Looking at the Forest not the Trees: Multiple Uses for Regression Trees in Surveys

Jaki S. McCarthy

USDA National Agricultural Statistics Service

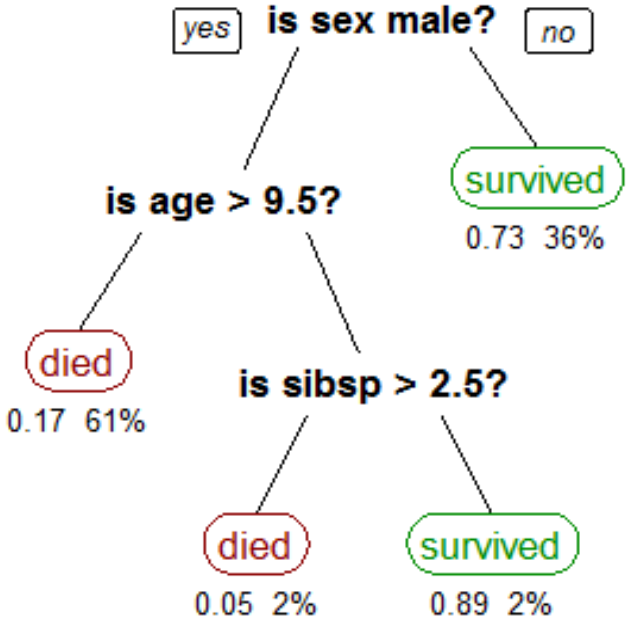
AAPOR Annual Conference 2018

Denver Colorado

Regression or Classification Trees

- Used to partition a (usually large) data set with respect to a target based on input variables
- Advances in computing power and availability of software make this possible with large datasets and many variables (Loh, 2014)

Predicting survival on the titanic



By Stephen Milborrow - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=14143467

Key strengths of classification trees

- No hypothetical models needed – results purely data driven
- Large numbers of predictor variables can be included
- Automatically selects important variables and cut points
- Interactions are inherent in the model
- Input variables can be correlated
- Missing data does not have to be imputed or records deleted
- Resulting trees can be easily interpreted

How can survey organizations leverage this modeling technique?

"By the fruits which it bears is the tree known."
-- **Jan Hus**

Lots of interesting statistical methods problems…..

- Impact of model parameters
- How to evaluate model performance
- How to validate models
- How to incorporate costs into models
- How to improve prediction through ensemble methods

But are there more applications for tree models in survey organizations?

NASS Uses of Classification Trees

- Survey weighting
- Data collection planning
- Classification of list frame units
- Identification of specific respondent subgroups
- Identification of important respondent characteristics

"The clearest way into the Universe is through a forest wilderness."
— John Muir

Nonresponse weighting

- Classification trees used to create weighting classes based on nonresponse propensities
  - Toth and Phipps, 2014; Lohr, et al, 2015; Buskirk and Kolenikov, 2015; Loh, et al, 2017
- In this case:
  - Target is response
  - Predictors are auxiliary variables available for all cases
  - Model is applied to new cases to create response propensity groups
  - Response propensity in each group used to create NR weights
  - Assumes that predictors are related to BOTH response and estimates of interest

NR weighting in the Census of Agriculture

- 2007 COA used classification trees to group records into weighting classes for NR weighting (Cecere, 2008)
- Inputs: frame data known prior to the census
- Records grouped into NR propensity “leaves”
- Nonresponse weights generated within those tree nodes

Trees for Adaptive Survey Design

- NR classification tree models were less effective than existing calibration models for NR weighting (Earp, Mitchell, Kott, Kreuter, 2012)
- But can be used to manage data collection (McCarthy, 2013; Toth and Phipps, 2012)
- NASS had developed models to classify farm operations:
  - Overall nonresponse propensity
  - Propensity to refuse cooperation; to remain a noncontact
  - Propensity to respond by mode
- Models can be used to adapt data collection strategies for subsets of the sample

But where else might classification trees be useful in survey organizations?

- Targets can include:
  - Nonresponse
  - Survey eligibility
  - Reporting errors
  - Land classification
  - What else?

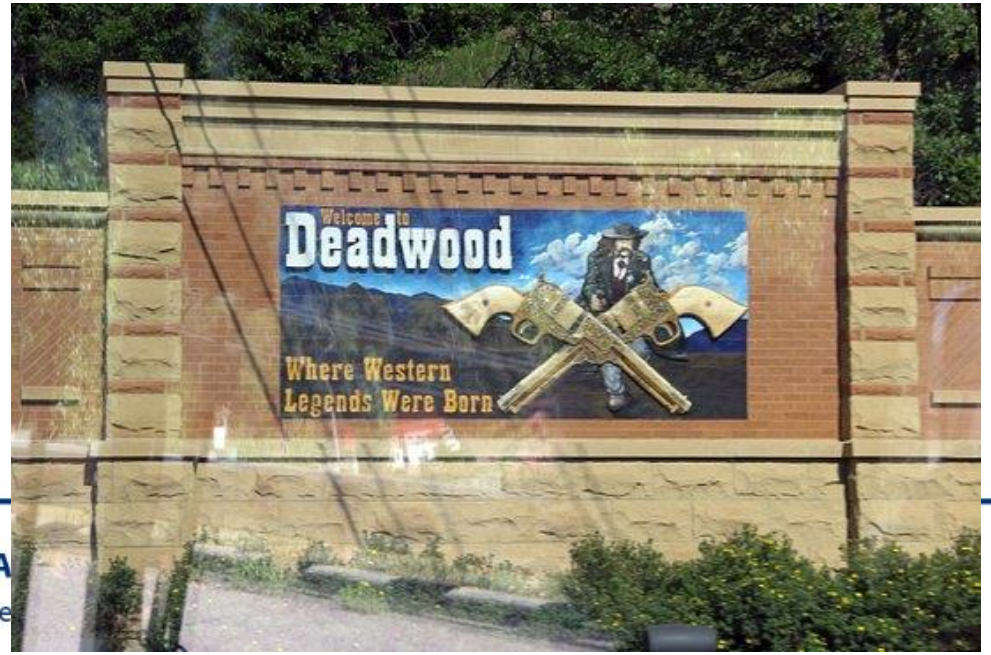
"There are rich counsels in the trees."
— Herbert P. Horne

List Frame Trimming

- For the Census of Agriculture, NASS maintains a large list frame
- Over 3 million records, to count just over 2 million farms
- Classification tree models used to trim records from the list frame (Garber, 2009)
- Records with known status (farm/non-farm) were used to build model
- Inputs were frame data
- Model then applied to records with unknown status to identify records with highest likelihood of being non-farms

Identifying List Frame Deadwood

- Classification trees used to identify sample units likely to be deadwood (i.e. out of business)
- Units that went from farm to non-farm status identified
- Inputs: farm characteristics, administrative info, response history

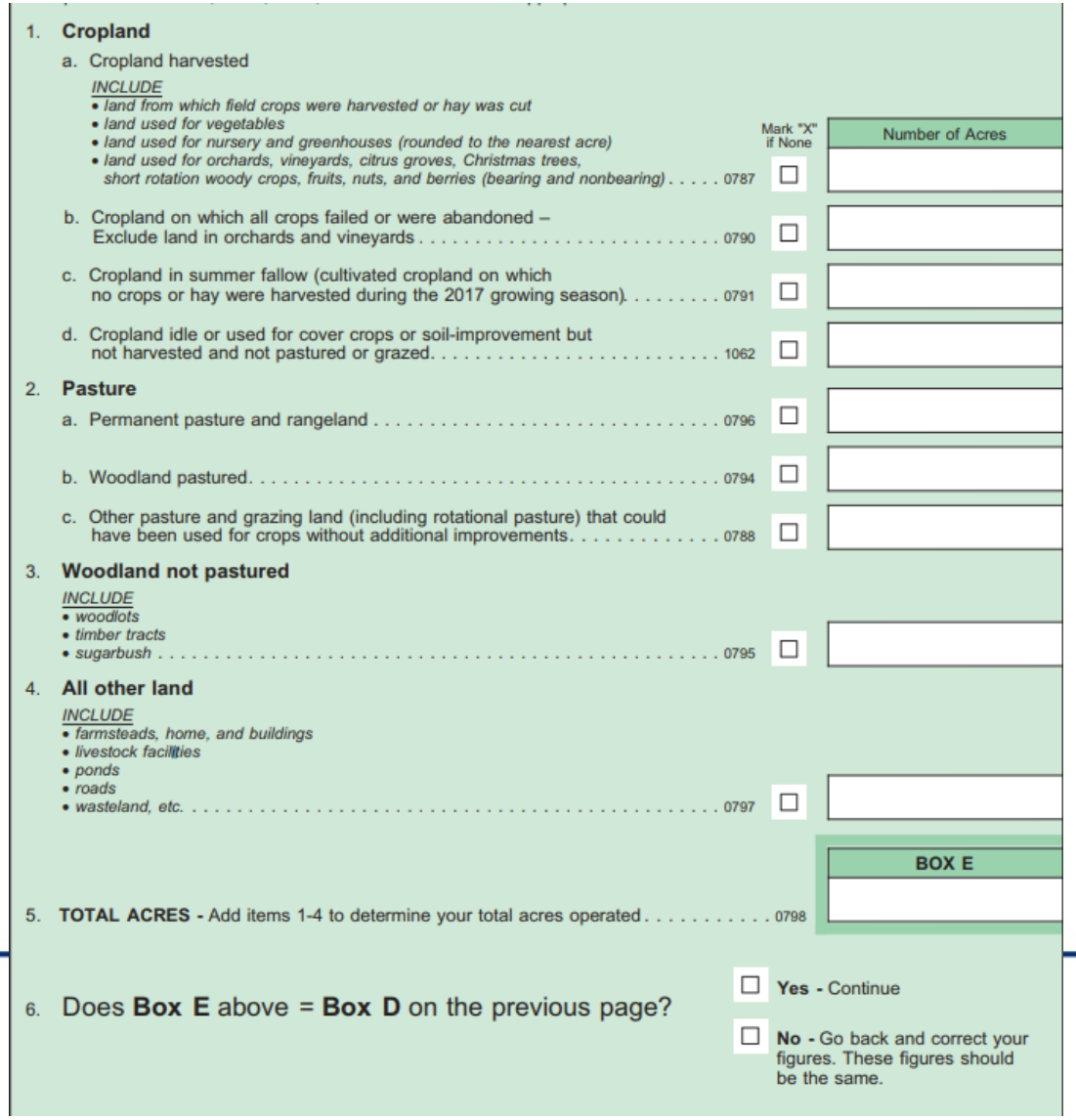


Identifying List Frame Deadwood

- For new survey samples, classify sample units as potential deadwood
- Use this model to conduct further resource intensive efforts
  - Field staff sent to verify actual status of these records
  - Some deadwood, but not all
  - Status resolved at much higher rate when targeted for verification

Identifying records with measurement errors

- Records with known errors can be classified
- For COA respondents, known errors included acres by type not equal to total acres
  - Predictors include other operation characteristics
  - Classification trees identified operations most likely to have errors



1. **Cropland**

   a. Cropland harvested

   INCLUDE
   - *land from which field crops were harvested or hay was cut*
   - *land used for vegetables*
   - *land used for nursery and greenhouses (rounded to the nearest acre)*
   - *land used for orchards, vineyards, citrus groves, Christmas trees, short rotation woody crops, fruits, nuts, and berries (bearing and nonbearing)* 0787

   Mark "X" if None | Number of Acres

   b. Cropland on which all crops failed or were abandoned – Exclude land in orchards and vineyards 0790

   c. Cropland in summer fallow (cultivated cropland on which no crops or hay were harvested during the 2017 growing season) 0791

   d. Cropland idle or used for cover crops or soil-improvement but not harvested and not pastured or grazed 1062

2. **Pasture**

   a. Permanent pasture and rangeland 0796

   b. Woodland pastured 0794

   c. Other pasture and grazing land (including rotational pasture) that could have been used for crops without additional improvements 0788

3. **Woodland not pastured**

   INCLUDE
   - *woodlots*
   - *timber tracts*
   - *sugarbush* 0795

4. **All other land**

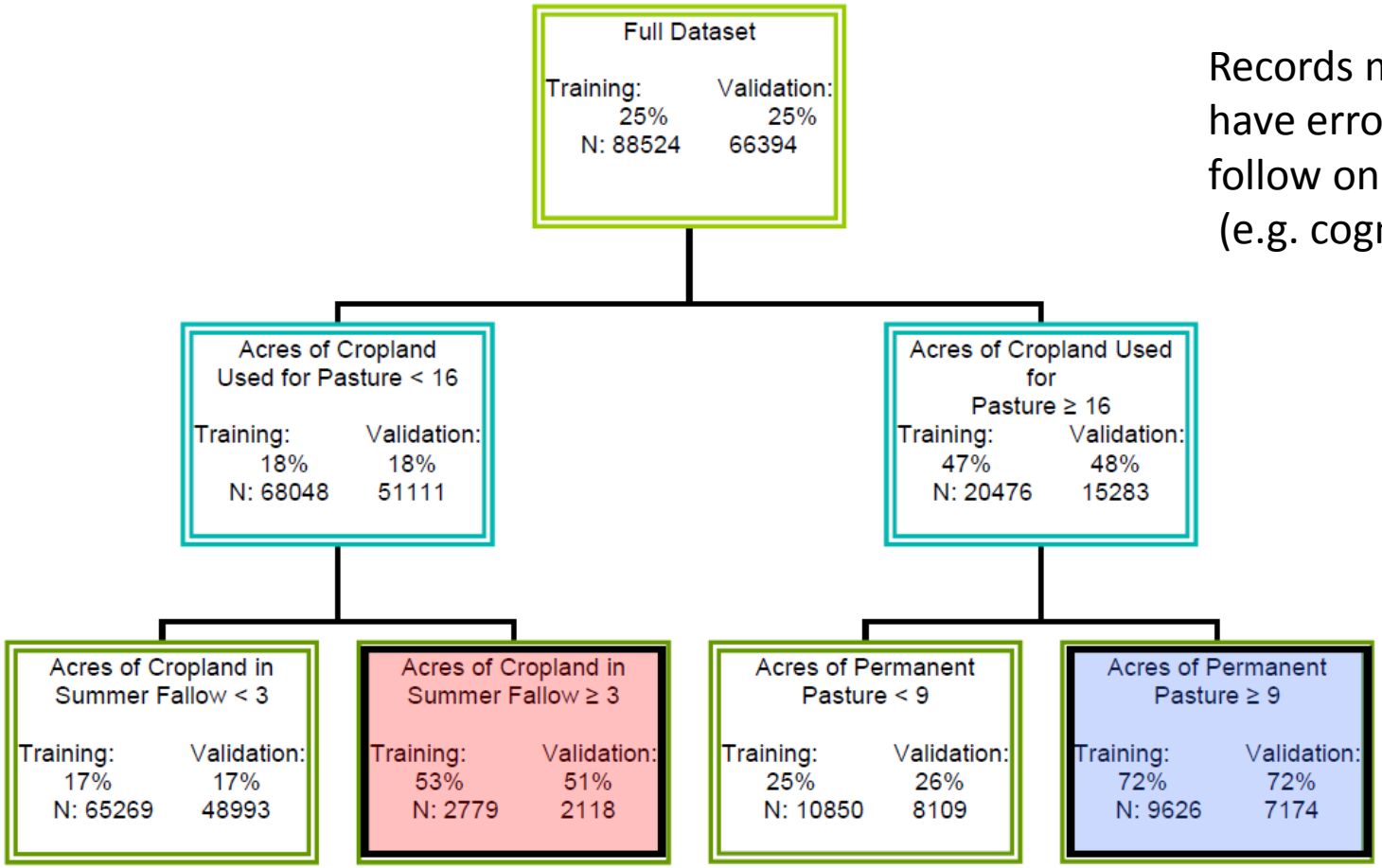
   INCLUDE
   - *farmsteads, home, and buildings*
   - *livestock facilities*
   - *ponds*
   - *roads*
   - *wasteland, etc.* 0797

   BOX E

5. **TOTAL ACRES** - Add items 1-4 to determine your total acres operated 0798

6. Does **Box E** above = **Box D** on the previous page?
   - **Yes** - Continue
   - **No** - Go back and correct your figures. These figures should be the same.

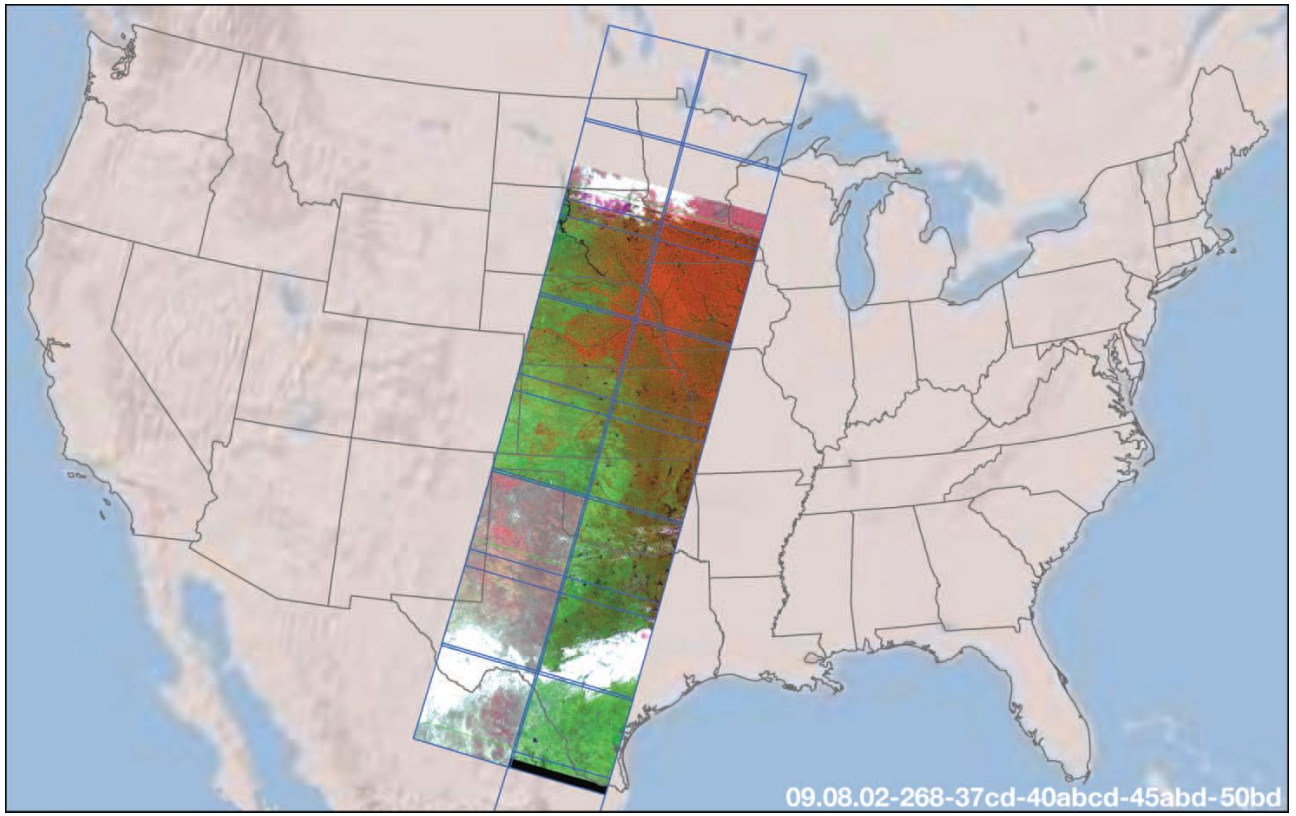
Classification tree example for errors in *Total Acres Operated*



Records most likely to have errors targeted in follow on activities (e.g. cognitive interviews)

Classifying satellite imagery

- NASS generates a GIS Cropland Data Layer (CDL) of crop acreage estimates similar to survey based estimates (Boryan, Yang, Mueller and Craig, 2011)
- Classification trees are used to classify map image pixels as specific crop types
- Targets are type of crop grown in known locations provided by farmers
- Inputs are satellite imagery data



- Models can then be used to classify all land in the US
- Crop acreage estimates can be calculated from the resulting CDL

Where else can we use classification trees?

- Trees are a powerful tool with distinct advantages over other models
  - No hypothesis needed!
  - Can examine high numbers of variables
  - Missing data does not need to be imputed and may be informative
  - Relationships do not have to be linear and higher order interactions are easily handled
  - Can identify small but important groups
  - Can be easily interpreted

"I think that I shall never see, A poem lovely as a tree….."
-- **Joyce Kilmer**

“This is your bravery test. You worked so hard and then a crazy-haired guy tells you to throw in a big ol' tree on top of it all.”