제2장

합계불일치 오류의 자동수정[1)]

이의규

제1절 서 론

1. 연구의 배경 및 필요성

자료수집 및 처리 단계에서 오류를 찾아내고 수정하는 에디팅(editing) 절차는 전통적인 통계조사에서의 전체 예산의 많은 부분을 차지하는 것으로 보고되고 있다(Granquist, 1997). 따라서 통계작성기관에서는 통계자료를 에디팅 함에 있어 작은 비용으로도 에디팅을 효율적으로 수행할 수 있는 새로운 에디팅 기법이 필요하다. 통계 선진국에서는 이미 응답자 재접촉, 내검 인력 및 예산상의 어려움을 극복하고자 에디팅과 관련하여 많은 연구가 수행되고 있으며 새로운 에디팅 기법들이 실무에서 적용되고 있다. 그 중 하나가 에디팅 업무의 자동화이다. 외국과의 조사환경이 현재 우리와 꼭 같지는 않더라도 향후 조사환경의 변화에 대비하여 자동 에디팅(automatic editing)에 대한 연구가 지속적으로 필요하다.

특히 사업체 대상 조사의 자동 에디팅은 매출액, 급여액, 종사자수, 영업비용 등 수량자료의 수리적 연관성 규칙을 이용하여 자동으로 수정되어야 할 항목을 선택하고 수정·대체하는 기법으로 통계 선진국에서는 오래전부터 예산 및 조사의 어려움을 극복하기 위해 활용되고 있으며 이를 통해 수작업으로 인한 과도한 에디팅의 단점을 보완하고 있다.

---

1) 본 연구는 네덜란드 통계청 소속 연구원 Sander Scholtus와 공동연구로 추진되었음

2. 연구목적 및 방향

본 연구의 목적은 서비스업조사 자료에서 합계 불일치 오류의 자동수정을 시도하고 그 유용성을 검토하고자 함이다. 일반적으로 경제조사에서 가장 중요한 자료 검토 중 하나는 종사자수의 합계, 사업실적의 합계, 자산 및 재고액의 합계 등 합계불일치 오류에 대한 검토이다. 따라서 각종 경제조사에서 합계일치 조건을 갖는 항목의 자동 에디팅 연구를 통해 에디팅 인력 및 시간을 절감하고 자료의 품질을 효율적으로 관리함을 궁극적인 목표로 한다. 특히 최종 점검에도 불구하고 특이 사항 없이 합계가 불일치한 자료에 대한 대응 절차가 제시된다.

본 연구는 에디팅의 자동화에 따른 위험을 최소화하고 적용가능성이 높은 합계일치 조건하에서의 간단한 오타의 자동수정에 무게를 둔다. 단순 오타로 인한 합계불일치 오류가 아닌 나머지 오류는 수학적 최적화 방법에 의하여 자동수정한다.

3. 연구의 범위 및 내용

본 연구는 2008년 서비스업조사의 내검(에디팅) 최종단계의 자료를 이용하여 종사자수의 합계 및 영업이익의 합계 불일치 오류에 대해 자동수정 기법을 적용하고 그 결과를 분석함을 연구범위로 한다.

본 연구에서는 등식조건하에서의 불일치 오류의 자동수정에 대한 이론적 배경 검토, 종사자수와 영업이익의 합계일치 필수 내검규칙의 분석, 합계불일치 오류의 자동수정 프로그램 및 패키지 사용 및 적용결과, 합계불일치 오류에 대한 체계적 자동수정 방법 및 절차, 최종자료와 비교・검토 등을 주요 내용으로 다룬다.

본 보고서는 다음과 같이 전개된다. 먼저 2절에서는 서비스업조사의 개요와 내검절차에 대해 간략하게 살펴본다. 3절에서는 펠레기-홀트(Fellegi-Holt) 기법과 선형계획법을 이용한 자동수정에 대해 검토한다(이의규, 2008, 2009). 이후 등식조건하에서의 단순 오타의 자동수정을 다룬다. 4절에서는 서비스업조사에서의 종사자수와 사업실적의 합계불일치 오류에 대한 자동수정기법을 적용하고 이에 대한 결과를 분석한다. 마지막 절에서는 연구의 결과요약과 향후의 연구 방향에 대해 언급한다.

제2절 서비스업 조사의 내검

1. 서비스업조사의 개요

서비스업조사의 목적은 서비스업에 대한 구조변화와 경영 실태를 파악하여 서비스산업관련 정책수립, 국내총생산(GDP)의 추계, 기업의 경영계획, 학술연구 등을 위한 기초자료를 제공하고자 함이다.

조사대상은 한국표준산업분류(KSIC)상 8개(E, J, L, N, P, Q, R, S) 산업대분류에 해당하는 사업체 중 표본으로 선정된 사업체이며, 조사항목은 기본항목 10개와 특성항목 4개로 구성되는데, 기본항목은 사업체명 및 소재지, 사업내용, 조직형태, 일일평균 영업시간, 사업체 정기 휴무일수, 연간영업개월수, 월평균 종사자수 및 연간급여액, 사업체 건물 연면적, 전자상거래 활용현황, 사업실적이며, 특성항목은 직능별 종사자수, 전산장비 보유대수, 무형자산 보유건수, 이용인원(고객)수이다.

조사방법은 지방통계청/사무(출장)소의 조사담당직원과 임시조사원이 직접 조사대상 사업체를 방문하여 면접조사 함을 원칙으로 한다. 면접이 곤란한 경우 응답자 직접기입 방식이나 인터넷 조사방식을 취한다.

2. 서비스업조사의 내검 절차

지방청에서는 조사 담당자간 1~2일의 교차내검과 조사표 전산입력후 전산내검을 실시하고 있으며, 본청에서는 종합내검(3개월), 지방청 질의조회, 수준분석(2개월)을 실시하여 자료내검 및 오류정정을 하고 있다.

조사표에서 얻어진 자료는 현지 입력・내검시 오류사항을 전산으로 검토하고 있으며 이때 점검코드는 필수적으로 만족해야만 입력이 가능한 필수코드와 점검이 필요한 선택코드로 나뉜다. 종사자수와 급여액, 매출액, 영업비용과 같이 금액에 관계된 오류코드와 내용은 <표 2-1>을 참조하기 바란다(통계청, 2008).

3. 현황 및 문제점

현재 어떤 레코드가 어떻게 해서 합계가 불일치한지를 전산으로 탐색하고 있으나, 이후의 처리는 주로 수작업에 의해 이루어진다. 특히 경제조사에서 합계일치 및 선형 등식 일치를 필요로 하는 항목(종사자수, 급여액, 사업실적, 영업비용, 유형・무형자산, 연말잔액, 재고액 등)은 항목간 선형 등식 관계를 필요로 하고 있으며 합계 항목이 다른

항목과 일치되어야 하는 경우도 존재한다. 입력·내검 프로그램 운용으로 오류레코드를 전산으로 검토하고 있으나, 내검에 통과하지 못한 레코드의 수정은 주로 조사원에 의한 수작업으로 이루어진다. 그런데 향후 촉박한 일정, 과다한 내검량, 응답 비협조 증가로 조사원에 의한 수정은 많은 비용과 시간이 소요되며 내검원 간 편차가 발생할 수 있다.

〈표 2-1〉 오류코드와 오류내용

유형	오류코드	오류내용	확인 여부
공통	F0601	(1)월평균종사자수 (가)남자합계(①자영업주+ … +⑤무급종사자) 불일치	필수
	F0602	(1)월평균종사자수 (나)여자합계(①자영업주+ … +⑤무급종사자) 불일치	필수
	F0603	(1)월평균종사자수 (다)계합계(①자영업주+ … +⑤무급종사자) 불일치	필수
	F0604	(1)월평균종사자수 합계 (가)남자+(나)여자=(다)계가 불일치	필수
	F0605	(1)월평균종사자수 ①영업주 (다)계{(가)남자+(나)여자}가 불일치	필수
	F0606	(1)월평균종사자수 ②무급가족종사자 (다)계 {(가)남자+(나)여자}가 불일치	필수
	F0607	(1)월평균종사자수 ③상용종사자 (다)계 {(가)남자+(나)여자}가 불일치	필수
	F0608	(1)월평균종사자수 ④임시일용종사자 (다)계 {(가)남자+(나)여자}가 불일치	필수
	F0609	(1)월평균종사자수 ⑤무급종사자 (다)계 {(가)남자+(나)여자}가 불일치	필수
	F0610	종사자수 누락	필수
	F0611	종사자수 1000명이상 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
	F0620	7. 월평균 종사자 ③상용종사자 월평균 급여액(5십만이하 또는 5백만원 이상) 확인[= (상용연간급여액/상용종사자)/영업개월수]	검토
조사표(3)	M0701	14-1, 14-2. 사업실적 (1)매출액 누락	필수
	M0702	14. 사업실적 (1)매출액이 1조원을 초과함	검토
	M0703	14-1 (2)영업비용합계{(①재료매입비+…⑤기타 영업비용)} 불일치	필수
	M0704	14 (1) 매출액 - (2)영업비용 ①재료매입비 0 이하임	검토
	M0705	14 (1) 매출액 - (2)영업비용 ②인건비가 0 이하임	검토
	M0706	(1)매출액-(2)영업비용 = (3)영업이익과 불일치	필수
	M0707	7. 연간급여액 합계가 14. ②인건비 보다 많음	필수
	M0708	14-2 (2)영업비용합계{(①재료매입비+…⑩기타 영업비용)} 불일치	필수
	M0709	7. 월평균종사자수 ③상용종사자, ④임시일용종사자, ⑤ 무급종사자수가 있는데 14. ②인건비 누락	검토
	M0710	7. 연간급여액 합계가 14. ②인건비 보다 많음	필수
	M0711-789	종사자당 월평균 매출액 (매출액/종사자수합계/연간영업개월수)이 과소 또는 과다 (산업분류별로 과소, 과다 한계값이 다름)	필수

자료: 통계청,「전산내검요령서」, 2008 (음영부분은 합계불일치 오류코드와 오류내용)

제3절 등식조건 하에서의 단순 오타의 자동수정

1. 에디팅의 개요

통계자료 에디팅(statistical data editing)은 자료 수집 및 처리 단계에서 오류를 찾아내고 이를 수정하는 과정을 말한다. 에디팅은 작업 방법에 따라 수작업 에디팅(manual editing)과 자동 에디팅(automatic editing)으로 구분한다. 그리고 에디팅의 대상에 따라 마이크로 에디팅(micro editing)과 매크로 에디팅(macro editing)으로 구분한다. 마이크로 에디팅은 개별적인 레코드 수준에서의 자료 점검인 반면 매크로 에디팅은 모든 레코드 수준에서의 자료 검토이다.

일반적으로 내검을 위해 설정하는 내검규칙에는 필수규칙(fatal edit, hard edit)과 선택규칙(query edit, soft edit)이 있다. 필수규칙은 반드시 해결되어야 할 점검사항이고 선택규칙은 다소 의심스러운 값의 범위에 있을 때 점검을 권고하는 규칙이다. 한편 과도한 선택 규칙은 업무 부담으로 새로운 오류를 유발할 수 있어 피해야 한다.

그리고 데이터의 오류는 체계적 오류(systematic error)와 랜덤 오류(random error)로 구분한다. 체계적 오류는 특정 항목에 대해 일관되게 나타나는 오류, 예를 들면 잘못된 단위로 응답하는 경우이다. 랜덤 오류는 구조적 원인이 아닌 우연적으로 나타나는 오류, 예를 들면 입력원에 의해 잘못 입력되는 경우를 말한다.

전형적인 자동 에디팅 절차는 먼저 측정단위 오류와 같은 체계적 오류를 연역적 알고리즘을 이용하여 자동 탐색 및 수정함으로 해결하고 이후 우연적 원인에 의한 랜덤 오류는 수학적 최적화 문제를 풀어 제거한다. 한편 랜덤 오류를 해결하기 위한 일반적인 방법은 펠레기-홀트(Fellegi-Holt) 방법으로 이는 모든 내검규칙을 동시에 만족하도록 수정되어야할 변수(항목)의 최소 집합을 찾는 것이다(Fellegi와 Holt, 1976).

2. Fellegi-Holt 기법의 리뷰

F-H 기법은 수학적 최적화에 기초한 대표적 에디팅 방법이다. F-H 방법은 조사 자료에 오류가 있는지를 판단하기 위해 조사 담당자에 의해 설정되는 내검규칙(edits)을 필요로 한다. 만약 설정된 내검규칙을 위반하면 어떤 변수를 대체해야 할지를 결정하는 자동화 전략이 필요한데, 주어진 정보를 최대한 보존하면서 모든 내검규칙을 만족하게 하는 최소의 수정할 변수를 찾아내자는 것이 F-H 전략이다. 이 F-H 방법은 종종 각 변수에 신뢰 가중치를 부여하여 변화되어야 할 변수의 신뢰 가중합을 최소화하는 해를 구하는 형식으로 일반화하여 사용한다. 캐나다 통계청의 Banff, 미국 센서스 국의 SPEER와

DISCRETE, 네덜란드 통계청의 SLICE는 일반화된 F-H를 기반으로 한다.

한편, 이해를 돕기 위해 다음과 같은 2개의 명시적 내검규칙(explicit edits)이 있다고 가정하자(각 변수는 음이 아닌 수).

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

여기서 하나의 레코드가 (X_1= 6, X_2 = 4, X_3 =8)로 코딩되었다 하자. 그러면 이 레코드는 두 번째 규칙을 위반한 레코드이다. 이때 두 개 이상의 변수를 모두 바꾸어서 성립이 가능할 수 있으나 최대한 자료를 보존한다는 원칙에서 X_3 하나만을 바꾸는 것이 합리적이다.

이와 같은 결론은 주어진 내검규칙 E_1과 E_2로부터 변수 X_2의 소거를 통해 다음과 같은 내재적 내검규칙(implicit edits)을 구함으로써 도출된다.

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

따라서 주어진 레코드의 전체 위배된 내검규칙은 E_2, E_3이다(<표 2-2> 참조).

〈표 2-2〉 위배된 내검규칙 행렬

	X_1	X_2	X_3	상태
E_1	1	1		합격
E_2		1	1	위배
E_3	1		1	위배

X_3는 위배된 내검규칙 E_2, E_3에 모두 포함되어 X_3를 바꾸어주는 것이 합리적이다. 즉 명시된 내검규칙으로부터는 어떤 변수를 바꾸어 주어야 할지가 명확하지 않으나 이처럼 추가된 내검규칙을 이용하면 자료의 오류위치를 효율적으로 판단할 수 있다. 더 나아가 X_3값을 미지수로 놓고 나머지 주어진 값을 조건식에 대입하여 풀면 $0 \leq X_3 \leq 4/3$ 일 때 모든 규칙을 만족한다. 따라서 $X_3 = 1$이 가능한 대체값이 된다.

F-H 방법의 장점은 오류자료의 수정할 항목을 결정할 때 모든 변수가 동시에 고려된다는 것이다. 또한 주어진 편집규칙으로부터 유도된 내재적 내검규칙(implied edits, implicit edits)이 오류자료의 변경할 변수들을 결정할 때 주요한 역할을 하며, 일반적인

If-Then-Else의 구조보다 효율적이고 내검규칙의 수정 또는 변경 시 그 관리가 용이하다 (Chen 등, 2002). 더욱이 각 변수에 신뢰 가중치를 부여하여 일반화가 가능하다.

반면 F-H 방법의 단점은 설정된 모든 내검규칙을 필수적으로 만족시켜야 하는 규칙(hard edits)으로 간주한다는 것이다. 또한 오류를 모두 랜덤오류로 인식한다. 특히 요구되는 내재적 내검규칙 수가 매우 많을 수 있으며, 이때 모든 내재적 규칙의 생성에 있어서 많은 시간이 소요된다(De Waal과 Coutinho, 2005).

3. 선형계획법을 이용한 수정

선형계획법을 이용한 수정방법은 오류위치포착 문제의 해를 얻는 더 간단하고 빠른 접근방법으로 원 관측값과 에디팅된 값과의 절대 차이값의 합을 최소화하는 방법이다 (De Waal, 2003).

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

여기서 다음과 같은 세 개의 변수 X_1, X_2, X_3와 2개의 에디팅 규칙이 있다고 가정하자.

$$\begin{aligned} X_1 &- X_2 \geq 0 \\ X_2 &- 3X_3 \geq 0 \end{aligned}$$

하나의 레코드가 X_1 = 6, X_2 = 4, X_3 = 8을 갖는다고 가정하자. 따라서 이 레코드는 두 번째 에디팅 규칙을 위배하는 레코드이다. 앞서 언급한 바와 같이 위 식은 다음과 같은 하나의 선형계획법(Linear Programming) 문제이다.

$$\min \left(| X_1 - 6 | + | X_2 - 4 | + | X_3 - 8 | \right)$$

제약조건식:

$$\begin{aligned} X_1 &- X_2 \geq 0 \\ X_2 &- 3X_3 \geq 0 \end{aligned}$$

한편 R 프로그램을 이용하면, X_1 = 6, X_2 = 4, X_3 = 1.33의 해를 얻는다. 다른 변수는 변화가 없는 반면 X_3는 8에서 1.33으로 바뀌었기 때문에 X_3 = 1.33(또는 1)이 하나의 가능한 해가 된다. 이는 앞의 F-H 방법에 의한 결과와 동일한 결과를 나타낸다.

한편 선형계획법을 이용한 최소한의 수정은 모든 에디팅규칙을 만족하면서 오류로 간주된 변수들을 수정하는 일치적인 대체이다. 연속형 자료인 경우 다음과 같이 거리함수를 최소화하는 문제로 일반화할 수 있다.

$$\min \sum_{i=1}^{n} w_i |\tilde{x}_i - x_i|$$

제약조건식:

$$E_i : a_{i,1}\tilde{x}_1 + a_{i,2}\tilde{x}_2 + ... + a_{i,n}\tilde{x}_n \geq b_i, \quad i = 1,2,...,m$$

$$\tilde{x}_j \geq 0, \quad j = 1,2,...,n$$

여기서 $\tilde{x}_i$: 미지의 에디팅된 값
x_i : 알려진 원 관측값
w_i : 각 항목의 신뢰 가중치

4. 균형 내검규칙 하에서의 간단한 오타의 자동수정

이제 만약 랜덤 오류가 $X_1 - X_2 = X_3$와 같이 선형 등식의 형태를 갖는 내검규칙(균형 내검규칙, balance edit)을 위반한다면 F-H 패러다임에 의해 사용되지 않는 오류 근거정보가 존재할 수 있다.

즉, 다음과 같은 세 개의 변수 X_1, X_2, X_3와 다음과 같은 1개의 균형내검 규칙이 있다고 가정하자.

$$X_1 - X_2 = X_3$$

하나의 레코드가 X_1 (매출액)= 353, X_2 (영업비용)= 283 X_3 (영업이익)= 115를 갖는다고 가정하자. 그러면 매출액 - 영업비용 = 영업이익이라는 내검규칙을 위반하게 된다. F-H 방법은 매출액이 353에서 398 또는 영업비용이 283에서 238 또는 영업이익이 115에

서 70으로의 수정을 가능한 해로 제시한다. 영업비용이 283에서 238은 자릿수 바뀜으로 자릿수 착오일 가능성이 높다(F-H는 이러한 정보를 이용하지 못함). 물론 영업비용에 신뢰 가중치를 낮게 부여하여 일반화된 F-H 방법을 적용하면 원하는 결과를 얻을 수 있으나, 실제로 많은 규칙이 존재하는 실제 경우에서 일반화 F-H 방법은 적용하기 어렵다. 따라서 단순오류의 대표적 패턴을 이용하여 먼저 오류를 해결하면 정확한 오류해결 뿐 아니라 나머지 오류에 대한 작업량을 줄일 수 있다.

간단한 예를 하나 더 들어 보자. 하나의 합계 내검규칙이 $X_1 + X_2 = X_3$라고 가정하자. 주어진 자료 X_1 = 100, X_2 = 20 X_3 = 300은 내검규칙을 위반하므로 오류자료이다. 이때, X_1의 100을 280으로 바꾸거나 X_2의 20을 200으로 바꾸거나 X_3의 300을 120으로 바꾸면 내검규칙이 성립한다. 그러나 X_1이나 X_3는 2개의 단위 숫자를 바꾸어야 하므로 설명력이 없고 X_2는 단지 0을 추가하면 되므로 가장 그럴 듯한 값이 된다.

Scholtus(2009)는 단순 오타의 대표적 패턴으로 이웃한 두 개의 자릿수 바뀜, 한 자릿수 늘어남, 한 자릿수 빠짐, 음수 기호가 빠지거나 들어가는 경우, 한 자릿수에서 다른 값으로 대체되는 경우의 5가지 오류를 수학적으로 표현한다. 이를 위해 먼저 모든 값은 다음과 같이 정의할 수 있다.

$$x = \sum_{j=0}^{M} \xi_j \cdot 10^j \, , \xi_j = 0,1,\cdots,9$$

① 두 개의 자릿수 바뀜

$$f_{ic}(x;k) = x + \xi_k \cdot (10^{k+1} - 10^k) + \xi_{k+1} \cdot (10^k - 10^{k+1}),\ k = 0,\cdots,M-1$$

예) $f_{ic}(4627;1) = 4267$

② 한 자릿수 늘어남

$$f_a(x;k,\xi) = \sum_{j=0}^{k-1} \xi_j \cdot 10^j + \xi \cdot 10^k + \sum_{j=k+1}^{M} \xi_{j-1} \cdot 10^j,\ k = 0,\cdots,M;\xi = 0,\cdots,9$$

예) $f_a(4627;1,8) = 46287$

③ 한 자릿수 빠짐

$$f_o(x;k)=\sum_{j=0}^{k-1}\xi_j\cdot 10^j+\sum_{j=k+1}^{M}\xi_j\cdot 10^{j-1},\ k=0,\cdots,M$$

예) $f_a(4627;1)=467$

④ 음수 기호가 빠지거나 들어감

$$f_m(x)=-x$$

예) $f_m(4627)=-4627$

⑤ 해당 자릿수에서 다른 숫자로 대체됨

$$f_r(x;k,\xi)=x+(\xi-\xi_k)\cdot 10^k,\ k=0,\cdots,M$$

예) $f_r(4627;1,8)=4687$

등식 내검규칙 e_r의 $a_{r,1}x_1+\cdots a_{r,n}x_n=0,\ (r=1,\cdots,m)$을 행렬로 표현하면 다음과 같다.

$$A\mathrm{x}=\mathbf{0}$$

여기서, $A=[a_{r,i}]_{m\times n}$

$$\mathrm{x}=[x_1,\cdots,x_n]'.$$

이때 각 내검규칙은 3개의 변수 첨자군으로 정의한다.

$$I_1^{(r)}=\{i:a_{r,i}>0\}$$
$$I_2^{(r)}=\{i:a_{r,i}<0\}$$
$$I_3^{(r)}=\{i:a_{r,i}=0\},\ r=1,\cdots,m$$

따라서

$$\sum_{i\in I_1^{(r)}} a_{r,i}x_i = -\sum_{i\in I_2^{(r)}} a_{r,i}x_i$$

이다. 그리고,

$$\overline{I_3}^{(r)} = I_1^{(r)} \cup I_2^{(r)}$$

는 내검규칙 e_r에 포함된 모든 변수의 첨자 집합을 의미한다.

한편 각 변수는 3개의 내검규칙 첨자군으로 정의할 수 있다.

$$\begin{aligned} R_1^{(i)} &= \{r : a_{r,i} > 0\} \\ R_2^{(i)} &= \{r : a_{r,i} < 0\} \\ R_3^{(i)} &= \{r : a_{r,i} = 0\},\ i = 1,\cdots,n \end{aligned}$$

그리고,

$$\overline{R}_3{}^{(i)} = R_1^{(i)} \cup R_2^{(i)}$$

는 변수 x_i를 포함하는 모든 내검규칙의 첨자 집합을 나타낸다. 그리고,

$$s_1^{(r)} = \sum_{i\in I_1^{(r)}} a_{r,i}x_i$$

$$s_2^{(r)} = -\sum_{i\in I_2^{(r)}} a_{r,i}x_i,\ \ r = 1,\cdots,m$$

이라 할 때,

$$\phi(r) = \begin{cases} 1\,, s_1^{(r)} = s_2^{(r)} \\ 0\,, s_1^{(r)} \neq s_2^{(r)} \end{cases}$$

로 정의하자. 그러면

$$E_1 = \{r : \phi(r) = 1\}$$
$$E_2 = \{r : \phi(r) = 0\}$$

E_1은 위반된 내검규칙의 첨자 집합을 표시하고, E_2는 만족하는 내검규칙의 첨자 집합을 나타낸다. 이제,

$$I_0 = \cap_{r \in E_2} I_3^{(r)}$$

는 만족하는 내검규칙들에 포함되지 않는 공통변수의 첨자 집합을 의미한다.

이미 만족된 내검규칙을 위반하지 않고 만족되는 내검규칙의 수를 증가하는 수정을 수행하고자 한다면 우리가 안전하게 변화시킬 수 있는 유일한 변수들은 I_0에 있는 것이다.

$i \in \overline{I}_3^{(r)}$이면서 e_r 규칙이 위배되면 x를

$$\tilde{x}_i^{(r)} = \frac{1}{a_{r,i}}\left(s_2^{(r)} - s_1^{(r)} + a_{r,i}x_i\right)$$

로 바꾸면 규칙은 만족된다.

이때 모든 $r \in \overline{R}_3^{(i)}$에 대해 $\tilde{x}_i^{(r)}$, $i \in I_0$의 값들을 얻을 수 있으며 각 값에 대해 $x_i = f\left(\tilde{x}_i^{(r)}\right)$인지를 확인한다. 만약 이 이 식이 성립된다면 오타로 인해 변화된 것으로 볼 수 있고 만약 5가지의 어떤 함수도 위 식이 성립되지 않는다면 현재의 값을 바꿀 수 없다.

제4절 자동수정의 적용 및 분석

적용자료는 2008년 기준 서비스업조사 자료(2009년 조사자료)로 총 43,463건으로 집계되었으며 2009년 수시과제 연구에서 서비스업조사 자료의 오류(합계불일치오류를 포함)에 대해 보고한 바 있다(이의규, 2010). 여기서는 합계 불일치 오류 수정의 자동화에 초점을 두고 자료를 분석한다.

1. 종사자수 합계 불일치 오류의 수정

가. 월평균종사자수

2008년 영업기간(1.1~12.31) 중에 근무한 월평균 종사자 수를 <표 2-3>과 같이 종사상 지위별 및 성별로 구분하여 기입한다.

〈표 2-3〉 종사자수 조사항목

	남자	여자	계
자영업주	x_1	x_2	x_3
무급가족종사자	x_4	x_5	x_6
상용종사자	x_7	x_8	x_9
임시・일용종사자	x_{10}	x_{11}	x_{12}
무급종사자	x_{13}	x_{14}	x_{15}
합계	x_{16}	x_{17}	x_{18}

나. 내검규칙(Edit rules)

종사자수 조사항목과 관련된 합계불일치 점검규칙은 다음과 같이 9개의 내검규칙이 주어진다. 자영업주 합계 불일치, 무급가족종사자 합계 불일치, 상용종사자 합계 불일치, 임시일용종사자 합계 불일치, 무급종사자 합계 불일치, 남자 합계 불일치, 여자 합계 불일치, 남여 합계 불일치, 계 합계 불일치(이는 중복되는 내검규칙임)이다. 이를 수식으로 표현하면 다음과 같다.

$$e_1 : x_1 + x_2 = x_3$$
$$e_2 : x_4 + x_5 = x_6$$
$$e_3 : x_7 + x_8 = x_9$$
$$e_4 : x_{10} + x_{11} = x_{12}$$
$$e_5 : x_{13} + x_{14} = x_{15}$$
$$e_6 : x_1 + x_4 + x_7 + x_{10} + x_{13} = x_{16}$$
$$e_7 : x_2 + x_5 + x_8 + x_{11} + x_{14} = x_{17}$$
$$e_8 : x_{16} + x_{17} = x_{18}$$
$$e_9 : x_3 + x_6 + x_9 + x_{12} + x_{15} = x_{18}$$

다. 종사자수 합계불일치 오류

종사자수 합계의 불일치 건수는 총 15건으로 남자 종사자수 합계 불일치, 상용종사자수의 남녀 합계불일치다. Status에서 FALSE는 오류를 나타낸다. 상용종사자수가 종사

자수보다 큰 경우가 2건(사업체번호: 3400105247, 1103203381)이며, 종사자수 합계 불일치 탐색 결과는 <표 2-4>와 같다. 표에서 각 내검규칙에서 TRUE는 불일치를 의미하고 FALSE는 일치를 나타낸다.

〈표 2-4〉 종사자수 합계불일치 탐색결과

	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	status
1*	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
2	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
3	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
4	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
5	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
6	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
7	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
8	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
9	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE
10	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE
11	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE
12	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
13	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
14	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE
15	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE

* 1번 레코드는 내검규칙 e_6(남자종사자수 합계불일치: TRUE), e_9(종사자수 합계불일치: TRUE)를 위배하여 오류 레코드임(FALSE)

오류자료의 탐색결과에 따른 종사자수 합계불일치 오류자료의 내용은 <표 2-5>와 같다.

〈표 2-5〉 종사자수 합계불일치 탐색결과

	사업체번호	자영남	자영여	자영합	가족남	가족여	가족합	상용남	상용여	상용합	임시남	임시여	임시합	무급남	무급여	무급합	남자합	여자합	전체합	상태
1	3405006435	1	0	1	0	0	0	13	13	26	0	0	0	0	0	0	13	13	26	오류
2	1100604786	1	0	1	0	0	0	9	7	17	0	0	0	0	0	0	10	7	17	오류
3	2608007136	1	0	1	0	0	0	6	3	9	2	8	10	0	0	0	8	11	19	오류
4	2403024564	0	0	0	0	0	0	66	2	68	0	0	60	0	0	0	66	2	68	오류
5	1104112818	0	0	0	0	0	0	238	100	338	0	0	40	0	0	0	238	100	338	오류
6	3900016606	0	0	0	0	0	0	13	2	15	41	14	57	0	0	0	54	18	72	오류
7	1100532638	0	0	0	0	0	0	623	405	1028	0	0	219	0	0	0	623	405	1028	오류
8	1108060943	0	0	0	0	0	0	45	5	50	0	0	4	0	0	0	45	5	50	오류
9	1105134258	0	0	0	0	0	0	19	13	32	0	0	0	15	14	29	24	27	51	오류
10	3400105247	0	0	0	0	0	0	103	9	112	6	1	7	0	0	0	100	8	108	오류
11	3603021041	0	0	0	0	0	0	59	22	81	4	6	10	0	80	80	63	28	91	오류
12	1103203381	0	0	0	0	0	0	72	18	93	0	0	2	0	0	0	72	18	90	오류
13	3100259164	0	0	0	0	0	0	1000	250	1250	0	0	30	0	0	0	1000	250	1250	오류
14	3702038242	0	0	0	0	0	0	51	12	63	0	0	0	0	180	180	51	182	233	오류
15	3306000288	0	0	0	0	0	0	11	6	17	5	11	16	0	85	85	16	17	33	오류

라. 단순오타의 자동수정

프로그램 R의 패키지를 이용하여 등식조건하에서의 단순오타의 자동수정을 실시한 결과가 <표 2-6>에 나타나있다. 총 15건 중 6건(4, 5, 6, 7, 8, 13번째 레코드)의 오류를 단순오타로 인식하였다. 4번째 레코드는 60→0, 5번째 레코드는 40→0, 6번째 레코드는 14→16, 7번째 레코드는 219→0, 8번째 레코드는 4→0, 그리고 13번째 레코드는 30→0로 수정되었다. 그런데 7번째 레코드는 x_{12}의 값을 219에서 0으로 바꾸면 완전하게 일치되므로 변수 당 허락되는 단위변화의 숫자를 3으로 설정한 결과이다(디폴트는 1).

〈표 2-6〉 correctTypos를 이용한 에디팅결과(종사자수)

	사업체번호	자영 남	자영 여	자영 합	가족 남	가족 여	가족 합	상용 남	상용 여	상용 합	임시 남	임시 여	임시 합	무급 남	무급 여	무급 합	남자 합	여자 합	전체 합	상태
1	3405006435	1	0	1	0	0	0	13	13	26	0	0	0	0	0	0	13	13	26	오류
2	1100604786	1	0	1	0	0	0	9	7	17	0	0	0	0	0	0	10	7	17	오류
3	2608007136	1	0	1	0	0	0	6	3	9	2	8	10	0	0	0	8	11	19	오류
4	2403024564	0	0	0	0	0	0	66	2	68	0	0	0	0	0	0	66	2	68	해결
5	1104112818	0	0	0	0	0	0	238	100	338	0	0	0	0	0	0	238	100	338	해결
6	3900016606	0	0	0	0	0	0	13	2	15	41	16	57	0	0	0	54	18	72	해결
7	1100532638	0	0	0	0	0	0	623	405	1028	0	0	0	0	0	0	623	405	1028	해결
8	1108060943	0	0	0	0	0	0	45	5	50	0	0	0	0	0	0	45	5	50	해결
9	1105134258	0	0	0	0	0	0	19	13	32	0	0	0	15	14	29	24	27	51	오류
10	3400105247	0	0	0	0	0	0	103	9	112	6	1	7	0	0	0	100	8	108	오류
11	3603021041	0	0	0	0	0	0	59	22	81	4	6	10	0	80	80	63	28	91	오류
12	1103203381	0	0	0	0	0	0	72	18	93	0	0	2	0	0	0	72	18	90	오류
13	3100259164	0	0	0	0	0	0	1000	250	1250	0	0	0	0	0	0	1000	250	1250	해결
14	3702038242	0	0	0	0	0	0	51	12	63	0	0	0	0	180	180	51	182	233	오류
15	3306000288	0	0	0	0	0	0	11	6	17	5	11	16	0	85	85	16	17	33	오류

마. 수학적 최적화 방법에 의한 자동수정

연역적 수정 알고리즘 패키지는 자료의 모든 불일치를 해결하지는 못한다. 완전한 자동에디팅을 위해서 나머지 불일치 오류는 선형계획법(LP)의 이용이나 Fellegi-Holt 패러다임에 근거한 오류위치탐색 알고리즘 등에 의해 해결한다.

한편 종사자수의 내합(각 종사형태별 남녀 종사자수의 합계)이 주변합이나 총합과 일치하지 않을 때는 내합을 근거로 자동수정되어야 한다. 먼저 주변합이 내합과 맞는 지

를 점검하고 다르다면 내합을 주변합으로 수정하여 주고, 다시 주변합이 총합과 맞는 지를 점검하여 다르다면 총합을 자동수정한다. 따라서 합계항목은 합계를 구성하는 항목에 의존하므로 구성항목에 더 큰 신뢰도를 부여한다. $x_3, x_6, x_9, x_{12}, x_{15}, x_{16}, x_{17}, x_{18}$(합계 항목)에 신뢰 가중치를 1을 부여하고 $x_1, x_2, x_4, x_5, x_7, x_8, x_{10}, x_{11}, x_{13}, x_{14}$(합계를 구성하는 항목)에 가중치 2를 부여한다.

이제 최소한의 수정을 제시하는 오류위치포착 문제는 다음과 같다.

$$\min \sum_{j=1}^{18} w_j |\tilde{x}_j - x_j|$$

여기서 수정되는 값 $\tilde{x}_j$는 다음을 만족하여야 한다.

$$\begin{aligned}
&\tilde{x}_1 + \tilde{x}_2 - \tilde{x}_3 = 0 \\
&\tilde{x}_4 + \tilde{x}_5 - \tilde{x}_6 = 0 \\
&\tilde{x}_7 + \tilde{x}_8 - \tilde{x}_9 = 0 \\
&\tilde{x}_{10} + \tilde{x}_{11} - \tilde{x}_{12} = 0 \\
&\tilde{x}_{13} + \tilde{x}_{14} - \tilde{x}_{15} = 0 \\
&\tilde{x}_{16} + \tilde{x}_{17} - \tilde{x}_{18} = 0 \\
&\tilde{x}_1 + \tilde{x}_4 + \tilde{x}_7 + \tilde{x}_{10} + \tilde{x}_{13} - \tilde{x}_{16} = 0 \\
&\tilde{x}_2 + \tilde{x}_5 + \tilde{x}_8 + \tilde{x}_{11} + \tilde{x}_{14} - \tilde{x}_{17} = 0 \\
&\tilde{x}_3 + \tilde{x}_6 + \tilde{x}_9 + \tilde{x}_{12} + \tilde{x}_{15} - \tilde{x}_{18} = 0 \\
&\tilde{x}_j \quad \geq \ 0, \qquad j = 1, \ldots 18
\end{aligned}$$

이 문제를 선형 문제로 작성하기 위해서는 y_j^+와 y_j^-를 이용하여 ($\tilde{x}_j = x_j + y_j^+ - y_j^-$) 이러한 수정 항목으로 다음과 같은 선형계획 문제로 변형하면 다음과 같다.

$$\min \sum_{j=1}^{18} w_j \left(y_j^+ + y_j^- \right)$$

제약식:

$$
\begin{aligned}
&y_1^+ - y_1^- + y_2^+ - y_2^- - y_3^+ + y_3^- = x_3 - x_1 - x_2 \\
&y_4^+ - y_4^- + y_5^+ - y_5^- - y_6^+ + y_6^- = x_6 - x_4 - x_5 \\
&y_7^+ - y_7^- + y_8^+ - y_8^- - y_9^+ + y_9^- = x_9 - x_7 - x_8 \\
&y_{10}^+ - y_{10}^- + y_{11}^+ - y_{11}^- - y_{12}^+ + y_{12}^- = x_{12} - x_{10} - x_{11} \\
&y_{13}^+ - y_{13}^- + y_{14}^+ - y_{14}^- - y_{15}^+ + y_{15}^- = x_{15} - x_{13} - x_{14} \\
&y_{16}^+ - y_{16}^- + y_{17}^+ - y_{17}^- - y_{18}^+ + y_{18}^- = x_{18} - x_{16} - x_{17}
\end{aligned}
$$

$$
\begin{aligned}
&y_1^+ - y_1^- + y_4^+ - y_4^- + \cdots + y_{13}^+ - y_{13}^- - y_{16}^+ + y_{16}^- = x_{16} - x_1 - x_4 - \cdots - x_{13} \\
&y_2^+ - y_2^- + y_5^+ - y_5^- + \cdots + y_{14}^+ - y_{14}^- - y_{17}^+ + y_{17}^- = x_{17} - x_2 - x_5 - \cdots - x_{14} \\
&y_3^+ - y_3^- + y_6^+ - y_6^- + \cdots + y_{15}^+ - y_{15}^- - y_{18}^+ + y_{18}^- = x_{18} - x_3 - x_6 - \cdots - x_{15} \\
&y_j^+ - y_j^- \geq -x_j \, (j = 1, \cdots, 18) \\
&y_j^+ \geq 0 \, (j = 1, \cdots, 18) \\
&y_j^- \geq 0 \, (j = 1, \cdots, 18)
\end{aligned}
$$

그러면 위 식은 다음과 같이 다시 간결하게 표현할 수 있다.

$$\min \boldsymbol{w}'\boldsymbol{z} \quad s.t. \quad A\boldsymbol{z} \triangle \boldsymbol{b}$$

$$
\begin{aligned}
&\boldsymbol{z} = \left(y_1^+, \cdots, y_{18}^+, y_1^-, \cdots, y_{18}^-\right)' \\
&\boldsymbol{w} = \left(w_1^+, \cdots, w_{18}^+, w_1^-, \cdots, w_{18}^-\right)'
\end{aligned}
$$

$\triangle$: 열 연산자($=, \geq$)
A : 계수행렬
b : 등식 우측의 벡터

프로그램 R에서 lpSolve 패키지를 내려 받아 선형계획법 문제를 해결할 수 있으며 그 수행결과는 <표 2-7>과 같다.

〈표 2-7〉 lpSolve를 이용한 에디팅결과(종사자수)

	사업체번호	자영남	자영여	자영합	가족남	가족여	가족합	상용남	상용여	상용합	임시남	임시여	임시합	무급남	무급여	무급합	남자합	여자합	전체합	상태
1	3405006435	1	0	1	0	0	0	13	13	26	0	0	0	0	0	0	14	13	27	해결
2	1100604786	1	0	1	0	0	0	9	7	16	0	0	0	0	0	0	10	7	17	해결
3	2608007136	1	0	1	0	0	0	6	3	9	2	8	10	0	0	0	9	11	20	해결
4	2403024564	0	0	0	0	0	0	66	2	68	0	0	0	0	0	0	66	2	68	해결
5	1104112818	0	0	0	0	0	0	238	100	338	0	0	0	0	0	0	238	100	338	해결
6	3900016606	0	0	0	0	0	0	13	2	15	41	16	57	0	0	0	54	18	72	해결
7	1100532638	0	0	0	0	0	0	623	405	1028	0	0	0	0	0	0	623	405	1028	해결
8	1108060943	0	0	0	0	0	0	45	5	50	0	0	0	0	0	0	45	5	50	해결
9	1105134258	0	0	0	0	0	0	19	13	32	0	0	0	15	14	29	34	27	61	해결
10	3400105247	0	0	0	0	0	0	103	9	112	6	1	7	0	0	0	109	10	119	해결
11	3603021041	0	0	0	0	0	0	59	22	81	4	6	10	0	80	80	63	108	171	해결
12	1103203381	0	0	0	0	0	0	72	18	90	0	0	0	0	0	0	72	18	90	해결
13	3100259164	0	0	0	0	0	0	1000	250	1250	0	0	0	0	0	0	1000	250	1250	해결
14	3702038242	0	0	0	0	0	0	51	12	63	0	0	0	0	180	180	51	192	243	해결
15	3306000288	0	0	0	0	0	0	11	6	17	5	11	16	0	85	85	16	102	118	해결

* 구성변수의 가중치는 2, 합변수의 가중치는 1

한편, F-H 패러다임을 기반으로 하는 네덜란드 통계청의 자동 에디팅 시스템 SLICE 수행 결과는 <표 2-8>에 나타내었다. 결과에서 15개의 불일치 자료는 모두 자동수정 되었으며, LP 방법에 의한 수행결과와 SLICE 수행결과는 정확하게 같은 결과가 나타났다(두 개의 목적함수가 다르기 때문에 항상 같은 결과를 보장하지는 못함).

〈표 2-8〉 SLICE를 이용한 에디팅결과(종사자수)

	Routine	id	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18
1	Input_Record	1	1	0	1	0	0	0	13	13	26	0	0	0	0	0	0	13	13	26
1	ErrorLocateSolution	1																x		x
1	Adapted_Record	1	1	0	1	0	0	0	13	13	26	0	0	0	0	0	0	14	13	27
2	Input_Record	2	1	0	1	0	0	0	9	7	17	0	0	0	0	0	0	10	7	17
2	ErrorLocateSolution	2									x									
2	Adapted_Record	2	1	0	1	0	0	0	9	7	16	0	0	0	0	0	0	10	7	17
3	Input_Record	3	1	0	1	0	0	0	6	3	9	2	8	10	0	0	0	8	11	19
3	ErrorLocateSolution	3																x		x
3	Adapted_Record	3	1	0	1	0	0	0	6	3	9	2	8	10	0	0	0	9	11	20
4	Input_Record	9	0	0	0	0	0	0	19	13	32	0	0	0	15	14	29	24	27	51
4	ErrorLocateSolution	9																x		x
4	Adapted_Record	9	0	0	0	0	0	0	19	13	32	0	0	0	15	14	29	34	27	61
5	Input_Record	10	0	0	0	0	0	0	103	9	112	6	1	7	0	0	0	100	8	108
5	ErrorLocateSolution	10																x	x	x
5	Adapted_Record	10	0	0	0	0	0	0	103	9	112	6	1	7	0	0	0	109	10	119
6	Input_Record	11	0	0	0	0	0	0	59	22	81	4	6	10	0	80	80	63	28	91
6	ErrorLocateSolution	11																	x	x
6	Adapted_Record	11	0	0	0	0	0	0	59	22	81	4	6	10	0	80	80	63	108	171
7	Input_Record	12	0	0	0	0	0	0	72	18	93	0	0	2	0	0	0	72	18	90
7	ErrorLocateSolution	12									x			x						
7	Adapted_Record	12	0	0	0	0	0	0	72	18	90	0	0	0	0	0	0	72	18	90
8	Input_Record	14	0	0	0	0	0	0	51	12	63	0	0	0	0	180	180	51	182	233
8	ErrorLocateSolution	14																	x	x
8	Adapted_Record	14	0	0	0	0	0	0	51	12	63	0	0	0	0	180	180	51	192	243
9	Input_Record	15	0	0	0	0	0	0	11	6	17	5	11	16	0	85	85	16	17	33
9	ErrorLocateSolution	15																	x	x
9	Adapted_Record	15	0	0	0	0	0	0	11	6	17	5	11	16	0	85	85	16	102	118

* 구성변수의 가중치는 2, 합변수의 가중치는 1

바. 최종자료와의 비교

2008년 기준 서비스업조사의 최종자료와 비교한 결과를 <표 2-9>에 나타내었다. 표에서 보면 종사자수의 합계불일치자료를 자동수정한 결과(위의 숫자)와 서비스업조사 최종자료(아래의 숫자)는 거의 일치한다. 2번과 6번 레코드에서만 약간의 차이가 나타나고 있다.

〈표 2-9〉 종사자수의 수정자료와 최종자료 비교

	사업체번호	자영 남	자영 여	자영 합	가족 남	가족 여	가족 합	상용 남	상용 여	상용 합	임시 남	임시 여	임시 합	무급 남	무급 여	무급 합	남자 합	여자 합	전체 합	비고
1	3405006435	1	0	1	0	0	0	13	13	26	0	0	0	0	0	0	**14**	**13**	**27**	동일
		1	0	1	0	0	0	13	13	26	0	0	0	0	0	0	**14**	**13**	**27**	
2	1100604786	1	0	1	0	0	0	9	7	**16**	0	0	0	0	0	0	10	7	17	유사
		0	1	1	0	0	0	9	7	**16**	0	0	0	0	0	0	9	8	17	
3	2608007136	1	0	1	0	0	0	6	3	9	2	8	10	0	0	0	**9**	11	**20**	동일
		1	0	1	0	0	0	6	3	9	2	8	10	0	0	0	**9**	11	**20**	
4	2403024564	0	0	0	0	0	0	66	2	68	0	0	**0**	0	0	0	66	2	68	동일
		0	0	0	0	0	0	66	2	68	0	0	**0**	0	0	0	66	2	68	
5	1104112818	0	0	0	0	0	0	238	100	338	0	0	**0**	0	0	0	238	100	338	동일
		0	0	0	0	0	0	238	100	338	0	0	**0**	0	0	0	238	100	338	
6	3900016606	0	0	0	0	0	0	13	2	15	41	**16**	57	0	0	0	54	18	72	유사
		0	0	0	0	0	0	13	2	15	41	14	54	0	0	0	54	16	70	
7	1100532638	0	0	0	0	0	0	623	405	1028	0	0	**0**	0	0	0	623	405	1028	동일
		0	0	0	0	0	0	623	405	1028	0	0	**0**	0	0	0	623	405	1028	
8	1108060943	0	0	0	0	0	0	45	5	50	0	0	**0**	0	0	0	45	5	50	동일
		0	0	0	0	0	0	45	5	50	0	0	**0**	0	0	0	45	5	50	
9	1105134258	0	0	0	0	0	0	19	13	32	0	0	0	15	14	29	**34**	27	**61**	동일
		0	0	0	0	0	0	19	13	32	0	0	0	15	14	29	**34**	27	**61**	
10	3400105247	0	0	0	0	0	0	103	9	112	6	1	7	0	0	0	**109**	**10**	**119**	동일
		0	0	0	0	0	0	103	9	112	6	1	7	0	0	0	**109**	**10**	**119**	
11	3603021041	0	0	0	0	0	0	59	22	81	4	6	10	0	80	80	63	**108**	**171**	동일
		0	0	0	0	0	0	59	22	81	4	6	10	0	80	80	63	**108**	**171**	
12	1103203381	0	0	0	0	0	0	72	18	**90**	0	0	**0**	0	0	0	72	18	90	동일
		0	0	0	0	0	0	72	18	**90**	0	0	**0**	0	0	0	72	18	90	
13	3100259164	0	0	0	0	0	0	1000	250	1250	0	0	**0**	0	0	0	1000	250	1250	동일
		0	0	0	0	0	0	1000	250	1250	0	0	**0**	0	0	0	1000	250	1250	
14	3702038242	0	0	0	0	0	0	51	12	63	0	0	0	0	180	180	51	**192**	**243**	동일
		0	0	0	0	0	0	51	12	63	0	0	0	0	180	180	51	**192**	**243**	
15	3306000288	0	0	0	0	0	0	11	6	17	5	11	16	0	85	85	16	**102**	**118**	동일
		0	0	0	0	0	0	11	6	17	5	11	16	0	85	85	16	**102**	**118**	

* 위의 숫자는 자동수정자료, 아래의 숫자는 최종자료임
* 대부분의 종사자수 수정자료와 최종자료 일치

3. 사업실적 합계 불일치 오류의 수정

가. 사업실적

2008년 기준 서비스업조사의 사업실적 조사항목 중 매출액은 2008년 1년간 용역 등을 제공하고 획득한 총영업수입액을 나타낸다. 영업비용은 영업활동으로 인하여 지출된 비용이며 재료매입비, 인건비, 임차료, 세금과공과, 감가상각비, 대손상각비, 기타비용 등으로 구성된다. 한편 영업이익은 매출액에서 영업비용을 차감한 금액이다.

나. 내검규칙(Edit rules)

따라서 사업실적 조사항목과 관련된 합계불일치 점검규칙은 다음과 같이 2개의 내검규칙이 주어진다. 첫째, 영업비용(x_2)은 재료매입비(x_3), 인건비(x_4), 임차료(x_5), 세금과공과(x_6), 감가상각비(x_7), 대손상각비(x_8), 기타비용(x_9)의 합계와 같아야 하며, 둘째로 영업이익(x_{10})은 매출액(x_1) - 영업비용(x_2)이어야 한다. 이를 수식으로 표현하면 다음과 같다.

$$e_1 : x_1 - x_2 = x_{10}$$
$$e_2 : x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 = x_2$$

다. 사업실적 합계불일치 오류

매출액에서 영업비용을 뺀 값과 영업이익이 불일치한 건수는 총 12건으로 나타났다. 참고로 영업이익이 음수인 경우는 2,539건이다. <표 2-10>을 보면 레코드 1은 두 개의 내검규칙이 모두 위배되어 오류레코드(FALSE)로 나타나고 레코드 2는 첫 번째 내검규칙이 위배되어 오류레코드로 검출됨을 나타낸다. 첫 번째 내검규칙인 매출액에서 영업비용을 뺀 값이 영업이익과 일치한지에 대한 점검 전에 영업비용이 재료매입비, 인건비, 임차료, 세금과공과, 감가상각비, 대손상각비, 기타비용의 합계와 일치하는 지 점검이 필요하다.

〈표 2-10〉 사업실적의 합계불일치 탐색결과

	e_1	e_2	status
1	TRUE	TRUE	FALSE
2	TRUE	FALSE	FALSE
3	TRUE	FALSE	FALSE
4	TRUE	FALSE	FALSE
5	TRUE	TRUE	FALSE
6	TRUE	TRUE	FALSE
7	TRUE	FALSE	FALSE
8	TRUE	TRUE	FALSE
9	TRUE	TRUE	FALSE
10	TRUE	TRUE	FALSE
11	TRUE	FALSE	FALSE
12	TRUE	FALSE	FALSE

오류자료의 탐색결과에 따른 사업실적 합계불일치 오류자료의 내용은 <표 2-11>과 같다.

〈표 2-11〉 사업실적 합계불일치 오류자료

	사업체 번호	매출액	영업 비용	재료 매입비	인건비	임차료	세금과 공과	감가 상각비	대손 상각비	기타 영업 비용	영업 이익	상태
1	3105026170	4667	77	0	1081	9	62	68	0	0	0	오류
2	2608007136	274	270	0	227	0	0	0	0	43	40	오류
3	1104135616	3800	39	0	0	39	0	0	0	0	0	오류
4	2300062240	10226	8363	5763	1232	0	24	189	0	1155	1862	오류
5	3102028469	8938	5781	2640	2704	33	4	5	0	0	0	오류
6	3501011097	2590	1	0	0	0	0	0	0	0	0	오류
7	1107127668	14200	12780	10020	350	25	11	38	0	2336	-1420	오류
8	1107037871	2570	25700	0	1798	0	0	0	0	0	0	오류
9	3203013247	1651	1072	0	0	19	0	0	0	0	0	오류
10	3202008871	5950	6044	31	4894	48	29	132	0	909	0	오류
11	1102032933	1114	2225	0	1353	0	52	16	0	804	929	오류
12	1108110072	3495	7848	3205	2103	480	25	251	69	1715	-4354	오류

라. 단순오타의 자동수정

프로그램 R의 패키지를 이용하여 등식조건하에서의 단순오타의 자동수정을 실시한 결과가 <표 2-12>에 나타나있다. 총 12건 중 2건(2, 7번째 레코드)의 오류를 단순오타로 인식하였다. 2번째 레코드의 40이 0으로, 7번째 레코드의 -1420이 1420으로 수정되었다. 나머지 레코드는 단순오타에 의한 오류로 보기 어렵고 이때 레코드의 불일치성은 수학적 최적화 문제로 해결한다.

〈표 2-12〉 correctTypos를 이용한 에디팅결과(사업실적)

	사업체 번호	매출액	영업 비용	재료 매입비	인건비	임차료	세금과 공과	감가 상각비	대손 상각비	기타 영업 비용	영업 이익	상태
1	3105026170	4667	77	0	1081	9	62	68	0	0	0	오류
2	2608007136	274	270	0	227	0	0	0	0	43	4	해결
3	1104135616	3800	39	0	0	39	0	0	0	0	0	오류
4	2300062240	10226	8363	5763	1232	0	24	189	0	1155	1862	오류
5	3102028469	8938	5781	2640	2704	33	4	5	0	0	0	오류
6	3501011097	2590	1	0	0	0	0	0	0	0	0	오류
7	1107127668	14200	12780	10020	350	25	11	38	0	2336	1420	해결
8	1107037871	2570	25700	0	1798	0	0	0	0	0	0	오류
9	3203013247	1651	1072	0	0	19	0	0	0	0	0	오류
10	3202008871	5950	6044	31	4894	48	29	132	0	909	0	오류
11	1102032933	1114	2225	0	1353	0	52	16	0	804	929	오류
12	1108110072	3495	7848	3205	2103	480	25	251	69	1715	-4354	오류

마. 수학적 최적화 방법에 의한 자동수정

앞에서 이미 언급한 바와 같이 단순오타의 자동수정은 자료의 모든 불일치를 해결하지는 못한다. 완전한 자동 에디팅을 위해서 나머지 불일치 오류는 선형계획법(LP)의 이용이나 Fellegi-Holt 패러다임에 근거한 오류위치탐색 알고리즘 등에 의해 해결한다.

사업실적에서 영업비용 합계는 영업비용을 구성하는 항목에 의존하고 영업이익은 영업비용 합계 항목에 의존하므로 영업비용을 구성하는 항목 > 영업비용 합계 > 영업이익

의 순으로 신뢰도를 부여하는 것이 합리적이다. 매출액과 영업비용을 구성하는 항목(x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_9)에 신뢰 가중치로 3을 부여하고 영업비용(x_2)에 가중치 2를, 그리고 영업이익(x_{10})은 가중치 1을 설정한다.

이제 최소한의 수정을 제시하는 오류위치포착 문제는 다음과 같다.

$$\min \sum_{j=1}^{10} w_j |\tilde{x}_j - x_j|$$

여기서 수정되는 값 $\tilde{x}_j$는 다음을 만족하여야 한다.

$$\tilde{x}_1 - \tilde{x}_2 - \tilde{x}_{10} = 0$$

$$\tilde{x}_3 + \tilde{x}_4 + \tilde{x}_5 + \tilde{x}_6 + \tilde{x}_7 + \tilde{x}_8 + \tilde{x}_9 - \tilde{x}_2 = 0$$

$$\tilde{x}_j \geq 0, \quad j = 1,...9$$

앞에서와 유사하게, 위의 문제는 y_j^+와 y_j^-를 이용하여 다음과 같은 선형계획 문제로 재표현할 수 있다.

$$\min \sum_{j=1}^{10} w_j \left(y_j^+ + y_j^-\right)$$

제약식:

$$y_1^+ - y_1^- + y_2^+ - y_2^- - y_{10}^+ + y_{10}^- = x_{10} + x_2 - x_1$$
$$y_3^+ - y_3^- + \cdots + y_9^+ - y_9^- - y_2^+ - y_2^- = x_2 - x_3 - \cdots - x_9$$

$$y_j^+ - y_j^- \geq -x_j \, (j = 1, \cdots, 10)$$
$$y_j^+ \geq 0 \, (j = 1, \cdots, 10)$$
$$y_j^- \geq 0 \, (j = 1, \cdots, 10)$$

프로그램 R에서 lpSolve 패키지를 내려 받아 선형계획법(lp) 문제를 해결할 수 있으며 그 수행결과는 <표 2-13>과 같다.

〈표 2-13〉 lpSolve를 이용한 에디팅결과(사업실적)

	사업체 번호	매출액	영업 비용	재료 매입비	인건비	임차료	세금과 공과	감가 상각비	대손 상각비	기타 영업 비용	영업 이익	상태
1	3105026170	4667	1220	0	1081	9	62	68	0	0	3447	해결
2	2608007136	274	270	0	227	0	0	0	0	43	4	해결
3	1104135616	3800	39	0	0	39	0	0	0	0	3761	해결
4	2300062240	10226	8363	5763	1232	0	24	189	0	1155	1863	해결
5	3102028469	8938	5386	2640	2704	33	4	5	0	0	3552	해결
6	3501011097	2590	0	0	0	0	0	0	0	0	2590	해결
7	1107127668	14200	12780	10020	350	25	11	38	0	2336	1420	해결
8	1107037871	2570	1798	0	1798	0	0	0	0	0	772	해결
9	3203013247	1651	19	0	0	19	0	0	0	0	1632	해결
10	3202008871	5950	6043	31	4894	48	29	132	0	909	-93	해결
11	1102032933	1114	2225	0	1353	0	52	16	0	804	-1111	해결
12	1108110072	3495	7848	3205	2103	480	25	251	69	1715	-4353	해결

* 매출액, 재료매입비~기타는 신뢰가중치 3, 영업비용은 2, 영업이익은 1을 부여

F-H 패러다임을 기반으로 하는 네덜란드 통계청의 자동 에디팅 시스템 SLICE 수행결과는 <표 2-14>에 나타내었다. 결과에서 10개의 불일치 자료는 모두 자동수정되었으며 LP방법에 의한 수행결과와 SLICE 수행결과는 정확하게 같은 결과가 나타났다. 역시 두 개의 목적함수가 다르기 때문에 항상 같은 결과가 나오지는 않으나 여기서는 같은 결과가 나왔다.

한편, 만약 SLICE 결과에서 두 개 이상의 해가 나타날 경우, 네덜란드 통계청은 첫 번째 해를 사용하는 단순한 선택기준을 이용하고 있다. 합리적이지 않아 보이지만 다행스럽게 그러한 상황은 많지 않으며 다른 선택기준을 사용해서 얻은 결과와 비교해 본 결과, 어떤 선택기준이 명백하게 좋다는 결과를 얻지는 못했다고 보고하였다.

〈표 2-14〉 SLICE를 이용한 에디팅결과(사업실적)

	Routine	id	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	Input_Record	1	4667	77	0	1081	9	62	68	0	0	0
1	ErrorLocateSolution	1		x								x
1	Adapted_Record	1	4667	1220	0	1081	9	62	68	0	0	3447
2	Input_Record	3	3800	39	0	0	39	0	0	0	0	0
2	ErrorLocateSolution	3										x
2	Adapted_Record	3	3800	39	0	0	39	0	0	0	0	3761
3	Input_Record	4	10226	8363	5763	1232	0	24	189	0	1155	1862
3	ErrorLocateSolution	4										x
3	Adapted_Record	4	10226	8363	5763	1232	0	24	189	0	1155	1863
4	Input_Record	5	8938	5781	2640	2704	33	4	5	0	0	0
4	ErrorLocateSolution	5		x								x
4	Adapted_Record	5	8938	5386	2640	2704	33	4	5	0	0	3552
5	Input_Record	6	2590	1	0	0	0	0	0	0	0	0
5	ErrorLocateSolution	6		x								x
5	Adapted_Record	6	2590	0	0	0	0	0	0	0	0	2590
6	Input_Record	8	2570	25700	0	1798	0	0	0	0	0	0
6	ErrorLocateSolution	8		x								x
6	Adapted_Record	8	2570	1798	0	1798	0	0	0	0	0	772
7	Input_Record	9	1651	1072	0	0	19	0	0	0	0	0
7	ErrorLocateSolution	9		x								x
7	Adapted_Record	9	1651	19	0	0	19	0	0	0	0	1632
8	Input_Record	10	5950	6044	31	4894	48	29	132	0	909	0
8	ErrorLocateSolution	10		x								x
8	Adapted_Record	10	5950	6043	31	4894	48	29	132	0	909	-93
9	Input_Record	11	1114	2225	0	1353	0	52	16	0	804	929
9	ErrorLocateSolution	11										x
9	Adapted_Record	11	1114	2225	0	1353	0	52	16	0	804	-1111
10	Input_Record	12	3495	7848	3205	2103	480	25	251	69	1715	-4354
10	ErrorLocateSolution	12										x
10	Adapted_Record	12	3495	7848	3205	2103	480	25	251	69	1715	-4353

* 매출액, 재료매입비~기타비용은 신뢰가중치 3, 영업비용은 2, 영업이익은 1을 부여

바. 최종자료와의 비교

2008년 기준 서비스업조사의 최종자료와 비교한 결과를 <표 2-15>에 나타내었다. 표에서 본사를 갖는 사업체의 사업실적을 제외하면 사업실적의 합계불일치자료를 자동수정한 결과(위의 숫자)와 서비스업조사 최종자료(아래의 숫자)는 대부분 일치한다. 2번과 6번 레코드에서만 약간의 차이가 나타나고 있다.

〈표 2-15〉 사업실적의 수정자료와 최종자료의 비교

	사업체 번호	매출액	영업비용	재료매입비	인건비	임차료	세금과공과	감가상각비	대손상각비	기타영업비용	영업이익	비고
1	3105026170	4667 17055	1220 15883	0	1081 3483	9 181	62 39	68 653	0 19	0 11508	3447 1172	조사불응 사업체 우편재조사
2	2608007136	274 274	270 260	0 0	227 227	0 0	0 0	0 0	0 0	43 33	4 14	
3	1104135616	3800 18559	39 16330	0 0	0 7781	39 802	0 153	0 1072	0 9	0 6513	3761 2229	본사와 구분 안 됨
4	2300062240	10226 10226	8363 8363	5763 5763	1232 1232	0 0	24 24	189 189	0 0	1155 1155	1863 1863	
5	3102028469	8938 7938	5386 5781	2640 2640	2704 2704	33 33	4 4	5 5	0 0	0 395	3552 2157	본사조사
6	3501011097	2590 2477	0 2176	0 0	0 982	0 15	0 4	0 297	0 6	0 872	2590 301	본사조사
7	1107127668	14200 14200	12780 12780	10020 10020	350 350	25 25	11 11	38 38	0 0	2336 2336	1420 1420	
8	1107037871	2570 11794	1798 11433	0 0	1798 1231	0 486	0 190	0 1749	0 0	0 7777	772 366	본사관리
9	3203013247	1651 1651	19 1072	0 0	0 600	19 19	0 0	0 0	0 0	0 453	1632 579	본사일괄 조사업체
10	3202008871	5950 5950	6043 6044	31 31	4894 4894	48 48	29 29	132 132	0 0	909 910	-93 -94	
11	1102032933	1114 1114	2225 2225	0 0	1353 1353	0 0	52 52	16 16	0 0	804 804	-1111 -1111	
12	1108110072	3495 3495	7848 7848	3205 3205	2103 2103	480 480	25 25	251 251	69 69	1715 1715	-4353 -4353	

제4절 결 론

1. 요약

수학적 최적화 기법은 등식조건 하에서 단순 오타의 속성에 대한 정보를 사용할 수 없다는 약점을 가진다. 따라서 우리는 합계일치조건 하에서의 단순오타의 속성을 이용한 자동수정 알고리즘을 서비스업 통계조사에 적용하였다.

종사자수의 경우 전체 합계불일치 오류 15개 중 6개가 단순오타로 인해 자동수정 되었으며 사업실적의 경우 전체 합계불일치 오류 12개 중 2개가 단순오타로 인식되어 자동수정 되었다. 한편 나머지 합계불일치 오류는 수학적 최적화 기법인 SLICE(네덜란드 통계청 자동에디팅 시스템)를 이용하거나 또는 선형계획 문제를 풀어 자동수정 하였다.

본사를 갖는 지사의 사업실적은 조사상의 어려움으로 인해 사업비용이 0으로 처리된 자료가 많아 이러한 자료는 자동수정 전에 특별한 자료로 간주하여 처리되어야 할 것이다. 사업실적을 본사에서 관리하는 사업체의 경우 매출액이나 영업비용이 부정확하게 입력되어 있다.

특히 결측치는 0이 아닌 결측치로 반드시 구분되어 표기되어야 한다. 오류위치포착에서 결측 변수는 자동적으로 신뢰가중치를 0을 가지게 되어 어떤 값도 대체될 수 있게 된다. 그러나 결측치가 0의 값으로 표기되면 실질적인 0으로 인식하게 되고 신뢰가중치는 0이 아닌 값을 가지게 되어 실질적인 0의 값이냐 아니냐에 따라 그 결과는 달라지기 때문이다.

2. 결언

현실의 통계조사에서는 부정확한 응답 및 입력오류 등으로 종종 오류를 갖게 되나 통계자료 사용자는 통계자료가 완전하고 논리적으로 문제가 없기를 기대한다. 한편, 오류를 수정하기 위해 모든 응답자를 재접촉 할 수 있으나 재접촉은 비용과 시간의 제약으로 항상 가능할 수는 없으며, 게다가 응답자는 조사에 협조하지 않을 수 있다. 오류레코드에 대해 재접촉/재조사가 힘들거나 이를 통해서도 해결되지 않을 때에는 최종적으로 내검규칙을 만족시키지 못한 항목 값은 수정되어야 할 경우가 존재한다.

우리는 최종 내검 단계의 자료에 대해 합계불일치 오류의 자동수정을 실시하고 그 결과를 검토하였다. 단순 오타로 인한 오류를 먼저 제거한 후, 나머지 오류는 항목 신뢰가중치를 갖는 수학적 최적화 기법을 이용하여 자동수정 하였다. 그 결과 합계일치 조건에서의 자동수정은 유용한 것으로 판단된다. 특히 균형 내검규칙 하에서의 단순오타로

인한 오류의 자동수정은 위험성이 적고 구현성이 용이하여 적용가능성이 매우 높다.

특히 재조사 및 재질의가 불가능 할 때, 자동수정은 적어도 최후의 대응책으로서 유용하다. 또한 적용자료는 최종단계의 자료를 이용하였으나, 최초 단계에서부터 자동내검을 실시하면 필수규칙을 위반한 레코드에 대해 자동으로 빠르게 수정위치 확인 및 수정 처리할 수 있으며, 더욱이 원자료로 복구가 가능하다는 장점이 있다.

한편 본 연구결과는 연간 경제조사나 경제총조사에서 유형자산, 무형자산, 연초재고, 연말재고 등의 합계불일치 오류 점검에 확대하여 활용 가능할 것이다. R 프로그램은 인터넷 웹사이트(http://www.r-project.org)에서 무료로 내려 받을 수 있으며 관련 패키지를 다운받아 실행 가능하다. 끝으로 합계 불일치 오류에 대한 명확하고 신속한 대응을 통해 내검 업무를 개선하고 에디팅의 자동화 영역을 넓히는 계기가 되기를 기대한다.

참고문헌

이의규외(2007), “사업체대상 조사의 자동내검기법”, 통계개발원.

이의규외(2008), “자동오류위치포착 및 수정방안”,「통계자료의 내검기법 연구」, 통계개발원.

이의규외(2009), “Fellegi-Holt 기법을 이용한 에디팅의 시도 및 분석”, 응용통계연구 22(4), 697-707.

이의규(2009), “자동내검기법의 적용방안”, 통계개발원.

이의규(2010), “자동내검기법의 적용 및 분석 -서비스업조사를 대상으로-”, 통계개발원.

통계청(2008), “도소매업 및 서비스업통계조사 조사지침서”, 내부자료.

통계청(2008), 「2005년 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 Wall, T.(2003), “Processing of Erroneous and Unsafe Data”, Ph. D. Thesis, Erasmus University Rotterdam.

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

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

Granquist, L.(1997), “The New View on Editing”, International Statistical Review, 65, 3, pp.381-387.

Scholtus, S.(2009), “Automatic Correction of Simple Typing Errors in Numerical Data with Balance Edits", Paper presented at the UNECE Work Session on Statistical Data Editing, Neuchatel..