조사동향

주택가격동향조사를 위한 데이터편집 사례연구

A Case Study of Data Editing for the Korean Housing Price Survey

박진우* · 박현주** · 김진억***
Jin－Woo Park · Hyun－Joo Park · Jin－Eok Kim

대규모 통계조사에서 수집된 데이터에는 오류나 결측값의 문제가 발생하기 마련이다. 조사, 데이터 입력, 데이터 처리 등의 과정에서 여러 가지 요인에 의해 이런 문제가 생길 수 있는데 이런 데이터를 방치한 채 통계를 생산할 경우 편향이나 다양한 분석에서의 불일치의 문제가 발생하게 되어 통계의 품질과 신뢰성을 떨어뜨릴 수 있으므로 수집된 데이터의 오류나 결측값을 찾아 수정하는 데이터편집은 매우 중요한 작업이다. 해외에서는 데이터편집의 문제를 공론화하여 다루고 있는 데 반해 우리나라에서 데이터편집에 관한 논의는 거의 없는 편이다. 본 연구의 목적은 주택가격동향조사를 위한 데이터편집의 사례를 소개함으로 데이터편집에 대한 논의의 폭을 넓히는 데 있다. 조사목적에 맞도록 편집규칙을 정하는 과정 및 관련 자료들을 소개하고, 온라인조사라는 조사방식에 맞는 입력 데이터편집방법을 마련하여 실시하는 예들을 소개하며, 마지막으로 출력 데이터편집에 의해 입력 편집에서 걸러지지 않은 오류나 문제들을 제거하는 방법도 소개한다.

주제어 : 편집규칙, 특이점, 입력 데이터편집, 출력 데이터편집, 사후층화, 가중값 조정

Large scale survey database may contain some erroneous data or missing data. Incomplete or erroneous data may be produced in the process of data collection or data capture. Since erroneous data can cause some bias and inconsistency, data editing, which is the procedure for detecting and adjusting individual errors in data records, is a very important work in statistical survey. In this paper, we introduce an editing process for the housing price survey to enhance discussions on that topic. We explain how to decide some appropriate edit rules and show some related data. Furthermore, we describe input editing

* 교신저자(corresponding author) : 수원대학교 통계정보학과 교수 박진우.
E-mail : jwpark@suwon.ac.kr
** 수원대학교 통계정보학과 석사과정
*** 국민은행 청약사업팀장

procedures which is appropriate for on-line survey and how to find and eliminate erroneous data through output editing.
key words : editing rule, suspicious data, input editing, output editing, post-stratification, weight adjustment

Ⅰ. 서 론

대규모 조사를 통한 통계를 생산하는 과정에서 자주 부딪히게 되는 문제 중의 하나는 오류나 결측을 포함하는 데이터의 처리 문제이다. 조사, 데이터 입력, 데이터 처리 등의 과정에서 여러 가지 이유들로 인해 오류가 발생할 수 있는데 이때 오류가 있는 데이터를 방치한 채 통계를 생산할 경우 여러 문제가 야기될 수 있다. 첫째, 추정에 큰 영향을 미치는 특이점이 존재할 경우 편향이 생기게 되는데 이와 같은 편향은 통계의 품질을 떨어뜨리는 요인이 된다. 둘째, 조사 데이터를 다양한 이용자들에게 제공하는 경우라면 이용자들이 특이점을 어떻게 취급하느냐에 따라 동일한 데이터를 가지고도 다른 결과를 산출할 수 있으므로 통계의 신뢰성에 의문을 가져올 수 있다. 많은 비용과 시간을 들이는 통계조사의 품질과 신뢰성이 몇몇 오류들에 의해 손상되어서는 안 될 것이다.

데이터 수집 및 처리 단계에서의 오류를 찾아내고 이를 수정하는 작업을 데이터편집(editing)이라고 한다(Granquist 1995). 오류가 있는 데이터를 방치함으로 야기될 수 있는 여러 문제들을 사전에 제거하기 위해 대부분의 통계기관에서는 데이터편집을 수행해 오고 있다. 특히 일회성의 조사가 아닌 계속조사의 경우 데이터편집은 더욱 중요한 문제로 부각된다. Granquist와 Kovar(1997)의 연구에 의하면, 각종 통계기관에서 데이터편집을 위해 소요되는 비용은 전체 조사비용의 20% 내지 40%에 이를 정도인 것을 알 수 있다. 이와 같이 데이터편집은 통계조사에서 중요한 비중을 차지하는 부분이다.

통계조사 과정에서 데이터를 수집하다 보면 항상 데이터편집의 필요성에 직면하게 된다. 데이터편집에 대한 체계적인 연구 결과가 발표되기 훨씬 이전부터도 통계를 생산하는 기관이라면 어디든지 나름의 방법으로 데이터편집을 수행해 왔었다고 할 수 있다. 1960년대 이후 데이터편집방법들을 공론화한 연구들이 나오기 시작하다가 Fellegi와 Holt(1976)에 의해 비로소 이론적으로 체계화되었다. Fellegi와 Holt는 데이터편집을 위한 편집규칙들을 논리적으로 분석하여 데이터편집을 위한 체계적인 알고리즘을 제시하였다. Fellegi와 Holt의 논문이 발표된 이래 세계 각국의 통계기관에서는 데이터편집을 위한 시스템들을 개발해 오고 있을 뿐 아니라 그러한 연구 결과 또한 활발하게 발표하고 있다(Granquist 1995; Waal & Quere 2003). 한편 Hidiroglou와 Berthelot(1986)는 주기적인 사업체조사에서 데이터편집 및 대체에 관한 연구를 하였다.

우리나라의 경우 데이터편집에 관한 논의나 연구가 상대적으로 매우 빈약한 편이다. 여러 기관들에서 중요한 조사에 대해 나름의 방법으로 데이터편집을 실시하고 있을 것이나 이에 대한 공개적인 논의가 제대로 이루어지고 있지 못 한 실정이다. 데이터편집과 관련된 주제가 부분적으로나마 다루어진 연구로는 류제복 외(2002), 이재원(2000), 박성현과 박진우(2004) 정도를 들 수 있다. 따라서 우리나라 조사환경에서의 데이터편집의 적용 가능성과 실태를 알 수 있는 다양한 사례연구가 필요하다고 할 수 있다.

본 논문의 목적은 국민은행에서 매월 수행하는 계속조사인 주택가격동향조사에서의 데이터편집사례를 소개하는 것이다. 국민은행에서는 주택가격동향 통계의 생산을 위해 데이터를 수집, 입력, 처리 등의 각 단계에서 발생할 수 있는 오류의 가능성을 체계적으로 방지하기 위해 일련의 데이터편집체제를 구축하여 작동시키고 있는데 그 구체적인 방법과 내용을 소개하고자 한다. 본 논문에서 다루는 사례는 주택가격동향조사라는 하나의 조사사례이지만 대부분의 계속조사 사례에서 공통적

으로 생길 수 있는 문제들을 다루고 있으므로 다른 여러 조사들의 데이터편집을 위한 좋은 참고사례가 될 것이다. 2절에서는 주택가격동향조사와 이 조사를 위한 데이터편집의 개요를 설명하며, 3절에서는 데이터편집 규칙을 설명한다. 4절에는 각각 입력 데이터편집(input data editing)과 출력 데이터편집(output data editing) 절차가 소개되며 마지막으로 결론을 내린다.

Ⅱ. 주택가격동향조사를 위한 데이터편집

2.1 주택가격동향조사

국민은행에서 매월 조사하여 작성, 발표하는 전국주택가격동향조사 통계는 전국 주택시장의 동향을 파악하고 분석할 목적으로 주택의 매매 및 전세가격의 변동 상황을 조사하여 작성하는 통계이다. 1986년 1월부터 시작한 이 조사는 2004년 8월 현재 아파트는 1개 특별시, 6개 광역시, 53개 市, 4개 郡, 90개 區에서 4,340개 평형을, 단독과 연립은 1개 특별시, 6개 광역시, 41개 市, 3개 郡, 90개 區에서 2,955 개의 표본주택을 대상으로 하고 있으며, 조사기준일은 매월 15일이 포함된 주의 월요일, 조사주기는 월 1회이다.

조사방법은 표본주택과 인접한 해당지역 부동산중개업소를 대상으로 직접 온라인상 조사표에 입력하는 자기기입식 조사를 기본으로 하고, 온라인 조사가 불가능한 부동산에 한하여 조사원이 전화 또는 팩스로 조사하는 방법을 이용하고 있다. 주요 조사항목으로는 주택매매가격, 주택전세가격을 들 수 있다.

조사 결과 작성되는 통계는 전국의 주택유형별(단독, 연립, 아파트), 규모별(대, 중, 소) 매매가격지수와 전세가격지수이다. 또한 세분화된 지역별, 주택유형별 매매가격지수와 전세가격지수도 발표되고 있

다. 다음의 <표 1>은 2004년 9월에 발표된 주택매매가격 종합지수 통계의 일부이다.

2.2 데이터편집 개요

국민은행의 데이터편집은 크게 데이터편집 규칙을 마련하는 데서부터 시작하여 다음으로 입력 데이터편집(input data editing)과 출력 데이터편집(output data editing)으로 구분하여 진행된다.

데이터편집 규칙이란 조사된 레코드가 오류인지의 여부를 판단하기 위한 규칙을 이르는데 논리적 편집규칙(logical edit)과 계량적 편집규칙(quantitative arithmetic edit)으로 구분된다. 주택가격조사를 위해 수집되는 데이터는 양적 데이터이므로 계량적 편집규칙을 마련해야 한다.

입력 데이터편집은 달리 마이크로 편집(micro-editing)이라고도 하는데 개별의 응답값을 조사하여 입력하는 단계에서 행하는 데이터편집

<표 0> 주택가격동향조사통계의 발표 양식

주택매매가격지수 ('03.09=100.0)

구 분		금 월 (2004.9)	전 월 (2004.8)	전년말 (2003.12)	전년동월 (2003.9)	증 감 률(%)		
						전월비	전년말비	전년동월비
종 합(전 국)		98.8	99.0	99.8	100.0	-0.2	-1.0	-1.2
지역별	서 울	100.0	100.3	100.1	100.0	-0.3	-0.1	0.0
	강 북	99.8	100.1	100.0	100.0	-0.3	-0.2	-0.2
	강 남	100.2	100.6	100.3	100.0	-0.4	0.0	0.2
	6개 광역시	98.1	98.2	99.2	100.0	-0.1	-1.1	-1.9
	수 도 권	98.7	99.1	100.2	100.0	-0.4	-1.4	-1.3
유형별	아 파 트	101.2	101.4	100.9	100.0	-0.1	0.3	1.2
	단 독	96.1	96.3	98.5	100.0	-0.2	-2.5	-3.9
	연 립	94.4	94.9	97.8	100.0	-0.5	-3.5	-5.6
규모별	대	99.9	100.2	100.5	100.0	-0.3	-0.6	-0.1
	중	99.4	99.6	100.1	100.0	-0.2	-0.7	-0.6
	소	98.1	98.4	99.2	100.0	-0.2	-1.1	-1.9

을 의미한다. 한편 출력 데이터편집은 매크로 편집(macro-editing)이라고도 하며, 만일의 경우 있을지 모를 심각한 오류를 미연에 방지하기 위해 간단한 기술통계량을 구해서 문제가 되는 레코드를 찾아내는 것이다. 주택가격동향조사를 위해서는 입력 데이터편집과 출력 데이터편집을 동시에 수행한다.

Ⅲ. 데이터편집 규칙

주택가격지수 생산을 위한 조사항목은 주택의 전세가격과 매매가격이며, 동일한 표본주택에 대해 매월 주기적으로 가격을 조사하게 된다. 주기적 조사에서의 계량적 편집규칙으로 대표적인 것은 Hidiroglou와 Berthelot(1986)의 편집규칙을 들 수 있다. Hidiroglou와 Berthelot의 편집규칙의 핵심적인 개념은 i번째 주택의 t시점의 조사가격과 $(t+1)$시점의 조사가격의 비(比)인 $r_i = x_i(t+1)/x_i(t)$가 일정 범위 내에 드는지를 파악한다는 데 있다. r_i가 $[\bar{r} - ks_r, \bar{r} + ks_r]$에 포함되지 않으면 편집규칙을 위배하는 것으로 간주하게 되는 것이다. 여기서 $\bar{r}$와 s_r은 각각 r_i들의 평균과 표본표준편차를 뜻한다. 여기서 k의 결정을 위한 객관적인 절차가 뚜렷이 없는 관계로 주관성의 논란이 야기되곤 하는데 일반적으로는 경험에 의해 결정된다.

주택가격조사에서는 편집규칙에 위배되는 데이터를 특이값이라는 용어로 표현하기로 한다. 우리나라의 주택가격은 계절성을 띤다. 즉, 봄과 가을의 이사철에는 가격변동의 폭이 큰 반면에 여름이나 겨울에는 특별한 변동이 없다. 따라서 이러한 계절성을 감안하지 않은 채 획일적으로 특이값을 검출하기 위한 한계를 정하는 것은 바람직하지 않다. 지역별, 시기별로 한계를 변화시켜 가는 것이 필요하며, 여기에는 과거의 시계열자료와 관리자의 경험, 국민은행의 주간주택가격동향조사 등을

참조하는 것이 바람직하다.

특이값의 한계를 너무 낮게 할 경우 편집규칙에 위배되는 레코드가 너무 많아져서 업무량이 증대하게 되며 이로 인해 조사비용도 늘어나게 된다. 반대로 특이값의 한계를 너무 높게 할 경우 특이값임에도 불구하고 편집규칙에 위배되지 않는 경우가 과다하게 발생하게 된다. 따라서 특이값의 한계를 적절하게 정하기 위해 다양한 검토를 하였다.

먼저 2003년 10월과 11월의 조사 데이터에 대해 서로 다른 한계를 적용시킬 경우 전체 표본 데이터 중 몇 퍼센트가 편집규칙에 위배되는지를 조사하였는데 그 결과가 <표 2>에 나와 있다. 10월은 주택시장에서 성수기라고 할 수 있고 11월은 10월에 비해 약간 거래가 한산해지는데 이런 사정이 <표 2>에 반영되어 나타나 있다. 성수기인 10월의 경우 전월의 조사가격과 현재의 조사가격의 변동의 폭이 11월에 비해 상대적으로 큰 편임을 알 수 있다. 아파트의 경우 그 정도가 심한 편인 데 반해 단독/연립주택은 그 정도가 미미한 편이다. <표 2>를 기초로 하여 전체 표본 중에서 편집규칙에 위배되어 재조사를 하는 표본의 수가 대략 5% 내외가 될 수 있도록 하는 한계를 정하기로 결정하였다.

<표 2> 검정기준에 따라 체크되는 유형별 특이값 개수

특이값 한계	2003년 10월		2004년 11월	
	아파트	단독/연립	아파트	단독/연립
15%	3.4%	2.1%	1.3%	1.8%
14%	4.2%	2.6%	1.8%	2.2%
13%	5.0%	2.8%	2.2%	2.5%
12%	6.4%	3.4%	2.7%	3.5%
11%	7.7%	4.7%	3.8%	5.0%
10%	10.4%	6.4%	5.4%	6.7%
9%	12.7%	7.5%	7.3%	8.2%
8%	15.5%	9.2%	9.8%	9.4%

앞에서 언급했던 것처럼 주택시장은 계절에 따라 사정이 달라지므로 매월 특이값 검출을 위한 한계를 획일적으로 정해 두는 것은 바람직하지 않다. 월별 주택가격의 동향을 보아 가며 유연하게 기준을 변경시켜 가는 것이 필요하다. 다음의 〈표 3〉에는 2003년 10월부터 2004년 8월까지 매월 적용한 특이값의 한계에 대한 기준과 해당 월에 특이값으로 판정되어 검출된 표본의 수를 나타낸 표이다. 이 표를 보면 특이값 판정을 위한 한계가 수시로 바뀌었음을 알 수 있다. 특히 2003년 10월에서 2004년 1월 사이 특이치 검증 기준에 상대적으로 많은 변화가 있었는데 그 까닭은 2003년 10월이 새로운 표본설계 이후 처음으로 조사가 이루어진 시점이기 때문이다. 처음에는 검증에서 검출되는 표본의 규모가 얼마나 될지 몰라 15%를 기준으로 잡았다가 현장의 조사상황을 검토해 가며 기준을 변동시켰기 때문이다. 매월 조사에서 특이값으로 검출되는 표본의 수는 전 주택유형을 합쳐서 200개에서 400개 내외가 되는데 이것은 전체 표본수에 비해 2.7%에서 5.5% 정도를 차지한다.

〈표 3〉 월별 특이값의 한계 및 검출된 표본수

주택유형	아파트		단독 / 연립	
조사시점	검증기준(%)	특이치	검증기준(%)	특이치
2003년 10월	15%	144	15%	60
11월	12%	117	12%	104
12월	12%	165	12%	156
2004년 1월	10%	189	10%	200
2월	10%	276	11%	131
3월	10%	151	10%	130
4월	9%	206	9%	113
5월	9%	172	8%	116
6월	9%	172	8%	208
7월	9%	164	9%	128
8월	9%	267	9%	165

Ⅳ. 입·출력 데이터편집

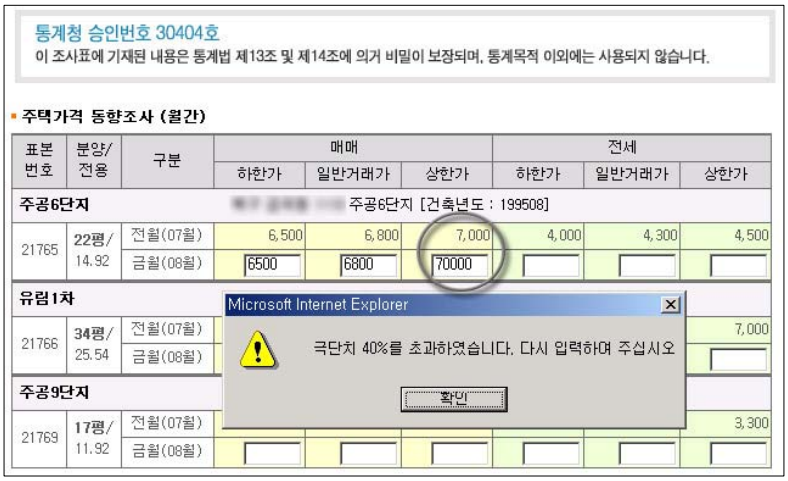
4.1 입력 데이터편집

입력 데이터편집은 크게 두 단계로 나누어 진행된다. 하나는 조사단계에서 응답 실수를 미연에 방지할 수 있게 예방하는 단계이고, 다른 하나는 수집된 응답 데이터에 대해 편집규칙을 적용하여 검토하는 단계이다.

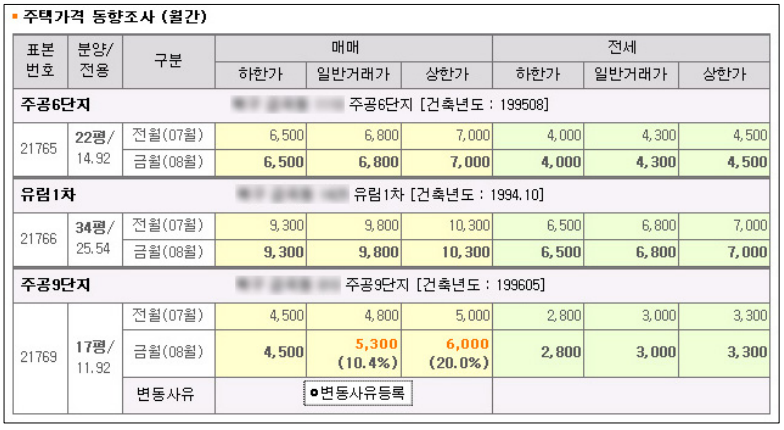
전국주택가격동향조사는 부동산중개업자들을 대상으로 하여 온라인 자기기입식으로 조사가 이루어지므로 응답자가 응답값을 입력하는 과정에서 실수를 할 개연성이 크다. 따라서 응답을 입력하는 단계에서의 실수를 원천적으로 봉쇄할 수 있도록 하기 위해 별도의 입력시스템을 개발하였다. 이 입력시스템을 이용하면, 응답자 입장에서는 응답 실수를 미연에 방지할 수 있게 되고, 조사관리자 입장에서는 별도로 코딩이나 자료입력을 할 필요가 없으므로 자료수집 단계에서 효율을 높일 수 있다.

〈그림 1〉은 주택가격조사를 위해 개발한 입력시스템의 자료입력 화면의 예이다. 응답자가 접속하면 각 응답자가 응답해야 할 표본주택과 그 주택에 대한 지난 달 응답가격이 화면에 나타나게 된다. 응답자는 빈 칸으로 되어 있는 금월의 칸에 각각 응답값을 입력하면 된다. 입력과정에서 응답자가 실수를 하여 7,000으로 응답할 것을 70,000으로 기입하였다고 하자. 그러면 입력시스템에서는 이러한 사항을 바로 찾아내어 〈그림 1〉에 나온 것처럼 입력오류가 발생하였음을 알리는 메시지를 자동으로 보내어 재입력하게 한다. 이럼으로써 입력오류를 미연에 예방할 수 있다. 시스템에서 자동으로 입력을 거부하게 되는 기준은 응답가격의 변동폭이 40%를 초과할 때이다. 물론 이 변동폭은 수시로 변경시킬 수 있는데 현재로서는 40%를 사용하고 있다.

다음으로는 응답한 가격의 전월(前月) 대비 변동폭이 40%를 초과하지는 않지만 그래도 변동폭이 특이값의 한계를 초과하여 나타나는 경우가 생기는지를 점검하기 위해 입력시스템은 자동으로 편집규칙을 가



〈그림 1〉 응답자 오류 방지 시스템

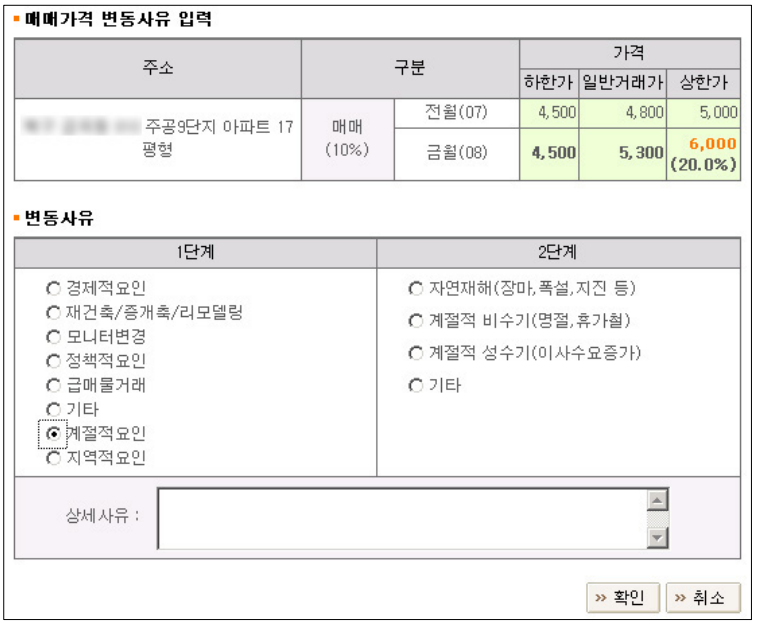


〈그림 2〉 특이값이 발생했을 때의 입력화면의 예

지고 응답값을 점검한다. 〈그림 2〉는 특이값으로 판명되는 경우에 나타나는 화면의 예이다. 한 응답자가 응답할 표본주택이 세 개가 있는데 그 중 세 번째 주택에 대해 응답가격이 전월에 비해 큰 폭으로 오른 상황이다. 이 경우 응답자가 입력을 마치면 시스템은 바로 〈그림 2〉의 화

면을 띄워 준다. 즉, 세 번째 주택의 응답가격의 전월 대비 변동률을 계산하여 보여주는데 실제 거래가격이 이렇게 많이 올랐다면 오르게 된 사유를 추가로 기입하여야만 조사가 완료될 수 있게 한 것이다. 정당한 사유를 입력하지 않을 경우 입력이 취소되고 다시 처음의 화면으로 돌아가게 된다. 이 때 사용되는 특이값의 한계값은 바로 앞의 <표 3>에 소개한 값이다.

다음의 <그림 3>은 변동사유를 기록하는 화면을 보여주고 있다. 특정 표본의 가격 변동폭이 특이값 검출 한계를 벗어났을 때 그 이유가 무엇인지를 기입하게 하는 화면인데 먼저 가장 일반적으로 나타나는 변동사유 항목들을 보기로 열거해 주고 보기에 나온 사유 외의 사유가 있을 때에는 직접 그 사유를 입력하도록 하였다. 대부분의 경우 보기에 열거된 항목들에 의해 가격변동이 일어난다.



<그림 3> 변동사유 등록화면

입력이 끝나면 입력된 정보는 자동으로 데이터베이스에 저장되는데 저장된 데이터에 대해 다시 한번 추가적인 데이터점검이 이루어진다. 조사관리자는 입력 시스템에 의해 특이값으로 검출된 데이터에 대해 등록된 변동사유를 살펴보고 미심쩍다고 생각될 때에는 해당 응답자에게 추가로 연락하여 확인하는 작업을 한다. 확인 작업을 거쳤음에도 불구하고 여전히 적절한 사유가 아니라고 판단되면 해당 응답을 무응답으로 처리하기도 하고 경우에 따라서는 별도로 조사한 가격정보로 대체하기도 한다. 대체를 하게 되는 경우 부적절한 정보를 제공한 부동산 중개업자를 다른 중개업자로 교체하는 작업도 병행한다.

4.2 출력 데이터편집

입력 데이터편집을 통해서 대부분의 입력오류를 미연에 방지할 뿐 아니라 전월 대비 가격의 증감이 큰 데이터를 검출하여 확인할 수 있다. 그러나 주택가격동향통계 생산을 위해서는 추가적으로 출력 데이터편집도 동시에 수행함으로써 통계의 질을 제고시키려 하고 있다. 출력 데이터편집은 입력 데이터편집 과정을 거쳐서 얻어진 데이터를 가지고 일차적으로 통계량을 계산한 후 얻어진 통계값 중에서 문제가 있다고 판단되면 그 통계값에 큰 영향을 미친 데이터를 역으로 찾아내어 확인하는 데이터편집 방법이다.

주택가격동향조사를 통해 생산되는 주택가격지수의 경우 전국의 각 시, 군, 구 단위의 통계까지 생산된다. 어느 지역의 지수가 다른 지역들의 지수에 비해 지나치게 두드러진 차이를 보이는 경우 해당 지역 데이터들을 점검하게 된다. 해당 지역에서 특이값으로 등록된 데이터에 대해 먼저 변동사유를 확인한 후 변동사유가 명확하다고 판단되면 다른 부동산싸이트들에 나온 가격정보들과 다시 비교한다. 이러한 비교를 통해 합리적인 가격이라고 판단되면 수용한다. 만일 변동사유가 명확하지

않거나, 타 부동산사이트와의 가격차가 현저할 경우에는 조사원이 재조사를 실시한다. 재조사 결과 이전의 응답정보가 부적절하다고 판단되면 재조사 가격을 사용하며 부적절한 정보를 제공한 응답자를 다른 응답자로 대체한다.

출력 데이터편집에 의해 문제로 드러나는 예들을 살펴보면 대부분 해당 지역의 전반적인 주택시장의 상황에 따라 가격의 변동이 일어난 것이 아니라 특정 표본의 개별적인 사정에 의해 변동이 일어난 것임을 알 수 있었다. 표본주택의 수리, 재건축 등으로 인해 갑자기 가격이 오르게 되는 것이다. 2004년 4월 주택가격동향조사 후에 출력 데이터편집을 하면서 확인한 사례를 보면, 전국 아파트 매매의 지수는 101.9 였는데 경상남도 진주시의 아파트매매 지수는 105.8 이 나와 이상이 있는 지역으로 여겨졌다. 입력된 원시 데이터를 찾아보니 경남 진주시의 표본 중에서 한 개의 아파트 단지가 큰 영향을 미치는 것으로 드러났다. 해당 아파트 단지의 재건축 승인을 얻기 위해 추진위원회가 결성되고 주민에 동의서를 받고 있어 매물도 부족하고 한달 사이에 집값이 너무 많이 오르게 된 경우였다.

출력 데이터편집 과정에서 검출된 데이터에 대해서는 각 상황별로 적절한 조치를 취해야 한다. 대표적인 조치로는 사후층화(post－stratification)를 통한 가중값 조정(weight adjustment)을 들 수 있다. 특정 주택의 가격에 큰 변동이 생기면 먼저 그러한 변동이 그 지역의 일반적인 경향인지 아니면 특정 표본에 국한되어 생긴 이례적인 상황인지를 파악한다. 일반적인 경향이라면 조사결과를 그대로 반영하여 통계를 생산하면 된다. 그렇지 않고 특정 표본에 국한되어 나타난 이례적인 상황이라면 사후층화를 하여 해당표본의 가중값을 축소시켜 반영한다.

2004년 6월 조사에서 서울시 종로구의 단독표본에 대한 사후층화를 한 사례를 소개하기로 한다. 종로구 숭인동이 통합개발계획 구역에 속해 있는데 이 중에서 1단계 개발지역으로 발표된 곳의 주택가격이 크게

올랐고 실제 오른 가격으로 거래들이 이루어졌다. 당시 전국의 종합 매매지수는 0.2% 하락했는 데 반해 종로구의 경우는 2.9% 상승한 것으로 나타났다. 종로구 전체의 주택가격이 오른 것이 아니라 숭인동의 특정지역만 상승했는데 상승지역에 속한 몇몇 표본주택의 영향으로 종로구 전체 지수가 높게 나타난 것이다. 이 때 재개발 지역에 해당되는 주택들의 규모를 파악한 후 종로구의 주택들을 재개발 지역과 그렇지 않은 지역으로 사후층화를 하였다. 사후층화를 하여 가중값을 재조정한 이후 구한 종로구의 매매지수는 전월 대비 0.7% 상승한 것으로 나타났다.

Ⅴ. 맺음말

조사를 통해 수집되는 원시 데이터 중에 오류가 있는 데이터가 포함될 경우 그것을 통해 얻어지는 통계의 품질에 문제가 생길 수 있다. 많은 비용과 노력을 들여 통계를 생산하게 되는데 오류가 있는 데이터를 사용함으로 인해 작성되는 통계의 신뢰성이 떨어지게 된다면 이것은 적지 않은 손실이다. 따라서 통계를 생산하는 모든 기관들은 데이터의 질을 점검하고 수정하는 데이터편집의 중요성과 필요성에 공감하고 있으며 또 나름의 방법으로 데이터편집을 실시하고 있다. 외국의 선진 통계기관들에서는 데이터편집을 위한 보다 합리적이고 효율적인 방법을 모색하기 위해 각자 경험한 사례들을 공론화하여 활발한 논의를 벌이고 있다.

각각의 통계조사는 문화와 조사목적 등에 따라 나름의 독특성을 지닌다. 따라서 보다 다양한 조사에 대한 데이터편집 사례들을 공유함으로 데이터편집에 대한 이해의 폭을 넓힐 수 있다. 그러나 우리나라의 경우 데이터편집에 관한 논의가 매우 빈약한 실정이며 더군다나 구체적

인 사례에 관한 연구는 거의 이루어지지 않았으므로 우리나라 상황에 맞는 데이터편집을 생각한다는 것은 어려운 현실이다.

본 논문에서는 국민은행의 주택가격동향조사를 위한 데이터편집의 사례를 소개하였다. 구체적으로 조사목적에 맞도록 편집규칙을 정하는 과정 및 관련 자료들을 소개하였고, 온라인조사라는 조사방식에 맞는 입력 데이터편집 방법을 마련하여 실시하는 예들을 소개하였다. 마지막으로 출력 데이터편집에 의해 입력 편집에서 걸러지지 않은 오류나 문제들을 제거하는 방법도 소개하였다.

본 연구에서는 주택가격조사의 데이터편집 사례를 자세히 소개하였는데 일반적인 계속조사의 과정을 망라한다는 점에서 다른 조사들의 데이터편집을 위한 모범적인 사례가 될 것이다. 사업체조사 등과 같이 계속조사이면서 조사변수가 양적인 변수인 경우 본 사례의 방법론이 대동소이하게 적용될 수 있을 것이기 때문이다. 관심변수가 양적변수인 일반적인 통계조사에서 쓰이는 데이터편집의 흐름을 망라하고 있으므로 좋은 참고자료가 될 수 있을 것으로 생각한다. 특히이 연구결과가 추후 데이터편집에 대한 연구와 논의를 활발하게 하는 데 일정 부분 기여를 할 수 있을 것으로 기대한다.

참고문헌

류제복 · 김영원 · 박진우 · 이재원. 2003. "Imputation Methods for the Population and Housing Census 2000 in Korea." 〈〈한국통계학회논문집〉〉 10(2): 575−583.

박성현 · 박진우. 2004. “표본조사 통계품질관리 가이드라인 연구.” 〈〈응용통계연구〉〉17(3): 557−571.

이재원. 2000. “무응답 및 오류 자료의 Imputation 적용 결과.” 〈〈무응답오차〉〉 제8장, 자유아카데미.

Fellegi, I. P. and Holt, D. 1976. “A Systematic Approach to Automatic

Edit and Imputation." *Journal of the American Statistical Association* 71:436−454.

Granquist, L. 1995. "Improving the Traditional Editing Process." *In Business Survey Methods* (eds. Cox et al.). John Wiley & Sons: 177−199.

Granquist, L. and Kovar, J. 1997. "Editing of Survey Data: How Much Is Enough?" *In Survey Measurement and Process Quality* (eds. Lyberg et al.). John Wiley & Sons:415−435.

Waal, T. and Quere, R. 2003. "A Fast and Simple Algorithm for Automatic Editing of Mixed Data." *Journal of Official Statistics* 19: 383−402.

Hidiroglou, M. A. and Berthelot, J. M. 1986. "Statistical Editing and Imputation for Periodic Business Surveys." *Survey Methodology* 12: 73−84.