연구노트

선거 개표방송에서 출구조사 자료를 활용한 중간 득표율 추정에 관한 연구*

Estimating the Interim Rate of Votes Earned Based on the Exit Poll Results during the Coverage of Ballot Results by Broadcasters

이윤동** · 박진우***
Yoon Dong Lee · Jinwoo Park

지상파 방송 3사에서 선거 개표방송을 할 때 사용하는 현재의 개표 집계방식은 각 개표소에서 집계된 개표결과를 단순 합산하여 발표하는 방식이다. 그런데 이 방식은 투표소별 개표 진도의 차이를 무시하는 방식이어서 불필요한 혼선을 초래할 여지가 있다. 방송사 입장에서는 이미 출구조사를 통해 얻은 지역별 데이터가 있는데도 불구하고 이 정보를 오후 6시 예측결과를 발표할 때에만 사용할 뿐이고, 이후 개표가 진행되는 동안에는 전혀 이용하지 않은 채 개표결과만을 단순 집계하여 발표한다. 본 논문에서는 베이지안(Bayesian) 기법을 도입하여 출구조사 자료와 개표결과를 통합하여 발표하는 방법을 제시하고자 한다. 이 방법을 사용함으로써 투표소별 개표 진도의 차이에서 생기는 혼선을 피할 수 있을 것으로 기대한다.

주제어 : 출구조사, 득표수, 득표율, 초기하분포, 다항분포, 사후분포

During major elections, three terrestrial broadcasting stations in Korea have covered the progresses of election results by announcing the simple sum of ballot counts of all ballot counting stations. The current approach, however, does not properly reflect the actual ballot count differences by ballot counting location, leading to cause unnecessary but possible confusions. In addition, the current coverage approach restricts the broadcasters from using regional poll data gained through exit polls by letting them to use the significant information on a one−off purpose to announce the

* 본 논문은 일부 2009학년도 서강대학교 교내연구비의 지원에 의하여 수행되었음.

** 서강대학교 경영학부 교수

*** 교신저자(corresponding author): 수원대학교 통계정보학과 교수 박진우.
E−mail: jwpark@suwon.ac.kr

initial prediction of the poll results and to fully disregard the exit poll results during the ballot counting process. Based on the understanding, this paper is designed to suggest a Bayesian approach to consolidate the exit poll results with the progressive ballot counting results and announce them as such. The suggested consolidation approach is expected to mitigate or avoid the possible confusions that may arise in connection with the different ballot counting paces by ballot counting station.

Keywords : exit poll, ballot counts, rate of votes earned, hyper-geometric distribution, multinomial distribution, posterior distribution

Ⅰ. 머리말

2010년 6월의 지방선거에서는 유례없이 치열한 박빙의 경쟁을 보인 지역들이 많았을 뿐 아니라, 사상 처음으로 도지사와 교육감 선거가 동시에 치러져서 방송 3사의 예측조사 결과에 대해 시민들의 이목이 많이 쏠렸다. 2010년 지방선거 예측을 위해 한국방송협회가 주관하고 KBS, MBC, SBS 지상파 방송 3사가 참여하는 방송사 공동 예측조사위원회(Korea Election Pool: KEP)를 구성한 바 있다(한국방송협회 2010). KEP에서 주관하고 방송 3사에 의해 발표된 출구조사 결과는 YTN이나 MBN의 전화조사를 기반으로 한 선거예측 조사결과와 많은 차이를 보였다. 방송 3사의 출구조사 결과는 실제 개표결과와 거의 차이가 없었고, 더 나아가 박빙지역의 당선자 예측에도 성공하는 개가를 이루었다.

가장 경합이 치열했던 서울시장 선거의 경우 실제개표의 결과가 오세훈 후보 47.4%, 한명숙 후보 46.8%의 득표율인 것으로 나타났는데, 출구조사의 예측 득표율은 오세훈 후보 47.4%, 한명숙 후보 47.2%로 개표결과와 거의 비슷하게 나왔다. 반면, 전화조사에 기초한 YTN이나 MBN의 예측은 실제개표의 결과와 차이가 나서 많은 비판을 받았고, 급기야 잘못된 예측에 대해 방송사 측에서 사과를 하는 지경에 이르게 되었다(김대영 2010).

출구조사의 화려한 성공에 묻혀 거의 관심 밖으로 밀려나긴 했지만, 2010년 지방선거 예측방송의 과정에서 통계학적 화두로 떠오를 수 있는 다른 주제가 있다. 그것은 다름 아닌 개표과정에서 있었던 혼선의 문제이다. 서울의 경우가 가장 대표적인 사례였는데,

출구조사에서는 오세훈 후보의 예상 득표율이 한명숙 후보에 비해 0.2%p 높은 것으로 예측되었다. 그러나 개표의 진행과정에서는 상당 시간 동안 한명숙 후보가 오세훈 후보를 앞지르는 상황이 전개되어 한명숙 후보 측에서는 한때 선거의 승리를 선언하는 일이 벌어지기도 하였다. 이러한 혼선이 생기게 된 근본 원인으로는 서울의 구별 개표 진도가 달랐던 것을 들 수 있다. 개표 초반에는 상대적으로 한명숙 후보 지지율이 높은 구들에 대한 개표가 빠르게 진행된 반면, 후반으로 갈수록 오세훈 후보 지지율이 높은 구들의 개표가 많이 이루어졌기 때문이다. 구별 개표 진도의 차이로 인해 일어나는 이러한 혼란을 사전에 방지할 수는 없었을까?

통계학적 관점에서 보면 현재의 개표 집계 방식은 문제가 있다. 현재의 방식은 각 개표소에서 집계된 개표결과의 단순 합계를 발표하는 방식인데, 이것은 투표소별 개표 진도의 차이를 무시하는 방식이다. 방송사 입장에서는 이미 출구조사를 통해 얻은 지역별 데이터가 있음에도 불구하고 이 정보를 오후 6시 예측결과 발표 시에만 이용할 뿐 이후에는 제대로 활용하지 못하는 것으로 생각된다. 김희연(2008)에 의하면, 2008년 대통령 선거 당시 SBS는 개표방송에서 분 단위의 집계정보를 이용하여 후보자별 당선확률을 계산한 바가 있다. KBS의 경우도 방송사가 자체 개발한 디시전-K(Decision-K)라는 선거예측시스템을 사용한 바 있다고 한다. 그러나 이런 문제가 공론화되어 적절한 이론적 방법론이 제시된 적은 없다.

본 논문에서는 베이지안(Bayesian) 기법을 도입하여 출구조사 자료와 개표결과를 통합하여 발표하는 방법을 제시하고자 한다. 이러한 연구는 선거의 개표과정에서 기존의 단순 집계방식을 보완하기 위한 이론적 대안을 제시하는 한편, 실제 투표소별 개표 진도의 차이에서 생기는 혼선을 피하는 데 도움을 줄 것으로 기대한다.

Ⅱ. 새로운 득표율 추정법

A, B, C 세 명의 후보가 한 선거구에 출마하는 경우를 가정한다. 선거구 내 특정 i번째 투표소(혹은 지역)에서 전체 N^i명의 유권자가 투표에 참여하며, 각 후보자별 득표수는 $\boldsymbol{N}^i = (N_A^i, N_B^i, N_C^i)$라고 표기한다. 논의를 단순화하기 위해 투표에서 무효는 없는 것으로 가정한다. 즉, $N^i = N_A^i + N_B^i + N_C^i$ 이다. 한편, 전체 n^i명의 표본을 대상으로 출구조사를 실시하였는데, 표본 응답자들 중 각 후보자에 대한 지지자수는

$\boldsymbol{n^i}=(n_A^i,\ n_B^i,\ n_C^i)$인 것으로 가정한다. 개표가 진행되는 특정 시점 t에서의 각 후보별 득표수를 $\boldsymbol{x_t^i}=(x_{A,t}^i,\ x_{B,t}^i,\ x_{C,t}^i)$로 나타내기로 하자. 개표개시 시점인 $t=0$일 때 $\boldsymbol{x_0^i}=\boldsymbol{0}$ 이 되며, 개표가 완료되는 시점 즉, $t=T$일 때 $\boldsymbol{x_T^i}=\boldsymbol{N}^i$가 성립된다.

시점 t에서의 득표수 $\boldsymbol{x_t^i}$와 득표율 $\boldsymbol{p_t^i}=\boldsymbol{x_t^i}/x_t^i$는 각각 통계적으로 두 가지 특성을 갖는다. 하나는, 각각 최종득표수 $\boldsymbol{N}^i=(N_A^i,N_B^i,N_C^i)$ 또는 최종득표율 $\boldsymbol{N}^i/N$ (혹은 t 시점 이후의 득표수와 득표율)에 대한 추정량으로서의 의미를 지닌다. 이때 각각의 통계량은 확률적 성격을 갖는다. 다른 하나는, t시점까지 공개된 정보에 의하여 설명되는 집합에 대한 모수(parameter)로서의 의미를 갖는다.

i번째 투표소에 대한 출구조사 결과는 각 후보별 득표율이 $\boldsymbol{n}^i/n^i$인 것으로 발표될 것이고, 이후 개표진행에 대한 보도는 매 집계시간별로 각 후보별 득표수 $\boldsymbol{x_t^i}$, 혹은 이를 해당 시점에서의 개표수 x_t^i로 나눈 득표율 $\boldsymbol{p_t^i}=\boldsymbol{x_t^i}/x_t^i$ 형태로 이루어지게 될 것이다. 또한 각 후보별 전체 지역 $i=1,\ 2,\ \ldots,\ m$에서의 득표수 합계는

$$\boldsymbol{x}_t=\sum_{i=1}^{m}\boldsymbol{x}_t^i=\sum_{i=1}^{m}x_t^i\cdot\boldsymbol{p_t^i}$$

로 집계되어 발표될 것이다. 그러나 이와 같이 단순한 방법에 의하여 집계되는 현행의 득표율과 득표수 계산방법은 앞 절에서 언급한 바와 같이, 특수한 경우에서는 득표율과 득표수에 대한 바른 정보를 제공하지 못할 우려가 있다. 특히 이와 같은 득표율과 득표수에 대한 집계방법은 출구조사에 의하여 얻은 정보와 개표 정보를 통합하지 못한 채 개표 정보만을 이용하고 만다는 문제점을 안고 있다.

이러한 문제점을 개선하기 위해 다양한 형태의 득표율 추정방법과 득표수 집계방법이 고려될 수 있다. 가령, $\boldsymbol{p_t^i}$에 대한 추정량 $\hat{\boldsymbol{p}}_t^i$의 대안으로 다음과 같은 추정량들을 생각해 보자.

$$\bar{\boldsymbol{p}}_t^i=\frac{\boldsymbol{n^i}+\boldsymbol{x_t^i}}{n^i+x_t^i}$$

$$\tilde{\boldsymbol{p}}_t^i=\frac{\boldsymbol{n^i}+(1-n^i/N^i)\,\boldsymbol{x_t^i}}{n^i+(1-n^i/N^i)\,x_t^i}$$

$$\widehat{\boldsymbol{p}}_t^i = \frac{1}{N^i}\left\{\boldsymbol{x}_t^i + (N^i - x_t^i)\left(\frac{\boldsymbol{n}^i + r\boldsymbol{x}_t^i}{n^i + r x_t^i}\right)\right\}$$

여기서 r은 어떤 적당한 상수이다.

위의 세 가지 추정량은 모두 개표 개시시점(즉, $t=0$)에서는 출구조사에서 얻은 득표율과 동일한 값을 가지며, 점차 개표가 진행됨에 따라 값이 모수인 $\boldsymbol{p}_t^i$에 근접하게 된다. 한편 개표가 종료되는 시점에 이를 경우, $\widetilde{\boldsymbol{p}}_t^i$와 $\widehat{\boldsymbol{p}}_t^i$는 모수인 $\boldsymbol{p}_t^i$와 동일한 값을 갖게 되는데 반해, $\overline{\boldsymbol{p}}_t^i$는 모수값과 일치하지 않는다. N^i에 비해 n^i가 매우 작아서 $n^i/N^i \approx 0$이 되는 경우, 개표가 종료되는 시점에 이르면 세 추정값은 거의 같은 값이 된다. 다만 후보자별 최종 득표수 집계는 단 한 표의 오차도 없이 매우 정확해야 하므로, $\overline{\boldsymbol{p}}_t^i$를 이용하는 추정량을 사용할 경우 자칫 문제가 생길 수도 있다.

Ⅲ. 추정량의 통계적 특성

이 절에서는 관심 지역 i가 특정 지역으로 고정되어 있는 경우만을 고려하여 모든 표기법에서 첨자 i를 생략하기로 한다. 또한 후보자를 A, B, C 세 명의 경우로 한정하지 않고 $j=1, 2, ..., J$와 같이 J명의 후보자가 출마했다고 가정하기로 한다. 아울러 각 후보자별로 출구조사 결과 얻은 표본 지지자수는 $n_j (j=1, \cdots, J)$로 나타내기로 한다. 또한 관심시점 t가 특정시점으로 고정되어 있다고 보아 시점을 의미하는 첨자 t를 굳이 표현할 필요가 없는 경우에는 $x_{j,t}$나 $\boldsymbol{x}_t = (x_{1,t}, ..., x_{J,t})$ 대신에 x_j나 $\boldsymbol{x}$로 표기하기로 한다.

모집단은 N명의 투표자로 구성되고, 각 투표자가 투표한 후보자는 $S_k \in \{1, ..., J\}$이다. 이때 선거와 그에 따른 개표의 목적은 각 후보자별 최종득표수를 알아내는 것인데 식으로 표현하면 다음과 같다.

$$N_j = \sum_{k=1}^{N} I(S_k = j)$$

여기서 $\sum_{j=1}^{J} N_j = N$이며, $j = 1,2,...,J$이다. 한편, 최종득표율 또한 관심 있는 모수로서 $p_j = N_j/N$, $j = 1,2,...,J$로 나타낼 수 있다.

출구조사와 개표의 목적은 최종득표수 $\boldsymbol{N}=(N_1, ..., N_J)$ 또는 최종득표율 $\boldsymbol{p}=(N_1, ..., N_J)/N$을 알아내는 것이다. 출구조사에 의하여 얻은 표본의 정보 n_j와 개표가 시작된 지 일정 시간이 경과된 후의 후보별 득표수 x_j는 다변량 초기하분포(multivariate hyper−geometric distribution)나 다항분포(multinomial distribution) 모형으로 설명될 수 있다. 다변량 초기하분포의 공액사전분포는 통합 다변량 초기하분포(unified multivariate hyper−geometric distribution)나 폴리야분포(Pólya distribution)인 반면, 다항분포의 공액사전분포는 디리쉬리분포(Dirichlet distribution)이다. 다변량 초기하분포와 다항분포에 대한 자세한 사항은 Steck & Zimmer(1968)과 Janardan(1976)를 참조하면 된다. 다변량 초기하분포 모형과 다항분포 모형 각각에서 얻은 사후분포의 유사성에 대해서는 Tuly et al.(2009)가 다룬 바 있다.

먼저 추정량 $\bar{p}_j = (n_j + x_j)/(n + x)$를 살펴보자. 이 추정량은 베이지안 추론법 관점에서 디리쉬리분포를 사전분포로 갖는 다항분포 모형 아래에서 사후분포의 평균이 된다. 아울러 $\bar{p}_j$는 다항분포 모형에서 모수 p_j의 최대우도추정량(maximum likelihood estimator)이기도 하다. 복원표본을 가정하는 다항분포 아래에서나 비복원표본을 가정하는 초기하분포 아래에서나 모비율 p_j에 대한 최대우도추정량은 동일하게 $\bar{p}_j = (n_j + x_j)/(n + x)$이다.

먼저 모수 $\boldsymbol{p} = (p_1, ..., p_J)$에 대한 사전분포로, $\boldsymbol{\alpha}=(\alpha_1, ..., \alpha_J)$를 모수로 갖는 디리쉬리분포 $DD(\boldsymbol{\alpha})$를 가정하자. 다음으로 출구조사에서 조사된 후보자별 표본 득표수 $\boldsymbol{n}=(n_1, ..., n_J)$와 개표에 의하여 얻게 되는 후보자별 득표수 $\boldsymbol{x} = (x_1, ..., x_J)$를 각각 다항분포로 가정하자. 이 경우 개표직전 출구조사에 의한 정보가 통합된 상태에서의 사후분포는 $DD(\boldsymbol{n}+\boldsymbol{\alpha})$가 되고, 개표가 이루어진 이후의 사후분포는 $DD(\boldsymbol{x}+\boldsymbol{n}+\boldsymbol{\alpha})$가 된다. 즉,

$$f(\boldsymbol{p}) \propto \prod_{j=1}^{J} p_j^{(\alpha_j - 1)}, \quad f(\boldsymbol{n}|\boldsymbol{p}) \propto \prod_{j=1}^{J} p_j^{n_j}, \quad f(\boldsymbol{x}|\boldsymbol{p}) \propto \prod_{j=1}^{J} p_j^{x_j}$$

라고 가정하면, 그 사후분포는

$$f(\boldsymbol{p}|\boldsymbol{x},\boldsymbol{n}) \propto \prod_{j=1}^{J} p_j^{(x_j+n_j+\alpha_j-1)}$$

가 된다. 따라서 사후분포의 평균은 $E(\boldsymbol{p}|\boldsymbol{x}) = (\boldsymbol{x}+\boldsymbol{n}+\boldsymbol{\alpha})/(x+n+\alpha)$가 된다. 여기에서 $\alpha = \sum_{j=1}^{J} \alpha_j$ 이다.

사전분포 $DD(\boldsymbol{\alpha})$를 가정할 때는 분포에 의하여 정의된 확률의 합이 1이 되는 정상적인 분포가 되기 위하여 $\alpha_j > 0$, $j = 1, 2, ..., J$ 가 만족되어야 하나, 사후분포가 정상적인 분포가 되도록 하는 조건만을 고려하는 경우 $x_j + n_j + \alpha_j > 0$, $j = 1, 2, ..., J$ 이기만 하면 충분하다. 만일 $\boldsymbol{\alpha} = \boldsymbol{0}$ 이라고 하면 $\bar{\boldsymbol{p}} = E(\boldsymbol{p}|\boldsymbol{x}) = (\boldsymbol{x}+\boldsymbol{n})/(x+n)$ 이다.

실제 개표과정이 유한모집단에 대한 비복원 추출과정이라는 점을 고려하여 이번에는 출구조사에서의 표본 득표수 $\boldsymbol{n} = (n_1, ..., n_J)$와 개표에 의한 득표수 $\boldsymbol{x} = (x_1, ..., x_J)$의 분포를 다항분포가 아닌 다변량 초기하분포로 가정하는 경우를 생각해 보자. 이때의 사후분포는 다음과 같이 나타낼 수 있다.

$$f(\boldsymbol{N}) \propto \prod_{j=1}^{J} \binom{a_j}{N_j} \text{ 이고 } f(\boldsymbol{x}|\boldsymbol{N}) \propto \prod_{j=1}^{J} \binom{N_j}{x_j} \text{ 라 하면, } f(\boldsymbol{N}|\boldsymbol{x}) \propto \prod_{j=1}^{J} \binom{a_j - x_j}{N_j - x_j}$$

이 경우 j번째 후보자의 득표수에 대한 사후추정량을 계산하면 다음과 같다.

$$E(N_j|\boldsymbol{x}) = x_j + (N-x)\frac{a_j - x_j}{a-x} = N\left[w\,\frac{x_j}{x} + (1-w)\,\frac{a_j}{a}\right]$$

여기서 $w = \frac{x}{N}\frac{a-N}{a-x}$ 이다. 또한 $a = N$ 인 경우 $w = 0$이 되고 $x \to N$(개표가 거의 완료)이면 $w \to 1$이 된다. 사후추정량 $E(N_j|\boldsymbol{x})$은 사전정보(출구조사)와 표본정보(개표결과)의 가중평균의 형태로 전통적인 베이지안 추론의 결과와 일치한다. 여기서 w가 0에 접근하면 사후추정량은 출구조사의 결과와 일치하게 되고, w가 1에 근사하면 개표결과와 일치하게 된다.

한편 사후추정량의 분산을 유도하면 다음과 같다.

$$Var(N_j|\boldsymbol{x})$$
$$= \frac{a-N}{N^2(a-x)(a-x+1)} \cdot \left[(N-x)x\,V_j(\boldsymbol{x}) + (N-1)a\,V_j(\boldsymbol{a}) + \frac{xa(N-x)}{a-x} D_j(\boldsymbol{x},\boldsymbol{a})\right]$$

여기서 $V_j(\boldsymbol{x}) = \frac{x_j}{x}\left(1-\frac{x_j}{x}\right)$, $V_j(\boldsymbol{a}) = \frac{a_j}{a}\left(1-\frac{a_j}{a}\right)$, $D_j(\boldsymbol{x},\boldsymbol{a}) = \left(\frac{x_j}{x}-\frac{a_j}{a}\right)^2$이다.

만일 $a_j = cn_j$ $(c \geqq N/n)$ 라 하면, a_j/a 는 출구조사에서 얻은 비율에 따라 n_j/n 과 같다. $c = N/n$, 즉 $a = N$인 경우, $E(N_j|\boldsymbol{x}) = N(n_j/n)$으로 개표에 의해 얻은 결과가 전혀 반영되지 않고 출구조사의 결과만을 계속 집계하는 경우가 된다. 반면 c가 매우 커지면 커질수록 w는 x/N에 수렴하고 결국,

$$E(N_j|\boldsymbol{x}) = N\left[\frac{x}{N}\frac{x_j}{x} + \left(1-\frac{x}{N}\right)\frac{n_j}{n}\right]$$

이 된다. 개표된 부분에 대하여는 개표결과를 반영하되 개표되지 않은 부분에 대하여는 출구조사의 결과가 나타날 것으로 예측하는 방법이다. $c = N/n$인 경우 $Var(N_j|\boldsymbol{x})$은 0이 되는 반면, c가 매우 커짐에 따라 증가하여 $(N-1)V_j(\boldsymbol{n})/N^2$로 수렴한다. $a_j = cn_j$ 라 할 때, $c = N/n$인 경우 $Var(N_j|\boldsymbol{x}) = 0$인 것은 사전확률에 대한 믿음이 너무 강하여 개표결과가 추정값에 반영되지 않는 경우에 해당하는 것으로, 바람직하지 않은 사전확률이 주어진 경우라고 할 수 있다. 사전확률의 영향은 가능한 배제하는 방향으로 사전확률이 설정되어야 하므로 c는 가능한 한 크게 잡는 것이 바람직하다.

또한 위의 결과는 a_j가 일단 한번 결정되면 바뀔 수 없다는 가정 하에서 사전확률분포의 모수를 $a_j = cn_j$라고 설정하고 득표수와 득표율 예측을 집계하는 방법이다. 그러나 본 연구에서 논의되고 있는 선거개표에서 득표수 집계 문제는 득표수가 집계되어 감에 따라 개표에 의하여 증가된 정보를 근거로 사전정보를 설정하고 위의 과정을 반복할 수 있는 경우에 해당한다. 그러므로 위에서 얻은 결과를 단순하게 그냥 사용하는 것보다는 개표된 정보에 의하여 사전확률분포를 바꿀 수 있게 한 후 이 사전확률에 의해 득표수와 득표율을 예측하게 하는 것이 더욱 합리적이다.

출구조사에 의한 정보 $\boldsymbol{n}$과 개표에 의한 득표수 정보 $\boldsymbol{x}$가 주어지는 경우 나머지 미개표 투표함에서의 득표율에 대한 추정은, 앞서 언급한 바와 같이 다항분포를 가정한 베이

지안 추론법이나 최대우도추정법에서 동일하게 $\bar{p}_j = (n_j + x_j)/(n+x)$로 주어지는 점에 착안하여, 사전확률모수를 a_j/a가 n_j/n이라고 하기보다 $(n_j + x_j)/(n+x)$, 혹은 이를 일반화 하여 적당한 상수 r에 대하여 $(n_j + rx_j)/(n+rx)$라고 설정하여 하는 것이 바람직하다. 즉 매우 큰 c에 대하여 $a_j = c(n_j + rx_j)$라 하자. 그러면 예상득표수는

$$E(N_j|\boldsymbol{x}) = N\left[\frac{x}{N}\frac{x_j}{x} + \left(1-\frac{x}{N}\right)\frac{n_j + rx_j}{n+rx}\right] \quad (1)$$

이 되고, 특정 지역 i에서의 득표율 $\hat{\boldsymbol{p}}^i = (\hat{p}_1^i, \ldots, \hat{p}_J^i)$는

$$\hat{\boldsymbol{p}}^i = \frac{\boldsymbol{x}^i}{N^i} + \left(1-\frac{x^i}{N^i}\right)\frac{\boldsymbol{n}^i + r\boldsymbol{x}^i}{n^i + rx^i}$$

가 된다. 앞서 언급한 바와 같이 베이지안 통계량은 출구조사(사전정보)와 개표결과(표본정보)의 가중평균 형태를 띤다. 즉 (1)식의 둘째 항은 $r \neq 0$일 때 출구조사와 개표결과가 혼합된 형태가 되므로, 추정량은 가중평균의 형태가 되는 것이다.

지역 i에 각각 x^i표가 개표된 경우, 전체 득표수 추정량은 다음과 같이 표현된다.

$$\boldsymbol{x} = \sum_{i=1}^{m} x^i \cdot \hat{\boldsymbol{p}}^i$$

위와 같은 설정에서 r이 매우 크면 단순집계방식과 동일해지고, r이 0인 경우는 앞서 언급한 사후분포의 평균과 정확하게 일치한다. r값을 어떻게 선정하는 것이 좋은지에 대해 검토해 보았으나 특별한 기준을 발견하기 어려웠다. 다만 r이 큰 경우 개표되는 값에 영향을 많이 받아 짧은 주기의 변동이 반영되는 반면, r이 작은 경우는 곡선의 변동이 보다 부드러운 형태를 보인다는 것을 관찰할 수 있다.

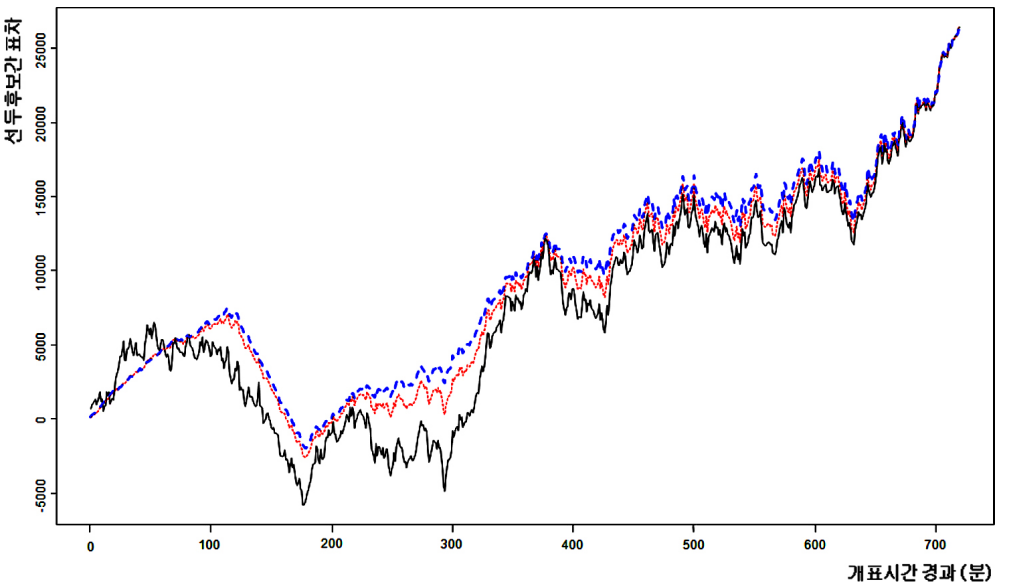
Ⅳ. 모의 실험

2010년 6월2일 실시된 서울시장선거에서의 최종 개표결과를 이용하여 모의실험을 수행하였다. 실제 6.2 서울지방선거 개표상황과 비슷하도록 25개 각 구별 개표진행 속도를

다르게 한 후, 출구조사 정보와 개표정보를 통합하여 득표수와 득표율을 집계하는 방법들을 비교하기로 한다. 이 모의실험에서는 단순집계방식과 $x\hat{\boldsymbol{p}}$을 이용한 집계 방법을 시간의 진행에 따라 비교한다. 실험의 단순화를 위하여, 오세훈과 한명숙 후보 이외의 후보자들에게 투표된 표는 모두 기타로 분류하였다.

실제 개표진행 상황에 대한 자세한 자료가 없는 까닭에 최종 득표수 결과를 근거로 시뮬레이션에 의하여 개표진행 모습을 재현하고, 그때 집계방법에 따라 그 차이가 어떻게 나타나는지를 확인하였다. 서울 25개 각 구에서 확률적으로 약간의 차이를 갖는 속도로 개표가 진행된다고 가정하였고, 서초·강남 지역의 개표가 개표개시 2시간 후에 1시간 동안 일시 정지된 상황을 가정하였다. 지난 6월 2일 실제 이와 같은 개표 지체가 발생하였다.

출구조사에 대한 정보를 참고하여, 출구조사 전체 응답자 약 13만 명에 대한 응답을 가상적으로 생성한 이후, 단순비례추출방법에 따라 투표자가 많은 지역에 많은 출구조사 표본을 할당하고, 최종득표율을 모수로 갖는 다항분포를 가정하여 모의 출구조사 표본을 생성하였다.



〈그림 1〉 개표 집계 방법에 따른, 선두 후보간 표차이 집계 모의실험 결과

〈그림 1〉은 개표진행 경과에 따라 $\hat{\boldsymbol{p}}^i$ 을 계산한 후 이로부터 득표수를 집계하는 경우의 진행과정을 보여 주고 있다. 그림에서 실선은 실제 득표수에서 1위 후보와 2위 후보의 득표수 차이를 보여 주고 있다. 반면 점선은 $r=0$인 경우와 $r=n^i/N^i$ 인 경우 각각에 대한 $\hat{\boldsymbol{p}}^i$ 을 계산한 후, 그로부터 득표수를 계산한 값을 나타낸다. $r=0$인 경우는 굵은 점선으로 나타나 있고, $r=n^i/N^i$ 인 경우는 가는 점선으로 나타나 있다.

〈그림 1〉에서 두 점선 간에는 큰 차이가 나타나지 않으나, 두 점선과 실선 사이에는 상당한 차이가 나타나고 있다. 단순 집계방식을 나타내는 실선을 보면, 개표개시 후 200분 경과 시점부터 300분에 이르는 시기에 1위 후보가 2위 후보에 비하여 5,000표 이상 뒤지는 것으로 나타났으나, 베이지안 추정방식에 의해 계산된 점선에서는 180분이 경과된 시점에서 잠시 동안만 1위 후보가 2위 후보에 비하여 약간 뒤지다가 이내 1위를 달리는 것을 관찰할 수 있다. 이는 개표 결과뿐 아니라 출구조사 결과를 동시에 고려하여 추정하는 베이지안 추정법이 단순 집계방식보다 개표과정에서 생기는 혼란을 줄이는 데 유용하다는 것을 나타낸다.

Ⅴ. 맺음말

동일한 선거구에서 투표소별로 후보자들에 대한 지지 양상이 극명하게 다를 경우, 개표방송에서 시간대별 득표율 순위는 투표소별 개표 진도에 민감한 영향을 받는다. 실제 2010년 서울시장선거 개표방송에서 이런 상황이 나타나 혼란을 겪은 바 있다. 본 연구에서는 개표가 진행되는 상황에서 보다 정확한 득표수 집계를 하기 위해 출구조사 정보를 사전분포로 활용하는 베이지안 추정법을 제시하였으며, 모의실험을 통해 새로운 추정법의 효용성을 보이기도 했다.

일부 방송사에서는 구체적인 방법론은 다를지언정 본 연구에서 제기하는 문제를 인식하고 나름의 검토를 한 것으로 보인다. 하지만 이 문제를 공론화하여 보다 이론적이고 체계적인 해결책을 모색하고자 하는 시도는 부족하였다. 개표가 진행되는 상황에서 개표결과와 출구조사 결과를 결합하는 문제를 통계적인 이론의 틀로 담아 공론화하였다는 점이 본 연구가 지니는 의의라고 할 것이다.

본 연구에서는 순수하게 통계이론적인 입장에서 보다 효과적인 추정법을 제시하는 데 관심을 두었다. 그러나 선거 개표방송에서 득표수 추정의 문제는 단순히 통계적인 문제의 틀을 벗어나 정치적으로 민감한 영향을 끼칠 수 있는 문제일 수 있다. 또한 방송사에서 충분한 사전 설명 없이 이 방법을 사용할 경우 기존의 개표방식에 익숙한 개표방송 시청자들에게 오히려 또 다른 혼선을 불러일으킬 수도 있다. 그러므로 본 연구에서 제안한 방법을 실제로 적용하기 위해서는 충분한 현장 시험과 시청자 홍보를 거쳐야 할 것이다.

참고문헌

김대영. 2010. "KEP(Korea Election Pool) 평가와 과제." ≪방송문화≫ 2010년 6월호: 28-33.

한국방송협회. 2010. ≪2010 지방선거 방송사 공동 예측조사위원회 백서≫.

김희연. 2008. ≪선거개표방송을 위한 당선 확률 예측≫. 고려대학교 석사학위논문.

Steck, G.P. and W.J. Zimmer. 1968. "The Relationship Between Neyman and Bayes Confidence Intervals for the Hypergeometric Parameter." *Technometrics* 10: 199-203.

Janardan, K.G. 1976. "Certain Estimation Problems for Multivariate Hypergeometric Models." *Annals of Institute of Statistical Mathematics* 28(Part A): 429-444.

Tuyl, F., R. Gerlach and K. Mengersen. 2009. "Posterior Predictive Arguments in Favor of the Bayes-Laplace Prior as the Consensus Prior for Binomial and Multinomial Parameters." *Bayesian Analysis* 4: 151-158.

<접수 2011/1/31, 수정 2011/2/25, 게재확정 2011/2/28>