A Modular Approach to Survey Editing and Imputation for Agriculture Statistics

OCTOBER 24, 2024

summit

Disclaimer

The findings and conclusions in this presentation 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**
- **Modularization**
  1. Edits and routing
  2. Imputation
  3. Error Resolution
- **Metrics**
- **Summary**

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.

Modularization

Building data through known datapoints

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

Explicit data editing rules coming from a combination of all existing code bases and from previous manual edits

An imputation model that fills in variables that cannot be calculated with existing data points

Identifies records that violate data validation rules and automatically corrects them using a linear programming model

Deterministic Edits and Routing

Deterministic Edits:
"If I know how many acres are owned and rented but the total land is missing, **then** I can calculate it."

LAND_OWNED > 0 & LAND_RENTED > 0 &
LAND_TOTAL == MISSING
THEN LAND_TOTAL := LAND_OWNED + LAND_RENTED

Routing:
"If you cannot calculate production since yield and production are missing **then** mark production for imputation."

LAND_HARVESTED > 0 & LAND_YIELD == MISSING &
LAND_PRODUCTION == MISSING
THEN LAND_PRODUCTION == IMPUTATION MARKER

Why begin with Deterministic Edits?

Deterministic Edits generate values based on known data. Routing then guides JIMMY by indicating what is expected for a given record and identifying any missing information.

Imputations

For a record $x = (x_o, x_m)$ with observed values x_o, missing values x_m

1. Given an estimate for the mean vector **$(\hat{\mu})$** and covariance matrix **$(\hat{\Sigma})$**, these can be arranged according to the missingness pattern:

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_o \\ \hat{\mu}_m \end{pmatrix}, \text{ and } \hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{00} & \hat{\Sigma}_{0m} \\ \hat{\Sigma}_{mo} & \hat{\Sigma}_{mm} \end{pmatrix}$$

2. The imputations estimate the missing values using the following (ordinary least squares-like) equation:

$$\hat{x}_m = \hat{\mu}_m + \hat{\Sigma}_{mo} * \text{inv}(\hat{\Sigma}_{00}) * (x_o - \hat{\mu}_0)$$

$$\hat{\beta} = (X'X)^{-1} (X'Y)$$

Why follow Deterministic Edits with Imputations?

Deterministic edits correct data based on specific rules, but some values may still be missing. Imputations use Previously Reported Data (PRD) to estimate these missing values, ensuring that both calculated and previously unavailable values are accounted for, resulting in a complete dataset.

Error Resolutions

Error correction rules

These rules are conditional statements in the USDA code that signal to an analyst that something is logically incorrect about the dataset.

Basis of Fellegi-Holt

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

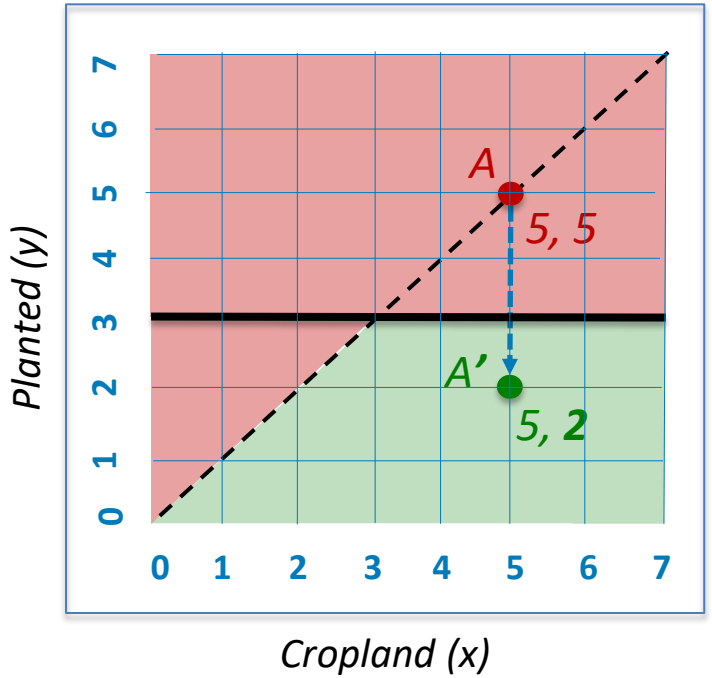
Implementing Fellegi-Holt

Rule 1: Cropland >= Planted (- - - -)

Rule 2: Planted < 3 (▬)

Why end with Error Resolutions?

Error Resolutions evaluate all values from Deterministic Edits and Imputation, ensuring they are checked holistically.



Metrics: September 2024 Agriculture Production Survey

What does this look like in practice?

Record with a Critical Error		% of Records with a Critical Error	
Pre-processing	***Post-processing***	***Pre-processing***	***Post-processing***
22,377	13,028	78.6%	43.9%

Critical Errors per Record

Pre-processing	Post-processing	Average Change
2.29	1.11	-1.17

Run Times

Total Records Proceeded*	Median Seconds Per Record
30,906 records	4.95

*IDEAL only processed a fraction of the entire September APS sample.

IDEAL Summary

Modularization

- Changes to the methodology of one script will not disrupt the functions of any others.

Things to Note

- 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

Nicole Schwartz

nicole.schwartz@summitllc.us

linkedin.com/in/nicole-schwartz