Imputation Methods for Nonresponse on Different Domains in a Business Survey

Frank Weideskog[1]

Abstract

Imputations often require access to auxiliary information or historical information from the survey itself to be able to provide any useful results. Imputations can substantially improve the quality of the statistics, but this requires that appropriate methods are used. In this presentation, where the focus is on business statistics, is the unit nonresponse estimated on an object level in a first step, but also estimations on different underlying domains are performed in a second step, using a two-step approach.

In a first step estimations from a set of imputation methods are calculated. The estimations from all imputation methods are compared to previous reported values for each estimated company. The estimated value for the method giving the minimum difference, are chosen the current period. Let Let $\hat{y}_{im}$ be the estimated total value for company i, period m. Also let $d_{im} = y_{im} - \hat{y}_{im}$ be the difference between the reported value y_{im} and the estimated value $\hat{y}_{im}$ for company i, period m. The average difference, for company i, of the absolute differences is considered.

The second step is divided into the situations where historical data is available or where historical data is missing. The allocation on domains is done using an allocation formula that is primarily determined by how the company previously reported and secondarily determined by how similar companies reported when historical reports are missing (model approach). The idea is to describe a two-step method approach and to show results that demonstrate the strength of this approach according to limit and measure bias in the estimations of nonresponse.

Key Words: nonresponse, imputation, business survey

1. Introduction

In many business surveys nonresponse occurs. In these cases often a lot of resources are added on costly re-contacts and written reminders sent to the non-reporting companies. In surveys where providers of statistical information (PSI's) are legally obliged to report, the nonrespondents also can end up in a prolonged penalty process with the County Administrative Court. Despite these efforts a not too insignificant loss often consists, although it may have been reduced slightly.

In order to estimate nonresponse a set of imputation methods can be used. Imputation can in this sense be regarded as a process of replacing missing data with substituted values. Unit nonresponse occurs when an object is missing and item nonresponse occurs when some variable values are missing for an object. Imputation refers mainly to limit the bias in the estimates. Imputation requires access to auxiliary information to be able to provide any useful results. Alternatively historical information from the survey itself can be used. Imputations can substantially improve the quality of the statistics, but this requires that appropriate methods are used. Imputations can be performed automatically (according to a decided algorithm), manually or by combinations of both approaches.

[1] Frank Weideskog, Process Department, Statistics Sweden, Karlavägen 100, 10451 Stockholm, email: frank.weideskog@scb.se.

This presentation regards unit nonresponse, where automatically estimations are performed. Situations where providers partly report are not covered in this text.

The unit nonresponse is estimated on an object level at a first step followed by estimations on different underlying domains at a second step (two-step approach):

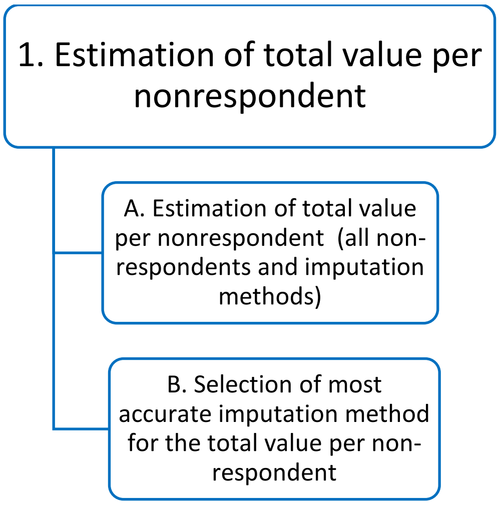
1. Estimation of total value per nonrespondent
2. Allocation of the estimated total value per nonrespondent on different domains

If the criteria when to use the methods are fulfilled, each nonrespondent will have estimations from the imputation methods in progress. The estimates among all imputation methods could be compared to how accurate the method is according to how precise previous months have been estimated.

The allocation of the estimated value on company level can either be done using previous reported observations (history is available), or using different approaches when history is not available.

2. Methodological approach

In a first step estimations from a set of imputation methods are performed. Estimations for previous periods are then compared to previous reported values for each estimated company. The estimated value for the method giving the minimum absolute average difference, are chosen for the current period.


1. Estimation of total value per nonrespondent

A. Estimation of total value per nonrespondent (all non-respondents and imputation methods)

B. Selection of most accurate imputation method for the total value per non-respondent

A number of different automatically performed imputation methods, such as the following categories of methods, can be used according to out approach:

- Projection and regression methods
- Imputation using auxiliary information
- Mean value imputation methods
- Growth rate methods

Not every option is applicable and if several options are applicable, they will not result in a similar estimate for the particular nonrespondent. The option to apply depends on the characteristic of data and on the availability of time series. For instance, a regression (or forecasting) approach requires a minimum number of observations.

The main reason for using more than one imputation method in our approach is to maximize the reliability in above all early estimates. If several methods are applicable, there is a need to define criteria to select only one of them.

2.1 Projection and regression methods

Projection methods

Exponential Smoothing (ES) covers a set of projection methods that might work in the business statistics area. The idea is that the series varies around some smooth curve that might be considered as the true level which may be time varying. The actual observations apart from this true level are also affected by other irregularities which mean that the true level is unobserved. ES is known to function well for forecasts over shorter periods. In those cases where nonresponse occurs for more periods further back in time, ES models may function worse.

Let y_{im} be the total of a main aggregate (for instance flow) for company *i*, period *m*.

y_{im} is then estimated by

$$\hat{y}_{fim} = \propto y_{i(m-1)} + (1-\propto)\, \hat{y}_{fi}(m-1) \qquad (1)$$

where $y_{i(m-1)}$ is the reported total of an aggregate for company *i*, period *m-1*, $\hat{y}_{fi}(m-1)$ is the estimated total of an aggregate for company *i*, period *m-1*, and α is a parameter value than can be set between 0 and 1, 0 < α < 1, where α = 0.2 is the default value in SAS (Proc Forecast procedure).

Formula (1) can also be added by a season component $\hat{s}_m$

y_{im} is then estimated by

$$\hat{y}_{fsim} = \hat{y}_{fim}\, \hat{s}_m \qquad (2)$$

where $\hat{y}_{fim}$ is the estimated total according to (1) for company *i* for current period *m* and $\hat{s}_m$ is the seasonal component for period *m*.

Using “Proc forecast” the parameter α is determined in advance, as a constant, which makes the estimates less precise. When using the procedure “Proc ESM” in SAS α automatically can be estimated (*see SAS user guide 9.4*). For multiplicative models the logarithmic transformation is done before the estimation. For this reason no zero values or negative values are permitted. Apart from the ES methods there are other methods to consider, such as ARIMA models and regression models, even though ARIMA models might not work well when the time series include zero values.

Regression models

A simple linear regression model for a total of a main aggregate (for instance flow) for company *i*, period *m*, y_{im} can be applied as:

$$\mathrm{y}_{\mathrm{im}} = \beta_0 + \beta_1 \cdot \mathrm{x}_{\mathrm{im}} + \mathrm{e}_{\mathrm{im}} \qquad (3)$$

Where x_{im} is an auxiliary variable value for company *i* period *m* and e_{im} is an error term assumed to be independent of earlier error terms (and also normally distributed).

The slope can be calculated as:

$$\beta_1 = \frac{\sum_{i=1}^{N} x_i{}^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} x_i y_i}{N \cdot \sum_{i=1}^{N} x_i{}^2 - \left(\sum_{i=1}^{N} x_i \right)^2}$$

and the intercept:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

The auxiliary variable *X* should be strongly correlated with the depending variable *Y*, the two sources must be transparent and the part of data that by definition show discrepancies should be excluded or trimmed.

Except for the common situation with an ordinary regression model, where another source is used as auxiliary variable, also an autoregressive (AR) model on data from the previous year may be of some help. Then the error term is assumed e_{im} to no longer be independent, but an AR series:

$$y_{im} = \beta_0 + \beta \cdot y - 1_{im} + e_{im} \tag{4}$$

For further reading:
Cochran,W (1977). Sampling Techniques. New York: J. Wiley & Sons.
Lundström, S, Särndal C (2001). Estimation in the presence of nonresponse and frame imperfections. Statistics Sweden
Särndal C, Swensson B, Wretman J (1992). Model Assisted Survey Sampling.

2.2 Imputations using auxiliary information

To utilize the auxiliary information one might particularly emphasize the importance of necessary exclusions or transformations of auxiliary data to avoid discrepancies caused by definition among the target variable and the auxiliary variable. Furthermore there should be a strong relationship between the target variable and the auxiliary variable giving accurate estimations using the auxiliary variable.

$$\hat{y}_{im} = x_{im} \tag{5}$$

where x_{im} is a value to an auxiliary variable for company *i*, period *m*.

2.3 Mean value imputation methods

The method can be used in different ways, with or without a seasonal component. Using this type of mean value imputation one should be cautious with setting up criteria for when the method should be used. If there are only a small number of reported previous periods to consider then the method might not be used. If the method is used for single imputation like in this situation this method should probably have lower priory than other imputation methods, and be used only when there are no other sources/methods for estimating the nonresponse.

y_{im} is estimated by

$$\hat{y}_{(a)im} = \bar{z}_{im} \;\; \hat{s}_m \tag{6}$$

where:

$\bar{z}_{im} = \frac{1}{n}\sum_{t=m-k}^{m-1} z_{it}$ is the mean value the last *k* periods, $\hat{s}_m$ is the seasonal component for period *m* and *n* is the number of non-missing observations for company *i* during the previous *k* periods.

2.4 Growth rate methods

The growth rate method is an alternative method that takes into account the temporal change (growth) in the data. The method is fairly simple, and is based on the growth in value between time points. The basic idea is to use data collected from companies both the current period *(t)* and a previous period *(t-k)* at the same time. A company must have received both periods to belong to the target group. Then the ratio between the incoming data for the current quarter and the incoming data for the previous quarter can be calculated. The ratio is then multiplied with the total value of the company for period *t-k.*

$$\hat{G}_t = \frac{\sum y^t_{i,t}}{\sum y^t_{i,t-k}}$$ = Growth rate factor between period *t* and *t-k* (7)

$$\hat{T}_{i,t} = y_{i,t-k} * \hat{G}_t$$ = Estimated value for company *i* and period *t* (8)

D^t_t Companies reporting period *t*

y^t_t Sum of values from companies belonging to D^t_t

$D^t_{t\text{-}k}$ Companies reporting period *t-k*

$y^t_{t\text{-}1}$ Sum of values from companies belonging to $D^t_{t\text{-}k}$

2.5 Selection of most accurate imputation method for total value per nonrespondent

In order to optimize the reliability of our methods the estimated values for each of the methods for previous periods are compared to the reported values. The idea is to choose the best method for each unique nonrespondent for a given period.

Let $\hat{y}_{im}$ be the estimated total value for company *i*, period *u*. Also let $d_{im} = y_{im} - \hat{y}_{im}$ be the difference between the reported value y_{im} and the estimated value $\hat{y}_{im}$ for company *i*, period *u*. The average difference, for company *i*, of the absolute differences is considered. The estimation of the method giving the lowest absolute differences is selected.

The average difference for the *p* previous periods, for company *i*, of the absolute differences can then be written as:

$$\bar{d}_i = \frac{\sum_{u=m-p}^{m-1} \left| d_{iu} \cdot V_{iu} \right|}{\sum_{u=m-p}^{m-1} V_{iu}} \qquad (9)$$

where $V_{iu} = 1$ if the company has both estimated and reported values for period u and

$V_{iu} = 0$ otherwise. If an estimated or reported value is lacking in any method or period for a company, the difference d_{iu} is set to ”missing value” for the period. The average difference $\bar{d}_i$ is only calculated if at least a decided number of the differences d_{iu} are different from ”missing value”. The smallest absolute difference, $\bar{d}_i$, among the compared imputation methods is regarded as best methods for the company the actual period, and will be selected as estimator.

If an auxiliary variable generally are close to the variable to estimate, and there is also a value for the auxiliary variable, then it could be prioritized in the selection of estimation method, preferably according to some quality criteria. If the auxiliary variable (X) do not deviate too much from the variable to estimate (*Y*), for an example according to the coefficient of variation (cv), it should be selected. The *cv* for a period is here defined as the standard deviation of the difference in the numerator and the mean value of the difference in the denominator, and can be defined as:

$$cv_i = \frac{s_{di}}{\bar{d}_i} \qquad (10)$$

2.6 Nonresponse error

According to the book “Margins of Error” by Duane F. Alwin, the nonresponse error in a survey is defined as “Error that results from the failure to obtain data from all population elements selected into the sample”

To estimate the nonresponse error we make the adoption that the latest reported value of a company is the true value.

The nonresponse bias *e* for company *i* period *j* can then be denoted as:

$$e_{ij} = y1_{ij} - ys_{ij} \qquad (11)$$

Where:

$y1_{ij}$ = Estimated nonresponse value for company *i* period *j* at the first time point
ys_{ij} = Reported value for company *i* period *j* at the last time point in the comparison, which did not report at first time point (true value)

The relative bias *e_relij* is defined as the nonresponse bias *e* divided by the reported value ys_{ij} for company *i* period *j* at the last time point.

The relative absolute bias, can then be denoted as:

$$\left|e\_rel_{ij}\right| = \frac{\sum_1^N \left|e_{ij}\right|}{\sum_1^N ys_{ij}} \qquad (12)$$

2.7 Example 1

To illustrate an example, in the Swedish FTG (Foreign Trade of Goods statistics) twelve different automatically imputation methods are used to estimate the nonresponse in the Intrastat system (EU-trade on goods) by both projection and regression methods, auxiliary information, and a simple mean value imputation method. As auxiliary information monthly VAT data on company level from the Swedish Tax Agency is used.

The VAT data is prioritized and directly selected as estimator, but only if the coefficient of variance (see equation 10) is low according to the difference between the Intrastat value and the VAT value.
If the methods compared in the selection step concern missing estimated values due to missing auxiliary data or when criteria is not fulfilled to use a method for a certain nonrespondent, estimations from the mean value method is used. The mean value method here is carefully performed and should only be used for smaller companies, serving as an additional method not included in the automatically selection step (see equation 9).

The relative nonresponse error, here defined as the relative absolute error is calculated on the three months in last quarter of arrivals 2014, and can be seen in table 1:

Table 1:
Relative nonresponse error of arrivals in the Swedish Intrastat survey 201410–201412 at first publication, automatic selection of most accurate imputation method.

Category of imputation method	Percent of value	Relative bias in percent	Nonresponse error in percent
Projection and regression methods	8.5	-2.5	68.4
Auxiliary information from VAT data	90.0	3.2	23.1
Mean value imputation method	1.5	-12.6	76.8
Total	**100.0**	**2.5**	**28.0**

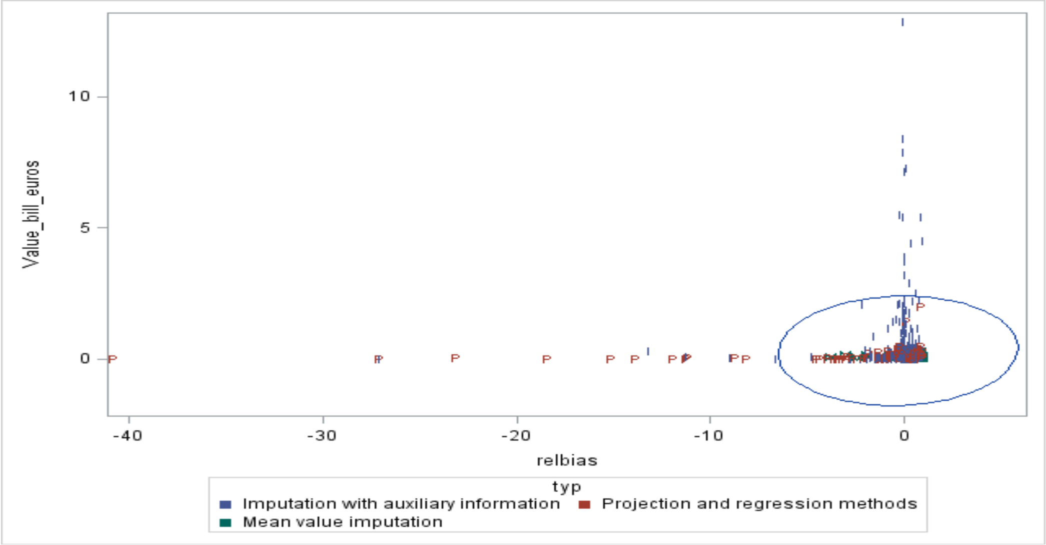
Table 2:
Relative nonresponse error of arrivals in the Swedish Intrastat survey 201410–201412 at first publication, without an automatic selection of most accurate imputation method.

Category of imputation method	Percent of value	Nonresponse error in percent, based on available estimations	Nonresponse error in percent, based on all estimations
Projection and regression methods	70.5	40.0	92.8
Auxiliary information from VAT data	90.5	23.2	44.0
Mean value imputation method	31.8	60.2	273.3

The relative nonresponse error of arrivals in Intrastat 201410–201412 is 28 percent, where 90 percent of the estimated value regard auxiliary information from VAT data giving most accurate estimations among the methods. However not all (90.5 percent) of the nonrespondents have reported VAT values at the first publishing. If only VAT-values had been used in the estimations would the relative nonresponse error increase to 44 percent (table 2). Table 2 also show the accuracy of the percent available estimations; for instance it can be found that the relative nonresponse error for the projection and regression methods amounts to 40 percent.

Figure 1 illustrate the spread on the estimates among the three categories of estimation methods; Projection and regression methods (P), Imputation by auxiliary information from VAT data (I) and the mean value imputation method (M). A prediction ellipse is helpful for detecting deviation from normality. Because the center of the ellipse is the sample mean, a prediction ellipse can give a visual indication of skewness in the data. In figure 1 a 95% prediction ellipse indicates a region that would contain about 95% of a new sample that is drawn from a bivariate normal population with mean and covariance matrices that are equal to the sample estimates.

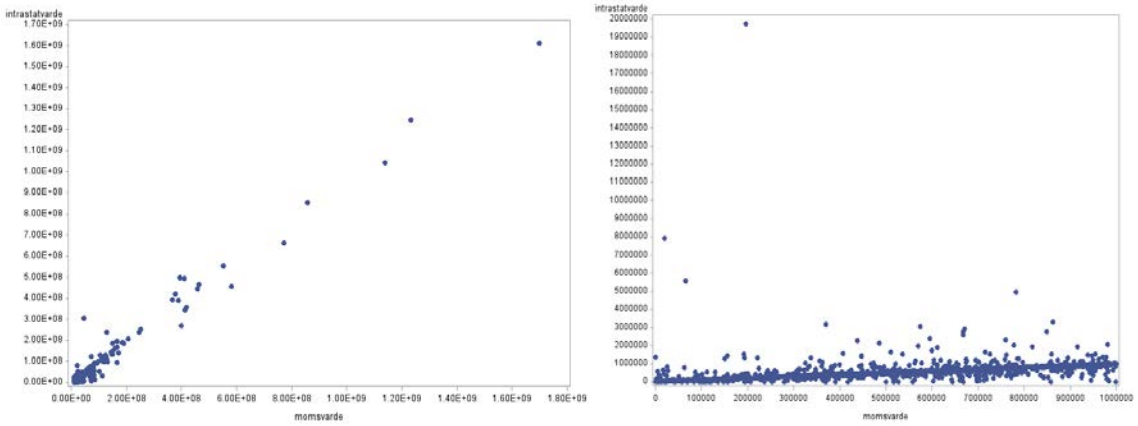
Figure 1
Scatter plot of imputations and relative bias for the three categories of methods



The correlation between the Intrastat value (y) and the VAT-value (x) is found to exceed 90 percent in both flows in comparison to all companies, indicating a very strong positive correlation between y and x. In Figure 2a-c, one can see the positive relationship between y and x.

Overall, it appears that the relationship between y and x is greater the larger the VAT value of the company (Figure 2). The correlation among the smallest companies is lower, while the largest companies show a very high correlation. It is likely that the relative difference is much greater among smaller companies than among larger.

Figure 2
Correlation between the Intrastat value (y) and the VAT value (x) in the arrivals 2014, size of providers



Group of largest providers

Group of smallest providers

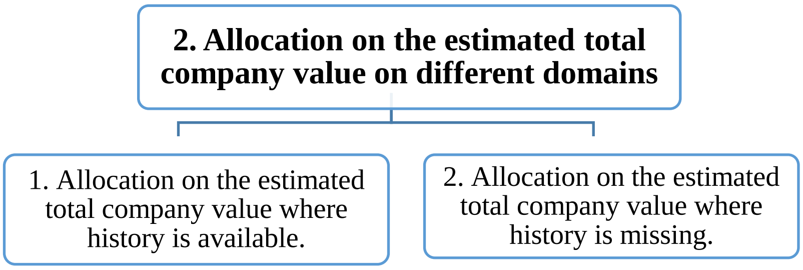
For further reading:
Weideskog, F. (2010). Improvement of the estimation methods in the Swedish Intrastat

system. Statistics Sweden

2.8 Allocation of the estimated total company value on different domains

The allocation on domains (step 2) can be done using an allocation formula that is primarily determined by how the company previously reported and secondarily determined by how similar companies reported when historical reports are missing (model approach).

The second step of the operation can be divided into the following situations:



2. Allocation on the estimated total company value on different domains

1. Allocation on the estimated total company value where history is available.

2. Allocation on the estimated total company value where history is missing.

As for the selection of estimation method on total company level, an automatic selection of best allocation method on a domain of interest could also be set up particularly in cases where history is missing. As an alternative method to compare with the method described in section 2.8.2, an allocation key based on reporting companies in the same size class as the nonrespondent could be used when calculating the allocation key.

If the allocation of the estimated total company regard cases where history is available one should take into account possible seasonal effects, and group the historical data in domains with pronounced seasonal effect or not before using it as auxiliary information in the estimations. For some companies an allocation key based on reported data concerning the last period earns as a good estimate, for others it might be better to use more historical periods or the same period for the previous year.

2.8.1 Allocation of the estimated total company value where history is available

The decision whether a company can use its own history or not, could be based on a requirement of a minimum number of reported periods. The condition could also be combined with a requirement criterion of a maximum of a number of most previous *p* periods to be included to calculate an allocation key for the company. For each company determined an allocation key where each value in the key is calculated as the ratio of the value of a domain, and the total value. These calculated keys can then be used for the company's estimated total value for the period to be estimated.

Suppose that company *i* has reported values for month *m-p* to month *m-1* and let $y_{ij(m-u)}$ denote the value that the company had for periods *m-u* to be domain *j*. The share of value that will be distributed on domain j, p_{ijm} is estimated by:

$$\hat{p}_{ijm} = \frac{\sum_{u=1}^{p} y_{ij(m-u)}}{\sum_{u=1}^{p}\sum_{j=1}^{J} y_{ij(m-u)}} \qquad (13)$$

The allocated value for company *i* in the domain *j* in period *m*, y_{ijm} is then estimated by

$$\hat{y}_{ijm} = p_{ijm}\hat{y}_{im} \qquad (14)$$

where $\hat{y}_{im}$ is the estimated total for company *i* in the current period to estimate.

2.8.2 Allocation of the estimated total company value where history is missing

In the situation where companies do not have sufficient historical information according to a set up requirement criterion, or have no previous reports at all, the total estimated value for all these companies could be allocated on domains according to a principle of "similar companies". The groups are made up of companies in the same domain and size class, where an allocation key for each group is produced. A measurement of size that could be used can thus be the annual turnover value of the company, the annual VAT value or the number of employed people in the company etc.

For each group of responding companies, an allocation key is determined with shares per domain. Every value in each key is calculated as a ratio between two sums for the company group, the collected value of the domain *j* as the numerator and the value of all collected values as the denominator. A group of responding companies is denoted by *g* and has a common key f_{gi} as shown below, where the sum over *i* refers to companies that belong to the group (group affiliation is shown by ∈).

$$f_{gj} = \frac{\sum_{i \in g} y_{ij}}{\sum_{i \in g} \sum_{j=1}^{J} y_{ij}} = \sum_{i \in g} w_i p_{ij} \quad \text{, for j=1,…..,J} \qquad (15)$$

where $w_i = \frac{t_{i}\cdot}{\sum_{i \in g} t_{i}\cdot}$ and $p_{ij} = \frac{y_{ij}}{\sum_{j=1}^{J} y_{ij}}$,and $t_i\cdot$ denotes the total value for company *i*,

thus $t_i \cdot = \sum_{j=1}^{J} y_{ij}$ (16)

The group key is a weighted average of the individual keys. Company *i* has a weight w_i that is proportional to its value, and is equal to its share of the group's total value.

Creating homogenous estimation groups

An allocation key between the reported value based on the companies classified branch of industry (NACE) in the numerator and the value on a domain linked to the NACE nomenclature (denominator) could be denoted as:

$$p_{ij} = \frac{y_{ij}}{\sum_j y_{ij}} \qquad (17)$$

where y_{ij} is the collected value according to the branch of industry or NACE combined with size class as indicated with *i* and the domain of interest (indicated with *j*). In order to create homogeneous groups based on p_{ij}, the cluster analysis procedure PROC FASTCLUS in SAS could be an option to use. The procedure find initial clusters with an iterative algorithm to minimize the sum of squared distances from a cluster mean. The method is based on sorting by next centroid. A set of points known as "cluster seeds" is chosen as the basis for clusters. Each observation can be assigned to the nearest cluster seed to form temporary clusters. The seeds can then be replaced by new temporary clusters, and the process repeated until no more changes occur in the clusters. The centroid method supply the distance between two clusters defined as Euclidean distance (squared distance) between their centroid and the mean value. The Euclidean distance can be expressed as:

$$D(x, y) = (\bar{y} - \bar{x})^2 . \qquad (18)$$

In the procedure a maximum number of seeds (and thus clusters) can be set according to a MAXCLUSTERS option. To determine the best number of clusters, three of the context relevant statistics, the R square value (R^2), F-pseudo-test and the Cubic Clustering Criteria (CCC) can be concerned. The R^2- value for each variable indicates how important the variable is for the cluster. The expected R^2 -value of the total measure under the uniform null hypothesis indicated, given that the variables are uncorrelated. To test the separation of all clusters and the number of selected clusters F-pseudo statistics are used:

$$\text{F-pseudo} = (R^2 / (C-1)) / (1 - R^2)) \qquad (19)$$

where:

c = number of clusters

n = number of observations.

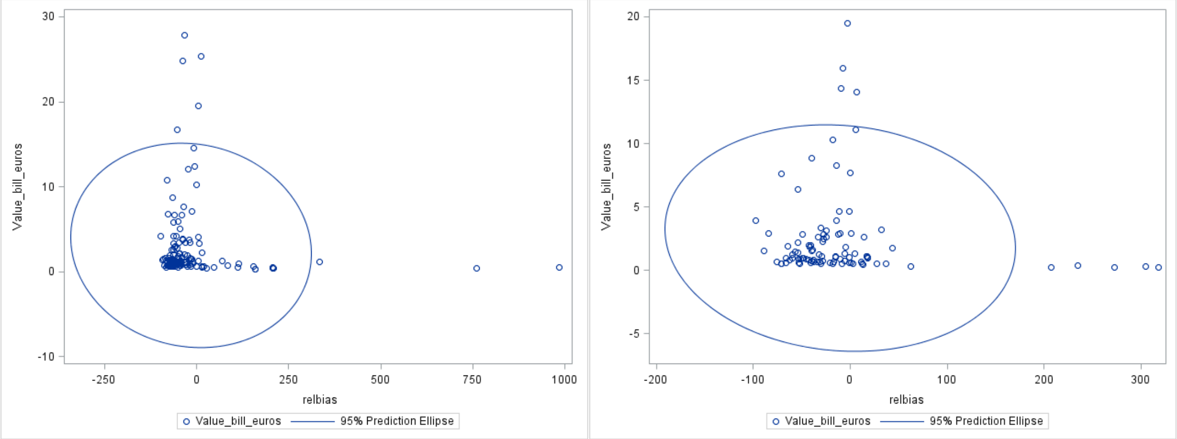
High F-values indicate that the number of selected clusters is acceptable and that the separation between all clusters is high.

For further reading the following sources can be suggested:
Sharma, S (1996). Applied Multivariate Techniques. J. Wiley & Sons.
Eurostat (2010) Summary Quality Report (for ETGS) - Edition 2010
Eurostat (2012) Quality Handbook (for ETGS) – Volume III
SAS Institute (2015). SAS user guide 9.4

2.9 Example 2

Another example from the Swedish Intrastat survey regards allocation of the estimated total company value on different domains, and the distinction whether the company have history or not. Figure 2 illustrate the spread on the estimates on CN2-level with a 95% prediction ellipse indicating a region that would contain about 95% of a new sample.

Figure 2
Scatter plot of estimated value and relative bias in the estimation of the nonresponse



Distributed trade per CN2 (all estimated trade) **Distributed trade per CN2 (trade with history)**

In the Swedish Intrastat survey the requirement is that at least three reported month are available during the last twelve months for history being concerned. In 2014 about 70 percent of the nonresponse value relates to company with history and 30 percent is estimated where history is missing in the arrivals of Intrastat.

The relative nonresponse error is based on the same months as the first practical example of arrivals 2014 (see section 2.7), and the domain of interest has been chosen to a two digit commodity code according to the Combined Nomenclature (CN).Very small estimated values on CN2-codes are excluded in the analysis.

Of the 52 CN2-codes in the analysis of all estimated nonresponse about half of the codes are in the interval ‘50-100 percent’ according to the relative nonresponse error (table 3). The lowest CN2-code correspond 6 percent relative nonresponse error and the highest 296 percent. The unweighted mean value for the relative nonresponse error on a CN2-code is 64 percent. Studying the relative nonresponse error only for the estimations regarding companies where history can be used the mean value for a CN2-code is shown to be estimated to 44 percent.

Table 3
Relative nonresponse error at the CN2-level in the Swedish Intrastat survey 201410-201412

Relative nonresponse error	Frequency	Percent
0 – 10 %	3	5.8
10 – 50 %	19	36.5
50 – 100 %	25	48.1
> 100 %	5	9.6
Sum of total	**52**	**100.0**

For further reading:
Weideskog, F. (2012). Improvement of the distribution keys for the estimated trade in the Swedish Intrastat system. Statistics Sweden.

3. Conclusions

This paper addresses appropriate imputation methods for nonrespondents on different domains in a business survey. The common idea is that a nonrespondent is estimated on total company value on each of the alternative methods as in a first step. Auxiliary information from another source showing high correlation with the value to report in the reference survey often works quite well as an estimator. Through an automatic selection the most reliable estimation method for the company for a given reporting period is chosen. Then in a second step the estimated total company value is allocated on different domains. At this stage, one can distinguish between companies own history of any criterion and those missing history. If there are no history a model approach will be needed, such as using the information about how a company is classified in the business register and compare and relate this classification to any appropriate reported domain. Since there are nonrespondents classified by NACE codes in the Swedish business register, it may be a good idea to use this information and perform a cluster analysis. The NACE aggregates are divided in various homogenous estimation groups where even size of the company is considered. Experience based on analysis of process data from the Swedish Intrastat survey show that the total nonresponse error (28 percent) is 1.5 times larger when using only one single method (the most accurate of the methods in the survey) instead of using the best method, according to an automatically selection from a set of estimation methods. To use our described practical approach one should be cautious with different criteria that are necessary to implement for the estimation models and the selection criteria of which model to choose. Above all the larger PSI's should be estimated with carefully set up criteria combined with manually controls of the output.

The relative nonresponse error using our practical method approach is on average around 60 percent on a selected main domain in the Swedish Intrastat survey. Studying the relative nonresponse error only for the estimations regarding companies where history can be used the corresponding figure is estimated to around 40 percent.

References

Cochran,W (1977). Sampling Techniques. New York: J. Wiley & Sons.

Särndal C, Swensson B, Wretman J (1992). Model Assisted Survey Sampling.

Lundström, S, Särndal C (2001). Estimation in the presence of nonresponse and frame imperfections. Statistics Sweden

Alwin.D (2007). Margins of Error. J. Wiley & Sons.

Groves.R (2004). Survey Errors and Survey Costs. J. Wiley & Sons.

Sharma, S (1996). Applied Multivariate Techniques. J. Wiley & Sons.

Eurostat (2011). European statistics code of practice

Eurostat (2010) Summary Quality Report (for ETGS) - Edition 2010

Eurostat (2012) Quality Handbook (for ETGS) – Volume III

Weideskog, F. (2010). Improvement of the estimation methods in the Swedish Intrastat system. Statistics Sweden

Weideskog, F. (2012). Improvement of the distribution keys for the estimated trade in the Swedish Intrastat system. Statistics Sweden.

SAS Institute (2015). SAS user guide 9.4