료에 적용하고 그 결과와 문제점을 분석한 바 있다. 본 연구에서는 F-H의 대안으로서 좀 더 간편한 자동 에디팅 방법을 서비스업조사에 적용한다.

본 보고서의 구성은 다음과 같다. 먼저 다음 절에서는 서비스업조사의 개요를 살펴본다. 특히 현재의 내검절차와 내검규칙에 대해 중점을 두어 검토한다. 3절에서는 수학적 최적화 기반의 F-H 기법을 간단하게 리뷰하고 절대차이 합을 최소화하는 선형계획법을 대안으로 제시한다. 4절에서는 이 선형계획법을 이용한 자동에디팅기법을 서비스업 조사에

적용하고 그 결과를 평가한다. 결론에서는 이러한 시도의 문제점과 향후 연구 방향을 제시한다.

제2절 서비스업조사의 개요와 내검절차

1. 서비스업조사의 개요

서비스업은 국내경제에서 차지하는 비중이 점점 커지고 있는 중요한 부문이다. 서비스업을 조사하는 목적은 서비스업에 대한 구조변화와 경영실태를 파악하여 서비스산업관련 정책수립, 국내총생산(GDP)의 추계, 기업의 경영계획, 학술연구 등을 위한 기초자료를 제공하고자 함이다. 서비스업 조사는 1988년에 최초 실시한 것으로 기록되고 있으며 2008년에 실시한 2007년 기준 조사는 18회 조사이다. 특히 17회 조사부터 서비스업조사 대상 전 업종에 대해 산업세세분류(시・도는 세분류) 단위까지 세분화하기 위해 표본규모를 확대한 바 있다. 이 서비스업조사는 통계법 제 17조 및 제 18조에 의해 지정된 지정통계이다(승인번호 제 10127호). 조사대상은 한국표준산업분류(KSIC)상 8개(E, J, L, N, P, Q, R, S) 산업대분류에 해당하는 사업체 중 표본으로 선정된 사업체이다 (<표 2-1> 및 <표 2-2> 참조).

<표 2-1> 서비스업조사의 대상

산업대분류	주요 사업
	하수・폐기물처리 원료재생 및 환경복원업
E	출판 영상 방송통신 및 정보서비스업 통신업 제외
J	부동산업 및 임대업 (61)
L	사업시설관리 및 사업지원 서비스업
N	교육서비스업 일반교습학원 및 기타 교육기관 등만 조사
P	보건업 및 사회복지 서비스업 (855~857)
Q	예술 스포츠 및 여가관련 서비스업
R	협회 및 단체 수리 및 기타 개인서비스업 협회 및 단체 제외
S	, (94)

〈표 2-2〉 서비스업조사의 개요

구 분	내 용
조사목적	- 정부 및 지방자치단체의 서비스산업 육성을 위한 정책자료 - 국내총생산(GDP) 및 지역소득(GRDP) 추계 산업연관표 작성 등 서비스산업 주요 경제지표 편제를 위한 기초자료 - 지역별 경제지표 개발 및 확충에 필요한 기초자료 - 기업의 경영계획 수립 및 대학·연구소 등의 연구자료 - 산업구조통계의 국제간 비교자료
법적근거	- 통계법 제17조 및 제18조에 의한 지정통계 제10127호
조사시기	- 조사기준일 : 조사기준년 12. 31. - 조사대상기간 : 조사기준년 연간 - 조사실시기간 : 매년 5~6월 경 1약 25일간 실시
조사대상	- 한국표준산업분류(KSIC) 상 E, J, L, N, P, Q, R, S 대분류에 해당하는 사업체 중 표본으로 선정된 사업체
조사항목	- 기본항목(10) : 개 사업체명 및 소재지 사업내용 조직형태 일평균 영업시간 사업체 정기 휴무일수 연간영업개월 수, 월평균 종사자 수 및 연간급여액 사업체 건물 연면적 전자상거래 활용현황 사업실적 등 - 특성항목(4) : 개 직능별 종사자 수 전산장비 보유대수 무형자산 보유건수 이용인원 고객 수 등
조사방법	- 지방통계청 사무소의 조사담당직원과 임시조사원이 직접 조사대상사업체를 방문하여 면접조사 함을 원칙으로 함(3) - 면접이 곤란한 경우 응답자 직접기입방식이나 인터넷 조사방식을 취함
결과공표	- 보고서 : 매년 말 서비스업통계 조사보고서 권 발간(3) - 국가통계포털(http://www.kosis.kr) 에 수록

한편 1993년 제6회 조사까지는 총사업체통계조사 결과를 이용하여 표본을 추출하였으나 1995년 조사부터는 사업체기초통계조사 결과를 이용하여 표본을 추출하고 있다. 산업세세분류별 및 시도별로 모집단 업종수가 10개 이하인 업종은 전수조사하며 사업체수가 많은 다른 업종은 종사자 규

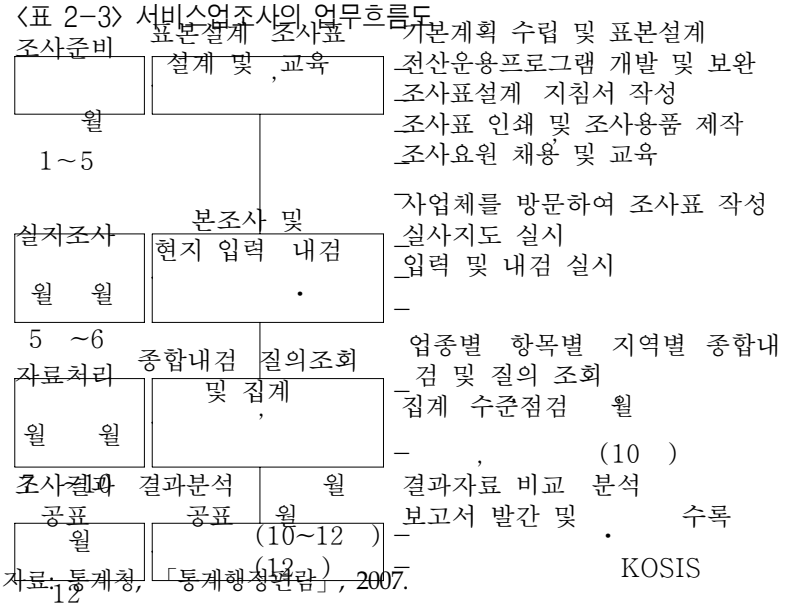
모별, 산업세세분류, 시·도 단위로 표본사업체를 추출하여 조사한다.

2. 서비스업조사의 내검절차

서비스업 조사의 자료처리방법에는 자료처리 단계별 소요인력과 작업처리 순서 그리고 내검 및 오류정정 절차가 나와 있다. 지방청에서는 조사담당자가 1~2일의 교차내검과 조사표 전산입력 후 전산내검을 실시하고 있으며 본청에서는 종합내검(3개월), 지방청 질의조회, 수준분석(2개월)을 실시하여 자료내검 및 오류정정을 하고 있다. 그리고 지방청 1,250명의 입력내검요원과 본청 종합내검요원 30명, 업종별 수준분석요원 6명의 인력이 이를 담당하고 있다[1](<표 2-3> 참조).

서비스업조사의 조사항목은 기본항목 10개로 구성된다. 사업체명 및 소재지, 사업내용, 조직형태, 일일평균 영업시간, 사업체 정기 휴무일수, 연간영업개월 수, 월평균 종사자 수 및 연간급여액, 사업체 건물 연면적, 전자상거래 활용현황, 사업실적이다.

〈표 2-3〉 서비스업조사의 업무흐름도



자료: 통계청, 「통계행정편람」, 2007.

내검인원 및 담당인원은 도소매업조사를 포함한 것임

1) .

이 중 금액과 관련된 조사항목은 월평균 종사자수 및 연간급여액, 매출액과 영업비용을 포함한 사업실적이다. 월평균 종사자 수 및 연간급여액항목은 월평균종사자수(자영업주+무급가족종사자+상용종사자+임시・일용종사자+무급종사자)로 조사하며 연간급여액(상용종사자연간급여액+임시・일용종사자연간급여액)이 조사된다. 사업실적은 매출액과 영업비용[재료매입비+인건비+임차료(+세금과공과+감가상각비+대손상각비+광고선전비+지급수수료)+수도광열비+기타영업비용] 그리고 매출액에서 영업비용을 뺀 영업이익으로 구성된다.

이와 같은 조사표에서 얻어진 자료는 현지 입력・내검시 오류사항을 전산으로 검토하고 있다. 이때 점검코드는 필수적으로 만족해야만 입력이 가능한 필수코드와 점검이 필요한 선택코드로 나뉜다. 오류는 범위초과 및 범위오류, 누락, 합계불일치, 항목간 불일치 오류 등으로 구분할 수 있다. 여기서 금액에 관계된 오류내용만을 발췌하면 <표 2-4>와 같다(<부록 1> 참조).

〈표 2-4〉 오류코드와 오류내용(일부)

오류코드	오류내용	확인여부
F0620	상용종사자 월평균 급여액 5십만원 이하 또는 5백만원 이상 확인(상용종사자 월평균 급여액=(상용연간급여액/상용종사자)/연간영업개월수)	검토
M0702	매출액이 1조원을 초과함	검토
M0704	매출액이 재료매입비보다 작음	검토
M0705	매출액이 인건비보다 작음	검토
M0707	연간급여액 합계가 인건비보다 큼	필수
M0711 ~ M0789	종사자당 월평균 매출액 (매출액/종사자수합계/연간영업개월수)이 과소 또는 과다 (산업분류별로 과소, 과다 한계값이 다름)	필수

자료: 통계청, 「전산내검요령서」, 2005

<표 2-4>를 살펴보면 상용종사자 급여액이 상용종사자수에 비해 매우 작거나 큰 경우, 매출액이 재료매입비나 인건비보다 작은 경우는 검토하여 이상이 없으면 통과시키나 급여액이 인건비보다 크거나 종사자당 월평균 매출액이 과소, 과대한 경우에는 반드시 해결되어야만 입력되도록 하고 있다.

제3절 F-H 기법의 리뷰와 대안

1. F-H기법의 리뷰

De Waal과 Coutinho(2005)는 자동 오류위치포착을 위한 방법으로 다음 세 가지를 제시하였다. 이상치 검색 기법(outlier detection techniques)에 의한 방법과 신경망(neural networks)에 의한 방법, 그리고 수학적 최적화 방법이다. 이 중 F-H 기법은 수학적 최적화에 기초한 대표적 방법이다. F-H는 컴퓨터의 발전과 더불어 자동에디팅 문제를 이론적으로 체계화하였으며 이후 각국에서 이를 이용한 자동에디팅 시스템 개발 및 연구 발표가 활발히 진행되어 왔다.

F-H 방법은 조사 자료에 오류가 있는지를 판단하기 위해 조사 담당자에 의해 설정되는 내검규칙(edits)을 필요로 한다. 자료의 형태가 범주형 자료(categorical data)인 경우에는 논리적 내검규칙(logical edits)을 부여하고 연속형 자료(continuous data)인 경우에는 산술적인 내검규칙(arithmetic edits)을 선정하여 오류 여부를 판단한다. 만약 하나의 레코드가 모든 항목에 기입은 되어 있지만 설정된 내검규칙을 위반하면 한 개 또는 그 이상의 항목은 부정확한 것으로 식별되어야 한다. 그런데 어떤 값이 부정확하고 대체되어야 하는지의 결정은 그리 쉽지 않다.

이때 어떤 변수를 대체해야 할지를 결정하는 자동화 전략이 필요한데 그것이 F-H 전략이다. 이는 주어진 정보를 최대한 보존하면서 모든 내검규칙을 만족하게 하는 최소 개의 수정할 변수를 찾아내자는 것이다. 자료의 정보를 수정하는 것은 매우 치명적일 수 있으므로 가능한 정보를 보존해야 한다는 원칙에 따른 것이다.

오류위치포착에 대한 이해를 돕기 위해 다음과 같은 2개의 명시적 내검규칙(explicit edits)이 주어졌다고 하자(각 변수는 음이 아닌 수).

$$E_1: \quad X_1 - X_2 \geq 0$$
$$E_2: \quad X_2 - 3X_3 \geq 0$$

이제 하나의 레코드가 (6, 4, 8)로 코딩되었다고 하자. 따라서 이 레코드는 두 번째 규칙을 위반한 레코드이다. 문제는 이때 어떤 필드(항목)를 수정하여야 최대한 정보를 유지하면서 모든 규칙을 만족하게 할 수 있는가이다. 이 레코드에서 X_2만을 바꾸어서는 성립이 안 된다. 역시 X_1만을 바꾸어서는 모든 내검규칙을 만족할 수 없다. 그러나 X_2를 1로 바꾼다면 모두 만족한다. 물론 두 개 이상의 변수를 모두 바꾸어서 X_3 성립이 가능할 수 있으나 최대한 자료를 보존한다는 원칙에서 X_2 하나만을 바꾸는 것이 합리적이라는 것이다. 이와 같은 결론은 주어진 내검규칙 E_1과 E_2로부터 변수 X_2의 소거를 통해 다음과 같은 식을 구함으로써 도출될 수 있다.

$$E_3: \quad X_1 - 3X_3 \geq 0$$

위의 E_3를 내재적 내검규칙(implicit edits)이라 한다. 위 식에 다시 레코드의 값을 각 규칙에 대입하면 주어진 레코드는 E_1와 E_3의 내검규칙을 만족하지 못하고 있음을 알 수 있다. 전체 위배된 내검규칙은 E_1, E_2, E_3가 된다. 그런데 E_2는 위배된 모든 내검규칙에 포함되어 있음을 볼 수 있다. 즉 명시된 내검규칙으로부터는 어떤 변수를 바꾸어 주어야 할 지가 명확하지 않으나 이처럼 추가된 내검규칙을 이용하면 자료의 오류위치를 효율적으로 판단할 수 있다. 더 나아가 X_3 값을 미지수로 놓고 나머지 주어진 값을 조건식에 대입하여 풀면 $X_3 \geq 4/3$일 때 모든 규칙을 만족하게 된다. 즉, X_3이 가능한 대체값이 될 수 있다. 이때 $X_3 = 4/3$이다.

이 F-H 방법의 가장 큰 특징은 오류자료의 수정할 항목을 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. 이 알고리즘은 일반적인 If-Then-Else의 구조보다 효율적이고 편집규칙의 수정 또는 변경 시 그 관리가 용이하다(Chen등, 2002). 또한 변수의 신뢰성 가중치를 부여할 수도 있어 여전히 유효한 방법으로 보고되고 있다(De Waal과 Coutinho, 2005).

그러나 설정된 모든 내검규칙을 필수적으로 만족시켜야 하는 규칙(hard edits)으로 간주한다는 것과 오류를 우연적 오류로 국한한다는 것이 단점으로 지적된다. 특히 요구되는 내재적 내검규칙 수가 매우 많을 수 있으며 이때 모든 내재적 규칙의 생성에 있어서 많은 시간이 소요될 수 있다는 것이다. 이러한 단점을 극복하기 위해 변수의 비(ratio) 규칙을 이용하는 방법이 제안된 바 있다. 이 방법은 내재적 내검규칙의 수가 선형규칙에 비해 줄어들고 속도는 빨라지나 변수가 모두 비음(non-negative) 인 경우에 국한한다(De Waal과 Coutinho, 2005).

2. F-H의 대안

많은 오류를 포함한 레코드에서 오류위치포착의 최적해를 찾고자 하는 알고리즘은 일반적으로 많은 계산시간과 기억용량을 필요로 한다. 따라서 오류위치포착에 대한 소프트웨어를 개발하는 것은 어려운 문제이다. 특히 Chernikova의 알고리즘을 구현하는 것은 더욱 어렵다.

Ton De Waal은 Survey Methodology에 오류위치포착과 Chernikova 알고리즘에 관한 논문, Solving the Error Localization Problem by Means of Vertex Generation(De Waal, 2003)을 발표한 바 있다. 저자는 알고리즘을 프로그램화하는 것이 여전히 어려운 문제라 지적하고 있다. F-H 패러다임을 응용한 매우 기초적인 것도 프로그램을 작성하는 데 몇 개월이 소요된다. 만약 내검규칙의 설정, 내검 분석, 이상치 탐색, 대체(imputation)에 대한 기능을 추가하고자 한다면 더욱 많은 시간을 필요로 할 것이며 단순한 예제 프로그램이 아니라 통계자료의 생산에서 실제 사용될 수 있는 소프트웨어를 개발하고자 한다면 엄청난 시간을 필요로 할 것이다. 여기서는 오류위치포착 문제의 해를 얻는 더 간단하고 빠른 접근방법을 소개한다.

F-H 패러다임을 프로그램화하는 간단한 대안은 원 관측값과 에디팅

된 값과의 절대 차이값의 합을 최소화하는 소프트웨어를 개발하는 것이다(De Waal, 2003). 즉,

$$\text{Minimize} \quad \sum_{i=1}^{n} | X_{raw,\,i} - X_{edit,\,i} |$$

$X_{raw,i}$는 알려진 원 관측값이고 $X_{edit,i}$는 미지의 에디팅된 값을 나타낸다. 이 목적함수는 $X_{edit,i}$가 내검규칙을 만족하는 제약식 조건하에서 최소화되어야 한다. 이 문제는 선형계획법 문제와 같이 표현될 수 있다. 선형계획법의 해를 구하는 문제는 여러 가지 소프트웨어를 이용하면 해결될 수 있으나 여기서는 무료로 다운로드받을 수 있는 프로그램 R을 이용하여 해결하였다.

간단한 예를 들어본다. 세 개의 변수 X_1, X_2, X_3와 다음과 같은 2개의 에디팅 규칙이 있다고 가정하자.

$$X_1 \geq X_2$$
$$2X_1 \geq X_3$$

이제 하나의 레코드가 X_1 = 5, X_2 = 30, X_3 = 80을 갖는다고 가정하자. 따라서 이 레코드는 두 개의 규칙을 모두 위배한다. 이때 오류위치포착과 수정을 위해 다음과 같은 함수를 고려한다.

$$\text{Minimize} \quad |X_1 - 5| + |X_2 - 30| + |X_3 - 80|$$

subject to

$$X_1 \geq X_2$$
$$2X_1 \geq X_3$$

앞서 언급한 바와 같이 위 식은 하나의 선형계획법(linear Programming) 문제이다. R 프로그램을 이용하면 X_1 = 40, X_2 = 30, X_3 = 80의 해를 얻는다. 오류위치포착(error localization)에 대한 문제의 해답은 첫 번째 변수 X_1의 값을 변화시키는 것이다. 다른 변수는 변화가 없는 반면 X_1은

은 5에서 40으로 크게 바뀌었기 때문이다. 그러나 반드시 선형계획법의 해로 나타난 40의 값을 사용할 필요는 없다. 이 경우 50을 더 적절한 값으로 볼 수도 있다. 왜냐하면 응답자의 단위 착오나 기입자의 입력오류일 가능성이 높기 때문이다.

한편 앞의 F-H 기법의 리뷰에서 제시한 예제를 살펴보자. 예제의 두 개의 규칙은 다음과 같았다.

$$X_1 - X_2 \geq 0$$
$$X_2 - 3X_3 \geq 0$$

앞에서와 같이 (6, 4, 8)의 레코드는 두 번째 규칙을 위배한다. 이때 선형계획법을 이용하여 풀면 (6, 4, 1.33)의 해를 갖는다. 즉, X_3 = 1.33 이 하나의 가능한 해가 되므로 앞과 동일한 결과를 얻는다.

결과에서 알 수 있듯이 원래의 레코드의 값들과 다른 최종값들에 대한 변수는 오류가 있는 것으로 간주된다. 이들 오류는 또한 적절한 대체방법에 의해 대체될 수 있다. 무응답 대체는 결측치를 추정하고 이 추정값을 대체하는 절차로 회귀대체법(regression imputation method)이나 도너대체법(donor imputation method)을 이용하여 대체할 수 있다. 그러나 흔히 무응답 대체는 에디팅 규칙을 고려하지 않은 채 추정하게 되어 대체 후 에디팅이 여전히 위배될 수 있다. 그러나 앞의 선형계획법은 모든 에디팅규칙을 만족하면서 목적함수를 최소화하기 때문에 오류로 간주된 변수들은 일관된 방법으로 대체될 수 있다. 또한 F-H는 소수 개의 오류를 갖는 레코드일 경우에만 보증할 수 있는 반면에 이 방법은 오류를 많이 포함한 레코드에 대해 효과적인 것으로 보고되고 있다.

정리하면, 연속형 자료인 경우 주어진 제약조건에서 다음과 같이 거리함수를 최소화하는 문제로 공식화할 수 있다.

$$\text{Minimize} \quad \sum_{i=1}^{n} w_i |X_i - \widetilde{X}_i|$$
$$\text{subject to}$$

$$E_i : a_{i,1}X_1 + a_{i,2}X_2 + ... + a_{i,n}X_n \geq b_i, \quad i = 1,2,...,m$$

$$X_j \geq 0, \quad j = 1,2,...,n$$

여기서 X_i는 알려진 원 관측값이고 X_n는 미지의 에디팅된 값을 나타낸다. 이 최소화 선형계획법의 문제는 심플렉스 알고리즘에 의해 그리 어렵지 않게 해를 구할 수 있다. 특히 각 항목의 신뢰도에 따라 가중치(w)를 부여하여 분석의 유연성을 높일 수 있다. 즉 어떤 항목이 다른 항목보다 정확하다고 판단된다면 더 큰 가중값을 부여할 수 있다.

제 4 절 자동내검기법의 적용

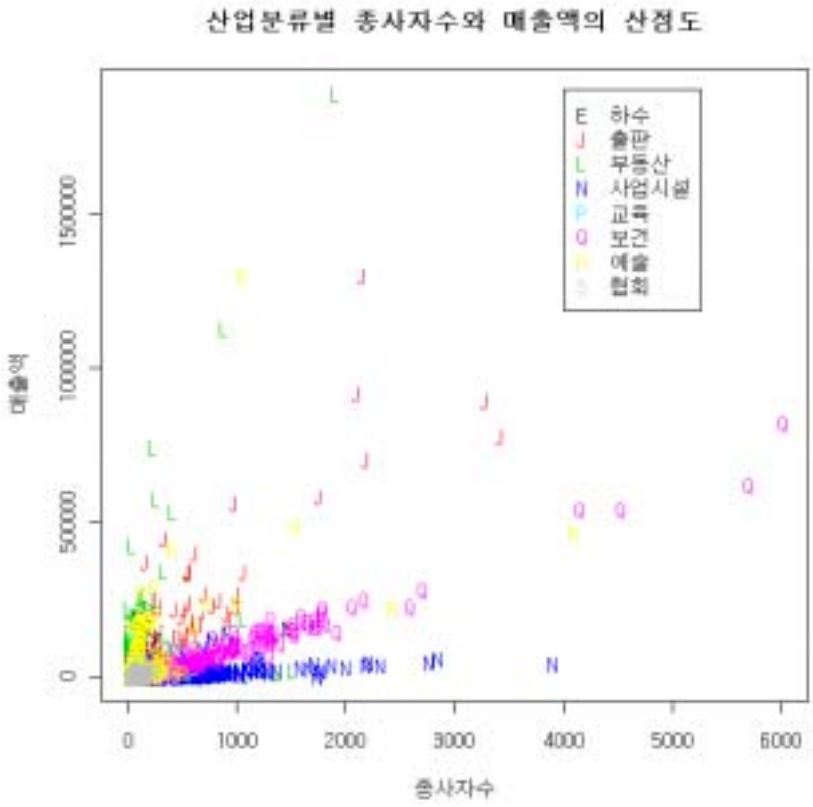
1. 항목 연관성 탐색

본 절에서는 그래프를 통해 항목 간 연관관계를 알아본다. 자료는 2007년 기준 서비스업통계 자료를 사용하였다. 한국표준산업분류(KSIC) 상의 대분류 E(하수・폐기물처리, 원료재생 및 환경복원업), J(출판・영상・방송통신 및 정보서비스업), L(부동산업 및 임대업), N(사업시설관리 및 사업지원 서비스업), P(교육서비스업), Q(보건업 및 사회복지서비스업), R(예술・스포츠 및 여가관련 서비스업), S(협회 및 단체, 수리 및 기타 개인 서비스업)에 해당하는 사업체 중에서 표본으로 선정된 약 50,000여개 사업체가 조사 대상이다. 분석결과 총 51,371건이 집계되었다. 다음 <표 2-5>는 산업분류별 조사범위를 나타낸다(<부록 2>참조).

<표 2-5> 산업분류별 조사범위

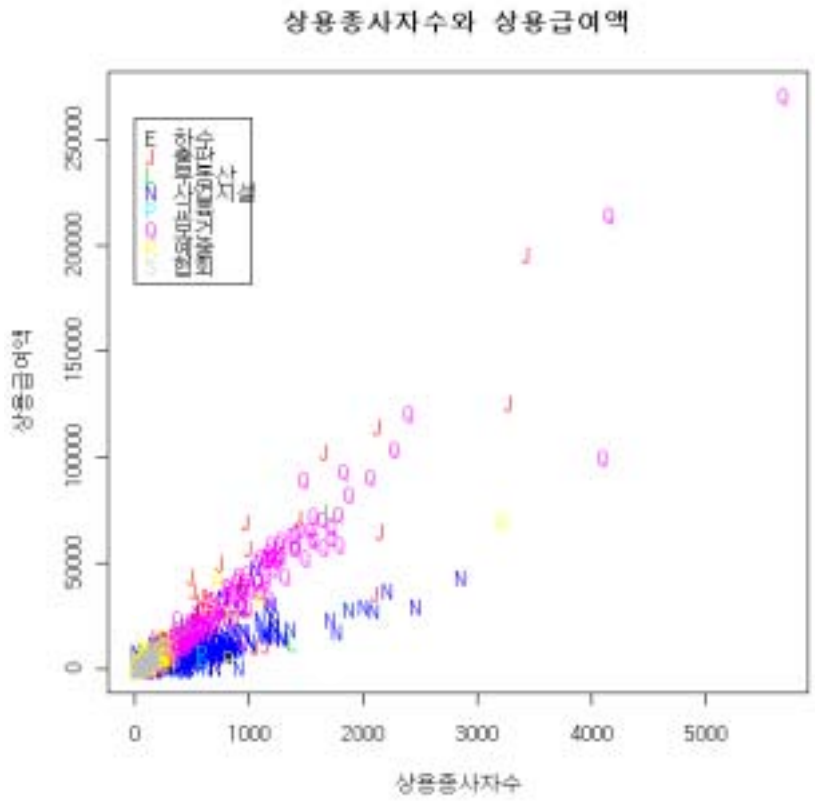
대분류	중분류	업종
E	37	하수・폐수 및 분뇨 처리업
	38	폐기물 수집운반 처리 및 원료재생업
	39	환경정화 및 복원업
J	58	출판・영상・방송통신 및 정보서비스업
	59	영상・오디오 기록물 제작 및 배급업
	60	방송업

	62	컴퓨터프로그래밍 시스템통합 및 관리업
	63	정보서비스업
L	68	부동산업
	69	임대업 부동산 제외
N	74	사업시설관리 및 조경 서비스업
	75	사업지원 서비스업
P	85	교육서비스업
Q	86	보건업
	87	사회복지 서비스업
R	90	창작 예술 및 여가관련 서비스업
	91	스포츠, 및 오락관련 서비스업
S	95	수리업
	96	기타 개인서비스업



[그림 2-1] 산업분류코드로 나타낸 종사자수와 매출액의 산점도

앞의 [그림 2-1]은 종사자수와 매출액을 산업분류별 문자로 구분하여 나타낸 산점도이다[2]. 예상한대로 업종별로 그 증가율이 달리 나타나고 있음을 알 수 있다. 부동산업 및 임대업(L)과 예술・스포츠 및 여가관련 서비스업(R)은 종사자수당 매출액 증가율이 가장 높으며 출판・영상・방송통신 및 정보서비스업(J)도 그 증가율이 높다. 반면 보건업 및 사회복지서비스업(Q)과 사업시설관리 및 사업지원 서비스업(N)은 상대적으로 매우 낮은 매출액 증가율을 보인다. 교육서비스업(P)과 협회 및 단체, 수리 및 기타 개인 서비스업(S)은 종사자수와 매출액이 작은 규모에 집중되어 있어 이 그래프로는 특성이 파악되지 않고 있다.



[그림 2-2] 산업분류코드로 나타낸 상용종사자수와 상용급여액의 산점도

영업월수가 1개월인 자료만 이용

2) 12

한편, 상용종사자수와 상용급여액의 산점도를 산업분류별 문자로 구분하여 나타내보면 앞의 [그림 2-2][3]와 같다. 역시 상용종사자수와 상용급여액간 양의 상관관계가 산업분류별로 달리 나타나고 있음을 그림을 통해 알 수 있다. 특히 출판・영상・방송통신 및 정보서비스업(J)이 가장 높으며 보건업 및 사회복지서비스업(Q)이 상대적으로 높다. 반면 사업시설관리 및 사업지원 서비스업(N)은 상대적으로 낮은 것으로 나타났다. 교육서비스업(P)과 협회 및 단체, 수리 및 기타 개인 서비스업(S)은 앞에서와 마찬가지로 상용종사자수와 연간급여액이 모두 작은 규모에 집중되어 있어 특성이 잘 드러나지 않고 있다.

상자그림은 자료가 대칭적일 때 자료의 몸통부분에서 멀리 떨어진 자료를 구분하고자 할 때 효과적이다. 그런데 일반적으로 경제자료는 자료의 속성상 우측으로 길게 늘어지는 분포를 갖게 된다. 이러한 분포에서 이상치를 탐색하는 상자그림은 적절하지 않다. 매우 큰 값들이 빈번히 존재할 때 로그값으로 변환하면 자료는 대칭이 되곤 한다. 따라서 로그변환 후 상자그림을 통해 이상치를 판단하는 것이 적절하다. 일반적으로 상자 길이의 아래 위 1.5배가 되는 가상선인 아래쪽 울타리와 위쪽 울타리로 이상치를 구분한다.

월평균상용종사자급여액(상용종사자급여액 합계를 상용종사자수로 나눈 후 이를 다시 영업월수로 나눈 값)의 로그 변환된 값의 상자그림을 그려보면, 아래쪽과 위쪽 울타리는 각각 -1.16, 1.92이다. 이 값을 원래의 값으로 변환하면 exp(-1.16)=0.313, exp(1.92)=6.82이다. 즉 월평균상용종사자급여액의 하한은 약 30만원, 상한은 약 6.8백만 원으로 볼 수 있다.

종사자당 월평균매출액(연간매출액을 종사자수로 나눈 후 이를 다시 영업월수로 나눈 값)의 로그 변환된 값의 상자그림에서는 아래쪽과 위쪽 울타리가 각각 -1.19, 3.06이다. 이 값을 원래의 값으로 돌리면 exp(-1.19)=0.30, exp(3.06)=21.33이다. 즉 종사자당 월평균매출액의 하한은 약 30만원, 상한은 약 21백만 원으로 볼 수 있다.

또한 종사자당 월평균급여액(급여액/종사자수합계/영업월수)의 상

영업월수가 개월인 자료만 이용

3) 12

자그림을 그려보면 상한은 3.15로 종사자당 월평균급여액이 약 3백만 원 이상은 드문 경우로 판단한다. 그러나 앞에서 언급하였듯이 경제자료는 속성상 큰 값을 흔히 갖게 되므로 큰 값들을 크기에 비례하여 축소하여 볼 필요가 있다. 그런데 종사자가 자영업주나 가족종사자 또는 무급종사자일 때는 급여액 합계가 0이다. 급여액 합계가 0인 자료가 상당히 많이 존재하므로 로그변환보다는 제곱근변환이 바람직하다. 제곱근으로 변환된 상자그림에서 상한은 2.8로 원래자료의 단위로 변환하면 2.8^2=7.84로 약 7.8백만 원이 된다.

이상과 같이 기존 내검규칙과 그래프를 통해 확인한 결과, 종사자수, 급여액, 매출액은 예상한 바와 같이 서로 연관성이 높음을 알 수 있다. 특히 종사자수는 급여액 및 매출액과 관련이 높다. 또한 종사자수와 매출액은 산업분류별로 다소 상이함을 보이고 있어 이러한 정보를 이용하면 좀 더 세밀한 자동내검이 가능할 것이다. 다음 절에서는 변수 간 관련규칙을 설정하고 앞에서 제안된 자동내검기법을 적용한다.

2. 선형계획법을 이용한 자동내검

이제 우리는 서비스업조사의 일부항목에 대해 다음과 같은 제약식을 구성할 수 있다.

1) 상용급여액 < 6.8×영업월수×상용종사자수
2) 상용급여액 > 0.3×영업월수×상용종사자수
3) 급여액 < 7.8×영업월수×종사자수
4) 급여액 > 0×영업월수×종사자수
5) 상용급여액 < 급여액
6) 매출액 < 21×영업월수×종사자수
7) 매출액 > 0.3×영업월수×종사자수
8) 매출액 > 인건비
9) 상용종사자수 < 종사자수
10) 급여액 < 인건비

여기서 상용종사자합계를 X_1, 상용종사자 연간급여액을 X_2, 종사자 합계를 X_3, 급여액을 X_4, 매출액을 X_5, 인건비를 X_6로 하자. 그러면 에디팅 규칙은 <표 2-6>과 같은 행렬로 표현할 수 있다. 이는 제약식 **AX** $\geq$ **0** 에서 행렬 **A**에 해당된다. 표에서 k는 해당 레코드의 영업월수이다. 즉 영업월수가 6개월이면 해당 레코드의 연간매출액이나 연간 상용종사자 급여액을 6으로 나누어 실제 월평균 값이 계산되도록 한 것이다.

〈표 2-6〉 점검규칙과 해당변수

변수 / 규칙	상용종사자 X_1	상용급여액 X_2	종사자 X_3	급여액 X_4	매출액 X_5	인건비 X_6
E1	6.8k	-1	0	0	0	0
E2	-0.3k	1	0	0	0	0
E3	0	0	7.8k	-1	0	0
E4	0	0	-0k	1	0	0
E5	0	-1	0	1	0	0
E6	0	0	21k	0	-1	0
E7	0	0	-0.3k	0	1	0
E8	0	0	0	0	1	-1
E9	-1	0	1	0	0	0
E10	0	0	0	-1	0	1

따라서 관측 자료에 근거한 이상치 판단기준을 고려하면, 선형계획은 다음과 같이 나타낼 수 있다.

$$\text{Minimize} \quad \sum_{i=1}^{6} w_i |x_i - \widetilde{x_i}|$$

제약조건식:

$$E_1 : 6.8kX_1 - X_2 > 0$$

$$E_2 : X_2 - 0.3kX_1 > 0$$

$$E_3 : 7.8kX_1 - X_2 > 0$$

$$E_4 : X_2 - 0kX_1 > 0$$

$$E_5 : X_4 - X_2 > 0$$

$$E_6 : 21kX_3 - X_5 > 0$$

$$E_7 : X_5 - 0.3kX_3 > 0$$

$$E_8 : X_5 - X_6 > 0$$

$$E_9 : X_3 - X_1 > 0$$

$$E_{10} : X_6 - X_4 > 0$$

선형계획의 문제해결을 위해 프로그램 R(http://www.r-project.org/)을 이용하였다. 여기서 상용종사자수와 종사자수는 다른 항목보다 신뢰도가 높은 것으로 파악되고 단위가 액수가 아닌 종사자 수이므로 신뢰가중치를 1000으로, 상용급여액, 급여액, 인건비는 50으로, 매출액은 1로 설정하였다.

3. 간단한 시뮬레이션

에디팅규칙을 만족하는 하나의 가상 레코드에서 하나 또는 두 개의 항목 값을 인위적으로 바꾸고, 작성된 프로그램을 실행한 후의 변화를 살펴보았다. 10명의 상용종사자, 3억 원의 상용급여액, 종사자수 13명, 급여액합계가 3억 3천만 원, 매출액 10억, 인건비 4억 원인 경우 앞의 선형계획법을 이용하여 작성된 프로그램 auto( )함수를 실행한 결과는 다음과 같다(이후 상용종사자수, 상용급여액, 종사자수, 급여액, 매출액, 인건비 순으로 기재). 실행결과, 이 레코드는 모든 내검규칙을 만족하기 때문에 실행 후에도 각 항목값에는 변화가 없다.

```
> auto(10, 300, 13, 330, 1000, 400)
  10 300 13 330 1000 400
```

이번에는 상용종사자수를 10에서 1000으로 입력하였을 때 실행결과이다. 상용종사자수 만을 가상적으로 크게 입력해 보았을 때 상용급여액 과소오류(E2)를 범하게 되며 또한 상용종사자수가 종사자수합계보다

큰 오류(E9)를 범하고 있다. 실행결과는 상용종사자수의 값을 급격하게 바꾸어 주고 있다. 이는 모든 점검규칙을 만족하기 위해서는 상용종사자수를 수정할 것을 알려 준다.

```
> auto(1000, 300, 13, 330, 1000, 400)
  상용급여액과소
  E2 필수  상용종사자수가 종사자수보다 큼
  E9[      ]
  13 300 13 330 1000 400
```

다음은 상용급여액만을 의도적으로 큰 값을 입력하였다 상용급여액과다 와 상용급여액이 급여액보다 큰 오류 를 나타낸다 실행 결과 상용급여액은 에서 으로 변화하여 상용급여액의 수정이 되어야 함을 보여주고 있다 (E1) (E7) 30000 330

```
> auto(10, 30000, 13, 330, 1000, 400)
  상용급여액과다
  E1 필수  상용급여액이 급여액보다 큼
  E7[      ]
  10 330 13 330 1000 400
```

급여액이 의도적으로 수정되었을 경우이다 급여액 과다오류 와 급여액이 인건비보다 큰 오류 를 갖게 되며 이때 급여액이 수정되어야 하며 구체적으로 인건비보다 (E3) (E10) 크지 않은 값으로 자동 수정된다

```
> auto(10, 300, 13, 33000, 1000, 400)
  급여액과다
  E3 필수  급여액이 인건비보다 큼
  E10[     ]
  10 300 13 400 1000 400
```

매출액에 인위적으로 비이상적인 값을 입력하였다 매출액 과다오류 가 발생하였으며 이에 따라 매출액이 변경되어야 함을 나타내고 있으며 (E5) 해당 종사자수에 비례한 상한 값을 제시하고 있다

> auto(10, 300, 13, 330, **100000**, 400)
E5 매출액과다

10 300 13 330 3284 400

다음은 인건비 오류입력일 경우이다 인건비가 에서 로 크게 변화하고 있어 이는 인건비를 수정하는 것이 적절하다는 것을 나타낸다 여기서 매출액 에서 은 변동이 없는 것으로 본다 .

40000 1001

. 1000 1001 .

> auto(10, 300, 13, 330, 1000, **40000**)
매출액이 인건비보다 작음
E8

10 300 13 330 1001 1001

이번에는 상용급여액과 급여액을 동시에 큰 값으로 입력하였다 역시 상용급여액과 급여액이 동시에 크게 변화한 값을 제시하고 있어 어들 항목이 수정되어야 함을 나타낸다

.

> auto(10, **3000**, 13, **33000**, 1000, 400)
상용급여액과다
E1 급여액과다
E3 필수 급여액이 인건비보다 큼
E10[ ]

10 398 13 400 1000 400

상용급여액과 매출액의 두 항목을 인위적으로 변화시킨 경우 세 가지의 오류가 발생하였다 상용급여액과다 매출액과다 그리고 상용급여액이 급여액보다 큰 오류이다 상용급여액과 매출액이 큰 변화를 보이고 있어 이들이 수정되어야 할 것이다 종사자수와 급여액이 다소 변화를 보이고 있으나 다른 두 개의 항목에 비해 상대적으로 매우 작다

.

> auto(10, **3000**, 13, 330, **100000**, 400)
상용급여액과다
E1 매출액과다
E5

E7[필수] 상용급여액이 급여액보다 큼

10 372 15 372 3745 400

상용급여액과 인건비를 함께 가상적으로 큰 값으로 입력한 경우이다 세 개의 오류를 범하고 있다 역시 상용급여액과 인건비가 매우 큰 변화를 보이고 있다

> auto(10, **3000**, 13, 330, 1000, **40000**)

E1 상용급여액과다

E7[필수] 상용급여액이 급여액보다 큼

E8 매출액이 인건비보다 작음

10 330 13 330 1001 1001

급여액과 인건비의 이상 입력인 경우 두 개의 점검규칙을 위배하고 있다 급여액과 인건비가 큰 변화를 보이고 있어 수정변수의 위치를 알려준다

> auto(10, 300, 13, **3300**, 1000, **40000**)

E3 급여액과다

E8 매출액이 인건비보다 작음

10 300 13 996 1002 1002

매출액과 인건비를 특이치로 입력한 경우에 내검기법을 적용하면 결과는 다음과 같다 두 개의 내검규칙을 위배한 것으로 나타났으며 매출액과 인건비가 큰 값으로 변경되어 있어 이 두 항목을 수정해야 한다

> auto(10, 300, 13, 330, **100000**, **40**)

E5 매출액과다

E10[필수] 급여액이 인건비보다 큼

10 300 14 330 3529 330

이상과 같이 가상자료를 이용하여 분석한 결과, 값이 크게 변화된 변수를 확인함으로써 수정되어야 할 변수를 결정할 수 있었다. 만약 이러한 결과가 실제로 내검 전 자료에 적용된다면 선택적이거나 필수적인 오류의 검색뿐 아니라 내검시 어떤 변수가 어떻게 수정되어야 할지를 참고할 수 있을 것이다.

4. 분석결과 및 평가

내재된 내검규칙의 생성은 수정하여야 할 변수를 선택하는 오류위치 포착의 단계에서 중요한 역할을 하고 있어 F-H에서는 반드시 필요하다. 그런데 앞에서도 언급한 바와 같이 이 내재된 내검규칙의 설정은 명시된 내검규칙이 조금만 증가하여도 매우 많아지고 이를 자동으로 결정하기에는 엄청난 시간이 필요하다는 문제를 가진다. 그러나 F-H와는 달리 이 선형계획법을 이용한 방법은 명시된 내검규칙으로부터 내재된 내검규칙을 생성할 필요가 없다.

선형계획법을 이용한 방법은 내재된 내검규칙의 생성을 필요로 하지 않을 뿐 아니라 수정절차가 동시에 이루어지고 있다. 더욱이 대체된 값은 내검규칙을 모두 만족하는 값이 되어 대체 후 내검을 달리 필요로 하지 않는다. 그러나 실제 기법을 적용한 결과, 보완되어야 할 부분이 필요하다. 즉 각 항목의 신뢰가중치를 설정하는 기준의 설정이나 더 구체적인 필수 점검규칙의 개발, 그리고 산업분류별 특성을 고려하는 방법 등 더 세밀한 작업을 필요로 한다.

한편 본 연구에서는 조사 자료로부터 각 내검규칙을 유도하였으나 실제로 좀 더 정확한 자료, 예를 들면 행정자료의 종사자수, 급여액, 매출액, 인건비의 관계에서 유도된 점검규칙을 사용하는 것을 생각해 볼 수 있다. 이는 실제 조사에서는 흔히 과소하게 조사되는 경향이 있으므로 그 경향을 알 수 있다면 한계값을 조정하여 수정값을 향상시킬 수 있을 것이다. 또한 산업분류별 종사자수와 매출액의 연관 정보를 이용하면 좀 더 정밀한 대체가 가능할 것이다. 예를 들면 오류 수정시 산업분류별 회귀(regression) 대체를 이용할 수 있을 것이다.

본 보고서에서는 자동으로 오류위치를 포착하고 내검규칙을 모두 만

족시키는 값으로 대체하는 자동내검기법을 시도하고 문제점을 분석하여 그 유용성을 제시하는 데 의미를 둔다. 실제로 단위착오로 인한 이상값이 들어온 경우 기대 효과가 클 것이다. 향후 실제자료를 적용하여 그 결과를 분석할 계획이다.

제5절 결 론

조사 자료의 수집을 국가통계기관의 제 1차적인 임무라 볼 때 통계자료의 품질관리는 매우 중요하다. 그렇기 때문에 통계작성기관에서는 최대한 과거자료나 연관성이 있는 다른 자료를 가지고 대조하는 자료내검을 수행한다. 그러나 지나치게 세밀한 내검은 오히려 시간적·경제적 어려움을 가중하게 될 뿐 아니라 철저한 내검에도 불구하고 오류가 여전히 존재할 수 있으며 근본적인 오류방지에 취약할 수 있다. 여러 선진국에서는 과도한 에디팅의 문제를 인식하고 에디팅(내검)과 임퓨테이션을 자동으로 수행하고 있어 효율적인 자료처리를 도모하고 있다.

선형계획법을 이용한 자동대체방법은 변수의 신뢰성, 변수간의 함수관계를 반영한 최소한의 응답대체라 할 수 있다. 이러한 자동오류위치 포착 및 수정은 현 조사환경에서 완전하게 수용하기는 어렵다하더라도 조사입력 및 내검에서 참조될 수 있다. 자동수정을 고려하지 않더라도 이러한 기법을 이용하면 수작업 내검시 도움이 될 것으로 판단한다. 특히 재접촉으로도 해결될 수 없는 모순된 자료에 대해 최후의 수단이 될 수 있다. 합리적이고 일관된 원칙을 가지고 이러한 자료를 수정한다면 그 수정논리의 정당성을 확보함과 동시에 수정 후에도 다시 원래대로 복원할 수 있다는 장점이 있다.

특히 예외가 인정되는 사업체의 경우는 내검에 걸리지 않도록 프로그램화하여 자동화할 수 있다. 또한 위험을 최소화하기 위해서는 좀 더 강한 내검규칙을 부여하고 이를 위배한 경우에만 자동 수정을 고려할 수 있을 것이다. 그러나 실제 분석결과에서 수정해야 할 변수가 정확하게 일치하지 않는 경우가 있다. 이는 이상치 탐색과정을 통한 내검규칙의 정확한 선정을 통해 개선될 수 있을 것으로 판단된다. 따라서 여전히

조사담당자의 경험에 의한 내검규칙의 설정은 매우 중요하다.

여러 통계 선진국의 조사환경은 우리와 차이가 있음이 분명하다. 그러나 자료의 품질을 강조하는 통계 선진국의 자동에디팅시스템은 무엇을 의미하는가. 자동에디팅시스템이 합리적이고 과학적이기 때문일 것이다. 어떤 구애됨 없이 재접촉을 하여 정확한 응답만 받을 수 있다면 그렇게 하는 것이 가장 좋을 것이다. 그러나 우리나라의 조사환경도 빠르게 변화하고 있으며 여러 가지 기법을 적용하여 자동화 할 수 있는 영역을 넓힐 필요가 있는 것 또한 분명하다.

본 연구에서는 내검규칙에 부합되는 수정값을 자동으로 제시할 수 있도록 프로그램을 작성하였으나 아직도 섬세한 자동내검을 위해서는 심도 있는 자동내검 방법론 연구가 진행될 필요가 있다. 자동내검 연구를 좀 더 깊이 있게 발전시키기 위해서는 이미 그동안 많은 시행착오를 겪었으며 오래된 경험을 축적한 미국, 캐나다, 네덜란드 등의 선진기법의 적극적인 수용도 자료처리의 선진화를 앞당길 수 있는 하나의 방안이 될 것이다.

참고문헌

김규성 (2008), 에디팅 품질관리 매뉴얼 한국통계학회.

박진우 박현주 김진억 (2005), 주택가격동향조사를 위한 데이터편집 사례연구 조사연구 권 호, , , 6 1 .

변종석 (2007), Introduction to Data Editing, Data Editing in Survey, 2007 통계의 날 기념 워크숍 한국조사연구학회

이의규와 심규호 (2007), 사업체대상 조사의 자동내검기법 통계개발원 “통계자료의 내검기법 연구」 통계개발원 .

통계청 (2008), 도소매업 및 서비스업통계조사 조사지침서 내부자료.

(2008), “ 년 기준 사업체기초통계조사 및 서비스업총조사 시범예행조사 전산내검 요령서」 내부자료 .

, .

Chen, B., Thibaudeau, Y., and Winkler, W. E.(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.

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

_____(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.

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.

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

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.

<부록 1> 서비스업조사의 오류코드와 오류내용

유형	오류코드	오류내용	확인여부
공통	F0104 (미사용)	사업체명부와 조사표의 산업분류가 조사대상범위(G,H,L,M,O,P,Q,R) 오류	필수
	F0105 (미사용)	표본번호 범위(1~3) 오류	필수
	F0201	사업체명부와 조사표의 사업체명이 다름	필수
	F0202	사업체명부와 조사표의 대표전화번호가 다름	필수
	F0203	대표자 성별범위(1~2) 오류	필수
	F0204	사업체명부와 조사표의 소재지가 다름	필수
	F0205	사업체명부와 조사표의 사업자등록번호가 다름	필수
	F0206	사업자등록번호 누락	검토
	F0207	사업자등록번호 자리수가 10이 아님	필수
	F0208	대표자 성별이 1인데, 자영업주 남+상용종사자 남=0인 경우	검토
	F0209	대표자 성별이 2인데, 자영업주 여+상용종사자 여=0인 경우	검토
	F0301	사업내용1 누락	필수
	F0302	사업내용1만 존재하는데 비중이 100%가 아님(필수)	필수
	F0303	취급상품1 누락	필수
	F0304	주요사업내용1, 비중1, 취급상품명1 입력항목 중 누락된 항목이 있음	필수
	F0305	주요사업내용2, 비율2, 취급상품명2 입력항목 중 누락된 항목이 있음	필수
	F0306	주요사업내용1이 100% 초과이거나 0이하임	필수
	F0307	주요사업내용2가 100% 초과이거나 0이하임	필수
	F0308	사업내용 비중1이 비중2보다 작거나 같음	필수
	F0309	주요사업내용 비율 합(=①+②)이 100% 초과임	필수
	F0310	주요사업내용1만 있는데 100% 아하임	필수
	F0401	조직형태 범위(①~④) 오류	필수
	F0402	① 개인사업체 또는 ② 비법인단체인데 재무제표상태 범위(1~2)오류	필수
	F0403	사업체명부와 조사표의 법인등록번호가 다름	검토
	F0404	③ 회사법인인데 법인등록번호가 누락	필수
	F0405	④ 회사이외법인인데 법인등록번호가 누락	검토
	F0406	① 개인사업체 또는 ② 비법인단체인데 법인등록번호가 입력됨	필수
	F0407	법인등록번호 자리수가 13이 아님	검토
	F0501	일일 평균 영업시간 범위(①~⑤) 오류	필수
	F0502	사업체 정기 휴무일수 범위(①~⑥) 오류	필수
	F0503	연간 영업개월 수 범위(1~12)오류	필수
	F0504	47122인 경우 일일 평균 영업시간 범위(①~④)로 표시된 오류	필수
	F0601	(1) 월 평균 종사자수 (가) 남자 합계(①자영업주 + … + ⑤ 무급종사자) 불일치	필수

F0602	(1) 월 평균 종사자수 (나) 여자 합계(① 자영업주 + … + ⑤ 무급종사자) 불일치	필수	
F0603	(1) 월 평균 종사자수 (다) 계 합계((① 자영업주 + … + ⑤ 무급종사자) 불일치	필수	
F0604	(1) 월 평균 종사자수 합계 (가) 남자 + (나) 여자 = (다) 계가 불일치	필수	
F0605	(1) 월 평균 종사자수 ① 자영업주 (다) 계{(가) 남자 + (나) 여자}가 불일치	필수	
F0606	(1) 월 평균 종사자수 ② 무급가족종사자 (다) 계 {(가) 남자 + (나) 여자}가 불일치	필수	
F0607	(1) 월 평균 종사자수 ③ 상용종사자 (다) 계 {(가) 남자 + (나) 여자}가 불일치	필수	
F0608	(1) 월 평균 종사자수 ④ 임시일용종사자 (다) 계 {(가) 남자 + (나) 여자}가 불일치	필수	
F0609	(1) 월 평균 종사자수 ⑤ 무급종사자 (다) 계 {(가) 남자 + (나) 여자}가 불일치	필수	
F0610	종사자수 누락	필수	
F0611	종사자수 1000명 이상 확인	검토	
F0612	(2) 연간급여액 합계(③상용종사자 + ④임시일용종사자) 불일치	필수	
F0613	(1) 월평균종사자수의 ③상용종사자수가 없는데 (2) 연간급여액 있음	필수	
F0614	(1) 월평균종사자수의 ③상용종사자수가 있는데 (2) 연간급여액 누락	필수	
F0615	(1) 월평균종사자수의 ④임시일용종사자수가 없는데 (2) 연간급여액 있음	검토	
F0616	(1) 월평균종사자수의 ④임시일용종사자수가 있는데 (2) 연간급여액 누락	검토	
F0617	3. 조직형태 ①개인사업체인데 7 월평균종사자 ①자영업주 없음	필수	
F0618	3. 조직형태가 ①개인사업체가 아닌데 7. 월평균종사자수 ①자영업주 또는 ②무급가족 종사자수가 있음	필수	
F0619	3. 조직형태가 ②비법인단체,③회사법인, ④회사이외법인인데 7. 월평균종사자수 ①자영업주 또는 ②무급가족종사자가 있음	필수	
F0620	7. 월평균 종사자 ③상용종사자 월평균 급여액(5십만 이하 또는 5백만원 이상) 확인[= (상용연간급여액/상용종사자)/영업개월수]	검토	
F0621	조사가능여부가 본사조사이고 7. 월평균종사자수 및 연간급여액에서 상용종사자와 임시일용 종사자가 있고 연간급여액이 없음	검토	
F0701	건물연면적 누락	필수	
F0702	(4) 건물연면적 합계{(1) 소유 + (2) 임차+(3)무상}가 불일치	필수	
F0801	활용 범위(①~⑤) 오류	필수	
F0803	1.(2) 홈페이지가 없는데 전자상거래 활용현황 (1) 활용에서 ①이 있음	검토	
F0901	응답자 성명 누락	필수	
F0902	응답자 전화번호 누락	필수	
F0903	조직형태가 ①또는 ②인데 응답자 부서 누락(확인)	검토	
F0904	조사방법란 범위(①~③)초과	필수	

조사표(3)	F0905	가방통계청검토자 누락	필수
	F0906	조사원 정보 누락	필수
	F0907	조직형태가 ③또는 ④인데 응답자 부서 누락(필수)	필수
	M0201	산업분류가 68, 69인데 종사자 1인당 건물연면적이 300㎡ 이상임	검토
	M0203	산업분류가 855, 856, 857인데 종사자 1인당 건물연면적이 550㎡ 이상임	검토
	M0204	산업분류가 86, 87인데 종사자 1인당 건물연면적이 600㎡ 이상임	검토
	M0205	산업분류가 58, 59, 60, 90, 91인데 종사자 1인당 건물연면적이 1,000㎡ 이상임	검토
	M0206	산업분류가 95, 96인데 종사자 1인당 건물연면적이 900㎡ 이상임	검토
	M0207	산업분류가 68, 69인데 종사자 1인당 건물연면적이 2.5㎡ 이하임	검토
	M0209	산업분류가 855, 856, 857인데 종사자 1인당 건물연면적이 5㎡ 이하임	검토
	M0210	산업분류가 86, 87인데 종사자 1인당 건물연면적이 6㎡ 이하임	검토
	M0211	산업분류가 58, 59, 60, 90, 91인데 종사자 1인당 건물연면적이 10㎡ 이하임	검토
	M0212	산업분류가 95, 96인데 종사자 1인당 건물연면적이 9㎡ 이하임	검토
	M0301	산업분류가 정보처리및컴퓨터관련업(582, 62, 63)인데 직능별종사자수가 누락됨	필수
	M0302	산업분류가 정보처리및컴퓨터관련업(582, 62, 63)이 아닌데 직능별종사자수가 기입됨	필수
	M0303	직능별 종사자수 ⑧합계(①연구원 + … + ⑦ 기타) 불일치	필수
	M0304	산업분류가 정보처리및컴퓨터관련업(582, 62, 63)인데 10. 직능별종사자수와 7-1. 월평균종사자합계가 불일치됨	필수
	M0401	산업분류가 정보처리및컴퓨터관련업(582, 62, 63)인데 11.전산장비 보유대수가 누락됨	검토
	M0402	산업분류가 정보처리및컴퓨터관련업(582, 62, 63)이 아닌데 11.전산장비 보유대수가 기입됨	필수
	M0403	전산장비 보유대수 합계(①대형서버 + … + ⑤ 기타) 불일치	필수
	M0501	무형자산 보유건수 합계(①산업재산권 + … + ④ 기타) 불일치	필수
	M0502	산업분류가 사업서비스업(582, 62, 63, 74, 75)이 아닌데 무형자산 보유건수가 입력됨	필수
	M0503	12. 무형자산보유건수 ④ 기타에 체크되어 있는데 ( )가 비어있음	필수
	M0601	산업분류가 90, 91인데 이용인원 해당유무가 누락됨	필수
	M0602	산업분류가 90, 91이 아닌데 이용인원 해당유무가 입력됨	필수
	M0603	산업분류가 90, 91이고 이용인원이 있음인데 월평균 또는 연간 이용인원이 입력 안됨	필수
	M0604	산업분류가 90, 91이고 이용인원이 없음인데 월평균 또는 연간 이용인원이 입력됨	필수
	M0605	서비스업의 이용인원(고객)수가 월평균 및 연간이용인원수 모두 입력됨	필수

M0606	산업분류가 90, 91이 아닌데 월평균 또는 연간 이용인원 수가 입력됨	필수
M0701	14-1, 14-2. 사업실적 (1) 매출액 누락	필수
M0702	14. 사업실적 (1) 매출액이 1조원을 초과함	검토
M0703	14-1 (2) 영업비용합계{(① 재료매입비+…⑤ 기타 영업비용)} 불일치	필수
M0704	14 (1) 매출액 - (2) 영업비용 ① 재료매입비 0 이하임	검토
M0705	14 (1) 매출액 - (2) 영업비용 ② 인건비가 0 이하임	검토
M0706	(1) 매출액-(2) 영업비용 = (3) 영업이익과 불일치	필수
M0707	7. 연간급여액 합계가 14. ② 인건비보다 많음	필수
M0708	14-2 (2) 영업비용합계{(① 재료매입비+…⑩ 기타 영업비용)} 불일치	필수
M0709	7. 월평균종사자수 ③ 상용종사자, ④ 임시일용종사자, ⑤ 무급종사자수가 있는데 14. ② 인건비 누락	검토
M0710	7. 연간급여액 합계가 14. ② 인건비보다 많음	필수
M0711	산업분류가 681인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)>1,000백만원]	필수
M0712	산업분류가 681인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)> 120백만원]	검토
M0713	산업분류가 681인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 5백만원]	검토
M0714	산업분류가 682, 69인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)> 110백만원]	필수
M0715	산업분류가 682, 69인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)> 50백만원]	검토
M0716	산업분류가 682, 69인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 1백만원]	검토
M0726	산업분류가 809인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)> 20백만원]	필수
M0727	산업분류가 809인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)> 10백만원]	검토
M0728	산업분류가 809인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.4백만원]	필수
M0729	산업분류가 809인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	검토
M0730	산업분류가 85, 86인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)>20백만원]	필수
M0731	산업분류가 85, 86인데 매출액 과다 재확인 [=(매출액/연간영업개월수/종사자수합계)>15백만원]	검토
M0732	산업분류가 85인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	필수
M0733	산업분류가 85인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 1백만원]	검토

	M0734	산업분류가 86인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.2백만원]	필수
	M0735	산업분류가 86인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	검토
	M0736	산업분류가 87인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)>150백만원]	필수
	M0737	산업분류가 87인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)>70백만원]	검토
	M0738	산업분류가 87인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.3백만원]	필수
	M0739	산업분류가 87인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 0.6백만원]	검토
	M0740	산업분류가 88인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)> 50백만원]	필수
	M0741	산업분류가 88인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)>20백만원]	검토
	M0742	산업분류가 88인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	필수
	M0743	산업분류가 901인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)> 30백만원]	필수
	M0744	산업분류가 901인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)> 20백만원]	검토
	M0745	산업분류가 901인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	필수
	M0746	산업분류가 901인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 0.8백만원]	검토
	M0747	산업분류가 9021,9022인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)>120백만원]	필수
	M0748	산업분류가 9021,9022인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)> 80백만원]	검토
	M0749	산업분류가 9021,9022인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	필수
	M0750	산업분류가 9021,9022인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 0.9백만원]	검토
	M0751	산업분류가 91,92인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)> 50백만원]	필수
	M0752	산업분류가 91,92인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)> 20백만원]	검토
	M0753	산업분류가 91인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.2백만원]	필수
	M0754	산업분류가 91인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	검토

	M0755	산업분류가 92인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	필수
	M0756	산업분류가 93인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)> 16백만원]	필수
	M0757	산업분류가 93인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)> 10백만원]	검토
	M0758	산업분류가 93인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	필수
	M0759	산업분류가 855~857인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)> 20백만원]	필수
	M0760	산업분류가 855~857인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)> 10백만원]	검토
	M0761	산업분류가 855~857인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.4백만원]	필수
	M0762	산업분류가 855~857인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	검토
	M0763	산업분류가 86, 87인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)>20백만원]	필수
	M0764	산업분류가 86, 87인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)>15백만원]	검토
	M0765	산업분류가 86인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	필수
	M0766	산업분류가 86인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 1백만원]	검토
	M0767	산업분류가 87인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.2백만원]	필수
	M0768	산업분류가 87인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	검토
	M0769	산업분류가 87인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)>150백만원]	필수
	M0770	산업분류가 87인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)>70백만원]	검토
	M0771	산업분류가 87인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.3백만원]	필수
	M0772	산업분류가 5911, 60, 90인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 0.6백만원]	검토
	M0773	산업분류가 91인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)> 50백만원]	필수
	M0774	산업분류가 91인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)>20백만원]	검토
	M0775	산업분류가 91인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	필수

	M0776	산업분류가 37인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)> 30백만원]	필수
	M0777	산업분류가 37인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)> 20백만원]	검토
	M0778	산업분류가 37인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	필수
	M0779	산업분류가 37인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 0.8백만원]	검토
	M0780	산업분류가 38,39인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)> 120백만원]	필수
	M0781	산업분류가 38,39인데 매출액 과다 재확인[= (매출액/연간영업개월수/종사자수합계)> 80백만원]	검토
	M0782	산업분류가 38,39인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	필수
	M0783	산업분류가 38,39인데 매출액 과소 재확인 [= (매출액/연간영업개월수/종사자수합계)< 0.9백만원]	검토
	M0784	산업분류가 95인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)> 50백만원]	필수
	M0785	산업분류가 95인데 매출액 과다 [= (매출액/연간영업개월수/종사자수합계)> 20백만원]	검토
	M0786	산업분류가 95인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	필수
	M0787	산업분류가 96인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)> 16백만원]	필수
	M0788	산업분류가 96인데 매출액 과다 재확인 [= (매출액/연간영업개월수/종사자수합계)> 10백만원]	검토
	M0789	산업분류가 96인데 매출액 과소 [= (매출액/연간영업개월수/종사자수합계)< 0.5백만원]	필수

<부록 2> 서비스업조사(2007년 기준)의 조사범위

E 하수 폐기물처리 원료재생 및 환경복원업
- 37 하수 폐수 및 분뇨 처리업
  - 370 하수 폐수 및 분뇨 처리업
- 38 폐기물 수집운반 처리 및 원료재생업
  - 381 폐기물 수집운반업
  - 382 폐기물 처리업
- 39 환경 정화 및 복원업
  - 390 환경 정화 및 복원업

J 출판 영상 방송통신 및 정보서비스업
- 58 출판업
  - 581 서적 잡지 및 기타 인쇄물 출판업
  - 582 소프트웨어 개발 및 공급업
- 59 영상 오디오 기록물 제작 및 배급업
  - 591 영화 비디오물 방송프로그램 제작 및 배급업
  - 592 오디오물 출판 및 원판 녹음업
- 60 방송업
  - 601 라디오 방송업
  - 602 텔레비전 방송업
- 62 컴퓨터 프로그래밍 시스템 통합 및 관리업
  - 620 컴퓨터 프로그래밍 시스템 통합 및 관리업
- 63 정보서비스업
  - 631 자료처리 호스팅 포털 및 기타 인터넷 정보매개서비스업
  - 639 기타 정보서비스업

L 부동산업 및 임대업
- 68 부동산업
  - 681 부동산 임대 및 공급업
  - 682 부동산 관련 서비스업
- 69 임대업 부동산 제외
  - 691 운송장비 임대업
  - 692 개인 및 가정용품 임대업

693 산업용 기계 및 장비 임대업
694 무형재산권 임대업

N 사업시설관리 및 사업지원 서비스업
74 사업시설 관리 및 조경 서비스업
741 사업시설 유지관리 서비스업
742 건물 산업설비 청소 및 방제 서비스업
743 조경 관리 및 유지 서비스업
75 사업지원 서비스업
751 인력공급 및 고용알선업
752 여행사 및 기타 여행보조 서비스업
753 경비 경호 및 탐정업
759 기타 사업지원 서비스업

P 교육서비스업
85 교육 서비스업
855 일반 교습 학원
856 기타 교육기관
857 교육지원 서비스업

Q 보건업 및 사회복지서비스업
86 보건업
861 병원
862 의원
863 공중 보건 의료업
869 기타 보건업
87 사회복지 서비스업
871 거주 복지시설 운영업
872 비거주 복지시설 운영업

R 예술 스포츠 및 여가관련 서비스업
90 창작 예술 및 여가관련 서비스업
901 창작 및 예술관련 서비스업
902 도서관 사적지 및 유사 여가관련 서비스업
91 스포츠 및 오락관련 서비스업

911 스포츠 서비스업
912 유원지 및 기타 오락관련 서비스업

S 협회 및 단체 수리 및 기타 개인 서비스업
95 수리업
951 기계 및 장비 수리업
952 자동차 및 모터사이클 수리업
953 개인 및 가정용품 수리업
96 기타 개인 서비스업
961 미용 욕탕 및 유사 서비스업
969 그외 기타 개인 서비스업