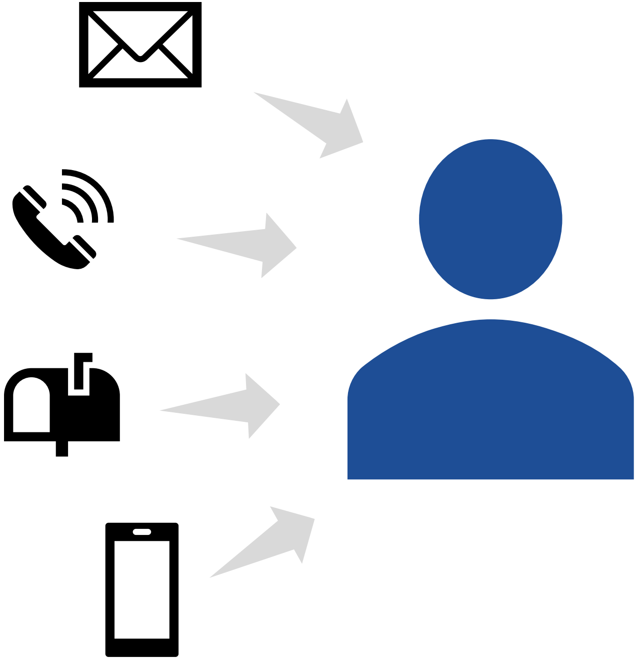
Using Cluster Analysis to Develop a Tailored Contacting Strategy

Jamie Wescott
Michael Duprey
Jerry Timbrook

Background

The Problem

- Researchers are challenged to find new methods for increasing survey response.
- Data collection effort cannot be increased indefinitely:
  - Limited project resources
  - Diminishing returns of additional contact attempts
  - Over-burdened sample members



Background

The Solution

- Find new ways to maximize the efficacy of each contact attempt.
- Contacting sample members using different modes (e.g., telephone, mail) is common practice, but all contact modes are not equally likely to result in response for all sample members (de Leeuw 2005).

Long term goal

Develop a **tailored contacting strategy** that identifies the optimal contacting approach for each SM:

- level of data collection effort (# of contacts)
- contact mode
- timing of contact attempts

Near term goal

Assess our ability to predict SMs' preferred contact times and whether these preferences are stable over time.

Research Questions

- Can clustering techniques be used to identify distinct groups of sample members that share similar characteristics?
- Does group membership have meaningful persistence across time? Or are they artifacts?
- Is there a simpler method to apply multiple interventions in succession (not remodeling each time)?
- Can cluster groups be used to inform individual data collection interventions?

About the Study

- Approach uses data from a nationally representative mixed-mode study of postsecondary students conducted in 2020 (base year).
- Pilot test of tailored contacting strategy conducted on follow-up study (subset of base year cohort) in 2022.

SAMPLE

Base year: ~170,000

Follow-up: ~37,000

DATA COLLECTION

Base year: 2020

Follow-up: 2022

SURVEY MODES

CATI & Web (mobile-friendly)

CONTACT MODES

Phone, Email, Hardcopy Mail, SMS

Sample Member Clustering Method

Our goal was to identify partitions among follow-up sample members that share similar characteristics (i.e., clusters).

Clustering Method

1. Subset the follow-up sample to base year respondents and cluster the cases using the k-prototypes algorithm.

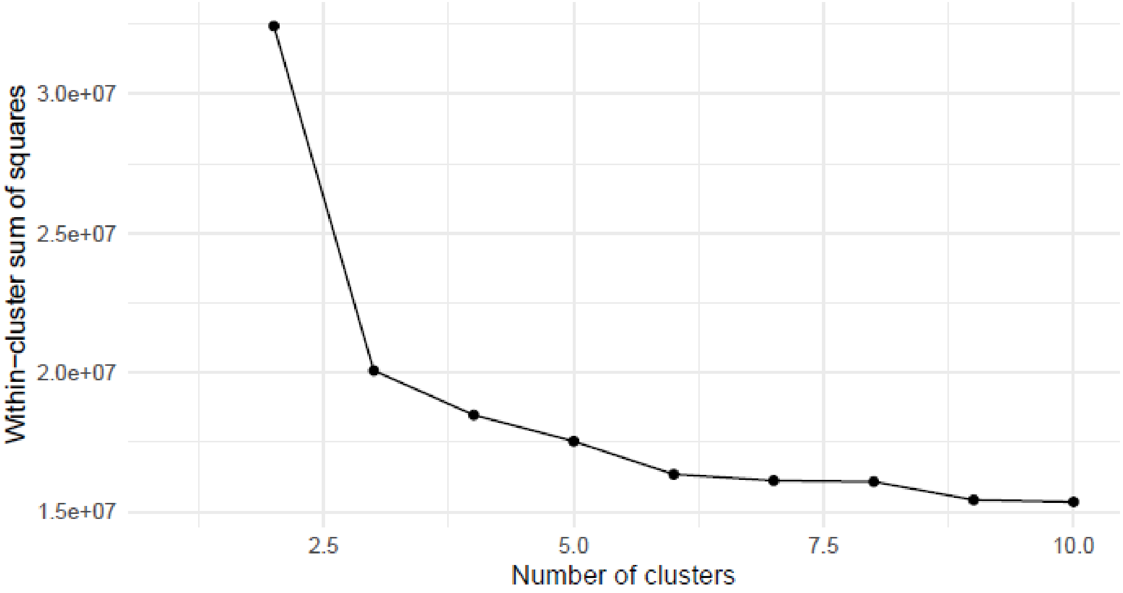
   Used base year data (e.g., race, ethnicity, sex, institution sector) and paradata (e.g., completion mode and time, contact status)

2. Fit a support-vector machine (SVM) model to these clusters.
3. Predict cluster membership for the full follow-up sample, including base year respondents and nonrespondents.

Sample Member Clustering Method

Number of clusters

- The within-cluster sum of squares was evaluated for k-prototypes clusters solutions of 1 through 10.
- The 6-cluster solution was selected as optimal.



Sample Member Clustering Method

- Cluster analysis resulted in 6 sample clusters.
- Pilot experiment tested the preferred contact time predicted based on cluster.

Cluster	Predicted Preferred Contact Time	Sample Size
1	Morning	1,379
2	Evening	5,223
3	Afternoon	6,969
4	Afternoon	4,486
5	Afternoon	4,362
6	Afternoon	1,093
Total		23,512

Experiment Design

Contact tailoring pilot experiment:

- Vary timing of 1st reminder email
    - Experiment group: approximately 3/5 of full follow-up sample
    - Excludes cases that began data collection early or responded prior to reminder 1
- Send email in morning (10:00 EDT), afternoon (2:00 EDT), or evening (6:00 EDT) time slots based on cluster group.
- Email sent on April 19, 2022

Experiment Design

- Sample members were split into intervention and observation groups.
- Email was sent in the morning, afternoon, or evening.

Intervention group

Email time assigned based on cluster

Observation group

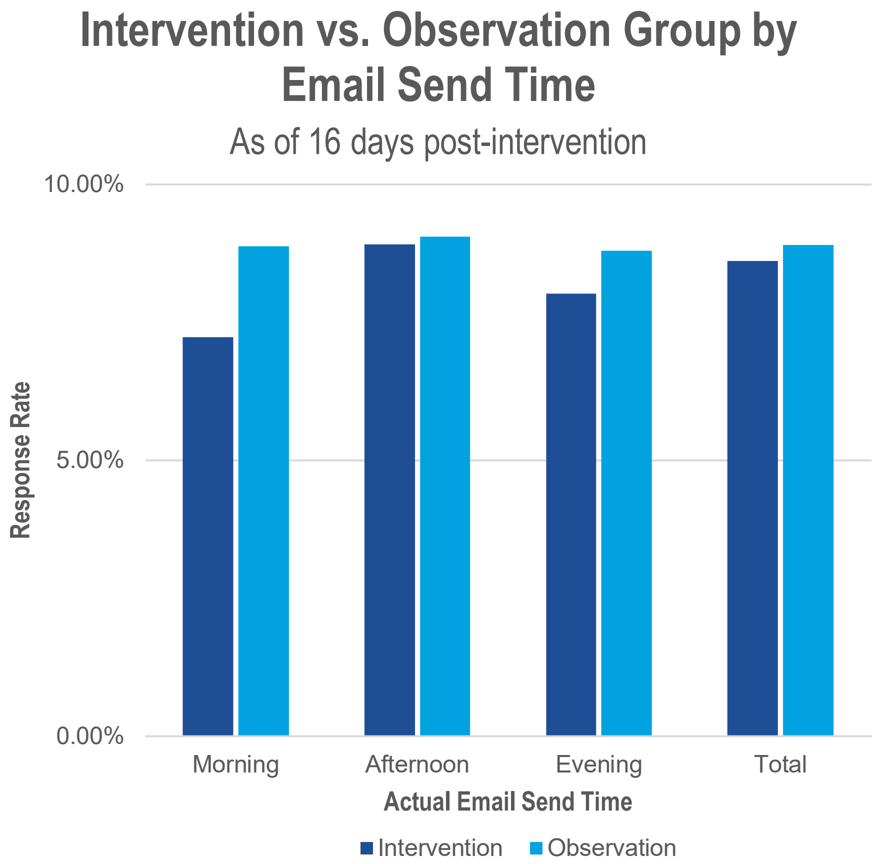
Email time assigned at random

	Intervention Group N = 11,932	Observation Group N = 11,930	
Email Time	Predicted & Actual Time	Predicted Time	Actual Time
Morning	678	701	3,923
Afternoon	8,454	8,456	3,914
Evening	2,630	2,593	3,913

Results

- No significant difference in response rate between intervention and observation groups (p < .05).

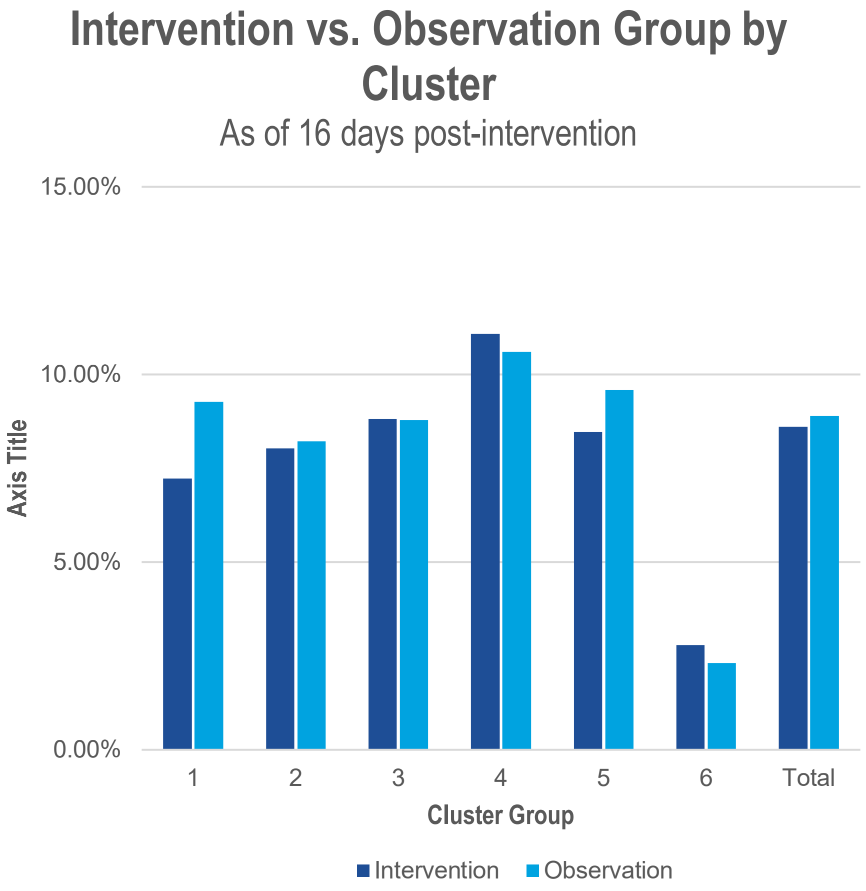
Actual Email Time	Intervention Group RR (%)	Observation Group RR (%)
Morning	7.23	8.87
Afternoon	8.91	9.04
Evening	8.02	8.79
Total	8.61	8.90



Results

- No significant difference in response rate between intervention and observation groups for any cluster (p < .05).

Cluster	Intervention Group RR (%)	Observation Group RR (%)
1	7.23	9.27
2	8.02	8.21
3	8.81	8.78
4	11.08	10.60
5	8.48	9.58
6	2.79	2.31
Total	8.61	8.90



Results

- Observation group email time was assigned at random.
- For approx. 1/3 of observation cases, actual and cluster-predicted time matched by chance (**blue** cells).
- No significant difference in response rate between matched and non-matched groups (p < .05).

	Observation Group RR (%) by Actual Email Time		
Cluster & Predicted Preferred Contact Time	Morning	Afternoon	Evening
1 (Morning)	**7.91**	10.13	9.64
2 (Evening)	9.06	8.69	**6.87**
3 (Afternoon)	7.73	**9.37**	9.23
4 (Afternoon)	10.94	**10.05**	10.85
5 (Afternoon)	10.01	**9.22**	9.49
6 (Afternoon)	1.91	**1.73**	3.16
Total	8.87	9.04	8.78

	RR (%)
Actual time MATCHES predicted time	8.52
Actual time DOES NOT MATCH predicted time	9.10

Conclusions

- Clustering analysis results indicate that it is possible to identify distinct groups of sample members that share similar characteristics.
- Our pilot attempt to convert these cluster groupings into an actionable contacting strategy based on time of day of contact was not successful.
- Possible explanations for null results:
  - Tailoring preferred time is not effective in isolation and must be combined with other interventions (e.g., day of the week, contact mode).
  - Time preference may not be stable over time or predictable a priori.

Thank you

Contact: Name| email: jwescott@rti.org

Appendix

k-prototypes algorithm

- Extends the more common k-means algorithm by defining a distance (psudo) metric between a sample member vector X_i and prototypic cluster vector Q_l, or prototype.
  - $x^r{}_{ij}$ and $q^r{}_{lj}$ are values of numerical variables (as vector elements), and $x^c{}_{ij}$ and $q^c{}_{lj}$ are values of categorical variables for sample member i and prototypic cluster l.
- The variables m_r and m_c represent the number of numerical and categorical variables respectively, while γ_l serves as a weight for categorical variables for cluster l.
- Prototypes are analogous to k-means centroids.

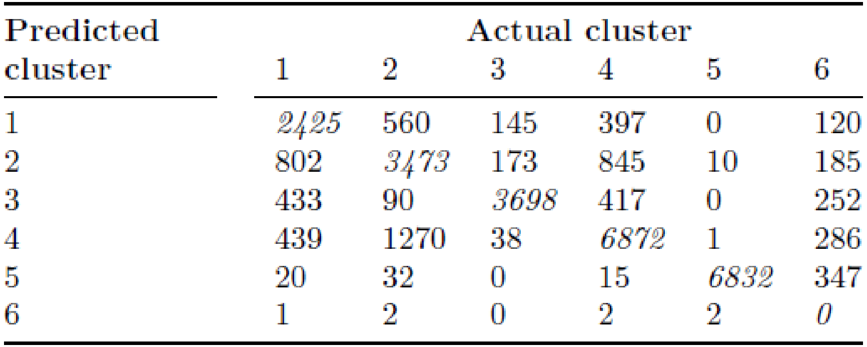
$$d(X_i, Q_l) = \sum_{j=1}^{m_r} (x^r_{ij} - q^r_{lj})^2 + \gamma_l \sum_{j=1}^{m_c} \delta(x^c_{ij}, q^c_{lj})$$

$$\delta(p, q) = \begin{cases} 0 & p = q \\ 1 & p \neq q \end{cases}$$

Appendix

Predicting Cluster Membership for BY Non-Respondents

- Uses SVM supervised learning model, with a linear kernel.
  - Covariates were derived from sampling frame data that are known a priori for all students.
- To evaluate model performance, we selected a random 30% sample of clustered cases for testing and trained the SVM on the remaining 70% of cases.
  - Table 3 provides a confusion matrix contingency table for actual and predicted cluster membership using these cases. Entries on the main diagonal represent true-positives.
- Overall, the model performs well for all clusters, except cluster 6.

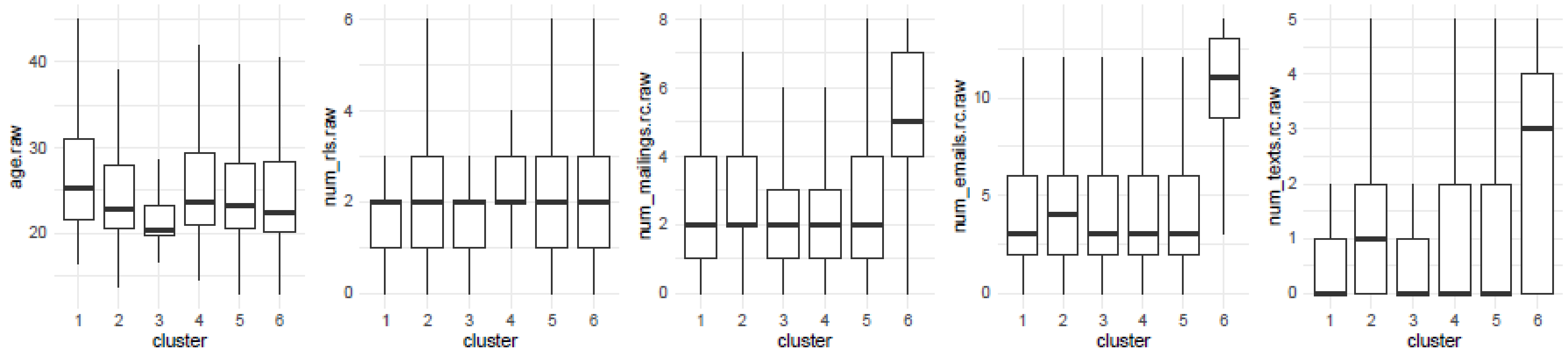


Predicted	Actual cluster					
cluster	1	2	3	4	5	6
1	*2425*	560	145	397	0	120
2	802	*3473*	173	845	10	185
3	433	90	*3698*	417	0	252
4	439	1270	38	*6872*	1	286
5	20	32	0	15	*6832*	347
6	1	2	0	2	2	*0*

Table 3: SVM confusion matrix

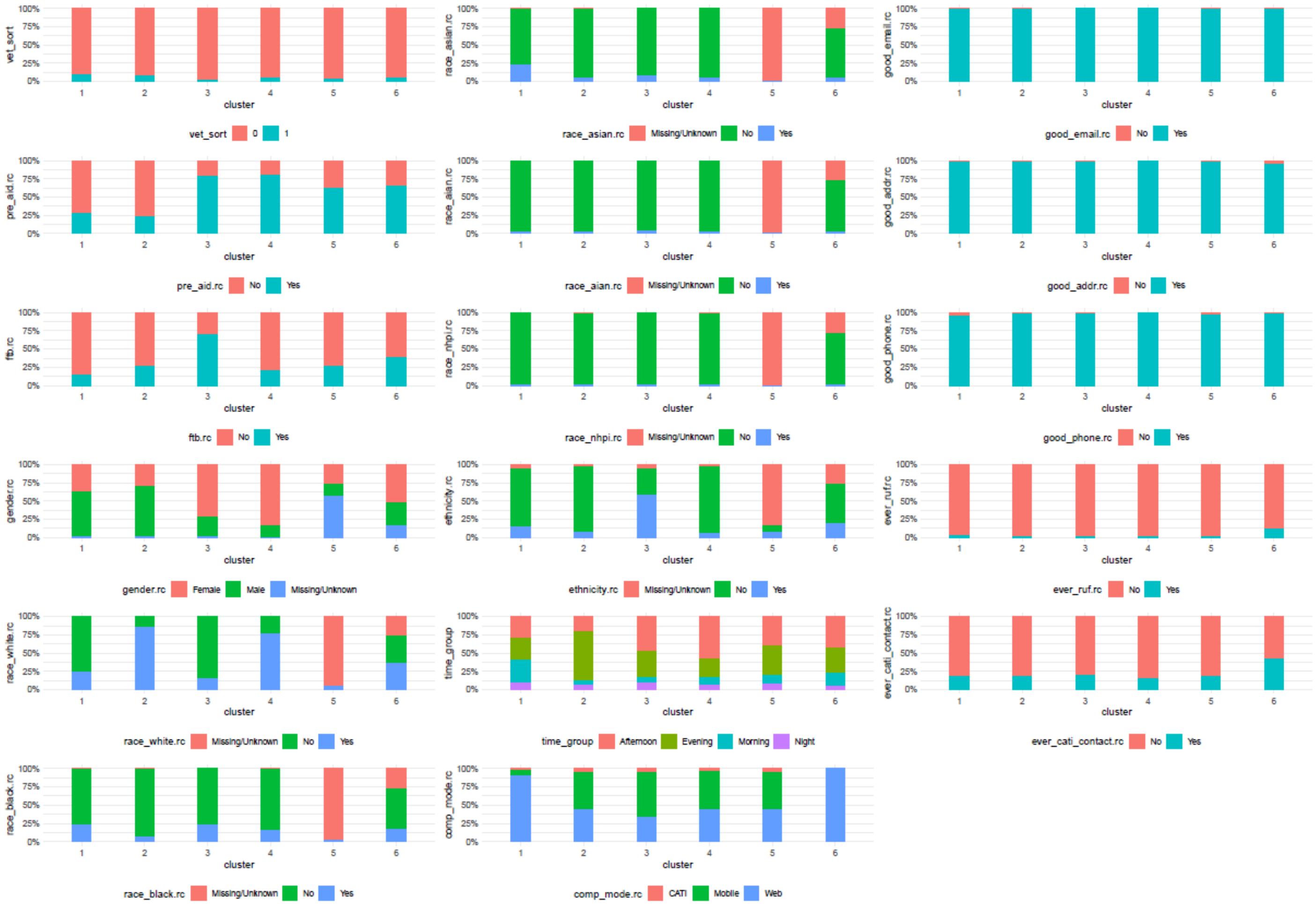
Appendix

Distribution of Numerical Variables by Cluster



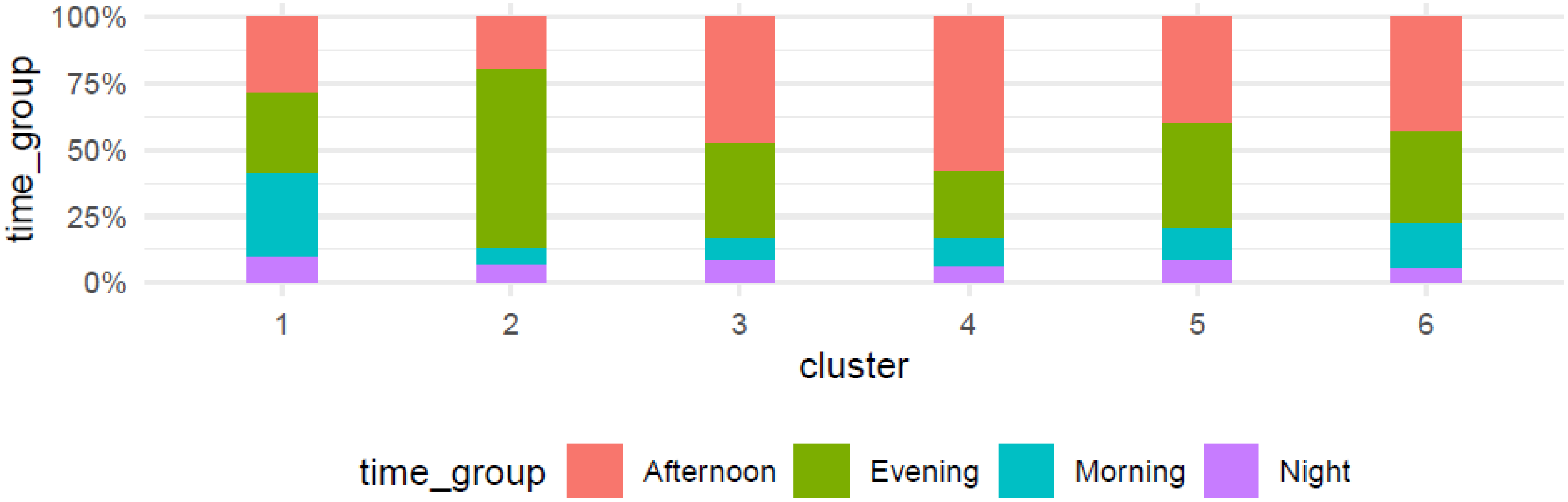
Appendix

Distribution of Categorical Variables by Cluster



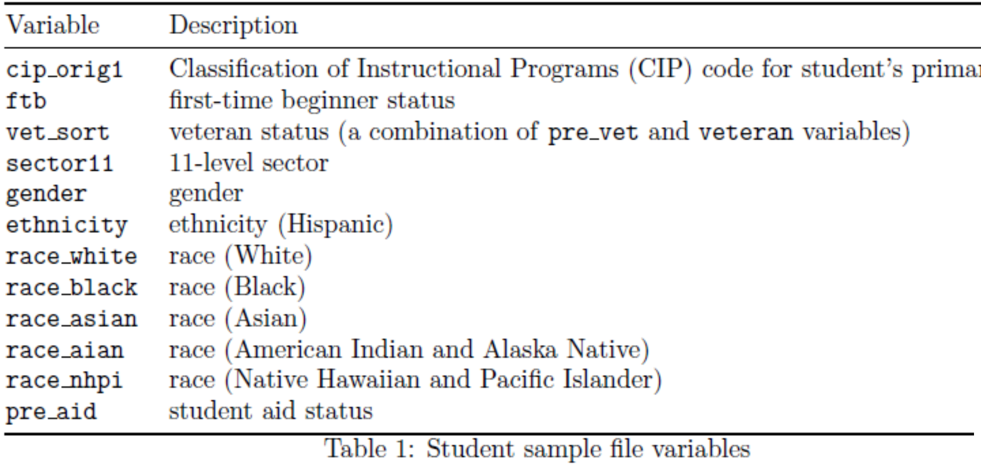
Appendix

Distribution of Categorical Variables by Cluster



Appendix

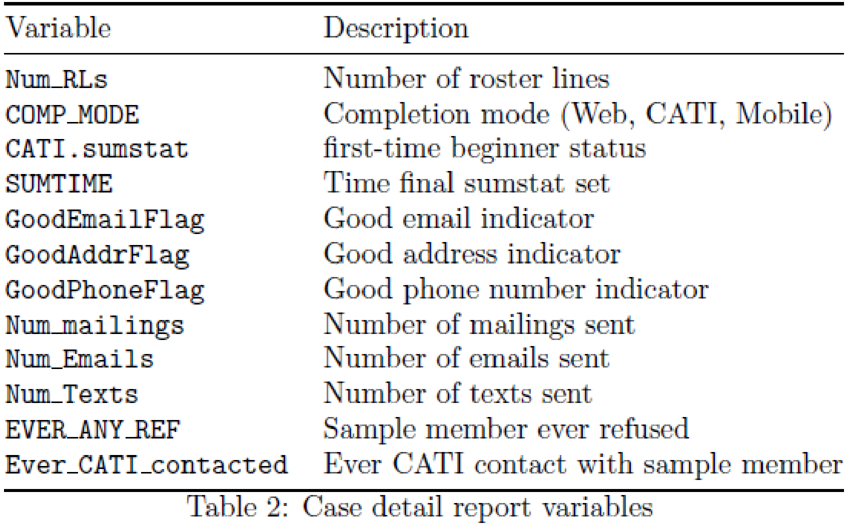
Frame data



Variable	Description
cip_orig1	Classification of Instructional Programs (CIP) code for student's primary major
ftb	first-time beginner status
vet_sort	veteran status (a combination of pre_vet and veteran variables)
sector11	11-level sector
gender	gender
ethnicity	ethnicity (Hispanic)
race_white	race (White)
race_black	race (Black)
race_asian	race (Asian)
race_aian	race (American Indian and Alaska Native)
race_nhpi	race (Native Hawaiian and Pacific Islander)
pre_aid	student aid status

Table 1: Student sample file variables

Paradata



Variable	Description
Num_RLs	Number of roster lines
COMP_MODE	Completion mode (Web, CATI, Mobile)
CATI.sumstat	first-time beginner status
SUMTIME	Time final sumstat set
GoodEmailFlag	Good email indicator
GoodAddrFlag	Good address indicator
GoodPhoneFlag	Good phone number indicator
Num_mailings	Number of mailings sent
Num_Emails	Number of emails sent
Num_Texts	Number of texts sent
EVER_ANY_REF	Sample member ever refused
Ever_CATI_contacted	Ever CATI contact with sample member

Table 2: Case detail report variables

Just for fun…

Increasing contact attempts
Making contact attempts more effective