Predicting Elections with Google Trends Data

Computational Social Science, WS 22/23, Ludwig-Maximilians-Universität

Sandra Vesterling

2023-03-28

1. Introduction

On a daily basis we search for information on the internet, which opens up the possibility to track our lives and thoughts. With the use of Google Trends one can collect this type of big data and process the information to make analyses about what people are interested in. The first field to make use of this possibility was the pharmaceutical field. If they can predict the outcome of diseases, for example the flu, then pharmaceutical companies can manage the inventory of drugs more efficiently, and even the public health could be strengthen if the government take action of these measurements (Jun et al., 2017).

The potential of predicting with the use of big data and Google Trends has reached other fields as well and one interesting area of subject is if it is possible to predict election outcomes. The purpose of this paper is to examine if predictions with Google Trends could have been adequate estimates of the general election in Sweden 2022.

The last general election in Sweden was held on the 11th of September 2022. It resulted in a change in government and a rather conservative party becoming the second largest party. It was also described by foreign media as "A historic shift" and "Sweden heading towards a new political era" (Asplund Catot, 2022).

The ability to predict election results accurately is crucial for political campaigns, media outlets, and the public to understand and interpret the pulse of the nation. Traditional methods of predicting election results include polling, surveying, and historical data analysis. However, these methods have limitations as they are time-consuming, expensive, and can suffer from sample bias. This has led to a growing interest in exploring alternative methods of predicting election outcomes, including the use of Google Trends data (Jun et al., 2017).

The idea behind using Google Trends for election prediction is that people's search patterns on Google can reveal their political interests, preferences, and sentiment. By analyzing the volume and frequency of certain search terms related to political parties, candidates, and issues, it is possible to infer the public's level of engagement and support towards these entities. Moreover, the real-time nature of Google Trends data allows for quick updates and adjustments, making it a potentially powerful tool for predicting election results (Jun et al., 2017).

However, the use of Google Trends data for election prediction is not without its challenges. One major concern is the issue of causality versus correlation. Just because people are searching for a particular candidate or party does not necessarily mean they will vote for them. Additionally, the demographic bias of internet users and the limited coverage of Google Trends in certain regions could skew the results. Hence, it is important to carefully design and interpret the analysis of Google Trends data to avoid drawing inaccurate conclusions.

In this paper, the potential of using Google Trends data to predict the results of the 2022 general election in Sweden will be explored. The search patterns related to political parties and issues in the time leading up to the election and compare them to the actual election results will be examined. The limitations and implications of using Google Trends for election prediction and provide recommendations for future research in this area will also be discussed.

2. Data

In this chapter the data used for answering the purpose of this paper are presented.

Google Trends Data

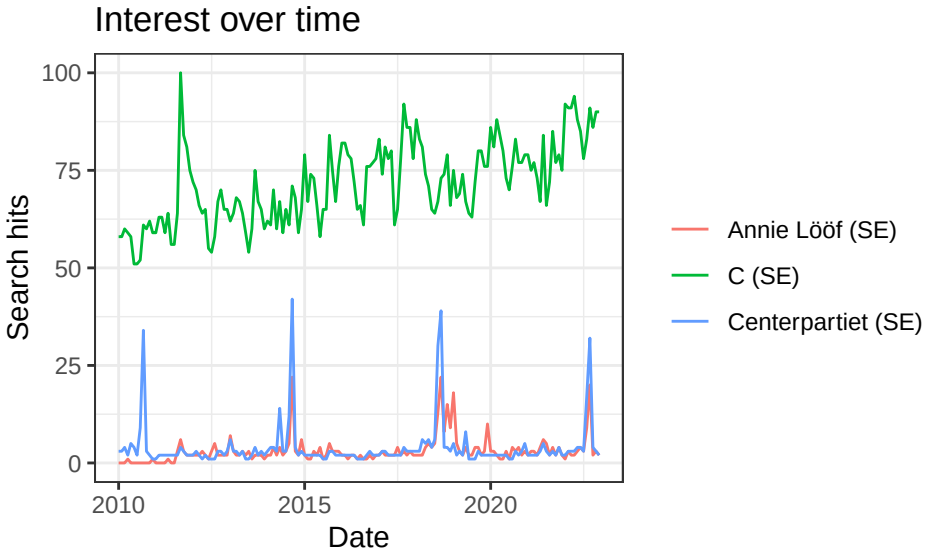
The data used in this paper is primary from Google Trends which provides access to a large sample of actual search requests made to Google. What is important to know about Google Trends is that only a sample of all searches are used in the database, not the entire data set. The results are also normalized to the time and location of a query, measuring the relative search interest of a keyword to the total search volume on Google. The resulting search interest for a term will then be an index value on a scale from 0 to 100 (Google, 2023).

When collecting the data for this paper, there are a few limitations that needs to be done. First is the time period of interest. A natural choice would be 30 days before the election, mainly because of two reasons: this is the day when the electoral register is determined and it is the last day for the parties to announce their participation in the election (Valmyndigheten, 2023). This leaves us with the period between the 12th of August 2022 and the 11th of September 2022.

Second, the keywords needs to be decided. There might be a very large number of words that are related to a party, so some kind of selection needs to be done along with some prior knowledge of the Swedish political system. Party names are obvious to include. Many parties also have a strong association with their abbreviation, but not necessarily all of them. Another way to get information about a party is to search for the respective party leader.

To examine whether a specific word related to a given party should be included in the set of keywords, it is examined if the search interest peaks around general elections for the specific word. The words examined are all party names and their acronym, and their current party leader. The time span for this examination is between January 1 2010 and December 1 2022. This includes the last four general elections held in Sweden, 2010, 2014, 2018 and 2022.

An example is shown below for Centerpartiet where "Centerpartiet" and "Annie Lööf" will be included but not "C".



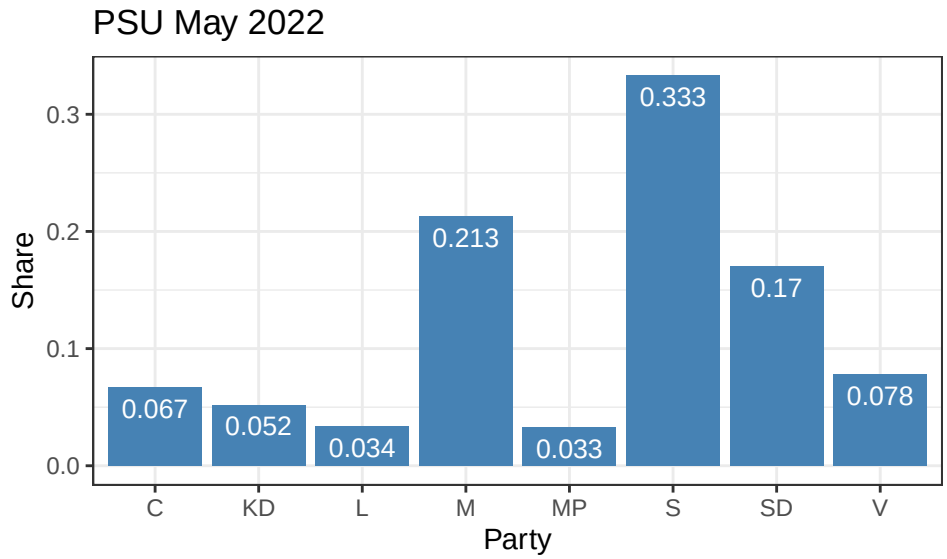
For all party names there are a strong variation in search interest around all these dates, except for "Liberalerna" who changed their name in November 2015. Since then the new name has peaked around the

two following elections, so the new name (Liberalerna) will be included but not the old name (Folkpartiet) (Svanström, 2017). Regarding the acronyms, there are only three that shows a variation in search interest around these dates. The values are SD for Sverigedemokraterna, KD for Kristdemokraterna and MP for Miljöpartiet. More difficult to examine are the names of the party leaders, since there has been a lot of changes in party leaders for many of the parties during the last years. All the current party leaders though have a peak in search interest around the last election (2022) and will therefore be included.

In summary, the following keywords are used: Moderaterna, Ulf Kristersson, Socialdemokraterna, Magdalena Andersson, Sverigedemokraterna, Jimmie Åkesson, SD, Centerpartiet, Annie Lööf, Liberalerna, Johan Pehrson, Kristdemokraterna, Ebba Busch, KD, Miljöpartiet, Märta Stenevi, Per Bolund, MP, Vänsterpartiet and Nooshi Dadgostar. A total of 20 keywords.

Party Preference Survey Data

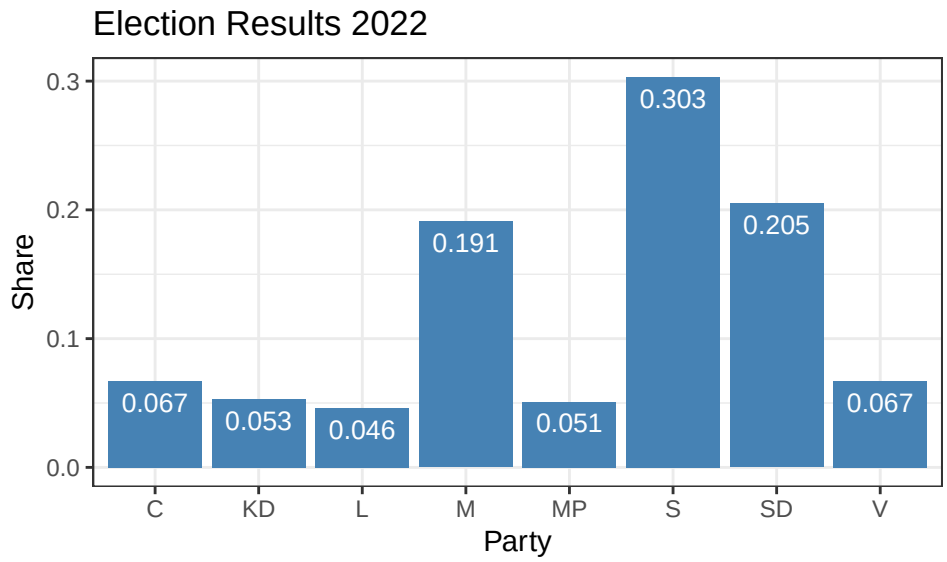
Google trends data is not the only data used in this paper. For comparison and model building the Party Preference Survey (PSU) provided by Statistics Sweden (SCB) will also be used. PSU is a survey published two times per year, in May and November, and presents "election results if an election were to be held today". Through a random selection process, approximately 9,000 individuals are selected to participate in the survey. The sample represents all individuals who are registered in Sweden and would have been eligible to vote in a national election during the current year. Specifically the PSU results from May 2022 will be used, because it is the last one produced before the election. The data is downloaded from their statistical database (SCB, PSU, 2022).



In the plot above the results from PSU in May 2022 is shown. The results are as follws: Socialdemokraterna (33.3%), Moderaterna (21.3%), Sverigedemokraterna (17%), Vänsterpartiet (7.8%), Centerpartiet (6.7%), Kristdemokraterna (5.2%), Liberalerna (3.4%) and Miljöpartiet (3.3%).

Real Election Results

The last data set to be included in this paper is the actual election results from the Swedish general election on September 11th 2022. The data is, as the PSU data, provided by Statistics Sweden, downloaded from their statistical database (SCB, Election, 2022).



The election results are as follows for the eight parliamentary parties: Socialdemokraterna (30.3%), Sverigedemokraterna (20.5%), Moderaterna (19.1%), Vänsterpartiet (6.7%), Centerpartiet (6.7%), Kristdemokraterna (5.3%), Miljöpartiet (5.1%) and Liberalerna (4.6%). Shown in the figure above.

3. Method

The methods used in this paper to evaluate if Google trends data could be used for predicting the Swedish general election 2022 are presented here. They are based on the assumption that voters are using Google to search for information on the party they are going to vote for. This means that the data can be treated as ordinary survey data, and be compared as such.

Google search interest calculations

To further understand the methodology behind using Google Trends data for election prediction, it is important to delve deeper into the formula and equations used. The search interest for each political party is measured by the sum of searches for all keywords related to the party, divided by the total amount of searches for all parties during the defined time period.

The formula used to calculate the sum of keywords for a given party can be further explained by the following equation:

$$\sum_{i=1}^{N} keyword_{i,p,t} = Party_{p,t}$$

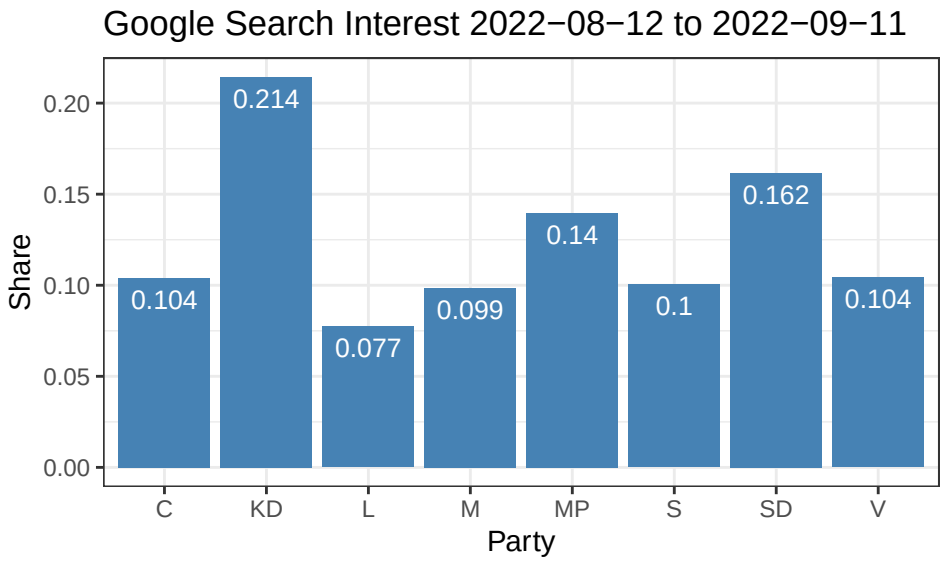
Where i represents the set of keywords related to the party p during the time interval t. For instance, if the party is the Socialdemokraterna, the set of keywords could include "Socialdemokraterna", "Magdalena Andersson" and "S". These keywords are chosen based on their relevance to the party and their frequency of use in search queries related to the party, as explained earlier. The sum of keywords for each party is then represented by $Party_{p,t}$. This represents the total number of searches for all the relevant keywords for party p during time period t.

To calculate the Google search interest for each party, the sum of keywords for that party is divided by the sum of keywords for all parties during the same time period. This is represented by the formula:

$$\frac{Party_{p,t}}{\sum_{p=1}^{M} Party_{p,t}} = Googlesearchinterest_{p,t}$$

Where M represents the total number of parties being analyzed. The resulting value represents the proportion of search interest related to a specific party during the defined time period.

The calculated Google search interest for each party is then shown in the figure below.



The Google search interest are as follows for the eight parliamentary parties: Kristdemokraterna (21.6%), Sverigedemokraterna (15.7%), Miljöpartiet (15.7%), Centerpartiet (10.4%), Socialdemokraterna (10.3%), Vänsterpartiet (10.3%), Moderaterna (10%) and Liberalerna (7.9%).

Modeling

When the Google search interest has been calculated two different linear models are fitted. Model 1 is a simple linear regression model that includes two covariates: the Google search interest and the PSU (previous election) results. The equation for Model 1 can be represented as:

Model 1:

$$Y = \beta_0 + \beta_1 * x_{Google} + \beta_2 * x_{PSU}$$

In this equation, Y represents the outcome of the election, such as the percentage of votes received by a particular party. x_{Google} represents the Google search interest for that party, while x_{PSU} represents the results from the PSU survey in May 2022. The β coefficients are the regression coefficients that represent the relationship between the predictor variables and the outcome variable.

Model 2 is a more complex linear regression model that includes an additional covariate: the interaction between the Google search interest and the PSU results. This model attempts to capture the relationship between changes in search interest over time and the effect it has on the election results. The equation for Model 2 can be represented as:

Model 2:

$$Y = \beta_0 + \beta_1 * x_{Google} + \beta_2 * x_{PSU} + \beta_3 * x_{Google*PSU}$$

In this equation, $x_{Google*PSU}$ represents the interaction term between the Google search interest and the PSU results. This interaction term captures how changes in search interest over time can have a greater impact on the election results, depending on the peoples party preferences a few months earlier.

Evaluation

To evaluate the two models different methods are used. In a linear regression analysis, the goal is to find the best-fitting model that explains the relationship between the response variable and the predictor variables.

One way to compare the fit of two models is to use an F-test, which compares the improvement in the fit of the model with additional variables to the expected improvement that would be obtained by chance.

The F-test calculates the ratio of the variance explained by the model to the unexplained variance. The F-statistic is calculated as the ratio of the regression mean square (MSR) to the residual mean square (MSE). The F-statistic follows an F-distribution, and the p-value associated with the F-statistic is used to determine the statistical significance of the model. The F-statistic is calculated as:

$$F_{test} = \frac{MSR}{MSE} = \frac{\sum_i \left(y_i - \bar{y}\right)^2 / (k-1)}{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 / (n-k)}$$

Where n is the number of observations and k is the number of predicting variables. Therefore, the F-test is used to determine whether the difference in the fit of the models is statistically significant, and if so, which model is a better fit to the data (Kleinbaum et al., 2014).

Another evaluation method that can be used to assess the fit of the models is the adjusted R-squared. The R-squared value represents the proportion of the variation in the response variable that is explained by the predictor variables in the model. However, as more predictor variables are added to the model, the R-squared value increases, even if the added variables do not contribute significantly to the model. The adjusted R-squared value takes into account the number of predictor variables in the model and adjusts the R-squared value accordingly. A higher adjusted R-squared value indicates a better fit of the model to the data. The mathematical formula is as follows (Kleinbaum et al., 2014):

$$R^2 = \frac{\sum_i \left(\hat{y}_i - \bar{y}\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2}$$

4. Results

This chapter presents the predictions for the 2022 Swedish general election from the two models described in chapter 4. The evaluation and comparison of the models will also be presented here.

Predictions

Two linear models were fitted to the data. Model 1 with two covariates, the google search interest and the PSU results. Model 2 with three covariates, the google search interest, the PSU results and the interaction between these.

Party	Election	Model_1	Model_2
C	0.067	0.070	0.070
KD	0.053	0.072	0.058
L	0.046	0.037	0.043
M	0.191	0.201	0.194
MP	0.051	0.044	0.036
S	0.303	0.309	0.300
SD	0.205	0.171	0.204
V	0.067	0.080	0.080

In the output above the predictions for the election results for all parties from both models are shown together with the real election results. Comparing the predicted values of each model to the actual values in the "Election" column can provide an initial indication of how well the models fit the data. In six of eight cases model 2 performs better than model 1.

Comparison and evaluation

To evaluate the performance of the models more comprehensively, other metrics than a visual inspection are presented in the following.

Model 1

```
##
## Call:
## lm(formula = Election ~ Google + PSU, data = merged_data)
##
## Residuals:
##         1         2         3         4         5         6         7         8
## -0.002945 -0.018503  0.009440 -0.009866  0.006825 -0.006329  0.034292 -0.012913
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.004676   0.025819  -0.181    0.863
## Google       0.136739   0.171788   0.796    0.462
## PSU          0.901704   0.071470  12.617 5.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01983 on 5 degrees of freedom
## Multiple R-squared:  0.9702, Adjusted R-squared:  0.9583
```

```
## F-statistic: 81.43 on 2 and 5 DF,  p-value: 0.0001531
```

In the model summary for model 1 in the output above it is shown that only the variable PSU is significant in the model, not the Google variable. The adjusted R-squared value of the model is 0.9572, which suggests that the model explains 95.72% of the variation in the response variable after adjusting for the number of predictors used. Furthermore, the F-statistic indicates that the overall model is statistically significant (p-value: 0.0001631), which means that at least one of the predictor variables has a significant effect on the Election outcome.

Model 2

```
##
## Call:
## lm(formula = Election ~ Google * PSU, data = merged_data)
##
## Residuals:
##         1         2         3         4         5         6         7         8
## -0.002730 -0.004928  0.002943 -0.002785  0.015084  0.003497  0.001460 -0.012541
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.04209    0.01872   2.248   0.0878 .
## Google      -0.30545    0.15093  -2.024   0.1130
## PSU          0.24890    0.18115   1.374   0.2414
## Google:PSU   6.13392    1.66405   3.686   0.0211 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01057 on 4 degrees of freedom
## Multiple R-squared:  0.9932, Adjusted R-squared:  0.9881
## F-statistic: 195.5 on 3 and 4 DF,  p-value: 8.585e-05
```

In the model summary for model 2 in the output above it is shown that only the interaction variable between PSU and Google search interest is significant in the model. The adjusted R-squared value of 0.9886 indicates that the model explains 98.86% of the variation in the response variable, which is a very high level of fit. The F-statistic of 203.4 with a very low p-value of 7.935e-05 suggests that the model as a whole is significant and the predictor variables included in the model are useful in explaining the election outcome. Additionally, the intercept and the Google and PSU variables are not significant, with p-values of 0.0663 and 0.0853, respectively, indicating that they do not have a significant independent effect on the election outcome. Overall, these results suggest that the interaction between the Google search interest and PSU results is the most important predictor variable for explaining the variation in the election outcome.

F-Test

```
## Analysis of Variance Table
##
## Model 1: Election ~ Google + PSU
## Model 2: Election ~ Google * PSU
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1      5 0.00196682
## 2      4 0.00044732  1 0.0015195 13.588 0.02109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This output shows the results of an ANOVA analysis comparing two linear regression models (Model 1 and Model 2) that explain the variation in the "Election" variable.

The results indicate that both models have significant F-tests (p-value < 0.05), indicating that they have a statistically significant effect on the variation in the election results. Furthermore, the results show that Model 2 is a better fit to the data than Model 1, as it has a lower residual sum of squares (RSS) and a higher F-value. This is evidenced by the significant difference in the sum of squares (Sum of Sq) between Model 1 and Model 2, and the p-value indicating that this difference is statistically significant ($p = 0.01842$).

Finally, the interaction term in Model 2 can be interpreted as indicating that there is an interaction effect between "Google" and "PSU" on the variation in "Election."

5. Discussion

This paper aims to examine whether Google Trends data could have been used for predicting the Swedish general election 2022. To do this two models were fitted, based on the assumption that voters of one party are searching for information about the party on Google, a strong yet crucial assumption.

The results from these models indicates that the second model, the model that includes the interaction term between the Google search interest and the results from the PSU in May 2022, performs better than the first model without interaction term. The second model has higher F-value and adjusted R-squared than model 1 and the interaction term is significant in the model.

Even though the model performs good, the variable with the Google search interest contributes little to the model and there are issues with both the Google Trends data and the methods used that needs to be adressed.

There is a clear selection problem with Google Trends Data. Google provides no background information about the persons behind the Google searches. There might be various reasons why or why not someone are using Google to gain information. For example, younger voters may be more likely to use Google to search for information, while older voters may rely on traditional media sources. This could create biases in the data, particularly if younger voters are more likely to support a certain party. As compared to the traditional PSU were the 9000 subjects are randomly selected, the Google data can not be assumed to be a random sample.

This raises questions about the representativeness of the Google search data and its ability to accurately capture the opinions and behaviors of the broader population. Additionally, there may be issues with the assumption that voters of one party are searching for information about that specific party on Google. It is possible that individuals are searching for information about a party for reasons other than support, such as curiosity or opposition.

Furthermore, while the second model shows promise, it is important to note that the variable with Google search interest contributes little to the model. This suggests that other factors may be more important in predicting election outcomes, and that the usefulness of Google Trends data may be limited in this context. As was shown, the PSU variable has a significant impact on the election result predictions which is not unexpected due to its large sample size and adequate methods. Even though it was performed a few months before the election, and not in direct connection as the Google search interest, it contributes better.

The choice of keywords to represent each party was made on background knowledge of the Swedish political system together with an examination of interest during earlier elections. This does not necessarily mean that all important keywords were selected correctly, there might be more words that should be included but were not captured by this selection method. The keywords play an important role in this context and developing a more decent method for this search would perhaps be beneficial for examining the role of Google Trends data in predicting elections.

The modeling of the data in this paper is rather simple, regarding both the methods used and the included variables. Both these aspects could be extended for a wider investigation. An obvious factor to take into consideration would be time. The time aspect in this paper was limited to a reasonable period due to important happenings on these dates but there is possibly a time-dependent effect that could be modeled more suited for the purpose.

Overall, the results of this study suggest that while Google Trends data may have potential as a predictor of election outcomes, there are significant limitations and potential biases that need to be considered. Further research is needed to fully understand the strengths and weaknesses of this approach and to determine the most effective ways to incorporate this data into election forecasting models.

References

Asplund Catot, C. (2022, September 12). Omvärldens ögon på det svenska rysarvalet. Europaportalen. Retrieved March 29, 2023, from https://www.europaportalen.se/2022/09/omvarldens-ogon-pa-det-svenska-rysarvalet

Google. (2023). FAQ about Google Trends Data - trends help. Google. Retrieved March 29, 2023, from https://support.google.com/trends/answer/4365533?hl=en

Jun,S-P. Yoo,H,S. Choi,S. (2017). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. Elsevier Inc.

Kalender för val till Riksdag, region- och kommunfullmäktige. Riksdag, region och kommun | Valmyndigheten. (2023, February 10). Retrieved March 29, 2023, from https://www.val.se/val-och-folkomrostningar/valkalendrar/valaret-2022.html

Kleinbaum,D.G., Kupper,L.L., Nizam,A. & Rosenberg,E.S. (2014). Applied regression analysis and other multivariable methods. Cengage Learning.

Official Statistics of Sweden, Election to the Riksdag - results by region and party etc. Number and percent. Year of election 1973 - 2022 (2022). Retrieved March 14, 2023, from https://www.statistikdatabasen.scb.se/pxweb/en/ssd/START___ME___ME0104___ME0104C/ME0104T3/.

Official Statistics of Sweden , Valresultat om det varit val idag (PSU) efter parti, tabellinnehåll och mätmånad (2022). Retrieved March 14, 2023, from https://www.statistikdatabasen.scb.se/pxweb/en/ssd/START__ME___ME0201___ME0201A/Vid10/.

Svanström, S. (2017, September 5). Folkpartiet Liberalerna byter namn. Liberalerna. Retrieved March 29, 2023, from https://www.liberalerna.se/nyheter/folkpartiet-liberalerna-byter-namn

Uddhammar, L. & Järnbert, M. (2022). Party preference survey (PSU). Statistiska Centralbyrån. Retrieved March 29, 2023, from https://www.scb.se/en/finding-statistics/statistics-by-subject-area/democracy/political-party-preferences/party-preference-survey-psu/