Assessment of Manual vs Automated Survey Editing and Imputation

Sean Rhodes*, Denise A. Abreu, Megan Lipke, Jennifer Maiwurm,

Darcy Miller, Linda J. Young

The findings and conclusions in this presentation are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

Introduction

- National Agricultural Statistics Service (NASS)
  - Provides timely, accurate, and useful statistics in service to U.S. agriculture
  - Conducts over 100 surveys and prepares over 400 reports on virtually every aspect of U.S. agriculture
- Census of Agriculture
  - Every five years
  - Detailed census of every farm and agricultural producer in the country

Crops APS

- Acreage, Production, and Stocks (APS)
- These quarters are tied closely to how crops develop across the U.S.
  - March: Planting Intentions
  - June: Planted Acres
  - September: Small Grains
  - December: Row Crops

- Stakeholders use results and estimates created by these surveys via producers' responses

Background

- NASS surveys have a three-step data cycle:
  - Data collection
  - Analysis
  - Publication
- Editing and Imputation
  - Data quality and consistency
  - Manual process: Blaise System
  - Edit logic

IDEAL

- Imputation, Deterministic Edits, Automation and Logic (IDEAL)
  - Editing and imputation are managed in separate steps
  - Greater flexibility
- Generalized System
  - Edit logic and imputation methods can be easily shared across surveys for the same set of variables
- Multiple Components
  - User interface
  - System architecture
  - R engine

What is JIMMY?

- R engine is named JIMMY
- Applies rules and make automated changes to data
  - Deterministic edits
  - Imputation
  - Fellegi-Holt
- Statistics Netherlands R packages (Van Der Loo and de Jonge)

Testing 2023 September Crop APS

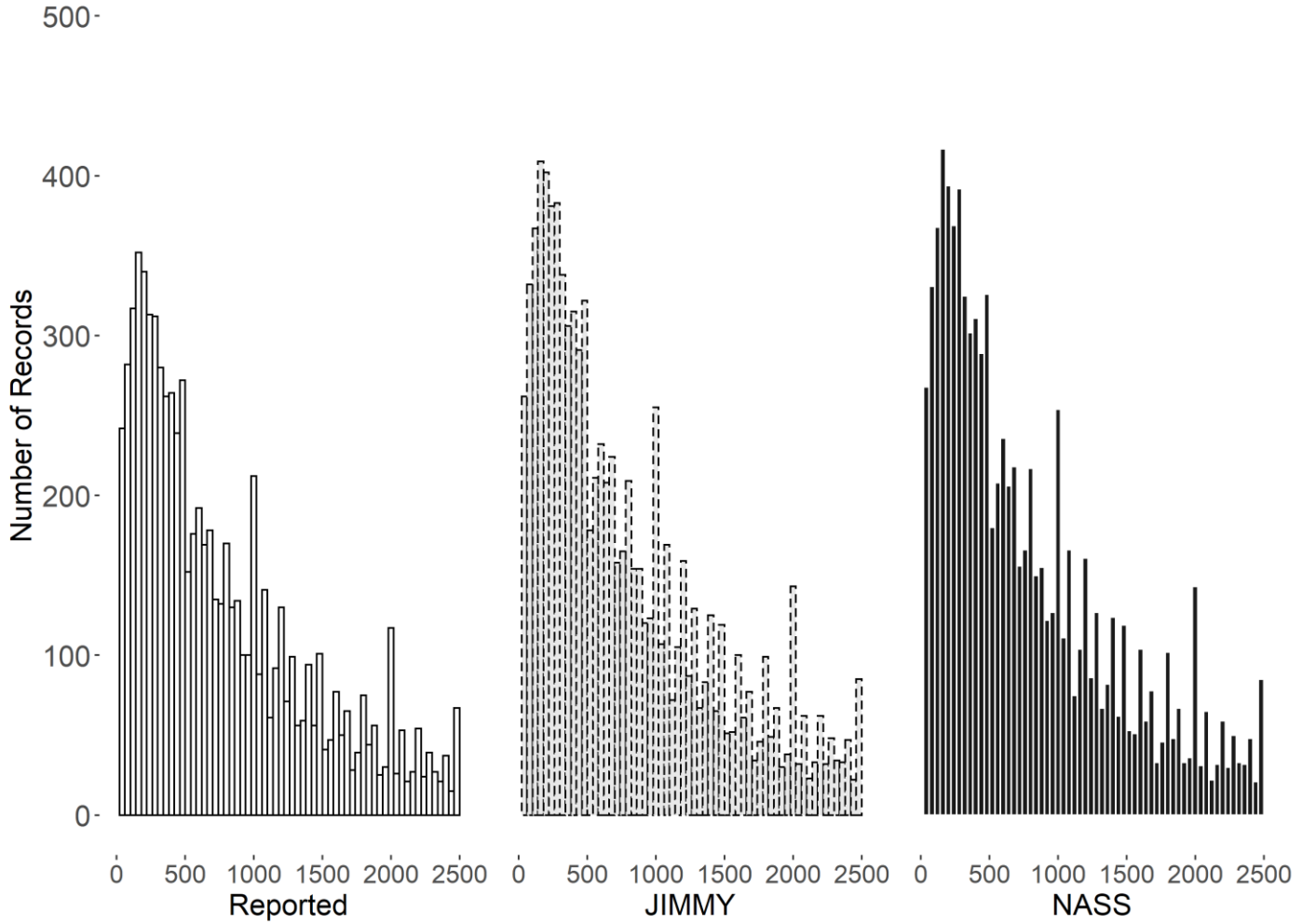
- Utilize previous testing methods
  - Reported, JIMMY, NASS
  - Compare distributions
  - Catalog edits
- New methods
  - Stratified sample
  - Assess
    - Data quality
    - Workload
  - Stakeholder analyst review

Assessing Quality of Data Processed

- Tested JIMMY on whole September Crops APS 2023 sample over 60,000 records
- 11,085 records were chosen
  - Corn Belt: Illinois, Iowa Minnesota, Missouri, Wisconsin
- Results
  - Single variable
  - Unit level
  - Qualitative responses

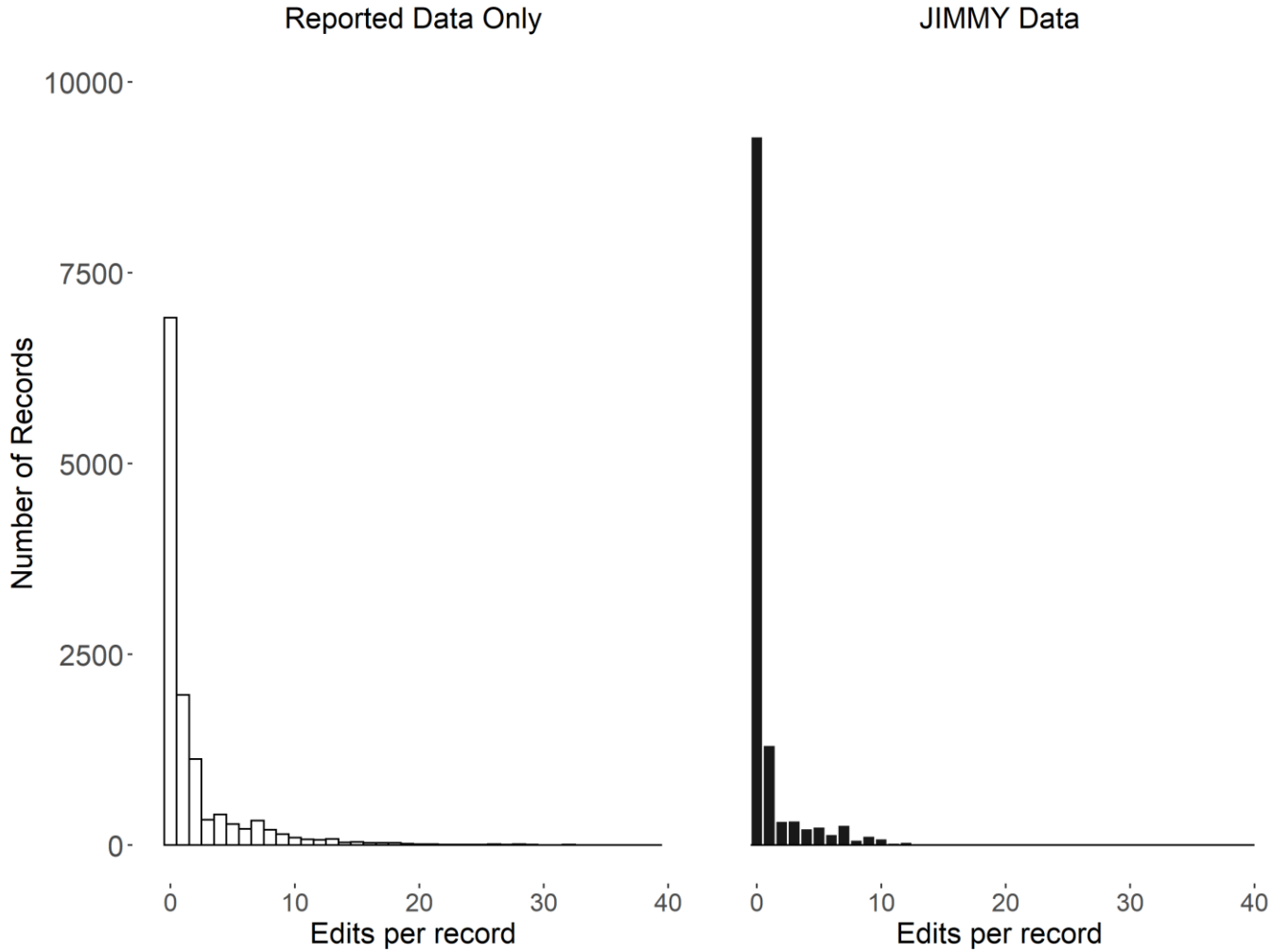
Results: Total Land Operated

- Similar distributions
- JIMMY data was produced by automated edits and imputation to the reported data
- Distributions do not follow a normal distribution



Results: Workload and Data Quality

- Number of edits per record
  - Reported
  - JIMMY
- Positively skewed towards zero
  - Require little to no manual editing
- Large number of records are closer to zero due the automatic editing



Results: Workload and Data Quality

- Reduction of edits required by an RFO analysts with and without JIMMY automatic editing
- Total number of edits were aggregated for the sample
- Almost 50% reduction in the number of edits due to JIMMY's automatic editing

Results: Workload and Data Quality

- RFO analysts completed reviews of 76 records over nine-week period
- Produced deeper insight into complex records
- For many records, analysts found JIMMY data to be reasonable
- For edits thought to be unreasonable, extensive insight was gained through discussion

RFO Weekly Review

Week	Number of Records
1	15
2	9
3	13
4	9
5	18
6	10
7	6
8	10
9	9

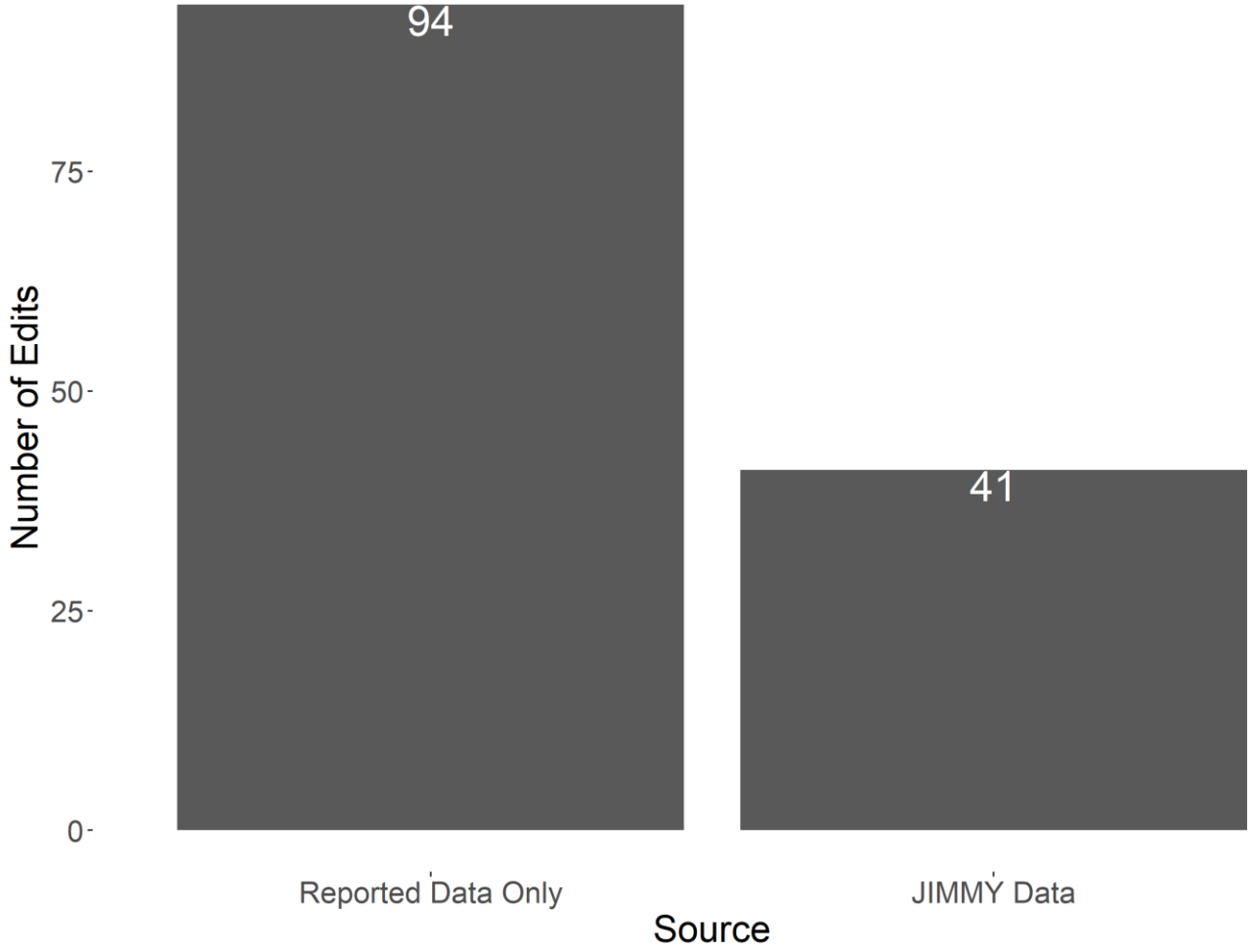
Conclusions

- Successes
  - Results of testing on historical data are promising
  - Results have shown incorporation of additional testing techniques have been advantageous
    - Stratified sample
    - Blaise system
    - RFO qualitative feedback

- Challenges
  - Number of edits to evaluate
  - Relationships between edits
  - Defining edits to be reasonable

September 2024 Results

- Random sample of 50 records from September APS 2024
- Total number of edits were aggregated for the sample
- Over 50% reduction in the number of edits due to JIMMY's automatic editing



Select References

- Manning, A. and Atkinson, D. (2009). “Toward a Comprehensive Editing and Imputation Structure for NASS – Integrating the Parts”. *USDA NASS RDD*. United Nations Statistical
- Dau, A. and Miller, D. (2018). “Dancing with the Software”. 2018 Joint Statistical Meetings. Vancouver, BC, Canada, 28 July – 2 August, 2018.
- E.de Jonge and M. van der Loo, "errorlocate: Locate Errors with Validation Rules," 2018.
- E. de Jonge and M. van der Loo, "validatetools: Checking and Simplifying Validation Rule Sets," 2019.
- Miller, D. (2021). “Growing a Modern Editing and Imputation System”. 2021 Federal Committee on Statistical Methodology Conference.
- Miller, D. and Young, Linda (2015). “Imputation at the National Agricultural Statistics Service”. United Nations Statistical Commission and Economic Commission for Europe, Conference for European Statisticians, Work Session on Statistical Data Editing. Budapest, Hungary, 14-16, September 2015.
- Lipke, M., Miller, D., Wagner, V., Brown, K. and Agnihotri, V. (2022). “Growing a Modern Edit and Imputation System”. United Nations Statistical Commission and Economic Commission for Europe, Conference for European Statisticians, Work Session on Statistical Data Editing. Virtual, 3-5 October 2022
- Laird, M., Maiwurm, J., Lipke, M., Miller, D., Denwiddie, M., Wagner, V., Brown, K. and Agnihotri, V. (2023) “Growing and Testing a Modern Edit and Imputation System". International Conference on Agricultural Statistics, Washington DC, United States, 17-19 May 2023

Thank you!

Sean.Rhodes@usda.gov