Developing a Student Contest for the ICES-V Conference

Katherine Jenny Thompson[1]

1. Introduction

As the fifth in the series of international conferences on establishment surveys, ICES-V is designed to look at key issues and challenges pertaining to establishment surveys. This conference introduced our first student contest: an Imputation /Missing Data Treatment Contest. Our objective in establishing the contest was to create interest and motivate innovation in the establishment survey field by inspiring students and the faculty they work with. Participants were provided with two simulated datasets that were incomplete due to unit nonresponse and were challenged to complete both data sets using some form of imputation. Submissions were judged on a variety of factors, including theoretical soundness, originality and effectiveness of methods, and clarity of explanation.

The contest has a story. It was born in June 2013 during a Program Committee (PC) meeting as part of a discussion on how the program could incorporate alternative ways of enticing new audiences to the conference. The Program Chair suggested a student-oriented contest as a possible venue.

Initially, the PC was skeptical. A high proportion of academic survey research focuses on household surveys. Part of this is due to the accessibility of demographic datasets. Realistic public use business datasets can be very difficult to find, largely due to confidentiality concerns. Business data populations are highly skewed, and the majority of a tabulated total in a given industry often comes from a small number of large businesses. Consequently, the risk of disclosure of confidential data can be quite high. Confidentiality concerns increase with the number of collected data items. At the same time, data users are most interested in studying the relationships between these items, e.g. building micro-economic models. It is therefore important to perturb these data so that the risk of disclosure is limited while retaining the usability of the data for a variety of analyses.

In the past decade, there has been great progress in developing and implementing disclosure avoidance methods such as noise infusion and synthetic data[2]. And there are limited datasets created from publicly released data such as the Census of Governments conducted by the U.S. Census Bureau. Nevertheless, it is fair to say that these datasets are not always easy to find. It is also fair to say that it is not always easy to produce a "shareable" synthetic or noise-infused dataset from official business statistics, as (1) the burden of proof of protection lies with the dataset creator, and this proof is rigorously

[1]Katherine Jenny Thompson, U.S. Census Bureau, ESMD HQ-35H175, 4600 Silver Hill Road, Washington DC 20233, email: Katherine.J.Thompson@census.gov. Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau.

[2] See Massell, Zayatz, and Funk (2006) and Abowd et al (2012) for a few examples of noise infusion and Kim, Karr,and Rieter (2015) and Dreschler (2012) for a few examples of synthetic business data applications. Note that this is a very limited bibliography and is just offered as to give a flavor of the field.

overseen internally (and on occasion by other agencies) and (2) sacrifices in quality at least for selected items often need to be made in the process.

This is a bit of a roundabout way of mentioning that I "just happened" to have two public use datasets[3] that could be used for such a contest, and that I "just happened" to mention this to the PC.

The possibility of developing an interesting and challenging student contest became much more feasible given the happy coincidence of readily available datasets. Various ideas were floated, with the most popular being an imputation or outlier detection challenge. A small subcommittee was tasked, comprising Pierre Lavallée (Statistics Canada), Michael Sinclair (Mathematica Policy Institute), and myself (U.S. Census Bureau); hereafter, I refer to our trio as the Contest Development Subcommittee. After internal discussion, we decided to incorporate a complex survey design into the contest as well as imputation, providing *samples* – not populations – with missing data and adding requirements on estimated reliability measures. Although this complicates the challenge, it made for a very realistic problem.

The contest designers hold a major advantage over any contest participant: we know the truth. Section 2 shares details on the development of the contest data, so that we will no longer be the sole proprietors of this knowledge. The contest template was developed in parallel with the contest data, and Section 3 discusses the process. I conclude with congratulations and thanks.

2. Developing the Data for the Contest

The datasets were modeled from two industries (Industry *XXX1* and Industry *XXX2*) surveyed by the Monthly Retail Trade Survey (MRTS) conducted by the U.S. Census Bureau. Each represents a multivariate population of sales and inventories from a single industry. Mulry, Oliver, and Kaputa (2014) provide more comprehensive details on the simulation methodology, briefly summarized below.

The population data were generated using the SIMDAT algorithm (Thompson 2000) with modeling cells equal to MRTS sampling strata and population size equal to the original frame size in each stratum. The SIMDAT algorithm is a nonparametric resampling algorithm that randomly generates synthesized multivariate observations from each unit plus m nearest neighbors; in this application, random sampling was performed with replacement to create a synthetic population from the sampled units. The application of the algorithm was slightly modified to satisfy disclosure avoidance criteria provided by the U.S. Census Bureau's Disclosure Review Board (DRB) and the Internal Revenue Service.

The original datasets contained observations from one point in time. However, we needed at least three separate measurements for each observation: one to use as a stratifying

[3] Being honest, this was not a coincidence. These datasets were originally generated for an internal research project but were subsequently approved for external release. Since then, these datasets have been quite useful for many outside-agency collaborations. For example, Andridge and Thompson (2015) used them to develop and test the gamma-formulation proxy-pattern mixture model software and Yang and Kim (2015) use them in a fractional imputation simulation study.

variable, one to use as the current period value (subject to nonresponse), and one to use as a possible predictor (prior value, not subject to nonresponse). For this, we generated a stationary series of length 20 for each observation within stratum (both variables) via the AR(1) model given by

$$y_{hi,t} - m_{hi} = \Phi_h\left(y_{h,t-1} - m_{hi}\right) + a_{hi,t}$$

where *h* indexes stratum, *i* indexes unit within stratum, and *t* indexes time (month) and

$y_{hi,t} - m_{hi} = 0$ and m_{hi} is the series mean
$a_{hi,t} \sim N(0, \sigma_h^2)$ is a white noise process
Φ_h = the sample-based estimate of lag one autocorrelation in the MRTS industry stratum

After generating the unit-level stationary series, we added the unit's initial value to each observation. When the process was complete, the original datasets were augmented with 19 additional pairs of observations. Of course the further the projection in the series, the more variable – and less realistic – the data. We designated the first observation of sales as the frame measure of size (MOS) and selected the 12th simulated month in the series as the current period data, mimicking a typical survey where there is a lag between the sample selection and its field implementation. By default, the 11th simulated month contained the prior period values. Table 1 presents the population data provided to the contestants.

Table 1: Population Data Characteristics ("Truth")

Industry *XXX1*			Industry *XXX2*		
Sales00	Total	48,346,053,043	Sales00	Total	1,677,378,977
	Mean	2,304,717		Mean	118,988
	Variance	4.38E+13		Variance	2.93134E+11
	Skewness	49.67		Skewness	55.94
Inventories00	Total	100,851,062,160	Inventories00	Total	1981167030
	Mean	4,807,697		Mean	140,538
	Variance	1.19E+14		Variance	1.63E+12
	Skewness	39.02		Skewness	63.82
Sales00 and Inventories00	Correlation	0.97	Sales00 and Inventories00	Correlation	0.75

Next, we applied the Lavallée-Hidiroglou stratification algorithm (Lavallée and Hidiroglou 1988) to each industry population to obtain six sampling strata per industry, with Stratum 6 containing certainty units (sampled with probability one). Sample sizes were determined by applying Neyman allocations on the frame MOS (sales) with a c.v. constraint of 0.01. Finally, we used PROC SURVEYSELECT® (SAS/STAT User's Guide, Second Edition) to select stratified simple random without replacement samples in each industry. This one-stage sample design is implemented by many of the business surveys conducted at the U.S. Census Bureau.

An imputation problem requires (1) missing data and (2) covariates for imputation. For simplicity, we decided to limit the contest data to unit nonresponse. Our response model

was designed so that larger units were more likely to respond: specifically, the certainty units were very likely to respond (Stratum 6) and the smallest units were the least likely to respond (Stratum 1). The response mechanisms differed by strata and were combinations of ignorable mechanisms (increasing in frame MOS) and missing not at random (MNAR) mechanisms (directly dependent on collected item). Certainty and large-unit size strata were more likely to have an ignorable response mechanism, whereas the small-unit size strata were more likely to have a non-ignorable response mechanism.

This response model is consistent with the literature. Because of their important contribution to survey totals, survey operation procedures are generally designed to increase the likelihood of obtaining valid responses from large businesses (Thompson and Oliver 2012), and research on collection methods and contact strategies has likewise been largely confined to obtaining accurate reported data from large businesses; see Thompson, Oliver, and Beck (2015) for a more comprehensive bibliography. Consequently, missingness is often less prevalent in large businesses survey data. At the other end of the spectrum, response burden can be quite high for smaller businesses, especially if the survey collects many items or has complex survey instruments (Thompson and Washington 2013 and Willimack and Nichols 2010). Equally important, they may perceive the burden of responding to the survey as a whole as being too high (Bavdaž 2010). This in turn leads to high levels of missingness.

Following Andridge and Thompson (2015), we randomly induced missingness in the sampled datasets with probability according to a logistic regression model

$$\text{logit}(\Pr(M = 1|Y, Z)) = \gamma_0 + \gamma_Z Z + \gamma_Y Y$$

with M = 1 indicating unit nonresponse, Y indicating the outcome variable (sales or inventories), and Z indicating the frame MOS. If γ_Z=0 in the model, then the response mechanism is not ignorable (MNAR); if γ_Y=0, then the response mechanism is ignorable (covariate dependent). We set $\gamma_0 = \log(p_h/q_h)$, where p_h is the targeted stratum response rate and estimated the other regression parameters from the sampled data. Final response propensities and response mechanisms are reported in Table 2.

The high correlation between MOS and sales makes it difficult to distinguish between an ignorable and nonignorable response mechanism in the affected strata; the distinction is clearer when inventories was used, but we limited these strata. For the record, every contest participant concluded that the data exhibited an ignorable response mechanism, with most concluding some form of missing-at-random. In this case, none of the participants guessed the truth.

Table 2: Final Response Propensities and Response Mechanisms in the Simulated Contest Data

Industry	Stratum	Response Mechanism	Independent Variable in Reg. Model	Targeted Unit Response Rate	Observed Unit Response Rate
XXX1	Overall			0.70	0.71
	Stratum1	MNAR	Sales00	0.48	0.50
	Stratum2	Ignorable	MOS	0.63	0.61
	Stratum3	MNAR	Inventories	0.77	0.76
	Stratum4	MNAR	Sales00	0.85	0.77
	Stratum5	MNAR	Sales00	0.88	0.89
	Stratum6	Ignorable	MOS	0.92	0.91
XXX2	Overall			0.70	0.71
	Stratum1	MNAR	Sales00	0.52	0.50
	Stratum2	MNAR	Inventories00	0.75	0.72
	Stratum3	Ignorable	MOS	0.77	0.76
	Stratum4	MNAR	Inventories00	0.60	0.58
	Stratum5	Ignorable	MOS	0.80	0.77
	Stratum6	MNAR	Sales00	0.90	0.89

Lastly, we created additional covariates to use for imputation besides prior period values. Administrative data are available for sales from the Business Register maintained at the U.S. Census Bureau and are often used in imputation. Our simulated administrative data values for inventories were designed to be highly correlated with the survey data, whereas the sales administrative data were less so. For both variables, the correspondence between administrative data value and reported data values are very high for the largest (certainty) units and for the smallest noncertainty (stratum 1) units with increasing variability in between the two extremes. Table 3 summarizes the correlations within stratum for each data item with its administrative data counterpart.

Table 3: Correlation within Stratum for Survey Data Item and Administrative Data Item

Industry	Stratum	Sales	Inventories	Industry	Stratum	Sales	Inventories
XXX1	1	0.80	0.83	XXX2	1	0.69	0.76
	2	0.43	0.56		2	0.59	0.80
	3	0.45	0.59		3	0.57	0.87
	4	0.26	0.62		4	0.36	0.76
	5	0.49	0.55		5	0.49	0.70
	6	1.00	0.97		6	1.00	1.00

For the certainty strata units, we generated administrative data values by adding a large random error (multiplied by a scaling factor) to the survey value. We generated administrative data values for the noncertainty units as

$$y_v^a = f_v\left[\beta y_v + \delta_0 e^{\delta_1 y_v^g} \in_1 + \in_2\right]$$

where y_v is the survey value for item v, β is a gamma-distributed random variable, δ_0 describes the variability when y_v is small, δ_0 and g control the onset and magnitude of y_v's influence on the variance, $\in_1$ and $\in_2$ are normally distributed errors, and f_v is a scaling factor. This model is a variation of the one developed by Steel and Fay (1995),

dropping an overall mean term from the prediction, incorporating a scaling factor not needed in the original model which predicted receipts from payroll, and utilizing normally distributed errors instead of gamma distributed errors. Overall, the modeled data behaved as expected. Unfortunately, there was a single negative value in one modeled dataset that was missed, despite our numerous quality checks.

That small error aside (and it can easily be corrected by taking the absolute value of the observation, setting it to a missing value, or substituting the corresponding survey data value), these simulated datasets are available for public use. Given the outstanding presentations shared at the ICES-V conference that resulted from examining these data, they are clearly useful and we hope that others will use them for their own research.

3. Developing the Contest Template

The final contest template is available at http://www.portal-stat.admin.ch/ices5/imputation-contest/. Its evolution is best described by the simple diagram presented in Figure 1.

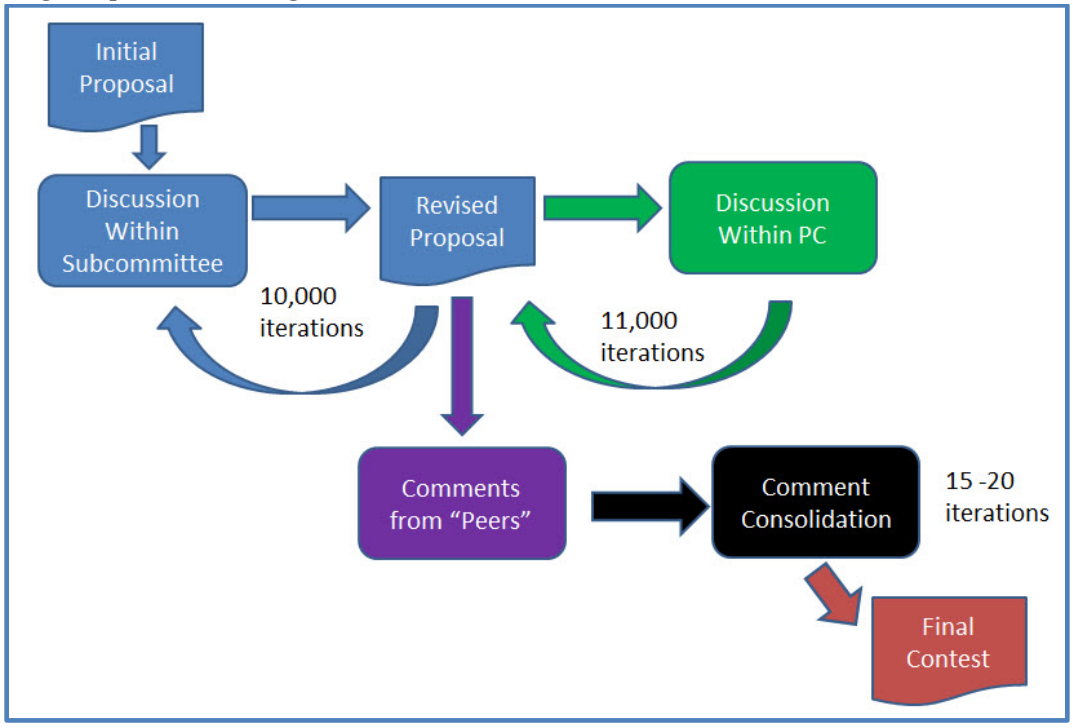

Figure 1: Artistic Rendering of Contest Template Development Process

More prosaically, the Contest Development Subcommittee drafted the first proposal. The development process engendered spirited dialogue, less about the form of the contest than the criteria for evaluating/judging the submissions. Actually, there was nearly complete consensus on the challenge itself. We agreed that we needed to prepare the datasets, "poke holes" in the datasets, and provide criteria for the imputed datasets (e.g., population totals, reliability constraints, preferred microdata properties). No one disliked the originally proposed criteria for the imputed data – and they never changed over any iteration. However, there were three very strong opinions on the criteria for judging the submissions. Each opinion was well developed and rational. Each was different.

Eventually after several rounds of revision, we decided to share the union of our criteria with the PC.

The PC reviewed the original proposal and provided useful comments. However, an eligibility criterion for the contest was to be a current undergraduate or graduate student at any level or a recent graduate that has held a degree for fifteen months or less as of June 2016. Sadly, the PC membership did not include anyone who could even pretend to claim that they had recently graduated, and we were concerned about a lack of empathy on our part. Collectively, we assembled a list of "peers" to help vet the contest proposal. These peers included colleagues in academia (our intended project sponsors), recent graduate school graduates, and – on occasion – coworkers that might share the same opinion as a subcommittee member.

This final review was especially helpful, providing perspectives completely lacking in the PC. It also led to a major change in our evaluation criteria. Originally, we planned to have each participant provide a fully imputed dataset, whose values would be checked against our "truth." Several of the Bayesian reviewers were very uncomfortable with that criterion, arguing that it essentially eliminated multiple imputation options. Consequently, we dropped the criterion, under protest from one subcommittee member who felt that the deletion would greatly complicate the submission judging process[4].

After a few editorial changes, the contest was released. As a group, we hoped that the contest problem would be sufficiently interesting to attract participants, although we privately admitted that none of us would have attempted the challenge ourselves while attending graduate school even with the enticement of a free trip to Switzerland.

4. Congratulations and Acknowledgements

At the risk of bragging, the Contest Development Subcommittee did an excellent job, if success is measured by (1) interest in the contest, (2) participation in the contest, or (3) quality of submissions. We were delighted to receive several submissions, all excellent. Initially, we hoped to save some efforts on behalf of our outside panel of judges by eliminating any entries that either violated the eligibility criteria or were especially lacking on any of the listed criteria for judging the contest. We could not eliminate any of the contest entries on either count.

As I've already stated, there were several people who provided invaluable aid in developing this contest. First, I thank my fellow subcommittee members for all of their efforts in the contest development and execution, crediting them with any innovative or coherent ideas. The collaboration was a pleasure and a learning experience, and I am grateful for our lively and enlightening discussions. Next, I thank Dr. David Haziza (Département de mathématiques et de statistique, Université de Montréal) and Dr. Hang J. Kim (Department of Mathematical Science, University of Cincinnati) for their careful review and insightful comments on each report. Their contributions were essential in ensuring the success and fairness of the contest. Note that Dr. Haziza and Dr. Kim provided independent reports and were each allotted a single vote, whereas our Contest

[4] He was probably right, but I am still personally grateful that I didn't have to check all of the datasets!

Development Subcommittee generated a consolidated review and shared a single vote. That said there was complete consensus among all the contest judges.

Special thanks are due to Laura Bechtel for her programming work and to Carma Hogue for creating a document template for the contest and for performing thorough editorial and content review. Of course, I am grateful to the ICES-V Program Committee and Organizing Committee for creating the opportunity for the contest as well as their useful suggestions and continued support. Lastly, I thank the Survey Methods Research Section and the Government Statistics Section of ASA for providing funds to support the contest.

Finally, I thank all of the students who participated. Selecting a single contest winner from a pool of outstanding reports was a difficult task, and I am grateful that we did not have to do it. The contest winners of the first student contest in the ICES series are listed below. I expect that readers will join me in the congratulations and enjoy three very different and clever solutions to the same problem.

Place	Author	Title
1st Place	Danhyang Lee Department of Statistics Iowa State University	Multivariate Regression Imputation Approach to the Analysis of Item Nonresponse in a Retail Trade Survey Data
2nd Place	Zhonglei Wang and Hejian Sang* Department of Statistics Iowa State University	Nonparametric Bootstrap to Generate Synthetic Population to Handle Complex Missing Data Problems
Honorable Mention	Julien Miron* and Audrey-Anne Vallée Institut de Statistique Université de Neuchâtel	Imputation Procedure and Inference In Presence of Imputed Data: Application To Industries

*Presenters

Acknowledgements

I thank Carol Caldwell, Carma Hogue, and Polly Phipps for their careful review of earlier versions of this manuscript.

References

Abowd, J.M., Gittings, R.K., McKinney, K.L., Stephens, B., Vilhuber, L. and Woodcock, S.D., 2012. Dynamically Consistent Noise Infusion and Partially Synthetic Data as Confidentiality Protection Measures For Related Time Series. *US Census Bureau Center for Economic Studies Paper No. CES-WP-12-13*.

Andridge, R. and Thompson, K.J. 2015. Assessing Nonresponse Bias in a Business Survey: Proxy Pattern-Mixture Analysis For Skewed Data. *The Annals of Applied Statistics* 9(4): 2237–2265. DOI: 10.1214/15-AOAS878.

Bavdaž, M. 2010. The Multidimensional Integral Business Survey Response Model. *Survey Methodology* 36:81–93.

Dreschler, J. 2012. New Data Dissemination Approaches in Old Europe – Synthetic Datasets for a German Establishment Survey. *Journal of Applied Statistics* 39(2): 243-265.

Lavallée, P. and Hidiroglou, M. 1988. On the Stratification of Skewed Populations. *Survey Methodology* 14: 33-43.

Kim, H., Karr, A. and Reiter, J. 2015. Statistical Disclosure Limitation in the Presence of Edit Rules. *Journal of Official Statistics* 31 (q): http://dx.doi.org/10.1515/JOS-2015-0006.

Massell, P., Zayatz, L., and Funk, J. 2006. Protecting the Confidentiality of Survey Tabular Data by Adding Noise to the Underlying Microdata: Application to the Commodity Flow Survey. Privacy in Statistical Databases. New York: Springer.

Mulry, M., Oliver, B., and Kaputa, S. 2014. Detecting and Treating Verified Influential Values in a Monthly Retail Trade Survey. *Journal of Official Statistics* 30: 721-747.

SAS/STAT(R) 9.2 User's Guide, Second Edition. Retrieved April 14, 2016, from https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#mixed_toc.htm.

Steel P. and Fay, R.E. 1995. Variance Estimation for Finite Populations with Imputed Data. *Proceedings of the Survey Research Section*, American Statistical Association, Alexandria, VA.

Thompson, J.R. 2000. Simulation: A Modeler's Approach. New York: John Wiley & Sons, 87-110.

Thompson, K.J. and Oliver, B. 2012. Response Rates in Business Surveys: Going Beyond the Usual Performance Measure. *Journal of Official Statistics* 28:221–37.

Thompson, K.J., Oliver, B., and Beck, J. 2015. An Analysis of the Mixed Collection Modes for Two Business Surveys Conducted by the U.S. Census Bureau. *The Public Opinion Quarterly* 79 (3): 769-789. doi:10.1093/poq/nfv013.

Thompson, K.J. and Washington, K.T. 2013. Challenges in the Treatment of Unit Nonresponse for Selected Business Surveys: A Case Study. *Survey Methods: Insights from the Field*. Available at http://surveyinsights.org/?p=2991.

Willimack, D. K., and Nichols, E.. 2010. A Hybrid Response Process Model for Business Surveys. *Journal of Official Statistics* 26:3–24.

Yang, S. and Kim, J.K. 2015. Fractional Imputation in Survey Sampling: A Comparative Review. Submitted to *Statistical Science*, with findings available from the *Proceedings of the Survey Research Section*, American Statistical Association, Alexandria, VA.