Investigating Different Polling Methods for Election Surveys

Michele Pascale

[1] University of Leeds, UK
[2] mm20mp@leeds.ac.uk

Abstract. *In order to ensure opinion polls are reflective of a general population, a broad demographic must be sampled, ideally at random. Because of this, these subsets must also be unbiased and well-informed in order to reduce potential opinion swaying and misrepresentation bias from media outlets. Polling companies hold major responsibility when it comes to both introducing and alleviating bias, especially when these results are being presented by large corporations and authorities to the public. Because of the broad spectrum of polling companies that currently exist, their methodologies and tactics are not uniform, and so it is advantageous to investigate the different methodologies and trends that a sample of these polls employ.*

1 Current Polling Methodologies

There remains to be seen a uniform method of assessing public opinion of election outcome and voting preference due to the vast array of different surveying methodologies that have been employed in the past. *"Formal"* methodologies (i.e. those that employ a systematic approach to ascertain public opinion) are usually employed by companies specialising in public opinion research, and have a much better understanding and method of evaluating the results that are produced in a more representative and unbiased manner. However, this is not always the case and the companies alone cannot be used as a metric of validity when it comes to the outcomes of these polls.

1.1 Probabilistic Sampling Methods

Probabilistic sampling is a sampling method that relies on an unbiased estimator for a given variable that is being investigated; for example surveying the probability of one candidate winning a majority election, given a population sample. The predominant advantage of probability-based sampling is that bias metrics, correlations and significance tests can be carried out in order to infer the efficacy of an employed survey. In particular, (Lavrakas 2008) outlines a plethora of modern probabilistic approaches to sampling and surveying populations, most of which are currently employed by polling companies when conducting their own quantitative research; a few examples of these are described below.

1.1.1 Random Digit Dialling

As outlined by Wolter et al. (2009), *Random Digit Dialling* (or *RDD* for short) is a probabilistic sampling technique used to randomly sample a collection of households, families and / or people via random digit generation, followed by selection of their telephone numbers. By proxy, researchers are able to quantify a collection of people (potentially a household) by sampling only one number; this is a particularly advantageous approach as numbers can be used to sample populations potentially greater than 1 person in different regions, and at random, which employs both an unbiased sample as well as groupings of people sharing similar interests and views based on households alone. These metrics, whilst not necessarily critical or always advantageous for researchers, provide an extra insight into the data collected that may be used if deemed necessary for further insight.

1.1.2 Registration Based Sampling

Registration Based Sampling (or *RBS* for short) is a sampling methodology that uses a database (usually pre-existing) consisting of personal information regarding registered voters; this generally includes their names, home addresses and telephone numbers. Unlike *RDD*, *RBS* does not leave the individual being surveyed anonymous and thus provides better outcome in terms of co-operation on the individuals behalf (Lavrakas 2008).

1.2 Non-Probabilistic Sampling Methods

Unlike probabilistic sampling methods, non-probabilistic (otherwise known as qualitative) research methods rely on individual collection, analysis and inference of non-numeric data.

1.2.1 Self-Selected Sampling

At its heart, *self-selected sampling* is a means of allowing individuals to voluntarily choose whether or not to take part in a survey; this does not necessarily have to be explicit based on the individual's consent, but rather can be implicit by nature of a given survey (i.e. choosing which religion, if any, that you feel that you most align with, is implicitly grouping an individual based on their religious beliefs whether they realise this or not.) As a result of this, there can be what is known as *self-selection bias* which arises when individuals explicitly decide whether or not to take part in a given survey, evidently leading to a biased dataset that is likely not representative of an entire target population (Lavrakas 2008).

1.2.2 Online / Internet Sampling

As outlined by (PRC 2021), *"online"* or internet surveys are relatively modern approaches to broad opinion inference from populations of interest. There still remains no general method for attaining personal information, such as telephone numbers or any aforementioned information for that matter, which serves as the initial point of contact between the surveyor and the individual at hand. As a result, internet polls can be seen as somewhat of a subsidiary of traditional polling methods, such as *RDD*, as

prior information about individuals must be known in order to contact them to complete an internet survey.

2 Comparison of Polling Methods

2.1 Advantages and Disadvantages of Probabilistic Methods

Although probabilistic sampling is the preferred method of research when conducting opinion polls, these samples are not without their flaws. In general, when comparing *RDD* and *RBS*, it is relatively difficult to pick a clear winner in terms of best general efficacy. In a report by Kennedy et al. (2018), the relative advantages and disadvantages were weighed and assessed; below are some figures illustrating some of the findings.

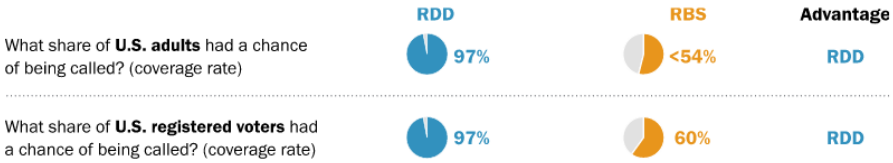

Fig. 1. Figure from Kennedy et al. (2018) denoting some advantages of RDD.

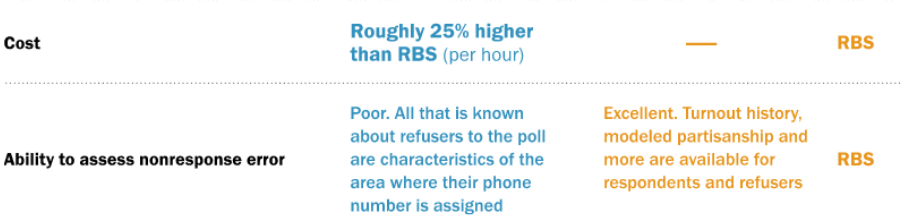

Fig. 2. Figure from Kennedy et al. (2018) denoting some advantages of RBS.

As can clearly be seen in Figure 1 and Figure 2, there are relative advantages and respective disadvantages to both approaches, and thus there needs to be consideration of each methodology based on what a given survey or poll is seeking to achieve.

2.2 Advantages and Disadvantages of Non-Probabilistic Methods

Rahman (2016) outlines that some of the major advantages involving qualitative methods include the depth of individual information that can be ascertained

from a sample; for example a person's feelings, opinions and experiences - all of which are deemed incredibly valuable in opinion polls. But arguably, there is also an inherent major downside to this method of sampling; for example, potential bias can be ascertained voluntarily (or even involuntarily) by the researchers, focusing too much on individual experience and less on a broad representation of a population. In particular, an educational study by Cumming (2001) displayed that when focusing on individual experience alone, the outcomes were far too narrow, pertaining to potential selection bias with regards to the individuals chosen. In a study involving six different countries, writing instructors with domain expertise were surveyed and when presented with the results he stated, "This sampling was selective and purposive, focused on instructors with high levels of expertise in each setting, rather than aiming to be representative of educators in the particular countries or institutions." Therefore, care must be taken when selecting individuals and data to survey qualitatively, especially for companies generating opinion polls, as bias is undeniably detrimental.

3 Initial Exploratory Analysis of Opinion Polls Data

In order to investigate factors that result in good outcomes from opinion polls, it is advantageous to perform exploratory data analysis on a dataset consisting of opinion polls, along with their other various metrics. By performing this analysis, hopefully there will be some insight gained into strategies and techniques that not only contribute to more accurate and representative polls, but also other insights, potentially into those subsets of individuals who are surveyed.

3.1 Data Description

The dataset that has been chosen is the *FiveThirtyEight* U.S. Presidential Election Polling Dataset, attained from the following URL: `https://projects.fivethirtyeight.com/2016-election-forecast/?ex_cid=rrpromo#plus`. The dataset is formatted as a `CSV` file with 26 variables in total, with the poll results aggregated from *HuffPost Pollster*, *RealClearPolitics*, polling firms and various news reports (each of which is outlined and linked in the `CSV` file); further to this, there are a total of 12625 row entries for each respective variable. More intricate details of the variables and the methods used to generate the polls are outlined here: `https://fivethirtyeight.com/features/a-users-guide-to-fivethirtyeights-2016-general-election-forecast/`

3.2 Data Exploration

Before any analysis, the dataset must be investigated to allow for a better understanding and to discover any initial issues. In particular, there are many empty variable entries, missing row entries and irrelevant or ambiguous variables - these will have to be adequately accounted for or entirely removed prior to any statistical inference, in order to ensure that the dataset remains cohesive and reflective of it's original population.

In particular, it's important to consider the following issues:

1. **Removal of candidate entries**
   - In this instance, it can be seen that McMullin has no entries within their columns and so this data can be removed entirely. This results in columns `rawpoll_mcmullin` and `adjpoll_mcmullin` being dropped from th `CSV` file.

   - Further to this, it is advantageous to also remove the columns containing another electoral candidate, namely Johnson, with variable headers `rawpoll_johnson` and `adjpoll_johnson` respectively.

2. **Removal of remaining empty variables and missing values**
   - Further to removal of empty or irrelevant candidates, it can be seen that the variable `multiversions` has only a few entries out of the vast dataset, so this can also be excluded as it doesn't seem to carry much information.

3. **Potential Conversion of Non-Numeric Values and Standardisation**
   - As can clearly be seen, many of the variables contain non-numeric data, namely: `cycle,branch,state,matchup`, for example. This may be detrimental when performing certain analyses that require numeric values only, so this needs to also be taken into consideration.

   - Similarly, it may be advantageous to standardise each of the numeric values within a given range in order to allow different techniques to be utilised (standard distributions for example and activation functions for example) but this may not necessarily be required in all cases, but is important to note.

3.3 Removal vs. Synthetic Generation of Missing Values

One inherent issue with removing values is that it drastically reduces the size of the dataset and therefore provides less examples to analyse, which is of utmost importance when it comes to training machine learning models and classifiers for example. Further to this, removal of missing values may result in a statistical disparity between certain metrics; namely, estimations for population mean and standard deviation that would have been observed will now be more statistically insignificant due to the reduction in the size of the sample. However, a multitude of methods exist such that missing data can be adequately handled, without removing values. In particular, *single imputation* is described as a family of methods such that missing or null values can be replaced, consisting of techniques such as *last-value replacement, mean replacement* and *single-regression replacement* (Curley et al. 2019). For example, *single-regression replacement* enlists regression analysis as a somewhat *"predictor"* of the missing value, based on prior values. However, a major issue with single regression replacement is that upon regression analysis the value is sampled only once, meaning that the

rate of occurrence of a value may not be accurately representative based on prior beliefs or results; in this instance, the predicted value will not be as informative and may be somewhat biased. Therefore, it is of utmost important to consider and weigh the benefits and risks of each type solution to missing-values.

3.4 Data Pre-Processing

3.4.1 Culling of Irrelevant Variables The `CSV` file was pre-processed using a `Python` script and the `pandas` library. For ease of use, the `CSV` file was converted into a `DataFrame` object prior to any pre-processing. Initially, all electoral candidates that were not being considered (as outlined in Section 3.2) had their columns dropped from the `dataframe`, reducing the variable count to 24. Further to this, variables `multiversions` and `url` were subsequently dropped, as they likely would not contribute to any significant statistical changes. The same holds for variables `cycle,branch,matchup` and `timestamp`. Finally, `forecastdate` was also removed, as it remained consistent throughout all entries and posed no real benefit to the data; thus leaving 17 variables in total, with the likelihood of more potentially being removed dependent on the analysis that is being considered.

3.4.2 Missing Value Removal In terms of missing values, it was weighed that any missing values are to be removed to ensure consistency within the data. Further to this, the number of row entries exceeds 12000, eluding to the fact that a removal of data-points should not result in a dataset that is not reflective of the populations surveyed, given the mass of data that is still available. The approach used to remove missing values was effectively to find any rows that contained any missing values and then omit them from the processed output `CSV` file. Upon inspection after the removal of values, the number of entries declined from 12625 to 11335; thus resulting in a reduction of $\approx 11\%$. It should be noted that, while this can be deemed as a substantial reduction in terms of variables, there is still be sufficient row / entry data to hopefully allow for sound statistical inference and analysis.

4 Inferential Analysis of Opinion Polls

The analysis required for this section was conducted using `R` and a variety of packages to allow visualisation and inference throughout. Unless otherwise mentioned, it should be assumed that any results were generated from `R` exclusively.

4.1 Polling Results by Poll Grades

When considering the validity of opinion polls, companies will strive to achieve a higher "grading" relative to competing polls, but this poses a question; does this constitute a more reflective prediction? Or is the higher grade affiliated with how

"popular" the current opinion of a polling outcome is? In order to investigate this, it is useful to analyse the relative predictions over time for each grading of poll.

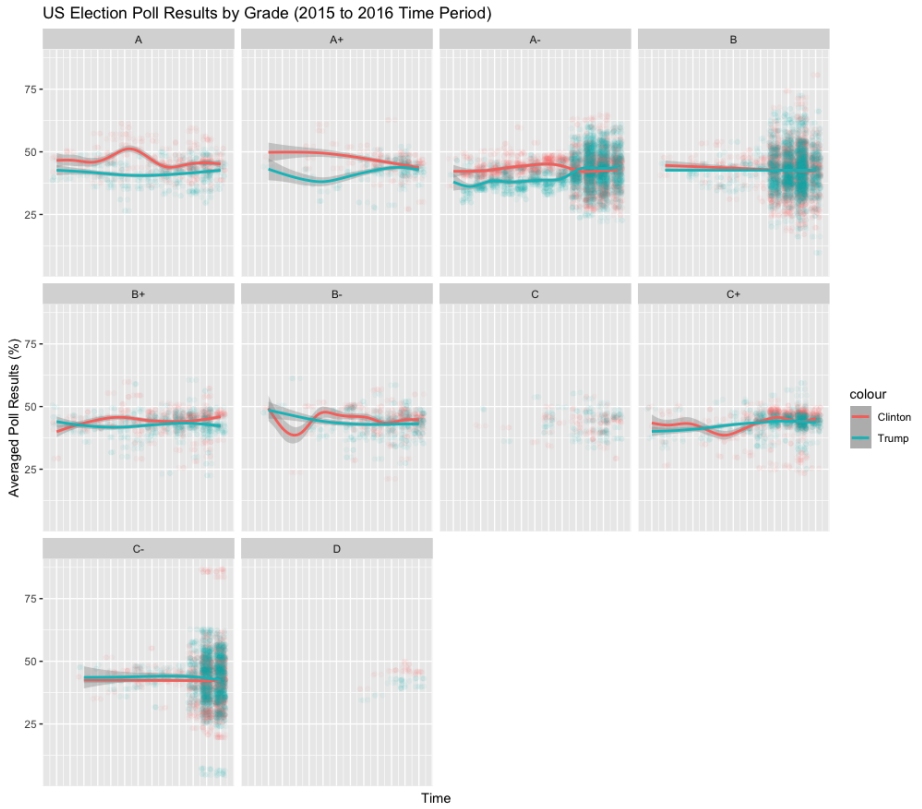

Fig. 3. Time-series plot of election polling results over time, grouped by relative grading.

Immediately, it can be seen in Figure 3 that there is a clear trend here - in particular the higher "graded" polls tend to give favour to Clinton more prominently, but as the poll ratings decline it can be seen that the advantage is not immediately obvious, tending between the two electoral candidates over time, which is what one might expect to see. Also, it is important to note here that likely as a result of data cleaning and potential class imbalance, the data for grades "C" and "D" respectively have no associated time series plots. This is due to no immediate trend being visible in the data due to the sparsity of "C" and "D" grade polls, again which is to be expected. This begs the question however, the higher ranked polls tend to favour Clinton more heavily, so potentially this may have an association with the polling location? For instance, higher ranked

polls may have been conducted either in specific states, pertaining to potential bias for Clinton. This hypothesis can be more easily tested via the means of a *chloropleth* diagram, which is displayed in the following section.

4.2 Polling Results by Poll Grades and Location

In order to infer whether location of the target poll has an influence on the poll rating and its preferred candidate, it's intuitive to investigate where these polls were employed to be able to see which demographics (based on U.S. states) were surveyed and whether this could pertain to potential bias. A *chloropleth heatmap diagram* is effectively a map with a projection of data onto constituent locations in the form of a heatmap. This allows the visualisation of which states constitute to a higher polling majority and allows inference of individual grades of polls, to ascertain if there are in fact any discrepancies.

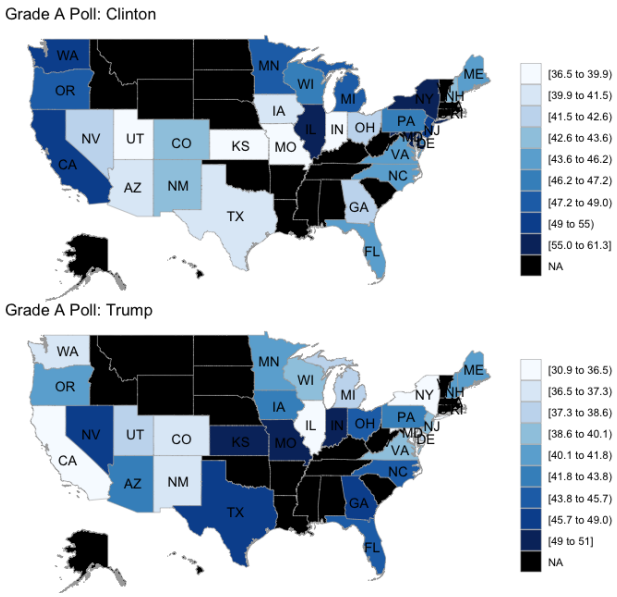

Fig. 4. Chloropleth heatmap for Grade A Polls, displaying adjusted polls per state for both Trump and Clinton.

Figure 4 displays that indeed there is evidently a Trump majority in key states such as `TX` and `AZ`, but this isn't necessarily reflective of all of the polls, which is more evidently apparent when comparing lower graded polls.

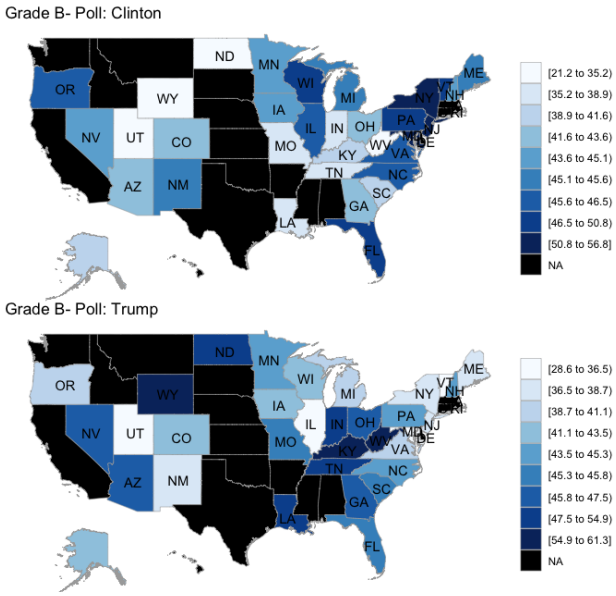

Fig. 5. Chloropleth heatmap for Grade B- Polls, displaying adjusted polls per state for both Trump and Clinton.

As can be seen in Figure 4, state `NC` displays a higher outcome for Clinton, whereas in Figure 5 there is clearly a majority for Trump. Further to this, it can be seen that in there is a new additional state added (`WY`) which is clearly a Trump majority, pertaining to the fact that there is initially a mild lead by Trump in the plot in Figure 3 as time elapses. Another item to note is that state `TX` has been removed in this poll, which gave Trump a lead in the Grade A polls, yet in the Grade B- polls, this has evidently been omitted, potentially pertaining to some selection-bias, selecting states to favour Clinton for example, on behalf of the pollsters.

4.3 Dimensionality Reduction via Principle Component Analysis

Analysing the distribution of polls by relative grade and location is an intuitive approach to visualise differences between how polling companies employ their own polling tactics; but unfortunately does not provide any formal insight into the impact each variable has. In order to assess the influence of polling variables to gain insight into what actually influences a poll (on a lower level), dimensionality reduction may be employed in order to display patterns from within the data. In particular, a method known as *Principle Component Analysis* may be employed, in order to both reduce dimensionality and see how each poll is influenced by each variable.

In Figure 6, the 7 numeric principle direction vectors can be visualised; effectively what these portray is their individual influence on the total variance of the data. The direction and length of each vector constitutes how much it contributes to each principle component. For example, we can see that the variables `rawpoll_trump,adjpoll_trump` and `rawpoll_clinton,adjpoll_clinton` contribute in almost opposite directions, but to the same principle component (namely, PC1.)

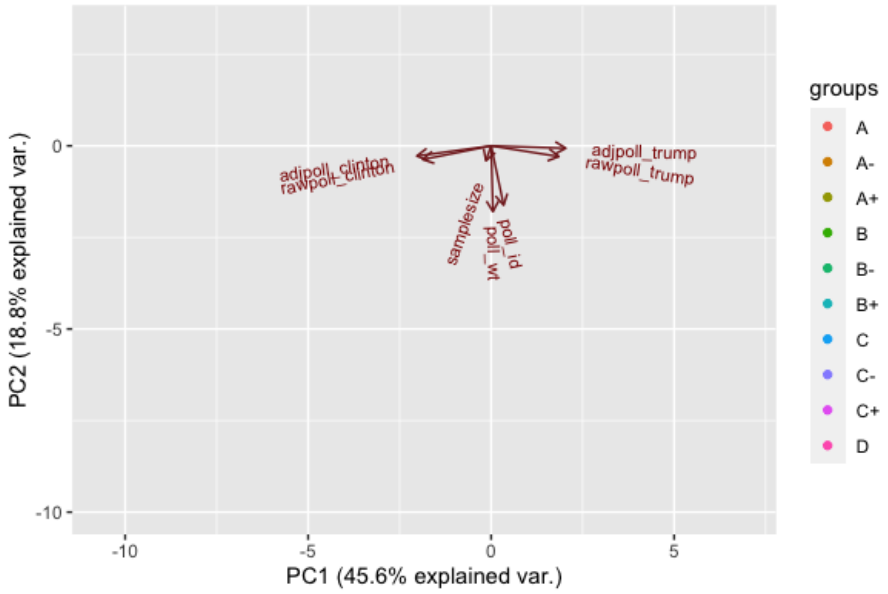

Fig. 6. Principle Component Vectors for Principle Components 1 and 2.

The main goal of principle component analysis is to reduce the dimension of the dataset markedly in order to allow visualisation and representation of variables and minimising the relative information loss. Here, the first two principle components contribute to 64.4% of the total variation from within the data and it is evident that the majority of this variation comes from the polling results of each candidate respectively. Further to this, it can be seen that over 98% of the data is represented by 5 principle components, thus reducing the 98% of the data from 7 dimensions to 5, as can be seen in Figure 7. It is also worth noting that principle component analysis functions on numerical data alone and some variables (such as `url`) are evidently non-numeric and likely will not attribute to a vast variation of the data in this example.

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6     PC7
Standard deviation     1.7868 1.1463 1.0018 0.8560 0.8007 0.3315 0.07711
Proportion of Variance 0.4561 0.1877 0.1434 0.1047 0.0916 0.0157 0.00085
Cumulative Proportion  0.4561 0.6438 0.7872 0.8919 0.9835 0.9991 1.00000
```

Fig. 7. Principle Components and their respective metrics.

More importantly, it is now easy to analyse the respective poll gradings and how they are influenced by each of the variables respectively by plotting the groups as a PCA bi-plot, as can be seen in Figure 8.

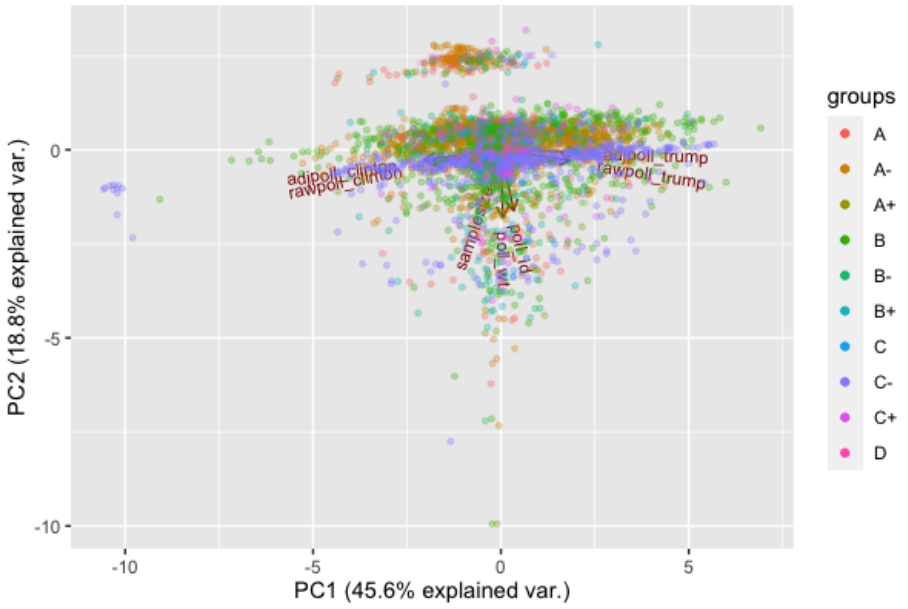

Fig. 8. Principle Component Vectors and Poll Groupings for Principle Components 1 and 2.

By visualising the groupings of poll grades, it's easy to see that the majority of the polls are influenced by the number of votes per candidate. But interestingly, there is a slight disparity above and below the central region whereby most of the polls reside; by looking at the polls in the lower half of the bi-plot there is a distinct lack of lower-graded polls, and these can be seen to reside slightly below the central region. These tend to be more influenced by sample-size alone and as has been displayed in the *chloropleth heatmaps* in Figures 4 and 5, this was likely the cause of some of the disparity. Further to this, the majority of

the high-scoring polls tend to reside centrally, but a small subset of polls reside above this region; in particular, a large subset of polls graded at "A-" are located in the upper fifth of the bi-plot, denoting a heavier influence from PC2. This pertains to these being both sensitive to both votes for each electoral candidate but also the sampling sizes, poll ID and weightings.

5 Concluding Remarks

Polling companies and researchers alike all strive to attain the same end goal of both a reflective and indicative polling survey when it comes to electoral research. A multitude of individuals are surveyed with a breadth of methods, both statistical and qualitative, each with their own benefits and risks; these risks must be adequately acknowledged and tested by polling companies prior to commencing any surveys to ensure integrity of their results. It has been shown that in general, electoral votes alone are not necessarily a good indicator when it comes to polling gradings, outcomes and general efficacy. Polling companies tend to survey a plethora of individuals, with some evidence in certain polls potentially leading to selection bias in terms of location surveyed, in favour of specific candidates. Having said this, this evidence is purely indicative and not formally proven and it is likely that statistical anomalies, data-handling issues, missing values and research bias can all influence the outcome of polls and the patterns that they both display and convey to fellow companies and media outlets. Characteristically, emphasis must be placed on research companies selecting individuals and groups of populations that both accurately reflect opinions and majorities, without introducing any bias, both implicitly or explicitly. The potential negative impacts of opinion swaying via the means of media conglomerates and large corporations is well known in the 21st century and researchers need to be mindful of this to not affect poll outcomes during time-critical events, such as major election campaigns. This report outlines a broad spectrum of methodologies and research pertaining to polling methods, including the selection of individuals to survey, but it is not without it's flaws. Given more time and by increasing the scope of this project, a deeper understanding could have been attained, focusing on micro-contributions to variation in data which would have likely produced more meaningful and detailed outlooks on election polls. Having said this, the strengths behind the project pertain to the results that have been produced and analysed. These provide insight that is not immediately intuitive and gives answers to many questions that are not necessarily commonplace when investigating opinion polls.

Bibliography

Cumming, A. (2001), 'E.s.l/e.f.l instructors' practices for writing assessment: Specific purposes or general purposes?', *Language Testing* **18**(2), 207–224.
URL: *https://doi.org/10.1177/026553220101800206*

Curley, C., Krause, R. M., Feiock, R. & Hawkins, C. V. (2019), 'Dealing with missing data: A comparative exploration of approaches using the integrated city sustainability database', *Urban Affairs Review* **55**(2), 591–615.
URL: *https://doi.org/10.1177/1078087417726394*

Kennedy, C., Hatley, N., Keeter, S., Mercer, A., Igielnik, R. & Traylor, F. (2018), 'Comparing random-digit dial and voter file surveys', *Pew Research Center Methods* .
URL: *https://www.pewresearch.org/methods/2018/10/09/comparing-survey-sampling-strategies-random-digit-dial-vs-voter-files/*

Lavrakas, P. J. (2008), *Encyclopedia of Survey Research Methods*, SAGE Publications.

PRC (2021), 'Collecting survey data', *Pew Research Center Methods* .
URL: *https://www.pewresearch.org/methods/u-s-survey-research/collecting-survey-data/*

Rahman, M. S. (2016), 'The advantages and disadvantages of using qualitative and quantitative approaches and methods in language "testing and assessment" research: A literature review', *Journal of Education and Learning* **6**(1), 102.
URL: *https://doi.org/10.5539/jel.v6n1p102*

Wolter, K., Chowdhury, S. & Kelly, J. (2009), Chapter 7 - design, conduct, and analysis of random-digit dialing surveys, *in* C. Rao, ed., 'Handbook of Statistics', Vol. 29 of *Handbook of Statistics*, Elsevier, pp. 125–154.
URL: *https://www.sciencedirect.com/science/article/pii/S0169716108000072*