Working Paper.
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (i): Selective editing / macro editing

An Assessment of Automatic Editing via the Contamination Model and Multiple Imputation

Prepared by Masayoshi Takahashi,[1] National Statistics Center, Japan

I. Introduction

1. The automatic data editing process is typically divided into two steps. The first step is error localization, where errors in the dataset are identified by some error detection methods. The second step is imputation, where errors are corrected by some imputation methods. The aim in this project is to propose a way to automate part of the editing process for the Japanese economic surveys. To attain this goal, we use the dataset of the Economic Census for Business Activity, which was conducted for the first time in Japanese history in February 2012. First, we contaminate the dataset by artificial errors. Next, in the error localization step, we apply R package SeleMix to detect the errors in the dataset. Then, in the error correction step, we employ the three competing multiple imputation algorithms, which are Markov chain Monte Carlo (MCMC), Fully Conditional Specification (FCS), and Expectation-Maximization with Bootstrapping (EMB). For this purpose, we use R packages Norm (MCMC), Mice (FCS), and Amelia II (EMB). Finally, we compare the performance of the error correction by these three algorithms against the true values in the dataset.

2. By way of organization, Section II discusses automatic editing. Section III explains how error is localized by SeleMix. Section IV introduces the competing multiple imputation software programs. Section V analyzes the dataset of the Economic Census for Business Activity by SeleMix, Amelia, Mice, and Norm. Section VI concludes.

II. Automatic Editing

3. The automation of the statistical data editing process involves two steps: (1) Error localization step; (2) error correction step. First, error is localized by some error detection techniques, where erroneous cells are identified. Next, error is corrected by some imputation techniques, where erroneous values are deleted and imputed (de Waal *et al.*, 2011).

4. There are several ways to localize random errors, such as statistical models, deterministic checking rules, and solution to a mathematical optimization problem. Statistical methods are further divided into outlier detection techniques and neural networks (de Waal *et al*, 2011). As Section III shows, this paper focuses on an outlier detection technique as an error localization method. There are also a variety of imputation techniques, but Takahashi and Ito (2012) showed the superiority of multiple imputation; thus, this paper focuses on multiple imputation as an error correction method.

[1] The views and opinions expressed in this paper are the author’s own, not necessarily those of the institution.

III. Error Localization: Selective Editing Using SeleMix

A. Contaminated Normal Distribution

5. Generally, a contaminated normal model with two peaks can be described by equation (1). In other words, variable x has a contaminated normal distribution if its distribution is composed of the following two parts: one is the normal distribution with mean μ and variance σ^2 which is generated by probability p; and the other part is some probability density function $g(x)$ which is generated by probability $1-p$.

$$f(x) = p(2\pi\sigma^2)^{-1/2} exp\left(-\frac{1}{2\sigma^2}[x-\mu]^2\right) + (1-p)g(x) \qquad (1)$$

6. If the variance of the distribution contaminated by the p.d.f. $g(x)$ is larger than σ^2, or the mean is completely different from μ, the observations obtained from the contaminated distribution are likely to be different from other observations; thus, they can be considered outliers (DeGroot and Schervish, 2002). For more detailed discussions on the contaminated normal model, see Buglielli *et al.* (2010), Guarnera *et al.* (2012), and Di Zio and Guarnera (2013).

B. R-Package SeleMix

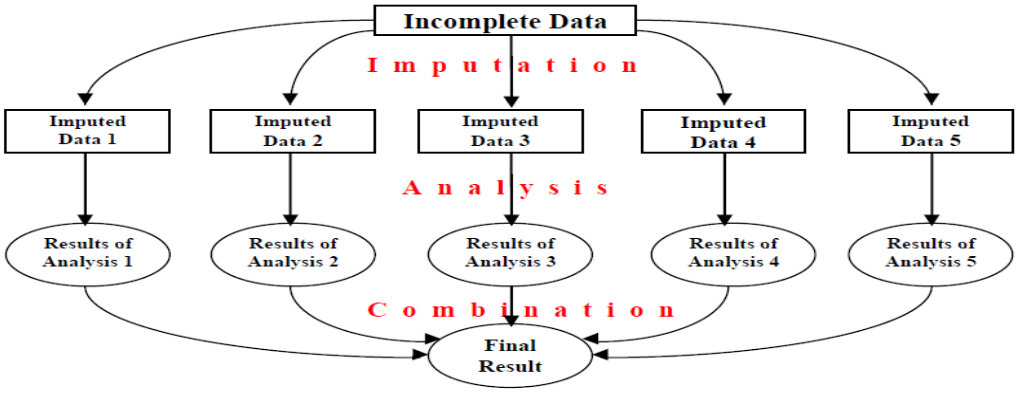
7. The model used in SeleMix pays a particular attention to random errors, which detects potentially influential outliers. It is a multivariate error model that estimates both the error probability and the influence of the error based on the assumption that, by inflating the variance of the error-free data, the distribution of the error data can be obtained. For more detailed discussions on SeleMix, see Buglielli *et al.* (2011) and Guarnera and Buglielli (2013).

IV. Error Correction: Competing Algorithms of Multiple Imputation

A. Multiple Imputation in a Nutshell

8. The origin of multiple imputation can be traced back to Rubin (1978), the basics of which are as follows (Rubin, 1987; King *et al.*, 2001; Takahashi and Ito, 2013). Multiple imputation substitutes missing values by *M* simulated values, where *M* is greater than 1. In order to attain this goal, we construct a posterior distribution of the missing data, conditional on the observed data, from which *M* simulated values are randomly drawn; thus, *M* multiply-imputed datasets are created, which reflect the uncertainty associated with imputation. We use each of these *M* multiply-imputed datasets separately for the purpose of statistical analyses. Then, we combine the results of the *M* statistical analyses to calculate a point estimate. The steps in multiple imputation (*M* = 5) are graphically shown in Figure 4.1.

Figure 4.1: Schematic Overview of Multiple Imputation



9. Here are the notations used in this paper. Let us define D as an $n \times p$ dataset, where n is the sample size and p is the number of variables. Under the conditions where D has no missing values, D is

normally distributed with mean vector μ and variance-covariance matrix Σ. In other words, $D \sim N_p(\mu, \Sigma)$. In our notation, *i* refers to observation index, where *i* ranges from 1 to *n*. Also, *j* refers to variable index, where *j* ranges from 1 to *p*. Let us further define *D* as $\{Y_1, \dots, Y_p\}$, where Y_j is the *j*-th column in *D* and Y_{-j} is the complement of Y_j, i.e., all columns in *D* except Y_j. *R* is defined as a response indicator matrix, where *R* has the same dimension as *D*. If *D* is observed, *R* takes the value of 1; if not, *R* takes the value of 0. Finally, Y_{obs} is the observed part of data and Y_{mis} is the missing part of data: $D = \{Y_{obs}, Y_{mis}\}$.

10. The assumption of multivariate normality implies that missing values are linearly modelled. Suppose that Y_{ij} is missing and that $Y_{i,-j}$ includes all of the observations in row *i* except variable Y_j. Then, the imputation model is defined as equation (2), where $\tilde{Y}_{ij}$ is a simulated value for a missing value, i.e., an imputed value. Here, ~ means random sampling from an appropriate posterior distribution, β is a regression coefficient, and ε stands for fundamental uncertainty.

$$\tilde{Y}_{ij} = Y_{i,-j}\tilde{\beta} + \tilde{\varepsilon}_i \tag{2}$$

11. In order to calculate regression coefficients, we need to know the mean, variance, and covariance. Fortunately, all of the necessary information can be found in μ and Σ. Were μ and Σ fully known, the value of the true regression coefficient β would be deterministically based on Y_j. If so, the imputation of missing values would be also deterministic. In this case, the likelihood function of complete data would be equation (3).

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(Y_i | \mu, \Sigma) \tag{3}$$

12. Unfortunately, almost no datasets are perfectly observed, i.e. some values are missing. Assuming that the mechanism of missingness is missing at random (MAR),[2] we form the likelihood of observed data Y_{obs}. Let $Y_{i,obs}$ an observed value of row *i* in *D*, $\mu_{i,obs}$ a subvector of μ, and $\Sigma_{i,obs}$ a submatrix of Σ. Since the marginal densities are normal, the likelihood function of observed data Y_{obs} is equation (4).

$$L(\mu, \Sigma | Y_{obs}) \propto \prod_{i=1}^{n} N\left(Y_{i,obs} | \mu_{i,obs}, \Sigma_{i,obs}\right) \tag{4}$$

13. The fact that μ and Σ are not fully known means that it is not possible to know β with certainty. Unlike $\hat{\beta}$ (ordinary least squares estimate of β), $\tilde{\beta}$ in equation (2) implies that estimation is uncertain. However, as Allison (2002) pointed out, the computation of equation (4) is not easy based on the traditional methods. Furthermore, it is not easy to randomly draw μ and Σ from this posterior distribution. In order to solve this problem, various computational algorithms have been proposed in the literature, which we will explain in the next section.

B. Markov chain Monte Carlo (MCMC): R-Package Norm

14. Rubin's version of multiple imputation is based on Markov chain Monte Carlo (MCMC), which is the well-known Bayesian computational algorithm (Rubin, 1987; Schafer, 1997). As an MCMC computational technique, data augmentation is often used, where imputations are generated from the conditional distribution of missing values (imputation step) and parameter values are generated from the posterior distribution (posterior step), and these two steps are repeated until convergence is attained (Schafer, 1997; Little and Rubin, 2002; Gill, 2008).[3] In terms of computer software programs, this algorithm is also known as joint modeling (JM), where we draw imputations from the conditional

[2] About the three assumptions of missingness, see Little and Rubin (2002). Specifically, Missing Completely At Random (MCAR) means $P(R|D) = P(R)$. Missing At Random (MAR) means $P(R|D) = P(R|Y_{obs})$. NonIgnorable (NI) means that $P(R|D)$ cannot be simplified and *R* is not independent of *D*.

[3] For more detailed information about MCMC, see Takahashi and Ito (2013).

distribution of the multivariate distribution for the missing data (van Buuren and Groothuis-Oudshoorn, 2011). R Package Norm 3.0.0 is an example of the software using this algorithm (Schafer, 2008).[4]

15. Norm was developed by Joseph L. Schafer (1997, 2008) of the Pennsylvania State University. Schafer is Rubin's disciple,[5] and we can say that Norm authentically implements Rubin's original multiple imputation. In order to conduct multiple imputation via MCMC, we first need to set the initial value. In Norm, we use the `emNorm` function, which creates the estimates based on the Expectation-Maximization (EM) algorithm. Also, note that multiple imputation utilizes random numbers, so that there is a need to set the seed by using the `set.seed` function for reproducibility.

```
emResult<-emNorm(data)
set.seed(1223)
```

16. `mcmcNorm` is the function to perform multiple imputation by MCMC. The right-hand side of `iter=` is the repetition number of a Markov chain. The right-hand side of `impute.every` saves the data for the designated numbers among the above repetitions. The following example saves every 1,000 data among the 5,000 repetitions, so that 5000 / 1000 = 5 multiply-imputed datasets are created ($M = 5$). The `summary` function returns the results.

```
mcmcResult<-mcmcNorm(emResult, iter=5000, impute.every=1000)
summary(mcmcResult)
```

17. The results of multiple imputation are saved as `mcmcResult$imp.list[[#]][,#]`, where [#] stands for the m-th imputation and [,#] stands for a variable number. In other words, the imputation from m = 2 for the variable in the third column is saved as `mcmcResult$imp.list[[2]][,3]`. In order to see the mean, we use the `mean` function, and in order to see the standard deviation, we use the `sd` function, as below. Also, if we want to use regression analysis, we simply use the `lm` function as below. After repeating these processes *M* times, simply combine the results using Rubin's rule (Takahashi and Ito, 2012).

```
mean(mcmcResult$imp.list[[#]][,#])
sd(mcmcResult$imp.list[[#]][,#])
summary(lm(mcmcResult$imp.list[[#]][,#]~mcmcResult$imp.list[[#]][,#]+m
cmcResult$imp.list[[#]][,#]))
```

C. Fully Conditional Specification (FCS): R-Package Mice

18. Fully Conditional Specification (FCS) is one of the most prominent competing algorithms against MCMC. In FCS, imputation is on a variable-by-variable basis, which means that each incomplete variable has its imputation model, under which missing values are iteratively imputed for each variable (van Buuren, 2012).[6] Theoretically, FCS is superior to JM in that it is possible to impute missing values even if we cannot prespecify an appropriate multivariate distribution. R Package MICE is an example of the software using the FCS algorithm (van Buuren and Groothuis-Oudshoorn, 2011).

19. MICE was developed by Stef van Buuren (2012) at Utrecht University in the Netherlands, which stands for Multivariate Imputation by Chained Equations. It is an incredibly flexible multiple imputation program. `mice` is the function for multiple imputation, where `data` is the name of the dataset to multiply-impute, the right-hand side of `m =` is the number of multiply-imputed datasets, the right-hand side of `seed =` is to set the seed, and `meth = "norm"` means that the imputation model is a Bayesian linear model (van Buuren and Groothuis-Oudshoom, 2011).

```
imp<-mice(data, m = #, seed = #, meth="norm")
```

[4] Note that Norm 3.0.0 works only in R 2.9.2 or before.

[5] Joseph L. Schafer obtained his Ph.D. in statistics with Donald B. Rubin as his advisor at Harvard University (Schafer, 1992).

[6] For more detailed information about FCS, see Takahashi and Ito (2013).

20. The imputed datasets created above are saved as `imp`, and we can conduct regression analysis using the `with` function. The `pool` function combines the results.

```
fit<-with(imp,lm(variable1~variable2+variable3))
summary(pool(fit))
```

21. Also, if we want to save the multiply-imputed datasets as a csv file, simply do as follows.

```
dataimp<-complete(imp, action="broad", include=FALSE)
dataimpdf<-data.frame(dataimp)
write.csv(dataimpdf,"micedata.csv")
```

D. Expectation-Maximization with Bootstrapping (EMB): R-Package Amelia

22. The Expectation-Maximization with Bootstrapping (EMB) is another competing multiple imputation algorithm, which is the combination of the traditional expectation-maximization and the non-parametric bootstrapping. In the EMB algorithm, the non-parametric bootstrapping method is used for estimation uncertainties by acquiring bootstrap subsamples of size *n*, which are randomly drawn from the incomplete dataset *M* times. After that, in each of these *M* bootstrap subsamples, the EM algorithm calculates *M* point estimates of μ and Σ, based on which *M* point estimates of $\tilde{\beta}$ will be calculated for imputation (Congdon, 2006; Honaker and King, 2010).[7] Unlike MCMC and FCS, the EMB algorithm can avoid the Cholesky decomposition[8] and can be expected to be computationally efficient (van Buuren, 2012). R package Amelia II is the software using this algorithm (Honaker, King, and Blackwell, 2011).

23. Amelia was developed by Gary King (2001) of Harvard University, which is expected to be computationally efficient.[9] First, we set the seed by using the `set.seed` function. After that, we create multiple imputation by using the `amelia` function, where `data` is the name of the dataset to multiply impute and the right-hand side of `m=` is the number of multiply-imputed datasets. The results of multiple imputation are stored in `a.out`. With the use of the `write.amelia` function, we can save them as a csv file (file name: outdata).

```
set.seed(#)
a.out<-amelia(data, m = #)
write.amelia(obj= a.out, file.stem = "outdata", orig.data = F,separate
= F)
```

24. In order for statistical analysis using multiply-imputed datasets, we can use R package Zelig by the `require` function as follows (Imai, King, and Lau, 2008).

```
require("Zelig")
z.out<-zelig(variable1~variable2+variable3, data = a.out$imputations,
model = "ls", cite = F)
summary(z.out)
```

[7] For more detailed information about EMB, see Takahashi and Ito (2012) and Takahashi and Ito (2013).

[8] The Cholesky decomposition is also known as the Cholesky factorization, which is defined as follows. If *A* is a positive symmetric definite matrix, i.e., $A = A'$, then there is a matrix *H* such that $A = HH'$, where *H* is lower triangular with positive diagonal elements (Leon, 2006).

[9] In 2001 when Amelia was first developed, its algorithm was EMis, which stands for Expectation-Maximization with Importance Sampling (King *et al.*, 2001). In 2010, it was reborn as Amelia II implemented with EMB for the purpose of further computational efficiency (Honaker and King, 2010).

V. Analysis Using the Japanese Economic Census

A. The Economic Census for Business Activity

25. The Economic Census in Japan aims to identify the actual conditions of business activities, identify the overarching industrial structure, and establish information on the population for a variety of statistical surveys for establishments and enterprises. The Economic Census for Business Activity is to reveal the economic activities among establishments and enterprises, by gathering information not only on the names and locations of establishments and enterprises, but also on various information such as the management structure, the number of workers, and the amount of turnover (Statistics Bureau of Japan, 2012). The Economic Census for Business Activity was conducted in February 2012 for the first time in Japanese history. Note that the results presented in this paper are analyzed by the National Statistics Center of Japan, using the preliminary dataset of the 2012 Economic Census for Business Activity. Also note that the views and opinions expressed in the analysis using this dataset are the author's own, not necessarily those of the institution.

B. Descriptions of Dataset

26. The dataset we used is Division E (Manufacturing in the industrial classification) of the Economic Census for Business Activity dataset. The number of complete observations in this dataset is 198,954. The variables used in this research are turnover, worker, and capital, among which turnover is the target variable for editing. Table 5.1 presents summary statistics of raw data. The means and medians are different in all of the variables. Also, the distances from the means to the 1st and 3rd quartiles are not equal in all of the variables. These facts indicate that they are not normally distributed.

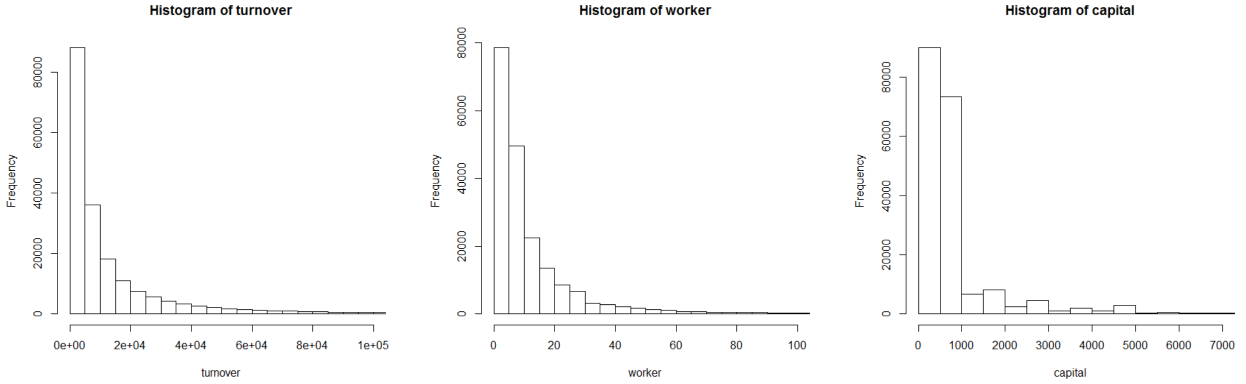
Table 5.1: Summary Statistics (Raw Data)

	1st Quartile	Median	Mean	3rd Quartile	sd
Turnover	2425.0	6200.0	28585.5	17840.0	234798.3
Worker	4.0	7.0	15.4	15.0	36.8
Capital	300.0	1000.0	1939.0	1000.0	18320.6

Note: The unit in turnover and capital is million yen. The unit in worker is person.

27. Figure 5.1 presents the histograms of these variables (raw data). In order to see the shapes of the histograms, the graphs are zoomed in. Clearly, none of them are normally distributed.

Figure 5.1: Histograms of Raw Data (Magnified)



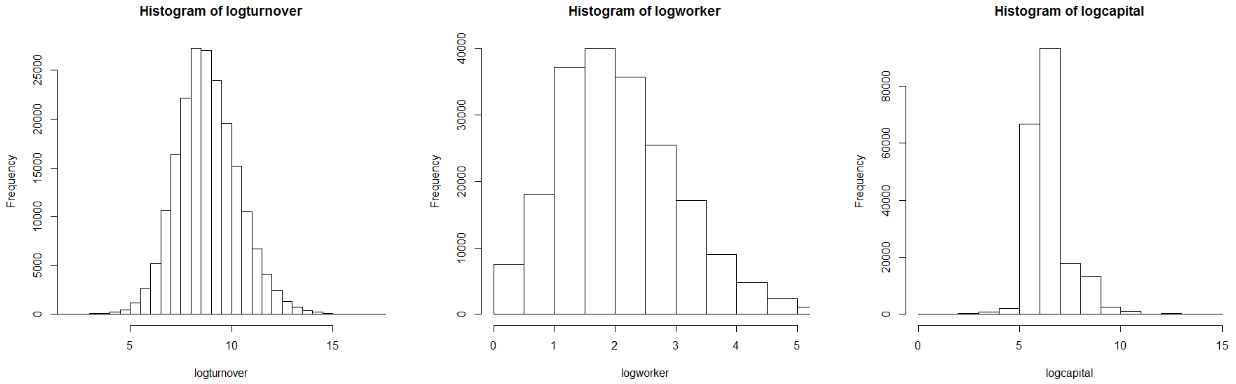
28. Since all of the variables are heavily skewed, in order to normalize the distributions, we transformed them by natural logarithm. Table 5.2 presents summary statistics of the three variables in log form. Now, the means and medians are quite similar, and the distances from the means to the 1st and 3rd quartiles are roughly the same, indicating that the distributions are close to normality.

Table 5.2: Summary Statistics (Natural Logarithm)

	1st Quartile	Median	Mean	3rd Quartile	sd
Turnover	7.794	8.732	8.822	9.789	1.531
Worker	1.386	1.946	2.060	2.708	1.052
Capital	5.704	6.908	6.575	6.908	1.012

29. Figure 5.2 presents the histograms of these variables (log-transformed data). In order to see the shapes of the histograms, the graphs are zoomed in. Now, the distributions are approximately symmetric around the means.

Figure 5.2: Histograms of Log-Transformed Data (Magnified)



C. Descriptions of Contamination (Artificial Errors)

30. Following Di Zio and Guarnera (2013), we contaminated turnover by swapping the first two digits of turnover, using the MAR assumption.[10] The percentage of contamination is about 17.8%, i.e., 35,395 observations are potentially random errors. However, not all swapping created errors. For example, if the original value is 4,485, swapping the first two digits creates 4,485, which is the same as the correct value. In this way, it is found that 3,550 are "correct errors," which are not, in fact, errors. As a result, among these 35,395 potential errors, 31,845 are truly errors. Summary statistics of the turnover variables are presented in Table 5.3. "Truth" refers to the turnover variable without errors, i.e., the original turnover variable in the dataset. "Truth + Error" refers to the contaminated turnover variable, which is the target variable in this research. "Truth – Error" refers to the portion of the original turnover variable, which was not contaminated. "Error" refers to the portion of the original turnover variable, which was contaminated.

Table 5.3: Summary Statistics of Turnover (Raw Data)

	1st Quartile	Median	Mean	3rd Quartile	sd
Truth	2425.0	6200.0	28585.5	17840.0	234798.3
Truth + Error	2800.0	6647.0	28782.5	17920.0	234796.4
Truth – Error	3824.0	8589.0	34350.5	22600.0	258594.1
Error	2100.0	6300.0	43681.5	24840.0	400176.1

Note: The unit in turnover is million yen.

31. Since we know that the variables in the dataset are heavily skewed, we log-transformed these turnover variables, whose summary statistics are presented in Table 5.4.

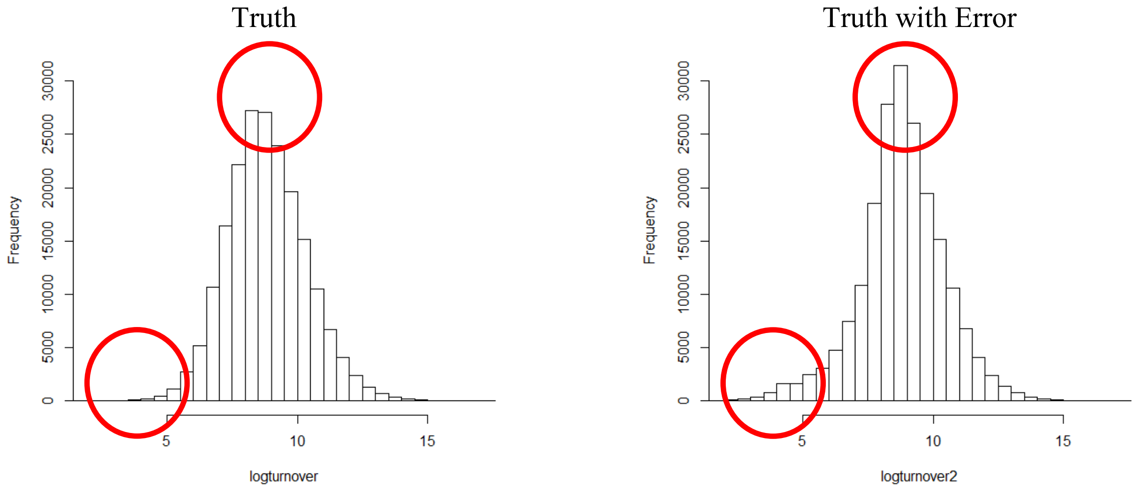
[10] This means that the probability of error occurrence is set to be high when the number of workers is small, but the probability of error occurrence is the same given the same number of workers. It is assumed that if the number of workers is small, the enterprise or establishment may not have enough resources (such as manpower) to accurately respond to surveys; thus, errors are expected to occur frequently.

Table 5.4: Summary Statistics of Turnover (Natural Logarithm)

	1st Quartile	Median	Mean	3rd Quartile	sd
Truth	7.794	8.732	8.822	9.789	1.531
Truth + Error	7.937	8.802	8.798	9.794	1.657
Truth – Error	8.250	9.060	9.198	10.030	1.360
Error	7.650	8.748	8.742	10.120	1.943

32. Figure 5.3 presents the histograms of the turnover variables in natural logarithm. We can see that the effects of contamination are noticeable in the left side and the center of the distributions (circled in red).

Figure 5.3: Histograms of Turnover (Natural Logarithm)



D. Results of Automatic Editing

33. R package SeleMix detected 30,250 observations as outliers. Among these 30,250 outliers, 15,914 are the errors that were introduced above. In order to detect influential outliers, we set t.sel=0.001; as a result, 15,150 outliers are identified as influential. Among these 15,150 influential outliers, 10,067 are the errors that were introduced above. Suppose that we manually inspected these 15,150 influential outliers to see which of them are actually errors. Then, we delete these 10,067 errors and impute them by multiple imputation. As multiple imputation programs, we use R packages Amelia, Mice, and Norm, which were described in Section IV. The number of multiply-imputed datasets (*M*) is set to 20.

34. Table 5.5 presents the results of regression analyses of turnover on worker (natural logarithm). In other words, the presented results are the intercept and slope followed by their associated standard errors in $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$, where $\hat{y}$ is log(turnover) and x_1 is log(worker). The inaccuracies in the error model are more or less corrected by deleting and multiply-imputing these 10,067 errors. For example, the true slope coefficient for logworker is 1.2257, but the error coefficient is 1.2620. The coefficients provided by the multiple imputation models are 1.1981 (Amelia) and 1.1980 (Mice). Therefore, the error is reduced by about 25%. The associated standard errors are now correctly estimated in the multiple imputation models. Among the multiple imputation models, unfortunately, Norm was not able to handle a large dataset; thus, its outputs are signified as NA (not applicable). Mice performed slightly better than Amelia, but their differences are almost negligible. The time it took to complete multiple imputation was 4 minutes 33 seconds by Amelia and 8 minutes 29 seconds by Mice. Therefore, Amelia is most computationally efficient among the three programs.

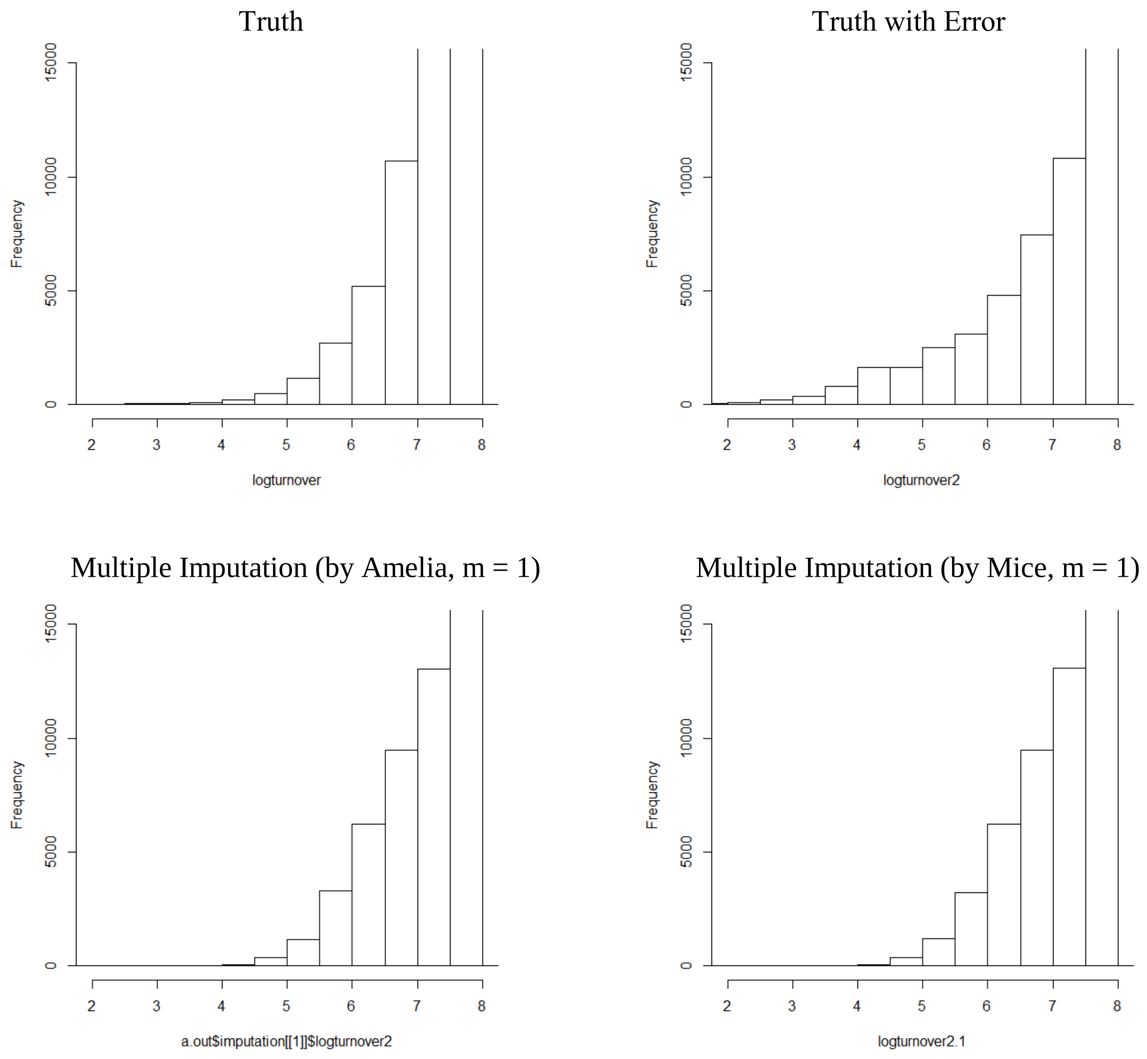
Table5.5: Results of Regression Analyses

	Truth	Error	Amelia	Mice	Norm
Intercept	6.2980	6.1989	6.3852	6.3854	NA
	(0.0041)	(0.0049)	(0.0044)	(0.0043)	
logworker	**1.2257**	**1.2620**	**1.1981**	**1.1980**	NA
	(0.0018)	**(0.0021)**	**(0.0019)**	**(0.0018)**	
n	198,954	198,954	198,954	198,954	188,887
# of errors	0	31,845	21,778	21,778	21,778
time			4m33s	8m29s	NA

Note: Reported values are coefficients (standard errors). The dependent variable is logturnover, which is turnover in natural logarithm. The independent variable is logworker, which is worker in natural logarithm. The dependent variable in the Error model is “Truth + Error” described in paragraph 30. “n” is the number of observations. The # of errors refers to the number of errors that remain in each dataset. Time is the time to complete multiple imputation.

35. Figure 5.4 presents the histograms of logturnover, which focuses on the left tails of the distributions. Here, we can clearly see that the errors in the tails are straightened out in the imputed datasets.

Figure 5.4: Tails of Histograms of Turnover (Natural Logarithm)[11]



Truth

Truth with Error

Multiple Imputation (by Amelia, m = 1)

Multiple Imputation (by Mice, m = 1)

[11] There are 19 other histograms for Amelia and Mice, respectively, but they are quite similar to the ones presented here.

VI. Conclusions

36. In this paper, we assessed a potential way to partly automate the editing process of economic surveys. For this purpose, we used the dataset of the Economic Census for Business Activity, which was the first overarching economic survey in Japanese history. As an error detection tool, we used SeleMix, which utilizes the contaminated normal model. As error correction tools, we used several multiple imputation programs (Amelia, Mice, and Norm), which are based on the EMB, FCS, and MCMC algorithms, respectively.

37. We showed that SeleMix was useful in identifying influential random errors and that multiple imputation was effective in correcting these errors. It was also found that, while the accuracy of imputation is roughly the same between Amelia and Mice, there is a difference in terms of computational efficiency. Specifically, Norm was unable to handle a large dataset (number of observations ≅ 200,000) while Amelia was quite fast in handling the same dataset.

References

[1] Allison, Paul D. (2002). *Missing Data*. CA: Sage Publications.

[2] Buglielli, M. Teresa, Marco Di Zio, and Ugo Guarnera. (2010). "Use of Contamination Models for Selective Editing," *European Conference on Quality in Survey Statistics*, Helsinki, Finland, 4-6 May 2010.

[3] Buglielli, M. Teresa, Marco Di Zio, Ugo Guarnera, and Francesca R. Pogelli. (2011). "An R Package for Selective Editing Based on a Latent Class Model," *Work Session on Statistical Data Editing, UNECE*, Ljubljana, Slovenia, 9-11 May 2011.

[4] Congdon, Peter. (2006). *Bayesian Statistical Modelling*, Second Edition. West Sussex: John Wiley & Sons Ltd.

[5] DeGroot, Morris H. and Mark J. Schervish. (2002). *Probability and Statistics*. Boston: Addison-Wesley.

[6] de Waal, Ton, Jeroen Pannekoek, and Sander Scholtus. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.

[7] Di Zio, Marco and Ugo Guarnera. (2013). "A Contamination Model for Selective Editing," *Journal of Official Statistics* vol.29, no.4, pp.539-555.

[8] Gill, Jeff. (2008). *Bayesian Methods—A Social Sciences Approach*, Second Edition. London: Chapman & Hall/CRC.

[9] Guarnera, Ugo and M. Teresa Buglielli. (2013). *Package 'SeleMix'*. http://cran.r-project.org/web/packages/SeleMix/SeleMix.pdf. Accessed February 26, 2014.

[10] Guarnera, Ugo, Orietta Luzi, Francesca Silvestri, M. Teresa Buglielli, Alessandra Nurra, and Giampiero Siesto. (2012). "Multivariate Selective Editing via Mixture Models: First Applications to Italian Structural Business Surveys," *Work Session on Statistical Data Editing*, *UNECE*, Oslo, Norway, 24-26 September 2012.

[11] Honaker, James and Gary King. (2010). "What to do About Missing Values in Time Series Cross-Section Data," *American Journal of Political Science* vol.54, no.2, pp.561–581.

[12] Honaker, James, Gary King, and Matthew Blackwell. (2011). "Amelia II: A Program for Missing Data," *Journal of Statistical Software* vol.45, no.7.

[13] Imai, Kosuke, Gary King, and Olivia Lau. (2008). "Toward A Common Framework for Statistical Analysis and Development," *Journal of Computational and Graphical Statistics* vol.17, no.4, pp.1-22.

[14] King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. (2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review* vol.95, no.1, pp.49-69.

[15] Leon, Steven J. (2006). *Linear Algebra with Applications*, Seventh Edition. Upper Saddle River, NJ: Pearson/Prentice Hall.

[16] Little, Roderick J. A. and Donald B. Rubin. (2002). *Statistical Analysis with Missing Data*, Second Edition. New Jersey: John Wiley & Sons.

[17] Rubin, Donald B. (1978). "Multiple Imputations in Sample Surveys — A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp.20-34.

[18] Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

[19] Schafer, Joseph L. (1992). "Algorithms for Multiple Imputation and Posterior Simulation From Incomplete Multivariate Data with Ignorable Nonresponse," Ph.D. Dissertation, Harvard University, Cambridge, MI.

[20] Schafer, Joseph L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall/CRC.

[21] Schafer, Joseph L. (2008). *NORM: Analysis of Incomplete Multivariate Data under a Normal Model, Version 3*. Software Package for R. University Park, PA: The Methodology Center, the Pennsylvania State University.

[22] Statistics Bureau of Japan. (2012). *Economic Census*. http://www.stat.go.jp/english/data/e-census.htm. Accessed February 26, 2014.

[23] Takahashi, Masayoshi and Takayuki Ito. (2012). "Multiple Imputation of Turnover in EDINET Data: Toward the Improvement of Imputation for the Economic Census," *Work Session on Statistical Data Editing, UNECE*, Oslo, Norway, September 24-26, 2012.

[24] Takahashi, Masayoshi and Takayuki Ito. (2013). "Multiple Imputation of Missing Values in Economic Surveys: Comparison of Competing Algorithms," *Proceedings of the 59th World Statistics Congress of the International Statistical Institute*, Hong Kong, China, 25-30 August 2013, pp.3240-3245.

[25] van Buuren, Stef and Karin Groothuis-Oudshoorn. (2011). "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software* vol.45, no.3.

[26] van Buuren, Stef. (2012). *Flexible Imputation of Missing Data*. London: Chapman & Hall/CRC.