제6장

경제조사 에디팅 매뉴얼

이의규

제1절 서 론

1. 연구의 필요성

경제조사는 일반적으로 사업체 또는 기업체를 대상으로 월간, 분기, 연간, 그 이상의 주기로 이루어진다. 월간이든 연간이든 경제조사 자료의 사용자는 높은 품질의 자료를 요구한다. 그러나 조사 자료는 종종 오류를 포함하게 된다. 왜냐하면 응답자가 고의든 아니든 잘못된 응답을 할 수도 있고, 질문의 내용을 이해하지 못했거나 대답하기 싫어해서 응답을 안 할 수도 있기 때문이다.

에디팅은 이러한 오류를 찾아내고 수정하는 일련의 활동으로 조사과정에서 필수적 절차이며 전체 조사과정에서 큰 부분을 차지한다. 통상적으로 각 경제조사마다 상세한 업무지침을 작성하여 에디팅 절차를 수행하고 있으나 경제조사 에디팅에 대한 이해와 효율적인 에디팅 방법론에 대한 지식이 충분하다고 말하기 어렵다.

따라서 에디팅에 대한 기본 개념과 에디팅 방법론의 실무 지식을 체계적으로 정리하여 경제조사 에디팅 업무의 효율성과 자료의 품질 향상에 도움이 되도록 경제조사 에디팅 매뉴얼을 작성할 필요가 있다.

2. 연구의 목적

본 연구는 경제조사의 에디팅에 대한 이해를 높이고 에디팅의 원리와 원칙에 입각하여 실무를 향상시키는 데 목적이 있다. 각 경제조사마다 에디팅 업무를 위한 세부적 지

침이 있으나 선진적인(효율적인) 에디팅 방법론에 대한 관심은 크지 않다.

전통적으로 업무담당자는 모든 자료의 완벽한 검토를 목표로 한다. 그러나 이러한 전통적인 에디팅 방식은 여러 연구에 의해 효율적이지 못하다는 것이 증명되었다(그렇다고 에디팅을 소홀하게 해도 좋다는 의미는 아니다). 실제로 신뢰 있는 공표치를 얻기 위해 모든 오류를 잡아낸다는 것이 불필요하다는 것이다. 왜냐하면 통계작성기관의 주요 결과는 전체 자료를 이용하여 집계되며, 표본 자료에 근거하고 있기 때문이다. 즉, 다음과 같은 이유로 모든 오류를 수정하는 것은 불필요하다.

- 개별 레코드의 작은 랜덤 오류는 전체 자료 집계시 상쇄된다.
- 그 조사가 표본조사라면 모든 수집된 자료가 완벽하게 수정되었다 할지라도 여전히 표본오차가 존재한다.
- 고품질 자료를 얻는 것은 가장 영향력 있는 오류를 제거하는 것만으로도 충분하다.

이와 같은 선진 에디팅 방법론의 이해는 에디팅 실무 적용에 있어서 중요한 지침이 될 것이다. 연구 목적은 크게 다음과 같이 두 가지로 요약된다.

- 경제조사 에디팅의 목적과 에디팅의 원칙을 이해한다.
- 에디팅의 방법론을 소개하고 예제를 통해 확인하며 실무에 적용한다.

3. 연구의 절차 및 방법

에디팅은 Fellegi & Holt(1976)의 자동에디팅 논문이 발표되면서 이전의 전통적인 수작업 에디팅에서 이론적으로 크게 발전할 수 있는 토대를 마련하였다. 이후 Granquist(1997)는 과도한 에디팅의 문제를 지적하고 좀 더 효율적인 에디팅의 필요성을 인식하고 지향하였다. 여러 연구결과에 의하면 세세한 에디팅은 투입되는 비용과 시간에 비하여 그 효과가 미비한 것으로 나타났다. 이러한 에디팅의 방법론은 캐나다와 네덜란드에서 큰 발전을 이루었다.

국내에서는 류제복외(2003), 박진우외(2005), 변종석(2007), 김규성(2008), 이기재(2011a, 2011b) 등의 에디팅 관련 논문을 찾아 볼 수 있으나 에디팅이 대규모 반복조사가 이루어지는 국가통계기관의 업무이며 주관적이고 이론적 전개가 적어 학계에서는 이 분야의 연구가 그리 활발하지 않다. 주로 결측치 처리에 대한 논문이 대부분이다.

한편, 통계청에서는 통계조사의 자동 내검[1]에 대한 연구(이의규외, 2007), Fellegi-Holt

1) 통계청에서는 자료의 오류를 찾아 수정하는 에디팅 절차를 내검(내용검토)이라 부름

기법을 이용한 에디팅의 시도 및 분석(이의규외, 2009c)을 통해 자동 에디팅에 대한 도입을 시도하였다. 또한 주기적 조사자료의 내검(이의규, 2010c)을 통해 매크로 에디팅 방법을 적용하였으며 선형계획법을 이용한 자동내검(이의규, 2009d, 2010a)을 서비스업통계조사 자료에 응용하였다. 최근에는 합계불일치 오류의 자동수정(이의규, 2011)을 발표한 바 있다.

본 매뉴얼 작성을 위해 그간의 에디팅 관련 연구결과(참고문헌 참조)와 국내외 에디팅 관련 문헌을 참고하였다. 캐나다 통계청의 Survey methods and practices(2003)와 호주통계청(ABS)[2]의 The editing guide(2007), 에디팅 품질관리 매뉴얼(2008)을 주로 참조하였다. 본 매뉴얼은 특히 경제조사에 초점을 맞추고, 품질관리 측면보다 에디팅 방법론적 실무 적용 측면에 중점을 두고 있음이 특징이다.

참고로 캐나다 통계청의 Survey methods and practices(2003)에서 에디팅 지침서와 호주 통계청의 에디팅 원칙을 발췌하여 서두에 소개한다.

- 에디팅은 전문적 지식을 가진 직원에 의해 수행되어야 함
- 에디팅은 조사의 여러 단계에 걸쳐 수행되어야 함
- 매 단계마다 적용되는 에디팅이 다른 단계에서 수행되는 에디팅과 충돌해서는 안 됨
- 조사의 품질 측정, 조사과정에 대한 정보를 제공하는 것으로 사용되어야 함
- 에디팅중 가정의 타당성을 검증하고 향후 에디팅 수정과 조사표 설계를 향상시키기 위해 사용되어야 함
- 수행된 에디팅 형식에 대한 정보와 조사 자료에 영향을 미치는 에디팅은 사용자들 사이에서 공유되어야 함
- 품질보증과 품질관리 과정은 에디팅 과정에서 드러난 오차를 최소화하고 교정할 수 있게 적용되어야 함

2) 호주통계청(ABS)은 에디팅의 개념, 내용, 원칙, 과정 등에 대한 매뉴얼을 마련하여 에디팅 실시하는데 에디팅은 본청 실사과 직원이 주로 담당(지방사무소에서는 조사표 스캔, 기본 코딩을 통해 입력을 하며 본청에서는 기본 자료가 정제된 후 이상치 검출 및 처리를 실시)

【ABS의 에디팅 원칙】

- 모든 수집 자료에 대한 에디팅 전략을 개발할 것 (Develop an editing strategy for every collection)
- 오류 예방을 실천할 것 (Practice error prevention)
- 중대한 예외 데이터에 초점을 둘 것 (Focus on important anomalies)
- 매크로와 마이크로 에디팅을 적절히 활용할 것 (Achieve the appropriate balance between micro and macro editing)
- 실제를 가장한 결과에 치중하지 말 것 (Don't force an outcome that makes the real world situation)
- 에디팅 처리에 따르는 제공자에 대한 부담을 관리할 것 (Manage the burden on respondents associated with resolving edits)
- 자동 및 수작업 절차를 적절히 활용할 것 (Achieve the appropriate balance between automatic and clerical processes)
- 최상의 업무관행에 뒤쳐지지 않도록 하고 지속적인 개선을 추구할 것 (Keep abreast of best practice, and apply continuous improvement)
- 결과물에 기대 및 정확도 요구 정도를 파악할 것 (Understand your output needs and precision requirements)
- 데이터의 특성을 파악할 것 (Understand characteristics of your data)
- 시간, 인력, 재원, 자원 및 시스템 제약을 파악할 것 (Understand your time, personnel, funding, and systems constraints)
- 에디팅 담당직원의 업무능력을 파악하고 역량을 강화할 것 (Understand and build skills of your editing staff)
- 품질 관련 정보를 수집·분석하고 공유할 것 (Capture, analyse and share information about quality)

4. 매뉴얼의 구성

본 매뉴얼은 에디팅의 입문서보다 실제적이고 구체적인 내용을 담고자 하였다. 따라서 경제조사별 업무지침을 분석하여 공통적으로 적용 가능한 방법론을 제시하였다. 좀 더 통계적 이론을 바탕으로 하여 동일 조건에서 균일한 작업이 이루어질 수 있고 경제조사에 모두 사용될 수 있으며, 비용이 많이 들지 않고 적시에 공표할 수 있는 에디팅 방법론을 소개하고자 하였다.

본 에디팅 매뉴얼은 제목에서 처럼 수치자료를 갖는 경제조사에 초점을 둔다. 제2절에서 에디팅의 전반적인 이해를 돕기 위한 에디팅의 개요 및 원칙에 대해 설명하고, 제3절에서 실제 에디팅 업무에 적용할 수 있는 에디팅의 방법론과 예제를 제시한다. 제4절에서 이상치의 탐색과 처리에 대해 살펴보고, 제5절에서 매뉴얼 작성에 대한 맺음말로 마친다.

제2절 에디팅 개요

1. 데이터 에디팅

가. 에디팅의 정의

에디팅은 조사의 전 과정에서 오류나 의심스러운 자료를 탐색하고 필요한 경우 수정하는 활동을 말한다. 여기서 중요한 것은 모든 오류가 식별되거나 수정되어야 하는 것이 아니라는 점이다. 에디팅은 매우 자원 집약적인 절차이기 때문에 모든 오류를 점검하는 것은 불가능하다. 그렇다면 에디팅은 어느 수준까지 해야 하는가? 에디팅 지침서에는 시간과 비용 등 자원의 합리적 사용 관점에서 판단할 것을 권고한다. 즉 전체조사 오차에 초점을 두고 자원을 절감하여 이를 조사 과정 개선에 자원을 이동하는 것이 필요하다.

나. 에디팅의 필요성

통계작성기관에서 에디팅은 없어서는 안 될 중요한 과정으로 인식되고 있다. 이는 데이터 에디팅이 다음과 같은 이유로 필요하기 때문이다.

- 통계자료가 목적에 부합(fit for purpose)하는지를 평가
- 통계자료의 정확성과 신뢰성 유지
- 적은 자원으로 많은 성과 달성
- 비용 및 시간 절감

다. 에디팅의 목적

에디팅의 목적은 누락과 무효 자료에 대한 확인 및 수정을 통해 자료의 신뢰성을 확보함으로써 조사결과의 편의를 감소시키고 통계자료의 품질을 향상시키는 것이다.

- 품질제공 : 통계 자료의 품질에 대한 정보 제공
- 조사개선 : 무응답 및 오류 패턴을 검토하여 향후 조사품질 향상에 기여하는 등 조사개선을 위한 정보 제공
- 자료정제 : 오류나 이상치 탐색, 처리 결정, 결측치 대체(imputation)를 통해 자료의 신뢰성 확보

라. 에디팅의 종류

에디팅은 에디팅의 대상에 따라 마이크로(미시적) 에디팅(micro editing)과 매크로(거시적) 에디팅(macro editing)으로 구분한다. 마이크로 에디팅은 개별 레코드의 검사를 통하여 자료가 논리에 맞는지 확인하고 교정하는 것이 목적이다. 예를들면, 전체 자료를 고려하는 것이 아니라 개별 레코드 내에서 항목간 연관규칙으로 점검하는 방식이다. 이는 입력단계에서 수행되어지므로 입력자료 에디팅(input data editing)이라고도 한다.

매크로 에디팅은 출력자료 에디팅(output data editing)이라고도 하는데 전체 자료의 분석을 통하여 자료에 문제가 없는 지를 확인하는 것이다. 매크로 에디팅은 자료의 수집이 이루어진 시점에서 주로 시행된다. 예를 들면 적합한 결과로부터 통계 모형의 이상치를 탐색하거나 산점도 등 그래프를 통해 이상치를 점검하는 절차를 말한다.

한편 작업방식에 따라 수작업 에디팅(manual editing), 쌍방향 에디팅(interactive editing), 자동 에디팅(automatic editing)으로 구분한다. 말 그대로 수동 에디팅은 전적으로 사람의 힘에 의지하는 에디팅이다. 쌍방향 에디팅은 컴퓨터의 도움을 받아 자료의 오류를 찾아내고 오류 확인이나 수정은 에디팅 인력에 의해 이루어지는 형태이다. 그리고 자동 에디팅은 에디팅 절차가 컴퓨터 프로그램에 의해 수행되는 방식이다. 에디팅 종류를 정리하면 아래와 같다.

- 마이크로 에디팅 : 개별 레코드 수준에서 유효성과 일관성 점검
  - 단위 레코드 분석을 통한 예외 자료를 식별하고 후속 결정 및 처리
- 매크로 에디팅 : 자료가 총계 수준에서 합리적인지를 점검
  - 집합 레코드 분석을 통한 예외 자료를 식별하고 후속 결정 및 처리

- 수작업 에디팅 : 말 그대로 전적으로 사람의 힘에 의지하는 에디팅
- 쌍방향 에디팅 : 컴퓨터의 도움으로 자료의 오류를 찾아내고 오류 확인이나 수정은 에디팅 요원에 의해 이루어진 형태
- 자동 에디팅 : 에디팅 절차가 컴퓨터 프로그램에 의해 수행되는 에디팅

마. 오류 점검형태의 종류

자료 에디팅 시 일반적으로 다음과 같은 5가지의 오류에 대해 점검한다.

- 타당성(legality) 점검 : 유효자료 여부를 점검
- 일치성(consistency) 점검 : 항목간 논리적 관계의 성립 여부를 점검
- 범위(range) 점검 : 자료가 허용 가능한 범위 내에 있는지를 점검
- 통계적(statistical) 점검 : 통계적 기법이나 모형을 통해 이상치를 점검
- 주관적(subjective) 점검 : 담당자의 경험이나 눈대중으로 자료를 점검

바. 오류의 종류

자료 오류는 다음과 같이 심각한 오류(fatal error)와 의심되는 오류(query error)로 구분할 수 있다. 의심되는 오류는 비용을 고려하여 점검을 결정하여야 한다. 따라서 점검이 반드시 필요하거나 매우 의심되는 자료로 판단할 기준이 필요하다. 특히 추정결과에 현저하게 영향을 미칠 가능성이 있을 때에만 수정해야 한다.

- 심각한 오류(fatal error)
  - 분명한 오류
  - 유효하지 않은 자료, 결측 자료, 모순되는 자료
  - 반드시 규명해야 하는 오류
- 의심되는 오류(query error)
  - 실수가 있다고 추정되는 오류
  - 주관적인 범위 밖의 자료

한편 데이터의 오류는 오류 원인에 따라 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.

- 체계적 오류(systematic error)
  - 응답자에 의해 일관적으로 보고되는 오류
  - 단위 측정오류
  - 일관적인 이해 부족 또는 잘못 해석하여 응답하는 오류
- 랜덤 오류(random error)
  - 비구조적인 문제로 야기된 오류
  - 우연적으로 발생되는 오류

랜덤오류를 해결하는 일반적 접근 방법으로 Fellegi-Holt 방법을 사용한다. 이 방법은 모든 에디팅 규칙을 만족시키는 최소 변수집합을 찾아 수정하는 방법이다.

사. 에디팅 규칙

하나의 레코드가 오류가 없는 지를 판단하기 위해 에디팅 규칙이 사용된다. 오류 레코드를 탐색하기 위한 에디팅 규칙은 주로 업무담당자에 의해 설정된다. 에디팅 규칙은 다음과 같은 형태로 구분할 수 있다.

- 필수 규칙(fatal edit, hard edit)
  - 반드시 수정되어야 할 자료를 검토하는 규칙
  - 예 : 매출액 > 0
- 의심 규칙(query edit, soft edit)
  - 오류 가능성이 있는 의심스러운 자료를 검토하는 규칙
  - 예 : 영업이익 > 0
- 균형 규칙(equality edit)
  - 항목의 합이 총계와 일치하는 지를 검토하는 규칙, 균형규칙은 필수 규칙
  - 예 : 매출액-영업비용=영업이익
- 비 규칙(ratio edit)
  - 두 변수의 비가 어떤 값보다 작거나 커야 하는 규칙
  - 예 : $\frac{\text{영업이익}}{\text{매출액}} \leq 0.5$

필수규칙은 반드시 해결되어야 할 점검사항이고 선택규칙[3]은 다소 의심스러운 값의 범위에 있을 때 점검을 권고하는 규칙이다. 지나치게 많은 선택 규칙은 업무 부담으로

새로운 오류를 유발할 수 있으며 과도한 에디팅을 유발하기 때문에 너무 많은 선택규칙을 설정하지 않도록 유의하여야 한다.

아. 에디팅의 단계별 활동

언제 에디팅을 실시하는가? 에디팅은 조사의 전 과정에서 이루어진다. 초기 수집단계에서부터 마지막 공표 때까지 수행되는 절차이다. 좋은 자료를 수집하기 위해서는 인터뷰가 이루어지는 단계에서 자료의 점검이 이루어져야 한다. 그러나 조사의 비용 및 시간의 제약이 존재한다. 통상적으로 에디팅은 이미 자료가 수집된 후 오류를 찾아내고 수정하는 절차로 인식된다. 에디팅 단계는 크게 면접 중, 자료입력 저장 전, 자료입력 저장 중, 자료입력 저장 후, 공표 전으로 나눌 수 있다.

자료 에디팅은 자료수집・입력과정에서 수행하는 에디팅과 자료입력 후의 에디팅으로 구분할 수 있다. 수집과정 중에서 수행하는 에디팅은 조사표 에디팅과 입력 중 에디팅이다. 조사표 응답시에 조사원이 에디팅을 수행한다. 적절한 에디팅을 위해서는 미리 에디팅 규칙을 숙지할 수 있도록 조사지침서에 수록하여야 한다. 한편, 컴퓨터 입력이나 스캐닝을 통해 자료를 입력하면 에디팅 규칙을 프로그램으로 작성하여 레코드마다 입력 중 에디팅을 실시한다.

입력 저장 후의 에디팅은 입력된 자료가 대부분 취합된 상태에서 에디팅을 실시한다. 먼저 우연적이 아닌 체계적 오류(systematic error)를 찾아 처리한다. 단위 측정오류와 같이 체계적 오류는 범위 점검이나 비율 점검을 사용하여 점검할 수 있다. 체계적 오류가 발생하면 조사표, 조사원 교육, 입력과정 등을 개선하여야 한다. 두 번째로, 영향력 있는 자료(influential error)를 찾아낸다. 추정치에 큰 영향을 주는 레코드를 찾아내고 이를 추적하여 조치한다. 마지막으로 매크로 에디팅을 적용한다. 이 단계에서는 이상치(outlier)를 탐색한다. 그래픽 에디팅 기법을 이용하여 이상치를 탐색하기도 한다.

- 면접중 에디팅 : 가장 바람직하나 면접시간이 길어지는 단점이 있음(CAI, 자발적 에디팅 조사표 설계 사용)
- 자료입력 저장 전의 에디팅 : 조사표 점검 등
- 자료입력 저장 중의 에디팅
  - 실시간 대화형 소프트웨어 활용 가능
  - 오류메시지가 나오면 입력과정을 멈추고 조치(수용, 임의 수정, 향후조치)가 있어야 입력이 계속된다는 단점이 있음

3) 통계청에서는 OK error라 부르는 오류를 점검하는 규칙에 해당

- 자료입력시의 에디팅은 자료입력 속도를 떨어뜨리므로 최소화하여야 함(타당성 점검과 간단한 일관성 점검 등)

- 자료입력 저장 후의 에디팅
  - 대부분의 에디팅이 이 시점에서 일어남
  - 항목간 일치성 점검, 이상치 탐색, 선별적 에디팅 등 에디팅 방법론이 적용
- 결과물 에디팅
  - 조사 공표 전의 에디팅 이전조사와 조사결과 비교

자. 에디팅의 영향

Granquist와 Kovar(1997)는 그들의 연구에 의하면 의심가는 자료의 에디팅 50%는 최종 총계 추정에서 효과가 거의 없거나 작다고 주장하였다. Hedlin(1993)은 에디팅 과정 중에 실시한 50%의 수정이 최후 수정값에 1% 미만의 변화를 나타냈다고 보고한 바 있다. 따라서 모든 자료를 에디팅하는 대신 영향력이 있는 자료를 선별하여 에디팅하는 선별적 에디팅(selective editing)의 사용이 필요하고 이를 통해 비용을 많이 들이지 않고 에디팅을 수행하고 있다.

또한 연구에 따르면 5%의 에디팅이 전체결과의 95%를 변화시키며, 의심스러운 값의 에디팅 적중률은 20~30%정도라고 한다. 한편, 관측값의 5~10% 수준에서 에디팅하는 것으로 충분하며 응답자 재조사는 원인 파악이 가능하다는 장점이 있으나 비용 부담이 크게 증가한다는 단점이 있다.

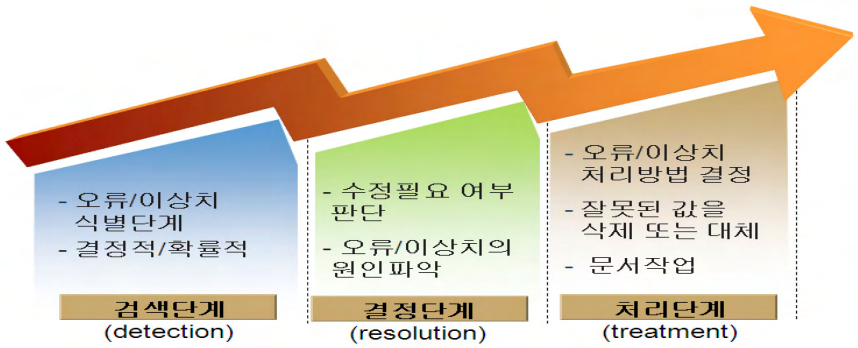
이러한 과도한 에디팅(over-editing)의 이유는 자료수집 이후에 품질문제를 다루는 것이 상대적으로 쉽게 느껴지기 때문이며 사전보다 사후에 오류를 고치는 것이 선호되기 때문이다. 과도한 에디팅은 다음과 같은 폐해가 있으므로 향후 에디팅 업무에서 선별적 에디팅, 자동 에디팅, 그래픽 에디팅 등 효율적인 에디팅 방법을 도입하여 과도한 에디팅을 방지해야 할 것이다.

- 무시해도 좋을 오류를 지나치게 에디팅하는 것은 자원 낭비
- 공표 지연, 시의성 및 적시성 저하
- 수정 후 새로운 오류 발생이 가능하며 참값을 왜곡시킬 수 있음
- 응답자 부담을 증가시키고 다음 조사를 악화시킬 가능성이 있음
- 투입한 비용 대비 그 효과가 미비함

2. 데이터 에디팅 절차 및 전략의 수립

가. 데이터 에디팅의 절차

에디팅의 과정은 오류레코드와 오류 레코드의 오류 필드를 찾는 단계와 오류 필드를 맞는 값(최소한 더 나은 값)으로 수정하거나 결측치를 추정값으로 대체하는 단계로 나눌 수 있다. 좀 더 세분하면, 에디팅은 검색, 결정, 처리의 3단계로 나눌 수 있다. 검색 단계에서는 이상치를 식별하는 단계이며, 결정 단계는 수정필요여부를 판단하고 이상치의 원인을 파악하는 단계이며, 처리 단계는 이상치 처리방법 결정, 이상치 정정, 그리고 문서화하는 단계이다.



[그림 6-1] 데이터 에디팅 절차

통계품질진단 에디팅 과정에서 오류가 발생한 경우에는 다음과 같은 규칙을 적용하여 수정한다.

- 단위 응답값 전체가 믿을 수 없는 수준인 경우에는 해당 데이터 버림
- 다른 보조정보나 다른 항목의 응답값에 의해 오류가 명확하게 고쳐질 수 있는 경우에는 수정-연역적 대체(deductive imputation)
- 수정이 적절치 않은 항목의 오류값은 무응답 처리
- 무응답에 대하여 대체(imputation)하는 것이 적절하다고 판단될 때에는 대체지침에 따라 처리

나. 데이터 에디팅 전략의 수립

데이터 에디팅의 전략을 수립하여 수행할 시 다음과 같은 이점이 있다(ABS, 2007).

- 직원의 근무환경 개선
- 효율성 제고
- 데이터 이해능력 제고
- 목적에 부합하는 품질 유지

호주 통계청(ABS)은 에디팅 전략 수립 시 다음과 같은 고려사항을 제시하고 있다.

- 고객요구사항
  - 자료수집 주요목적, 주요고객과 데이터의 이용, 주요 산출물
  - 데이터 항목의 우선순위, 내부적 우선순위, 산출물 우선순위
  - 정확성/품질 요구사항, 시의성 요구사항
- 운영상 제약요소
  - 자원의 이용가능성, 주요일정, 시스템, 업무부담
- 에디팅 대상 데이터의 서술
  - 데이터의 특성, 비교데이터 소스, 데이터에 대한 예상
- 마이크로 에디팅 과정과 방법
  - 계획과 준비, 초기 에디팅, 선택적 에디팅, 자동에디팅 및 임퓨테이션
- 매크로 에디팅 과정과 방법
  - 계획과 준비, 데이터 평가, 공표를 위한 데이터 승인
- 시행에 영향을 주는 다른 요소들
  - 이해관계자의 책임, 신규사원의 모집과 훈련, 시스템 요구사항, 업무절차 평가
- 시간계획표의 운영

3. 데이터 에디팅 결과표의 작성

캐나다 통계청은 사업체조사에서 각 응답 레코드에 에디팅 규칙을 적용하고 각 에디팅 규칙에 대해 합격, 결측, 불합격 레코드를 분석하고 있다(Banff, 2007). 이 보고서는 5개의 요약표를 소개하고 있는데 에디팅 규칙을 조정하는데 이들을 사용하고 있다.

요약표에 대한 간략한 설명과 예로서 2010년 농업총조사를 위한 4차 시험조사의 내검 수행 전 조사자료의 분석 결과를 제시한다. 조사 가구는 총 3,526가구이며 각 레코드

는 총 350개의 필드로 구성되어 있다. 여기서는 경지, 노지작물, 시설면적, 시설작물, 버섯, 과수, 시군구 작물, 친환경 재배 작물, 가축 및 목초지, 농업 판매 및 경영형태, 생산자조직 참여현황에 관한 항목에 대하여 95개의 내검지침서의 내검규칙을 이용하여 다음 소개되는 5개의 요약표 중 합격/결측/불합격 레코드 요약, 내검규칙별 불합격 빈도표, 불합격 내검규칙수의 분포표 3가지를 제시한다.

가. 각 에디팅 규칙에 대한 합격/결측/불합격 레코드의 수

각 에디팅 규칙에 대해 합격, 결측, 불합격한 레코드의 수를 요약하여 표로 작성한다. 이 표로부터 어떤 에디팅 규칙이 다른 에디팅 규칙보다 더 빈번하게 레코드가 불합격하거나 결측되는 경향이 있음을 결정할 수 있어 직관적으로 매우 유용하다.

<표 6-1>을 보면 특이하게 13번 에디팅 규칙에 위배된 경우가 1,726건으로 전체 레코드의 49%, 약 절반이 위배된 것으로 나타났다. 13번 에디팅 규칙은 “논벼면적이 있는데 유기비료 사용 면적이 없는 경우”로 선택적인 에디팅 규칙이다. 이러한 결과를 선택적인 에디팅 규칙으로 점검하는 것은 그리 적절하지 않은 것으로 보인다.

〈표 6-1〉 각 에디팅 규칙별 합격/결측/불합격 빈도

에디팅 규칙	오류코드	합격	결측	불합격	에디팅 규칙	오류코드	합격	결측	불합격
1	AB001DISIIC	3,474	0	52	49	AI004DISIIT	3,526	0	0
2	AB002D0SITT	3,517	0	9	50	AI005D0SIMT	3,329	0	30
3	AB003D0SITT	3,489	0	37	51	AJ001D0SITT	3,526	0	0
4	AB004DISIIC	3,457	0	69	52	AL012D0SIMT	3,494	0	32
5	AB011DISIIC	3,449	0	77	53	AL013D0SIMT	3,513	0	13
6	AB012D0SITT	3,525	0	1	54	AL014D0SIMT	3,525	0	1
7	AC001D0SITT	3,499	0	27	55	AL016D0SIMD	3,516	0	10
8	AC002D0SITT	3,511	0	15	56	AL017D0SITT	3,516	0	10
9	AC011D0SIMD	3,442	0	84	57	AL019D0SITT	3,471	0	55
10	AC012D0SITT	3,509	0	17	58	AL020D0SITT	3,522	0	4
11	AC021D0SITT	3,502	0	24	59	AL021D0SIMT	3,495	0	31
12	AC022D0SIMD	3,445	0	81	60	AL022D0SIMT	3,522	0	4
13	AC031D0SITT	1,800	0	1,726	61	AL023D0SITT	3,524	0	2
14	AC032D0SIMD	3,484	0	42	62	AL024D0SIMT	3,513	0	13
15	AC041D0SITT	3,511	0	15	63	AL026D0SIMT	3,523	0	3
16	AC042D0SITT	3,526	0	0	64	AL027D0SIMT	3,526	0	0
17	AC043D0SITT	3,517	0	9	65	AL028D0SIMT	3,524	0	2
18	AC044D0SIMT	3,513	0	13	66	AL029D0SITT	3,525	0	1
19	AC051D0SITT	3,511	0	15	67	AL030D0SITT	3,525	0	1
20	AC052D0SITT	3,525	0	1	68	AL031D0SITT	3,526	0	0

21	AC053D0SITT	3,523	0	3	69	AL032D0SITT	3,526	0	0
22	AC054D0SIMT	3,523	0	3	70	AL033D0SITT	3,525	0	1
23	AC061D0SITT	3,526	0	0	71	AL034D0SITT	3,525	0	1
24	AC062D0SITT	3,518	0	8	72	AL035D0SITT	3,409	0	117
25	AC063D0SITT	3,510	0	16	73	AL037D0SITT	3,488	0	38
26	AC064D0SIMT	3,507	0	19	74	AL038D0SITT	3,320	0	5
27	AC071DISRCT	3,526	0	0	75	AL039D0SITT	3,525	0	1
28	AC072D0SITT	3,521	0	5	76	AL040D0SITT	3,526	0	0
29	AD001DISIIC	3,437	0	89	77	AL041D0SITT	3,526	0	0
30	AD002D0SIMT	3,524	0	2	78	AL042D0SITT	3,526	0	0
31	AD003D0SIMT	3,483	0	43	79	AL043D0SITT	3,525	0	1
32	AE001D0SITT	3,519	0	7	80	AL044D0SITT	3,523	0	3
33	AE011D0SITT	3,521	0	5	81	AL052D0SIMT	3,490	0	36
34	AF001D0SITT	3,525	0	1	82	AL053D0SIMC	3,501	0	25
35	AF011D0SITT	3,526	0	0	83	AL054D0SITT	3,504	0	21
36	AF012D0SIMT	3,520	0	6	84	AL061D0SIMT	3,471	0	55
37	AG001DISIID	3,525	0	1	85	AL062D0SIMT	3,525	0	1
38	AG002DISIIT	3,526	0	0	86	AL063D0SIMT	3,526	0	0
39	AG003DISIIT	3,492	0	34	87	AL064D0SIMT	3,526	0	0
40	AG004DISIIT	3,526	0	0	88	AL065D0SIMT	3,524	0	2
41	AG005D0SITT	3,523	0	3	89	AL066D0SIMT	3,526	0	0
42	AG006D0SIMT	3,458	0	68	90	AL067D0SIMT	3,515	0	11
43	AG011D0SIMD	3,487	0	39	91	AM002D0SITT	3,523	0	3
44	AH001DISIIT	3,524	0	2	92	AM003D0SITT	3,526	0	0
45	AH002DISIID	3,523	0	3	93	AM004D0SIMT	3,526	0	0
46	AI001D0SITT	3,511	0	15	94	AM005D0SIMT	3,526	0	0
47	AI002D0SIMT	3,522	0	4	95	AM006D0SIMT	3,526	0	0
48	AI003DISIIT	3,514	0	12					

나. 합격/결측/불합격 에디팅 규칙수의 분포

이 요약표는 합격, 결측, 불합격된 에디팅 규칙 도수 분포를 보여준다. 하나의 에디팅 규칙도 합격(결측, 불합격)되지 않거나 한 번, 두 번, ..., 전체의 에디팅 규칙에 합격(결측, 불합격)한 레코드의 빈도로 구성된다. 각 레코드는 각 열에서 한 번 씩은 들어가므로 각 합격, 결측, 불합격 열의 총수는 레코드의 수와 같다. 이 표는 모든 에디팅 규칙을 합격하는 레코드의 수와 소수 개의 규칙을 위반하는 레코드의 수가 적절하게 구성되고 있는지 그리고 모든 에디팅 규칙을 합격하는 레코드가 대부분 구성되고 있는지, 모든 에디팅 규칙을 위반하는 레코드의 수가 희소하게 구성되고 있는지 등을 검토할 수 있다.

<표 6-2>를 보면 불합격된 에디팅 규칙이 하나도 없는 레코드가 1,228로 앞의 표에서 본 바와 같다. 그런데 불합격 레코드 2,298건 중 1개의 에디팅 규칙을 위배한 경우가

1,806건으로 78.6%이고 2개 이하의 에디팅 규칙을 위배한 경우는 2,085건으로 불합격 레코드 중 90.7%로 나타났으며 95개의 에디팅 규칙 중 9개를 위배한 레코드는 4건이 존재한다(10개 이상을 위반한 경우는 없음). 이 결과에서 보면 대부분이 1-2개의 에디팅 규칙을 위배하고 있어 조사 자료의 정도(精度)가 상당히 높음을 알 수 있다.

〈표 6-2〉 불합격된 에디팅 규칙수의 분포

에디팅 규칙수	0	1	2	3	4	5	6	7	8	9	10
레코드(3,526)	1,228	1,806	279	105	49	24	18	8	5	4	0

다. 합격/결측/불합격 레코드의 수

이 표는 매우 기본적인 표로서 합격, 결측 및 불합격 레코드의 수를 요약한다. 합격, 결측 및 불합격된 레코드의 수를 더하면 당연히 총 레코드의 수가 되므로 전체 레코드에서 얼마나 많은 레코드가 합격인지, 불합격인지, 그리고 결측된 항목 값을 구성하고 있는지를 쉽게 파악할 수 있다. 여기서 불합격 레코드의 수는 하나 이상의 점검규칙을 위반하는 경우 뿐만 아니라 하나 이상의 결측된 값을 동시에 포함할 수 있음을 유념하여야 한다.

<표 6-3>에서 전체 3,526개의 레코드 중 1,228개가 95개의 에디팅 규칙 모두를 통과하였다. 즉 34.8%의 통과율을 보였다. 불합격 레코드는 나머지 2,298개(약 65.2%)로 나타난다.

〈표 6-3〉 합격/결측/불합격 레코드 수

구분	레코드 수	백분율(%)
합격된 레코드	1,228	34.8
결측이 있는 레코드	0	0.0
불합격된 레코드	2,298	65.2
전체	3,526	100.0

라. 합격/결측/불합격 에디팅 규칙에 포함되는 변수의 총수

이 표는 합격, 결측 및 불합격될 때 에디팅 규칙에 포함된 각 항목의 빈도를 변수별로 보여주는 표이다. 추가로 해당 변수를 포함하지 않은 에디팅 규칙의 변수의 빈도와 해당 변수를 포함하는 에디팅 규칙의 수를 도표화 한다. 이 표로부터 어떤 변수가 다른 변수보다 불합격하거나 결측되는 에디팅 규칙에 포함되는 경향이 있는지 판단할 수 있다.

마. 레코드 결과에 영향을 미치는 각 변수의 합격/결측/불합격 수

레코드의 최종결과에 영향을 미치는 각 변수의 빈도를 보여주는 표이다. 최종 레코드의 상태가 합격이면 레코드는 모든 에디팅 규칙을 통과하고 모든 항목은 합격이다. 만약 최종 상태가 결측이라면 에디팅 규칙 상태값이 결측과 합격이 가능하다. 결측 상태를 갖는 에디팅 규칙에 항목이 포함되면 결측이고 결측 상태를 갖지 않는 항목은 적용불가로 카운트한다. 최종 상태가 불합격이라면 적어도 하나의 에디팅 규칙은 불합격이고 하나 이상의 에디팅 규칙상태가 결측일 수 있다. 이때에는 불합격과 관련된 에디팅 규칙을 갖는 항목은 불합격이고 그렇지 않은 항목은 적용불가로 카운트한다. 각 레코드는 각 항목에 대해 한 번 씩 더해지므로 행 합계는 레코드의 총수가 된다. 이 표로부터 어떤 변수가 다른 변수보다 결측/불합격 레코드에 영향을 미치고 있는지를 결정할 수 있다.

위와 같이 에디팅 규칙이 자료에 적용될 때 실패율을 관찰함으로써 에디팅 규칙의 적절성을 평가할 수 있다. 하나의 특별한 에디팅 규칙에 대해 높은 실패율은 에디팅 규칙의 검토가 요구된다. 또한 한 항목이 높은 에디팅 실패율에 포함된다면 현장조사 수집절차에서 검토가 필요하며 조사표의 개선으로 이어질 수 있다. 한편, 에디팅 결과 요약표는 얼마나 많은 자료가 수정되어야 하고 대체되어야할 지를 사전에 추정할 수 있게 해준다.

4. 경제조사 에디팅 업무의 검토

통계청의 경제조사는 5년에 한 번 실시하는 경제센서스, 매년 조사하는 건설업조사, 광업제조업 조사, 기업활동조사, 농어업법인조사, 도소매업조사, 서비스업조사, 사업체기초통계조사, 운수업조사, 전문과학기술서비스조사와 매월 동향파악을 목적으로 하는 건설경기동향조사, 광업제조업동향조사, 기계수주동향조사, 사이버쇼핑동향조사, 서비스업동향조사(도소매업부문, 서비스업부문)가 있으며 전자상거래동향조사는 분기에 한 번씩 조사된다(<표 6-4> 참조).

이와 같은 경제조사에서는 종사자수, 급여액, 매출액, 영업비용 등 사업실적이 주요항목이다. 에디팅에는 유효성 점검, 범위 점검, 일치성 점검, 수준 점검 등이 종합적으로 이루지고 있다. 중요도 점수와 같은 객관적 수치의 우선순위 부여 및 이에 따른 선별적 에디팅이 필요하다. 의심되는 자료의 상하한 값이 주관적 경험적으로 설정되고 있어 자료의 분포에 따른 이상치 탐색방법을 적용하는 것이 필요하며, 그래픽 에디팅 방법으로 데이터의 특성을 이해하고 시각화를 함으로써 합리적인 에디팅 업무량을 결정할 필요가 있다.

〈표 6-4〉 통계청 조사주기별 경제조사

총조사	연간조사	월간조사
경제센서스	건설업조사	건설경기동향조사
	광업제조업조사	광업제조업동향조사
	기업활동조사	기계수주동향조사
	농어업법인조사	사이버쇼핑동향조사
	도소매업조사	서비스업동향조사(도소매업부문)
	서비스업조사	서비스업동향조사(서비스업부문)
	사업체기초통계조사	전자상거래동향조사_분기
	운수업조사	
	전문과학기술서비스조사	

통계청에서는 각 조사마다 입력 내검 프로그램을 통해 조사표를 입력하게 되는데 담당자가 에디팅 규칙을 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 컴퓨터를 이용한 자동화된 방식으로 진행되나 수정은 전적으로 조사원 또는 에디팅 요원 등 사람의 힘에 의해 수행되고 있다.

부언하면, 통계청의 입력 내검 프로그램은 수작업의 용이한 수정을 위해 위배된 규칙에 대응되는 오류코드를 알려주는 것이지만 자동 에디팅 시스템은 최소한의 항목값 수정을 통해 에디팅 규칙을 만족하도록 자동으로 오류위치를 포착하여 수정한다는 점에서 구별된다. 따라서 일괄적으로 수정이 가능한 부분은 자동화 작업이 필요하다.

한편 데이터 에디팅의 제 일차 목적은 조사개선인바, 에디팅 결과표 작성 및 평가를 통해 향후 오류방지 및 조사개선을 이끌어내야 할 것이다.

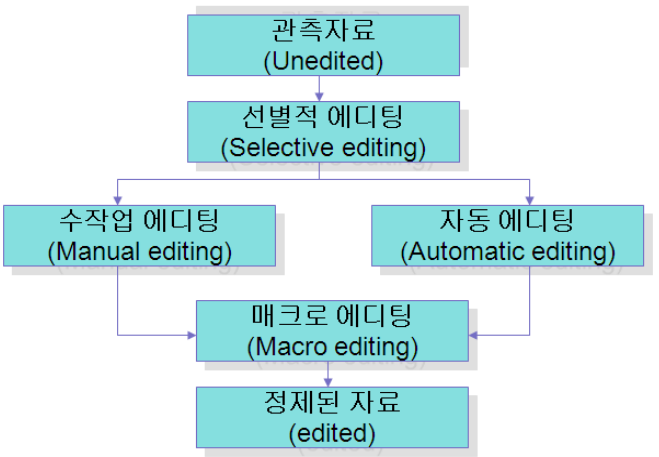
제3절 미시적 에디팅 방법론

네덜란드 통계청은 2000년 3월부터 2003년 2월까지 영국, 핀란드, 스위스, 이탈리아, 덴마크 등과 함께 데이터 에디팅 연구 개발 프로젝트에 참여하였다(Pannekoek와 De Waal, 2003). 이 프로젝트의 결과로써 [그림 6-2]와 같은 에디팅 절차를 제시하였다.

먼저 숫자 단위 착오와 같은 명백한 오류를 수정하고, 레코드의 중요성에 따라 선별

적 에디팅을 적용한다. 중요 레코드는 수작업 에디팅을 시도하고, 중요도가 낮은 레코드는 자동 에디팅을 실시한다. 이후 에디팅된 전체 자료를 거시적으로 검토한다.

이 장에서는 이러한 에디팅 절차에 따라 선별적 에디팅(selective editing), 쌍방향 에디팅(interactive editing), 자동에디팅(automatic editing) 순으로 소개하고 예제를 제시한다. 전체 레코드 수준에서의 매크로 에디팅은 제4장에서 이상치 및 결측치 처리와 함께 소개한다.



[그림 6-2] 네덜란드의 데이터 에디팅 절차

1. 선별적 에디팅(selective editing)

선별적 에디팅(selective editing)[4]은 개별 조사단위가 추정결과에 미치는 상대적 중요도를 고려하여 영향력 있는 자료에 수작업 에디팅을 실시하는 방법으로 경제조사 에디팅에서 많이 쓰이고 있다.

전통적인 방식에 따른 에디팅에서 개별 조사단위가 추정결과에 미치는 상대적 중요도에 대한 고려 없이 처리되는 비효율성을 개선한 것이다. 이 방법은 오차의 대부분이 소수 사업체의 오류에 기인한다는 것인데 이에 대한 근거는 앞 장의 과도한 에디팅에서 이미 설명된 바 있다. 이 방법은 경제조사와 같이 수량 자료에 효과적이며 센서스나 표본조사에서 모두 사용 가능하다.

4) 호주통계청(ABS)은 선택적 에디팅을 유의미 에디팅(significance editing)이라고 부르며 독자적으로 발전시키고 있다[McDaritt et al. (1992) “The AWE Significance Editing Study”].

선별적 에디팅은 자료를 에디팅 시 영향을 주는 정도에 따라 점수(score)를 부여하고 이를 기준으로 중대한(critical) 부류와 경미한(non-critical) 부류로 구분한다. 중대한 부류는 모든 에디팅 규칙을 적용하여 수작업으로 에디팅하는 반면 경미한 부류는 에디팅을 적용하지 않거나 자동 에디팅을 적용한다. 예를 들면, 매출액이나 급여와 같이 선택된 항목에 대하여 점수를 부여하고 각 사업체에 대해서는 항목점수를 합산한다. 사업체 중 기준 점수보다 높은 사업체만을 수작업으로 에디팅 한다. 선별적 에디팅은 구체적으로 에디팅을 수행하기 전에 어느 정도의 에디팅을 하면 얼마의 효과를 가져 올 수 있는 지를 예상할 수 있게 하는 기능도 존재한다.

선별적 에디팅은 주요한 사업체의 오류를 찾아내어 수정하므로 비용 및 시간 절감 뿐만 아니라 지나친 재접촉을 지양하여 향후 조사 참여 거부나 무응답을 방지할 수 있는 장점이 있다. 사후 접촉은 응답부담 뿐만 아니라 응답내용을 다시 회상하기 어렵고 추가 비용이 발생해 바람직하지 않다. 한편 전통적인 에디팅에 익숙한 담당자나 사용자의 거부감이 발생할 수 있으며 항목간 모순된 자료가 존재할 수 있는 단점이 있다.

또한 선별적 에디팅은 결과 품질의 손실 없이 비용을 절감할 수 있고 가장 영향력 있는 레코드에 초점을 둠으로써 자료 품질이 개선될 수 있으며 조사과정 시간을 줄임으로써 시의성을 향상시킬 수 있다. 그러나 마이크로 수준에서의 품질은 떨어질 수 있는 단점이 있으나 예산 문제로 볼 때 장점이 더 많으며 실제 수행사례를 보면 큰 차이가 없는 것으로 보고되고 있다. 더욱이 재접촉 및 재조사를 줄임으로써 응답부담이 경감된다. 선별적 에디팅의 장점과 단점을 정리하면 다음 <표 6-5>와 같다.

〈표 6-5〉 선별적 에디팅의 장점과 단점

장점	단점
비용절감	개별단위수준의 자료품질 저하
자료품질 향상	불일치 자료의 잔존으로 품질저평가 우려
적시성 향상	작은 영역에서 비표집오차 발생 가능
응답부담 경감	관계자 및 사용자의 거부감

선별적 에디팅의 절차는 네 개의 단계로 나누어 설명할 수 있다. 첫 번째 단계는 명백한 오류를 처리하는 단계로 일단 검출되면 수정한다. 두 번째 단계는 영향력 있는 오류를 선택하는 것이다. 세 번째 단계에서는 영향력이 있는 오류는 담당자에 의해 해결되어야 하고 때로 응답자는 재접촉 될 것이다. 영향력이 적은 오류는 가능한 효율적으로 검출되고 수정되어야 한다. 자동 에디팅은 이러한 영향력이 적은 오류를 처리하는데 가

장 많이 쓰이는 방법이다. 네 번째 단계인 검증 단계는 담당자에 의해 수행된다. 예를 들면 현재 자료에 근거한 공표수치와 전년도의 공표수치를 비교하는 매크로 에디팅을 사용한다. 이 단계에서는 개별 레코드의 수정보다 전체 결과에 더 초점을 둔다. 선별적 에디팅의 절차를 요약하면 다음과 같다.

- 단위 착오, 부호 오류 등 명백한(체계적) 오류 수정
- 주류와 비주류 레코드를 나누기 위해 선별적 에디팅 적용
- 주류의 레코드는 쌍방향으로 에디팅하고, 비주류의 레코드는 자동으로 에디팅
- 매크로 에디팅으로 공표수치 검증

선별적 에디팅 시 비주류에 대해 자동 에디팅을 하는 것이 과도하다고 하는 주장이 있다. 네덜란드 통계청(De Waal, 2008)은 선별적 에디팅과 자동 에디팅 단계가 중복될 수 있으나 다음과 같은 이유로 선별적 에디팅이 사용되면서 자동에디팅의 사용이 필요하다. 첫 번째 이유는 비록 각 오류가 영향력이 적다고 해도 비주류 레코드 오류의 합이 공표수치에 영향을 줄 수 있기 때문이다. 두 번째는 비주류 레코드는 내적으로 불일치하므로 만약 에디팅이 되지 않은 채 마이크로 자료가 공표될 때 문제를 일으킬 수 있기 때문이다. 마지막으로 자동 에디팅은 선별적 에디팅 절차의 품질을 점검하는 메커니즘을 제공하는데, 선별적 에디팅이 잘 실행되었다면 쌍방향 에디팅을 위해 선택되지 않은 레코드는 수정이 필요치 않거나 아주 미비하게 수정될 것이기 때문이다. 정리하면 다음과 같다.

- 경미한 부류의 오류 합이 최종 값에 영향을 줄 가능성 제거
- 마이크로 자료 제공시 내적 불일치 오류 문제에 대응
- 수행된 선별적 에디팅에 대한 검증

선별적 에디팅을 수행하기 위하여 다음 5가지의 단계가 필요하다. 각 단계에 대해 다음에서 간략하게 설명한다(ABS, 2007).

가. 주요 항목과 영역 결정

첫 번째 단계는 주요 조사항목과 에디팅을 적용할 주요 영역(domains)을 정하는 것이다. 주요항목(key items)이란 주요 산출물에 영향을 크게 주는 항목을 말한다. 주요항목의 수는 점수방법(scoring method)에 영향을 준다. 주요항목의 수가 5-6개를 넘는 것은 피한다. 영역(domains)이란 산출되는 통계의 어느 수준에 중점을 두어 에디팅을 할 것인가를 결정하는 요소이다. 예를 들어 에디팅의 영역을 시도 수준에 중점을 둘 것인지, 영리/

비영리분야에 중점을 둘 것인지, 산업별 시도별로 세분된 통계에 중점을 둘 것인지 결정하는 것이다. 이상치를 가지고 있는 응답 자료가 충분히 선정될 수 있도록 각 영역에서 샘플을 충분히 뽑아야 한다.

나. 예상치의 산출

경제조사에서 선별적 에디팅을 적용하기 위해서는 먼저 다음의 사항을 정할 수 있는가를 판단해야 한다.

- 각 주요 항목에 대하여 사업체별 예상치를 구할 수 있는가?
- 각 주요항목에 대한 예상추정치를 주요 영역(key domain)별로 구할 수 있는가?

예상치는 반복조사의 경우 과거 조사자료나 다른 경로에서 얻은 보조자료를 이용할 수 있다. 만약 과거자료나 보조자료를 사용할 수 없는 경우에는 현재 사용되고 있는 자료를 가지고 임퓨테이션을 하여 구한다.

주요 영역에 대한 예상추정치는 과거 자료, 현재 자료 및 그 밖의 정보를 이용하여 대략적으로 계산된 추정치이다. 이는 주요 추정치 크기의 상대적 순위를 얻고자 함이다.

다. 점수와 순위 계산

다음 단계는 각각의 주요 항목에 대하여 점수를 계산하는 것이다. 주요 항목에 대한 점수는 에디팅에 영향을 주는 크기를 고려하기 위해 각 영역(domain)의 예상 추정치의 크기와 관련된 가중 응답치와 예상치 사이의 차이로 다음과 같이 표현된다.

$$score = \frac{|\text{가중치} \times (\text{관측치} - \text{예상치})|}{\text{예상목표추정치}} \times 100$$

항목 점수의 예는 다음과 같다. 어떤 경제조사의 한 사업체가 과거조사에서는 1,988명의 직원이 종사(간접고용 형태)한 것으로 나타났으나 현 조사에서 보고된 종사자수는 31,000명이라고 하자. 이 사업체의 가중치는 3.3이고 시도 영역에서 추정되는 종사자 수가 483,700명이라고 가정하면 이 사업체의 종사자수 항목의 점수는 다음과 같이 계산된다.

$$score(\text{종사자수}) = \frac{|3.3 \times (31,000 - 1,988)|}{483,700} \times 100 = 19.8(\%)$$

이는 응답 항목이 에디팅 되는 경우 종사자수(간접고용) 항목의 예상추정치가 19.8% 까지 변화될 수 있음을 나타낸다.

만약 주요 항목이 여러 개인 경우에는 각 항목점수를 합하여 사업체 순위를 만들 수 있다. <표 6-6>은 주요 항목이 8개인 경우 응답 사업체의 우선순위를 예로 보여주고 있다. 이 사업체는 응답 사업체 순위가 2위이고 8개의 각 항목점수별 순위가 기록되어 있다. 점수와 순위가 높을수록 기여도가 크다는 것을 나타낸다.

〈표 6-6〉 각 항목 점수를 나타내는 사업체 세부내용

Providers to be Edited to Achieve 80% of Total Editing Benefit (maximum rank = 800)
Rank=2 UNITID=Quality Engineering P/L Stratum=30095 Selection Weight=3.33 New Unit?=Y

Item	Description	Reported Value	Expected Value	Benefit (%)	Item Rank
DEMPINDP	Indirect Employment	31000.0	1988.0	19.80	1
DLABCOST	Labour Costs	117798000.0	213674670.0	3.54	2
DINCPERM	Income from Permanent Placements	2449000.0	6585096.0	1.99	8
DINCTEMP	Income from Temporary Placements	144523000.0	125248622.0	1.04	2
DDIREMPT	Direct Employment	224.0	121.0	0.76	23
NUMTEMJP	Temporary Job Placements	30000.0	16527.0	0.31	61
NUMPERJP	Permanent Job Placements	130.0	683.0	0.30	31
DTOTEXP	Total Expenditure	315624000.0	145721308.0	0.26	7

라. 업무량의 결정

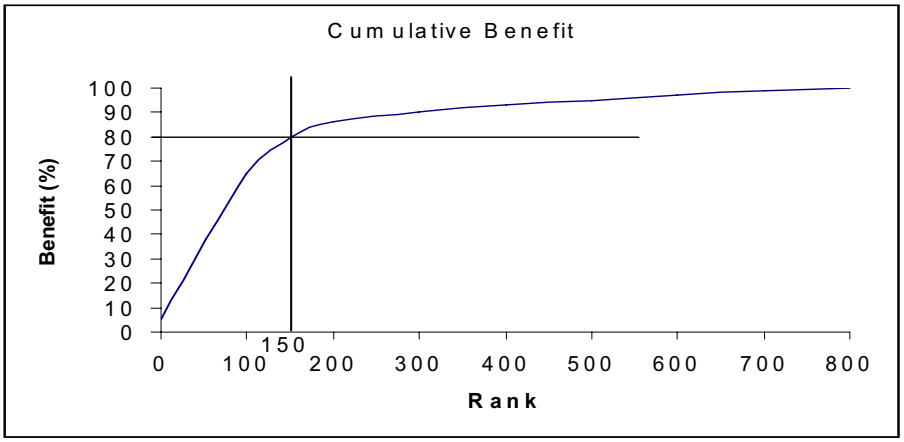
순위가 정해지고 나면 이제 우선적으로 에디팅을 할 항목이나 사업체의 업무량을 정해야 한다. 업무량 구분점(cutoff)은 사전적으로 정해지거나 쌍방향적으로 결정될 수도 있다. 사전적 설정 방법은 응답 전에 업무량의 구분점을 정하는 것을 말한다. 만약 합리적인 예상값이나 추정치를 얻을 수 있다면 첫 응답값을 받자마자 에디팅을 시작할 수 있다. 예를 들어 1.0이 구분점으로 설정된다면 1.0이상의 점수를 얻는 항목이나 사업체가 에디팅 대상으로 선정된다. 이 방법의 단점은 에디팅 대상으로 선정되는 사업체의 수가 바뀔 수도 있어 선정된 사업체 에디팅으로부터 얻는 기대 이익이 달라질 수 있다는 점이다.

또 다른 방법은 누적 이익이 특정한 수가 되도록(예를 들어 90%) 에디팅 항목이나 사업체를 선정하는 것이다. 이 방법은 목표로 하는 누적 이익을 얻을 때까지 요구되는

최소한의 업무량을 정하는 것이다. 그러나 누적 이익의 고정에 따라 에디팅을 해야 하는 사업체 수가 변하게 된다.

구분점을 미리 정할 수 없는 경우에 비용/편익 그래프(cost/benefit graphs)를 이용하여 구분점을 정할 수 있다. 비용/편익 그래프는 각 사업체에 대하여 동일한 에디팅 비용이 들어간다고 가정 할 때 에디팅 비용에 대한 누적기대편익을 백분율로 나타낸다. 이상적인 구분점은 에디팅으로 인한 한계 편익이 줄어드는 비용/편익 그래프의 기울기가 평평해 지는 점에서 정해질 것이다. [그림 6-3]은 비용/편익 그래프의 예이다. 전체 사업체 중 단지 18%(150개)의 에디팅으로 80%의 편익을 달성할 수 있음을 보여준다.

선별한 응답 사업체의 에디팅만으로 만족할 만한 결과를 얻기 어려운 항목에 대해서는 개별 항목별로 비용/편익 그래프를 그리고 각 항목에 대하여 구분점을 할당함으로써 에디팅 대상 업체를 추가적으로 선정할 수도 있다.



[그림 6-3] 비용-편익 그래프

마. 점검 및 재검토 과정

선별적 에디팅 과정은 수시로 점검하고 재검토해야 한다. 너무 많은 응답 사업체를 에디팅 하고 있는건 아닌지 검토해야 하며 사용자 요구의 변화, 기술발전, 감축예산, 모집단의 변화 등을 선별적 에디팅에 반영하여야 한다. 선별적 에디팅의 누적 효과에 대하여 검토하는 것 또한 필요하다. 예를 들어, 응답자가 이전의 자료를 복사하여 응답한 경우 과거자료를 이용하여 산출한 점수는 거의 0에 가까우므로 구분점(cutoff) 이하 응답 사업체 중 이러한 사업체에 대하여 에디팅을 실시하여 사전 설정한 이 방법이 제대로 기능하고 있는지를 검토할 필요가 있다.

예제: 선별적 에디팅

주관적인 검출방법은 담당자가 주관적으로 이상치를 판단하는 반면 객관적인 검출방법은 자동으로 이상치를 검출하고 이상의 크기를 나타낸다. 다음 예제[5]는 마이크로 에디팅 단계에서 이상치를 검출하는 선별적 에디팅 방법의 예를 나타낸다.

서비스업통계 담당부서원이라고 가정하자. 초기 에디팅 절차는 이미 마친 <표 6-7>의 자료에서 총비용과 총 종사자수의 에디팅에 초점을 둔다. 15개의 레코드가 다음 표에 제시되었다. 표에는 각 사업체에 대한 기댓값과 시도수준의 기대목표추정치가 포함되어 있다. 이제 이상치 처리 결정을 하기위해 판단이 필요하다.

<표 6-7>의 자료에서 주관적 분석에 근거하여 볼 때 A14617 사업체의 총비용은 단위 숫자의 착오로 보여 재검토를 결정할 사업체로 판단된다. 일단 가장 중요한 이상치를 식별하기 위해 에디팅 점수와 상대이익의 계산을 포함한 선택적 에디팅 방법의 사용에 대해 살펴보자. <표 6-8>에서 비용과 종사자수에 대한 에디팅 점수가 계산되고 합쳐져 사업체 점수가 생성되었다. 이 사업체에 대한 이상치 수정의 상대 편익이 계산되었다. 특별하게 관심을 끄는 사업체가 있는가?

〈표 6-7〉 기댓값을 포함하는 원자료

식별번호	표본층	시도	승수	총비용	총비용 (기댓값)	총종사자수	총종사자수 (기대값)
A14617	3	7	2	1305093000	1318507	23	22
A21155	4	7	1	4450700	4849476	63	43
A23259	1	7	13	109184	128300	14	1
A45534	3	7	2	1349037	2038371	31	29
A46649	2	7	7	708960	178076	24	6
A48265	3	7	2	3141067	3216099	65	26
A78930	2	7	7	1253310	1287736	33	16
A14935	2	7	7	240903	206700	7	7
A19899	3	7	2	1719305	1331145	22	17
A28095	3	7	2	678073	656319	40	18
A50654	1	7	13	89694	96718	1	1
A58828	1	7	13	440872	469281	6	9
A70704	2	7	7	407579	473939	11	10
A79167	1	7	13	136245	150967	3	1
A81593	1	7	13	186053	207758	6	5

5) 호주 전문가 초청강연회 강의자료(통계개발원, 2010)

<표 6-8>에서 아주 이상한 값을 수정한 후 다시 사업체 점수에 따라 순위를 매긴다. 그림에서 비용-편익 그래프를 이용하여 적절한 cut-off를 선택할 수 있을 것이다. 어떤 사업체가 수작업 에디팅을 필요로 하는 이상치를 갖는가?

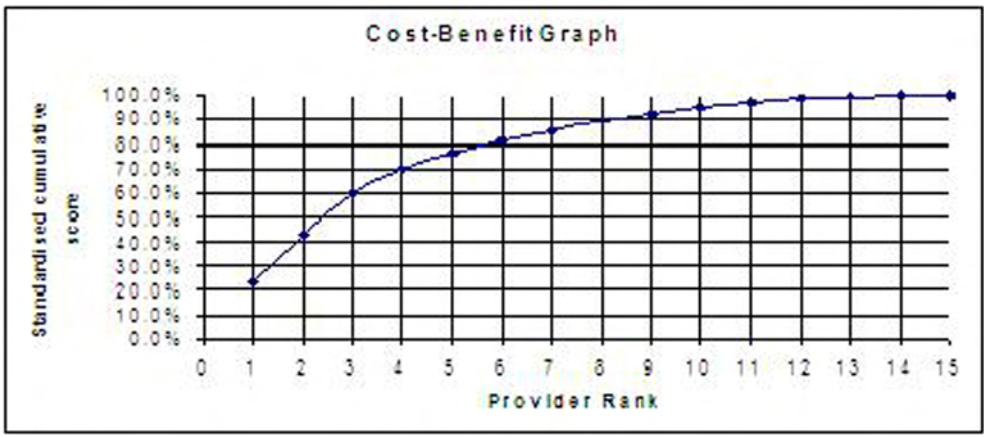
〈표 6-8〉 에디팅 점수(원자료)

식별 번호	표본층	승수	총비용	기댓값 (총비용)	총종사자수	기댓값 (총종사자수)	스코어 (총비용)	스코어 (총종사자수)	스코어 (사업체)	상대 편익
A14617	3	2	1,305,093,000	1,318,507	23	22	4881.04%	0.17%	4,881%	98.8%
A21155	4	1	4,450,700	4,849,476	63	43	0.75%	1.67%	2%	0.0%
A23259	1	13	109,184	128,300	14	1	0.47%	14.08%	14%	0.3%
A45534	3	2	1,349,037	2,038,371	31	29	2.58%	0.33%	3%	0.1%
A46649	2	7	708,960	178,076	24	6	6.96%	10.50%	13%	0.3%
A48265	3	2	3,141,067	3216,099	65	26	0.28%	6.50%	7%	0.1%
A78930	2	7	1,253,310	1,287,736	33	16	0.45%	9.92%	10%	0.3%
A14935	2	7	240,903	206,700	7	7	0.45%	0.00%	0%	0.0%
A19899	3	2	1,719,305	1,331,145	22	17	1.45%	0.83%	2%	0.0%
A28095	3	2	678,073	656,319	40	18	0.08%	3.67%	4%	0.1%
A50654	1	13	89,694	96,718	1	1	0.17%	0.00%	0%	0.0%
A58828	1	13	440,872	469,281	6	9	0.69%	3.25%	3%	0.1%
A70704	2	7	407,579	473,939	11	10	0.87%	0.58%	1%	0.0%
A79167	1	13	136,245	150,967	3	1	0.36%	2.17%	2%	0.0%
A81593	1	13	186,053	207,758	6	5	0.53%	1.08%	1%	0.0%

- 총비용 기대목표값(Expected target estimate for total expenditure) = 53,422,000
- 총종사자수 기대목표값(Expected target estimate for total employment) = 1,200
- $\frac{|13\times(186,053-207,758)|}{53,422,000}\times 100 = 0.53(\%)$
- $\frac{|13\times(6-5)|}{1200}\times 100 = 1.08(\%)$
- $\sqrt{0.53^2+1.08^2} = 1$
- $\frac{\text{사업체 점수}}{\Sigma\text{사업체 점수}}\times 100 = 0.0(\%)$

〈표 6-9〉 에디팅 점수(명백한 오류의 수정 후)

식별번호	표본층	승수	총비용	기댓값(총비용)	총종사자수	기댓값(총종사주수)	점수(기댓값)	점수(총종사자수)	점수(사업체)	상대편익	순위
A14617	3	2	1,305,093	1,318,507	23	22	0.05%	0.17%	0%	0.3%	
A21155	4	1	4,450,700	4,849,476	63	43	0.75%	1.67%	2%	3.0%	
A23259	1	13	109,184	128,300	14	1	0.47%	14.08%	14%	22.9%	1
A45534	3	2	1,349,037	2,038,371	31	29	2.58%	0.33%	3%	4.2%	
A46649	2	7	708,960	178,076	24	6	6.96%	10.50%	13%	20.5%	
A48265	3	2	3,141,067	3,216,099	65	26	0.28%	6.50%	7%	10.6%	
A78930	2	7	1,253,310	1,287,736	33	16	0.45%	9.92%	10%	16.2%	
A14935	2	7	240,903	206,700	7	7	0.45%	0.00%	0%	0.7%	
A19899	3	2	1,719,305	1,331,145	22	17	1.45%	0.83%	2%	2.7%	
A28095	3	2	678,073	656,319	40	18	0.08%	3.67%	4%	6.0%	
A50654	1	13	89,694	96,718	1	1	0.17%	0.00%	0%	0.3%	15
A58828	1	13	440,872	469,281	6	9	0.69%	3.25%	3%	5.4%	
A70704	2	7	407,579	473,939	11	10	0.87%	0.58%	1%	1.7%	
A79167	1	13	136,245	150,967	3	1	0.36%	2.17%	2%	3.6%	
A81593	1	13	186,053	207,758	6	5	0.53%	1.08%	1%	2.0%	



[그림 6-4] 비용-편익 그래프(명백한 오류 수정 후)

[그림 6-4]에서 누적이익이 최소 80%를 만족하는 사업체 순위는 6이다. 따라서 우리는 6개의 사업체를 수작업으로 에디팅함으로써 80%의 이익을 가질 수 있다. 6위 밖의 사업체는 에디팅을 해도 기대되는 변화가 미비하다고 본다.

2. 쌍방향 에디팅(interactive editing)

쌍방향 에디팅은 컴퓨터를 이용하여 각 레코드에서 위배된 에디팅 규칙을 나타낸다. 담당자는 이 결과를 가지고 재접촉 등 해당 레코드를 검토한다. 자료 수정 후 다시 입력하고 모든 레코드가 에디팅 규칙을 통과할 때까지 반복적으로 수행된다. 즉 쌍방향 에디팅의 근본적인 문제점은 모든 레코드가 에디팅 된다는 것이며, 수작업 수정 시 일관성을 검토하지 못한다는 것이다. 수정된 레코드는 여전히 에디팅 규칙을 만족하지 못할 수 있고 에디팅이 반복될 수 있어 시간과 비용이 많이 든다.

블레이즈(Blaise)와 같은 조사처리 시스템 사용 시에는 설정된 에디팅 규칙이 자료 입력시 또는 후에 점검될 수 있고 바로 수정될 수 있다. 쌍방향 에디팅 또는 컴퓨터 보조 에디팅이라 불리는 블레이즈는 위배된 에디팅 규칙과 연관된 필드를 계산하므로 오류를 식별하는 데 도움을 준다. 자료 수정을 위해 조사표나 스캔된 조사표를 검토한다.

블레이즈는 CAPI, CATI, CASI, CAWI를 위한 하나의 입력 내검 시스템이다. 자료수집 단계에서의 에디팅은 질문에 타당하지 않은 응답 또는 두 개 이상의 질문간 모순된 응답이 있을 때 블레이즈는 바로 알려주게 되고 응답자에게 다시 질문함으로써 해결될 수 있다.

그러나 CAPI, CATI, CASI, CAWI가 자료를 수집하는 이상적인 방법처럼 보일 수 있으나 단점도 있다. 첫 번째 단점은 비용이 많이 든다는 것이다. 둘째는 CATI와 CAPI는 응답자가 인터뷰하는 동안 질문에 응답할 수 있어야 한다는 전제조건이 필요하다. 가구나 개인에 관한 조사에서는 응답자가 질문에 빨리 응답할 수 있으나 사업체 조사에서는 정확한 답을 빨리 줄 수 없다. 위와 같은 이유로 CAPI와 CATI는 주로 가구조사에서 자료를 수집할 때 사용된다.

특히 CASI와 CAWI는 실용적인 자료수집 방법이지만 응답이 모순되었다는 것을 계속 보고하면 응답자는 귀찮아지고 이후의 질문에 응답을 그만둘지 모른다는 실제적인 문제가 있다. 또한 서베이 응답자 그룹이 특정된다는 통계적 문제와 CASI나 CAWI에 의해 수집된 자료가 종이조사에 의해 수집된 자료보다 높은 통계 품질을 가진다.

오류 또는 무응답이 발생한 응답자에 대해 재접촉/재조사의 필요성이 적거나 불가능한 경우, 최종적으로 에디팅 규칙을 만족시키지 못한 항목 값은 수정되어야 한다. 적은 노력으로 정보의 손실을 막기 위해 가능한 한 최소의 항목을 수정함으로써 모든 에디팅 규칙을 만족시키는 전략이 필요하다.

3. 자동 에디팅(automatic editing)

가. 자동 에디팅 개요

경제 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. 통계 선진국에서는 예산 및 조사의 어려움을 극복하고, 수작업으로 인한 과도한 에디팅의 단점을 보완하기 위해 사업체 조사에 자동 에디팅을 활용하고 있다.

수작업 에디팅은 앞에서 언급한 바와 같이 매우 노동집약적인 과정으로 자료 처리 시간, 노동력, 비용, 자료의 양이 적은 경우에는 적합하나 반복조사, 자료의 양이 많은 경우에는 부적절하다. 자동에디팅은 초기 비용이 많이 든다는 단점이 있다. 자동화의 이러한 단점에도 불구하고 가능하면 오류 검색과 오류의 수정이 모두 자동화되어야 한다. 왜냐하면 자동에디팅은 일단 구축되면 향후 조사처리 과정에서 시간과 비용을 절감할 뿐만 아니라 일관적이고 정확한 처리 결과를 획득한다는 이점이 있기 때문이다.

한편 조사원이 현장에서 오류수정을 수행하는 것은 수집 초기단계의 정보에 근거해서 수정을 할 수 있다는 장점이 있다. 그러나 적지 않은 에디팅 작업량과 촉박한 일정은 재접촉의 질을 보증할 수 없을 가능성이 있으며 응답자가 불응하거나 여러 상황으로 인해 확인이 불가능한 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 조사부담 역시 클 것이다. 이상 수작업과 자동에디팅을 비교·정리하면 <표 6-10>과 같다.

〈표 6-10〉 수작업 에디팅과 자동 에디팅의 비교

구분	수작업 에디팅	자동 에디팅
단점	노동집약적 에디팅 시 조사원간 편차 발생 응답자 및 조사원 부담 증가	초기비용 소요
장점	오류원인 파악	향후 시간과 비용절감 데이터 재생 가능
적합한 조사	소규모, 일회성 조사	대규모, 반복 조사

자동에디팅은 사람이 세밀하게 작업하는 부분을 인정하지 않는 것이 아니라 컴퓨터를 이용하여 가능한 수고를 덜자는 것이다. 이는 옷감이 손상되기 쉬운 세탁물은 손으로 세탁하고 옷감에 신경을 쓰지 않아도 될 세탁물은 자동으로 세탁하는 것에 비유할 수 있다. 또한 더러움이 심한 와이셔츠의 깃이나 소매를 사람이 특별하게 처리하고 이를 자동세탁기로 처리한다면 손세탁의 수고를 상당히 덜어 주는 것과 같다고 하겠다.

나. Fellegi-Holt 기법

전형적인 자동 에디팅 절차는 먼저 측정단위 오류와 같은 체계적 오류를 연역적 알고리즘을 이용하여 자동으로 탐색하거나 수정함으로써 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 에디팅 규칙을 동시에 만족하도록 수정되어야할 변수(항목)의 최소 집합을 찾는 것이다(Fellegi와 Holt, 1976).

F-H 기법은 수학적 최적화에 기초한 대표적 에디팅 방법이다. F-H 방법은 조사 자료에 오류가 있는지를 판단하기 위해 조사 담당자에 의해 설정되는 에디팅 규칙(edits)을 필요로 한다. 만약 설정된 에디팅 규칙을 위반하면 대체해야 하는 변수를 결정하는 자동화 전략이 필요한데, 주어진 정보를 최대한 보존하면서 모든 에디팅 규칙을 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.

■ 간단한 예제 1

이해를 돕기 위해 다음과 같은 2개의 명시적 에디팅 규칙(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 =8로 코딩되었다고 하자. 그러면 이 레코드는 두 번째 규칙을 위반한 레코드이다. 이때 두 개 이상의 변수를 모두 바꾸면 성립이 가능할 수 있으나 최대한 자료를 보존한다는 원칙에서 X_3 하나만 바꾸는 것이 합리적이다.

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

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

따라서 주어진 레코드의 전체 위배된 에디팅 규칙은 E_2, E_3이다(<표 6-11> 참조).

〈표 6-11〉 위배된 에디팅 규칙 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 에디팅 규칙 E_2, E_3에 모두 포함되므로 X_3를 바꾸어주는 것이 합리적이다. 즉 명시된 에디팅 규칙으로부터는 어떤 변수를 바꾸어 주어야 할지가 명확하지 않으나 이처럼 추가된 에디팅 규칙을 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 에디팅 규칙(implied edits, implicit edits)이 오류 자료의 변경할 변수들을 결정할 때 중요한 역할을 하며, 일반적인 If-Then-Else의 구조보다 효율적이고 에디팅 규칙의 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 에디팅 규칙을 필수적으로 만족시켜야 하는 규칙(hard edits)으로 간주한다는 것이다. 또한 오류를 모두 랜덤오류로 인식한다. 특히 요구되는 내재적 에디팅 규칙 수가 매우 많을 수 있으며, 이때 모든 내재적 규칙의 생성에 있어서 많은 시간이 소요된다(De Waal과 Coutinho, 2005).

다. 선형계획법을 이용한 수정

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

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

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

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

간단한 예제 2

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

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

하나의 레코드가 X_1 = 6, X_2 = 4, X_3 = 8을 갖는다고 가정할 경우 이 레코드는 두 번째 에디팅 규칙을 위배한다. 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 :

$$X_1 - X_2 \geq 0$$

$$X_2 - 3X_3 \geq 0$$

여기서 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 에디팅 규칙을 만족하면서 오류로 간주된 변수들을 수정하는 일치적(consistent imputation) 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

라. 등식조건 하에서의 간단한 오타의 자동수정

경제조사에서 가장 중요한 자료 검토 중 하나는 합계불일치 오류에 대한 검토이며 이때 자동 에디팅을 위해 사용되는 수학적 최적화 기법은 단순 오타의 속성에 대한 정보를 사용할 수 없는 약점이 있다. 특히 등식 조건하에서의 단순오타로 인한 오류의 자동 수정은 실행이 쉽고 위험성이 적어 유용하다. 경제조사의 종사자수, 자산, 재고 등의 합계불일치 오류의 점검에 활용이 가능하다.

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

■ 간단한 예제 3

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

$$X_1 \quad - \quad X_2 \quad = \quad X_3$$

하나의 레코드가 X_1 (매출액) = 353, X_2 (영업비용) = 283, X_3 (영업이익) = 115를 갖는다고 가정할 경우 매출액 - 영업비용 = 영업이익이라는 에디팅 규칙을 위반하게 된다. F-H 방법은 매출액이 353에서 398 또는 영업비용이 283에서 238 또는 영업이익이 115에서 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 방

법은 적용하기 어렵다. 따라서 단순오류의 대표적 패턴을 이용하여 먼저 오류를 해결하면 정확한 오류해결 뿐만 아니라 나머지 오류에 대한 작업량을 줄일 수 있다.

간단한 예제 4

하나의 합계 에디팅 규칙이 $X_1 + X_2 = X_3$라고 가정하자. 주어진 자료 X_1 = 100, X_2 = 20, X_3 = 300은 에디팅 규칙을 위반하므로 오류자료이다. 이때, X_1의 100을 280으로 바꾸거나 X_2의 20을 200으로 바꾸거나 X_3의 300을 120으로 바꾸면 에디팅 규칙이 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 다음과 같이 수학적으로 정의하였다.

① 두 개의 자릿수 바뀜: $f_{ic}(4627;1) = 4267$

② 한 자릿수 늘어남: $f_a(4627;1,8) = 46287$

③ 한 자릿수 빠짐: $f_a(4627;1) = 467$

④ 음수 기호가 빠지거나 들어감: $f_m(4627) = -4627$

⑤ 해당 자릿수에서 다른 숫자로 대체됨: $f_r(4627;1,8) = 4687$

만약 이 식이 성립된다면 오타로 인해 변화된 것으로 볼 수 있고 만약 5가지의 어떤 함수도 위 식이 성립되지 않는다면 현재의 값을 바꿀 수 없다.

예제: 종사자수 합계불일치 오류의 수정

적용 자료는 2008년 기준 서비스업조사 자료(2009년)이며 총 43,463건이다.

- 월평균 종사자수

2008년 영업기간(1.1~12.31) 중에 근무한 월평균 종사자 수를 <표 6-12>와 같이 종사상 지위별 및 성별로 구분하여 나타낸다.

〈표 6-12〉 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}

- 에디팅 규칙(Edit rules)

종사자수 조사항목과 관련된 합계불일치 점검규칙은 다음과 같이 9개의 에디팅 규칙이 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 에디팅 규칙임)이다. 이를 수식으로 표현하면 다음과 같다.

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

- 종사자수 합계불일치 오류

오류자료의 탐색결과에 따른 종사자수 합계불일치 오류자료의 내용은 <표 6-13>과 같다. 종사자수 합계의 불일치 건수는 총 15건으로 남자 종사자수 합계 불일치, 상용종사자수의 남녀 합계불일치다. 상용종사자수가 종사자수보다 큰 경우가 2건(사업체번호: 3400105247, 1103203381)이 나타나고 있다.

〈표 6-13〉 종사자수 합계불일치 탐색결과

	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	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

- 단순오타의 자동수정

프로그램 R의 패키지를 이용하여 등식조건하에서 단순오타의 자동수정을 실시한 결과가 <표 6-14>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→0으로 수정되었다. 그런데 7번째 레코드는 x_{12}의 값을 219에서 0으로 바꾸면 완전하게 일치되므로 변수 당 허락되는 단위변화의 숫자를 3으로 설정한 결과이다(디폴트는 1).

〈표 6-14〉 correctTypos를 이용한 에디팅 결과(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

• 수학적 최적화 방법에 의한 자동수정

연역적 수정 알고리즘 패키지는 자료의 모든 불일치를 해결하지 못한다. 완전한 자동에디팅을 위해서 나머지 불일치 오류는 선형계획법(LP)의 이용이나 Fellegi-Holt 패러다임에 근거한 오류위치 탐색 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 지를 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 가중치 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.

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

프로그램 R에서 lpSolve 패키지를 내려 받아 선형계획법 문제를 해결할 수 있으며 그 수행결과는 <표 6-15>와 같다.

종사자수의 합계불일치자료를 자동수정한 결과와 서비스업조사 최종자료는 거의 일

치하였다. 이상의 자동오류 수정은 재조사 및 재질의가 필요치 않거나 불가능할 때 유용하다. 경제조사에서 유형자산, 무형자산, 연초재고, 연말재고 등의 합계 불일치 오류 점검에 확대하여 활용 가능할 것이다.

〈표 6-15〉 lpSolve를 이용한 에디팅 결과(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	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

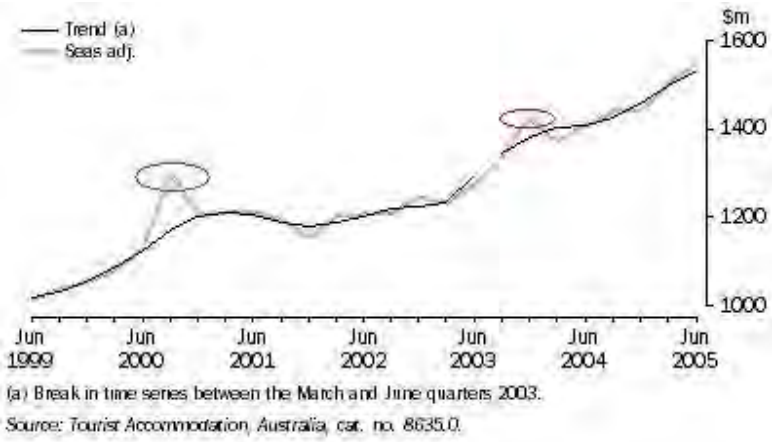
제4절 거시적 에디팅 방법론

최종 통계결과는 현실적으로 납득이 되어야 한다. 거시적 에디팅(매크로 에디팅)은 최종 결과물이 이해할 만하고 설명될 수 있는 값인지를 확인하기 위한 절차이다. 과거 자료에 대한 경향을 잘 살펴보거나 다른 자료와 비교해 보고 다양한 관계에 대한 타당성을 검토하며 전반적인 일관성을 판단한다. 담당자는 통계결과물의 적정 여부를 판단하여 그 결과가 이해하기 어려운 경우 그 원인을 밝혀내는 것이다. 거시적 에디팅을 통하여 검출된 이상치를 해결할 때 조사 자료에 포함된 오류뿐만 아니라 시스템, 절차, 방법론상 오류 등 가능한 원인을 모두 고려해야 한다.

1. 매크로 에디팅(Macro-editing)

■ 예제: 주관적 에디팅의 적용

호주에서 숙박 수입에 대한 계절 조정치와 추세 추정치(기대치로 사용)를 비교하여 적용된 매크로 에디팅 사례를 보여준다([그림 6-5] 참조). 이들 추정치 사이에 커다란 차이(2000년과 2003년의 숙박 수입이 비정상적으로 증가)가 있는 시기가 있는데, 그 이유는 각각 시드니 올림픽과 럭비 월드컵으로 인하여 숙박 수요가 늘었기 때문이다.



[그림 6-5] 호주의 숙박수입, 계절 조정값과 추세치

마이크로 에디팅은 개별 레코드에서 오류를 찾는 방법이라고 앞에서 정의한 바 있다. 한편 전통적인 마이크로 에디팅은 총계수준에서의 추정치 품질에 대한 개별 레코드의 중요성을 고려하지 않아 과도한 에디팅으로 가는 경향이 있다. 매크로 에디팅은 총 자료의 분석을 통해 자료의 오류를 찾는 방법으로 비용-편익 효율성 원칙에 근거하여 조사 추정치에 잠재적인 영향력에 근거한 에디팅으로 개별 레코드에서의 의심스러운 자료를 식별하는 절차를 말한다(Granquist, 1995).

다시 말해, 매크로 에디팅 방법은 개별 레코드에서 의심스러운 자료를 식별하기 위해 조사 추정치에 대한 잠재적인 영향력을 검토한다. 매크로 에디팅은 선별적 에디팅이나 자동 에디팅으로 감지하지 못한 오류를 검출할 수 있다. 매크로 에디팅과 마이크로 에디팅은 서로 보완적으로 마이크로 에디팅은 매크로 에디팅보다 많은 오류를 검출하지만 매크로에디팅은 영향력이 있는 오류를 추적할 수 있다.

매크로 에디팅은 선별적 에디팅의 한 형태이다. 매크로 에디팅과 선별적 에디팅의 차이점은 에디팅 과정에서 수행되는 시점 차이다. 선별적 에디팅은 자료가 수집되고 있는 에디팅 과정의 초기 단계에서 이루어지고 매크로 에디팅은 자료가 거의 수집된 마지막 단계에서 이루어진다. 선별적 에디팅은 개별 레코드에 문제가 없는 지를 검토하고 매크로 에디팅은 전체 자료가 문제가 없는 지를 검토한다. 매크로 에디팅은 크게 두 가지의 형태로 구분한다.

총계 방법은 먼저 기본 영역 수준에서 의심스러운 총계를 식별한다. 이 총계에 속한 개별 레코드의 자료를 검사한다. 의심스러운 총계에 속한 레코드의 목록은 상자그림과 같은 그래픽 툴을 사용하여 검사한다. 좀 더 체계적인 자료 점검을 위해 과거자료에 근거하여 채택 영역이 계산될 수 있다. 한편 그래픽 방법은 강력하고 유연하면서 이해하기 쉽고 시각적으로 자료의 영향력을 볼 수 있는 방법이다. 통계산출물이 적절한 것인지를 평가하는 시각적인 매크로 에디팅 기법을 말한다.

가. 총계 방법(aggregation method)

총계 방법은 공표될 숫자에 문제가 없는지 검증하는 것이다. 이것은 공표 결과를 이전 공표 결과, 행정자료, 다른 관련 정보와 비교하는 것으로 수행된다. 만약 이상한 결과치가 관측되면 마이크로 에디팅 절차가 이 결과치에 영향을 주는 개별 레코드와 필드에 적용된다. 예를들면 이상한 결과값은 다음과 같이 점검될 수 있다.

$$\left|\frac{Y-\hat{Y}}{\hat{Y}}\right|\times 100 > p$$

Y는 점검되는 결과치이고 $\hat{Y}$은 기대 공표 결과치이고 p는 임의의 백분율이다. 즉 공표치가 기대 공표치 대비 p% 벗어난다면 공표치에 해당되는 마이크로 데이터는 마이크로 에디팅 과정을 수행하게 된다. 매크로 에디팅 관련 소프트 패키지는 모수 추정치에 대한 개별 관측치의 영향력 추정이 가능하다. 가장 영향력이 있는 관측치부터 시작하여 각 개별 레코드는 쌍방향으로 점검되고 수정된다. 기대치와 큰 차이가 없을 때까지 에디팅 과정이 수행되는데 변화가 생길 때마다 공표 수치를 재 추정함으로써 영향력이 점검될 수 있다.

나. 분포 방법(distribution method)

매크로 에디팅의 두 번째 형태는 분포 방법이다. 에디팅이 요구되는 데이터 셋 또는 참조 데이터 셋이 먼저 각 변수의 분포를 결정하는 데 사용된다. 다음은 모든 개별 값은

이 분포와 비교된다. 전형적으로 위치와 분산 측도가 계산된다. 예외적인 값을 포함한 레코드는 후속 검토와 수정을 위한 대상이 된다.

EDA(exploratory data analysis)는 변수의 분포를 분석하는 여러 가지 기법을 제공하는 통계학의 한 분야이다. 많은 EDA 기법은 매크로 에디팅에 적용될 수 있다. EDA 옹호자는 그래픽 방법 사용의 중요성을 강조한다. 이러한 방법은 수치적 기법보다 변수의 움직임에서 많은 통찰력을 제공할 수 있다. 자료의 분포 그래프는 만약 수치만 계산되어지면 밝힐 수 없는 뜻밖의 특성을 보여 줄 수 있다.

매크로 에디팅에 대한 EDA 기법의 적용은 많은 논문에서 찾아 볼 수 있다. 상자그림, 산점도 등 전통적인 EDA 방법에서 진보된 EDA 방법까지 다양하다. 한편 패턴을 더 쉽게 구별하기 위해 자료의 변수변환이 적용되기도 한다. 매크로 에디팅을 위한 소프트웨어는 상자그림, 산점도외 회귀분석, 시계열 분석, 이상치 탐색, 다변량 이상치 탐색 등을 제공한다.

매크로 에디팅의 단점은 에디팅을 위해 요구되는 시간과 자원이 예측되기 어렵다. 또한 거의 모든 자료가 도착되고 과정을 위해 준비될 때까지 매크로 에디팅 과정을 기다려야 한다. 그래픽 매크로 에디팅은 상대적으로 많은 양의 자료를 동시에 검토할 수 있게 한다. 그러나 많은 주요 변수가 있는 그래픽 매크로에디팅은 보통 가장 적절한 에디팅 방법은 아니다. 한편 총계방식의 가장 중요한 단점은 이 방법이 의심스럽지 않다고 본 공표치에 기여하는 레코드가 여전히 영향력있는 오류, 검출되지 않고 수정되지 않은 오류를 여전히 포함한다는 것이다. 이것은 편향된 공표 결과를 초래할 수 있다.

매크로 에디팅의 장점은 에디팅 규칙을 필요로 하지 않는다는 것이다. 비 에디팅 규칙 등 에디팅 규칙을 설정하는 것은 어렵고 시간이 많이 든다. 그러나 이것이 장점이 됨과 동시에 위험스러운 부분이 있다. 마이크로 에디팅 방법과 에디팅 규칙이 설정되면 레코드의 판단이 분명하다. 그러나 규칙이 설정되지 않으면 의심스러운 레코드와 그렇지 않은 레코드의 판단이 담당자의 주관적 판단에 맡기어 진다.

자동 에디팅과 대체는 그 결과가 재생산될 수 있다. 그러나 매크로 에디팅과 쌍방향 에디팅은 그렇지 않다. 이는 매크로 에디팅과 쌍방향 에디팅에 주관적 판단이 존재함을 의미한다. 담당자가 바뀌면, 같은 담당자라도 조사할 때마다 다른 결과를 초래할 수 있다. 많은 오류를 포함하고 있을 때 마이크로 에디팅은 매크로 에디팅보다 효율적이다. 매크로 에디팅은 마이크로 에디팅에 의해 상당한 품질의 자료가 확보될 때까지 유보한다.

마이크로 에디팅을 유지하는 주장은 이 방법이 레코드가 내부적으로 일치되는 것을 확인하는 유일한 방법이기 때문이다. 체계적인 오류의 자동 수정은 매크로 에디팅 전에 수행되어야만 한다. 이러한 유형의 마이크로 에디팅은 비용도 많이 들지 않고 매크로 에

디팅 단계에서 사용되는 총계 추정치와 분포를 향상시킬 수 있다. 이상의 마이크로 에디팅과 매크로 에디팅의 특성을 비교하면 다음과 같다.

〈표 6-16〉 마이크로 에디팅과 매크로 에디팅의 비교

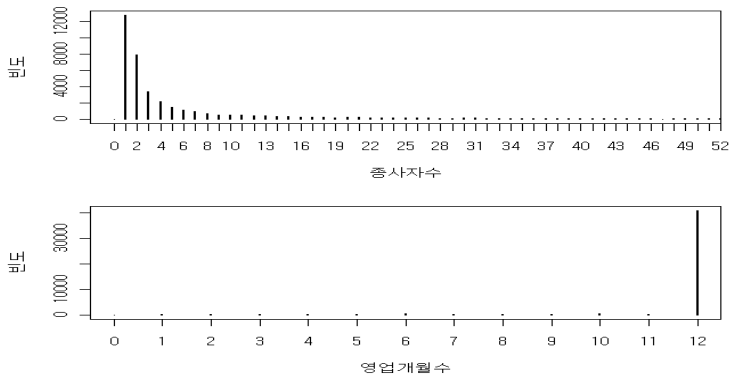
구분	마이크로 에디팅	매크로 에디팅
에디팅 시점	자료 수집시	자료수집 완료 후
에디팅 규칙	필요	불필요
의심레코드 판단	객관적	주관적
재생산성	가능	일반적으로 불가능
내적 일치성	확보	미확보

예제: 그래프를 이용한 자료탐색

2008년 기준 서비스업 조사 자료(총 43,463건)를 그래프 방법을 이용하여 각 주요항목의 분포를 살펴보고 이상치를 탐색하였다.

- **종사자수와 영업개월수**

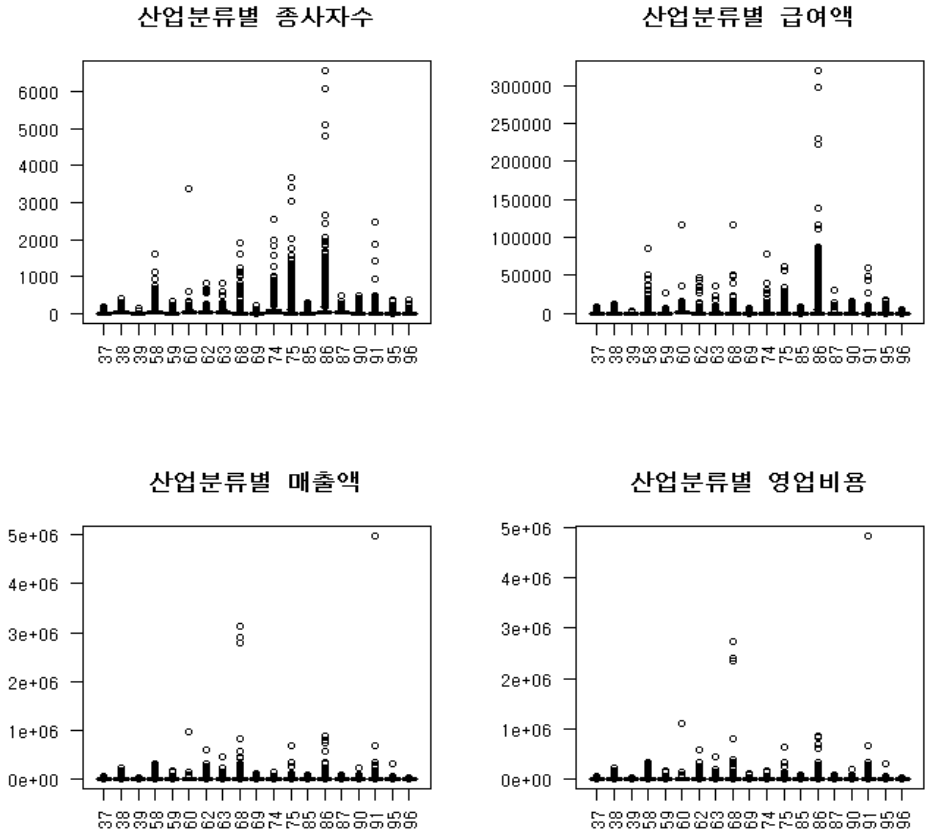
[그림 6-6]은 종사자수와 영업개월수의 분포를 보여주고 있다. 종사자수별로 그 분포를 살펴본 결과, 4인 이하인 경우가 26,265건으로 전체에서 60%를 차지하고 있으며 1인 사업체인 경우는 12,774건으로 약 30%이다. 종사자수는 중요 항목으로 정확한 검토가 필요하다. 종사자수가 0인 경우가 2건(사업체번호는 3407004475와 3300049275임)으로 필수규칙을 위반하였다. 사업실적이 작은 것으로 미루어보아 종사자수는 1-2인으로 판단된다. 한편 사업체의 영업개월수의 분포를 보면 대부분이 12개월로 나타나고 있다(40,903건). 영업개월수가 0인 경우가 1건(사업체번호는 1102070684)이 나타나고 있어 검토가 필요하다.



[그림 6-6] 종사자수와 영업개월수의 분포

- **산업분류별 주요항목: 종사자수, 급여액, 매출액, 영업비용**

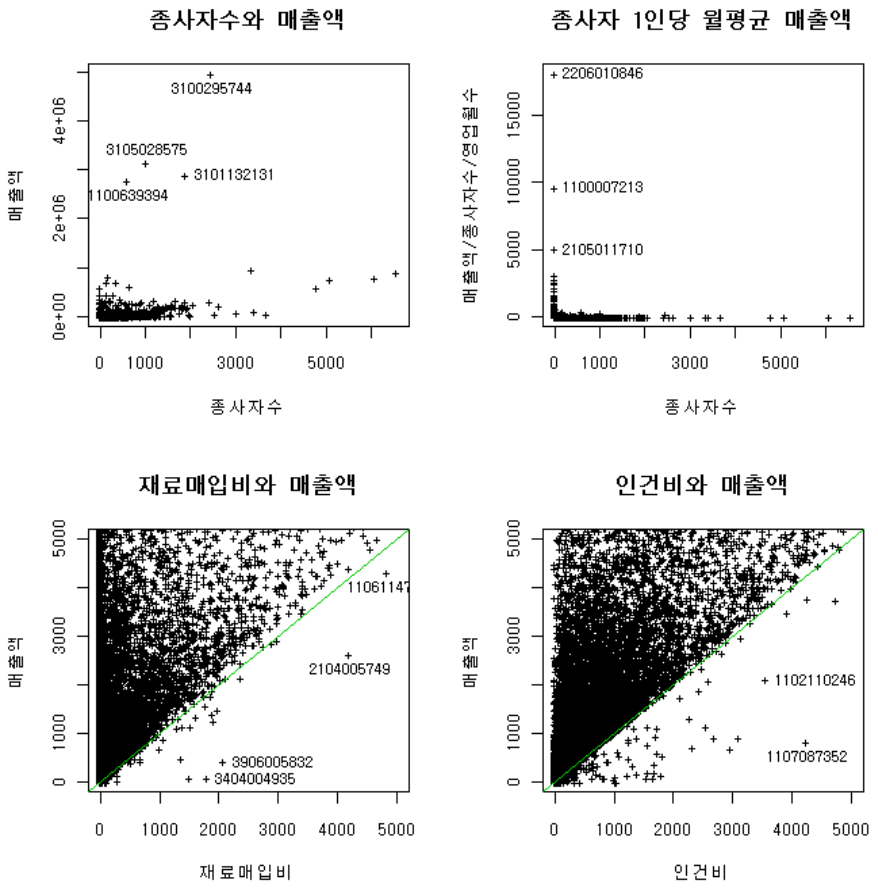
산업분류별 주요항목인 종사자수, 급여액, 매출액, 영업비용에 대해 상자그림을 살펴보면 [그림 6-7]과 같다. 종사자수의 최대값은 산업분류 86인 사업체의 6,578명으로 나타나고 있다. 급여액의 최대값은 역시 산업분류 86인 사업체의 3천 2백억 원으로 나타나고 있으며 74와 75 분류의 사업체는 비슷한 규모의 종사자를 가진 사업체에 비해 급여액이 낮은 것으로 보인다. 종사자수와 급여액, 영업비용과 매출액은 깊은 연관성을 가지며, 규모가 큰 사업체의 사업실적이 영업비용과 서로 일치되고 있음을 알 수 있다.



[그림 6-7] 산업분류별 주요 항목의 상자그림

- **매출액 관련 산점도의 검토**

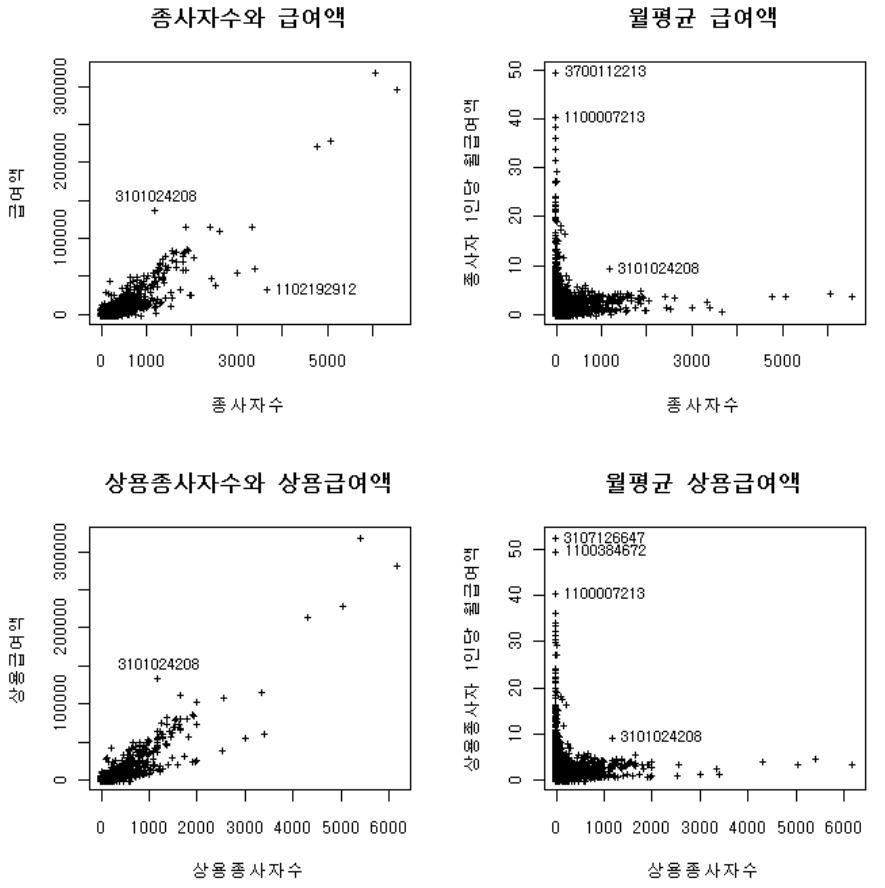
[그림 6-8]에서 종사자수와 매출액의 산점도를 살펴보면, 4개의 사업체가 종사자 규모에 비해 매출액의 규모가 크게 나타나 검토가 필요하다. 매출액이 0이거나 음수인 경우는 없는 것으로 나타났다. 종사자 1인당 월평균 매출액을 보면, 1인당 월평균 매출액이 2백억, 1백억, 50억인 경우가 나타나 검토가 필요하나 대부분 종사자수가 매우 작은 경우에서 나타나고 있다. 재료매입비와 매출액의 산점도를 보면, 매출액이 재료매입비보다 작은 경우는 120건이다. 인건비와 매출액의 산점도를 보면, 매출액이 인건비보다 작은 사업체는 225건으로 집계된다. 특히 매출액이 인건비의 10%도 안 되는 사업체는 12건으로 나타나 검토가 필요하다.



[그림 6-8] 매출액 관련 산점도

- **급여액 관련 산점도의 검토**

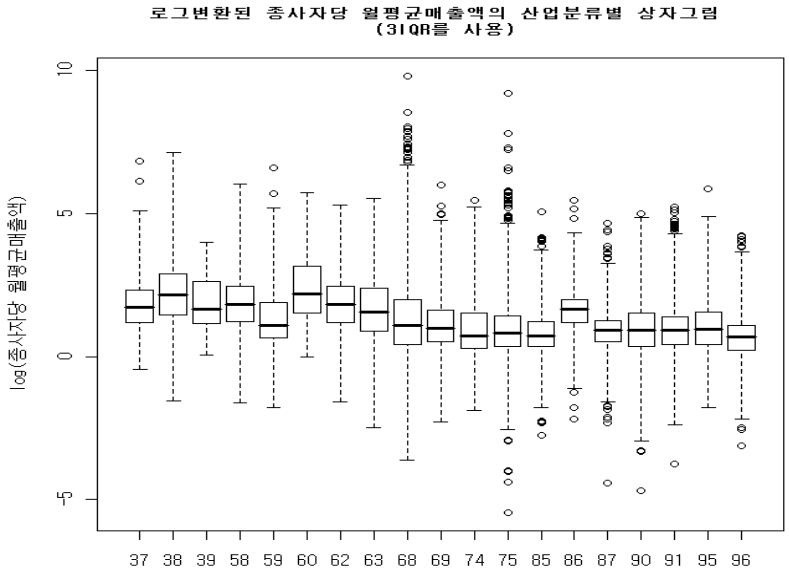
[그림 6-9]의 종사자수와 연간급여액의 산점도를 살펴보면, 상당한 양의 선형관계(크게 두 개의 선형관계)가 나타나며 선형관계를 다소 벗어난 두 개의 사업체가 보인다. 급여액이 0인 경우는 16,429건(인건비가 0인 경우는 16,014건)으로 나타났다. 1인당 월평균 급여액이 5천만 원인 경우가 나타나며 2천만 원을 넘는 경우가 상당수 존재하고 있으나 종사자수가 매우 작은 경우에서 대부분 나타나고 있다. 3101024208 사업체는 약 1,000명 종사자수 규모의 다른 사업체에 비해 높은 급여액(약 1천만 원)을 나타내고 있다. 한편 상용급여액 관련 산점도는 급여액 관련 산점도와 매우 유사하게 나타나고 있다.



[그림 6-9] 급여액 관련 산점도

- **월평균 매출액의 상자그림**

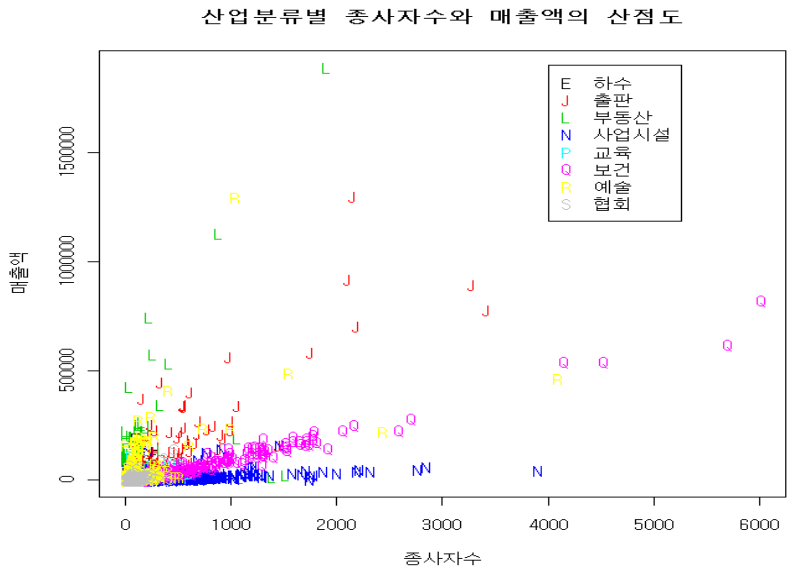
한편, 상자그림은 자료가 대칭적일 때 자료의 몸통부분에서 멀리 떨어진 자료를 구분하고자 할 때 효과적이다. 그런데 일반적으로 경제자료는 자료의 속성상 우측으로 길게 늘어지는 분포를 갖게 된다. 이러한 분포에서 이상치를 탐색하는 상자그림은 적절하지 않다. 매우 큰 값들이 빈번히 존재할 때 로그값으로 변환하면 자료는 대칭이 되곤 한다. 따라서 로그 변환 후 상자그림을 통해 이상치를 판단하는 것이 적절하다. 일반적으로 상자 길이의 아래 위 1.5배가 되는 가상선인 아래쪽 울타리와 위쪽 울타리로 이상치를 구분한다.



[그림 6-10] 로그 변환된 종사자당 월평균 매출액의 산업분류별 상자그림(3•IQR 사용)

1인당 월평균 매출액의 규모는 큰 값 쪽으로 치우쳐 있는 분포를 가지게 되어 로그변환한 후 산업분류별로 상자그림을 나타내었다([그림 6-10]). 상자그림의 이상치 한계선으로는 3*IQR을 사용하여 극단 이상치를 구분하였다. 그림에서 중위수는 산업분류 60에서 가장 높고 74에서 가장 낮으며, 68과 75 분류에서 자료가 넓게 퍼져 있음을 알 수 있다. 68과 75, 87과 91에서 다소 극단 상한 이상치가 존재하고 극단 하한 이상치는 75, 85, 86, 87, 90, 91, 96에서 나타나고 있다. 극단 이상치의 상한과 하한이 각 산업분류별로 상이함을 알 수 있어 종사자당 월평균 매출액의 점검은 산업분류별로 구분하여 검토가 필요하다.

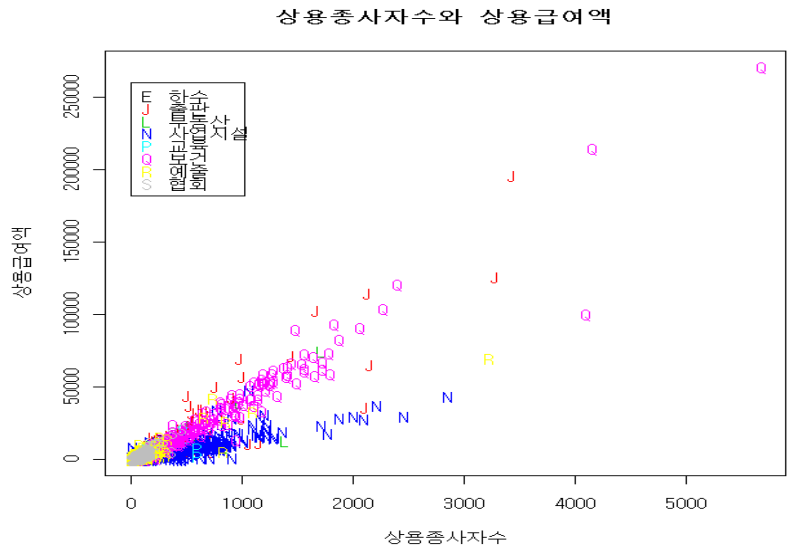
[그림 6-11]은 종사자수와 매출액을 산업분류별 문자로 구분하여 나타낸 산점도이다[6]. 예상한대로 업종별로 그 증가율이 달리 나타나고 있음을 알 수 있다. 부동산업 및 임대업(L)과 예술・스포츠 및 여가관련 서비스업(R)은 종사자수당 매출액 증가율이 가장 높으며 출판・영상・방송통신 및 정보서비스업(J)도 그 증가율이 높다. 반면 보건업 및 사회복지서비스업(Q)과 사업시설관리 및 사업지원 서비스업(N)은 상대적으로 매우 낮은 매출액 증가율을 보인다. 교육서비스업(P)과 협회 및 단체, 수리 및 기타 개인 서비스업(S)은 종사자수와 매출액이 작은 규모에 집중되어 있다.



[그림 6-11]산업분류코드로 나타낸 종사자수와 매출액의 산점도

한편, 상용종사자수와 상용급여액의 산점도를 산업분류별 문자로 구분하여 나타내보면 [그림 6-12][7]와 같다. 역시 상용종사자수와 상용급여액간 양의 상관관계가 산업분류별로 달리 나타나고 있음을 그림을 통해 알 수 있다. 특히 출판・영상・방송통신 및 정보서비스업(J)가 가장 높으며 보건업 및 사회복지서비스업(Q)이 상대적으로 높다. 반면 사업시설관리 및 사업지원 서비스업(N)은 상대적으로 낮은 것으로 나타났다. 앞에서와 마찬가지로 교육서비스업(P)과 협회 및 단체, 수리 및 기타 개인 서비스업(S)은 상용종사자수와 연간급여액이 모두 작은 규모에 집중되어 있다.

6) 영업월수가 12개월인 자료만 이용
7) 영업월수가 12개월인 자료만 이용



[그림 6-12] 산업분류코드로 나타낸 상용종사자수와 상용급여액의 산점도

2. 이상치 검색 및 처리 방법

이상치는 거리 측도에 의해 평가되는 극단적 관측값이다. 이상치(outlier)는 조사의 결과에 큰 영향을 주거나 오류인 까닭으로 검토되어야 한다. 일반적으로 조사업무 담당자는 자료 수집과정에서 나타날 수 있는 이러한 이상치를 검출한다. 일단 의심스러운 건수가 검출되면 원 조사표를 검토하거나 응답자를 재접촉하여 자료가 정확한 것인지를 확인하여야 한다. 따라서 주어진 시한 내에 가능한 많은 오류를 수정하기 위해서는 가장 오류일 가능성이 높은 건수를 정확하게 식별하는 것이 필요한데 이 과정이 효율적 에디팅의 중요한 부분이 된다.

한편, 자료의 이상치, 패턴, 자료 속에 내재된 관계는 순수하게 수치·분석적인 방법만으로 찾아내기 쉽지 않다. 그래픽 에디팅 (graphical editing)은 사람의 시각적 인지력을 이용하여 이들을 쉽게 검출하는 방법으로 그동안 많은 국가 통계 기관에서 통계조사의 비용을 줄이는 효율적인 도구로 사용되어 왔다(Houston, 1993; Esposito, 1994; Engstrom, 2005). 이는 그래프를 이용하는 방법이 에디팅 과정을 개선하고 관리하는데 도움이 되기 때문이다.

가. EDA 방법

탐색적 자료 분석(Exploratory Data Analysis) 기법은 자료에서 탐색해내기 힘든 것

을 용이하게 찾는 통계적 방법이다. 탐색적 자료 분석에서 가장 흔히 쓰는 그래프는 산점도와 상자그림이라 할 수 있다. 이것들은 대부분의 자료와 구별되는 자료를 도식화할 때 흔히 쓰인다. 여기서는 주기적인 경제조사 자료와 관련하여 이들의 간략한 쓰임새를 요약한다.

- 산점도(Scatter plot)

산점도는 두 변수 간 관계를 살피고자 할 때 흔히 사용되는 방법으로서 간단하지만 매우 유용한 도구이다. 그런데 현재와 과거의 자료에 대한 산점도의 경우, 기울기가 1인 직선에 근접한 자료는 현재의 자료가 과거 자료와 일치하는 사업체를 의미하며 이 선에서 멀어질수록 불일치하다는 것을 나타내므로 사업체조사에서 유용하게 쓰일 수 있다(Bienias 등, 1994).

- 상자그림(Box plot)

상자그림은 최소값, 제1사분위수, 중위수, 제3사분위수, 최대값의 다섯 개의 숫자를 상자 형태의 그림으로 자료를 요약·표현하여 이상치를 파악하고자 할 때 유용하게 사용된다. 이는 자료의 중심과 퍼짐, 그리고 형태를 간결하게 나타내는 도구로서 자료가 대칭적이라면 이러한 이상치 검토를 필요로 하는 자료를 도식화하는데 좋은 도구가 된다. 그러나 사업체조사에서 흔히 볼 수 있는 경제 자료는 자료 속성상 오른쪽으로 길게 늘어진 형태로 나타나곤 한다. 만약 자료가 이와 같이 한 쪽으로 치우치는 속성을 갖는다면, 꼬리 쪽의 많은 자료가 실제로는 그 분포에서 일어날 수 있는 값이지만 이상치로 표시된다. 따라서 이러한 경우에 원래의 값에 대해 상자그림을 이용하여 이상치를 발견하는 것은 그리 유의하지 않을 수 있다.

한편, 현재년도 출하액과 과거년도 출하액의 비율(ratio)을 검토하여 이상치를 판단한다고 하자. 즉 비율 자료들의 중위수에서 멀어질수록 이러한 자료는 검토되어야 마땅할 것이므로 비율에 대한 상자그림이 유용할 것이다. 그런데 예를 들어 비율 자료들의 중위수가 0.8이라 하자. 만약 일반적으로 비율의 중심이 1이라 생각하고 0.8과 1.2로 검출조건을 설정하여 이상치를 검색한다면 이는 더 많은 자료가 검출될 뿐만 아니라 이상치를 구별하는데 도움이 되지 못한다. 그러므로 현재와 과거의 비율에서 이상치를 판별할 때는 자료의 분포를 반영하는 것이 필요하다. 다음 절에서는 비율 자료의 상자그림과 유사한 원리를 갖는 Hidiroglou와 Berthelot에 의해 고안된 방법을 좀 더 자세히 살펴보기로 한다.

나. Hidiroglou-Bertherlot 방법

연간조사나 월간조사와 같이 연속적인 조사에서 이상치를 검출하기 위한 방법으로 Hidiroglou와 Berthelot(1986)에 의해 고안된 방법을 대표적으로 사용한다. 이 방법은 현재

의 자료에서 한 항목과 그와 관련된 항목의 비율(ratio)이나, 한 항목의 현재자료와 과거자료와의 비율이 일정 범위 안에 포함되는 지를 판단하는 방법이다. 범위는 변환 비율의 대표값을 중심으로 상한과 하한 값으로 정하여진다. 그러나 한계값을 결정하는 데 있어서 사용자가 설정하는 승수가 포함되어 주관이 완전히 배제되지 않는다는 단점이 있다(박진우외, 2005).

Hidiroglou와 Berthelot에 의해 고안된 방법은 다음 세 가지 방법으로 구분하여 사용될 수 있다. 이 세 가지 방법 중 적절한 선택은 자료에 따라 결정된다.

- 비율을 이용하는 방법(Ratio method)
  - 신뢰할 수 있는 보조 정보가 있을 때 사용
- 과거자료를 이용하는 방법(historical trend method)
  - 과거의 자료가 보조 정보로 취해질 때
  - 비율을 이용하는 방법의 특별한 경우
- 현 자료를 이용하는 방법(Current method)
  - 어떤 보조정보나 과거자료가 없는 경우

과거자료를 이용하는 방법은 각 선택된 변수 x_i 에 대해 다음과 같이 수행된다(Banff Support Team, 2007).

1) $r_i = \frac{x_{i,t}}{x_{i,t-1}}$ 를 계산한다.

$x_{i,t}$: i번째 레코드의 현재 값 $(x_{i,t} > 0)$

$x_{i,t-1}$: i번째 레코드의 과거 값 $(x_{i,t-1} > 0)$

2) s_i 값으로 변환한다.

$$s_i = \begin{cases} 1 - \frac{r_m}{r_i} \ , & 0 < r_i < r_m \\ \frac{r_i}{r_m} - 1 \ , & r_i \geq r_m \end{cases}$$

여기서 r_m은 r_i의 중위수(median)이다. s_i 는 분포의 양쪽 부분에서 이상치가 동등하게 검출될 수 있도록 해 준다. 이는 비율(ratio)의 값이 1보다 크면 과거와 비교해 증가한 것으로, 1보다 작으면 감소한 것을 의미하나 1을 중심으로 대칭이 되지 않는다. 예를 들어 과거자료의 값이 150이고 현재자료의 값이 50이라면 비율은 0.333이고 현재자료의

값이 150이고 과거자료의 값이 50이라면 비율은 3이다. 즉 같은 100의 변화를 갖더라도 비율의 값은 1을 중심으로 대칭이 되지 않는다. 그러나 변화된 비율 s_i는 0에 대해 좌우 대칭이 되도록 한다.

3) 효과(effect) $e_i = s_i[\max(x_i, y_i)]^u$ 를 계산한다.

사용자가 u값을 정할 수 있는데, 예를 들면 $u = 0$은 관측값이 크거나 작거나 똑같이 취급하고(이 경우 e_i 는 s_i 값과 같음), $u = 1$은 큰 단위의 작은 변동에 더 큰 의미를 부여한다.

4) 제1사분위수 e_{q1} , 중위수 e_m , 그리고 제3사분위수 e_{q3} 를 계산하여

$$e_{D_L} = Max(e_m - e_{q1}, |K|)$$
$$e_{D_U} = Max(e_{q3} - e_m, |K|)$$

를 구한다. 각 사분위수가 너무 가까우면 사용자가 설정한 중위수의 일정배수 K를 채택하게 한다. 이제 극단 이상치를 식별하는 범위는 아래와 같다.

$$e_i < e_m - c_1 e_{D_L} \text{ 또는 } e_i > e_m + c_1 e_{D_U}$$

그리고 보통 이상치를 식별하는 범위는 아래와 같다.

$$e_m - c_1 e_{D_L} \le e_i < e_m - c_0 e_{D_L} \text{ 또는 } e_m + c_0 e_{D_U} < e_i \le e_m + c_1 e_{D_U}$$

여기서 사용자 설정 승수 c_1 과 c_0를 설정시 c_1 은 c_0 보다 커야한다 (일반적으로 c_1 = 6, c_0= 3).

■ 예제: 이상치 검출(H-B 방법)

현재 광업·제조업 조사에서는 주요항목별로 전년도 대비 변화를 검토하고 있으며 증감률을 지역별로 달리 설정하여 검토함으로써 지역별로 에디팅 업무량 조정이 가능하게 되었다(통계청, 2008a, 2008b). 그러나 여전히 검색조건을 반복하여 결정하는 번거로움이 있을 뿐만 아니라 전체자료의 움직임이나 패턴을 쉽게 이해할 수 없다. 또한 레코드를 개별적으로 검토하는 방식 대신 여러 레코드를 종합적으로 검토하는 것은 총 결과에 미치는 개별 레코드의 영향을 가늠하는데 있어서도 중요하다. 더 나아가 증감률의 검토 기준 설정시 자료의 중심과 산포 등의 정보를 반영할 수 있다.

다음 예제는 2006년과 2007년도에 각각 실시한 2005년과 2006년 기준 광업·제조업 조사 자료[8]에 EDA와 H-B기법을 적용한 결과이다. 적용될 주요항목은 종사자수, 급여액, 출하액, 주요비용이다. 조사 자료에서 사업체 고유번호가 두 조사에 존재하는 경우

8) 자료 이용의 한계로 이미 에디팅이 완료된 자료를 사용하였음.

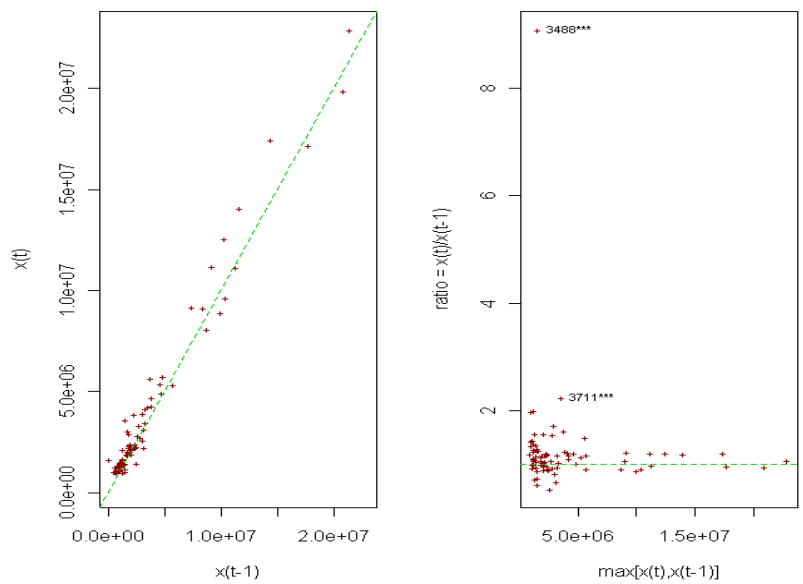
에 자료를 병합하였다. 이 결과 102,406건의 사업체가 일치되었다. 이 자료에서 종사자수, 급여액, 출하액, 주요비용 항목의 두 시점 간 자료에 대해 각각 H-B기법과 그래픽 방법을 적용하였다. 통계 프로그램 R(http://www.r-project.org/)을 이용하여 작성하였다.

- **주요항목별 현 • 전 시점 자료 간 산점도와 비율**

먼저 주요항목별로 현재년도와 과거년도의 두 자료 간 산점도와 비율을 살펴본다. 여기서는 출하액, 주요비용 항목에서 규모가 가장 큰 범위에 속하는 사업체에 대해 산점도와 비율 그래프를 통하여 이상치를 파악한다.

[그림 6-13]은 2005년도와 2006년도의 광업 • 제조업 조사에서 2006년도에 1조 이상 출하액을 갖는 사업체를 대상으로 한 결과이다. 좌측이 출하액 간 산점도이며 우측이 과거 출하액과의 비율을 보여준다. 그림에서 점들이 직선 근처에 흩어져 있어 2006년도의 출하액이 2005년도와 비교하여 대체로 큰 변동이 없는 것으로 나타난다. 그러나 오른쪽 그림에서 보면 상대적으로 작은 출하액에서 그들의 비율이 매우 큰 점들이 발견되고 있다. 이러한 비율의 그림을 통해 보면 산점도에서 잘 볼 수 없는 특이한 점을 쉽게 발견할 수 있다.

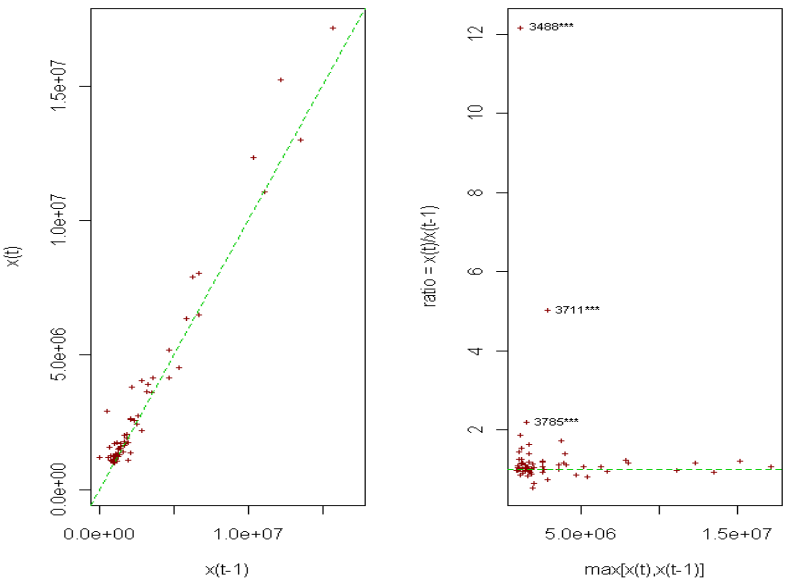
특히 사업체 중 3488*** 사업체는 다른 사업체와 달리 2005년도에 비해 약 9배의 출하액을 나타내고 있다. 또한 3711*** 사업체는 약 2배 이상의 출하액 증가를 보이고 있어 다른 사업체와는 확연히 구분되고 있다.



[그림 6-13] 출하액간의 산점도와 비율 (2006년도 출하액 1조 이상)

[그림 6-14]는 2006년도에 주요비용이 1조 이상인 사업체를 대상으로 하여 그린 산점도와 과거 주요비용과의 비율을 보여준다. 2006년도에 주요비용이 1조 이상으로 조사된 사업체의 전년도 주요비용과의 비율을 보면 3488*** 사업체가 다른 사업체와는 매우 다르게 2005년도에 비해 약 12배의 주요비용의 증가를 보이고 있으며 3711*** 사업체도 약 5배의 증가를 보이고 있다.

그런데 앞에서 3488*** 사업체와 3711*** 사업체는 출하액에서도 그 증가가 두드러지게 나타난 사업체로서 주요비용의 증가가 예상되는 사업체이다. 따라서 두 개의 사업체는 모두 항목 간 일치된 결과를 보이고 있다. 이를 통하여 한 항목의 현전년도의 이상변동은 다른 관련된 항목의 현전년도의 변동으로 재검토되어질 수 있음을 알 수 있다.



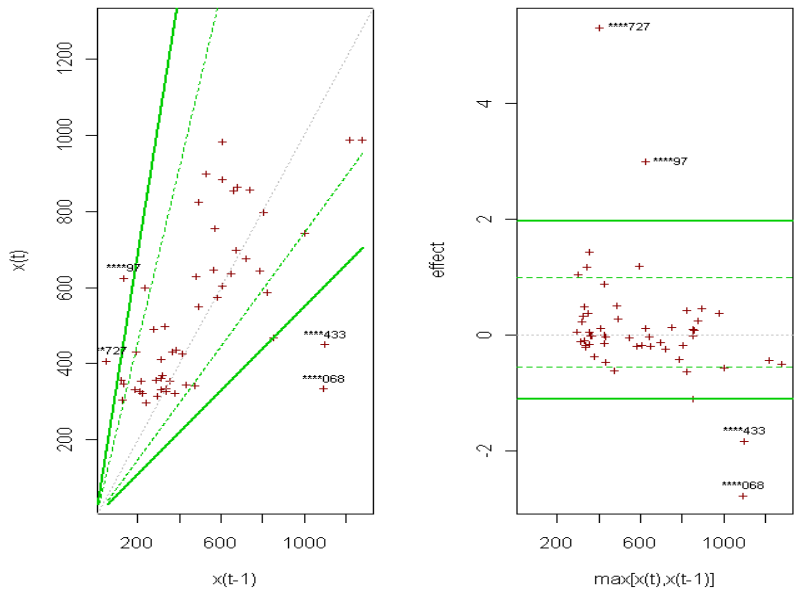
[그림 6-14]주요비용간의 산점도와 비율 (2006년도 출하액 1조 이상)

- **특정지역의 현·전 시점 자료 간 산점도와 이상치 검출**

앞에서는 전국자료의 각 주요항목에 대하여 과거년도와의 비교를 통해 특이점을 검색하였으나 여기서는 특정 행정구역별 자료를 선택하여 살펴본다. 주요항목에 대한 산점도와 비율에 대한 도표는 각 행정구역별로 적용될 수 있다. 하나의 예로서 ○○시 ○○구 사업체를 각 규모별로 주요항목에 대해 산점도를 작성하였다. 이와 더불어 앞 절에서

소개된 H-B 한계값을 그림 위에 제시하였다. H-B 한계값을 계산 시, 사용자가 설정하는 승수 c_1 과 c_0는 각각 6과 3으로 고정하였다.

[그림 6-15]의 왼쪽 그림은 2006년도 급여액이 3억 이상 10억 미만인 사업체의 경우 현・전 급여액 간 산점도와 그 위에 H-B 방법에 의해 계산된 한계선을 표시한 그림이다. 오른쪽 그림은 현전 급여액의 효과(effect)를 나타낸 그림이다. 앞의 전국자료에서는 현・전시점 자료의 비율을 도식화하였으나 이 그림에서는 변환된 비율인 효과 $e_i = s_i[\max(x_i, y_i)]^u$ 의 값을 도식화하였다. 여기서는 $u = 0$을 적용한 경우의 결과이며 $u = 0$일 경우는 $e_i = s_i$가 된다.

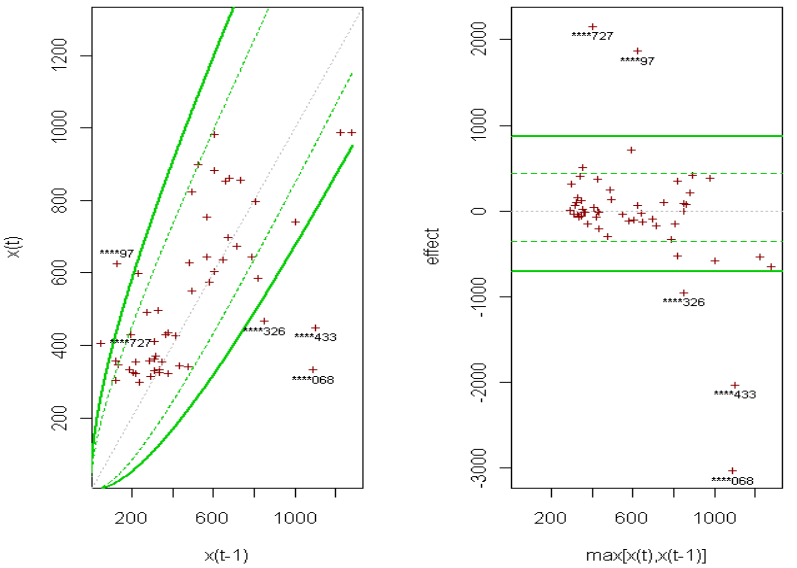


[그림 6-15] 특정지역 급여액 간 산점도와 효과 (2006년도 급여액 3억 이상 10억 미만)

그림에서 보듯이 ****727 사업체는 56백만원에서 408백만원으로, ****97 사업체는 136백만원에서 627백만원으로 각각 상한선 밖에 있다. ****433 사업체는 1,107백만원에서 452백만원으로, ****068 사업체는 1,096백만원에서 336백만원으로 감소하여 하한선 밖에 나타나고 있다. 즉 이들 사업체는 급여액에 있어서 전년도에 비해 급격히 증가하거나 감소한 사업체임을 보여주고 있다.

[그림 6-15]의 왼쪽 그림은 산점도 위에 $u = 1$을 적용한 경우의 한계선을 표시한 결과이고 오른쪽은 $u = 1$일 때의 효과값이다. [그림 6-16]과 거의 비슷한 결과를 보여 주고

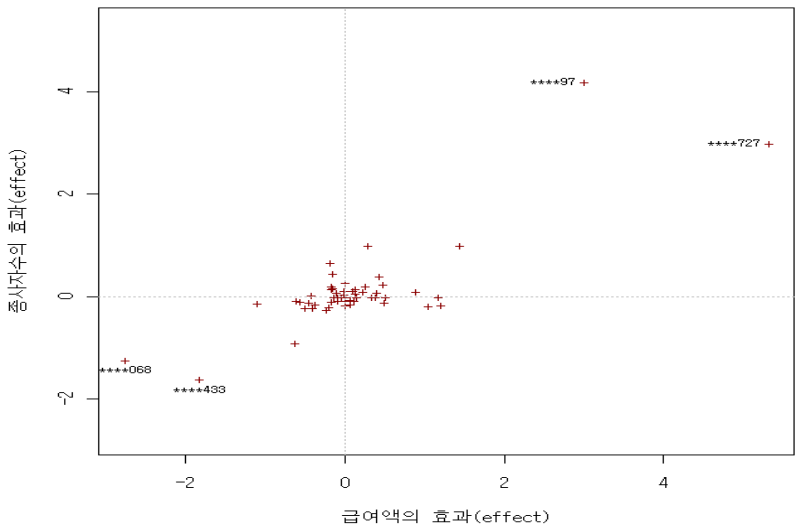
있음을 알 수 있다. 그러나 한계선은 더 이상 직선이 아닌 곡선으로 표현되며 규모가 커질 때 하한과 상한을 좁혀 줌으로써 이상치가 민감하게 검출되도록 한다. 이 결과로 ****326 사업체가 하한 밖에 있어 이상치가 추가로 구별되고 있다. 따라서 규모를 고려하여 좀 더 민감하게 이상치를 구별하고자 하는 경우에 이 방법을 사용할 수 있을 것이다.



[그림 6-16] u=1인 경우의 한계선과 효과(2006년도 급여액 3억 이상 10억 미만)

이제 [그림 6-17]을 보자. 이 그래프는 급여액의 효과값을 가로축에, 종사자수의 효과값을 세로축으로 하여 작성된 산점도이다. 그림을 살펴보면, 우측 상단의 두점은 급여액과 종사자수 항목 모두에서 큰 효과값을 가지고 있음을 나타낸다. 즉 급여액의 특이한 증가는 종사자수의 특이한 증가로 일치자료일 가능성이 매우 높음을 시사한다. 또한 좌측 하단의 두 점 역시 두 개의 항목 모두에서 특이하게 감소하고 있어 서로 일치하고 있다.

이는 한 항목의 과거 값과 비교하였을 때는 특이자료로 판단되나 관련된 다른 항목의 변화를 살펴보면 실제 가능한 자료라는 것을 의미한다. 즉 이 그래프를 이용하면 재검토의 시간과 비용을 줄일 수 있으며 더 나아가 이상치로 제거될 위험을 피할 수 있다. 한편 왼쪽 상단 부분이나 오른쪽 하단부분에 사업체가 위치한다면 이러한 사업체에 대해서는 신중하게 검토해야 할 것이다.



[그림 6-17] 효과 대 효과 산점도

- 토의

이상에서 우리는 특정 영역에서 선택한 각 사업체의 주요항목별 이상치를 산점도와 H-B 방법을 이용하여 검출하였는데, 그 밖의 지역 및 규모에 대해서도 유사한 방법으로 이를 적용할 수 있을 것이다. 특히 하나의 항목의 이상치는 그 항목과 관련된 다른 항목의 현·전 효과값과의 그래프를 통하여 살펴보면 시각적으로 자료의 내용을 재검토할 수 있다.

이러한 그래픽 에디팅은 자료의 분포와 자료에 내재된 움직임을 함께 반영한 결과라는 점에서 중요한 의미를 갖는다. 즉 경기변동이나 특수 지역, 품목 상황이 이상치 탐색에 반영될 수 있다. 따라서 에디팅의 판단기준이 증감률 ±50%와 같이 더 이상 대칭적으로 나타나지 않는다.

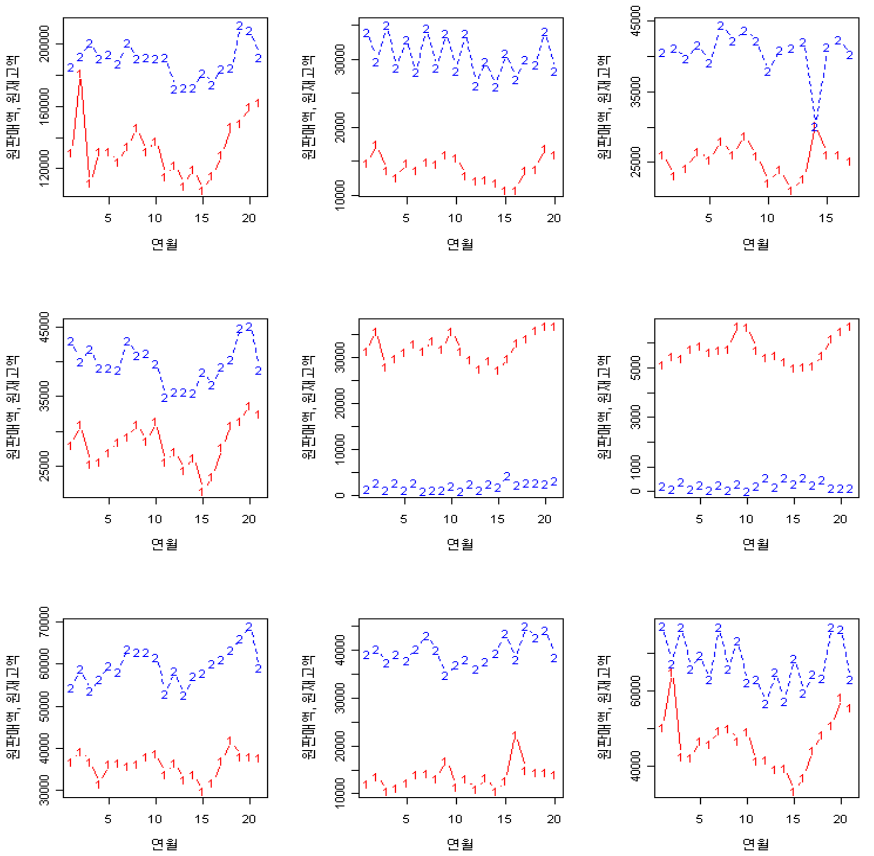
한편 화면에 출력된 이상치를 컴퓨터에서 클릭할 경우 바로 그 레코드의 구체적인 값들을 볼 수 있도록 시스템을 구현한다면 더욱 손쉬운 에디팅 작업이 진행될 수 있을 것이다.

예제: 서비스업통계 도소매업 월별 시계열(2008년 1월~2009년 9월)

서비스업통계 도소매업 월별 재고액과 판매액의 시계열 그림을 시각적으로 이용하면 자료의 특성과 패턴을 이해하고 자료 에디팅을 용이하게 할 수 있을 것이다. 특히 재고액의 이상치는 전월자료를 이용한 H-B 이상치 판단기법과 그래픽을 이용한 방법을 통

해 검출한다면 자료의 품질관리에 도움이 될 것으로 판단된다.

[그림 6-18]에서 첫 번째와 마지막 시계열 그래프의 앞부분에서 상당한 폭의 변화가 있음을 시각적으로 확인할 수 있다. 또한 두 번째 시계열 그래프에서는 재고액의 월별 변화가 규칙적으로 변하고 있음은 특이하다.



[그림 6-18] 각 사업체의 재고액과 판매액 시계열 그림

다. 이상치 처리 방법

앞에서 하나의 변수 또는 밀접한 관련이 있는 두 변수를 이용한 이상치에 대한 검출 방법을 소개하고 이상치를 검출하였다. 이상치(outlier)는 추정 가중치와 연계하여 모집단의 다른 사업체를 대표하지 못한다고 생각되는 비정상적인 값을 의미한다. 가중치가

이상치에 적용되면 관심 변수에 대하여 과다추정이나 과소추정을 하게 된다.

잘못 층화된 사업체는 추정치를 만들 때 배제하면 안 된다. 왜냐하면 그 사업체는 모집단의 일부분을 이루고 있고 모집단에 대한 정보를 제공하기 때문이다. 또한 사업체는 그 자체에 대한 정보도 제공하고 있다. 중요한 문제는 그 사업체가 동일한 표본 층에서 다른 사업체를 대표하고 있는지의 여부이다.

일반적으로 이상치는 가중값 조정, 값 조정, 동시 조정, 로버스트 추정을 통해 처리한다. 이러한 이상치의 처리는 추정량의 비편향성에 손상을 주지만 분산을 줄여주는 역할을 한다. 실제 데이터 분석과 모의실험을 통해서 주어진 자료에 알맞은 이상치 검출 및 처리방법을 찾게 된다. 표본조사에서 이상치의 영향을 줄이기 위하여 가장 널리 사용되는 두 가지 방법은 예외적인 이상치(surprise outliering) 방법과 윈저화(winsorization) 방법이다. 이 절에서는 이 두 가지 처리방법을 살펴본다.

1) Surprise outliering 방법

Surprise outlier란 사업체에서 조사된 값이 정확하게 보고된 값이지만 동일한 표본층에 있는 다른 사업체를 대표하지 못한 경우를 일컫는다. 추정치를 계산할 때 예외적인 이상치는 단지 1의 가중치 갖도록 처리한다. surprise outlier를 나타내는 사업체의 나머지 가중치는 이상치를 포함하고 있는 표본층의 다른 사업체에 배분된다. 따라서 추정 항목은 가중치 조정에 의하여 영향을 받는다.

조사된 사업체 자료를 surprise outlier로 해야할지 결정하기 위해서는 사업체가 surprise outlier로 처리될 때 추정치에 미치는 영향을 고려하는 것이 도움이 된다. 또한 추정치 편의(bias)를 최소화할 수 있도록 현 시점의 예외인정 이상치의 개수와 추정치에 미치는 영향 정도를 이전에 실시한 조사결과와 비교분석해야 한다.

2) Winsorization 방법

Winsorization 방법이란 이상치를 찾아내어 보다 합리적인 값으로 대체해 주는 이상치 처리방법을 말한다. Winsorization은 이상치를 다른 적절한 값으로 수정하는 것이며 추정가중치를 바꾸지 않는다. 먼저 모집단에 대한 과거정보를 이용하여 관심 대상이 되는 항목에 대하여 구분값(cutoff)을 정하고 이 구분값보다 큰 모든 값은 사전적으로 정해진 구분값에 가까운 값으로 대체한다.

Winsorization 방법이 모든 조사에 대하여 적절한 방법이라고 할 수는 없다. 이 방법은 관심변수에 대하여 과거자료가 있어야 적용할 수 있으며 관심변수가 비교적 오랜 시간에 걸쳐 안정적이고 예측 가능할 때 사용하는 것이 좋다.

3. 결측치 대체 방법

무응답은 조사단위가 응답되지 않는 단위 무응답과 일부 항목이 응답되지 않는 항목 무응답으로 나뉘며 단위 무응답의 경우 재조사나 무응답 가중치 조정을 통하여 처리하고 항목무응답의 경우 주로 대체를 통하여 처리한다.

응답군의 특성과 무응답군의 특성이 같다면 문제가 없으나 일반적으로 무응답으로 인해 편향 추정의 우려가 있으며 표본수의 감소로 인한 추정의 효율이 저하될 수 있다. 무응답 대체의 목적은 이러한 편향을 보정하고 추정 효율을 높이고자 하는 데 있다. 결측치의 대체 방법에는 확률적 대체법과 결정론적 대체법으로 나누어지는데, 결정론적 대체는 주어진 응답자의 자료에 대해 하나의 대체값으로 결정되는 반면 확률적 대체는 랜덤성을 부여하여 대체될 때마다 다른 값으로 대체된다. 예를 들면 비 추정 재고액에 무작위 잔차를 더함으로써 결정론적 대체를 확률적으로 만들어 주변 분포를 보존하는 방법이다.

모평균의 추정에 관심이 있는 경우 회귀대체법 등 결정론적 대체법을 사용할 수 있으나 모분산 등 다른 모수의 추정에도 관심이 있는 경우에는 확률적 대체로 주변 분포를 보존해야 한다.

- 평균값대체

대체군내에서 평균값으로 대체하는 방법이다. 대체군은 대체를 필요로 하는 응답자와 같은 범주에 있는 것으로 판단되는 응답자로 구성되며 동일한 유형, 업종, 규모를 바탕으로 대체군을 형성한다. 부가정보가 없을 때 극소수의 레코드만을 대체 시 사용한다. 이 방법은 편향을 일으키지는 않으나 분포가 파괴될 수 있다.

- 비추정대체

부가적 정보 사용 가능시, 둘 또는 그 이상의 변수들 간에 비를 만들어 사용한다. 예를 들면, 재고액=a·판매액+e로 재고액은 판매액과 선형관계를 의미한다. 또 다른 예를 들면,

$$\text{해당사업체의 재고 대체액} = \frac{\text{동일 산업분류내 재고액 평균}}{\text{동일 산업분류내 응답된 판매액 평균}} * \text{해당사업체의 판매액}$$

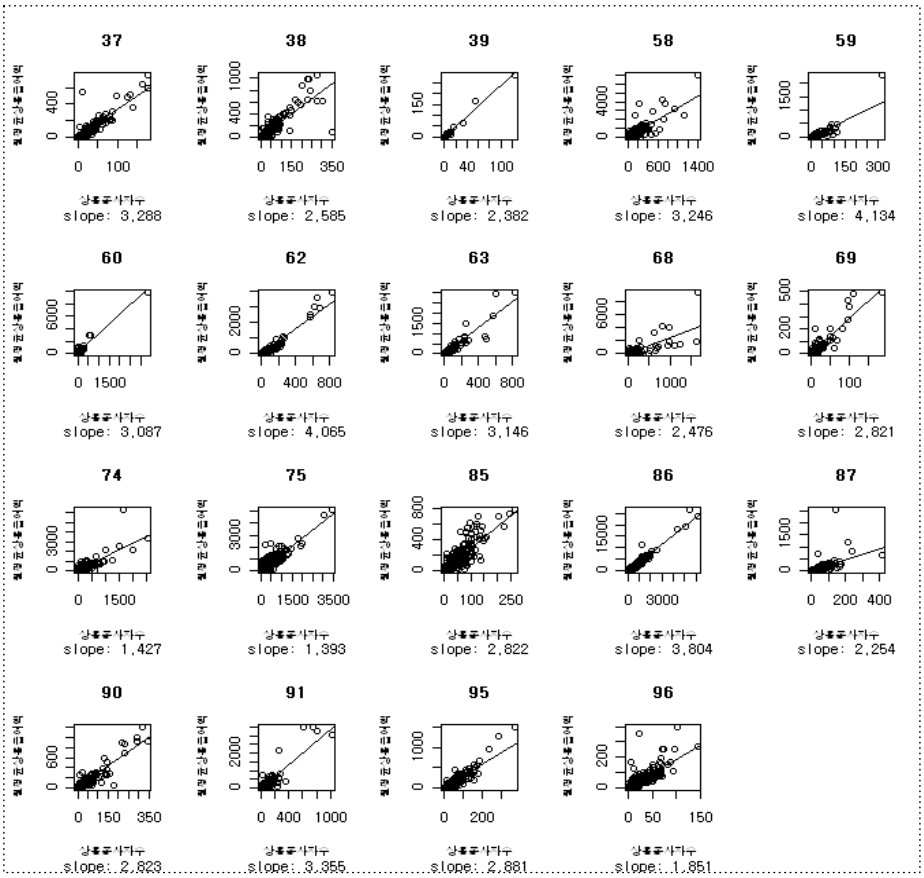
으로 대체할 수 있다. 이 방법이 사업체조사에서 가장 많이 사용되는 방법이다.

- 회귀대체

회귀대체 방법은 회귀직선을 추정하고 추정된 회귀직선을 이용하여 무응답을 대체하는 방법이다. 예를 들면 재고액을 반응변수로 하고 판매액을 설명변수로 하여 회귀방정식을 구한다. 보조변수가 2개 이상이면 중회귀모형을 적합한 후 회귀대체를 수행한다.

예제: 회귀식을 이용한 대체(2008년 기준 서비스업 조사)

월평균 상용급여액을 상용종사자수로 원점을 지나는 회귀식 모형을 가정하고 산업분류별로 회귀계수를 추정하였다. [그림 6-19]에서 각 산업분류별 기울기가 추정되었으며, 상용종사자수가 있는데 연간상용급여액이 누락(필수)된 경우 이 회귀식을 이용하여 상용급여액을 추정・제시하였다(이의규, 2010b).



[그림 6-19] 산업분류별 상용종사자 1인당 월평균 급여액의 회귀계수 추정

- 핫덱 결측치 대체

사업체 조사보다는 가계조사 자료의 무응답 대체법으로 많이 사용하는 방법이다. 광범위한 정량자료 조사에서는 정량적 자료 중에서 대응변수를 통해 제공레코드를 찾는 것이 바람직할 수도 있다.

참고로 Survey methods and practices(캐나다 통계청, 2003)에서 제시된 결측치 대체 지침서의 내용을 발췌하여 소개한다.

- 대체된 레코드는 에디팅할 수 없었던 레코드와 유사해야 한다.
- 좋은 결측치 대체는 평가목적을 위해 피검준비를 해야 한다.
- 결측치 대체된 레코드는 모든 에디팅 과정을 충족시켜야 한다.
- 결측치 대체법은 대체될 자료의 유형을 고려하면서 신중히 선택되어야 한다.
- 결측치 대체법은 가급적 무응답 편향을 줄이고 항목간의 관계를 보존하는데 목적을 두어야 한다.
- 결측치 대체 시스템은 사전에 숙고, 특화, 프로그램화와 시험의 단계를 거쳐야 한다.
- 그 과정은 자동화, 객관화, 재생가능하고 효율적이어야 한다.
- 결측치 대체 시스템은 어떠한 형태의 결측 또는 비일치 변수라도 다룰 수 있어야 한다.
- 제공자 결측치 대체법의 경우, 대체 레코드는 선택된 제공레코드와 밀접하게 닮아 있어야 한다.

제5절 맺음말

시간과 예산이 충분하다면 실제 응답한 자료를 얻어내는 것이 가장 좋을 것이다. 그러나 현실적으로 모든 오류를 찾아내고 수정하는 것은 불가능하다. 따라서 우리는 제한된 자원으로 효율적인 에디팅 업무 전략 수립이 필요하다. 해외 통계 선진국에서는 이미 예산감축과 조사환경 악화로 지금까지 앞에서 제시된 선별적 에디팅과 자동 에디팅을 적절히 병행하여 효율성을 높이고 있다.

또한 에디팅의 개요에서 언급하였듯이 데이터 에디팅은 오류를 찾아내고 수정하는 것이 본연의 목적은 아니다. 에디팅의 근본 목적은 오류의 원인을 찾아내어 조사과정을 개선하는 것이고 두 번째는 자료의 품질에 대한 정보를 제공하기 위함이다. 특히 사후 수정보다는 사전 예방이 비용면이나 시간적인 측면에서 바람직하다 하겠다. 따라서 에디팅 결과보고서 등을 작성하고 평가하여 조사의 개선을 위해 노력해야 할 것이다.

우리는 본 매뉴얼에서 선별적 에디팅, 쌍방향 에디팅, 자동 에디팅, 매크로 에디팅, 그래픽 에디팅 등 다양한 에디팅 방법론을 소개하고 실무에 응용할 수 있도록 실무 예제와 함께 살펴보았다. 본 매뉴얼에서는 주로 에디팅의 이해와 에디팅 방법론에 중점을

두고 있어 이상치와 결측치를 대체하는 방법에 대해서 간략하게 설명하였다. 본 에디팅 매뉴얼은 이러한 처리방법에 대한 내용이 추가되고 보완되어야 할 것이다.

한편 국내외에서는 행정자료의 사용이 증가하고 있어 향후 행정자료의 효율적 이용 및 연계와 관련된 에디팅 연구에 좀 더 많은 관심이 필요할 것이다. 본 매뉴얼이 경제조사에 있어서 에디팅의 이해와 응용에 조금이라도 도움이 되기를 기대한다.

참고문헌

김규성(2008), 에디팅 품질관리 매뉴얼, 한국통계학회.

박진우외(2005), 주택가격동향조사를 위한 데이터편집 사례연구, 조사연구, 6권 1호.

류제복외(2003), Imputation Methods for the Population and Hosing Census 2000 in Korea", 한국통계학회논문집 10(2), 575-583.

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

이기재(2011a), "사업체 조사에서 특이치의 검출 및 처리방안", 국가통계포럼, 한국통계학회 국가통계연구회/통계개발원.

이기재(2011b), "사업체 조사 선택적 에디팅", 제1회 국가통계 방법론 심포지엄, 통계개발원.

이의규외(2007), "사업체대상 조사의 자동내검기법", 통계개발원.

이의규외(2008), "수량적 연관규칙 위배자료의 자동오류위치포착", 통계개발원.

이의규외(2009a), "자동오류위치포착 및 수정방안",「통계자료의 내검기법 연구」, 통계개발원.

이의규외(2009b), "그래픽내검기법을 이용한 내검효율성 제고", 「통계자료의 내검기법 연구」, 통계개발원.

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

이의규(2009d), "자동내검기법의 적용방안 -서비스업조사를 대상으로-", 통계개발원.

이의규(2010a), "자동내검기법의 적용 및 분석 -서비스업조사를 대상으로-", 통계개발원.

이의규(2010b), "도소매업 재고액 무응답 대체", 통계개발원.

이의규(2010c), "주기적 조사 자료의 내검: 그래프 활용을 중심으로", 통계연구 제 15권 제1호, 16-27

이의규(2011), "합계 불일치 오류의 자동수정", 통계개발원.

통계개발원(2010), "호주의 에디팅 방법과 사례", 해외 전문가 초청 강연회, 통계개발원.

통계청(2008a), 「2007년 기준 광업·제조업 통계조사 조사지침서」.

통계청(2008b), 「2007년 기준 광업·제조업 통계조사 입력시스템 이용자 지침서」.

Austrailian Bureau of Statistics(2007), The editing guide, Beta v4.

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

Bienias, L. Julia, D.M. Lassman, S.A. Scheleur, and H. Hogan(1994), "Improving Outlier Detection in Two Establishment Surveys", UN/ECE Work Session on Statistical Data Editing.

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.(2003), “Processing of Erroneous and Unsafe Data”, Ph. D. Thesis, Erasmus University Rotterdam.
De Waal, T.(2008), "An overview of statistical data editing", discussion paper, Statistics Netherlands.
De Wall, T. and Coutinho, W. (2005), “Automatic Editing for Business Surveys: An Assessment of Selected Algorithms”, International Statistical Review, 73, 1, 73-102.
Engström, P. and C. Ängsved(2005), "A Graphical Macro-Editing Application", UN/ECE Work Session on Statistical Data Editing.
Esposito, R., J.K. Fox, D. Lin, and K. Tidemann(1994), "ARIES: A Visual Path in the Investigation of Statistical Data", Journal of Computational and Graphical Statistics, Vol. 3, No. 2, pp. 113-125.
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. (1995), “Improving the Traditional Editing Process”, Business Survey Methods, eds. B.G. Cox, D.A. Binder, B.N. Chinappa, A. Christianson, M.J. Colledge, and P.S. Kott, New York: Wiley, 385-401.
Granquist, L. (1997), “The New View on Editing”, International Statistical Review, 65, 3, 381-387.
Granquist, L. and Kovar, J.G.(1997), “Editing of Survey data: How much is enough? in Survey Measurement and Process Quality”, New York: Wiley, 415-435.
Hedlin, D. (1993), “A Comparison of Raw and Edited Data of the Manufacturing Survey", unpublished report, Statistics sweden, Stockholm, Sweden.
Hidiroglou, M.A. and J.M. Berthelot(1986), “Statistical Editing and Imputation for Periodic Business Surveys”, Survey Methodology, 12, pp.73-84.
Houston, G. and A.G. Bruce(1993), “gred : Interactive graphical editing for business surveys”, journal of official statistics vol.9.no.1, 1993, pp. 81-90.
Pannekoek, J. and De Waal, T. (2003), “Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the Euredit Project”, Discussion paper 03011, Statistics Netherlands, Voorburg.
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.
Statistics Canada(2003), Survey methods and practices, Ottawa.