The Korean Journal of Applied Statistics
2021, Vol. 34, No. 3, 375–388

DOI: https://doi.org/10.5351/KJAS.2021.34.3.375
Print ISSN1225-066X / Online ISSN 2383-5818

Mixed-effects zero-inflated Poisson regression for analyzing the spread of COVID-19 in Daejeon

Gwanghee Kim[a], Eunjee Lee[1,a]

[a]Department of Information and Statistics, Chungnam National University

Abstract

This paper aims to help prevent the spread of COVID-19 by analyzing confirmed cases of COVID-19 in Daejeon. A high volume of visitors, downtown areas, and psychological fatigue with prolonged social distancing were considered as risk factors associated with the spread of COVID-19. We considered the weekly confirmed cases in each administrative district as a response variable. Explanatory variables were the number of passengers getting off at a bus station in each administrative district and the elapsed time since the Korean government had imposed distancing in daily life. We employed a mixed-effects zero-inflated Poisson regression model because the number of cases was repeatedly measured with excess zero-count data. We conducted k-means clustering to identify three groups of administrative districts having different characteristics in terms of the number of bars, the population size, and the distance to the closest college. Considering that the number of confirmed cases might vary depending on districts' characteristics, the clustering information was incorporated as a categorical explanatory variable. We found that Covid-19 was more prevalent as population size increased and a district is downtown. As the number of passengers getting off at a downtown district increased, the confirmed cases significantly increased.

Keywords: COVID-19, zero-inflated Poisson model, mixed-effects, Daejeon

1. 서론

2020년 1월 중국 우한지역을 중심으로 발생한 COVID-19는 2020년 1월 20일에 국내 첫 확진자가 발생한 이후 2020년 10월 5일 기준 국내에서는 24,164명, 전 세계에서는 35,393,654명의 확진자가 발생한 감염력이 높은 전염병이다. 대전광역시에서는 2020년 2월 22일에 COVID-19 환자가 최초 발생하였고, 한동안 다른 광역 지방자치단체에 비해 인구대비 적은 확진자 수를 유지하였다. 그러나 2020년 6월 15일 이후 방문판매 등의 요인으로 확진자가 크게 증가하여 2020년 6월 1일 기준 46명이었던 대전광역시의 확진자 수는 2020년 6월 30일 기준 121명까지 늘어나 한 달 동안 매우 가파른 상승세를 보였다. 2020년 6월 말부터 크게 증가했던 대전광역시의 확진자 수는 2020년 7월 18일 이후로 증가 추세가 비교적 진정된 모습을 보였었다. 그러나 전국적으로 2020년 9월에 COVID-19 확진자가 일 400명대로 크게 증가하였고, 대전광역시의 확진자 역시 2020년 8월 10

Lee's work has been partially supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (MIST)(NRF-2018R1C1B5086268).

[1] Corresponding author: Department of Information and Statistics, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon, 34134, Korea. E-mail: eunjee2@gmail.com

Published 30 June 2021 / journal homepage: http://kjas.or.kr

일을 기점으로 다시 발생하기 시작하여 2020년 10월 16일까지 249명의 확진자가 추가로 발생하였다. 이처럼 언제든지 바이러스가 재유행할 수 있고 대전광역시에서 나타난 확진자 증가 현상이 타지역에서도 나타날 수 있기 때문에, 잠재적 위험요인을 규명하여 이를 방지할 대책 마련에 도움이 되고자 본 연구를 계획하게 되었다.

본 연구에서는 대전광역시의 가파른 COVID-19 확진자 증가의 원인이 시민들의 잦은 이동, 장기간 지속한 사회적 거리두기로 인한 피로와 상대적으로 적은 확진자 수로 인한 방심에 있다고 보고, 그와 관련 있는 위험요인을 탐색하여 영향을 분석한다. 또한 행정동의 특징에 따라 위험요인의 효과가 달라질 수 있는 점을 고려한다. 따라서 각 행정동의 확진자 수를 반응변수로, 생활 속 거리두기로 전환된 시점으로부터 흐른 시간(단위: 주), 행정동의 버스 하차 인원을 설명변수로 하여 이들의 관계를 모형화한다. 또, 유사한 성격을 갖는 행정동을 세 집단으로 군집화하고, 각 집단의 특성에 따라 버스 하차 인원이 확진자 수에 미치는 영향이 어떻게 달라지는지 살펴본다.

연구에 필요한 데이터 구축을 위해 Kaggle의『Data Science for COVID-19 in South Korea』에서 (지역별 누적 확진자 데이터를 사용하였고, 이 중 대전지역의 데이터에는 대전광역시 코로나19 웹사이트에 게시된 확진자 정보를 참고하여 거주지 정보 등을 추가하였다. 방역당국의 지침이 생활 속 거리두기(사회적 거리두기 1단계)로 전환된 2020년 5월 6일이 포함된 주의 첫날인 2020년 5월 4일부터 2020년 7월 5일까지 대전광역시 내 63개 행정동에 대한 확진자 수를 수집하였고, 확진자의 발생을 설명하기 위해 각 행정동의 인구, 행정동별 버스 하차 인원, 주점업 사업체의 수 데이터를 사용하였다. 행정동의 인구는 행정안전부의 주민등록 인구 및 세대 현황(2020년 6월 기준) 자료를 사용하였고, 버스 하차 인원은 대전 교통데이터 DW 시스템의 행정구역별 승하차 인원 자료를 사용하였다. 주점업 사업체의 수 정보는 통계청에서 실시하는『전국사업체조사(2018년)』의 자료를 사용하였다.

COVID-19의 확산을 예측하고, 치료 및 예방법을 만들기 위해 COVID-19와 관련된 다양한 연구가 진행되고 있다. 대한감염학회가 제공하는 COVID-19 국내 저자 논문 정리 자료에 따르면 2020년 7월 30일을 기준으로 역학, 진단, 치료 등 7개의 주제에서 278개의 논문이 국내/외 학술지에 게재되어 있다. 특히, 더욱 효과적인 방역정책의 수립을 위해 역학 및 임상양상, 모델링에 대한 많은 연구가 진행되었다. 대표적으로 Moon 등 (2020)은 서울 지역의 확진자 수 정보를 기초로 시간에 따라 변화하는 감염재생산수를 측정하여 방역정책의 효과를 확인하고자 하였다.

행정동별 확진자 수를 주 단위로 분리하였을 때, 어떤 주에는 모든 행정동에서 확진자가 발생하지 않아 포아송분포로 기대되는 0보다 더 많은 0이 관측될 수 있다. 이를 설명하기 위해 일반적으로 포아송 분포에서 기대되는 0보다 더 많은 0이 관측되었을 때 사용할 수 있는 모형인 영과잉 포아송 회귀모형을 사용할 수 있다. Lambert (1992)는 공분산을 고려한 영과잉 포아송 회귀모형을 제안하였고, Hall (2000), Min과 Agresti (2005)는 랜덤효과를 고려한 영과잉 포아송 회귀모형이 랜덤효과를 고려하지 않은 모형보다 반복 측정된 자료에 더 잘 적합함을 보였다. Wang 등 (2002, 2003)는 환자가 병원에 머무르는 기간에 대해 관찰 값 사이에 상관관계가 존재할 때 각 관찰 값에 랜덤효과를 고려한 혼합효과 영과잉 포아송 회귀모형을 적합할 수 있으며, 평균보다 분산이 크게 나타나는 과대산포문제가 발생했을 때는 혼합효과 영과잉 음이항 회귀모형을 적합할 수 있음을 보였다. COVID-19 확진자 정보는 시군구와 같은 행정구역 단위로 제공되고 매일 확진자 수가 갱신되는 특징이 있다. Han과 Kim(2015)은 부산지역의 갑상선암 발생 자료를 분석하여 자료가 공간정보와 함께 제공되고, 소지역 및 희귀질환 자료와 같은 영과잉 가산 자료일 때 공간에 의한 랜덤효과를 고려한 영과잉 포아송 모형을 사용하는 것이 더 적절하다고 주장하였고, Zhu 등 (2015)는 시간에 따라 변화하는 자료에 혼합효과를 고려한 영과잉 모형을 적합할 수 있음을 보이고 이를 알코올 의존증 치료 연구에 적용하였다. 기존의 영과잉 포아송 모형 연구는 주로 빈도주의적 접근 방법이 지배적이지만 Kim과 Hwang (2018)은 베이지안 추론 방법을 토대로 랜덤효과를 포함한 영과잉 포아송 모형을 연구하였다.

2. 본론

2.1. 데이터 소개와 군집분석

본 연구에서는 2020년 5월 4일부터 2020년 7월 5일까지 대전 지역 내에서 발생한 확진자의 거주지 정보를 바탕으로 확진자가 크게 증가한 원인을 살펴보고자 하였다. 대전광역시에서는 2020년 2월 22일 첫 확진자가 발생한 이후 다른 지역에 비해 비교적 적은 수의 확진자를 유지하였다. 방역지침이 생활 속 거리두기로 전환된 2020년 5월 6일 기준 대전광역시의 인구 10만 명당 확진자 수는 2.72명으로, 광역시 중에서는 광주 다음으로 낮은 수치를 기록했다. 그러나 2020년 6월에 큰 폭으로 증가하여 2020년 6월 30일의 인구 10만 명당 확진자 수는 7.95명으로 증가하였다.

Figure 1의 점선은 생활 속 거리두기로 전환된 5월 2주 차를 표시한 것이다. 대전광역시의 그래프를 보면 누적 확진자가 6월에 크게 증가하는 것을 확인할 수 있다. 2020년 5월 4일부터 2020년 7월 5일까지 대전광역시의 신규 확진자 수는 99명이었으며, 이 중 해외입국자 및 다른 지역에서 감염된 경우는 분석에서 제외하여 총 89명을 반응변수로 사용하였다.

주민등록 인구통계 등 공공데이터는 행정동 단위로 제공되는 반면 확진자 거주지는 법정동 단위로 제공되어 법정동을 행정동으로 재분류하고, 법정동이 분할된 행정동(둔산1동, 둔산2동 등)은 하나의 동으로 재정의하였다. 하나의 법정동을 복수의 행정동이 관리하는 경우, 어떤 행정동에 더 많은 면적이 포함되어 있는지를 지도상 경계로 확인하여 임의로 분류하였다. 확진자가 발생하고 복수의 행정동이 존재하는 경우에만 재분류를 진행하였는데, 이 조건에 해당하는 장대동, 하기동, 봉명동, 죽동 중 장대동, 봉명동, 죽동은 온천동(온천 1동과 온천 2동의 합)으로, 하기동은 노은동(노은 1동과 노은 2동의 합)으로 분류하였다.

본 연구에서는 신도심, 구도심, 주택가 등과 같은 행정동의 특징에 따라 버스 하차 인원이 확진자 수에 미치는 영향이 다를 수 있다는 점에 착안하여 이를 모형에 반영하고자 하였다. 행정동의 특징을 나타내는 변수로 대학까지의 거리, 인구수, 주점업 사업체 수를 선택하였고, 이들과 하차량 변수와의 교호작용을 모형에 포함할 수 있다. 다만 행정동의 특징을 나타내는 세 개의 변수들이 모두 연속형이어서 하차량 변수와 교호작용을 고려할 경우 이를 해석하기 어렵고 많은 수의 교호작용항을 모형에 포함해야 한다는 문제가 있었다. 따라서 본 연구에서는 데이터 기반으로 행정동을 분류하기 위해 위의 세 변수를 이용하여 $k-$평균 군집분석을 수행하였고, 군집분석 결과를 가변수화 하여 모형에 포함하고자 한다. 대전 내 대학에 대한 위치 정보는 대전광역시청 홈페이지에서 얻었으며, 도로명 주소의 위 · 경도 변환은 다울주소전환 서비스(www.dawuljuso.com)를 이용하였다. 행정동으로부터 가장 가까운 대학까지의 거리는 행정동의 주민센터를 기준점으로 각 대학까지의 거리를 하버사인 공식(haversine formula)을 사용하여 계산한 후 최솟값을 선택하였다. 하버사인 공식은 지구의 곡률을 고려하여 두 위경도 사이의 거리를 계산하는 공식으로, 거리(d)와 지구 반지름(r)이 주어졌을 때 하버사인 공식은 다음과 같다.

$$\begin{aligned}
\text{have}(\Theta) &= \text{have}(\varphi_i - \varphi_j) + \cos(\varphi_i)\cos(\varphi_j)\text{have}(\lambda_j - \lambda_i) \\
\text{have}(\theta) &= \sin^2\left(\frac{\theta}{2}\right) = f\,\frac{1-\cos(\theta)}{2}.
\end{aligned}$$

여기서 (φ_i, λ_i)와 (φ_j, λ_j)는 각각 i지점과 j지점의 위도와 경도이다. 거리를 구하기 위해 아크하버사인을 곱하면 다음과 같은 공식이 나온다.

$$\begin{aligned}
d &= \text{rarchav}(h) = 2r\arcsin\left(\sqrt{h}\right) \\
d &= 2r\arcsin\left(\sqrt{\text{have}(\varphi_j - \varphi_i) + \cos(\varphi_i)\cos(\varphi_j)\text{have}(\lambda_j - \lambda_i)}\right) \\
&= 2r\arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_j - \varphi_i}{2}\right) + \cos(\varphi_i)\cos(\varphi_j)\sin^2\left(\frac{\lambda_j - \lambda_i}{2}\right)}\right).
\end{aligned}$$

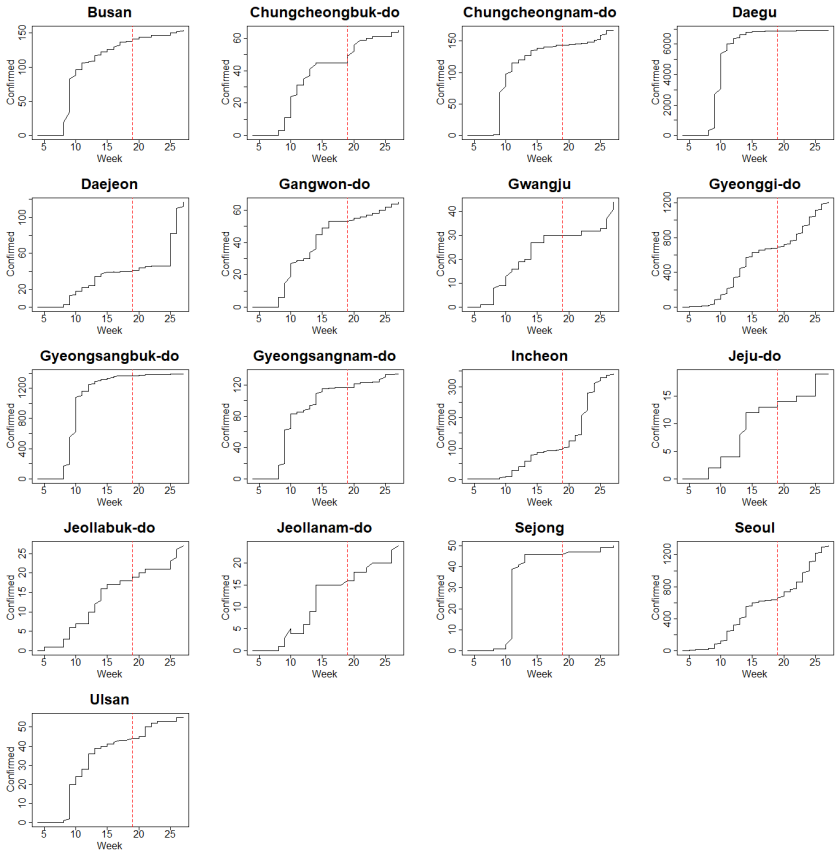

Figure 1: *COVID-19 Cases in Korea by City/Province until June 30, 2020.*

Notes: While the confirmed cases persistently increased in Seoul, Incheon, and Gyeonggi areas, the occurrence rates of Daegu and Gyeongbuk areas began to be low after the sharp increase. The confirmed cases of Daejeon exponentially increased in June 2020.

위의 공식을 사용하여 주민센터와 각 대학까지의 거리를 계산하였다. 여러 개의 행정동이 하나의 동으로 합쳐진 경우에는 각 행정동의 주민센터 위치와 인구의 가중평균값인 인구중심점을 구한 후 같은 방법으로 계산하였다. 인구중심점은 지역 내 거주하는 모든 인구가 고르게 분포되어 있다는 가정 하에서의 무게중심으로 (Kim 등, 2018), 본 연구에서는 재정의된 행정동 내에 다수의 주민센터가 존재할 때 대학까지의 거리를 계산하기 위한 기준점으로 사용되었다.

$$\bar{x_j} = \frac{\sum_i x_{ij} p_{ij}}{\sum_i p_{ij}}, \quad \bar{y_j} = \frac{\sum_i y_{ij} p_{ij}}{\sum_i p_{ij}}, \quad \begin{cases} i = 1, 2, \ldots, n_j, \\ j = \text{재정의 후 하나로 합쳐진 행정동.} \end{cases}$$

$(\bar{x_j}, \bar{y_j})$는 여러 개의 행정동이 합쳐졌을 때 인구중심점의 좌표이고, n_j는 재정의 과정을 거치며 합쳐진 행정동의 수이다. p_{ij}는 재정의 이전 행정동의 인구수이다. 여러 개의 행정동이 합쳐진 j번째 행정동의 인구중심점은 행정동의 인구가 균일하게 분포하고 있다고 가정하고, 재정의 이전 행정동의 주민센터 좌표와 인구수의 가중평균을 사용한다.

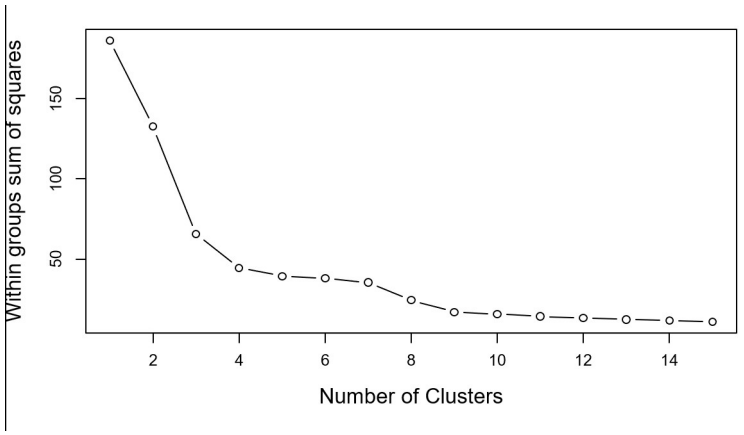

Figure 2: *Cluster size decision by a scree plot.*

Table 1: Clustering result : table

집단	행정동	크기
H	관저동, 노은동, 둔산동, 온천동	4
M	가수원동, 가양동, 가장동, 갈마동, 관평동, 괴정동, 내동, 대동, 대사동, 대청동, 대화동, 대흥동, 도마동, 만년동, 목동, 문창동, 문화동, 법동, 변동, 복수동, 부사동, 비래동, 산성동, 삼성동, 석교동, 성남동, 송촌동, 신성동, 신인동, 오류동, 오정동, 용두동, 용문동, 용운동, 용전동, 원신흥동, 월평동, 유천동, 은행선화동, 자양동, 전민동, 정림동, 중리동, 중앙동, 중촌동, 진잠동, 탄방동, 태평동, 판암동, 홍도동, 회덕동, 효동	52
L	구즉동, 기성동, 덕암동, 목상동, 산내동, 석봉동, 신탄진동	7

R의 NbClust 패키지를 사용하여 $k-$평균 군집분석을 위한 최적 군집 수를 확인하였다. 최적 군집 수를 결정하기 위해 제곱오차의 scree plot에서 기울기가 급격히 완만해지기 시작하는 지점을 확인하고, 해당 지점의 군집 수를 최적 군집 수로 결정하였다. 군집화 변수로는 각 행정동의 인구, 가장 가까운 대학까지의 최소거리, 해당 행정동의 주점업 사업체 수를 사용하였다. 단위에 의한 영향을 최소화하기 위해 각 변수의 값을 표준화한 자료를 사용하였다.

Figure 2는 군집의 수에 따른 제곱오차의 값을 scree plot으로 나타낸 것이다. $k = 3$인 지점에서 그래프의 기울기가 완만해지는 것이 확인되어 행정동을 3개의 군집으로 분류하였고, 결과를 Table 1에 요약하였다. 각 군집의 이름은 매우 번화한 집단을 H (high) 집단, 유동인구와 거주 인구가 적고 유입요인이 거의 없는 집단을 L (low) 집단, 그사이에 위치한 집단을 M (medium) 집단으로 명명하였다. 그리고 군집분석의 결과를 대전광역시 법정동 지도를 사용하여 Figure 3으로 표현하였다. 검정색으로 표시된 지역이 H 집단을, 회색으로 표시된 지역은 M 집단을, 밝은 회색으로 표시된 지역은 L 집단을 나타낸다. H 집단은 대전시의 좁은 영역에 밀집되어 있고 M 집단은 그 주위를 둘러싸고 있는 형태를, L 집단은 대전시의 주변부에 위치한 모습을 보인다. H 집단에는 4개의 행정동, M 집단에는 52개의 행정동, L 집단에는 7개의 행정동이 포함되었는데, 이러한 결과는 대전시의 특수한 상황을 잘 반영하고 있다. 예를 들어, 2000년대 초반 중앙행정기관이 대전으로 대거 이전하였고 대덕연구개발특구가 활발히 조성되어 해당 기관들이 자리 잡은 H 집단을 중심으로 신도심이 형성되었다. M 집단은 H 집단에 접근성이 높은 지역과 구도심으로 이루어져 있고 거주하기에 적절한 인프라가 구축되어 있다. 대전의 경계에 위치한 L 집단은 상대적으로 최근에 대전시에 편입된 행정동으로, 개발이 덜 되어있고 공장 또는 농업지역이다.

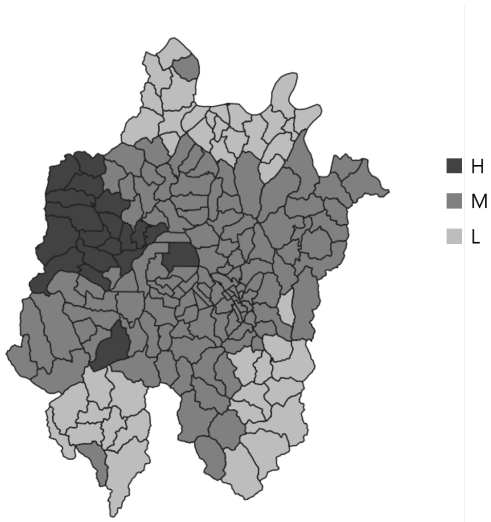

Figure 3: *Clustering result : map.*

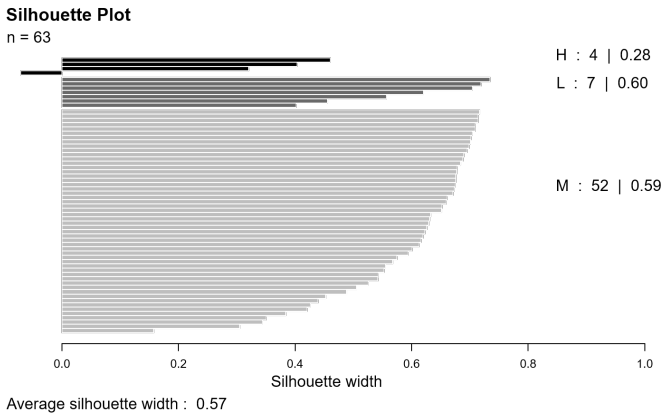

Figure 4: *Assessment of the clustering result by silhouette analysis.*

Figure 4은 각 군집에 대한 실루엣 분석 결과이다. 실루엣 계수는 각 집단의 중심점과 하나의 군집 내 자료가 다른 군집과 비교하여 얼마나 유사한지를 측정하는 값으로, i 번째 개체에 대해 다음과 같이 정의한다.

$$s(i) = \frac{(b(i) - a(i))}{(\max(a(i), b(i)))}.$$

이는 -1과 1 사이의 값으로, 해당 개체가 적절한 군집에 속한 경우 1에 가까운 값을 갖고, 그렇지 않으면 -1에 가까운 값을 갖는다. 실루엣 분석 결과 실루엣 계수의 평균이 0.57로 계산되어 Table 1의 군집결과가 적절하다고 판단하였다. 단, 실루엣 분석에서 관저동은 H 집단에 적절하지 않은 이상치로 탐지되었다.

Figure 5를 보면 총 인구수와 대학까지의 최소거리에 의한 군집은 3개로 잘 분리되지만 Figure 6에서 H 집단과 M집단이 서로 겹치면서 H집단으로 분류된 관저동보다 M집단의 행정동들이 H집단의 중심점과 더 가까워지게 된다. 이로 인해 관저동의 실루엣계수가 음수로 계산되었다.

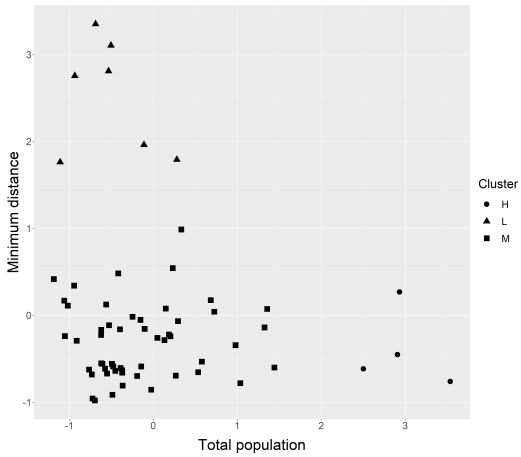

Figure 5: *A scatter plot of the minimum distance to university and the total population with the clustering results.*

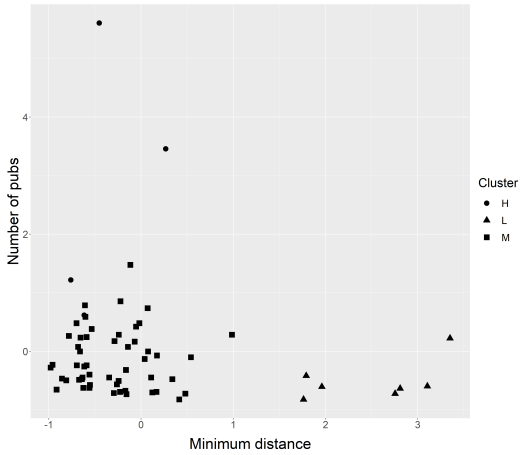

Figure 6: *A scatter plot of the number of pubs and the minimum distance to university with the clustering results.*

Table 2: Descriptive statistics of variables for each group

변수	Total	H 집단	M 집단	L 집단
버스 하차 인원	27.9 (±30.6)	99.9 (±66.2)	24.5 (±19.0)	12.6 (±5.27)
인구수	23.2 (±17.3)	75.1 (±6.54)	20.4 (±11.3)	14.2 (±7.75)
행정동 집단	63 (100%)	4 (6%)	52 (83%)	7 (11%)

Table 2는 각 군집에 대한 변수들의 평균과 표준편차를 나타낸다. 행정동과 법정동의 통일을 위해 재정의 과정을 거친 총 63개의 행정동에 대해, 각 행정동의 주 평균 버스 하차 인원은 27,929명이고, 평균 인구수는 23,191명이다. H 집단에 속한 행정동은 총 4개로, 이 집단에 속한 행정동의 주 평균 하차 인원은 99,872명이고, 평균 인구수는 75,079명이다. H 집단은 시민들의 이동이 빈번하고 인구 밀도가 높은 곳으로, 바이러스의 전파와 확산이 쉬운 특성을 가진다. M 집단에 속한 행정동은 총 52개이며, 이 집단에 속한 행정동의 주 평균

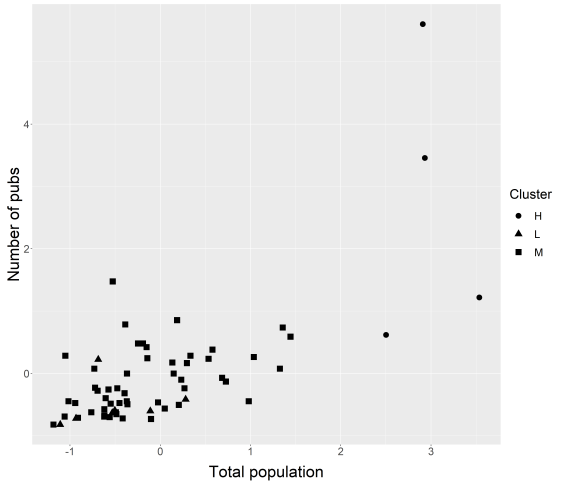

Figure 7: *A scatter plot of the number of pubs and the total population with the clustering results.*

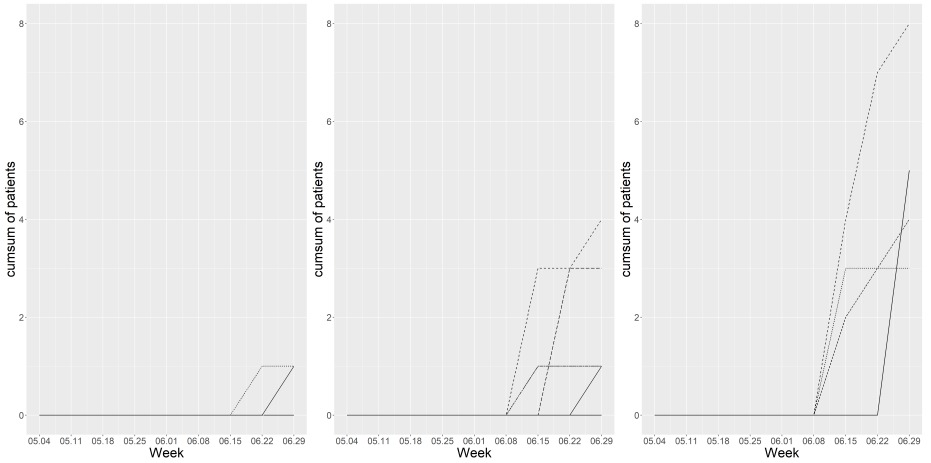

Figure 8: *The number of cumulative patients per week. The left, middle, and right panels show the L, M, and H groups' cumulative patient profiles.*

하차 인원은 24,457명이고, 평균 인구수는 20,408명이다. L 집단에 속한 행정동은 총 7개로, 이 집단에 속한 행정동의 주 평균 하차 인원은 12,613명이고, 평균 인구수는 14,215명이다. 평균 인구수보다 주 평균 하차 인원이 적고, 다른 지역보다 이동량과 인구 밀도가 낮은 것으로 보아 L 집단은 인구의 이동이 적은 외곽지역으로 볼 수 있다. M 집단에 대다수의 행정동이 속해있어서 버스 하차 인원이나 인구수의 편차도 다른 집단에 비해 크게 나타난다. 다만 H 집단의 버스 하차 인원은 같은 군집 내에서도 매우 크게 변화하는 모습을 보여 인구수가 비슷한 같은 군집 내에서도 유동인구의 차이가 존재한다는 것을 확인할 수 있다.

Figure 8은 집단별로 환자 발생 추이를 나타낸 것이다. L 집단에 속한 7개 행정동 중 2개의 행정동에서만 확진자가 발생하였고, 그 추이는 왼쪽 그래프에서 확인할 수 있다. 확진자가 발생하지 않은 행정동은 y축이 0인 직선으로 나타나 있다. 52개의 행정동이 속해있는 M 집단에서 확진자가 발생한 곳은 네 곳이었고, 그 추이

는 중간 그래프에 나타나 있다. L 집단과 마찬가지로 확진자가 발생하지 않은 행정동은 y축이 0인 직선으로 나타나 있다. 오른쪽 그래프는 H 집단에 속한 4개 행정동의 누적 확진자 발생 추이를 나타낸다. 각 행정동에서 확진자가 많이 발생할 확률이 낮기 때문에 포아송 분포를 따른다고 가정할 수 있다. 특히 행정동별 확진자 수를 주 단위로 분리하였을 때, Figure 8에서 볼 수 있듯이 어떤 주에는 모든 행정동에서 확진자가 아예 발생하지 않았으므로 포아송 분포로 기대되는 0보다 더 많은 0이 관측될 수 있다. 따라서 이러한 현상을 설명하기 위한 영과잉 포아송 모형을 사용할 필요가 있다. 일반적으로 자료 간 독립성을 가정하는 경우와는 달리 주 단위의 확진자 수는 시간상으로 연속된 관측치이기 때문에 가까운 관측치일수록 더 유사한 값을 가지는 자기 상관이 존재할 수 있다. 이런 경우 설명변수를 포함한 고정효과와 시간효과를 포함한 랜덤효과를 함께 고려하는 혼합효과모형을 이용할 수 있다. 다음 장에서 영과잉 포아송 모형과 영과잉 혼합효과모형을 설명한다.

2.2. 영과잉 포아송 모형

반응변수 Y_i가 음이 아닌 정수값을 갖는 영과잉 이산형 자료의 형태를 띌 때, Y_i의 확률분포는 0의 값이 발생하는 경우와 0보다 큰 정수값을 갖는 포아송 분포가 혼합된 혼합분포 구조를 가지고 있다.

$$Y_{ij} \sim \begin{cases} 0, \\ \text{Poisson}(\lambda), \end{cases}$$

반응변수 $Y_i\,(i = 1, \ldots, n)$가 영과잉 포아송 모형을 따를 때, 다음과 같은 확률질량함수(probability mass function)를 가진다.

$$P(Y_i = y_i|\lambda, p) = \begin{cases} p + (1-p)e^{-\lambda}, & \text{if } y_i = 0, \\ (1-p)\dfrac{e^{-\lambda}\lambda^{y_i}}{y_i!}, & \text{if } y_i > 0. \end{cases}$$

여기서 $0 \le p \le 1$는 0의 값에서 주어지는 임의의 확률이며, $\lambda > 0$는 포아송 분포의 평균이다. 이 모형에서 0의 값은 두 개의 분포에서 각각 발생하고 있음을 알 수 있다. 즉 영과잉 상태에서 발생하는 경우와 포아송 분포를 통해서 발생하는 경우로 나뉘어진다.

2.2.1. 영과잉 포아송 회귀모형

위에서 정의한 영과잉 포아송 모형에 공변량을 고려하면 영과잉 포아송 회귀모형이 정의된다.

$$\text{logit}(\pi_i) = \alpha_0 + \alpha_1 X_{11i} + \cdots + \alpha_m X_{1mi},$$
$$\log(\lambda_i) = \beta_0 + \beta_1 X_{21i} + \cdots + \beta_l X_{2li}.$$

여기서 X_{1i}는 제로 단계에서의 m개의 공변량으로 이루어진 벡터이고, X_{2i}는 포아송 분포 단계에서의 l개의 공변량으로 이루어진 벡터이다. 또한 α와 β는 각각 거기에 상응하는 계수 벡터이다.

2.2.2. 랜덤효과를 포함하는 영과잉 포아송 회귀모형

영과잉 포아송 회귀모형에서 $\text{logit}(\pi_{ij})$과 $\log(\lambda_{ij})$는 각각 베르누이 분포에서의 성공확률과 포아송분포의 평균에 대한 연결함수이다. i번째 그룹의 j번째 개체의 반응변수를 $Y_{ij}(i = 1, \ldots, n,\ j = 1, \ldots, m)$라고 정의할 때, 다음과 같이 식을 정의할 수 있다.

$$\text{logit}(\pi_{ij}) = \alpha Z_{ij} + u_i,$$
$$\log(\lambda_{ij}) = \beta X_{ij} + v_i.$$

Table 3: The estimation results for the ZIP with mixed effects model

	Variables	Estimate (SE)	p-value
log-link	Intercept	–0.816 (1.756)	0.642
	sum_out	–0.036 (0.020)	0.077
	group L	0.449 (1.842)	0.807
	group H	–3.256 (1.618)	0.044
	time	–0.226 (0.188)	0.231
	total_pop	0.080 (0.022)	<0.001
	sum_out:group L	–0.149 (0.151)	0.325
	sum_out:group H	0.043 (0.022)	0.053
logit-link	Intercept	9.688 (2.707)	<0.001
	time	–1.735 (0.412)	<0.001
	total_pop	0.035 (0.037)	0.350

X와 Z는 설명변수들의 행렬이고 α와 β는 각각에 대한 회귀계수 벡터를, u_i와 v_i는 각 모형의 부분에서 랜덤효과를 나타낸다. 이 때 행렬 X와 Z는 서로 다른 설명변수들을 나타낼 수도 있다. 각 부분의 랜덤효과항에 대한 분포는 다음과 같이 정의된다.

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(0, \Sigma\right), \qquad \Sigma = \begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2. \end{bmatrix}$$

여기서 Σ는 랜덤효과항의 공분산행렬을 의미한다. 각 관찰값의 랜덤효과항이 평균이 0이고 분산이 Σ인 이변량정규분포를 따른다고 가정한다. 고정효과의 랜덤효과항과 영과잉을 설명하는 부분의 랜덤효과항은 서로 상관관계가 존재할 수 있다.

2.3. 대전지역 COVID-19 데이터에 대한 적용

본 연구에서는 대전광역시의 COVID-19 확진자 수에 영향을 미치는 위험요인을 탐색하고 그 영향을 분석하기 위해 앞서 소개한 데이터에 랜덤효과를 포함하는 영과잉 포아송 회귀모형을 적용한다. 반응변수로 각 행정동의 주 단위 확진자 수를, 설명변수로 생활 속 거리두기로 전환된 시점으로부터 흐른 시간(time, 단위: 주), 행정동의 인구수(total_pop)와 버스 하차 인원(sum_out), 그리고 행정동 집단(group L, group H)을 사용한다. group L과 group H은 행정동 집단에 대한 가변수로 group L은 L 집단에 대한 표시자를, group H은 H 집단에 대한 표시자를 나타낸다. 본 연구에서 인구수는 군집분석과 회귀모형에서 동시에 사용되었다. 특히 회귀모형에서의 인구수는 모형의 설명력을 높이고 확진자 발생에 미치는 잠재적 효과를 조정하기 위해 설명변수로써 사용되었다. 가령 인구수가 많은 지역일수록 버스 하차 인원도 증가하고 코로나 확진 빈도도 높다. 이렇듯 인구수가 교락효과를 가지므로 버스 하차 인원과 행정동 집단이 코로나 발생에 미치는 순수한 영향을 추정하기 위해 인구수를 설명변수로 포함하였다.

행정동 집단의 특성에 따라 버스 하차 인원이 확진자 수에 미치는 영향이 달라지는지 살펴보기 위해 행정동 집단과 하차 인원과의 교호항을 고려한다. 시간과 행정동 집단 변수 사이의 교호항은 유의수준 5%에서 유의하지 않아 제거하였다. R의 GLMMadaptive 패키지를 사용하여 혼합효과 영과잉 포아송 회귀모형을 적합하였다. 모수의 추정은 quasi-Newton 방법을 사용하였다.

Table 3은 랜덤효과를 포함하는 영과잉 포아송 회귀모형의 추정된 회귀계수와 표준오차, 그리고 $p-$값을 나타낸다. H집단과 버스 하차인원의 상호작용항의 유의확률이 0.053로 나타나 유의수준 0.05에서 유의하지는 않으나 그 경계에 있어 하차랑의 효과가 행정동의 특징과 결부되어 있을 가능성을 시사하였다. 따라서

Table 4: Model assessment of a ZIP model with or without mixed effects. Their AIC and BIC values are evaluated

Model	AIC	BIC
ZIP model	372.38	420.12
ZIP model with mixed effects	360.06	390.07

집단과 버스 하차인원의 모든 상호 작용항을 최종 모형에 포함하였고, hierarchy rule에 따라 집단과 하차인원의 주효과를 모형에 포함시켜 해석하였다. H 집단인 경우, 버스 하차인원이 천 명 증가할 때, 확진자 수가 $\exp(-0.036 + 0.043) = 1.007$배 증가하고, L 집단인 경우, 버스 하차인원이 천 명 증가할 때, 확진자 수가 $\exp(-0.036 - 0.149) = 0.83$배 증가, 즉, 17% 감소한다. M 집단인 경우, 버스 하차인원이 천 명 증가할 때, 확진자 수가 $\exp(-0.036) = 0.96$배 증가, 즉, 4% 감소한다. 버스 하차인원이 천 명일 때, M 집단에 비해 H 집단의 확진자 수는 평균적으로 $\exp(-3.256 + 0.043) = 0.04$배, 즉 96% 적다. 버스 하차인원이 천 명일 때, M 집단에 비해 L 집단의 확진자 수는 평균적으로 $\exp(0.449 - 0.149) = 1.34$배, 즉 34% 많다. 상호작용항을 해석하면 버스 하차인원이 천 명 증가할 때, M집단에 비해 대학까지의 거리가 가깝고, 주점업 사업체의 수가 많은 H집단의 확진자가 $\exp(0.043) = 1.04$배, 즉, 4%증가한다고 해석할 수 있다.

흥미로운 점은 모형에서 행정동 집단의 효과가 예상한 것과 다르게 나타났다는 것이다. 유동인구가 많은 H 집단의 확진자 수가 그렇지 않은 M 지역에 비해 많을 것으로 예상하였는데, 분석 결과는 그 반대였다. 예를 들어 버스 하차 인원이 평균 하차량인 28,000명으로 동일할 때, L 집단의 확진자 수가 M 집단보다 $\exp(-2.8 \times 0.149 + 0.449) = 1.032$배 많고, H 집단의 확진자수는 M 집단의 $\exp(2.8 \times 0.043 - 3.256) = 0.043$배로 약 96% 적다고 할 수 있다. 이런 현상이 나타난 이유는 확실하지 않으나, 7월 9일 대전광역시 허태정 시장의 코로나19 상황브리핑에서 유추해볼 수 있다. 이 브리핑에서 대전시는 유흥주점, 단란주점 등 고위험 시설에 대한 집합 제한 행정조치를 유지하겠다고 밝혔는데, 실제로 주점업 사업체가 많은 H 집단에 속한 행정동에 시 차원에서 더욱 강도 높은 방역 대책을 적용한 결과 이런 현상이 발생했을 가능성이 있다. 즉, 고위험 지역에 대해서는 시 차원에서 상대적으로 강한 제재가 존재하였기 때문에 인구수와 버스 하차량이 통제되었을 때, 고위험 지역인 H 집단의 확진자가 더 적게 나타날 수 있는 것이다.

2.4. 모형의 평가

Liu 등 (2020)은 혼합효과를 고려한 영과잉 포아송 모형의 적합도를 판단할 수 있는 통계적 검정법을 제안하였고, 이를 바탕으로 모형의 가정들을 만족하는지 확인하기 위해 DHARMa패키지를 사용하여 모형진단을 수행하였다. 관측된 값과 시뮬레이션으로 도출된 값을 비교하였으며, 만약 모형이 데이터를 잘 적합한다면 두 값은 동일한 분포를 따를 것이라고 기대할 수 있다. 잔차는 0과 1 사이의 값으로 변환되며, 이 값이 균등분포를 따를 것이라고 기대한다.

Figure 9를 보면 QQplot이 y=x그래프와 유사한 모양을 보이는 것으로 보아 관측된 값과 시뮬레이션으로 계산된 값이 서로 같은 분포를 가지고 있는 것으로 보인다. 같이 제시된 잔차그림의 경우에도 분포가 수평선에서 약간 벗어난 모양을 보이지만 대체로 수평에 가까운 모습을 보여 모형에 큰 문제가 없다는 결론을 내렸다.

모형의 적합도를 평가하기 위해 랜덤효과를 포함한 모형과 그렇지 않은 모형의 Akaike information criterion (AIC)와 Bayesian information criterion (BIC)를 비교하였다. 그 결과 랜덤효과를 포함한 ZIP 모형이 랜덤효과를 고려하지 않은 ZIP 모형보다 작은 AIC와 BIC 값을 가져 자료를 더 잘 적합하는 것으로 나타났다.

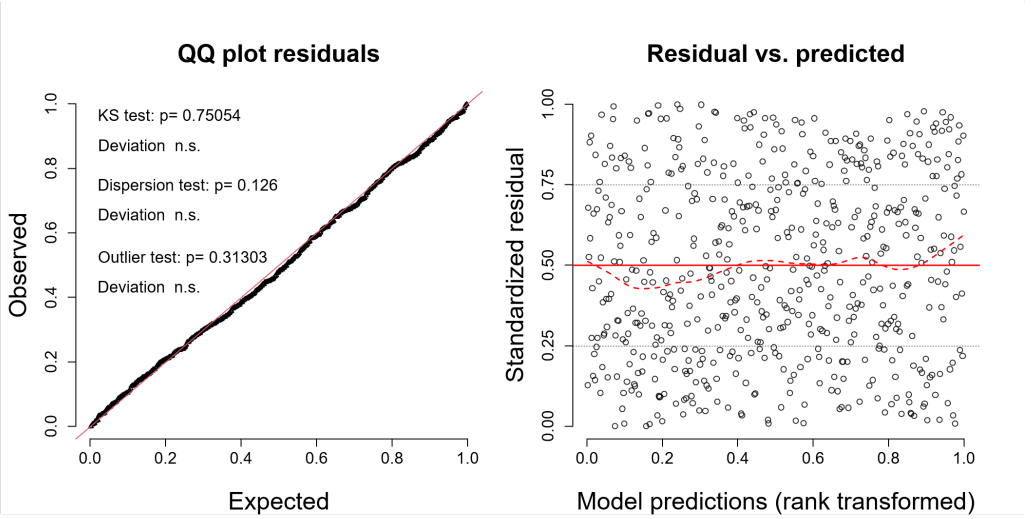

Figure 9: *Diagnostics of assumptions for the ZIP model with mixed effects.*

3. 결론

본 연구에서는 5월 초부터 7월 초까지의 대전광역시 COVID-19 확진자 자료를 이용하여 행정동별 주 단위 확진자 발생 수를 반응변수로 두고, 버스 하차량과 인구수, 행정동의 특성이 확진자의 발생에 어떤 영향을 미치는지 모형을 통해 확인하였다. 행정동을 특징에 따라 범주화하기 위해 군집분석을 통해 3개의 군집을 만들었고, 각 시점의f 자료가 이전 시점의 자료에 영향을 받을 수 있다는 것을 고려하여 혼합효과 영과잉 포아송 회귀모형을 자료에 적합하였다. 모형 적합 결과 COVID-19 전파 위험성이 높은 H 집단과 각 행정동의 인구수의 회귀계수가 유의수준 0.05에서 통계적으로 유의한 것으로 나타났다. 그리고 관심 변수인 버스 하차 인원과 군집과의 상호작용에 대해 H 집단과 버스 하차 인원의 상호작용의 유의확률이 0.053으로 계산되었다. 이 결과를 바탕으로 H 집단의 행정동에서 버스 하차 인원이 증가했을 때, 유의수준 10%에서 확진자가 유의하게 증가하는 것을 확인할 수 있었다. 또한, 고정된 인구수와 버스 하차량에 대해 H 집단의 확진자 수는 M 집단의 확진자 수보다 훨씬 적을 것으로 기대되었는데 이는 코로나 고위험 지역에 대한 시 차원의 강력한 대응이 효과를 발휘한 것으로 해석할 수 있다.

본 연구의 한계점으로 개인정보 문제로 인해 확진자의 동선이 공개되지 않거나 삭제되어 확진자들의 실제 생활지역을 알기 어렵다는 점이 있다. 추후 연구를 위해 제한적으로 확진자의 동선이 공개된다면 이 정보를 바탕으로 확진자의 생활지역을 특정하여 분석에 활용할 수 있을 것이다. 또한, 대전지역의 유동인구 자료가 공개되지 않아 버스의 버스 하차 인원으로 인구 이동을 간접적으로 측정하였기 때문에 해당 지역에 방문한 사람의 수나 목적 등을 알 수 없었다. 이후에 관련 자료가 보강된다면 시간대와 연령에 따른 이동정보를 바탕으로 더욱 심층적인 분석을 진행할 수 있을 것이다.

References

Agresti A (2020). *An Introduction to Categorical Data Analysis*, Wiley.

Hall DB (2000). Zero-inflated poisson and binomial regression with random effects: a case study, *Biometrics*, **54**, 1030–1039.

Han JH and Kim CH (2015). Zero inflated Poisson model for spatial data, *The Korean Journal of Applied Statis-*

tics, **28**, 231–239.

Kim JT, Lee NY, Oh MA, and Lee SI (2018). A study on the transition of population movement and centroid in Korea, *Journal of The Korean Official Statistics*, **23**, 1–23.

Kim YK and Hwang BS (2018). A Bayesian zero-inflated Poisson regression model with random effects with application to smoking behavior, *The Korean Journal of Applied Statistics*, **31**, 287–301.

Lambert D (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34**, 1–14.

Liu J, Yanyuan M, and Jill J (2020). A goodness-of-fit test for zero-inflated Poisson mixed effects models in tree abundance studies, *Computational Statistics & Data Analysis*, **144.**

Moon, *et al.* (2020). Time-variant reproductive number of COVID-19 in Seoul, Korea, *Epidemiology and Health*, **42**.

Wang K, Yau KK, and Lee AH (2002). A zero-inflated Poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays, *Computer Methods and Programs in Biomedicine*, **68**, 195–203.

Min Y and Agresti A (2005). Random effect models for repeated measures of zero-inflated count data, *Statistical Modelling*, **5**, 1–19.

Zhu H, Luo S, and Stacia MD (2015). Zero-inflated count models for longitudinal measurements with heterogeneous random effects, *Statistical Methods in Medical Research*, **26**, 1774–1786.

Received February 17, 2021; Revised April 8, 2021; Accepted May 26, 2021

혼합효과 영과잉 포아송 회귀모형을 이용한 대전광역시 코로나 발생 동향 분석

김광희[a], 이은지[1,a]

[a]충남대학교 정보통계학과

요 약

본 연구는 대전광역시에서 나타난 확진자 증가 현상을 분석하여 COVID-19의 확산을 방지할 대책 마련에 도움이 되고자 계획되었다. 확진자 증가의 원인이 시민들의 잦은 이동과 장기간 지속한 사회적 거리두기로 인한 피로와 방심에 있다고 보고, 각 행정동의 주별 확진자 수를 반응변수로, 생활 속 거리두기로 전환된 시점으로부터 흐른 시간, 행정동의 버스 하차 인원을 설명변수로 하여 이들의 관계를 모형화하였다. 행정동별 확진자 수가 주 단위로 반복측정 되었고, 포아송분포로 기대되는 0보다 더 많은 0이 관측될 수 있기 때문에 혼합효과 영과잉 포아송 회귀모형을 적용하였다. 행정동의 성격에 따라 확진자 발생 동향이 다를 수 있어서서 서로 유사한 성격을 갖는 행정동을 군집화하여 이를 범주형 설명변수로 사용하였다. 또한 버스 하차 인원의 효과가 행정동의 성격에 따라 달라질 수 있다는 점을 고려하여 두 변수 간의 교호작용항을 포함하였고 상대적으로 번화한 행정동에서 그 효과가 유의한 것으로 나타났다 (유의수준=0.1). 모형 적합 결과 인구수의 증가와 번화한 행정동이라는 요인, 그리고 버스 하차 인원의 증가가 확진자 수의 증가와 중요한 연관 관계를 가진다는 것을 보였다. 한편, 추정된 모형에 따르면 인구수와 버스 하차량이 고정되었을 때 번화한 집단의 확진자 수가 그렇지 않은 집단에 비해 훨씬 적을 것으로 기대되었는데, 이는 코로나 고위험 지역에 대한 시 차원의 강력한 대응이 효과를 발휘한 것으로 해석할 수 있다.

주요용어: 코로나19, 혼합효과, 영과잉포아송회귀모형, 대전광역시

이 논문은 2018년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2018R1C1B5086268).

[1]교신저자: (34134) 대전광역시 유성구 대학로 99, 충남대학교 정보통계학과. E-mail: eunjee@cnu.ac.kr