# Bayesian Data Editing for Continuous Microdata

Hang Kim[1,2], Jerome Reiter[1], Alan Karr[2], and Lawrence Cox[2]

Duke University[1] and National Institute of Statical Sciences[2]

2013 FCSM Research Conference
Washington, DC
November 6, 2013

Research supported by the National Science Foundation [SES-11-31897]

# Contents

# Contents

# Statistical data editing

procedure of detecting and correcting errors in records to improve data quality

**Manual editing vs. Automatic editing**

- Manual editing spends high costs and times with a large number of records
- Some complex feature of data can be found by using computer power
- Automatic editing can replace manual editing while preserving (or improving) the released data quality

**Two steps of automatic editing**

1. Error localization step[*1]
   identifies erroneous records and fields to be corrected for the records
2. Imputation step
   replaces the identified fields with more accurate data

[*1] De Waal, Pannekoek, and Scholtus (2011)

# Automatic error localization

identifies **erroneous records** and fields to be corrected for the records

**Two mathematical approaches of error localization**[*1]

1. Statistical modeling-based approach
    - identify **unusual records** (outliers) under a statistical model
    - extensively discussed in the literature but scarcely used in practice
2. Mathematical optimization-based approach
    - use logical conditions, *edit rules* to find **inconsistent records**
    - often based on the (generalized) Fellegi-Holt paradigm (Fellegi and Holt 1976)
    - best-known and most-used methods by statistical agencies

      ex) SPEER (U.S. Census Bureau),
      AGGIES (National Agricultural Statistics Service),
      Banff (used to be called GEIS, Statistics Canada), ...

[*1] De Waal, Pannekoek, and Scholtus (2011)

# Edit rules for continuous data (using in optimization-based approach)

"Edit rule is a logical condition to the value of a data field which must be met if the data is to be considered correct" (UNECE 2000)

$\tilde{\boldsymbol{x}}_i = \{\tilde{x}_{i1}, \ldots, \tilde{x}_{ip}\}$: record $i$ with reported values of $p$ fields and $q$ balance edits

- Range restriction

$$L_j \leq \tilde{x}_{ij} \leq U_j \qquad \text{where } j = 1, \ldots, p$$

- Ratio edit

$$L_{jj'} \leq \frac{\tilde{x}_{ij}}{\tilde{x}_{ij'}} \leq U_{jj'} \qquad \text{where } j \neq j'$$

- Balance edit

$$\sum_{j \in C_l} \tilde{x}_{ij} = \tilde{x}_{is_l} \qquad \text{where } l = 1, \ldots, q$$

  - $C_l$: the set of indices for *reported components*
  - $s_l$: the index for *the reported sum*

# The (generalized) Fellegi-Holt algorithm

1. Find all implicit edit rules from user-specified edits
2. Define the latent variable $s_{ij}$

   If $s_{ij} = 1$, field $j$ is "flagged" and replaced with a reasonable value

   If $s_{ij} = 0$, field $j$ is released without editing
3. **Minimal set of (weighted) fields to impute criterion**

   (in short, minimum change criterion). Solve the optimization problem to find the values of $\{s_{i1}, \ldots, s_{ip}\}$ which minimizes

   $$\sum_{j=1}^{p} w_j s_{ij}$$

   where $w_j$ is the reliability weight of field $j$
4. Blank fields $j$ with $s_{ij} = 1$ and impute them by imputation methods (e.g. Hot-deck imputation)

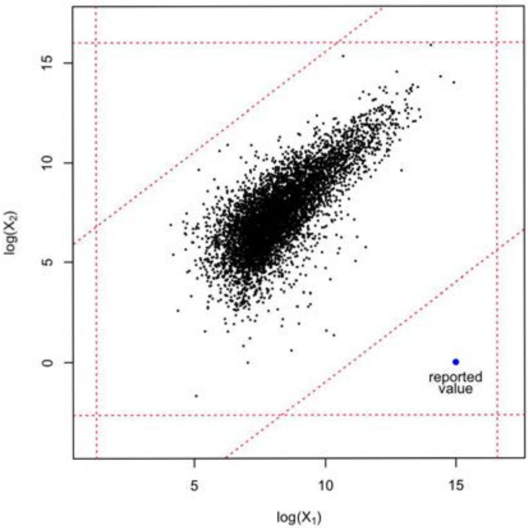
# Issues for the F-H type optimization-based approach

1. No closed form of a feasible region
   It is difficult to find all implied edits from balance edits and ratio edits (despite marginal solution of Draper and Winkler 1997)
2. Risk of the minimum change criterion
   especially when the number of erroneous fields is greeter than the assumed minimum number
3. Using simple imputation methods
   The usual imputation methods, such as Hot-deck imputation or regression imputation, may fail to find a complex, multivariate feature of data
4. Unknown statistical quality
   The optimization-based approach does not measure uncertainty introduced by data editing procedure

*[1] the Nearest-neighbor Imputation Methodology

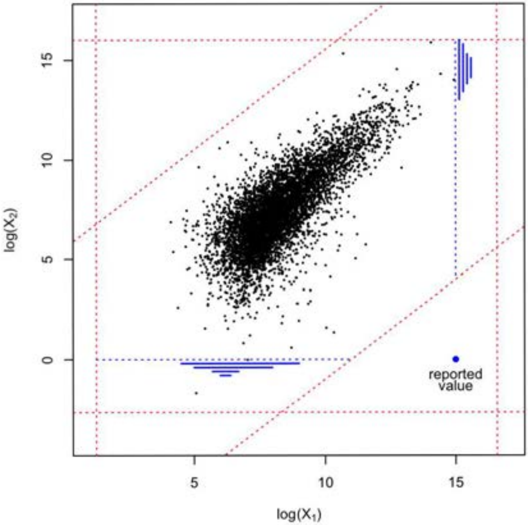
# When is the minimum change criterion harmful?

Want to edit the reported value (blue dot) given the error-free values (black dots)



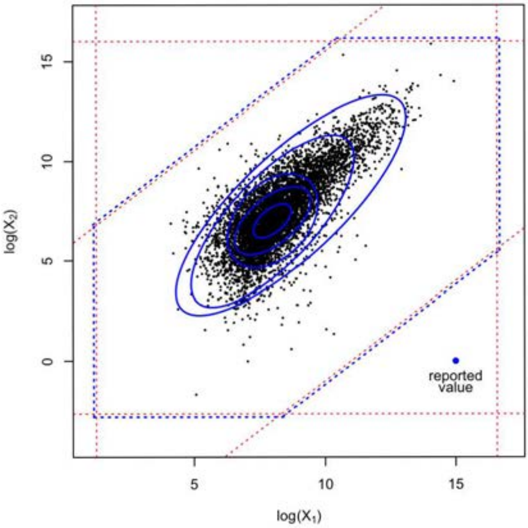
# When is the minimum change criterion harmful?

Supports of the imputed value (blue lines) under the minimum change criterion



# When is the minimum change criterion harmful?

Supports of the imputed value without the minimum change criterion



# Issues for the F-H type optimization-based approach

1. No closed form of feasible region

   It is difficult to find all implied edits from balance edits and ratio edits (despite marginal solution of Draper and Winkler 1997)

2. Risk of the minimum change criterion

   especially when the number of erroneous fields is greeter than the assumed minimum number

3. Using simple imputation methods

   The usual imputation methods, such as Hot-deck imputation, may fail to find a complex, multivariate feature of data

4. Unknown statistical quality

   The optimization-based approach does not measure uncertainty introduced by data editing procedure

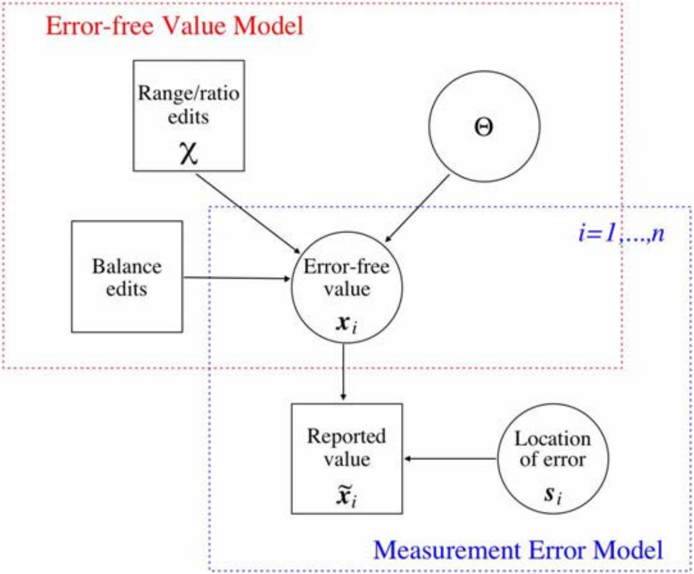
*1 the Nearest-neighbor Imputation Methodology

# Contents

# Key features of Bayesian data editing approach

1. Modeling the latent structure of reported records by introducing latent variables for
    1. unobserved error-free (true) values
    2. unobserved location of errors
2. Incorporating *a priori* knowledge for reliability of data fields (if any)
3. Using nonparametric Bayesian imputation methods
4. Using multiply imputed values or posterior distributions for inference drawn from MCMC

# Framework of the Bayesian data editing Model



# Measurement error model

$\boldsymbol{s}_i = (s_{i1}, \ldots, s_{ip})$: latent variables to indicate error location

- $s_{ij} = 1$ if field $i$ needs be corrected
- $s_{ij} = 0$ otherwise

**Model for reported value $\tilde{\boldsymbol{x}}_i$**

$$f(\tilde{\boldsymbol{x}}_i | \boldsymbol{x}_i, \boldsymbol{s}) = f\left(\tilde{\boldsymbol{x}}_i^1 | \boldsymbol{x}_i\right) \prod_{\{j: s_{ij}=0\}} I\left[\tilde{x}_{ij} = x_{ij}\right]$$

1. $\tilde{\boldsymbol{x}}_i^1 \stackrel{\text{def}}{=} \{\tilde{x}_{ij} : s_{ij} = 1, j = 1, \ldots, p\}$
   $\rightarrow f\left(\tilde{\boldsymbol{x}}_i^1 | \boldsymbol{x}_i\right)$: $(p - \sum_j s_{ij})$-dimensional density for the erroneous values
2. $f\left(\tilde{\boldsymbol{x}}_i^1 | \boldsymbol{x}_i\right)$ can be any form of probability distribution that models the measurement error generating process (if any)

# Error-free value model

$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$: latent error-free values for record $i$ with reported values $\tilde{\boldsymbol{x}}_i$

**Model for $\boldsymbol{x}_i$ with inequality constraints and $q$ balance edits**

$$f(\boldsymbol{x}_i|\theta) = f\left(\boldsymbol{x}_{i,C}|\theta\right) \cdot \prod_{l=1}^{q} I\left[\sum_{j \in C_l} x_{ij} = x_{is_l}\right] \cdot I\left[\boldsymbol{x}_i \in \mathcal{X}\right]$$

1. $\boldsymbol{x}_{i,C} \stackrel{\text{def}}{=} \{x_{ij} : j \in C_l, l = 1, \ldots, q\}$

   $\rightarrow f(\boldsymbol{x}_{i,C}|\theta)$: $(p-q)$-dim. density for latent values for reported components

2. $I[\cdot] = 1$ if the statement is true and $I[\cdot] = 0$ otherwise

   $\rightarrow$ Calculate the latent value for reported sum by balance edit

3. $\mathcal{X}$: the set of convex regions with the inequality constraints (range restrictions and ratio edits)

   $\rightarrow$ All latent error-free values must satisfy range restrictions and ratio edits

# Prior distributions

**For error-free values $\boldsymbol{x}_i$ for $i = 1, \ldots, n$**

- use Dirichlet Process Gaussian mixture model
  
  → to reflect complex joint distributional features based on observed data with minimum level of *a priori* distributional assumption

$$f\left(\boldsymbol{x}_{i,C}|\theta\right) \propto \sum_{k=1}^{K} \pi_k N(\boldsymbol{x}_{i,C}; \boldsymbol{\mu}_k, \Sigma_k)$$

$$\pi_k \sim \text{DirichletProcess}$$

**For error location variables $\boldsymbol{s}_i$**

- reflect *a priori* knowledge about fields' reliability

  ex) If an agency finds that field 1 is twice as reliable as field 2, one may assume $\tau_1 = 1/3$ and $\tau_2 = 2/3$ where the prior distribution of $\boldsymbol{s}_i$ is that $\boldsymbol{s}_i = (1, 0)$ with prob. $\tau_1$ and $\boldsymbol{s}_i = (0, 1)$ with prob. $\tau_2$

**For edit-failing records $\tilde{\boldsymbol{x}}_{ij}$**

- assume uniform distribution over the space where an edit is violated if there is no available information about error-generating process

  → to minimize the impact of model misspecification

# Contents

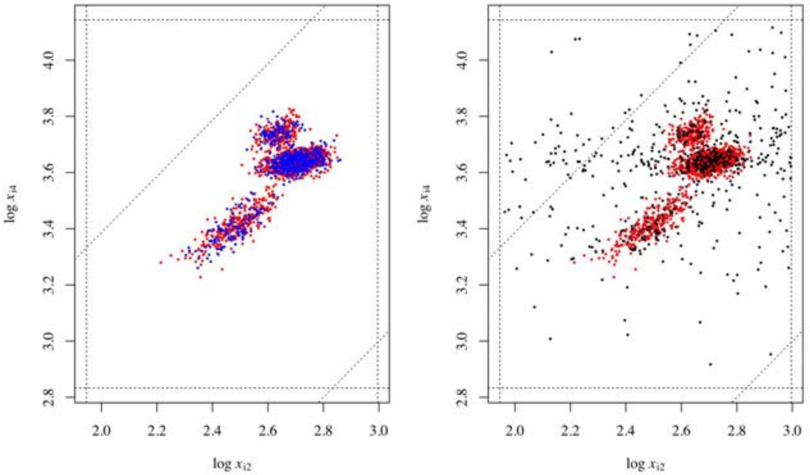
## Simulated data

- introduce edits for $p = 8$ fields
  - range restrictions for each field
  - ratio edits for some pairs of fields
  - $q = 2$ balance edits, i.e., $x_{i1} + x_{i2} + x_{i3} = x_{i4}$ and $x_{i5} + x_{i6} = x_{i7}$
- generate $n = 2000$ error-free values $\boldsymbol{x}_i$ from mixture of three normal dist'n
- for 600 out of 2000 records, introduce edit-failing records $\tilde{\boldsymbol{x}}_i (= \boldsymbol{x}_i)$ which are uniformly distributed over a compact region where at least an edit is violated

## Implemented methods for comparison

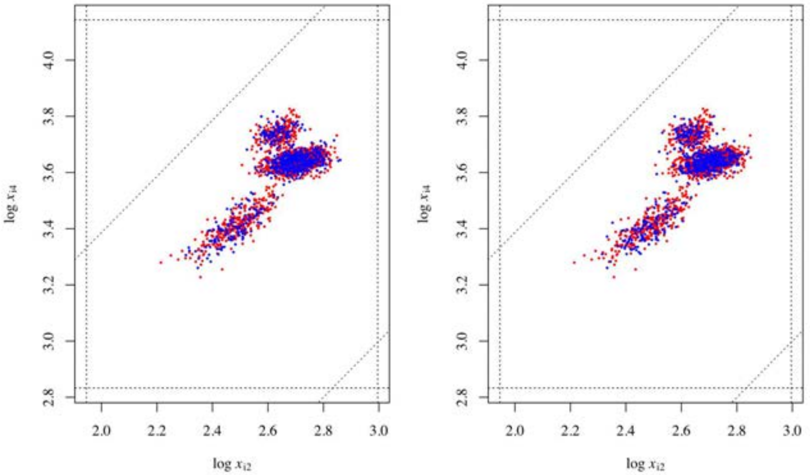
1. Bayesian editing method
2. Bayesian editing method with the minimum change criterion by restricting the support of $\boldsymbol{s}_i$
3. F-H based editing process currently used by agencies (not included)

# Simulated error-free values $\boldsymbol{x}_i$ and reported values $\tilde{\boldsymbol{x}}_i$



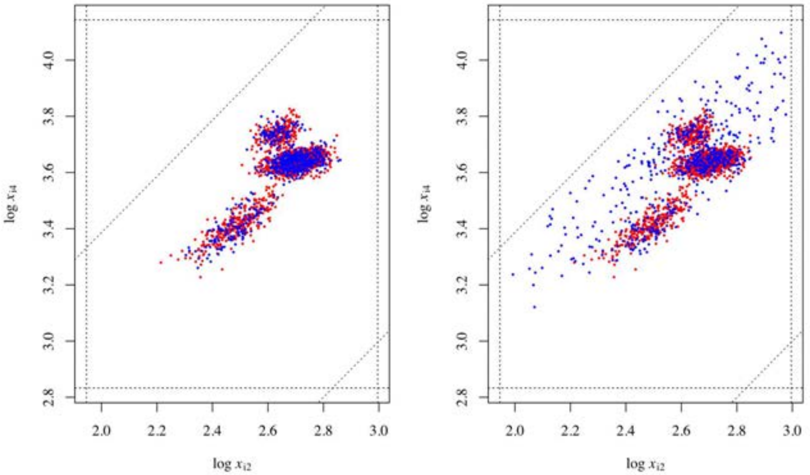

- Left: Error-free values $\boldsymbol{x}_i$
- Right: Reported values $\tilde{\boldsymbol{x}}_i$ when $s_{ij} = 0$ (red dots) or $s_{ij} = 1$ (black dots)

# Result of the Bayesian editing method




- Right: Error-free values $\boldsymbol{x}_i$
- Left: Edited values when $s_{ij} = 1$ (blue) and unchanged values when $s_{ij} = 0$ (red)

# Result of Bayesian editing with the minimum change criterion




- Right: Error-free values $\boldsymbol{x}_i$
- Left: Edited values when $s_{ij} = 1$ (blue) and unchanged values when $s_{ij} = 0$ (red)

# F-H based editing process used by agencies

Some difficulties to implement a real editing process with the simulation data for comparison purpose

For example, the current editing process of the Census of Manufactures

- needs reliability weights for error localization
- cannot find a closed form of feasible region with balance edits and ratio edits
- use different imputation methods for fields

# Contents

# Concluding remarks

The proposed approach replaces the two step optimization-based approach with a single probability based, data-driven approach in which

1. a stochastic model to identify values plausibly in error unlike the F-H routines is suggested
   - reflecting uncertainty over the unknown faulty values when making corrected data
2. a flexible joint probability (DP Gaussian) model captures more complex associations than typical hot deck imputation schemes
3. imputed values from the model are guaranteed to satisfy all linear constraints (balance and ratio edits)

# Future research

- Application to the Census of Manufactures data (in progress)
- Study of measurement error models reflecting real error-generating mechanism
- Contemplation of the role of edit rules

  Can it be replaced by a statistical outlier model?

# References


Lisa R. Draper and William E. Winkler (1997)
Balancing and Ratio Editing With the New Speer System
Research Report RR97/05, Statistical Research Division, US Bureau of the Census

Ivan P Fellegi and D Holt (1976)
A Systematic Approach to Automatic Edit and Imputation
*Journal of the American Statistical Association*, 71, 17–35

Hang J. Kim, Jerome P Reiter, Quanli Wang, Lawrence H Cox, and Alan F Karr (2013)
Multiple Imputation of Missing or Faulty Values Under Linear Constraints
Technical Report No. 181, National Institute of Statistical Sciences

United Nations Statistical Commission and Economic Commission for Europe (UNECE, 2000)
Glossary of Terms on Statistical Data Editing
presented at Conference of European Statisticians

Ton De Waal, Jeroen Pannekoek, and Sander Scholtus (2011)
Handbook of Statistical Data Editing and Imputation
Wiley

# Thank you!

**Hang J. Kim**

*hangkim@niss.org*

Triangle Census Research Network (TCRN)

http://sites.duke.edu/tcrn/

Duke University and National Institute of Statistical Sciences

# Appendix 1. Additional practical assumptions for measurement error model

Let $\mathcal{B}$ be an arbitrary support such that $\tilde{x}_i \in \mathcal{B}$ for all records $i = 1, \ldots, n$

**Uniform measurement error model**

$$f(\tilde{\boldsymbol{x}}_i | \boldsymbol{x}_i, \boldsymbol{s}_i) = \text{Unif}\left(\tilde{\boldsymbol{x}}_i^1 \in \mathcal{B}^1\right) \prod_{\{j: s_{ij}=0\}} I\left[\tilde{x}_{ij} - x_{ij}\right]$$

where

- $\tilde{\boldsymbol{x}}_i^1 \overset{\text{def}}{=} \{\tilde{x}_{ij} : s_{ij} = 1, j = 1, \ldots, p\}$
- $\mathcal{B}^1$: subspace of $\mathcal{B}$ on $(\sum_j s_{ij})$-dimension corresponding to the fields with errors

# Appendix 1. Additional practical assumptions for measurement error model (cont.)

## Additional practical assumptions for measurement error model

1. When $\tilde{\boldsymbol{x}}_i$ satisfies all edit rules,

$$f(\tilde{\boldsymbol{x}}_i|\boldsymbol{x}_i, \boldsymbol{s}_i) = \prod_{j=1}^{p} \delta\left(\tilde{x}_{ij} - x_{ij}\right).$$

2. When all inequality constraints but some balance edits are satisfied,

$$f(\tilde{\boldsymbol{x}}_i \mid \boldsymbol{x}_i, \boldsymbol{s}_i) = \text{Unif}\left(\tilde{\boldsymbol{x}}_i^1; \tilde{\boldsymbol{x}}_i \in \mathcal{X}\right) \prod_{\{j: s_{ij}=0\}} \delta\left(\tilde{x}_{ij} - x_{ij}\right).$$

3. When at least one inequality constraint is violated,

$$f(\tilde{\boldsymbol{x}}_i \mid \boldsymbol{x}_i, \boldsymbol{s}_i) = \text{Unif}\left(\tilde{\boldsymbol{x}}_i^1; \tilde{\boldsymbol{x}}_i \notin \mathcal{X}\right) \prod_{\{j: s_{ij}=0\}} \delta\left(\tilde{x}_{ij} - x_{ij}\right).$$

*Note* $\mathcal{X} \subset \mathcal{B}$