Communications for Statistical Applications and Methods
2019, Vol. 26, No. 3, 261–272

https://doi.org/10.29220/CSAM.2019.26.3.261
Print ISSN 2287-7843 / Online ISSN 2383-4757

Interval prediction on the sum of binary random variables indexed by a graph

Seongoh Park[a], Kyu S. Hahn[b], Johan Lim[a], Won Son[1,c]

[a]Department of Statistics, Seoul National University, Korea;
[b]Department of Communication, Seoul National University, Korea; [c]The Bank of Korea, Korea

Abstract

In this paper, we propose a procedure to build a prediction interval of the sum of dependent binary random variables over a graph to account for the dependence among binary variables. Our main interest is to find a prediction interval of the weighted sum of dependent binary random variables indexed by a graph. This problem is motivated by the prediction problem of various elections including Korean National Assembly and US presidential election. Traditional and popular approaches to construct the prediction interval of the seats won by major parties are normal approximation by the CLT and Monte Carlo method by generating many independent Bernoulli random variables assuming that those binary random variables are independent and the success probabilities are known constants. However, in practice, the survey results (also the exit polls) on the election are random and hardly independent to each other. They are more often spatially correlated random variables. To take this into account, we suggest a spatial auto-regressive (AR) model for the surveyed success probabilities, and propose a residual based bootstrap procedure to construct the prediction interval of the sum of the binary outcomes. Finally, we apply the procedure to building the prediction intervals of the number of legislative seats won by each party from the exit poll data in the 19^{th} and 20^{th} Korea National Assembly elections.

Keywords: binary sum, exit poll, graph indexed variables, Korea National Assembly election, prediction interval, residual bootstrap, spatial auto-regressive model

1. Introduction

In this paper, we are interested in the interval prediction of the sum of dependent binary random variables indexed by a graph. Suppose an undirected graph $G = (V, E)$ is given, where V is the set of vertices and E is the set of edges $e = (v, w)$ with $v, w \in V$, and observe data $\{(p_v, X_v), v \in V\}$ on the graph. The observation p_v of P_v is the success probability for the binary outcome on the vertex v and X_v is an exogenous covariate that influences P_v. The statistic we are interested in is the weighted sum of binary random variables Y_vs,

$$\mathrm{T} = \sum_{v \in V} w_v Y_v, \tag{1.1}$$

where w_v is a pre-decided weight for the vertex v and, given $\{(P_v, X_v), v \in V\}$, Y_vs are independently from a Bernoulli distribution with success probabilities p_vs. We construct $100(1-\alpha)\%$ prediction interval of T when we have observations $\{(p_v, X_v), v \in V\}$.

[1] Corresponding author: Economist, The Bank of Korea, 67, Sejong-daero, Jung-gu, Seoul 04514, Korea.
E-mail: son.won@gmail.com

Published 31 May 2019 / journal homepage: http://csam.or.kr

The problem above often arises in prediction problem in various elections in many countries. One example is the United States Electoral College for the US presidential election, where v is the index of the state which forms the graph along with its spatial location; w_v is the number of Electoral College (EC) vote of the state v; P_v is the winning probability of a candidate of interest at state v that is available from the survey or exit poll; and X_v is the extra covariate that can influence on P_v. We are interested in predicting the number of EC votes won by the candidate using the survey results. Another example, from which this paper is actually motivated, is the election for the Korean National Assembly (KNA), where v is the election district that forms a graph with its spatial location; w_v is set as 1 (one seat for one district); and T is the number of congress seats won by a party of interest in the election. The success probabilities P_vs are evaluated from the exit poll, and X_vs are chosen to explain the regionalism (whose existence in Korean politics is known for many decades). Here, we are interested in making interval prediction for the number of congressional seats taken by a party.

A traditional and popular method to construct the prediction interval of T is by assuming that P_vs are known constants and generating many independent (over the vertex set V) Bernoulli random variables with success probabilities P_vs. This method has been the practice for the exit polls for KNA election since 2008's election. However, the data example in Section 4 shows that the independent Bernoulli method disregards both the randomness of P_vs and the spatial correlations among them. In sequel, it underestimates the variance of T and results in a shorter prediction interval. The variance of T is decomposed as

$$\begin{aligned}\mathrm{var}(\mathrm{T}) &= \mathrm{var}\left\{\mathrm{E}(\mathrm{T}|\mathbf{P})\right\} + \mathrm{E}\left\{\mathrm{var}(\mathrm{T}|\mathbf{P})\right\} \\ &= \mathrm{var}\left(\sum_{v\in V} P_v\right) + \mathrm{E}\left\{\sum_{v\in V} P_v\,(1-P_v)\right\},\end{aligned}$$

where the existing methods disregard the first term and approximate the second to $\sum_{v\in V} p_v(1-p_v)$ with the surveyed supporting rate p_vs. Thus, the existing methods underestimate the variance of T, if P_vs are random.

This is one of many reasons for the failure of the exit poll for the KNA election. We remark that the exit poll for the KNA election starts in year 2004 (the 17^{th} election) and the prediction interval for the number of the congressional seats is first given in year 2008 (the 18^{th} election).

In this paper, we propose a new method to build a prediction interval of the sum statistic T. Our new way is a resampling based procedure. It assumes the spatial auto-regressive (AR) model for P_vs (more precisely for $Z_v = \log\{P_v/(1-P_v)\}$) and adapts the residual bootstrap to get re-samples of T. If $P_v = 0$ or 1, the logarithm is not well-defined. We may use the perturbed logit function defined by $\log(a/(1-a))$ if $P_v = 0$ and $\log((1-a)/a)$ if $P_v = 1$ where a is a very small constant such as $1e-10$. The prediction interval is directly from the resamples.

The new method is applied to the exit polls of the 19^{th} and 20^{th} KNA elections and compared with the independent Bernoulli method. The new method provides a wider interval than the independent Bernoulli method as expected, but contains the true number of seats (observed as the outcomes of the election) within the interval in both the 19^{th} and 20^{th} elections.

The remainder of the paper is organized as follows. In Section 2, we introduce the spatial AR model assumed for the observations $\{(p_v, X_v), v \in V\}$ and the iterative procedure to estimate the model parameters. The bootstrap procedure to build the prediction interval for T is proposed in Section 3. In Section 4, we start the section with a brief introduction to the exit poll and the history of the exit poll for the KNA election. We then apply our method to the exit polls of the 19^{th} and 20^{th} KNA elections, and compare the results to those of the independent Bernoulli method. In Section 5, we conclude the

paper with a brief summary of the work.

2. Spatial model and estimation

In this section, we introduce the spatial AR model and the estimation procedure of the model parameters. Recall that $G = (V, E)$ is the graph on which $\{(P_v, X_v), v \in V\}$ are defined. Suppose we let $Z_v = \log\{P_v/(1 - P_v)\}$, and $N(v) = \{w \in V | (v, w) \in E\}$ be the neighborhood of the vertex v on the graph. For Z_v, we assume the spatial AR model, that is,

$$\begin{aligned} Z_v &= \beta_0 + X_v^{\mathrm{T}}\boldsymbol{\beta} + U_v, \\ U_v &= \sum_{w \in N(v)} \rho_{v.w} U_w + \epsilon_v, \end{aligned} \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$ (assume the dimension of the exogenous covariate vector is p) is the regression coefficient vector, $\rho_{v.w} \in \mathbb{R}$ is the spatial AR coefficient, and ϵ_vs are independent measurement errors having mean 0 and variance σ_ϵ^2. The model above is one of popular models popular models in spatial regression (Anselin, 1988; Baltagi *et al.*, 2003; Arnold and Wied, 2010). The model has too many parameters and is rarely estimable from the observation. In practice, we make further structural assumption on the spatial AR coefficients $\{\rho_{v.w},\ v, w \in V\}$. Some examples are (i) coefficients are constant as $\rho_{v.w} = \rho$ for all $v, w \in V$, (ii) $\rho_{v.w} = \rho/d_v$ with $d_v = |\{w \in V,\ (v, w) \in E\}|$, or (iii) $\rho_{v.w} = \rho_1 \mathrm{I}\{w \in N_1(v)\} + \rho_2 \mathrm{I}\{w \in N_2(v)\}$ with different neighborhoods $N_1(v), N_2(v)$ of v. In below, for notational simplicity, we assume $\rho_{v.w} = \rho$ for all $v, w \in V$, that is, the case (i), together with an assumption that $\mathrm{I} - \mathrm{A}(\rho)$ is positive-definite. The condition in terms of the matrix is typically made in many literatures (Anselin, 1988; Lee, 2002), which naturally restricts the feasible range of ρ. It is required to guarantee a well-defined covariance matrix and to define Z_v well from U_v in (2.1), or vice versa.

When the observations $\{(p_v, X_v), v \in V\}$ (equivalently, $\{(z_v, X_v), v \in V\}$) are given, a computationally attractive procedure to estimate the model parameters is the iterative least square method (Hordijk, 1974; Cochrane and Orcutt, 1949; Ord, 1975), which iteratively updates the estimate of $(\beta_0, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ and $(\rho, \sigma_\epsilon^2)$. In the first step, given the estimate of ρ and σ_ϵ^2, the variance-covariance matrix of $\mathbf{U} = (U_1, \ldots, U_{|V|})^{\mathrm{T}}$ in (2.1) is

$$\mathrm{var}(\mathbf{U}) = (\mathrm{I} - \mathrm{A}(\rho))^{-2}\sigma_\epsilon^2,$$

where $\mathrm{A}(\rho)$ is a $|V| \times |V|$ symmetric adjacency matrix having ρ at the $(u, v)^{th}$ element if $(u, v) \in E$, and 0, otherwise. The estimate of the regression coefficient can then be obtained by solving the generalized least squares problem as

$$\left(\hat{\beta}_0, \hat{\boldsymbol{\beta}}^{\mathrm{T}}\right)^{\mathrm{T}} = \left\{\mathbf{X}^{\mathrm{T}}\mathrm{var}(\mathbf{U})^{-1}\mathbf{X}\right\}^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{var}(\mathbf{U})^{-1}\mathbf{z}, \tag{2.2}$$

where $\mathbf{X}$ is a $|V| \times (p + 1)$ matrix whose v^{th} row is $(1, X_v^{\mathrm{T}})$ and $\mathbf{z} = (z_1, z_2, \ldots, z_{|V|})^{\mathrm{T}}$. In the second step, to update ρ, we do the multiple regression, where the response is the residual from the first step $\tilde{U}_v = z_v - \hat{\beta}_0 - X_v^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ and the covariate $\sum_{w \in N(v)} \tilde{U}_v$ for $v \in V$. Here, the covariate vector can be changed depending on the structural assumption on spatial AR coefficients $\rho_{v.w}$s. We iterate the above two steps until the parameter estimates converge.

3. Prediction interval with spatial bootstrap

In this section, we introduce the new method to build the $100(1-\alpha)\%$ prediction interval for the sum statistic using the spatial bootstrap. For simplicity of notation, we take off hats from parameter estimates unless it is confusing. Our procedure begins by computing residuals

$$\epsilon_v = z_v - \beta_0 - X_v^{\mathrm{T}}\boldsymbol{\beta} - \rho \sum_{w\in N(v)} \left(z_w - \beta_0 - X_w^{\mathrm{T}}\boldsymbol{\beta}\right), \quad v \in V. \tag{3.1}$$

We let ϵ_v^* be a bootstrapped residual which is sampled uniformly from $\{\epsilon_w, w \in V\}$ for every $v \in V$. Then, a resampled copy can be obtained

$$\mathbf{z}^* = \mathbf{X}\begin{pmatrix}\beta_0\\ \boldsymbol{\beta}\end{pmatrix} + (\mathrm{I} - \mathrm{A}(\rho))^{-1}\boldsymbol{\epsilon}^*, \tag{3.2}$$

where $\boldsymbol{\epsilon}^* = (\epsilon_1^*, \ldots, \epsilon_{|V|}^*)^{\mathrm{T}}$, which is a vector representation of equation (2.1). Now, given success probabilities $\{p_v^*,\ v \in V\}$, the sum of Bernoulli variables can be evaluated empirically, which is easy to implement using existing software. Exploiting the Monte Carlo method, we obtain $|V|$ many Bernoulli trials $\{Y_v^*,\ v \in V\}$, the sum of which is $\mathrm{T}^* = \sum_{v\in V} Y_v^*$. We repeat this procedure as many times as required, say B times. To make it clear, it is summarized in Algorithm 1.

We propose a method to build the prediction interval of T using the empirical $100(\alpha/2)^{th}$ and $100(1-\alpha/2)^{th}$ percentiles of the bootstrapped sums $\mathrm{T}^{(b)} = \sum_{v\in V} Y_v^{(b)}$, $b = 1, 2, \ldots, B$, denoted by $\mathrm{T}_{\alpha/2}$ and $\mathrm{T}_{1-\alpha/2}$. Thus, our proposal is

$$\left[\mathrm{T}_{\frac{\alpha}{2}},\ \mathrm{T}_{1-\frac{\alpha}{2}}\right]. \tag{3.3}$$

Our proposal can be understood as a procedure with a t-type statistic that is

$$t_{\mathrm{ind}} = \frac{\mathrm{T} - \hat{\mu}(\mathrm{T})}{\widehat{\mathrm{se.ind}}(\mathrm{T})} = \frac{\mathrm{T} - \sum_{v\in V} p_v}{\sqrt{\sum_{v\in V} p_v(1-p_v)}}, \tag{3.4}$$

where T is the number of seats we want to predict, $\hat{\mu}(\mathrm{T}) = \sum_{v\in V} p_v$ is the estimated expected number of seats from the observations $\{p_v, v \in V\}$, and $\widehat{\mathrm{se.ind}}(\mathrm{T})$ is the estimated standard deviation of T under the misspecified independent assumption among P_vs. The distribution of the above t_{ind} is unknown even either under normality assumption on P_vs or in asymptotic due to the correlation of P_vs over the graph $G = (V, E)$. We propose the spatial bootstrap samples $\{p_v^{(b)}, v \in V\}$ to approximate the distribution t_{ind}. Using $\{\mathrm{T}^{(b)}, b = 1, 2, \ldots, B\}$ obtained as described in Algorithm 1, we evaluate the statistics

$$t_{\mathrm{ind}}^{(b)} = \frac{\mathrm{T}^{(b)} - \hat{\mu}(\mathrm{T})}{\widehat{\mathrm{se.ind}}(\mathrm{T})}, \quad b = 1, 2, \ldots, B,$$

where $\hat{\mu}(\mathrm{T})$ and $\widehat{\mathrm{se.ind}}(\mathrm{T})$ are those defined in (3.4). Suppose t_u and t_ℓ are the upper, lower $100(\alpha/2)^{th}$ empirical quantiles of $\{t_{\mathrm{ind}}^{(b)}, b = 1, 2, \ldots, B\}$. Then, our proposal (3.3) is equivalent to

$$\left[\hat{\mu}(\mathrm{T}) + t_\ell\,\widehat{\mathrm{se.ind}}(\mathrm{T}),\ \hat{\mu}(\mathrm{T}) + t_u\,\widehat{\mathrm{se.ind}}(\mathrm{T})\right] = \left[\sum_{v\in V} p_v + t_\ell\sqrt{\sum_{v\in V} p_v(1-p_v)},\ \sum_{v\in V} p_v + t_u\sqrt{\sum_{v\in V} p_v(1-p_v)}\right].$$

Algorithm 1 Monte Carlo method with bootstrap sampling

Input: $\{(p_v, X_v), v \in V\}$

1: Estimate $\beta_0, \boldsymbol{\beta}, \rho$ with the input as described in Section 2, and compute its residual $\boldsymbol{\epsilon}$ based on equation (3.1).
2: **for** $b = 1, \ldots, B$ **do**
3: **for** $v \in V$ **do**
4: Get a bootstrap sample ϵ_v^* by sampling uniformly from $\{\epsilon_v\}_{v \in V}$.
5: Obtain z_v^* using equation (3.2) and transform back to success probabilities p_v^*.
6: Run the Bernoulli trial Y_v^* with probability p_v^*.
7: **end for**
8: Compute their sum $\mathrm{T}^{(b)} = \sum_{v \in V} Y_v^*$.
9: **end for**

Output: $\{\mathrm{T}^{(b)} : b = 1, \ldots, B\}$: the bootstrap distribution of T.

4. Application to exit polls of KNA election

In this section, we apply the method proposed in Section 2 and 3 to building the prediction interval of the number of seats of each party from the exit poll data in two most recent Korean National Assembly elections, the 19^{th} and 20^{th} elections in year 2012 and 2016. However, we only show the results of the 19^{th} election, in which the exit poll fails to predict the true number of seats for several reasons (Kwak *et al.*, 2013), and defer the results of the 20^{th} election to the Appendix.

4.1. The exit polls and the KNA election

We start with a brief introduction of election exit polls that have been widely used in the U.S. since the 1970s. Their use has now expanded to other democracies (Mitosfky, 1991, 1995; Greiner and Quinn, 2010; Wang *et al.*, 2015). Exit polls are interviews conducted with voters as they leave polling booths. The expansion of exit polls has been driven by broadcasters' competition to be the first to declare election results. In the exit poll, predicting the outcomes of presidential elections is fairly straightforward because the popular vote count is of primary concern. However, popular votes are of secondary concern in legislative elections. Instead, what is of key interest is the number of seats each party wins whereas the outcomes of individual races are of often secondary importance (Brown and Payne, 1975; Curtice and Firth, 2008; Bafumi *et al.*, 2010). Therefore, news media's post-election announcement of exit poll results focus on the number of seats predicted to be won by each party. Given this objective, the news media therefore aims to find an approximate prediction interval of the number of seats, not only its point estimate.

The exit poll data we analyze in this section is obtained during the 19^{th} KNA election held on April, 17, 2012. The KNA is a 300-member unicameral legislature. The KNA election has been held every four years since the promulgation of the March 1988 Electoral Law. The KNA is a Supplementary Member system, where 246 members are elected from constituencies while 54 members are elected at the national level through proportional representation (PR). In the 2012 KNA election, three broadcasting networks (KBS, MBC, and SBS) jointly conducted exit polls. They contract three major polling firms and hire approximately 13,000 interviewers and 500 overseers. Total expenditure by the three networks exceeded seven million U.S. dollars.

The pollsters employ a two-stage cluster design (Mendenhall *et al.*, 1971). In the first stage, random samples of 2,484 polling stations are selected across the nation. In the second stage, one

in every five voters is sampled for a face-to-face interview at each polling station. Interviews are conducted between 6am. and 5pm., whereas the polling booths close at 6pm. The exit polls end early because the networks need time to tabulate the votes so that the results can be announced immediately upon the closure of polling stations. Aside from respondents' voting choice, their age and gender are recorded. In the 19^{th} election, this procedure yielded 674,819 completed interviews. The overall non-response rate was 17.4%. In their post-election announcement, three networks report that the station-specific margin of error ranged between ±2.2% and ±5.1% across 2,484 individual polling stations.

4.2. The spatial graph structure and model

In our exit poll example, the 246 election districts consist of the vertex set V and their spatial locations define the edge set. If two election districts $u, v \in V$ are spatially neighboring each other, we define $e(u, v) = 1$ and 0 otherwise. About the neighboring system, we further divide it into two types depending on the administrative district (the case (iii) mentioned in Section 2). Here, we consider the metropolitan/province level administrative district. There are 17 metropolitan/province level administrative districts, each of which contains from one to dozens of election districts; each election district is for 1 number of legislative seat. The two types of neighboring system are written as follows. First, let

$$N(v) = \{w \in V, e(v, w) = e(w, v) = 1\}$$

be the set of all districts neighboring with v. We define the within and between neighborhoods, denoted as $N_{\text{wtn}}(v)$ and $N_{\text{btw}}(v)$, by

$$\begin{aligned} N_{\text{wtn}}(v) &= \{w \in N(v),\ w, v \text{ are in the same administrative district}\}, \\ N_{\text{btw}}(v) &= \{w \in N(v),\ w, v \text{ are not in the same administrative district}\}. \end{aligned}$$

According to the above two types of neighboring system, we define the spatial model (2.1) for the exit poll data. We have the wining probability of a given party for each district v evaluated from the exit poll data, say $p_v, v \in V$. We define $z_v = \log\{p_v/(1 - p_v)\}$ and assume the model (2.1) with the graph $G = (V, E)$ defined above. In the model (2.1), X_v denotes 16 dummy variables indicating 17 administrative districts, where Seoul district is set as the reference district. The spatial coefficient $\rho_{v.w}$ is defined as

$$\rho_{v.w} = \begin{cases} \rho_{\text{wtn}}, & \text{if } w \in N_{\text{wtn}}(v), \\ \rho_{\text{btw}}, & \text{if } w \in N_{\text{btw}}(v), \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the model for spatial latent variables in (2.1) becomes

$$U_v = \rho_{\text{wtn}} \sum_{w \in N_{\text{wtn}}(v)} U_w + \rho_{\text{btw}} \sum_{w \in N_{\text{btw}}(v)} U_w + \epsilon_v. \tag{4.1}$$

Let A_{wtn} be the adjacency matrix in $\mathbb{R}^{|V| \times |V|}$ for the within neighborhood so that it has ρ_{wtn} at (v, w) if $v \in N_{\text{wtn}}(w)$ or $w \in N_{\text{wtn}}(v)$, and 0 otherwise, and A_{btw} defined in the same manner. A pair of spatial correlations $(\rho_{\text{wtn}}, \rho_{\text{btw}})$ should range in the set where $\mathrm{I} - A_{\text{wtn}} - A_{\text{btw}}$ is positive definite.

Table 1: Prediction result from the exit poll data by three broadcasting systems in the 19^{th} KNA election.

	KBS	SBS	MBC	True
SNP	(131, 147)	(126, 151)	(130, 153)	152
MTP	(131, 147)	(128, 150)	(128, 153)	127

KNA = Korean National Assembly; SNP = Saenuri Party; MTP = Minju Tonghap Party.

4.3. Existing methods for prediction interval

Two existing methods that are known to be used in practice are the normal approximation (NoA) and the Monte Carlo (MCind) approximation by Huh (2008) under the assumption of independence among the observed p_vs.

It is followed by the classical central limit theorem that the sum of n independent (heterogeneous) Bernoulli variables behave like the Gaussian random variable when n is sufficiently large and some regularity conditions are assumed. Consequently, based on the asymptotic normality of the pivotal statistic (3.4), the confidence interval for T of level $1-\alpha$ is given by

$$\left[\hat{\mu}(\mathrm{T}) - z_{\frac{\alpha}{2}}\,\widehat{\mathrm{se.ind}}(\mathrm{T}),\ \hat{\mu}(\mathrm{T}) + z_{\frac{\alpha}{2}}\,\widehat{\mathrm{se.ind}}(\mathrm{T})\right],$$

where z_γ $(0 < \gamma < 1)$ is the $100(1-\gamma)^{th}$ quantile of the standard normal distribution.

On the other hand, MCind approximation is based on independent Bernoulli random samples $\{Y_v^{(b).\mathrm{ind}}, v \in V\}$ with observed success probabilities $\{p_v, v \in V\}$. The suggested prediction interval of level $1-\alpha$ has its endpoints by $100(\alpha/2), 100(1-\alpha/2)^{th}$ quantiles of $\mathrm{T}^{(b).\mathrm{ind}} = \sum_{v\in V} Y_v^{(b).\mathrm{ind}}$, $b = 1, 2, \ldots, B$. However, unlike ours, this does not consider the variability and spatial coherency in $\{p_v : v \in V\}$, and its length tends to be shorter than ours.

4.4. Results: the 19^{th} election

In the 19^{th} election, two major parties, Saenuri Party (SNP) and Minju Tonghap Party (MTP), competed with each other (there are a few more parties, but their numbers of seats are too small to be included); subsequently, the two parties won 152 and 127 number of seats, respectively. Table 1 reports the prediction using the exit poll data by three broadcasting systems. The table also shows that KBS predicted SNP and MTP would win seats between (131, 147) and (131, 147), respectively; MBC announced that the two parties would win seats between (126, 151) and (128, 150), respectively; finally, SBS predicted that the two parties would win seats between (130, 153) and (128, 153), respectively. Here, the prediction intervals are officially based on independent Bernoulli trials. The three broadcasting networks use the same data from the exit poll after the same adjustment procedure for the non-response. However, three networks do additional adjustment to the results and their final results are different. Their adjustments for the non-response are not precise and introduce a significant bias in both point and interval predictions. In the analysis below, we use the adjustment procedure which the authors develop for the MBC system during the 20^{th} KNA election.

We apply the proposed spatial bootstrap method (SB) to build the prediction intervals as well as compare the results to the NoA and MCind. In the analysis below, three methods are applied to each political party, SNP and MTP, independently, to evaluate the expected number of seats and its prediction interval for each party. The size of the Monte Carlo samples for the MCind and SB is set at 10,000.

Table 2 reports the predictions on the number of seats by three methods. We find that the prediction interval by the SB method wider than the two existing methods. The SB accounts for both the spatial dependence and the effects of exogenous administrative districts which makes the interval wider than

Table 2: 95% prediction interval of the number of seats

	NoA	MCind	SB	True
SNP	(137.6, 152.4)	(137.0, 153.0)	(132.0, 158.0)	152
MTP	(128.8, 142.9)	(128.0, 144.0)	(116.0, 142.0)	127

NoA = normal approximation; MCind = Monte Carlo approximation under independence assumption; SB = spatial bootstrap procedure; SNP = Saenuri Party; MTP = Minju Tonghap Party.

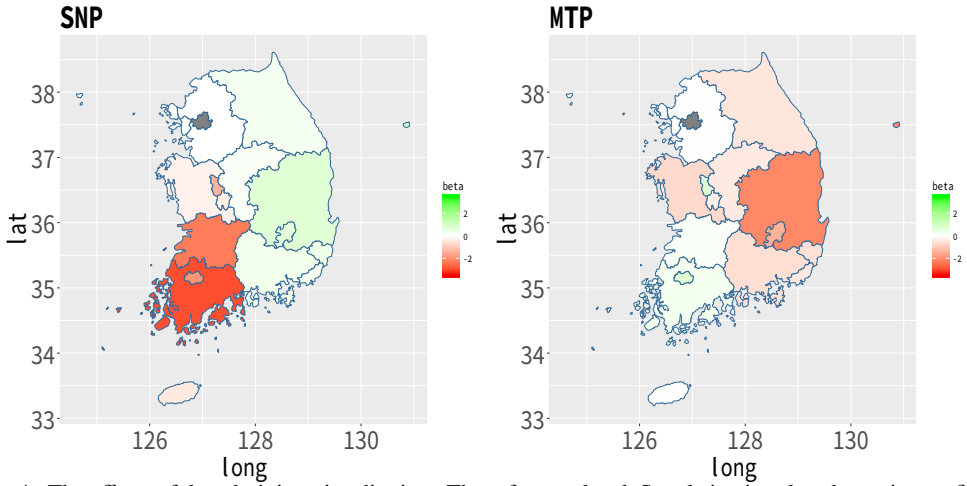

Figure 1: *The effects of the administrative districts. The reference level, Seoul city, is colored gray in two figures. Its estimate is set as* 0.

Table 3: Summary statistics for the spatial coefficients

Party	nhd-type	est	s.e.	p-value
SNP	ρ_{wtn}	−0.057	0.055	0.297
	ρ_{btw}	0.092	0.092	0.321
MTP	ρ_{wtn}	−0.003	0.042	0.943
	ρ_{btw}	−0.300	0.029	<0.001

"est" and "s.e." are, respectively, an average and a standard deviation of B estimates from bootstrapped samples. P-values are calculated based on the normal distribution. SNP = Saenuri Party; MTP = Minju Tonghap Party.

those based on the independent assumption without considering the administrative districts' effect; consequently, the true numbers of seats fall within the intervals by the SB.

The covariate effects, the effects of administrative districts, are plotted in Figure 1, where the coefficient (effect) of administrative district "Seoul" is fixed as 0 for the comparison. The figure shows that the SNP has positive effect in the east and south east part of Korea, whereas the MTP does in the south west part of Korea.

The estimated spatial coefficients are reported in Table 3, where the standard errors are estimated from the bootstrap replications and the p-values are based on the normal distribution. The results show no significant within-spatial-correlation once we account for the covariate effect of the administrative districts. In particular, for the MTP, the covariate effects of the administrative districts are different among districts (Figure 1). The two neighboring election districts from different administrative districts are marginally correlated, but look negatively correlated once we adjust the effects of the administrative districts. We conjecture this would be the reason of the negative estimate of $\hat{\rho}_{\text{btw}}$ of the MTP.

Table 4: 95% prediction interval of the number of seats in small area

	NoA	MCind	SB	True
Belt-SNP	(1.4, 5.7)	(1.0, 6.0)	(4.0, 8.0)	5
Belt-TMP	(2.3, 6.6)	(2.0, 7.0)	(0.0, 4.0)	3
Seoul-SNP	(10.5, 18.4)	(10.0, 18.0)	(9.0, 24.0)	16
Seoul-TMP	(27.4, 35.2)	(27.0, 35.0)	(22.0, 37.0)	30

In the table "NoA" and "MCind" stand for the normal approximation and Monte Carlo approximation under independence assumption. "SB" stands for the spatial bootstrap procedure. "Belt-PARTY" refers to the outcome of PARTY in the Nakdonggang belt and "Seoul-PARTY" in Seoul city.

4.5. Small area prediction

In the election, it is often a particular interest to see a specific region with small number of electoral precincts such as the Nakdonggang River belt that refers to 8 districts around the Nakdonggang River in western Busan. The seat prediction in the small area $W \subset V$ is straightforward since the results can be derived as a byproduct of Algorithm 1. Replacing V by its subset W in the algorithm provides the desired result. Table 4 shows the predictive interval for the Nakdonggang River belt and Seoul city with 8 and 48 legislative seats.

5. Summary

The primary goal of exit polls and legislative elections is to predict the number of seats won by major parties. However, they are not very accurate despite the large amount of financial resources dispatched to conducting exit polls. Furthermore, no formal procedures are suggested to build the prediction intervals of the number of seats won by each party. In this work, we recast the problem into a more general problem: the prediction of the sum of binary random variables on the graph when their success probabilities are observable. We consider the AR regression model to account for the effect of exogenous covariates on the graph and the spatial dependence over the graph. We propose a spatial bootstrap procedure to build the prediction interval of the sum along with the AR regression model. We apply our procedure to the exit poll data from the 19^{th} KNA election in 2012 and the 20^{th} election in 2016 (for the 20^{th} election, see Appendix).

Appendix: The 20^{th} KNA election

The 20^{th} Korean National Assembly Election was held April, 13, 2016. In the 20^{th} KNA, 253 members are elected from constituencies while 47 members are elected at the national level through Proportional Representation (PR); the numbers changed from the 19^{th} KNA. All other details of the exit poll are similar to the 19^{th} election. However, due to the failure in the 19^{th} KNA election, three broadcasting systems (KBS, MBC, and SBS) developed and used their own method to predict the number of seats including the adjustment method for the non-response. Thus, the number of predicted seats show a larger variability over three broadcasting systems.

Three major parties - Saenuri Party (SNP), The Minju Party (TMP), and Gukmin Party (GMP) - run for the election for the 20^{th} KNA. As in the 19^{th} election, there are many other minor parties, but we disregard them in the analysis. Table A.1 shows that SNP, TMP, and GMP ultimately won 122, 123, and 38 seats, respectively. The table also shows that KBS predicted that SNP, TMP, and GMP would win seats between $(121, 143)$, $(101, 123)$, and $(34, 41)$, respectively; MBC announced that the three parties would win seats between $(118, 136)$, $(107, 128)$, and $(32, 42)$, respectively; finally, SBS predicted that the three parties would win seats between $(123, 147)$, $(97, 120)$, and $(31, 43)$, respec-

tively. The prediction intervals are officially based on independent Bernoulli trials. Here, we remark that the three broadcasting networks use the same data from the exit poll, but their adjustments of non-response are not equal. Consequently, they report different predicted number of seats and its intervals.

We apply the proposed spatial bootstrap method to the 20^{th} KNA election and compare the results to those of the normal approximation and the independent Monte Carlo approximation. In the analysis below, three methods are applied to each political party, SNP, TMP, and GMP, independently, to evaluate the expected number of seats and its prediction interval. Table A.2 and A.3, and Figure A.1 reports the results. Here, unlike the 19^{th} election, the prediction intervals by all three methods contain the true number of seats won by each party.

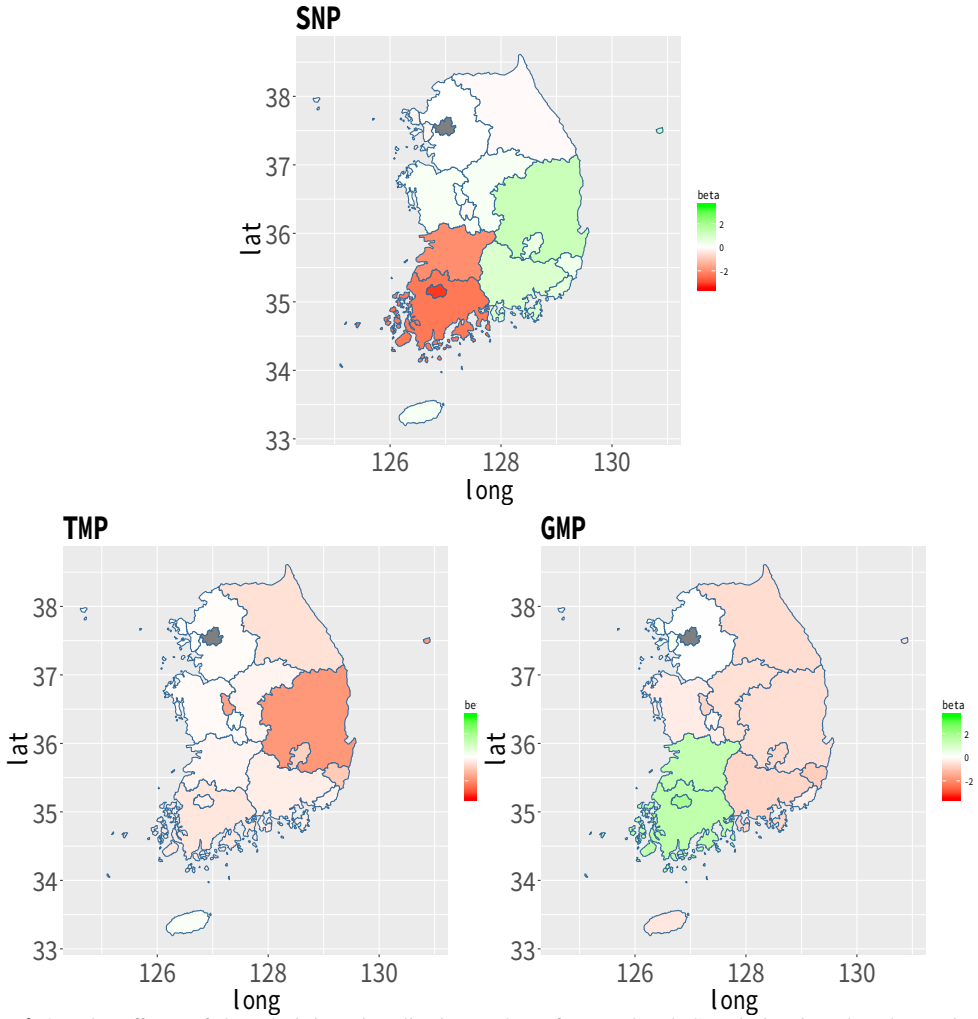

Figure A.1: *The effects of the administrative districts. The reference level, Seoul city, is colored gray in three figures. Its estimate is set as* 0.

Table A.1: Predicted numbers of seats by three major broadcasting companies in Korea in the 20^{th} election

	KBS	SBS	MBC	True
SNP	(121, 143)	(123, 147)	(118, 136)	122
TMP	(101, 123)	(97, 120)	(107, 128)	123
GMP	(34, 41)	(31, 43)	(32, 42)	38

SNP = Saenuri Party; TMP = Minju Party; GMP = Gukmin Party.

Table A.2: 95% prediction interval of the number of seats

	NoA	MCind	SB	True
SNP	(117.2, 134.1)	(116.0, 135.0)	(110.0, 159.0)	122
TMP	(110.5, 127.7)	(109.0, 129.0)	(95.0, 141.0)	123
GMP	(33.8, 40.4)	(33.0, 41.0)	(29.0, 41.0)	38

In the table "NoA" and "MCind" stand for the normal approximation and Monte Carlo approximation under independence assumption. "SB" stands for the spatial bootstrap procedure. SNP = Saenuri Party; TMP = Minju Party; GMP = Gukmin Party.

Table A.3: Summary statistics for the spatial coefficients

party	nhd-type	est	s.e.	p-value
SNP	ρ_{wtn}	0.065	0.053	0.217
	ρ_{btw}	0.189	0.039	<0.001
TMP	ρ_{wtn}	−0.013	0.054	0.803
	ρ_{btw}	−0.198	0.045	<0.001
GMP	ρ_{wtn}	0.013	0.076	0.868
	ρ_{btw}	0.145	0.081	0.075

"est" and "s.e." are, respectively, an average and a standard deviation of B estimates from bootstrapped samples. P-values are calculated based on the normal distribution. SNP = Saenuri Party; TMP = Minju party; GMP = Gukmin party.

References

Anselin L (1988). *Spatial Econometrics*, Dordrecht: Kluwer Academic Publishing.

Arnold M and Wied D (2010). Improved GMM estimation of the spatial autoregressive error model, *Economic Letters*, **108**, 65–68.

Bafumi J, Erikson RS, and Wlezien C (2010). Ideological balancing, generic polls and midterm congressional elections, *Journal of Politics*, **72**, 705–719.

Baltagi BH, Song SH, and Koh W (2003). Testing panel data regression models with spatial error correlation, *Journal of Econometrics*, **117**, 123–150.

Brown P and Payne C (1975). Election night forecasting, *Journal of the Royal Statistical Society, Series A*, **138**, 463–498.

Cochrane D and Orcutt GH (1949). *Application of least squares regression to relationships containing auto-correlated Error Terms*, **44**(245), 32–61.

Curtice J and Firth D (2008). Exit polling in a cold climate: the BBC-ITV experience in Britain in 2005, *Journal of the Royal Statistical Society, Series A*, **171**, 509–539.

Greiner DJ and Quinn KM (2010). Exit polling and racial bloc voting: Combining individual-level and R × C ecological data, *The Annals of Applied Statistics*, **4**, 1774–1796.

Hordijk L (1974). Spatial correlation in the disturbances of a linear interregional model, *Regional and Urban Economics*, **4**, 117–140.

Huh MH (2008). Predicting major political parties' number of seats in general election: the case of 2004 general election of Korea, *Korean Association for Survey Research*, **9**, 87–100.

Kawk J and Choi B (2014). A comparison study for accuracy of exit poll based on nonresponse model,

Journal of the Korean Data and Information Science Society, **25**, 53–64.
Kwak ES, Kim JY, and Kim YW (2013) Analysis of forecasting error of the exit poll for the general election of 2012 in Korea, *The Korean Association for Survey Research*, **11**, 33–55.
Lee L (2002). Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models, *Econometric Theory*, **18**, 252–277.
Mendenhall W, Scheaffer RL, and Ott L (1971). *Elementary Survey Sampling*, Wadsworth Publishing Company.
Mitosfky WJ (1991). *A Short History of Exit Polls*, Sage, Newbury Park, CA.
Mitosfky WJ (1995). *A Review of the 1992 VRS Exit Polls*, Westview Press, Boulder, Colorado.
Ord K (1975). Estimation methods for models of spatial interaction, *Journal of the American Statistical Association*, **70**, 120–126.
Wang W, Rothschild D, Goel S, and Gelman A (2015). Forecasting elections with non-representative polls, *International Journal of Forecasting*, **31**, 980–991.

Received October 15, 2018; Revised March 7, 2019; Accepted March 28, 2019