Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Growing Regression Trees that Use Sampling Frame Covariates to Explore Response Burden for Use in Survey Design

by Yeng Xiong, Laura T. Bechtel, Colt S. Viehdorfer and Diane K. Willimack

Release date: October 29, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Growing Regression Trees that Use Sampling Frame Covariates to Explore Response Burden for Use in Survey Design

Yeng Xiong, Laura T. Bechtel, Colt S. Viehdorfer, and Diane K. Willimack [1]

Abstract

The Economic Directorate of the U.S. Census Bureau is developing coordinated design and sample selection procedures for the Annual Integrated Economic Survey. The unified sample will replace the directorate's existing practice of independently developing sampling frames and sampling procedures for a suite of separate annual surveys, which optimizes sample design features at the cost of increased response burden. Size attributes of business populations, e.g., revenues and employment, are highly skewed. A high percentage of companies operate in more than one industry. Therefore, many companies are sampled into multiple surveys compounding the response burden, especially for "medium sized" companies.

This component of response burden is reduced by selecting a single coordinated sample but will not be completely alleviated. Response burden is a function of several factors, including (1) questionnaire length and complexity, (2) accessibility of data, (3) expected number of repeated measures, and (4) frequency of collection. The sample design can have profound effects on the third and fourth factors. To help inform decisions about the integrated sample design, we use regression trees to identify covariates from the sampling frame that are related to response burden. Using historic frame and response data from four independently sampled surveys, we test a variety of algorithms, then grow regression trees that explain relationships between expected levels of response burden (as measured by response rate) and frame covariates common to more than one survey. We validate initial findings by cross-validation, examining results over time. Finally, we make recommendations on how to incorporate our robust findings into the coordinated sample design.

Key Words: response burden; establishment surveys; regression trees; coordinated sampling.

1. Introduction

Under recommendation from the National Academy of Sciences (National Academies of Sciences, Engineering, and Medicine, 2018), the Economic Directorate of the U.S. Census Bureau is forming an integrated (coordinated) sample selection system, named Annual Integrated Economic Survey (AIES), for six of the Directorate's annual surveys. The scope of each survey tends to focus on a specific industry of the economy, such as manufacturing or retail, but these surveys often have some overlapping content. The sampling frame for each survey is independently developed and maintained, which optimizes sample design features at the cost of increased response burden. A high percentage of companies operate in more than one industry, and they can be selected to receive multiple surveys. The AIES design will replace this inefficient process with a unified sample that factors in response burden. This research, aimed at exploring the relationship between response burden (as measured by response rate) and covariates available on the sampling frame for the annual surveys, is a small part of the AIES transformation and the first step in developing a burden metric for the sample redesign.

For this research we focus on four of the six annual surveys whose sampling procedures are being integrated together: Annual Survey of Manufactures (ASM), Annual Retail Trade Survey (ARTS), Annual Wholesale Trade Survey (AWTS), and Annual Capital Expenditures Survey (ACES). Responses to the surveys are required by law for sampled

[1]Yeng Xiong, U.S. Census Bureau, 4600 Silver Hill Rd, Washington, DC, U.S., 20233 (yeng.xiong@census.gov); Laura T. Bechtel, U.S. Census Bureau, (laura.bechtel@census.gov); Colt S. Viehdorfer, U.S. Census Bureau (colt.s.viehdorfer@census.gov); Diane K. Willimack, U.S. Census Bureau (diane.k.willimack@census.gov)

companies. ASM draws its sample from manufacturing establishments stratified by industry and questionnaire eligibility. Its sample is selected with probability proportional to receipts. ARTS, AWTS, and ACES are stratified simple random samples at the company level. The population of ARTS is specific to retail companies and AWTS to wholesale. Both ARTS and AWTS are stratified by industry and size. The companies sampled for ACES span all non-agricultural industries and are first split into employer and nonemployer frames. Employer units are then stratified by industry and size whereas nonemployer units are classified into one of four strata: sole proprietorships, corporations and partnerships, companies registered for employer identification number but had no employees/payroll, and companies with no employees/payroll but may have them by the data collection period. ASM[2], ARTS[3], and AWTS[4] are longitudinal surveys with a sample cycle of at least five years, and ACES[5] is a cross-sectional survey.

2. Response Burden

Response burden is believed to contribute to nonresponse and adversely affect data quality. U.S. legislation requires that potential respondents be informed of the expected response burden, defined in terms of an estimated amount of time to complete a survey (including reading instructions and gathering data). However, the research literature indicates several attributes of response burden beyond simply an amount of time. Willeboordse (1997) categorizes burden into four dimensions with two alternative aspects within each: objective vs. subjective, gross vs. net, maximalist vs. minimalist, and imposed vs. accepted. The dimensions provide context to aid examination and evaluation of burden. Time, a common measure of burden, falls under objective, which measures response burden with calculable quantities, whereas subjective refers to perceived burden, which is experiential, *e.g.*, how easy or difficult did respondents find the survey. Gross considers the cost of burden, while net takes account of offsetting benefits of survey response. Maximalist burden considers the additional steps taken when filling out a questionnaire such as record formation, preparation tasks, and post-collection re-contacts; minimalist only acknowledges questionnaire completion. Imposed burden supposes that all units in the sample experience burden. Accepted assumes that respondents have consented to and taken on the burden. For this research, burden is defined as objective, gross, maximalist, and imposed.

Drawing on this definition of burden, the research literature identifies several factors associated with objective burden and/or perceived burden (Dale and Haraldsen, 2007; Giesen, et.al., 2018). These factors are categorized based on whether they account for motivation, which may be associated with perceived benefits of survey response; data collection procedures, which are usually considered to be associated with the cost side; or organizational complexity, also assumed to be positively correlated with burden. A frame covariate—covariate that is accessible during sampling—is then connected, if possible, to one of the factors. Motivating factors that can impact response burden may include whether the respondent generally trusts the U.S. Census Bureau and whether the respondent finds the economic statistics produced to be useful. However, frame covariates that can be empirically associated with these motivating factors are unknown. In fact, determining motivating factors generally requires additional follow up that may or may not be useful until the next sample selection.

Characteristics of the questionnaire itself have been shown to be associated with response burden (Haraldsen and Jones, 2007). Therefore, having information about the questionnaire (*e.g.*, number of questions, availability of data in records, readability, layout, collection mode, etc.) may be helpful. This information, though, is not on the frame; they are metadata generated from the questionnaire and may not be easily defined. Consider the number of questions in a survey. A survey can adjust the number of questions shown to a unit depending on its answer to certain questions. That is, some units can have fewer questions than others answering the same survey. Timing and reporting routines are also contributing factors. Depending on the time of year a questionnaire is administered, companies may have other reporting responsibilities that may prevent them from responding. If they have multiple surveys or limited personnel to handle the request, they may not respond. These types of metadata are not included in this analysis but may be considered in future analysis.

---

[2] https://www.census.gov/programs-surveys/asm/technical-documentation/methodology.html

[3] https://www.census.gov/programs-surveys/arts/technical-documentation/methodology.html

[4] https://www.census.gov/programs-surveys/awts/technical-documentation/methodology.html

[5] https://www.census.gov/programs-surveys/aces/technical-documentation/methodology.html

While sampling frames typically lack covariates reflecting motivation and data collection, they do have several considered to be associated with the organizational complexity of a unit, believed to play a substantive role in the business survey response process (Willimack and Nichols, 2010; Bavdaz, 2010) and contribute to response burden. For instance, as businesses grow, their structure may become more complex through adding new locations, operating in multiple industries, and dividing into departments, perhaps even adapting or redefining their legal arrangement. Data is then further distributed throughout the organization, and burden will consequently increase because respondents may need to consult multiple sources to answer survey questions. Yet they develop reporting routines to minimize their own response burden. Acknowledging that the relationships of these variables with one another is hypothetical, as is their association with response burden, these factors and associated sampling frame covariates are defined in Table 2-1.

Table 2-1
Frame covariates used as explanatory variables and corresponding factors

Factor	Frame Covariate	Covariate Description	Available for Surveys
Organizational structure	TYPE	Indicator of units having one establishment (SU) or more (MU)	ASM, ARTS, AWTS, ACES
	STRATUM	Factors of the sampling stratum unit is in[6]	ACES
	LFO	Factors of the legal form of organization	ACES
Industry diversity	NO_NAICS	Number of industries company operates in	ASM, ACES
Size	NO_ESTAB	Number of establishments in company	ASM, AWTS
	EMP_TYPE	Factor of employer and nonemployer	ACES
	PAYROLL	Annual payroll for company (dollar)	ASM, ACES
	PAYROLL_ESTAB	Annual payroll for establishment (dollar)	ASM
	RECEIPT	Annual receipts for company (dollar)	ARTS, AWTS
	CERT	Indicator of units being sampled with certainty (*i.e.*, selection probability = 1)	ASM, ARTS, AWTS, ACES
Reporting Routines	INPREVSAMP	Indicator of unit being sampled before	ACES

3. Methodology

Response is the dependent variable and acts as a proxy for response burden. This assumes that respondents have accepted the burden and nonrespondents are too burdened to respond. To explore the relationship between response burden and the frame covariates in Table 2-1, a regression tree is formed from a series of splits that grow branches and nodes based on adding successive covariates most associated with response. The tree will stop growing once a complexity parameter is reached or no improvement can be made to minimize the error between response and response rate (Breiman et. al., 1983). The final nodes in a tree, referred to as terminal nodes, represent units with homogeneous response patterns. The nodes with lower response rates are assumed to reflect higher relative burden.

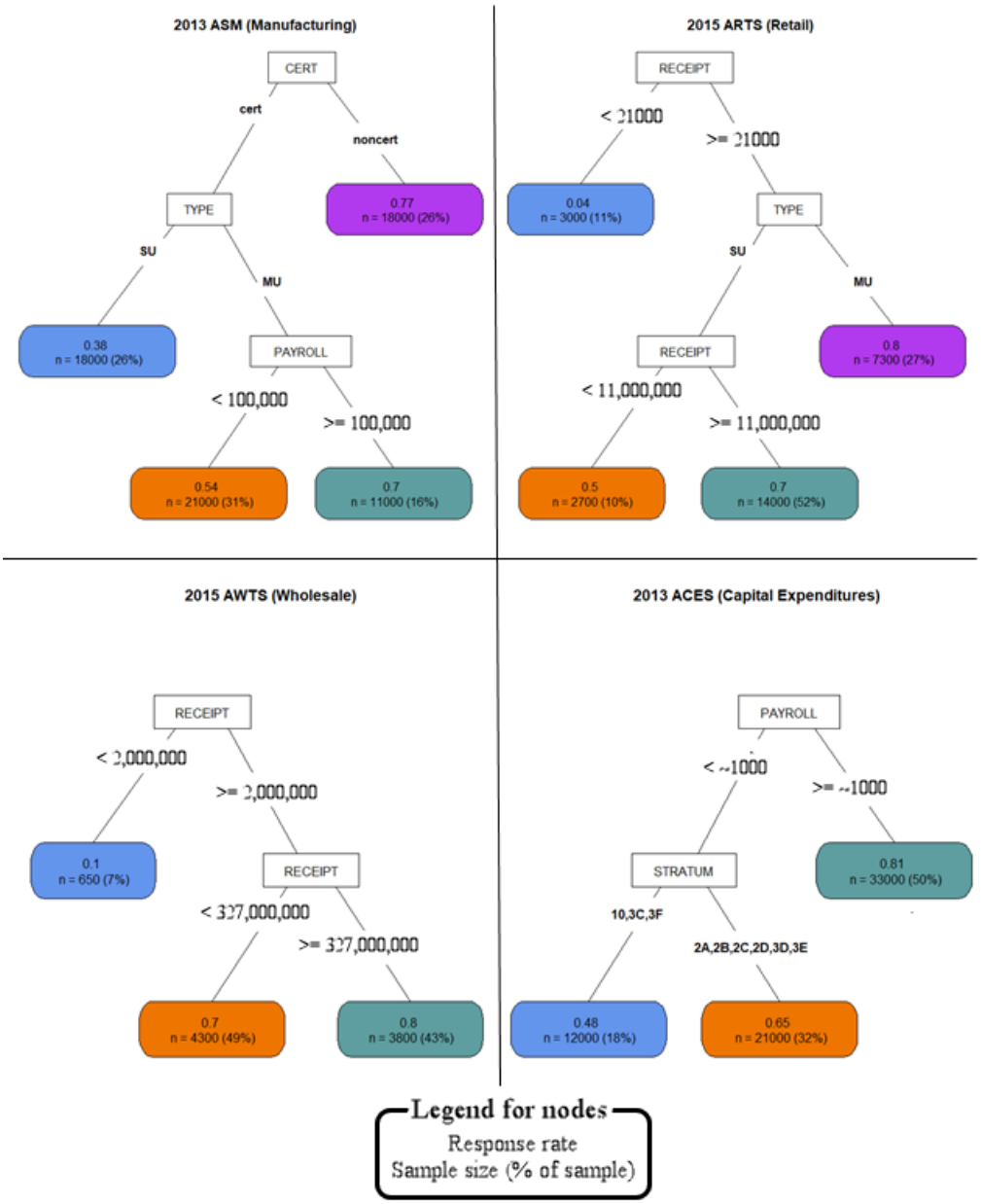
The R package, *rpart*, is used for developing the regression trees (Therneau and Atkinson, 2019). Tree growth is not limited for any of the surveys, and the trees are unweighted since response burden only pertains to the sampled units. For ASM and ACES, data from 2013 to 2019 are used for the analysis. ASM data is collected as part of the Economic Census for years ending in 2 or 7. This is the case with the 2017 ASM data presented here. The 2015 to 2019 data are available for ARTS and AWTS. For this research, the regression trees analyze response burden within a survey and not across surveys. With our longitudinal surveys ASM, ARTS, and AWTS, a base regression tree is grown with the data from the first year, which is the last year of a sample cycle, and evaluated on the subsequent years when a new sample is selected (ASM 2019 is the start of another new sample). For ACES, which has a new sample selected every year, the 2013 data is used to grow a base regression tree and the 2014-2019 data are used for evaluation (ACES 2018 used ACES 2017 employer data).

[6] ACES employer units are stratified by industry and size; ACES nonemployer units are stratified based on legal form of organization.

4. Results

The base regression trees are plotted in Figure 4-1. The colored nodes at the bottom of a regression tree correspond to the expected burden levels of units with the characteristics defined by the covariates. For ASM, the first split is on the certainty covariate. The certainty units (units that must be included in sample) are more burdened than the noncertainty units. The certainty units can be diverse in terms of size and structure. This results in certainty units being further divided by the covariates type and payroll. The low response rate observed among the certainty single units implies they are the most burdened. The certainty multi units are "medium" burdened with the smaller units, in terms of payroll, more burdened than the larger units. The least burdened are the noncertainty units.

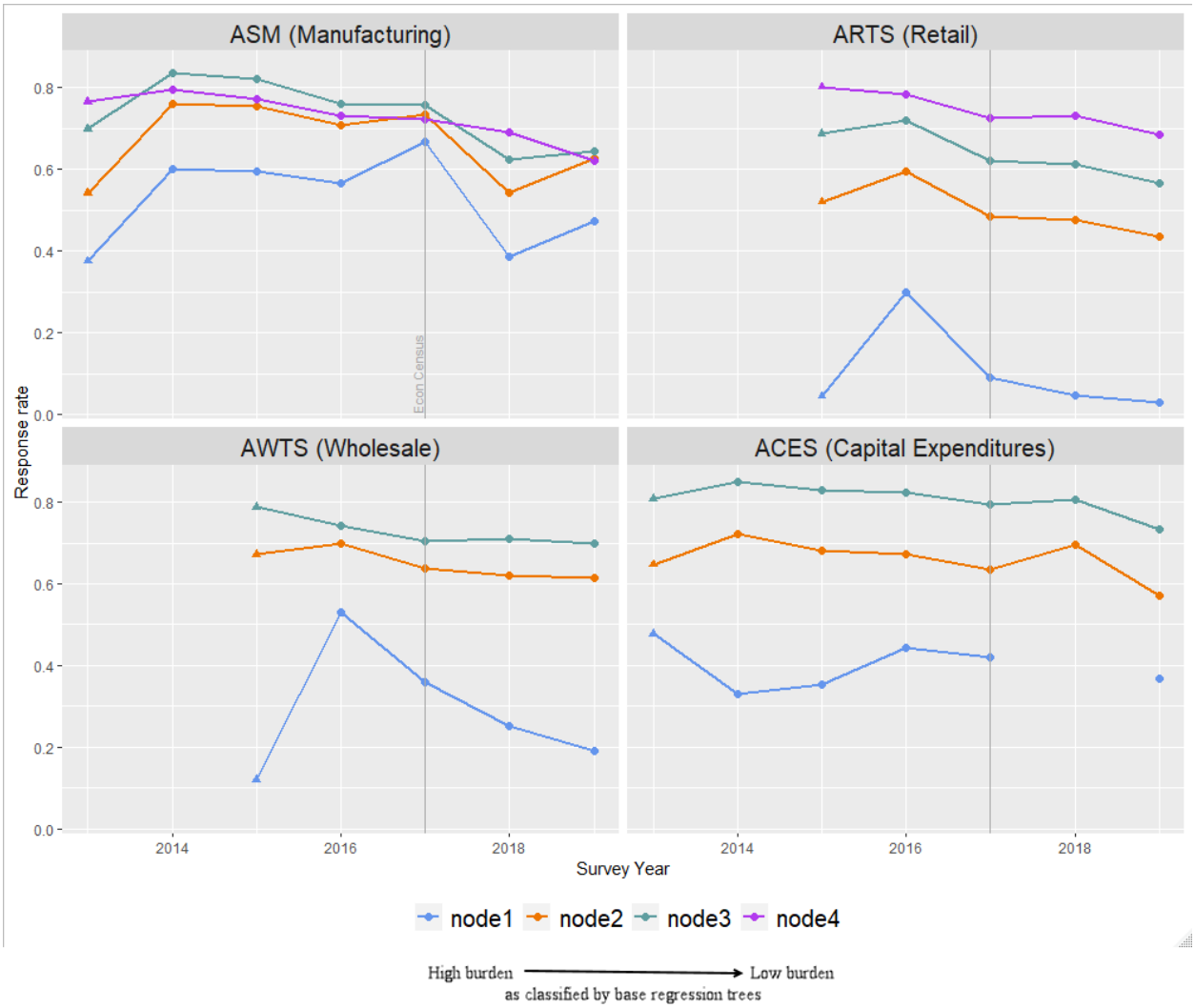
Figure 4-1
Base regression trees characterizing different levels of burden, by survey
(Data Source: ASM 2013, ARTS 2015, AWTS 2015, ACES 2013)



A similar story is seen in the ARTS regression tree in which the size and organizational structure are the key covariates for explaining burden. The smallest units are the most burdened, followed by the larger single units, who are again partitioned by size. The largest multi-units are the least burdened. The AWTS tree is dominated by size. The smaller units continue to be more burdened than the larger units.

The first split in the ACES tree is based on size with smaller units being more burdened. The smaller units are then subdivided into two groups. The first group consists of the small certainties (stratum 10) and the start-up companies (strata 3C,3F). This seems an unusual pairing, but upon further consideration, it makes sense that they would be the most burdened. Smaller units typically do not have the resources and personnel to handle a survey request and being a certainty will compound that burden. As a result, they are placed into the same class as the start-ups. In the second group are the smaller noncertainty units (strata 2A-2D) and the smaller nonemployer units such as corporations and partnerships (stratum 3D) and sole proprietorships (stratum 3E). The least burdened are the larger units.

Figure 4-2
Response rates of groups with different burden levels identified by regression trees over time, by survey
(Data Source: ASM 2013-2019, ARTS 2015-2019, AWTS 2015-2019, ACES 2013-2019)



Using the terminal nodes identified by the base trees in Figure 4-1, Figure 4-2 charts the response rates for these groupings over time. While the numerical values of the response rates are changing, the rank order of the rates, which

is color-coded to denote a relative burden level as classified by its base regression tree, is generally consistent over time. This pattern is common to all the surveys. An exception to this is the ASM purple node, which represents the noncertainty manufacturing units. The response burden is expected to be different between the certainty and noncertainty units, and this may be providing more evidence of that. In the future, it may be necessary to analyze the noncertainty units separately from the certainty units.

5. Conclusions

In this analysis we demonstrate that the regression tree analysis can discern different levels of response, serving as proxies for burden levels, that possibly holds up over time. Our empirical results also show a relationship between response burden and size that is prevalent across surveys. This is not surprising; larger units often have dedicated personnel to fill out surveys from federal government agencies whereas smaller units lack this resource. However, though not empirically detected in the current analysis, the relationship between response burden and size is not always so linear. Smaller units are less likely to be sampled for multiple surveys and reporting is usually more straight-forward for them. Incorporating additional metadata, such as those discussed in Section 2, that reflects the data collection procedures may help to better describe the complex relationship between response burden and size. Another limitation is using response rate as an indicator of survey burden. The high response rate of larger units is procedurally inflated because there is generally a more stringent follow up routine with them than with smaller units. Supplementing response with information about respondent behavior (e.g., response time, if an extension is requested) may help to better realize burden. For future research, we are looking to develop a procedure, such as a standardized burden metric based on results from this analysis, that can be used to account for response burden in AIES when coordinating sampling across the Economic Directorate's annual surveys.

Acknowledgements

The authors thank Alfred Dave Tuttle, James Hunt, Ian Thomas, Valerie Mastalski, Steven Roman, Susan Pozzanghera, Jeremy Knutson, Joseph Barth, and Magdalena Ramos, for their careful review and constructive comments on earlier versions of this paper.

References

Bavdaz, M. (2010), "The multidimensional integral business survey response model", *Survey Methodology*, 36, pp. 81-93.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. (1983), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Dale, T. and G. Haraldsen (eds.). (2007), *Handbook for Monitoring and Evaluating Business Response Burdens*, Luxemburg: Eurostat.

Giesen, D., M. Vella, C. F. Brady, P. Brown, D. Ravindra, and A. Vaasen-Otten. (2018), "Response Burden Management for Establishment Surveys at Four National Statistical Institutes", *Journal of Official Statistics*, 34, pp. 397-418.

Haraldsen, G. and J. Jones. (2007), "Paper and Web Questionnaires Seen from the Business Respondent's Perspective", *Proceedings of the Third International Conference on Establishment Surveys, American Statistical Association*, pp. 1040-1047.

National Academies of Sciences, Engineering, and Medicine. (2018), *Reengineering the Census Bureau's Annual Economic Surveys*, Washington, DC: The National Academies Press, consensus study available at https://www.nap.edu/read/25098/chapter/1.

Therneau, T. M. and E. J. Atkinson. (2019), “An Introduction to Recursive Partitioning using the RPART Routines”, unpublished report available at http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

Willeboordse, A. (1997), “Minimizing Response Burden”, in A. Willeboordes (eds.) *Handbook on Design and Implementation of Business Surveys*, Luxembourg: Eurostat, pp. 111-118.

Willimack, D. K. and E. Nichols. (2010), “A Hybrid Response Process Model for Business Surveys”, *Journal of Official Statistics*, 26, pp. 3-24.