Cleaning Out the Gutter:

Identifying and Eliminating Deadwood from a Sampling Frame Using Trees

March 2018

Andrew J. Dau
Gavin R. Corral, Jodie M. Sprague, Linda J. Young
United States Department of Agriculture
National Agricultural Statistics Service
USDA/NASS

USDA NASS

- Over 400 reports annually
  - Census of Agriculture every 5 years
- Reports driven by surveys
- Surveys driven by sampling frames
  - List frame

Maintaining the Sampling Frame

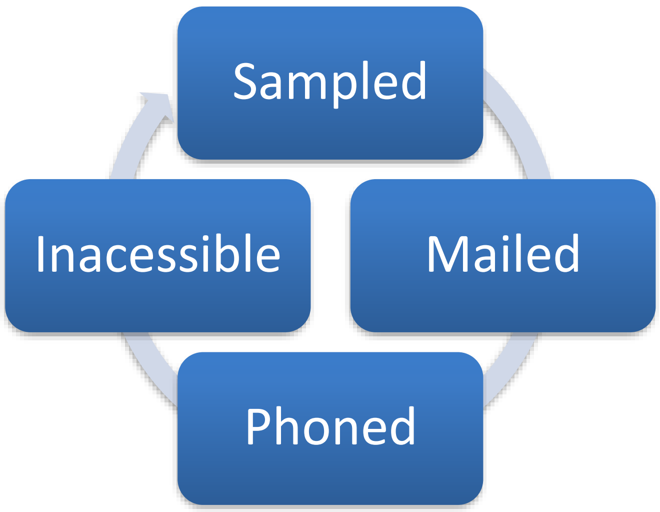
- Processes for adding to frame are on-going.
- Frames age/deteriorate over time.
- Aging records create deadwood.
  - Records that are in business on the frame, but in reality are out of business

Bowling…and “Deadwood”

Source: www.ncaa.com

What's the Problem With Deadwood?

- Impacts on estimates.



- Higher inaccessible rate/ lower overall response rate.
- Can remain on sampling frame for long time.
- Costs → Inflated Samples

How to Identify Deadwood?

- Not easy to predict.
- Despite best efforts, never 100% accurate.
- Can we build a predictive model?
  - 70+ of covariates available

Goal

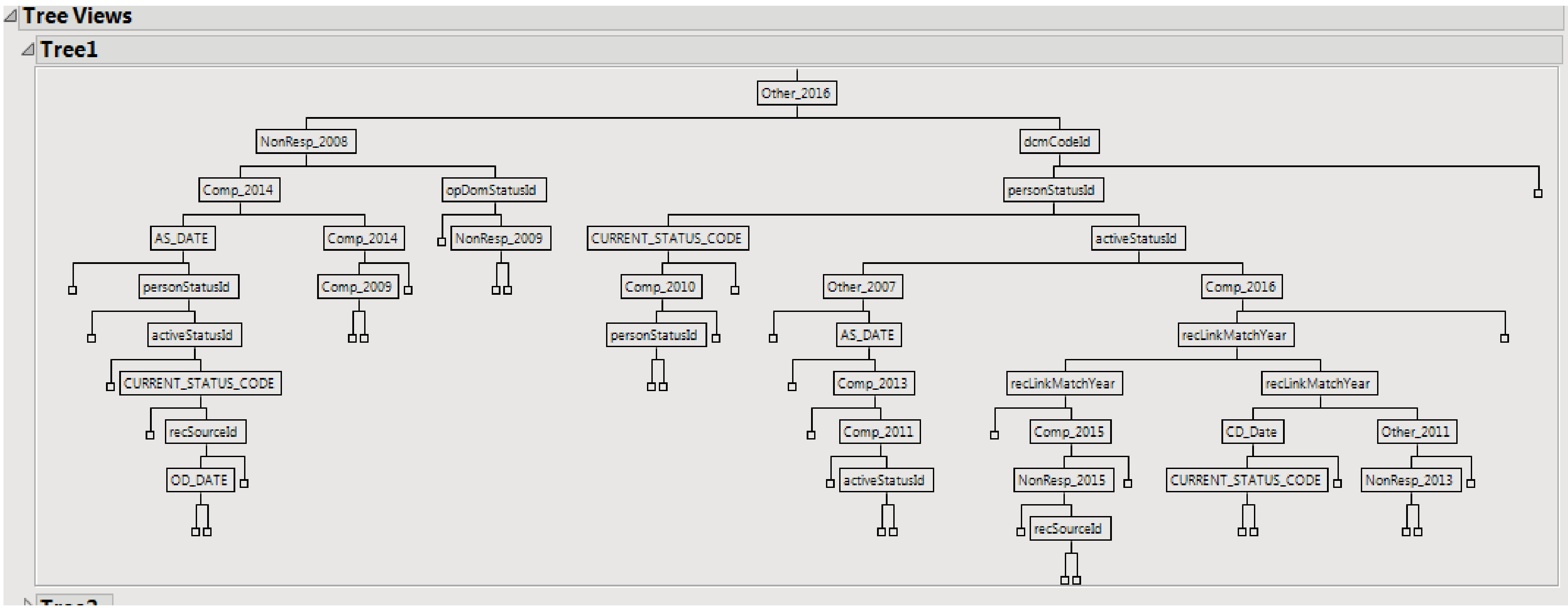
- Build a predictive model which can aid in identifying deadwood thereby maintaining an up-to-date list frame.

Classification and Regression Trees

- "Classification and regression trees are machine-learning methods for constructing prediction models from data." (Loh,2011)

- Boosted Trees - SAS JMP

The Model…An Example



Model Development

- Previous Survey Data
  - What kinds of operations were in-business?
  - What kinds of operations were out-of-business? (deadwood)

- Create binary indicator

- Model Comparison → R^2, ROC, & Confusion Matrix

What's in Our Model?

- Most recent administrative linkage
- Most recent sampling frame data update
- Death Index
- Previous Response History
- Age
- Location
- Ag Census Response

Model Output

- The model creates propensity scores, indicating the likelihood of a record being deadwood.

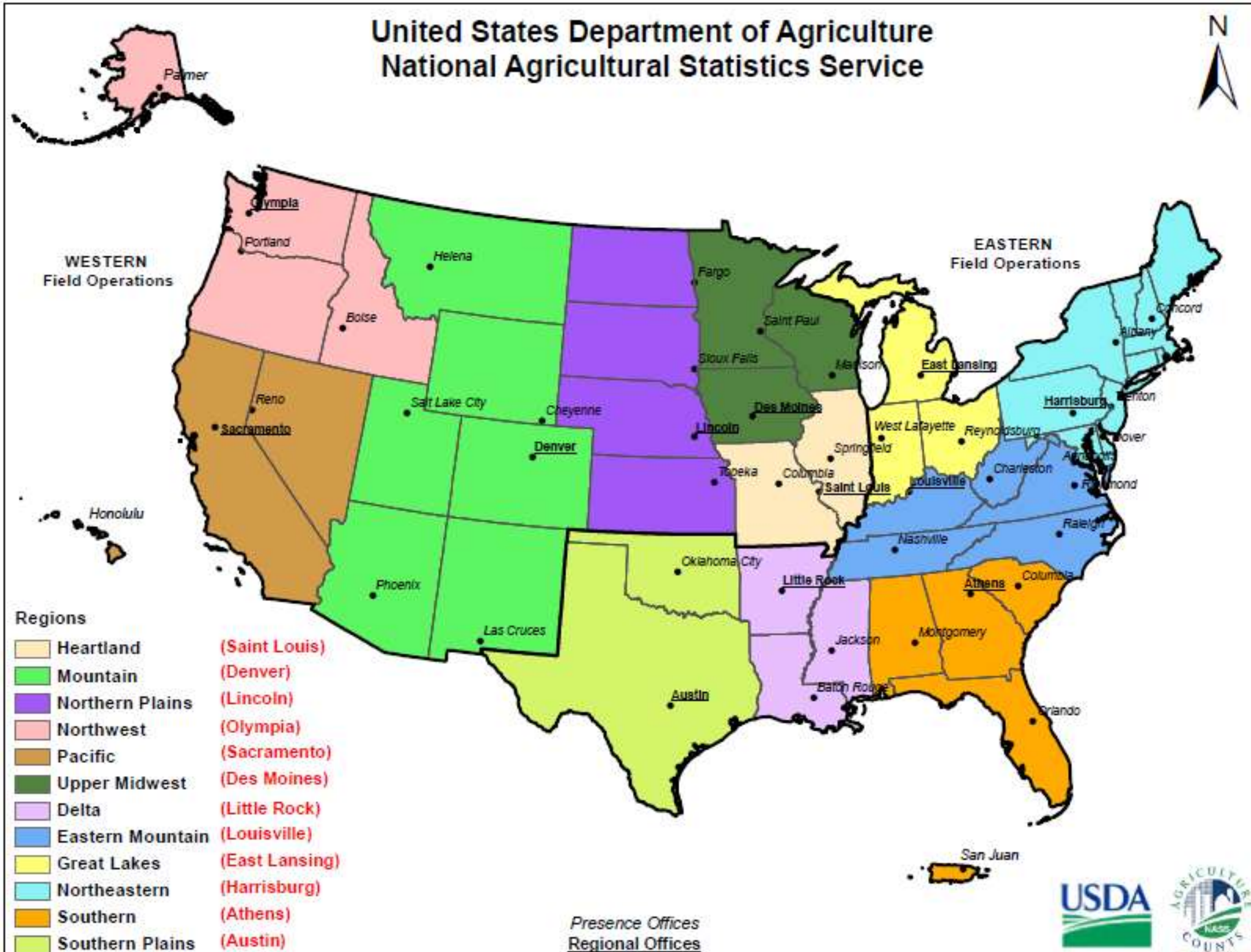
Prob(deadwood= = 1)	Most Likely deadwood
0.0018551978	0
0.0060186538	0
0.9177965625	1
0.00984204	0
0.0114227775	0
0.0018398113	0

The Process

1. Predict likelihood of deadwood for each record in a survey sample.

2. Request face-to-face enumeration during survey process.

3. Verify operating status, complete survey.


United States Department of Agriculture
National Agricultural Statistics Service
N
WESTERN
Field Operations
EASTERN
Field Operations
Palmer
Olympia
Portland
Helena
Boise
Reno
Sacramento
Salt Lake City
Cheyenne
Denver
Honolulu
Phoenix
Las Cruces
Fargo
Saint Paul
Sioux Falls
Madison
Des Moines
Lincoln
Topeka
Columbia
Saint Louis
Springfield
East Lansing
West Lafayette
Reynoldsburg
Louisville
Charleston
Harrisburg
Concord
Albany
Dover
Richmond
Raleigh
Nashville
Oklahoma City
Little Rock
Jackson
Baton Rouge
Austin
Montgomery
Athens
Columbia
Orlando
San Juan
Regions
Heartland (Saint Louis)
Mountain (Denver)
Northern Plains (Lincoln)
Northwest (Olympia)
Pacific (Sacramento)
Upper Midwest (Des Moines)
Delta (Little Rock)
Eastern Mountain (Louisville)
Great Lakes (East Lansing)
Northeastern (Harrisburg)
Southern (Athens)
Southern Plains (Austin)
Presence Offices
Regional Offices
USDA

September – Acreage, Production, and Stocks Survey (APS)

348 Potential Deadwood Records Identified

4 Regions, Boots on Ground

76 Records

8 Regions, No indication of Deadwood

272 Records

September APS Results

Region	Records	Inaccessible	Deadwood
Targeted 4 Regions	76	21%**	29%**
Non-Targeted 8 Regions	272	39%**	2%**

*Proportions significantly different at .01 level

Are a lot of the inaccessible records in the non-targeted 8 regions actually deadwood?

Small Grain County Estimates Survey (Crops CE)

1098 Potential Deadwood Records Identified

4 Regions, Boots on Ground

356 Records

8 Regions, No indication of Deadwood

742 Records

Small Grain CE Results

Region	Records	Inaccessible	Deadwood
Targeted 4 Regions	356	20%**	38%**
Non-Targeted 8 Regions	742	39%**	18%**

*Proportions significantly different at .01 level

Once again, are a lot of the inaccessible records in the non-targeted 8 regions actually deadwood?

September Recap

- Targeted regions had higher out-of-business (deadwood) rates and lower inaccessible rates.
- All indications point towards expanding the boots on the ground data collection to all 12 regions.

Additional Results

Survey	Year	Deadwood Removed	Deadwood ID'd	Deadwood (%)	Inaccessible (%)
15 Surveys	2016-2018	3,442	8,779	39.21%	25.28%

Conclusion and Future Steps

- The model is accurately identifying a high rate of deadwood records.
- Continue process of identifying potential deadwood at a survey level.
- Approved Decision Memorandum – Jan 24, 2018

Acknowledgements

Response Rate Research Team and Deadwood Sub-team

– *Dan Boostrom*
– *Gavin Corral*
– *Cheryl Ito*
– *Troy Marshall*
– *Barbara Rater*
– *Jodie Sprague*
– *Robyn Sirkis*
– *Gerald Tillman*
– *Linda Young*

References

- Loh, Wei-Yin. "Classification and Regression Trees." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1 (2011): 14-23. Web.
- *JMP: User Guide*. Cary, North Carolina.--: SAS Institute, 2005. Print.
- Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. New York, NY: Springer, 2016.
- Corral, G. & Dau, A. (2017). *Identifying Out of Business Records on the NASS List Frame Using Boosted Regression Trees*. In JSM Proceedings.