Estimating Propensity of Survey Response by Mode Type Using Regression Trees

Gavin Corral[1], Tyler Wilson[2], Andrew Dau[3]

[123]USDA National Agricultural Statistics Service, 1400 Independence Avenue SW,
Washington DC, 20250

Abstract

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) produces over 400 publications annually. NASS employs multiple data sources when estimating the propensity of survey response, including USDA agencies and the Census Bureau. In an effort to assist regional field offices (RFOs) and increase response rates, several new propensity score models based on Logistic Regression, Bootstrap Forest, and Boosted Regression models have been developed. The models using the Bootstrap Forest approach outperformed other methods. This paper outlines two models: Model 1 (more expensive) estimates the probability of response by field enumeration and; Model 2 (less expensive) estimates the probability of response by mail or computer assisted telephone interviewing (CATI). The potential for using these models with measures of impact as part of the data collection strategy to increase response rates of NASS surveys and increase overall data collection efficiency is also discussed.

Key Words: Propensity, Scores, Impact, Machine Learning, Boosted Trees, Classification Trees, Classification

"The Findings and Conclusions in This Preliminary Publication Have Not Been Formally Disseminated by the U. S. Department of Agriculture and Should Not Be Construed to Represent Any Agency Determination or Policy."

1. Introduction

The National Agricultural Statistics Service (NASS) has had a growing interest in using response propensity modeling (RPM) via classification Trees to target respondents based on their likelihood to respond. Only recently, has a respondent’s likelihood to respond been explored alongside one’s preferred mode of response.

In survey methods, RPM fits under the umbrella terms of Responsive Design and Adaptive Design (Groves and Heeringa 2006; Tourangeau et al. 2017; Lavrakas et al. 2018). However, recent RPM methods have been tailored to specific agency and organizational needs within the parameters of each survey with oftentimes the goal of increasing response rates, decreasing costs, and producing more accurate estimates (Lavrakas et al. 2018). This research explores RPM in terms of increasing response rates and decreasing costs with the intent of continuing into how RPM can produce more accurate estimates.

Response propensity scores at NASS are derived using several variations of classification or partition tree models. Classification models do not require strict specification rules such as interaction and missing data (Earp et al. 2014). And generally speaking, results from the models, can be interpreted as a series of hierarchal rules or breakpoints (Phipps and Toth 2012). Recent upgrades in SAS JMP 12 and 13 Pro have made comparing different classification techniques such as random or partition Trees, Bootstrap Trees, and Boosted Trees a simple, yet effective, tool for finding a useful model within a short, operational timeline.

This RPM research explores the possibility of using two models when planning data collection procedures.

- Model #1 – Identify records most likely to be complete via Field Enumeration
- Model #2 – Identify records most likely to be complete via Mail/CATI

Model 1 provides an indicator on likelihood to respond via an expensive mode (field), while Model 2 provides an indicator on likelihood to respond via an inexpensive mode. These models would assist survey administrators and regional offices decide on the most effective data collection strategy for each respondent in addition to providing agency administrators insight into the allocation of funds.

The initial dataset contains over 700 variables such as previous response history aggregated over time, census bureau variables relating to spatial demographics and an assortment of NASS metadata. This paper discusses the research methods, results, and potential uses of modeling by inexpensive and expensive mode for each respondent in the Crops Acreage, Production and Stocks (APS) Survey. Crops APS is conducted four times a year and provides users with estimates of crop acreage, yields and production, and quantities of grain and oilseeds stored on farms (NASS 2018).

2. Methods

JMP Pro 13 software was utilized for all statistical analysis in this project. Logistic Regression, Boosted Trees, and Random Forests approaches were considered, evaluated, and compared for the modelling of propensity scores. Measures of fit, misclassification rates, and sorting efficiency were used to determine the most useful model. For this study we collapsed response variables according to whether or not they required field enumeration (model 1) or not (model 2).

Logistic Regression variables were selected based on their significance (p-value). Both the Boosted Tree and Bootstrap Forest models were fitted using the variables column contributions (>0.02). After the reduced models were fit, each model was evaluated against the others based on misclassification, G^2, and area under the curve AUC or the ROC curve from the validation set.

Once the reduced models were fit, a validation column was created with the JMP dataset to use for model comparison. For model comparison, only the results based on the test set were considered (shown in output as "validation" set).

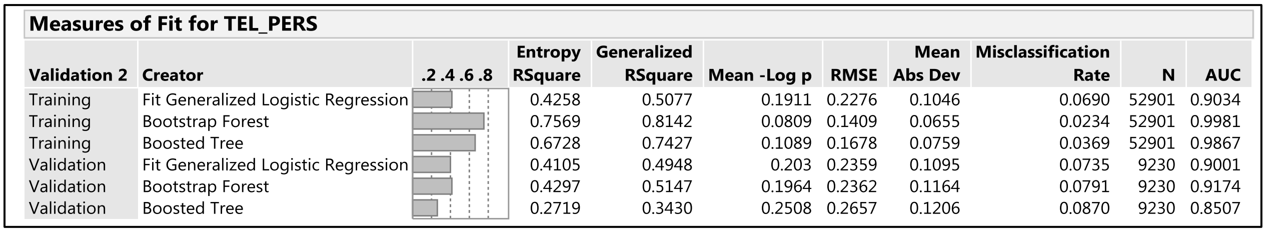
3. Results

Using the model comparison feature, propensity scores for Models 1 and 2 were derived using Boosted Tree, Bootstrap Forest, and Logistic Regression approaches. Their measures of fit are presented in this chapter for the field enumeration model and the mail or CATI model. The results of best fit model for each mode is highlighted and a confusion matrix is presented in an effort to understand the Type 1 and 2 errors expected. Although over 700 variables were explored while building these models, only variables from NASS relating to specific response history information were ultimately used.

3.1 Model #1 Field Enumeration

In *Table 1* below, a Bootstrap Forest, Boosted Tree, and generalized Logistic Regression model are presented for field enumeration. For the validation sets, both the entropy and generalized r-square are highest for the Bootstrap Forest procedure. The Logistic Regression model has the lowest misclassification rate followed closely by the Bootstrap Forest.

Table 1. Measures of fit for each of the three models considered to estimate propensity to respond via field enumeration



Measures of Fit for TEL_PERS

Validation 2	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Training	Fit Generalized Logistic Regression		0.4258	0.5077	0.1911	0.2276	0.1046	0.0690	52901	0.9034
Training	Bootstrap Forest		0.7569	0.8142	0.0809	0.1409	0.0655	0.0234	52901	0.9981
Training	Boosted Tree		0.6728	0.7427	0.1089	0.1678	0.0759	0.0369	52901	0.9867
Validation	Fit Generalized Logistic Regression		0.4105	0.4948	0.203	0.2359	0.1095	0.0735	9230	0.9001
Validation	Bootstrap Forest		0.4297	0.5147	0.1964	0.2362	0.1164	0.0791	9230	0.9174
Validation	Boosted Tree		0.2719	0.3430	0.2508	0.2657	0.1206	0.0870	9230	0.8507

The Receiver Operating Curve (ROC) lines for the validation sets in *Figure 1* illustrate that all three models classify response by field well (>0.9), however the green line (Bootstrap Forest) shows the most area underneath the curve. This implies that the Bootstrap Forest model is most efficient at sorting the data.

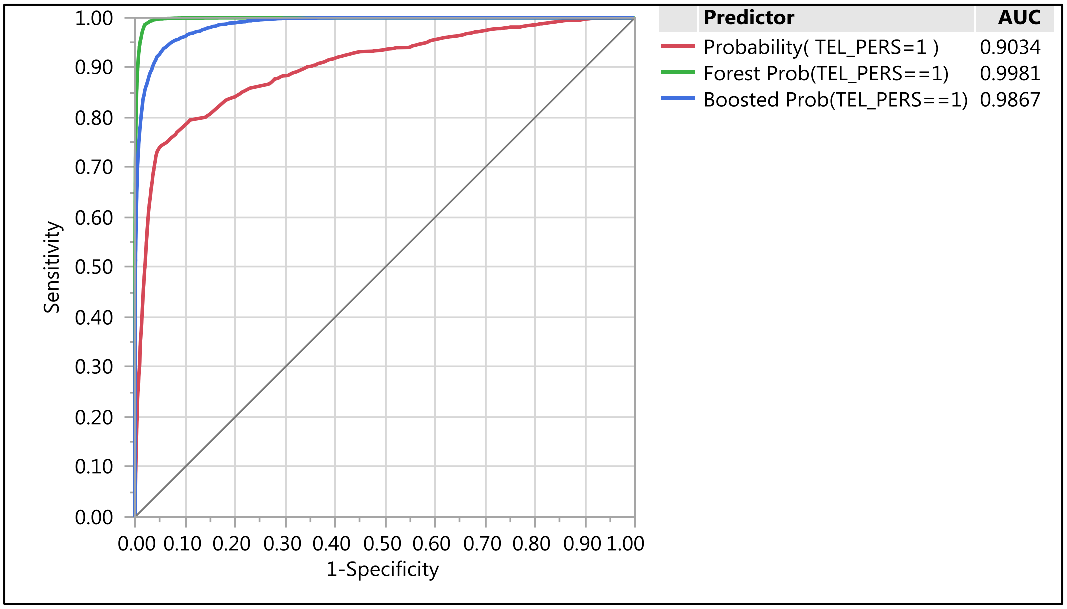
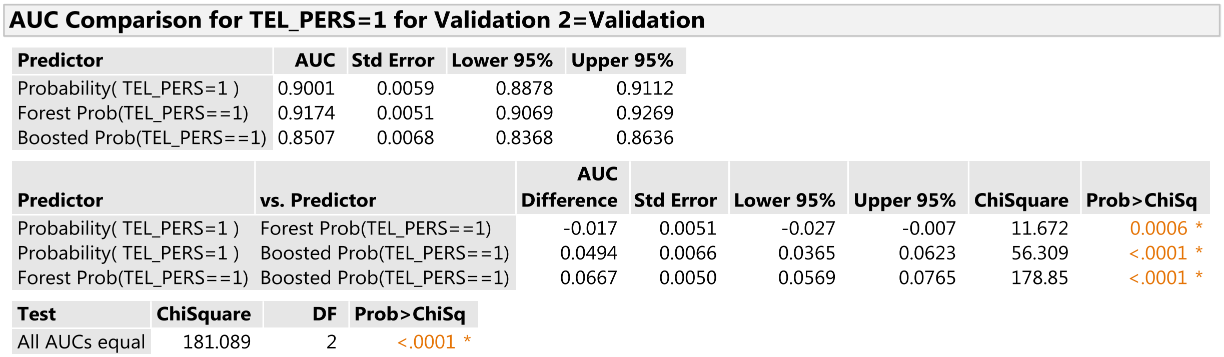


Figure 1. Illustrates the receiver operating curves for the three candidate models.

According to the measures of fit, the Bootstrap Forest model was the chosen model. Furthermore, significance testing of each models (*Table 2*) area under the ROC curve provides evidence that the Bootstrap Forest model outperforms the other methods.

Table 2 AUC comparison for models predicting whether or not a record will respond vie field enumeration.



AUC Comparison for TEL_PERS=1 for Validation 2=Validation

Predictor	AUC	Std Error	Lower 95%	Upper 95%
Probability(TEL_PERS=1)	0.9001	0.0059	0.8878	0.9112
Forest Prob(TEL_PERS==1)	0.9174	0.0051	0.9069	0.9269
Boosted Prob(TEL_PERS==1)	0.8507	0.0068	0.8368	0.8636

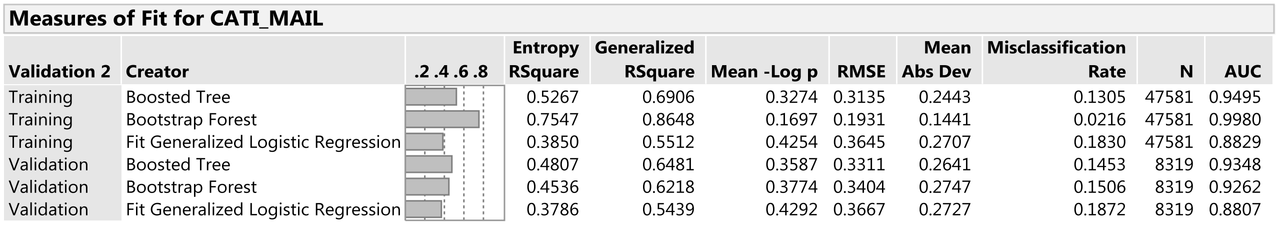
Predictor	vs. Predictor	AUC Difference	Std Error	Lower 95%	Upper 95%	ChiSquare	Prob>ChiSq
Probability(TEL_PERS=1)	Forest Prob(TEL_PERS==1)	-0.017	0.0051	-0.027	-0.007	11.672	0.0006 *
Probability(TEL_PERS=1)	Boosted Prob(TEL_PERS==1)	0.0494	0.0066	0.0365	0.0623	56.309	<.0001 *
Forest Prob(TEL_PERS==1)	Boosted Prob(TEL_PERS==1)	0.0667	0.0050	0.0569	0.0765	178.85	<.0001 *

Test	ChiSquare	DF	Prob>ChiSq
All AUCs equal	181.089	2	<.0001 *

3.1 Model #2 Mail or CATI

Table 3 below illustrates the measures of fit for each model by training and validation set. Boosted tree model has the highest R^2, Entropy R^2, and misclassification of the three models.

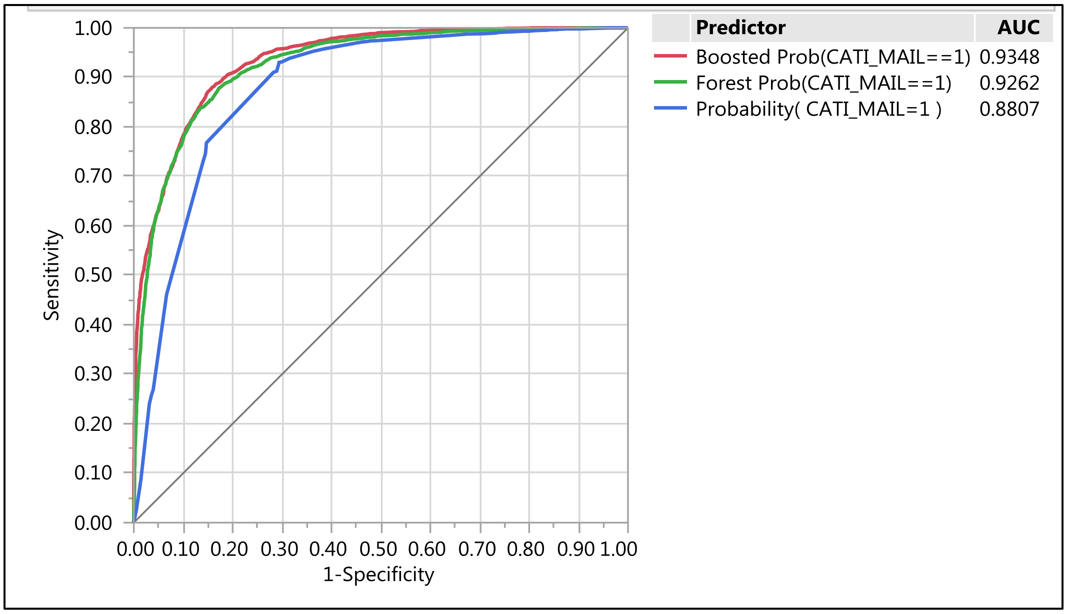
Table 3. Measures of fit for each of the three models considered to estimate propensity to respond via Mail or CATI



Measures of Fit for CATI_MAIL

Validation 2	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N	AUC
Training	Boosted Tree		0.5267	0.6906	0.3274	0.3135	0.2443	0.1305	47581	0.9495
Training	Bootstrap Forest		0.7547	0.8648	0.1697	0.1931	0.1441	0.0216	47581	0.9980
Training	Fit Generalized Logistic Regression		0.3850	0.5512	0.4254	0.3645	0.2707	0.1830	47581	0.8829
Validation	Boosted Tree		0.4807	0.6481	0.3587	0.3311	0.2641	0.1453	8319	0.9348
Validation	Bootstrap Forest		0.4536	0.6218	0.3774	0.3404	0.2747	0.1506	8319	0.9262
Validation	Fit Generalized Logistic Regression		0.3786	0.5439	0.4292	0.3667	0.2727	0.1872	8319	0.8807

Figure 2 below illustrates how the ROC curves from the validation sets of each model compare to each other. It is apparent in Figure 2, that the red line (Boosted Tree model) has the highest AUC at 0.9348.

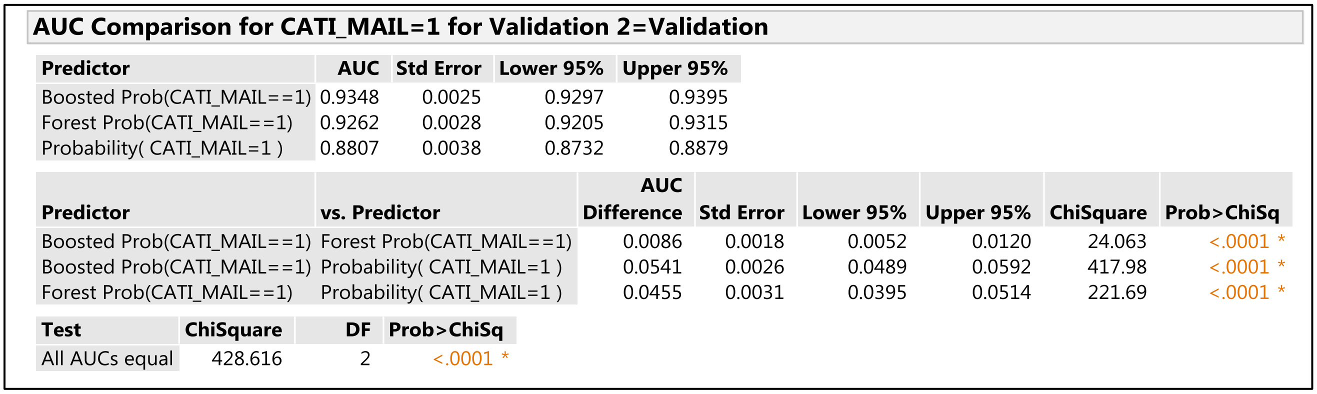


	Predictor	AUC
🟥	Boosted Prob(CATI_MAIL==1)	0.9348
🟩	Forest Prob(CATI_MAIL==1)	0.9262
🟦	Probability(CATI_MAIL=1)	0.8807

Figure 2. Illustrates the receiver operating curves for the three candidate CATI/Mail models

The significance testing of the AUC for each model in *Table 4* below provides evidence that each model performs differently. Furthermore, the results seen in *Table 4* show that the Boosted tree model outperformed both the Bootstrap Forest and Logistic Regression methods for predicting response via non field enumeration.

Table 4 AUC comparison for models predicting whether or not a record will respond via Mail or CATI



AUC Comparison for CATI_MAIL=1 for Validation 2=Validation

Predictor	AUC	Std Error	Lower 95%	Upper 95%
Boosted Prob(CATI_MAIL==1)	0.9348	0.0025	0.9297	0.9395
Forest Prob(CATI_MAIL==1)	0.9262	0.0028	0.9205	0.9315
Probability(CATI_MAIL=1)	0.8807	0.0038	0.8732	0.8879

Predictor	vs. Predictor	AUC Difference	Std Error	Lower 95%	Upper 95%	ChiSquare	Prob>ChiSq
Boosted Prob(CATI_MAIL==1)	Forest Prob(CATI_MAIL==1)	0.0086	0.0018	0.0052	0.0120	24.063	<.0001 *
Boosted Prob(CATI_MAIL==1)	Probability(CATI_MAIL=1)	0.0541	0.0026	0.0489	0.0592	417.98	<.0001 *
Forest Prob(CATI_MAIL==1)	Probability(CATI_MAIL=1)	0.0455	0.0031	0.0395	0.0514	221.69	<.0001 *

Test	ChiSquare	DF	Prob>ChiSq
All AUCs equal	428.616	2	<.0001 *

3.4 Application of Propensity Scores

Using the model output, we examined how the model *would* have predicted information for 2017 June Crops APS by comparing the prediction information to the final results.

Table 5. Sample & Model Frequencies

Sample Size (List frame)	*Model #1 (Most Likely Field)*	*Model #2 (Most Likely MAIL/CATI)*
69,722	5,483 (7.9%)	29,301 (42.0%)

As seen in *Table 5* above, almost half of the total sample were found as likely (>.5) to be completed via either by Field or Mail/CATI. Between the two models, there remained a small (140) overlap of records that showed a high likelihood of completion in both models. For those, field offices would be advised to use the least expensive mode (Mail/CATI).

Using the 5,483 records identified by the first model (likely field completion), we observed the data collection codes assigned to these records. *Table 2* simplifies these observations by whether or not a field interview code was assigned and if the survey was completed.

Table 6. DCMS Code Assignments vs. Final Disposition for Model #1

	Response	
Field Data Collection Code	Complete	Not Complete
Yes	2599 (73.34%)	945 (26.66%)
No	1198 (61.78%)	741 (38.22%)

Table 6 shows that records specified for field collection had a response rate of 73.34 percent. When data collection was anything other than field data collection, the

response rate was 61.78 percent. If all the records identified by Model 1 as likely field completions (n=5,483 were set to field data collection *and* assuming the response rate of these records was 73.34% as suggested in *Table 6,* the total number of complete records would equate to 4021 (5483*.7334). This is an increase of 224 records compared to methods without adherence to the propensity model. Although 221 completions are substantial, especially for reports such as county estimates or rare commodities, this is only a 0.3% increase in response rate relative to the entire June APS sample.

For propensity Model 2, some 29,301 records were identified that would likely (>0.5) complete the survey via mail or telephone. According to *Table 7* below, without adherence to the propensity model, a majority (~96%) of the records were already assigned to mail or CATI. The remaining 4% of records were sent directly to the field.

Table 7. DCMS Code Assignments vs. Final Disposition for Model #2

	Response	
MAIL/CATI Data Collection DCMS Code	Complete	Not Complete
Yes	20,001 (71.16%)	8,108 (28.84%)
No	807 (67.70%)	385 (32.30%)

Table 7 illustrates that if all survey respondents were set to the model's suggestion of mail or CATI, a higher response rate may have been observed. Again assuming a 71.16% response rate on the entire subset of likely mail or telephone records (n=29,301) was obtained, we would have an additional 43 completions. However, this result relative to the entire June 2017 sample, would provide only a slight bump to the response rate.

4. Conclusion

This research compared three different RPM approaches to obtain two well-fit models for expensive data collection procedures (field enumeration) and inexpensive data collection procedures (Mail/CATI). A Bootstrap Forest model proved most effective for field enumeration while a Boosted tree model proved most effective for Mail/CATI. Both models had high AUC scores (>0.90).

Initial investigation on how these models will increase response rates are minimal – as adherence to the models within observational data only increased response rates by a small percentage of the overall total. However, any response rate increase is viewed as beneficial as these processes will eventually be hardcoded in the background of data collections and require little to no additional time required by field office personnel.

In addition, these models could provide survey administrators a more accurate outlook when preplanning and allocating resources for expensive and inexpensive data collection procedures. Oftentimes, the decision to use expensive versus inexpensive data collection procedures is basely primarily on the money available at a given time. These models may add an increased understanding of what type of response rates we can expect when a certain percentage of funds are allocated to expensive and inexpensive modes of response.

However, a limitation of this research, especially when broaching the subject of cost-benefit in survey data collection, is the determination of impact or leverage a survey response has on the final estimate. Without incorporating impact into these propensity models by mode, this research remains incomplete. Moving forward, impact measures are being examined that can ultimately govern the use of propensity models in planning survey data collection.

Acknowledgements

This research was done with the support and guidance of numerous individuals within USDA NASS including the Response Rate Research Team, Linda Young, Valbona Bejleri, and Barbara Rater.

References

Earp, M., Mitchell, M., McCarthy, J., & Kreuter, F. (2014). “Modeling nonresponse in establishment surveys: Using an ensemble tree model to create nonresponse propensity scores and detect potential bias in an agricultural survey.” *Journal of Official Statistics*, *30*(4), 701-719.

Groves, Robert M., and Steven Heeringa. 2006. "Responsive design for household surveys: tools for actively controlling survey errors and costs." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(Part 3): 439-457.

Lavrakas, P., Jackson, M., McPhee, C. 2018. “The Use of Response Propensity Modeling (RPM) for Allocating Differential Survey Recruitment Strategies: Purpose, Rationale, and Implementation.” Survey Practice Vol. 11., Issue 2.

Phipps, P. and D. Toth. 2012. “Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model with Linked Administrative Data.” Annals of Applied Statistics 6: 772–794. DOI: http://dx.doi.org/10.1214/11-AOAS521.

Tourangeau, R., Brick, J M., Lohr, S. and J. Li. 2017. “Adaptive and responsive survey designs: a review and assessment.” *Journal of the Royal Statistical Society A, 180*(Part 1): 203–223.