Hot Deck Imputation: How Many Times Can or Should a Donor Be Used?

Laura Bechtel[1], Yarissa Gonzalez[2], Katherine Jenny Thompson[3]
[1,2,3]U.S. Census Bureau, 4600 Silver Hill Rd, DC 20233

Abstract

Hot deck imputation provides a flexible approach to dealing with missing data that retains multivariate relationships without making explicit parametric model assumptions. Instead, hot deck methods impute missing values (recipients) using reported values (donors) from a similar unit. Under ideal conditions, each imputation cell contains many candidate donors for each recipient. Unfortunately, in practice, this is not always the case. In many instances, there are more recipients than donors within an imputation cell, so that the same donor record may be selected for different recipients. Oftentimes, a limit is set on the number of times a donor can be used. The optimal choice for this limit can be a topic of large debate and generally is determined subjectively by subject-matter experts.

In this paper, we explore a more objective way to determine the optimal choice for donor-use limits when implementing random hot deck imputation (HDR), concentrating primarily on the highly skewed distributions typical of establishment surveys. Through simulation, we investigate the interplay of several factors that may affect the optimality of this limit, such as varying sample sizes (donors, recipients, and total units), varying response mechanisms, and varying response rates. We evaluate the results by comparing the bias and the goodness-of-fit of imputed data obtained from varying donor limit scenarios.

Key Words: Hot Deck Imputation, Simulation, Nonresponse, Donor Limit

1. Introduction

Various surveys within the Economic Directorate at the U.S. Census Bureau[1] employ some version of hot deck imputation to account for unit or item nonresponse. Hot deck imputation provides a flexible approach to dealing with missing data that retains multivariate relationships without making explicit parametric model assumptions. Instead, hot deck methods impute missing values (recipients) using reported values (donors) from a similar unit.

Under ideal conditions, each imputation cell contains many candidate donors for each recipient. Unfortunately, in practice, this is not always the case. In many instances, there are more recipients than donors within an imputation cell, so that the same donor record may be selected for different recipients. Oftentimes, a limit is set on the number of times a donor can be used. The optimal choice for this limit can be a topic of large debate and generally is determined subjectively by subject-matter experts. This phenomenon is not

[1] Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau.

unique to just the Economic Directorate at the U.S. Census Bureau; as Andridge and Little (2010) point out, "The optimal choice of [the donor limit] is an interesting topic for research…" They also indicate that the choice is more than likely related to an accuracy-precision trade off. There is limited literature available offering guidance on setting this limit. For example, Joenssen and Bankhofer (2012) look into several different donor limit schemes applied to multivariate normal data for two different versions of hot deck, random and nearest neighbor. The response rates considered in their study were 80-percent or higher, which often leads to more donors than recipients in the studied imputation cells. They found that for random hot deck, which is the focus of our research, the choice of a donor limit was "frivolous." In establishment surveys, we generally do not have normal data. We are particularly interested in the case of more recipients than donors, which we find in sparsely reported detail data. Our research expands upon Joenssen and Bankhofer's (2012) findings by utilizing highly skewed simulated multivariate data modeled from two business populations. The multivariate data includes a total variable, total receipts (RCPTOT), and detail variables that sum to the total, $\text{Detail}_1 - \text{Detail}_5$. We specifically address the case where there are more recipients than donors. Additionally, we look at several different donor limiting schemes including a different dynamic donor limit that accounts for the ratio of donors to recipients (nonrespondents) as suggested in Kalton and Kish (1981).

In Section 2 we present the studied random hot deck methodology and the donor limits we will evaluate. In Section 3 we present our simulation design and evaluation criteria. Our results are then discussed in Section 4. We conclude in Section 5.

2. Hot Deck Imputation

Hot deck imputation procedures use reported values from the current sample to impute for missing values. Sample units are classified into disjoint groups (imputation cells) based on variables *available for all units in the sample* that are correlated with the missing values. This classification implicitly assumes that nonrespondents and respondents are from the same distribution within each imputation cell, i.e., that the response mechanism is missing-at-random within imputation cell (Ford, 1983). Units are classified as either donors or recipients. Donors are observations that provided usable values for the variable(s) of interest, whereas recipients did not provide usable value(s) and need an imputed value(s) provided by a donor.

There are many forms of hot deck imputation that have various ways of matching donors with recipients. In this paper, we focus on random hot deck imputation (HDR), the simplest form of hot deck imputation. With this method, a donor is selected at random for each recipient within an imputation cell. This can be thought of as taking a simple random sample of size r from d where r is the number of recipients and d is the number of donors in the imputation cell (Kalton and Kish 1981).

When selecting the random sample, the sampling can be done with or without replacement. Ideally, there are more donors than recipients so that without replacement random sampling can be used to minimize variance (Kalton and Kish 1981). However, when there are fewer donors than recipients sampling without replacement will not assign a donor to every recipient. In this situation, a backup method (e.g., cold-deck or warm-deck) and/or cell collapsing can be used. If the former is used, then the backup method should be considered to be a viable alternative; in the latter, collapsing criteria should

preserve the missing at random cell properties. If neither approach is feasible, the other alternative is to allow a donor to be used more than once, that is to use with-replacement sampling. To avoid overuse of any particular donor when with-replacement sampling is used, often donor limits are used in practice. These limits fix the maximum number of selections for each donor, removing it from the set of donors available for imputation when the limit is achieved.

For our research, we investigate the effects various donor limits have on the parameters of interest. We begin with the two most "extreme" options: without-replacement sampling and (unrestricted) with replacement sampling. Without-replacement sampling assigns a donor limit of one (DL1). Once all donors have been selected, the backup method is implemented for the remaining recipients. For RCPTOT our backup method is the cell mean:

$$RCPTOT_c^{recipient} = \left(\frac{\sum_{k=1}^{d_c} RCPTOT_k}{d_c}\right)$$

where c is the imputation cell, k is the *kth* donor, and d_c is the number of donors in imputation cell c. For the details, we use a ratio imputation method:

$$Detail_j^{recipient} = RCPTOT^{recipient}\left(\frac{\sum_{k=1}^{d_c} Detail_{j,k}}{\sum_{k=1}^{d_c} RCPTOT_{j,k}}\right)$$

where j is the *jth* detail and $RCPTOT^{recipient}$ is the population value of RCPTOT for the recipient. Using the population value could result in a much more precise imputed value than we would observe in practice, where the recipient's value of RCPTOT could have been imputed. DL1 is equivalent to a simple random without replacement sample of size r from d and employing a backup method when no donors remain. To contrast this, we look at no donor limit (DLN), which is equivalent to taking a sample of size r with replacement from the d.

Commonly in practice if DL1 seems to be relying on the backup method too often or if a backup method is unavailable, an arbitrary donor limit is assigned based on trial and error or subjective opinion. In a similar vein, we consider a donor limit of two (DL2), which we chose based on the minimum response rate we considered in this study, 40-percent.

Of course, the performance of random hot deck imputation with subjectively determined donor limits can vary considerably between data sets. Using a dynamic (data-driven) approach to determine donor limits as proposed by Joennsen and Bankhofer (2012) or Kalton and Kish (1981) is consequently appealing. Here, we considered two dynamic limits modified from Kish and Kalton (1981).The dynamic donor limit (DLD) sets the donor limit to $\left\lceil \frac{r}{d} \right\rceil$ where r is the number of recipients and d is the number of donors. Note that if there are more donors than recipients this limit is set to one and is equivalent to a simple random sample without replacement. One nice attribute about this limit is the ceiling function usage ensures that a backup method is never used. Oftentimes there is reluctance to use a single donor more than one time while there are unused donors; a downside of this approach is that it can result in some donors being used multiple times while others may not be used at all.

The second donor limit approach combines the DLD donor limit with systematic selection of donors, as discussed in Kalton and Kish (1981). We denote this as the Dynamic Donor Limit – Cycle through Donors (DLC) approach. If there are more donors than recipients, the donor limit is one, and donors are selected without replacement. If

there are more recipients than donors, DLC will minimize the number of donors that are used $\left\lceil \frac{r}{d} \right\rceil$ times by systematically selecting each donor once at each donor selection cycle.

3. Analysis

3.1 Simulation Design

In this section, we describe the simulation design we used to evaluate the effects of varying donor limits on statistics of interest over repeated samples. For the simulation, we used two populations that were modeled from historic 2012 Economic Census data in two industries. These data sets contain two total variables: total annual payroll (PAYROLL) and total receipts (RCPTOT). Additionally, there are five detail variables (Detail_1 – Detail_5) that sum to RCPTOT, i.e., $RCPTOT = \sum_{j=1}^{5} Detail_j$. Each detail item j represents the amount of the total receipts in the specified category, and an establishment may report values in any combination of details, as long as the additivity constraint is satisfied. Thus, zero is a legitimate value for any detail item. This historic data used for modeling contains many more detail items. However, many of the requested detail items are rarely reported. For modeling, each detail item was ranked within industry by percentage of industry total receipts and numbered accordingly: Detail_1 contains the highest percentage of total industry receipts, Detail_2 contains the second highest percentage, etc. , That said, the majority of the total receipts in an industry are reported by one to four detail items. To facilitate modeling, the value for the remaining detail items was consolidated into a single "balance of details" item denoted Detail_5. Consequently, the distribution of Detail_5 is the combination of many different distributions and is very difficult to model or impute.

For our analysis, we start with the simplest scenario, by not incorporating sampling into our simulations. Instead, we induce nonresponse into each population and our samples are effectively the respondent observations from a "census." We randomly induced nonresponse 1000 times using PAYROLL as an auxiliary variable for modeling a covariate dependent response mechanism. This violates the missing at random assumption for hot deck discussed in Section 2. However, it is more realistic for a business program, where larger units are more likely to respond (Thompson and Oliver 2012).

Following Andridge and Thompson (2015), we induce nonresponse via the logistic regression model

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

with H= 1 indicating item nonresponse for either total receipts (RCPTOT) or all detail variables (Detail_1 – Detail_5), and Z indicating PAYROLL. This generates simulated data with the ignorable covariate dependent response mechanism. We set $\gamma_0 = \log(p/(1-p))$, where p is the targeted response rate and the regression parameter is estimated from the population data. We conducted the analysis for three different overall response rates: 40, 60, and 80-percent.

After independently inducing nonresponse, we implement hot deck imputation. Each industry contains two imputation cells (larger and smaller establishments). Performing hot deck for a single variable is straightforward: match donor to recipient, and then

substitute the donor's value in the missing recipient's record. The procedure for detail variables is a little trickier because of the additivity constraint. To preserve the multivariate relationship (multinomial distribution) of details, the donor record's distribution of details is applied to the recipient record as

$$Detail_j^{recipient} = RCPTOT^{recipient}\left(\frac{Detail_j^{donor}}{RCPTOT^{donor}}\right)$$

where j indexes the five different details. This requires a valid RCPTOT value for all units but ensures the additivity requirement.

3.2 Evaluation Criteria

In this section, we present our evaluation criteria. Primarily, the Census Bureau products are tabulations. Consequently, minimizing the bias is very important, and macro-analysis is preferred. However, many programs publish frequency distributions of non-zero (positive) detail values (establishment counts by detail). Consequently, realistic imputed micro-data are very useful.

3.2.1 Macro-Data Evaluation Criteria

To assess the accuracy of the totals produced by each of the different donor limit schemes we looked at absolute relative bias. Let x_i be the population total for item *i* and $\hat{x}_{simp}$ be the estimated total for the completed simulated nonresponse data set *s*, donor limit *m*, and response rate *p*. We define the absolute relative bias as:

$$B(x)_{imp} = \left| \frac{\sum_{s=1}^{1000} \frac{\hat{x}_{simp}}{1000} - x_i}{x_i} \right|$$

To assess the precision we used Mean Squared Error (MSE) of the estimates produced by each of the different donor limiting hot deck imputation implementations:

$$MSE(x)_{imp} = \frac{\left(\sum_{s=1}^{1000} \hat{x}_{simp} - x_i\right)^2}{1000}$$

3.2.2 Micro-Data Evaluation Criteria

To evaluate the quality of the micro-data, we conducted two different types of analysis. To assess the fit of the imputed distributions of details compared to the true population distributions, we implemented the Kolmogorov-Smirnov goodness-of-fit tests (K-S Test). To compare the establishment counts between the imputed and true populations, we conducted chi-squared goodness-of-fit tests. For each of the *j* details where *j*=1…5, we conduct chi-squared goodness-of-fit tests for each donor limit *m* and each response rate, *p*. Let e_j be the proportion of establishments that reported a nonzero value for detail *j* in the population $\hat{e}_j^{smp}$ be the estimated proportion of establishments that reported a nonzero value from the completed simulated nonresponse data set *s* where *s*=1…1000, the goodness-of-fit test hypothesis is:

$$H_0: \hat{e}_j^{smp} = e_j$$
$$H_A: \hat{e}_j^{smp} \neq e_j$$

All tests are conducted at the 10% significance level. Recall that with goodness-of-fit tests, the objective is to **fail** to reject the null hypothesis.

4. Simulation Results

4.1 Macro-Data Analysis

Recall, for our macro-data analysis we looked at both absolute relative bias and MSE of the total estimates for each of the considered donor limit schemes. For reference, we also looked at these statistics when solely employing the backup method to impute for all recipients. Figure 1 presents the absolute relative bias for each of the detail items from one of the two considered populations. The conclusions from the two populations were the same; graphs for the second population are available on demand. When looking at the 40-percent response rate (more recipients than donors), DL1 has a much smaller absolute relative bias than the other studied donor limits. However, recall from Section 2 that the available backup method for this simulation study is likely to yield more precise estimates than the hot deck. In general, the backup method is usually much worse than the implemented method and is a last resort. Thus, the low-response rate results shown below for DL1 may be overly optimistic, especially given the comparable results between the considered hot deck variations when the response rates increase.

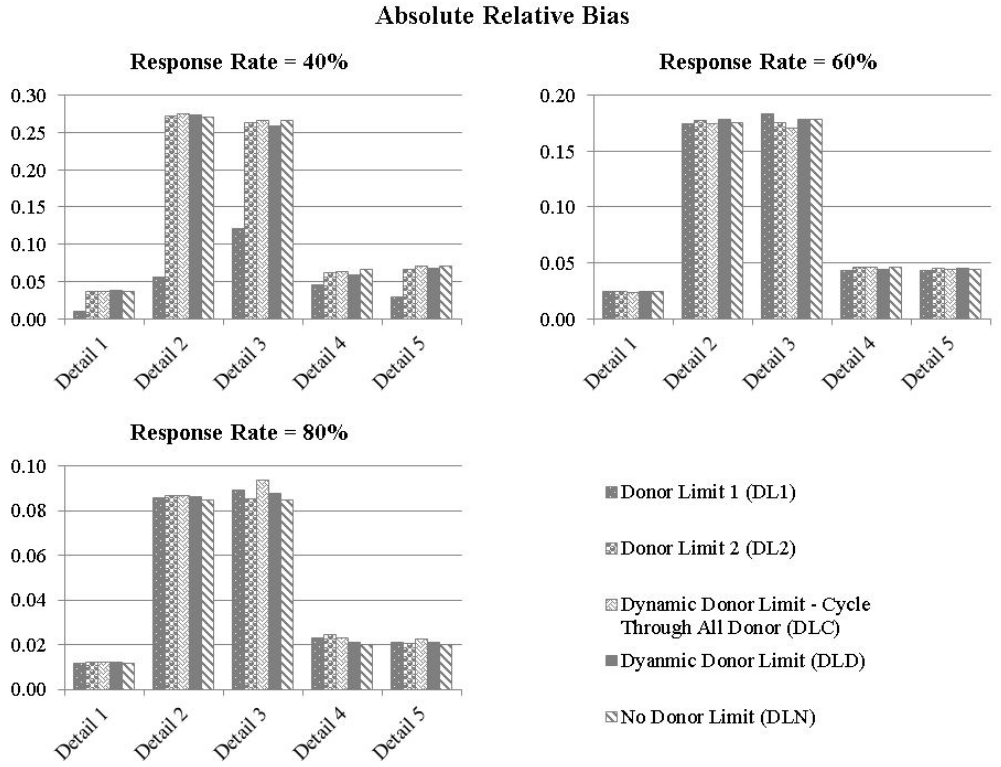

Figure 1: Absolute Relative Bias of Details in One Population

Figure 2 presents the MSE results, which are very similar to the absolute relative bias results. For the 40-percent response rate, the MSE of solely using the backup method was much lower when compared with the hot deck methods. However, the same cautions

apply in the interpretation, as there is no additional imputation variance when the backup method is applied, which in turn leads to a reduced estimate of variance.

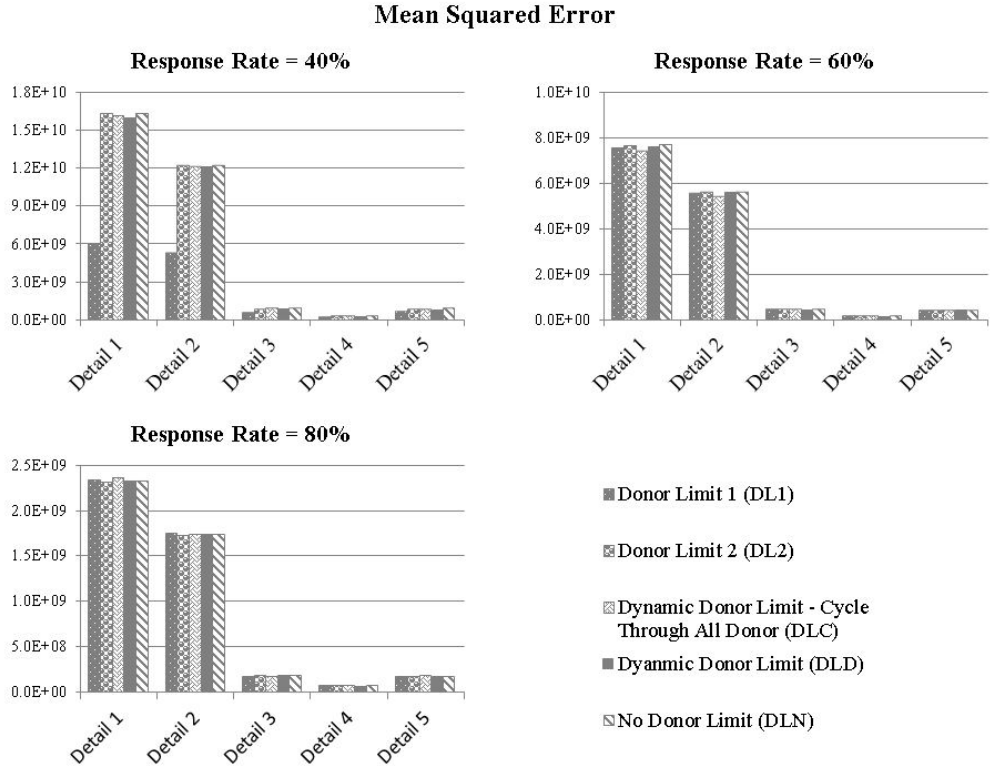

Figure 2: Mean Squared Error of Details in One Population

In Figure 3 we present the absolute relative bias and MSE values for RCPTOT for each of the three considered response rates. Unlike the detail variable results, we see the DL1 absolute relative bias and MSE values are very close to those values observed for all of the other donor limit schemes, even with the low response rate of 40-percent. This is because the backup method for RCPTOT performed similarly to the hot deck methods. These results when considered with the previous results in Figures 1 and 2, demonstrate the way that DL1 is influenced by the properties of the backup method.

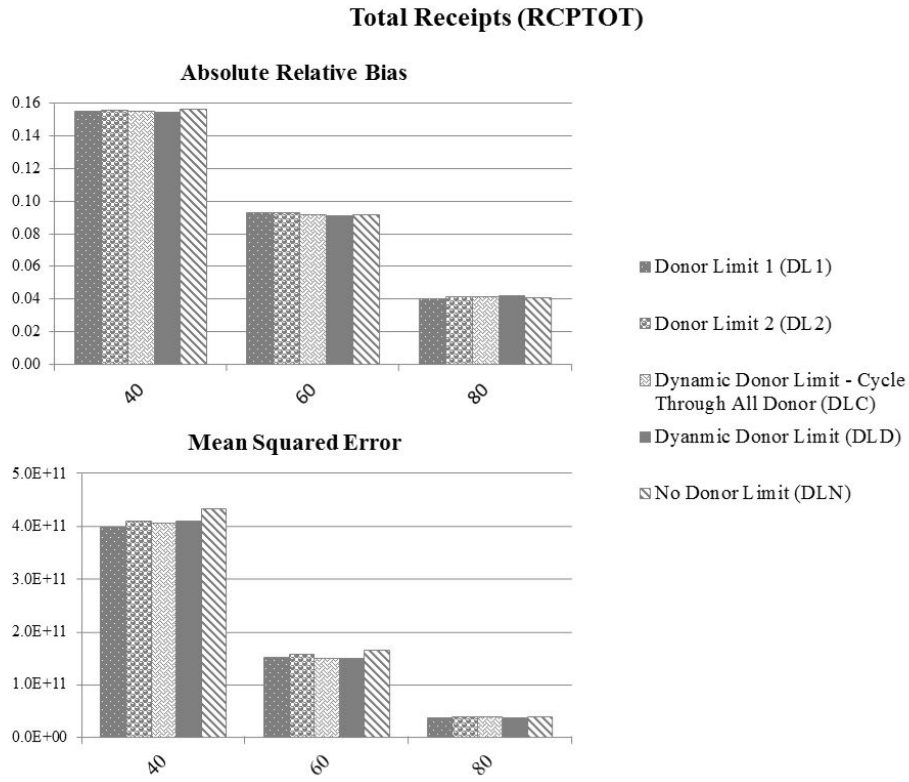

Figure 3: Absolute Relative Bias and MSE for RCPTOT in One Population

All of the considered donor limit schemes result in totals with comparable statistical properties with the exception of DL1. When there are more recipients than donors DL1 can be heavily influenced for better or worse by the backup method.

4.2 Micro-Data Analysis

One of the appealing features of hot deck imputation is that it provides the user with realistic micro-data. In this section, we present the results of goodness-of-fit tests that assess the properties of the micro-data. We first present the results of the Kolmogorov-Smirnov tests where the null hypothesis is that the empirical distribution of the hot deck completed data set is the same as the empirical distribution of the population. Rejecting the null hypothesis indicates that the empirical distributions are not the same. We performed the K-S tests for each of the hot deck completed data sets for each considered response rate. Figure 4 presents the proportion of K-S tests out of 1000 where we reject the null hypothesis for each donor limit scheme. When the response rate is 40-percent, then at most two-thirds of the recipients will be imputed via hot deck with the DL1 approach, with the remaining one third imputed via the backup method. Consequently, there are two separate imputed empirical distributions, which when combined may not resemble the empirical distribution of the population. In our application, the two imputed empirical distributions are clearly dissimilar, as demonstrated by the consistent rejection of the goodness-of-fit tests. In contrast, the empirical distributions obtained using the remaining donor limits look very similar, and the goodness-of-fit tests are rejected less than 15-percent of the time. This indicates that hot deck limits that do not rely on the backup method as frequently as DL1 tend to give fairly realistic micro-data distributions within a population. When the response increases to 60-percent, all donor limit schemes

rarely reject the null hypothesis and there is really no notable difference between them. When the response rate is 80-percent the null hypothesis was not rejected at all for any of the considered donor limits and no graph is presented in Figure 4. Looking at the graph for RCPTOT, we see very similar patterns to what we observed with the detail items.

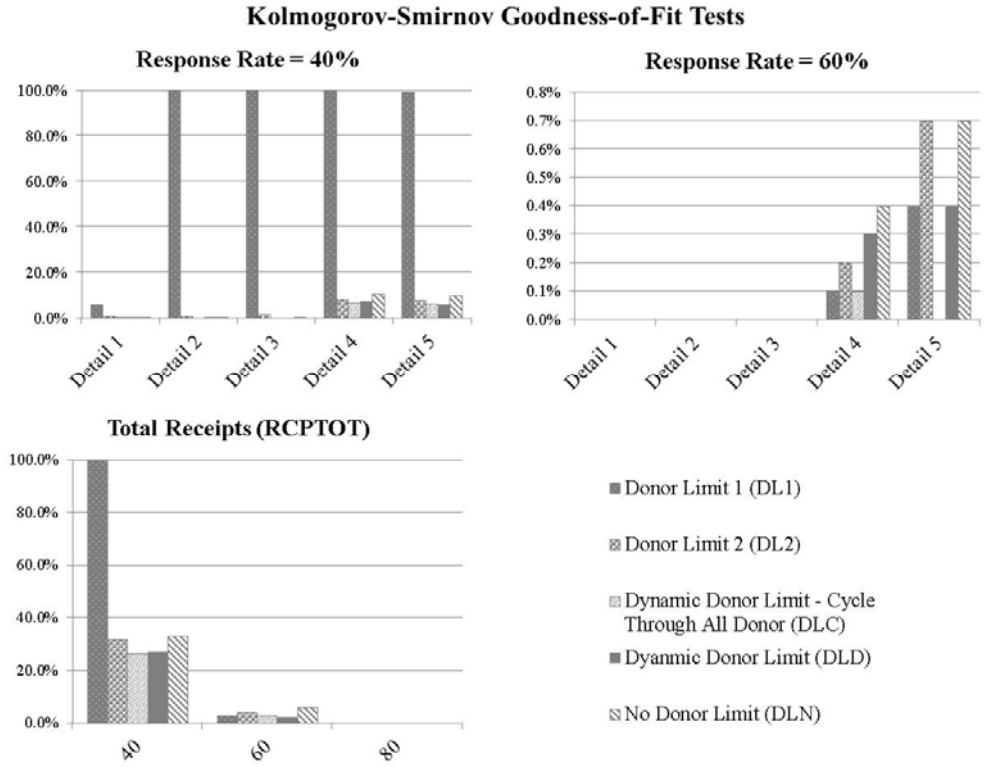

Figure 4: K-S Test Percent of Rejected Goodness-of-Fit Tests

Besides values, counts are often of interest, specifically the number of establishments that reported (nonzero) receipts for each detail item [note: these tests are not meaningful for total receipts in our context given that a value is assigned to every establishment]. To assess the resulting $Detail_1$- $Detail_5$ establishment counts for each considered donor limit, we look at chi-squared goodness-of-fit tests as described in Section 3.2. As with the K-S Tests, rejecting the null hypothesis indicates that the estimated establishment counts are not statistically the same as the population counts. Figure 5 presents the proportion of chi-squared goodness-of-fit tests out of 1000 where we reject the null hypothesis for each donor limit scheme. For all of the details, except for $Detail_1$, when the response rate is 40-percent, we reject the null hypothesis most of the time for DL1. For $Detail_1$ the null hypothesis is rejected less than 20-percent of the time. We observe this difference among the details, because the proportion of units that reported legitimate zero values for each detail varies within the population. For $Detail_1$, a zero value was reported by less than three-percent of the establishments (out of more than 900 establishments). $Detail_2$ – $Detail_5$, however, have a higher percentage of legitimate zeroes reported. The backup method does not impute a zero-value instead imputing a small value for every recipient. Consequently, the use of the backup method for $Detail_1$ does not impact the establishment counts as much as it does for $Detail_2$ – $Detail_5$. Looking at the other donor limits when the response rate is 40-percent, we note that the dynamic donor limits (DLC and DLD) appear to yield the best fits. As the response rate increases, the dynamic donor limits

continue to perform most consistently. Note that the differences in results with the 80-percent response rates are slightly exaggerated due to scale: the null hypothesis is rejected less than 3-percent of the time in all scenarios.

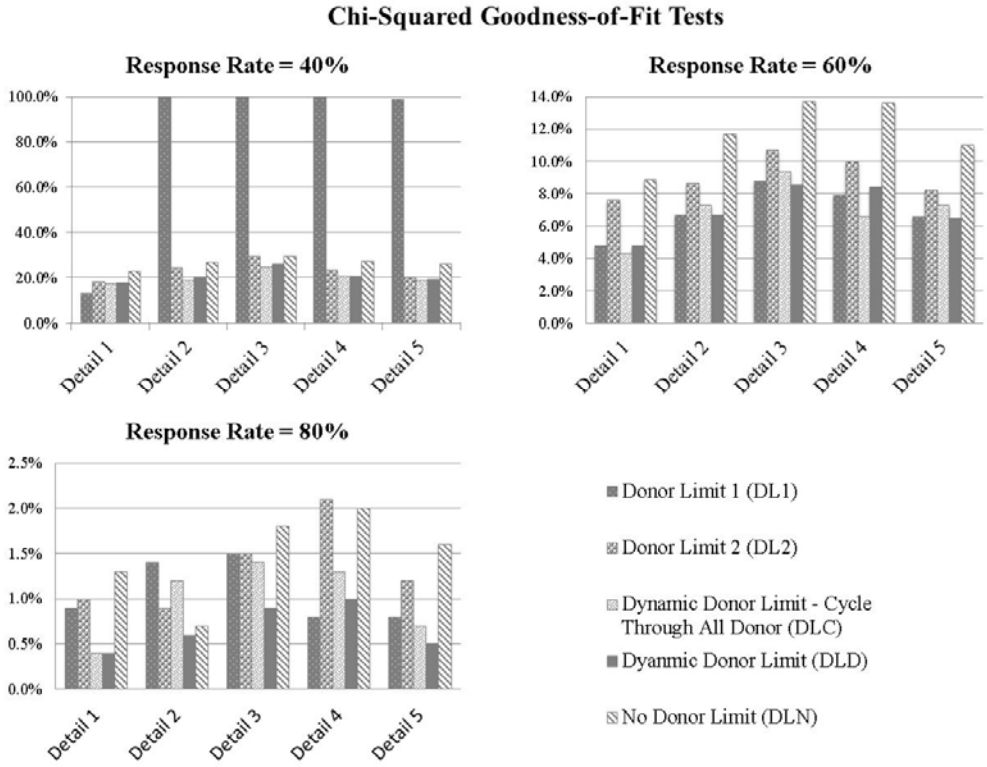

Figure 5: Chi-Squared Test Percent of Rejected Goodness-of-Fit Tests

5. Conclusion

Research is often conducted under somewhat "ideal" conditions. In the case of hot deck imputation research, often there are a large number of donors that greatly exceeds the number of recipients and the donors are a random subsample. In our applications, this situation rarely occurs especially when a large number of details are collected. In this context, the choice of donor limit is no longer **"**frivolous" as suggested in Joenssen and Bankhofer (2012). Using unrestricted donor limits with random hot deck will increase the imputation variance (Kalton and Kish 1981); using a donor limit of one creates hybrid-imputed distributions within the same imputation cell. Neither seems optimal.

Moreover, for DL1 as the response rate moves from 50-percent towards zero, the backup imputation choice becomes more influential. This is true also for other donor limit schemes where the limit is an arbitrary constant. With DL1 and a low donor-to-recipient ratio (i.e., less than 0.5), the backup method becomes the primary method. In this case, the feasibility and statistical properties of the backup method model must be evaluated. In general, one would not use a hot deck method if a viable and reliable backup method existed, unless micro-data distributions were of primary concern.

On the other hand, allowing the donors to be used more than once – even at the cost of increased variability – could eliminate the need for a backup method. However, increased variability may not be acceptable by survey practitioners. The dynamic limits we investigated seem to be a feasible compromise between the two extremes of donor limits we considered. Furthermore, when considering statistical properties of the micro-data (highly skewed) regardless of the donor-recipient ratio there seems to be an indication that either of the dynamic donor limits considered will generally result in more realistic micro-data when compared with the other studied donor limits. When considering the establishment counts and the chi-squared goodness-of-fit tests, the dynamic limits resulted in consistent results across response rates.

For the statistician implementing random hot deck imputation, especially where there are many imputation cells with varying response rates and/or an unreliable or non-existent backup method, the dynamic limits presented here are a good option. Both dynamic limits perform similarly, provide a reasonable compromise of the two extremes DL1 and DLN, and result in precise establishment count estimates consistently across response rates. Furthermore, using either of the dynamic limits sidesteps the debate over what the arbitrary donor limit should be.

Acknowledgements

We would like to thank Carma Hogue and Bonnie Kegan for their thorough reviews of this research.

References

Andridge, R. R. and Little, R. J.A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*, 78(1), 40-64.

Andridge, R., and Thompson, K. J. (2015). Assessing Nonresponse Bias in a Business Survey: Proxy Pattern-Mixture Analysis for Skewed Data. *Annals of Applied Statistics*, 9(4), 2237--2265.

Ford, Barry L. (1983). An Overview of Hot-Deck Procedures. In W. Madow, H. Nisselson, I. Okin (Ed.), *Incomplete Data in Sample Surveys ,Vol. 2,Theory and Bibliographies*, pp185 – 207. New York, NY: Academic Press, Inc.

Joenssen, D. W. and Bankhofer, U. (2012). Donor Limited Hot Deck Imputation: Effects on Parameter Estimation. *Journal of Theoretical and Applied Computer Science*, 6(3), 58-70.

Kalton, G. and Kish, L. (1981). Two Efficient Random Imputation Procedures. *Proceedings of Survey Research Method Section*, American Statistical Association, Alexandria, VA.

Thompson, K. J. and Oliver, B. (2012). Response Rates in Business Surveys: Going Beyond the Usual Performance Measure. *Journal of Official Statistics*, 28(2), pp. 221-237.