The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.2.215
30(2), 215–231

Autologistic models with an application to US presidential primaries considering spatial and temporal dependence

Ho Jeong Yeom[a] · Won Kyung Lee · So Young Sohn[a,1]

[a]Department of Information and Industrial Engineering, Yonsei University

(Received January 5, 2017; Revised February 22, 2017; Accepted March 16, 2017)

Abstract

The US presidential primaries take place sequentially in different places with a time lag. However, they have not attracted as much attention in terms of modelling as the US presidential election has. This study applied several autologistic models to find the relation between the outcome of the primary election for a Democrat candidate with socioeconomic attributes in consideration of spatial and temporal dependence. According to the result applied to the 2016 election data at the county level, Hillary Clinton was supported by people in counties with high population rates of old age, Black, female and Hispanic. In addition, spatial dependence was observed, representing that people were likely to support the same candidate who was supported from neighboring counties. Positive auto-correlation was also observed in the time-series of the election outcome. Among several autologistic models of this study, the model specifying the effect of Super Tuesday had the best fit.

Keywords: US president primary election, autologistic model, spatial dependence, temporal dependence

1. 개요

미국에서는 전당 대회에 참석하여 후보를 결정하는 대의원들의 수를 결정하기 위해 전당 대회 이전에 몇 달간의 예비 선거(primary election)과 당원 대회(caucus)가 진행된다. 특히, 예비 선거를 통해 일반 유권자들이 지지하는 후보와 당원 대회를 통해 해당 선거구가 지지하는 후보를 파악할 수 있다. 또한, 예비경선을 통해 각 지역별 유권자 성격에 따른 후보 선택 성향에 대해 살펴볼 수 있으며, 추후 대통령 선거인단 선출 시 특정 후보 지지에 대해 유추해볼 수 있다.

미국 대통령 선거와 관련해서 여러 가지 연구 (Abramowitz, 2008; Lewis-Beck, 2005; Norpoth, 2004)들이 진행되었지만, 선거 과정을 충분히 고려하지 않고 대통령 선거의 최종 결과와 관계가 있는 변

This research, 'Geospatial Big Data Management, Analysis and Service Platform Technology Development', was supported by The Ministry of Land, Infrastructure and Transport (MOLIT), Korea, under the national spatial information research program supervised by the Korea Agency for Infrastructure Technology Advancement (KAIA)(17NSIP-B081011-04).

[1]Corresponding author: Department of Information and Industrial Engineering, Yonsei University, 50, Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea. E-mail: sohns@yonsei.ac.kr

수를 탐색하여 선거 예측 모형을 개발하는 것이 대부분이었다. 미국에서 대통령이 선출되기까지 걸리는 긴 과정 중에서 예비 선거는 선거인단 선출 전 민심을 확인할 수 있는 단계이다. 따라서, 예비 선거에 대한 연구는 대통령 선거 결과 예측 관련 변수들을 확인해볼 수 있는 기회가 될 수 있다.

예비 선거는 한 번에 모든 선거가 이루어지지 않고 계획된 일정대로 정해진 날짜에 특정 선거구에서 선거를 실시하게 되어, 지역별로 시간 차이(time lag)가 발생하게 된다. 이 부분에서 경선과 대선의 차이점이 발생하며, 예비 선거 결과를 연구하는데 있어서는 대선과 달리 투표 결과의 변수로 시공간적인 요소를 고려할 수 있다. 시공간 분석을 통해 특정 선거구의 투표 결과가 이웃한 선거구의 결과와 연관이 있거나, 이전에 투표를 실시한 선거구의 투표 결과가 다음 시기에 투표를 실시하는 선거구의 결과와 연관이 있는지 확인할 수 있다. 일반적으로 시공간 분석은 패널데이터를 활용하나, 미국 예비 경선의 경우 시기적으로 투표하는 지역이 다르기 때문에 완벽한 패널데이터의 양상을 갖추고 있지 못한 특징이 있다. 일반적으로 공간분석모형에서는 공간상 이웃과의 물리적 거리를 반영하는데 예비 경선의 경우에는 선거구별 선거 날짜의 차이를 공간상의 거리로써 반영해 볼 수가 있다.

본 연구에서는 시공간적인 특징을 갖는 예비 경선 상황을 반영하여 지역별로 선거 결과와 관계가 유의한 사회경제적 변수를 규명하기 위해, 거리와 시차 관점의 공간 관계를 동시에 반영한 자기로지스틱 회귀모형(autologistic regression)을 제안한다. 또한 시공간 상관관계를 고려하지 않은 로지스틱 회귀모형, 그리고 공간 의존성을 고려한 자기로지스틱 회귀 모형과 비교하여 시공간 모델링의 중요성을 살펴본다.

본 논문의 구성은 다음과 같다. 제 2장에서는 미국 선거 결과 예측 관련 기존 연구를 고찰한다. 제 3장에서는 본 연구 진행 방법과 분석에 사용된 데이터 설명 및 연구에서 적용하려는 연구 모형을 설명한다. 4장에서는 분석 결과와 시사점을 제시한다. 제 5장에는 결론과 추후 연구분야를 제시한다.

2. 선행 연구

2.1. 미국 선거 관련 선행 연구

Abramowitz (2008)는 2008년 대선 후보자의 득표율 예측을 위해 갤럽(Gallup)이 조사한 대통령 후보 지지율, 현직 대통령의 2/4분기 실질 국내총생산(GDP) 성장률, 여당의 임기 기간 변수를 선택한 선형 회귀모형을 사용하였다. Lewis-Beck과 Tien (2008) 또한 2008년 대선 결과 예측을 연구했으며, 이들은 갤럽이 조사한 대통령 후보 지지율, 현직 대통령의 상반기 GNP, 일자리 성장률 및 현직 정당 여부를 변수로 선택하여 1952년부터 2004년까지의 데이터를 다중선형회귀모형에 적용하여 추정된 결과로 선거를 예측하였다. Norpoth (2004)는 후보자의 예비선거 지지율과 이전 두 차례의 후보자의 당의 득표율을 변수로 한 자기 회귀 모형(autoregressive model)으로 선거 결과를 예측하였다.

선거 결과 예측을 목표로 했던 연구들 외에, 예비 선거에서 후보자를 선택하게 되는 요인을 분석한 연구들도 있었다. Bartels (1987)는 모멘텀 효과(momentum effect)를 기반으로 선거에서 후보자들에 대한 초기 정보가 부족한 상태에서는 아이오와 코크스나 뉴햄프셔 프라이머리 혹은 2월 초의 슈퍼화요일 때 예상보다 많은 지지를 받게 된 후보를 지지하는 경향이 있음을 설명하였고, 이러한 현상이 예비 선거에서 후보자를 선택하는데 관련이 있다고 주장하였다. Norrander (1993)은 이러한 모멘텀 효과가 슈퍼화요일에 예상보다 많은 지지를 받을 수도 있는 뜻밖의 후보, 즉 다크호스 후보가 선거에서 유리하게 작용했음을 확인했으며, Steger (2007)은 모멘텀 효과가 공화당보다 민주당의 선거 결과와 더 큰 연관성이 있음을 주장하였다.

2.2. 사회경제학 변수와 선거

기본적으로 피선거권을 가지는 사람들이 갖는 지역적인 특징이 있기 때문에, 사회경제적 자료가 선거와의 연관성이 있을 것이라는 가설 하에, 선거 결과와의 연관성을 탐색한 기존의 연구들이 있었다.

2008년 미국 대선에서 버락 오바마의 당선은 백인들의 지지에 큰 영향을 받았으며, 이에 Kim (2012)은 어떤 특성을 가진 백인들이 버락 오바마를 지지하였는지에 대한 연구를 진행하였다. 또한, 미국 인구의 큰 비중을 차지하는 백인, 아프리카계 미국인(African American) 외에 히스패닉(Hispanic)인의 수가 증가한 점에서 히스패닉 및 흑인 유권자가 선거 결과와 어떠한 관계가 있는지를 파악하였다. Kim (2013)은 히스패닉인과 2004년 대통령 선거 결과와의 관계를 파악하였고, 미국이 간접 선거 제도를 따른다는 점에서 일부 히스패닉 계가 조지 부시 후보를 지지하게 된 것이 그가 당선되는 데는 영향을 주었지만, 전체적으로는 히스패닉 계는 존 케리 후보를 더 많이 지지했던 것으로 나타나 히스패닉 인구의 선거 결과에 대한 유의성이 나타나지 않았다. Cho (2009)는 2008년 1, 2, 3, 4월에 민주당 후보자에 대한 지지도 여론조사를 약 1,500명을 대상으로 수행하였고, 여론조사 데이터를 분석하여 지지율과 사회경제학적 요소간의 관계를 파악하였고, 성별, 연령, 교육, 흑인 변수가 지지율에 유의한 변수로 파악되었다. 다른 변수인 소득수준과 히스패닉 인구는 지지율에 유의하지 않았다. 2008년의 민주당 예비선거 후보자가 버락 오바마와 힐러리 클린턴이었고, 오바마의 경우는 흑인의 지지가, 힐러리의 경우는 여성의 지지가 큰 관계가 있었다는 것을 보여주었다.

2.3. 자기로지스틱 회귀모형(autologistic regression)을 적용한 연구 사례

공간적으로 이웃한 지역 간의 공간 의존성을 분석할 수 있는 방법으로 자기로지스틱 모형(autologistic model)이 제안되었고 (Besag, 1974) 다양한 분야에서 이 모형이 적용되고 있다. 특히 자기로지스틱 모형은 기후 관점에서 식물 혹은 종의 분포에 대한 공간의 연관성을 알아보기 위해 많은 연구에서 적용되었다 (Austin, 2002; He 등, 2003; Wu와 Huffer, 1997). Buckland과 Elston (1993)는 스코틀랜드 그램피언(Grampian)주의 야생 사슴 출현 분포를 연구하였는데, 지역의 고도(altitude), 진흙 혹은 숲 지역(pinewood, mires), 좌표값(easting, northing)을 공변량으로 고려하고 해당 지역의 사슴 출몰 여부를 종속 변수로 하여 로지스틱 회귀모형으로 모형화하였다. Augustin 등 (1996)는 Buckland과 Elston (1993)의 연구에서 자기상관(autocorrelation) 변수를 추가한 자기로지스틱 모형을 적용하여 새롭게 결과를 분석하였다. Gumpertz 등 (1997)는 피망(bell pepper)을 대상으로 특정 병원균(phytophthora) 전염의 공간적 패턴을 파악하기 위해 토양 관련 변수와 함께 autologistic regression을 사용하였다. 종속변수인 전염 여부에 대해 토양수의 성분과, 연구대상 지역을 16 by 16의 격자로 나누어 행 열로 인접한 지역의 감염 개체 여부와 대각 지역의 감염 개체 여부에 대해 공간적으로 병균의 전염이 경향성을 보이는 것인지를 살펴보았다.

3. 연구 진행

본 연구에서는 민주당 예비 선거 결과와 유의한 관계가 있는 사회경제적 변수 규명에 있어 시공간 관계를 반영하는 데 주안점이 있다. 민주당의 두 후보자인 힐러리 클린턴 또는 버니 샌더스의 당선 여부를 종속변수로, 각 카운티(county)의 사회경제적 지표, 공간 의존성, 시간 의존성을 설명변수로 고려한다. 공간적 자기상관관계 파악을 위한 Join-count statistic를 통해 설명변수의 유의성을 확인한다. 또한, 각 카운티의 시간 및 공간 속성을 고려하기 위해 가중 행렬을 구성하여 시간 의존성과 공간 의존성을 모형에 적용할 수 있도록 하였다. 앞서 결정한 사회경제적 변수와 시공간 선거 결과 연관성을 살펴볼 수 있도록 로지스틱 회귀모형과 여러 자기 로지스틱 회귀모형을 고려한다.

3.1. 연구 데이터

미국에는 푸에르 토리코와 미국령을 제외하고 총 51개 주 3,142개 카운티가 존재하며, 지역 간 인접도를 계산하는데 있어 북아메리카 중심의 미 대륙과 멀리 떨어져 있는 알래스카와 하와이 지역은 분석에서 제

Table 3.1. 2016 US Democrat primary election schedule

t_i	Date	State
1	Feb 1^{st}	Iowa
2	Feb 9^{th}	New Hampshire
3	Feb 20^{th}	Nevada
4	Feb 27^{th}	South Carolina
5	Mar 1^{st} (Super Tuesday)	Alabama, Arkansas, Colorado, Georgia, Massachusetts, Minnesota, Oklahoma, Tennessee, Texas, Vermont, Virginia
6	Mar 5^{th}	Kansas, Louisiana, Maine, Nebraska
7	Mar 8^{th}	Michigan, Mississippi
8	Mar 15^{th}	Florida, Illinois, Missouri, North Carolina, Ohio
9	Mar 22^{th}	Arizona, Idaho, Utah
10	Mar 26^{th}	(Alaska, Hawaii) Washington
11	Apr 5^{th}	Wisconsin
12	Apr 9^{th}	Wyoming
13	Apr 19^{th}	New York
14	Apr 26^{th}	Connecticut, Delaware, Maryland, Pennsylvania, Rhode Island
15	May 3^{rd}	Indiana
16	May 10^{th}	West Virginia
17	May 17^{th}	Kentucky, Oregon
18	Jun 7^{th}	California, Montana, New Jersey, New Mexico, North Dakota, South Dakota
19	Jun 14^{th}	District of Colombia

외하고 3,108개의 카운티에 대해 분석을 수행한다. 카운티의 경계를 기반으로 하여 공간 상관관계와 의존성을, 각 카운티별 예비 선거가 실시되었던 날짜를 기반으로 시간 의존성을 관찰한다.

본 연구에서 고려하는 미국의 대통령 선거는 간선제이며, 선거인단 독점 방식과 같은 특징을 가지고 있는 것으로 잘 알려져 있는데, 대통령 선거 이전의 예비 선거는 또 다른 방식으로 진행된다. 예비 선거는 투표를 통해 대의원 수를 확보하는 것으로, 선거를 실시하는 카운티에서 당에서 지지하는 후보자를 투표하면, 주 단위로 할당되어 있는 대의원 수를 후보들의 득표 비율에 비례하게 확보하며, 전체 합계가 먼저 과반을 달성하거나 가장 많은 대의원 수를 확보하면 대통령 후보로써 선출된다. 주에 속한 각 카운티들의 득표 수를 모두 합한 것으로 대의원 수를 확보하기 때문에 특정 카운티에서 많은 지지를 받는 것이 그 주에서 더 많은 표를 얻을 수 있는 중요한 요인 중 하나다.

또한 예비 선거는, 많은 유권자들이 동시에 표를 행사하기 때문에 한 번에 표결하는 방식으로 진행되는 선거가 아닌, 지역별로 선거일자가 다르며 시간차를 두고 지역별로 선거가 진행되는 방식이다. 따라서, 먼저 선거를 실시한 지역의 투표 결과가 이후 진행될 주변 지역의 투표 결과에 영향을 미칠 수도 있다. 이는 모멘텀 효과처럼 슈퍼화요일의 결과가 다음 선거 결과에 영향을 주게 되는 것과 같이 생각할 수 있다. 시간 차이는 Table 3.1의 민주당 예비 선거 일정을 참고하여 시점 혹은 시간에 따라 이산적(discrete) 혹은 연속적(continuous)인 방법을 생각해볼 수 있다.

종속변수로 사용될 민주당의 예비 선거 결과는 버니 샌더스와 힐러리 클린턴 두 후보에 대해서 버니 샌더스가 더 많은 표를 얻어 당선된 카운티는 0, 힐러리 클린턴이 더 많은 표를 얻어 당선된 카운티는 1로 설정한다. 총 3,108개의 카운티에서 버니 샌더스가 당선된 카운티는 1,460개, 힐러리 클린턴이 당선된 카운티는 1,648개로 구성되어 있다.

Table 3.2. Independent variables

Name	Variable
AGE65	Persons 65 years and over, percent, 2014
Female	Female persons, percent, 2014
Black	Black or African American alone, percent, 2014
Hispanic	Hispanic or Latino, percent, 2014
Income	Per capita money income in past 12 months, 2009–2013
Poverty	Persons below poverty level, percent, 2009–2013
Firms	Total number of firms, 2007
Retail	Retail sales per capita, 2007
Land	Land area in square miles, 2010
Popul	Population per square mile, 2010
Time	Spatially lagged dependent variable with time weight matrix and primary results
Spatial	Spatially lagged dependent variable with spatial weight matrix and primary results

연구에 필요한 카운티 단위의 인구통계학 및 사회경제적 데이터(demographic and socioeconomic data)는 US Department of Commerce의 Bureau of the Census (http://www.census.gov/quickfacts/)를 통해서, 51개 변수에 대한 자료를 제공받을 수 있다. 인구에 대한 비율 데이터들로는 연령대별인구, 여성인구, 인종별, 교육수준별, 최저 소득 인구 등이 있으며, 그 외에 1인당 평균 소득, 인종 별 회사 소유 수, 인구 밀도 등이 있다. Table 3.2는 Bureau of the Census에서 제공하는 사회경제적 요인들의 데이터 중에서 본 연구에 적용할 변수에 대해, 변수명과 그 의미를 설명하고 있다.

3.2. 공간 상관분석

민주당의 선거 결과와 이웃한 카운티들 간의 상관관계를 알아보기 위해 공간 자기상관(spatial autocorrelation)을 관찰한다. Kim 등 (2016)의 연구와 같이 종속변수가 연속형인 경우에는 Moran's I와 공간 회귀모형을 적용한 분석이 가능하다. 이와 다르게 종속변수가 선거 당선 여부와 같이 변수가 범주형인 경우에는 join-count 통계량을 통해 공간적인 패턴을 파악할 수 있다 (Cliff와 Ord, 1981). 이 때 공간 자기상관을 계산하기 위해서는 먼저 공간 가중치 행렬(spatial weight matrix)를 구성해야 한다. Join-count 분석을 통해 공간 의존성을 확인하였다면, 그 이후에 공간 의존성을 고려한 자기로지스틱 회귀모형에 적용하여 어느 후보를 더 지지했는가에 대한 모형을 추정할 수 있다. 이 때 시간의 의존성 여부를 확인하기 위해 공간과 비슷한 개념의 시간 가중치 행렬이 필요하다. 시간 가중치 행렬도 따로 적용하여 시간 의존성도 확인함으로써 시간과 공간을 고려한 자기로지스틱 회귀모형을 나타낼 수 있다.

3.2.1. Join-count 분석 이진(binary) 변수에 대해서 유의한 공간적 패턴이 있는지를 알아보기 위해 Join-count 통계량이 사용된다. 각 카운티마다 힐러리 클린턴이 당선되거나 버니 샌더스가 당선되는 두 가지 경우가 발생될 수 있고, 전자의 경우를 C, 후자의 경우를 S라 하자. 이 때, 카운티의 공통된 경계를 공유하는 이웃의 선거 결과를 같이 고려하면 세 가지 패턴의 결과가 관찰될 수 있다. 즉, 인접한 두 지역에서 모두 힐러리 클린턴이 당선되었을 경우 "CC", 서로 다른 후보가 당선되었을 경우 "CS", 그리고 모두 버니 샌더스가 당선되었을 경우 "SS"라는 패턴으로 총 세 가지 종류의 패턴이 있다. 이 중 이웃 간 서로 다른 후보가 당선되는 "CS"패턴의 횟수에 주목하는데, 먼저 지역별로 선거 결과가 서로 독립이어서 지역별로 더 지지하는 후보자가 누구인지는 공간적으로 랜덤하게 분포되어 있다는 귀무가설하에, 이웃 간에 다른 후보가 당선되는 "CS"패턴의 기대빈도수를 실제 선거 결과와 비교하여 가설 검정하

여 선거 결과의 공간적 패턴을 파악할 수 있다. 검정 통계량이 기각역에 포함된다면 귀무가설을 기각하고 대립가설이 채택되며, 지역별로 지지하는 후보자가 공간적으로 규칙이 있음을 확인할 수 있다.

먼저, 선거 결과가 공간적으로 랜덤하게 분포되어 있을 경우에 기대되는 이웃 간 선거 결과가 다르게 나오는 패턴의 횟수를 구한다. 서로 다른 후보가 당선되는 패턴의 기대되는 횟수와 분산은 다음과 같다. 이 때 w_{ij}는 공간 가중치 행렬의 원소, P_{CS}는 두 인접한 카운티 간에 서로 다른 후보가 당선되는 확률이다.

$$E(\mathrm{CS}) = \frac{1}{2}\Sigma_i\Sigma_j w_{ij} P_{\mathrm{CS}}, \qquad \sigma^2_{\mathrm{CS}} = E\left(\mathrm{CS}^2\right) - \{E(\mathrm{CS})\}^2 .$$

$E(\mathrm{CS})$를 추정하기 위해 먼저 힐러리 클린턴 당선 지역이 전체 카운티에서 차지하는 비율 P_C, 버니 샌더스 당선 지역이 차지하는 비율 P_S를 추정한다.

$$P_C = \frac{n_C}{n}, \qquad P_S = \frac{n - n_C}{n} = 1 - P_C.$$

전체 n개 카운티에서 힐러리 클린턴이 당선된 카운티의 수를 n_C, 버니 샌더스가 당선된 카운티의 수를 $n - n_C$라 한다. 여기서 선거 결과가 무작위로 분포되므로 각 지역에서 특정 후보가 당선되는 사건은 서로 독립이므로 두 인접한 카운티 간에 서로 다른 후보가 당선되는 확률 P_{CS}는 다음과 같다.

$$P_{\mathrm{CS}} = P_C\,(1 - P_C) + (1 - P_C)P_C = 2P_C(1 - P_C).$$

다음으로, 실제 선거 결과 분포도에서 이러한 "CS"패턴이 얼마나 많은 경우가 있는지를 도출한다. CS join-count 통계량은 다음과 같이 정의할 수 있다.

$$\mathrm{CS} = \frac{1}{2}\sum_i\sum_j w_{ij}(y_i - y_j)^2.$$

이 때, y_i는 i 카운티의 민주당의 후보 당선 여부를 이진 변수로 나타낸 것이며 $y_i = 0$일 때는 버니 샌더스, $y_i = 1$일 때는 힐러리 클린턴이 당선된 것을 의미한다. w_{ij}는 i 카운티와 j 카운티의 이웃 관계를 나타내는 가중치로 공간 가중치 행렬에 기반된다. 공간 가중치는 크게 경계(boundary) 기반과 거리(distance) 기반, 두 가지 방법으로 나누어지는데, 본 연구에서는 경계 기반 가중치의 방법 중에서 공통된 경계를 공유하는 카운티들을 이웃으로 설정하여 가중치를 부여하는 방법인 Rook contiguity를 적용하였다. 한편 w_{ij}는 다음과 같이 나타낼 수 있다.

$$w_{ij} = \begin{cases} 1, & l_{ij} \neq 0, \\ 0, & l_{ij} = 0, \end{cases} \qquad l_{ij} = \text{지역 } i\text{와 } j \text{ 사이가 공유하는 경계의 길이.}$$

CS join-count 통계량은 정규분포를 따르는 것으로 가정, 다음과 같은 검정 통계량 $z(\mathrm{CS}) = (\mathrm{CS} - E(\mathrm{CS}))/\sigma_{\mathrm{CS}}$을 도출하고 z값이 특정 유의수준에서 작을 경우 공간적으로 같은 선거 결과가 밀집되어 있는 것을 의미하며, z값이 유의하게 높을 경우 이웃한 지역들의 선거 결과가 반대로 나타나는 패턴이 생기는 것을 추론할 수 있다. z값이 기각역에 속하지 않는 경우에는 귀무가설을 기각할 수 없으며 공간적 자기상관관계가 없다고 판단할 수 있다 (Cliff와 Ord, 1981).

3.2.2. 자기로지스틱 모형(autologistic model) 범주형 분류(categorical classification)의 한 방법으로 로지스틱 회귀분석이 사용될 수 있다. 선거 예측에 있어서도 후보자의 당선 여부를 종속변수로 설정한다면 로지스틱 회귀분석을 적용할 수 있다. 본 연구에서는 다음과 같이 사회경제적 변수들을 설명변수로 하고 선거 결과를 종속변수로 한 로지스틱 회귀모형을 첫 번째 연구 모형으로 선택한다.

$$\log\frac{P(y_i = 1)}{P(y_i = 0)} = \sum_{k=1}^{p}\beta_k x_{ik}, \tag{3.1}$$

y_i는 지역 i에서 지지하는 후보자를 나타내는 반응변수, x_{ik}는 지역 i에 대한 총 p개 중 k번째 설명변수, β_k는 그 설명변수에 대해 추정된 회귀계수를 나타낸다.

만약 반응 변수가 서로 독립이 아니고 유의한 공간적 상관성이 있을 경우, Besag (1972, 1974)가 제시한 다음의 공간 자기로지스틱 모형이 적용될 수 있다.

$$\log \frac{P(y_i = 1)}{P(y_i = 0)} = \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{j \neq i} \eta_{ij} y_i,$$

$\eta = \{\eta_{ij}\}$는 의존계수로서 여러 지역의 반응 변수들 간 공간적 효과가 없을 경우 0, 양 혹은 음의 효과가 있을 경우에는 0이 아닌 값을 가진다. Pairwise 의존성을 가정하면 $Y = \{y_1, y_2, \ldots, y_n\}$의 결합 확률질량함수는

$$\pi(Y|\theta) = c(\theta)^{-1} \exp\left(Y^T X_i \beta + \frac{\eta}{2} Y^T A Y\right)$$

와 같은 형태가 된다. 여기서 θ는 모수로 이루어진 벡터 $(\beta_1, \beta_2, \ldots, \beta_p, \eta)^T$이고, $c(\theta)$는 intractable 정상화(normalizing) 함수이다. 그리고 $X_i\beta$는 $\sum_{k=1}^{p} \beta_k x_{ik}$를 나타내는 벡터, A는 인접행렬(adjacency matrix)이며 A_{ij}는 지역 i와 j의 인접 여부를 나타낸다.

한편, 결합 확률질량함수 식에서 $Q(Y|\theta) = Y^T X_i \beta + (\eta/2) Y^T A Y$로 두면 $\pi(Y|\theta)$는

$$\pi(Y|\theta) = \frac{\exp\left(Q(Y|\theta)\right)}{\sum\limits_{Y} \exp\left(Q(Y|\theta)\right)}$$

로 나타낼 수 있다.

Caragea와 Kaiser (2009)는 Besag가 제시한 모형에서 모수가 서로 교락(confound)되어 모형 식별에 문제가 있음을 지적하고 centered autologistic 모형을 다음과 같이 제시하였다 (Park, 2015). 본 연구에서는 고전의 autologistic model을 보완한 모형인 centered autologistic model을 두 번째 모형으로 선택하였다.

$$\log \frac{P(y_i = 1)}{P(y_i = 0)} = \sum_{k=1}^{p} \beta_k x_{ik} + \sum_{j \neq i} \eta_{ij} (y_j - \mu_j), \tag{3.2}$$

이 때 μ_j는 $\eta = 0$인 경우 반응변수 Y에 대한 기대값이며 아래와 같이 나타낸다.

$$\mu_j = E\left(Y_j | \eta = 0\right) = \frac{\exp(X_j\beta)}{1 + \exp(X_j\beta)}.$$

Centered autologistic model에 대해서 반응변수 Y에 대한 결합 확률질량함수는 다음과 같다.

$$\pi(Y|\theta) = c(\theta)^{-1} \exp\left(Y^T X_i \beta - \eta Y^T A \mu + \frac{\eta}{2} Y^T A Y\right),$$

이 때, μ는 μ_j를 벡터 형식으로 나타낸 것이다.

반응 변수의 공간적 의존성을 고려한 autologistic model에서 선거 지역이 가지는 시간적 속성과 후보자 지지 결과와의 연관성도 함께 분석하기 위해 본 연구에서는 자기로지스틱 모형에 시간 가중치 행렬을 적용하는 모형을 고려한다. 따라서 반응변수에 대한 시간적 의존성을 나타내는 항을 추가하며, 이 $p+1$번째 설명변수의 계수를 β_{p+1}라고 한다면 다음과 같은 세 번째 모형을 생각할 수 있다.

$$\log \frac{P(y_i = 1)}{P(y_i = 0)} = \sum_{k=1}^{p} \beta_k x_{ik} + \beta_{p+1} \mathrm{TIMEone}_i + \sum_{j \neq i} \eta_{ij} (y_j - \mu_j). \tag{3.3}$$

TIMEone$_i$은 앞서 공간 상관관계를 확인하기 위해 사용했던 공간 가중치 행렬에서 시간 차이로 지역 간의 가중치를 설정하는 시간 가중치 행렬을 사용하여 나타낸다. TIMEone$_i$의 시간 가중치 행렬은 투표 결과가 이전 시점에 실시한 선거와 연관성을 가지는 것으로 가정하여 이산적인 시간차를 고려하였다. 카운티 i의 선거 시점을 t_i이라고 한다면, 바로 이전 시점에 투표한 지역과의 관계를 1, 이외의 지역과는 0으로 설정했다. TIMEone$_i$은 다음과 같이 나타낼 수 있다.

$$\text{TIMEone}_i = \frac{\sum_j u_{ij} y_i}{\sum_j u_{ij}}, \quad u_{ij} = \begin{cases} 1, & t_i - t_j = 1, \\ 0, & \text{else}, \end{cases}$$

이때의 u_{ij}는 시간 가중치 행렬의 원소이다. TIMEone$_i$는 u_{ij}로 지역들 간의 시간관계를 설정한 경우에 대해서, 지역 i에 이웃한 지역들 j들 중에서 힐러리 클린턴을 지지한 지역의 비율을 의미하며, 이에 대한 모형의 계수를 추정하여 시간 의존성의 유의성과 반응 변수와의 관계를 확인할 수 있다.

시간의 의존성을 나타내기 위한 변수인 TIMEone$_i$은 실제 선거의 형식에 맞게 다양한 시간 가중치 행렬을 설정한 값으로 변수를 선택할 수 있다. Bartels (1987)이 주장한 모멘텀 효과를 참고하여 예비 선거 초반에 실시하게 되는 아이오와나 뉴햄프셔 주 등 4곳의 선거 결과는 슈퍼 화요일 선거 결과와, 슈퍼 화요일에 실시한 선거 결과가 이후의 투표와 관련이 있다는 가정하에 다른 시간 가중치 행렬을 적용한 TIMEtwo$_i$를 고려하였다. v_{ij}는 그 때의 시간 가중치 행렬의 원소이다. TIMEone$_i$대신 TIMEtwo$_i$를 추가한 네 번째 모형은 다음과 같이 나타낼 수 있다.

$$\log \frac{P(y_i = 1)}{P(y_i = 0)} = \sum_{k=1}^{p} \beta_k x_{ik} + \beta_{p+1} \text{TIMEtwo}_i + \sum_{j \neq i} \eta_{ij} (y_j - \mu_j), \tag{3.4}$$

$$\text{TIMEtwo}_i = \frac{\sum_j v_{ij} y_i}{\sum_j v_{ij}}, \quad v_{ij} = \begin{cases} 1, & t_i - t_j = 1 \cup \{(t_i - t_j > 0) \cap (t_i \leq 5)\} \cup \{(t_i - t_j > 0) \cap (t_j = 5)\}, \\ 0, & \text{else}, \end{cases}$$

일반적으로 로지스틱 회귀분석과 같은 경우 모수 추정을 위하여 최대 우도 추정법(maximum likelihood estimation)을 사용하나, 자기로지스틱 회귀분석의 경우에는 최대 우도 추정법에 의한 추정된 계수가 정확하지 않으며, 이를 보완한 방법으로 $\theta = (\beta_1, \beta_2, \ldots, \beta_p, \eta)^T$의 추정을 위해 유사-가능도 최대화 방법(maximum pseudo-likelihood estimation) 또는 마르코브 체인 몬테 카를로 방법(Markov Chain-Monte Carlo method)을 적용한다(Sherman 등, 2006). 본 연구에서는 유사-가능도 최대화 방법을 적용하여 자기로지스틱 회귀분석 모형 (3.2)의 모수를 다음과 같이 추정한다. 즉,

$$\tilde{\theta} = \arg\max l_{\text{PL}}(\theta)$$

이고, $l_{\text{PL}}(\theta) = \sum_i \log[\exp(y_i(X_i\beta + \eta \sum_{ji} y_j))/(1 + \exp(X_i\beta + \eta \sum_{ji} y_j))]$이다.

Centered 자기로지스틱 모형에서의 $l_{\text{PL}}(\theta)$가 다음과 같이 구할 수 있다.

$$l_{\text{PL}}(\theta) = Y^T (X\beta + \eta A(Y - \mu)) - \sum_i \log\left(1 + \exp\left(X_i\beta + \eta A_i(Y - \mu)\right)\right) \tag{3.5}$$

모형 (3.3), (3.4)의 경우 TIMEone$_i$, TIMEtwo$_i$와 같이 시간 의존성을 나타내는 변수를 모형 (3.2)의 $l_{\text{PL}}(\theta)$에 대한 식 (3.5)의 X와 같이 취급하여 모수를 추정하였다.

Table 4.1. Join count statistics for US counties

	E(CS)	CS	Standard error	z-value	p-value
Statistic	4353.42	1571.00	46.59	−59.73	<0.0001

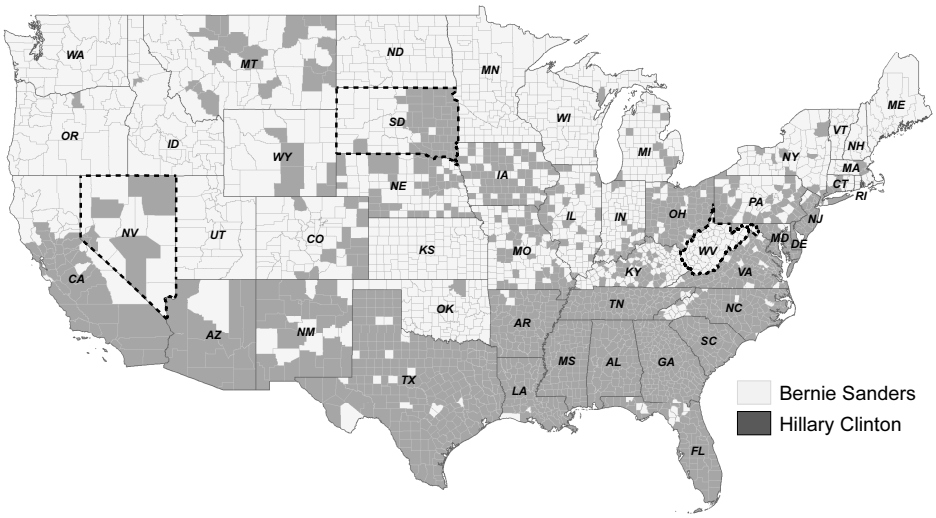

Figure 4.1. Quantile map of US primary results.

4. 분석 결과

4.1. 연구 모형 실행 결과

Join count statistics을 통해 미국 3,108개의 카운티들간의 공간적 상관관계를 파악하였으며, 그 결과는 Table 4.1에 제시하였다. 공간 가중치 행렬에서 서로 이웃한 관계를 가지는 8,736개 카운티 중에서 이웃한 카운티의 선거 결과와 다른 경우는 1,571개가 있었다. 이웃 간에 다른 후보가 당선되는 "CS" 패턴의 기대 빈도수와 그의 표준편차는 각각 4353.42, 46.59로 계산되었으며 이 때의 z-value는 −59.73가 된다.

결과적으로 민주당의 예비 선거 결과는 같은 후보자를 지지하는 카운티들끼리 군집을 이루고 있는 공간적인 패턴이 유의하다는 것을 설명해주고 있다. 선거 결과는 이웃한 카운티들끼리 같은 후보자를 지지하게 되는 공간적인 의존성이 존재한다는 것을 의미한다.

Figure 4.1에서 선거 결과를 토대로 더 많은 지지를 받은 후보자에 대한 지역별 분포도를 나타내었다. 대부분은 같은 주(state)내에서의 선거 결과가 비슷한 경향을 가지고 있다. 어두운 지역은 힐러리 클린턴, 밝은 지역은 버니 샌더스가 당선된 지역이다. 주로 북쪽 지역의 경우는 버니 샌더스, 남동쪽에는 힐러리 클린턴을 지지하는 카운티가 많이 있음을 확인할 수 있다. 힐러리 클린턴의 경우 큰 도시가 있는 주에서 버니 샌더스의 경우 그 외의 주에서 우세를 점하고 있는 것을 관찰할 수 있다. West Virginia의 경우에는 주변 주의 카운티들이 주로 힐러리 클린턴을 지지하는 카운티들로 둘러싸여 있지만 버니 샌더스를 지지하고 있음을 알 수 있었고, Nevada의 경우는 버니 샌더스를 지지하는 지역에 바로 접해 있지만 반대로 힐러리 클린턴을 지지하는 선거 결과가 나타나는 특징이 있었다. South Dakota의 경우는 주 내에서 서쪽은 버니 샌더스, 동쪽은 힐러리 클린턴을 지지하고 있는 다른 주와는 다른 특징을 보였다.

Table 4.2. Variance inflation factors about 10 independent variables of training data

Name	AGE65	Female	Black	Hispanic	Income	Poverty	Firms	Retail	Land	Popul
Variable	1.25	1.18	1.68	1.24	2.90	3.11	1.49	1.23	1.09	1.338

Table 4.3. Model 1: logistic regression

	Estimate	Std.	Error	z-value
(Intercept)	−10.660000*	1.384	−7.701	1.35E−14
AGE65	0.118100*	759.375	8.526	2.00E−16
Female	0.163400*	4258.865	5.955	2.60E−09
Black	0.237700*	403.027	17.067	2.00E−16
Hispanic	0.052080*	134.987	10.845	2.00E−16
Income	−0.000040*	1299.099	−2.011	0.0443
Poverty	−0.000630	793.874	−0.041	0.9672
Firms	−0.000001	102.625	−0.105	0.9167
Retail	−0.000030*	348.453	−2.484	0.0130
Land	−0.000090	143.916	−1.811	0.0702
Popul	0.000110	124.443	0.701	0.4834

*: p-value ≤ 0.05.

Join count statistics의 결과를 통해 예비 선거 결과는 공간 의존성이 존재하며 양의 상관관계가 있음을 파악할 수 있다. 따라서 공간변수를 고려한 자기로지스틱 회귀분석을 통해 어느 후보자를 지지하는데 있어서 공간 의존성이 어떤 의미를 갖는지 확인해 볼 수 있다.

4.2. 모형 추정

앞에서 제시했던 네 가지 모형을 실행하기 위해, 다수의 데이터를 모형의 훈련용(training)으로 적용하고, 나머지 데이터로 모형의 검증(validation)을 통해 모형을 비교하려 하였다. 따라서 3,108개의 카운티 데이터를 분할하는데, 선거를 실시한 전체 일자 19번 중 첫 번째인 2월 1일부터 16번째인 5월 10일까지의 2,664개 카운티의 데이터를 훈련용으로, 5월 17일, 6월 7일 그리고 6월 14일에 선거한 444개의 카운티의 데이터를 검증용으로 설정하였다. 먼저 훈련용 데이터로 추정한 계수들과 그 모형들을 분석하고, 도출된 모형을 검증용 데이터에 적용하여 모형의 설명력을 파악하였다.

먼저 모형 추정을 위해 사용될 훈련용 데이터의 10개의 설명변수에 대해 각 변수들끼리의 상관성을 확인하기 위해 분산팽창요인(variance inflation factor; VIF)을 계산하였다. Table 4.2에서 각 변수들에 대한 VIF를 확인할 수 있으며, 대부분 값이 1에 가까우며, 비교적 값이 큰 income, poverty변수의 경우도 대략 3에 가까운 값이 VIF값이 10이상인 변수는 없는 것으로 나타났다. 따라서 본 연구에서는 10개의 사회경제적 변수를 모두 선택하였다.

Table 4.3는 공간 의존성을 고려하지 않으며, 사회경제적 요소만을 설명변수로 고려하여 선거 결과에 대한 로지스틱 회귀분석을 실행한 결과이다. 65세 이상 인구, 여성 인구, 흑인 또는 히스패닉 유권자가 많은 지역일수록 버니 샌더스를 지지할 확률보다 힐러리 클린턴을 지지할 확률이 높음을 의미하며, 특히 흑인의 경우는 오즈비(odds ratio)가 $e^{0.2377} = 1.268$로 가장 관계가 두드러졌다. 1인당 수입과 1인당 소비는 어떤 후보자에게 투표하게 되는 것과 관계가 매우 적은 편이지만 1인당 수입과 소비가 클수록 힐러리 클린턴을 지지할 확률보다는 버니 샌더스를 지지할 확률이 높음을 의미한다.

Table 4.4는 로지스틱 회귀분석에 적용한 설명변수에 공간 의존성을 고려한 공차변수를 추가한 자기로지스틱 회귀분석의 결과이다. 자기로지스틱 회귀분석 결과, 로지스틱 회귀분석에서 유의하지 않았던

Table 4.4. Model 2: autologistic regression

	Estimate	Lower	Upper
(Intercept)	−6.143000*	−9.426000	−2.374000
AGE65	0.111100*	0.077320	0.149700
Female	0.081300*	0.019880	0.142500
Black	0.213500*	0.189500	0.237100
Hispanic	0.050740*	0.037160	0.062690
Income	−0.000005	−0.000044	0.000026
Poverty	−0.048610*	−0.081910	−0.023360
Firms	0.000008	−0.000002	0.000017
Retail	−0.000004	−0.000021	0.000017
Land	0.000271*	0.000063	0.000427
Popul	−0.000047	−0.000134	0.000563
Spatial	1.090000*	1.051000	1.167000

*: p-value ≤ 0.05.

Table 4.5. Model 3: autologistic regression with time weight matrix

	Estimate	Lower	Upper
(Intercept)	−6.619000*	−10.2200000	−2.578000
AGE65	0.111600*	0.0773000	0.144000
Female	0.080580*	0.0020740	0.163000
Black	0.211200*	0.1825000	0.241600
Hispanic	0.048760*	0.0352600	0.063040
Income	−0.000005	−0.0000470	0.000026
Poverty	−0.051170*	−0.0868800	−0.023940
Firms	0.000009	−0.0000025	0.000025
Retail	−0.000006	−0.0000310	0.000018
Land	0.000280*	0.0001500	0.000433
Popul	−0.000048	−0.0000900	0.000300
Time	0.956100*	0.4093000	1.765000
Spatial	1.069000*	1.0270000	1.161000

*: p-value ≤ 0.05.

poverty level이 유의한 변수로 파악되었으며, 반대로 로지스틱 회귀분석에서 유의했던 1인당 소비는 유의하지 않은 점을 알 수 있다. 로지스틱 회귀분석과 마찬가지로 65세 이상 인구, 여성 인구, 흑인 및 히스패닉 유권자가 많은 지역일수록 버니 샌더스를 지지할 확률보다 힐러리 클린턴을 지지할 확률이 높음을 파악하였다. 지역 면적은 모형에서 추정된 계수의 값은 적지만 넓은 지역일수록 버니 샌더스보다는 힐러리 클린턴을 지지함을 보여주고 있다. 모형에서 추정된 계수가 음의 부호를 갖는 변수는 poverty이며, 이는 가난한 인구의 비율이 많을수록 힐러리 클린턴보다는 버니 샌더스를 지지함을 파악할 수 있었다. 공간 의존성을 나타내는 spatial의 계수는 1.09이며, $e^{1.09} = 2.974$로 힐러리 클린턴을 지지하는 이웃 지역이 많을수록 odds가 2.974배씩 증가하게 되는 것을 알 수 있다.

Table 4.5는 자기로지스틱 모형에 이전 시점의 선거 결과가 다음 선거와 관련이 있는 시간 의존성을 고려한 회귀분석 결과를 보여주고 있다. 분석 결과, 65세 이상 인구, 여성 인구, 흑인 또는 히스패닉 유권자가 많은 지역일수록, 넓은 지역일수록 버니 샌더스보다는 힐러리 클린턴을 지지함을 보여주고 있으며, 가난한 인구 비율이 높을수록 힐러리 클린턴보다는 버니 샌더스를 지지함을 파악할 수 있었다. 시간 의

Table 4.6. Model 4: Autologistic regression with time weight matrix 2

	Estimate	Lower	Upper
(Intercept)	−4.030000*	−7.9800000	−0.820000
AGE65	0.113000*	0.0787000	0.155000
Female	0.080000*	0.0008000	0.171000
Black	0.215000*	0.1910000	0.256000
Hispanic	0.050400*	0.0363000	0.065200
Income	−0.000005	−0.0000500	0.000032
Poverty	−0.046700*	−0.0940000	−0.009760
Firms	0.000009*	0.0000003	0.000017
Retail	−0.000003	−0.0000250	0.000014
Land	0.000257*	0.0001070	0.000433
Popul	−0.000050	−0.0001200	0.000550
Time	−3.120000*	−4.0700000	−2.240000
Spatial	1.100000*	1.0600000	1.160000

*: p-value ≤ 0.05.

존성을 의미하는 time의 계수는 0.9561, 공간 의존성을 나타내는 spatial의 계수는 1.069로 추정되었다. 이전 시점에 진행되었던 선거에서 힐러리 클린턴이 당선된 지역이 많을수록 오즈비가 $e^{0.9561} = 2.602$로 이전 시점에 힐러리 클린턴이 우세한 카운티가 많을수록 이후의 카운티에서도 힐러리 클린턴이 우세할 확률이 높아지는 것으로 나타났다.

Table 4.6는 자기로지스틱 모형에 이전 시점 및 슈퍼 화요일의 선거 결과에 대한 시간 의존성을 고려한 회귀분석의 추정 결과를 정리하였다. 세 번째 모형에서 시간 가중치 행렬을 다르게 설정한 이번 모형은 결과 중 Firms변수와 Time변수에서 차이점이 나타났다. 앞선 모형에서는 유의하지 않았던 사업장 수가 많을수록 힐러리 클린턴을 지지하게 되는 약한 양의 상관관계로 나타났다. 슈퍼 화요일 선거 결과의 영향력을 크게 고려한 시간 의존성은 세 번째 모형에서는 양의 상관관계로 나타났던 것과는 다르게 음의 상관관계로 나타났다.

4.3. 모형의 적합성(goodness-of-fit) 비교

예비 선거 결과와 설명 변수들간의 상관관계를 분석한 네 가지 회귀분석 모형을 비교하기 위해 모형의 정확성(오분류율)과 ROC curve를 제시하였다. Figure 4.2에는 훈련용 데이터로 추정한 네 가지 모형에 대한 ROC curve를 하나의 plot에 나타내었고, 검은색으로 표시된 로지스틱 모형의 경우는 나머지 자기로지스틱 모형과 비교할 때 area under the curve (AUC)가 작은 것을 확인할 수 있다.

Table 4.7에는 훈련용 데이터에 대한 각 모형의 AUC값과 가장 적은 오분류율(misclassification rate)과 그 때 π_i의 임계값(threshold value)을 정리하였다. AUC값과 오분류율 모두 시간을 고려하지 않고 공간 의존성을 고려한 자기로지스틱 모형이 가장 좋은 모형으로 파악되었다.

Figure 4.3에는 검증용 데이터로 추정한 네 가지 모형에 대한 ROC curve를 하나의 plot에 나타냈다. Figure 4.2에서 관찰한 ROC curve와 마찬가지로 로지스틱 모형은 나머지 모형과 비교하여 좋은 결과를 보여주지 못하였다.

Table 4.8에는 검증용 데이터에 대한 각 모형의 AUC값과 가장 적은 오분류율과 그 때 π_i의 임계값을 정리하였다. 슈퍼 화요일에 대한 모멘텀 효과를 고려했던 네 번째 모형의 AUC값이 가장 컸으며, 오분류율 역시 슈퍼 화요일에 대한 모멘텀 효과를 고려한 자기로지스틱 모형인 네 번째 모형이 가장 적음을 확인할 수 있었다.

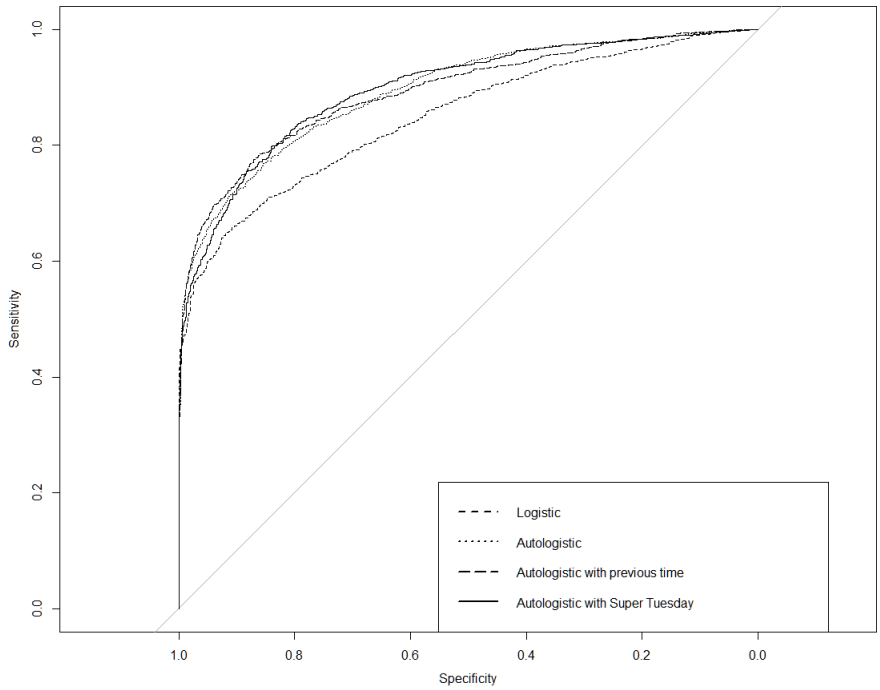

Figure 4.2. ROC curve of four models about training data. Dashed line is logistic model, dotted line is autologistic model, long dash line is autologistic model with previous time and solid line is autologistic model with Super Tuesday.

Table 4.7. AUC and misclassification rate (threshold): comparison of the four models with the training data

	Logistic	Autologistic	Autologistic with previous time	Autologistic with super Tuesday
AUC-Training	0.8489	0.9691	0.9659	0.9620
Misclassification rate (threshold)	0.2305 (0.61)	0.0930 (0.57)	0.1017 (0.55)	0.1059 (0.55)

5. 결론

본 연구에서는 2016년 미국 민주당 예비 선거의 지역에서 더 많은 지지를 받은 후보자를 사회경제적 변수, 공간 의존성 변수, 시간 의존성 변수와의 연관성을 고려한 로지스틱 및 공간 자기상관 로지스틱 모형을 적용하여 어떤 유권자들이 어느 후보자를 더 지지하게 되는지 알아보았다.

사회경제적 변수들 중 연령, 성별, 인종의 경우는 선행연구에서도 선거 결과와 연관성이 있음을 확인되어왔다. 연구 분석 결과, 각 카운티의 65세 이상 인구 비율, 여성 비율, 흑인 비율이 높을수록 버니 샌더스를 지지할 확률보다는 힐러리 클린턴을 지지할 확률이 높음을 파악할 수 있었다. 이외에 1인당 수입, 1인당 지출, 저소득 인구 비율, 사업장 수는 모형에 따라 유의성에 차이가 있었으나, 유의한 상관관계의 대부분은 음의 상관관계를 가졌다. 즉 1인당 수입, 지출, 저소득 인구 비율이 높을수록, 사업장 수가 많을수록 힐러리 클린턴을 지지할 확률보다는 버니 샌더스를 지지할 확률이 높음을 확인할 수 있었다.

본 연구에서는 민주당의 예비 선거 결과가 공간적인 상관관계가 있으며, 무작위로 분산되어 있는 형태가

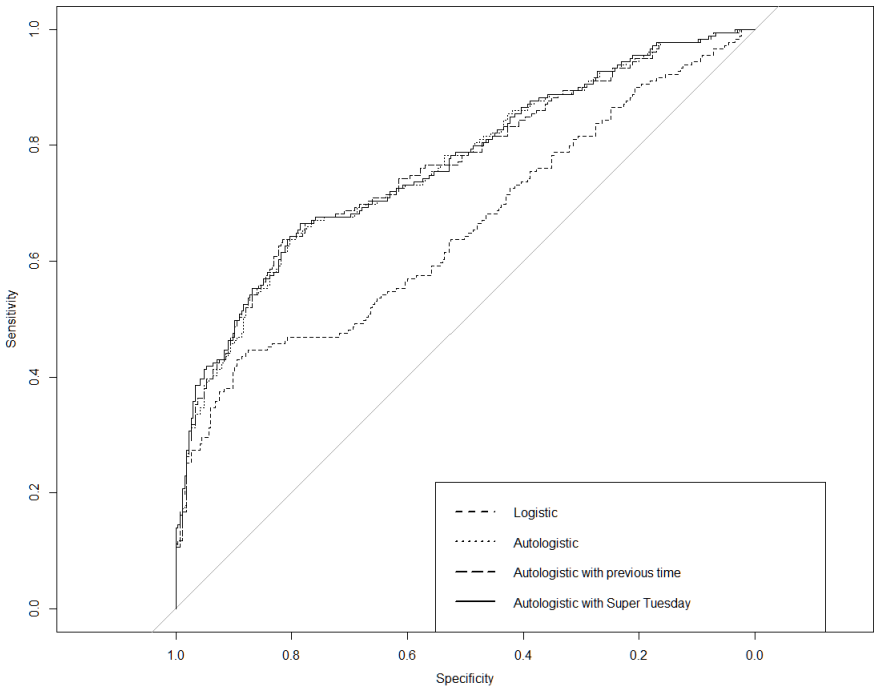

Figure 4.3. ROC curve of four models with the validation data. Dashed line for logistic model, dotted line for autologistic model, long dash line for autologistic model with previous time, and solid line for autologistic model with Super Tuesday.

Table 4.8. AUC and misclassification rate (threshold): comparison of the four models with the validation data

	Logistic	Autologistic	Autologistic with previous time	Autologistic with super Tuesday
AUC-Training	0.6547	0.9093	0.9004	0.9120
Misclassification rate (threshold)	0.2928 (0.54)	0.1464 (0.57)	0.1577 (0.58)	0.1441 (0.57)

아닌 두 결과가 상반된 지역에서 군집을 이루고 있음을 확인하였다. 자기 로지스틱 모형에서 추정된 공간 의존성 변수의 계수도 양의 값으로 계산된 것으로 보아, 힐러리 클린턴을 지지했던 이웃 카운티가 많은 지역은 힐러리 클린턴을, 버니 샌더스를 지지했던 이웃 카운티가 많은 지역은 버니 샌더스를 지지했음을 알 수 있다.

시간 차이를 두고 실시하는 예비 선거는 이전에 있었던 선거 결과가 이후에 진행된 선거 결과와 연관이 있음을 의미하는 시간 의존성도 존재하였다. 이전 선거에서 힐러리 클린턴이 당선된 카운티가 많은 경우에는 이후에도 힐러리 클린턴에게 투표할 확률이 높을 것으로 추정되어 초반에 지지율을 확보하는 것이 이후의 득표에도 도움을 줄 수 있게 된다고 볼 수 있다. 하지만 가장 많은 선거인단이 결정되어 예비 선거에서 가장 중요한 날로 꼽히는 슈퍼 화요일의 결과가 이후의 선거 결과에 영향을 준다는 모멘텀 효과를 적용한 경우는 오히려 이전에 지지하지 않았던 후보에게 투표하게 될 확률이 높게 나타났다. 민주당의 후보자는 슈퍼 화요일의 선거 결과가 좋지 않아도 이후에 충분히 득표하며 당선 가능성을 높일 수 있는 것이다.

본 연구에서 사용한 모형은 두 후보자가 더 우세했던 지역을 분류한 것인데, 이 결과를 이용하여 많은 표가 걸려 있는 주요 지역의 결과를 참고하거나 각 지역별로 더 우세했던 지역들과 할당된 대의원 수를 곱하여 기대 득표수를 계산함으로써 선거 예측을 해볼 수 있을 것이다. 또한, 본 연구에서의 시간 및 공간 의존성을 확인하기 위해 모형에 적용되는 가중치 행렬은 연구 목적에 따라 혹은 더 근거 있는 연구를 위해 다양하게 적용할 수 있다. 이번 연구에서 제시한 것보다 실제 선거 형태와 가까운 행렬을 제시하여 분석에 적용할 수 있다면, 시간 및 공간 의존성의 설명력이 더 커질 것이며 선거와 연관된 변수 탐색과 결과 예측에 도움을 줄 수 있을 것이다.

References

Abramowitz, A. I. (2008). Forecasting the 2008 presidential election with the time-for-change model, *PS: Political Science and Politics*, **41**, 691–695.

Augustin, N. H., Mugglestone, M. A., and Buckland, S. T. (1996). An autologistic model for the spatial distribution of wildlife, *Journal of Applied Ecology*, **33**, 339–347.

Austin, M. P. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling, *Ecological Modelling*, **157**, 101–118.

Bartels, L. M. (1987). Candidate choice and the dynamics of the presidential nominating process, *American Journal of Political Science*, **31**, 1–30.

Besag, J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data, *Journal of the Royal Statistical Society Series B (Methodological)*, **34**, 75–83.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society B (Methodological)*, **36**, 192–236.

Buckland, S. T. and Elston, D. A. (1993). Empirical models for the spatial distribution of wildlife, *Journal of Applied Ecology*, **30**, 478–495.

Caragea, P. C. and Kaiser, M. S. (2009). Autologistic models with interpretable parameters, *Journal of Agricultural, Biological, and Environmental Statistics*, **14**, 281–300.

Cho, S. D. (2009). Voter choice in primary: the 2008 democratic primary election in the United States, *Korean Political Science Review*, **43**, 193–214.

Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes: Models & Applications*, Pion, London.

Gumpertz, M. L., Graham, J. M., and Ristaino, J. B. (1997). Autologistic model of spatial pattern of Phytophthora epidemic in bell pepper: effects of soil variables on disease presence, *Journal of Agricultural, Biological, and Environmental Statistics*, **2**, 131–156.

He, F., Zhou, J., and Zhu, H. (2003). Autologistic regression model for the distribution of vegetation, *Journal of Agricultural, Biological, and Environmental Statistics*, **8**, 205–222.

Kim, D. H., Kang, K. Y., and Sohn, S. Y. (2016). Spatial pattern analysis of CO2 emission in Seoul metropolitan city based on a geographically weighted regression, *Journal of the Korean Institute of Industrial Engineers*, **42**, 96–111.

Kim, D. Y. (2012). White voters' choice in the 2008 U.S. presidential election, *Journal of International Area Studies*, **16**, 3–22.

Kim, J. W. (2013). The "44% controversy" over the 2004 presidential election, *The Korean Journal of American History*, **38**, 219–248.

Lewis-Beck, M. S. (2005). Election forecasting: principles and practice, *The British Journal of Politics and International Relations*, **7**, 145–164.

Lewis-Beck, M. S. and Tien, C. (2008). Forecasting presidential elections: when to change the model, *International Journal of Forecasting*, **24**, 227–236.

Norpoth, H. (2004). From primary to general election: a forecast of the presidential vote, *Political Science and Politics*, **37**, 737–740.

Norrander, B. (1993). Nomination choices: caucus and primary outcomes 1976–88, *American Journal of Political Science*, **37**, 343–364.

Park, J. (2015). Review of spatial linear mixed models for non-gaussian outcomes, *Korean Journal of*

Applied Statistics, **28**, 353–360.

Sherman, M., Apanasovich, T. V., and Carroll, R. J. (2006). On estimation in binary autologistic spatial models, *Journal of Statistical Computation and Simulation*, **76**, 167–179.

Steger, W. P. (2007). Who wins nominations and why? An updated forecast of the presidential primary vote, *Political Research Quarterly*, **60**, 91–99.

Wu, H. and Huffer, F. R. W. (1997). Modelling the distribution of plant species using the autologistic regression model, *Environmental and Ecological Statistics*, **4**, 31–48.

미국 대통령 예비선거에 적용한 시공간 의존성을 고려한 자기로지스틱 회귀모형 연구

염호정[a] · 이원경[a] · 손소영[a,1]

[a]연세대학교 정보산업공학과

(2017년 1월 5일 접수, 2017년 2월 22일 수정, 2017년 3월 16일 채택)

요 약

미국 대통령 예선은 선거인단이 시차를 두고 여러 회에 걸쳐 진행되는 특징이 있음에도 많은 연구가 진행되지 않았다. 본 연구에서는 다양한 자기로지스틱 모형을 통해 미국 대통령 예비선거 결과와 사회경제적 변수간의 시공간 의존성의 관계를 파악하고자 한다. 2016년 데이터에 적용한 분석결과 각 카운티의 노년층, 흑인, 여성 그리고 히스패닉 인구 비율이 높은 지역일수록 힐러리 클린턴을 지지할 확률이 높은 것으로 나타났다. 또한, 주변 카운티에서 많은 지지를 받은 후보가 이웃 지역에서도 많이 지지를 받을 확률이 높고 이전 선거에서 많은 지지를 받는 것과 다음 선거 지역의 결과 간의 상관관계도 확인되었다. 시공간 의존성을 알아보기 위한 모형 중에서 슈퍼화요일의 선거 결과가 이후 선거와 관련이 있다고 가정한 모형의 설명력이 가장 높은 것으로 판명되었다.

주요용어: 미국 대통령 예비선거, autologistic 모형, 공간 의존성, 시간 의존성

본 연구는 '국토교통부 국토공간정보연구사업 국토공간정보의 빅데이터 관리, 분석 및 서비스 플랫폼 기술개발(17NSIP-B081011-04)과제'의 연구비 지원에 의해 연구되었음.

[1]교신저자: (03722) 서울특별시 서대문구 연세로 50, 연세대학교 정보산업공학과. E-mail: sohns@yonsei.ac.kr