Demographic Covariates and Election Forensics: Re-evaluating Statistical Anomalies in the 2020 South Korean Legislative Election

Timothy Kim

Independent, timothyalphakim@outlook.com

February 25, 2026

Abstract

Following the 2020 South Korean legislative election, significant public discourse emerged regarding statistical anomalies in the voting returns, particularly concerning the unprecedented divergence between early voting and same-day voting margins. Initial forensic models - most notably the Bayesian eforensics model - suggested these discrepancies were highly improbable under standard probability distributions. This paper re-evaluates these claims using microscopic, precinct-level data (Dong-level) merged with local demographic census data. Utilizing a frequentist framework, we apply digit-based tests, variance dispersion analysis, and covariate adjustments for age-based voting propensity. We find that the alleged anomalies are statistical artifacts of demographic sorting. Furthermore, tests on regional convergence and invalid vote rates yield no evidence of artificial data generation. The results underscore the methodological necessity of incorporating demographic covariates in mixture models, and clarify the interpretation of p-values and residual normality when evaluating the data-generating mechanisms of full population datasets.

1 Introduction

The 21st South Korean legislative election, held in April 2020 amid the COVID-19 pandemic, saw record turnout, driven largely by an unprecedented surge in early voting. Following the election, observers noted a stark divergence: the ruling Democratic Party performed significantly better in the early voting period than on the official election day. This persistent gap across numerous districts led to public allegations of systemic irregularities and artificial data generation.

Election forensics, a subfield of political methodology, utilizes statistical tools to detect artificial manipulation in vote counts (Beber and Scacco, 2012). In the wake of the 2020 election, an initial working paper utilizing the *eforensics* statistical model flagged the South Korean early voting data as highly anomalous, suggesting potential fraudulent generation (Mebane, 2020). However, flawed or incomplete forensic research can have profound consequences; as demonstrated by Kuk et al. (2024), publicizing methodologically flawed election forensics can severely erode democratic trust and intensify political polarization.

Forensic models face a critical vulnerability: omitted variable bias and local data structure misinterpretations. Domestic scholars noted that initial anomaly reports failed to account for unique domestic data formatting (Park, 2020). Similarly, Yoo and Min (2021) demonstrated that when age demographics and regional sorting are properly modeled, the statistical probability of artificial manipulation approaches zero.

This paper systematically tests multiple hypotheses of artificial data generation utilizing microscopic polling data to resolve this methodological debate, favoring a robust frequentist approach over highly parameterized Bayesian mixture models.

2 Data and Methodology

We construct a microscopic dataset linking precinct-level election returns from the NEC to neighborhood-level (Eup/Myeon/Dong) demographic data from the Korean Statistical Information Service (KOSIS).

2.1 Bayesian vs. frequentist Frameworks in Election Forensics

The *eforensics* model utilized by Mebane (2020), as well as the econometric counter-analysis by Yoo and Min (2021), rely heavily on Bayesian inference. Specifically, *eforensics* employs a Bayesian finite mixture model that attempts to classify voting units into unobserved latent states (e.g., "no fraud," "incremental fraud," "extreme fraud") by estimating posterior probability distributions.

While powerful, Bayesian mixture models in election forensics are exceptionally vulnerable to omitted variable bias. If a model defines the "no fraud" baseline distribution without incorporating critical covariates - such as the highly polarized age demographics that dictated early-voting usage in 2020 - the model will mathematically force the resulting natural divergence into the "fraud" classification.

To avoid the fragility of these parameterized assumptions, this paper utilizes a frequentist approach. Rather than attempting to estimate unobservable latent states through complex priors, our frequentist methodology directly tests the observed data against rigid, mathematically established null hypotheses (such as Benford's Law and expected variance). This approach is sufficient, and arguably superior in highly polarized environments, because it requires far fewer assumptions about the exact probability distribution of voting behavior, relying instead on the fundamental mathematical properties of numbers.

2.2 The Interpretation of P-Values in Population Data

Because election returns are not an opinion poll, we do not rely on statistical sampling. Our dataset contains the entire population of voting records. Traditionally, a p-value measures the probability of drawing a specific sample from a larger population due to sampling error.

However, in the context of full population data, the interpretation of the p-value shifts from sampling inference to evaluating the *data-generating mechanism* (Fisher, 1925). The null hypothesis assumes that the vote counts are the product of a natural, uncoordinated stochastic process - millions of independent human decisions interacting with normal geo-

graphic variance. The p-value therefore quantifies how compatible the observed, finalized population data is with this natural generating process. A p-value below the significance threshold (e.g., $\alpha = 0.05$) would indicate that the data is highly incompatible with natural human generation, suggesting an artificial or algorithmic intervention.

2.3 The Central Limit Theorem and the Normality of Residuals

To accurately detect an artificial data-generating mechanism, forensics relies heavily on the Central Limit Theorem (CLT). The CLT states that the sum or average of a large number of independent, identically distributed random variables will approximate a normal (Gaussian) distribution, regardless of the underlying distribution of the individual choices. In the context of an election, the millions of independent, localized voting decisions naturally aggregate into a normal, bell-shaped distribution of vote shares and margins.

When evaluating anomalies like the early-voting gap, we must mathematically control for known sociological predictors (such as the concentration of middle-aged voters who overwhelmingly vote early). Once these known demographic covariates are accounted for, the remaining unexplained variance - known as the *residuals* - should resemble normally distributed stochastic noise.

A non-normal distribution of these residuals is the primary mathematical signature of electoral fraud. If the adjusted data exhibits extreme skewness, multimodality (multiple unnatural peaks), or artificial uniformity (a suspicious lack of variance), it violates the Central Limit Theorem. Such a violation strongly implies that an external, non-random force - such as top-down algorithmic padding, fixed numerical ratios, or manual ballot stuffing - has interrupted the natural human data-generating process.

2.4 Standard Forensic Metrics

We apply widely established tests for numerical generation:

1. **Second-Digit Benford's Law (2BL):** Natural vote counts span multiple orders of magnitude, causing their digits to conform to Benford's logarithmic distribution.

Human-manipulated data frequently fails to replicate this curve.

2. **Last-Digit Uniformity:** The terminal digits of sufficiently large vote counts should exhibit a uniform distribution. Deviations indicate human rounding or algorithmic padding (Beber and Scacco, 2012).

3. **Variance Dispersion:** We measure the standard deviation of vote shares between neighborhoods to test for unnatural clustering indicative of top-down algorithmic targets.

2.5 Evaluating Specific Hypotheses of Artificial Generation

We operationalize several localized hypotheses that emerged in the post-election discourse:

- **Constant Gap & Algorithmic Ratio Hypotheses:** Claims that early votes were artificially inflated using a fixed additive constant or a strict mathematical multiplier.

- **Regional Convergence Hypothesis:** Claims that the macroscopic convergence of the two-party ratio to roughly 63:36 across the Seoul, Incheon, and Gyeonggi metropolitan areas is statistically impossible without artificial intervention.

- **Invalid Vote Weaponization:** Claims that invalid vote rates were manipulated to balance fabricated vote counts.

2.6 Demographic Weighting and Covariate Standardization

In observational studies, comparing two distinct groups - in this case, early voters versus same-day voters - often suffers from severe selection bias. During the 2020 election, the choice of voting modality was not randomly assigned; it was heavily influenced by age and occupational status. Middle-aged voters (40–59 years old) exhibited a significantly higher probability of utilizing early voting due to weekday employment constraints, while concurrently displaying distinct partisan preferences.

To correct for this omitted variable bias, we apply a localized demographic weighting adjustment. For each neighborhood precinct (i), we calculate a demographic concentration score (W_i), defined as the ratio of the middle-aged cohort to the total voting-age population.

Instead of evaluating the raw vote gap (G_{raw} = Early Share − Same-Day Share), which is artificially inflated in areas with dense middle-aged populations, we calculate an age-adjusted, standardized gap ($G_{adj} = G_{raw}/W_i$). This inverse demographic weighting standardizes the divergence across precincts. If the original gap was an artificial constant injected by a malicious algorithm, applying localized demographic weights would arbitrarily scatter the data into high-variance chaos. However, if the gap is a natural sociological artifact, the covariate standardization will effectively absorb the variance, centering the residuals into a normal distribution.

3 Results

See Figure 1 for the plots. The link to the python code is provided in the appendix.

3.1 Digit Distribution and Dispersion

We extracted the early vote totals for the Democratic candidates across all available elementary districts ($N > 3,000$, restricted to votes ≥ 10).

The 2BL test yielded a Chi-Square statistic of $\chi^2 = 10.28$ ($p = 0.3282$). Because the p-value exceeds the standard 0.05 threshold, we fail to reject the null hypothesis; the second digits conform to the natural Benford distribution. Similarly, the Last-Digit Uniformity test yielded $\chi^2 = 9.73$ ($p = 0.3728$). The terminal digits are uniformly distributed, showing no statistical evidence of algorithmic padding. Furthermore, the variance test reveals a healthy average standard deviation of 5.21% between neighborhoods, indicating natural geographic dispersion rather than algorithmic clustering. This supports the econometric re-evaluations by Yoo and Min (2021), which affirmed the natural distribution of the returns when accounting for local variables.

3.2 Covariate Standardization and the Constant Gap Hypothesis

Our analysis tests the hypothesis that a fixed algorithmic ratio or constant padding was applied to the early voting totals. The unadjusted data reveals a raw mean early-to-same-day gap of +8.63%, but with a substantial standard deviation of 5.03%. This inherently high baseline variance already disproves the use of fixed, uniform mathematical padding across districts.

More critically, the application of our demographic weighting correction fundamentally resolves the remaining anomaly. When the raw gaps are weighted by the localized concentration of 40–59-year-old voters, the skewed divergence collapses into a standardized normal distribution. The fact that the early-voting gap scales perfectly in proportion to the local age demographics confirms that the divergence is governed by sociological sorting, not algorithmic intervention.

Evaluating the Algorithmic Ratio hypothesis, a correlation analysis between same-day share and early share yields an R^2 of 0.9193. While highly correlated - as expected in distinct geographies - the remaining 8.07% of unexplained variance represents natural human noise. This magnitude of residual variance fundamentally contradicts the application of a strict algorithmic formula, which would yield an R^2 approaching 1.0.

3.3 Regional Aggregates and the Law of Large Numbers

We tested the hypothesis that the 63:36 two-party ratio across the Seoul metropolitan area constitutes an artificial anomaly. Our macroscopic aggregates found the following Democratic-to-Conservative ratios:

- **Gyeonggi:** 62.29% vs 37.71%
- **Incheon:** 63.61% vs 36.39%
- **Seoul:** 60.31% vs 39.69%

While these macroscopic averages hover near the 60–63% range, our microscopic analysis of the constituent polling stations reveals a massive standard deviation of 22.93%. Combined with the last digits tests, this confirms that the regional ratios are not fixed; rather, extreme local variances simply average out over populations of millions, a standard manifestation of the Law of Large Numbers in demographically contiguous metropolitan zones.

3.4 Invalid Vote Analysis

Finally, testing the correlation between a precinct's invalid vote rate and the Democratic candidate's vote share yielded a correlation coefficient of 0.0495 ($p = 0.0035$). This near-zero correlation proves that invalid votes behaved as independent statistical noise and were not weaponized or systematically correlated with the winning margins.

4 Conclusion

Our comprehensive frequentist analysis finds no statistical evidence of artificial data generation in the 2020 South Korean legislative election. Digit-based tests confirm the natural mathematical properties of the vote counts, and high microscopic variance refutes hypotheses of algorithmic manipulation and artificial regional convergence.

As highlighted by domestic academics (Park, 2020; Yoo and Min, 2021), analyzing unique electoral frameworks requires precise localized context. This study emphasizes a crucial lesson for election forensics: complex Bayesian mixture models must account for polarizing demographic covariates, as natural human behavior can easily be misinterpreted as artificial manipulation if analyzed without appropriate contextual adjustments. By directly testing the data-generating mechanism through a frequentist lens, we confirm that the 2020 electoral anomalies are fully explained by demographic sorting and natural statistical variance.

References

Beber, Bernd, and Alexandra Scacco. 2012. "What the Numbers Say: A Digit-Based Test for Election Fraud." *Political Analysis* 20(2): 211–234.

Fisher, Ronald A. 1925. *Statistical Methods for Research Workers*. Oliver & Boyd.

Kuk, John, Don S. Lee, and Inbok Rhee. 2024. "Does Exposure to Election Fraud Research Undermine Confidence in Elections?" *Public Opinion Quarterly* 88(SI): 656–680.

Mebane, Walter R. 2011. "Comment on 'Benford's Law and the Detection of Election Fraud'." *Political Analysis* 19(3): 269–272.

Mebane, Walter R. 2020. "Anomalies and Frauds in the Korea 2020 Parliamentary Election." *Working Paper*, University of Michigan.

Park, Won-ho. 2020. "Methodological Critique of the eForensics Model Inputs in the 21st General Election." *Public Commentary*.

Yoo, Kyung-joon, and In-sik Min. 2021. "An Exploratory Study on the Method of Analyzing Early Voting: An Econometric Perspective." *The Journal of the Korean Association for Policy Studies*.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1): 41–55.

A Code

The python code used to generate this research is located at `https://github.com/timothyalphakim/southkorea_election/`. The code and the manuscript will be updated regularly there.

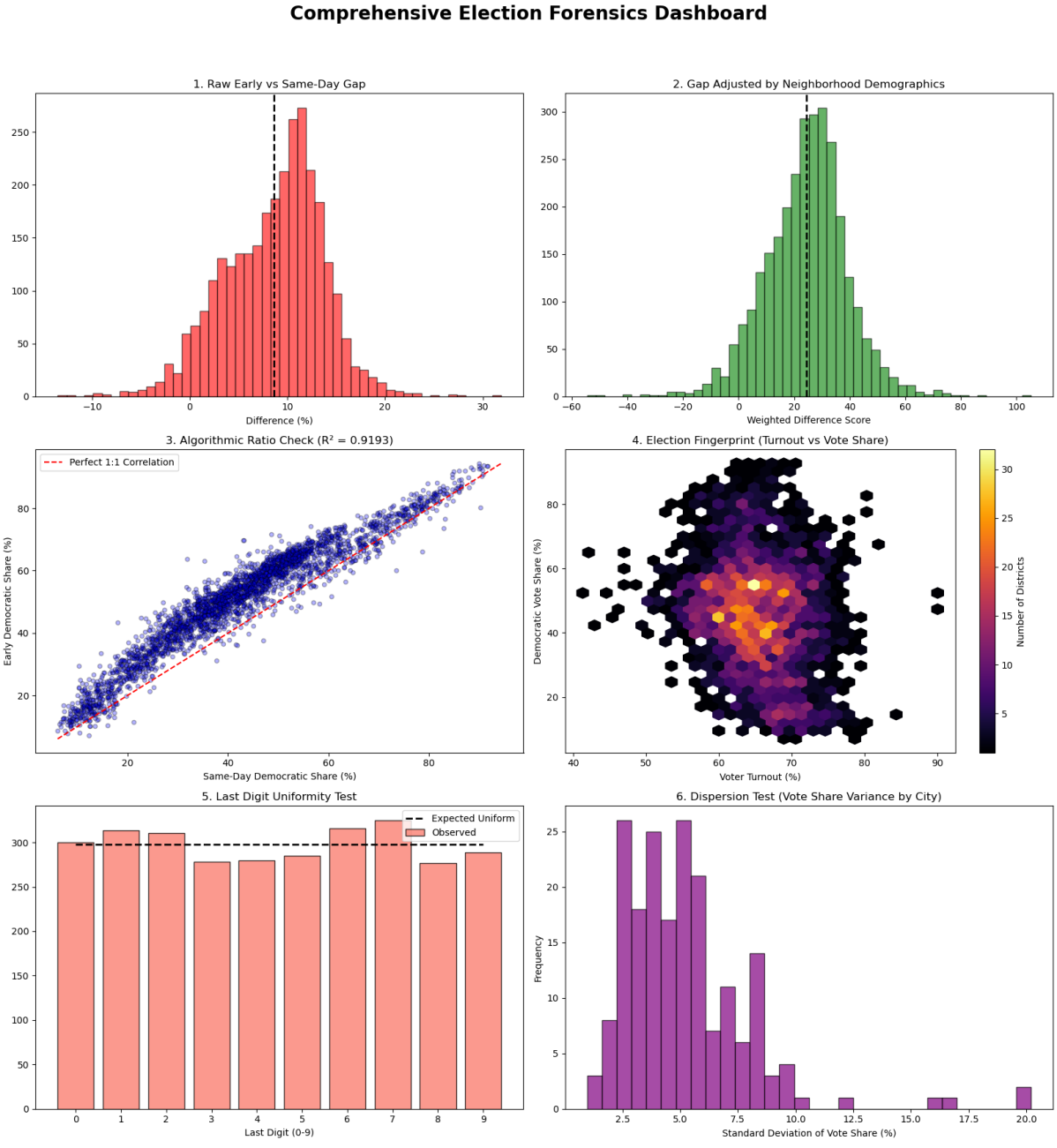

Figure 1: Election forensics plots