Fitting a Bayesian Fay-Herriot Model

Nathan B. Cruze

United States Department of Agriculture
National Agricultural Statistics Service (NASS)
Research and Development Division

Washington, DC
October 25, 2018

Disclaimer

The Findings and Conclusions in This Preliminary Presentation Have Not Been Formally Disseminated by the U.S. Department of Agriculture and Should Not Be Construed to Represent Any Agency Determination or Policy.

Overview

- NASS interest in small area estimation (SAE)
- The Fay and Herriot (1979) model
- Case study: county estimates of planted corn, Illinois 2014
  - Computation in R and JAGS

Small Area Estimation (SAE) Literature

"A domain is regarded as 'small' if the domain-specific sample is not large enough to support [survey] estimates of adequate precision."–Rao and Molina (2015)

Regression and mixed-modeling approaches in SAE literature

- Shrinkage–improve estimates with other information
- Utility of auxiliary data as covariate
- Variance-bias trade off

Two common models

1. Unit-level models, e.g., Battese et al. (1988)
   - USDA NASS (formerly SRS) as source of data/funding
2. Area-level models, e.g., Fay and Herriot (1979)

NASS Interest In SAE

Iwig (1996): USDA's involvement in county estimates in 1917

Published estimates used by:

- Agricultural sector
- Financial institutions
- Research institutions
- Government and USDA

Published estimates used for:

- County loan rates
- Crop insurance
- County-level revenue guarantee

National Academies of Sciences, Engineering, and Medicine (2017)

- Consensus estimates: Board review of survey and other data
- Currently published without measures of uncertainty
- Recommends transition to system of model-based estimates

Fay-Herriot (Area-Level) Model

Fay and Herriot (1979)–improved upon per capita income estimates with following model

$$\hat{\theta}_j = \theta_j + e_j, \quad j = 1, \ldots, m \textit{ counties} \qquad (1)$$

$$\theta_j = \boldsymbol{x}_j' \boldsymbol{\beta} + u_j \qquad (2)$$

Adding Eqs. 1 and 2

$$\hat{\theta}_j = \boldsymbol{x}_j' \boldsymbol{\beta} + u_j + e_j$$

- $\hat{\theta}_j$, direct estimate
- $E(e_j|\theta_j) = 0$
- $V(e_j|\theta_j) = \hat{\sigma}_j^2$, estimated variance
- $\boldsymbol{x}_j$, known covariates
- u_j, area random effect
- $u_j \overset{iid}{\sim} (0, \sigma_u^2)$

Fay-Herriot Formulated As Bayesian Hierarchical Model

'Recipe' for hierarchical Bayesian model as in Cressie and Wikle (2011)

Data model:

$$\hat{\theta}_j | \theta_j, \boldsymbol{\beta} \overset{ind}{\sim} N(\theta_j, \hat{\sigma}_j^2) \tag{3}$$

Process model:

$$\theta_j | \boldsymbol{\beta}, \sigma_u^2 \overset{iid}{\sim} N(\boldsymbol{x}_j' \boldsymbol{\beta}, \sigma_u^2) \tag{4}$$

Prior distributions on $\boldsymbol{\beta}$ and σ_u^2

- Browne and Draper (2006), Gelman (2006): $\sigma_u^2 \sim ?$
- We will specify $\sigma_u^2 \sim Unif(0, 10^8)$, $\boldsymbol{\beta} \overset{iid}{\sim} MVN(\mathbf{0}, 10^6 \boldsymbol{I})$

Goal: Obtain posterior summaries about county totals, θ_j

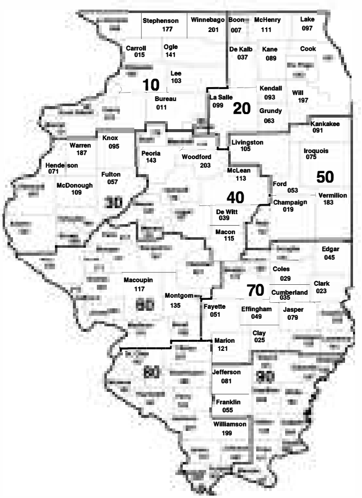
County Agricultural Production Survey (CAPS)

Case study in Cruze et al. (2016)

Illinois planted corn

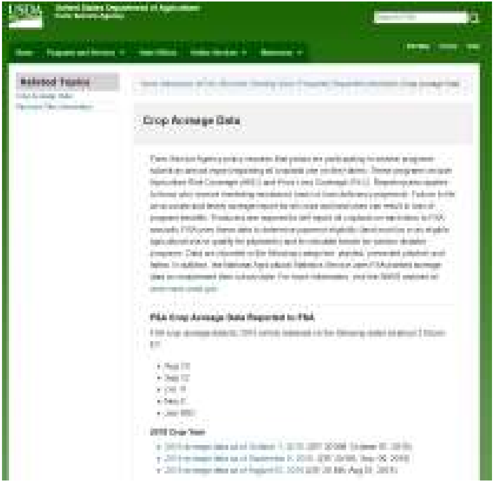
- 9 Ag. Statistics Districts
- 102 counties
- a major producer of corn
- End-of-season survey
  - Direct estimates of totals
  - Estimated sampling variances

	Min	Median	Max
n reports	2	47	93
CV (%)	9.1	19.2	92.3



https://www.nass.usda.gov/Charts_and_Maps/Crops_County/indexpdf.php

Covariate x_1: USDA Farm Service Agency (FSA) Acreage



https://www.fsa.usda.gov/news-room/efoia/electronic-reading-room/frequently-requested-information/crop-acreage-data/index

- FSA administers farm support programs
- Enrollment popular, not compulsory
- Data self-reported at FSA office
- Administrative vs. physical county

Covariate x_2: NOAA Climate Division March Precipitation

Weather as auxiliary variable

- March: Planting ‘intentions’
- April: Illinois planting
- **Could rainfall in March affect planting?**
- One-to-one mapping: ASD and climate division
- Repeat value for all counties within ASD

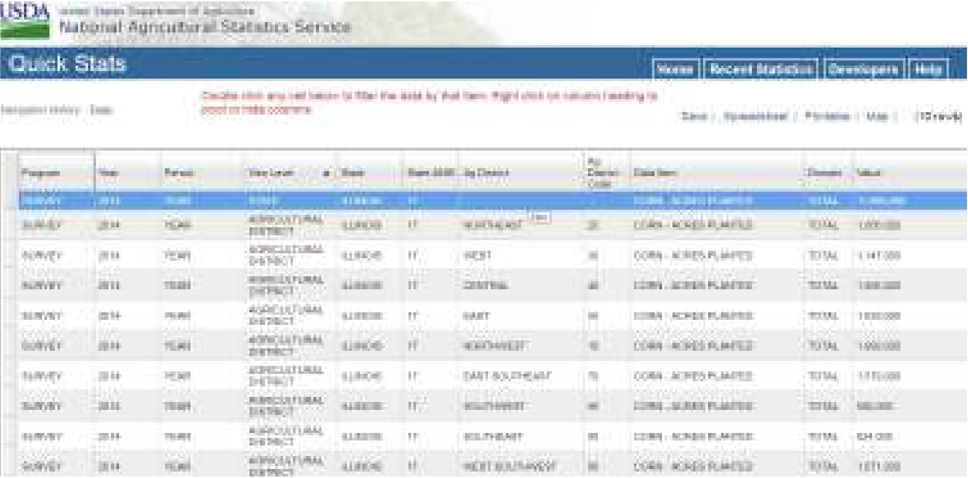
ASD	Precip (in)
10	1.08
20	1.35
30	1.27
40	1.66
50	1.50
60	1.36
70	1.46
80	1.69
90	2.00

Source: ftp://ftp.ncdc.noaa.gov/pub/data/cirs/climdiv
Details in Vose et al. (2014)

NASS Official Statistics

From prior publication: Illinois 2014, 11.9 million acres of corn planted

- Require: State-ASD-county benchmarking of estimates



State/district: https://quickstats.nass.usda.gov/results/3A17F375-B762-37BD-8C03-D581DC8F7A85
County: https://quickstats.nass.usda.gov/results/478D1A7B-E680-3E5E-95E4-9A59F938A256

JAGS Model

```
##### Assume this source saved in C:/Your Directory Name/Your_JAGS_model.R
model{
        for(j in 1:m){      #Looping over counties, m=102 for Illinois

        #Defines `data model'-note-JAGS uses precision
            thetahat[j] ~ dnorm(theta[j],  1/vhat.dir[j])

        #Defines `process model'
            theta[j] ~ dnorm(beta0+beta1*X1[j]+beta2*X2[j], sigma2u.inv)
        }

    ## Priors:
    sigma2u ~ dunif(0, 10^8)
    sigma2u.inv <- pow(sigma2u, -1)      #Again, precision

    beta0~dnorm(0,.000001)               #Again, precision
    beta1~dnorm(0,.000001)
    beta2~dnorm(0,.000001)
}
```

- Note data, process, prior structure from earlier slide
- Note distributions parameterized in terms of precision
- Read into R script as stored R source code or as text string

Online resources http://www.sumsar.net/blog/2013/06/three-ways-to-run-bayesian-models-in-r/

A Pseudo-Code R Script

```
##### Loading some libraries--assumes functioning JAGS installation
library(rjags)
library(R2jags)

##### Your data import and wrangling  go here
##### We'll actually fit a model scaled by 'size' (# reports)
thetahat<-DirInd/Size              #### Survey Estimate
vhat.dir<-VarDirInd/Size^2         #### Estimated Survey Variance
X1<-rma_pccm/size                  #### RMA data
X2<-test$pcpn.3                    #### NOAA March Precipitation

################ Initialize Model
set.seed([illegible]); m=[illegible]        #### Set seed; define number of counties

#### Initialize sampler--Plausible initial value
#### for sigma2u based on least squares
init.sig <- (summary(init.lm.coef)$sigma^2)

#### Distinguish data inputs and parameters
jags.data <- list("thetahat","vhat.dir", "X1","X2","m")
jags.params <- c("theta","sigma2u","beta0","beta1","beta2")

jags.inits <- function(){list("sigma2u" = init.sig)}    #### Function for initial value

#### Execute model: assumes JAGS as source code; object returned is an R list object
jags(jags.data , jags.inits , jags.params, "C:/Your Directory Name/Your_JAGS_model.R",
    n.chains = [illegible], n.iter = [illegible], n.burnin = [illegible])
```

Analysis of JAGS Model Output

Posterior summaries of parameters–based on 3,000 saved iterates

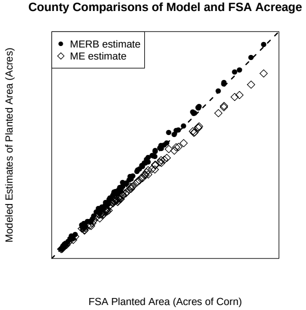
- Posterior means, standard deviations, quantiles, potential scale reduction factors, effective sample sizes, pD, DIC

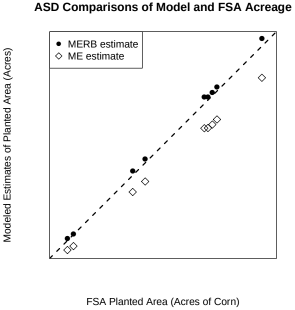
```
Inference for Bugs model at C:/Your Directory Name/Your_JAGS_model.R
 3 chains, each with 10000 iterations (first 1000 discarded), n.thin = 9
 n.sims = 3000 iterations saved
            mu.vect   sd.vect     2.5%       25%       50%       75%     97.5%  Rhat n.eff
beta0        97.024   205.223 -297.362   -39.365    94.004   235.130   492.579 1.002  1500
beta1         0.865     0.037    0.790     0.841     0.865     0.891     0.937 1.005   830
beta2       -48.553   118.049 -276.194  -126.387   -48.104    28.315   183.179 1.001  2300
sigma2u   20223.038 11544.842 3252.631 11870.939 18247.001 26419.969 47345.031 1.039    84
theta[1]   3399.432   163.965 3083.123  3296.654  3399.326  3505.508  3719.588 1.002  3000
theta[2]   1982.413   153.739 1690.704  1885.191  1977.139  2076.279  2302.119 1.001  3000
theta[3]   2621.446   149.324 2320.691  2525.084  2620.279  2713.351  2925.278 1.001  3000
theta[4]   1296.049   141.511 1014.616  1209.529  1291.823  1383.444  1582.351 1.001  3000
theta[5]   3456.315   157.861 3120.367  3359.261  3458.199  3557.888  3754.838 1.002  1900
```

- Transform back to acreage scale
- Ratio benchmarking–inject benchmarking factor back into chains as in Erciulescu et al. (2018)

Results: Models With and Without Benchmarking

- Modeled estimates (ME) may not satisfy benchmarking
- Ratio-benchmarked estimates (MERB) are consistent with state targets and improve agreement with external sources

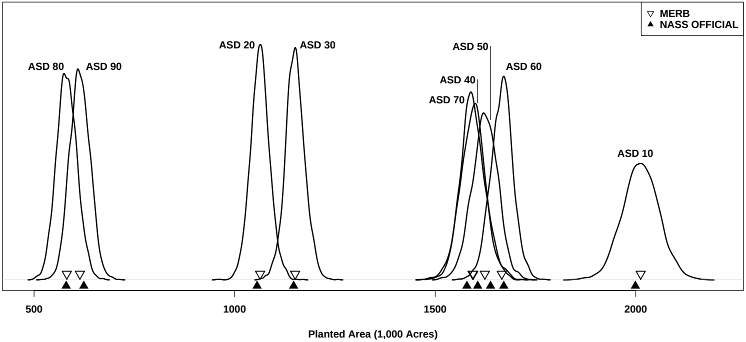




Results: Posterior Distributions of ASD-Level Acreages

Used county-level inputs to produce county-level estimates

- **Idea:** derive ASD-level estimates from Monte Carlo iterates
- Sum corresponding draws from county posterior distributions
  - Compute means and variances from aggregated chains



Results: Relative Variability of Survey Versus Model

Obtain estimates and measures of uncertainty for counties and districts

- Recall the goal of SAE–increased precision!

CV (%) of CAPS Survey Estimates

	Min	*Q1*	*Median*	*Mean*	*Q3*	*Max*
County	9.1	16.6	19.2	22.2	23.5	92.3
District	4.4	5.6	6.8	6.6	7.2	8.7

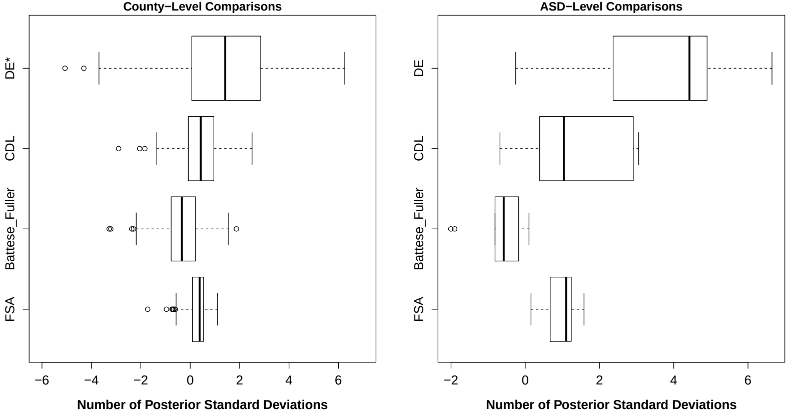
CV (%) of MERB Estimates

	Min	*Q1*	*Median*	*Mean*	*Q3*	*Max*
County	3.6	5.6	7.2	9.0	10.5	31.2
District	1.7	2.0	2.1	2.5	2.3	4.4

Results: Comparison to Other Sources

For counties and districts, compute 'standard score'

- (model estimate-other source)/model standard error
- Direct Estimates, Cropland Data Layer, Battese-Fuller, FSA



Conclusions

Discussed Bayesian formulation of Fay-Herriot model motivated by NASS applications

Other R packages facilitate Bayesian small area estimation

- 'BayesSAE' by Chengchun Shi
- 'hbsae' by Harm Jan Boonstra
- May be bound by limited choice of prior distributions
- Transformations of data may be needed

Proc MCMC in SAS added 'Random' statement as of version 9.3

Thanks to Andreea Erciulescu (NISS) and Balgobin Nandram (WPI) for three years of adventures in small area estimation!

References

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Browne, W. J. and Draper, D. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514.

Cressie, N. and Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ.

Cruze, N., Erciulescu, A., Nandram, B., Barboza, W., and Young, L. (2016). Developments in Model-Based Estimation of County-Level Agricultural Estimates. In *Proceedings of the Fifth International Congress on Establishment Surveys*. American Statistical Association, Geneva.

Erciulescu, A. L., Cruze, N. B., and Nandram, B. (2018). Model-based county level crop estimates incorporating auxiliary sources of information. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. doi:10.1111/rssa.12390.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366):269–277.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534.

Iwig, W. (1996). The National Agricultural Statistics Service County Estimates Program. In Schaible, W., editor, *Indirect Estimators in U.S. Federal Programs*, chapter 7, pages 129–144. Springer, New York.

National Academies of Sciences, Engineering, and Medicine (2017). Improving Crop Estimates by Integrating Multiple Data Sources. The National Academies Press, Washington, DC.

Rao, J. and Molina, I. (2015). *Small Area Estimation*. In Wiley Online Library: Books. Wiley, 2nd edition.

Vose, R. S., Applequist, S., Squires, M., Durre, I., Menne, M. J., Williams, C. N., Fenimore, C., Gleason, K., and Arndt, D. (2014). Improved Historical Temperature and Precipitation Time Series for U.S. Climate Divisions. *Journal of Applied Meteorology and Climatology*, 53(5):1232–1251.