KOALA: Estimating coalition probabilities in multi-party electoral systems

Alexander Bauer[1], Andreas Bender[1], André Klima[1], Helmut Küchenhoff[1]

[1] Department of Statistics, Ludwig-Maximilians-Universität, Munich, Germany

E-mail for correspondence: Alexander.Bauer@stat.uni-muenchen.de

Abstract: We present a survey-based approach for the estimation of probabilities for specific coalition majorities in multi-party electoral systems. A Bayesian Multinomial-Dirichlet model with Monte Carlo simulation is used for estimation. The method is based on opinion polls conducted by established pollster institutes and is accompanied by a pooling approach to summarize multiple current surveys, accounting for dependencies between institutes. The method estimates sample uncertainty-based probabilities, if the election was held today. An implementation of the method in R is freely available.

Keywords: Election analysis; Opinion polls; Election reporting; Multinomial-Dirichlet; Pooling.

1 Introduction and data

Election polls as surveys conducted by different pollster institutes try to represent the public opinion based on a finite sample. Current reporting on such surveys is most often broken down to the observed shares, proper reporting of sample uncertainty is rarely done (cf. QUELLE). In our opinion, the focus in survey reporting in multi-party electoral systems should be shifted towards the most relevant question, i.e. how probable majorities for specific coalitions are. We present our KOALA (Coalition Analysis) approach to estimate such probabilities to bring more value to opinion poll-based reporting.

As database we use opinion polls conducted by established pollster institutes, quantifying the electoral behaviour *if an election was held today*. Our approach is to be differentiated from prediction-aimed methods (cf. POLLYVOTE or NORPOTH & GSCHWEND, 2010). We focus on the question

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018.

of quantifying *current* majority situations, not taking into consideration potential shifts until election day.

A Bayesian Multinomial-Dirichlet model with Monte Carlo simulations is used for estimation. Also, a pooling approach is presented to summarize the results of multiple current opinion polls to reduce sample uncertainty. All methods were implemented in `R` and are available in the open-source package `coalitions` on GitHub (QUELLE). Also, the website `koala.stat.uni-muenchen.de` visualizes estimated coalition probabilities and is used to communicate the results to the general public (QUELLE).

2 Pooling approach

In the presence of multiple published opinion polls, pooling is used to summarize the observed results in order to reduce sample uncertainty. To assure a reliable pooling regarding the current public opinion, we only use polls published within the past 14 days and only use the most recent survey published by each pollster institute.

Based on the multinomial distribution of the voter number X_{ij} of party j in poll i with underlying true party share θ_j, pooling over multiple polls representing independent random samples would lead to a multinomial distribution for the summed number of votes $\sum_i X_{ij}$:

$$\sum_i X_{i1}, \ldots, \sum_i X_{ik} \sim Multinomial(\sum_i n_i, \theta_1, \ldots, \theta_k).$$

Further investigations however show that polls published by the main German pollster institutes show a certain amount of correlatedness and the independency assumption does not hold. To account for this, we adjust the resulting multinomial distribution by using an *effective sample size* (cf. QUELLE). Quantification of pairwise correlation is done based on the variance of the difference between two polls. The following equation holds for two independent random sample polls A and B:

$$\Leftrightarrow Cov(X_{Aj}, X_{Bj}) = \frac{1}{2} \cdot (Var(X_{Aj}) + Var(X_{Bj}) - Var(X_{Aj} - X_{Bj}))\,.$$

We take $Var(X_{Aj})$ and $Var(X_{Bj})$ as the theoretical variances of the binomially distributed, observed voter numbers and estimate $Var(X_{Aj} - X_{Bj})$ based on the observed differences between the party shares. Having done so, one can estimate the covariance $Cov(X_{Aj}, X_{Bj})$ and accordingly also the correlation. As the binomial distribution is directly proportional to the sample size, the effective sample size n_{eff} can be defined as the ratio between the estimated variance for the pooled sample and the theoretical variance of a sample of size one:

$$n_{\text{eff}} = \frac{Var(\text{pooled})}{Var(\text{sample of size 1})},$$

with, in the case of two surveys,

$$Var(\text{pooled}) = Var(X_A + X_B) = Var(X_A) + Var(X_B) + 2Cov(X_A, X_B)$$

and Var(sample of size 1) the theoretical variance of the pooled share. Looking at the party-specific correlations between 20 surveys conducted by the two most regular German pollster institutes, Emnid and Forsa, we on average end up with a medium high correlation, using mean party shares and sample sizes per institute for the theoretical variances. Other institute comparisons were not performed as too few surveys were conducted over comparable timeframes. For simplicity, we set the correlation used in our methodology to 0.5. For calculating n_{eff} we base the calculation on the result of the biggest party, as the specific party choice only marginally affects n_{eff}.

3 Calculation of probabilities

For estimating probabilities for specific coalitions, we choose a Multinomial-Dirichlet model with an uninformative prior for the true party shares θ_j (QUELLE):

$$\theta = (\theta_1, \ldots, \theta_k)^{\mathrm{T}} \sim \mathit{Dirichlet}(\alpha_1, \ldots, \alpha_k),$$

$$\text{with} \qquad \alpha_1 = \ldots = \alpha_k = \frac{1}{2}$$

Basing the Bayesian model on one (pooled) survey the posterior also results in a Dirichlet distribution with parameters $\alpha_j = x_j + \frac{1}{2}$ for each party j. Probabilities for specific events can now be estimated using Monte Carlo simulations of random election outcomes. After calculating to which effective number of seats in parliament the simulated results lead, one can easily calculate probabilities by taking the share of times some specified event occurs. To adjust for observed survey shares generally being only published as rounded numbers, before applying the Bayesian model we add uniformly distributed random numbers $r_\gamma \sim U[-\gamma, \gamma]$ to the shares (e.g. $\gamma = 0.5\%$) and rescale the sum to 100%.

For visualizing the probability together with underlying uncertainty for a specific coalition majority we recommend using a density plot as shown in Figure 1.

4 Conclusion

We presented an approach to estimate probabilities for specific election outcomes based on publicly available opinion polls. Pooling allows for the inclusion of information from multiple surveys. Visualizing the results on a publicly available website for chosen elections, our long-term goal is to make proper uncertainty assessment standard in general opinion poll-based reporting.

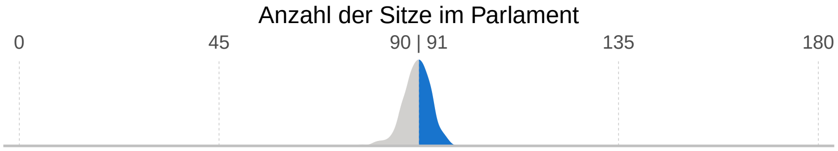

FIGURE 1. Density of $1,000$ simulated shares of parliament seats for the coalition CSU/FDP in the Bavarian state election. The blue area marks simulations with a majority.

Acknowledgments: Special Thanks to ...

References

Diggle, P.J., Liang, K-Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman & Hall.

Henderson, C.R. (1973). Sire evaluation and genetic trends. In: *Proceedings of the Animal Breeding and Genetics Symposium in Honour of Dr. L. Lush*, Champaign, Illinois, 10 – 41.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B*, **58**, 619 – 678.

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects (with Discussion). *Statistical Science*, **6**, 15 – 51.