On Implementing a New Imputation Method into Production in the 2017 Economic Census Illustrated through Selected Vignettes

Katherine Jenny Thompson and Laura Bechtel
Economic Statistical Methods Division

The views expressed in this presentation are those of the authors and not necessarily those of the U.S. Census Bureau
General Set Up

BIG Problem

Research Team

Solution

Implementation Team

Solution

Production
The Next Three Presentations

BIG Problem

Research Team

New Method(s)

Implementation Team

Solution

Production
The Next Three Presentations

BIG Problem

New

Research Team

New Method(s)

Implementation Team

Additional Requirements

Integration challenges

Existing Processes

Storage

Production
My Discussion

Methods for treating missing product data

Research Team

Implementation Team

Production

Solution

New
Major Changes for the 2017 Economic Census

- North American Product Classification (NAPCS) - nearly 3,000 broad products and 5,000 detail products
  - Missing data treatment problem
  - New data item(s) problem
  - New economy-wide product-by-industry table problem

- All electronic data collection

- Only one option for reporting product sales data – rounded to $1,000
Example of Product Questions

- Retail Trade Industry – New Car Dealers
ITEM 22: DETAIL OF SALES, SHIPMENTS, RECEIPTS, OR REVENUE

Of the $,000.00 of Sales, Shipments, Receipts, or Revenue reported in Item 5, what was the value for each product or service?

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Retail sales of automobiles and light-duty trucks (Include all outright sales plus transportation charges, dealer preparation charges, and dealer-installed options. Deduct discounts granted to the purchaser as an increase in trade-in allowance over fair market value and manufacturers' rebates and incentives.) (Report rental of automobiles and light-duty trucks on line 9, leasing of automobiles and light-duty trucks on line 10, and maintenance and repair services and the value of service contracts for automobiles and light-duty trucks on line 11. Report wholesale sales of new and used automobiles on line 3 and wholesale sales of new and used light-duty trucks on line 4.)</td>
<td></td>
</tr>
<tr>
<td>a. Retail sales of new passenger cars (except fleet)</td>
<td>$     ,000.00</td>
</tr>
<tr>
<td>b. Retail sales of new passenger cars -- fleet</td>
<td>$     ,000.00</td>
</tr>
<tr>
<td>c. Retail sales of new vans and light-duty trucks, including minivans, cargo vans, sport utility vehicles (SUVs), and light passenger trucks (except fleet)</td>
<td>$     ,000.00</td>
</tr>
<tr>
<td>d. Retail sales of new vans and light-duty trucks, including minivans, cargo vans, sport utility vehicles (SUVs), and light passenger trucks -- fleet</td>
<td>$     ,000.00</td>
</tr>
<tr>
<td>e. Retail sales of used passenger cars (Include sales of passenger cars previously rented or leased.)</td>
<td>$     ,000.00</td>
</tr>
<tr>
<td>f. Retail sales of used vans and light-duty trucks, including minivans, cargo vans, sport utility vehicles (SUVs), and light passenger trucks (Include sales of vans and light-duty trucks previously rented or leased.)</td>
<td>$     ,000.00</td>
</tr>
<tr>
<td>g. Retail sales of all other powered transportation vehicles</td>
<td>$     ,000.00</td>
</tr>
</tbody>
</table>

Subtotal: $     ,000.00
5. Retail sales of automotive parts, supplies, and accessories
   (Report parts installed in repair for automobiles and light-duty trucks on line 11.)

   a. Retail sales of automotive lubricants, including oils, greases, etc.  
      $__ ,000.00

   b. Retail sales of new automobile and light-duty truck tires and tubes  
      $__ ,000.00

   c. Retail sales of new medium- and heavy-duty truck tires, including industrial, off-the-road, and farm tractor tires  
      $__ ,000.00

   d. Retail sales of retreaded or used automobile and light-duty truck tires  
      $__ ,000.00

   e. Retail sales of retreaded or used medium- and heavy-duty truck tires, including industrial, off-the-road, and farm tractor tires  
      $__ ,000.00

   f. Retail sales of automotive parts, new and rebuilt, including wheels (except batteries)  
      $__ ,000.00

   g. Retail sales of automotive parts, used, including wheels (except batteries)  
      $__ ,000.00

   h. Retail sales of automotive batteries  
      $__ ,000.00

   i. Retail sales of automotive accessories, including safety- and comfort-related items  
      $__ ,000.00

   j. Retail sales of automotive supplies, including appearance and maintenance chemicals, automotive paint, antifreeze, functional fluids, etc.  
      (Report automotive lubricants, including oils and greases, on line 5a.)  
      $__ ,000.00

   k. Retail sales of automotive audio equipment, components, parts, and accessories (Include radios, stereos, compact disc players, mp3 players [audio only], and other sound reinforcement and recording equipment.)  
      $__ ,000.00

   Subtotal  
      $__ ,000.00
Notes:

- Respondents can “write-in” products that are not pre-listed
- All **detail** product values should sum to their associated broad product value
- All **broad** product values should sum to the reported total value of sales (within a tolerance) for the establishment
Missing Data Treatment

**Research Team**

- \( \approx 1,000 \) industries
- \( \approx 8,000 \) products
  - Broad products
  - Detail products
- Sample data (i.e., sampling weights)
- Imputation cells
  - Cell collapsing rules
  - Minimum number of donors
- Restrictions on value (\( > \$1,000 \))
Research Challenges

- Team Composition (next slide)
- Short time frame (≈ 6-9 months)
- Magnitude of the problem
  - ≈ 1,000 industries and ≈ 8,000 products
- Historical data limitations (new collection)
Research Team Composition

- Methodologists
  - Very limited experience with Economic Census
  - Even less experience with “products”

- Subject matter experts
  - Very limited experience with imputation methodology

- NO Programmers/IT Specialists
Missing Data Treatment

Research Team
- \( \approx 1,000 \) industries
- \( \approx 8,000 \) products
  - Broad products
  - Detail products
- Sample data (i.e., sampling weights)
- Imputation cells
  - Cell collapsing rules
  - Minimum number of donors
- Restrictions on value (\( > \$1,000 \))
Missing Data Treatment

Research Team

- 1,000 25 industries
- 8,000 Top 10 products
  - Broad products
  - Detail products
- Sample data (i.e., sampling weights)
- LOCAL Imputation cells
  - Cell collapsing rules
  - Minimum number of donors
- Restrictions on value (→ $1,000)
Quick Summary

- Research conducted under restricted conditions
- SAS code developed in-house
  - Not ready for prime time!
- Recommendation: Hot Deck
  - Nearest neighbor or random
  - Unaddressed production requirements
Implementation Team

- Overlap
  - Team leader: Subject matter expert
  - Consultants: Methodologists (4)
- New members
  - Subject Matter Experts
  - Programmers
  - Methodologists (1 new)
Implementation Team

Common members (overlap)

Research Team
# Implementation Team

<table>
<thead>
<tr>
<th>Topic</th>
<th>Team Members</th>
<th>Knowledge Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic Census Processing</td>
<td>Subject Matter</td>
<td>Expert</td>
</tr>
<tr>
<td></td>
<td>Methodologists</td>
<td>Varied</td>
</tr>
<tr>
<td></td>
<td>Programmers/IT</td>
<td>Expert</td>
</tr>
<tr>
<td>Hot Deck Imputation</td>
<td>Subject Matter</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Methodologists</td>
<td>Expert</td>
</tr>
<tr>
<td></td>
<td>Programmers/IT</td>
<td>Low</td>
</tr>
</tbody>
</table>
Missing Data Treatment

Research Team
• 1,000 25 industries
• 8,000 Top 10 products
  • Broad products
  • Detail products
• Sample data (i.e., sampling weights)
• LOCAL Imputation cells
  • Cell collapsing rules
  • Minimum number of donors
• Restrictions on value (→ $1,000)

Implementation Team
• ≈ 1,000 industries
• ≈ 8,000 products
  • Broad products
  • Detail products
• Sample data (i.e., sampling weights)
• Imputation cells
  • Cell collapsing rules
  • Minimum number of donors
• Restrictions on value (> $1,000)
Missing Data Treatment

Research Team
- 1,000 industries
- 8,000 Top 10 products
  - Broad products
  - Detail products
- Sample data (i.e., sampling weights)
- LOCAL Imputation cells
  - Cell collapsing rules
  - Minimum number of donors
- Restrictions on value (>$1,000)

Implementation Team
- ≈ 1,000 industries
- ≈ 8,000 products
  - Broad products
  - Detail products
- Sample data (i.e., sampling weights)
- Imputation cells
  - Cell collapsing rules
  - Minimum number of donors
- Restrictions on value (> $1,000)
- “Must” products for industries
- Choice of hot deck method by industry
- Maximizing use of reported data
- Backup methods
- Processing time...
## Missing Data Treatment

### Research Team
- **1,000** 25 industries
- **8,000** Top 10 products
  - Broad products
  - Detail products
- Sample data (i.e., sampling weights)
- LOCAL Imputation cells
  - Cell collapsing rules
  - Minimum number of donors
- Restrictions on value (≥ $1,000)

### Implementation Team
- **≈ 1,000** industries
- **≈ 8,000** products
  - Broad products
  - Detail products
- Sample data (i.e., sampling weights)
- Imputation cells
  - Cell collapsing rules
  - Minimum number of donors
- Restrictions on value (＞ $1,000)
- “Must” products for industries
- Choice of hot deck method by industry
- Maximizing use of reported data
- Backup methods
- Processing time ...
Vignettes

1. Processing time

2. Imputation Cell Collapsing/Minimal Donor

3. Imputation for Detail Products
Vignette 1: Processing Time

- Programmer concerns
  - Prohibitive processing time
  - Complex coding

- Addressed by
  - Testing methods
    - Test deck of 2.4 million donor records (with over 20 million products) and 1.1 million full recipients
    - Independent validation of production code
  - Collaborative development of specifications
Vignette 2: Imputation Cells

- Needed
  - Imputation cell definitions
  - Collapsing rules
  - Minimum donors

- Limited research

- Education Process
Simple Example

- Cell Collapsing
  - Ideal = Industry By Tax Status By Unit Type
  - Collapse 1 = Industry By Tax Status
    - DROP Unit Type
  - Base = Industry
    - DROP Unit Type and Tax Status

- Minimum cell count = 5
  - Base = 1
## Compute Donor Counts
### Ideal Cells

<table>
<thead>
<tr>
<th>Industry</th>
<th>Tax Exempt Status</th>
<th>Unit Type</th>
<th># of Establishments</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAAAAA</td>
<td>Taxable</td>
<td>SU</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Taxable</td>
<td>MU</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>SU</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>MU</td>
<td>2</td>
</tr>
<tr>
<td>BBB BBBB</td>
<td>Taxable</td>
<td>SU</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Taxable</td>
<td>MU</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>SU</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>MU</td>
<td>4</td>
</tr>
<tr>
<td>CCCCCC</td>
<td>Taxable</td>
<td>SU</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Taxable</td>
<td>MU</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>SU</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>MU</td>
<td>2</td>
</tr>
</tbody>
</table>
## Determine Usage of Ideal Cells

<table>
<thead>
<tr>
<th>Industry</th>
<th>Tax Exempt Status</th>
<th>Unit Type</th>
<th># of Establishments</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAAAAA</td>
<td>Taxable</td>
<td>SU</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Taxable</td>
<td>MU</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>SU</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>MU</td>
<td>2</td>
</tr>
<tr>
<td>BBBBBBB</td>
<td>Taxable</td>
<td>SU</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Taxable</td>
<td>MU</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>SU</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>MU</td>
<td>4</td>
</tr>
<tr>
<td>CCCCCCC</td>
<td>Taxable</td>
<td>SU</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Taxable</td>
<td>MU</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>SU</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>MU</td>
<td>2</td>
</tr>
</tbody>
</table>

Less than 5 Establishments ⇒ Collapse
**Compute Donor Counts for Collapse 1 Cells**

<table>
<thead>
<tr>
<th>Industry</th>
<th>Tax Exempt Status</th>
<th># of Establishments</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAAAAA</td>
<td>Taxable</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>4</td>
</tr>
<tr>
<td>BBB BBB</td>
<td>Taxable</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>6</td>
</tr>
<tr>
<td>CCCCCC</td>
<td>Taxable</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>202</td>
</tr>
</tbody>
</table>

STILL Less than 5 Establishments ⇒ Collapse

All Counts ≥ 5 Establishments ⇒ Use Collapse 1 cells
## Compute Donor Counts for Base Cells

<table>
<thead>
<tr>
<th>Industry</th>
<th># of Establishments</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAAAA</td>
<td>8</td>
</tr>
<tr>
<td>BBBBBB</td>
<td></td>
</tr>
<tr>
<td>CCCCCC</td>
<td></td>
</tr>
</tbody>
</table>

All Counts $\geq 1$ Establishments $\Rightarrow$ Use base cells
## Final Cells for Hot Deck

<table>
<thead>
<tr>
<th>Industry</th>
<th>Tax Exempt Status</th>
<th>Unit Type</th>
<th>Hot Deck Cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAAAA</td>
<td>ALL</td>
<td>ALL</td>
<td>Base</td>
</tr>
<tr>
<td>BBBBBBB</td>
<td>Taxable</td>
<td>SU</td>
<td>Ideal</td>
</tr>
<tr>
<td></td>
<td>Taxable</td>
<td>MU</td>
<td>Ideal</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>ALL</td>
<td>Collapse 1</td>
</tr>
<tr>
<td>CCCCCCC</td>
<td>Taxable</td>
<td>ALL</td>
<td>Collapse 1</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>ALL</td>
<td>Collapse 1</td>
</tr>
</tbody>
</table>
## Final Cells for Hot Deck

<table>
<thead>
<tr>
<th>Industry</th>
<th>Tax Exempt Status</th>
<th>Unit Type</th>
<th>Hot Deck Cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAAAAA</td>
<td>ALL</td>
<td>ALL</td>
<td>Base</td>
</tr>
<tr>
<td>BBBB</td>
<td>Taxable</td>
<td>SU</td>
<td>Ideal</td>
</tr>
<tr>
<td></td>
<td>Taxable</td>
<td>MU</td>
<td>Ideal</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>ALL</td>
<td>Collapse 1</td>
</tr>
<tr>
<td>CCCCC</td>
<td>Taxable</td>
<td>ALL</td>
<td>Collapse 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### The Contention Point

<table>
<thead>
<tr>
<th>Industry</th>
<th>Tax Exempt Status</th>
<th>Unit Type</th>
<th># of Establishments</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCCCCC</td>
<td>Taxable</td>
<td>SU</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Taxable</td>
<td>MU</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>SU</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>MU</td>
<td>2</td>
</tr>
</tbody>
</table>

Subject Matter Experts and Programmers Contention
- NO need to collapse for SU unit type
- “NOT FAIR”
## Alternate Proposal

<table>
<thead>
<tr>
<th>Industry</th>
<th>Tax Exempt Status</th>
<th>Unit Type</th>
<th>Imputation Cell</th>
<th># of Establishments</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCCCCCC</td>
<td>Taxable</td>
<td>SU</td>
<td>Ideal</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Taxable</td>
<td>MU</td>
<td>Collapse 1</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>SU</td>
<td>Ideal</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>Exempt</td>
<td>MU</td>
<td>Collapse 1</td>
<td>202</td>
</tr>
</tbody>
</table>

### Methodology Concerns (Severe)
- Imputation cells no longer disjoint
  - Affects variance estimation
- Hurts probability of selecting MU donors
Impasse?

FOR ALTERNATIVE
Subject matter experts
Programmers/IT specialists
Methodologists (2.5)

AGAINST ALTERNATIVE
Methodologists (2.5)
Compromise/Decision

- Alternative Proposal

- Minimum number of donors = 1

- Lessons learned:
  - Put in measures to avoid unacceptable compromises
  - Include programmers and subject matter experts at the research stage
Vignette 3: Detail Products

- Not considered during research stage

- Limited historic data for research
  - Businesses more likely to report broad products than detail products
  - Different types of details by industry

- Subject matter experts wanted to maximize use of valid reported data in imputation
Imputation of Detail Products

- Group establishments into types
- Use category average (ratio) imputation
  - Statistical model frequently used for business data
  - NOT part of the research for product data
- “Hot deck” imputation – all products (broad & detail)
# Establishment Groups

<table>
<thead>
<tr>
<th>Donors</th>
<th>Broad products usable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>All required Detail products balance to Broad products</td>
</tr>
<tr>
<td>Partial</td>
<td>Some usable Detail products</td>
</tr>
<tr>
<td>Minimal</td>
<td>No usable Detail products</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recipients</th>
<th>Missing products</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>Need Broad products and Detail products</td>
</tr>
<tr>
<td>Partial</td>
<td>Need some (designated) Detail products</td>
</tr>
<tr>
<td>Minimal</td>
<td>Need all designated Detail products</td>
</tr>
</tbody>
</table>

| Ineligible | All products usable, but not “typical”; excluded from donor pool                   |
Complete Donor Example

Sales = BP 1 = DP1₁ + DP1₂ + DP1₃ + BP 2 = DP2₁ + DP2₂ + BP 3 = DP3₁ + DP3₂ + DP3₃ + BP 4
Partial Donor Example

Sales

= BP 1 = DP1₁ + DP1₂ + DP1₃

+ BP 2 = DP3₁ + DP3₂ + DP3₃

No usable detail products for broad product 2

BP 4
Mimimal Donor Example

No usable detail products for any broad product
“Completing” Partial Donors

Missing detailed products for BL 2 and 3

Category Averages for BL 2 and 3

Average proportion that each detail product contributes to a broad product

Completed record is now a donor
# Establishment Groups

<table>
<thead>
<tr>
<th>Donors</th>
<th>Broad products usable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>All required Detail products add to Broad products</td>
</tr>
<tr>
<td>Partial</td>
<td>Some usable Detail products</td>
</tr>
<tr>
<td>Minimal</td>
<td>No usable Detail products</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Recipients</th>
<th>Missing products</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>Need Broad products and Detail products</td>
</tr>
<tr>
<td>Partial</td>
<td>Need some (designated) Detail products</td>
</tr>
<tr>
<td>Minimal</td>
<td>Need all designated Detail products</td>
</tr>
</tbody>
</table>

**Ineligible**

| All products usable, but not “typical”; excluded from donor pool |

These units are both donors and recipients!
Going Back to Original Picture

BIG Problem

Research Team

Solution

Implementation Team

Solution

Production
Summary

ADVANTAGES

DISADVANTAGES
Summary

ADVANTAGES

• Workable solutions
• Buy-in on methods
• Shared understanding
• Research ideas for 2022

DISADVANTAGES

• A few “less than perfect” decisions
• Many extra meetings
Thank you

Contacts:

Katherine.J.Thompson@census.gov
Laura.Bechtel@census.gov