부동산학연구 제19집 제4호, 2013. 12, pp. 153～164

Journal of the Korea Real Estate Analysts Association
Vol.19, No.4, 2013. 12, pp. 153～164

매월 조사되는 주택 가격 변동률의 이상치 탐색 방법에 관한 연구

Outlier Detection Methods for Monthly Rate of Housing Price

육 태 미 (Youk, Taemi)*
방 송 희 (Bang, Songhee)**
이 재 성 (Lee, Jaesung)***

< Abstract >

In statistical theory, an outlier is a value that is numerically distant from overall pattern of a distribution. It may be a meaningful observation, but it comes from an error in survey, data entry and process in most cases. Detection and handling are needed because outliers by errors debase the statistical quality leading to biased parameter estimation.

Generally, traditional Box-plot or Z-score are very useful for univariate outlier detection and a Median rule could be applied in the non-Gaussian case. These methods calculate the tolerance interval that defines the range of acceptable observation values. Outlier detection for periodic surveys would consider the past view, because it is based on a ratio of value comparing the current time with previous time. If time period, however, is short, a state to get many unchanged values can occur. In this case, the ratio is centered at 1, and therefore outlier detection method reflecting this factor is required.

This paper considers Quartile Method with power transformation and Hidiroglou-Berthelot(1986) method that is efficient in periodic data. The methods were applied to housing sales price. We suggest an outlier detection method for real-world data. In addition, we also analyzed data using Tukey Algorithm of United Kingdom's office of National Statistic(ONS).

주 제 어 : 가격비율, 자료편집, 사분위수 방법, H-B 방법, Tukey 알고리즘
key word : Price Ratio, Data Editing, Quartile Method, Hidiriglou-Berthelot Method, Tukey Algorithm

* 고려대학교 통계학과 박사과정, xoal84@korea.ac.kr (주저자)
** 한성대학교 부동산경영학과 초빙교수, 21172@naver.com, (교신저자)
*** 고려대학교 통계학과 석사, tianod@korea.ac.kr

I. 서론

어떠한 목적에 의해 구축된 데이터는 일정한 패턴을 가지는 것이 일반적이다. 이상치(outlier)는 구축된 자료에서 일반적인 행태를 보이는 나머지 값들과 수치적으로 분리되는 값을 말한다. 이러한 이상치는 실제로 어떤 현상의 큰 변화를 대변해 주는 의미 있는 자료일 수도 있지만 대부분의 경우에는 조사, 자료입력, 처리 과정의 오류에서 나타난다. 오류에 의한 이상치는 추정의 편향을 야기하게 되어 대규모 통계조사의 질을 떨어트리며 특히 주기적인 조사의 경우에 그 영향력은 더 높다.

데이터 수집 및 처리 단계에서 발생하는 오류를 찾아내고 이를 수정하는 작업을 데이터 편집(data editing)이라고 하며, 이상치의 탐색은 이러한 데이터 편집 과정 중 하나이다(Granquist, 1995). 이상치 탐색에 관한 연구는 자료 사용 목적에 따라 다양하게 시도되고 있다. 일반적으로 일변량의 경우 자료가 정규성을 만족하면 상자그림(box-plot)(Tukey, 1977)이나 Z-score를 이용하여 중심으로부터 거리가 먼 자료를 이상치로 판단한다. Barnerjee and Iglewicz(2007)은 대표본에서 여러 모수적 분포 하의 이상치 경계를 제시하였으며, Carling(2000)은 오른쪽으로 꼬리가 긴(right-skewed) 분포에서 중위수 규칙(median rule)을 제안하였다. 그리고 Brys et al.(2003)은 편향된 분포에서 강건한 Medcouple (MC)을 통해 이상치의 경계를 고려하였고, Hubert and Vandervieren(2008)은 이를 보완하여 편향된 방향에 따른 MC의 적절한 상수 적용에 대해 제시하였다. 또한 자료들 간의 거리(distance)를 계산하거나 군집을 통해 다른 자료들과 구별되는 값을 분석하기도 하는데, 다변량 자료에서는 자료가 얼마나 깊고(deep) 중심화(central)되어 있는 지를 측정하는 깊이 함수(depth function)를 이용하여 이상치를 구별한다(Liu, 1990; Liu and Singh, 1993; Serfling, 2002).

주기적인 조사는 관심대상이 시간의 흐름에 따라 어떻게 변하는지를 관찰하여 이후의 행태에 대한 예측을 가능하게 해준다. 가격 조사의 경우에 조사주기의 단위가 짧다면 전시점 대비 현시점의 변동이 크지 않고 그 비율 분포가 현시점에 재화 가격의 상승 또는 하락에 따라 꼬리의 방향이 결정되는 편향된 분포를 갖게 된다. 특히 주택 가격은 월 또는 주간 단위로 조사될 때 그 성격상 가격변동이 자료의 극히 일부에서만 나타나게 되어, 비율 분포가 극단적으로 1에 중심되어 있는 형상을 띄게 된다. 이러한 특징을 가지고 있는 주기적인 주택 가격 조사의 이상치 탐색에 관해 현재까지 진행되어진 연구는 거의 전무하다.

본 연구는 매월단위로 조사되는 주택가격 자료에서 이상치를 탐색 하는 방법들에 대한 논의를 목적으로 한다. 전시점 가격 자료에 대한 현시점의 가격 비율을 이용하여 다양한 이상치 탐색 기법을 적용해 보고 그 타당성을 살펴봄으로써 합리적으로 시도할 수 있는 이상치 탐색 방법을 제시하고자 하였다.

2장에서 여러 가지 이상치 탐색 방법론에 관련된 선행연구를 살펴보고, 3장에서는 실증 자료에 활용에 대한 논의를 하였다. 그리고 4장에서는 전국주택가격동향조사의 자료를 통해 방법론을 적용시키고 5장에서는 결론 및 한계점을 논의한다.

II. 선행연구의 검토

비모수적 이상치 탐색의 방법론들은 일반적으로 허용구간(tolerance interval)을 산출하고 그 구간을 벗어나는 관측값들을 이상치로 판단한다. 현재 국내·외 통계 기관들에서는 일변량 자료의 이상치 탐색 방법으로 사분위수 방법(Quartile Method), H-B(Hidiroglou-Berthelot) 방법, Tukey 알고리즘 등의 다양한 기술을 사용하고 있다. 이러한 방법들은 자료의 흩어짐의 정도를 고려하여 중심의 위치에서 상대적으로 먼 관측값을 이상치로 판별하게 되고 과정의 일정 부분에서 유사하기 때문에 퍼포먼스의 차이는 크게 구별되지 않는다.

1. 사분위수 방법(Quartile Method, QM)

사분위수(q_1: 제 1사분위수, q_2: 중위수, q_3: 제 3사분위수)를 이용하여 허용구간을 계산하는 이상치 탐색법은 가장 흔히 사용하는 방법이다. 관측값 x_i에 대하여 분위수를 이용한 범위의 하한과 상한을 각각 L_r, U_r이라고 하면, 허용구간을 다음과 같이 정의할 수 있다.

$$(q_2 - c_L L_x, q_2 + c_U U_x).$$

여기서 c_L과 c_U는 정해진 상수로, 만약 자료의 분포가 대칭이라면 동일한 값으로 설정 할 수 있다. 범위의 하한과 상한은 자료가 집중되어 범위를 좁게 잡는 것을 방지하기 위해서 최소 허용구간(minimum tolerance interval)으로 표현하며, 그 식은 다음과 같다.

$$L_x = \max(q_2 - q_1, |aq_2|)$$
$$U_x = \max(q_3 - q_2, |aq_2|).$$

여기서 a는 0과 1사이의 값으로, Lee et al. (1992)은 대부분의 응용에서 0.05가 적당함을 밝혔다.

2. Hidiroglou-Berthelot(H-B) 방법

Hidiroglou and Berthelot(1986)는 특히 주기적인 조사의 자료에 적용할 수 있는 방법을 제안하였다. i번째 관측값의 t시점에 대한 $t-1$시점의 값의 비율을 $r_i = p_i^t / p_i^{t-1}$라고 하자. Hidiroglou-Berthelot(H-B) 방법은 먼저

$$s_i = \begin{cases} 1 - (q_2/r_i) & , 0 < r_i < q_2 \\ (r_i/q_2) - 1 & , r_i \geq q_2 \end{cases}$$

로 정의되는 s_i를 사용하여 자료를 0을 중심으로 하는 대칭 변환한다. 여기서 q_2는 r_i $(i = 1,...,n)$의 중위수이다. 만약, $q_2 = 1$이고 모든 i에 대해서 $r_i \approx 1$이면 위의 변환은 로그변환(logarithmic transformation, $p = 0$)과 유사해진다. 그리고 변환된 자료 s_i를 이용하여 자료의 효과를 반영한 e_i를 계산한다. 효과 e_i는 자료의 크기를 추가하는 것으로

$$e_i = s_i \left\{ Max\left(p_i^t, p_i^{t-1}\right) \right\}^U$$

이고, U는 채택 범위의 형태를 결정한다. U가 0이면 자료의 크기를 모두 무시하게 되어 효과가 크거나 작거나 동일하게 취급하게 되고, 1

이면 이상치 모집단이 큰 값들로 구성될 것이다. 효과 e_i를 통한 최소 허용 구간이 이상치 탐색의 기준이 된다.

H-B 방법은 이상치를 판별하는데 있어서 각 관측값의 효과를 반영해 준다는 장점이 있고 그 반영의 정도를 연구자가 조정가능 하게 한다. 하지만 단계별 적절한 모수의 오설정은 이상치 탐색의 실패를 야기할 수 있다.

3. Tukey Algorithm(TA)

Tukey 알고리즘은 United Kingdom's Office of National Statistic(ONS)에서 1987년 이후로 CPI지수를 작성할 때 이용한 방법이다. 본 연구에서는 ONS CPI technical manual(2006) 버전의 알고리즘을 고려하였고 그 방법은 아래와 같은 단계로 진행된다.

1. 가격비율 r_i를 오름차순으로 정렬하고 값이 1인 관측값을 제거
2. 비율 분포의 상·하위에서 각 5%의 값을 제거
3. 단계 1,2 에 의해서 제거되고 남은 가격비율 자료를 d라고 정의함. 그리고 $\overline{r_d}$, $\overline{r_L}$, $\overline{r_U}$를 각각 d의 산술평균, d의 중위수 아래쪽 자료의 산술평균, d의 중위수 위쪽 자료의 산술평균이라 하고 계산 함
4. 허용구간을 아래와 같이 계산 함
$\left(\overline{r_d} - c_L\left(\overline{r_d} - \overline{r_L}\right), \overline{r_d} + c_U\left(\overline{r_U} - \overline{r_d}\right)\right)$

TA는 가격비율이 1인 값을 제거하고 허용구간을 구하기 때문에 자료가 안정적으로 사용이 되고 허용구간이 평균에 의해서 많이 움직이지 않는다. 그러나 자료의 일부만을 사용하였기에 결과의 정확성을 보장하기는 힘들며 특히 표본수가 작은 경우에는 적절하지 못한 방법이 될 것이다.

III. 이상치 탐색 방법

주기적인 조사의 경우에는 전시점 대비 현시점 자료의 변동이 다른 값들에 비해서 상대적으로 크다면 이를 이상치라고 의심해 볼 수 있다. 그런데 조사주기가 짧은 경우 가격 비율은 일반적인 경우와 다르게 편향되거나 한 값에 집중된 분포를 보이기 때문에 자료의 특성을 반영한 이상치 탐색 방법이 필수적이다.

1. 자료 변환

사분위수 방법은 월별로 변화하는 자료의 이상치 탐색 방법으로 쓰이기에는 무리가 있다. 가격 비율은 편향되어 있을 가능성이 높고, 사분위수 방법은 분포의 꼬리 방향에 따라 민감하게 되어 가면화 효과(masking effect)가 일어나게 된다. 이에 쉽게 쓰일 수 있는 대안은 멱변환(power transformation)을 이용하여 관측값을 대칭분포에 가깝도록 바꾼 후 범위를 설정하는 것이다. 멱변환은 p에 대한 함수 꼴로

$$T_p(x) = \begin{cases} x^p & , p > 0 \\ \log(x) & , p = 0 \\ -x^p & , p < 0 \end{cases}$$

이다. 적절한 p의 추정을 위해 탐색적 자료 분석(Exploratory Data Analysis, EDA)방법인 Hoaglin et al.(1983)의 변환 그림 등을 사용할 수

있다(Thompson and Sigman, 1999). 사분위수 방법은 간단하지만 변환을 통하여 적용 가능한 분포로 변경하는 데에 어려움이 존재한다.

2. 실증 자료의 활용

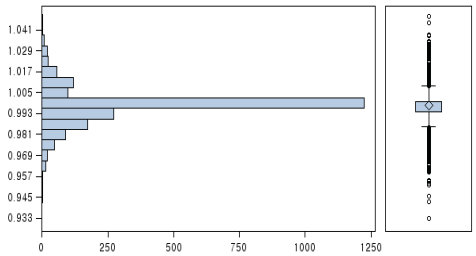
전시점 가격 자료에 대한 현시점의 가격 비율을 이용하여 이상치를 탐색하는 본 연구에서는 짧은 주기의 비율에서 나타날 수 있는 분포가 주요 관심사이다. 월별로 조사되는 표본 주택 가격의 전월 대비 당월의 비율은 대부분 1의 값을 갖게 되어 적어도 두 개의 사분위수가 동일할 가능성이 높아진다. 이러한 경우에 H-B방법을 통한 이상치 탐색의 과정에서 사용되는 효과 e_i의 허용구간은 정해진 상수값에만 의존하고 분포를 적절히 활용하지 못하게 된다.

이상치의 정의를 전월 대비 현월 주택가격의 변동이 큰 관측값이라고 본다면, 가격 변동이 없는 주택을 이상치 모집단에서 제외하고 가격 변동이 있는 주택에서 이상치를 탐색하는 것을 고려할 수 있다. 이 방법은 주택이 안정되는 시기에는 효율적으로 사용될 수 있지만 가격이 전반적으로 급변하는 시기에는 오히려 분포의 왜곡을 가져올 수 있음을 유의하여야 한다.

〈표 1〉 서울시 아파트 표본 기초 통계량

	매매가격 (단위: 천원)	전월 대비 당월 가격비율
평균	479,787	0.9979
최솟값	119,000	0.9330
1사분위수	290,000	0.9941
중위수	395,000	1.0000
3사분위수	550,000	1.0000
최댓값	3,050,000	1.0492

〈그림 1〉 가격 비율의 분포



〈표 2〉 가격 변화가 없는 표본주택의 비율

조사월	%	조사월	%
'12.2	75.9	'13.1	51.1
'12.3	65.8	'13.2	58.8
'12.4	63.9	'13.3	56.6
'12.5	65.4	'13.4	43.6
'12.6	54.8	'13.5	50.9
'12.7	57.6	'13.6	51.1
'12.8	45.9		
'12.9	51.7		
'12.10	51.0		
'12.11	57.4		
'12.12	57.5		

IV. 실증분석

1. 전국주택가격동향조사 및 자료 개요

전국주택가격동향조사[1]는 아파트, 연립주택, 단독주택에 대해서 선정된 표본을 조사하는 매매 및 전세가격 동향 통계이다. 월간 조사의 표본은 2013년 6월 기준 아파트 14,334호, 연립주택 2,646호, 단독주택 2,227호로 총 19,207개이며, 표본으로 선정된 주택의 매매가격과 전세가격을 인근 실거래사례를 참고하여 조사한다. 가격은

1) 당초 KB국민은행에서 조사·공표하던 전국주택가격동향조사는 제2차 국가통계위원회의 결정(2010.6)에 따라서 2013년 1월부터 한국감정원으로 작성기관이 변경되었다.

〈표 3〉 이상치 판별 비교

	표본 개수		Box-plot[2]		Z-score[3]		Median Rule[4]		QM[5]	
			이상치 개수		이상치 개수		이상치 개수		이상치 개수	
전체	2,182		159	(7.3%)	45	(2.1%)	346	(15.9%)	85	(3.9%)
변동[1] ±2.5%이상	110	(5.0%)	110	(5.0%)	45	(2.1%)	110	(5.0%)	85	(3.9%)
변동 ±3%이상	60	(2.7%)	60	(2.7%)	39	(1.8%)	60	(2.7%)	60	(2.7%)
변동 ±4%이상	12	(0.5%)	12	(0.5%)	12	(0.5%)	12	(0.5%)	12	(0.5%)
변동 ±5%이상	3	(0.1%)	3	(0.1%)	3	(0.1%)	3	(0.1%)	3	(0.1%)

[1]변동: 전월 매매가격 대비 변동률,
[2]Box plot: 바깥울타리(= $[q_1 - 3IQR, q_3 + 3IQR]$)를 벗어나는 극단 이상치,
[3]Z-score: |Z-score|〉3 인 이상치
[4]Median Rule: 허용구간= $[q_2 - 2.3IQR, q_2 + 2.3IQR]$,
[5]QM: 변환을 이용한 사분위수 방법

매월 15일이 포함된 일주일 중 월요일이 기준으로 조사되며, 주택 가격지수(housing price index)는 익월 1일에 공표된다.

분석을 위하여 본 논문에서는 전국주택가격동향조사의 "서울특별시", "아파트"의 2013년 6월 월 조사된 2,182개 표본의 매매가격을 사용하였다.

<표 1>과 <그림 1>을 보면 전월(5월) 대비 당월(6월)의 가격 비율은 대체로 1주위에 밀집되어 있는 분포임을 확인할 수 있다. 특히 1의 값을 갖는 주택이 1,114개로 약 51.1%에 해당되어 첨도가 높은 형상의 띄고 있고 주택의 가격은 월 단위로 변화가 크지 않은 것을 확인시켜준다. 또한 2012년 이후 전월 대비 당월의 가격 비율이 1의 값을 갖는 비율은 전체의 43.6~75.9% 임을 볼 수 있다<표 2>. 2012년 이후 매월 가격 비율의 분포는 유사하고, 가격변동은 비교적 안정적으로 일어나고 있는 형태이다.

2. 변환을 이용한 QM

주택 가격지수는 주기적으로 조사되는 표본의 매매가격으로 작성되는데, 이 경우 이상치는 당월 조사가격만으로 판단하기 어렵다. 따라서 일반적으로 전월 대비 당월 가격비율을 통해서 나머지 값들과 수치적으로 분리되는 값만을 이상치로 분류하게 된다. 가격비율을 위한 이상치는 다양한 성격으로 정의 될 수 있지만, 본 연구에서는 나머지 값들과 수치적으로 분리되는 값만을 고려하여 탐색 방법들 간의 이상치 판별 결과를 비교하였다.

먼저, 가격비율은 정규성을 갖는 분포를 위해 EDA 관점에서 적절한 멱변환을 선택하여 일반적으로 사용되는 Box-plot, Z-score, 중위수 규칙과 QM을 비교하였고, 결과는 <표 3>에 나타내었다.

QM은 변동률 2.5~3% 사이에서 허용구간의 경계가 결정되어 자료의 3.9%를 이상치로 판별한다. 네 가지 방법은 모두 변동률이 4% 이상을 갖는 자료는 이상치로 분류하였다. Z-score를 이용하는 방법은 자료가 정규분포를 따르지 않거나 표본수가 작은 경우에는 알맞지 않을 수 있고 극단값에 영향을 크게 받는다. 또한, 중위수 규칙은 가장 이상치 비율이 높지만 이는 IQR의 척도와 관계가 있으며 더 큰 척도를 고려한다면 이상치 비율은 낮아질 수 있다. 사분위수가 모두 동일한

자료이라면 QM은 자료의 형태가 아니라 범위 경계를 위한 상수에만 크게 영향을 받게 되기 때문에 그 사용이 부적절하다.

3. H-B 방법과 TA

H-B 방법과 TA의 적용을 위한 자료는 가격 형성에 유의한 영향을 주는 요인인 규모를 고려하였다. 규모가 85m^2 미만인 경우와 이상인 경우의 두 개의 집단으로 나누고 각각의 집단에 대해서 이상치 탐색을 시행하였다.

첫 번째로 H-B방법 적용을 위한 효과 e_i변환의 모수 U를 일반적으로 0.3과 0.5사이의 값으로 사용하는데(Richard Belcher, 2003), U가 큰 값을 선택하면 고가의 주택들만 이상치로 선택될 위험이 있어 그 영향력을 낮추기 위하여 U는 0.3으로 선택하였다. 효과 e_i의 분포는 0주위에 밀집되고 중위수와 3사분위수가 동일하게 0의 값을 갖게 되어, 상한 경계값은 0이 된다. 즉, 매매가격이 상승한 주택은 모두 이상치로 분류되는 오류가 발생하게 된다. 또한, 하한 경계값을 결정하는 상수 c_L의 선택의 주관적인 부분을 배제할 수 없기 때문에 적절한 허용 한계범위를 계산하기 어려워진다. 이때 자료를 도식화하여 시각적으로 벗어난 관측값을 확인하거나 과거자료를 토대로 한 경험적 선택으로 상수값 설정의 기준을 마련할 수 있다.

<표 4>와 <그림 2>는 H-B방법의 결과로 상한 경계값이 낮아 전체 자료의 68.9%를 이상치로 탐색하게 된다. 만약 효과가 0에 밀집되고 1사분위수와 중위수가 동일한 자료라면 하한 경계값이 높아져 값이 작은 쪽에서 많은 자료를 이상치를 판단할 것이고, 1사분위수와 3사분위수가 동일한 중앙 극집중 분포 자료라면 변동이 일어나는 모든 효과를 이상치로 판단하게 된다.

가격비율이 1인 값이 대부분인 전국주택가격 동향조사의 자료에 H-B 방법을 직접 적용하기에 무리가 있어 이를 해결하기 위해서 TA의 가격비율이 1인 값을 제거하는 단계를 응용하여 H-B방법에 적용하였다. 즉, 전월 대비 가격변동이 일어나지 않는 자료를 제거하고 가격비율의 중위수를 구하는 변형된 H-B 방법을 시행하였다. <표 5>와 <그림 3>은 변형된 H-B 방법의 결과이고 실제로 전체 자료의 군집에서 벗어나는 몇 개의 관측값이 적절하게 이상치로 분리되었음을 확인할 수 있다. 가격비율 일부를 제거하기 전과 후에 허용구간의 경계가 달라지게 되어 전체 자료의 1.3%만이 이상치로 구분이 되었다.

두 번째로 전월 대비 가격비율의 변동이 없는 자료를 제외하고 허용구간의 경계를 결정하는 TA 방법에서 사용되는 3가지 산술평균은 <표 6>와 같으며, 이 방법에 의해 85m^2 미만은 자료의 0.36%가, 85m^2 이상은 자료의 1.52%가 이상치로 판별된다. <그림 4>는 변형된 H-B 방법과 TA의 공통 이상치를 표현한 그림이다. 변형된 H-B 방법은 상한경계의 위쪽에 존재하는 관측값이 TA 방법보다 상대적으로 적음을 확인 할 수 있으나 이는 e_i변환의 모수 U의 영향으로, 모수 값이 1에 가까워질수록 큰 값에서 이상치가 증가하게 된다. 또한, TA는 하한 경계와 상한 경계를 벗어나는 관측값이 변형된 H-B 방법보다 비교적 대칭적이다.

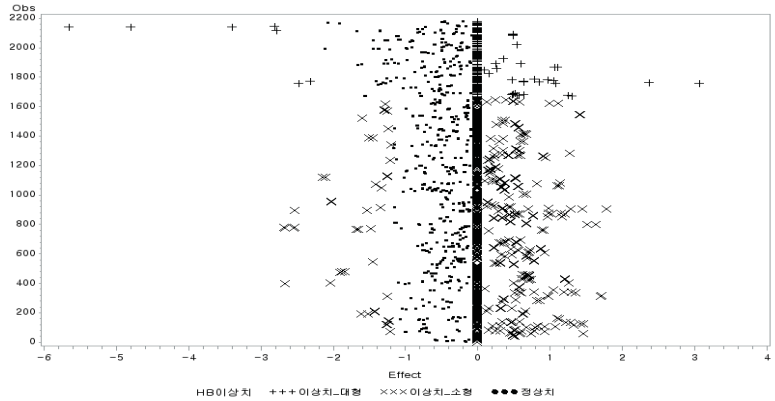
QM과 TA는 가격비율의 퍼짐만을 사용하였다면 H-B 방법은 가격비율에 실제 주택가격을 고려하여 이상치를 결정했다는 것에서 차이점이 있다. 주택 가격지수는 전체 주택시장의 흐름을 보

〈표 4〉 H-B 방법의 이상치 경계값

규모	표본수	e_i의 이상치 기준		이상치 수
		하한 경계값	상한 경계값	
85m2 미만	1,654	-1.194	0	1,186 (71.7%)
85m2 이상	528	-2.239	0	318 (60.2%)

* 85m2 미만: c_L=c_U=5/ 85m2 이상: c_L=c_U=4.8

〈그림 2〉 H-B 방법

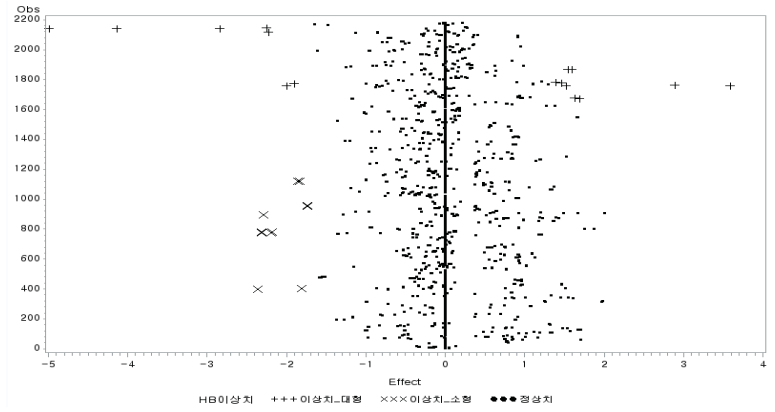


〈표 5〉 변형된 H-B 방법의 이상치 경계값

규모	표본수	e_i의 이상치 기준		이상치 수
		하한 경계값	상한 경계값	
$85m^2$ 미만	1,654	-1.641	2.943	11 (0.67%)
$85m^2$ 이상	528	-1.864	1.307	18 (3.41%)

* $85m^2$ 미만: c_L=c_U=5/ $85m^2$ 이상: c_L=c_U=4.8

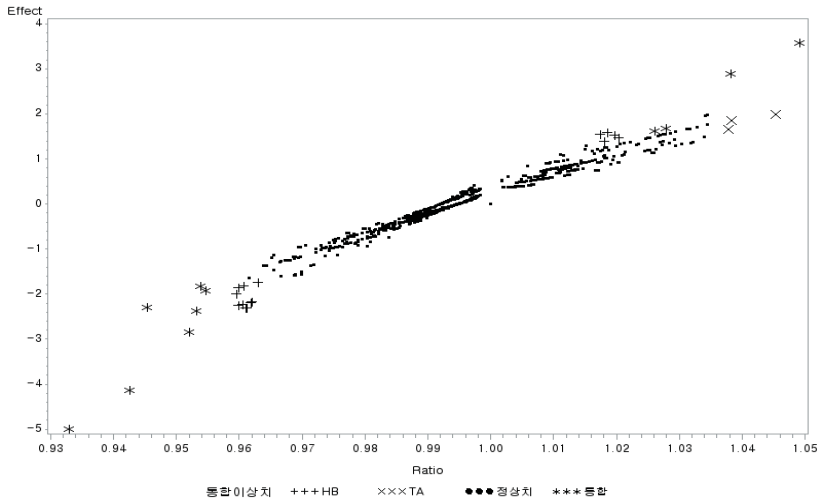
〈그림 3〉 변형된 H-B 방법



〈표 6〉 TA의 이상치 경계값

규모	$(\overline{r_d},\ \overline{r_U},\ \overline{r_L})$	표본수	r_i의 이상치 기준		이상치 수
			하한 경계값	상한 경계값	
85m^2 미만	(0.988, 0.997, 1.006)	1,654	0.958	1.036	6 (0.36%)
85m^2 이상	(0.985, 0.991, 0.998)	528	0.959	1.024	8 (1.52%)

〈그림 4〉 변형된 H-B 방법과 TA의 비교



기 위한 지표로, 주택의 가치를 반영하여 이상치를 탐색하는 H-B 방법이 다른 방법들보다 더 의미가 있을 것이라고 판단된다. 그리고 현재 월간 조사로 생성되는 주택 가격지수에서는 주기 특성에 의해 가격 변동이 없는 주택이 많기 때문에 변형된 H-B 방법의 사용이 효율적일 것이다.

4. 이상치 처리

이상치로 탐색이 된 자료의 원인이 조사, 자료 입력 및 처리에서 생겨난 오류라면 이러한 값들을 처리해 주는 과정이 필수적이다. 이상치를 고려하지 않고 조사된 자료를 분석한다면 모수 추정에 편향(bias)을 가져올 수 있다. 현재 전국주택가격동향조사에서는 재조사를 원칙으로 하고 있으나 일반적으로는 이상치 제거나 대체(imputation) 방법을 많이 사용하고 있다. 공표되고 있는 주택 가격지수는 기준월 대비 해당월 가격비율의 기하평균인 제본스(Jevons)지수[2]를 사용하고 있는데, 수치적 이상치의 영향력을 확인하기 위하여 이상치 제거 전과 후의 전월대비 지수 변동률을 비교해보고 단순 대체 방법 중 간단

2) 지수 추정치가 불편추정치가 아니라는 단점이 있지만 지수가 갖추어야 하는 주요 공리를 모두 만족하고 표본의 크기가 충분히 클 경우 일치추정량이기 때문에 국제적으로 권고하고 있다. 제본스 지수는 다음과 같이 두 시점의 가격비율을 기하평균하여 산정한다.

$HPI=\prod_{i=1}^{n}\left(p_i^t/p_i^0\right)^{n^{-1}}$, 여기서 p_i^0: i번째 표본의 기준월의 가격, p_i^t: i번째 표본의 해당월의 가격이다.

〈표 7〉 주택 가격지수의 전월대비 변동률

		전체 자료			이상치제거 (변형된 H-B)			이상치제거 (TA)		
		n	변동률	CV(%)	n	변동률	CV(%)	n	변동률	CV(%)
서울전체		2,182	-0.21	0.075	2,153	-0.21	0.073	2,168	-0.20	0.075
	85m2 미만	1,654	-0.15	0.085	1,643	-0.16	0.083	1,648	-0.14	0.085
	85m^2 이상	528	-0.42	0.147	510	-0.37	0.145	520	-0.37	0.147
		중위수대체[1] (H-B&TA)[2]			중위수대체 (변형된 H-B)			중위수대체 (TA)		
			변동률	CV(%)		변동률	CV(%)		변동률	CV(%)
서울전체			-0.21	0.075		-0.19	0.075		-0.21	0.074
	85m^2 미만		-0.14	0.085		-0.12	0.086		-0.15	0.085
	85m^2 이상		-0.40	0.145		-0.41	0.143		-0.40	0.145

[1]중위수대체: 이상치 제거 후, 시군구 읍면동과 규모별 중위수 대체,
[2]H-B&TA: 변형된 H-B 방법과 TA의 통합 이상치

한 중위수 대체를 추가적으로 진행하였다. <표 7>을 보면, 이상치의 비율이 모두 1.5% 미만 이므로 이상치 제거 전과 후의 변동률은 크게 차이가 없음을 확인할 수 있다. 그리고 H-B 방법은 하한경계 아래에 존재하는 관측값의 개수가 많아 가격비율이 작은값들이 이상치로 탈락되어 전체 자료를 사용했을 때 보다 중위수대체를 한 경우에서 전월대비 변동률이 높아 졌음을 확인할 수 있다. 반면에 TA는 상한과 하한경계 바깥쪽의 이상치의 비율이 유사하여 중위수 대체 후의 결과가 H-B 방법보다 전체자료에 더 근접한 변동률을 보여준다. 규모별 변동률은 85m^2 이상인 경우의 표본수가 상대적으로 적기 때문에 이상치의 영향을 많이 받고 있음을 알 수 있다.

V. 결론

전국주택가격동향조사에서는 매매 가격지수를 기준시점 대비 현시점 가격비율의 기하평균으로 산정하고 있다. 기하평균은 변화율이나 성장률과 같이 비율의 중심위치 척도에 유용하지만 이상치에 영향을 받을 가능성이 존재한다. 현재 조사에서는 자료의 분포를 고려하지 않고 상·하위 일정 부분만을 이상치로 판단하여 재조사를 한다. 본 연구에서는 짧은 주기의 조사에서 흔히 나타날 수 있는 전시점 대비 해당시점의 비율이 한 값에 과도하게 집중되어 있는 자료를 위한 여러 가지 이상치 탐색법을 고려하였다.

Box-plot, Z-score, 중위수 규칙과 같은 일반적인 방법은 적용이 간편하다는 장점이 있지만 대칭형 단봉분포(unimodal distribution)가 아니거나 표본수가 작을 경우에는 이상치를 탐색에 실패할 수 있다. 최소 허용구간에 의존하는 QM과 H-B 방법은 사분위수들이 적어도 두개가 동일 할 경우에는 자료의 성격을 제대로 반영하지 못하고 적절한 상한과 하한 경계의 계산이 불가능해 진다.

이러한 경우를 보완하고자 본 논문에서는 가격에 변동이 없는 관측값은 이상치 고려 대상에서 배제하는 방법으로 변형된 H-B 방법과 TA 분석을 제안하였다. 제안된 방법은 자료의 일부

분만을 사용하기에 결과의 정확성을 보장하기는 무리가 있지만 충분한 자료가 확보되는 경우라면 효율적으로 사용 가능 할 것이다.

일련의 과정들을 통해서 탐색된 수치적으로 분리되는 값들이 모두 실제로 이상치라고 단정할 수는 없기 때문에 처리 전 원인분석이 선행되어야 한다. 만약 어떠한 현상의 변화를 대변하는 중요한 값이라면 정상치로 재분류하여 분석을 시행하고, 그렇지 않다면 이상치를 제거하거나 관측값을 수정하는 등의 작업을 진행할 수 있다. 재조사를 통해서 오류값을 정정하거나 대체방법론을 적용하는 것이 이에 속한다. 또한, 자료 분석 시 가중치를 변경하거나 이상치에 크게 의존하지 않는 강건한 추정법(robust estimation)을 사용할 수 있다.

본 연구에서는 주기적으로 시행되는 전국주택가격동향조사의 매매가격 비율에 대한 여러 가지 이상치 탐색방법의 적용 및 비교를 통하여 현 시점에서 효율적으로 사용할 수 있는 방법을 제시하였다.

주관적으로 판단해오던 이상치를 비모수적 방법을 고려하여 객관적인 수치 이상치로 판별해 낼 수 있다는 데에서 본 연구는 큰 의미가 있다. 또한, 순위(rank)를 기초로 하는 이상치 탐색법과 관련변수를 이용한 다변량 이상치 탐색 방법 적용의 토대가 될 수 있을 것이다. 그러나 제안한 이상치 판단 방법은 허용구간을 이용하는 탐색방법들에서 문제 시 되는 경험적 판단이 필요한 상수의 설정문제와 중심과의 거리 개념의 이상치가 실제 이상치 인지에 대한 확인이 불가하다는 한계가 있다.

마지막으로, 본 연구에서는 주택가격동향조사라는 하나의 사례를 다루었지만 대부분의 계속조사의 업무 수행 시 이상치의 판단 문제는 공통적으로 발생할 수 있으므로 주택가격동향조사 이외에 계속조사를 바탕으로 한 통계작성 시에도 자료의 목적에 따라 적용이 가능할 것으로 보인다.

논문접수일 : 2013년 8월 12일
논문심사일 : 2013년 8월 19일
게재확정일 : 2013년 12월 20일

참고문헌

1. Barnerjee, S. and Iglewicz, B., “A simple univariate outlier identification procedure designed for large samples,” *Communication in Statistics - Simulation and Computation*, Vol. 36, 2007, pp.249-263.
2. Brys, G., Hubert, M., and Stuyf, A., “A comparison of some new measures of skewness,” in: *Developments in Roubust Statistics* (ICORS 2001), eds. Dutter, R., Filzmoser, P., Gather, U., and Rousseeuw, P. J., 2003, pp.98-113, Heidelberg: Springer-Verlag.
3. Carling, K., "Resistant outlier rules and the non-Gaussian case," *Computational Statistics and Data Analysis*, Vol. 33, 2000, pp.249-258.
4. Granquist, L., “Improving the traditional editing process,” in: *Business Survey Methods* (eds. Cox et al.). John Wiley & Sons, Chichester, 1995.
5. Hoaglin, D. C., Mosteller, F., and Tukey, J. W., (eds.) *Understanding robust and exploratory data analysis*. New York: Wiley, 1983.
6. Hidiroglou, M. A. and J. M. Berthelot, “Statistical editing and imputation for periodic business surveys,” *Survey Methodology*, Vol. 12, 1986, pp.73-84.
7. Hubert, M. and Vandervieren, E., “An adjusted boxplot for skewed distributions,” *Computational Statistical Data Analysis*, Vol. 52, 2008, pp.5186-5201.
8. John W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.
9. Lee, H., Ghangurde, P. D., Mach, L., and Yung, W., “*Outliers in Sample Surveys,”* Statistics Canada, August 1992.
10. Liu, R. Y., “On a notion of data depth based on random simplices,” Annals of Statistics, Vol. 18, 1990, pp.405-414.
11. Liu, R. Y. and Singh, K., “A quality index based on data depth and multivariate rank tests,” *Journal of the American Statistical Association*, Vol. 88, 1993, pp.252-260.
12. Serfling, R., “A depth function and a scale curve based on spatial quantiles,” in: Statistical data analysis based on the LI-Norm and related methods, ed. Dodge Y., Basel: Birkhauser, 2002, pp.25-38.
13. Thompson, K. J. and Sigman, R. S., “Statistical methods for developing ratio edit tolerances for economic data,” *Journal of Official Statistics,* Vol. 15, 1999, pp.517-535.