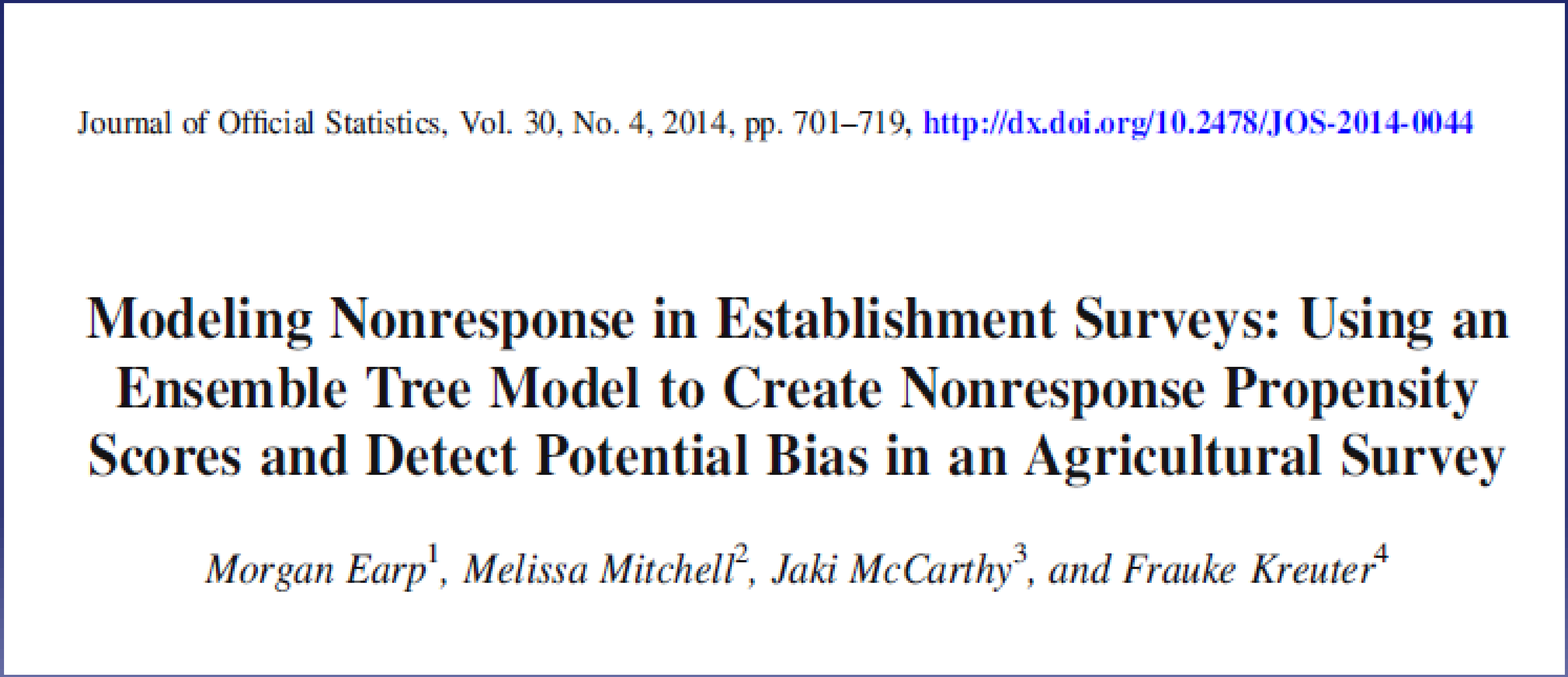


Journal of Official Statistics, Vol. 30, No. 4, 2014, pp. 701–719, http://dx.doi.org/10.2478/JOS-2014-0044

Modeling Nonresponse in Establishment Surveys: Using an Ensemble Tree Model to Create Nonresponse Propensity Scores and Detect Potential Bias in an Agricultural Survey

Morgan Earp[1], *Melissa Mitchell*[2], *Jaki McCarthy*[3], *and Frauke Kreuter*[4]

Purpose

- Developed classification trees to identify hardcore nonrespondents
- Assessed relationship between classification tree nonresponse propensity and actual nonresponse
- Created 10 classes based on classification tree nonresponse propensities to assess and compare nonresponse bias

Motivation

- Attempt to reduce nonresponse bias, by identifying and targeting influential nonrespondents prior to survey administration

ARMS Nonresponse Rates

Forecast
Outline
Motivation
Problem Statement
Purpose
Related Work
Methods
Results
Summary
Future Work

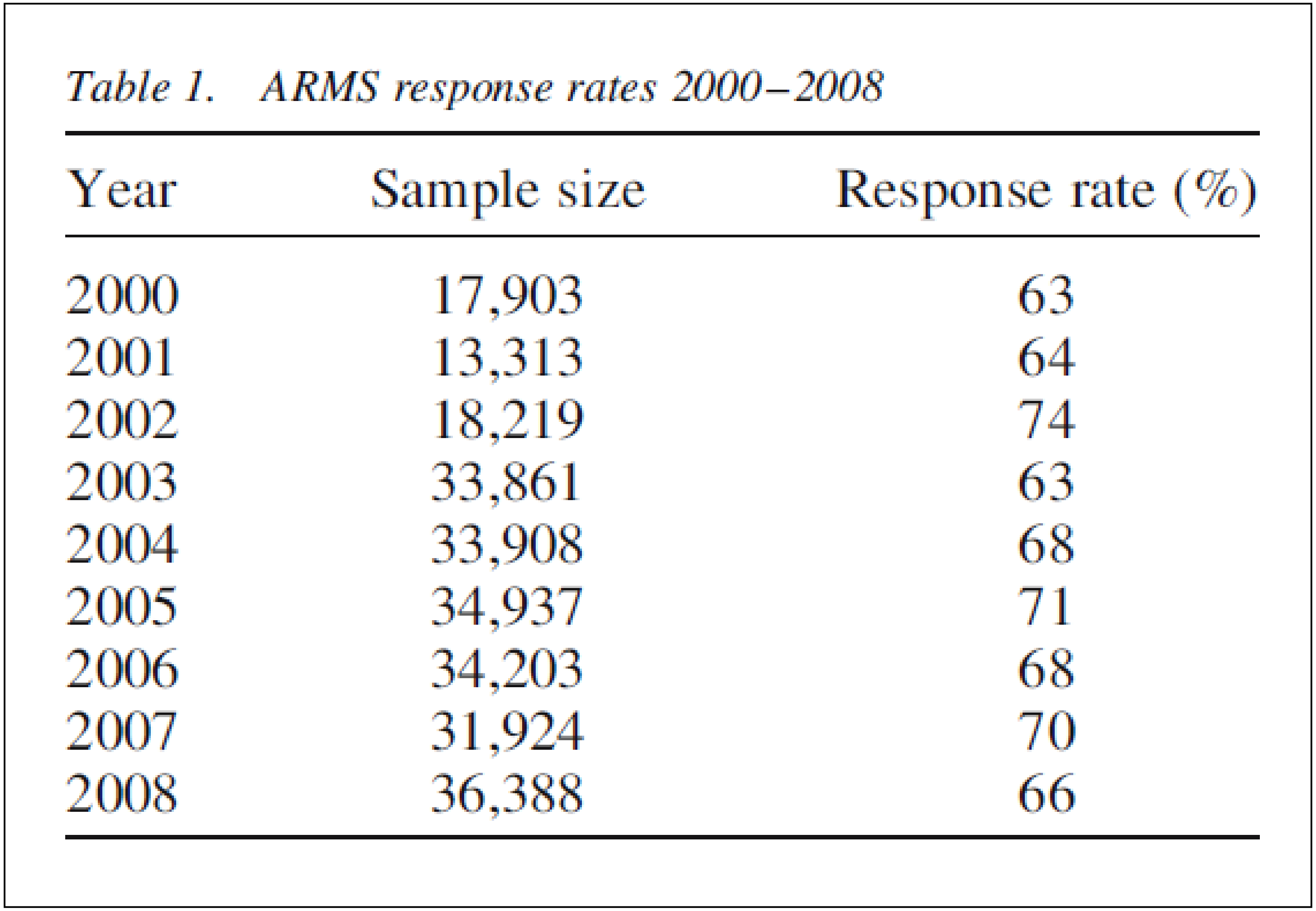


Table 1. ARMS response rates 2000–2008

Year	Sample size	Response rate (%)
2000	17,903	63
2001	13,313	64
2002	18,219	74
2003	33,861	63
2004	33,908	68
2005	34,937	71
2006	34,203	68
2007	31,924	70
2008	36,388	66

(p. 703)

Methods

- Used an ensemble of classification trees to identify likely nonrespondents

- Used nonresponse propensity deciles to classify nonrespondents and assessed bias using the relative difference of the mean

Forecast
Outline
Motivation
Problem Statement
Purpose
Related Work
Methods
Results
Summary
Future Work

BLS

Classification Trees

- A “data mining” technique which segments a dataset using a series of simple rules to maximize dichotomies
- Creates subsets of records exhibiting a higher percentage of the “target”(respondent or nonrespondent)

Forecast
Outline
Motivation
Problem Statement
Purpose
Related Work
Methods
Results
Summary
Future Work

Splitting Criteria

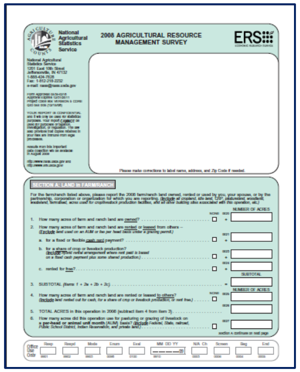
Forecast
Outline
Motivation
Problem Statement
Purpose
Related Work
Methods
Results
Summary
Future Work

- Optimal Splitting Criteria
  - Significance Testing
    - Uses the *p* value as the stopping rule after applying a Bonferroni adjustment to mitigate bias toward variables w/ many values
      - Interval (*F* test)
      - Nominal (Chi-Square)
  - Variance Reduction
    - Measures the reduction in entropy, after adjusting for ordinal differences
      - Ordinal (Entropy)

Classification Tree Proxy Data

- Imported Census of Agriculture (COA) response history for the ARMS III 2000-2008 Samples (n = 254,632)

- Imported and matched 2002 COA data to be used as proxies of these operations characteristics
  - 78% match rate for 2002



National Agricultural Statistics Service

2008 AGRICULTURAL RESOURCE MANAGEMENT SURVEY

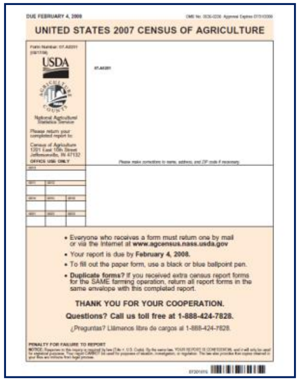
ERS

National Agricultural Statistics Service
1201 East 10th Street
Jeffersonville, IN 47132
1-888-424-7828
Fax: 1-812-218-2232

Please make corrections to label name, address, and Zip Code if needed.

SECTION A: LAND in FARM/RANCH

1. How many acres of farm and ranch land are owned?
2. How many acres of farm and ranch land are rented or leased from others –
   a. for a fixed or flexible cash rent payment?
   b. for a share of crop or livestock production?
   c. rented for free?
3. SUBTOTAL (items 1 + 2a + 2b + 2c)
4. How many acres of farm and ranch land are rented or leased to others?
5. TOTAL ACRES in this operation in 2008 (subtract item 4 from item 3)
6. How many acres did this operation use for pasturing or grazing of livestock on a per-head or animal unit month (AUM) basis?



DUE FEBRUARY 4, 2008

UNITED STATES 2007 CENSUS OF AGRICULTURE

USDA

National Agricultural Statistics Service

Please return your completed report to:

Census of Agriculture
1201 East 10th Street
Jeffersonville, IN 47132

OFFICE USE ONLY

Please make corrections to name, address, and ZIP code if necessary.

- Everyone who receives a form must return one by mail or via the Internet at **www.agcensus.nass.usda.gov**
- Your report is due by **February 4, 2008.**
- To fill out the paper form, use a black or blue ballpoint pen.
- **Duplicate forms?** If you received extra census report forms for the SAME farming operation, return all report forms in the same envelope with this completed report.

THANK YOU FOR YOUR COOPERATION.

Questions? Call us toll free at 1-888-424-7828.

¿Preguntas? Llámenos libre de cargos al 1-888-424-7828.

PENALTY FOR FAILURE TO REPORT

Types of Proxy Data

- Proxy data included 70 COA variables significantly related to ARMS nonresponse
  - Operator Demographics
  - Farm Type
  - Size
  - Commodities Raised
  - Expenses
  - Location

BLS

Example Tree

ARMS III Matched Sample (Training Data)

37%
n =79,616

Sum of Poultry Inventory Data

< 4
38%
n= 71,644

≥ 4
30%
n =7,972

Total Value of Products Sold + Government Payments

< $110,005
27%
n= 30,904

≥ $110,005
45%
n =40,740

Total Sales – Not Under Production Contract (NUPC)

< $844,879
41%
n= 31,211

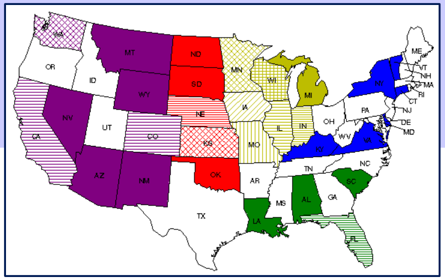
≥ $844,879
57%
n =9,529

States

AL, AZ, CA, CO, CT, DE, FL, IL, IN, IA, KS, KY, LA, MI, MN, MO, MT, NE, NV, NM, NY, ND, OK, SC, SD, VT, WA, & WY

Yes
70%
n=3,346

No
52%
n =1,118



BLS

Analyses

- Assessed the relationship between classification propensity scores and nonresponse rates using logistic regression
- Assessed the relationship between classification propensity scores and nonresponse bias by plotting the relative bias of the mean by classification propensity score decile

Variables

- Inputs
  - Classification Tree Propensity Score
    - ARMS 2000-2008 nonresponse
    - Census 2002 operation characteristics
- Controls
  - Total Sales & Total Acres Operated
    - Census 2007
- Target
  - ARMS 2009 Nonresponse

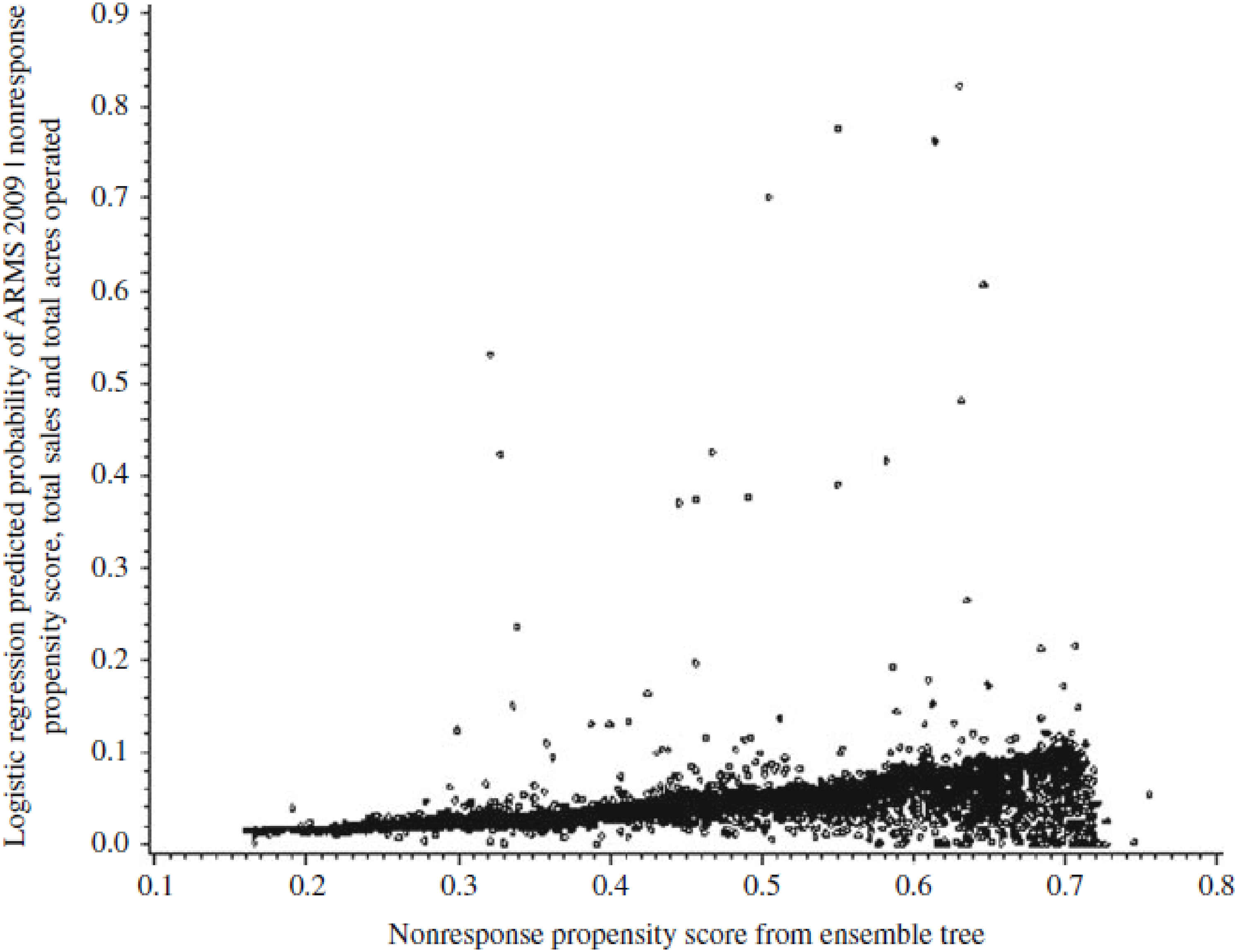


Fig. 1. Plot of the logistic regression predicted probability of 2009 ARMS nonresponse given the ensemble tree nonresponse propensity score, 2007 total sales, and 2007 total acres operated, by the ensemble tree nonresponse propensity score

Logistic Regression Results

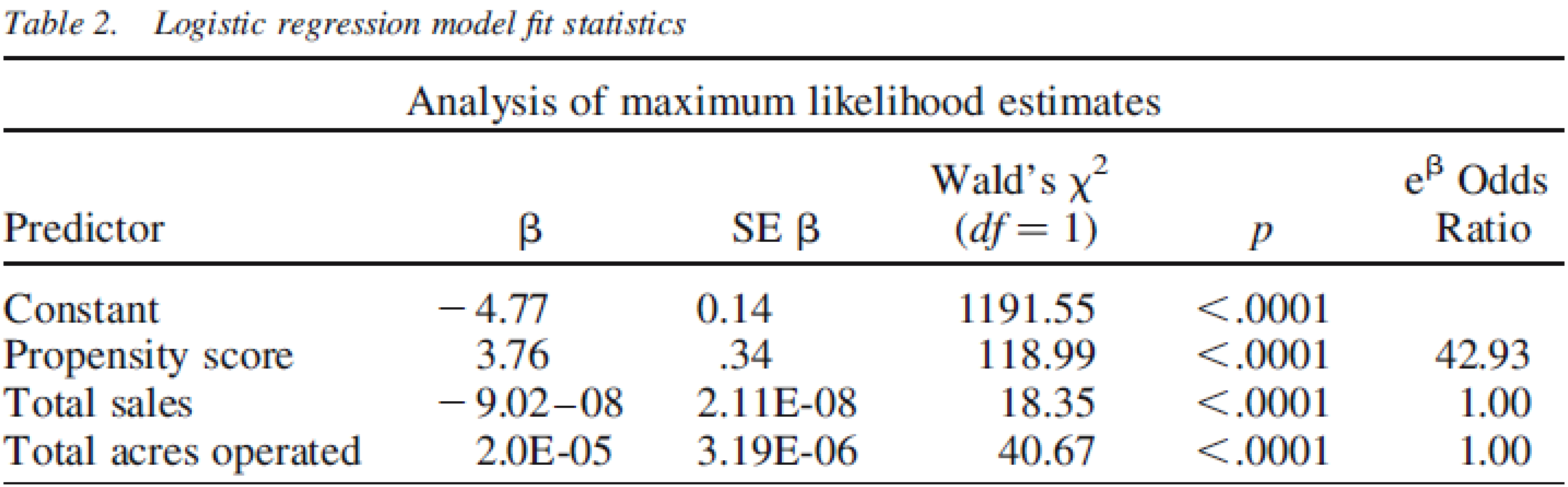


Table 2. Logistic regression model fit statistics

Analysis of maximum likelihood estimates					
Predictor	β	SE β	Wald's χ^2 ($df = 1$)	p	e^{β} Odds Ratio
Constant	− 4.77	0.14	1191.55	<.0001	
Propensity score	3.76	.34	118.99	<.0001	42.93
Total sales	− 9.02−08	2.11E-08	18.35	<.0001	1.00
Total acres operated	2.0E-05	3.19E-06	40.67	<.0001	1.00

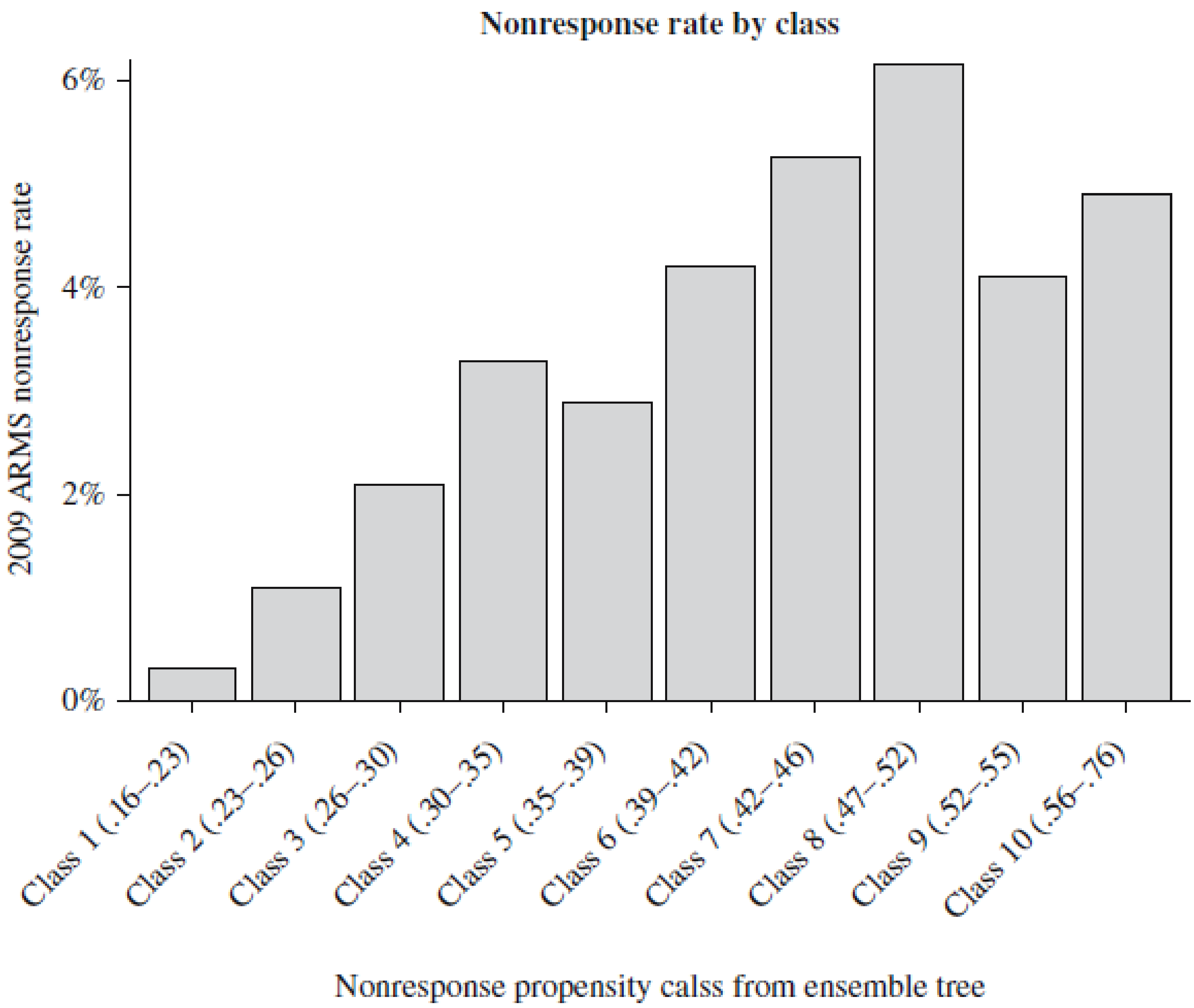
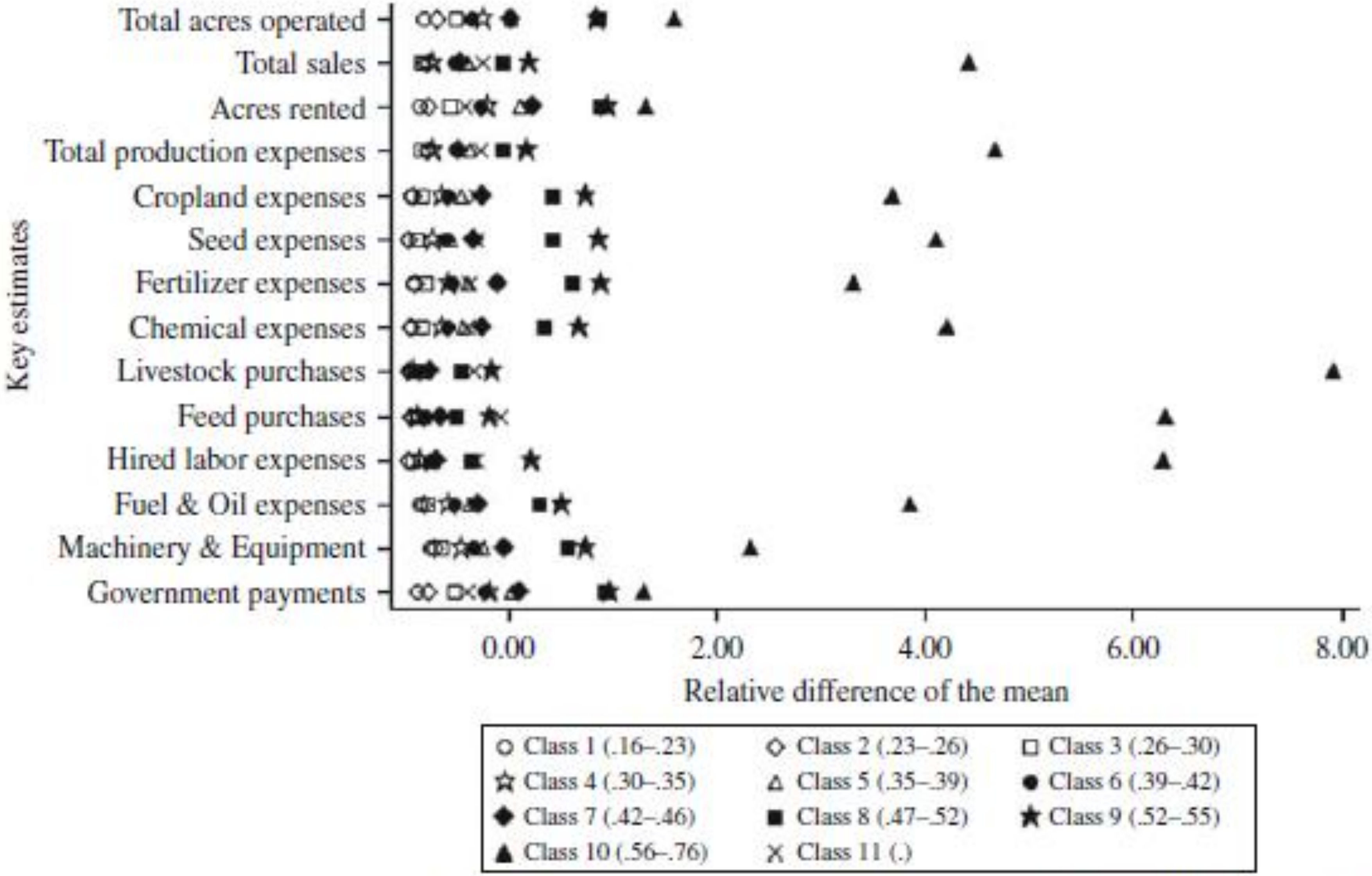


Fig. 2. *ARMS 2009 nonresponse rate by ensemble tree nonresponse propensity class*



Relative difference of the mean = [(class mean – overall mean)/overall mean]

Fig. 3. Relative difference of the mean for key estimates by nonresponse propensity class

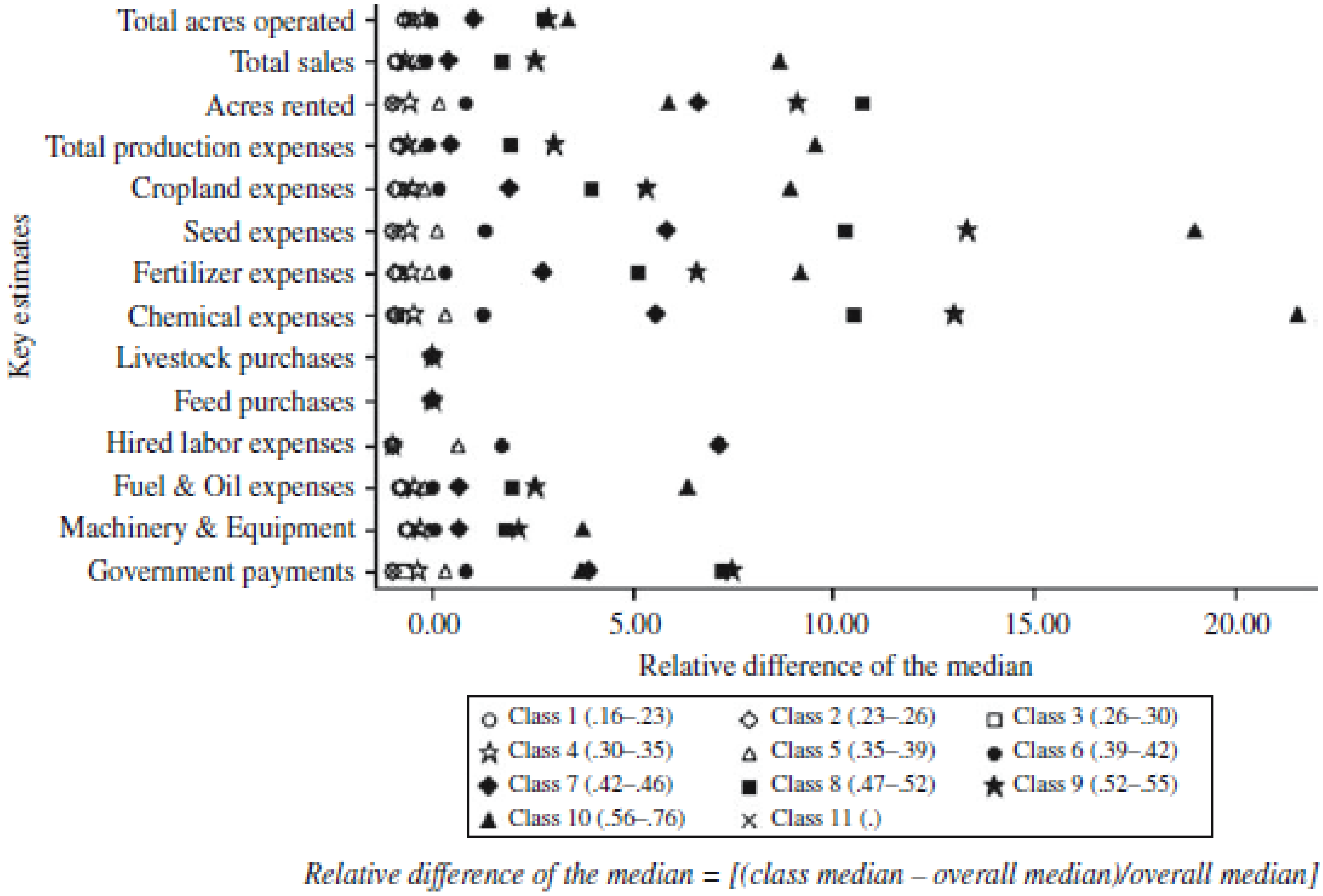


Fig. 4. Relative difference of the median for key estimates by nonresponse propensity class

Conclusion

- Easily identify characteristics associate w/ nonresponse
- Can ensure that each variable is considered once in the overall average model
- These propensity scores were positively correlated with the amount of potential bias across several key estimates

Conclusion

- We would like to compare this tree method w/ random forests
- They are currently being used to pre-score samples prior to data collection to ensure that those farms that are least likely to respond and most likely to bias estimates as a result receive special attention.

Contact Information

Contact Information

Morgan Earp

Earp.Morgan@bls.gov