

Rule-Based Data Validation and Reconciliation of Survey Responses

AUGUST 7, 2024



Disclaimer

The findings and conclusions in this review of literature are those of the author and should not be construed to represent any official USDA or U.S. government determination or policy.

Road map

- **Introduction to work**
- **Issues statistical agencies face**
- **Innovations for addressing issues**
- **IDEAL**
- **Next steps**



Introduction and motivation



- Each year, the U.S. Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) conducts more than 300 surveys to understand and enumerate every aspect of U.S. agriculture.



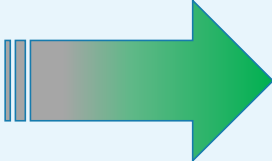
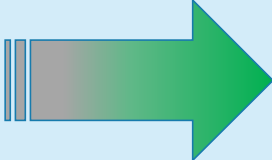
- Ensuring that survey responses are **valid, reliable, and internally consistent** is vital to publishing accurate official statistics:
 - The quality of survey responses varies with survey and respondent.
 - A significant amount of manual labor is required to edit and impute missing or incorrect survey responses.



- As part of an agencywide modernization effort, NASS is looking at **automating the editing and imputation processes** to improve the quality, consistency, and efficiency of its survey data processing. This will be done through the Imputation, Deterministic Edits, And Logic (IDEAL) engine.



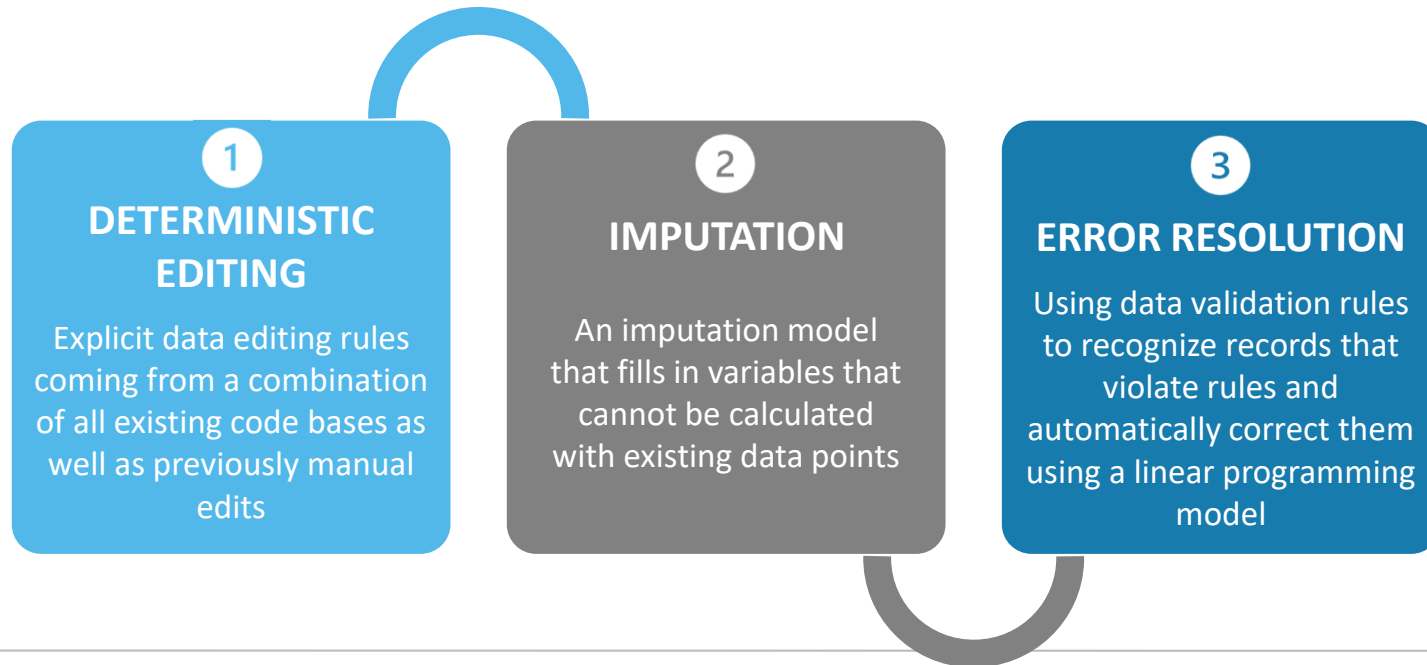
IDEAL resolves issues statistical agencies face

Issue		Innovation
<p><i>Aging code bases</i></p> <ul style="list-style-type: none">• Obsolete coding language• Falling behind innovations, such as new algorithms and cloud computing		<p><i>Error resolution</i></p> <ul style="list-style-type: none">• Automatically correct data to enhance speed and consistency
<p><i>Decentralization</i></p> <ul style="list-style-type: none">• Multiple editing systems that could be one• Lack of consistency		<p><i>Centralization</i></p> <ul style="list-style-type: none">• Develop a singular ruleset and treat it as its own input• Automate edits done by hand



IDEAL comprises three interrelated components

IDEAL is the R script-based data engine that can process USDA NASS datasets.
It has three main components:



Error resolution: Fellegi-Holt's principle of parsimony

Basis of Fellegi-Holt

Implement an edit by correcting the smallest number of items possible by the smallest amount.

Implementing Fellegi-Holt

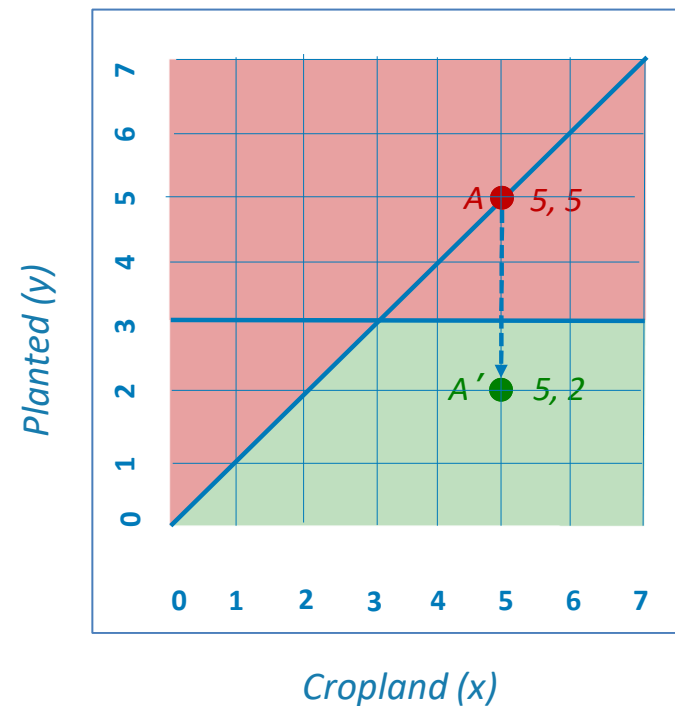
Rules:

Rule 1: Cropland \geq Planted

Rule 2: Planted < 3

Data point:

A: (5,5) **A':** (5,2)



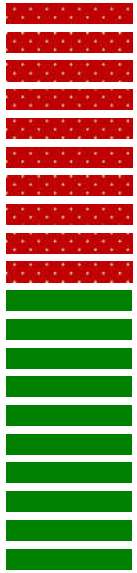
IDEAL is efficient and effective

Parallel Cloud Processing

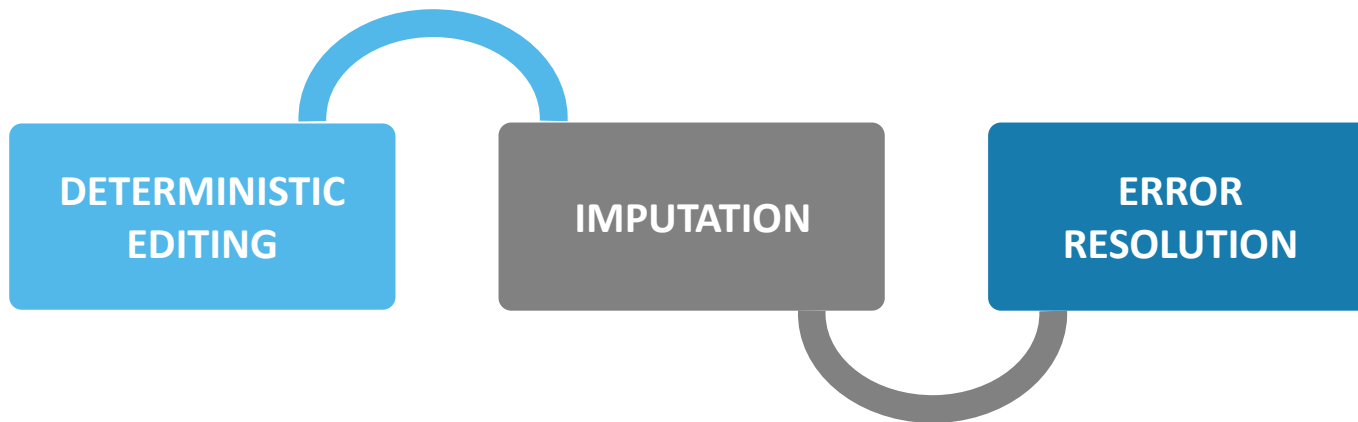
Batches of 1,000 records can be processed simultaneously in around 1 hour, scaling **up to 60,000 records** within the same timeframe.

50%

Clean Rate

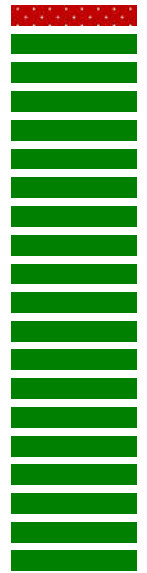


Before IDEAL



90%

Clean Rate



After IDEAL



Items to consider

Interplay between academic ideals and practical challenges

- Speed and timing of process, availability of rules
 - More than 70,000 records
 - Approximately 400 possible variables

Lessons learned

- Managing business rules is challenging, especially over a span of more than 30 years and with numerous analysts involved.
- While automatic error correction is highly effective, analysts are still essential for addressing the most complex cases.



Special Thanks

- **Joe Parsons**
- **Linda Young**
- **Lance Honig**
- **Denise Abreu**
- **Vikas Agnihotri**
- **Karl Brown**
- **Megan Lipke**
- **Jennifer Maiwurm**
- **Darcy Miller**
- **Sean Rhodes**



Contacts

Gunnar Ingle



gunnar.ingle@summitllc.us



[linkedin.com/in/gunnar-ingle-a8b92b145](https://www.linkedin.com/in/gunnar-ingle-a8b92b145)

Albert Lee, PhD



albert.lee@summitllc.us



[linkedin.com/in/albert-lee-90051b5](https://www.linkedin.com/in/albert-lee-90051b5)

