

스캐너 데이터에서의 이상점 탐지 기법

임요한^a, 이성임^b, 김영래^c, 김상균^d, 손원^e, 황희진^f

스캐너 데이터란 판매 시점에 스캐너에 의해 기록된 상거래 내역 데이터를 말한다. 스캐너 데이터는 상품명, 가격, 수량 등의 상세한 정보를 포함하고 있어 소비자물가지수 작성 등 공식통계 편제에도 활용할 수 있다. 이미 여러 해외 국가들이 스캐너 데이터를 소비자물가지수 작성에 사용하고 있으며 최근 우리나라에서도 관련 연구가 시작되었다. 스캐너 데이터를 실제로 사용하기 위해 선행되어야 할 중요한 과제 중 하나는 수집된 데이터에서 비정상적인 가격 관측값들을 탐지하고 이들을 사전에 제거하는 작업이다. 이상점 탐지를 위하여 Quartile 방법, Tukey 알고리즘 등의 다양한 방법이 제안되어 사용되고 있으나, 이들 방법은 가격 정보만 활용하고 있을 뿐, 가격과 밀접한 관계를 가지고 있는 판매 수량 정보를 고려하지 않고 있다.

본고에서는 판매 수량 정보를 반영한 새로운 스캐너 데이터 이상점 탐지 방법을 제안하고, 제안된 방법의 적절성과 효용성을 모의실험과 실제 데이터 분석 결과를 통해 알아본다.

I. 서론

II. 연구방법

1. 이상점 탐지 기존 연구
2. 관리 한계선 설정을 위한 분산 함수 추정
3. 거래량을 고려한 이상점 탐지

III. 모의실험

1. 모의실험 설계
2. 모의실험 결과

IV. 실증분석

1. 데이터 소개
2. 이상점 탐지 결과

V. 결론

a 서울대학교 통계학과(e-mail: johanlim@snu.ac.kr, phone: 02-880-2625)

b 단국대학교 응용통계학과(e-mail: silee@dankook.ac.kr, phone: 031-8005-3259)

c 서울대학교 통계학과 학생연구원

d 단국대학교 응용통계학과 학생연구원

e 단국대학교 응용통계학과(e-mail: son.won@dankook.ac.kr, phone: 031-8005-3269)

f 한국은행 경제통계국(e-mail: hjhwang@bok.or.kr, phone: 02-759-4329)

* 본 연구의 내용은 집필자들의 개인 의견으로 한국은행의 공식견해를 나타내는 것은 아님.

** 본 연구에 사용된 스캐너 데이터는 대한상공회의소로부터 입수하였음.

*** 본 연구에 기반한 영문 원고는 <https://arxiv.org/abs/1912.01832>에서 열람할 수 있고 해외 학술지에서 심사 중임.

I. 서론

스캐너 데이터란 소매점에서 판매되는 제품에 부착된 바코드를 판매 시점에 스캐너로 읽어 기록한 데이터로 POS(Point of Sale) 데이터라고도 한다(ILO, 2004). 최근 편의점 등 소규모 상점에도 판매관리 시스템이 갖춰지고 품목, 가격 및 수량 등 실제 거래에 대한 세부 정보가 수집되고 있어 스캐너 데이터의 활용 가치가 높은 것으로 평가된다.

소비자물가지수(CPI)는 거래 규모가 큰 대표 상품을 표본 조사하여 작성하고 있는데, 거래되는 상품과 업태가 다양해짐에 따라 스캐너 데이터와 같은 빅데이터를 활용한 물가지수 작성기법 연구가 활발히 진행되고 있다. 노르웨이, 스위스, 스웨덴, 호주 등 주요국 통계작성 기관에서는 물가지수를 작성할 때 실제로 스캐너 데이터를 이용하고 있으며(Mayhew, 2017), 영국(ONS, 2016), 일본(Abe & Tonogi, 2010) 등에서도 소비자물가지수 작성에 스캐너 데이터를 활용하기 위한 연구를 진행하고 있다.

스캐너 데이터의 활용에는 몇 가지 제약요인이 존재하는데, 그 중 대표적인 것이 이상점의 존재이다. 스캐너 데이터의 경우 판매 수량 또는 가격의 단위 착오 등으로 인한 이상점이 발생할 수 있다. 데이터 정제와 관련된 이러한 문제는 대부분의 실제 데이터 분석에서 중요하게 다뤄져야 할 문제이며, 특히 공식 통계의 작성에 있어서는 더 민감한 문제라고 할 수 있다. 예를 들어, 이상점이 다수 존재하여 비정상적인 가격을 지수 작성에 사용하거나 반대로 정상적인 가격 변동을 이상점으로 간주하여 제거한다면 소비자물가지수의 왜곡을 초래하게 되고 이를 기반으로 한 정부의 경제정책 또한 부정적인 영향을 받을 수 있기 때문이다.

이처럼 스캐너 데이터를 활용하기 위해서는 이상점을 정확하게 식별해내는 것이 매우 중요한 작업이라 할 수 있다. 그러나 스캐너 데이터는 수많은 상점에서 발생한 다양한 상품의 거래 정보로 구성된 대용량 자료인 만큼, 이를 일일이 검토하여 이상점 여부를 판단하는 데에는 많은 노력과 시간이 소요된다. 따라서 이상점을 정확하게 식별해낼 수 있는 통계적 기법의 개발이 요청되고 있다.

현재 이상점 탐지를 위해 사용되고 있는 방법으로 Quartile 방법과 Tukey 알고리즘, 그리고 이 방법들을 일부 변형한 방법들이 있다. ONS는 Tukey 알고리즘, 캐나다 통계국은 Quartile 방법, 미국 관세청은 Resistant Fences(RF) 방법을 사용하는 등 통계작성 기관에 따라 다양한 이상점 탐지 기법을 사용하고 있다(Rais, 2008). 이 방법들은 가격 변화율의 분포를 이용해 사전에 정한 임계값을 벗어난 관측값들을 이상점으로 판단한다는 공통점이 있다.

유럽의 경우 현재 가격보다 300% 이상 높거나 25% 수준보다 낮은 가격으로 거래된 것으로 관측되면 이를 이상점으로 판정하는 것이 대표적인 예라 할 수 있다(Eurostat, 2017).

이러한 기존의 이상점 탐지 기법들은 모두 품목의 단위 가격만을 사용한다는 한계가 있다. 만약 어떤 상품의 가격이 하락할 때 판매량이 크게 증가한다면, 이는 해당 상품이 일시적으로 할인된 가격에 판매된 것이라고 생각할 수 있다. 이처럼 가격과 수량이 밀접한 연관관계에 있다는 점을 고려할 때, 이상점 판정을 위해 가격뿐만 아니라 수량 정보도 활용할 수 있음을 알 수 있다. 본고에서는 스캐너 데이터에 포함되어 있는 가격과 판매량 정보를 모두 고려하여 기존의 방법보다 합리적인 이상점 탐지기법을 제안하고자 한다.

본고는 다음과 같이 구성된다. 먼저 II장에서는 기존의 연구방법에 대하여 고찰하고 판매량 정보를 함께 고려한 새로운 이상치 탐지 기법을 제시한다. III장에서는 기존 연구방법과 새롭게 제시된 연구방법을 비교하기 위해 모의실험 결과를 소개한다. IV장에서는 새롭게 제안한 방법을 실제 스캐너 데이터에 적용해 본다. 마지막으로 V장에서는 모의실험과 실제 자료 분석 결과를 평가한 후 향후 과제에 대해 논의한다.

II. 연구방법

1. 이상점 탐지 기준 연구

CPI 작성에 사용되고 있는 기존의 이상점 탐지기법들은 일반적으로 가격변화율이 정상적인 범위를 벗어나 크게 변화된 점을 이상점으로 정의하고 이를 탐지하기 위해 비모수적인 접근법을 사용하고 있다. 한 상품의 t 시점의 거래 가격을 P_t 라 하면 거래 가격의 변화율 R_t 는 t 시점의 가격을 $(t-1)$ 시점의 가격으로 나눈 비율 $R_t = P_t/P_{t-1}$ 로 정의한다.

기존의 방법들은 변화율 R_t 에 대해 허용 한계선(Tolerance Limit)을 정하고, 이를 넘어서는 변화율을 이상점으로 판단하는 방법을 사용한다. 본고에서는 이를 이상점을 관리하기 위한 관리 한계선(Control Limit)으로 부르기로 한다. 기존 이상점 탐지 방법의 대표적인 예로 Quartile 방법과 Tukey 알고리즘 방법이 있는데 이들 방법은 다음과 같다.

가. Quartile 방법(Quartile Method)

Quartile 방법은 R_t 의 사분위수를 이용하여 관리 한계선을 결정한다. 제 1, 2, 3사분위수를 각각 Q_1, Q_2, Q_3 라 할 때, 변화율에 대한 관리 상한선(Q_U)과 관리 하한선(Q_L)은 다음과 같이 정의된다.

$$\text{관리 상한선: } Q_U = Q_2 + c_u(Q_3 - Q_2)$$

$$\text{관리 하한선: } Q_L = Q_2 - c_l(Q_2 - Q_1)$$

이 방법은 데이터로부터 유연하게 관리 한계선을 정하므로 R_t 가 대칭이 아닌 경우에도 실제 데이터의 분포에 적합한 관리 한계선을 설정할 수 있도록 한다. 관리 한계선을 벗어나는 이상점은 조절모수 c_u 와 c_l 에 의해 결정되므로 이들 조절모수의 결정이 중요하다. Quartile 방법의 관리 한계선을 이해하기 위해, R_t 가 정규분포를 따르며 $c_u = c_l = 4.5$ 라고 가정한다면 관리 한계선은 대략 $(Q_2 - 3\sigma, Q_2 + 3\sigma)$ 로 주어진다. 이것은 전체 관측값들 중 약 99.73%만을 정상으로 판정하고 나머지 0.27%는 이상점으로 판단하게 된다는 것을 의미한다. 이 경우 가격 변화율에 특별한 문제가 없더라도 0.27%에 해당하는 관측값들은 항상

이상점으로 탐지될 수 있으므로 가짜 알람률(false alarm rate)이 0.27%인 기법이라 할 수 있다.

Quartile 방법은 관측값의 분포가 비대칭인 경우에도 이상점 탐지를 효율적으로 할 수 있도록 유연하게 한계선을 정의하는 장점이 있지만 단점 또한 존재한다. 첫 번째 문제는 가격 변동이 자주 발생하지 않는 경우에는 $Q_1 \approx Q_2 \approx Q_3$ 가 되어, c_u 와 c_l 을 크게 설정하더라도 지나치게 많은 데이터가 이상점으로 판정된다는 것이다.

두 번째 문제는 분포가 오른쪽으로 치우친 데이터에 존재하는 가면 효과(Masking Effect)이다. 가면 효과란 특정 이상점으로 인하여 다른 이상점이 드러나지 않는 현상으로 이 경우 오른쪽 꼬리 부분의 이상점에 대해서는 민감하게 반응하고 반대로 왼쪽 꼬리 부분의 이상점에 대해서는 둔감하게 반응하게 된다(Rais, 2008). 가면 효과는 데이터 변환을 통해 해결할 수 있으며, 일반적으로 자연로그 변환을 이용해 가면 효과를 해결할 수 있는 것으로 알려져 있다(Saïdi et al., 2005; Thompson et al., 1999).

나. Tukey 알고리즘(Tukey Algorithm)

Tukey가 처음 제안한 방법으로, Quartile 방법과 달리 가격 변동이 거의 발생하지 않는 경우에도 이상점 판단에 유용하게 사용할 수 있는 장점이 있다. Tukey 알고리즘은 가격 변동이 없는($R_t = 1$) 데이터는 제거하고, R_t 에 대한 표본을 다시 구축하여, 이를 Tukey 표본이라 부른다. Tukey 표본을 $\{R_1^t, R_2^t, \dots, R_n^t\}$ 이라 할 때, 관리 한계선은 다음과 같이 정의한다.

$$\text{관리 상한선: } T_U = \bar{R}^t + 2.5(\bar{R}_U^t - \bar{R}^t)$$

$$\text{관리 하한선: } T_L = \bar{R}^t - 2.5(\bar{R}^t - \bar{R}_L^t)$$

이때, \bar{R}^t 는 Tukey 표본의 표본평균이며, \bar{R}_U^t 은 \bar{R}^t 보다 큰 Tukey 표본의 평균, \bar{R}_L^t 은 \bar{R}^t 보다 작은 Tukey 표본의 평균을 의미한다. Tukey 알고리즘은 사전에 불필요한 데이터를 제거하기 때문에 이상점이 일부 존재하는 경우에도 강건한(robust) 한계선을 구할 수 있는 방법으로 알려져 있다(ONS, 2010). 그러나 데이터의 일부만 사용하여 이상점을 판단하기 때문에 Tukey 표본의 크기가 크지 않을 경우 이상점 탐지의 정확도가 떨어질 수도 있다(Rais, 2008).

2. 관리 한계선 설정을 위한 분산 함수 추정

앞 절에서 살펴본 기존 이상점 탐지 기법들은 가격 변화율 R_t 의 분포만 고려하고 있으며 거래량은 이상점 탐지에 사용되지 않는다. 본 절에서는 가격 변화율과 거래량이 밀접한 관계가 있음에 착안하여 가격변화율의 분포함수를 거래량의 함수로 표현하고 이를 추정하여 활용하는 새로운 이상점 탐지 방법을 제안하고자 한다. 이 절에서는 가격 변화율의 비대칭성을 반영하기 위해 R_t 를 로그 변환한 데이터 $Y_t = \log(R_t) = \log(P_t/P_{t-1})$ 를 고려하기로 하고 Y_t 의 변동성이 큰 경우와 작은 경우로 나누어서 각각의 경우의 분산함수 추정 방안을 제안한다.

가. 가격변화율의 변동성이 큰 경우의 분산함수 추정

먼저 가격변화율의 변동성이 큰 경우의 분산함수 추정 방법을 살펴본다. 시점 t 의 거래 가격을 P_t 라 하면 $Y_t = \log(R_t) = \log(P_t/P_{t-1})$ 로 이전 시점에 대한 가격 변화율의 로그 값으로 정의된다. 시점 t 의 거래량을 V_t 라 하면 Y_t 의 확률분포는 다음과 같이 표현할 수 있다.

$$Y_t = \mu(V_t, V_{t-1}) + \sigma(V_t, V_{t-1})\epsilon_t \quad (2.1)$$

여기서 오차항 ϵ_t 는 기댓값이 0이고 분산이 1인 분포를 따른다고 가정한다. 본 연구의 목적인 이상점 탐지를 위해서는 가격에 체계적인 변화가 없는 정상상태, 즉 $\mu(V_t, V_{t-1}) = 0$ 에서 Y_t 의 분포를 추정해야 한다. 따라서 이상점을 탐지하기 위한 관리 한계선을 정하는 문제는 정상적인 분포 안에서 관측 가능한 범위를 정하는 문제가 되고, 이것은 모형 (2.1) 하에서 Y_t 의 산포 $\sigma^2(V_t, V_{t-1})$ 를 추정하는 문제로 단순화된다. 시점 t 에서 관측 가능한 데이터는 (Y_t, V_t, V_{t-1}) 이고 $\mu(V_t, V_{t-1}) = 0$ 의 가정하에 모형 (2.1)의 양변을 제곱하면

$$\begin{aligned} Y_t^2 &= \sigma^2(V_t, V_{t-1})\epsilon_t^2 \\ &= \sigma^2(V_t, V_{t-1}) + \sigma^2(V_t, V_{t-1})(\epsilon_t^2 - 1) \\ &= \sigma^2(V_t, V_{t-1}) + \sigma^2(V_t, V_{t-1})u_t \end{aligned} \quad (2.2)$$

의 관계를 얻는다. 이때, u_t 는 기댓값이 0이고 분산이 2인 비대칭 확률변수가 된다. 관계

식 (2.2)로부터 Y_t 의 분산함수 추정은 설명변수 $\tilde{V}_t = (V_t, V_{t-1})$ 를 이용하여 Y_t^2 의 평균 함수를 추정하는 문제로 볼 수 있고 본 연구에서는 가장 보편적인 추정량인 커널 추정량 $\hat{f}(\tilde{V}_t)$ 을 사용한다.

$$\hat{\sigma}^2(\tilde{V}_t) \equiv \hat{f}(\tilde{V}_t) = \frac{\sum_{i \in T} K_h(\tilde{V}_t, \tilde{V}_i) Y_i^2}{\sum_{i \in T} K_h(\tilde{V}_t, \tilde{V}_i)} \quad (2.3)$$

이때 T 는 추정에 사용되는 과거 데이터의 집합이고, K_h 는 커널 함수로, 여기서는 가우시안 커널을 두 번 미분해 구한 이차 가우시안 커널을 사용한다.

$$\nabla^2 G_h = \frac{\partial^2 G_h}{\partial x^2} + \frac{\partial^2 G_h}{\partial y^2}$$

이때 $G_h(V_1, V_2) = \exp\left(-\frac{\|V_1 - V_2\|}{h}\right)$ 이다. 위와 같이 계산된 커널 추정량 $\hat{f}(\tilde{V}_t)$ 을 $\sigma^2(V_t, V_{t-1})$ 함수의 추정량으로 사용한다. 이때 커널 추정은 R 소프트웨어의 'np' 패키지의 'npreg' 함수를 이용하였고, bandwidth h 는

$$CV(h) = \sum_{t \in T} (Y_t^2 - \hat{f}_{-t}(\tilde{V}_t))^2$$

를 최소화하도록 교차검증(cross validation)을 통해 결정하였다. 이때 $\hat{f}_{-t}(\tilde{V}_t)$ 는 t 시점의 관측값을 제외하고 계산한 커널 추정량으로, 커널 함수를 K_h 로 표현하면 다음과 같이 계산된다.

$$\hat{f}_{-t}(\tilde{V}_t) = \frac{\sum_{i \in T_t} K_h(\tilde{V}_t, \tilde{V}_i) Y_i^2}{\sum_{i \in T_t} K_h(\tilde{V}_t, \tilde{V}_i)} \quad (2.4)$$

나. 가격변화율의 변동성이 작은 경우의 분산함수 추정

다음으로 가격변화율의 변동성이 작은 경우의 분산함수 추정 방법을 살펴본다. 변동성이 작은 경우에는 분산함수 추정을 위해 필요한 충분한 관측값을 확보하기 어려우므로 변동성이 클 때와 달리 이전 시점의 가격이 아닌 표준가격을 활용한다. 즉, 시점 t 에서의 가격을 P_t , 표준가격을 P_s 라 하면 $Y_t = \log(R_t) = \log(P_t/P_s)$ 로 나타난다. 시점 t 에서의 거래량을 V_t 라 하면 Y_t 의 확률분포는 다음과 같이 표현된다.

$$Y_t = \mu(V_t) + \sigma(V_t)\epsilon_t \quad (2.5)$$

여기서 오차항 ϵ_t 는 변동성이 클 때와 마찬가지로 기댓값이 0이고 분산이 1인 분포를 따른다고 가정하고 이상점 탐지를 위해서 정상상태($\mu(V_t) = 0$)에서 Y_t 의 분포를 추정한다. 또한, 이상점을 탐지하기 위한 관리 한계선을 정하는 문제는 정상적인 분포 안에서 관측 가능한 범위를 정하는 문제로, 모형 (2.5) 하에서는 Y_t 의 산포 $\sigma^2(V_t)$ 를 추정하는 문제로 단순화된다.

시점 t 에서 관측 가능한 데이터는 (Y_t, V_t) 이고 $\mu(V_t) = 0$ 의 가정 하에 모형 (2.5)의 양변을 제곱하면

$$\begin{aligned} Y_t^2 &= \sigma^2(V_t)\epsilon_t^2 \\ &= \sigma^2(V_t) + \sigma^2(V_t)(\epsilon_t^2 - 1) \\ &= \sigma^2(V_t) + \sigma^2(V_t)u_t \end{aligned} \quad (2.6)$$

의 관계식을 얻는다. 이때, u_t 는 변동성이 클 때와 마찬가지로 기댓값이 0이고 분산이 2인 비대칭 확률변수가 된다.

관계식 식 (2.6)으로부터 분산 추정은 설명변수 V_t 를 이용하여 Y_t^2 의 평균함수를 추정하는 문제로 볼 수 있으며, 커널 추정량 $\hat{f}(V_t)$ 을 사용하여 구한다.

$$\hat{\sigma}^2(V_t) \equiv \hat{f}(V_t) = \frac{\sum_{i \in T} K_h(V_t, V_i) Y_i^2}{\sum_{i \in T} K_h(V_t, V_i)} \quad (2.7)$$

이때, T 는 추정에 사용되는 과거 데이터의 집합이고 K_h 는 이차 가우시안 커널 함수이다. bandwidth h 는 변동성이 큰 가격변화율의 분산함수에서와 동일한 방법으로 구할 수 있다.

3. 거래량을 고려한 이상점 탐지

이 절에서는 위에서 제안한 방법으로 추정된 분산함수를 통해 이상점을 탐지하는 방법을 소개한다. 기존의 이상점 탐지 방법인 Quartile 방법과 Tukey 알고리즘이 가격 변화율만 고려하여 이상점을 정의한 데 반해, 실제로는 가격이 거래량과 관련되어 있다는 사실에 주목하여, 이 절에서는 가격의 산포를 설정할 때 거래량을 반영할 것을 제안한다.

먼저 가격변화율의 변동성이 큰 경우, t 시점에 새로운 데이터 Y_t, V_t 가 관측되면 관측값 (V_t, V_{t-1}) 을 분산함수 추정식 식 (2.3)에 대입하여 분산 추정량 $\hat{\sigma}^2(V_t, V_{t-1})$ 을 계산하고, 이를 이용해 관리 상한선과 하한선을 다음과 같이 계산한다.

$$\begin{aligned} \text{관리 상한선: } & 3\hat{\sigma}(V_t, V_{t-1}) \\ \text{관리 하한선: } & -3\hat{\sigma}(V_t, V_{t-1}) \end{aligned} \quad (2.8)$$

가격변화율의 변동성이 작은 경우에는 t 시점에 새로운 데이터 Y_t, V_t 가 관측되면 관측값 V_t 를 분산함수 식 (2.7)에 대입하여 분산 추정량 $\hat{\sigma}^2(V_t)$ 을 계산하고, 이를 이용해 관리 상한선과 하한선을 다음과 같이 계산한다.

$$\begin{aligned} \text{관리 상한선: } & 3\hat{\sigma}(V_t) \\ \text{관리 하한선: } & -3\hat{\sigma}(V_t) \end{aligned} \quad (2.9)$$

위 관리 한계선 식 (2.8)과 식 (2.9)에서 상수 3은 Y_t 가 정규분포일 경우 제1종 오류의 크기를 0.27%로 한다.

III. 모의실험

1. 모의실험 설계

이 절에서는 새롭게 제안한 이상점 탐지 기법의 성능을 확인하기 위한 모의실험 결과를 소개한다. 식 (2.8)과 식 (2.9)에서 제안된 Variable 방법과 i) 거래량을 고려하지 않고 과거 자료의 표본분산으로 분산을 추정한 Constant 방법, ii) 분산함수를 알고 있다고 가정한 Oracle 방법, 그리고 기존에 가격 변화율을 이용해서 만든 iii) Tukey 알고리즘 방법 및 iv) Quartile 방법을 모의실험을 통해 비교해 보고 이상점 탐지에 있어 이들 방법이 가지는 장 단점을 확인해 본다.

모의실험을 위한 자료는 다음과 같이 생성한다. 과거 데이터로 사용될 학습데이터 (training data)의 개수는 $m = 100$ 일 때와 $m = 300$ 으로 하였으며, 평가데이터(test data)의 개수는 $n = 300$ 으로 고정하였다. 또한, 평가데이터 중 10%의 시점을 랜덤하게 선택하여 Y_t 의 값을 평행이동한 후 이를 이상점으로 정의하고, 이 시점을 모은 집합을 J 라고 정의한다. 학습데이터와 평가데이터는 각 시점에서의 가격과 거래량 (P_t, V_t) 으로 구성되었으며, 거래량 V_t 는 카이제곱 분포로부터 생성하였다. 가격 P_t 는 변동성이 클 때(식 (2.1))와 변동성이 작을 때(식 (2.5))로 나누어, 다음과 같이 생성하였다.

$$V_t \stackrel{i.i.d.}{\sim} 1 + \chi^2(5)$$
$$P_t = \begin{cases} P_{t-1} \exp(\sigma(V_t, V_{t-1})\epsilon_t + \delta_t) & : \text{변동성이 클 때} \\ P_{t-1} \exp(\sigma(V_t)\epsilon_t + \delta_t) & : \text{변동성이 작을 때} \end{cases} \quad (\text{단, } P_0 = 1)$$

이때 $\epsilon_t \stackrel{i.i.d.}{\sim} N(0,1)$ 이고, δ_t 는 이상점에서 0이 아닌 값을 가지는 가변수(dummy variable)이며, 다음과 같이 정의된다.

$$\delta_t = \begin{cases} 2 & \text{for } t \in J \\ 0 & \text{for } t \notin J \end{cases}$$

또한, 분산 $\sigma^2(V_t)$ 의 형태로는 아래와 같은 세 경우를 고려한다¹⁾.

$$\begin{aligned} \text{Case 1: } \sigma^2(V_t, V_{t-1}) &= 1 & (3.1) \\ \text{Case 2: } \sigma^2(V_t, V_{t-1}) &= \frac{1}{46} V_t^2 \\ \text{Case 3: } \sigma^2(V_t, V_{t-1}) &= \frac{1}{92} (V_t + V_{t-1})^2 \end{aligned}$$

분산의 형태 변화에 따른 이상점 탐지의 성능을 알아보기 위해, 분산 $\sigma^2(V_t)$ 이 거래 수량과 관련 없는 경우(Case 1), 현재 시점의 거래 수량에만 관련이 있는 경우(Case 2), 그리고 한 시점 이전까지의 거래 수량까지 관련 있는 경우(Case 3)를 각각 가정하였다. 위의 가정으로부터 학습데이터와 평가데이터 (P_t, V_t)를 각각 m 개, n 개 생성하였다. 단, 변동성이 작을 때는 이전 시점의 가격이 아닌 표준가격을 사용하기 때문에 식 (3.2)와 같이 Case 1과 Case 2만 고려하였다.

$$\begin{aligned} \text{Case 1: } \sigma^2(V_t, V_s) &= 1 & (3.2) \\ \text{Case 2: } \sigma^2(V_t, V_s) &= \frac{1}{46} V_t^2 \end{aligned}$$

이 가정들 중에서 Case 1의 경우 거래량에 관계없이 분산이 항상 일정하다고 가정하고 있으므로 비교 대상이 되는 방법들 중 거래량을 고려하지 않고 과거 자료의 표본분산으로 분산을 추정하는 Constant 방법에 유리하고, Case 2와 Case 3의 경우에는 Constant 방법에 상대적으로 불리한 반면 본고에서 제안한 Variable 방법에는 유리한 가정이라 볼 수 있다.

2. 모의실험 결과

각 이상점 탐지 기법의 성능을 평가하기 위해 사용한 척도들은 다음과 같이 정의된다. 먼저 진양성(IP, True Positive)은 실제로 양성(P, Positive)인 관측값을 양성으로 판별하는 경우를 뜻하며 이 모의실험에서는 이상점일 때 이상점이라고 바르게 판단한 경우를 말한다. 이와 반대로 위양성(FP, False Positive)은 실제로 이상점이 아닐 때, 이상점이라고 판별한 경우로 정의된다. 진음성(IN, True Negative)은 실제로 이상점이 아닐 때 (N, Negative) 이상점

1) Case 2와 Case 3의 계수 $\frac{1}{46}$ 과 $\frac{1}{92}$ 은 $E(V_t^2) = 46$ 으로부터 결정되었다.

이 아니라고 판별한 경우이고, 위음성(FN, False negative)은 실제로 이상점이지만 이상점이 아니라고 판별한 경우이다. 이들을 이용해 이상점 탐지 기법의 평가를 위한 척도로 민감도 (SEN, Sensitivity), 특이도(SPE, Specificity), 정확도(ACC, Accuracy)를 계산할 수 있고, 그 정의는 다음과 같다.

$$SEN = \frac{TP}{P} = \frac{TP}{(TP+FN)} \quad (3.3)$$

$$SPE = \frac{TN}{N} = \frac{TN}{(TN+FP)}$$

$$ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN}$$

가. 가격변화율의 변동성이 큰 경우에 대한 모의실험

<표 1>과 <그림 1>은 가격변화율의 변동성이 큰 경우 식 (3.1)과 같은 $\sigma^2(V_t, V_{t-1})$ 에 대한 가정 Case 1, Case 2, Case 3에서의 모의실험 결과를 나타낸다. 각각 $m = 100$, $m = 300$ 개의 학습데이터(training data)를 적용하여 정확도 및 민감도와 특이도를 비교하였다.

먼저 <표 1>을 통해 정확도를 살펴보면 분산함수를 사전에 정확히 알고 있다고 가정하는 경우(Oracle)를 제외하면 본고에서 제안한 Variable 방법의 정확도가 대체로 가장 높게 나타났다. 다만, 학습데이터의 개수가 $m = 100$ 개인 경우에는 분산 $\sigma^2(V_t)$ 이 거래 수량과 관련 없다고 가정한 Case 1에서 Variable 방법의 정확도가 Quartile 방법 또는 Constant 방법에 비해 낮게 나타났으나 학습데이터가 $m = 300$ 개로 늘어나면 정확도가 비슷해지는 것으로 관찰되었다.

<표 1> 이상점 탐지 기법별 정확도(가격변화율 변동성이 큰 경우)

		이상점 탐지 기법				
		Quartile	Tukey	Constant	Variable	Oracle
$m = 100$ $n = 300$	Case 1	0.913	0.897	0.913	0.909	0.911
	Case 2	0.902	0.894	0.899	0.922	0.941
	Case 3	0.887	0.866	0.895	0.897	0.914
$m = 300$ $n = 300$	Case 1	0.913	0.908	0.913	0.913	0.914
	Case 2	0.900	0.890	0.898	0.931	0.939
	Case 3	0.892	0.874	0.896	0.904	0.914

<그림 1>은 각 경우에 대한 이상점 탐지 기법들의 민감도와 특이도를 보여준다. 각 그림에 표시된 오차막대(Error Bar)는 표준오차를 이용한 신뢰구간($\pm 2\sigma$)을 나타낸다.

학습데이터의 개수가 $m = 100$ 개와 $m = 300$ 개로 달라진다 하더라도 실험 결과에는 큰 차이 없이 기존 방법중 Tukey 방법이 다른 방법들에 비해 민감도가 높고 특이도는 낮은 것으로 나타났다. 이는 Tukey 방법의 이상점으로 판단하는 기준이 상대적으로 낮게 책정되어 있어 다른 방법들에 비해 더 많은 관측값들을 이상점이라고 탐지하는 경향이 있음을 의미한다. 실제로 이상점이라 판별한 관측값의 수, 즉 진양성과 위양성의 합이 더 큰 것을 확인할 수 있다.

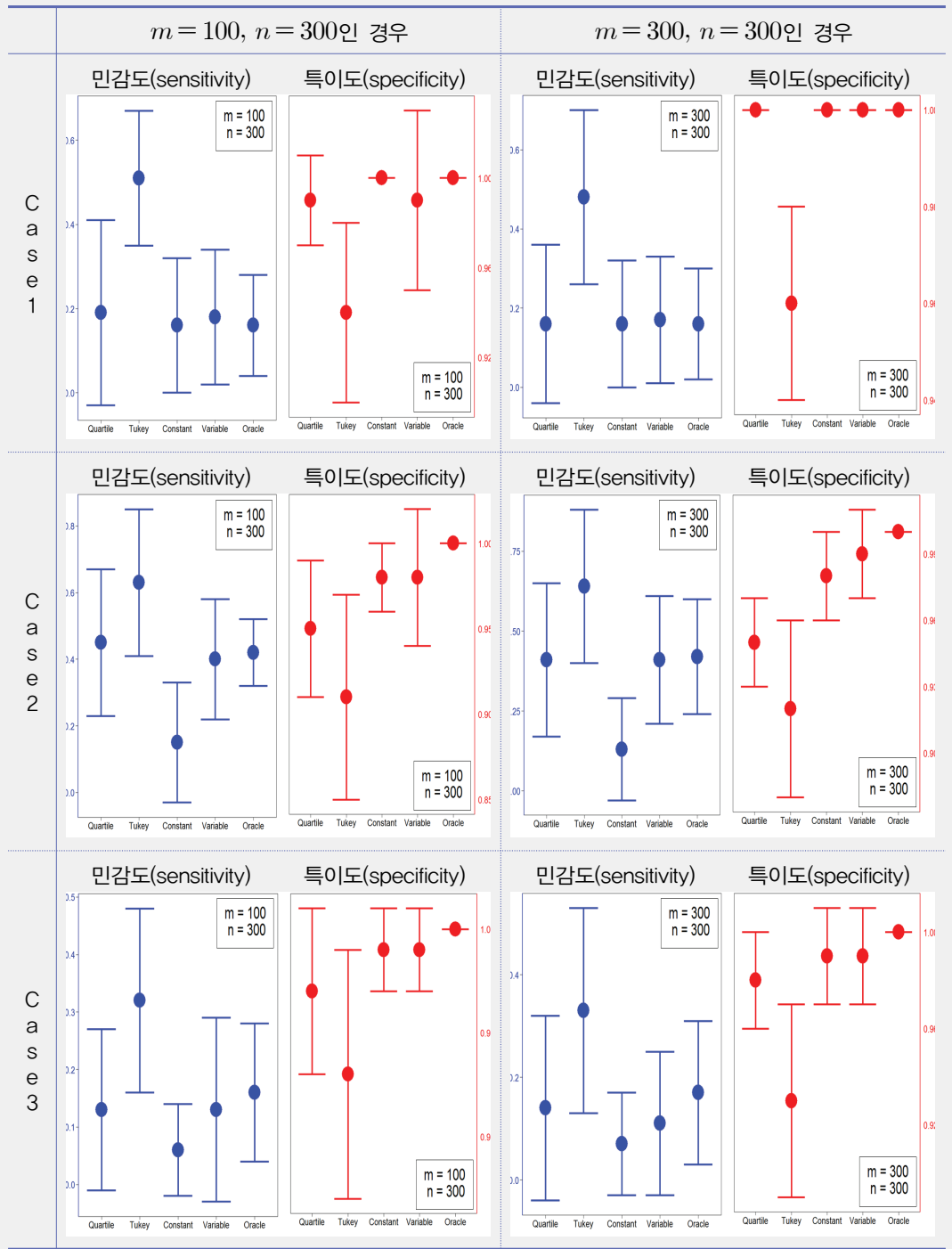
분산함수의 형태에 대한 가정별로 나누어, 먼저 분산함수의 형태를 상수로 가정한 경우 (Case 1)에 대해 살펴보자. 이 경우에는 분산을 상수함수로 추정하는 방법들의 성능이 더 좋은 것으로 나타날 수 있는데, 새로 제안한 Variable 방법이 Quartile 방법과 과거 자료의 표본분산으로 분산을 추정한 Constant 방법 등은 물론 분산함수를 정확히 알고 있다고 가정하는 Oracle 방법의 경우와 비교했을 때에도 그 성능이 비슷하게 나타남을 알 수 있다. 다만, 특이도의 경우 $m = 100$ 일 때는 Variable 방법이 Constant 방법보다 낮았으나 $m = 300$ 으로 증가하면 Constant 및 Oracle 방법과 같은 수준으로 높아졌다.

분산함수가 현재 시점의 거래량의 함수로 주어진 경우(Case 2)와 현재 시점의 거래량과 한 시점 이전의 거래량의 함수로 주어진 경우(Case 3)를 살펴보면, 본고에서 제안한 Variable 방법이 분산함수를 정확히 알고 있다고 가정하는 Oracle 방법의 경우에는 미치지 못하지만 다른 방법론들과 비교했을 때에는 더 뛰어난 정확도를 보이는 것을 확인할 수 있다. 특히, Quartile 방법, Tukey 방법 등 기존 방법들보다 특이도가 높게 나타나 이상점이 아닌 관측값을 이상점이 아니라고 옳게 판단할 가능성이 크다는 것을 알 수 있다. 분산함수를 상수함수로 추정하는 Constant 방법에 비해서는 민감도 측면에서 더 좋은 결과를 보였다. 즉, 실제 이상점인 관측값을 이상점으로 잘 판단할 가능성이 Constant 방법에 비해 높게 나타났다.

이상과 같은 실험 결과를 통해 보았을 때 본고에서 새롭게 제안한 분산함수 추정을 이용한 Variable 방법이 가격변화율의 분산에 변화가 있는 경우는 물론, 분산이 항상 일정한 경우에도 기존의 다른 방법들과 가격변화율의 표본분산을 이용한 Constant 방법에 비해 더 좋은 성능을 보여줄 수 있다.

<그림 1>

이상점 탐지 기법별 민감도와 특이도 비교
(가격변화율 변동성이 큰 경우)



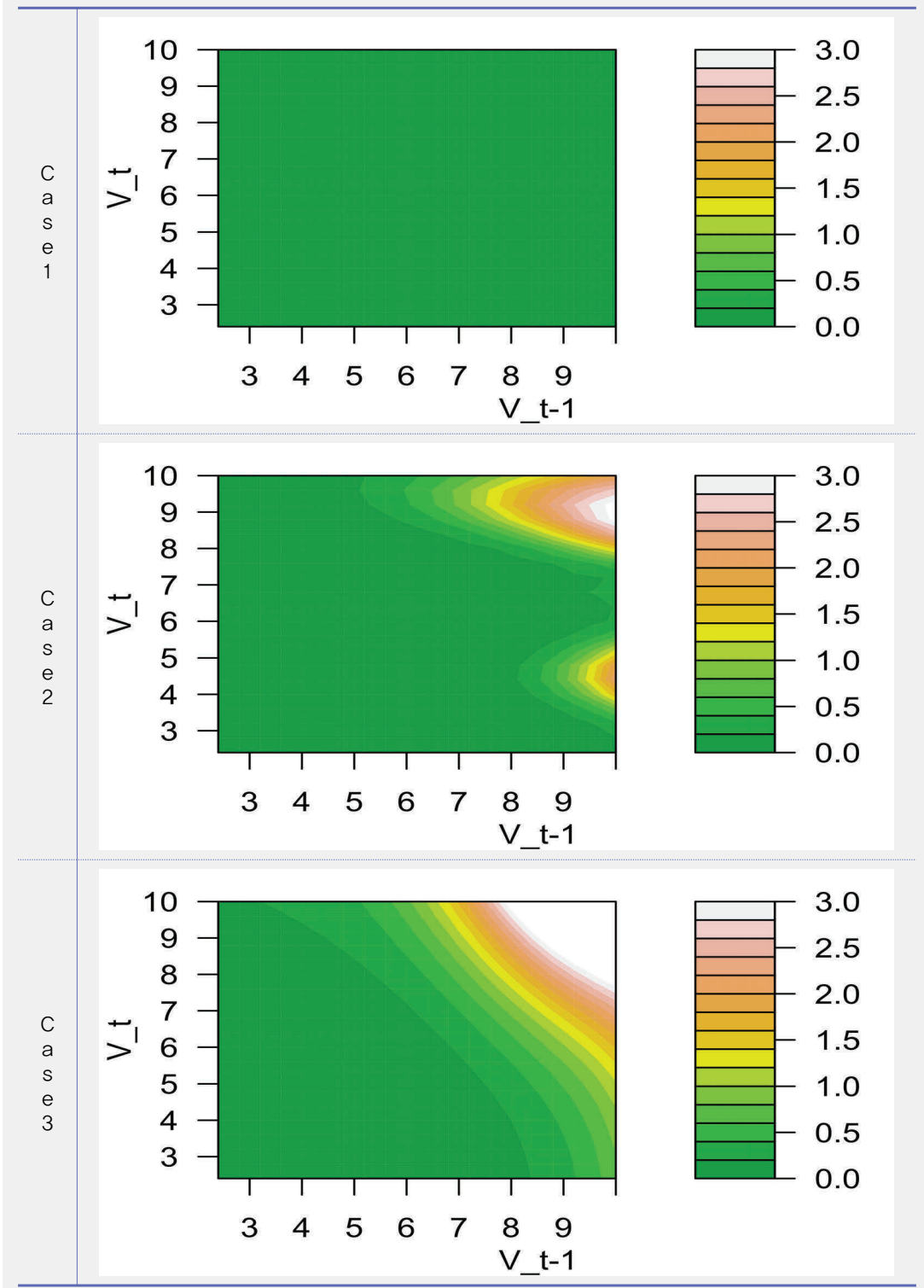
본 연구에서 제시한 분산함수 추정식 식 (2.3)이 정확한지 알아보기 위해, 같은 모의실험 설계 하에 분산함수의 추정량을 계산해 보았다. (V_t, V_{t-1}) 두 개의 설명변수로 분산함수를 추정한 경우에 편의(bias)의 제곱과 추정량의 분산의 합인 평균제곱 오차(MSE, Mean squared error)를 계산하여 <그림 2>와 같이 등고선으로 나타냈다.

먼저 분산이 항상 일정한 *Case 1*의 경우를 살펴보면 식 (2.3)의 커널 추정에서 Bandwidth h 가 매우 큰 값을 가져 분산함수를 거의 상수로 추정하고, 따라서 모든 (V_t, V_{t-1}) 에 대해 평균제곱오차가 거의 비슷하게 나타나며 *Case 2*, *Case 3*과 비교하면 모든 부분에서 추정이 잘되고 있는 것을 확인할 수 있다.

*Case 2*와 *Case 3*의 경우에는 V_t 또는 V_{t-1} 이 큰 경우 분산 추정량에 대한 평균제곱 오차가 크게 나타나는 것을 확인할 수 있는데, 모의실험 단계에서 $V_t \stackrel{i.i.d}{\sim} 1 + \chi^2(5)$ 로 생성했기 때문에, V_t 가 10 정도 되는 큰 값에서는 자료가 상대적으로 적게 생성되어 커널을 이용한 비모수 추정의 정확성이 떨어지기 때문이라고 이해할 수 있다.

<그림 2>

분산 추정량의 평균제곱오차(가격변화율 변동성이 큰 경우)



나. 가격변화율의 변동성이 작은 경우에 대한 모의실험

<표 2>와 <그림 3>은 가격변화율의 변동성이 작은 경우에 대한 분산함수식 (3.2)의 *Case 1*, *Case 2*에 대한 모의실험 결과를 보여준다.

먼저 <표 2>에서 정확도를 살펴보면 대체로 분산함수를 사전에 정확히 알고 있다고 가정하는 Oracle 방법의 경우, 본고에서 제안한 Variable 방법, Quartile 방법과 Constant 방법, Tukey 방법의 순으로 정확도가 높게 나타났다. 다만, 분산 $\sigma^2(V_t)$ 이 거래 수량과 관련 없다고 가정한 *Case 1*에서는 Tukey 방법을 제외한 다른 방법들의 정확도가 분산함수를 사전에 정확히 알고 있다고 가정하는 Oracle 방법의 경우에 비해 미세하게 높게 나타나기도 하였다. 이는 분산함수가 고정된 *Case 1*의 경우에는 분산함수에 대한 사전정보가 실제 분산함수를 추정하는 데 큰 도움이 되지 않기 때문인 것으로 판단된다. 앞에서 살펴본 바와 같이 Tukey 방법은 상대적으로 많은 관측값을 이상점이라고 판정하는 경향이 있어 정확도가 가장 낮게 나타났다.

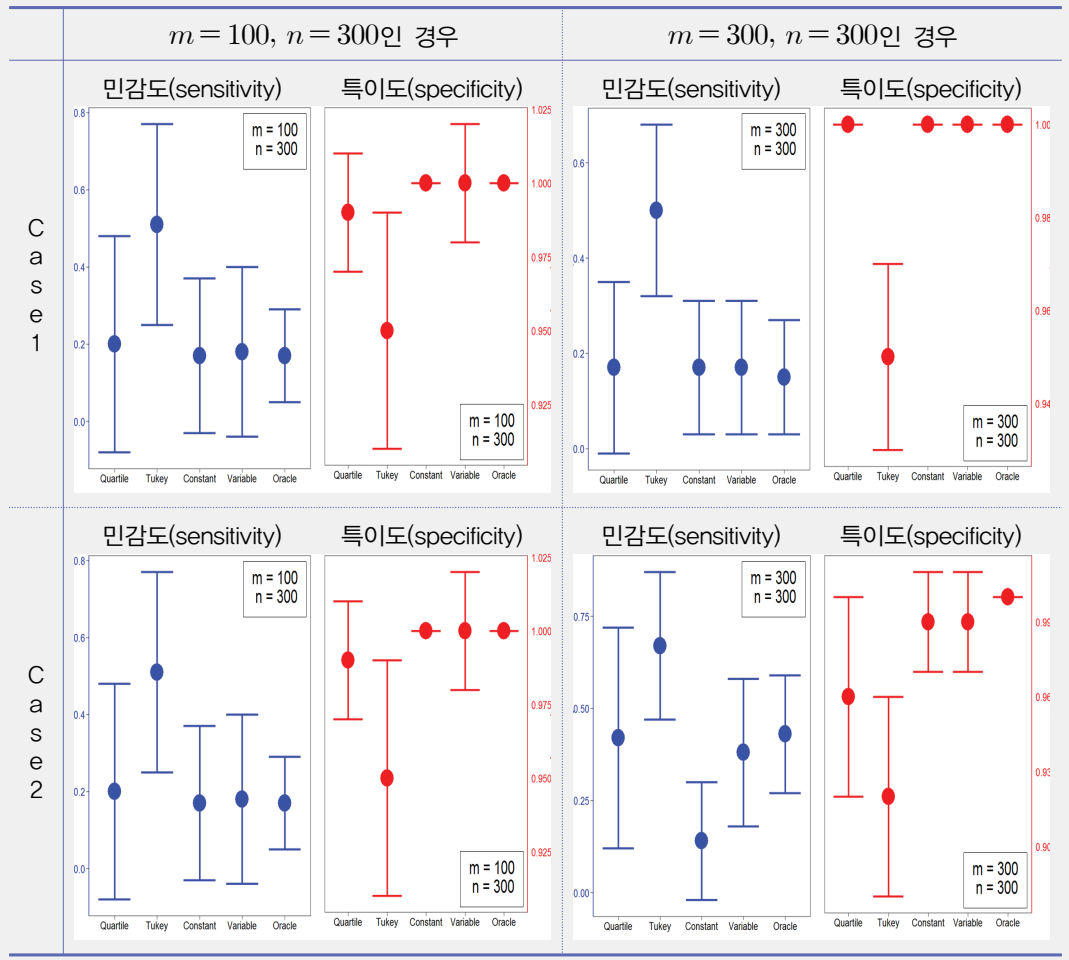
다음으로 <그림 3>에 나타난 민감도와 특이도를 살펴보면 가격변화율의 변동성이 큰 경우의 모의실험과 동일하게 *Case 1*에서는 새로 제안한 Variable 방법이 Quartile 방법과 Constant 방법은 물론 Oracle 방법과도 비슷한 성능을 보였다. 또한 분산함수가 현재 시점의 거래량의 함수로 주어진 *Case 2*에서는 본고에서 제안한 Variable 방법이 분산함수를 정확히 알고 있다고 가정하는 경우(Oracle)에는 미치지 못하지만 다른 방법론들과 비교했을 때에는 더 뛰어난 성능을 보이는 것을 확인할 수 있었다. Quartile, Tukey 등 기존 방법들보다는 특이도 측면에서, Constant 방법에 비해서는 민감도 측면에서 더 좋은 성능을 보였다. 즉, 기존 방법들에 비해서는 이상점이 아닌 관측값의 식별에 있어서, Constant 방법에 비해서는 실제 이상점인 관측값의 식별에 있어서 뛰어난 성능을 보여주었다.

<표 2> 이상점 탐지 기법별 정확도(가격변화율 변동성이 작은 경우)

		이상점 탐지 기법				
		Quartile	Tukey	Constant	Variable	Oracle
$m = 100$ $n = 300$	<i>Case 1</i>	0.914	0.906	0.914	0.914	0.914
	<i>Case 2</i>	0.897	0.888	0.898	0.928	0.940
$m = 300$ $n = 300$	<i>Case 1</i>	0.913	0.908	0.914	0.913	0.912
	<i>Case 2</i>	0.902	0.895	0.901	0.933	0.940

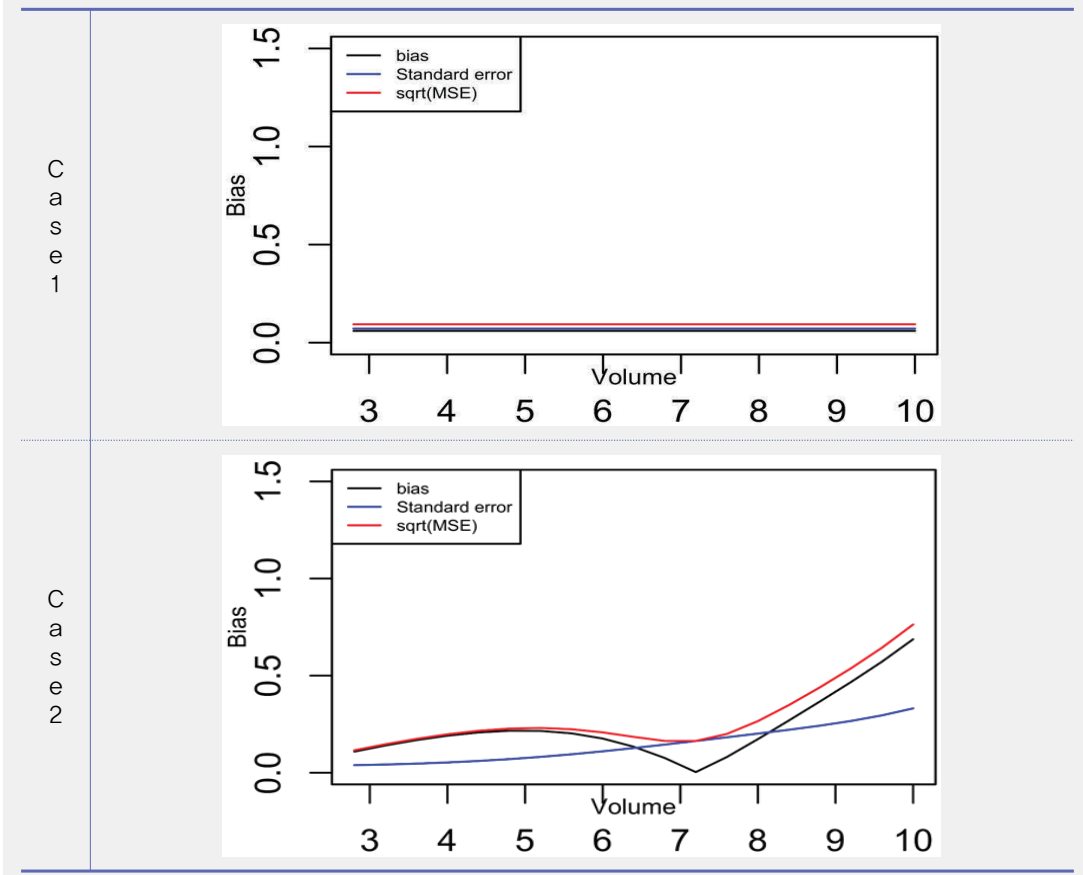
<그림 3>

이상점 탐지 기법별 민감도와 특이도 비교
(가격변화율 변동성이 작은 경우)



마지막으로 본 연구에서 제시한 분산함수 추정식 식 (2.7)의 정확성을 살펴보기 위해 분산함수 추정량의 편의와 분산, 그리고 평균제곱오차(MSE)를 계산하여 <그림 4>에 나타내 보았다. Case 1의 경우를 보면 V_t 에 관계없이 편의, 표준오차, 평균제곱오차가 일정하게 나타나는 것을 알 수 있는데, 이는 식 (2.7)의 커널 추정에서 Bandwidth h 가 매우 큰 값을 가져, 분산함수를 상수로 추정하기 때문이다. 다음으로 Case 2의 경우에는 Case 1과 달리 V_t 에 따라 값이 달라지는 것을 알 수 있는데 V_t 가 커질수록 표준오차는 증가하며, 편의는 V_t 가 7일 때 가장 작고 이후 급격하게 증가하는 것으로 나타났다. 그에 따라 평균제곱오차 역시 V_t 가 7일 때 낮고 이후 증가하는 것을 확인할 수 있다.

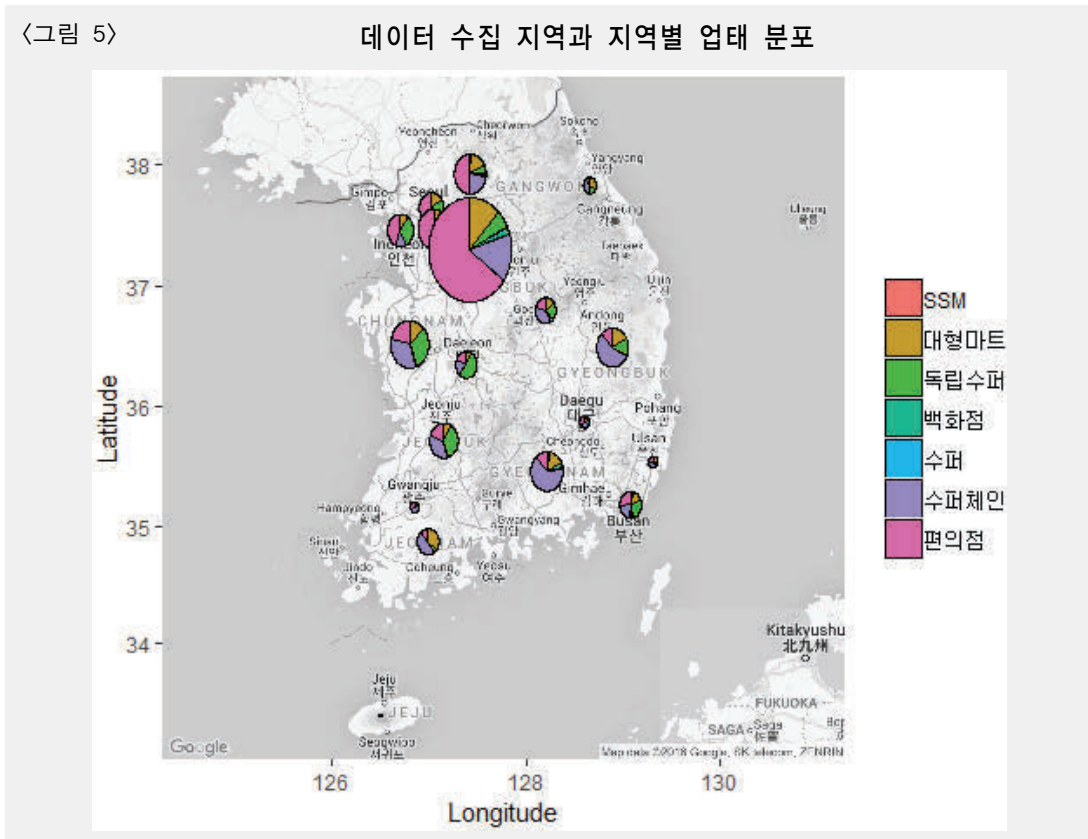
<그림 4> 분산 추정량의 평균제곱오차(가격변화율 변동성이 작은 경우)



IV. 실증분석

1. 데이터 소개

이 절에서는 본고에서 제안된 방법을 실제 스캐너 데이터에 적용하여 가격변화율에 대한 이상점을 탐지해본다. 실증분석에 사용된 스캐너 데이터는 대한상공회의소에서 2013년부터 2017년까지 전국 약 2천 개 매장으로부터 수집되었으며 20만 개 이상 품목의 주별 거래량, 거래금액 등의 정보가 포함되어 있다. 매장의 종류는 <그림 5>에서와 같이 SSM, 대형마트, 독립슈퍼, 백화점, 편의점 등 7개 업태로 나누어져 있다. 본 연구에서는 우유, 맥주, 아이스크림, 라면 등의 품목에서 일부 상품을 선택하여 이상점을 식별해 본다.



2. 이상점 탐지 결과

가. 우유

먼저 기존 방법들과 새롭게 제안된 방법을 적용하여 우유 제품 A의 거래가격에 대한 이상점을 식별하고 그 결과를 비교해 본다. 새롭게 제안된 Variable 방법을 적용하기 위해서 우유 제품 A의 2013년 거래가격 데이터로부터 분산함수를 추정하였다. 단, 업태에 따라 품목의 거래가격 산포가 다르게 나타날 수 있으므로²⁾ 업태별로 분산함수를 추정하였다. 이를 기반으로 2014년 중 주별 거래가격 데이터에 대한 이상점 여부를 온라인 모니터링³⁾ 해 보았다. 가격변동의 빈도가 낮을 경우 이상점 탐지 기법들간의 차이를 확인하는 것이 쉽지 않을 수 있기 때문에, 가격 변화가 상대적으로 빈번하게 발생하였던 업체⁴⁾의 거래가격을 대상으로 실증분석을 진행하였다.

1) 가격변화율의 변동성이 큰 경우에 대한 이상점 모니터링

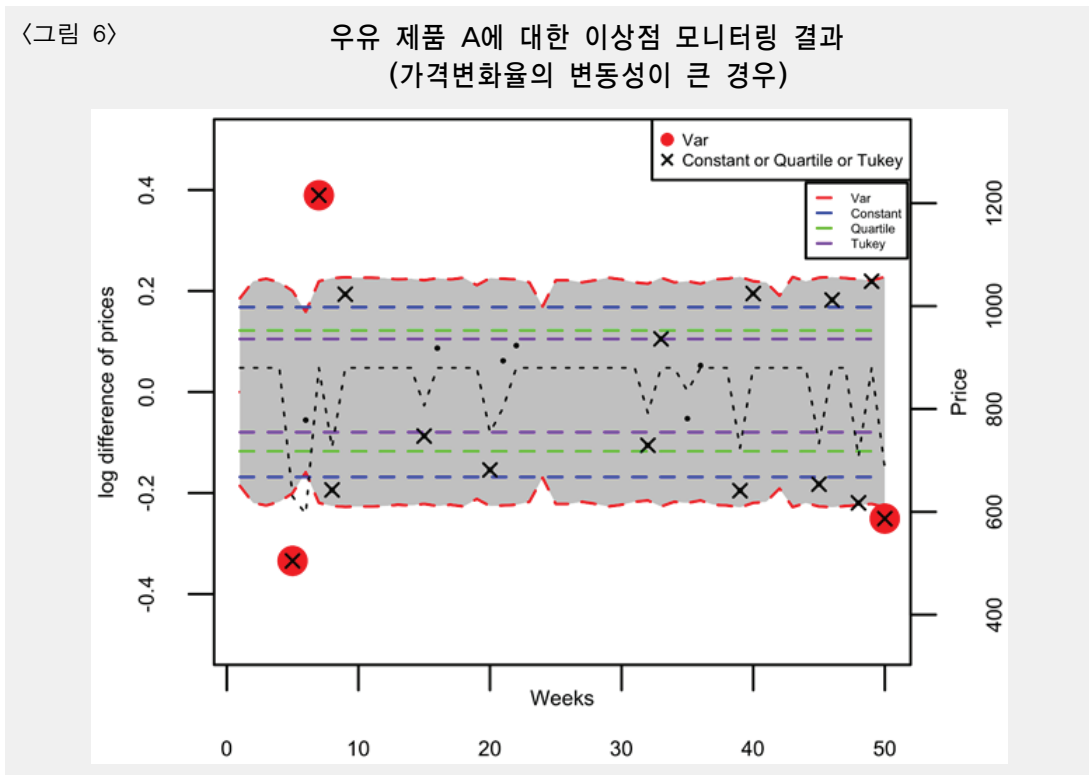
<그림 6>은 기존 방법들(Tukey, Quartile)과 분산함수를 상수로 추정한 Constant 방법, 그리고 분산함수를 이용해 추정한 Variable 방법들로부터 우유 A의 로그 가격변화율과 이상점을 나타낸 관리도이다. 가로축은 데이터가 수집되는 시점을, 세로축은 로그 가격변화율을 나타낸다. 관리도에서 검은색 점선은 매주 판매된 우유의 평균가격을, 붉은 선은 본고에서 제안된 Variable 방법으로 구한 관리 한계선을 나타낸다. 또한, •은 가격 변화가 있는 경우 이전 시점 대비 로그 가격변화율을 나타내고, ◯은 Variable 방법을 통해 탐지한 이상점을 나타내며, ×는 기존 방법들(Tukey, Quartile)과 Constant 방법을 통해 탐지한 이상점을 가리킨다.

관리도를 살펴보면 모의실험결과에서 알 수 있는 바와 같이 Variable 방법은 관리 한계선이 더 넓은 폭을 가져 이상점을 식별하는 데 있어 보수적인 경향을 보였다. 이에 따라 기존 연구로 찾은 이상점은 15개, Variable 방법으로 찾은 이상점은 3개로 나타났다. 구체적으로는 먼저 모든 방법이 이상점으로 판단한 6주차와 7주차의 경우 거래가격이 850원에서

2) 예를 들어 편의점이나 백화점의 경우 거래가격과 거래량 사이에 별다른 상관관계가 없는 것으로 관찰되었다.
3) 기존 데이터로부터 얻은 정보를 사용하여 새롭게 추가되는 데이터에 대해 실시간 모니터링하는 방법을 말한다.
4) 서울 남부의 한 슈퍼체인과 경상남도의 한 백화점을 대상으로 하였다.

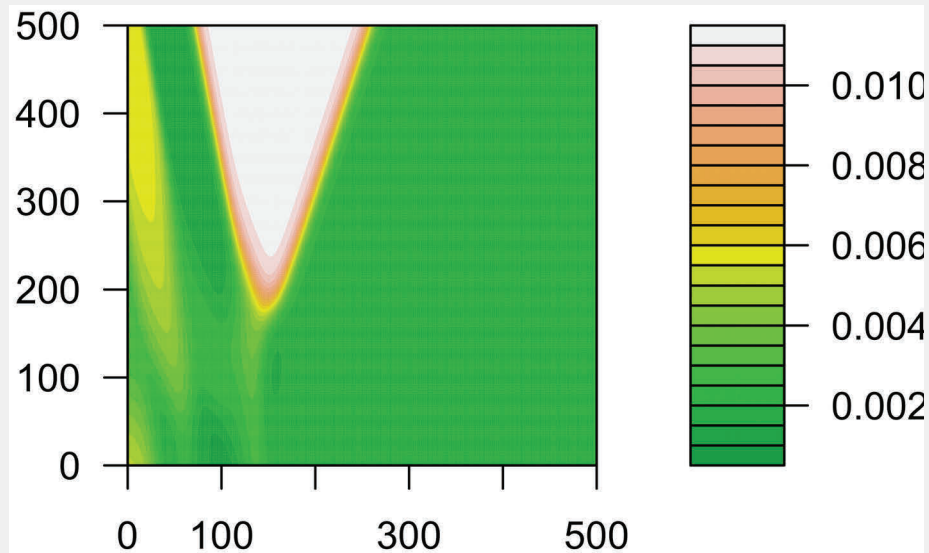
600원으로 일시적으로 하락한 후 다시 850원으로 상승한 시점에 해당한다.

기존 방법들로는 이상점으로 식별되었지만 Variable 방법으로는 이상점이 아닌 것으로 판별된 점들의 경우 약 100원 내외로 비슷한 폭의 가격하락과 상승이 거듭되어 나타나는 점을 고려하였을 때 일시적으로 되풀이되는 가격할인 행사에 따른 가격 변동으로 추정해 볼 수 있다. 이와 같이, 본고에서 제시한 방법이 실제로도 이상점 식별에 있어서 더 좋은 성능을 보임을 알 수 있다.



<그림 7>은 우유 제품 A의 분산함수를 그린 등고선 그래프다. 가로축은 $(t-1)$ 시점에서의 거래량, 세로축은 t 시점에서의 거래량이며, 색이 진할수록 분산이 작고 색이 밝을수록 분산이 큰 것을 나타낸다. 전 시점의 거래량이 100~200개에서 현시점의 거래량이 200개 이상으로 증가했을 때 가격이 상대적으로 크게 변화하여 가격변화율의 분산이 크게 추정되었다.

〈그림 7〉 우유 제품 A의 로그 가격변화율에 대한 분산추정량
(가격변화율의 변동성이 큰 경우)



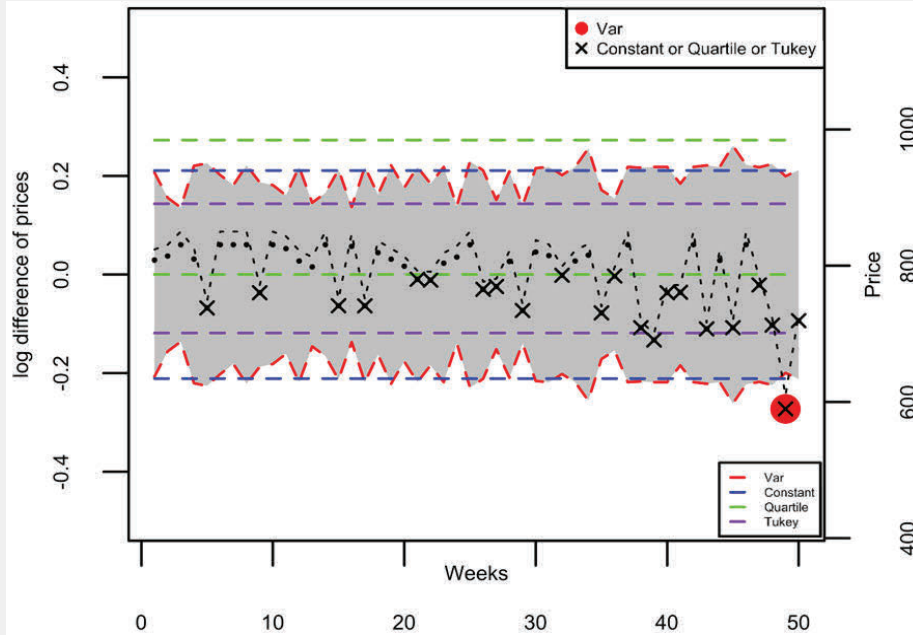
2) 가격변화율의 변동성이 작은 경우에 대한 이상점 모니터링

〈그림 8〉의 관리도에서 확인할 수 있는 바와 같이 가격변화율의 변동성이 작은 경우에도 분산함수를 적용해 추정한 Variable 방법이 기존 이상점 식별 방법들(Tukey, Quartile)과 분산함수를 상수로 추정한 Constant 방법에 비해 이상점을 적게 찾는 것을 알 수 있다. 기존 방법들로부터는 22개의 이상점이 식별된 반면, Variable 방법으로는 1개의 이상점만이 식별되었다. 〈그림 8〉에서 Quartile 방법의 관리 하한선이 0으로 나타났는데, 이는 관리한계선을 계산하기 위해 사용한 2013년 데이터의 제 1사분위수(Q_1)와 제 2사분위수(Q_2)가 모두 0으로 계산되었기 때문이다.

〈그림 9〉는 우유 제품 A의 분산추정량에 대한 그래프로, 가로축은 t 시점에서의 거래량, 세로축은 분산추정량을 나타낸다. 판매량에 따라 분산추정량의 변화가 발생하며, 판매량이 약 50개일 때와 약 200개일 때 가격 변동폭이 크게 추정되었다. 판매량이 320개 이상일 때부터는 분산추정량이 급격히 커지며 최대 판매량이 400개를 넘지 않기 때문에 약 400개 이후부터는 분산추정량이 유지되는 것으로 나타났다.

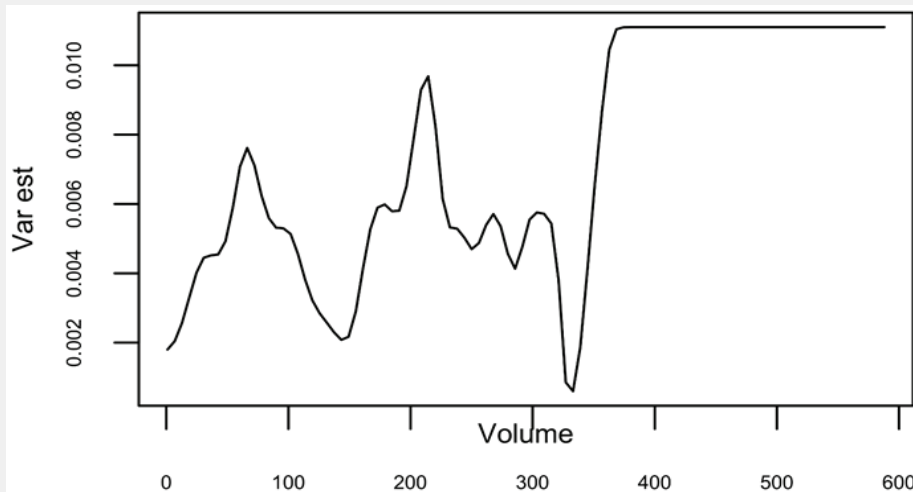
<그림 8>

우유 제품 A에 대한 이상점 모니터링 결과
(가격변화율의 변동성이 작은 경우)



<그림 9>

우유 제품 A의 로그 가격변화율에 대한 분산추정량
(가격변화율의 변동성이 작은 경우)



나. 기타 품목

가격 변동성은 품목의 특성에 따라서도 다르게 나타날 수 있다. 예를 들어 아이스크림의 경우 계절에 따라 판매량과 가격이 영향을 받을 수 있다. 반대로 계절적인 요인에 가격이 영향을 받지 않는 경우도 많다. 계절적인 요인에 의해 가격 변동폭이 클 것으로 예상되는 품목으로 아이스크림과 맥주, 계절적인 요인에 영향을 많이 받지 않을 것으로 예상되는 품목으로 라면을 선정하여 이상점을 탐지해 보고, 기존 방법과 새롭게 제안한 방법에 의한 이상점 식별 결과를 비교해 보았다.

1) 가격 변동성이 큰 경우 : 아이스크림, 맥주

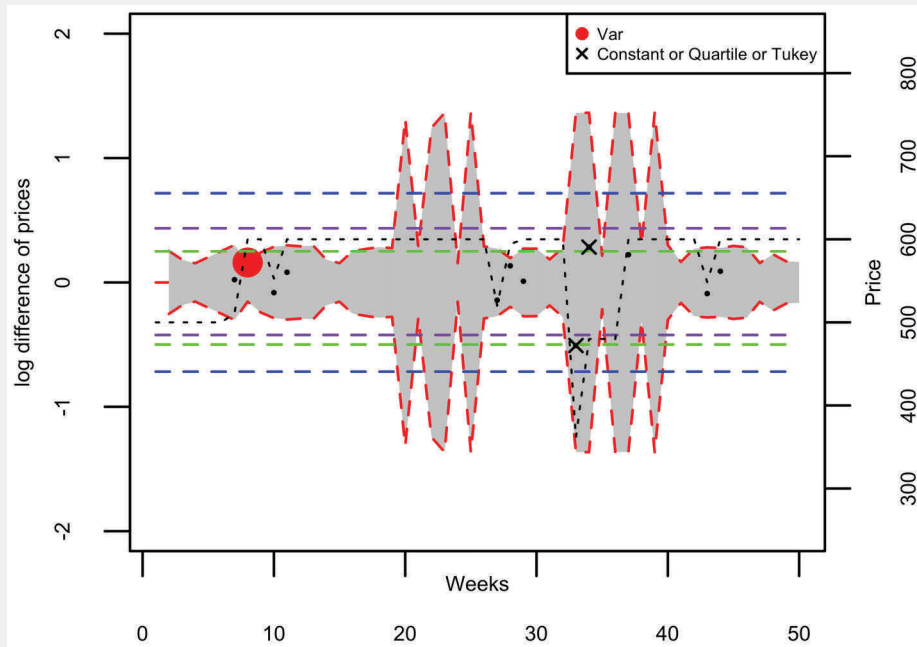
<그림 10>은 아이스크림 제품 B의 거래가격 데이터⁵⁾에 대한 관리도이다. Variable 방법은 8주차의 거래가격을 이상점으로 판단했고, 기존 방법들은 이상점으로 판단하지 않았다. 이 시점은 가격이 500원에서 600원으로 상승하였고, 이후 가격이 계속 600원으로 유지된 점에서 짐작할 수 있듯이 기존 판매가격이 상승한 시점을 정확히 찾아낸 것으로 볼 수 있다.

33주차와 34주차에는 기존 방법들과 달리 본 연구에서 제안한 Variable 방법만 이상점이 아닌 것으로 판단했다. 가격은 600원에서 350원으로 하락하였고, 거래수량은 188개, 221개로 증가하여 이 업체의 평균 거래량 약 112개에 비해 상대적으로 많은 거래가 일어난 것을 확인할 수 있다. 이후 다시 가격이 회복되었음을 감안할 때 당시의 가격 하락은 일시적인 가격할인 행사에 따른 것으로 추정해 볼 수 있다. 본 연구에서 제안한 Variable 방법은 거래 수량이 증가함에 따라 가격의 변동성이 크게 추정되어 관리 한계선의 폭이 넓게 설정되기 때문에 다른 이상점 식별 방법들과 달리 위와 같은 관측값을 이상점이 아닌 것으로 판별하였다. 이와 같은 현상은 <그림 11>의 분산추정량에 대한 등고선 그림에서도 확인할 수 있다. 전 시점과 현 시점의 거래량이 약 200개 부분에서, 분산 추정량의 크기가 크고, 따라서 관리 상한과 관리 하한이 넓게 정해지는 것을 확인할 수 있다.

5) 경상남도의 SSM에서 수집된 가격 데이터를 사용하였다.

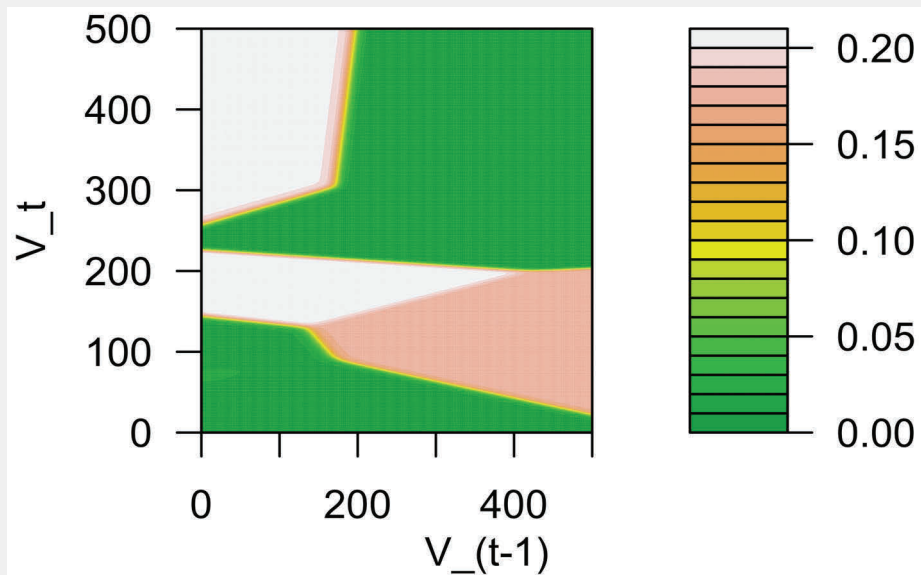
<그림 10>

아이스크림 제품 B에 대한 이상점 모니터링 결과



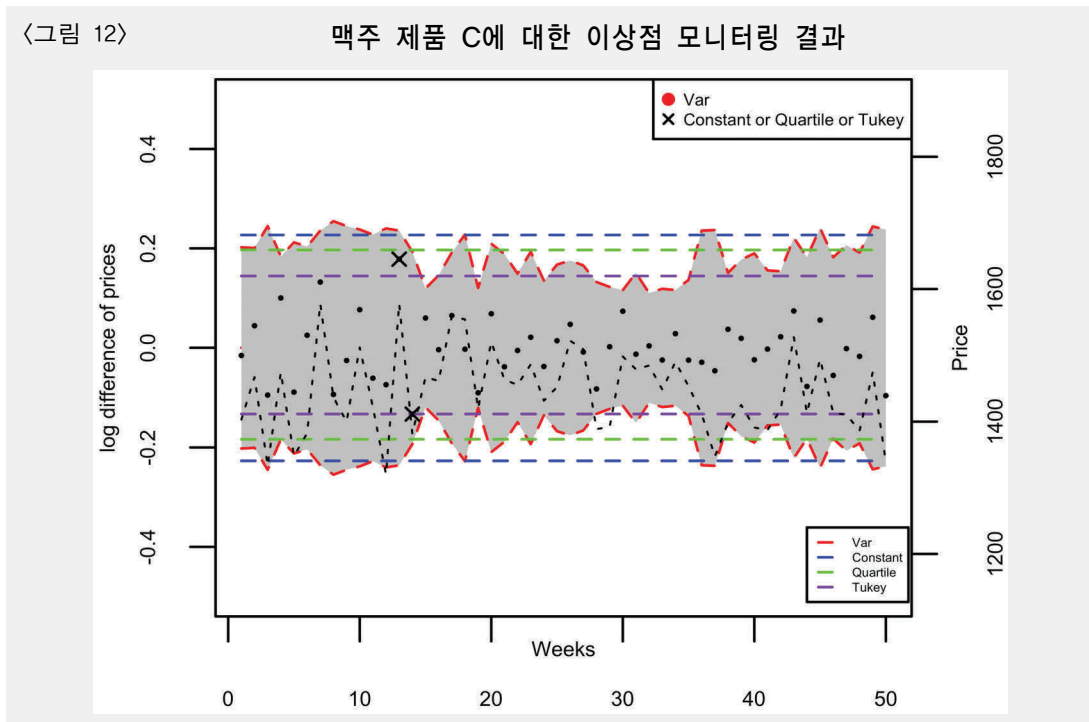
<그림 11>

아이스크림 제품 B의 로그 가격변화율에 대한 분산추정량



<그림 12>는 맥주 제품 C의 거래 데이터⁶⁾에 대한 관리도로 기존 방법에 의해서만 이상점 2개가 식별되었음을 볼 수 있다. 맥주 제품 C의 경우 가격 변화가 매우 빈번하게 일어나고 있지만, 대체로 1400~1500원 사이에서 오르내리고 있어 이러한 가격변화를 이상점으로 보기 어렵다는 것을 확인할 수 있다. 이상점으로 식별된 13주차와 14주차의 가격도 기존 방법 중 상대적으로 많은 관측값들을 이상점으로 판정하여 특이도가 높은 방법인 Tukey 방법에 의해서만 이상점으로 식별되는 것을 확인할 수 있다.

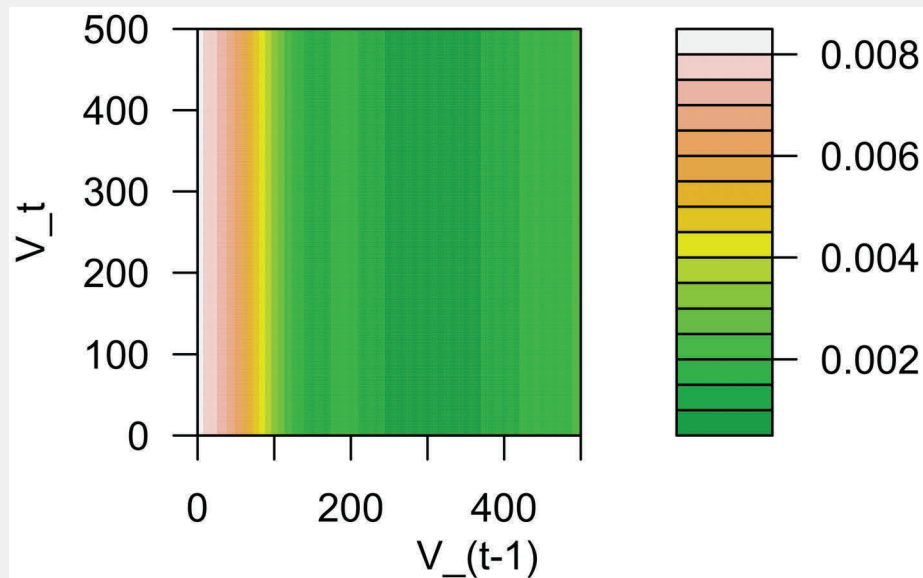
<그림 13>은 맥주 제품 C의 분산 추정량을 그린 등고선 그림으로, 전 시점의 거래량이 100개 이하로 작은 시점에서 분산을 크게 추정하는 것을 확인할 수 있다.



6) 경상북도의 편의점에서 판매된 맥주의 거래 데이터를 사용하였다.

<그림 13>

맥주 제품 C의 로그 가격변화율에 대한 분산 추정량



2) 가격변동성이 작은 경우 : 라면

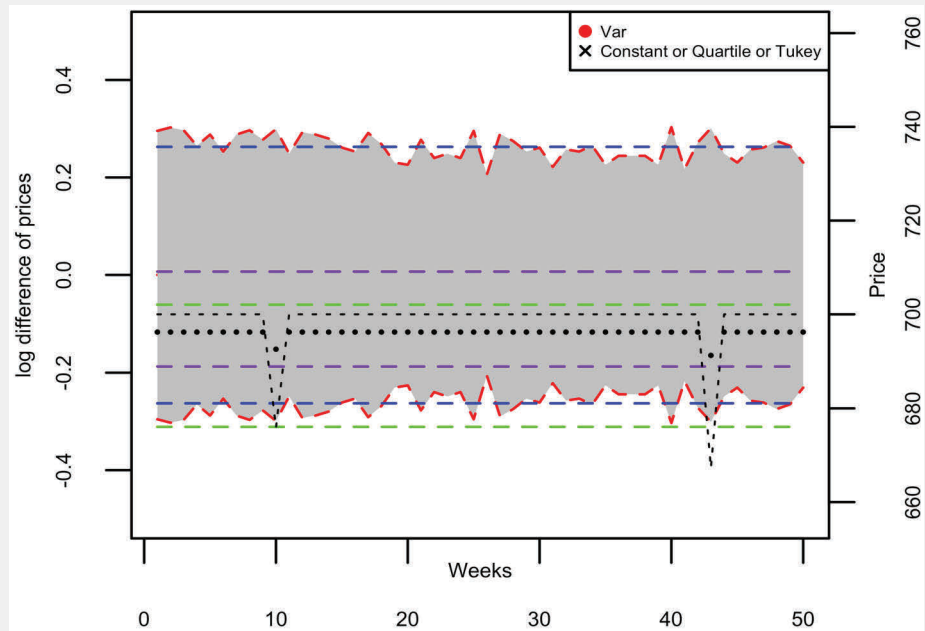
<그림 14>는 라면 제품 D의 거래가격 데이터⁷⁾에 대한 관리도이다. 라면 가격은 큰 변화 없이 대부분 거의 일정한 수준을 유지하였고, 10주차와 43주차에만 가격이 20원 정도 하락한 것을 확인할 수 있다. 이처럼 가격 변동이 거의 없는 경우에는 표준가격을 기준으로 로그 가격변화율을 이용해 이상점 탐지를 시행하는 것이 바람직하다. 거의 가격 변동이 없으므로, 모든 방법들이 관측값들을 이상점이 아닌 것으로 판별하였지만, 보라색 점선으로 표시된 Tukey 방법의 관리하한을 고려하였을 때, 조금만 더 가격 변화가 있었다면 이상점으로 판별되었을 가능성이 있음을 확인할 수 있다.

<그림 15>의 분산 추정량을 보면, 거래량이 약 50개 정도일 때 가장 큰 분산을 가지며, 그보다 거래량이 늘어나거나 줄어들수록 분산 추정량이 작아지는 것을 확인할 수 있다. 200개 이후로는 큰 분산값이 유지되는데, 이는 거래량이 200개가 넘는 자료가 거의 없어 정확한 추정이 이루어지지 않았기 때문일 것으로 추정된다.

7) 울산광역시시의 한 슈퍼체인에서 판매된 라면의 거래 데이터를 사용하였다.

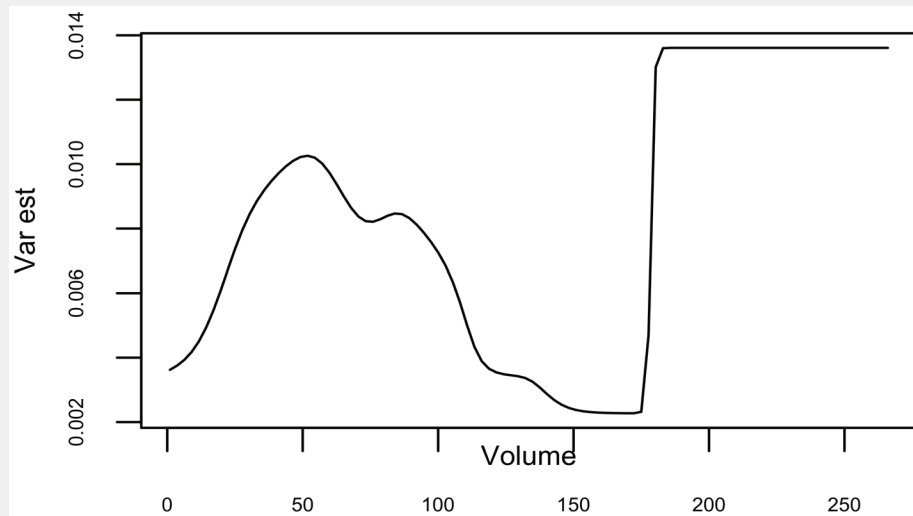
〈그림 14〉

라면 제품 D에 대한 이상점 모니터링 결과



〈그림 15〉

라면 제품 D의 로그 가격변화율에 대한 분산 추정량



V. 결론

본고에서는 스캐너 데이터의 거래가격 변화율 모니터링을 통해 거래가격의 이상점을 효율적으로 탐지하기 위한 절차를 제안하였다. 특히, 기존의 모니터링 방법들은 스캐너 데이터의 가격 정보만 사용할 뿐 가격에 상당한 영향을 줄 것이라 생각되는 판매 수량에 대한 정보는 고려하고 있지 않아 이를 개선한 절차를 제시하였다. 새로이 제안된 모니터링 방법은 거래가격과 판매량 사이의 연관성에 기반하여 거래가격의 분산을 판매량의 함수로 추정하였다. 분산함수의 추정을 위해서는 여러 추정방법이 제안되어왔는데 본 연구에서는 가장 보편적으로 사용되는 비모수 커널 회귀방법을 적용하였다. 또한, 이러한 분산함수의 추정식을 이용하여 판매량을 고려하였을 때 거래가격의 이상점 여부를 판단하는 온라인 모니터링 절차를 고안하였다.

모의실험 결과를 통해 판매량을 고려하여 가격변화율에 대한 산포를 추정하는 새로 제안된 방법이 기존의 이상점 탐지 기법들에 비해 높은 정확도를 지니고 있음을 알 수 있었으며 특히 특이도가 높은 편으로 나타났다. 이러한 특성은 거래량이 크게 증가하거나 감소할 때 거래가격의 하락이나 상승폭이 더 크게 나타날 수 있음을 의미하며, 판매량 변화가 큰 시점에는 이상점을 더 보수적으로 판단하게 된다. 실증분석을 통해서도 가격 변동의 특성이 다른 다양한 품목들의 거래가격 데이터를 사용하여 판매가격의 이상점을 식별해보았으며 새롭게 제안된 이상점 식별 방법을 통해 효과적으로 이상점을 탐지할 수 있음을 확인할 수 있었다.

참고문헌

- Abe, N. & Tonogi, A., “Micro and Macro price Dynamics in Daily Data”, *Journal of Monetary Economics*, 57(6), 2010, 716-728.
- Eurostat, “Practical Guide for Processing Supermarket Scanner Data”, 2017.
- International Labour Office, “Consumer Price Index Manual: Theory and Practice”, 2004.
- Mayhew, M., “A Comparison of Index Number Methodology Used on UK Web Scraped Price Data”, ONS methodology working paper series number 12. 2017.
- Office for National Statistics, “Consumer Price Indices Technical Manual”, 2010.
- Office for National Statistics, “Research Indices Using Web Scraped Price Data”, 2016.
- Rais, S., “Outlier Detection for Statistics Canada’s Consumer Price Index”, Business Survey Methods Division, Statistics Canada, 2008.
- Saïdi, A. & Rubin-Bleuer, S., “Detection of Outliers in the Canadian Consumer Price Index”, Business Survey Methods Division, Statistics Canada, 2005.
- Thompson, K., & Sigman, S., “Statistical Methods for Developing Ratio Edit Tolerances for Economic Data”, *Journal of Official Statistics*, Vol. 15, No 4, 1999.