WP. 31
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (vi): New and emerging methods

MASS IMPUTATION FOR BUILDING A NUMERICAL STATISTICAL DATABASE

Supporting Paper

Prepared by Natalie Shlomo, Southampton Statistical Sciences Research Institute, University of Southampton; Ton De Waal and Jeroen Pannekoek, Statistics Netherlands

I. INTRODUCTION

1. Statistical agencies are making more use of administrative data in their statistical processes with the aim of improving the accuracy and reliability of their statistical data while reducing response burden and cutting costs. Administrative data, however, are collected for non-statistical purposes and suffer from general quality problems such as differences in definitions, timeliness and under and over coverage. To incorporate administrative data in statistical systems, agencies are investing resources in the research and development of edit and imputation processes for administrative data and the enhancement of statistical databases.

2. Assuming that survey data can be directly linked to the administrative data, the question of how to construct a complete statistical database from multiple data sources has been researched by several statistical agencies. In particular, work by Kroese, Renssen and Trijssenaar, 2000 compared mass imputation techniques for imputing non-sampled units to a procedure based on 'repeated weighting' for obtaining a set of mutually consistent population tables of interest. As the authors pointed out, the requirements for the statistical system should ensure:

- Reliable estimates with respect to using sound design and model based theory of estimation,
- Logically consistent estimates, i.e. edit constraints preserved

3. The focus of the work by Kroese, Renssen and Trijssenaar, 2000 was to develop a social statistic database that could be used to replace the traditional Census in the Netherlands. Their conclusion was that considering the large number of variables in the social statistic database, most of which are categorical, the 'repeated weighting' technique would ensure reliable estimates and better preserve the relationships between variables. Houbiers, 2004 described the 'repeated weighting' technique to construct the social statistic database. In that paper, and in a subsequent paper by Van de Laar, 2004, the problem of failed edit constraints were briefly discussed.

4. In contrast to a social statistic database, creating a statistical database for businesses is more often carried out by mass imputation techniques (see Kovar and Whitridge, 1995 and references therein for more details). We assume the existence of a business register typically derived from administrative sources, such as VAT and Social Security. In addition, we assume that business surveys are carried out and that the sampled units in the surveys can be directly linked to the business register. It is clear that using 'repeated weighting' techniques to obtain a series of consistent tables for business statistics might not be adaptable to the case of magnitude tables which are typically produced as business statistics outputs. In addition, 'repeated weighting' would suffer from the usual problems associated with

regression estimators for business statistics where the models are highly sensitive to errors in the covariates. Mass imputation, on the other hand, has been used in some countries to create a business statistics database when missing-at-random arguments can be assumed. The method makes more use of auxiliary variables for developing imputation models compared to survey weighting and can be made robust to outliers and other potential errors in the covariates. In addition, the variables are typically numerical and the number of variables small. The edit constraints are simple to define and can be expressed in terms of linear restrictions. One example of an edit constraint:
$a_1x_1 + ... + a_nx_n + b \geq 0$ where $a_j \quad (j = 1,..,n)$ and b are constants which define the edit.

5. In Pannekoek, Shlomo and DeWaal, 2008 algorithms are described that impute for numerical variables while benchmarking totals and preserving edit constraints. In this paper, we apply one of the algorithms to a generated simulation dataset based on a sequential imputation approach and using a modified regression model. The aim is to compare resulting estimates to the 'true' original values known in the simulation dataset. We demonstrate how this approach can be used to carry out mass imputation to create a complete business statistics database under missing at random assumptions, and show that the estimates obtained from the mass-imputed database are superior to using single survey weighted estimates. The technique ensures reliable estimates since it uses a sound model-based approach to estimation where sufficient statistics are preserved and correlations approximately preserved. The consistency of the estimates is guaranteed as well since the technique preserves edit constraints in the data.

6. In Section 2 we review the modified regression model that we will use in the simulation study which benchmarks totals and preserves edit constraints. In Section 3 we describe the simulation study and compare the estimates obtained from the mass-imputed dataset and the weighted survey estimates obtained from a single survey to the 'true' values in the generated dataset. We conclude in Section 4 with a discussion.

II. MODIFIED REGRESSION ALGORITHM FOR MASS-IMPUTATION

A. Adjusted Predicted Mean Imputation

7. The idea of this algorithm is to obtain predicted mean imputations that satisfy the sum constraint and then adjust these imputations such that they also satisfy the interval constraints. To illustrate this idea we use a simple regression model with one predictor but generalisation to multiple regression models is straightforward.

A.1 Introducing some notation by the example of standard regression imputation

8. Suppose that we want to impute a target column $\mathbf{x}_t$ using as a predictor a column $\mathbf{x}_p$. The standard regression imputation approach is based on the model:

$$\mathbf{x}_t = \beta_0 + \beta\mathbf{x}_p + \varepsilon,$$

We assume that the predictor is either completely observed or already imputed, so there are no missing values in the predictor anymore. There are of course missing values in $\mathbf{x}_t$ and to estimate the model we can only use the records for which both $\mathbf{x}_t$ and $\mathbf{x}_p$ are observed. The data matrix for estimation consists of the columns $\mathbf{x}_{t.obs}, \mathbf{x}_{p.obs}$, where *obs* denote the records with $\mathbf{x}_t$ observed (and *mis* will denote the opposite) With the OLS estimators of the parameters, $\hat{\beta}_0$ and $\hat{\beta}$ we obtain predictions for the missing values in $\mathbf{x}_t$ using

$$\hat{\mathbf{x}}_{t.mis} = \hat{\beta}_0 + \hat{\beta}\mathbf{x}_{p.mis},$$

where $\mathbf{x}_{p.mis}$ contains the x_p-values for the records with $\mathbf{x}_t$ missing and $\hat{\mathbf{x}}_{t.mis}$ are the predictions for the missing $\mathbf{x}_t$-values in those records. The imputed column $\tilde{\mathbf{x}}_t$ consists of the observed values and the predicted values filled in for the missing values $\tilde{\mathbf{x}}_t = (\mathbf{x}_{t.obs}^T, \hat{\mathbf{x}}_{t.mis}^T)^T$, where T denotes the transpose.

9. These imputed values will not satisfy the sum constraint but a slightly modified regression approach can ensure that they do and will be described next.

A.2 Extending the standard regression imputation to satisfy the sum-constraint

10. This approach adds to the observed data the known totals of the missing data for the target variable as well as the predictor. These totals are $X_{p.mis} = X_p - \sum_i x_{p.obs,i}$ and $X_{t.mis} = X_t - \sum_i x_{t.obs,i}$, respectively. The total $X_{t.mis}$ is added to the column $\mathbf{x}_{t.obs}$ and the total $X_{p.mis}$ is added to the column $\mathbf{x}_{p.obs}$. Furthermore, the regression model is extended with a separate constant term for the record with the totals of the missing data. The model for these observed data can then be written as

$$\begin{aligned} \mathbf{x}_{t.obs} &= \beta_0 + \beta \mathbf{x}_{p.obs} + \varepsilon \\ X_{t.mis} &= \beta_1 m + \beta X_{p.mis} \end{aligned} \tag{1}$$

with *m* the number of records with missing values. We apply OLS to estimate the model parameters which will be used to predict and impute the missing values in $\mathbf{x}_t$, i.e.

$$\hat{\mathbf{x}}_{t.mis} = \hat{\beta}_1 + \hat{\beta}\mathbf{x}_{p.mis}, \tag{2}$$

and so the sum of the predicted values will equal

$$\hat{X}_{t.mis} = \sum_i \hat{x}_{t.mis,i} = m\hat{\beta}_1 + \hat{\beta} X_{p.mis}$$

11. In order to demonstrate the property of this model that the imputed values will sum up to the known total, we re-express the model for the observed with the known totals added as

$$\begin{bmatrix} \mathbf{x}_{t.obs} \\ X_{t.mis} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{x}_{p.obs} \\ 0 & m & X_{p.mis} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ 0 \end{bmatrix}$$

or

$$\begin{bmatrix} \mathbf{x}_{t.obs} \\ X_{t.mis} \end{bmatrix} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

If this model is estimated by ordinary least squares (OLS) estimation, the residuals are orthogonal to each of the columns of the model matrix **Z**. Thus, for the second column we obtain $m(X_{t.mis} - \hat{X}_{t.mis}) = 0$ and hence $\hat{X}_{t.mis} = \sum_i \hat{x}_{t.mis,i} = X_{t.mis}$ which implies that the sum of the imputed values equals the known value of this total.

A.3 Adjusting the regression imputations to satisfy the sum constraint and the interval constraints

12. Since the interval constraints have not been considered in obtaining the predicted values, it can be expected that a number of these predictions are not within their admissible intervals. One way to remedy this situation is to calculate adjusted predicted values defined by

$$\hat{\mathbf{x}}^{adj}_{t.mis} = \hat{\mathbf{x}}_{t.mis} + \mathbf{a}_t, \tag{3}$$

such that the adjusted predictions satisfy both the sum constraint (which is equivalent to $\sum_i a_{t,i} = 0$) and the interval constraints and the adjustments are as small as possible. One way to find such a value for $\mathbf{a}_t$ is to solve the quadratic programming problem

minimise $\mathbf{a}_t^T \mathbf{a}_t$, subject to $\mathbf{1}^T \mathbf{a}_t = 0$ and $\mathbf{l}_t \le \hat{\mathbf{x}}_{t.mis} + \mathbf{a}_t \le \mathbf{u}_t$,

or, we can minimize the sum of the absolute values of the $a_{t,i}$ instead and solve the resulting linear programming problem.

13. As a simple alternative we may consider the following algorithm which alternates between adjusting to satisfy the interval constraints and adjusting to satisfy the sum constraint.

14. This algorithm starts with $\mathbf{a}_t^{(0)} = 0$ and the predictions (3) satisfy the sum constraint but not necessarily the interval constraints. Each prediction outside its admissible interval will then be moved to the closest boundary value by an appropriate adjustment, which is the smallest possible adjustment to satisfy the interval constraints, i.e.

$$a_{t,i}^{(1)} = l_{t,i} - \hat{x}_{t.mis,i} \quad \text{if} \quad \hat{x}_{t.mis,i} < l_{t,i} \tag{4a}$$

$$a_{t,i}^{(1)} = u_{t,i} - \hat{x}_{t.mis,i} \quad \text{if} \quad \hat{x}_{t.mis,i} > u_{t,i} \tag{4b}$$

$$a_{t,i}^{(1)} = 0 \quad \text{if} \quad l_{t,i} \le \hat{x}_{t.mis,i} \le u_{t,i} \tag{4c}$$

15. The adjusted values $\hat{\mathbf{x}}_{t.mis}^{adj}$ will now satisfy the interval constraints but almost surely not the sum constraint, which is equivalent to saying that the $a_{t,i}^{(1)}$ do not sum to zero. To obtain adjustments that also preserve the sum constraint, we divide the *m*-units in three set; L_t, U_t, O_t, with numbers of elements m_L, m_U, m_O, according to whether the current adjusted value $\hat{\mathbf{x}}_{t.mis}^{adj}$ is on the lower boundary, upper boundary or neither boundary. Let the current sum of the $a_{t,i}^{(1)}$ be $S_t^{(1)}$ then, sum-to-zero adjustments can be obtained as

$$a_{t,i}^{(2)} = a_{t,i}^{(1)} - S_t^{(1)}/(m_U + m_O) \text{ for all } i \in U_t \cup O_t \text{ if } S_t^{(1)} > 0 \tag{5a}$$

or

$$a_{t,i}^{(2)} = a_{t,i}^{(1)} + S_t^{(1)}/(m_L + m_O) \text{ for all } i \in L_t \cup O_t \text{ if } S_t^{(1)} < 0 \tag{5b}$$

16. We add or subtract a constant to the $a_{t,i}^{(1)}$ to make them sum to zero, thereby taking care not to subtract anything from $a_{t,i}^{(1)}$'s that already set the $\hat{\mathbf{x}}_{t.mis}^{adj}$ on their lower boundary and not to add anything to $a_{t,i}^{(1)}$'s that already set the $\hat{\mathbf{x}}_{t.mis}^{adj}$ on their upper boundary. After this step it may be that some of the $a_{t,i}^{(2)}$ cause their corresponding $\hat{\mathbf{x}}_{t.mis}^{adj}$ to cross their interval boundaries. In that case both steps (4) and (5) must be repeated.

B. Regression Imputation with Random Residuals

17. It is well known that in general predictive mean imputations show less variability than the true values that they are replacing. In order to better preserve the variance of the true data, random residuals can be added to the predicted means. The adjusted predictive mean imputations considered in the previous section will also be hampered by this drawback because these adjustments are intended to be as close as possible to the predicted means and not to reflect the variance of the original data.

18. In order to better preserve the variance of the true data we start with the predicted values $\hat{\mathbf{x}}_{t.mis}$ obtained from (2) that already satisfy the sum constraint, and our purpose is to add random residuals to these predicted means such that the distribution of the data is better preserved and in addition both the interval and sum constraints are satisfied. These residuals serve the same purpose (satisfying the constraints) as the adjustments $a_{t,i}$ but in contrast to the $a_{t,}$ they are not as close as possible to the predicted means, they are intended to also reflect the true variability around these predicted means

19. A simple way to obtain residuals is to draw each of the *m* residuals by Acceptance/Rejection sampling from a normal distribution with mean zero and variance equal to the residual variance of the regression model. This means, by repeatedly drawing from this normal distribution until a residual is drawn that satisfies the interval constraint.

20. The residuals obtained by this AR-sampling may not sum to zero so that the imputed values do not satisfy the sum constraint. We may then adjust these residuals to sum to zero by the "shift" operation according to (5) after which it may be necessary to again adjust some of the residuals to also satisfy the interval constraint.

III. SIMULATION STUDY

A. Generating the Data

21. We generated 10,000 records for our simulated business statistics database with 4 continuous variables X_1, X_2, X_3 and P, and a categorical variable GR with 18 categories. All variables were generated using the Normal Distribution and hence linear regression modelling is appropriate. In practice, numerical variables for business statistics tend to be skewed and transformations are generally applied, such as the log transformation. The algorithm in section 2 can be modified to ensure benchmarked totals under a log transformation of the dependent variable as seen in Pannekoek, Shlomo and De Waal, 2008 but no transformation will be carried out in the simulation study. Table 1 presents the original correlation structure of the generated database used for the simulation study.

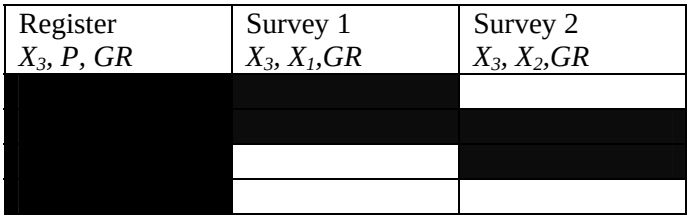
Table 1: Correlations between variables in simulated database

	X_1	X_2	X_3	P
X_1	1.00	0.56	0.47	0.41
X_2	0.56	1.00	0.83	0.60
X_3	0.47	0.83	1.00	0.71
P	0.41	0.60	0.71	1.00

22. In our simulation, we assume the structure of the business statistics database as depicted in Figure 1. We assume that the database contains two business surveys each with 1,000 units which are linked directly to a business register containing all units in the population (10,000 units). There are 500 units overlapping in both Survey 1 and Survey 2. We assume no unit non-response in this simulation and therefore need to impute for the non-sampled units only. Since the samples are drawn randomly, we can assume that imputation for the non-sampled units would follow missing-at-random assumptions. The register contains variables X_3, P, GR, Survey 1 contains X_3, X_1,GR and Survey 2 X_3, X_2,GR. We assume 2 scenarios to calculate the final weights in each of the surveys:

1. Post-stratification in each group GR where survey weights are defined as the total number of businesses in the register in the group divided by the number of businesses in the sample in the group.
2. Since variable X_3 is known in both the register and the surveys, a weight based on a ratio estimator can be applied. The survey weight is defined as the total of X_3 in the group divided by the total of X_3 in the sample in the group.

Figure 1: Structure of the simulated business statistics database



23. We note that to apply the modified regression technique for the mass-imputation of a business statistics database, we typically would not know the true totals of X_1 and X_2 in order to carry out the

benchmarking as described in section 2. As in the case of 'repeated weighting', we can use the weighted estimates from the respective surveys to obtain the totals for benchmarking. In this case, the variance of estimated parameters based on benchmarking totals that are themselves survey estimates needs to be modified to reflect this extra variation. Variance estimation for a mass-imputed dataset will not be addressed in this paper.

24. We assume the following edits in the database:

(1) $X_2 > X_3$
(2) $X_1 > 2X_2$
(3) $X_1 > 2X_3$

which can be expressed as: $X_1 > 2X_2 > 2X_3$.

B. Results of the Mass Imputation on the Simulated Dataset

25. In the first step of the modified regression technique we start with a prediction for X_2 based on the covariates P and X_3. We use the weighted survey estimate of the total of X_2 derived from Survey 2 for the benchmarking under both survey weight scenarios proposed in section III.A. We draw random residuals and adjust them so that the benchmarked total is preserved and the interval edit constraints satisfied. In the second step of the modified regression technique, we predict X_1 based on the covariates P and X_3 as before and also add the imputed X_2 from the previous stage. We use the weighted survey estimate of the total of X_1 from Survey 1 for the benchmarking. Again, we draw random residuals and adjust them so that the benchmarked total is preserved and edit constraints satisfied. We repeat the process multiple times under the sequential imputation approach.

26. Table 2 presents the percent relative difference of the correlations obtained from the mass-imputed business statistics database to the correlations of the 'true' generated variables in the simulation database (see table 1). We used the weighting scheme based on post-stratified weights to estimate the totals for benchmarking in the modified regression approach. Similar results were obtained when using survey weights based on the ratio estimator on variable X_3. As can be seen, the impact on the correlation structure in the mass-imputed business statistic dataset is not severely compromised compared to the 'true' correlations. Further analysis of the correlation structure within each group *GR* (not shown) also shows that the correlations are approximately preserved in the mass-imputed business statistics database.

Table 2: Percent relative difference in correlations between 'true' variables in the generated dataset and mass-imputed variables in the business statistics database

Variable	X_1	X_2	X_3	P
X_1	0	11.2	0	-5.3
X_2	11.2	0	1.2	-1.7
X_3	0	1.2	0	0
P	-5.3	-1.7	0	0

27. In Tables 3a and 3b, we compare estimates of the totals X_1 and X_2 across the grouping variable *GR* to the 'true' values of the totals obtained from the generated database. The estimates are obtained by aggregating the values of the variables in the mass–imputed business statistics database and by a weighted survey estimate of the total using a single survey under the different weighting scenarios (see Section III.A). Table 3a presents the percent relative differences for the aggregated variable X_1 in each group and Table 3b presents the percent relative differences for aggregated variable X_2. The shaded areas in the tables represent the groups that show improvement under the mass-imputation approach, i.e. the estimate is closer to the 'true' value. As can be seen in Tables 3, many more estimated

totals obtained by aggregating the mass-imputed business statistics database are closer to the 'true' value for each of the variables compared to the weighted estimates from the respective surveys.

Table 3a: Percent relative difference between 'true' totals to weighted survey estimates (according to weighting scenario) and totals derived from the mass- imputed business statistics database for X_1 (shaded areas show lower differences for the in mass-imputation approach)

Group *GR*	'True' Total X_1	Ratio Estimate Weights		Post Stratified Weights	
		Weighted survey estimate	Aggregated from imputed database	Weighted survey estimate	Aggregated from imputed database
Total	12,021,426	-0.48	-0.48	-0.14	-0.14
1	428,624	-3.41	0.52	-1.04	0.29
2	755,558	0.37	0.27	-0.32	0.79
3	589,707	1.40	0.77	0.93	1.48
4	737,671	0.38	0.37	0.76	0.99
5	571,375	-4.70	0.14	1.63	-0.03
6	607,233	1.10	0.11	-0.04	0.25
7	635,275	-0.12	0.24	-0.67	0.40
8	720,261	-1.35	-0.46	-0.34	0.00
9	723,848	-0.19	-0.12	0.29	0.31
10	729,229	0.85	-0.73	0.76	-0.42
11	708,591	0.44	-0.53	-0.41	-0.38
12	697,730	-5.70	-1.31	-2.58	-0.60
13	679,615	0.73	-1.69	0.05	-1.35
14	597,881	-3.72	-0.62	0.60	0.13
15	807,319	2.13	-0.91	-1.15	-0.64
16	635,074	1.81	-0.96	0.17	-0.78
17	587,479	-0.47	-1.69	-0.33	-1.28
18	808,949	-0.44	-1.45	-0.41	-1.21

28. Another quality measure for comparing the survey weighted estimates and the estimates obtained from the mass-imputed business statistics database to the 'true' values is to assess the impact of the estimation on a magnitude table. Assume we want to create a magnitude table spanned by the grouping variable *GR* (18 categories) and a categorized *P* variable (9 categories). In each cell of the table, we want the total X_1 and X_2, respectively. Knowing the 'true' values in the generated database, we can create the 'true' magnitude table and compare the cell values to the same table produced from the survey weighted estimates and the mass-imputed business statistics database. We use a quality measure defined by summing over the absolute relative difference in each of the cells of the table, calculated by: $ARD = \sum_c \frac{|D(c) - D(c')|}{D(c)}$ where $D(c)$ is the 'true' cell in the magnitude table and $D(c')$ is the estimated cell in the table depending on the estimation method: weighted survey estimates or estimates from the mass-imputed business statistics database. Table 4 presents the results of this comparison when using the post-stratified weighting scenario. Results for the ratio estimator weights are similar. As can be seen from the table, the mass-imputed business statistics database is far superior in producing a magnitude table that is closer to the 'true' magnitude table compared to using a single survey and weighted cell totals.

Table 3b: Percent relative difference between 'true' totals to weighted survey estimates (according to weighting scenario) and totals derived from the mass-imputed business statistics database for X_2 (shaded areas show lower differences for the in mass-imputation approach)

Group GR	True Variable X_2	Ratio Estimate Weights		Post Stratified Weights	
		Weighted survey estimate	Aggregated from imputed database	Weighted survey estimate	Aggregated from imputed database
Total	3,002,924	-0.03	-0.03	-0.13	-0.13
1	99,848	1.19	0.72	0.07	1.12
2	180,368	4.63	0.46	0.61	0.81
3	143,609	1.26	0.05	0.35	0.21
4	181,089	0.01	-0.48	-1.47	-0.31
5	140,708	1.20	0.29	0.25	0.30
6	149,260	-0.35	0.42	-1.37	0.19
7	157,303	-0.87	0.09	0.71	0.26
8	178,287	-1.26	0.39	-0.29	-0.04
9	180,927	-0.53	0.28	0.18	0.13
10	181,712	-0.21	-0.10	0.24	-0.29
11	177,533	-0.02	0.34	-0.35	0.02
12	174,843	-3.27	-0.33	-2.55	-0.44
13	170,982	2.06	-0.53	-1.18	-0.19
14	152,994	-1.06	-0.28	0.36	-0.63
15	206,497	-0.09	-0.20	-0.29	-0.61
16	163,496	-0.91	-0.47	0.71	-0.87
17	151,204	0.94	-0.43	0.59	-0.64
18	212,256	-2.05	-0.32	1.25	-0.55

Table 4: *ARD* measure comparing 'true' magnitude table spanned by *GR* (18 categories) and categorized *P* (9 categories) with survey weighted estimates and aggregations from the mass-imputed business statistics database

Cells contain sum of X_1		Cells contain sum of X_2	
Weighted survey estimate	Aggregated from imputed database	Weighted survey estimate	Aggregated from imputed database
347.5	1.92	467.2	1.63

IV. DISCUSSION

29. In this paper, we demonstrate how the algorithms developed in Pannekoek, Shlomo and De Waal (2008) for imputing missing values under benchmarked totals and edit constraints can be used to create mass-imputed numerical statistical databases which work particularly well for business statistics. We have shown in a simulation study that under the assumptions of missing at random, mass-imputation for combining register and survey data is superior to using single survey estimates for producing statistical outputs. Future work will be directed to applications on real datasets where the correlation structure may not be strong and the transformations of variables to obtain normality assumptions may induce bias. In addition, we will investigate variance estimation for estimates produced from the imputed statistical

database which need to take into account the quality and efficiency of the regression models and the variability induced by benchmarked totals that are derived from survey sample estimates.

REFERENCES

Houbiers, M. (2004). *Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands*. JOS, Vol. 20, No. 1, 55-75.

Kovar, J. and Whitridge, P. (1995*). Imputation of Business Survey Data*. Business Survey Methods (Cox, Binder, Chinnappa, Christianson, Colledge and Kott, eds.)
New York: Wiley, 403-423.

Kroese, B., Renssen, R., and Trijssenaar, M. (2000). *Weighting or Imputation: Constructing a Consistent Set of Estimates Based on Data from Different Sources*. Special Issue of Integrating Administrative Registers and Household Surveys, Volume 15. Statistics Netherlands

Pannekoek, J., Shlomo, N. and De Waal, T. (2008). *Calibrated Imputation of Numerical Data Under Linear Edit Restrictions*. Presented at the UNECE Work Session on Statistical Data Editing, Vienna. http://www.unece.org/stats/documents/2008/04/sde/wp.23.e.pdf

Van de Laar, R. (2004). *Edit Rules and the Strategy of Consistent Table Estimation*. Discussion Paper 04013 (version1). Statistics Netherlands.