

# Examining Sensitivity of Small Area Inferences to Uncertainty about Sampling Error Variances

William Bell

U.S. Census Bureau

[William.R.Bell@census.gov](mailto:William.R.Bell@census.gov)

## Disclaimer

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

# 1. Introduction

**Area level model (Fay and Herriot 1979):**

$$\begin{aligned}y_i &= Y_i + e_i \\ &= (\mathbf{x}'_i \boldsymbol{\beta} + u_i) + e_i\end{aligned}$$

- $y_i$  = direct survey estimate of population quantity  $Y_i$  for area  $i$
- $e_i$  = sampling error in  $y_i$  as an estimate of  $Y_i$
- $\mathbf{x}_i$  = vector of regression variables for area  $i$
- $\boldsymbol{\beta}$  = vector of regression parameters
- $u_i$  = area  $i$  random effect (model error).

## Model:

$$\begin{aligned}y_i &= Y_i + e_i \\ &= (\mathbf{x}'_i \boldsymbol{\beta} + u_i) + e_i\end{aligned}$$

## Standard Model Assumptions:

- $u_i \sim i.i.d. N(0, \sigma_u^2)$  (and independent of  $e_i$ )
- $e_i \sim \text{ind. } N(0, v_i)$
- $v_i$  are known.

Note: We really only have estimates  $\hat{v}_i$  of  $v_i$ .

What happens when  $\hat{v}_i \neq v_i$ ?

- How much does  $E[(Y_i - \hat{Y}_i)^2]$  increase?
- How much do we misstate  $E[(Y_i - \hat{Y}_i)^2]$ ?
- Can we (partly) address these issues by modeling  $\hat{v}_i$ ?

# Outline of Talk

1. Introduction ✓
2. Rough calculations (for large  $m$ ) of consequences of  $\hat{v}_i \neq v_i$ :
  - percent increase in MSE
  - percent misstatement of MSE.
3. Literature review on dealing with  $\hat{v}_i \neq v_i$ .
4. Empirical example – SAIPE model for age 5-17 state poverty rates and their variances

## 2. Rough calculations of consequences of $\hat{v}_i \neq v_i$

Consider simple case where  $\beta$  and  $\sigma_u^2$  are known ( $m$  very large), but  $v_i$  are unknown, estimated by  $\hat{v}_i$ . Let

$$\begin{aligned}\tilde{Y}_i &= h_i y_i + (1 - h_i) \mathbf{x}'_i \beta \\ \hat{Y}_i &= \hat{h}_i y_i + (1 - \hat{h}_i) \mathbf{x}'_i \beta\end{aligned}$$

where

$$h_i = \frac{\sigma_u^2}{\sigma_u^2 + v_i} = \left(1 + \frac{v_i}{\sigma_u^2}\right)^{-1} \quad \hat{h}_i = \frac{\sigma_u^2}{\sigma_u^2 + \hat{v}_i} = \left(1 + \frac{\hat{v}_i}{\sigma_u^2}\right)^{-1}.$$

Then the MSE of  $\hat{Y}_i$  conditional on  $\hat{v}_i$  is

$$E[(Y_i - \hat{Y}_i)^2 | \hat{v}_i] = E[(Y_i - \tilde{Y}_i)^2] + E[(\tilde{Y}_i - \hat{Y}_i)^2 | \hat{v}_i].$$

The MSE of  $\tilde{Y}_i$  is  $\sigma_u^2(1 - h_i)$ . The reported MSE of  $\hat{Y}_i$  is  $\sigma_u^2(1 - \hat{h}_i)$ .

After a little algebra, we have that

$$\begin{aligned} \text{MSE pct diff} &\equiv 100 \times \frac{\text{MSE}(Y_i - \hat{Y}_i) - \text{MSE}(Y_i - \tilde{Y}_i)}{\text{MSE}(Y_i - \tilde{Y}_i)} \\ &= 100 \times \frac{(h_i - \hat{h}_i)^2}{h_i(1 - h_i)}. \end{aligned}$$

$$\begin{aligned} \text{MSE relbias} &= 100 \times \frac{\text{reported MSE}(Y_i - \hat{Y}_i) - \text{actual MSE}(Y_i - \hat{Y}_i)}{\text{actual MSE}(Y_i - \hat{Y}_i)} \\ &= 100 \times \left\{ \frac{\sigma_u^2(1 - \hat{h}_i)}{\sigma_u^2(1 - h_i) + (h_i - \hat{h}_i)^2(\sigma_u^2 + v_i)} - 1 \right\} \\ &= 100 \times \left\{ \frac{h_i(1 - \hat{h}_i)}{h_i(1 - h_i) + (h_i - \hat{h}_i)^2} - 1 \right\}. \end{aligned}$$



We examine MSE pct diff and MSE relbias for multiplicative errors in  $\hat{v}_i$  as an estimate of  $v_i$ :

underestimation factors:  $\hat{v}_i/v_i = \frac{3}{4}, \frac{1}{2}, \frac{1}{4}$

overestimation factors:  $\hat{v}_i/v_i = \frac{4}{3}, 2, 4$ .

For each of the above values of  $\hat{v}_i/v_i$ , plot MSE pct diff and MSE relbias for values of  $v_i/\sigma_u^2$  from .02, ..., 1, ..., 50 (on log scale).

Fig. 1. Percent difference in MSE and percent bias in reported MSE from using estimated versus true sampling variance

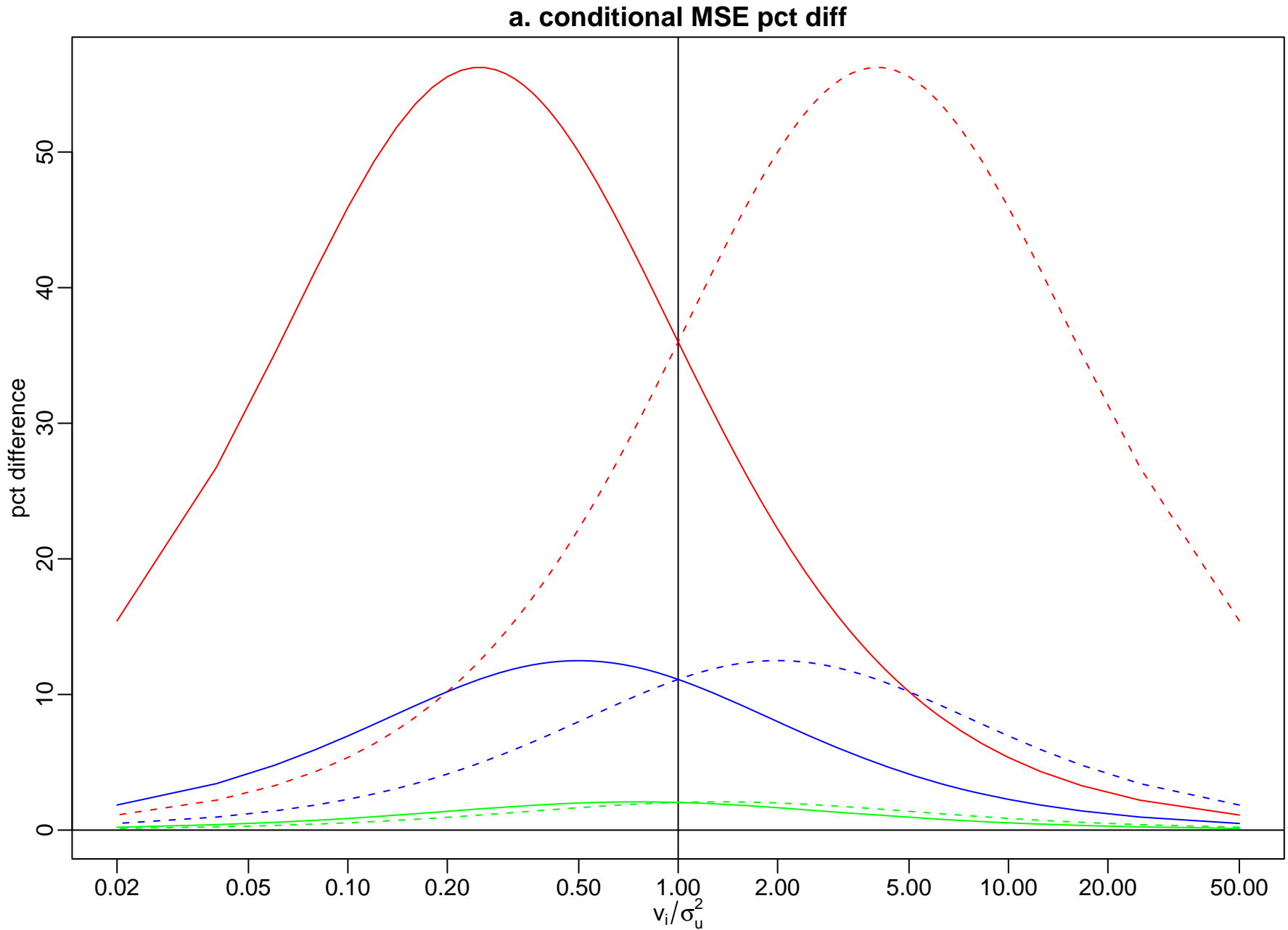
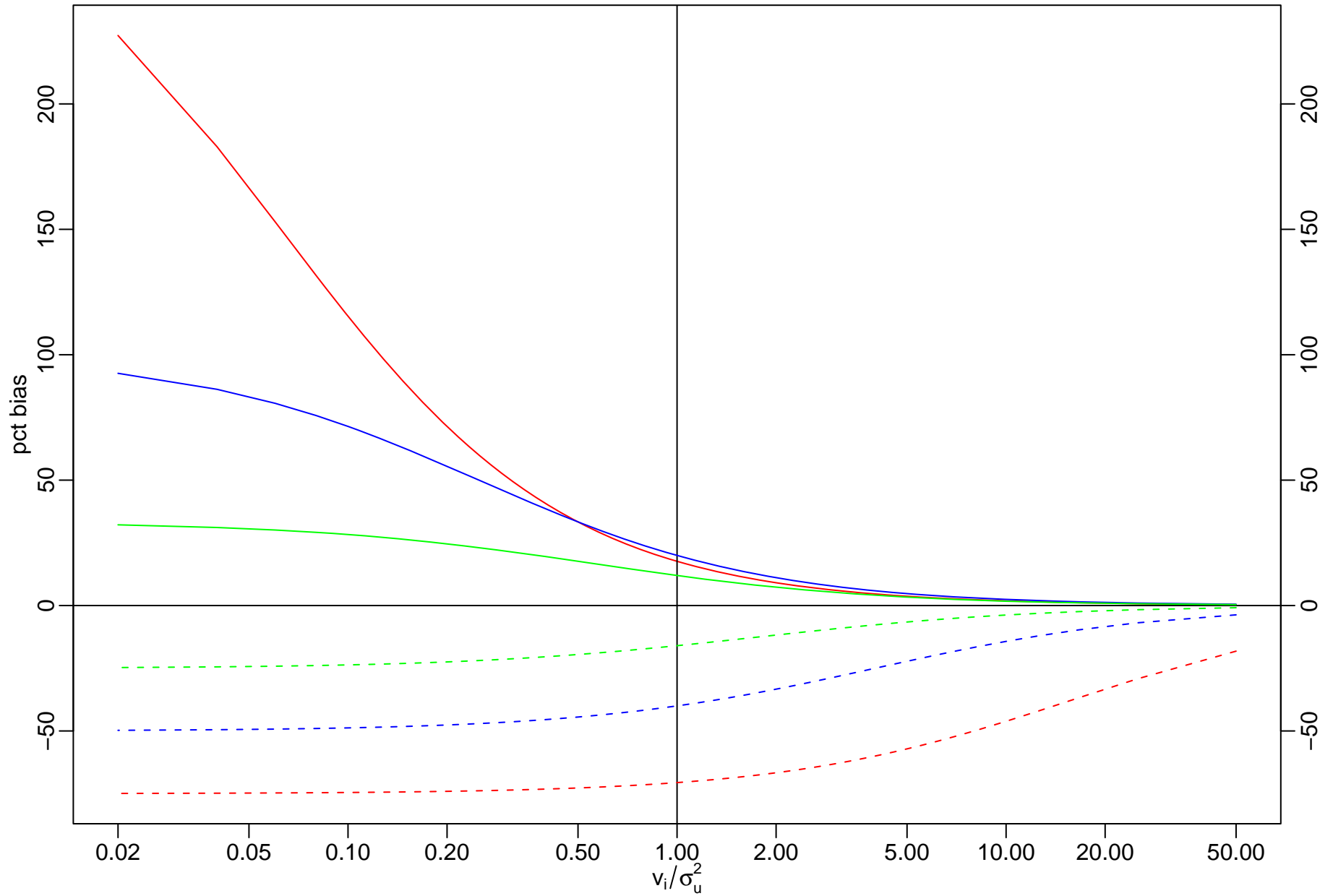


Fig. 1. Percent difference in MSE and percent bias in reported MSE  
from using estimated versus true sampling variance

**b. conditional MSE pct bias**



## Conclusions for large $v_i/\sigma_u^2$ :

- Underestimation of  $v_i$  is the more severe problem for both MSE pct diff and MSE relbias.

MSE increase is due to  $\hat{h}_i > h_i$ , so too much weight given to  $y_i$ .

## Conclusions for small $v_i/\sigma_u^2$ :

- Overestimation of  $v_i$  is the more severe problem for MSE pct diff.
- MSE relbias is very severe from either severe under- or overestimation of  $v_i$ .

Since large errors in  $\hat{v}_i$  seem more likely when  $v_i/\sigma_u^2$  is large, our general conclusion is:

The largest potential problem comes from  
severe underestimation of  $v_i$  when  $v_i/\sigma_u^2$  is large.

Given an assumed distribution of  $\hat{v}_i$ , unconditional versions of MSE pct diff and MSE relbias can be computed as

$$\text{MSE pct diff} = 100 \times \frac{E[(h_i - \hat{h}_i)^2]}{h_i(1 - h_i)}.$$

$$\text{MSE relbias} = 100 \times \left\{ \frac{h_i E(1 - \hat{h}_i)}{h_i(1 - h_i) + E[(h_i - \hat{h}_i)^2]} - 1 \right\}.$$

We do this (by numerical integration) assuming  $\hat{v}_i \sim v_i \chi_d^2/d$  for three values of  $d$  (6, 16, 80):

**Table 1. 5% and 95% points for the  $\chi_d^2/d$  distribution**

$d$	5% point	95% point
6	.27	2.10
16	.50	1.64
80	.75	1.27

Fig. 1. Percent difference in MSE and percent bias in reported MSE  
from using estimated versus true sampling variance

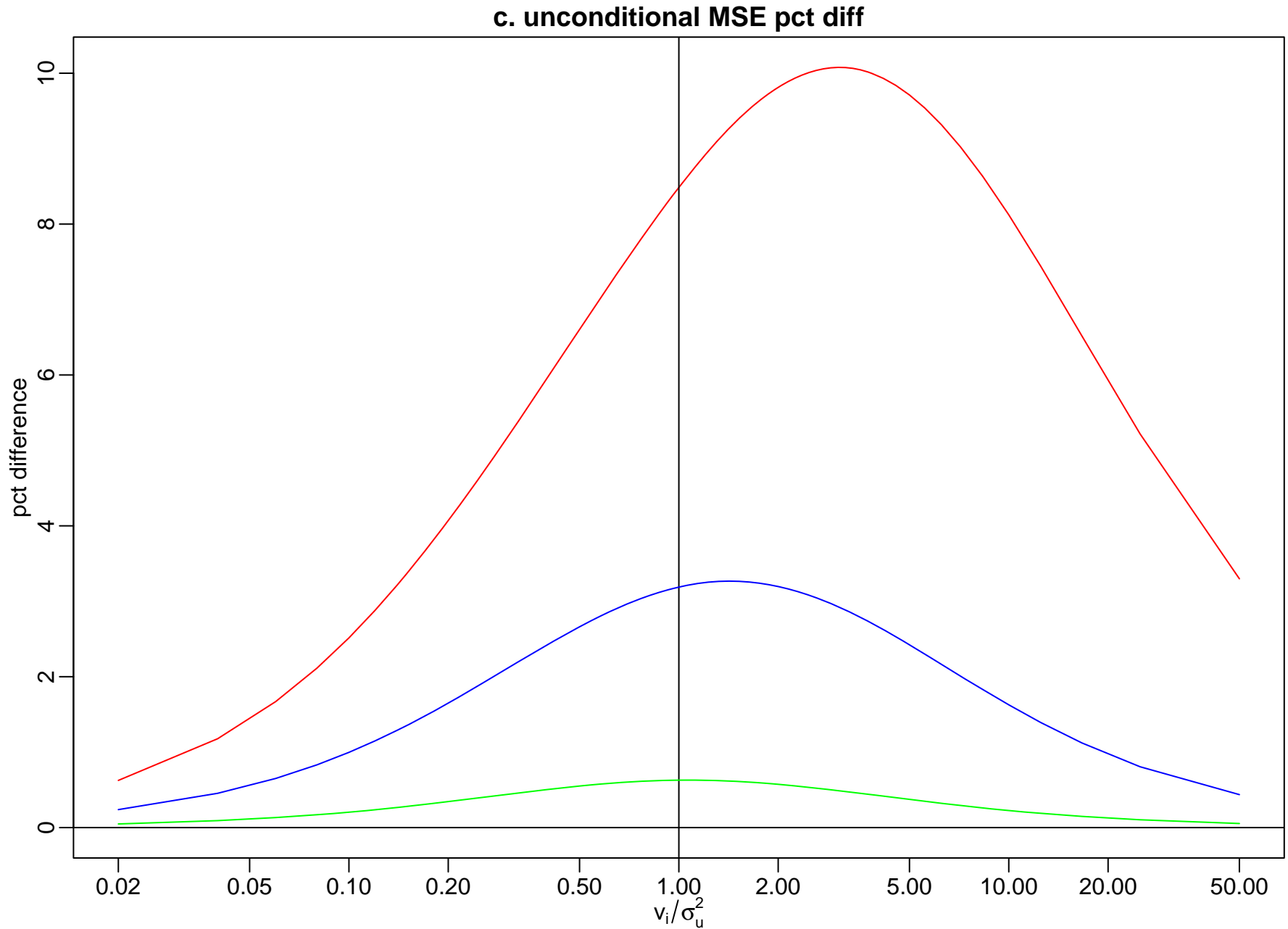
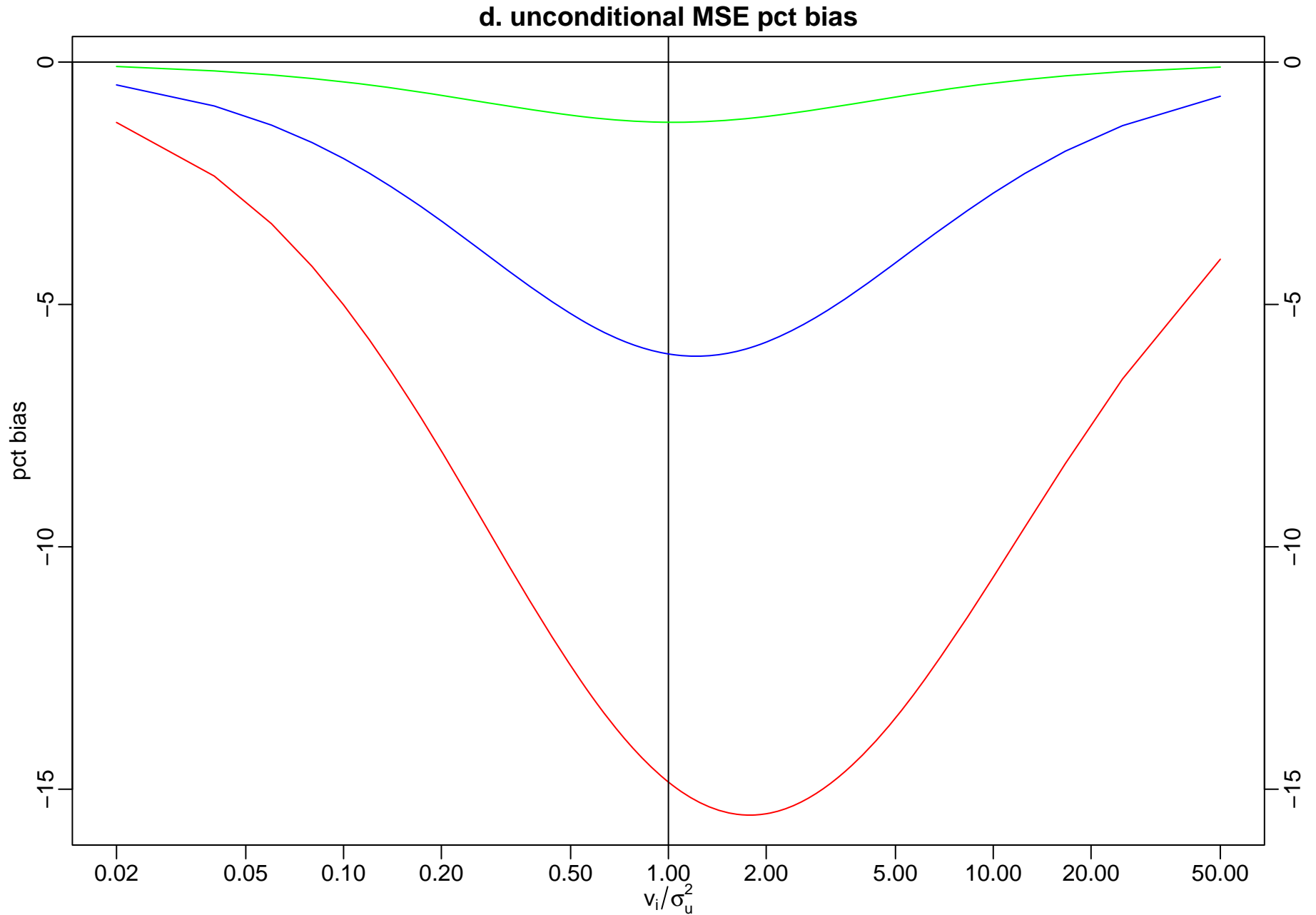


Fig. 1. Percent difference in MSE and percent bias in reported MSE  
from using estimated versus true sampling variance





### 3. Literature Review – Dealing with $\hat{v}_i \neq v_i$

- Approximate MSE results when  $v_i$  are estimated
- Modeling the  $\hat{v}_i$  to improve them (“small area variance modeling”)

Philosophy: If the direct survey point estimates need to be improved by small area modeling, so do the direct survey variance estimates.

## Approximate MSE results when $v_i$ are estimated

Assume  $E(\hat{v}_i) = v_i$  with  $\hat{v}_i \perp u_i$  and  $\hat{v}_i \perp e_i$ . Wang and Fuller (2003, Theorem 1) show that

$$\begin{aligned} \text{MSE}(Y_i - \hat{Y}_i) &\approx \sigma_u^2(1 - h_i) + (1 - h_i)^2 \mathbf{x}_i' V(\hat{\beta}) \mathbf{x}_i \\ &\quad + (\sigma_u^2 + v_i)^{-3} \{ \sigma_u^4 V(\hat{v}_i) + v_i^2 V_A(\hat{\sigma}_u^2) \}. \end{aligned}$$

They develop two estimators of the MSE.

Rivest and Vandal (2003) provide an essentially similar MSE estimator.

Note: The assumptions  $\hat{v}_i \perp u_i$  and  $\hat{v}_i \perp e_i$  may not be satisfied in practice.

Interesting features of these results:

- Wang and Fuller's result is asymptotic in both

$m = \#$  of small areas, and

$d =$  degrees of freedom of  $\hat{v}_i$ .

- Rivest and Vandal's result assumes  $\hat{v}_i$  are approximately normal.

Simulation results by Wang and Fuller and by Rivest and Vandal suggest MSE estimators work pretty well in many of the cases considered (including  $m$  and  $d$  not so large).

- Exception: Wang and Fuller find results are poor when  $v_i/\sigma_u^2$  is very large.

## Small area variance modeling

Start with a generalized variance function (GVF) – some examples:

normal:  $v_i = \gamma/n_i$

lognormal:  $v_i = \gamma Y_i^2/n_i$       relvariance  $\approx \text{Var}(\log(y_i))$   
is constant over areas  $i$

binomial:  $v_i = \gamma Y_i(1 - Y_i)/n_i$       ( $Y_i =$  population proportion  
for area  $i$ ;  $\gamma =$  design effect)

Here  $n_i =$  some measure of sample size, and  $\gamma =$  GVF parameter to be estimated.

Note: GVF could depend on other covariates related to the sample design.

## Small area variance modeling

Note different perspectives on use of GVF:

- classical survey sampling – fit GVF as an approximation to direct variance estimates when providing the latter is difficult (e.g., for a large number of estimates)
- small area variance modeling – fit GVF as a model to improve on imprecise direct variance estimates.

Some references to GVF fitting/modeling: Wolter (1985); Valliant (1987); recent JSM proceedings papers by staff of U.S. Bureau of Labor Statistics; talk this afternoon by Sam Hawala.

## Two issues with GVF modeling

Issue 1:  $v_i$  may depend on the unknown true value  $Y_i$

**Frequentist solution:** Substitute something (what?) for unknown  $Y_i$

$x_i' \hat{\beta}$  or EB estimate of  $Y_i$

What not to substitute for unknown  $Y_i$ ?

$y_i$  (direct survey estimates)

**Bayesian solution:** Let  $v_i$  depend on  $Y_i$  via MCMC

Liu, Lahiri, and Kalton (2007); You (2008)

$$\text{GVF: } v_i = \frac{Y_i(1-Y_i)}{n_i} \times deff_i$$

## Two issues with GVF modeling

Issue 2: What if  $n_j$  is large and  $v_j \approx \hat{v}_j$  (and so both are small) for some  $j$ ?

**Solution:** Include random area effects in the variance model.

Otto and Bell (1995); Arora and Lahiri (1997);

Gershunskaya and Lahiri (2005); You and Chapman (2006);

Talk this afternoon by Jerry Maples.

Note: This issue does not arise if all area sample sizes are small, which can occur (e.g., in time series modeling of repeated survey estimates).

## Small area variance modeling with random effects

**Working model:**

$$\hat{v}_i | \omega_i \sim \text{ind.} \quad \omega_i \tilde{v}_i \frac{\chi_{d_i}^2}{d_i}$$

$$\omega_i^{-1} \sim \text{i.i.d. Gamma}(\delta + 1, \delta^{-1})$$

where  $\tilde{v}_i \equiv \tilde{v}_i(n_i, Y_i, z_i, \gamma)$  is a GVF, and  $v_i = \omega_i \tilde{v}_i$  is the true sampling variance. Otto and Bell (1995) developed a multivariate (Wishart) version.

We fit this model to the estimated sampling variances  $\hat{v}_i$ .

Question: What is  $d_i$ ?



## Small area variance modeling with random effects

### Implications of the working model:

1.  $E(\omega_i) = 1$

2. As  $\delta \rightarrow \infty$ ,  $\omega_i \rightarrow 1$  (no area random effects)

As  $\delta \rightarrow 0$ ,  $\omega_i$  become fixed, unrelated area effects.

3.  $\hat{v}_i \sim \frac{\delta}{\delta+1} v_i F(d_i, 2(\delta + 1))$

4.  $\omega_i^{-1} | \hat{v}_i \sim \text{Gamma}\left(\delta + 1 + \frac{d_i}{2}, \left(\delta + \frac{d_i \hat{v}_i}{2 \tilde{v}_i}\right)^{-1}\right)$

5.  $E(v_i | \hat{v}_i) = \left(\delta + \frac{d_i}{2}\right)^{-1} \left(\delta \tilde{v}_i + \frac{d_i}{2} \hat{v}_i\right)$ .

## 4. Empirical example

SAIPE model for state 5-17 poverty rates (CPS data)

$$\begin{aligned}y_{it} &= Y_{it} + e_{it} \\ &= (\mathbf{x}'_{it}\beta_t + u_{it}) + e_{it}\end{aligned}$$

- $y_{it}$  = CPS direct survey estimate of population 5-17 poverty rate ( $Y_{it}$ ) for state  $i = 1, \dots, 51$  in year  $t = 1995, \dots, 1998$
- $u_{it}$  = state  $i$ , year  $t$  random effect  $\sim$  ind.  $N(0, \sigma_{u,t}^2)$ , and independent of  $e_{it}$
- $e_{it}$  = survey errors  $\sim$  ind.  $N(0, v_{it})$

- $\mathbf{x}_{it}$  = vector of regression variables for state  $i$ , year  $t$ :
  - pseudo state poverty rate from tax return information and also tax “nonfiler rate”
  - food stamp participation rate
  - “census residuals” (from regressing previous census estimate on other elements of  $\mathbf{x}_{it}$  for the census year)
- $\beta_t$  = vector of regression parameters for year  $t$ .

## SAIPE state sampling error model

**Data:**  $C_i = 4 \times 4$  direct estimated sampling covariance matrix for state  $i$  for 1995, ..., 1998

**Model:** Assumes  $E(C_i) = V_i$  (true sampling covariance matrix) with

$$C_i | \omega_i, \{Y_{it}\} \sim \omega_i \mathbf{W}_i / d$$

$\mathbf{W}_i \sim \text{Wishart}(d, \widetilde{\mathbf{V}}_i(\eta))$  independent over  $i$

$$\omega_i^{-1} \sim i.i.d. \text{ Gamma}(\delta + 1, \delta^{-1})$$

where the true sampling error covariance matrix for state  $i$  is

$$V_i = \omega_i \widetilde{\mathbf{V}}_i(\eta) = \omega_i \mathbf{D}_i(\gamma) \mathbf{R}(\phi) \mathbf{D}_i(\gamma).$$

$\mathbf{D}_i(\gamma) = 4 \times 4$  diagonal matrix with entries given by square roots of

$$\tilde{v}_{it} = \text{GVF}_{it} \equiv \gamma Y_{it}(1 - Y_{it})/n_{it}$$

$\mathbf{R}(\phi) = 4 \times 4$  correlation matrix depending on parameters  $\phi = (\phi_1, \phi_2)$   
(AR(2) model)

$\omega_i =$  state  $i$  random effects on sampling variances

$$E(\omega_i) = 1 \text{ for all values of } \delta$$

$$\delta \rightarrow \infty \Rightarrow \omega_i = 1 \text{ for all } i$$

$$\delta \rightarrow 0 \Rightarrow \omega_i = \text{fixed state effects}$$

### Sampling error model parameter MLEs

parameter	$d$	$\gamma$	$\phi_1$	$\phi_2$	$\delta$
MLE	16.7	5.9	.32	-.02	40
std. error	.98	.20	.02	.02	—

We shall examine results using four alternative estimates of the estimated sampling error variances  $\hat{v}_{it}$ :

1.  $\hat{v}_{it} =$  direct survey variance estimates,
2.  $\hat{v}_{it} = \text{GVF}_{it}(\hat{\eta})$ , i.e., the fitted GVF with no state effects ( $\omega_i = 1$ ), which results as  $\delta \rightarrow \infty$ .
3.  $\hat{v}_{it} = \hat{\omega}_i \text{GVF}_{it}(\hat{\eta})$  where  $\hat{\omega}_i = E(\omega_i | \mathbf{C}_i, \hat{\eta})$  are the predictions of  $\omega_i$  from the sampling error model at its MLEs (including  $\hat{\delta} = 40$ )
4.  $\hat{v}_{it} = \tilde{\omega}_i \text{GVF}_{it}(\hat{\eta})$  where the  $\tilde{\omega}_i$  are fixed state effects obtained as  $\delta \rightarrow 0$

Fig. 2. Sampling error variances of age 5–17 state poverty ratios, 1995–1998

Direct estimates (points) and fitted GVF's with and without state effects

(state effects: green = fixed, red = random, blue = none)

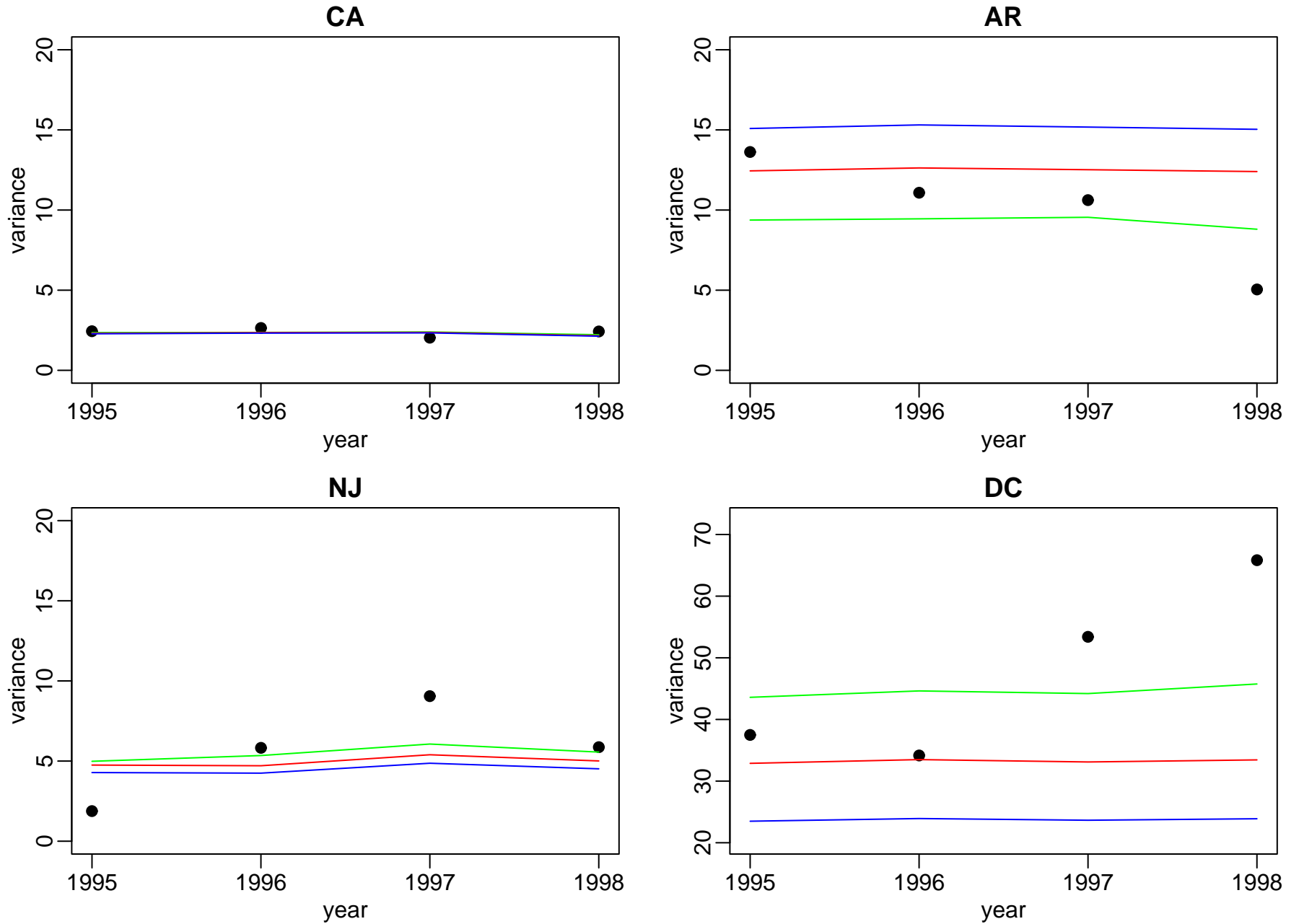




Fig. 3. Alternative predictions of age 5–17 state poverty ratios, 1995–1998

Direct estimates (points) and Bayesian predictions using alternative sampling variances (dotted ~ use of direct variances, solid ~ use of GVs with state effects: green = fixed, red = random, blue = none)

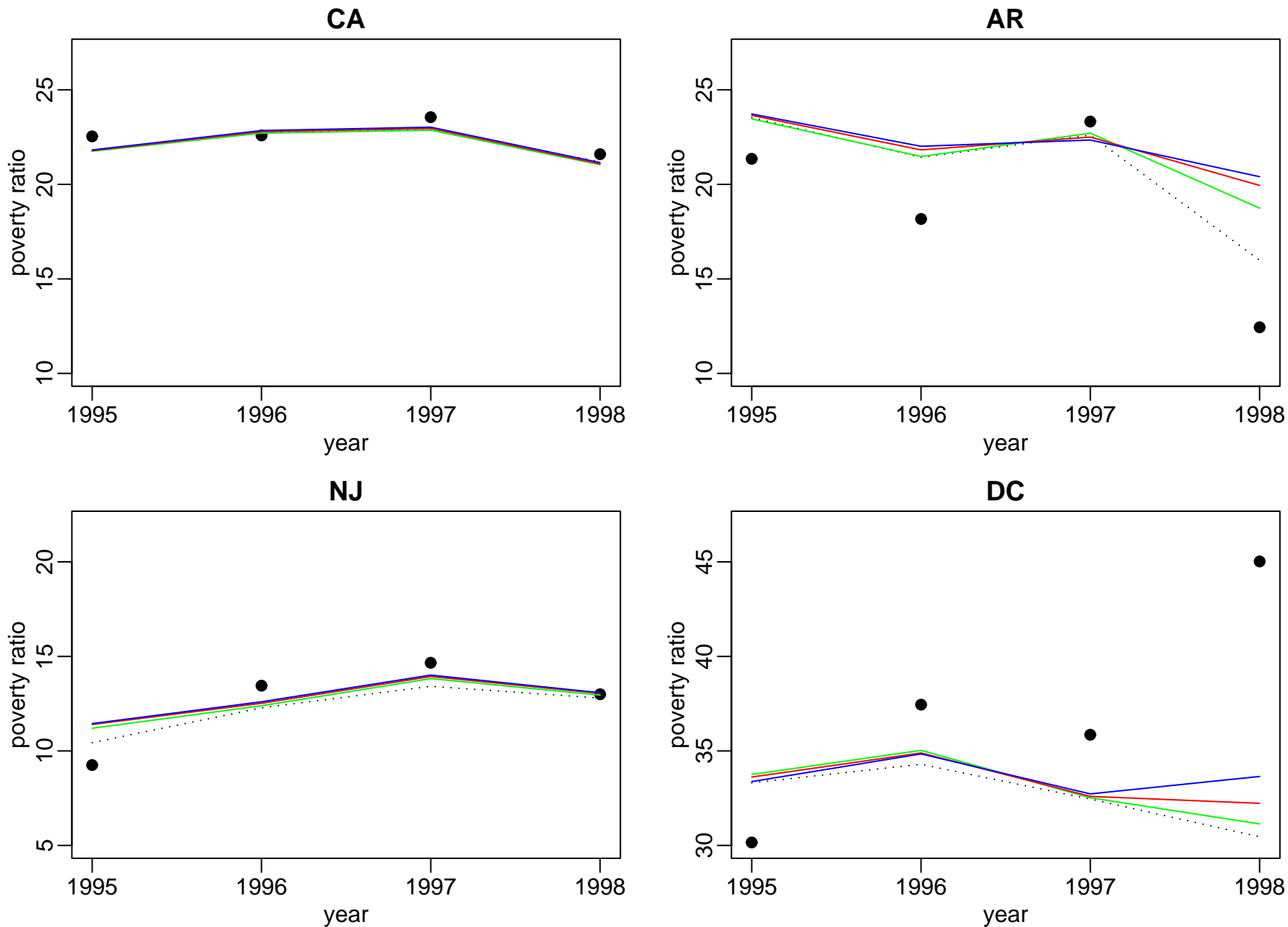
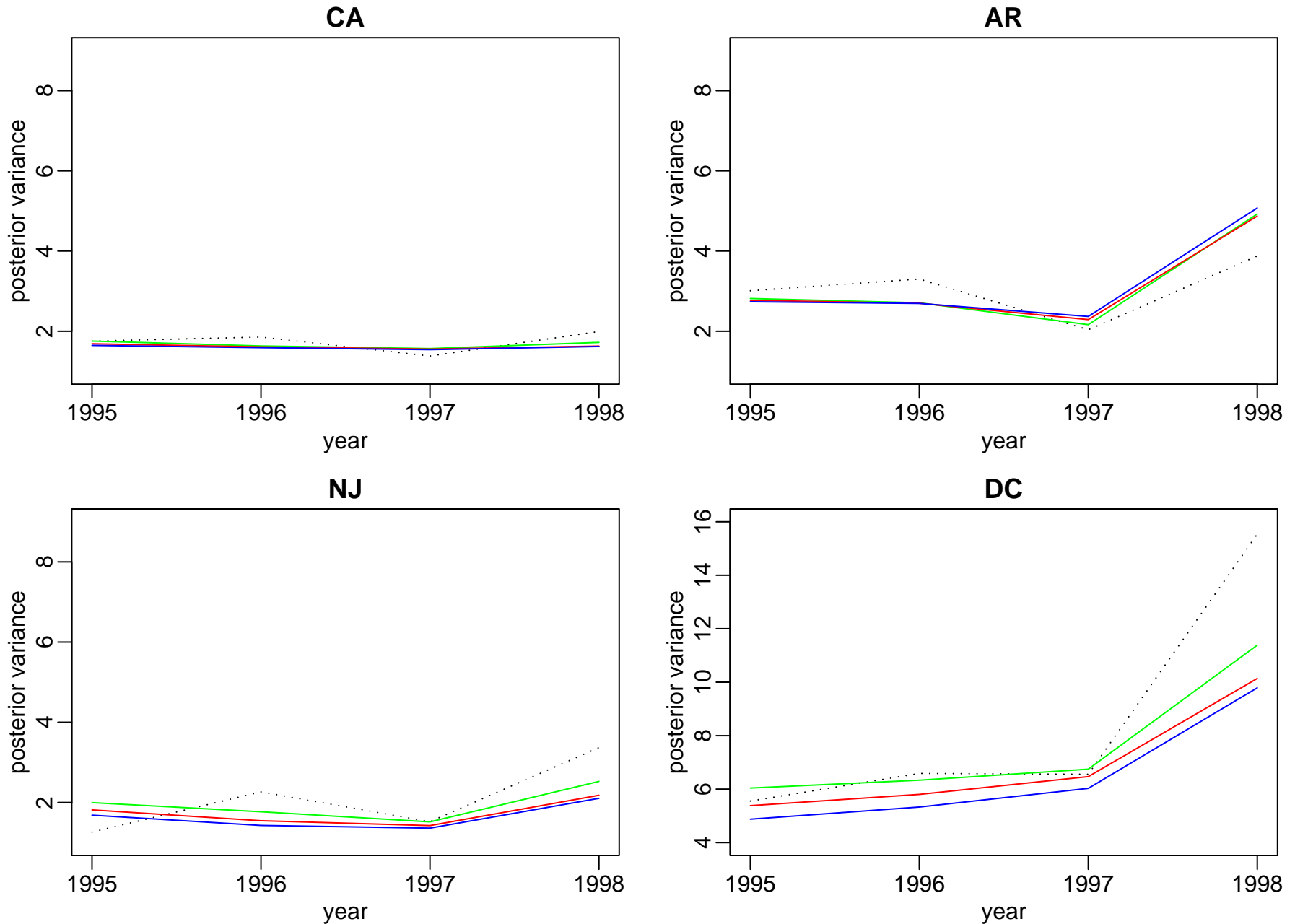


Fig. 4. Alternative posterior variances of age 5–17 state poverty ratios, 1995–1998

Posterior variances using alternative sampling variances

(dotted ~ use of direct variances, solid ~ use of GVs with state effects: green = fixed, red = random, blue = none)



## General conclusions:

1. Most serious potential problems from use of direct sampling variance estimates in small area models comes from severe underestimation of the sampling variance when it is large.
2. The worst problems with use of direct sampling variances can possibly be addressed by modeling the variances.
3. More research on this topic is needed.