Rule-Based Data Validation and Reconciliation of Survey Responses

OCTOBER 25, 2023

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 and Motivation**
- **Error-correction process**
  - Defining error-correction rules
  - Automated error corrections
- **Transforming data to improve error correction**
  - Imputing missing values
  - Augmenting edit rules
- **Error-correction performance**

Introduction and motivation

- Each year, the U.S. Department of Agriculture's National Agricultural Statistics Service (NASS) conducts more than 100 surveys to understand and enumerate every aspect of agriculture in the United States.

- 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 agency-wide modernization effort, NASS is looking at **automating the editing and imputation processes** to improve the quality, consistency, and efficiency of its survey data processing.

Benefits to NASS

- **Saves time**
  - Automates many edits that analysts routinely and consistently make.
  - Frees NASS analysts to pursue more difficult cases—further improving data quality.
- **Improves consistency**
  - Uses an algorithm in comparison to personalized edits and imputations.
  - Allows for consistency across surveys, regions, administrators, and time.
- **Makes rules catalog explicit to more users**
  - Condenses entire rules universe into a singular file with consistent structure.
  - Centralizes and organizes each rule catalog to facilitate consistent updates and management.

Before error correction—deterministic edits and imputation

- **Deterministic edits**
  - Each survey has a host of edit rules, for example:
    - "If I know how many acres are owned and rented but the total land is missing, I can calculate it."
    - LAND_OWNED > 0 & LAND_RENTED > 0 & LAND_TOTAL == MISSING
      THEN LAND_TOTAL := LAND_OWNED + LAND_RENTED

- **Imputation**
  - The goal is to have values in the ballpark, which are then fixed in error correction.
  - Mean imputation using one draw from a multivariate normal.
    - Uses historical information

Defining 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.

Examples:

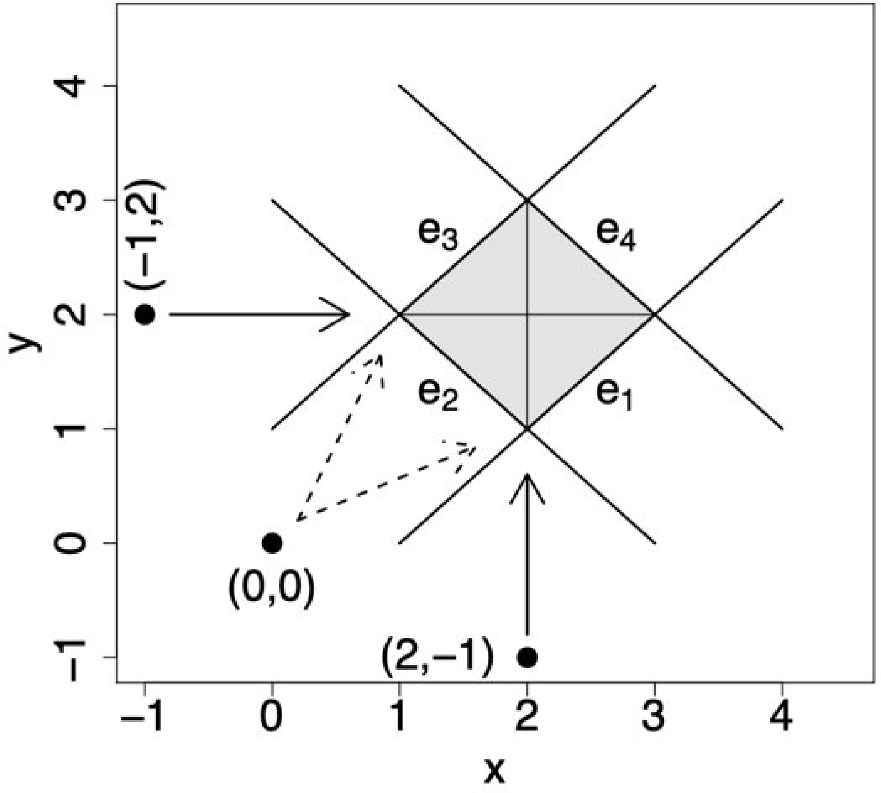
```
If Farm planted Crop A
Then
Acres_Planted_CropA >= Acres_Harvested_CropA
```

or

```
If Farm has rented acreage
Then
Acres_Cultivated <= Acres_Owned + Net_Acres_Rented
```

Fellegi-Holt's principle of parsimony

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



Source: Statistics Netherlands (2011)

R packages and implementation

- **R package**
  - *Validate*
    - Used to declare data validation rules and confront data to find violated rules in records.
  - *Errorlocate*
    - Uses the *lpSolveAPI* to solve the liner problem and output solution values.

- **Implementation**
  - Issues
    - Linear rules are required for R packages.
    - Rules must be explicit.
    - Nonlinear functions including rounding.
  - Solutions
    - Multiplication: Log values.
    - Range Check.

Error-correction performance

- **Dataset**
  - Over 30,000 Records
  - 150 + Variables
- **Results from error correction**
  - 151,000 + values that get an error correction
  - 21% of values dirty before error correction
  - 7% of values dirty after error correction

Further thoughts

- **Repeatable process**
- **Interplay between academic ideals and practical challenges:**
  - Speed and timing of process / availability of rules.
- **Lessons learned**
  - Business rule management is difficult, especially over more than 30 years and many analysts.
    - Code parsers are necessary but not sufficient.
    - Error rules are frequently not independent of deterministic edits.
  - With human editors, code is not only source of rules.
  - Automatic error correction is very good, but analysts are needed for the worst cases.

Special Thanks

- Linda Young
- Joe Parsons
- Denise Abreu
- Vikas Agnihotri
- Karl Brown
- Megan Lipke
- Jennifer Maiwurm
- Darcy Miller
- Sean Rhodes
- Vito Wagner