Mobile Surveys are a New Mode

Tobias Konitzer*[1], Stephanie Eckman†[2], and David Rothschild‡[3]

[1]Stanford University

[2]RTI International

[3]Microsoft Research

February 26, 2017

Abstract

Survey researchers regularly trade-off the costs and benefits of different modes: face-to-face, mail, telephone, and web surveys. This paper makes the case that conducting surveys via applications installed on mobile devices such as smartphones and tablets is a new mode altogether. The mobile mode shares many characteristics of the web mode, but also has unique advantages and disadvantages. We discuss what makes mobile surveys different from web surveys with an emphasis on the passive data that mobile surveys can collect. We detail what forms of passive data are available via the mobile mode and how they can be used to increase the utility of survey data.

Keywords— web surveys, mobile devices, passive data, mode

*tobiask@stanford.edu

†seckman@rti.org

‡davidmr@microsoft.com

1 Introduction

"The use of multimode data collection apps is not simply the next stage in evolution of [computer assisted interviewing], but rather a species unto itself." (6)

The internet offers inexpensive survey data collection via browser-based questionnaires. We refer to these as *web surveys*. Many web surveys are in fact completed on a mobile device (3). However, using a mobile device to complete a web survey can introduce coverage, nonresponse, and measurement error (5; 8). Rather than designing web surveys to work better on mobile browsers, we can develop an application that runs on a smartphone or tablet and serves up survey questions. We call these *mobile surveys*.

In this paper we draw on two recent mobile surveys to demonstrate the power of the mode and the passive data it can collect. We argue that mobile surveys are a new mode of survey data collection, with their own advantages and disadvantages. Survey researchers who take advantage of the benefits offered by the new mobile mode can realize substantial benefits, but should also be aware of their drawbacks.

2 Illustrative Examples

To demonstrate the unique capabilities of mobile surveys, we conducted two surveys using the Pollfish application, which is embedded in third-party mobile applications. The third-party applications display advertisements to users, and some of those contain a poll request. When a user clicks the request, the Pollfish application launches (see Figure 1). In the US, Pollfish has 10 million monthly active registered users. Globally, the number of active users is greater than 300 million.[1]

The first time a user takes a Pollfish survey, the application requests that she accept

[1] `http://www.pollfish.com`.

the Terms of Service and then asks age and gender. Pollfish builds up a database of user attributes to target users for later surveys. For example, a client may want to survey only persons 18 to 25: the application can send the survey request only to eligible persons in its database. Each mobile device receives at most one survey request per day. The application also collects passive data from the mobile device, which we discuss in more detail in Section 3. The Terms of Service detail the data collected.[2]

Pollfish is not the only mobile survey application. Alternatives which may offer similar survey and passive data collection include Confirmit SODA and Unomer.

Survey 1

Our primary survey ran on October 21, 2015 and collected data from 1,977 adult respondents. All users were eligible for the survey. The instrument contained 14 questions.[3] Age and gender were also known for all respondents. No response rate calculation is possible for this type of survey, because we do not know how many saw the advertisement but did not click. All respondents were eligible for random drawings of $10 Amazon gift cards.

The respondent sample overrepresents women and the young. Women are 66% of respondents and are the majority in each age category. The modal age category for women is 25-34 and for men it is 18-24.

Survey 2

On the evening of September 19, 2016, we targeted Pollfish users who were in Chicago, Philadelphia, Dallas and New York at the time.[4] The survey questions asked whether the respondent had watched TV in the last 30 minutes and whether she had seen the Monday

[2]See `http://www.pollfish.com/terms/respondent` for more information.

[3]The text of the questions is given in the Appendix A.

[4]The location of each user was determined through passively collected location data, described in the next section.

Night Football game: Philadelphia Eagles at Chicago Bears.[5] We received 99 responses before the evening ended.

3 Available Passive Data

One of the great strengths of the mobile mode is that it allows researchers to collect passive data from the device, in addition to the survey responses. The passive data we have for the survey respondents are: applications installed, operating system, service provider and device location. Device location and installed applications were delivered to us under a non-disclosure agreement for research purposes only.[6] We have the full set of passive data only for Survey 1 respondents.

Applications installed

We have data on installed applications for all 1,977 respondents. Respondents have installed 16,183 unique applications. The number of applications per person varies from 1 to 329 (median 24). Facebook and Pandora are the most common. We also have access to a classification of installed applications. There are 92 unique categories in the data set. The most common is Tools, followed by Shopping, Entertainment, Social, and Communication. Neither variable tells us how often applications are used.

Operating System

Nearly all of the respondents in our data set use the Android operating system (98.3%). The Pollfish application is embedded in many more Android than iOS applications.

[5] See Appendix B for question wording.

[6] Details of precautions to honor privacy of respondents are outlined in Stanford's IRB documentation, IRB-36007.

Cellular service provider

The passive data contain the name of the user's service provider. The respondents use 65 distinct providers, though some appear to be duplicates: "SUN" and "Sun," for example. In 14.5% of cases, the given provider is "Unknown."

Device Location

The latitude and longitude of the mobile device's location can be derived in two ways. If the GPS is on and communicating with satellites, the device can provide a coordinate that is quite accurate (7). Otherwise, the IP address of the phone can be matched to a location. IP locations are less accurate (10), and can be thrown off by the use of proxies or virtual private networks (VPNs), that make a computer appear to be elsewhere.

Pollfish collected and delivered 943,776 latitude/longitude coordinates from our respondents, spanning May 5, 2015 to April 26, 2016, several months before and after the survey itself. Pollfish collects coordinates whenever a third-party application in which it is embedded is open, making them eligible for a survey, not only during survey administration. Overall 61% of all the coordinates came from GPS signals and 39% from IP addresses. We received no coordinates for one respondent.

For the cases with coordinates, the number of readings per respondent varied from two to 25,808 (median 102). The percent of respondents' coordinates from the more accurate GPS reading varied from 0% to 98%. For 29% of respondents, no coordinates came from GPS.

We show the whereabouts of a randomly chosen individual in Figure 2.[7] We discuss how these data can be useful to researchers below.

[7]We have used the toner-lines version from Stamen Maps to hide this respondent's true location, for privacy reasons. See `http://maps.stamen.com/#toner/12/37.7706/-122.3782` [accessed 2016-12-09].

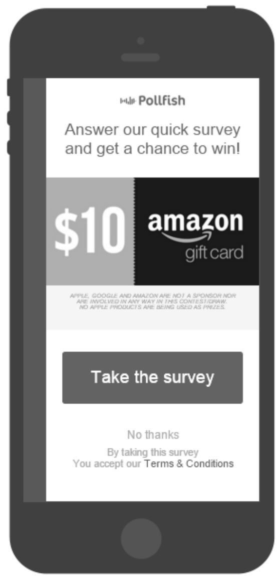

Figure 1: Pollfish Survey Mock-up

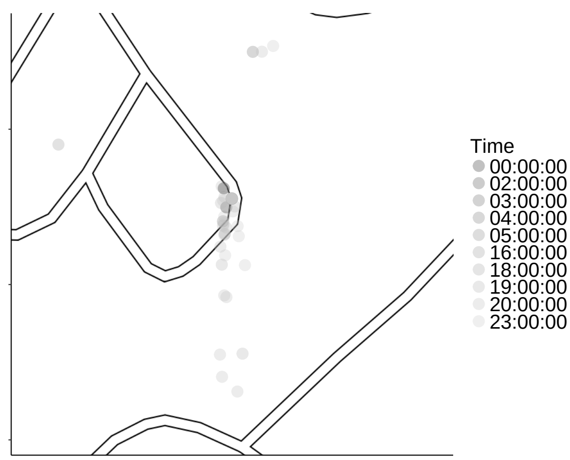

Figure 2: Locations of Example User, by Hour of Day

Web surveys cannot collect most of these variables. They can capture user agent strings from the browsers used to complete the survey, which contain the IP address of the device used, the operating system, and the browser name and version (3). The IP address can be mapped to an approximate location of the respondent at the time of the survey. The coordinates available to mobile surveys are not only more accurate (when they come from the GPS device), they are also available much more frequently, resulting in a rich panel of respondent travel. In addition, no information about installed applications or cellular providers is available to web surveys.

4 Advantages of Mobile Mode

These unique passive data are useful in both collection and analysis of survey data.

Home Location Imputation

The location data collected by the application allow us to deduce the likely location of respondents' homes. We imputed home location with the modal latitude and longitude coordinates between 7pm and 5am local time: 68% of our respondents had location readings in this time window. We discuss below how the imputed home location can be used.

We compared this location to the location at the time of the survey and flagged respondents who seemed to be at home when responding (allowing for an error tolerance of three decimal points, or 360 feet[8]). The survey also asked respondents where they "usually answer mobile polls." 74% of respondents whom we predict were home while taking our survey reported that they usually take polls at home. We did not ask respondents where they were at the time of our survey, because this question was deemed too invasive by the

[8]https://www.lightmanufacturingsystems.com/heliostats/support/decimal-latitude-longitude-accuracy/

IRB. So, without asking directly for where they are, we cannot validate this approach fully, but the 74% accuracy rate is indicative of the potential for granular geo-identification.

Web surveys, on the other hand, must ask respondents for an approximate home location, which some would not provide, and which would likely be coarser (e.g. a ZIP code), or rely on the more imprecise IP-based geo-identification.

Imputation of Respondent Demographics

Pollfish records age and gender of all respondents, but researchers are often interested in race, ethnicity, education and other characteristics.[9] Of course, a survey could ask for these characteristics, but best practices suggest that questionnaires completed via mobile devices should be kept short and for some categories of questions derived characteristics are more accurate than revealed characteristics (5).

We have explored the possibility of imputing education, race, political knowledge and political engagement. We built two imputation models for each variable and validate them with survey data. The first is a naive approach that predicts based on age and gender. The second approach takes advantage of all the available passive data as well as American Community Survey data for the Census Block Group where the imputed home location lies. See Appendix C for more details on these models.

As shown in Figure 3, imputation using passive data does particularly well imputing race: Surveys may wish to consider not asking the race question when they have access to passive data. The imputation model using passive data also makes substantial improvements to predicting education. Both of these variables are geographically clustered, which explains why the model with passive data does well. This approach also outperforms the naive model in predicting political engagement, likely due to the passive data on installed

[9] Pollfish continues to add demographics about its respondents, but researchers will always want more variables than are available.

applications – those who are more politically engaged, for example, may have news applications installed.

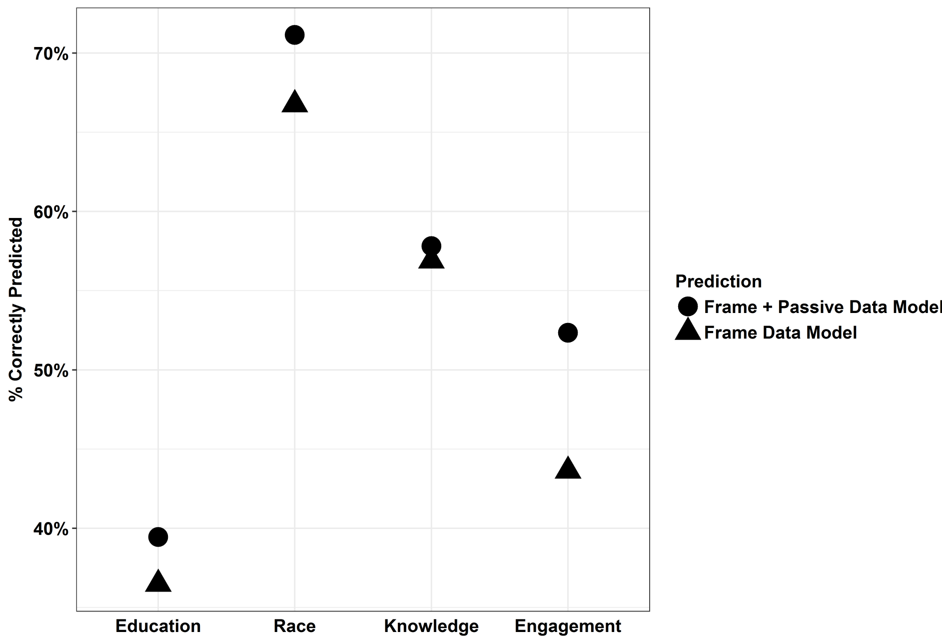

Figure 3: Accuracy of Two Approaches to Imputation, by Variable

Location-Based Targeting

We can also leverage location information to target polls to respondents based on their current whereabouts, as in Survey 2, conducted over one evening in four cities. As expected, respondents in the two cities participating in the football game were more likely to watch the game (Table 1). Such granular geographic targeting paired with rapid response could be used to quickly capture public opinion after an attack, protest or disaster, or measure the short-term effects of exposure to political ads.

Microtargetting operates on an even finer level: for example, if a user enters a bar, the application could pop up questions about drinking behavior. Targeting could also be done

Did you watch/Are you watching Monday Night Football?	Participating City		Non-Participating City	
	n	Column %	n	Column %
Yes, watching it now/caught most of it	17	31.5	5	11.1
Yes, caught some of it	18	33.3	9	20.2
No	19	35.2	31	68.9
Total	54		45	

Table 1: Exposure to Monday Night Football, by City Participation (Unweighted)

retrospectively, for example, to survey persons who were in Manhattan during Hurricane Sandy. In addition to improving sampling efficiency, microgtargetting could also improve measurement: surveys could be sent to respondents who are at the imputed home location, if the survey contains sensitive questions respondents may not want to answer when at work or in public. Pollfish currently offers some microtargetting and is likely to expand its capabilities. Again, web surveys would not be capable of such targeting due to the absence of both fine-grained and historic location data.

5 Disadvantages of Mobile Surveys

In addition to the above advantages that the mobile mode offers over web, there are also disadvantages.

Representation

Both web and mobile surveys suffer from undercoverage. Those without access to the internet are undercovered in web surveys and those without smartphones are undercovered in the mobile mode. From 2011 to 2015, smartphone penetration nearly doubled from 35% to 68%, while desktop and laptop ownership increased only slightly (1), indicating that the mobile mode is able to cover nearly as many people as the web mode.

However, to participate in a mobile survey, respondents need not only have a device but must also install the survey application (or a third party application in which it is embedded). In a recent experiment, 61% of long-time panel participants downloaded a survey application – the compliance rate in the general population would likely be lower. Response rates to the mobile survey were lower than those for the web survey, but the mobile and web respondents showed only small demographic differences. Younger respondents were more likely to respond in the mobile mode and older respondents in the web mode. There were no significant differences in gender, race, and education, but participation and response rates among those who were registered to vote were lower in the mobile than in the web mode. Only a few substantive variables, relating to how mobile devices were used, differed between the two groups (9).

At this point, web surveys can reach a larger and broader population than mobile surveys, but that may not always be the case. Neither mode captures the general population.

Privacy

The collection of passive data raises new concerns about privacy. Survey researchers have always collected personal data from respondents, and these data are linked to addresses and names, though usually only temporarily. However, the passive data available for each respondent in the mboile mode are much richer. We can determine rather precisely where respondents live and work and what their interests are. These data open up powerful opportunities to learn more about our respondents and reduce survey length and burden. But they can also be misused, and the field should develop strong privacy protection policies to minimize risk to those who volunteer their data for research purposes.

6 Discussion

Surveys via applications installed on mobile devices are not a type of web survey, but a new mode. The set of respondents who can be reached by mobile mode surveys is smaller and less representative of the population, but the rich passive data collected via the app may make up for this shortcoming in some research contexts. These passive data can be used to impute respondents' characteristics, leading to shorter surveys. They allow for geographic targeting of respondents which improves sampling efficiency and can perhaps decrease measurement error. At the same time, the privacy implications of collecting these passive data cannot be ignored.

We have discussed only those advantages and disadvantages of the mobile mode that follow from the combination of survey and passive data. The mode offers other advantages as well, such as consistent design across platforms and devices, and the ability to work offline (see 2, for a list of other advantages and disadvantages).

Survey modes will continue to evolve. Extrapolating from recent trends, it is likely mobile surveys will be an increasing share of the market. Thus, as we move into the future of ever-connected devices and adapt our survey practices in response to this changing world, we should be aware of the unique advantages and disadvantages offered by this unique survey mode.

References

Anderson, Monica (Pew Research Center). 2015. "Technology Device Ownership: 2015.". **URL:** *http://www.pewinternet.org/2015/10/29/technology-device-ownership-2015/*

Buskirk, Trent and Charles Andrus. 2012. "Smart Surveys for Smart Phones: Exploring Various Approaches for Conducting Online Mobile Surveys via Smartphones." *Survey Practice* 5(1).

Callegaro, Mario. 2010. "Do You Know Which Device Your Respondent Has Used to Take Your Online Survey?" *Survey Practice* 3(6).

Chalmers, R Philip. 2012. "mirt: A multidimensional item response theory package for the R environment." *Journal of Statistical Software* 48(6):1–29.

Couper, Mick P., Christopher Antoun and Aigul Mavletova. 2017. Mobile Web Surveys: A Total Survey Error Perspective. In *Total Survey Error in Practice*, ed. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, C. Tucker and B. West. Wiley.

Link, Michael W., Nielsen Joe Murphy, Michael F. Schober, Trent D. Buskirk, Jennifer Hunter Childs and Casey Langer Tesfaye. 2014. "Mobile Technologies for Conducting, Augmenting and Potentially Replacing Surveys: Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research.". **URL:** *https://www.aapor.org/getattachment/Education-Resources/Reports/REVISED_Mobile_Technology_R*

McNamara, Joel. 1981. *GPS For Dummies.* Wiley Publishing: Hoboken, NJ.

Peterson, Gregg, Jamie Griffen, John LaFrance and JiaoJiao. Li. 2017. Smartphone Participation in Web Surveys: Choosing between the Potential for Coverage, Nonresponse and Measurement Error. In *Total Survey Error in Practice*, ed. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, C. Tucker and B. West. Wiley.

Pew Research Center. 2015. "App vs. Web for Surveys of Smartphone Users.".
URL: *http://www.pewresearch.org/2015/04/01/app-vs-web-for-surveys-of-smartphone-users/*

Poese, Ingmar, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet and Bamba Gueye. 2011. "IP Geolocation Databases: Unreliable?" *SIGCOMM Comput. Commun. Rev.* 41(2):53–56.

Samejima, Fumiko. 1969. "Estimation of Latent Ability Using a Response Pattern of Graded Scores." *Psychometrika Monographs* 34(4).

Zou, Hui and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.

A Questions and Response Options in Survey 1

1. What is the highest level of education that you have completed? (single answer, set order): Less than high school / High school / Some college / College / Graduate school
2. What is your race or ethnic group? (multiple answers, set order): White / Hispanic or Latino / Black or African-American / Native American or American Indian / Asian/Pacific Islander / Other
3. Where do you most commonly answer polls on your phone? (single answer, shuffle order): Home / Work / Commuting / Out-and-about
4. Which of these programs/stations do you watch regularly? (multiple answers, shuffle order/last fixed): ABC/NBC/CBS Nightly News / Fox News / CNN / MSNBC / NFL Football / Empire / The Voice / Big Bang Theory / Dancing with the Stars / NCIS / Walking Dead / Sportscenter / How to Get Away with Murder / Other/None
5. Which of these blogs/newspapers do you read regularly (online or offline)? (multiple answers, shuffle order/last fixed): New York Times / Washington Post / Wall Street Journal / USA Today / Los Angeles Times / Daily News of New York / New York Post / Huffington Post / TMZ / Business Insider / Mashable / Gizmodo / LifeHacker / Gawker / Other/None
6. How often do you talk about politics with other people? (single answer/set order): All the time / Daily / A few times a week / Once per week / Once per month / Rarely
7. Who is the Chief Justice of Supreme Court? (single answer, shuffle order/last fixed): John Roberts / Paul Ryan / William Rehnquist / Samuel Alito / Antonin Scalia /

Anthony Kennedy / Do not know

8. In which countries does ISIS currently control meaningful territory? (multiple answers, shuffle order/last fixed): Syria / Ukraine / Greece / Iraq / Iran / Turkey / Saudi Arabia / Indonesia / Libya / Pakistan / Do not know

9. What is the maximum number of terms to which the president of the USA can be elected? (single answer/set order): 1 / 2 / 3 / 4 / No limit / Do not know

10. On which of the following does the U.S. federal government currently spend the least? (single answer, shuffle order/last fixed): Foreign aid / Medicare / National defense / Social Security / Interest on debt / Transportation / Do not know

11. What month of the year were you born? (single answer/set order): January / February / March / April / May / June / July / August / September / October / November / December / Rather not say

12. What day of the month were you born? (single answer/set order): 1 / 2 / 3 / 4 / 5 / 6 / 7 / 8 / 9 / 10 / 11 / 12 / 13/ 14 / 15 / 16 / 17 / 18 / 19 / 20 / 21 / 22 / 23 / 24 / 25 / 26 / 27 / 28 / 39 / 30 / 31 / Rather not say

13. What is your political party affiliation? (single answer/reverse order): Strong Democrat / Weak Democrat / Lean Democrat/ Independent / Independent/Other / Lean Republican/Independent / Weak Republican / Strong Republican / Not sure

B Questions and Response Options in Survey 2

1. How important is this election relative to other recent presidential elections? (single answer/set order): Most Important / More Important / Average Importance / Less Important /Least Important

2. How do you plan on engaging in the presidential election? (multiple answer/random order): Vote / Volunteer / Donate / Talk to friends / Nothing

3. Your opinion of Donald Trump is (single answer/set order): Very Favorable / Somewhat Favorable / Somewhat Unfavorable / Very Unfavorable / No Opinion/Not Sure

4. Your opinion of Hillary Clinton is (single answer/set order): Very Favorable / Somewhat Favorable / Somewhat Unfavorable / Very Unfavorable / No Opinion/Not Sure

5. Have you been watching TV over the last 30 minutes? (single answer/set order): Yes / No

6. Did you catch any of the Eagles at Bears NFL game? (single answer/set order): Yes, watching it now/caught most of it / Yes, caught some of it / No

7. How many political TV advertisements have you seen today? (single answer/set order): 0 / 1 / 2 / 3 / 4 / 5+

8. Are there any political TV advertisement you have seen this year that you really remember? (single answer/set order): Yes, mainly Clinton ads / Yes, mainly Trump ads / Yes, mix of Trump and Clinton ads / Not really

9. How often do you consume information about the presidential election? (single answer/set order): All day/every day / Daily / Weekly / Not much at all

10. How would you vote if the election were held today? (single answer/set order): Definitely Democrat Hillary Clinton / Lean Democrat Hillary Clinton / Lean Republican Donald Trump / Definitely Republican Donald Trump / Other

C Details on Imputation Models

We used two approaches to imputing five respondent characteristics. The first approach uses only frame data, and the second uses both frame and passive data.

We test the accuracy in predicting hold-out data – We divided our data into test (20%), or N=395 responses, and training data (80%), using all our 1,977 responses.

The variables we impute are: race, education, political knowledge and political engagement. To build the political knowledge index, we simply tale the mean of the knowledge-questions mentioned in the Appendix A. For the political engagement index, we do not want to assume that every item contributes equally to the underlying scale. To compensate, we used indicators of readership of *New York Times*, *Wall Street Journal*, *Washington Post* and an ordinal measure of political discussion to scale a graded 2-parameter Polytomous Item Response Theory (IRT) model. We then derive latent scores of political engagement, as specified in Equation 1. This model permits a larger number of categories ($k > 2$) and treats the probability of scoring in category k as the probability of responding in (or above) this category minus the probability of responding in (or above) the next category $k + 1$ (11; 4).

$$
\begin{aligned}
P(x_{ij} = k|z_i) &= Pr(x_{ij} \geq k|\alpha_{jk}, \beta_j, z_i) - Pr(x_{ij} \geq k+1|\alpha_{j,k+1}, \beta_j, z_i) \\
Pr(x_{ij} \geq k|\alpha_{jk}, \beta_j, z_i) &= \frac{1}{1 + e^{-(\alpha_{jk} + \beta_j * z_i)}} \\
Pr(x_{ij} \geq k+1|\alpha_{j,k+1}, \beta_j, z_i) &= \frac{1}{1 + e^{-(\alpha_{j,k+1} + \beta_j * z_i)}}
\end{aligned}
\tag{1}
$$

where $P(x_{ij} = k)$ is the probability that individual i falls in the k^{th} response category for item j, z_i is the standing of individual i on the latent trait dimension, α_{jk} denotes the question- and category-specific difficulty parameter, and β_j are the discrimination parameters. Conceptually, the difficulty parameters here represent the cut-off points in the cumulative probability scale; the value for difficulty parameter α_{jk} represents the average latent trait score for a 50% chance of assigning either a rating of $(k, k-1, ..., k-(k-1)$ or a rating of $(k+1, k+2, ..., p)$ to item j. We identify the model by constraining the distribution of the latent traits to standard-normal.

Frame Data Imputation Models

These models use age and gender to predict the four attributes. Because our variables are polytomous in nature. We fit a multinomial regression model to predict class probabilities. To compute accuracy of our prediction, we impute the class with the highest posterior class probability to our hold out data, and compute binary accuracy.

Frame + Passive Data Imputation Models

For this model, we use respondents' gender and age as above. In addition, we derive the home latitude/longitude by taking the modal readings between 07:00 p.m. and 05:00 a.m. local time per respondent. We then identify the home Census Block Group of each respondent, and include in the model the proportion White, proportion Black, proportion Hispanic, proportion Other, proportion male/female less-than-high-school, high-school, some-college, college and post-college for that block group. The data are from the 2014 ACS 5-year estimates for education and race, treating Hispanics as pure Hispanics, non-Hispanic Whites as Whites, non-Hispanic Blacks as Blacks, and non-Hispanic Others as Others. We also use binary indicators for installed applications, operating system, cellphone model and respondents' home Census Block Group. In total, we rely on 18,476 predictors. Because we have more covariates than observations, we use elastic net shrinkage models (12), and optimize tuning parameters via 10-fold cross-validation.

Again, because our variables are polytomous in nature, we fit a multinomial version of the elastic net model to predict class probabilities. To compute accuracy of our prediction, we impute the class with the highest posterior class probability to our hold out data, and compute binary accuracy.

References

Chalmers, Philip. 2012. "Mirt: A Multidimensional Item Response Theory Package for the R Environment." *Journal of Statistical Software 48(6).*

Samejima, Fumiko. 1969. "Estimation of Latent Ability using a Response Pattern of Graded Scores." *Psychometrika Monograph Supplement 34.*