Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Neuchâtel, Switzerland, 18-20 September 2018)

U.S. Federal Committee on Statistical Methodology Working Group on Transparent Quality Reporting of Editing and Imputation when Integrating Data from Multiple Data Sources

Prepared by Wendy L. Martinez, U.S. Bureau of Labor Statistics, United States

I. Introduction

1. At the request of the U.S. Chief Statistician Nancy Potok, a working group was convened by the U.S. Interagency Council on Statistical Policy (ICSP) and the U.S. Federal Committee on Statistical Methodology (FCSM) in September 2017. The group's objective was to propose quality standards for use with statistical information products and services that are based on multiple data sources. The group considered a wide range of issues in the development of such standards. Principal attention was directed toward standards related to transparency regarding the quality of input data sources; statistical methodology; and output data products and services. To reach their objective, the working group convened three workshops focusing on these areas. This paper will synthesize some of the concepts and recommendations for transparent quality reporting that were identified in these workshops. Given the goal of this UNECE working session, the main focus of the paper will be on findings and recommendations in the area of editing and imputation.

II. Background

A. Integrated Data Quality Initiative

2. The Interagency Council on Statistical Policy (ICSP) [1] is chaired by the U.S. Chief Statistician. The ICSP began operating officially in 1995. It is composed of the heads of the thirteen U.S. principal federal statistical agencies [2] and the Environmental Protection Agency Chief Statistician. Most of the members if the ICSP have management responsibility for statistical programs in their respective agencies. Their advice and cooperation are essential for effective implementation of OMB (Office of Management and Budget) statistical policy decisions and for planning improvements in federal statistical programs. The mission of the ICSP is to coordinate statistical work when activities and issues cut across agencies, to exchange information about agency programs and activities, and to provide advice to OMB on statistical matters.

3. The Federal Committee on Statistical Methodology (FCSM) is an interagency group whose mission is to improve the quality of federal statistics [3]. The FCSM was created by the OMB's Office of Statistical Policy to advise the ICSP on methodological and statistical issues that affect the quality of federal data. The FCSM members are career federal employees selected by OMB based on their individual expertise and interest in statistical methods. In addition to their advisory role, the mission of the FCSM is to:

- Compile, assess, and disseminate information on statistical or survey methods and practices for federal statistical agencies;

- Provide recommendations on issues of statistical methodology such as measurement, analysis, survey methods, survey errors, data collection methods and technologies, record linkage, disclosure limitation, and dissemination of information that affect federal statistical programs and improve data quality, including timeliness, accuracy, relevance, utility, accessibility, and cost effectiveness;
- Provide a forum for statisticians in different federal agencies to discuss issues affecting federal statistical programs; and
- Promote and support cooperative research across agencies on issues relevant to federal statistics.

4. Along with the Washington Statistical Society (WSS) [4], the FCSM sponsored three workshops focusing on documenting the quality of integrated data. As federal agencies, state and local governments, and private researchers develop new approaches to maximize the information available from the ever-growing body of existing data, integrated data sets are becoming more common. Integrated data include data from multiple input data sources and may include data files created via record linkage or other combination methods (e.g. concatenation and geographic linkages). They are also likely to combine data whose original purpose is different from the analytic need motivating the use of integrated data. Input data may be similar in terms of data types or may be composed of disparate types; e.g. survey data, administrative data, and less structured or unstructured data.

5. While the use of integrated data provides opportunities to learn more from our data, it also poses a number of challenges. Several of these are related to three key aspects of data quality: 1) the quality of the input data sources; 2) how the input data were processed into an integrated data set and the resulting data structure; and 3) how to convey information about the quality of the integrated data, any analyses performed, and the information drawn from those data. These challenges are compounded to the extent that the input data may be in formats that federal agencies and the statistical community are less familiar with in terms of quality metrics, and analyses of the integrated data may require analytic decisions guided by fewer well-accepted standards than agencies have typically followed.

B. Transparency

6. The working group defined *transparency* as follows. *Transparency* (sometimes referred to as *openness*) is the availability of documentation for a given estimate or data set that identifies the data sources and potential error associated with the methods of data collection, processing, and estimation. Descriptions of potential error should include the magnitudes of error whenever possible. Three levels of transparency are recommended:

a. A succinct, strategic summary for decision makers and the public who are not formally trained in statistical methods or data processing and analysis;

b. A more detailed summary for trained analysts who are infrequent or novice users of the estimate or data set; and

c. A detailed assessment for experienced users of the estimate or data set who are extending the data, developing in-depth analyses of the data, or developing ways to improve the estimate or data set.

C. Workshop Overviews

7. Each of the three workshops sponsored by FCSM and the WSS focused on one of the three key aspects of data quality: input data sources, data processing, and data outputs. The workshops provided information about best practices for documenting these three aspects of data quality. The ultimate purpose of these workshops was to inform a report that will provide recommendations to OMB for creating integrated data quality standards. The framework of the workshops was based in part on the U.S. National Academies of Science Engineering and Medicine (NAS) report *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps* [5] that described how we might move beyond existing data quality structures and address the quality considerations associated with integrating multiple data sources.

8. The first workshop looked at what information needs to be conveyed about the quality of the input data sets. There are well developed standards conveying data quality information about census and survey data. Less well documented are sets of quality metrics and information to provide potential users of administrative data, less structured data such as data collected through automated sensor systems, and unstructured data such as those captured through imagery or raw text in various formats. Emphasis was placed on how to evaluate and document the quality of input data from these latter types of data, though census and survey data documentation was included as a starting framework.

9. The second workshop focused on the processing of integrated data (including modelling procedures) and issues encountered during integration. Topics included those often associated with processing single data sets, such as how inconsistencies in the data are identified and corrected, how imputation is conducted, and how weighting is developed. However, these were discussed in the context of applying these concepts to the combined input data. Discussion also considered issues of data quality unique to integrated data, such as processes for linking individual records across data sets and how to blend input files when they are contributing unique cases to the resulting data set.

10. The third workshop discussed the quality of the resulting integrated data set and any estimates derived from it. Much of the information about the quality of the resulting data is derived from information about the quality of the input data and what processes were applied to blend them together. However, there are a number of unique metrics that need to be developed and conveyed to consumers of the integrated data. These include information about what sources were used and how much each contributes to an estimate, what processes were used to produce the estimate including descriptions of any key errors that may influence the interpretation of the results, and finally some quantitative measure of all combined errors.

11. Workshop organizers and speakers were asked to consider these two aspects of transparent reporting of the quality of integrated data:

a. **Fitness for use**: What quality features are important for data users to understand when either considering to use a particular integrated data source or evaluating the strengths and weaknesses of the resulting analysis?

b. **Communication with stakeholders**: What is the best way to communicate these predominant quality features to stakeholders, whether they be technical specialists, substantive data users, or the general public?

12. The next three sections summarize each of the three workshops. Along with the presentations at the workshop, a report written by faculty and graduate students from the (U.S.) University of Maryland served as the main source of information provided in these sections [6]. Slides for workshops presentations are available on the WSS website [7].

III. Data Inputs

13. Workshop 1 took place on December 1, 2017, where the goal was to identify quality issues associated with non-traditional data before they are integrated with survey data. The first part of the workshop looked at administrative data that is structured similarly to survey data. The second session of the workshop focused on semi-structured or unstructured data. For example, this type of data could come from automated sensory systems, images, or raw text. Each of the speakers discussed their experiences assessing data quality when working with auxiliary data sources.

14. Here is a brief summary of some presentations given in Session One, which focused on structured and administrative data sources:

a. Steven B. Cohen, Research Triangle Institute, *The Utility and Limitations in Administrative Data for Medical Care Expenditure Analysis*: Dr. Cohen described how data on medical expenditures could be integrated with survey data for analysis. Utilizing administrative records on medical

expenditures enables deeper exploration of rare conditions without added burden on the respondent. He proposed several factors to consider when determining the quality and fitness for use of a data source. These included the purpose of the data being gathered, the degree of documentation for the data, and the suitability with the research or analysis using the integrated data.

b. Michael Berning and David Sheppard, U.S. Census Bureau, *Quality of Administrative Records as Source Data*: The speakers mentioned the increasing number of administrative records available in the Census Bureau and the importance of distinguishing between high and low quality sources. They described several factors, such as relevance, accessibility, coherence, interpretability, accuracy, and institutional environment.

c. Bonnie Murphy and Crystal Konny, U.S. Bureau of Labor Statistics, *Quality Considerations for Administrative Data Used for the Producer Price Index (PPI) and consumer Price Index (CPI) Development*: These two speakers summarized how the PPI and CPI programs use alternative data sources to improve and augment price statistics. Among other things, administrative data are used for supplementing, validating, or imputing survey data. The U.S. Bureau of Labor Statistics (BLS) utilizes several data quality metrics to decide if an alternative data source is fit for use. These include coverage, accuracy, and reliability. They caution that using outside data sources removes quality control from the BLS, which limits transparency and the ability to quantify error.

15. Here is a brief summary of some presentations given in Session Two, which looked at less-structured data sources:

a. David Johnson, USDA National Agricultural Statistics Services, *Data Quality of Satellite Imagery for Studying Complex Systems*: David Johnson described satellite imagery being combined with additional data sources adding information on crop production and area reports. The unstructured imagery (data) can improve the quality and timeliness of the data in the reports. Measurements of uncertainty with this type of data are not well established, making it difficult to be transparent about data quality.

b. Roberto Rigobon, Massachusetts Institute of Technology, *Web-scraped Data, Considerations of Quality Issues for Federal Statistics*: Roberto Rigobon discussed using web-scraped data with federal data and statistics. He presented the advantages and disadvantages of using this type of data. He cautioned that data users should be award of these to determine if they are fit for use. Advantages of web-scraping are (1) it can be a non-intrusive collection method; (2) computer programs can read the information the researcher wants; and (3) it is quick and inexpensive to collect, since the process is automated. Disadvantages are (1) it might not be representative; (2) there are concerns about reliability and availability; (3) there could be an increased need for data editing with this type of data; and (4) data creation could involve unknown processing (e.g., cleansing, modelling) before it is posted on the web and subsequently scraped.

IV. Data Processing

16. Workshop 2 took place on January 25, 2018. The goal of this second workshop was to identify quality issues focusing on how input data are combined to form an integrated data set, and how the integrated data are structured. The aspects of data processing covered by the four sessions were record linkage, harmonization across data sources, statistical matching or data fusion, and disclosure avoidance.

17. The organizers had a larger list of potential topics to cover in the workshop that included multiple frames, combining aggregate statistics or estimates (e.g., small area estimation), dimension reduction or feature extraction, editing and imputation, adjusting for representativeness, estimation, and metadata. These topics were deemed lower priority because they are better understood and some were covered by other workshops.

18. Rebecca Steorts (Duke University and the U.S. Census Bureau) gave a talk on entity resolution. She defined it “as the practice of joining multiple data sets by removing duplicate entities, often in the

absence of a unique identifier." [6] The key issues with entity resolution are whether the entity being matched is the same; how they can be done in a fast and automated way; and what quality measures should be used. She mentioned de-duplication as an approach for entity resolution, along with record linkage. Steorts stressed the importance of robust methods for entity resolution and ways to measure robustness. She said the field needs more open-source software for entity resolution, additional approaches that promote reproducibility, improved evaluation metrics, and more transparency. The discussant in this session (William Winkler, U.S. Census Bureau) suggested several standard metrics to report data quality: precision, recall, and the reduction ratio.

19. The second session was on the harmonization of data across multiple sources.

a. Ben Reist (U.S. Census Bureau) gave a talk on using survey estimates to assess the quality of administrative data, especially when the survey is thought to be of better quality. This is in contrast to the more typical approach of using administrative data to improve surveys.

b. Don Jang (NORC at the University of Chicago) defined data harmonization as "the process of mapping and synchronizing data derived from multiple sources into a coherent data file for analysis." [6] Challenges for harmonization include (1) the process of linking records, (2) the data can vary in what they report and whom they represent, and (3) there are no general measures of data quality to evaluate harmonized data.

20. Session three is perhaps the one most relevant to this UNECE working group, as it was on combining data by statistical matching, imputation, and modelling. There was one speaker – Jerry Reiter from Duke University. He discussed various models and methods for statistical matching, which is used to integrate data sets in the absence of unique identifiers. These approaches could also be used to merge data sets that have no overlapping observations.

21. For instance, say we want to understand the association between variables Y and Z when we do not observe them together. One data set has variables X and Y, and another has variables Y and Z. In this case, the joint distribution of Y and X cannot be estimated from the data alone, and we need additional information to conduct the matching. This external information could be assumptions made about the association, another data set where Y and Z are observed together, or constraints on the associations based on domain knowledge. Some assumptions are necessary, even with this additional information. The most common one is conditional independence – Y is independent of Z, given X.

22. Reiter mentioned several approaches to statistical matching and imputation: (1) nearest neighbour hot deck, (2) regression modelling predicting Y from X to impute missing values of Y in the second data set; and (3) the use of auxiliary data. He listed the benefits and disadvantages of these approaches, and he concluded his talk by listing what quality measures and information agencies should report when integrating data. These are summarized here:

- Metadata for the files
- What X variables were employed
- Steps taken to harmonize X variables
- Edits that were done
- What matching method was used
- Assumptions used in building models
- Goodness-of-fit on models
- Results of sensitivity analysis
- Any potential selection biases when using auxiliary data

23. The last session focused on disclosure avoidance. There was one speaker – Latanya Sweeney (Harvard University) who gave a talk on the trade-off between privacy and utility. She gave some wonderful examples, one of which described how she was able to identify the governor of Massachusetts in a data set on the utilization of health care, which was released to the public. The data producers did not think it would compromise any individual's privacy. Perhaps, that was the case with the single data set. However she was able to merge the data with voter registration data, which was available for purchase.

She matched on zip code, birth data, and gender to uniquely identify the governor of Massachusetts. Sweeney proposed an idea to improve disclosure prevention in a similar way to data encryption. In other words, vulnerabilities are exposed in the current method, so improvements are developed to address the issues.

V. Data Outputs

24. Workshop 3 took place on February 26, 2018, and it focused on issues for conveying information on data quality associated with the final integrated data set and resulting estimates. Speakers noted that consumers of data products have varying levels of understanding and knowledge. Thus, the type, amount or level of detail provided by producers might not be suitable, appropriate, or necessary for all users. The levels of transparency discussed were

- **High**: This would be for academics, subject-matter experts, or agency professionals.
- **Moderate**: This would be appropriate for policy-makers, journalists, or students.
- **Low**: This is for the general public.

25. Paul Biemer (Research Triangle Institute) gave a talk on assessing and improving the accuracy of estimators. One approach he described was based on the Total Survey Error (TSE) model, which was originally developed for survey data. His approach incorporated integrated data sets and hybrid estimates based on them. He mentioned the importance of enabling data users to follow the logic and assumptions, so they can decide if the data products are fit for use.

26. This workshop also included information on work done outside the U.S. federal statistical system. John Czajka (Mathematica) presented several ways agencies could report on the quality of integrated data based on international standards and data quality profiles.

VI. Summary

27. Summaries of the three workshops and key themes are given below.

a. Workshop 1 was concerned with identifying data quality standards and issues for non-survey data. One session looked at administrative data, which is usually structured, and another session looked at less-structured data sets. An important point made in the sessions was that administrative data sources have many things in common with traditional survey data with respect to the assessment and reporting of data quality, while this is less true of less or unstructured data. Speakers were clear that data producers should provide information about the original reason for collecting the data and should be transparent regarding the strengths and weaknesses of the technology used to collect the data. This enables data users to understand the error properties of integrated data.

b. Workshop 2 concentrated on processing and modelling using data integrated from disparate sources with an emphasis on record linkage, statistical matching and data fusion, harmonization across data sources, and disclosure avoidance. Speakers emphasized the importance of treating harmonization as a separate process and one that should be planned for at the survey design stage. Regarding transparent reporting in editing and imputation, agencies should provide metadata, information on harmonization and edits, modelling assumptions, and an assessment of model fits. The importance of sensitivity analysis was stressed.

c. Workshop 3 looked at issues associated with the transparent reporting on quality of the output data, resulting estimates, and information gleaned from the integrated data. Frauke Kreuter (Joint Program on Survey Methodology, University of Maryland) summarized the main points from the workshops. Some key points brought up included: (1) the importance of assessing the quality of products at the level of estimates rather than the entire data set, (2) the need for collaboration, (3) the burden of assessing data quality will likely move from data collection to processing and

harmonization, (4) the metrics of data quality could depend on the research questions being asked, and (5) the need for general data quality measurements and reporting principles.

28. The overarching themes discussed by speakers and participants at all three workshops were the importance of transparent data quality reporting and clear communication with users. Data producers must be transparent about each step of the process, starting from the motivation for collecting the data in the first place, through the steps of harmonization and matching, to producing the final estimates and products. There was much discussion on the different levels of transparency, depending on the type of data user (e.g., level of knowledge, use).

29. It became clear that further exploration of specific important areas brought up in the three workshops would enable the group to better inform the ICSP and OMB. Thus, the group decided to hold two additional workshops in September 2018. One workshop will focus on sensitivity analysis associated with processing integrated data and the second will look at metadata requirements. Important takeaways from these two workshops will be included in the presentation at this UNECE meeting.

30. In addition, the workshop organizers and speakers built on a wide range of literature that is developing rapidly in this area, e.g., [8], [9] and references cited therein. The preparer of this paper also wishes to acknowledge the contributions from the members of the FCSM Working Group on Transparent Quality Reporting in the Integration of Multiple Data Sources. John Eltinge of the U.S. Census Bureau is the Chair. Other members of the working group and their affiliation are listed here.

Bureau of Justice Statistics
John Popham

Bureau of Labor Statistics
Wendy Martinez
Nicole Nestoriak
Polly Phipps

Bureau of Transportation Statistics
Rolf Schmitt

Census Bureau
Paul Marck
Joseph Schafer

Economic Research Service
Mark Prell

National Agricultural Statistics Service
Linda Young

National Center for Education Statistics
Chris Chapman
Richard Reeves

National Center for Health Statistics
Irma Arispe
Renee Gindi
Lisa Mirel
Carolyn Neal
Jennifer Parker

Statistics of Income Division, IRS
Brian Balkovic
Tamara Rib

Veterans Health Administration
Alon Bon-Ari
Ned Confer
Katherine Hoggatt
David Maron
Brian Sauer
Elanni Streja

References

[1] https://obamawhitehouse.archives.gov/omb/inforeg_statpolicy/bb-structure-federal-statistical-system

[2] https://nces.ed.gov/FCSM/agencies.asp

[3] https://nces.ed.gov/FCSM/index.asp

[4] http://washstat.org/

[5] National Academies of Sciences, Engineering, and Medicine. 2017. *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24893 or http://nap.edu/24893

[6] *Findings from the Integrated Data Workshops Hosted by the Federal Committee on Statistical Methodology and Washington Statistical Society*. 2018. Alexandra Brown, Andrew Caporaso, Katharine G. Abraham, and Frauke Kreuter, University of Maryland.

[7] http://washstat.org/presentations/

[8] "From multiple modes for surveys to multiple data sources for estimates." 2014. Connie Citro. *Survey Methodology*, **40**:137-161, https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14128-eng.pdf?st=PflT6FYU

[9] "Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models," 2017. D. L. Oberski, A. Kirchner, A. Eckman, and F. Kreuter, *Journal of the American Statistical Association*, **112**:1477-1489. https://www.tandfonline.com/doi/pdf/10.1080/01621459.2017.1302338