Analysis of a Variable Region Size in Regional Voting

Jared Hirt
University of Northern British Columbia

Dr. Liang Chen
University of Northern British Columbia

***Abstract*—Choosing the wrong region size in a regional election can result in electing a candidate that is not preferred by the majority of the population. We propose a method of conducting regional elections where the election is conducted several times on the same sample, but with different region sizes considered. A candidate is then elected if more than half of the regions vote for them. This approach has the benefit of increased stability present in regional elections, but also mitigates the chance of choosing an inappropriate region size. Our proposed model offers a formula for calculating the stability of an election with multiple region sizes, and with a set amount of blocks of concentrated noise of different sizes.**

***Index Terms*—National Voting, Regional Voting, Image Processing, Stability, Noise, Concentrated Noise, Electoral College, Artificial Intelligence.**

I. Introduction

Voting is a common way to make distributed decisions. This holds relevance to Computer Science as elections are commonly used to reach a consensus in many computer systems. This is used in Computer Vision to determine what object a picture of something is, in Large Language Models to determine what is true and what is disinformation, and in data science to make conclusions from multiple sources of data.

For the public, voting is often seen in governments where they hold regular elections to elect a new leader. The simplest form of voting is a direct vote. The direct vote has voters cast a vote for a candidate, and the candidate with the highest number of votes wins. The regional voting system is another commonly used system. The regional voting system has voters segmented into different regions. The regions then hold a direct vote election to determine which candidate wins the region. The candidate who wins the highest number of regions is then elected. As found by Chen[1], regional voting is more stable, which means that given concentrated areas of noise, the candidate supported by the majority of the population is more likely to be elected.

A concentrated region of noise is defined as a cluster in a sample that does not reflect the population and vote against the majority candidate due to sampling errors or environmental factors. An example of this could be seen if there was a region that was slightly in favour of candidate A over candidate B. If there were 3 voting stations in the region, and two of them were in areas with a higher density of supporters of candidate B, it would constitute result in a higher representation of voters for candidate B. An example of a population with concentrated noise is shown in figure 1.

The next definition is stability. Stability is the probability that given a sample, support rate for the majority candidate, and the amount of concentrated noise blocks of each size; What is the probability that the majority candidate is elected.

When using a regional model, you need to intelligently choose the size of the regions balancing two properties. The first property is when you increase the region size there is a highly likelihood that the region is polluted with noise. The second property is that when you decrease the region size, there is a larger chance that an unpolluted region will vote against the candidate that is preferred by the majority.

This paper intends to improve on the model in three ways. Firstly, we consider all possible regions of the size specified, allowing for edge wrap around to allow fair representation of voters on the edge and corner. This is shown in figure 2.

Secondly, we account for a set number of blocks of

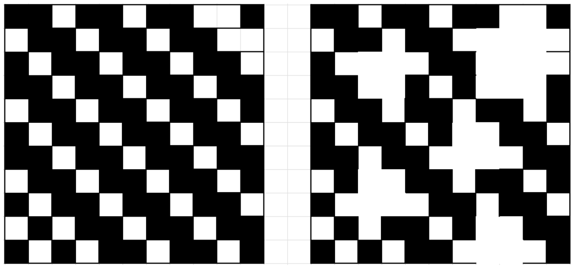

Fig. 1. The example on the left shows a sample without areas of concentrated noise. The two candidates are represented by the white and black cells within the 11×11 square. The majority candidate is black which has 78 cells compared to white which has 43 cells. The sample on the left is the same area but with 4 blocks of 2×2 concentrated noise, and 1 block of 3×3 concentrated noise. Given a direct vote this would lead to 58 votes for black and 63 votes for white. This would falsely elect white purely due to the blocks of concentrated noise. The goal of using a regional voting system is to break the vote into many regions which would forfeit the vote of regions polluted by noise, but making unpolluted regions more likely to vote for the majority candidate.

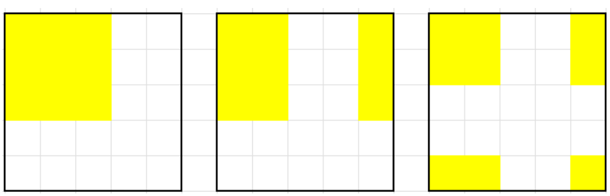

Fig. 2. Valid regions of size 3×3 with a population size of 5×5

concentrated noise of different sizes. This enables analysis in different scenarios, where the size and quantity of blocks of concentrated noise may differ.

Finally, we conduct the election several times with different region sizes and we then choose the candidate with the most regions won. We hypothesize that by allowing regions of multiple sizes you can mitigate the risk of choosing an inappropriate region size given the noise.

II. Model and Assumptions

A. Model

We define the following:

- N: Population size of $n \times n$
- $\vec{r}$: a vector containing the region sizes
  - ex. <3, 5, 7, ...>, this represents an election where we consider the region sizes of $3, 5, 7, \ldots$.
- t: the number of region sizes
- $\vec{m}$: a vector containing the quantity of square noise blocks of different sizes
  - ex.<0, 4, 1...>, this represents zero 1×1 noise blocks, four 2×2 noise blocks, one 3×3 noise blocks, and so on.
- k: the maximum noise block size
- p: The support rate of the population for candidate A

For this simplified model we will treat the population as if it were a perfect square of size $n \times n$. The population is then affected by concentrated noise which forms in square blocks. The exact number and size of noise region is found within $\vec{m}$. We then conduct a regional election using the region sizes in $\vec{r}$, and considering every possible region of that size located continuously in the population. We also allow a region to wrap around the edges as shown in figure 2. The winning candidate is then chosen by the candidate that won the largest number of regions.

B. Assumptions

- We assume that when a region is polluted, even if only by a tiny bit that it votes against the candidate A.
- We assume that the support for a candidate is uniformly distributed before noise is applied to the model.

III. Main Theorems

Lemma III.1. *Number of regions of size $r \times r$*

Given a region of size $r \times r$ it can be shown that there are N regions, where $N = n \times n$, provided that $r < n$.

All regions can be found by placing the top left of the region on every cell in the population. Each of these regions will contain a unique combination of cells, and there will be a total of N of them.

Lemma III.2. *Number of regions polluted per noise block.*

Given a noise block width and length i and a region width and length r, the number of polluted regions from this noise block is:

$$(r+i-1)^2 \tag{1}$$

This can be seen by putting the noise block and the region on a grid. If you put the noise block i cells away from the left side of the region then it will not overlap. However if you put it somewhere in the range of [0, i-1] cells away the regions will overlap. Similar logic follows for placing noise blocks above the region.

This creates a square where if the top left corner of the noise block is placed, then the region will be polluted.

Lemma III.3. *Probability of a region being pollution free for all noise blocks.*

Given a vector of noise block sizes $\vec{m}$ the probability of a region of size r being pollution free is:

$$\left[\prod_{i=1}^{k}\left(1-\frac{(r+i-1)^2}{N}\right)^{m_i}\right] \tag{2}$$

This calculates the probability of a region being pollution free by subtracting the number of polluted regions per region size divided by the total number of regions. Where the number of polluted regions is found from equation 2.

We then raise that to the power of how many regions are of that size to get the probability of a region being pollution free given the noise block size of i.

We then iterate through every i in the vector of noise blocks $\vec{m}$ multiplying the results to get the total probability

Theorem III.1. *Stability given a single region size*

$$\sum_{x=\lceil\frac{N}{2}\rceil}^{N}\left[\binom{N}{x}\left[\prod_{i=1}^{k}\left(1-\frac{(r+i-1)^2}{N}\right)^{m_i}\right]^{x}\left(1-\left[\prod_{i=1}^{k}\left(1-\frac{(r+i-1)^2}{N}\right)^{m_i}\right]\right)^{N-x}\left[\sum_{j=\lceil\frac{N}{2}\rceil}^{x}\binom{x}{j}p^{j}(1-p)^{x-j}\right]\right] \tag{3}$$

Proof. We get the probability of a region of size r being pollution free from equation 2.

We then use the binomial distribution formula as shown in equation 4.

$$\sum_{x=l}^{u}\binom{n}{i}p^{i}(1-p)^{n-i} \tag{4}$$

We then substitute in equation 2 for p, N for n, and $\lceil N/2\rceil$ for l. $\lceil N/2\rceil$ is chosen for the lower bound as that is the minimum number of regions that must be won.

We then multiply this with the probability of $\lceil N/2\rceil$ or more regions being chosen out of x total regions using the binomial distribution formula as shown in equation 4.

□

Lemma III.4. *Probability of exactly y regions of size r voting for candidate A*

$$\sum_{x=0}^{N}\left[\binom{N}{x}\left[\prod_{i=1}^{k}\left(1-\frac{(r+i-1)^2}{N}\right)^{m_i}\right]^{x}\left(1-\left[\prod_{i=1}^{k}\left(1-\frac{(r+i-1)^2}{N}\right)^{m_i}\right]\right)^{N-x}\binom{x}{y}p^{y}(1-p)^{x-y}\right] \tag{5}$$

Proof. We consider Theorem III.1 which shows the probability of candidate A being elected given a vector of noise blocks $\vec{m}$.

We then modify the formula by replacing the lower bound with 0 as we want to consider less than the majority number of regions won.

We then change the probability of $\lceil N/2\rceil$ or more regions voting for candidate A for all j up to x, to exactly y of x regions voting for candidate A.

We then define this as equation 6.

$$P(y,r) \tag{6}$$

□

Theorem III.2. *Probability of electing candidate A*

$$\sum_{x=\lceil \frac{Nt}{2} \rceil}^{Nt} \left[\sum_{y_1=0}^{N} P(y_1, r_1) \left[\sum_{y_2=0}^{N} P(y_2, r_2) \cdots \right.\right.$$
$$\left[\sum_{y_{t-1}=0}^{N} P(y_{t-1}, r_{t-1}) P(x - y_1 - y_2 \cdots - y_{t-1}, r_t) \right.$$
$$\left.\left.\left. \right] \cdots \right]\right] \tag{7}$$

Proof. We begin deriving this equation by applying equation 6 to every region size in $\vec{r}$. We then consider every possible number of regions voting for A and finding the probability. We then consider all cases where the y in for every region size adds up to x. We then perform a summation for all x where x is enough regions to win the election. □

Lemma III.5. *The expected mean of the combined probability distribution*

$$\mu = N \sum_{j=1}^{t} \left[\left[\prod_{i=1}^{k} \left(1 - \frac{(r_j + i - 1)^2}{N} \right)^{m_i} \right] \sum_{x=\lceil \frac{r_j^2}{2} \rceil}^{r_j^2} \binom{r_j^2}{x} p^x (1-p)^{r_j^2 - x} \right] \tag{8}$$

The equation 8 utilizes the mean of a binomial distribution with n trials and probability p is found in equation 9

$$\mu = np \tag{9}$$

We then add up the means of each region size r with N trials per r. Where the probability is found from equation 2 multiplied by the probability of a region of size r voting for candidate A.

Lemma III.6. *Approximation of the p term in equation 9.*

The inner term of the probability distribution can become computationally complex for large region sizes. This can be alleviated by approximating the probability of an unpolluted region of size r_j electing candidate A. This is done by approximating the binomial distribution with a normal distribution where the Z score is found in equation 10.

$$Z = \frac{\lceil \frac{r_j^2}{2} \rceil - (r_j^2)p}{\sqrt{(r_j^2)p(1-p)}} \tag{10}$$

Lemma III.7. *The variance of the combined probability distribution*

$$\sigma^2 = N \sum_{j=1}^{t} \left[\left(\left(\left[\prod_{i=1}^{k} \left(1 - \frac{(r_j + i - 1)^2}{N} \right)^{m_i} \right] \sum_{x=\lceil \frac{r_j^2}{2} \rceil}^{r_j^2} \binom{r_j^2}{x} p^x (1-p)^{r_j^2 - x} \right) \left(1 - \left(\left[\prod_{i=1}^{k} \left(1 - \frac{(r_j + i - 1)^2}{N} \right)^{m_i} \right] \sum_{x=\lceil \frac{r_j^2}{2} \rceil}^{r_j^2} \binom{r_j^2}{x} p^x (1-p)^{r_j^2 - x} \right) \right) \right) \right] \tag{11}$$

This follows the formula for the variance of a binomial distribution which is shown in equation 12.

$$\sigma^2 = np(1-p) \tag{12}$$

We can then sum the variance of each different probability distribution given by Theorem III.2. The p is then found the same way as in lemma III.5, and can be approximated in the same way as lemma III.6 for large region sizes.

Theorem III.3. *Approximation of stability given a vector $\vec{m}$ of noise block sizes and a vector $\vec{r}$ of region sizes.*

$$Z = \frac{\frac{Nt}{2} - \mu}{\sigma} \tag{13}$$

The mean and standard deviation can be found using lemmas III.6 and III.7. The threshold is set for $Nt/2$ as that is the number of regions a candidate must win in order to win the election.

IV. Results

We can create a sample election to compare the stability between the traditional single region size model and the improved variable region size model. We create the sample with the following parameters.

- $N = 2^{30}$
- $p = 0.501$

We then consider the case where we use the region sizes of length and width 100, 200, and 300.

We will then perform 4 different trials for each regions size by using different amount of noise blocks of length and width 100, 200, 300, and 400. We will then compare whether the addition of multiple region sizes improves the stability on average.

A. increase in width 100 noise blocks

In this trial we will consider only noise blocks of size 100×100. We will start with 0 noise blocks then increase by increments of 1. We will then plot the stability of each region size.

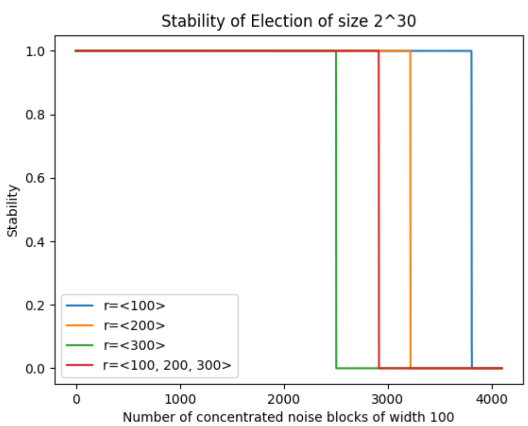

Fig. 3. This sample election increases the number of noise blocks of size 100×100. This is the smallest noise block size considered in this paper, and shows that the smaller region sizes perform better with the larger region sizes performing worse.

B. increase in width 200 noise blocks

In this trial we will consider only noise blocks of size 200×200. We will start with 0 noise blocks then increase by increments of 1. We will then plot the stability of each region size.

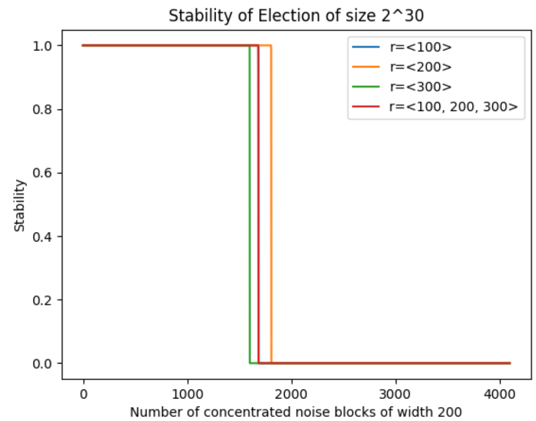

Fig. 4. This sample election increases the number of noise blocks of size 200×200. With this larger size we see the region size of 200 become more optimal than the region size of 100. This follows the trend of larger region sizes becoming more optimal as the size of the noise blocks increases.

C. increase in width 300 noise blocks

In this trial we will consider only noise blocks of size 300×300. We will start with 0 noise blocks then increase by increments of 1. We will then plot the stability of each region size.

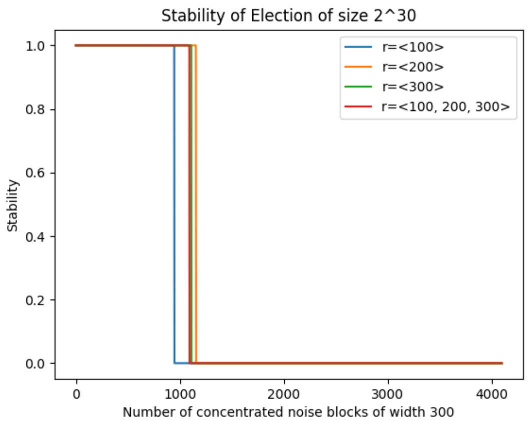

Fig. 5. This sample election increases the number of noise blocks of size 300×300. With this noise block size he region size of 200 stays the most optimal. However, effectiveness of regions of size 100 drastically decreases while the effectiveness of regions of size 300 increases.

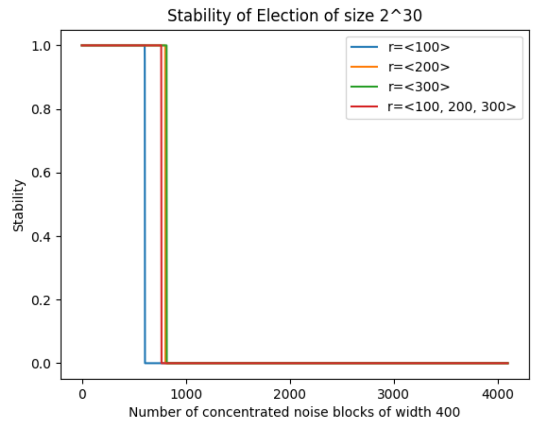

Fig. 6. This sample election increases the number of noise blocks of size 400×400. This is the largest noise block size we will be considering in this paper. With this noise block size, the region size of 300 narrowly becomes optimal, overtaking the region size of 200. Larger region sizes becoming more optimal follows the trends noted earlier of the larger the noise block size, the larger the optimal region size.

D. *increase in width 400 noise blocks*

In this trial we will consider only noise blocks of size 400×400. We will start with 0 noise blocks then increase by increments of 1. We will then plot the stability of each region size.

E. *Final Results*

Throughout the trials, different region sizes performed differently depending on the characteristics of the concentrated noise. Each region size in the election had a range of values for which an election with that region size would be stable. Below is a chart containing the maximum stable values as found in the experiment.

	$< 100 >$	$< 200 >$	$< 300 >$	$< 100, 200, 300 >$
100	3805	3216	2502	2913
200	1685	1806	1599	1682
300	946	1154	1110	1091
400	605	801	815	764

Fig. 7. This table shows the region sizes along the top, with the noise block size on the left. The values represent the highest number of noise blocks such that the region size remains stable. This follows a trend of larger region sizes becoming more optimal as the noise block sizes become larger.

We can then further process these values to give a percentage compared to the best found region size. This is found by taking the highest stable noise size for a region size, then dividing it by the maximum found amongst all region sizes. This is necessary for analysis, as larger noise blocks contain more noise per block, so only comparing the number of noise blocks would not be a meaningful comparison.

	$< 100 >$	$< 200 >$	$< 300 >$	$< 100, 200, 300 >$
100	100%	85%	66%	77%
200	93%	100%	89%	93%
300	82%	100%	96%	95%
400	74%	98%	100%	94%

Fig. 8. This table shows how large the range of stability is per region size. This compares the range of stability per region size with the range of stability of the best region size per experiment.

We then perform statistical analysis on the percentages. Which yield the following statistical values.

	$< 100 >$	$< 200 >$	$< 300 >$	$< 100, 200, 300 >$
Average	87%	96%	88%	89%
Median	88%	99%	92%	93%
Max	100%	100%	100%	95%
Min	74%	85%	66%	77%
Standard Deviation	11%	7%	15%	9%

Fig. 9. This table showcases various methods of statistical analysis on the table in figure 8. An important factor to note in these trials is that the region size of 200 is better than combining the region sizes of 100, 200, and 300. Similar results would occur if the noise block sizes tested were smaller or larger, except the region size of 100 or 300 would become the best.

These statistical results reaffirm that with an series of elections with different noise block sizes, there is an optimal size. However, using the multi-region size model you get results better than the worst, and worse than the best. This reduces the likelihood of choosing a bad region size for an election based on the characteristics of the concentrated noise.

V. Discussions

We hypothesized that by allowing the election to occur multiple times with different region sizes it allows you to conduct a stabler election on average. As shown by the experiments we see that when multiple region sizes are considered, it has a range of stability greater than the worst region size, but worse than the best region size. We also note the following characteristics from the experiments conducted above:

- If the size of noise blocks get smaller then the optimal region size gets smaller, as shown by figures 3 - 6.
- If the size of noise blocks get bigger then the optimal region size gets larger, as shown by figures 3 - 6.
- The stability of an election given set noise is almost always 1 or 0, as shown in figures 3 - 6.
- Choosing the best region size involves choosing a large enough size that it increases the probability of an unpolluted region voting for the correct candidate, as well as not making the region size too large such that a higher number of regions are polluted.
- Conducting the election with multiple region sizes mitigates the risk of choosing a wrong region size.

Based on these observations, it makes sense to conduct elections with the proposed model when there is uncertainty about the characteristics of concentrated noise in the poll, as you reduce the impact of choosing an incorrect region size for the size and amounts of noise.

The reason for why using multiple region sizes is more stable than the worst region size, and less stable than the best region size is similar to that of taking the mean of a set of data. When you take the mean of a set of data, the result is larger than the smallest element in the set, and smaller than the larger element of the set.

VI. Conclusion

Elections are important for any democratic society, however elections can often be plagued with areas of concentrated noise. The effect of these areas of concentrated noise can be mitigated by using a regional voting system as opposed to a direct voting system. When conducting the regional voting you need to choose a region size such that it is small enough to be unlikely to be polluted by noise, but large enough to have a high probability of voting for the candidate the majority of the population supports.

This research improves upon traditional regional voting by proposing that an election where characteristics about concentrated noise is unknown, is conducted multiple times with different region sizes. This model ensures that the range of stability for the election is greater than that of the worst region chosen. This mitigates the effect of choosing a bad region size and that negatively influencing the results of the election.

The contributions of this paper are:

- Creating a model to calculate the regions of stability of an election when you consider multiple region sizes and multiple noise block sizes.
- Showing that performing an election with multiple region sizes results in an election that is not as stable as the best region size, but stabler than the worst region size.

VII. Future Work

For future work, further analysis on when the stability switches from 1 to 0 should be done. Additionally, further analysis is needed for how to choose the optimal region sizes for an election when some characteristics of the noise are known, and whether weighting some region sizes more than others produces more statistically meaningful results. Finally, additional research is needed on how varying amounts of noise blocks of different sizes affects which region size is optimal.

References

[1] L. Chen and N. Tokuda, "A general stability analysis on regional and national voting schemes against noise—why is an electoral college more stable than a direct popular election?" *Artificial Intelligence*, vol. 163, no. 1, pp. 47–66, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370204001560