서울시민카드 사용자 유입과 이탈에 대한 시계열 이상치 탐지와 DTW 클러스터링

김경희[a]

성균관대학교

요 약

본 연구에서는 성별과 연령대에 따라 서울시민카드 이용자의 유입과 이탈에 대해 분석한다. 먼저 ARMA 모형 (Auto Regressive Moving Average Model) 적합에서 얻은 추정 값를 활용해 자료에 대한 단변량 시계열 이상치 탐지 (outlier detection)을 시행한다. 이후 DTW (Dynamic Time Warping) 클러스터링을 통해 묶인 군집을 파악한다. 성별과 연령에 따른 서울시민카드 주별 이용 빈도에 대한 자료의 이상치 탐지 결과와 DTW 클러스터링 결과를 기반으로 이용자의 유입 및 탈퇴에 대한 이해를 높여 서울시민카드 사용 발전에 도움이 되고자 한다.

Keywords: 서울시민카드, ARMA 모형, 이상치 탐지, DTW 클러스터링

1. 서론

서울시민카드는 서울도서관, 서울시립미술관 등의 시립시설과 문화, 체육센터 등의 공공시설을 여러 회원가입을 거치지 않고 하나의 통합 모바일 카드로 이용할 수 있게 하는 취지로 만들어진 서울시 공공 앱이다 (김소연, 2017). 이 카드는 공공시설의 위치, 운영정보나 공연, 전시 등의 관람 정보 등을 제공하며, 모바일 쿠폰과 민간제휴 할인혜택 제공 및 공공시설 예약 신청까지 다양한 서비스를 포함한다. 따라서 서울시민카드는 서울시의 여러 시설 이용에 필요한 다양한 정보를 빠르고 편리하게 알 수

[a] 03063, 서울. 종로구, 성균관로 25-2, 성균관대학교 응용통계연구소, 연구원, 통계학과, 석사과정, kkh97122647@gmail.com

있으며 시설들에서 제공하는 이벤트들에 손쉽게 참여할 수 있다는 장점이 있다. 또한 최근 서울시 개인 인증 수단과 결제 수단(네이버페이, 제로페이 등)의 기능이 추가되어 서울시 내에서 전자지갑의 역할도 가능하게 되었다. 더 나아가 서울시는 서울시민카드에 추가적인 공공시설 연계 확대, 서비스 혁신, 혜택 증가를 통해 다양한 공공서비스를 시민들이 편리하게 제공받도록 하여 이용률을 높이겠다는 목표를 밝혔다 (김현정, 2019).

본 연구는 서울시민카드 일별 성별/연령별 이용현황정보를 시계열 이상치 탐지(time series outlier detection)방법과 DTW (Dynamic Time Warpping)클러스터링 방법에 적용해 분석한다. 이상치 탐지 결과를 바탕으로 이용자들의 앱 접속 빈도의 급격한 변화가 있는 시점들을 파악하고, DTW 클러스터링을 통해 유사 시계열 패턴을 보이는 이용자들을 군집으로 묶은 후 각 군집 특성을 알아본다. 본 분석에서 얻은 정보를 차후 이용가능 대상자들의 지속적인 유입과 이용 빈도를 높이는 데 활용해 서울시민카드 공공 앱의 목표를 달성하는 데 도움이 되고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 이상치 종류를 정의하고 이상치 탐지 절차와 함께 단변량 시계열 이상치 탐지를 소개한다, 다음으로 DTW 거리 척도를 소개하고 DTW 거리에 기반한 클러스터링을 알아본다. 3장에서는 서울시민카드 일별 접속 데이터에 대해 2장에서 소개한 시계열 이상치 탐지 방법을 적용하여 이상치를 탐지하고 DTW 클러스터링을 적용한 후 분석 결과에 대해 설명한다. 마지막으로 4장에서는 본 논문의 결론과 시사점에 대해 다룬다.

2. 방법론

2.1. 시계열 이상치 탐지

시계열 이상치는 특정 시점에만 영향을 주는 점 이상치 (point outlier)와 특정 시점 이후 지속적인 영향을 주는 구간 이상치 (subsequence outlier)로 구분된다 (Ane Blázquez-García, 2021). 점 이상치에는 AO (additive outlier)가 있고, 연속 이상치에는 TC (temporary outlier)와 LS (level shift)가 있다. 그림1은 AO, TC, LS가 발생한 시계열 형태를 개괄적으로 보여준다.

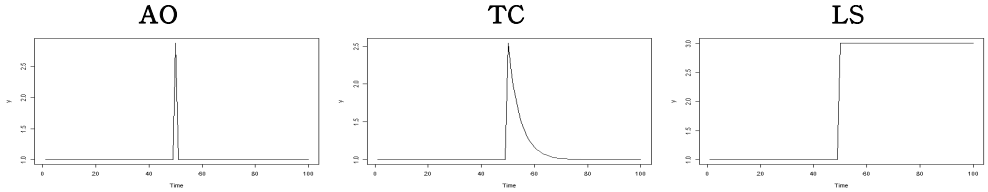

그림1. 이상치 그래프

t는 시간을 의미하고 d는 임의의 특정 시점을 뜻할 때, AO는 t=d 시점에서만 이상점 효과 (outlier effect)를 보인다. TC는 t=d에서 이상점의 효과가 나타나고 해당 시점 이후 서서히 이상점 효과가 사라지는 형태를 보인다. LS는 t=d에서 이상점 효과를 보이고 해당 시점 이후에도 효과가 유지되는 형태를 보인다.

본 연구에선 Tsay (1988)가 제안한 시계열 이상치 탐지 방법론을 적용한다. 이 방법은 Box and Jenkins (1976)의 ARMA (Autoregressive Moving Average)모형 적합을 기반으로 시행된다. 적합한 식에서 이상점 효과 ω_0와 이상점 효과의 분산 var(ω_0)을 추정하여 이상치 종류에 따른 검정 통계량을 계산한다. 이후 계산된 검정통계량이 임계값 보다 클 때 이상치라고 판단한다.

먼저 관찰된 시계열을 $\{Y_t\}$라 할 때, 이상점이 있는 시계열은 다음과 같이 모델링할 수 있다 (Tsay, 1988).

$$Y_t = \omega_0 \frac{A(B)}{G(B)H(B)} \xi_t^{(d)} + Z_t,\ Z_t \sim WN(0,\sigma^2) \tag{2.1}$$

여기서 B는 $BZ_t = Z_{t-1}$로 자료를 한 시점을 뒤로 옮기는 효과를 내는 후향 연산자 (backshift operator)다. $\xi_t^{(d)}$는 t=d 시점에 이상치가 존재하면 1의 값을 가지고 존재하지 않으면 0의 값을 가지는 지시함수 (indicator function)이며, ω_0는 이상점 효과 크기 값이고 $\frac{A(B)}{G(B)H(B)}$는 이상점 효과의 패턴을 나타낸다. t=d 시점에서의 ω_0와 $\frac{A(B)}{G(B)H(B)}$는 이상치 종류에 따라 아래 식 (2.2)와 같이 가정한다 (이경민, 2021).

$$\begin{aligned}
&\text{AO: } \omega_{0,d} = \omega_{AO,d},\ \frac{A(B)}{G(B)H(B)} = 1 \\
&\text{TC: } \omega_{0,d} = \omega_{TC,d},\ \frac{A(B)}{G(B)H(B)} = \frac{1}{(1-\delta B)} \text{ 단, } 0 < \delta < 1 \\
&\text{LS: } \omega_{0,d} = \omega_{LS,d},\ \frac{A(B)}{G(B)H(B)} = \frac{1}{(1-\delta B)} \text{ 단, } \delta = 1
\end{aligned} \tag{2.2}$$

TC와 LS의 $\frac{A(B)}{G(B)H(B)}$에 포함된 초매개모수 (hyperparameter) δ는 TC의 경우 0에서 1 사이 값을 가지고 LS의 경우 $\delta = 1$의 값을 가진다. δ는 이상점 효과 지속 기간을 조정하는 역할을 하며 δ가 1에 가까울수록 이상점이 발생한 이후의 기간에도 지속적인 이상점 효과를 주고 0에 가까울수록 한 시점에만 영향을 미침을 의미한다.

다음으로 식 (2.1)을 Chang (1982)와 Hillmer et al. (1983)에 따라 새로운 모수 x_t, y_t로 정의하면

$$y_t = \omega_0 x_t + a_{t,},\ a_t \sim WN(0,\sigma_a^2)$$
$$\text{단, } y_t = \frac{\Phi(B)}{\theta(B)} Y_{t,}\ x_t = \frac{\Phi(B)}{\theta(B)} \frac{\omega(B)}{\delta(B)} \zeta_t^{(d)},\ a_t = \frac{\Phi(B)}{\theta(B)} Z_t \quad (2.3)$$

식 (2.3)로 나타낼 수 있다. 이 모형은 상수항이 없는 일차 선형회귀로 최소제곱법을 통해 t=d 시점에서 이상점 효과 ω_0를 추정한 값은 식 (2.4)과 같다.

$$\hat{\omega}_{i,d} = \frac{\sum_{t=d}^{n} y_t x_{i,t}}{\sum_{t=d}^{n} x_t^2},\ Var(\hat{\omega}_{i,d}) = \frac{\sigma_a^2}{\sum_{t=d}^{n} x_{i,t}^2},\ \text{n: 시계열 길이} \quad (2.4)$$

Tsay (1988)이 제안한 이상치 탐지 검정 통계량은 t=d 시점에서 $\lambda_{i,d} = \frac{\hat{\omega}_{i,d}}{\sqrt{Var(\hat{\omega}_{i,d})}}$, i = AO, TC, LS이다. 최종적으로 이상점 탐지의 절차는 다음과 같이 구성된다.

Step 1. 관찰된 자료 Y_t에 대해 ARMA를 적합하고 선형회귀를 사용해 이상치 종류에 따라 $\frac{A(B)}{G(B)H(B)}$를 설정한 후 시계열의 모든 시점에 대해 계수의 $\omega_{i,d}$ 추정값과 분산을 계산한다.

Step 2. 주어진 임계값 C에 대해 $\max\{\lambda_{i.d} | i \in \{AO,TC,LS\}\} > C$ 를 만족하는 i가 있으면 t=d 시점엔 이상치가 존재하며 이때의 i가 t = d의 시점의 이상점 종류라고 판단할 수 있다. 임계값 C는 Tsay (1988)에서 제시한 table 1.에 따라 설정하였다.

2.2. 동적 타임 워핑(DTW: Dynamic Time Warping) 클러스터링

클러스터링 (Clustering)은 관측치들의 유사도를 고려하여 군집을 나누는 분석이다. 클러스터링은 크게 계층적 군집화 (hierarchical clustering)와 비계층적 군집화 (nonhierarchical clustering)으로 나뉜다. 계층적 군집은 처음에 n개의 군집으로부터 시작하여 점차 군집의 개수를 줄여나가는 방법이다. 비계층적 군집은 분할적 군집으로도 불리는데 이는 군집의 계층을 고려하지 않고 평면적으로 군집을 형성하는 방법이다 (김재희, 2011).

그림2. 유클리디안 거리와 DTW 거리 짝 대응 비교

클러스터링에서 관측치의 유사도를 측정하기 위해 다양한 거리 척도를 활용하는데, 자료의 형태가 시계열인 경우 유클리디안 거리 (ED; Euclidean Distance)와 동적 타임워핑 거리 (DTW distance; Dynamic Time Warpping distance)를 사용해 클러스터링 하는 방법이 있다. ED의 경우 순차적으로 시계열들 간 짝을 대응시켜 거리를 계산하는 반면, DTW는 이웃시점들의 거리들 중 가장 가까운 시점과 짝을 연결해 거리를 계산한다. 짝을 대응시키는 모습을 Keogh, E., Ratanamahatana, C (2005)에서 그림2와 같이 표현하였다.

두 시계열 $\boldsymbol{X} = (x_1, x_2, \ldots, x_m)$, $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)$이 주어졌을 때, 유클리디안 거리는 두 시계열의 길이가 같다는 n=m 가정 하에 식 (2.5)로 정의한다.

$$ED(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2.5}$$

유클리디안 거리는 계산이 간편하다는 장점이 있으나, 두 시계열의 길이가 동일해야 계산 가능하며 두 시계열 내의 시간 지연 (delay)나 평균 차 (shift)가 존재할 경우 큰 값을 도출해 두 시계열의 유사성을 유연하게 알아내는데 어려움이 있다.

동적 타임워핑 거리는 위 한계점들을 보완한 거리 측정 방법이다. DTW는 시계열 데이터에서 두 데이터의 유사성을 측정하는 거리 척도로 이웃 시점들의 값까지 비교하여 거리의 합이 가장 적은 경우를 거리로 측정한다 (Kate, 2015). 위 시계열 $\boldsymbol{X}$와 $\boldsymbol{Y}$가 주어졌을 때, DTW 거리는 아래와 같이 식 (2.6)으로 표현된다.

$$\begin{aligned} DTW(\boldsymbol{X}, \boldsymbol{Y}) &= D(m,n) \\ D(i,j) &= dist(x_i, y_j) + \min\begin{cases} D(i,j-1) \\ D(i-1,j) \\ D(i-1,j-1) \end{cases} \\ \text{단, } dist(x_i, y_j) &= |x_i - y_j|, \ \text{i=1,...,m, j=1,...,n} \end{aligned} \tag{2.6}$$

이때, $D(0,0) = 0$이고 0이 아닌 모든 i와 j에 대해 $D(i,0) = D(0,j)$의 값은 ∞으로 둔다. $DTW(\boldsymbol{X}, \boldsymbol{Y})$ 결과 값이 작을수록 유사도가 더 높다고 판단된다. DTW는 시계열

들의 길이가 다를 때 짝의 개수를 맞추지 않아도 계산 가능하다는 장점이 있다. 또한, 시계열 내에서 shift가 존재하는 경우 시계열들의 유클리디안 거리를 계산하면 시계열 패턴이 거의 유사함에도 큰 거리 값을 도출하는 단점이 있다. 이런 경우에 DTW 거리 기반 클러스터링을 하는 것이 더 적절하다고 알려져 있다.

3. 실증 자료 분석

3.1. 자료 설명 및 전처리

본 논문에서 분석하는 자료는 서울특별시에서 제공하는 서울시민카드 일별 성별/연령별 통계정보[a]이며, 서울특별시 서울시민카드 공공 앱 일별 접속 빈도가 연령대와 성별에 따라 구분하여 기입되어 있다. 본 연구에서는 20대부터 50대 연령층을 추출해 2020년 1월 23일부터 2021년 12월 22일까지 700일에 해당하는 데이터를 주별 평균 내어 사용하였다. 따라서, 성별에 따른 20대, 30대, 40대, 50대 접속 자료를 얻어 길이가 100인 8개의 시계열 데이터를 최종 생성하였다.

3.2. 기초통계분석

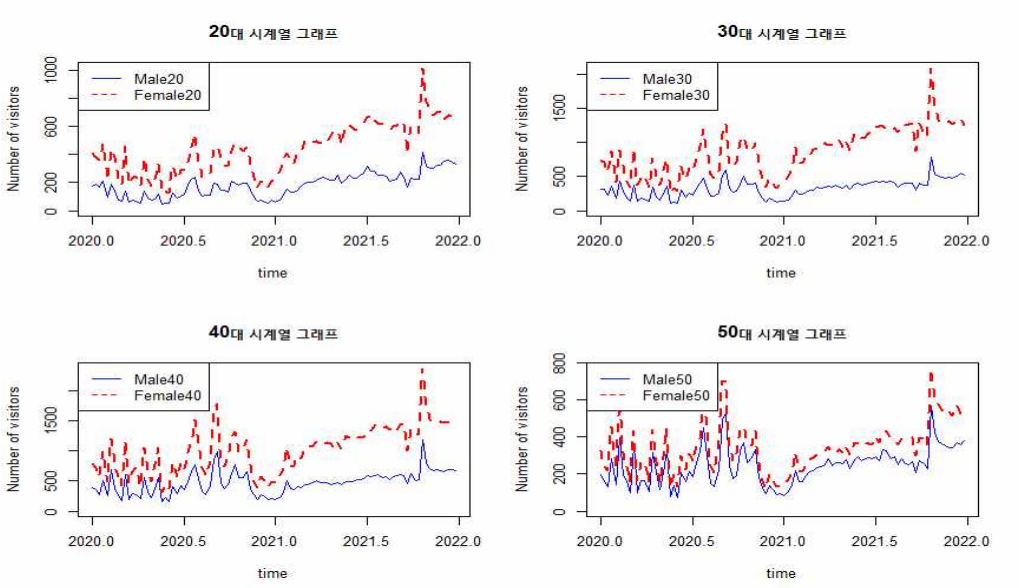

그림3: 성별과 연령대에 따른 시계열 그래프

[a] 데이터 출처: http://data.seoul.go.kr/dataList/OA-20886/S/1/datasetView.do

그림3의 왼쪽 위 그래프부터 순서대로 성별에 따른 20대, 30대, 40대, 50대의 시계열 그래프를 나타냈다. 가장 큰 특징으로는 모든 연령층에 대해 여자가 남자보다 이용 빈도가 높았다. 다음으로는 다양한 연령층 사이에서 30대와 40대가 가장 많은 접속 수를 보이고 있었다. 또한 모든 성별과 연령층에 대해 2020년도에는 접속 빈도가 고정적이지 않고 증가와 감소를 반복했으나 2021년에는 비교적 낮게 접속 빈도가 유지되다가 하반기부터 이용 양상의 변화를 보였다.

분기	20남	20여	30남	30여
Q1	106.70	276.39	232.91	553.41
Q2	142.95	335.58	307.51	729.90
Q3	193.77	478.51	318.88	901.42
Q4	280.50	651.98	449.18	1273.77
	40남	**40여**	**50남**	**50여**
Q1	342.90	688.78	188.50	278.31
Q2	474.98	915.90	242.19	336.07
Q3	435.22	1015.38	228.70	299.36
Q4	634.34	1450.36	318.45	459.06

표1: 25주기 성별과 연령층에 따른 평균 이용 빈도

표1은 2020년 1월 23일부터 25주 단위로 분기 별 성별과 연령 층에 따른 평균 이용 빈도를 나타낸 표이다. 구체적인 날짜는 Q1: 2020/1/23~2020/7/8, Q2: 2020/7/9~2020/12/30, Q3: 2020/12/31~2021/6/23, Q4: 2021/6/24~2021/12/22이다. 대부분의 시계열에서 이용 빈도가 시간이 흐름에 따라 증가하는 것으로 보이나, 40대 남성과 50대 남성, 여성의 경우 Q3에서 이용 감소세를 보인다. 그러나 Q3에서 감소세를 보인 연령층을 포함하여 모든 성별, 연령층이 Q4에서 Q2에 대비해 급증한 이용 빈도를 보인다.

3.3. 이상치 탐지 결과

Time	20남	20여	30남	30여	40남	40여	50남	50여
4					AO			
5	AO	AO						AO
6			AO		AO		AO	
10	AO	AO	AO	AO	AO			
15					AO			
19							AO	
20			TC		TC	TC		TC
29			AO		AO	AO	AO	AO
30	LS							
34			TC	TC	TC	TC	TC	TC
39			LS					
44			LS		LS			
54		LS		LS		LS	LS	
91	TC	TC	AO	AO	TC	TC	TC	TC

표2: 이상치 탐지 결과

표2에서 왼쪽 열에 명시된 time은 2020년 1월 23일-2020년 1월 29일이 1이며, 주 단위로 1씩 증가하게 설정하였다. 전체 시간 구간은 [1, 100]이며, 그 중 이상치가 발생했던 시점들만 표2에 반영해 작성하였다. δ는 일시적으로 영향을 줄 수 있도록 0.7로 설정하였다 (이경민, 2021).

표2의 초기 시점 4, 5, 6, 10, 15에서 40대 여성을 제외한 모든 성별과 연령층에서 AO 이상치들이 다수 발생했음을 볼 수 있다. 해당 시점들은 서울시민카드를 통한 현대아울렛 H.point 신규 가입 이벤트, 메가박스 영화 이벤트, GS25 음료 할인 등 여러 단기성 이벤트들의 빈도가 높았던 기간이다. 이 결과를 그림4와 함께 해석하면 대부분의 성별, 연령층에서 해당 이벤트들의 효과가 즉각적인 사용자 유입으로 반영되고 있음을 볼 수 있었다.

반면, 시점 20에서 30대 남성, 40대 남성과 여성, 50대 여성을 대상으로 TC가 발생했는데, 이 경우 그림4과 함께 해석하면 이용 빈도가 감소하는 이상치로 해석 가능하다. 이는, 초반에 진행하던 단기성 이벤트들이 종료되고 새로운 이벤트가 없어 이용자 이탈로 이어진 것으로 파악된다.

2020년 9월 10일-2020년 9월 16일인 시점 34을 포함해 그 이후로 TC, LS와 같이 장기적으로 해당 시점에서 이용자들의 어플리케이션 이용 빈도 변화가 지속되었다는 이상치들이 나타났다. 특히, 시점 34는 서울패스 개인 인증을 도입한 시기이며 시점 44는 제로페이를 서울시민카드에 추가하면서 결제기능도 추가되었다는 점을 고려하면 어플리케이션의 활용도가 다양해지면서 이용자들이 꾸준히 유입되었음을 볼 수 있다.

더 나아가 시점 91에 대해 알아보면 해당 시점엔 서울관광할인패스가 2021년 10월 18일부터 도입되었다. 서울관광할인패스란 서울의 주요 관광지 및 체험시설 121개소에서 최대 50% 이상 할인을 받을 수 있는 패스를 무료로 다운로드 가능하게 한 이용권이다. 시점 91은 전 연령에 대하여 이용빈도 증가를 보였다. 특히, AO인 30대를 제외하고 대부분의 연령층에서 서울관광할인패스의 효과를 통해 이용 빈도가 지속적으로 증가하는 양상을 보인 TC로 반영되었음을 볼 수 있다.

따라서 초기에 활발히 진행되었던 단기성 쿠폰 이벤트의 경우 즉각적인 이용자 유입에는 효과적이나 지속적인 유입을 기대하기는 어려웠다. 반면, 서울시민카드라는 특수성을 적극 활용하여 서울패스, 제로페이, 서울관광할인패스 등 타 어플리케이션에서 제공하기 어려운 서비스들이 통합 확대되었을 때 본 어플리케이션의 이용자 수가 꾸준하게 증가할 수 있음을 볼 수 있었다. 각 연령에 대해 단기성 이벤트는 남성이 여성보다 반응이 더 컸고, 그에 따라 이벤트가 줄어들었을 때 시점 20과 같이 이탈 가능성도 훨씬 큰 것으로 보인다. 이때 성별 관계없이 전 연령에게 효과적이였던 서울시민카드 내 서비스 업데이트가 본질적인 이용자 수를 확보, 유지할 수 있을 것으로 기대된다. 단, 기초통계분석의 그림3과 분기별 평균인 표1와 같이 2021년 상반기의 경우 비교적 낮은 접속 빈도를 보였는데 이는 실제로 이상치 탐지 결과 표2와 함께 고려해보면 해당 기간 동안 어플리케이션의 이벤트나 업데이트가 많지 않았던 것으로 설명할 수 있다.

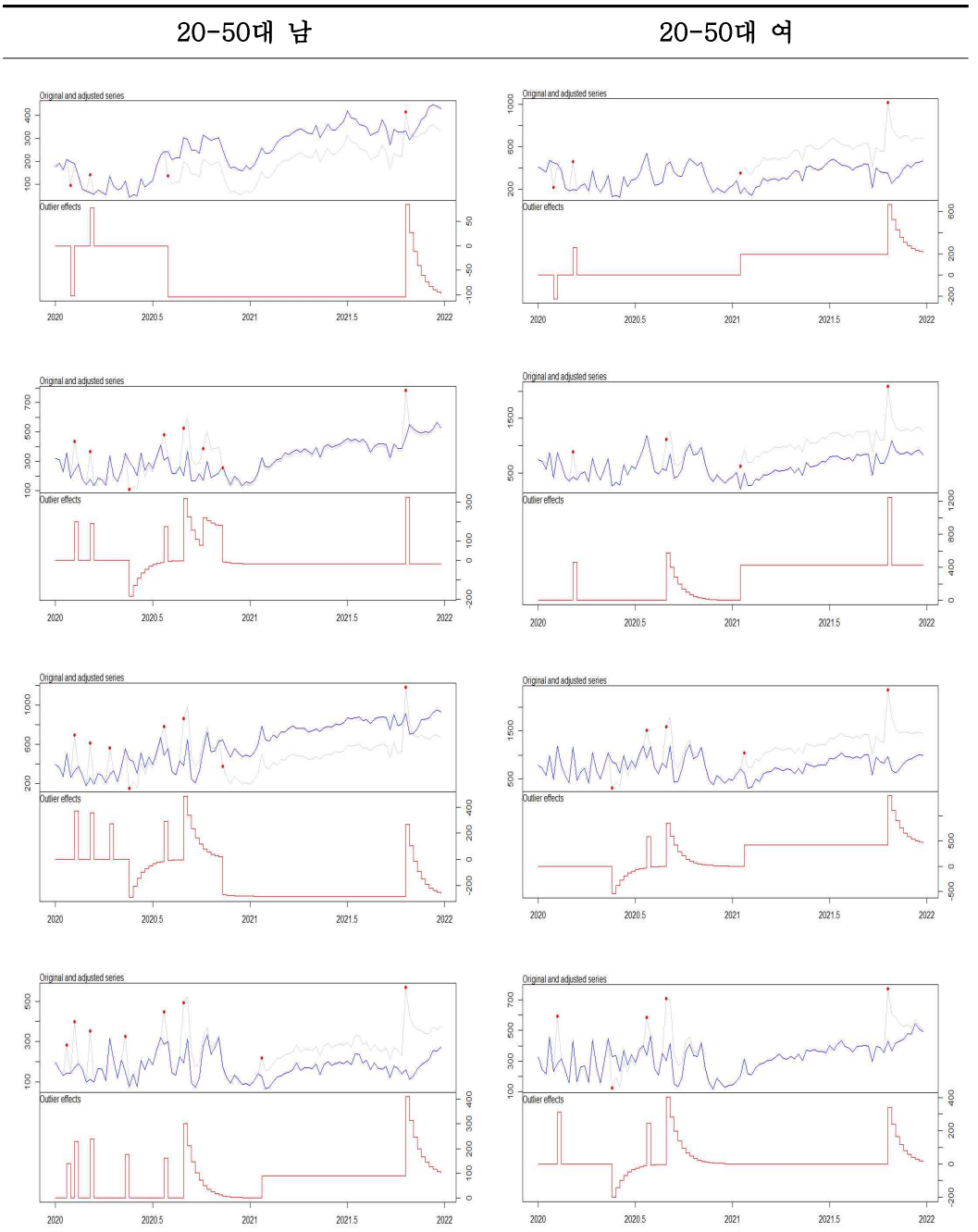

그림4: 이상치 위치와 이상점 효과 그래프

그림 4의 각 그래프에서 상단에 위치한 그래프의 파란색 실선이 auto.arima를 적용해 얻은 예측 시계열, 회색 실선이 실제 시계열 그리고 빨간 점이 이상치를 나타낸다. 하단에 위치한 그래프의 빨간색 실선은 이상점 효과를 나타낸 것이다. 하단 그래프의 경우, 수직으로 증가하였다가 떨어지면 AO, 서서히 감소하면 TC 그리고 수평 이동이

보이면 LS로 해석할 수 있다.

3.4. DTW 클러스터링 결과

본 연구의 시계열 자료들은 서로 길이가 같으나 어느 정도의 평행 이동 (shift)가 존재할 수 있기에 DTW 거리에 기반해 시계열 클러스터링하였다. 또한 계층을 나누지 않고 평면적으로 군집을 나누기 위해 분할적 군집화 방법으로 계산되었다.

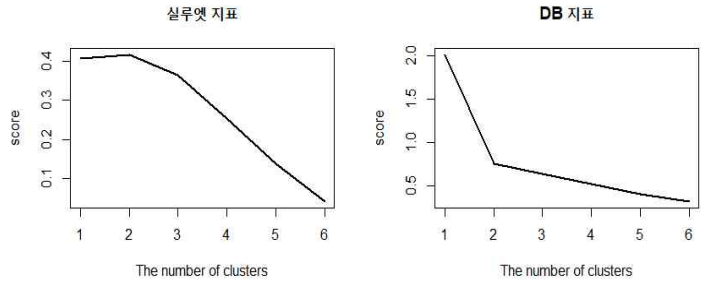

그림5. 실루엣 지표와 DB 지표 스코어

군집의 개수를 설정할 때 Kaufman L (1990)가 제안한 실루엣 지표 (sillhouette index)와 David L. Davies와 Donald W. Bouldin (1979)가 제안한 DB 지표 (db index에 기반해 군집의 개수를 설정하였다. 실루엣 지표는 다른 군집과 거리를 표준화한 값이고 높을수록 좋으며, DB 지표는 다른 군집과 분리 정도의 비율로 계산되는 값으로 낮을수록 좋다. 그림 5의 두 지표를 함께 고려하여 최적의 군집의 개수인 3개로 설정 후 분석하였다. 클러스터링 결과로 다음과 같은 세 개의 군집을 구분했다.

군집1: 30대 여성, 40대 여성

군집2: 20대 남성, 30대 남성, 50대 남성

군집3: 20대 여성, 40대 남성, 50대 여성

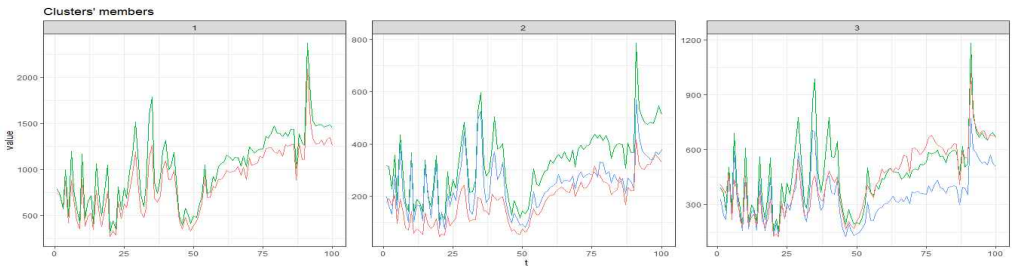

그림6: 군집별 시계열 그래프

그림6은 군집에 따라 묶인 시계열 그래프들을 보여준다. 그림 6에서 각 군집마다 이용 빈도 증감 위치나 형태가 비슷함을 볼 수 있다. 또한, 섹션 3.3에서 분석한 이상치 탐지 결과와 비교하면, 군집1은 꾸준한 접속을 보이는 군집, 군집2는 이탈 가능성이 있는 군집 그리고 군집3은 이벤트나 업데이트에 반응하는 군집으로 해석할 수 있다. 30대 여성과 40대 여성이 포함된 군집1은 서울시민카드 이용량이 타 군집과 구분되게 크게 나타났으며 이벤트와 업데이트에 즉각적인 반응을 보인다. 20대 남성, 30대 남성과 50대 남성이 포함된 군집2는 이상치 효과가 전반적으로 존재하나, 군집2에 포함된 대상들의 최대 이용량은 800에 웃돌며 타 군집에 비해 가장 적은 이용량을 기록했다. 20대 여성, 40대 남성과 50대 여성이 있는 군집 3은 군집1만큼 이용량이 많은 것은 아니지만 이벤트나 업데이트가 발생할 때 어느 정도 반응을 보인다.

차후 각 군집에서 공통으로 반응했던 이벤트나 업데이트 및 서비스에 대한 이해를 바탕으로 서울시민카드 이용자들이 필요로 하는 적절한 서비스를 고려할 수 있다.

4. 결론 및 시사점

본 논문에서는 시계열 단변량 이상치 탐지 방법론과 DTW 클러스터링을 적용해 성별과 연령대별 이상치 시점을 분석했고 비슷한 시계열 양상을 나타내는 군집 결과를 얻었다. 분석 결과를 통해 단기성 이벤트에는 즉각적인 이용자 유입이 발생하였으나 단발성에 미친다. 또한 단기성 이벤트에 잘 반응하는 집단의 경우 이용자 접속 빈도를 늘리는 데는 효과적이나, 이벤트가 종료된 경우 쉽게 이탈하는 형상를 볼 수 있었다. 반면 내부 서비스 업데이트 중 서울시민카드의 성격과 맞는 서울관광할인패스와 같은 중장기적 이벤트와 결제기능 및 실용적인 서비스들의 추가 도입은 이용자들의 지속적인 유입을 이끈다는 것을 확인할 수 있었다. 본 분석에서 얻은 정보로 이용 대상자들의 신규 유입과 이용 빈도를 높이는 데 활용될 수 있을 것으로 판단된다.

본 연구에서 단변량 이상치 탐지 기법으로 이상점들을 알아보았으나 단변량 이상치 탐지의 경우 과잉 추정과 연속적인 이상점들을 계속해 추출하는 단점이 있다. 이러한 상황에서, 다른 시계열들을 함께 고려할 수 있는 다변량 이상치 탐지 방법은 좋은 대안일 수 있다. 또한, 계절이나 사회현상을 반영할 수 있는 변수들도 추가해 함께 분석한다면 이용자들의 유입과 이탈에 대한 이해를 높일 수 있을 것으로 예상하므로 이에 대한 후속 연구가 필요할 것이다.

References

김소연. (2017년 12월 10일). 서울시, 공공시설 통합 앱 '서울시민카드' 출시. 한국일보. https://www.hankookilbo.com/News/Read/201712101248310879

김현정. (2019년 4월 7일). 서울시민카드 2.0, 뭐가 달라지나? 메트로신문. https://www.metroseoul.co.kr/article/2019040700099

김재희 (2011), R 다변량 통계분석, 교우사.

이경민, 백창룡. (2021). 장기억 시계열 모형의 이상점 탐지 연구, 한국데이터정보과학회지, 32(6), 1205-1218.

Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A. Lozano. (2021). A Review on Outlier/Anomaly Detection in Time Series Data, ACM Comput. Surv. 54, 3, Article 56, p. 33.

Bell, W. R. and Hillmer, S. C. (1983), Modeling time series with calender variation, Journal of the American Statistical Association, 78, 526-534.

Box, George E. P. and Jenkins, Gwilym M. (1976). Time Series Analysis: Forecasting and Control., San Francisco: Holden-Day.

Chang, 1. (1982). Outliers in time series, unpublished Ph.D. Dissertation, University of Wisconsin, Madison.

D. L. Davies and D. W. Bouldin (1979). A Cluster Separation Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224-227.

Kate, R.J. (2016), Using dynamic time warping distances as features for improved time series classification, Data Min Knowl Disc 30, 283 - 312.

Keogh, E., Ratanamahatana, C. (2005). Exact indexing of dynamic time warping, Knowl Inf Syst 7, 358 - 386.

Leonard Kaufman, Peter J. Rousseeuw (1990). Finding groups in data : An introduction to cluster analysis, Hoboken, NJ: Wiley-Interscience. p. 87.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series, Journal of Forecasting.