

Using American Community Survey Data
to Estimate Local Area Unemployment Characteristics

F. Jay Breidt
Colorado State University

IMS-ASA'S SRMS JOINT MINI MEETING
ON CURRENT TRENDS IN SAMPLE SURVEYS
AND OFFICIAL STATISTICS
Calcutta, January 2004

Research Support for Surveys Over Time

- US Environmental Protection Agency
 - support for surveys of aquatic resources
 - this talk not reviewed or endorsed by EPA
- US Bureau of Labor Statistics
 - Local Area Unemployment Statistics program
 - this talk not reviewed or endorsed by BLS

Local Area Unemployment Statistics

- US Bureau of Labor Statistics program providing:
 - labor force, employment, unemployment, and unemployment rate
 - both annual averages and monthly estimates
- Small domain estimation:
 - some 6,700 geographic areas
 - census regions and divisions, states, counties, metropolitan statistical areas, labor market areas, etc.
- Uses: allocation of > **\$43 billion** in federal funds

LAUS Estimates

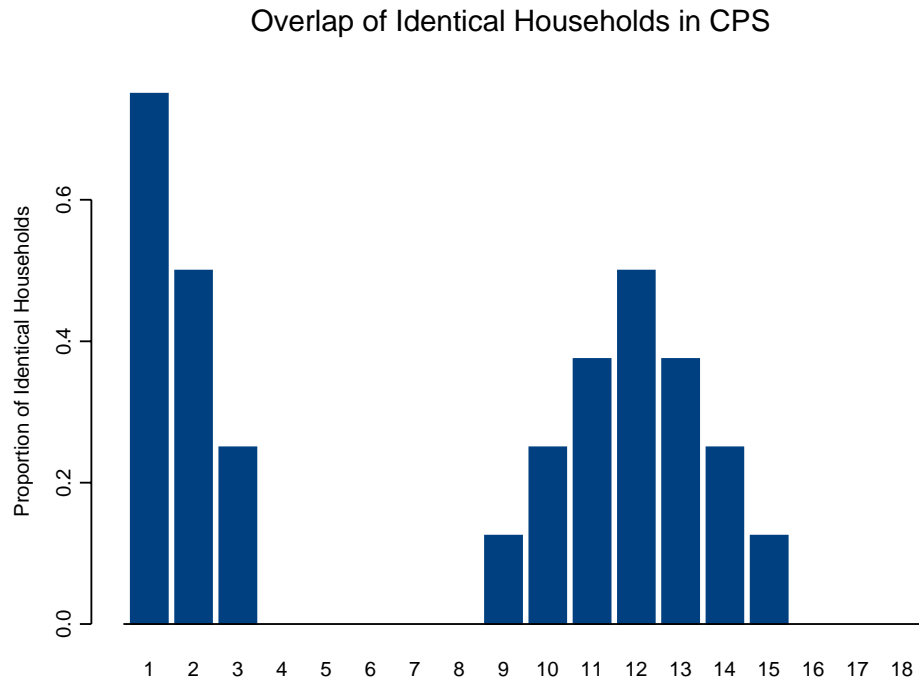
- Sub-state level LAUS estimates:
 - indirect estimates, largely synthetic
 - combines Current Employment Statistics data, demographic info
- “State” level LAUS estimates:
 - all states, DC, Los Angeles-Long Beach, New York
 - basic data from **Current Population Survey**
 - indirect estimates from time series models
- Focus on state level and a bit smaller

Current Population Survey

- **Purpose:** primary source of information on labor force characteristics of the US population
 - employment, unemployment, earnings, etc.
 - demographic characteristics: age, sex, race, etc.
- **Design:**
 - monthly sample of 50K households
 - 4–8–4 rotation: 4 months in, 8 months out, 4 in
 - efficient estimates of month-to-month and year-to-year change

4-8-4 Rotation Scheme of CPS

- Induces month-to-month overlap of identical households



Indirect Estimates from Time Series Models

- Uses CPS data plus covariates
 - Current Employment Statistics, unemployment insurance claims
- Employs **signal-plus-noise** models (Scott and Smith, 1974; Bell and Hillmer, 1990)

$$\begin{aligned}\text{direct estimate} &= (\text{signal}) + (\text{sampling error}) \\ &= (\text{regression}) + (\text{trend}) + (\text{seasonal}) \\ &\quad + (\text{sampling error})\end{aligned}$$

- Borrows strength across time
 - not possible to slice up 50K 6700 ways!

American Community Survey

- US Decennial Census has two parts:
 - **Short Form:** counts the population
 - **Long Form:** demographic, housing, social, and economic information from 1-in-6 sample of households
- ACS will replace Long Form for 2010 Census
 - 3M households per year / 250K monthly
 - drawn anew each month, avoiding repeats for 5 years
 - aimed at timely small domain estimates
 - intended for use in federal funding allocations

Improvements in LAUS Using ACS?

- Can ACS be usefully combined with CPS for small domain estimates?
- ACS contains labor force information
 - 5 times CPS sample size: 250K vs. 50K
 - much smaller sampling error
- Nonsampling error concerns
 - different instrument, definitions
 - different modes of data collection
 - different nonresponse followup

Implementation of ACS

- Timeline:
 - **Demonstration**: 1996–1998 (4, 8, 10 counties)
 - **Comparison**: 1999–ongoing until full implementation (31 counties)
 - **Full implementation** nationwide, pending Congressional funding
- Not fully implemented: little real data in what follows
 - strict confidentiality restrictions: Title 13, US Code
 - \$250K fine, 5 years in prison, or both

Outline

- Start with empirical model for CPS
 - well-studied state-space formulation (Tiller, Pfeffermann, etc.)
 - basic structural model for **signal = trend + seasonal**
 - correlated sampling error induced by rotating sampling design
- Write down hypothetical model for ACS
 - reflect sampling and nonsampling error
 - reference week inconsistencies, nonresponse
 - nonresponse followup protocols: mode and delay effects
- Combine CPS and ACS using Kalman recursions
 - compare MSE for signal and trend, with and without ACS data

Basic Structural Model for CPS

- Basic structural model: $\theta_t = T_t + \sum_j S_{jt}$

$$T_t = T_{t-1} + R_{t-1} + V_t^T$$

$$R_t = R_{t-1} + V_t^R$$

$$S_{jt} = \cos \omega_j S_{j,t-1} + \sin \omega_j S_{j,t-1}^* + V_{jt}^S$$

$$S_{jt}^* = -\sin \omega_j S_{j,t-1} + \cos \omega_j S_{j,t-1}^* + V_{jt}^{*S}$$

- $\{V_t^T\} \sim \text{WN}(0, \sigma_T^2)$, $\{V_t^R\} \sim \text{WN}(0, \sigma_R^2)$,
 $\{V_t^S\} \sim \text{WN}(0, \sigma_S^2)$, $\{V_t^{*S}\} \sim \text{WN}(0, \sigma_S^2)$

Correlation of CPS Sampling Error

- Correlations induced by month-to-month overlap
 - theoretically corresponds to MA(15)
- But replacement households are obtained from nearby addresses
 - induces other dependence, modeled as AR(2), uncorrelated with MA(15)
 - MA(15)+AR(2)=ARMA(2,17)
- Approximate via AR(15):

$$e_t = \phi_1 e_{t-1} + \cdots + \phi_{15} e_{t-15} + V_t^e$$

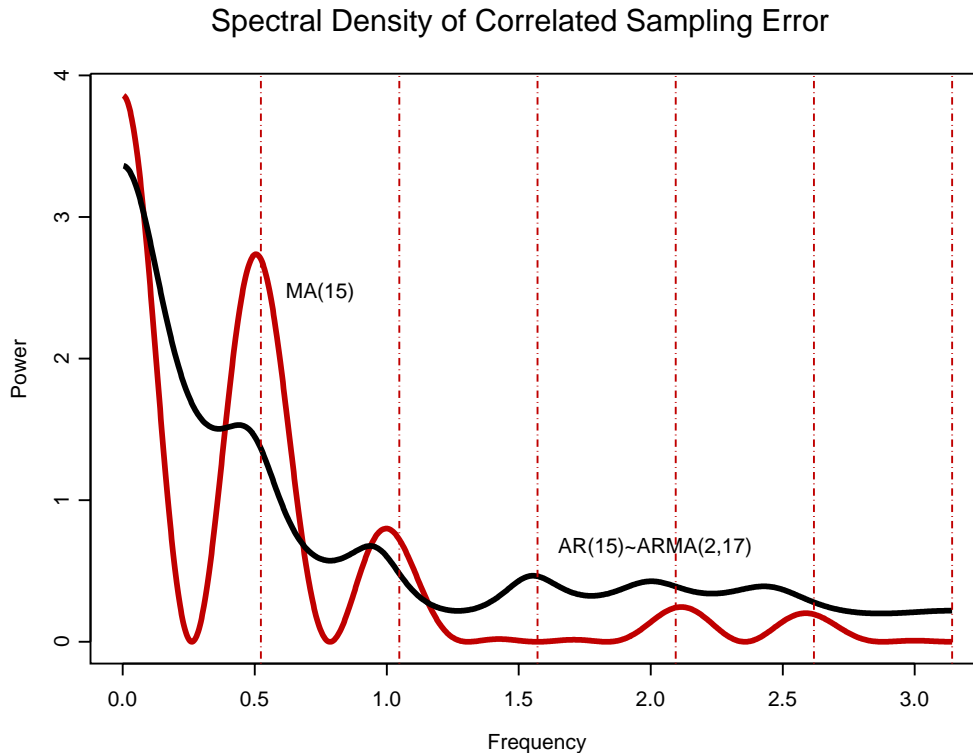
Illustrative Data Series

- Use Black Male Youth (aged 16–19) employed-to-population ratio
- Convenient sample size: fairly small, with large sampling errors
- “Known” hyperparameters:

$$\sigma_T^2, \sigma_R^2, \sigma_S^2, \phi_1, \dots, \phi_{15}, \sigma_e^2$$

Spectrum of CPS Sampling Error

- Power at low frequencies and seasonal frequencies is confounded with trend, seasonal



Turn Now to ACS Model

- CPS model complete
 - ignore covariates and changing design variance
- Consider total error in ACS:
 - CPS population* \neq *ACS pop*: moving reference window
 - ACS population* \neq *sample*: sampling error
 - sample* \neq *respondents*: nonresponse
 - responses* \neq *true values*: mode effects, delay effects

ACS Sampling and Follow-up

- **Month 1:** all sampled households receive questionnaire
 - some households respond by mail
- **Month 2:**
 - non-responding households get assigned to CATI pool
 - mail response may still come in late
- **Month 3:**
 - CAPI follow-up: 1/3 of nonrespondents
 - CAPI or non-CAPI households may still mail back

Markov Model for Nonresponse

- Follow-up across months

	Start	CATI	CAPI followup	No CAPI followup	MR	TR	PR	NR
Start		$1 - \mu_1$			μ_1			
CATI			$(1/3)(1 - \mu_2 - \tau)$	$(2/3)(1 - \mu_2 - \tau)$	μ_2	τ		
CAPI					μ_3		δ	$1 - \mu_3 - \delta$
No CAPI					μ_3			$1 - \mu_3$
MR					1			
TR						1		
PR							1	
NR								1

$$= \begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

Absorbing Probabilities

- From elementary properties of Markov chains:

$$\begin{bmatrix} \text{Pr}(\text{mail}) \\ \text{Pr}(\text{CATI}) \\ \text{Pr}(\text{CAPI}) \\ \text{Pr}(\text{non-resp.}) \end{bmatrix}' = [1 \ 0 \ 0 \ 0](\mathbf{I} - \mathbf{Q})^{-1}\mathbf{R} := \begin{bmatrix} 0.626 \\ 0.082 \\ 0.091 \\ 0.201 \end{bmatrix}'$$

– empirical values from 1996 ACS pilot counties

– constrain to solve nonlinear system: assume $\mu_j = \mu^j$

$$\mu = 0.4734, \tau = 0.1557, \delta = 0.8385$$

CPS-ACS Difference in Reference Week

- **Current Population Survey:**
 - reference week: calendar week (Sunday–Saturday) containing 12th of month
 - sampling week: week containing 19th of month
- **American Community Survey:** *LAST WEEK, did this person do ANY work for either pay or profit?*
 - reference week: week prior to data collection
 - sampling week: any week in three-month period

Moving Reference Window

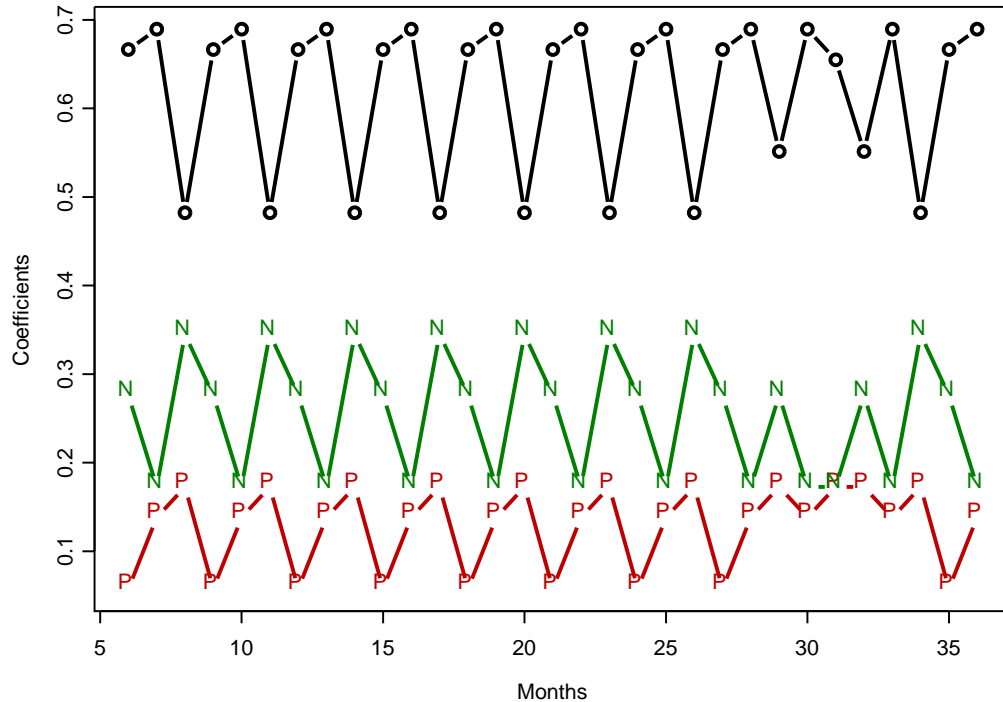
Obs. Month	Reference Week	True Signal
$t - 1$	contains 12th	θ_{t-1}
$t - 1$		linear interpolation
\vdots		\vdots
t		linear interpolation
\vdots		\vdots
t	contains 12th	θ_t
\vdots		\vdots
t		linear interpolation
\vdots		\vdots
$t + 1$		linear interpolation
$t + 1$	contains 12th	θ_{t+1}
t	monthly avg.	$\psi_t^P \theta_{t-1} + (1 - \psi_t^P - \psi_t^N) \theta_t + \psi_t^N \theta_{t+1}$

Moving Reference Window Assumptions

1. CPS unbiasedly estimates the truth
 2. Between reference weeks, **signal = linear interpolant**
 3. Response is uniform within month, so monthly average signal can be computed
- Deviations from assumptions?
 - additional, correlated non-sampling error
 - possible trend and seasonalities

Moving Reference Window Weights

- Bottom to top: ψ_t^P , ψ_t^N , $1 - \psi_t^P - \psi_t^N$



ANOVA Model for ACS Measurements

- Observation i in **month** t via **mode** g after h -month **delay** (sampling month was $t - h$):

$$\begin{aligned} A_{tghi} = & \psi_t^P \theta_{t-1} + (1 - \psi_t^P - \psi_t^N) \theta_t + \psi_t^N \theta_{t+1} \\ & + (\alpha_g + a_{tg}) + (\beta_h + b_{t-h}) \\ & + d_{tgh} + e_{tghi} \end{aligned}$$

- **month** effect is moving reference window
- **mode** effect and **month*mode** interaction
- **delay** effect and **month*delay** panel effect
- three-way interaction and sampling error

Why No mode*delay Interaction?

- **mode** effects $\alpha_1, \alpha_2, \alpha_3$ are unconstrained: 3 df
- **delay** effects $\beta_0, \beta_1, \beta_2$ have one identifiability constraint: 2 df
- **mode*delay** interaction (Mail0, Mail1, Mail2, Telephone1, Personal2): 5 df

Why Include Three-Way Interaction?

- Three-way interaction of `month*mode*delay`
- d_{t11} : **all** these people responded in month t by mail after one month delay
- d_{t21} : at least **some** of these people **never** would have responded by mail, but were “caught” by CATI

ACS Summary Statistics

- Compute means over each month*mode*delay

		sample	
mode	delay	size	mean
Mail	0	N_{t10}	\bar{A}_{t10}
Mail	1	N_{t11}	\bar{A}_{t11}
Mail	2	N_{t12}	\bar{A}_{t12}
CATI	1	N_{t21}	\bar{A}_{t21}
CAPI	2	N_{t32}	\bar{A}_{t32}

- Build state-space model for complete observation vector

$$\mathbf{Y}_t = (C_t, \mathbf{A}'_t)'$$

- Let Kalman filter combine these optimally

State-Space Model

- Combined state vector:

$$\mathbf{X}_t = (T_t, R_t, \mathbf{S}'_{t+1}, \mathbf{S}^*{}'_{t+1}, \theta_t, \theta_{t-1}, \mathbf{e}_t, \boldsymbol{\alpha}, \mathbf{a}_{t2}, \mathbf{a}_{t3}, \boldsymbol{\beta}, \mathbf{b}_t)'$$

- Combined state-space model:

$$\begin{bmatrix} C_t \\ \bar{A}_{t10} \\ \bar{A}_{t11} \\ \bar{A}_{t12} \\ \bar{A}_{t21} \\ \bar{A}_{t32} \end{bmatrix} = \mathbf{Y}_t = \mathbf{G}_t \mathbf{X}_t + \mathbf{W}_t, \quad \{\mathbf{W}_t\} \sim (\mathbf{0}, \mathbf{R}_t(\mathbf{N}_t))$$

$$\mathbf{X}_{t+1} = \mathbf{F}_t \mathbf{X}_t + \mathbf{V}_t, \quad \{\mathbf{V}_t\} \sim (\mathbf{0}, \mathbf{Q}_t)$$

Kalman Recursions

- Innovation:

$$\begin{aligned} \mathbf{I}_t &= \mathbf{Y}_t - \mathbf{G}_t \hat{\mathbf{X}}_t \\ \Delta_t &= \mathbf{G}_t \Omega_t \mathbf{G}'_t + \mathbf{R}_t (\mathbf{N}_t) \end{aligned}$$

- Filter: (under normality, $E[\mathbf{X}_t | \mathbf{Y}_1, \dots, \mathbf{Y}_t]$)

$$\begin{aligned} \mathbf{X}_{t|t} &= \hat{\mathbf{X}}_t + \Omega_t \mathbf{G}'_t \Delta_t^{-1} \mathbf{I}_t \\ \Omega_{t|t} &= \Omega_t - \Omega_t \mathbf{G}'_t \Delta_t^{-1} \mathbf{G}_t \Omega_t \end{aligned}$$

- Predict:

$$\begin{aligned} \hat{\mathbf{X}}_{t+1} &= \mathbf{F}_t \mathbf{X}_{t|t} \\ \Omega_{t+1} &= \mathbf{F}_t \Omega_{t|t} \mathbf{F}'_t + \mathbf{Q}_t \end{aligned}$$

Initializing the Covariance Recursions

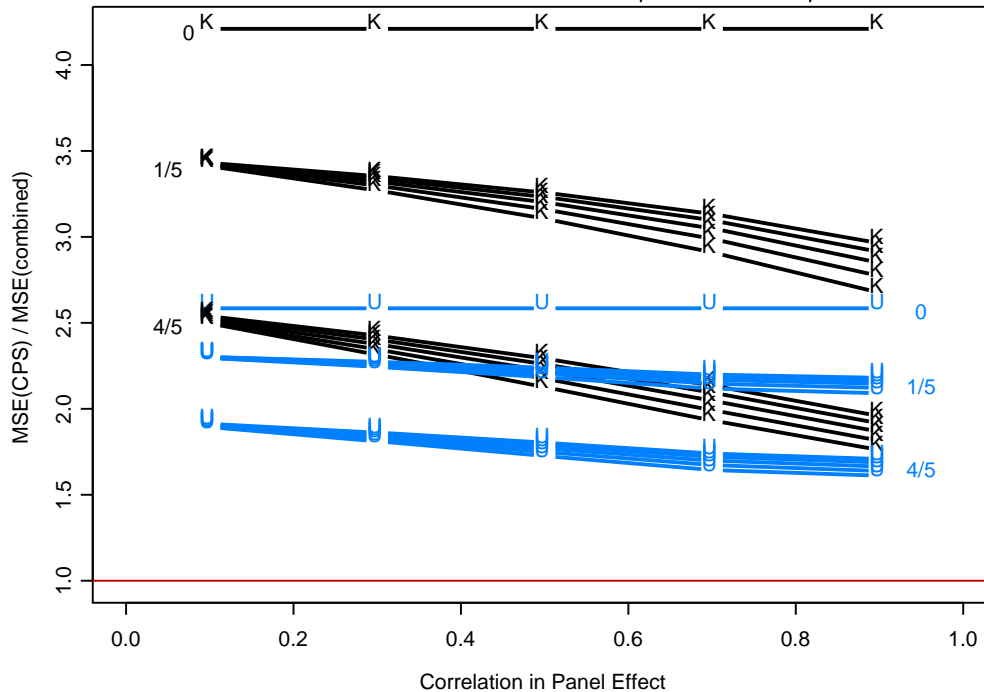
- Need $\mathbf{\Omega}_1$ to start recursions
- Nonstationary random effects: large prior variance
- Stationary random effects: prior variance is marginal variance
- Fixed effects: known (0 prior variance) or unknown (large prior variance)

Some Numerical Experiments

- Assume $(\text{ACS sampling var}) = (\text{CPS sampling var})/5$
 - ACS sample size is 5 times greater
- Set $(\text{Mail0 nonsampling var}) = k(\text{CPS sampling var})$
- Then *if response was complete*:
 - $k = 0$: no nonsampling variance
 - $k = 1/5$: ACS nonsampling var equals sampling var
 - $k = 4/5$: ACS total var equals CPS sampling var

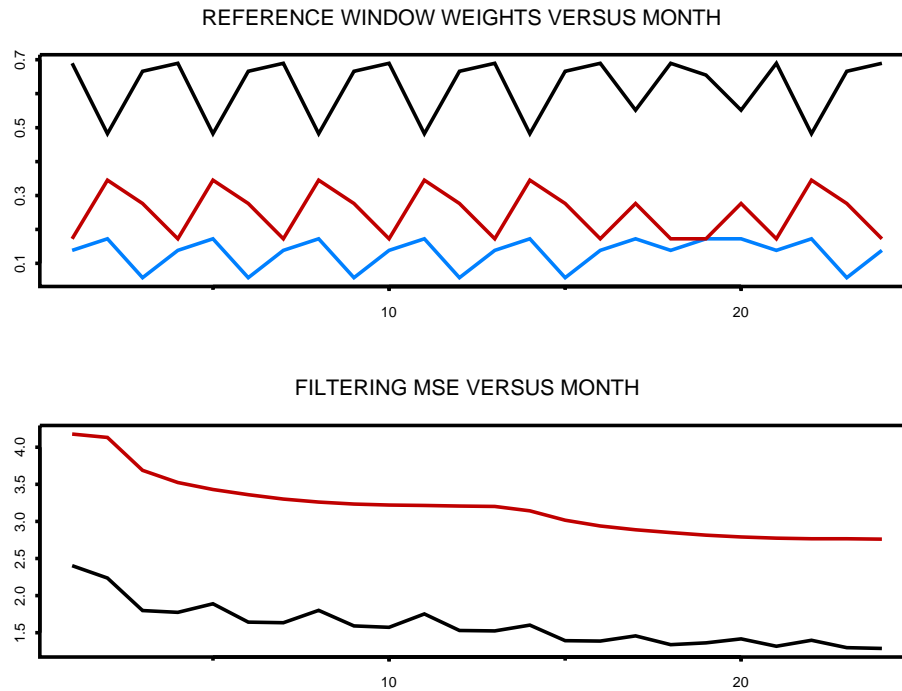
Big Picture: Effects on MSE Ratios

- **K**nown or **U**nknown fixed effects; $\sigma_{dgh}^2 = 0$
- ACS nonsampling error = 0, 1/5, or 4/5 of CPS sampling error



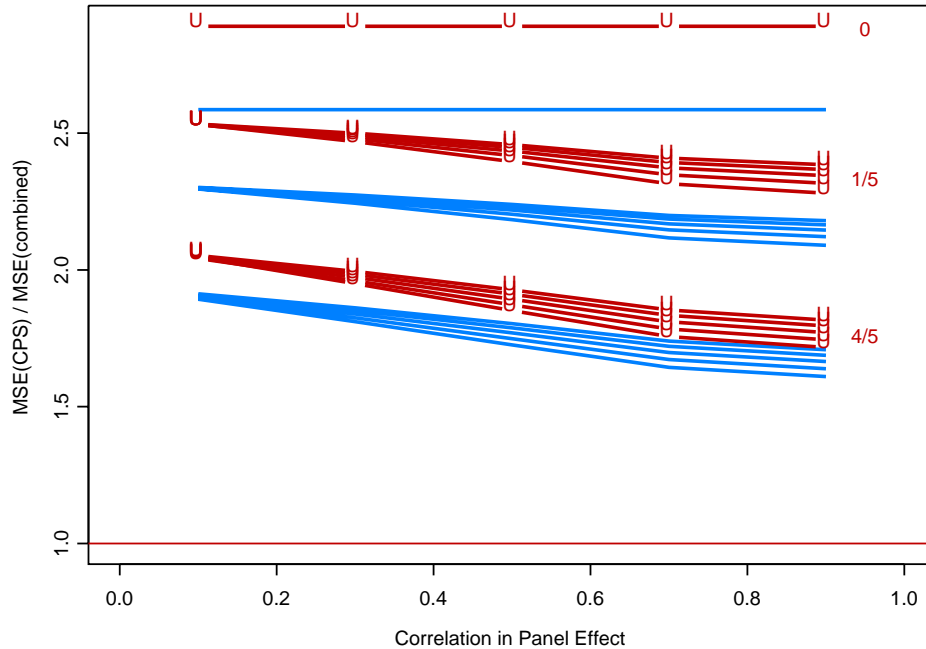
How Is MSE Affected by Moving Reference Window?

- One particular parameterization:



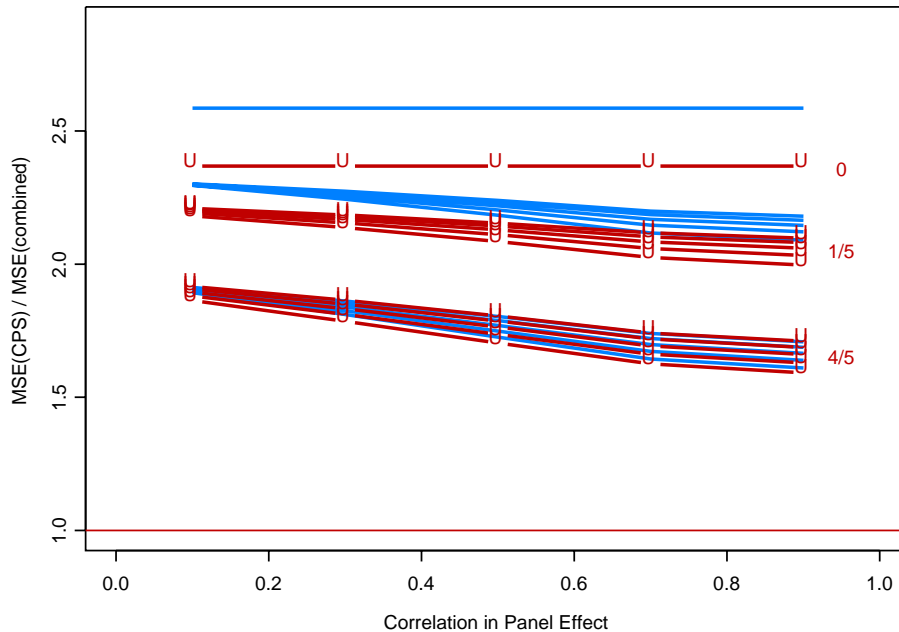
Effect of Moving Reference Window

- Focus on unknown fixed effects (blue from big picture)
- Red lines: set $\psi_t^P = \psi_t^N = 0$ (no moving reference window)



Effect of Lower Mail Response

- Cut mail response in half to 31.3% (still $> 23\%$ in Starr Co TX)



Other Influential Factors

- Less correlated nonsampling error is easier
 - decrease in panel effects or interviewer effects implies increase in efficiency
- More trend, less seasonal is easier
 - σ_T^2 increases implies efficiency increases
 - σ_R^2 increases implies efficiency increases
 - σ_S^2 increases implies efficiency *decreases*

Indirect Estimation Across Space and Time

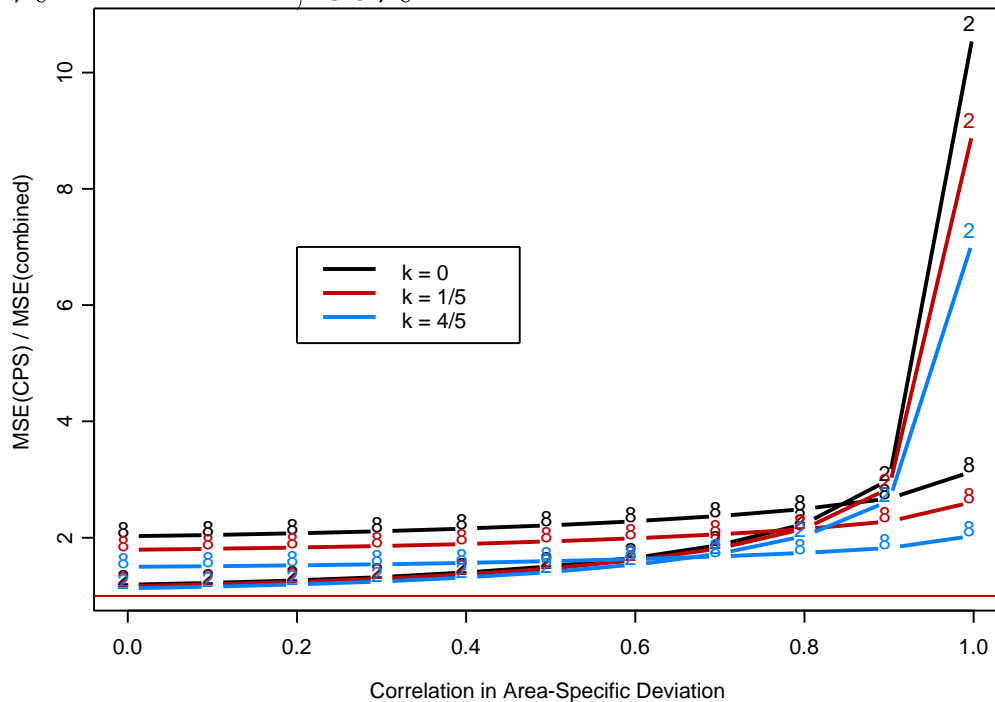
- For area s of size M_s , $\theta_{st} = \theta_t + D_{st}$ where

$$\{D_{st}\} \sim \left(0, \frac{\sigma_D^2}{M_s} \left(1 - \frac{M_s}{M} \right) \right)$$
$$\frac{\sum_s M_s \theta_{st}}{\sum_s M_s} = \theta_t + \frac{\sum_s M_s D_{st}}{\sum_s M_s} = \theta_t$$

- Consider a “disaggregation” problem:
 - obtain ACS summary stats for each small area: $\{\mathbf{A}_{st}\}$
 - obtain CPS estimate only for large area: C_t
- Compare CPS-only synthetic to ACS/CPS composite

Efficiency for Case of Two Small Areas

- AR model for area-specific deviation: $D_{st} = \rho D_{s,t-1} + V_{st}^D$
- 20% small area, 80% small area



Summary and Future Research

- KF MSE computations for hypothetical models
 - account for at least some important factors
 - fairly wide range of nonsampling error specifications
 - in all cases, potentially large gains from ACS
- Directions for further study
 - moving reference window characteristics
 - non-homogeneous transition probabilities (weekly?)
 - delay/panel and mode effects
 - hyperparameter estimation; accounting for uncertainty (Rao 2003, §5.4, 8.3, 10.8)