KOALA: A new paradigm for election coverage

An opinion poll based “now-cast” of probabilities of events in multi-party electoral systems

Alexander Bauer

Statistical Consulting Unit StaBLab, LMU Munich

DAGStat | March 20, 2019 | Munich

KOALA

KoalitionsAnalyse

Alexander Bauer

Statistical Consulting Unit StaBLab, LMU Munich

DAGStat | March 20, 2019 | Munich

KOALA

KoalitionsAnalyse

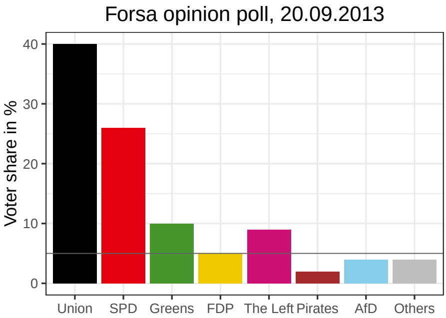
Collaborators

Dr. Andreas Bender	Nuffield Department of Clinical Medicine, University of Oxford, United Kingdom
Dr. André Klima	StaBLab, LMU Munich
Prof. Dr. Helmut Küchenhoff	StaBLab, LMU Munich

Outline

Outline

❶ Motivation



Questions of interest

- Which parties will pass the 5% hurdle and enter the parliament?
- Which parties will form the governing coalition?

❶ Motivation

Reported voter shares (Forsa, last pre-election poll 2013)

Union	SPD	Greens	FDP	The Left	Pirates	AfD	Others
40%	26%	10%	5%	9%	2%	4%	5%

Redistributed voter shares (based on 5% hurdle)

Union	SPD	Greens	FDP	The Left	Pirates	AfD	Others
44.44%	28.89%	11.11%	5.56%	10.00%	-	-	-

Media reports...

- sometimes state "Union-FDP miss joint majority with a seat share of 50%"
  ⇒ Problem 1: This completely neglects sample uncertainty
- usually report sample uncertainty à la "a 2.5% margin of error"
  ⇒ Problem 2: Uncertainty is hard to grasp

❶ Motivation

Reported voter shares (Forsa, last pre-election poll 2013)

Union	SPD	Greens	FDP	The Left	Pirates	AfD	Others
40%	26%	10%	5%	9%	2%	4%	5%

Redistributed voter shares (based on 5% hurdle)

Union	SPD	Greens	FDP	The Left	Pirates	AfD	Others
44.44%	28.89%	11.11%	5.56%	10.00%	-	-	-

Media reports...

- sometimes state "Union-FDP miss joint majority with a seat share of 50%"
  ⇒ Problem 1: This completely neglects sample uncertainty
- usually report sample uncertainty à la "a 2.5% margin of error"
  ⇒ Problem 2: Uncertainty is hard to grasp

❶ Motivation

Reported voter shares (Forsa, last pre-election poll 2013)

Union	SPD	Greens	FDP	The Left	Pirates	AfD	Others
40%	26%	10%	5%	9%	2%	4%	5%

Redistributed voter shares (based on 5% hurdle)

Union	SPD	Greens	FDP	The Left	Pirates	AfD	Others
44.44%	28.89%	11.11%	5.56%	10.00%	-	-	-

Media reports...

- sometimes state "Union-FDP miss joint majority with a seat share of 50%"

  ⇒ Problem 1: This completely neglects sample uncertainty
- usually report sample uncertainty à la "a 2.5% margin of error"

  ⇒ Problem 2: Uncertainty is hard to grasp

❶ Motivation

Reported voter shares (Forsa, last pre-election poll 2013)

Union	SPD	Greens	FDP	The Left	Pirates	AfD	Others
40%	26%	10%	5%	9%	2%	4%	5%

Redistributed voter shares (based on 5% hurdle)

Union	SPD	Greens	FDP	The Left	Pirates	AfD	Others
44.44%	28.89%	11.11%	5.56%	10.00%	-	-	-

Media reports...

- sometimes state "Union-FDP miss joint majority with a seat share of 50%"
  ⇒ Problem 1: This completely neglects sample uncertainty
- usually report sample uncertainty à la "a 2.5% margin of error"
  ⇒ Problem 2: Uncertainty is hard to grasp

❶ Motivation

We aim to do now-casting

We communicate sample uncertainty in a more natural way by calculating **event probabilities** that fully reflect sample uncertainty.

We do not aim to do for-casting

- Our approach simply communicates sample uncertainty in a novel way
- Also, a relevant share of voters is still undecided shortly before election day (Küchenhoff et al., 2018)

❶ Motivation

We aim to do now-casting

We communicate sample uncertainty in a more natural way by calculating **event probabilities** that fully reflect sample uncertainty.

We do not aim to do for-casting

- Our approach simply communicates sample uncertainty in a novel way
- Also, a relevant share of voters is still undecided shortly before election day (Küchenhoff et al., 2018)

Outline

❷ Methods

Estimating probabilities of events (POEs)

Given one opinion poll with sample size n:

$$\boldsymbol{X} = (X_1, \ldots, X_P)^T \sim Multinomial(n, \theta_1, \ldots, \theta_P),$$

with voter counts X_j and the true percentage of voters θ_j per party j.

A **Dirichlet posterior distribution** results for $\theta|\boldsymbol{x}$:

$$\theta|\mathbf{x} \sim Dirichlet(x_1 + 1/2, \ldots, x_P + 1/2),$$

based on an **uninformative Dirichlet prior** (Gelman et al., 2013)

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)^T \sim Dirichlet(\alpha_1, \ldots, \alpha_P),$$

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

❷ Methods

Estimating probabilities of events (POEs)

Given one opinion poll with sample size n:

$$\boldsymbol{X} = (X_1, \ldots, X_P)^T \sim Multinomial(n, \theta_1, \ldots, \theta_P),$$

with voter counts X_j and the true percentage of voters θ_j per party j.

A **Dirichlet posterior distribution** results for $\boldsymbol{\theta}|\boldsymbol{x}$:

$$\boldsymbol{\theta}|\mathbf{x} \sim Dirichlet(x_1 + 1/2, \ldots, x_P + 1/2),$$

based on an **uninformative Dirichlet prior** (Gelman et al., 2013)

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)^T \sim Dirichlet(\alpha_1, \ldots, \alpha_P),$$

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

❷ Methods

Estimating probabilities of events (POEs)

Given one opinion poll with sample size n:

$$\boldsymbol{X} = (X_1, \ldots, X_P)^T \sim Multinomial(n, \theta_1, \ldots, \theta_P),$$

with voter counts X_j and the true percentage of voters θ_j per party j.

A **Dirichlet posterior distribution** results for $\boldsymbol{\theta}|\boldsymbol{x}$:

$$\boldsymbol{\theta}|\mathbf{x} \sim Dirichlet(x_1 + 1/2, \ldots, x_P + 1/2),$$

based on an **uninformative Dirichlet prior** (Gelman et al., 2013)

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)^T \sim Dirichlet(\alpha_1, \ldots, \alpha_P),$$

with $$\alpha_1 = \ldots = \alpha_P = \frac{1}{2}.$$

❷ Methods

Estimating probabilities of events (POEs)

Given the **posterior distribution of voter shares** we can use
Monte Carlo simulations to estimate POEs:

1. Simulate 10 000 election outcomes from the posterior
   (adding uniformly distributed random noise to account for rounding errors)
2. If necessary: Redistribute voter shares to get obtained seats in parliament
3. $\text{POE} = \frac{\#\text{events}}{\text{number of simulations}}$

Example

Given the Forsa poll, the coalition of Union-FDP obtained a majority of seats in 2 633 of 10 000 simulations

$\Rightarrow$ POE $\approx 26\%$

❷ Methods

Estimating probabilities of events (POEs)

Given the **posterior distribution of voter shares** we can use **Monte Carlo simulations** to estimate POEs:

1. Simulate 10 000 election outcomes from the posterior
   (adding uniformly distributed random noise to account for rounding errors)
2. If necessary: Redistribute voter shares to get obtained seats in parliament
3. $\text{POE} = \frac{\#\text{events}}{\text{number of simulations}}$

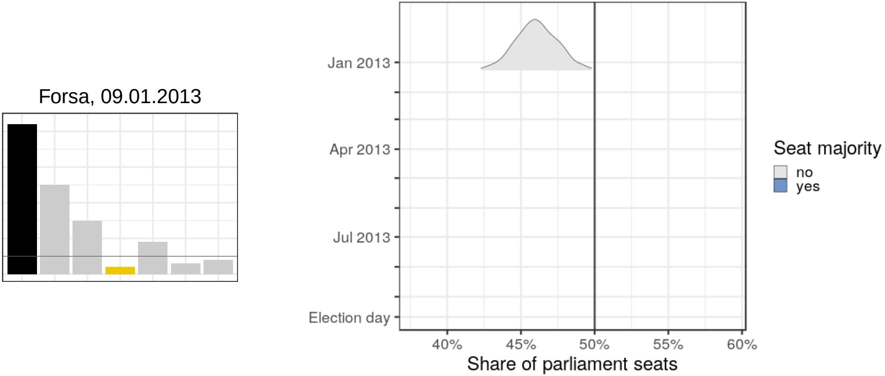
Example

Given the Forsa poll, the coalition of Union-FDP obtained a majority of seats in 2 633 of 10 000 simulations

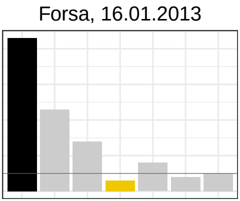
$\Rightarrow \text{POE} \approx 26\%$

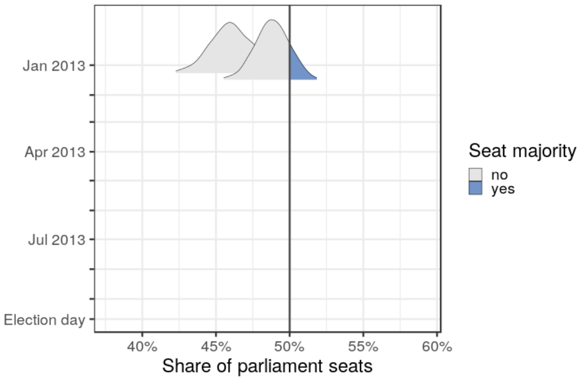
❷ Methods

Visualization using ridgeline plots (Wilke, 2017)



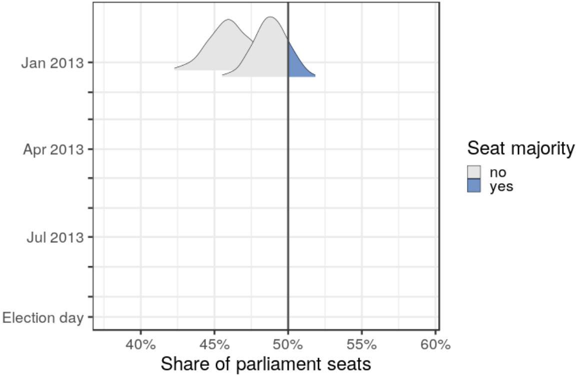
Visualization using ridgeline plots (Wilke, 2017)





❷ Methods

Visualization using ridgeline plots (Wilke, 2017)

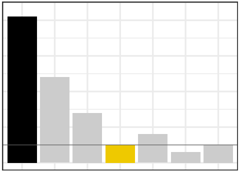


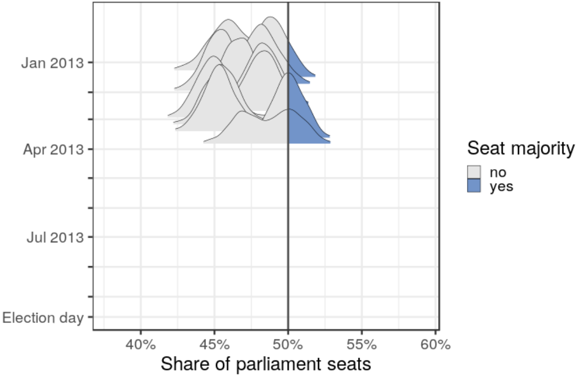
I'm a .gif, click me (in Adobe Acrobat)!

❷ Methods

Visualization using ridgeline plots (Wilke, 2017)

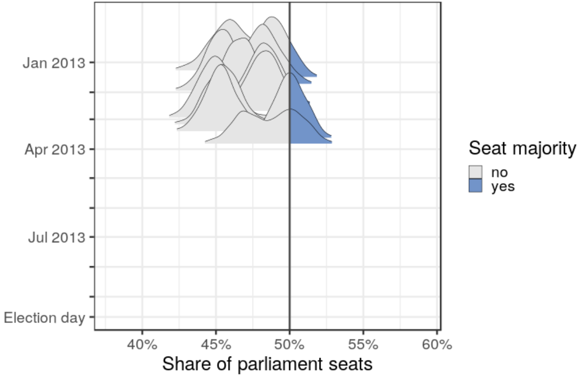
Forsa, 26.03.2013





❷ Methods

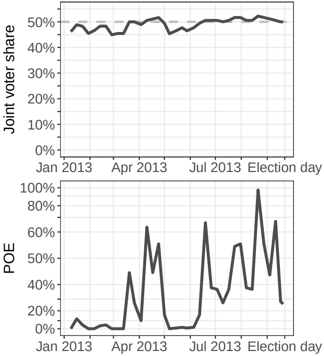
Visualization using ridgeline plots (Wilke, 2017)

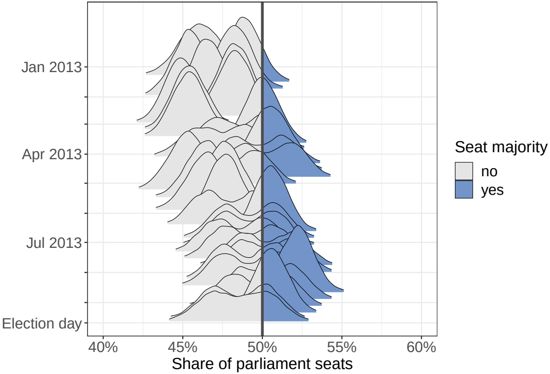


I'm a .gif, click me (in Adobe Acrobat)!

❷ Methods

Visualization using ridgeline plots (Wilke, 2017)





❷ Methods

Pooling

We aggregate multiple polls to reduce sample uncertainty.
In case of multiple random samples:

$$\left(\sum_i X_{i1}, \ldots, \sum_i X_{iP}\right)^T \sim Multinomial\left(\sum_i n_i, \theta_1, \ldots, \theta_P\right).$$

We account for correlations between polling agencies by using an **effective sample size** (Hanley et al., 2003).

Example

Pooling two polls with 1 500 and 2 000 respondents we get an effective sample size of $n_{\mathbf{eff}} = 2\,341$ (based on a strongest party share of 40%).

❷ Methods

Pooling

We aggregate multiple polls to reduce sample uncertainty.
In case of multiple random samples:

$$\left(\sum_i X_{i1}, \ldots, \sum_i X_{iP}\right)^T \sim \mathit{Multinomial}\left(\sum_i n_i, \theta_1, \ldots, \theta_P\right).$$

We account for correlations between polling agencies by using an **effective sample size** (Hanley et al., 2003).

Example

Pooling two polls with $1\,500$ and $2\,000$ respondents we get an effective sample size of $n_{\mathbf{eff}} = 2\,341$ (based on a strongest party share of 40%).

❷ Methods

Pooling

We aggregate multiple polls to reduce sample uncertainty.
In case of multiple random samples:

$$\left(\sum_i X_{i1}, \ldots, \sum_i X_{iP}\right)^T \sim Multinomial\left(\sum_i n_i, \theta_1, \ldots, \theta_P\right).$$

We account for correlations between polling agencies by using an **effective sample size** (Hanley et al., 2003).

Pooling in practice

- We only pool surveys published in the last 14 days
- We only include one survey per polling agency

Outline

❸ Technical implementation

R package `coalitions`

Functionality

- Scrape wahlrecht.de for (new) polls
- Calculate pooled sample
- Sample from posterior distribution
- Redistribute votes below 5% threshold
  and calculate parliament seats (e.g. based on method Sainte-Laguë-Schepers)
- Calculate coalition probabilities

More on github.com/adibender/coalitions

❸ Technical implementation

Web-Interface

Communicating the results

1. Website koala.stat.uni-muenchen.de

   ⇒ Automatic updates scraping data from wahlrecht.de
2. Twitter @KOALA_LMU

   ⇒ Automatic tweets of new results
3. Blog koala-blog.netlify.com

Technical implementation in R

- User interface was built with the shiny package
- Server is based on Shiny Server Open Source
- Tweets are sent with the twitteR package

❸ Technical implementation

Web-Interface

Communicating the results

1. Website koala.stat.uni-muenchen.de
   ⇒ Automatic updates scraping data from wahlrecht.de
2. Twitter @KOALA_LMU
   ⇒ Automatic tweets of new results
3. Blog koala-blog.netlify.com

Technical implementation in R

- User interface was built with the shiny package
- Server is based on Shiny Server Open Source
- Tweets are sent with the twitteR package

Outline

❹ Conclusion

The KOALA approach

- New paradigm for opinion poll coverage
- Bayesian approach to now-cast POEs
- Sample uncertainty is reduced by pooling multiple polls
- Communication to the general public

- **Keep in mind:** We calculate now-casts, not for-casts

References

KOALA

Bauer, A., Bender, A., Klima, A., and Küchenhoff, H. (2018) KOALA: A new paradigm for election coverage. *arXiv.org*. URL http://arxiv.org/abs/1807.09665.

Bender, A., and Bauer, A. (2018) coalitions: Coalition probabilities in multi-party democracies. *The Journal of Open Source Software*. doi: 10.21105/joss.00606.

URL http://joss.theoj.org/papers/10.21105/joss.00606.

Methods

Gelman, A. et al. (2013) *Bayesian data analysis*, volume 3. CRC press Boca Raton, FL.

Hanley, J. A. et al. (2003) Statistical analysis of correlated data using generalized estimating equations: an orientation. *American journal of epidemiology*, 157(4), 364–375.

Küchenhoff, H. et al. (2018) *Universitätsstudie zur Bayernwahl USBW 18 (München – Passau – Regensburg). Erste Ergebnisse – Oktober 2018.*

URL https://www.stablab.stat.uni-muenchen.de/lehre/pdfs/usbw18.pdf.

Further software

Chang, W. et al. (2017) *shiny: Web Application Framework for R*.
URL https://CRAN.R-project.org/package=shiny. R package version 1.0.5.

Wilke, C. O. (2017) *ggridges: Ridgeline Plots in 'ggplot2'*.
URL https://CRAN.R-project.org/package=ggridges. R package version 0.4.1.