R&D, Attrition and Multiple Imputation in BRDIS

by

Juana Sanchez
University of California, Los Angeles

Sydney Noelle Kahmann
University of California, Los Angeles

CES 17-13 February, 2017

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Republication in whole or part must be cleared with the authors.

To obtain information about the series, see www.census.gov/ces or contact J. David Brown, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 5K034A, 4600 Silver Hill Road, Washington, DC 20233, CES.Papers.List@census.gov. To subscribe to the series, please click here.

Abstract

Multiple imputation in business establishment surveys like BRDIS, an annual business survey in which some companies are sampled every year or multiple years, may enhance the estimates of total R&D in addition to helping researchers estimate models with subpopulations of small sample size. Considering a panel of BRDIS companies throughout the years 2008 to 2013 linked to LBD data, this paper uses the conclusions obtained with missing data visualization and other explorations to come up with a strategy to conduct multiple imputation appropriate to address the item nonresponse in R&D expenditures. Because survey design characteristics are behind much of the item and unit nonresponse, multiple imputation of missing data in BRDIS changes the estimates of total R&D significantly and alters the conclusions reached by models of the determinants of R&D investment obtained with complete case analysis.

Keyword*:* Multiple Imputation, R&D, attrition, unit nonresponse, item nonresponse, MICE, Stata MI, visualization, BRDIS, LBD

JEL Classification: 032, 033, 034

---

[*]The research in this paper was conducted while the first author was Special Sworn Status researcher of the U.S. Census Bureau at the Center for Economic Studies. Research results and conclusions expressed are those of the author and do not necessarily reflect the views of the Census Bureau. This paper has been screened to insure that no confidential data are revealed. University of California Los Angeles, Department of Statistics. 1825 Math Sciences Building. Box 951554. Los Angeles, CA 90095-1554. Phone: (310) 825-1318. California Census Research Data Center, (CCRDC - UCLA), 4284 Public Affairs Building, Box 957236 and 337 Charles E. Young Dr, East Los Angeles, CA 90095. Sanchez: jsanchez@stat.ucla.edu. Kahmann: skahmann@ucla.edu.

.

1 Introduction

Analysis of consumer surveys has suggested that multiple imputation (MI) corrects for biases that occur in estimates based on complete case analysis (CCA), and results in gains in efficiency as well [30] [3] [22]. Analysis of business establishment survey data has reached similar conclusions: answers with MI, answers with imputations currently employed by government agencies,and answers with CCA have been found to be different [35] [36]. MI of business establishment survey data presents many challenges [5] [1], particularly when MI is motivated by the need for internal wide-ranging economic analyses and model building, which necessitate the examination of variable relationships not distorted by the imputation process. MI of the Agricultural Resource Management Survey (ARMS) data by the National Agricultural Statistical Service (NASS) is an example of that [25] [21]. But item nonresponse and unit nonresponse decrease survey data quality and affect nonresponse error and inference. Addressing this problem is of the utmost importance.

It could be argued that wide-range MI of the whole data is not needed in the Business Research and Development and Innovation Survey (BRDIS) [2][37] [39], since the sole purpose of this survey is to provide univariate estimates of total R&D expenditures (henceforth R&D) and related quantities. Besides, large companies are responsible for more than three fourths of R&D [40] [37] and interest centers around them, for which data quality is believed to be better [8][7]. However, external researchers accessing the survey data via the network of Census Research Data Centers (RDCs) use BRDIS for economic analysis and multivariate model buildin. But most importantly, there is the possibility that MI of the target R&D variable helps provide more accurate estimates of R&D than CCA, particularly if nonresponse is highly correlated with survey design variables.

In this paper, we present an approach to conducting MI of R&D expenditures in BRDIS in order to determine how model parameter estimates of the determinants of R&D and estimates of total R&D for the year 2013 are affected by MI. BRDIS data are linked to the Longitudinal Business Database (LBD) [12] in order to base the model on variables that are both available to all survey units and are not provided by BRDIS. At the time of writing this paper, released BRDIS data corresponded to the years 2008-2013.

Conducting MI blindly is not recommended. The data should indicate how it should be done [26] [17] [16] [19] [29] [24]. In Section 3 we use simulated data to illustrate the insights into the patterns of missing values that can be gained. Visualization methods that show those patterns and their relations to other variables are conducted with the software R [23]. We use VIM [14] and Amelia packages [10]. Due to disclosure limitations and software availability at the RDCs, we can not present details of such analysis with BRDIS data. However, the first author of this paper conducted similar analysis of BRDIS microdata using other procedures that helped capture similar insights. Results from the latter are presented in Section 4.

Two facts support use of MI of R&D in BRDIS. First, many companies in BRDIS are sampled between 2 and 6 years. These companies may have the same annual unit nonresponse rate as the overall BRDIS population (approximately 30% according to NSF BRDIS Methodology Reports) [38], but the item nonresponse rate due to unit nonresponse is higher than for the whole set of companies combined. The larger the number of BRDIS surveys in which a company is called to participate, the larger the item nonresponse rate due to unit nonresponse. This rate is an indication of temporary attrition, i.e., companies avoiding the response burden in a given year. Because

companies avoid response burden at apparently randomly chosen years, the published aggregate rate of unit and item nonresponse for all companies may stay constant year after year. Instead of indicating constant survey quality, a closer disaggregated look at the response rates suggests response burden avoidance is more prevalent among the companies that matter most for the estimates of R&D. In fact, in BRDIS, item nonresponse in a given year helps predict unit nonresponse in the future. In cases like this, reported R&D values for the same company or for companies with the same seniority are used to multiply impute values in years the company is avoiding burden, resulting in higher estimates of total R&D in the target year.

Secondly, because the target variable in BRDIS is R&D expenditures, all companies surveyed in BRDIS are expected to report it, whether it is 0 or larger than 0. Linking BRDIS to the Longitudinal Business Database (LBD) [12], it is possible to notice that as the number of years being surveyed increases, payroll, proportion of multiunit companies, diversification, geographic scope, employment, age, R&D expenditures and manufacturing increase according to several summary statistics such as mean, median, minimum and maximum. In addition to that, more years in BRDIS also means more companies in the known-R&D stratum annually. We use past and current values of those variables, together with survey design variables (such as stratum and weight) and values of R&D expenditures to multiply impute the missing values of R&D even if unit nonresponse is the reason for the missingness. In this paper, the number of years that a given company is surveyed is measured by a variable referred to hereafter as *count*. R&D item nonresponse depends on *count*, regardless of the value of R&D, hence the missing at random (MAR) assumption is appropriate for BRDIS survey data [29] [6].

Three software packages are used in this paper to explore and conduct MI with the simulated data: R MICE [33] [34], Stata MI [31] and R MI [9]. The final imputation and analysis of BRDIS was conducted with Stata MI and Stata SVY. Parameter estimates and total R&D estimates, calculated with and without multiple imputation, are presented in Section 5. Qualitative comparisons of those results with the ones obtained using alternative methods can also be found in this section. The results presented in this paper support earlier findings using a different imputation method and different model [27].

The paper ends with conclusions and recommendations for further research in Section 6.

2 BRDIS and LBD

This paper uses BRDIS data linked with the LBD [12] [2]. BRDIS is survey data and LBD is administrative data. The first author did the linking separately for multi and single business units using appropriate identifiers [28]. Not only does LBD contribute auxiliary variables for multiple imputation, but it also plays a very important role in multivariate models of R&D, by providing exogenous business metrics not requested in BRDIS and testing long held hypotheses about the determinants of R&D. The rest of this section describes in more detail the data used in this paper.

Table 1 displays the variables included in the imputation and regression models [27]. The table displays the mean (Mean) and standard deviation (Std) of the economic variables listed, without weighting or imputation. The other columns of Table 1 can be explained as follows, with a letter y indicating effect. Logistic regression analysis was done to determine which of the economic variables affect the probability of missingness of R&D (column ProbM), and which of them relate to the probability of being an R&D performer (column ProbRD). The last two columns of the table indicate whether the variable was included in the R&D statistical model (InRM) and/or the

Table 1: Summary statistics and model inclusion for variables appearing in the regression models or imputation models. Unweighted. Source: BRDIS and LBD 2009-2013

Var name	**Mean**	**Std**	ProbM	ProbRD	InRM	InIM
R&D expense	11002	168517			y	y
Multi unit	0.32	0.5		y	y	y
Number of states	2.7	6			y	y
Number of NAICS	2	3.6		y	y	y
Annual payroll (in $1000)	60768	544263		y	y	y
R&D establishments	0.14	3			y	y
Age of oldest est	22	12	y	y	y	y
Years in BRDIS	2.5	2				
Industry				y	y	y
Stratum			y	y		y
Sampling weight						y

imputation model (InIM).

As can be seen in Table 1, the probability of missingness in R&D (column ProbM) is significantly related to the age of the oldest establishment in the company, the stratum, and the survey instrument type. The probability of expenditures on R&D (column ProbRD) is associated significantly with the number of North American Industrial Classification System (NAICS) sectors in which the company operates, whether the company is multiunit or single unit, the survey instrument received, the age of the oldest establishment, the stratum, payroll, and industry.

It must be pointed out that before imputation, the highest correlations of R&D expense is with count (0.11), Number of states (0.21), number of NAICS (0.23), total payroll (0.38), R&D establishments (0.20), and total employment (0.17).

2.1 BRDIS

BRDIS is an annual survey of about 40,000 for-profit, non agricultural companies with at least 5 employees or payroll larger than $250,000, at least one establishment that is in business during the survey calendar year and located in the 50 states of the U.S. or DC, and classified in selected industries with a particular focus on companies that perform R&D in the U.S. (global R&D is not included). BRDIS is administered annually by the National Science Foundation (NSF) and the U.S. Census Bureau, replacing the Survey of Industrial Research and Development in 2008 [39]. The data user community is broad, and includes the U.S. Bureau of Economic Analysis, businesses, the NSF and academic researchers. The years of BRDIS data used in this paper are: 2009-2013 for the estimation and 2008-2013 for the missing data visualization and other qualitative summaries. At the time of writing this paper, the only releases of BRDIS data available to researchers at the RDC were 2008-2013.

The survey is the primary source of information on U.S. business R&D expenditures, R&D work force, R&D management and intellectual property. The innovative capacity of the U.S. is measured almost exclusively by R&D outlays (somewhat over 2.8% of GDP in the U.S.A. in 2010) because R&D expenditures activities, when accumulated, are believed to create the stock of knowledge [4].

Hence BRDIS is a very important economic survey. Response is mandatory and confidential under Title 13 U.S. Code, thus the unit response rate is high. In 2008, the overall unit response rate was 77.4%, and the unit response rate for the top 500 domestic R&D performing companies was 92.6% [39]. The survey frame is extracted from the U.S. Business Register (BR).

There is one sampling unit (SU) per enterprise, covering all the establishments under common ownership or control that operate in the US (60% single unit and 30% multiple unit). The SU is assigned to the U.S. NAICS 2007 industry group in which it has the largest proportion of sales relative to R&D Measure of Size (MOS), although allocation to industry group is later recoded based on business codes reported by the company in BRDIS. For a given company with more than one establishment, the prior year's annual payroll and employment data for its active establishments are summed to the company level. Multiunit employment and payroll imputation in the BR are done before sampling for BRDIS. In multiunit companies, if an establishment was not in the 50 states or the District of Columbia, the establishment was treated as if it did not exist. Measure of size does not include global R&D in BRDIS. For example, multinational corporations in the BRDIS sample frame are assigned industry codes based on their local operations only (their global operations or outsourcing are not taken into account).

All of the above considerations result in a sampling frame of BRDIS that consists of three major strata: (a) known positive R&D in the last 5 years (two treatments). These are companies who are known from prior BRDIS surveys or other sources to have known R&D , and with measure of size their most recently reported domestic R&D; (b) Known zero R&D in the last 5 years. These are companies known from previous R&D surveys or other sources to have zero domestic R&D; (c) Unknown R&D or unkwnown group (two treatments), which consists of companies about which nothing is known of their R&D expenditures. Relatively speaking, the largest group is stratum (c), followed by stratum (a) and (b) [38].

After allocating to strata based on MOS, the companies are allocated to about 60 business strata corresponding to 60 industry groups.

Although the model estimates presented in Section 5 account for the main strata, the 60 business strata are not accounted in the Stata SVY methodology employed in this paper. Instead, industry is divided into three categories: manufacturing, services and research establishment. The estimated model control for industry by including it in the imputation and statistical models.

Business surveys in the United States are usually not integrated. BRDIS, like most business establishment surveys, collects hard data from companies for which records are available, but it does not collect all the information that would be relevant to a multivariate analysis of the relation between R&D and company characteristics. After all, the survey is not done to help researchers, but to obtain univariate population estimates. It is because of this that active cases of BRDIS must be linked to active LBD establishments to obtain auxiliary variables not provided by BRDIS. These auxiliary variables, in turn, are very useful when conducting multiple imputation, because they are available for all BRDIS units.

2.2 LBD

The LBD is a longitudinal census of business establishments and companies in the U.S. with paid employees. LBD is comprised of survey and administrative records. The LBD covers all industries and all U.S. States [12]. The multiunit nature of the company, the legal form of organization, the age of the company, the number of establishments, the states where the company operates, the zip codes where it has establishments, the number of research establishments and payroll and

Table 2: Data as it comes in BRDIS

ID	COUNT	MYOBS	YEAR	R&D	UR
222	3	1	2008	20000	1
222	3	2	2010	15000	1
222	3	3	2012	.	1
541	3	1	2008	1000	1
541	3	2	2009	.	0
541	3	3	2010	.	1

employment, all are variables either measured by LBD or that can be put together using information in LBD. The multivariate analysis of this paper presented in Section 5 seeks to determine the effects of those variables and industry on R&D.

After linkage with BRDIS, establishment data of LBD was compressed to obtain one company aggregate record because BRDIS contains one record per enterprise.

3 Analysis and visualization of missing data

Before conducting MI it is important to explore the patterns of missing data and to determine whether the data are missing completely at random, missing at random or missing not at random [29] [6] [11]. In this task it helps to use all of the available data, that is, survey data for the years 2008-2013. At the time of writing of this paper, the latest release of BRDIS survey data was the one for 2013. This section describes our methodology. The target variable is R&D expense (heretofore R&D), which is the value that companies report when asked for the total R&D paid for by the company in the survey year [2].

The first author appended all BRDIS data sets (years 2008 to 2013) and classified the data by the number of years in which the ID appears in the appended data sets. A variable called *count* was then created, with values from 1 to 6, telling whether the company was sampled one year, two years or any year up to 6. For example, a company with *count*=3 was sampled three times. Because, say, a company that is sampled three years could have been sampled in 2008, 2010, 2012 and other company sampled three years could have been sampled in 2008, 2009, 2010, a variable called *myobs* is 1 for the first year, 2 for the second year and 3 if the third year. Thus the order of survey participation is preserved, and all companies appearing three times are treated the same way, regardless of the years in which they appear. The artificial data in Table 2 for companies that appeared three years in BRDIS between 2008 and 2013 illustrates the data set up. UR in Table 2 tells us whether the company is unit responder (UR=1) or not (UR=0). Note that the data in Table 2 has no resemblance to BRDIS data at all, it is completely made up to illustrate our methodology. In that table, company with ID 541 was a unit and item responder in 2008, a unit nonresponder and item nonresponder in year 2009 and a unit responder and item nonresponder in 2010.

For companies that appear more than one year, the variable of interest, R&D, and other variables related to it, are reshaped from long to wide. This is done to make the missing data patterns exploration and visualization easier. Thus, for example, for the companies that appear three times

Table 3: Reshaping the data

ID	COUNT	MYOBS	year	RD1	RD2	RD3	UR1	UR2	UR3
222	3	1	2008	20000	15000	.	1	1	1
541	3	1	2008	.	.	689	1	0	1

in Table 2, we end up with one observation, as indicated in Table 3.

RD1, RD2 and RD3 for ID 222 represent the value of R&D in year 2008, 2010 and 2012. The variable year represents the first year observed. Variables UR1, UR2 and UR3 represent the value of UR each survey year.

To illustrate the methodology for missing data patterns exploration and visualization, we simulated a data set with a format like that in Table 3. This data set consists of 6 values of R&D (rd1 to rd6), multiunit dummy variable (mu=1 if multiunit) and industry variable (industry) with three categories (research company, manufacturing and service). The second author used this data set to show the different visualizations explored and to come up with strategies for the first author to analyze the actual BRDIS data. This simulated data is intended to represent simulated companies that have value of *count* = 6. Again, the similarity of the data to BRDIS data is remote, but the data is rich enough to present realistic patterns of missing values for *count=6* companies. As in BRDIS, the proportion of observations missing does not increase by year. That is, each year there is the same proportion of missingness, the same proportion of zeros and the same proportion of RD > 0. Companies alternatively take breaks from BRDIS in a random fashion. As in BRDIS for *count*=6, much more than one-fourth of companies have all observations complete, about that much have a mixed pattern of missingness (some first year, some second, etc.) and the small remaining group have a monotone pattern (missing the last, missing the two last, etc) with less than ten percent missing all of the observations.

In BRDIS, item nornresponse for the 6 year participants is due mostly to unit nonresponse and the proportion of item nonresponse is higher in the 6th year. For any *count*, missingness is much larger in the service industries (newspapers, transportation, real state, health industries).

3.1 *Count* as classifier

It makes sense to visualize R&D's missing data patterns separately by *count* because companies within a given *count* are similar, on average, in R&D expense, payroll, employment, and any other measure of company size, with larger *count* representing larger values of those measures. As *count* increases, so do the averages of those variables. Similarly, *count* is very closely related to the survey design variables *stratm* and *sicrcd*. BRDIS uses a complex stratified design. Once the three major strata in the sampling frame are identified (known R&D MOS greater than zero, known zero R&D MOS, unknown R&D and MOS payroll), different sampling methods are used: either take all with certainty, PPS (Probability Proportional to Size) or SRS (simple random sampling) [38]. The item nonresponse for R&D, every year, is higher among the unknown stratum, followed by the zero stratum and then the known stratum. In the latter, the rate of missingness has been increasing in the last years. Companies with larger *count* are more likely to be in the known R&D stratum

Industry is a second layer of stratification that is closely related to *count*. There are about 60 industry groups forming the substrata in BRDIS. Sample and collection unit are defined as

Table 4: Industrial composition of BRDIS (2013). Source: BRDIS Annual Report 2013. NSF.

Industry	**Target Population**	**BRDIS**	**% of sample**	**Certainty**	**Non-certainty**
Manufacturing NAICS 31-33	145,925	19,294	42.79%	9,882	9,412
Other NAICS 21-23, 42-81	1,826,034	25,795	57.2%	14,775	11,020
All	1,971,959	45,089		24,657	20,432

the portion of the fully consolidated company that is located in the U.S. The industry groups go through a careful recoding each year, based on business codes reported by the company, to help the NSF more closely tie the R&D to an industry. In broad terms, the industrial composition of BRDIS companies is as seen in Table 4 [38]. *Count* also plays a role in industry classification: the larger the *count*, the larger the proportion of manufacturing firms.

Because *count* is so correlated with most of the variables in BRDIS, it makes sense to use the above described variables in imputation models for R&D expense and to impute by *count*.

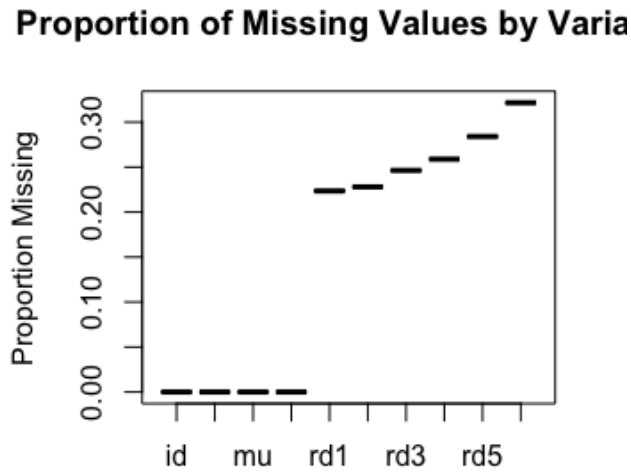
3.2 Identifying unit and item nonresponse in BRDIS

In BRDIS, companies that returned the survey and responded with positive amount of R&D expense *wotrd* or funded R&D or reported worldwide sales, domestic sales, worldwide employees or domestic employees are considered unit responders [38] and that is indicated by variable *UR* as 1 (respondent) or 0 (nonrespondent). The (unweighted) response rate (URR) published by NSF/Census Bureau in its methodology reports is based on this criterion, and was, for example, 73.6% for BRDIS in 2013 [38]. Because of the broadness of this unit response criterion, if the company is a unit responder, it is still possible that the item of most interest, *wotrd*, was not recorded or reported. In that case, the missing value in the R&D expense variable is an item nonresponse and enters into the published quantity response rate also published by NSF in the annual methodology reports. Although target survey variables, like R&D, are edited, they do not get their values imputed by the Census Bureau/NSF. That leaves a substantial amount of unit responders without a value for R&D, that is, with item nonresponse for our target variable. The latter is common in surveys, and, as was mentioned earlier, data production approaches it annually.

3.3 Missing data visualization methods

With *count* being so correlated with measures of size and survey design variables, missing data analysis began by looking at groups of companies characterized by their value of *count*. Indeed, item nonresponse due to unit nonresponse increases with *count* even though item nonresponse decreases with *count*. The second author used data visualization software to offer further insight appropriate for the reshaped data. Because of disclosure avoidance, the simulated data for *count*=6

Figure 1: Proportion of missing values by variables in the simulated data set



is used to illustrate how those methods could potentially be used. The methods could not be used with BRDIS data because the software was not available to researchers in the RDC network at the time of writing this paper. But, using them with the simulated data offered many suggestions for what to look for in the actual BRDIS data and the first author conducted similar analysis using other procedures.

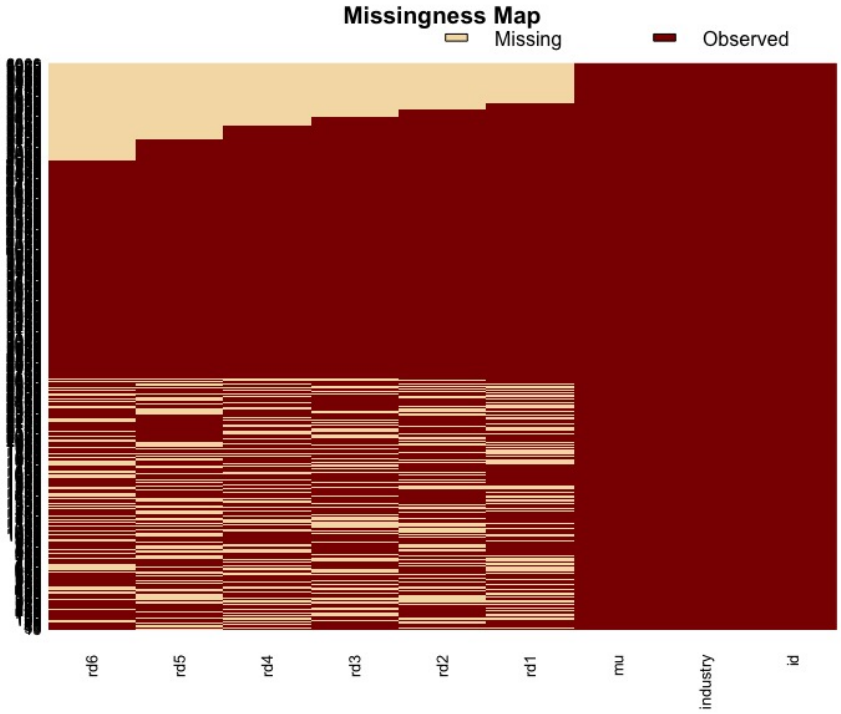
The first step in analyzing missing data is to evaluate the number of missing observations per variable. Most software packages for multiple imputation will summarize this in a table. A quick look with a plot, however, could be more revealing. Using R [23], Figure 1 shows the proportion of missing observations in each of the variables of our simulated data set. In our reshaped data, a similar plot would help reveal whether the proportion of missing values increases with time, thus indicating a general phenomenon of attrition. In companies that have *count* 6, the proportion of missing values have increased over time in the simulated data set.

But it is more interesting to see whether missing values in some variable seem to appear when there are missing values in another variable. Different visualizations of the missing data patterns can then be observed using several R packages. The package Amelia's command *missmap* allows us to see a matrix with two colors, one for missing observations (light color) and one for observed ones (dark color). Figure 2 shows the plot for the simulated data [10].

The plot shows that there is a big group of companies with all values observed (the large chunk in the middle), another big group with intermittent missing values, and then a small group of companies that have a monotone pattern: missing all, missing the last 5, the last 4, and so on. Notice that the horizontal axis is labeled from *rd6* (the one with most missing values) to *rd1*. A major problem with the Amelia graph is that the observation numbers are displayed on the vertical axis.

The R package VIM (Visualization and Imputation of Missing Values) allows for the visualization of missing values as well [14] [32]. The plot presented in Figure 3 allows us to see how

Figure 2: The missing data patterns in the simulated missing data. Top ones are the first observations, bottom ones later observations.

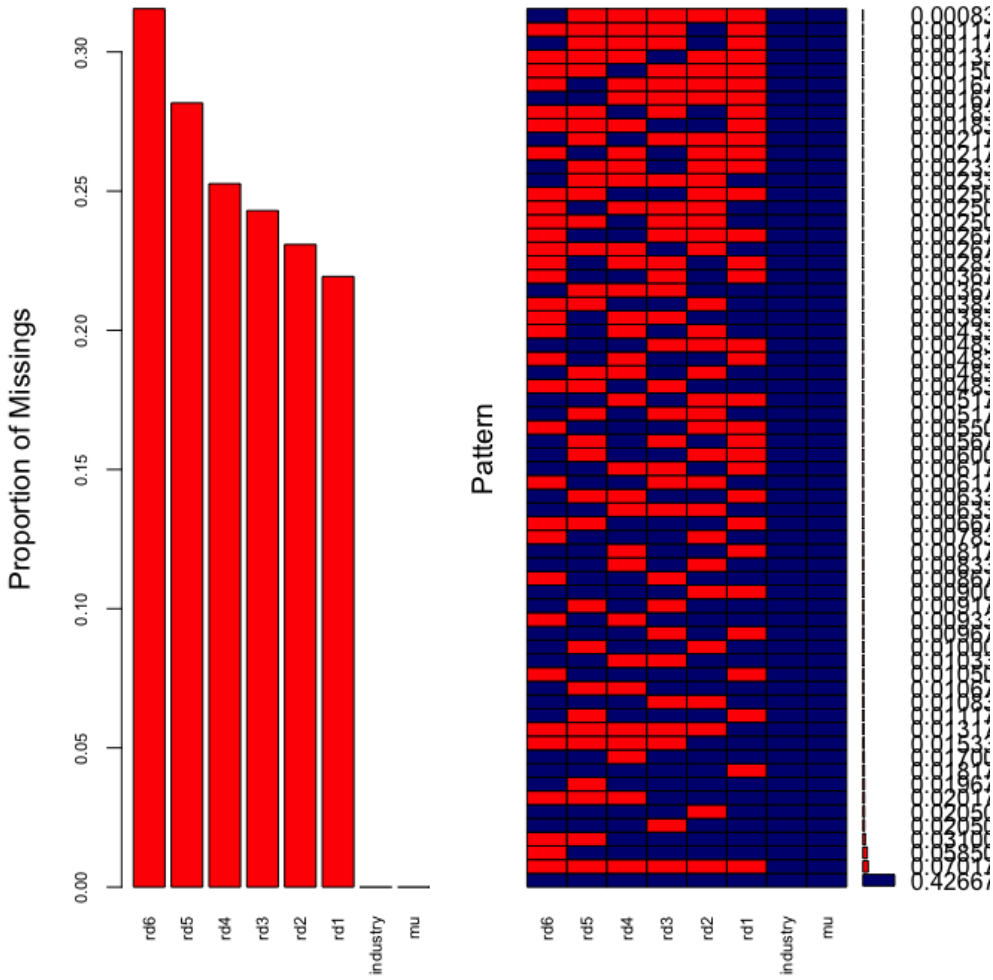


many missing values are contained in each variable and the frequency of the different patterns of missing values in combinations of variables. The barplot on the left-hand-side shows the proportion of missing values in each of the variables. On the right hand side, all existing patterns of missing and non-missing values in the observations are visualized, ordered by their frequency with the most frequent at the bottom. Notice that the group of companies with all observed values (all blue) and the group with all missing values (all red) are most frequent, and all other groups are next. In addition, the frequencies of the different combinations are represented by a small bar plot. Variables are sorted by the number of missing values and combinations are sorted by the frequency of occurrence to give more power to finding the structure of missing values.

Another way to visualize the missing data in VIM is the matrix plot. The matrix plot visualizes all cells of the data matrix by rectangles, similar to heat maps. Matrix plots are very powerful for finding the structure of missing values if the observations are sorted according to a selected variable. The matrix plot for the simulated data can be seen in Figure 4.

To obtain additional information about what type of values of R&D are associated with the missingness, we do spinograms as in Figure 5. The horizontal axis is scaled according to the relative frequency of the bins, i.e. the widths of the bars reflect the frequencies rather than their height. On the vertical axis, the proportion of missing and observed values in other variables can be displayed. The proportion of missing values in the variable of interest can be represented by an extra bar. For example, Figure 5 shows that the proportion of missing values of the *rd2* variable is smaller for the few large companies that have between 40 and 50 values of rd. That would imply the R&D

Figure 3: Aggregation plot of the simulated data with the VIM package in R.



missingness is related to R&D, thus we would have missing not at random.

On the other hand, showing that the level of a third variable such as *industry* or *mu* affects the missingness in R&D would imply missing at random. The opposite would imply missing completely at random (if, in addition, there is no relation between missingness and R&D). Figure 6 shows a case of no relationship between industry and R&D missingness.

In Figure 6, since the height of each cell corresponds to the proportion of missing/observed values in rd1 as a function of industry, it is now possible to compare the proportions of missing values across the different industries. Significant differences in these proportions indicate a MAR situation, which should be considered, e.g., when generating close-to-reality scenarios for missing data in simulation studies.

Further investigations can be conducted that could lead to insight about the missing observations. For example, parallel box plots in VIM for a continuous variable, showing the conditional distributions according to another variable with values recoded as missing or non-missing can be

Figure 4: An alternative way to present the simulated missing data patterns by ID using the matrixplot command in VIM

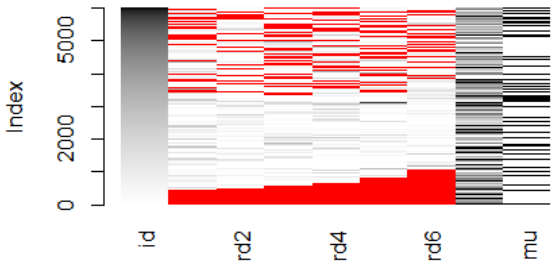
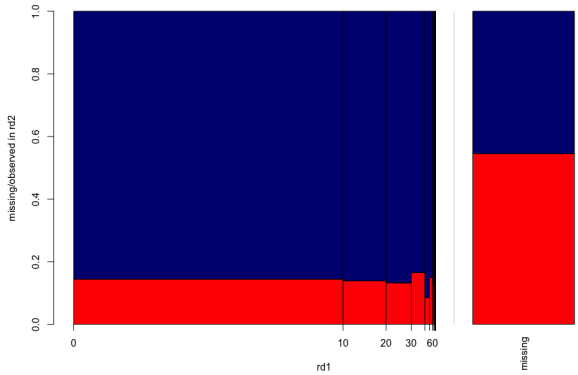


Figure 5: Spinogram of missingness of simulated RD2 as a function of RD1's size.



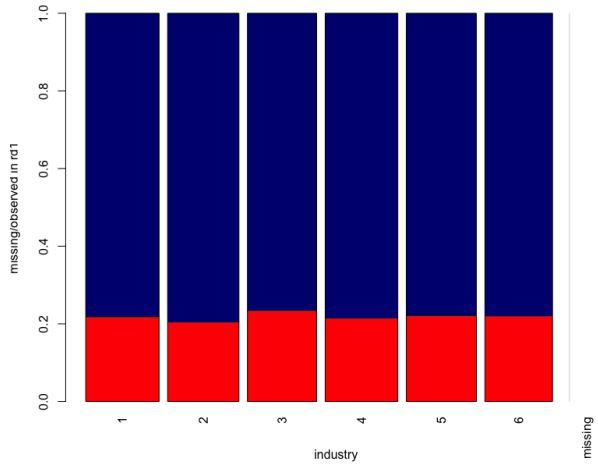
compared by multiple parallel boxplots. This plot is especially useful to explore whether one continuous variable explains the distribution of missing values in any another variable. Figure 7 shows the distributions of *rd2* to *rd6* given missing status of *rd1*. Take for example, the *rd2* boxplots. The red boxplot of *rd2* is the distribution of *rd2* for those observations in which *rd1* is missing. The blue box for *rd2* shows the distribution of *rd2* for those observations in which *rd1* is not missing. According to Figure 7 none of *rd2* to *rd6* help explain the missingness of *rd1* because the red and blue plots are very similar. This makes sense for the simulated data set.

The marginplots in VIM allow for two by two variables, and one can see the relations in both variables. Figure 8 shows the distributions of *rd1* for missing and observed values of *rd2* are very similar. The same can be said about the distributions of *rd2* for missing and observed values of *rd1* (the latter in the horizontal box plots).

The descriptions of missingness given so far are literal in the sense that the plots tell us whether the data are missing or not. The R package MI, however, attempts to cluster the companies by their multivariate missingness [9]. The plot obtained with this software is shown in Figure 9.

The *rd1-rd6* variables are independent in the simulated data set, which is not the case in BRDIS. The missing data patterns in the simulated data were created artificially but they are not too far

Figure 6: Spinogram of missingness of simulated *rd1* as a function of *industry*.



from those in BRDIS, in the sense that the rd variables are skewed and the relative proportion of missing pattern types are close to what they are in the actual data. The most interesting group from the point of view of imputation is the groups of firms that take any pattern except the following NNNNNN, YYYYYY, YYYYYN, YYYYNN, YYYNNN, YYNNNN,YNNNN. That leaves 57 possible patterns such as those presented below, where Y represents response and N nonresponse.

```
"NNNNNN" "YNNNNN" "NYNNNN" "YYNNNN" "NNYNNN" "YNYNNN"
"NYYNNN" "YYYNNN" "NNNYNN" "YNNYNN" "NYNYNN" "YYNYNN"
"NNYYNN" "YNYYNN" "NYYYNN" "YYYYNN" "NNNNYN" "YNNNYN"
"NYNNYN" "YYNNYN" "NNYNYN" "YNYNYN" "NYYNYN" "YYYNYN"
"NNNYYN" "YNNYYN" "NYNYYN" "YYNYYN" "NNYYYN" "YNYYYN"
"NYYYYN" "YYYYYN" "NNNNNY" "YNNNNY" "NYNNNY" "YYNNNY"
"NNYNNY" "YNYNNY" "NYYNNY" "YYYNNY" "NNNYNY" "YNNYNY"
"NYNYNY" "YYNYNY" "NNYYNY" "YNYYNY" "NYYYNY" "YYYYNY"
"NNNNYY" "YNNNYY" "NYNNYY" "YYNNYY" "NNYNYY" "YNYNYY"
"NYYNYY" "YYYNYY" "NNNYYY" "YNNYYY" "NYNYYY" "YYNYYY"
"NNYYYY" "YNYYYY" "NYYYYY" "YYYYYY"
```

It is this group of companies that lends support to our practice of imputing using other years‘ values of R&D and by *count*. A pattern like YYYNNN could represent permanent attrition or company death. But YNYYNY just reflects temporary attrition. It would make sense for a company like the latter to impute using the information in the Y years. That is what we do with BRDIS.

The plots seen have helped to not only visualize the missing data patterns, but also to see some reasons for nonresponse while raising some questions. First, why keep surveying companies that repeatedly do not respond (pattern NNNNNN)? Certainly the cause for nonresponse among those who are chronic nonresponders must be different than the reason for other companies. Second, companies that alternate response and nonresponse suggest that the cause of the nonresponse for them is fatigue due to the burden of having to respond to the same survey each year. Why not use a survey design that reduces that burden like, for example, sequential Poisson sampling?

Figure 7: Parallel box plots of simulated data in VIM.

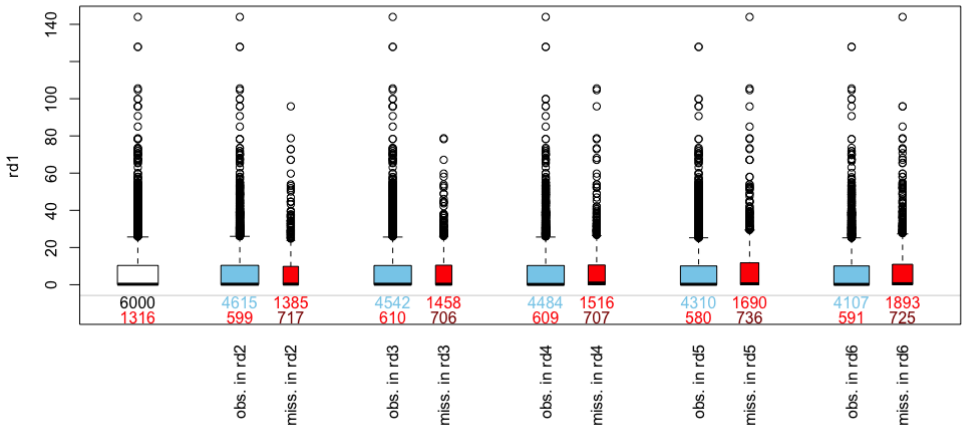


Table 5: Recoding the simulated data

ID	COUNT	MYOBS	year	RD1	RD2	RD3	UR1	UR2	UR3
222	3	1	2008	0	0	1	1	1	1
541	3	1	2008	1	1	0	1	0	1

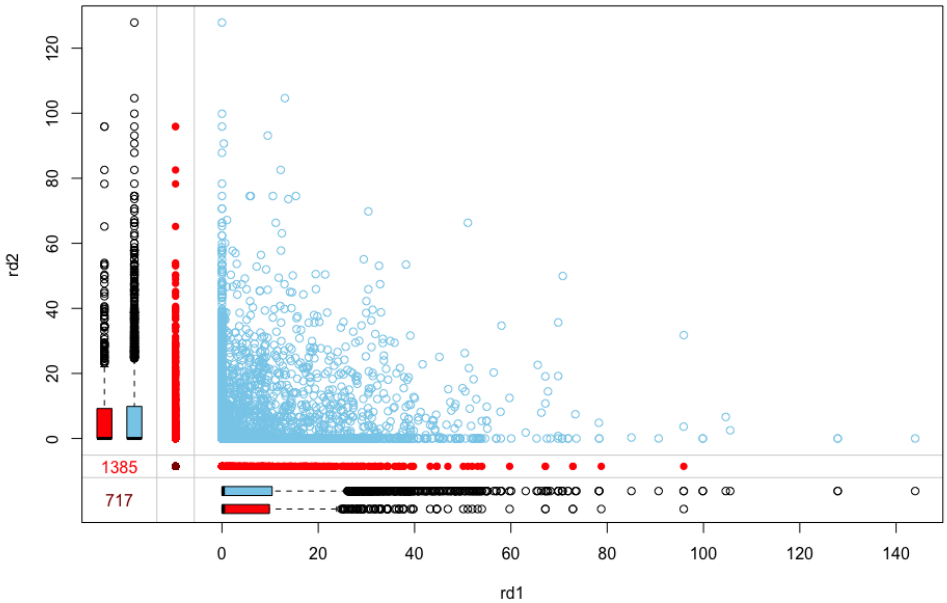
And those that disappear for good after several consecutive years of participation, in a monotone fashion, cannot be companies that closed, because the data include only active companies. Item nonresponse is a good predictor of future unit nonresponse so this information could be used to prevent the latter.

3.4 Does past item non-response help predict future unit non-response?

Studies of the relationship between item nonresponse and subsequent unit nonresponse have found a fairly strong positive relationship between them [20]. To investigate whether this is the case, the simulated data was further processed. First, the data in the format of Table 3 was recoded as seen in Table 5.

With the data in the format of Table 5 for *count*=6, we modeled unit nonresponse propensity for *rd6* as a function of *rd1-rd5* and *ur1-ur5*. A common way to do this in the response literature is binary logistic regression. By modeling the probability of unit nonresponse in the last year, j, as a function of unit nonresponse and item non response in period j-1, we can test the hypothesis that item nonresponse helps predict future unit nonresponse. Based on the analysis done in this data, item nonresponse in past recent periods significantly predicts unit nonresponse in the current period. In other words, item nonresponse in an otherwise compliant company can be seen as an indication that next year this company may not return the survey. This is a sign of fatigue due to survey response burden. Table 6 shows the results of the logistic model used to predict the unit

Figure 8: Marginplots of simulated data using VIM.



nonresponse in year 6 using item and unit nonresponse in the past 5 years, controlling for *industry*. These are results for the simulated data. All variables are binary. The *rd* variable is 1 when the R&D is nonmissing, and the *ur* variable is 1 when there is a unit response. The simulated data has more than 3 industry groups. The model is saying that item response in years 3, 4, 5 will increase the log odds of unit response in year 6 and viceversa, item nonresponse in those years is a predictor of unit nonresponse in year 6.

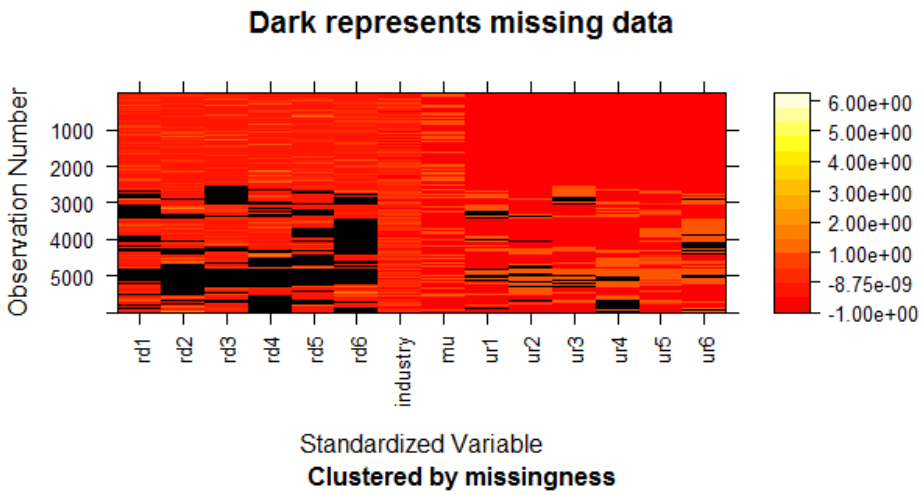
Disclosure avoidance restrictions prevents us from presenting detailed results like those in Table 6. However, we can say that in logistic regression models for *count*=6 using the actual BRDIS data, it was found that unit nonresponse in year 6 for companies having *count* 6 is more likely to be higher for those with unit nonresponse in years 3, 4, 5, and item nonresponse in year 5. Thus, BRDIS companies signal a forthcoming unit response in year 6 by engaging in item nonresponse the previous year. Item nonresponse helps predict unit nonresponse.

4 Multiple imputation of R&D in BRDIS.

The missing data visualization presented in Section 3 illustrates how we can obtain a large amount of information about missing data patterns and causes of missing data if we look at all the available data under the proper lens. That lens, in BRDIS, is the value of *count*. *Count*, the number of years that a company has been surveyed, is a proxy for firm size, age, industry, payroll, employment, survey variables and R&D. Bringing what was learned from the simulated data into the analysis of the actual BRDIS/LBD data, with different procedures, revealed the following:

- The proportion of observations with missing R&D does not increase by year. That is, each year there is approximately the same proportion of missing R&D, the same proportion of zero

Figure 9: This plot was produced using *image(mi data frame)*. It does not describe the pattern of missing values but statistical clustering of the missingness based on the multivariate normal distribution. It also shows outliers, values of the variables that are beyond three standard deviations from the mean of the variable (lighter values).



R&D and the same proportion of R&D > 0. Firms take turns in taking breaks from BRDIS answering but in a random fashion, as in the patterns indicated earlier in Section 3. Some choose one year, some choose another.

- In any given year, almost all the unit responders complete their R&D field. The large item nonresponse rate is mostly due to unit nonresponse, and that phenomenon is more prevalent as *count* increases. This is true of all years.
- Item nornresponse for the 6 year participants is due mostly to unit nonresponse.
- Companies that are sampled many times may have the same unit nonresponse rate as the overall population, but the proportion of their item nonresponse due to unit nonresponse is higher the higher count is.
- Looking at 6 year participants, a small percent are companies that never returned the survey.
- For companies with *count* 6, the proportion of missing values is higher in the 6th year. For any *count*, missingness is much larger in the service industries (newspapers, transportation, real estate, health industries, among others).
- The proportion of item nonresponse given *count*, decreases with *count*. This happens every year.

These findings reinforce the conclusion that multiple imputation of BRDIS data must be done by *count*.

Thus, in the jargon of multiple imputation, item and unit nonresponse in R&D expense can be considered missing at random (MAR) conditional on the predictors used in the imputation models. This means that the missingness does not depend on the measured R&D expense itself but on other

Table 6: Logistic Regression to predict log odds of unit nonresponse in 2013. Simulated data

Variable	*Estimate*	*Std.Error*	z value	p-value
Int	-4.72113	0.25885	-18.239	$< 2e-16$
rd1	0.19396	0.20847	0.930	0.3522
ur1	-1.47888	1.02041	-1.449	0.1473
rd2	-0.17345	0.22329	-0.777	0.4373
ur2	0.07051	0.61618	0.114	0.9089
rd3	0.50979	0.21560	2.365	0.0181
ur3	-14.70949	419.48244	-0.035	0.9720
rd4	0.48744	0.21697	2.247	0.0247
ur4	-0.50800	0.60648	-0.838	$2.28e-09$
rd5	1.29181	0.21615	5.976	$2.67e-12$
ur5	0.18507	0.44277	0.418	0.6760

variables. In the case of a complex survey like BRDIS, missingness is closely tied up to the survey design characteristics, in particular, the strata [27] but more generally to *count*, as discussed in this paper. Therefore, it is important to conduct the MI by *count*. If we do so, MI values will be based on companies with characteristics similar to those of the company being imputed.

With multiple imputation, more than one value of the nonresponse is imputed, and the analysis is performed on each imputation sample to obtain a pooled unique estimate. The variance of summary statistics between and within imputation samples is then used to incorporate imputation variance (nonsampling error) into the calculations and obtain more accurate estimates of standard errors and significance tests [29] [6] [11] [26] [17] [19] [18].

We use an operational approach to determine whether multiple imputation affects estimates produced and their standard errors. Multiple imputation of item and unit nonresponse is a way of addressing survey data quality, in particular, nonsampling errors [27].

4.1 Statistical analysis

Most researchers want to do analysis for a specific group of the data, i.e. impute and obtain estimates for models fitted to subpopulations that interest them. For example, some may be interested in the largest 200 R&D investors [8]. Others may be interested in companies in Ohio. We are interested in estimating total R&D in the year 2013. We are also interested in estimating the parameters of a model that predicts 2013 R&D expenses (a BRDIS variable) as a function of several variables described next. Total R&D is estimated annually by the NSF/Census Bureau. The last release of BRDIS data is 2013.

The independent variables in the model were summarized in Table 1 and are further described now. The names given to them here are the ones used in the Results tables of Section 5. Total payroll (*paytotal*) is created by aggregating the payroll of multiunit companies reported in LBD and the payroll reported for single units. A multiunit or single unit dummy variable (*mu*) in LBD is used to report whether the company is multiunit (mu=1) or single unit (mu=0). Industry sectors (*industry*) is a classification variable created from the sicrcd variable in BRDIS that identifies whether the company is a research company or is in manufacturing (not research) or service. The number of research establishments (*rdesttotal*) is based on sicrcd in BRDIS. The number of NAICS

(*nnaics*) is obtained by counting the number of bestnaics in LBD for multinunit companies. Single unit companies have only one bestnaics. The number of states where the company operates (*nstate*) is obtained by counting the number of different states of the establishments in LBD. Note that this contains only the 50 states and DC. The age of the company (*agemax*) is obtained from LBD and represents the oldest establishment‘s age. The age of the establishment is constructed with LBD data using the last year that the company was observed minus the first year observed.

To MI and analyze subpopulations while accounting for the survey features such as strata and weights, without requiring too much programming, it is convenient to have good software that allows such customization. SAS 9.4 would be appropriate for the whole task of MI and survey based estimation. The R package MICE, closely tied to VIM visualization and Lattice visualization [34] [33] offers another possibility. R‘s MICE can conduct the imputation, but setting up R MICE to do the imputation, by *count* and then to do survey-specific subpopulation estimation, is only possible by first getting the fitted R&D obtained by pooling all MI estimates and then doing the survey analysis separately. Additional programming and use of other packages would be required. The R package MI proved to be too difficult to understand, and had a poor interface for our analysis. An additional concern is that we found the MICE package and Stata gave different results for the pooled regression estimates.

After experimenting with several of the software programs mentioned, we concluded that the best possible scenario for the analysis of attrition and for multiple imputation is to use VIM, Amelia, and Lattice to visualize patterns and then to use Stata MI imputations with SVY structure variables and weights to do several types of imputation and subpopulation estimation. SAS 9.4 is also a good candidate for the imputation and estimation. The first author used Stata 14 for the imputation and estimation. The rest of the paper describes how that was done and gives the results obtained.

5 Results

MI in Stata assumes missing at random. *MI impute* creates imputations by simulating from a (approximate) Bayesian posterior predictive distribution of the missing data. It can do so taking into account the survey weights and stratification. Moreover, it can do Predictive Mean Matching (PMM) like MICE. With PMM in Stata, you need to decide how many nearest neighbors to include in the set of possible donors, otherwise it defaults defaults to one nearest neighbor, knn(1).

5.1 MICE with PMM

We use MICE with PMM to impute the data. The PMM method is a stochastic regression technique in which a missing value on a variable is replaced with the value from a donor - a respondent whose regression-predicted score is closest to the regression-predicted score of the respondent for whom the value is missing. Because actual values of the value to be imputed are assigned, it is appropriate for imputing discrete and continuous measures. PMM is superior to both mean imputation and deterministic regression methods with regard to standard error estimation [15] [13]. PMM combines the standard linear regression and the nearest-neighbor imputation approaches. It uses the normal linear regression to obtain linear predictions. It uses linear prediction as a distance measure to form the set of nearest neighbors (possible donors) consisting of the complete values. Finally, it randomly draws an imputed value from this set. By drawing from the observed data, PMM preserves the distribution of the observed values in the missing part of the data, which makes it more robust than

Table 7: Estimates of total weighted R&D and average weighted R&D for 2013, without multiple imputation (CCA). Subpopulation study for 2013 using ($N_{2009-2013} = 110000$; subpopulation N=23000). Three industry categories are used as control: research, manufacturing (not research) and service. The last two were used as independent variables and only the manufacturing (non-research) was statistically significant with a large effect. ($p < 0.01$). The service category has a negative effect that is not significant.

Variable	Coef.	Std. Err.	t	$P > \|t\|$	95% Conf. Int
1.mu	3246.022	4805.34	0.68	0.499	$(-6172.403, 12664.45)$
paytotal	0.1426	0.0567	2.51	0.012	(0.0313, 0.2538)
agemax	-372.2174	89.0009	-4.18	0.000	$(-546.6584, -197.7764)$
nnaics	2684.272	3477.208	0.77	0.440	$(-4131.025, 9499.57)$
nstate	-617.9197	1403.648	-0.44	0.660	$(-3369.058, 2133.219)$
rdesttotal	3251.362	3792.992	0.86	0.391	$(-4182.87, 10685.59)$
constant	871.5226	5845.309	0.15	0.881	$(-10585.23, 12328.28)$

Table 8: Stata MI with PMM ($N_{2009-2013} = 145000, N_{2013} = 30000$). Number of burn in iterations=10, datasets=5, nearest neighbors=5. Multiple regression results with MI and subpopulation analysis for 2013. Uses all the data for the estimation of standard errors, but only 2013 for the regression coefficient estimates.

Variable	Coef.	Std. Err.	t	$P > \|t\|$	DF	% increase s.e.
1.mu	3613.862	3768.045	0.96	0.338	110956.6	0.00
paytotal	0.1378	0.0411	3.36	0.001	110810.4	0.01
agemax	-355.5897	69.3308	-5.13	0.000	106687.9	0.06
nnaics	2037.27	2760.437	0.74	0.461	110231.4	0.02
nstate	23.9638	1121.139	0.02	0.983	107453.1	0.05
rdesttotal	4114.519	4379.975	0.94	0.348	110987.3	0.00
constant	-195.6143	6125.513	-0.03	0.975	110637.7	0.02

the fully parametric linear regression approach [15]. Because PMM draws its imputed values from the observed values, it has the property that the imputed values will never be outside the range of the observed values. This makes it very useful for bounded variables. It can also be used for some non-continuous distributions. However, PMM is not appropriate if you have reason to believe the unobserved values are outside the range of the observed values.

The imputation model imputes R&D as a function of all the other variables mentioned above and the survey variables representing weight, strata and industry. The imputation is done by count. All the observations for the years 2009-2013 for the corresponding count are used to impute any R&D value. Thus, although we are interested in analyzing the data of the year 2013, the imputation stage uses all available information.

5.2 Analysis

We first analyze the data without imputation. Table 7 shows results of the regression model that predicts R&D from the variables listed in the table. Only *payroll*, *age* and *manufacturing* (not

Table 9: Imputation Variance information related to Table 8

	Within	Between	Total	RVI	FMI	Rel efficiency
1.mu	$1.4e+07$	999.758	$1.4e+07$	0.0000	0.0001	0.9999
paytotal	0.0017	$3.1e-07$	0.0017	0.0002	0.0002	0.9999
agemax	4801.06	4.7560	4806.77	0.0012	0.0012	0.9997
nnaics	$7.6e+06$	3045.8	$7.6e+06$	0.0005	0.0005	0.9999
nstate	$1.3e+06$	1121.8	$1.3e+06$	0.0011	0.0011	0.9997
rdesttotal	$1.9e+07$	83.0194	$1.9e+07$	$5.2e-06$	$5.2e-06$	0.9999
cons	$3.8e+07$	10014.5	$3.8e+07$	0.0003	0.0003	0.9999

shown) appear to be significant.

We next look at the results for the same model obtained with the imputed data using PMM. Table 8 gives those results. We can see that, compared to the results without MI, the same variables are statistically significant. The coefficients of the model are different in level but not in sign or significance, except for the variable *nstate*. The standard errors of the estimates are smaller in the model that uses imputed data. Multiple imputation output gives the additional information related to the multiple imputation effects on variances. The percentage increase in standard error due to missing data ranges between 0 and 6 percent. These can be seen in Table Table 8 and 9. Repetition of the analysis with more iterations, more imputed datasets and different nearest neighbors confirm these results.

5.3 Estimates of total and average R&D

As indicated earlier, MI may affect not only multivariate estimates of relations between variables but also the estimates for which the survey was created, namely estimates of total R&D. Table 10 shows the estimates obtained for total R&D and average R&D with and without imputation. As we can see in the results, total R&D estimate is higher with MI. The estimate of total R&D expenditures with multiple imputation is $3.81e+08$ with standard error $3.18e+07$, and 95% confidence interval for the total of $3.18e+08$ to $4.43e+08$.

5.4 Additional results

Although not shown here, we must point out that the estimates of total 2013 R&D obtained separately for each *count* subgroup show different conclusions.

- First, there is a lot of variability in the estimates of total R&D by *count*, as would be expected. There is also a lot of variability in the estimates of average R&D.
- Second, there are extreme differences between the estimates obtained without and with imputation for all *counts*. Thus the differences between estimates of model parameters and total estimate of R&D are more pronounced when the analysis is done for subpopulations of companies with smaller sample size than it is for the whole 2013 sample.
- All the expected variables such as *multiunit, age, payroll* and *industry* play a role in explaining R&D if the company has participated many years (4 or more) in the survey but not if few

Table 10: Univariate estimates of Total and Average R&D without and with imputation plus Imputation Variance. Year 2013. BRDIS 2009-2013.

			No imputation			
Statistic	Estimate	s.e.	95% CI LB	95% CI UB		
Total R&D	$2.61e+08$	$2.49e+07$	$2.12e+08$	$3.10e+08$		
Average R&D	11513	1098.121	9360.691	13665.31		
			Multiple Imputation			
	Estimate	s.e.	95% CI LB	95% CI UB		
Total R&D	$3.81e+08$	$3.18e+07$	$3.18e+08$	$4.43e+08$		
Average R&D	12640.4	1054.216	10574.15	14706.65		
			Multiple Imputation Diagnostics			
	Within	Between	Total	RVI	FMI	Relative efficiency
			Multiple imputation by count, adjusted weights			
Total R&D	$1.0e+15$	$7.1e+11$	$1.0e+15$	0.0008	0.0008	0.9998
Average R&D	$1.1e+06$	778.537	$1.1e+06$	0.0008	0.0008	0.9998

years. In the latter case, only *age* is significant. Thus, the results presented in our Tables, which are for all companies participating in BRDIS in 2013, mask the differences across companies with different values of the *count* variable.

The conclusions reached in the economic literature concerning the determinants of R&D like those in the pool of independent variables used here, which are often obtained using data for the largest companies, do not apply to medium sized and smaller companies.

6 Conclusions

Several investigations of BRDIS data, guided by preliminary analysis of missing data patterns using simulated data, led us to the conclusion that missing data patterns in BRDIS are not happening completely at random or not at random. Measuring the nonresponse rate for the entire sample each year and comparing it to the rate of the subset of companies that participated in the survey all years, helped determine that the number of years that a company has participated in the survey has an impact on nonresponse rates. This led us to further investigations that led to the conclusion that as the number of years participating in BRDIS increase, the item nonresponse rate decreases, the most important economic variables increase, the rate of item nonresponse due to unit nonresponse increases, the percentage in the known R&D stratum increases and the percentage of manufacturing companies increases. Visualization of missing data patterns with simulated data provided other insights that confirmed the closed ties that temporary attrition has to survey design characteristics. Those conclusions led us in turn to conducting the MI of BRDIS survey data by *count*, the number

of years that the company participates in BRDIS, and including all the available information about a company to impute its R&D value.

The estimate of total R&D in year 2013, the year of interest, is higher with the multiply imputed data. Parameter estimates of the relations between payroll, multiunit nature, industry and other variables are also different under multiple imputation than those obtained with CCA. Most significantly, disaggregated analysis by *count* shows that the conclusions of the literature on the determinants of R&D apply only to large companies, i.e., the companies that are surveyed more often.

The analysis of item nonresponse presented in this paper suggests that it might be beneficial to BRDIS estimates of total R&D to move to a Sequential Interval Poisson sampling design in order to obtain higher response rates. NSF/Census Bureau change some aspect of the the BRDIS survey each year: in 2013, the threshold was changed, other years the industry specification changes. Perhaps by paying more attention to what is happening over time to the response rates, and switching to a sequential Poisson type of sampling, response burden will decrease and the estimates of total R&D will increase. Worrying about missing data on a year by year basis and ignoring temporary attrition may be obscuring the fact that the response strategy of BRDIS companies is affecting the estimates obtained, and that strategy is in turn motivated by the sampling design.

A direction for future research is to conduct domain analysis of specific subgroups to see if the results found here replicate when the sample size is smaller, the item nonresponse rate is larger or smaller, the imputation model uses industry specific variables and there is more than one variable with missing values.

Knowing that a survey data quality could be due to the survey design, can be critical for conducting a meaningful post-data collection analysis of non-response and more informed data analysis.

Acknowledgments

This research was partially supported by Professional Development Awards 2013-2016 and other support received from the University of California Los Angeles. We thank Dennis Li, who contributed to improving the article with his bibliographic and editorial assistance.

References

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

[2] C. Bureau/NCSES. Business Research and Development and Innovation Survey (BRDIS), 2010. http://www.census.gov/econhelp/brdis.

[3] Y. Chen and D. Fu. Measuring income inequality using survey data: The case of china. *Journal of Economic Inequality*, 13(2):299–307, 2015.

[4] W. M. Cohen. Fifty years of empirical studies of innovative activity and performance. pages 129–213, 2010. in Hall, B. H. and Rosenberg, N.(ed.), *Economics of Innovation*. North Holland.

[5] J. Drechsler. Multiple imputation in practice–A case study using a complex german establishment survey. . *Advances in Statistical Analysis*, 95(1):1–26, 2011.

[6] C. K. Enders. *Applied Missing Data Analysis.* The Guilford Press, 2010.

[7] L. Foster and C. Grim. Characteristics of the Top R&D Performing companies in the U.S.: Evidence from the Survey of Industrial R&D. *Census Bureau, Center for Economic Studies Discussion Paper CES-WP-10-33*, 2010.

[8] L. Foster, C. Grim, and N. Zolas. A Portrait of Firms that Invest in R&D. *Census Bureau, Center for Economic Studies Discussion Paper CES-WP-16-41*, 2016.

[9] A. Gelman and J. Hill. Opening Windows to the Black Box. *Journal of Statistical Software*, 40, 2011.

[10] J. Honaker, G. King, and M. Blackwell. Amelia II: A Program for Missing Data. *Journal of Statistical Software*, 45(7):1–47, 2011.

[11] IDRE. Missing Data Techniques with Stata. 2016.

[12] R. S. Jarmin and J. Miranda. The Longitudinal Business Database, 2012. Center for Economic Studies, U.S. Census Bureau. Working Paper 02-17.

[13] G. Kalton and D. Kasprzyk. The Treatment of Missing Survey Data. *Survey Methodology*, 12:1–16, 1986.

[14] A. Kowarik and M. Templ. Imputation with the R package VIM. *Journal of Statistical Software*, 74(7):1–16, 2016.

[15] L. R. Landerman, K. C. Land, and C. F. Pieper. An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values. *Sociological Methods & Research*, 26(1):3–33, 1997.

[16] J. F. Lineback and E. B. Fink. Recent Developments in Assessing and Mitigating Non-response Bias. *ICES-IV Proceedings*, 2012.

[17] R. Little. Missing Data Adjustments in Large Surveys. *Journal of Business and Economic Statistics*, 6(3):287–296, 1988.

[18] R. J. Little. Missing Data in Census Bureau Surveys. *Proceedings of the Second Annual Census Bureau Research Conference. Washingto, DC: U.S. Department of Commerce, Bureau of the Census.*, pages 442–454, 1986.

[19] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data, 2nd edition.* Wiley and Sons, 1987.

[20] B. J. Meekins and R. L. Sangster. Predicting Wave Nonresponse from Prior Wave Data Quality. *Proceedings of the American Statistical Association*, 2014.

[21] D. Miller. Imputations Reaction to Data: Exploring the Boundaries and Utility of IVEware and Iterative Sequential Regression (ISR). *Proceedings of the Fifth International Conference on Establishment Surveys, June 20-23, 2016, Geneva, Switzerland: American Statistical Association.*, 2016.

[22] D. A. Penn. Estimating missing values from the general social survey: An application of multiple imputation. *Social Science Quarterly*, 88(2):573–584, 2007.

[23] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2016.

[24] T. Raghunathan. *Missing data analysis in practice.* CRC Press, 2015.

[25] M. W. Robbins, S. K. Ghosh, and J. D. Habiger. Imputation in High-Dimensional Economic Data as Applied to the Agricultural Resource Management Survey. *Journal of the American Statistical Association*, 108:81–95, 2013.

[26] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* John Wiley &Sons, 1987.

[27] J. Sanchez. Sensitivity of Econometric Estimates to Item Non-response Adjustment. *Proceedings of the Fifth International Conference on Establishment Surveys*, 2016. June 20-23, 2016, Geneva, Switzerland: American Statistical Association.

[28] J. Sanchez. On Matching of LBD and BRDIS firms. September, 2015. U.S. Census Bureau, Center for Economic Studies technical memo. For internal use.

[29] J. Schafer. Multiple Imputation: a primer. *Statistical Methods in Medical Research*, 8:3–15, 1999.

[30] N. Schenker, T. E. Raghunathan, P. Chiu, D. M. Makuc, G. Zhang, , and A. J. Cohen. Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association, 101(475)*, pages 924–933, 2006.

[31] StataCorp. *Stata Statistical Software: Release 14.* College Station, TX: StataCorp LP., 2015.

[32] M. Templ, A. Alfons, and P. Filzmoser. Exploring Incomplete Data Using Visualization Techniques. *Adv. Data Anal Classif*, 6:29–47, 2012.

[33] S. Van Buuren. *Flexible Imputation of Missing Data.* Chapman & Hall/CRC, 2012.

[34] S. Van Buuren and K. Groothuis-Oudshoorn. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software.*

[35] T. White, J. Reiter, and A. Petrin. Plant-level Productivity and Imputation of Missing Data in the Census of Manufactures. *Center for Economic Studies, U.S. Census Bureau, Working Paper CES 11-02*, 2011.

[36] T. K. White, J. P. Reiter, and A. Petrin. Imputation in U.S. manufacturing data and its implications for productivity dispersion. 2016.

[37] R. Wolfe. U.S. Businesses Report 2008 Worldwide R&D Expense of $330 Billion: Findings from New NSF Survey. *Infobrief NSF 10-322*, 2010. Arlington,VA:National Science Foundation, Division of Science Resources Statistics.

[38] R. M. Wolfe. 2013 Business R&D and Innovation Survey (BRDIS) Methodology Report. *National Center for Science and Engineering Statistics, National Science Foundation.*

[39] R. M. Wolfe. NSF Announces New U.S. Business R&D and Innovation Survey. *NCSES Infobrief. NSF 09-304, National Science Foundation*, 2010.

[40] R. M. Wolfe. Businesses Spent $341 Billion on R&D Performed in the United States in 2014. *NCSES Infobrief. NSF 16-315, National Science Foundation*, 2016.