Some Comments On Bayesian Methods for Surveys

Nathan B. Cruze

National Agricultural Statistics Service (NASS)
United States Department of Agriculture
nathan.cruze@nass.usda.gov

FCSM 2018
Washington, D.C.
March 9, 2018

Bayesian Methods and the Survey Process

Motivated by household AND establishment surveys

- Williams: 'Bayesian Estimation Under Informative Sampling with Unattenuated Dependence'
- Savitsky: 'Scalable Bayes Clustering for Outlier Detection Under Informative Sampling'
- Benecha: 'Investigating Covariate Selection for a Bayesian Crop Yield Forecasting Model'
- Gregory: 'Are We Underestimating Food Insecurity?'

Four talks emphasizing

- design, weighting, and analysis (Williams, Savitsky)
- model-based estimation (Benecha, Gregory)

Common Themes

1. Informative sampling and the pseudo-posterior distribution
2. Shrinkage
   - Outlier nomination: posterior distribution contracts on a point
   - Yield as weighted average of survey and Stage 2 model inputs
3. 'Known' unknowns
   - Effects of included or omitted covariates in retrospect
   - Partial identifiability of underreporting of food insecurity
4. Markov chain Monte Carlo computation
   - Gibbs samplers
   - Approximate maximum a posteriori (MAP) partitioning with `growclusters`
   - NUTS Hamiltonian Monte Carlo algorithm in STAN

Informative Sampling

A sampling design is informative whenever distribution in sample is different than distribution in population

- Unit inclusion probabilities correlated with response
- Analytic inference requires special attention
- Many informative designs
  - Single stage, fixed-size, stratified design of Current Employment Statistics (CES)
  - Five-stage, geographically-indexed sampling design of National Survey on Drug Use and Health (NSDUH)
  - Clustering, sorting, etc.

Question: What about joint inclusion probabilities and higher order dependencies?

The Pseudo-Posterior Distribution

As in Savitsky and Toth (2016)

$$\hat{\pi}(\lambda|\boldsymbol{y_o}, \tilde{\boldsymbol{w}}) \propto \underbrace{\left[\prod_{i=1}^{n} p(y_{o,i}|\lambda)^{\tilde{w}_i}\right]}_{pseudo-likelihood} \pi(\lambda)$$

- Likelihood is from sample–goal is inference about parameters of a (super)population
- Note the role of informative sampling design weight
- Act on pseudo-likelihood in Bayesian manner

Bayesian Estimation Under Informative Sampling

Consistent estimation of population parameters

- Theoretical arguments for adjustments based on first order inclusion probabilities
  - Asymptotic factorization
  - Asymptotic independence
- Unattenuated dependence is feature of many survey designs
- Why do first order inclusion probabilities still seem to work?
- **A Bayesian interpretation or justification for broad class of analysis models already applied in practice**

Scalable Outlier Detection Under Informative Sampling

Bayesian motivation recast as optimization problem–**scalability!**

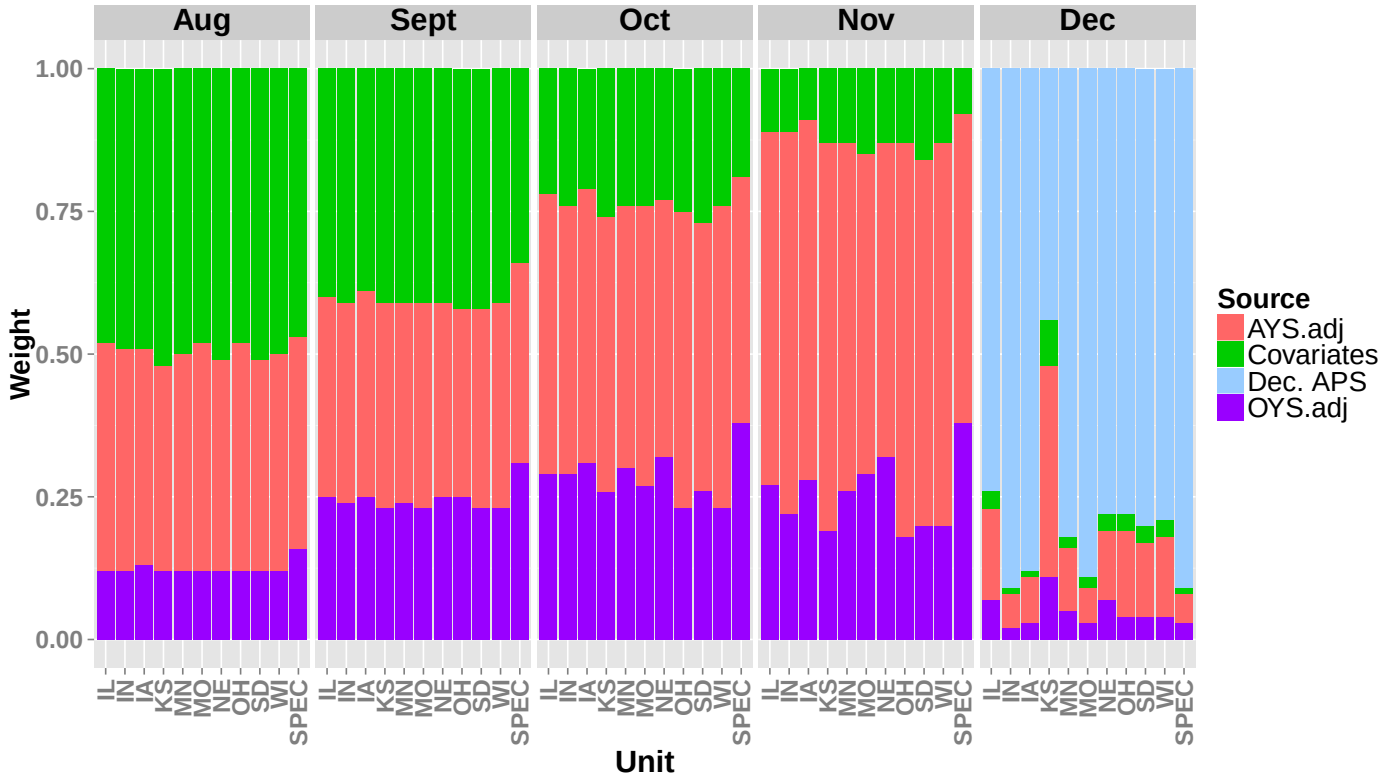
$$\underset{K,s,M}{argmin} \sum_{p=1}^{K} \sum_{i:s_i=p} \tilde{w}_i ||\boldsymbol{x_i} - \boldsymbol{\mu}_p||^2 + K\lambda$$

- Informative design: a MAP partition and cluster centers with respect to *population* rather than the sample
- λ prevents case of each establishment as own cluster–algorithm will terminate with interpretable clusters
- **Nominated outliers are relatively few establishments remaining in 'small' clusters**

Crop Yield Forecasting

Question: Where do you get your priors?

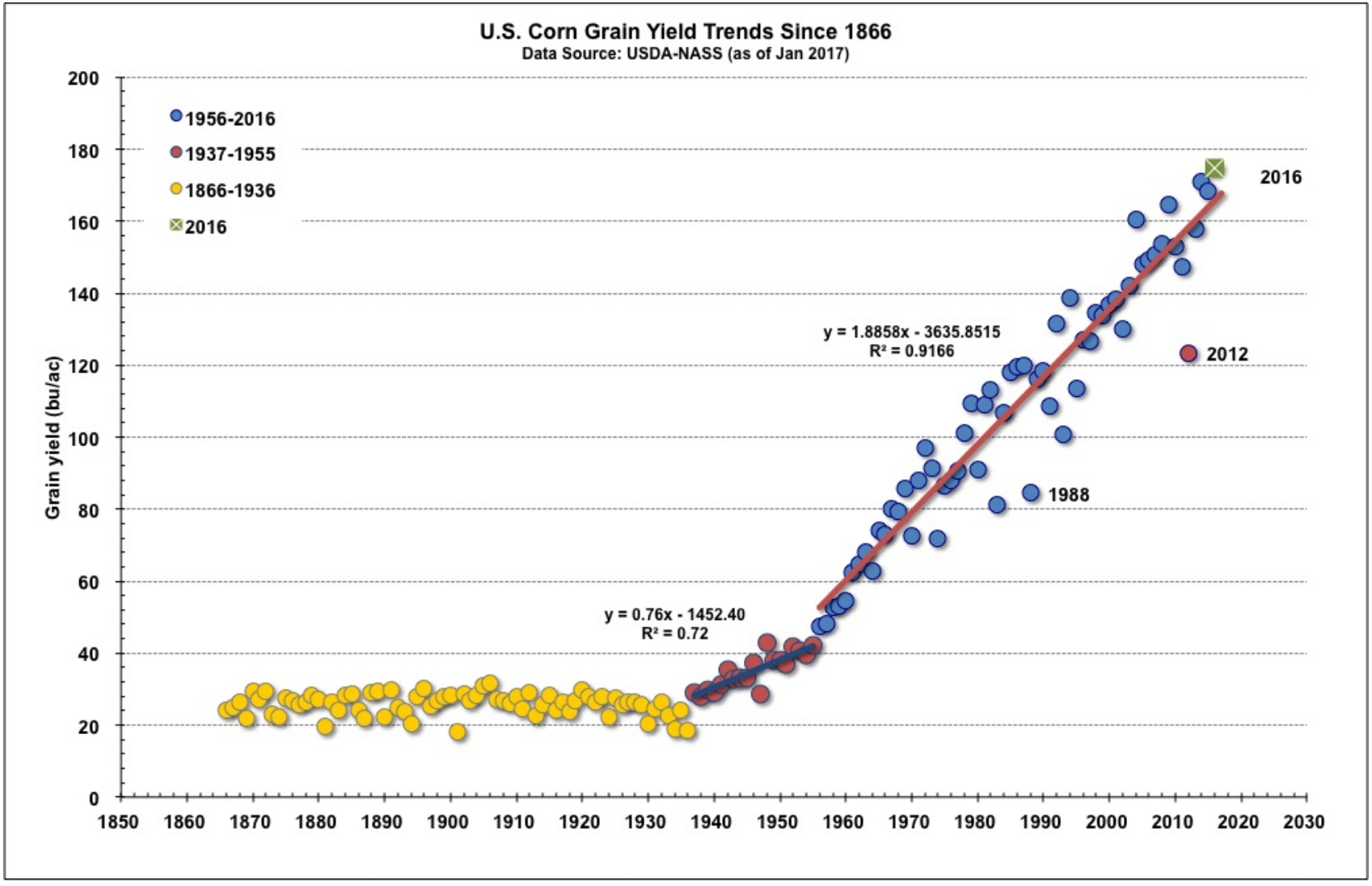
- Non-informative, conjugate family priors
- Mean of posterior distribution of μ_t as weighted average



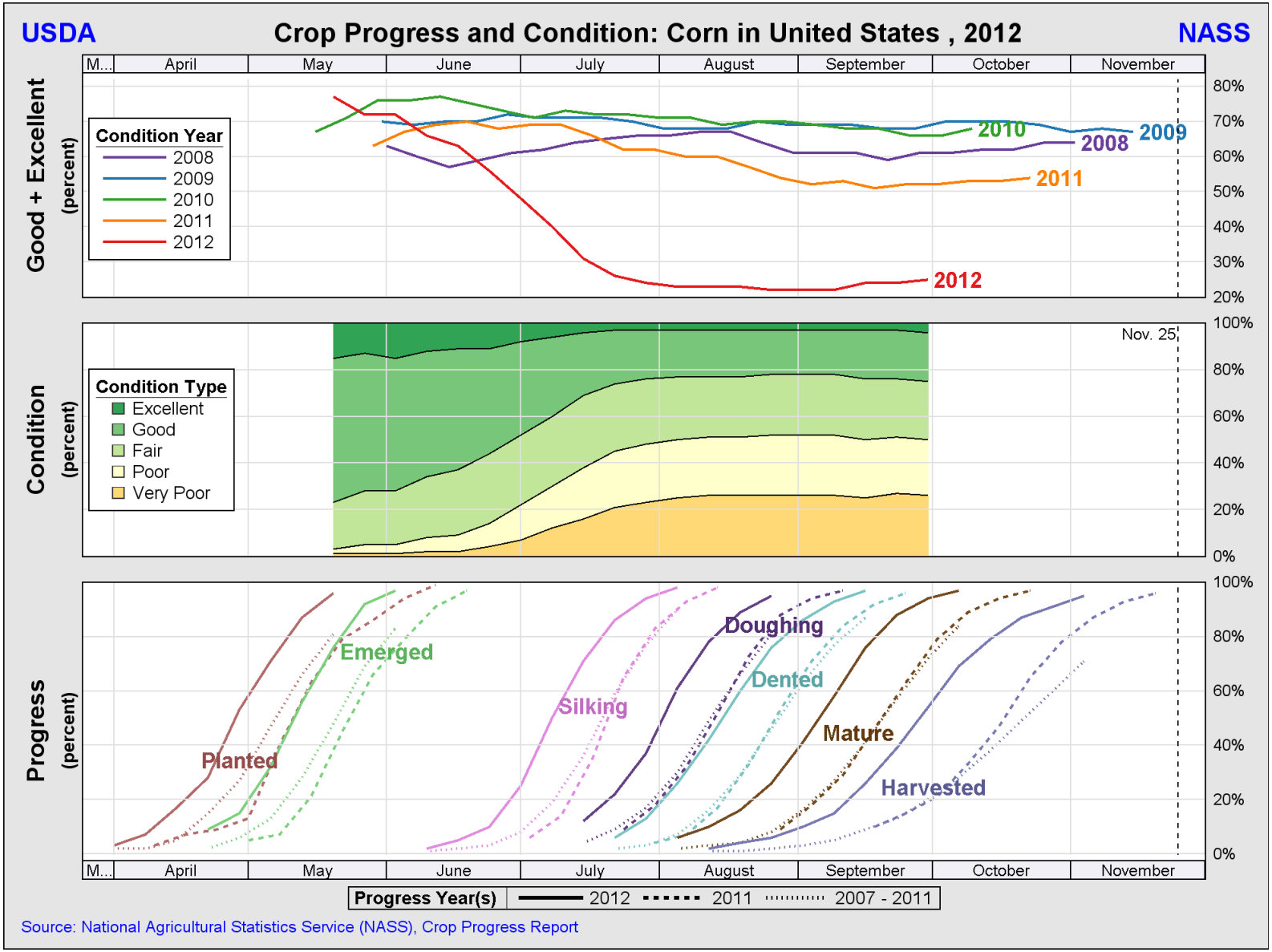
Question: Where do you get your Stage 2 model?

Crop Yield Forecasting

- Origins of corn model found in Wang et al. (2012); Nandram, Berg, and Barboza (2014); Adrian (2012)
- 'Current' model inputs are similar to other USDA research (Wescott and Jewison 2013)
  - Weather adjusted trends
  - In hindsight, severity of 2012 drought
- See Irwin and Good (2016) for outsider's perspective on NASS yield forecasting



Source: R.L. Nielsen (Purdue) https://www.agry.purdue.edu/ext/corn/news/timeless/YieldTrends.html



Source: USDA NASS

https://www.nass.usda.gov/Charts_and_Maps/Crop_Progress_&_Condition/2012/US_2012.pdf

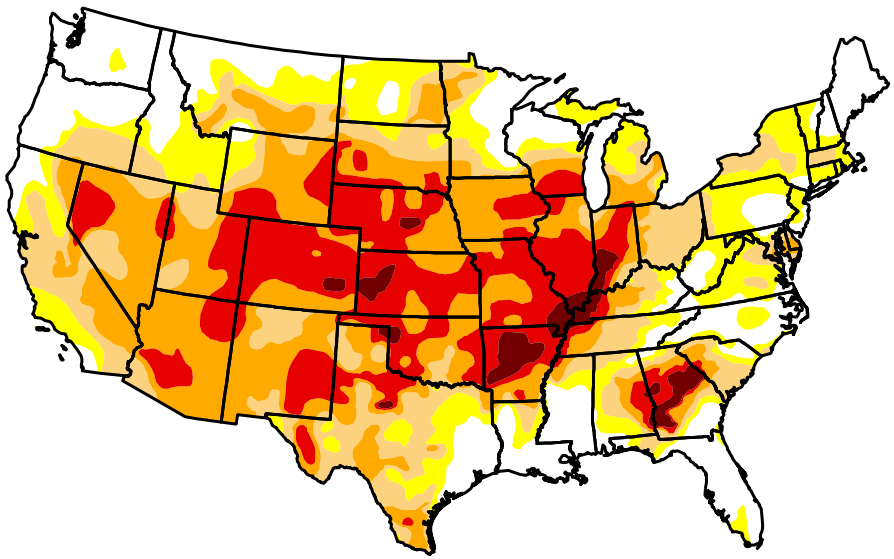
U.S. Drought Monitor

CONUS

July 24, 2012

(Released Thursday, Jul. 26, 2012)

Valid 7 a.m. EST



Drought Conditions (Percent Area)

	None	D0-D4	D1-D4	D2-D4	D3-D4	D4
Current	19.92	80.08	63.86	45.57	20.57	2.38
Last Week *7/17/2012*	19.25	80.75	63.54	42.23	13.53	0.99
3 Months Ago *4/24/2012*	40.58	59.42	37.07	19.95	6.65	1.86
Start of Calendar Year *1/3/2012*	50.41	49.59	31.90	18.83	10.18	3.32
Start of Water Year *9/27/2011*	56.45	43.55	29.13	23.44	17.80	11.37
One Year Ago *7/26/2011*	59.11	40.89	29.50	23.86	18.19	11.04

Intensity:

- D0 Abnormally Dry
- D1 Moderate Drought
- D2 Severe Drought
- D3 Extreme Drought
- D4 Exceptional Drought

The Drought Monitor focuses on broad-scale conditions. Local conditions may vary. See accompanying text summary for forecast statements.

Author(s):
Richard Heim
NCDC/NOAA

http://droughtmonitor.unl.edu/

Source: UNL Drought Monitor
http://droughtmonitor.unl.edu/data/pdf/20120724/20120724_conus_trd.pdf

Are We Underestimating Food Security?

"...identification is not an all-or-nothing concept and...models that do not point identify parameters of interest can, and typically do, contain valuable information about these parameters." (Tamer 2010, p. 168)

1. Estimable contrasts in design of experiments
   - $\widehat{\tau_{trt} - \tau_{ctrl}}$ AND NOT $\hat{\tau}_{trt} - \hat{\tau}_{ctrl}$

2. Set identifiability of β in error-in-variables linear model
   $\beta \in (\beta_{yx}, \frac{1}{\beta_{xy}})$

$$\begin{aligned} y &= \beta x^* + \epsilon \\ x &= x^* + \eta \\ (\epsilon, \eta, x^*)^T &\sim N((0, 0, \mu_{x^*})^T, diag(\sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_{x^*})) \end{aligned}$$

3. Factoring joint posterior in the Bayesian setting
   Define ζ, identified part; η, unidentified part

$$\pi(\zeta, \eta | y) = \pi(\zeta | y)\pi(\eta | \zeta)$$

Are We Underestimating Food Security?

Item Response Theory (IRT) and the four parameter model

$$P(Y_{ij} = 1) = c_j + (d_j - c_j)F(-\alpha_j(\theta_i - \beta_j))$$

- Current Population Survey–Food Security Supplement
- Recast problem as latent variable problem

$$\pi(\alpha, \beta, c, d|data) = \pi(\alpha, \beta|data)\pi(c, d|\alpha, \beta)$$

- Probable values (of c,d) within the feasible set of values
- **A 'menu' of options for discussing presence and degree of misreporting of food insecurity**

References

Adrian (2012) 'A Model-Based Approach to Forecasting Corn and Soybean Yields.' *Proceedings of the Fourth International Conference on Establishment Surveys,* June 11-14, 2012, Montreal, Quebec, Canada: American Statistical Association. http://www.amstat.org/meetings/ices/2012/papers/302190.pdf. [accessed Feb 22, 2018]

Irwin, S., and Good, D. (2016) 'Opening Up the Black Box: More on the USDA Corn Yield Forecasting Methodology.' *Farmdoc Daily* (6):162, Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign, August 26, 2016. http://farmdocdaily.illinois.edu/2016/08/opening-up-the-black-box-more-usda-corn-yield.html [accessed Feb 22, 2018]

Nandram, B., Berg, E., and Barboza, W. (2014) 'A Hierarchical Bayesian Model for Forecasting State-Level Corn Yield.' *Environmental and Ecological Statistics* 21: 507-530.

Savitsky, T.D., and Toth, D. (2016), 'Bayesian Estimation Under Informative Sampling', *Electronic Journal of Statistics* 10(1), 1677-1708.

Tamer, E. (2010), 'Partial Identification in Econometrics', *The Annual Review of Economics* 2, 167-195.

Wang, J.C., Holan, S.H., Nandram, B., Barboza, W., Toto, C., and Anderson, E. (2012) 'A Bayesian Approach to Estimating Agricultural Yield Based on Multiple Repeated Surveys.' *Journal of Agricultural, Biological and Environmental Statistics* 17, 84-106.