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
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)
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

\[
\hat{\theta}_j = \theta_j + e_j, \quad j = 1, \ldots, m \text{ counties} \tag{1}
\]

\[
\theta_j = x'_j \beta + u_j \tag{2}
\]

Adding Eqs. 1 and 2

\[
\hat{\theta}_j = x'_j \beta + u_j + e_j
\]

- \(\hat{\theta}_j\), direct estimate
- \(E(e_j|\theta_j) = 0\)
- \(V(e_j|\theta_j) = \hat{\sigma}_j^2\), estimated variance
- \(x_j\), known covariates
- \(u_j\), area random effect
- \(u_j \sim iid (0, \sigma_u^2)\)
Fay-Herriot Formulated As Bayesian Hierarchical Model

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

Data model:
\[ \hat{\theta}_j | \theta_j, \beta \overset{\text{ind}}{\sim} N(\theta_j, \hat{\sigma}_j^2) \] (3)

Process model:
\[ \theta_j | \beta, \sigma_u^2 \overset{iid}{\sim} N(x'_j \beta, \sigma_u^2) \] (4)

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

Goal: Obtain posterior summaries about county totals, \( \theta_j \)
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

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>n reports</td>
<td>2</td>
<td>47</td>
<td>93</td>
</tr>
<tr>
<td>CV (%)</td>
<td>9.1</td>
<td>19.2</td>
<td>92.3</td>
</tr>
</tbody>
</table>

Covariate $x_1$: USDA Farm Service Agency (FSA) Acreage

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

[Link to USDA FSA website for crop acreage data]

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

<table>
<thead>
<tr>
<th>ASD</th>
<th>Precip (in)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.08</td>
</tr>
<tr>
<td>20</td>
<td>1.35</td>
</tr>
<tr>
<td>30</td>
<td>1.27</td>
</tr>
<tr>
<td>40</td>
<td>1.66</td>
</tr>
<tr>
<td>50</td>
<td>1.50</td>
</tr>
<tr>
<td>60</td>
<td>1.36</td>
</tr>
<tr>
<td>70</td>
<td>1.46</td>
</tr>
<tr>
<td>80</td>
<td>1.69</td>
</tr>
<tr>
<td>90</td>
<td>2.00</td>
</tr>
</tbody>
</table>

Details in Vose et al. (2014)
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

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

```r
# Priors:
sigma2u ~ dunif(0, 10^8)
sigma2u.inv <- pow(sigma2u, -1) # Again, precision
beta0~dnorm(0, 0.000001) # Again, precision
beta1~dnorm(0, 0.000001)
beta2~dnorm(0, 0.000001)
```
A Pseudo-Code R Script

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

### Your data import and wrangling go here

### we'll actually fit a model, called by "size" (a vector)
that.dat <- DirMult/Size
that.dire <- VarDirInd/size
zic <- size / Nsize
s2 <- test & prop.

### Initialize Model
set.seed(21101); m = CMC

### Initialize sampler - plausible initial value
### for sd's based on least squares
init.sig <- summary(init_lm.coef$sigmas^2)

### Distinguish data inputs and parameters
jags.data <- list("that.dat", "that.dire", "zic", "size", "m")
jags.params <- c("theta", "alpha1", "beta1", "beta2")

jags.init <- function(){list("sigmas2" = init.sig)}

### Execute model; assumes JAGS as output code; object returned is a list object
jags <- jags.data, jags.init, jags.params, "~/

n.chains = 3, in.iter = 1000, n.burnin = 1000
"
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

- 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

![Graph showing posterior distributions of planted area across different ASD levels]
Results: Relative Variability of Survey Versus Model

Obtain estimates and measures of uncertainty for counties and districts

- Recall the goal of SAE–increased precision!

<table>
<thead>
<tr>
<th>CV (%) of CAPS Survey Estimates</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>County</td>
<td>9.1</td>
<td>16.6</td>
<td>19.2</td>
<td>22.2</td>
<td>23.5</td>
<td>92.3</td>
</tr>
<tr>
<td>District</td>
<td>4.4</td>
<td>5.6</td>
<td>6.8</td>
<td>6.6</td>
<td>7.2</td>
<td>8.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CV (%) of MERB Estimates</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>County</td>
<td>3.6</td>
<td>5.6</td>
<td>7.2</td>
<td>9.0</td>
<td>10.5</td>
<td>31.2</td>
</tr>
<tr>
<td>District</td>
<td>1.7</td>
<td>2.0</td>
<td>2.1</td>
<td>2.5</td>
<td>2.3</td>
<td>4.4</td>
</tr>
</tbody>
</table>
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

County-Level Comparisons

ASD-Level Comparisons
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


