Quality of Data Processing

Wendy L. Martinez

FCSM 2018

March 7, 2018

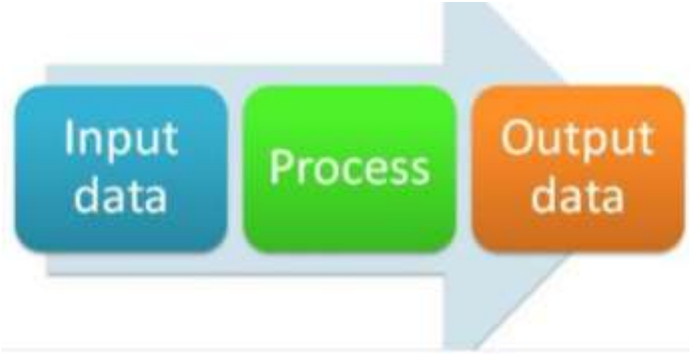
Disclaimer

- The findings and views expressed here are those of the author(s) and do not necessarily reflect the policies of the Bureau of Labor Statistics (BLS) or the Federal Government
- **Source**: Workshop 2 speakers, summary document by Alexandra Brown and Andrew Caporaso (JPSM)
- However… all mistakes are mine.

Members of Subgroup

- Joe Schafer, Census Bureau (Lead)
- Wendy Martinez, BLS
- Brian Sauer, Veterans Administration
- Lisa Mirel, National Center for Health Statistics

Three Workshops



Workshop 1: Quality of Input Data

December 1, 2017

Workshop 2: Quality of Data Processing

January 25, 2018

Workshop 3: Quality of Output Data / Synthesis

February 26, 2018

Questions to be Addressed

- In context of integrated data, what should be communicated to users of the final data products:
- **Fitness for use**:
  - Quality features when deciding to use a data source
  - Quality features to understand strengths and weaknesses of final product
- **Communication**: Best way to communicate quality features to diverse audience

BLS

Data Processing – Integrated Data

- **Record linkage**: exact match, privacy-preserving.
- **Using multiple frames**: drawing samples from two or more frames to improve coverage or reduce costs.
- **Statistical matching**: Joining two or more non-overlapping samples by variables shared in common, then applying modeling or imputation techniques to handle missing values.

Data Processing – Integrated Data

- **Models for combining statistics**: Combining estimates from different sources at national, ,subnational or subpopulation levels, as in small-area estimation.
- **Dimension reduction**: Techniques for summarizing unstructured data (e.g., images, free-form text)
- **Harmonization**: Combining information across data sets in the presence of mode effects, differing definitions or granularities.

BLS

Data Processing – Integrated Data

- **Edit and Imputation**: Other types of cleaning after data sources are combined.
- **Adjusting for Representativeness**: Making combined data more representative of the intended population.
- **Estimation**: Computing estimates of population quantities and associated measures of uncertainty

Data Processing – Integrated Data

- **Disclosure Avoidance**: Techniques for preventing re-identification of de-anonymization of individual records
- **Provenance and Curation of Metadata**: Preserving information about data sources, dictionaries, audit trails, etc.

Prioritizing the Topics

Which of these topics are

- substantially more complicated or qualitatively different when combining multiple data sources?
- less familiar to statisticians and methodologists?
- not well covered by existing standards for quality and transparency?
- not as well covered by existing literature (e.g. on Small Area Estimation or Total Survey Error)?
- not already covered in Workshop 1?

Prioritization of Topics

Topic	Priority (L/H)
1. Record linkage	**H**
2. Multiple frames	L
3. Statistical matching / data fusion	**H**
4. Combining aggregate statistics or estimates (as in SAE)	L
5. Dimension reduction / feature extraction	L
6. Harmonization across data sources	**H**
7. Edit and imputation	L
8. Adjusting for representativeness	L
9. Estimation	L
10. Disclosure avoidance	**H**
11. Provenance / curation of metadata	L

Workshop 2 – Speakers

- Record Linkage
  - Rebecca Steorts, Duke University
  - William Winkler, Census
- Harmonization of Data Across Sources
  - Ben Reist, Census
  - Don Jang, NORC
  - Scott Holan, University of Missouri

Workshop 2 – Speakers

- Combining Data by Statistical Matching, Imputation, and Modeling
  - Jerry Reiter, Duke University
  - Ed Mulrow, NORC
- Disclosure Avoidance: Frameworks, Techniques, and Quality Issues
  - Latanya Sweeney, Harvard University
  - John Abowd, Census

Record Linkage

- Rebecca Steorts talked about **entity resolution**.
- Defined as practice of joining multiple data sets by removing duplicate entries, often in the absence of a unique identifier.
- **Issues**:
  - Entity is same across data sets?
  - Matching in a quick and automated way
  - Metrics to evaluate quality of the match

Record Linkage

- One approach to entity resolution is de-duplication – first combining into single data set.
- Another is record linkage with researcher reviewing record linkage uncertainty of graphical structure – requires quadratic number of comparisons.
- Both approaches typically match on a unique identifier, if exists.
- Exact matching – features of records are compared.
- How close do they have to be for a match?
- Systematic method for evaluation needed.

Record Linkage Metrics

- Recall = 1 – False Negative Rate
- Precision = 1 – False Positive Rate
- Computational run time and complexity
- Robustness
  - Choices of training/testing data
  - Tuning parameters
  - Models

Record Linkage

- **Take-Away Messages**
  - Need for high-quality data sets where true matches are known
  - Transparency – statistical agencies showing what they are producing and how they do it
  - Additive error (Winkler) – 5% error in each of two linked data sets and a 5% matching error, the resulting data set has 15% error

BLS

Harmonization

- **Harmonization** is “the process of mapping and synchronizing data derived from multiple sources into a coherent data file for analysis.” (Jang)
- **Challenges**:
  - Data sources are hard to link
  - Data can vary in who/what they represent
  - No universal data quality measures to evaluate harmonized data
  - Integration and harmonization requires significant resources

Harmonization

- Ben Reist: Using survey estimates to assess the quality of administrative record data.
- Treating survey data as the ‘gold standard’ is a strong assumption.
- Can be used to adjust/improve estimates from administrative records

Harmonization

- Don Jang: Example with the Scientists and Engineers Statistical Data System – NSF
- Leverages estimates from 3 surveys.
- Harmonization is implemented at the question level – naming, formats, coding and editing rules are standardized across surveys.
- Response rates also have to be coordinated for weighting.

Statistical Matching

- Jerry Reiter: Statistical matching is used to blend data sets without unique identifiers.
- May be used to match data sets without overlapping observations.
- Goal – Learn associations Y and Z
- One file contains X and Y, X and Z.
- Joint distribution cannot be estimated from data alone.

Statistical Matching

- Some form of external information is needed.
  - Assumptions made about association between Y and Z given X – most common is conditional independence.
  - Another data set with Y and Z
  - Constraints on associations from other sources

Statistical Matching

- Quality measures to report:
  - What assumptions were made
  - What models were used
  - Quality of model fit
  - Results of sensitivity analysis
  - Provide metadata for files used
  - Steps taken to harmonize X variables (e.g., asked in similar ways?)
  - Edits performed
  - Potential for selection bias
  - …

Disclosure Avoidance

- Latanya Sweeney focused on protecting privacy while preserving data utility.
- 1997 – Sweeney was able to re-identify the governor of Massachusetts:
  - Data on health care utilization – public-use data file not compromising privacy
  - Voter registration data available for purchase

Disclosure Avoidance

- Matched on overlapping fields: Zip code, birth date, and gender
- In 1990 Census data, 87% of Americans are unique based on date of birth, gender, and zip code
- Suggests improving disclosure prevention where people expose vulnerability in current approach and develop method to address it.

Disclosure Avoidance

- Should report what disclosure prevention methods were applied.
- John Abowd suggests one introduce random noise that is statistically independent of any of the other distributions used.
- Necessary but not sufficient condition to prevent disclosure.

Summary Messages

- Data harmonization is a fundamental first step in blending multiple data sources.
- Data producers must be transparent about each step:
  - Original need to collect data
  - Harmonization steps
  - Matching procedures
  - Models used and assumptions
  - Evaluation techniques used
  - How privacy was maintained
- Decisions captured in metadata – users can judge utility

Contact Information

Wendy Martinez

Office of Survey Methods Research
Bureau of Labor Statistics
202-691-7400
martinez.wendy@bls.gov