Best prediction combining several sources of information from survey data and big data

Jae-Kwang Kim

Iowa State University

Cox Award Talk
June 23, 2015

1 Joint work with Zhonglei Wang (ISU) and Nathan Cruz (NASS)
Gertrude M. Cox

- Born in Dayton, Iowa (1900).
- B.S and M.S. at Iowa State College (1929, 1931).
- Worked as an assistant in the Statistical Laboratory at Iowa State College (1933-1939).
- Assistant professor of Statistics at Iowa State University (1939).
- Move to NC State University (1940), Director of RTI (1960).
1 Introduction

2 Hierarchical Structural model

3 Application to NASS project

4 Conclusion
Motivation

- Cooperative agreement with National Agricultural Statistical Service (NASS) in USDA.
- Want to incorporate information from several sources
  1. JAS (June Area Survey): Survey estimates obtained from an area frame sample.
  2. FSA (Farm Service Agency): Self-reported administrative data.
  3. CDL (Cropland Data Layer): Classification of satellite imagery data.

Summary

<table>
<thead>
<tr>
<th>Source</th>
<th>Observation</th>
<th>Main source of Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAS</td>
<td>direct measurement</td>
<td>Sampling error</td>
</tr>
<tr>
<td>FSA</td>
<td>Self-reported data</td>
<td>Coverage error</td>
</tr>
<tr>
<td>CDL</td>
<td>Satellite image classification</td>
<td>Measurement error</td>
</tr>
</tbody>
</table>
**Figure:** The 2009 cropland data layer products. The legend identifies aggregated agricultural and non-agricultural land cover categories by decreasing acreage.
Area level model is used to combine information from the three sources.

- We observe three estimates for each area.

- Measurement: crop acres (for each crop)

- $Y_i$: true crop acres for area $i$ (Unobserved)

- $\hat{Y}_i$: estimate from JAS (subject to sampling error)

- $X_{1i}$: estimate from FSA data (subject to coverage error)

- $X_{2i}$: estimate from CDL data (subject to measurement error, or classification error)

- The analysis unit is a sub-state area, called a district.

- Three estimates are correlated.
Figure: Relationship between JAS and CDL.
The goal is to predict $Y_i (=\text{true crop acres})$ using the observation of $\hat{Y}_i (=\text{JAS})$, $X_{i1} (=\text{FSA})$, and $X_{i2} (=\text{CDL})$.

Area level model is a useful tool for combining information from different sources by making an area level matching.

Area level model consists of two parts:

1. Sampling error model: relationship between $\hat{Y}_i$ and $Y_i$.
2. Structural error model: relationship between $Y_i$ and $(X_{i1}, X_{i2})$. 
Area level model

- **Sampling Error Model:**
  \[ \hat{Y}_i \sim N(Y_i, V_i) \]

  where \( V_i = V(\hat{Y}_i) \), the sampling variance of \( \hat{Y}_i \), is consistently estimated from JAS sample data.

- **Structural Error Model:** Explain \( Y_i \) in terms of \((X_{i1}, X_{i2})\).
  \[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i \]

  where \( e_i \) is a model error term. It is often called Fay-Herriot model approach.

- Instead of Fay-Herriot model approach, one can also consider measurement error model approach which explains \((X_{i1}, X_{2i})\) in terms of \( Y_i \).
Area level model: Fay-Herriot model approach

Figure: A Directed Acyclic Graph (DAG) for classical area level models \( (X = (X_1, X_2)) \).

(1): Sampling error model (known),
(2): Structural error model (known up to \( \theta \)).
Combining two models

Three components

1. Data: \((\hat{Y}_i, X_{i1}, X_{i2})\)
2. Latent variable: \(Y_i\)
3. Parameter: \(\theta\) that characterizes the structural error model

\[
f(Y_i \mid X_{i1}, X_{i2}) = f_1(Y_i \mid X_{i1}, X_{i2}; \theta)
\]

Sampling error model is a model for the unbiased measurement \(\hat{Y}_i\) for the latent variable \(Y_i\):

\[
g(\hat{Y}_i \mid Y_i)
\]
Prediction model

- Prediction model = sampling error model + structural error model
- Bayes formula for prediction model

\[ p(Y_i \mid \hat{Y}_i, X_{i1}, X_{i2}) \propto g(\hat{Y}_i \mid Y_i) f(Y_i \mid X_{i1}, X_{i2}), \]

where \( g(\cdot) \) is the sampling error model and \( f(\cdot) \) is the structural error model.
- \( g(\cdot) \): assumed to be known.
- \( f(\cdot) \): known up to parameter \( \theta \).
## Comparison

<table>
<thead>
<tr>
<th></th>
<th>Bayesian</th>
<th>Frequentist</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
<td>Posterior distribution $p(\text{latent, } \theta \mid \text{data})$</td>
<td>Prediction model $p(\text{latent} \mid \text{data, } \theta)$</td>
</tr>
<tr>
<td><strong>Parameter Est’n</strong></td>
<td>Data augmentation</td>
<td>EM algorithm</td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td>I-step</td>
<td>E-step</td>
</tr>
<tr>
<td><strong>Parameter update</strong></td>
<td>P-step</td>
<td>M-step</td>
</tr>
<tr>
<td><strong>Best prediction</strong></td>
<td>$E(\text{latent} \mid \text{data})$</td>
<td>$E(\text{latent} \mid \text{data, } \hat{\theta})$</td>
</tr>
</tbody>
</table>
Parameter estimation (frequentist approach)

- Obtain the prediction model using Bayes formula
- EM algorithm: Update the parameters

\[ \hat{\theta}^{(t+1)} = \arg_{\theta} \max \sum_{i} E\{\log f(Y_i | X_{i1}, X_{i2}; \theta) \mid \hat{Y}_i, X_{i1}, X_{i2}; \hat{\theta}^{(t)}\} \]

where the conditional expectation is with respect to the prediction model evaluated at the current parameter \( \hat{\theta}^{(t)} \).
Prediction vs Parameter estimation

Figure: EM algorithm

\[ \hat{Y} \rightarrow \hat{\theta} \rightarrow Y \]

E-step

M-step
Prediction (frequentist approach)

- Best prediction: Expectation from the prediction model at $\theta = \hat{\theta}$

$$\hat{Y}_i^* = E\{Y_i \mid \hat{Y}_i, X_{i1}, X_{i2}; \hat{\theta}\}$$

- If $f(Y_i \mid X_{i1}, X_{i2})$ is a normal distribution then

$$\hat{Y}_i^* = \alpha_i \hat{Y}_i + (1 - \alpha_i)E(Y_i \mid X_{i1}, X_{i2}; \hat{\theta})$$

for some $\alpha_i$ where

$$\alpha_i = \frac{V(Y_i \mid X_{i1}, X_{i2}; \hat{\theta})}{V(\hat{Y}_i) + V(Y_i \mid X_{i1}, X_{i2}; \hat{\theta})}.$$  

This is often called composite estimator.
Prediction (frequentist approach)

- Note that
  \[ \hat{Y}_i^* - Y_i = \left\{ \hat{Y}_i^* - \tilde{Y}_i^* \right\} + \left\{ \tilde{Y}_i^* - Y_i \right\} \]
  
  where \( \tilde{Y}_i^* = E\{ Y_i \mid \hat{Y}_i, X_{i1}, X_{i2}; \theta \} \).

- Mean Squared Prediction Error:
  \[ V\{ \hat{Y}_i^* - Y_i \} = V \left\{ \hat{Y}_i^* - \tilde{Y}_i^* \right\} + V \left\{ \tilde{Y}_i^* - Y_i \right\}. \]

  The first term is often small. The second term is estimated by \( V\{ Y_i \mid \hat{Y}_i, X_{i1}, X_{i2}; \theta \} \) evaluated at \( \theta = \hat{\theta} \).

- If \( f(Y_i \mid X_{i1}, X_{i2}) \) is a normal distribution then
  \[ V\{ Y_i \mid \hat{Y}_i, X_{i1}, X_{i2}; \theta \} = \alpha_i V(\hat{Y}_i \mid Y_i) \leq V(\hat{Y}_i \mid Y_i). \]
Summary

- **Basic Steps**
  1. Model specification: Structural error model + Sampling error model
  2. Parameter estimation: EM algorithm
  3. Prediction: Bayes theorem
  4. Estimation of Mean Squared Prediction Error (MSPE)

- Either frequentist approach or Bayesian approach can be used.

- Often called **Small Area Estimation** in survey sampling.
1 Introduction

2 Hierarchical Structural model

3 Application to NASS project

4 Conclusion
The best predictor of $Y_i$, discussed in Section 1, provides a smaller variance than the direct estimator $\hat{Y}_i$ because

$$\begin{aligned}
V\{\hat{Y}_i^* - Y_i\} &= V\{\tilde{Y}_i^* - Y_i\} \\
&= E\{V(Y_i | \hat{Y}_i, X_{i1}, X_{i2})\} \\
&\leq E\{V(Y_i | \hat{Y}_i)\} \\
&= V(\hat{Y}_i - Y_i),
\end{aligned}$$

where $\tilde{Y}_i^* = E\{Y_i | \hat{Y}_i, X_{i1}, X_{i2}; \theta\}$. 
Motivation (Cont’d)

- However, such phenomenon does not hold for predicting $Y_h = \sum_{i \in S_h} Y_i$, where $S_h$ is the set of districts in state $h$. In fact,

$$V\{\hat{Y}_h^* - Y_h\} = V\{\hat{Y}_h^* - \tilde{Y}_h^*(\theta_h)\} + V\{\tilde{Y}_h^*(\theta_h) - Y_h\}$$

and the first term is no longer small when $\hat{\theta}_h$ is computed from the same sample in state $h$.

- Thus, to obtain an improved predictor for the state-level crop acres, we use a two-level model that borrows strength from other states.
Hierarchical structural model

- Use a two-level structural error model.
  - Level 1 (Within-state model):
    \[ Y_{hi} = X_{hi} \beta_h + u_{hi} \]
    where \( X_{hi} = (1, X_{hi1}, X_{hi2}) \) and \( \beta_h = (\beta_{h0}, \beta_{h1}, \beta_{h2})' \).
  - Level 2 (between-state model):
    \[ \beta_h \sim N(\beta_2, \Sigma_2) \]
    i.e.
    \[
    \begin{pmatrix}
    \beta_{h0} \\
    \beta_{h1} \\
    \beta_{h2}
    \end{pmatrix}
    \sim
    N
    \left[
    \begin{pmatrix}
    \beta_0 \\
    \beta_1 \\
    \beta_2
    \end{pmatrix},
    \begin{pmatrix}
    \sigma_{00} & \sigma_{01} & \sigma_{02} \\
    \sigma_{10} & \sigma_{11} & \sigma_{12} \\
    \sigma_{20} & \sigma_{21} & \sigma_{22}
    \end{pmatrix}
    \right].
    \]
  - Sampling error model remains the same. \( \hat{Y}_{hi} \sim N(Y_{hi}, V_{hi}) \).
Two-level structural error model

- Level 1 model
  \[ Y_{hi} \sim f_1(Y_{hi} | X_{hi}; \theta_h) \]
  and \( \hat{Y}_{hi} \) is a measurement for \( Y_{hi} \).

- Level 2 model
  \[ \theta_h \sim f_2(\theta_h | Z_h; \zeta) \]
  where \( Z_h \) is the state-specific covariate. An example of \( Z_h \) is a classification of states into groups (major crop states / minor crop states).
Figure: A DAG for two-level Fay-Harriott model.

(1): Sampling error model,
(2): level-one structural error model,
(3): level-two structural error model
Hierarchical Structural model

Frequentist approach to multi-level model estimation

- Three steps for parameter estimation in each level
  1. **Summarization**: Find a measurement for latent variable to obtain the sampling error model.
  2. **Combine**: Find a prediction model for latent variable by combining the sampling error model and the structural error model.
  3. **Learning**: Estimate the parameters from data.

- Apply the three steps in level one model and then do these in level two model.
Hierarchical Structural model

## Structures in Two level model

<table>
<thead>
<tr>
<th>Model</th>
<th>Measurement (Data summary)</th>
<th>Parameter</th>
<th>Latent variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level one</td>
<td>$\hat{Y}<em>h = (\hat{Y}</em>{h1}, \cdots, \hat{Y}_{hn_h})$</td>
<td>$\theta_h$</td>
<td>$Y_h = (Y_{h1}, \cdots, Y_{hn_h})$</td>
</tr>
<tr>
<td>Level two</td>
<td>$\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_H)$</td>
<td>$\zeta$</td>
<td>$\theta = (\theta_1, \cdots, \theta_H)$</td>
</tr>
</tbody>
</table>
Parameter estimation for Level 1 model

- $\theta_h$: parameter in $f_1(Y_{hi} \mid X_{hi}; \theta_h)$
- **Summarization**: Find a measurement $\hat{Y}_{hi}$ for $Y_{hi}$ and obtain the sampling error model

$$\hat{Y}_{hi} \mid Y_{hi} \sim g_1(\hat{Y}_{hi} \mid Y_{hi})$$

- **Combine** the two models using Bayes formula

$$p_1(Y_{hi} \mid X_{hi}, \hat{Y}_{hi}; \theta_h) \propto g_1(\hat{Y}_{hi} \mid Y_{hi})f_1(Y_{hi} \mid X_{hi}; \theta_h)$$

- **Learning**: Parameter is estimated using EM algorithm

$$\hat{\theta}_h^{(t+1)} = \arg \max_{\theta_h} \sum_i E\{\log f_1(Y_{hi} \mid X_{hi}; \theta_h) \mid X_{hi}, \hat{Y}_{hi}; \hat{\theta}_h^{(t)}\}.$$
Hierarchical Structural model

Parameter estimation for Level 1 model (Cont’d)

**Figure:** EM algorithm for level 1 model

![EM algorithm for level 1 model](image-url)
Parameter estimation for Level 2 model

- \( \zeta \): parameter in \( f_2(\theta_h \mid Z_h; \zeta) \)

- **Summarization**: Find a measurement \( \hat{\theta}_h \) for \( \theta_h \) and obtain the sampling error model

\[
\hat{\theta}_h \mid \theta_h \sim g_2(\hat{\theta}_h \mid \theta_h)
\]

- **Combine** the two models using Bayes formula

\[
p_2(\theta_h \mid Z_h, \hat{\theta}_h; \zeta) \propto g_2(\hat{\theta}_h \mid \theta_h)f_2(\theta_h \mid Z_h; \zeta)
\]

- **Learning**: Parameter is estimated using EM algorithm

\[
\hat{\zeta}^{(t+1)} = \arg\max_{\zeta} \sum_h E\{\log f_2(\theta_{hi} \mid Z_h; \zeta) \mid Z_h, \hat{\theta}_h; \hat{\zeta}^{(t)}\}.
\]
Parameter estimation for Level 2 model (Cont’d)

Figure: EM algorithm for level 2 model
Best Prediction

- Prediction under Level 1 model
  - Latent variable: $Y_{hi}$
  - Best prediction
    $$\tilde{Y}^*_h(\theta) = E\{Y_{hi} \mid \hat{Y}_{hi}, X_{hi}; \theta_h\}$$
  - Under single level model, we would use $\hat{Y}^*_{hi} = \tilde{Y}^*_h(\hat{\theta}_h)$, where $\hat{\theta}_h$ is the MLE of $\theta_h$ under the level one model.

- Prediction under two level model: compute
  $$\tilde{Y}^{**}_{hi}(\zeta) = E\{\tilde{Y}^*_h(\theta) \mid Z_h, \hat{\theta}_h; \zeta\}$$
  and use
  $$\hat{Y}^{**}_{hi}(\zeta) = \tilde{Y}^{**}_{hi}(\hat{\zeta}).$$
Estimation of MSPE

For \( \hat{Y}^{**} = \sum_i \hat{Y}_{hi} \), we can show that

\[
E\{ (\hat{Y}^{**} - Y_h)^2 \} = \sum_i \alpha_{hi} V_{hi} + \left\{ \sum_i \sum_j (1 - \alpha_{hi})(1 - \alpha_{hj}) q_{hij} \right\}
\]

where \( V_{hi} = \hat{V}(\hat{Y}_{hi}) \)

\[
\alpha_{hi} = \frac{V(Y_{hi} | X_{hi})}{V_{hi} + V(Y_{hi} | X_{hi})}
\]

and

\[
q_{hij} = X_{hi} \left\{ \Sigma_2^{-1} + V(\hat{\beta}_h)^{-1} \right\}^{-1} X'_{hi}.
\]
1 Introduction

2 Hierarchical Structural model

3 Application to NASS project

4 Conclusion
Logistic regression model

- $M_{hi}$: total crop acres in district ($hi$). (This is available to us.)
- Since $\bar{Y}_{hi} = Y_{hi}/M_{hi}$ is a proportion, we may use

$$\bar{Y}_{hi} = p(\beta_{h0} + \beta_{h1}\bar{X}_{hi1} + \beta_{h2}\bar{X}_{hi2}) + e_{hi}$$

(1)

where $p(x) = \{1 + \exp(-x)\}^{-1}$ and

$$e_{hi} \sim (0, \psi_{hi}p_{hi}(1 - p_{hi}))$$

for some $\psi_{h} > 0$ and $p_{hi} = p(\beta_{h0} + \beta_{h1}\bar{X}_{hi1} + \beta_{h2}\bar{X}_{hi2})$.

- Berg and Fuller (2014) also considered model (1) in the single-level model approach.
- Thus, we extend Berg and Fuller (2014) approach to two-level model.
Application to NASS data

- We are interested in predicting the acres of soybean for each state.
- 14 major states: KY, WI, IL, SD, AR, KS, ND, MI, OH, MO, IN, IA, MN, NE
- 13 minor states: AL, NC, NY, GA, LA, TX, MX, TN, PA, OK, MD, SC, VA
- Two-level FHM approach
  1. Level One model: Logistic regression model
  2. Level Two model: normal model within major / minor groups

\[
\begin{pmatrix}
\beta_{h0} \\
\beta_{h1} \\
\beta_{h2}
\end{pmatrix}
\sim N
\begin{bmatrix}
\begin{pmatrix}
\beta_0 \\
\beta_1 \\
\beta_2
\end{pmatrix},
\begin{pmatrix}
\sigma_{00} & \sigma_{01} & \sigma_{02} \\
\sigma_{11} & \sigma_{12} & \\
& \sigma_{22}
\end{pmatrix}
\end{bmatrix}.
\]
For model diagnostics, we computed standardized residual, given by

\[ e_{l,hi} = \frac{\hat{Y}_{hi} - \hat{p}_hi}{\sqrt{V_{hi} + \psi_h \hat{p}_hi(1 - \hat{p}_hi)}}, \]

where \( \hat{p}_hi = p(\hat{\beta}_h0 + \hat{\beta}_h1 \bar{X}_{hi1} + \hat{\beta}_h2 \bar{X}_{hi2}). \)
Explanatory data analysis—Logistic two-level model

Figure: Plot of standardized residual $e_{l, hi}$ on the fitted value. The left one is for major crop states, and the right one is for minor crop states.
Results—2013 soybeans (Major Crop States)

Estimation results for soybeans of year 2013 based on unadjusted FSA (Major case)

Method: JAS FSA CDL Logis_S

Kim (ISU)

Best prediction from big data
Results—2013 soybeans (Minor Crop States)

Estimation results for soybeans of year 2013 based on unadjusted FSA (Minor case)

Method: JAS, FSA, CDL, Logis_S

Kim (ISU)  Best prediction from big data
The following summaries are obtained.

- Mean of the standard errors.
- Mean of relative efficiency:

\[ s_{model} / s_{JAS}, \]

where \( s_{model} \) and \( s_{JAS} \) are the standard errors of the model estimator and JAS, respectively.
Summary statistics for the two models (Cont’d)

<table>
<thead>
<tr>
<th>Year</th>
<th>Crop</th>
<th>s.e</th>
<th>R.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>corn</td>
<td>1.687</td>
<td>0.768</td>
</tr>
<tr>
<td></td>
<td>soybeans</td>
<td>1.567</td>
<td>0.816</td>
</tr>
<tr>
<td></td>
<td>winterwheat</td>
<td>1.086</td>
<td>0.733</td>
</tr>
<tr>
<td>2012</td>
<td>corn</td>
<td>1.827</td>
<td>0.819</td>
</tr>
<tr>
<td></td>
<td>soybeans</td>
<td>1.611</td>
<td>0.741</td>
</tr>
<tr>
<td></td>
<td>winterwheat</td>
<td>1.157</td>
<td>0.754</td>
</tr>
<tr>
<td>2013</td>
<td>corn</td>
<td>2.058</td>
<td>0.750</td>
</tr>
<tr>
<td></td>
<td>soybeans</td>
<td>1.508</td>
<td>0.685</td>
</tr>
<tr>
<td></td>
<td>winterwheat</td>
<td>1.143</td>
<td>0.770</td>
</tr>
</tbody>
</table>
1. Introduction

2. Hierarchical Structural model

3. Application to NASS project

4. Conclusion
Discussion

- Two-level structural model for “borrowing strength” from other states.
  - Model specification
  - Parameter estimation
  - Prediction
  - MSPE estimation
- Frequentist approach to hierarchical Fay-Herriot model.
- Extension of the proposed method to Measurement error model is under development.
The end

thank you!