Westat

Improving Lives Through Research®

Using Machine Learning to Identify Careless Web Respondents

Ting Yan, Gizem Korkmaz, David Cantor, Kevin Wilson, Olivia He, Rashi Saluja

FedCASIC 2024

Acknowledgement

- This work was supported by a grant awarded by the National Science Foundation [NSF-2050809] to Ting Yan (PI) and David Cantor (Co-PI).

Web surveys

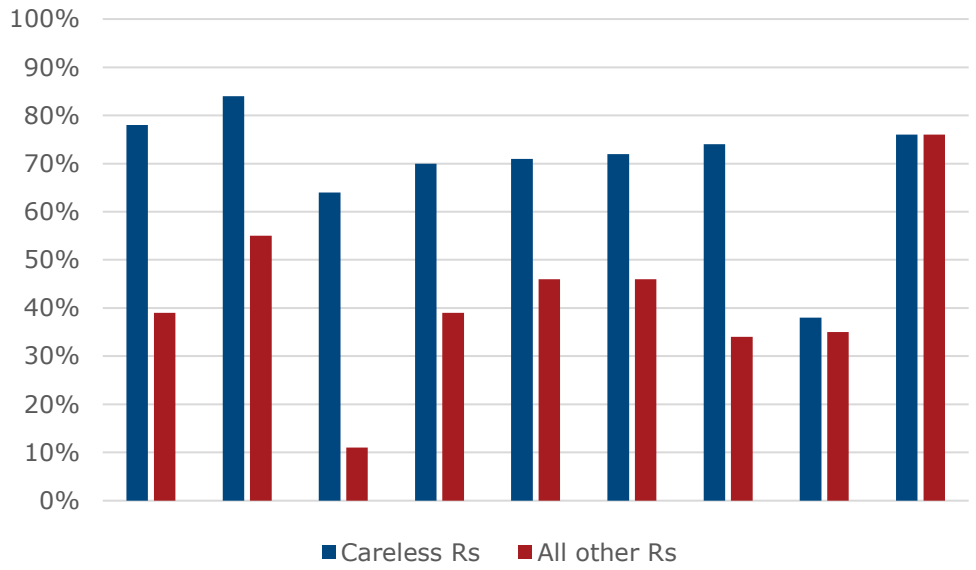
- Web surveys are a popular mode of data collection
- Known data quality issues
  - When used in a single-mode design, coverage error due to exclusion of people without internet access/or device
  - Nonprobability web panels worse inference than probability web panels
  - Nonresponse error when used both in a single-mode or in a multimode design
  - Measurement error when used both in a single-mode or in a multimode design)
    - Two types of web respondents of concern

Two types of web respondents of concern

- Fraudulent respondents (Kennedy et al. 2021; Puleston, 2019)
  - Bot
  - People living outside targeted area (or fake respondents)
  - Duplicate IPs/Multi-completers
  - Ghost respondents
- Careless respondents (Kennedy et al., 2021; Puleston, 2019; Jones et al., 2015)
  - Also called inattentive/insincere/bogus/satisficing respondents
  - Respondents do not read questions carefully, do not spend time and effort to carefully answer questions, multitask, not motivated
  - Focus of the talk

Impact of careless respondents

Kennedy et al. (2021): Table 6



Identifying careless respondents

- During and/or after data collection
  - Attention checks/instructional manipulation checks/traps (Gummer et al., 2021)
  - Speeding (Conrad et al., 2017)
  - Low-incidence question, inconsistent answers (Jones et al., 2015)
  - Proxy indicators of data quality examined alone or together
    - Straightlining/nondifferentiated answers, extreme responses, midpoint, acquiescence, missing data rate
  - Open-ended questions (Kennedy et al., 2021)
  - Response entropy (Tawa, 2021)

This talk

- We explored using machine learning to identify careless respondents
  - Four unsupervised clustering methods
    - K-means clustering
    - Hierarchical clustering
    - Density-based spatial clustering of applications with noise (DBSCAN)
    - Mean shift clustering
- How does each clustering method work?
- Do they converge?

Data

- National Study of Social, Economic, and Health Experiences (NSSEHE)
  - Tracks changes in opinions, life style, and health of Americans
  - Experiments to investigate mechanisms account for panel conditioning
- A sample of 8000 registered voters in two states
  - Invited to participate in four waves of web surveys through mailings, emails, and text messages
- Fourth wave data collection between February 2023 to March 2023
  - a total of 947 completes at a response rate of 71.4%

Data (2)

- Variables used in clustering methods
  - Whether or not R failed the trap questions
  - Whether or not R reported multitasking
  - Whether or not R answered too fast
  - Item nonresponse rate
  - Extreme response rate
  - Middle response rate
  - Response entropy

Identifying careless respondents

Variables used	Cases flagged
Whether or not R failed trap questions	5% failed at least one trap question
Whether or not R reported multitasking	25% reported multitasking
Whether or not R answered too fast	5% fastest
Item nonresponse rate	7% with item nonresponse rate >=5%
Extreme response rate	8% with extreme response rate>=50%
Middle response rate	1% with middle response rate >=50%
Response entropy	10% with largest and smallest 5%

Results

K-Means Clustering

Variables used	Cluster 1 (n=482)	Cluster 2 (n=465)
Whether or not R failed trap questions	5.2%	5.6%
Whether or not R reported multitasking*	**30.2%**	**20.3%**
Whether or not R answered too fast	5.6%	4.5%
Item nonresponse rate>=5%	5.8%	7.8%
Extreme response rate>=50%*	**0%**	**16.1%**
Middle response rate>=50%*	**2.1%**	**0%**
Response entropy too large or too small	10.0%	10.3%

*$p<.05$

Hierarchical clustering

Variables used	Cluster 1 (n=48)	Cluster 2 (n=899)
Whether or not R failed trap questions	8.3%	5.2%
Whether or not R reported multitasking	25.0%	25.3%
Whether or not R answered too fast*	**100%**	**0%**
Item nonresponse rate>=5%	10.4%	6.6%
Extreme response rate>=50%	10.4%	7.8%
Middle response rate>=50%	2.1%	1.0%
Response entropy too large or too small	12.5%	10.0%

*$p<.05$

DBSCAN clustering

Variables used	Cluster 1 (n=98)	Cluster 2 (n=849)
Whether or not R failed trap questions*	**52.0%**	**0%**
Whether or not R reported multitasking	25.0%	25.3%
Whether or not R answered too fast*	**49.0%**	**0%**
Item nonresponse rate>=5%*	**13.3%**	**6.0%**
Extreme response rate>=50%	11.2%	7.5%
Middle response rate>=50%	2.0%	0.9%
Response entropy too large or too small	10.2%	10.1%

*$p<.05$

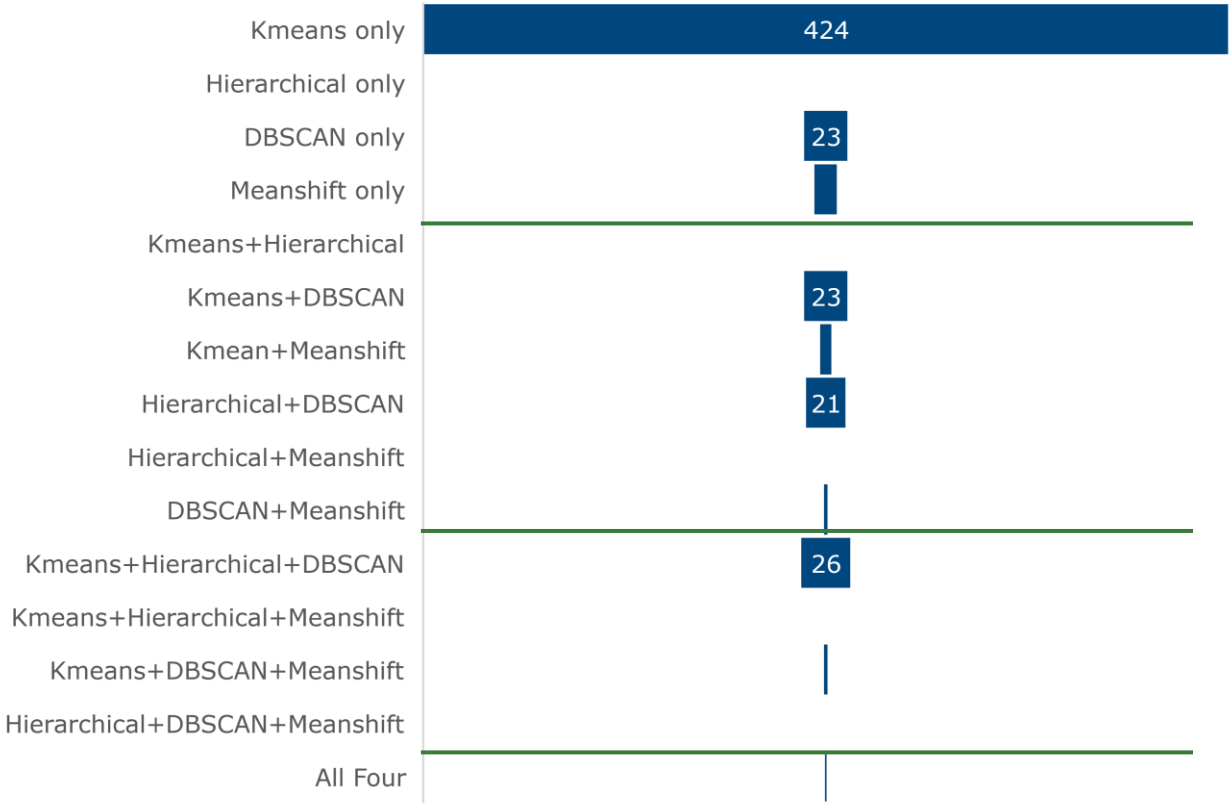
Mean shift clustering

Variables used	Cluster 1 (n=23)	Cluster 2 (n=924)
Whether or not R failed trap questions^	***13.0%***	***5.2%***
Whether or not R reported multitasking	31.8%	25.1%
Whether or not R answered too fast	4.4%	5.1%
Item nonresponse rate>=5%*	**21.7%**	**6.4%**
Extreme response rate>=50%	13.0%	7.8%
Middle response rate>=50%	0%	1.1%
Response entropy too large or too small*	**60.9%**	**8.9%**

*$p<.05$; ^ $p<.10$

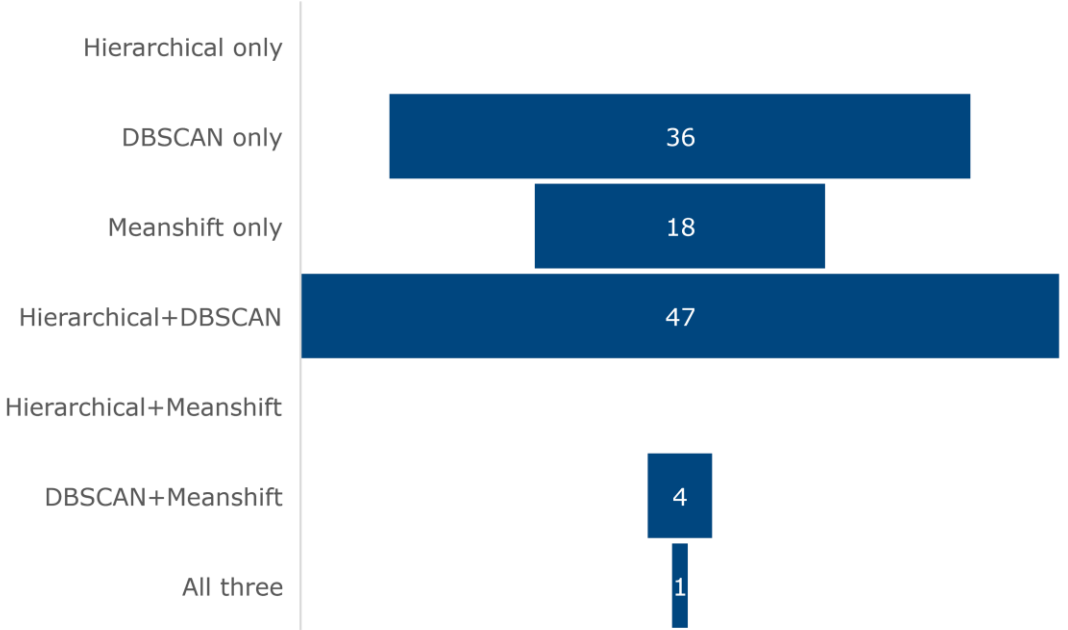
Convergence of four clustering methods

- 85% careless Rs were flagged by only one method
- 10% flagged by two methods
- 5% by three methods
- 1 case flagged by all four methods
- Hierarchical part of DBSCAN



Convergence of four clustering methods

- 51% careless Rs were flagged by only one method
- 48% flagged by two methods
- 1 case flagged by all three methods
- Hierarchical part of DBSCAN



Discussion

Conclusions and Discussion

- Clustering methods can be used to identify careless respondents
  - Different number of respondents were identified
    - K-Means>DBSCAN>Hierarchical>Mean Shift
    - Overlaps between DBSCAN and Hierarchical
  - Different variables determined clusters
    - K-Means: Multitasking, ERS, MRS
    - Hierarchical: speeding
    - DBSCAN: trap, speeding, item nonresponse
    - Mean Shift: trap, item nonresponse, response entropy

Conclusions and Discussion (2)

- More research needed to evaluate and validate the methods
  - Number of clusters?
    - K-Means and hierarchical methods produce solutions of more than 2 clusters
  - Are clusters meaningful and different from each other?
  - Who are in each cluster?
  - Can we use these clusters to inform detection of careless respondents during data collection?
  - Can we predict careless respondents based on these clusters?

Thank you

tingyan@westat.com

westat.com