Coverage Bias in Election Research using Data from Social Media

Authors:
Hannah Bucher (hannah.bucher@gesis.org)
Marie Kühn (marie.kuehn@gesis.org)
Joss Roßmann (joss.roßmann@gesis.org)
Social Media Data in Election Research

- **predicting election outcomes** by looking at the frequency of party mentions on twitter or using more sophisticated methods (e.g. Tumasjan, Sprenger et al. 2010 or Oliveira, Bermejo et al. 2017 for Spain and the USA, for an overview, see Gayo-Avello 2013 or Phillips, Dowling et al. 2017)

- **predicting public opinion or political orientation** analyzing twitter content (for an overview, see Phillips, Dowling et al. 2017)
Previous Research

- some mainly descriptive research investigating the composition of demographics, opinions and personality traits on different social media platforms (e.g. Greenwood 2016, Barberá and Rivero 2015, Ruths and Pfeffer 2014 or Ryan and Xenos 2011) with a focus on twitter (for an overview see Jungherr 2016) using social media data

- some collections of possible data quality issues and concerns for social media data (e.g. Ruths and Pfeffer 2014 or Gayo-Avello 2013)

- attempts to characterize users of different social media platforms using survey data:
  - Rainie, Smith et al. 2012 looking at political participation on social media in the USA amongst mostly social media users only
  - Blank and Lutz 2017 looking only at internet users and mostly at demographic variables in Great Britain
  - Blank 2017 looking at twitter users and nonusers
Research Questions

1. who is underrepresented in analyses based on social media data?

2. to what extent does this misrepresentation of certain parts of the population bias common estimates in election research?
Data

- data: GLES cross-sectional surveys before and after the German federal election 2017 (Roßteutscher et al. 2018, doi:10.4232/1.13139)

<table>
<thead>
<tr>
<th>Year</th>
<th>mode</th>
<th>n</th>
<th>Sampling strategy</th>
<th>Social media</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017</td>
<td>CAPI</td>
<td>4294</td>
<td>stratified multistage sample (oversampling for East Germany)</td>
<td>Facebook, WhatsApp, YouTube, Twitter, Google+</td>
</tr>
</tbody>
</table>
## Research Design

<table>
<thead>
<tr>
<th>Model</th>
<th>Subset analyzed</th>
<th>Model type</th>
<th>Dependent Variable</th>
<th>Independent Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>all</td>
<td>multinominal logistic</td>
<td>internet and social media usage</td>
<td>gender, age</td>
</tr>
<tr>
<td>2</td>
<td>social media users</td>
<td>logistic</td>
<td>social media usage by platform: Facebook, WhatsApp, YouTube, Twitter, Google+</td>
<td>educational level, marital status, household size, net income</td>
</tr>
<tr>
<td>3</td>
<td>social media users</td>
<td>linear OLS</td>
<td>days of internet usage per week</td>
<td>region (east/west), migration, economic situation (prospective)</td>
</tr>
<tr>
<td>4</td>
<td>social media users</td>
<td>linear OLS</td>
<td>number of social media platforms used</td>
<td>justice (ego), recall (turnout 2013), party identification</td>
</tr>
<tr>
<td>5</td>
<td>social media users</td>
<td>logistic</td>
<td>form of social media usage: commenting and posting</td>
<td>left-right self-assessment, ego positions, populism (Akkerman et al.)</td>
</tr>
</tbody>
</table>
## Research Design

<table>
<thead>
<tr>
<th>Model</th>
<th>Subset analyzed</th>
<th>Model type</th>
<th>Dependent Variable</th>
<th>Independent Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>all</td>
<td>multinomial logistic</td>
<td>internet and social media usage</td>
<td>gender, age</td>
</tr>
<tr>
<td>2</td>
<td>social media users</td>
<td>logistic</td>
<td>social media usage by platform: Facebook, WhatsApp, YouTube, Twitter, Google+</td>
<td>educational level, marital status, household size, net income</td>
</tr>
<tr>
<td>3</td>
<td>social media users</td>
<td>linear OLS</td>
<td>days of internet usage per week</td>
<td>region (east/west), migration, economic situation (prospective)</td>
</tr>
<tr>
<td>4</td>
<td>social media users</td>
<td>linear OLS</td>
<td>number of social media platforms used</td>
<td>justice (ego), recall (turnout 2013), party identification</td>
</tr>
<tr>
<td>5</td>
<td>social media users</td>
<td>logistic</td>
<td>form of social media usage: commenting and posting</td>
<td>left-right self-assessment, ego positions, populism (Akkerman et al.)</td>
</tr>
</tbody>
</table>
Differences in sociodemographics Model 1

- Differences in gender (f.)
- Differences in age (30-<45, 45-<60, 60+)
- Differences in educational level (unassignable, isced0U/1, isced2, isced3, isced5, isced6)
- Differences in east-west (w.)
- Differences in marital status (married, divorced/widowed)
- Differences in household size (2-4p., 5+p.)
- Differences in net income (low-income group, low-middle-income group, middle-income group, high-income group)
- Differences in subjective class affiliation (lower class, working class, central middle class, upper middle class, upper class)
- Differences in migrant background (y.)

Ref.: social media user
N = 3120
Pseudo R² = 0.29
Differences in attitudes Model 1

Ref.: social media user
N = 3120
Pseudo $R^2 = 0.29$
Summary Model 1

- the populations differ on key demographics
- different mechanisms driving internet usage and social media usage
- evidence for different attitudes between online and offline population and different assessment of own prospective economic situation between social media users and nonusers
# Research Design

<table>
<thead>
<tr>
<th>Model</th>
<th>Subset analyzed</th>
<th>Model type</th>
<th>Dependent Variable</th>
<th>Independent Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>all</td>
<td>multinomial logistic</td>
<td>internet and social media usage</td>
<td>gender, age</td>
</tr>
<tr>
<td>2</td>
<td>social media users</td>
<td>logistic</td>
<td>social media usage by platform: Facebook, WhatsApp, YouTube, Twitter, Google+</td>
<td>educational level, marital status, household size, net income</td>
</tr>
<tr>
<td>3</td>
<td>social media users</td>
<td>linear OLS</td>
<td>days of internet usage per week</td>
<td>region (east/west), migration, economic situation (prospective)</td>
</tr>
<tr>
<td>4</td>
<td>social media users</td>
<td>linear OLS</td>
<td>number of social media platforms used</td>
<td>justice (ego), recall (turnout 2013), party identification</td>
</tr>
<tr>
<td>5</td>
<td>social media users</td>
<td>logistic</td>
<td>form of social media usage: commenting and posting</td>
<td>left-right self-assessment, ego positions, populism (Akkerman et al.)</td>
</tr>
</tbody>
</table>
Differences in sociodemographics Model 2

Subpopulation Social Media Users

DV: facebook (y/n)
N= 2145
Pseudo R² = 0.15

DV: whatsapp (y/n)
N= 2136
Pseudo R² = 0.11

DV: youtube (y/n)
N= 2145
Pseudo R² = 0.11

DV: twitter (y/n)
N= 2136
Pseudo R² = 0.13

DV: google+ (y/n)
N= 2145
Pseudo R² = 0.04
Differences in attitudes Model 2

Subpopulation Social Media Users

DV: facebook (y/n)  
N= 2145  
Pseudo R² = 0.15

DV: whatsapp (y/n)  
N= 2136  
Pseudo R² = 0.11

DV: youtube (y/n)  
N= 2145  
Pseudo R² = 0.11

DV: twitter (y/n)  
N= 2136  
Pseudo R² = 0.13

DV: google+ (y/n)  
N= 2145  
Pseudo R² = 0.04
Summary Model 2

- different demographic selectivities for different social media platforms
- possible difference in attitude measures especially among Facebook, WhatsApp and Google+ users
## Research Design

<table>
<thead>
<tr>
<th>Model</th>
<th>Subset analyzed</th>
<th>Model type</th>
<th>Dependent Variable</th>
<th>Independent Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>all</td>
<td>multinomial logistic</td>
<td>internet and social media usage</td>
<td>gender</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>age</td>
</tr>
<tr>
<td>2</td>
<td>social media users</td>
<td>logistic</td>
<td>social media usage by platform: Facebook, WhatsApp, YouTube, Twitter, Google+</td>
<td>educational level</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>marital status</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>household size</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>net income</td>
</tr>
<tr>
<td>3</td>
<td>social media users</td>
<td>linear OLS</td>
<td>days of internet usage per week</td>
<td>region (east/west)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>migration</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>economic situation (prospective)</td>
</tr>
<tr>
<td>4</td>
<td>social media users</td>
<td>linear OLS</td>
<td>number of social media platforms used</td>
<td>justice (ego)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>recall (turnout 2013)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>party identification</td>
</tr>
<tr>
<td>5</td>
<td>social media users</td>
<td>logistic</td>
<td>form of social media usage: commenting and posting</td>
<td>left-right self-assessment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ego positions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>populism (Akkerman et al.)</td>
</tr>
</tbody>
</table>
Differences in sociodemographics Model 3

Subpopulation Social Media Users

- gender (f.)
- age
  - 30-45
  - 45-60
  - 60+
- educational level
  - unassignable
  - isced0/1
  - isced2
  - isced3
  - isced5
  - isced6
- migrant background (y.)
- marital status
  - married
  - divorced/widowed
- household size
  - 2-4p.
  - 5+p.
- net income
  - low-income group
  - low-middle-income group
  - middle-income group
  - high-income group
- subjective class affiliation
  - lower class
  - working class
  - central middle class
  - upper middle class
  - upper class
- east-west (w.)

N= 2145
R² = 0.12
Differences in attitudes Model 3

- Recall (turnout 2013)
- Political interest
  - Economic situation (prospective)
    - Substantially better
    - Slightly better
    - Slightly worse
    - Considerably worse
  - Justice (ego)
    - Much less than the fair share
    - A little less than the fair share
    - A little more than the fair share
    - Much more than the fair share
- Party identification (y.)
- Ego positions
  - Left-right self-assessment
  - Socio-economic dimension
  - Libertarian-authoritarian dimension
  - Climate change
- Populism

Subpopulation Social Media Users

N = 2145
R² = 0.12
Summary Model 3

- internet usage subject to demographic selectivities
- risk of bias on political interest
## Research Design

<table>
<thead>
<tr>
<th>Model</th>
<th>Subset analyzed</th>
<th>Model type</th>
<th>Dependent Variable</th>
<th>Independent Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>all</td>
<td>multinomial logistic</td>
<td>internet and social media usage</td>
<td>gender</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>age</td>
</tr>
<tr>
<td>2</td>
<td>social media users</td>
<td>logistic</td>
<td>social media usage by platform: Facebook, WhatsApp, YouTube, Twitter, Google+</td>
<td>educational level</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>marital status</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>household size</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>net income</td>
</tr>
<tr>
<td>3</td>
<td>social media users</td>
<td>linear OLS</td>
<td>days of internet usage per week</td>
<td>region (east/west)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>migration</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>economic situation (prospective)</td>
</tr>
<tr>
<td>4</td>
<td>social media users</td>
<td>linear OLS</td>
<td>number of social media platforms used</td>
<td>justice (ego)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>recall (turnout 2013)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>party identification</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>left-right self-assessment</td>
</tr>
<tr>
<td>5</td>
<td>social media users</td>
<td>logistic</td>
<td>form of social media usage: commenting and posting</td>
<td>populism (Akkerman et al.)</td>
</tr>
</tbody>
</table>
Differences in sociodemographics Model 4

- gender (f.)
- age
  - 30<45
  - 45<60
  - 60+
- educational level
  - unassignable
  - isced0/1
  - isced2
  - isced3
  - isced5
  - isced6
- migrant background (y.)
- marital status
  - married
  - divorced/widowed
- household size
  - 2-4p
  - 5+p
- net income
  - low-income group
  - low-middle-income group
  - middle-income group
  - high-income group
- subjective class affiliation
  - lower class
  - working class
  - central middle class
  - upper middle class
  - upper class
- east-west (w.)

N= 2145
R² = 0.18

Subpopulation Social Media Users
Differences in attitudes Model 4

Subpopulation Social Media Users

N = 2145
R² = 0.18
Summary Model 4

- demographic selectivities
- possible bias on self perception items
<table>
<thead>
<tr>
<th>Model</th>
<th>Subset analyzed</th>
<th>Model type</th>
<th>Dependent Variable</th>
<th>Independent Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>all</td>
<td>multinomial logistic</td>
<td>internet and social media usage</td>
<td>gender, age</td>
</tr>
<tr>
<td>2</td>
<td>social media users</td>
<td>logistic</td>
<td>social media usage by platform: Facebook, WhatsApp, YouTube, Twitter, Google+</td>
<td>educational level, marital status, household size, net income</td>
</tr>
<tr>
<td>3</td>
<td>social media users</td>
<td>linear OLS</td>
<td>days of internet usage per week</td>
<td>region (east/west), migration</td>
</tr>
<tr>
<td>4</td>
<td>social media users</td>
<td>linear OLS</td>
<td>number of social media platforms used</td>
<td>justice (ego), recall (turnout 2013), party identification</td>
</tr>
<tr>
<td>5</td>
<td>social media users</td>
<td>logistic</td>
<td>form of social media usage: commenting and posting</td>
<td>left-right self-assessment, ego positions, populism (Akkerman et al.)</td>
</tr>
</tbody>
</table>
Differences in sociodemographics Model 5

Subpopulation Social Media Users

DV: postings (y/n)
N = 2136
Pseudo R² = 0.14

DV: comments (y/n)
N = 2088
Pseudo R² = 0.16
Differences in attitudes Model 5

Subpopulation Social Media Users

DV: postings (y/n)
N= 2136
Pseudo R² = 0.14

DV: comments (y/n)
N= 2088
Pseudo R² = 0.16
Summary Model 5

- demographic selectivities
- possible bias in economic self-perception and political interest
Overall Results

- demographic selectivities in all considered selection steps
- this concerns especially age, gender, marital status and educational level
- possible bias in attitudinal variables mainly due to non-coverage of offliners, selection of social media platform and frequency and kind of internet usage
- possible bias in self perceived economic situation due to social media usage (general, specific and number of social media platforms)
- all in all, especially younger persons with a good actual and/or perceived economic situation and politically interested persons are likely to be overrepresented
- nature and direction of bias varies between different selection steps
Conclusions

• significant risk of bias due to non-coverage of offliners on demographic and attitudinal variables
• further demographic selectivities due to exclusion of onliners who do not use (certain) social media platforms
• additionally, possible bias due to different selectivities in certain attitudinal variables and measures of self-perception controlling for demographic differences
Limits and Discussion

- the data used only includes adults eligible to vote in Germany in 2017 -> inferences can only be made to that population
- limits inherent in available variables
- low incidences for some characteristics
- next steps
  - look at more variables
  - try to identify different groups of online and social media usage
  - fit common models in election research and identify differences in estimates due to online and social media usages
Literature


Literature


Data

Roßteutscher, Sigrid; Schmitt-Beck, Rüdiger; Schoen, Harald; Weßels, Bernhard; Wolf, Christof; Bieber, Ina; Stövsand, Lars-Christopher; Dietz, Melanie; Scherer, Philipp; Wagner, Aiko; Melcher, Reinhold; Giebler, Heiko (2018): Pre- and Post-election Cross Section (Cumulation) (GLES 2017). GESIS Data Archive, Cologne. ZA6802 Data file Version 3.0.0, doi:10.4232/1.13139