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

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

National Statistics Center (Japan)
Masayoshi Takahashi

National Statistics Center
NSTAC

Notes: The views and opinions expressed in this presentation are the authors' own, not necessarily those of the institution.

Outline

1. Introduction
2. Automatic Editing
3. Error Localization: Selective Editing Using SeleMix
4. Error Correction: Competing Algorithms of Multiple Imputation
5. Analysis Using the Japanese Economic Census
6. Conclusions

National Statistics Center
NSTAC

What is in this Project?

- Propose a way to automate part of the editing process for economic surveys
- Use the Economic Census for Business Activity data
- Contaminate the data by artificial errors
- Apply R-package SeleMix for error detection
- Employ three multiple imputation algorithms for error correction

Two Steps in Automatic Editing

- Error Localization Step
  - Erroneous cells are identified
  - Variety of error detection techniques
  - Focus on outlier detection
- Error Correction Step
  - Erroneous values are deleted and imputed
  - Variety of imputation techniques
  - Focus on multiple imputation

Contaminated Normal Distribution

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

- Variable x has a contaminated normal distribution if its distribution is of:
  - Normal distribution with mean μ and variance σ^2 which is generated by probability p
  - Some probability density function $g(x)$ which is generated by probability 1-p

Contaminated Normal Distribution

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

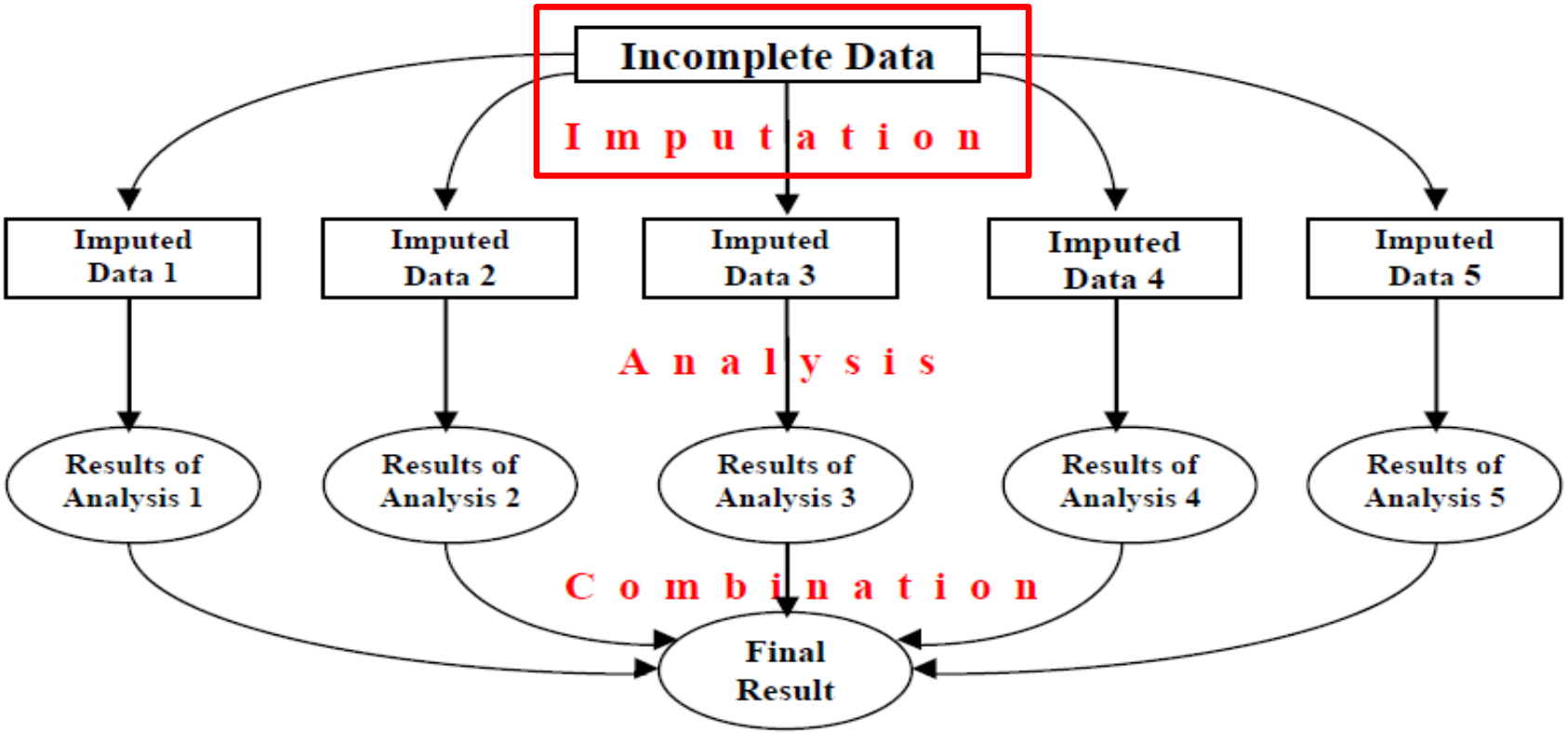
- Variable *x* has a contaminated normal distribution if its distribution is of:
  - Normal distribution with mean μ and variance σ^2 which is generated by probability *p*
  - Some probability density function *g*(*x*) which is generated by probability 1-*p*

R-Package SeleMix

- Pays attention to random errors, which detects potentially influential outliers
- A multivariate error model that estimates both the error probability and the influence of the error based on the contaminated normal model
- For more info. on SeleMix, see Buglielli *et al.* (2011), Guarnera and Buglielli (2013), and other papers by ISTAT

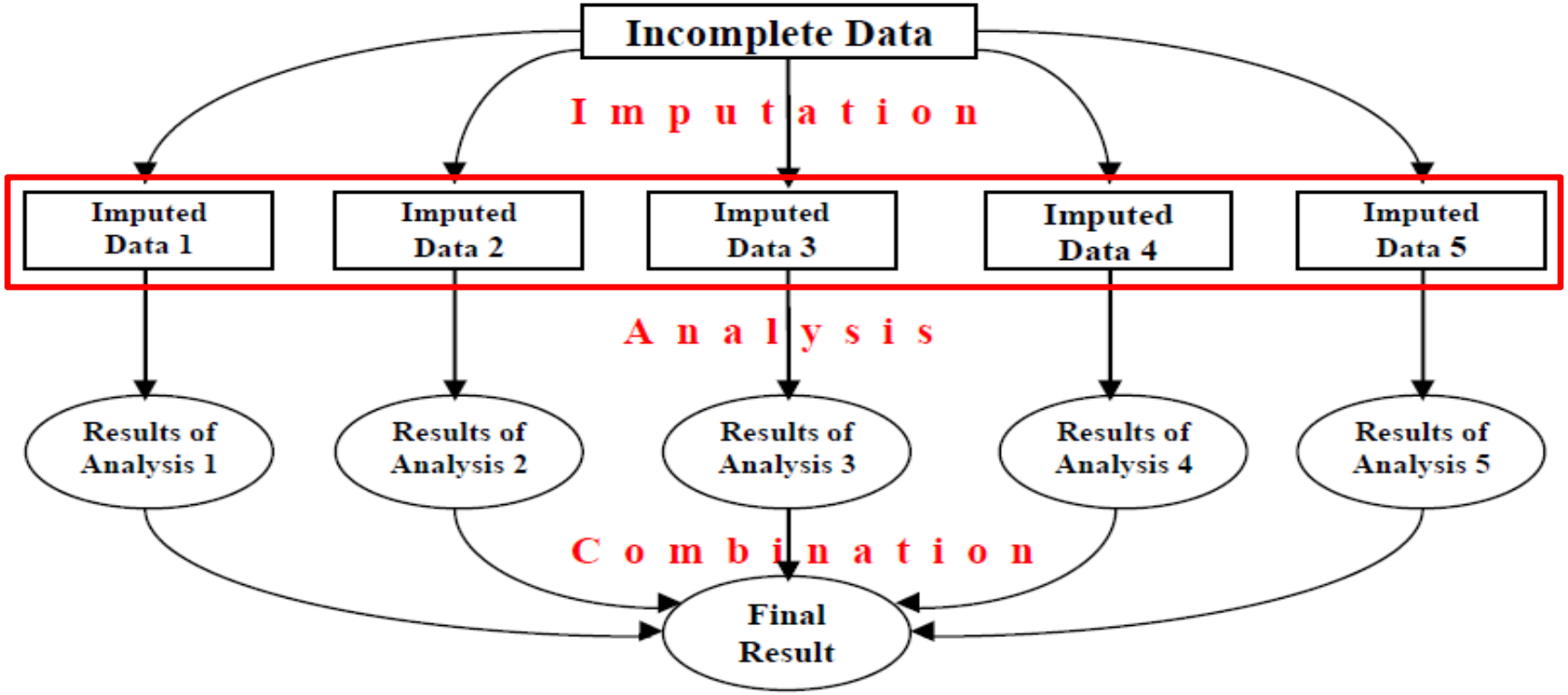
Multiple Imputation in a Nutshell

Construct a posterior distribution of the missing data, conditional on the observed data



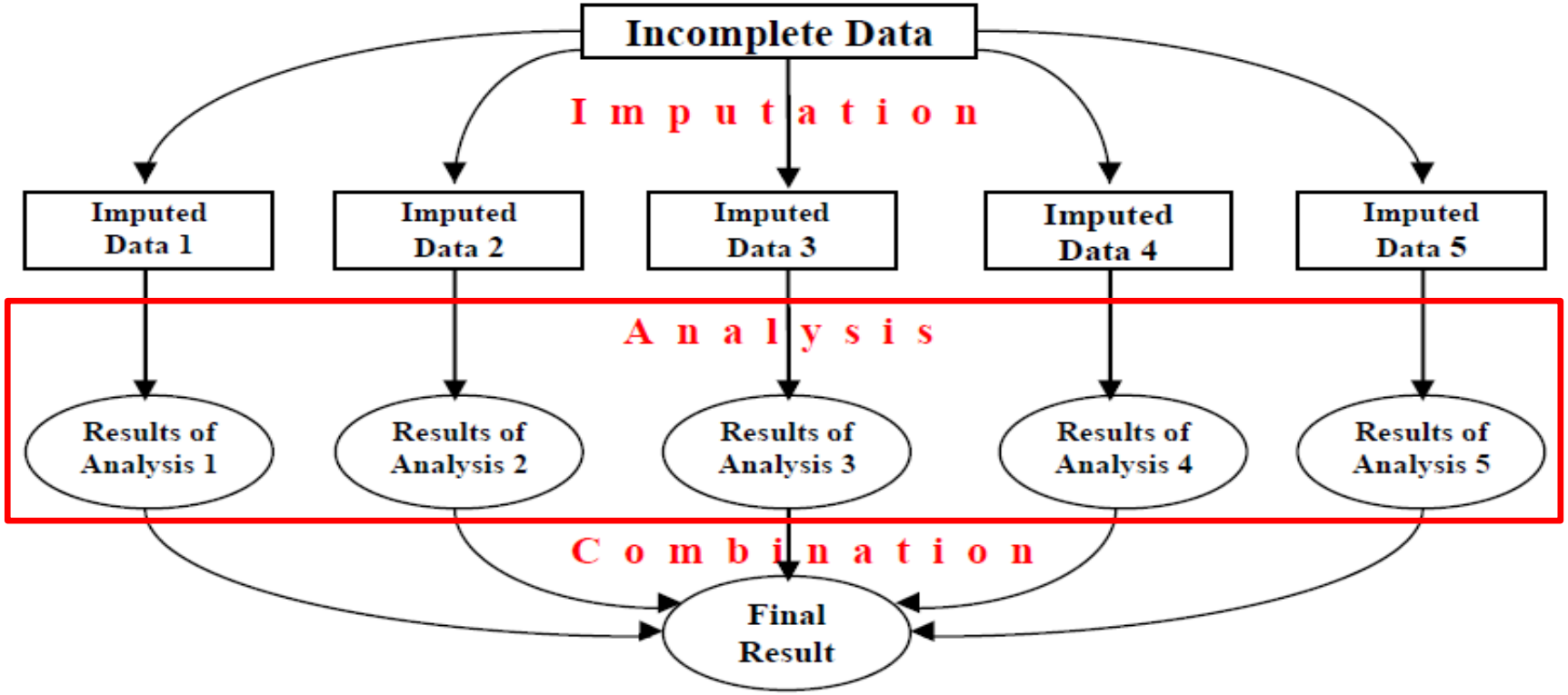
Multiple Imputation in a Nutshell

M multiply-imputed datasets are created, which reflect the uncertainty associated with imputation



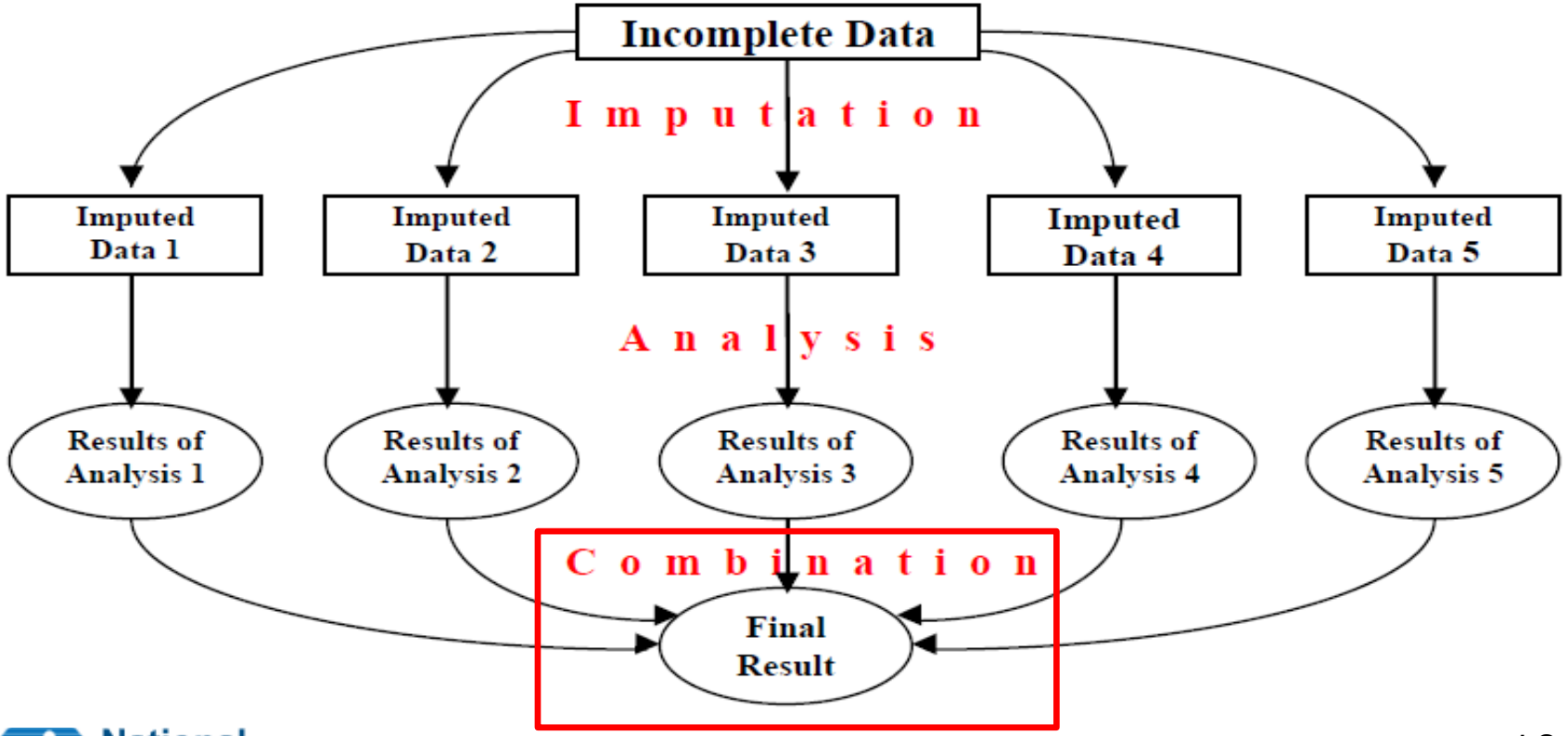
Multiple Imputation in a Nutshell

Use each of these *M* multiply-imputed datasets separately for the purpose of statistical analyses



Multiple Imputation in a Nutshell

Combine the results of the *M* statistical analyses to calculate a point estimate



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

- Mechanism
  - I-Step: Imputations are generated from the conditional distribution of missing values
  - P-Step: Parameter values are generated from the posterior distribution
  - Repeat until convergence is attained
- R-Package Norm
  - Developed by Joseph L. Schafer (Rubin's disciple) of the Pennsylvania State University
  - Authentically implements Rubin's original version of multiple imputation

Fully Conditional Specification (FCS): R-Package Mice

- Mechanism
  - Imputation is on a variable-by-variable basis
  - Each incomplete variable has its imputation model, under which missing values are iteratively imputed for each variable
- MICE
  - Developed by Stef van Buuren at Utrecht University in the Netherlands
  - Multivariate Imputation by Chained Equations
  - Flexible multiple imputation program

EMB Algorithm: R-Package Amelia

- Mechanism
  - EMB stands for the Expectation-Maximization with Bootstrapping algorithm
  - Combination of the traditional expectation-maximization and the non-parametric bootstrapping
- Amelia
  - Developed by Gary King of Harvard University.
  - Expected to be computationally efficient
- See Takahashi and Ito (2012, pp.3-4)

The Economic Census for Business Activity in Japan

- Aims to
  - identify the actual conditions of business activities
  - identify the overarching industrial structure
  - establish information on the population for a variety of statistical surveys for establishments and enterprises
- Conducted in February 2012 for the first time in Japanese history

Note that the results presented in this presentation were 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.

Descriptions of Data

- Division E
  - Manufacturing sector in the industrial classification of the Economic Census for Business Activity
- Number of complete observations
  - 198,954(approximately 200,000)
- Variables
  - Turnover: target variable for editing
  - Worker
  - Capital

} explanatory variables

Summary Statistics (Raw Data)

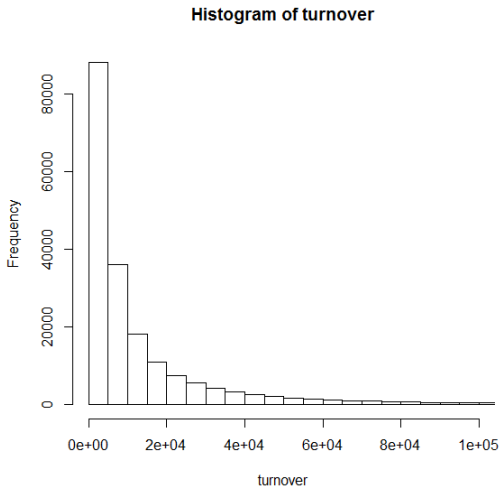
Table 5.1

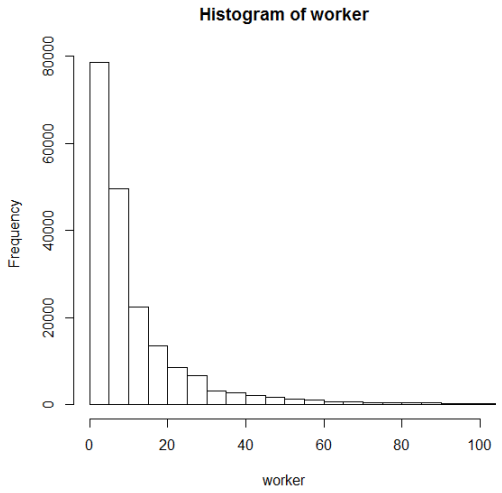
	1st Quartile	Median	Mean	3rd Quartile	sd
Turnover	2425.0	6200.0	28585.5	17840.0	234798.3
Worker	4.0	7.0	15.4	15.0	36.8
Capital	300.0	1000.0	1939.0	1000.0	18320.6

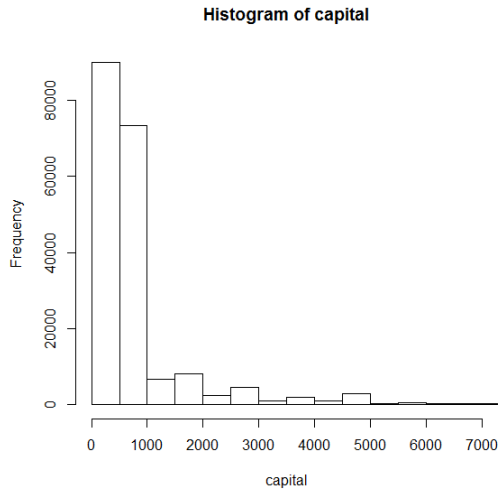
Note: The unit in turnover and capital is million yen. The unit in worker is person.

Histograms of Raw Data (Magnified)

Figure 5.1







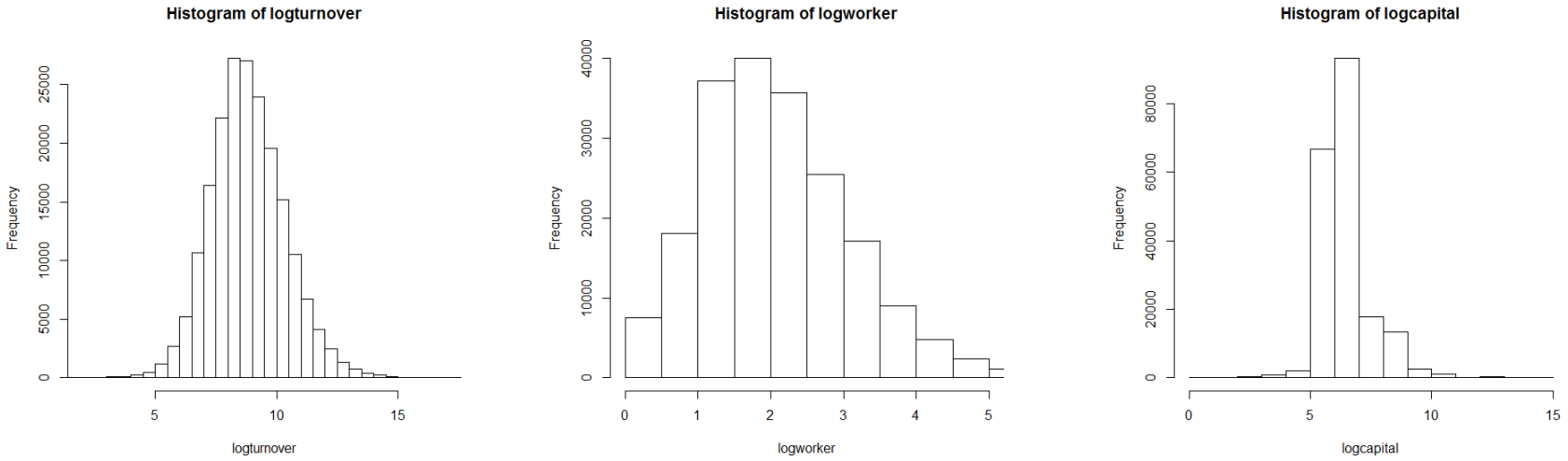
Summary Statistics (Natural Logarithm)

Table 5.2

	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

Histograms of Log-Transformed Data (Magnified)

Figure 5.2



National Statistics Center
NSTAC

Descriptions of Contamination (Artificial Errors)

- Following Di Zio and Guarnera (2013)
  - Contaminated turnover by swapping the first two digits of turnover
  - Used the MAR assumption
- Percentage of contamination
  - About 17.8%=35,395 obs. are potential errors
  - However, not all swapping created errors
    - Ex) 4,485; 11,548; 3,359 etc
    - 3,550 obs. are “correct errors”
    - 31,845 obs. are truly errors

Results of Automatic Editing: Error Detection

- R-package SeleMix detected 30,250 observations as outliers
  - Among these 30,250 outliers, 15,150 outliers are identified as influential
  - Among these 15,150 influential outliers, 10,067 are the errors
  - Assuming that these 10,067 errors were manually confirmed to be erroneous, delete them for error correction by multiple imputation

Results of Automatic Editing: Error Correction

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)**	
Abs. Diff.		**0.0363**	**0.0276**	**0.0277**	
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: The results via multiple imputation are the combined results using 20 multiply-imputed datasets

Summary of the Paper

- Assessed a way to partly automate the editing process of economic surveys
- Used the dataset of the Japanese Economic Census for Business Activity
- Error detection
  - Used SeleMix, which utilizes the contaminated normal model
- Error correction
  - Used several multiple imputation programs (Amelia, Mice, and Norm), which are based on the EMB, FCS, and MCMC algorithms

Findings

- SeleMix was useful in identifying influential errors
- Multiple imputation was effective in correcting these errors
- The accuracy of imputation is roughly the same between Amelia and Mice
- There is a difference in terms of computational efficiency
  - Norm was unable to handle a large dataset
  - Amelia was quite fast in computation

References 1

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.

References 2

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.

References 3

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.

Thank you